UQ – Abgabe3 Gruppe D (Bernhard, Ivanov, Schröppel)

Problem.

Während in der letzten Abgabe bereits bestimmt werden konnte, wie stark die Korrelation zwischen Veränderung eines Inputparameters und Änderung der Werte einer Zielfunktion ist, beschäftigt sich diese Abgabe damit, zu Quantifizieren welchen Anteil der Gesamtvariant der einer Kenngröße (Quantity of Interest kurz QoI) von den einzelnen Inputparametern bedingt wird. Die gewählte Methode hierfür sind die Sobol-total-effect Indizes. Diese total-effect Sobol Indizes setzen dabei die zu erwartende Varianz die auftreten würde wenn nur der untersuchte Parameter variabel wäre ins Verhältnis zur Varianz der QoI. Ein Nachteil, der durch diese Methode entsteht, ist der Verlust der Information in welche Richtung der Parameter das Ergebnis beeinflusst. Während ein positiver beziehungsweise negativer Partialrangkorrelationskoeffizient bedeutet, dass ein größerer Wert für den Parameter ein größeres beziehungsweise kleineres Ergebnis bedeutet, fehlt diese Information bei den Sobol Indizes. Dies liegt an den Eigenschaften der Varianz welche nur positive Werte annehmen kann.

Sobol-Methode.

Die Sobol-Methode basiert auf der Zerlegung der Varianz der QoI in Summanden von Varianzen der Eingabeparameter in zunehmender Dimensionalität und bestimmt somit den Beitrag der einzelnen Eingabeparameter und ihrer Wechselwirkungen zur Gesamtvarianz der QoI. Dabei ist das Ziel der Sobol-Sensitivitätsanalyse zu bestimmen, welcher Anteil der Varianz in der QoI von den einzelnen Eingabeparametern abhängt. Dabei wird sowohl der direkte Einfluss des einzelnen Parameters als auch die Wechselwirkung mit anderen Parametern untersucht. Bei der Zerlegung der Ausgabevarianz in einer Sobol-Empfindlichkeitsanalyse wird dasselbe Prinzip wie bei der klassischen Varianzanalyse verwendet. Es lässt sich sagen, dass Sobolindizes erster Ordnung als Haupteffekt interpretiert werden, um den Teil des additiven Beitrags eines einzelnen Parameters zur Ausgangsvarianz zu messen. Sobolindizes zweiter Ordnung werden verwendet, um den Teil des Beitrags von Parameterinteraktionen zur Ausgangsvarianz zu messen. Die Empfindlichkeitsindizes totaler Ordnung berücksichtigen sowohl die Haupteffekte, die Effekte zweiter Ordnung als auch die Effekte höherer Ordnung, was die Auswertung über den gesamten Bereich des Parameterraums beinhaltet. Betrachten wir nun einmal die Formel:

$$S_{T_i} = \frac{\mathbf{V}\mathrm{ar}(Y) - \mathbf{V}\mathrm{ar}_{\mathbf{X}_{\sim i}}(\mathbf{E}_{X_i}[Y|\mathbf{X}_{\sim i}])}{\mathbf{V}\mathrm{ar}(Y)}$$

$$= 1 - \frac{\mathbf{V}\mathrm{ar}_{\mathbf{X}_{\sim i}}(\mathbf{E}_{X_i}[Y|\mathbf{X}_{\sim i}])}{\mathbf{V}\mathrm{ar}(Y)}.$$

So erkennen wir, dass der Ausdruck des Zählers nichts anderes darstellt als die Empfindlichkeitsindizes erster Ordnung aller Parameter ohne $\mathbf{X_i}$. Anders ausgedrückt, die erwartete Varianz die übrig bleiben würde, wenn wir alle Variablen außer $\mathbf{X_i}$ wüssten.
Dies Wissend ergibt sich somit für den ganzen Ausdruck die Bedeutung, dass dieser den Anteil in dem $\mathbf{X_i}$ eine Rolle spielt festhält und damit sozusagen alle Kombinationen und Korrelationen von Variablen mit in die Wertung aufnimmt.

