# UQ – Abgabe2 Gruppe D (Bernhard, Ivanov, Schröppel)

## Problem

In der ersten Abgabe wurde eine erste primitive Untersuchung des Einflusses von Parameterveränderungen auf bestimmte Zielgrößen (Quantities of Interest, QoI) am Beispiel des SEIR-Modells durchgeführt. Die vom Robert-Koch-Institut für die zur Modellierung der Covid-19 Pandemie vorgeschlagenen Parameter wurden jeweils zufällig um bis zu p-Prozent verändert. Anschließend wurde das Modell für diese zufälligen Parametervektoren durchgerechnet und anhand der Ergebnisse konnte beispielsweise die Verteilung der Anzahl der kumulativ infizierten Personen nach 60 Wochen untersucht werden. Dies erlaubte erste Rückschlüsse darauf, mit welchem Spektrum an Ergebnissen zu rechnen ist, falls eine gewisse Unsicherheit in der Genauigkeit der Parameter besteht. Es gibt jedoch keinen Rückschluss darauf, welche Parameter in welchem Maße für das Ergebnis verantwortlich sind und damit gegebenenfalls besonders genau zu bestimmen sind, um ein möglichst gutes Modellergebnis zu erhalten.

Durch gezielte Veränderung einzelner Inputparameter konnte dennoch herausgefunden werden, dass das SEIR-Modell deutlich sensibler auf Veränderungen der Transmissionsrate Beta und Erholungsrate Gamma als auf Veränderungen in der Übergangsrate Alpha oder der Anzahl anfänglich Infizierter I0 reagiert. Hierbei handelt es sich jedoch nur um eine qualitative Aussage. Außerdem wurden diese Untersuchungen immer nur an einer gewissen Stelle im Parameterraum durchgeführt. Es wäre möglich, dass beispielsweise eine p-prozentige Veränderung eines Parameters an einer anderen Stelle (d.h. Veränderung zu einer anderen Basis), ein stark verändertes Ergebnis für diese primitive Sensitivitätsanalyse liefern würde. Durch einfaches Ausprobieren lässt sich vor allem nicht quantifizieren, welchen Einfluss die einzelnen Parameter tatsächlich auf das Modell haben. Ziel dieser Aufgabe ist es also eine mathematisch bestimmbare Größe für den Zusammenhang zwischen einzelnen Inputparametern und der aus dem Modellergebnis abgeleiteten QoI zu bestimmen.

## Lösungsansatz

Eine mögliche Methode zur Quantifizierung der Abhängigkeiten ist die Bestimmung der linearen Abhängigkeiten zwischen den Variablen. Diese Größe wird als (Pearson-)Korrelationskoeffizient bezeichnet. Die klassische Berechnung des Korrelationskoeffizienten hat jedoch ein paar entscheidende Nachteile. Würde man einfach die Korrelation zwischen einem Inputparameter und den Outputwerten (QoI) bestimmen, so würde man nicht berücksichtigen, dass bei einem Modell wie dem SEIR-Modell das Modellergebnis und damit die QoI von mehr als einem Parameter abhängig ist. Um die Korrelation besser zu bestimmen, wird der Partialkorrelationskoeffizient (PCC) berechnet. Dabei werden sowohl die Werte des zu untersuchenden Inputparameters als auch die Outputwerte zuerst um die lineare Abhängigkeit von den restlichen Inputparametern bereinigt. Sowohl die Werte des zu untersuchenden Inputparameters als auch die Outputwerte werden dazu mittels linearer Regression möglichst gut durch die anderen Parameter angenähert. Das Residuum zwischen der Regressionsgerade (entspricht dem Anteil, der durch

die restlichen Parameter bestimmt wird) und dem tatsächlichen Wert entspricht dem Anteil des Inputparameters, der nicht bereits durch die anderen Inputparameter bestimmt ist (d.h. der unabhängige Teil) beziehungsweise entspricht es dem Teil des Outputwerts, der tatsächlich durch den untersuchten Parameter bestimmt wird. Anschließend wird der Korrelationskoeffizient zwischen den Residuen des Inputparameters und der Outputwerte berechnet. Durch die Berechnung der Partialkorrelationskoeffizienten kann die lineare Abhängigkeit zwischen einem Inputparameter und dem Outputwert unter Ausschluss linearer Einflüsse der übrigen Parameter bestimmt werden.

