Datenvalidierung mittels Simulation stochastischer Kennzahlensysteme

Veit Köppen

koeppen@wiwiss.fu-berlin.de

Institut für Wirtschaftsinformatik, Freie Universität Berlin

Zusammenfassung

Gegenstand meiner Forschung ist die Datenvalidierung (im Sinne der Datenkonsistenz) im Fall vollspezifizierter stochastischer, nichtlinearer Gleichungssysteme mittels Simulation. Neben den stochastischen Variablen bestehen die untersuchten Kennzahlensysteme aus Bilanz- oder Definitionsgleichungen. Dabei sind die Operatoren diejenigen der Grundrechenarten. Aufgrund von Multiplikation und Division sind die Systeme nichtlinear.

Ein Teilbereich meiner bisherigen Forschung ist der stochastische Abgleich von Daten mit einem stochastischen Gleichungssystem. Im Folgenden wird die untersuchte Modellklasse dargestellt. Daran schließt sich die Vorstellung des entwickelten Algorithmus an.

Modellklasse (Eine-Gleichung / Drei-Variablen-Fall)

Modell $\mathcal{M}$:

$$x_1 = \xi_1 + v_1$$
$$x_2 = \xi_2 + v_2$$
$$z = H \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + w$$
$$\begin{bmatrix} v_1 \\ v_2 \\ w \end{bmatrix} \sim F\left(0, \Sigma_{vw} = \begin{bmatrix} \Sigma_{vv} & 0 \\ 0 & \Sigma_{ww} \end{bmatrix}\right)$$

Abbildung 1a: Strukturelles Zustandsraummodell $\mathcal{M}$

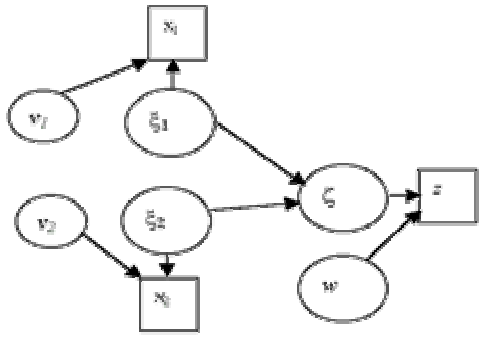

Abbildung 1b: Modellgraph von $\mathcal{M}$

Die Vektoren $x \in R^p$ und $z \in R^q$ sind die Messwerte. Diese sind fehlerbehaftet. Die wahren Werte ξ und $\zeta = H(\xi)$ können nicht beobachtet werden. Es wird angenommen, dass alle Gleichungen separierbar sind, d.h. eindeutig nach jeder auftretenden Variablen aufgelöst werden können. Die abhängige Variable wird dabei als Variable der linken Seite (LHS) bezeichnet und die unabhängigen Variablen als Werte der rechten Seite (RHS).

SamPro Algorithmus

SamPro (Sampling and Projection) ist ein Algorithmus zur Ziehung von beliebig verteilten Zufallszahlen und zur Schätzung von nichtbeobachtbaren Zustandsvariablen des Modells $\mathcal{M}$. Dabei kann der Algorithmus Inkonsistenzen in den Daten im Vergleich zum Modell ermitteln. Eine Fundamentalannahme ist, dass das Modell $\mathcal{M}$ korrekt ist. Das Ziehen von Zufallszahlen aus beliebigen Dichten wird dabei durch den Metropolis-Hastings-Algorithmus ermöglicht. Multivariate Verteilungen können dabei berücksichtigt werden.

Der SamPro Algorithmus ist wie folgt strukturiert:

1. Sampling der RHS; mittels Metropolis-Hastings-Algorithmus.
2. Simulation der Verteilungen aller LHS mittels Modell $\mathcal{M}$.
3. Projektion im Produktraum aufgespannt durch alle $k \in N$ Schätzer für jede einzelne Variable des Kennziffernsystems.

Liegen für eine Variable k simulierte Schätzungen vor, so wird in einem ersten Schritt für jede Schätzung das α- und das (1-α)-Quantil $\left(\underline{q} \text{ und } \overline{q}\right)$ berechnet. Wenn der Schnitt dieser k Quantilintervalle leer ist, so ist das Gleichungssystem streng inkonsistent. Andernfalls wird das Intervall $I_q = \left[\underline{q}_{\max}, \overline{q}_{\min}\right]$ mit $\max\left\{\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_k\right\} = \underline{q}_{\max}$ und $\min\left\{\overline{q}_1, \overline{q}_2, \ldots, \overline{q}_k\right\} = \overline{q}_{\min}$ berechnet und für die Projektion genutzt. Abbildung 2 verdeutlicht das Intervall I_q, das als Menge konsistenter Modellwerte aufgefasst wird. Anschließend wird aus dem mehrdimensionalen Raum $range(z) \times range(x)$ auf den Unterraum projiziert, der z.B. durch $x_1 = x_2$ gegeben ist (siehe Abb. 3).

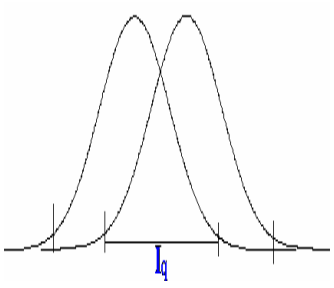

Abbildung 2: Intervall I_q

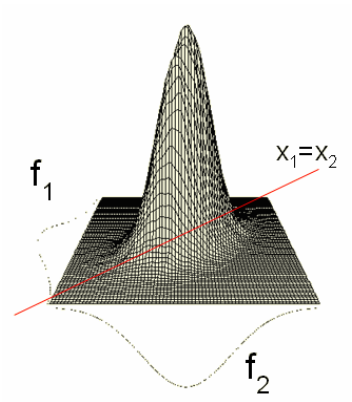

Abbildung 3: Projektion

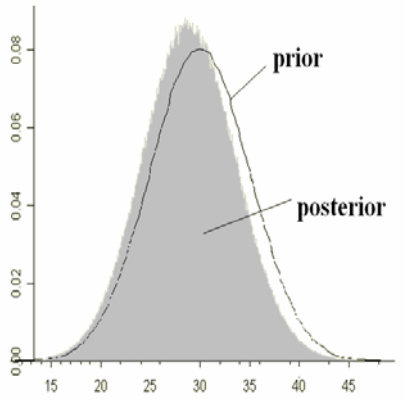

Abbildung 4: Vergleich prior und posterior

Simulationsergebnisse

Durch den SamPro Algorithmus lassen sich zum Modell $\mathcal{M}$ existierende Inkonsistenzen zwischen Variablen aufdecken. Darüber hinaus lassen sich Schätzer für fehlerbehaftete Variablen berechnen bzw. fehlende Werte schätzen. Dabei sind Varianzreduktion und Verschiebung der Mittelwerte der Posterior- im Vergleich zu den der Prior-Verteilungen möglich. Die Verschiebungen sind proportional zur Varianz des jeweiligen Messfehlers (siehe Abb. 4).

Literatur

Hastings, W.K.: Monte Carlo sampling methods using Markov Chains and their applications. Biometrica, 57 (1), 97-109, 1970.

Köppen, V.; Hausmann, A., Lenz, H.-J.: Simulation – A Support for Controllers Decision Process. In Proceedings of ICTM 2005: Challenges and Prospects, Melaka, 1055-1070, 2005.

Köppen, V.; Lenz, H.-J.: Simulation of Non-linear Stochastic Equation Systems, Proceeding of the Fifth Workshop on Simulation, St. Petersburg, 373- 378, 2005.