Um noch ein wenig konkreter zu werden, was die Implementierung des Verfahrens angeht, widmen wir uns zunächst noch einmal der Formel für die Bestimmung der Indizes Totalen Effekts, welche wir oben schon einmal umgeschrieben wiederfinden konnten.

$$S_{T_i} := \frac{\mathbf{E}_{\mathbf{X}_{\sim i}}[\mathbf{V}\mathrm{ar}_{X_i}(Y|\mathbf{X}_{\sim i})]}{\mathbf{V}\mathrm{ar}(Y)}.$$

Dabei betrachten wir uns den Zähler hinsichtlich der Implementierung genauer.

$$\mathbf{E}_{\mathbf{X}_{\sim i}}[\mathbf{V}\mathrm{ar}_{X_i}(Y|\mathbf{X}_{\sim i})]$$

Zur Berechnung benötigen wir einen treuen Schätzer hinsichtlich der Varianz und des Erwartungswertes. Wie wir bereits Wissen haben wir bereits vor einigen Wochen in der Vorlesung einen treuen Schätzer für die Varianz von Y falls alle Parameter außer X_i bekannt sind hergeleitet, und zwar:

$$\frac{1}{2}(Z_1 - Z_2)^2$$

Z_1 und Z_2 entsprechen dabei zwei Funktionsauswertungen mit verschiedenen Werten für den Parameter X_i und identischen Parametern $X_{\sim i}$
Nun fehlt uns nur noch ein Schätzer für den Erwartungswert, dieser Varianz den wir wie folgt definieren:

$$\frac{1}{N}\sum_{j=1}^{N}\mathbf{V}\mathrm{ar}_{X_i}(\mathcal{G}(X_{1,j},\ldots,X_{i-1,j},X_i,X_{i+1,j},\ldots,X_{n,j})\,|$$
$$|\,\mathbf{X}_{\sim i} = (X_{1,j},\ldots,X_{i-1,j},X_{i+1,j},\ldots,X_{n,j})^\top),$$

Es handelt sich dabei um die Monte-Carlo Summe der einzelnen Varianzen. Damit ergibt sich der folgende Schätzer, welcher auch Jansen – Schätzer genannt wird.

$$\frac{1}{2N}\sum_{j=1}^{N}(\mathcal{G}(A)_j - \mathcal{G}(A_B^{(i)})_j)^2.$$

Dabei sind A und A_B Matrizen, die nach dem folgenden Schema implementiert werden:

Man sampelt zwei Matrizen A, B welche Uniform im Intervall [0,1] verteilt sind und erzeugt daraus eine Matrix A_B welche am Index 1 die erste Spalte aus B hat und den Rest aus A. Für Index 2 die zweite Spalte aus B hat und den Rest aus A. Das geschieht, wie Sie an diesem Beispiel sehen für alle Indexwerte.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \\ a_{4,1} & a_{4,2} & a_{4,3} \end{pmatrix} \qquad A_B^{(1)} = \begin{pmatrix} b_{1,1} & a_{1,2} & a_{1,3} \\ b_{2,1} & a_{2,2} & a_{2,3} \\ b_{3,1} & a_{3,2} & a_{3,3} \\ b_{4,1} & a_{4,2} & a_{4,3} \end{pmatrix}$$

$$A_B^{(2)} = \begin{pmatrix} a_{1,1} & b_{1,2} & a_{1,3} \\ a_{2,1} & b_{2,2} & a_{2,3} \\ a_{3,1} & b_{3,2} & a_{3,3} \\ a_{4,1} & b_{4,2} & a_{4,3} \end{pmatrix} \qquad A_B^{(3)} = \begin{pmatrix} a_{1,1} & a_{1,2} & b_{1,3} \\ a_{2,1} & a_{2,2} & b_{2,3} \\ a_{3,1} & a_{2,2} & b_{3,3} \\ a_{4,1} & a_{2,2} & b_{4,3} \end{pmatrix}$$

Dabei wird recht deutlich was wir bei dem Schätzer für den Erwartungswert schon gesehen haben. Wir haben eine Summe aus bedingten Varianzen, wo wir alle Werte außer X_i kennen, was man hier an den A_B Zeilenweise deutlich ablesen kann.

