

Chapitre 4

-

Annexe D

Sélection des statistiques descriptives

Gurvan Hermange

Contenu de l'annexe

Au chapitre 4, nous avons indiqué que la construction du sous-ensemble de statistiques descriptives, noté S_A , reposait sur l'intuition sur laquelle se base la méthode de projection de Fearnhead et Prangle, à savoir qu'une statistique descriptive devrait être choisie pour sa capacité à estimer un paramètre.

Nous détaillons dans cette annexe les choix effectués pour aboutir aux statistiques descriptives utilisées pour la calibration de notre modèle.

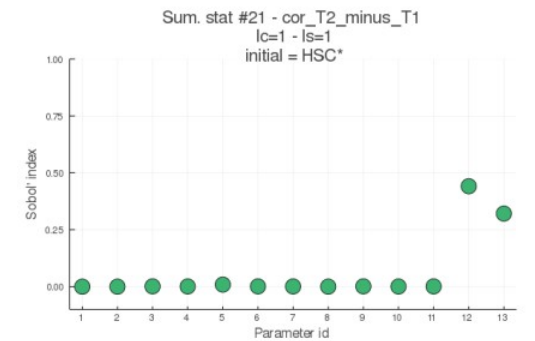
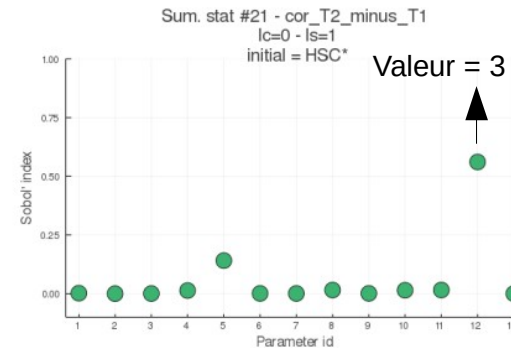
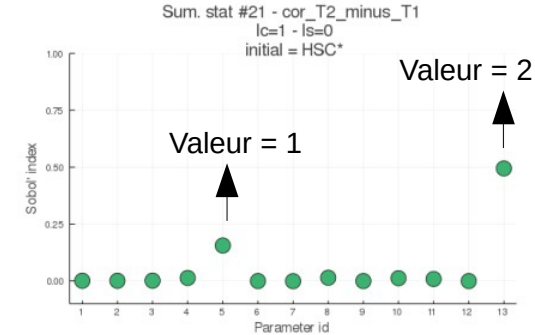
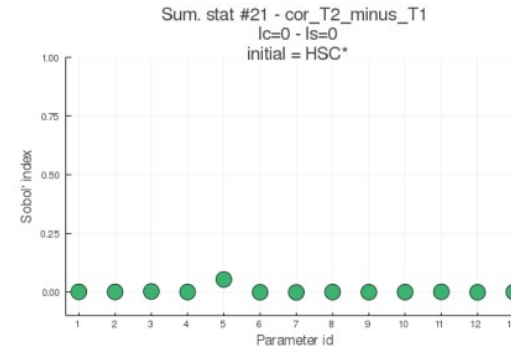
1. Intuition de la méthode

- Nous souhaitons avoir autant de statistiques descriptives que de paramètres à estimer (à savoir 11)
- Nous souhaitons que les statistiques descriptives soient choisies en fonction de leur capacité à estimer un paramètre
- C'est-à-dire que nous choisissons d'associer, à chaque paramètre, une statistique descriptive qui lui est sensible

2) Étude de chaque statistique descriptive

Exemple :

- Pour chaque statistique descriptive, et chaque condition initiale (la cellule de départ étant soit une HSC*, MPP ou HPC), nous regardons l'indice de Sobol estimé pour chaque paramètre (voir annexe B).
- Pour chaque triplet (condition initiale, paramètre, statistique descriptive), nous associons une valeur entière (variable catégorielle) comprise entre 0 et 4 indiquant si la statistique descriptive est plus (4) ou moins (0) sensible au paramètre (en regardant l'indice de Sobol (noté S dans la suite) sur les 4 différents modèles, correspondant chacun à une hypothèse avec/sans concordance/synchronicité).
- Le critère est le suivant :
 - $S \sim 0 \rightarrow 0$
 - $0 < S < 0.25 \rightarrow 1$
 - $0.25 < S < 0.5 \rightarrow 2$
 - $0.5 < S < 0.75 \rightarrow 3$
 - $0.75 < S < 1 \rightarrow 4$



