Verslag Python Challenge 2

Aiko Decaluwe, Fien Dewit, Viktor van Nieuwenhuize (Groep 4)
Opleiding: Tweede Bachelor Fysica en Sterrenkunde
Vak: Statistiek en Gegevensverwerking Datum: 1 December 2020

1 Inleiding

De gegeven dataset bestaat uit waarden die uit een Gammaverdeling met een bepaalde k en θ zijn getrokken. Deze Gammaverdeling is gedefinieerd als:

$$f(y;k,\theta) = \frac{y^{k-1}e^{y/\theta}}{\theta^k \Gamma(k)}$$

Om de schatters van k en θ te bepalen bestaan er verschillende methoden. We vergelijken de methode van de momenten (MM) en de Maximum Log Likelihood methode (MLLH). Verder zullen we ook bootstrappen en deze waarden in functie van het aantal getrokken waarden (N) uitzetten. Ten slotte wordt de gebootstrapte waarde voor N_{max} vergeleken met de geschatte waarden via de MM- en MLLH-schatters.

2 Berekening schatters en variantie

2.1 Methode van de momenten

We maken gebruik van de methode van de momenten om de schatters voor k en θ te bepalen in functie van de variabelen $\overline{x}$ en $\overline{x^2}$. We berekenen dus de verwachtingswaarde van de gammaverdeling voor $\overline{x}$ en $\overline{x^2}$:

$$\begin{aligned}
M_1 = \overline{x} =< x > &= \int_0^{+\infty} x \frac{x^{\hat{k}-1}e^{x/\hat{\theta}}}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}dx \\
&= \frac{1}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}\int_0^{+\infty} x^{\hat{k}} \cdot (-\hat{\theta})de^{-x/\hat{\theta}} \\
&= \frac{\hat{\theta}}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}\left(-\left[e^{-x/\hat{\theta}}x^{\hat{k}}\right]\Big|_0^{+\infty} + \int_0^{+\infty} e^{-x/\hat{\theta}}dx^{\hat{k}}\right) \\
&= \hat{k}\hat{\theta}\int_0^{+\infty} \frac{x^{\hat{k}-1}e^{x/\hat{\theta}}}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}dx \\
&= \hat{k}\hat{\theta} \\
M_2 = \overline{x^2} =< x^2 > &= \int_0^{+\infty} x^2 \frac{x^{\hat{k}-1}e^{x/\hat{\theta}}}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}dx \\
&= \frac{1}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}\int_0^{+\infty} x^{\hat{k}+1} \cdot (-\hat{\theta})de^{-x/\hat{\theta}} \\
&= \frac{\hat{\theta}}{\hat{\theta}^{\hat{k}}\Gamma(\hat{k})}\left(-\left[e^{-x/\hat{\theta}}x^{\hat{k}+1}\right]\Big|_0^{+\infty} + \int_0^{+\infty} (k+1)e^{-x/\hat{\theta}}x^{\hat{k}}dx\right) \\
&= \hat{\theta}(\hat{k}+1)\cdot M_1 \\
&= \hat{\theta}^2\hat{k}(\hat{k}+1)
\end{aligned}$$

Uit de integralen van de verwachtingswaarden voor $\overline{x}$ en $\overline{x^2}$ bekomen we 2 vergelijkingen:

$$\begin{cases} \overline{x} &= \hat{k}\hat{\theta} \\ \overline{x^2} &= \hat{\theta}^2\hat{k}(\hat{k}+1) \end{cases}$$

Deze kunnen we oplossen naar een schatter voor k en een schatter voor θ.

$$\hat{k} = \frac{\overline{x}^2}{\overline{x^2} - \overline{x}^2}$$
$$\hat{\theta} = \frac{\overline{x^2} - \overline{x}^2}{\overline{x}}$$

De variantie van deze schatters wordt gegeven door de formule:

$$\mathrm{cov}(\hat{\theta}_\mathrm{i}, \hat{\theta}_\mathrm{j}) = \sum_{l,m} \frac{\partial \hat{\theta}_i}{\partial \hat{e}_l} \frac{\partial \hat{\theta}_j}{\partial \hat{e}_m} \hat{\mathrm{cov}}(\hat{\mathrm{e}}_\mathrm{l}, \hat{\mathrm{e}}_\mathrm{m})$$