Der für die Durchführung einer zuverlässigen Sobol-Sensitivitätsanalyse erforderliche Stichprobenumfang hängt von zwei Hauptfaktoren ab: von der Komplexität des Modells und von der Anzahl der ausgewerteten Parameter. Obwohl es keinen allgemeinen Konsens über die optimale Anzahl der zu generierenden Parametersätze gibt, gilt die allgemeine Faustregel, dass für eine größere Anzahl an Modellparametern, auch mehr Parametersätze benötigt werden. Für ein komplexes Modell mit einer großen Anzahl von unsicheren Parametern (z. B. 20 Parameter) sollten beispielsweise mindestens 100.000 Modellevaluierungen durchgeführt werden.

Implementierung.

Die Implementierung des SEIR-Modells sowie die Funktionen für das Latin-Hypercube-Sampling (LHS) wurden aus der vorangegangenen Abgabe übernommen. Einzig die Methode zur Verifikation der Ergebnisse des LHS wurde für eine bessere Verständlichkeit neu implementiert. Für jeden Parameter werden die gesampelten Werte zunächst sortiert. Anschließend wird überprüft, ob jeder der Sampelwerte innerhalb des zugehörigen Subintervalls liegt. Die Untergrenze dieses Subintervalls wird bestimmt, in dem auf die Untergrenze des Samplingintervalls, i-mal die Breite der Subintervalle addiert wird. Der Wert i entspricht dem Index des untersuchten Samplingwerts. Die Breite der Subintervalle wird bestimmt, in dem man die Breite des gesamten Samplingintervalls durch die Anzahl der Samples teilt.

Für die Berechnung der Sobol-Indizes wurde zunächst die Funktion create_sobol_matrices(A,B) implementiert. Diese Funktion erstellt aus den beiden Matrizen A und B die entsprechenden Matrizen A_B. Zunächst werden die Matrizen wieder transponiert wodurch die Samplingwerte für je einen Parameter in je einer Zeile stehen. Anschließend wird für jeden Parameter einmal die Matrix A kopiert und die Zeile mit den Werten des entsprechenden Parameters durch die entsprechende Zeile der Matrix B ersetzt. Die erstellten Matrizen werden in einem Array abgespeichert welches abschließend zurückgegeben wird. Zur Verifikation dieser Funktion, wurde eine Funktion implementiert welche die beiden Matrizen A und B sowie die Liste der Sobol-Matrizen M als Parameter übergeben bekommt. Es wird überprüft, ob in jeder der Sobol-Matrizen die Werte der Parameter gleich den Werten der Matrix A beziehungsweise B sind. Für die i-te Matrix muss dabei gelten, dass die Werte in Spalte i gleich den Werten der Spalte i der Matrix B sind und gleichzeitig alle anderen Werten den Werten aus Matrix A entsprechen.
Bei der Implementierung der Funktion compute_total_indices(f_A, f_AB) welche die Sobol Indizes berechnen soll wurde sich an dem oben bereits genannten Jansens-Estimator orientiert.