Les valeurs sont renseignées dans un fichier :

Summary statistics		Cellule de départ (p)		1 – HSC								
num	label	Indice paramètre		P1 → 1	mu1	sig	P2 → 2	P3 → 3	mu3	mu4	rho_s	rho_c
21	cor_T2 minus T1					1					3	2

3) Synthèse des résultats

Summary statistics		Cellule de départ (p)	1 – HSC								2 – MPP								3 – HPC										
num	label	Indice paramètre	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c
	4 mean_T2_1			1				1	1								1	1								2	2		
	5 std_T2_1				1									1										1					
	6 skewness_T2_1																												
	7 kurtosis_T2_1																												
	8 mean_T2_2			1	1			1	1								1	1								2	2		
	9 std_T2_2				2									2										2					
	10 skewness_T2_2													1															
	11 kurtosis_T2_2																												
	12 mean_T2			1				1	1								1	1								2	2		
	13 std_T2				2									2										2					
	14 skewness_T2																												
	15 kurtosis_T2																												
	16 cor_T2				3					1				3					1					3		1	1	1	
	17 mean_T2_1_minus_T1			1				1	1								1	1								2	2		
	18 std_T2_1_minus_T1				2									2			1	1						2		2	2		
	19 mean_T2_2_minus_T1			1				1	1								1	1								2	2		
	20 std_T2_2_minus_T1				3									3										3					
	21 cor_T2_minus_T1				1					3	2			1					3	2				1				3	1
	22 mean_colony_size								3									3									3		
	23 std_colony_size								4									4									4		
	24 mean_nb_cells_gen1																												
	25 mean_nb_cells_gen2			1	1			1	1					1			1	1											
	26 mean_nb_cells_gen3							1	2								1	2											
	27 mean_nb_cells_gen4							1	2								1	3											

Les résultats de la procédure précédente sont reportés dans un fichier Excel (selection_stats_descriptives.xlsx).

À noter qu'on ne considère que les paramètres sélectionnés à l'issue de l'analyse de sensibilité. À noter également qu'on exclut les statistiques descriptives associées à l'expérience MultiGen, pour la condition initiale HPC, pour laquelle nous n'avons pas d'observations expérimentales

4) Choix du couple (stat descriptive, paramètre)

- Pour chaque paramètre (et condition initiale), on regarde la valeur maximale (pour la variable catégorielle) obtenue sur les statistiques descriptives
- Pour un paramètre, on choisira alors de lui associer une statistique descriptive appartenant à celles pour lesquelles l'indice maximal a pu être obtenu.
- Plusieurs choix sont souvent possibles, dans ce cas on choisit empiriquement la statistique descriptive, en essayant au global :
 - d'équilibrer entre celles qui sont calculées à partir d'une cellule initiale HSC*, et celles calculées à partir d'une cellule initiale MPP
 - De favoriser des statistiques descriptives qui sont sensibles à plusieurs paramètres

Cellule de départ (p)	1 – HSC									2 – MPP									3 – HPC								
Indice paramètre	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c	P1-1	mu1	sig	P2-2	P3-3	mu3	mu4	rho_s	rho_c
MAX	4	1	3	2	2	1	4	3	2	0	0	3	3	2	1	4	3	2	0	0	3	0	0	2	4	3	1
Stat. Descriptive choisie	54	25	16	-	-	-	49	21	-	-	-	-	63	61	-	-	-	21	-	-	-	-	-	18	-	-	-

5) Cas de I_s et I_c

- I_s (et I_c) valent 0 ou 1 suivant si l'hypothèse d'une synchronicité (et concordance) est considérée
- Ces indicatrices peuvent être traitées comme des paramètres
- On ajoute deux statistiques descriptives à l'ensemble S_A , choisies par leur capacité à discriminer entre les hypothèses.
- Pour cela, on regarde les figures de l'annexe A, en choisissant des statistiques descriptives telles qu'on obtient des différences dans la distribution de leurs valeurs (sur l'ensemble des paramètres explorés par échantillonnage LHS), suivant les hypothèses de concordance / synchronicité faites
- On choisit ainsi de considérer également la statistique descriptive #45 (partant d'un MPP) et #65 (partant d'une HSC*)