Hierin worden de verschillende afgeleiden gegeven door:

$$\frac{\partial \hat{\theta}}{\partial \overline{x}} = \frac{-\overline{x}^2 - \overline{x^2}}{\overline{x}^2} \qquad \frac{\partial \hat{\theta}}{\partial \overline{x^2}} = \frac{1}{\overline{x}} \qquad \frac{\partial \hat{k}}{\partial \overline{x}} = \frac{2\overline{x}\overline{x^2}}{(\overline{x^2} - \overline{x}^2)^2} \qquad \frac{\partial \hat{k}}{\partial \overline{x^2}} = \frac{-\overline{x}^2}{(\overline{x^2} - \overline{x}^2)^2}$$

De uitdrukkingen voor de covariantie van de schatters werden niet manueel uitgerekend. De vergelijkingen voor de afgeleiden van de schatters $\hat{k}$ en $\hat{\theta}$ naar de momenten $\overline{x}$ en $\overline{x^2}$ werden in python geschreven en met deze uitkomsten vormden we de uitdrukking voor de covarianties. De uitdrukkingen voor de covariantie van de momenten werden ook apart uitgerekend.

$$\mathrm{cov}(\hat{k}, \hat{k}) = \left(\frac{\partial \hat{k}}{\partial \overline{x}}\right)^2 \hat{\mathrm{cov}}(\overline{x}, \overline{x}) + \left(\frac{\partial \hat{k}}{\partial \overline{x^2}}\right)^2 \hat{\mathrm{cov}}(\overline{x^2}, \overline{x^2}) + 2\left(\frac{\partial \hat{k}}{\partial \overline{x}}\right)\left(\frac{\partial \hat{k}}{\partial \overline{x^2}}\right) \hat{\mathrm{cov}}(\overline{x}, \overline{x^2})$$
$$\mathrm{cov}(\hat{\theta}, \hat{\theta}) = \left(\frac{\partial \hat{\theta}}{\partial \overline{x}}\right)^2 \hat{\mathrm{cov}}(\overline{x}, \overline{x}) + \left(\frac{\partial \hat{\theta}}{\partial \overline{x^2}}\right)^2 \hat{\mathrm{cov}}(\overline{x^2}, \overline{x^2}) + 2\left(\frac{\partial \hat{\theta}}{\partial \overline{x}}\right)\left(\frac{\partial \hat{\theta}}{\partial \overline{x^2}}\right) \hat{\mathrm{cov}}(\overline{x}, \overline{x^2})$$
$$\begin{aligned}\mathrm{cov}(\hat{k}, \hat{\theta}) = &\left(\frac{\partial \hat{k}}{\partial \overline{x}}\right)\left(\frac{\partial \hat{\theta}}{\partial \overline{x}}\right) \hat{\mathrm{cov}}(\overline{x}, \overline{x}) + \left(\frac{\partial \hat{k}}{\partial \overline{x^2}}\right)\left(\frac{\partial \hat{\theta}}{\partial \overline{x^2}}\right) \hat{\mathrm{cov}}(\overline{x^2}, \overline{x^2}) + \left(\frac{\partial \hat{k}}{\partial \overline{x}}\right)\left(\frac{\partial \hat{\theta}}{\partial \overline{x^2}}\right) \hat{\mathrm{cov}}(\overline{x}, \overline{x^2}) \\ &+ \left(\frac{\partial \hat{k}}{\partial \overline{x^2}}\right)\left(\frac{\partial \hat{\theta}}{\partial \overline{x}}\right) \hat{\mathrm{cov}}(\overline{x}, \overline{x^2})\end{aligned}$$

De covarianties van de momenten $\overline{x}$ en $\overline{x^2}$ kan berekend worden via:

$$\hat{\mathrm{cov}}(\hat{e}_l, \hat{e}_m) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (g_l(x_i) - \overline{g_l})(g_m(x_i) - \overline{g_m})$$