$$s_{Ti} = \frac{1}{2N}\sum_{j=1}^{N}\left(f\_A_j - f\_AB_j^i\right)^2$$

f_A beziehungsweise f_AB ist dabei eine Liste beziehungsweise Matrix mit den Werten der entsprechenden QoIs berechnet mit den Parametern der Matrix A beziehungsweise Liste von Matrizen AB. $f\_AB_j^i$ beschreibt dabei den Wert der QoI für eine Auswertung in Zeile j der Sobol-Matrix bei der die Spalte i von Matrix A durch die entsprechende Spalte aus Matrix B ersetzt wurde.
Zur Verifikation wurde überprüft, ob die durch die oben beschriebene Funktion berechneten Sobol-total-effect-indices mit den, für zwei einfache Funktionen, von Hand berechneten Indizes übereinstimmen. Die erste Untersuchte Zielfunktion ist dabei $Y=X_1+X_2$. Es ist offensichtlich, dass beide Parameter denselben Einfluss auf den Wert der Zielfunktion haben. (Unter der Annahme, dass beide mit der identischen Gleichverteilung gesampelt wurden.) Beide Sobol-total-effect-indices sollten also einen identischen Wert annehmen. Da in dieser Funktion nur „First order effects“ auftreten und die Parameter sich nicht gegenseitig beeinflussen, entsprechen die beiden total-effect Indizes gleichzeitig den First-order-indizes. Daraus ergibt sich die Bedingung, dass die Summe der Indizes auch 1 ergeben muss. Daraus folgt, dass der Wert für die Sobolindizes gleich 0.5 sein sollte. Da die Sobol-indizes durch den Jansens-Estimator nur angenähert werden und abhängig von den gesampelten Werten auch vom Zielwert 0.5 abweichen können, werden Werte im Intervall [0.45, 0.55] als gültig akzeptiert. Die zweite untersuchte Funktion ist $Y=X_1*X_2$. Auch hierbei üben beide Parameter den selben Einfluss auf Y aus wodurch die beiden Sobol Indizes identisch sind. Außerdem tritt hier auch eine gegenseitige Beeinflussung der Parameter auf. Dies hat zur Folge, dass die Summe der beiden Indizes größer 1 ist.
Zum Abschluss der Implementierung wurden die neu erstellten Funktionen auf das SEIR Modell angewandt. Zunächst wurden zweimal 1000 Samples für die Matrizen A und B erzeugt. Aus diesen Matrizen wurden dann die 4 Sobol-Matrizen erstellt und für diese Matrizen sowie die Matrix A die entsprechenden Werte für die beiden QoIs berechnet. Abschließend wurden die Sobol-Indizes berechnet und die Ergebnisse in einem Barplot dargestellt.

Interpretation

Wie bereits beschrieben, sind die Werte für Sobol-Indizes stets positiv (oder 0). Dadurch geht die Information verloren, in welche Richtung der Parameter die QoI beeinflusst. Der berechnete Sobol-total-effect-Index für einen entsprechenden Parameter gibt an, welchen Teil der Gesamtvarianz durch den jeweiligen Parameter bedingt wird.

<u>Vergleich für Finale kumulative Fälle (QoI1):</u>

```
sobol_indices are: 0.84 for beta, 0.03 for alpha,
0.27 for gamma and 0.00 for I0 and sum up to 1.142
```

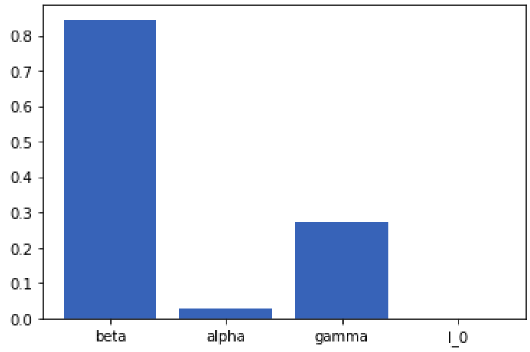
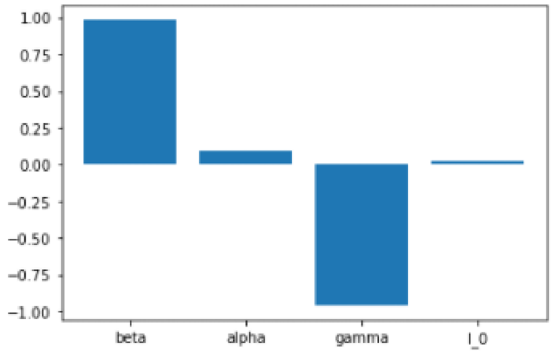