Bei nichtlinearen Zusammenhängen ist es jedoch gegebenenfalls sinnvoller, nicht die linearen Abhängigkeiten zwischen Parameter und QoI zu bestimmen, sondern die monotone Abhängigkeit. Diese Abhängigkeit wird durch Partialrangkorrelationskoeffizienten (PRCC) beschrieben. Hierbei wird nicht die lineare Abhängigkeit zwischen den Residuen von Parameter und Outputwert berechnet, sondern die Korrelation zwischen den Residuen der Ränge der Parameterwerte beziehungsweise Outputwerte. Es wird also der Partialkorrelationskoeffizient zwischen den Rängen der Werte des zu untersuchenden Inputparameters und der Ränge Outputwerte berechnet. Der Rang eines Wertes gibt dabei an, an welchem Index der Wert in einer sortierten Liste stehen würde.
Durch diese Methode kann der Zusammenhang der Monotonie zwischen zwei Variablen bestimmt werden.

Für die Berechnung der Partialkorrelationskoeffizienten wurde folgende Formel aus dem Skript verwendet

$$\hat{\rho}_{X_1 \to Y} = \hat{\rho}_{X_1^*, Y^*} = \frac{\sum_{i=1}^{N} x_{1,i}^* y_i^*}{\sqrt{\sum_{i=1}^{N} x_{1,i}^{*\,2}} \sqrt{\sum_{i=1}^{N} y_i^{*2}}}.$$

$x_1^*$ sowie $y^*$ sind dabei die Residuen zwischen $x_1$ Werten und der Regressionsgerade bzw. der y-Werte und der Regressionsgeraden. Für die Berechnung von Partialrangkorrelationskoeffizienten, muss der Partialkorrelationskoeffizient zwischen den Rängen von x1 und y bestimmt werden.

Für eine Anwendung auf das SEIR-Modell müssen zunächst wieder Samples des Parametervektors generiert werden. Diese Samples werden jetzt jedoch mit der verbesserten Methode des Latin-Hypercube-Samplings erzeugt. Die Achsen des Hyperwürfels werden dazu in N (N=Anzahl Samples) gleichgroße Subintervalle aufgeteilt (Gilt nur für gleichverteilte Zufallsvariablen; Es wird angenommen, dass die Parameter innerhalb des gegebenen Intervalls gleichverteilt sind). Für mehrdimensionale Samples muss eine zufällige Permutation der Anordnung der Subintervalle für jede Achse ausgewählt werden. Ziel des LHS ist es, dass die Werte für die Parametersamples so gewählt werden, dass in jedem der Subintervalle exakt ein Sample liegt. Ein Vorteil dieser Samplingmethode liegt darin, dass die Samples innerhalb des Hyperwürfels gleichmäßiger (Gilt nur für gleichverteilte ZV) verteilt liegen. Dadurch soll eine bessere Annäherung der Samples an die tatsächliche Verteilung des Parameters erzielt werden.

Nachdem die Samples für den Parametervektor des SEIR-Modells erzeugt wurden, wurde das Modell für alle Parametervektoren berechnet. Es wurde dabei auch wieder die Anzahl kumulativer Fälle (C) mitberechnet.

Aus diesen Ergebnissen konnten nun die QoIs (Quantities of Interest) bestimmt werden. Es wurden dabei die kumulative Zahl der Infizierten nach 60 Wochen sowie die Woche, in der die meisten Infektionen vorliegen, ermittelt. Anschließend konnten die Partialrangkorrelationskoeffizienten zwischen den einzelnen Parametern und den beiden QoIs bestimmt und dargestellt werden.

Die Implementierung erfolgte wie in Abgabe 1 in der Programmiersprache Python.

Die Funktionen zur Berechnung des SEIR-Modells wurden aus der vorangegangenen Aufgabe übernommen. Um Parametersamplings nach den Vorgaben des Latin Hypercube Samplings zu generieren, wird für jede der Achsen des Hyperwürfels zunächst die Größe der Subintervalle, in denen die jeweiligen Samples liegen müssen, bestimmt. Durch die Funktion linspace() aus der numpy Bibliothek wurden anschließend die unteren Intervallgrenzen dieser einzelnen Subintervalle bestimmt und in einem Array gespeichert. Dieses Array wurde dann mithilfe der numpy Funktion shuffle() durch eine zufällige Permutation des Arrays überschrieben. Anschließend konnte in jedem dieser Subintervalle [Intervalluntergrenze, Intervalluntergrenze + Subintervallbreite] ein gleichverteilter Zufallswert bestimmt werden. Auf diese Weise wird in jedem der Subintervalle genau ein Parametersample erzeugt.