Waaruit dan volgt dat:

$$\hat{\mathrm{cov}}(\overline{x}, \overline{x^2}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})(x_i^2 - \overline{x^2})$$
$$\hat{\mathrm{cov}}(\overline{x}, \overline{x}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2$$
$$\hat{\mathrm{cov}}(\overline{x^2}, \overline{x^2}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i^2 - \overline{x^2})^2$$

Vervolgens kunnen we de correlatie tussen $\hat{\theta}$ en $\hat{k}$ bereken met behulp van de formule voor de variantie. Ook dit werd niet volledig met de hand uitgerekend, maar volgt gewoon uit de vorige uitkomsten.

$$\rho_{\hat{\theta}\hat{k}} = \frac{\mathrm{cov}(\hat{\theta}, \hat{\mathrm{k}})}{\sqrt{\mathrm{cov}(\hat{\theta}, \hat{\theta})\,\mathrm{cov}(\hat{k}, \hat{k})}}$$

2.2 Maximum Log Likelihood Methode

De likelihoodvergelijking van de Gammaverdeling is:

$$\mathcal{L}(y;k,\theta) = \prod_{i=1}^{N} \frac{y_i^{k-1} e^{-y/\theta}}{\theta^k \Gamma(k)}$$

Indien men hiervan het natuurlijk logaritme neemt, verkrijgt men het volgende:

$$\begin{aligned}
\ln\Big(\mathcal{L}(y;k,\theta)\Big) &= \sum_{i}^{N} \ln\left(\frac{y_i^{k-1} e^{-y/\theta}}{\theta^k \Gamma(k)}\right) \\
&= \sum_{i}^{N} \left[\ln\Big(y_i^{k-1}\Big) - \frac{y_i}{\theta} - \ln\Big(\theta^k\Big) - \ln\Big(\Gamma(k)\Big)\right]
\end{aligned}$$

Om vergelijkingen voor de schatters $\hat{k}$ en $\hat{\theta}$ te vinden berekenen we het maximum van deze log likelihood. Dit doen we door de partiële afgeleide naar de schatters te nemen en dit gelijk te stellen aan 0.

1) Afleiden naar k

$$\begin{aligned}
\frac{\partial}{\partial k}\left(\ln\Big(\mathcal{L}(y;k,\theta)\Big)\right)\Bigg|_{k=\hat{k}} &= \sum_{i}^{N}\left[\ln\Big(y_i\Big) - \ln\Big(\theta\Big) - \frac{\Gamma'(k)}{\Gamma(k)}\right]\Bigg|_{k=\hat{k}} \\
\iff 0 &= \sum_{i}^{N}\left[\ln\Big(y_i\Big) - \ln\Big(\hat{\theta}\Big) - \frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})}\right] \\
\iff 0 &= \overline{\ln\Big(y\Big)} - \ln\Big(\hat{\theta}\Big) - \psi^{(0)}(\hat{k})
\end{aligned}$$

2) Afleiden naar θ

$$\begin{aligned}
&\frac{\partial}{\partial \hat{\theta}}\left(\ln\Big(\mathcal{L}(y;k,\theta)\Big)\right)\Bigg|_{\theta=\hat{\theta}} = \sum_{i}^{N}\left[\frac{y_i}{\theta^2} - \frac{k\theta^{k-1}}{\theta^k}\right]\Bigg|_{\theta=\hat{\theta}} \\
&\iff 0 = \sum_{i}^{N}\left[\frac{y_i}{\hat{\theta}^2} - \frac{\hat{k}}{\hat{\theta}}\right] \\
&\iff \frac{1}{N}\sum_{i}^{N} y_i = \hat{k}\hat{\theta} \\
&\iff \overline{y} = \hat{k}\hat{\theta}
\end{aligned}$$

Door af te leiden naar beide schatters wordt er een stelsel van 2 vergelijkingen verkregen. Uit de vergelijking voor de afgeleide naar $\hat{\theta}$ kon de schatter $\hat{\theta}$ worden gehaald. Dit was echter niet het geval voor de vergelijking die we verkregen nadat de log likelihood vergelijking afgeleid werd naar $\hat{k}$. Deze vergelijking werd opgelost met een numerieke methode.