Abbildung 1: Sobol-Total-Effekt-Indizes für QoI1



PRCCs für QoI1

Es fällt auf, dass der Parameter Beta sowohl den höchsten PRCC als auch den höchsten Sobol Index (0.84) aufweist. Der PRCC liefert dabei die Information, dass höhere Werte für Beta fast immer zu einer größeren Zahl an kumulativ Infizierten nach 60 Wochen führt. Allerdings wird erst durch Betrachtung des zugehörigen Sobol-Index klar, dass diese Änderung in der QoI deutlich größer ist als dies bei anderen Parametern der Fall wäre. Es wird dadurch eindeutig, dass die Transmissionsrate Beta der entscheidende Parameter des Modells ist. Für Alpha und I0 sind sowohl PRCC als auch Sobol-Index fast 0. (Sobol Indizes gerundet: 0.03, 0.00) Diese Parameter sind für die Zahl der kumulativ Infizierten nach 60 Wochen also sehr unbedeutend. Bei der Betrachtung des Parameters Gamma fällt auf, dass dieser einen stark negativen PRCC und einen mittelgroßen Sobol-Index (0.27) aufweist. Durch den PRCC lässt sich bestimmen, dass höhere Werte für Gamma fast immer zu einer geringeren Anzahl Infizierter führen. Der Sobol-Index zeigt jedoch, dass diese Änderung relativ klein ist. Werden also Beta und Gamma mit denselben Faktor vergrößert, so ist damit zu rechnen, dass der Anstieg der QoI der durch das größere Beta bedingt, wird größer ist als die Reduktion der QoI durch ein größeres Gamma. Die Effekte von Beta überwiegen also die Effekte von Gamma. Dies macht insofern Sinn, da die Transmissionsrate Beta der Entscheidende Parameter einer Pandemie ist. Selbst wenn die Menschen sehr schnell wieder gesund werden würden (hohe recovery rate Gamma), würden Sie bei einer hohen Transmissionsrate auch in dieser Zeit weitere Leute anstecken. Die höhere Recoveryrate führt dann nur dazu, dass die gesamte Pandemie schneller vorüber ist. Es werden zwar etwas weniger Leute angesteckt allerdings breitet sich die Pandemie trotzdem durch einen großen Teil der Bevölkerung aus. Wie lange es dauert, bis eine infizierte Person auch ansteckend ist spielt insgesamt in der Betrachtung der Anzahl der kumulativ Infizierten kaum eine Rolle. Auch wie viele Menschen zu Beginn infiziert sind spielt eigentlich keine Rolle. Dieser Wert legt nur fest ab welchem Zeitpunkt der Pandemie diese modelliert wird. Jede Pandemie startet

irgendwann mit einem Infizierten, d.h. ein I_0 > 1 führt nur dazu, dass der vorhergegangene Verlauf nicht betrachtet wird.
Um zu quantifizieren, wie groß der gegenseitige Einfluss der Parameter aufeinander ist, wird die Definition der first-order und total-effect Indizes benötigt. Während first-order Indizes nur additive Effekte und damit keine Einflüsse auf andere Parameter bestimmen, messen die total-effect-indices auch die gegenseitigen Einflüsse. Da sich die First-order-Indizes stets zu 1 aufsummieren, kann der Teil der Summe der total-effect-indices der über 1 liegt als der Teil betrachtet werden, der durch die gegenseitigen Einflüsse der Variablen aufeinander zustande kommt. Für den Fall der kumulativ Infizierten beträgt die Summe der Total-Effekt-Indizes 1.142 . Das bedeutet, dass die gegenseitigen Wechselwirkungen der Parameter aufeinander mit 0.142 quantifiziert werden können. Durch diese Betrachtungsweise waren keine weiteren Modell- oder Sobol-Indexberechnungen nötig.