Anschließend wurde die Funktion partial_corrcoef(X, y) implementiert. Die Matrix X stellt die erzeugten Parametervektoren dar. Jeder Zeilenvektor der Matrix repräsentiert dabei einen der gesampelten Parametervektoren. Der Funktionsparameter y ist ein numpy-array, welches den Wert der QoI zu den Parametervektoren der Matrix X, beinhaltet. Der Partialkorrelationskoeffizient wird für jeden der Modellparameter (jeder Spaltenvektor von X) einzeln berechnet. Zunächst werden die gesampelten Werte für den zu untersuchenden Modellparameter aus der Matrix kopiert und in dieser entfernt. Anschließend können mittels der Methode der kleinsten Quadrate, die in der Bibliothek numpy.linalg unter dem Funktionsnamen lstsq() definiert ist, die Parameter der Regressionsgeraden bestimmt werden. Es wird dabei die Regressionsgerade für die Werte des zu untersuchenden Parameters sowie für die Werte der QoI bestimmt. Der Abstand zwischen tatsächlichem Wert und Wert der Regressionsgeraden an dieser Stelle gibt den Anteil an, welcher tatsächlich durch Änderung des Parameterwerts zustande kommt und nicht bereits durch die anderen Parameter vorbestimmt ist. Um möglichst gute Ergebnisse für die Regressionsgerade zu erhalten, wird vor der Berechnung der Regressionskoeffizienten noch zu jedem Parametervektor ein konstanter Wert 1 hinzugefügt. Dadurch kann eine Konstante in der Regressionsgeraden bestimmt werden.

Für die Berechnung von Partialrangkorrelationskoeffizienten kann die Funktion partial_corrcoef(X, y) wiederverwendet werden. Anstatt der gesampelten Werte müssen die Matrix X und der Vektor y die Ränge der jeweiligen Werte enthalten. Die Ränge der Werte in einem Array können mithilfe der numpy Funktion searchsorted bestimmt werden. Searchsorted liefert den Index, an dem ein Wert in einer sortierten Liste eingefügt werden würde. Dieser Index entspricht dem Rang des Wertes.

Abschließend wurden die Partialrangkorrelationskoeffizienten zwischen den Parametersamples und den beiden QoIs bestimmt. Die Bestimmung der beiden QoIs aus den Modellergebnissen erfolgt analog zu Abgabe 1. Mithilfe der in einer Datei gespeicherten Parametersamples und den zugehörigen QoIs können dann die

Partialrangkorrelationskoeffizienten berechnet werden. Dargestellt werden die Ergebnisse in einem Barplot.

# Interpretation/Ergebnisse:

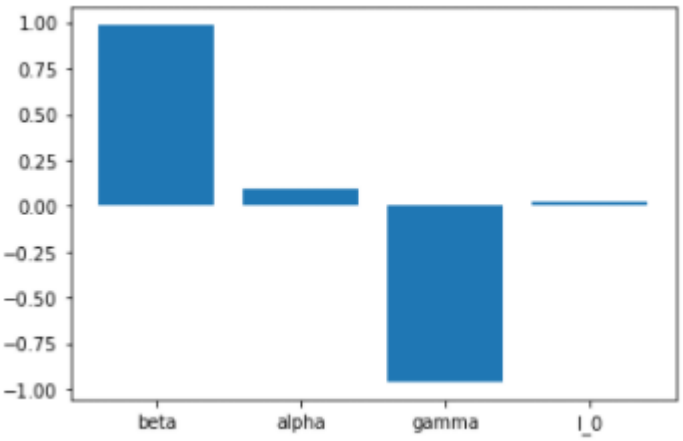


*Abbildung 1: PRCCs - Final cumulative cases*

Wie Abb.1 zeigt, hängt die finale Anzahl kumulativer Fälle im Wesentlichen von den Parametern Beta und Gamma ab. Beta hat dabei mit einem Partialrangkorrelationskoeffizienten (PRCC) von 0,98 den stärksten Einfluss. Monoton steigende Beta-Werte führen also fast immer zu einer monoton steigenden Anzahl an kumulativen Fällen nach 60 Wochen. Eine höhere Transmissionsrate führt also zu einer größeren Anzahl an Menschen, die sich im Verlauf der Pandemie mit der Krankheit infizieren. Beinahe gegensätzlich verhält es sich mit der Erholungsrate Gamma, welche einen PRCC von -0,96 aufweist. Steigt die Erholungsrate was bedeutet, dass die Menschen schneller wieder gesund werden, so führt dies zu einer geringeren Anzahl an insgesamt infizierten Personen im Laufe der Pandemie. Die Übergangsrate Alpha (PRCC 0,1) sowie die Zahl anfänglich infizierter I0 (PRCC 0,02) haben dahingegen kaum Einfluss auf die Zahl kumulativer Fälle.