Hiervoor hebben we de verschillende "Nonlinear solvers" getest op de performantie. De ene methode zal sneller zijn voor de ene vergelijking en een andere methode zal sneller zijn voor een andere vergelijking. Aangezien dit heel vaak wordt herhaald zal zelfs een kleine verbetering in snelheid een grote invloed hebben op de snelheid om het script uit te voeren. Uiteindelijk kwamen we uit op een gelijkspel tussen 'broyden1' en 'anderson'

$$\begin{cases}
0 &= \overline{\ln\Big(y\Big)} - \ln\Big(\hat{\theta}\Big) - \psi^{(0)}(\hat{k}) \\
\overline{y} &= \hat{k}\hat{\theta}
\end{cases}$$

Wanneer de vergelijking voor $\hat{\theta}$ wordt ingevuld in vergelijking (1) krijgen we een vergelijking die enkel afhankelijk is van de schatter $\hat{k}$. Deze vergelijking kan dan numeriek opgelost worden.

$$\overline{\ln\Big(y\Big)} - \ln\Big(\overline{y}\Big) + \ln\Big(\hat{k}\Big) - \psi^{(0)}(\hat{k}) = 0$$

Om de variantie te bepalen van de schatters kan de volgende vergelijking worden gebruikt:

$$\mathrm{V}(\hat{a})^{-1} = -\frac{d^2}{da^2}\Big[\ln(\mathcal{L})\Big]\Big|_{a=\hat{a}}$$

1) Variantie $\hat{k}$:

$$\begin{aligned}\mathrm{V}(\hat{k})^{-1} &= -\frac{\partial}{\partial k}\sum_i^N \left[\ln\Big(y_i\Big) - \ln\Big(\hat{\theta}\Big) - \frac{\Gamma'(k)}{\Gamma(k)}\right]\Bigg|_{k=\hat{k}} \\ &= N\psi^{(1)}(\hat{k})\end{aligned}$$

2) De variantie van $\hat{\theta}$

$$\begin{aligned}\mathrm{V}(\theta)^{-1} &= -\frac{\partial}{\partial \theta}\sum_{i=1}^N \left[\frac{y_i}{\theta^2} - \frac{\hat{k}}{\theta}\right]\Bigg|_{\theta=\hat{\theta}} \\ &= N\frac{\hat{k}}{\hat{\theta}^2}\end{aligned}$$

De covariantie kan berekend worden via:

$$\begin{aligned}\mathrm{cov}^{-1}(\hat{k},\hat{\theta}) &= -\frac{\partial^2}{\partial k \partial \theta}\Big[\ln(\mathcal{L})\Big]\Big|_{\substack{k=\hat{k}\\ \theta=\hat{\theta}}} \\ &= -\frac{\partial}{\partial \theta}\sum_i^N \left[\ln\Big(y_i\Big) - \ln\Big(\theta\Big) - \frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})}\right]\Bigg|_{\theta=\hat{\theta}} \\ &= \frac{N}{\hat{\theta}} \\ &= -\frac{\partial}{\partial \theta}\sum_{i=1}^N \left[\frac{y_i}{\hat{\theta}^2} - \frac{k}{\hat{\theta}}\right]\Bigg|_{k=\hat{k}} \\ &= \mathrm{cov}^{-1}(\hat{\theta},\hat{k})\end{aligned}$$

Hierna kunnen deze uitkomsten in een matrix worden gezet om de inverse covariantiematrix te bepalen.

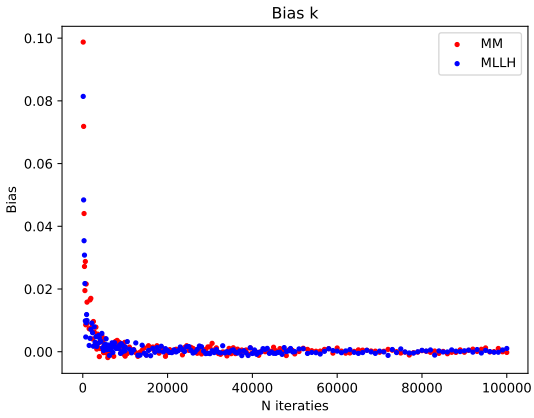
$$V^{-1} = \begin{bmatrix} N\psi^{(1)}(\hat{k}) & \frac{N}{\hat{\theta}} \\ \frac{N}{\hat{\theta}} & N\frac{\hat{k}}{\hat{\theta}^2}\end{bmatrix}$$