<u>Vergleich Woche mit maximaler Anzahl Infizierter (QoI2):</u>

```
sobol_indices are: 0.84 for beta, 0.13 for alpha,
 0.13 for gamma and 0.01 for I0 and sum up to 1.103
```

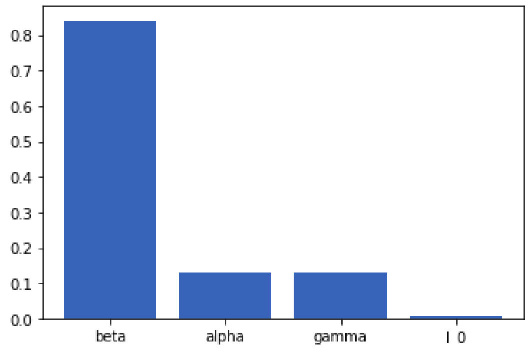
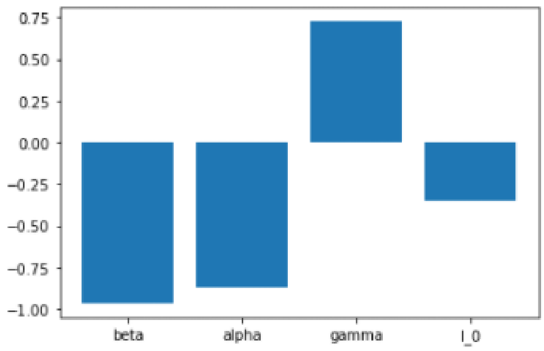


Abbildung 2: Sobol-Total-Effekt-Indizes für QoI2



PRCCs für QoI2

Auch bei der Betrachtung in welcher Woche die meisten Infektionen auftreten, fällt der Parameter Beta durch seinen hohen (negativen) PRCC als auch durch den hohen Sobol-Index (0.84) auf. Auch für diese QoI ist also die Transmissionsrate entscheidend. Der Unterschied zur Betrachtung der Kumulativ Infizierten liegt darin, dass ein höheres Beta hier zu einem geringeren Ergebnis, also einer früheren Woche, führt. Je ansteckender die Krankheit ist, desto schneller ist auch der Punkt erreicht, ab dem die Zahl der Neuinfektionen wieder abflacht, da bereits ein Teil der Population genesen und damit immun ist. Den Gegenteiligen Effekt hat eine schnellere Genesung wodurch es länger dauert, bis genug Personen infiziert waren und die Zahl der Infektionen rückläufig ist. Dies erkennt man auch am PRCC für Gamma. Höhere Gammas führen zu einer späteren Woche mit maximal Infizierten. Die Analyse des Sobol Index verrät jedoch, dass dieser Effekt weit weniger zum Tragen kommt als der vorher genannte Effekt durch eine Veränderung von Beta. Es ist außerdem zu erkennen, dass in diesem Fall sogar Alpha denselben Einfluss auf die Woche maximal infizierter ausübt wie Gamma. In diesem Fall führt ein höheres Alpha ebenfalls dazu, dass die Woche mit den maximal infizierten früher erreicht wird. Ein geringeres Alpha trägt ebenso dazu bei, dass sich die Krankheit schneller ausbreitet, wodurch auch wieder der Punkt, ab dem die Neuinfektionen abflachen schneller erreicht wird. Der Parameter I_0 ist auch für diese QoI ohne große Bedeutung. Eine Veränderung von I_0 hat praktisch keinen Einfluss auf die Woche der Peak-Infektionen.

Die Summe aller Sobol-Total-Effekt-Indizes ist 1.103. Die gegenseitigen Einflüsse der Parameter untereinander lassen sich somit auf 0.103 quantifizieren und sind damit geringer als bei der Untersuchung der kumulativ Infizierten.