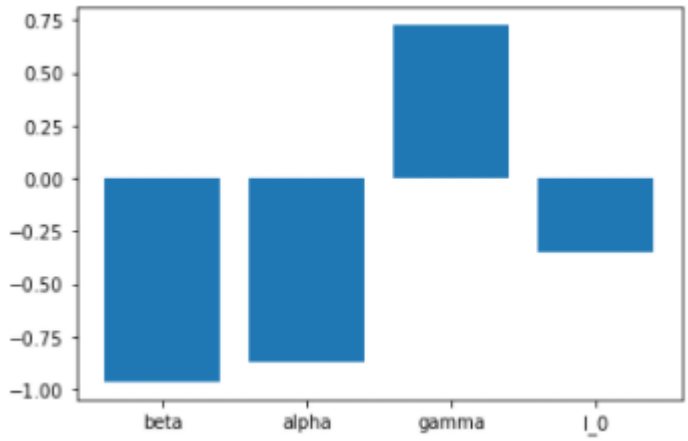


*Abbildung 2: PRCCs - Week of peak infections*

Auf die Woche, in welcher die meisten Infektionen stattfinden, hat ebenfalls die Transmissionsrate Beta den größten Einfluss (PRCC -0,97). Höhere Transmissionsraten führen also dazu, dass der Peak in der Zahl der Infizierten bereits früher auftritt. Auch ein Anstieg in der Übergangsrate Alpha (PRCC -0,88) führt zu einem früheren Peak. Mit einem PRCC von 0,72 verschiebt eine höhere Erholungsrate in vielen Fällen den Zeitpunkt der maximalen Infektionen nach hinten. Die unbedeutsamste Variable für diese Kennzahl ist wie bei der Analyse der kumulativen Fälle wieder die Zahl anfänglich Infizierter. Allerdings hat sie mit einem PRCC von -0,35 durchaus noch Einfluss auf die Kenngröße, nur nicht so stark wie die anderen Parameter.

Insgesamt fällt auf, dass jede QoI des Modells eigene Parametersensitivitäten besitzt. Es kann daher eigentlich nicht über die Sensitivität des SEIR-Modells gesprochen werden, sondern immer nur über die Sensitivitäten einzelner Kenngrößen, die das SEIR Modell liefert. Auffällig ist dabei auch, dass unterschiedliche Parameter für unterschiedliche Kenngrößen sehr unterschiedliche Einflüsse haben können. Die Übertragungsrate Alpha spielt bei der Zahl der kumulativ infizierten nach 60 Wochen nur eine sehr untergeordnete Rolle, wohingegen sie bei Betrachtung der Woche mit den meisten Infektionen maßgeblich ist. Auch die „Richtung“, in welche die Parameter die QoI letztlich beeinflussen, ist nicht identisch. Während Beta bei der ersten Untersuchung einen stark positiven Partialrangkorrelationskoeffizienten aufweist, ist dieser bei der zweiten Untersuchung stark negativ.

In der ersten primitiven Sensitivitätsanalyse für das gesamte Modell, wurden die beiden Parameter Beta und Gamma als maßgeblich herausgearbeitet. Dies bestätigt sich ebenfalls bei genauerer Untersuchung für zwei gewählte Kenngrößen. Eine mögliche Handlungsempfehlung die beiden Parameter Beta und Gamma möglichst gut zu ermitteln und Alpha und I0 nur ungefähr zu schätzen, hätte jedoch zur Folge gehabt, dass die Woche der Peak Infektionen nur sehr ungenau bestimmt worden wäre. Aus den neu ermittelten PRCCs können jetzt auch neue Handlungsempfehlungen abgeleitet werden. Wird eine möglichst präzise Vorhersage der Woche mit den meisten Infektionen gewünscht, so ist es unerlässlich neben Beta und Gamma auch die Übergangsrate Alpha möglichst gut zu bestimmen. Es empfiehlt sich außerdem grundsätzlich zuerst eine Sensitivitätsanalyse der gewünschten Kenngrößen durchzuführen, um Erkenntnisse zu gewinnen, welche Parameter für welche Kenngröße welchen Einfluss haben und deshalb eventuell besonders genau modelliert werden sollten.