Om de echte variantie te berekenen moet deze matrix worden geïnverteerd. Dit wordt volledig in python gedaan en niet expliciet uitgerekend in het verslag. Deze matrixinversie creëert ook een afhankelijkheid voor $\hat{k}$ en $\hat{\theta}$ voor zowel de variantie van $\hat{k}$ als de covariantie. Iets wat niet aanwezig is in de inverse variantie en covariantie.

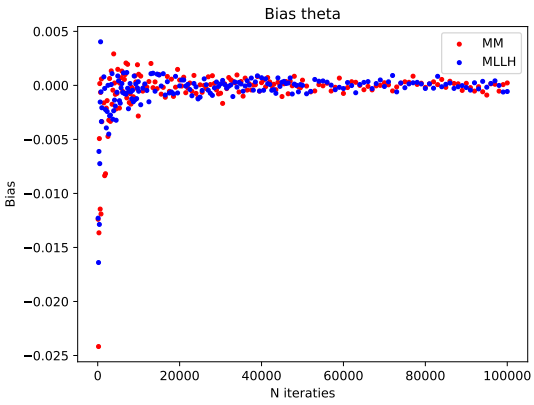
Verder werd er ook gebruik gemaakt van de polygamma functie $\psi^{(n)}(x)$. Waarbij $\psi^{(0)}(x)$ de digamma functie is.

3 Vergelijken van MM en MLLH-methode

Nu bestuderen we de eigenschappen van $\hat{k}$ en $\hat{\theta}$ als functie van het aantal getrokken samples. Dit wordt gedaan aan de hand van bootstrappen. Hierbij nemen we N willekeurige samples uit de set metingen en passen we de schatters $\hat{k}$ en $\hat{\theta}$ toe. Dit werd dan M keer herhaald. op basis van de M resultaten die we uitkwamen, werd de bias, $V(\hat{k})$, $V(\hat{\theta})$ en $\rho(\hat{k},\hat{\theta})$ bepaald. Wij kozen $M = 1000$.
Dit werd herhaald voor enkele metingen tussen 1 en de hoeveelheid data N_{max}. Hoe groter de waarde voor N, hoe kleiner het verschil tussen N en $N-1$ werd. Hierdoor was het mogelijk om de stappen te vergroten naarmate N groter werd, wat ons wat winst in snelheid opleverde.
In figuur 1, 2 en 3 zijn de plots te zien voor de bias, de variantie en de correlatiecoëfficiënt in functie van N.

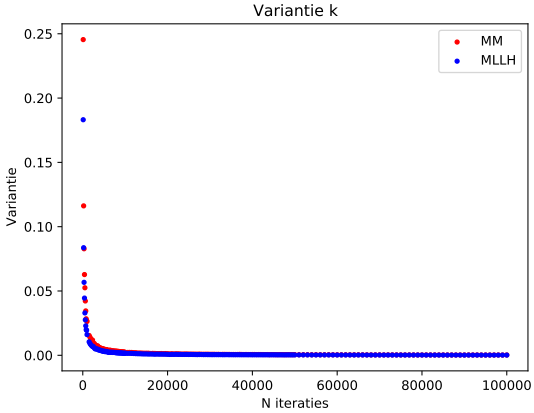


(a) *Bias van $\hat{k}$ in functie van N_{max}*

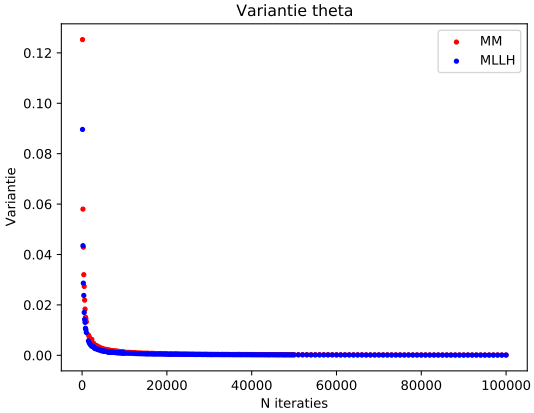


(b) *Bias van $\hat{\theta}$ in functie van N_{max}*

Figuur 1: *Bias voor $\hat{k}$ en $\hat{\theta}$ voor MM en MLLH*

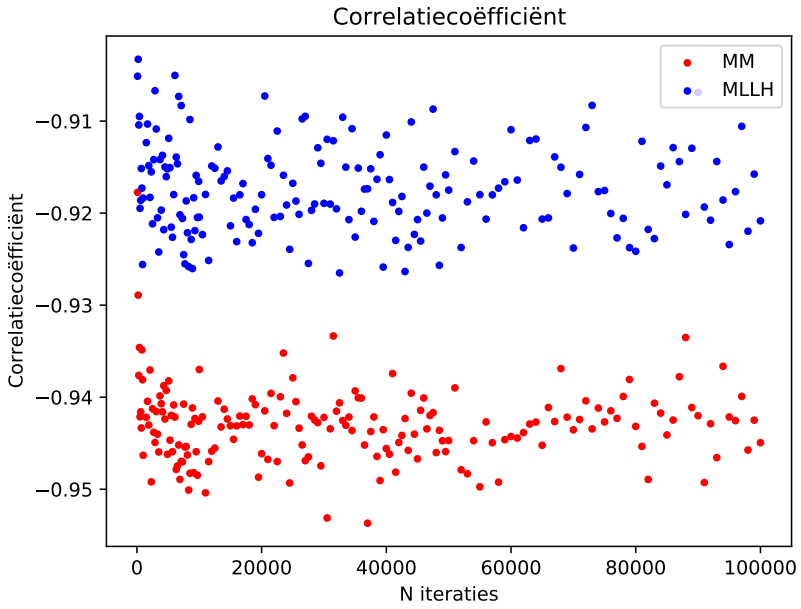


(a) *Variantie van $\hat{k}$ in functie van N_{max}*



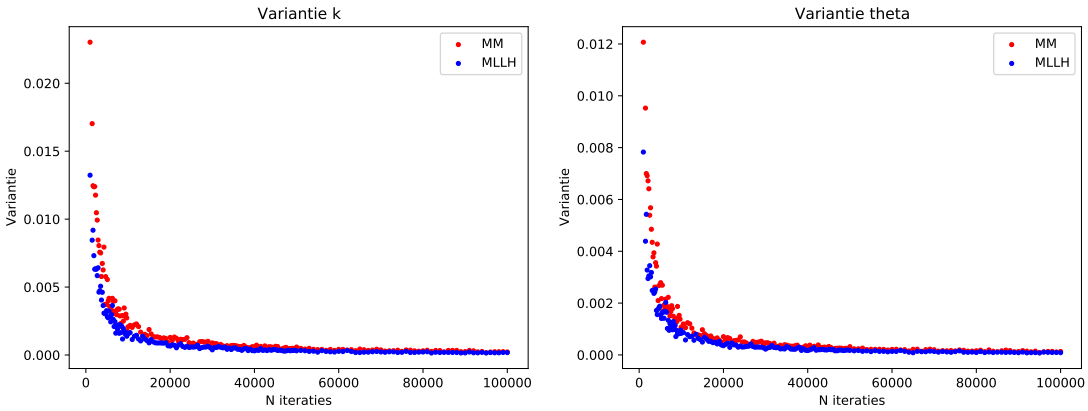
(b) *Variantie van $\hat{\theta}$ in functie van N_{max}*

Figuur 2: *Variantie voor $\hat{k}$ en $\hat{\theta}$ voor MM en MLLH*



Figuur 3: *Correlatiecoëfficiënt voor $\hat{k}$ en $\hat{\theta}$ in functie van N_{max}*

In figuur 1 zien we de biassen van $\hat{k}$ en $\hat{\theta}$. Er is weinig verschil op te merken tussen de bias voor de schatters bij de methode van de momenten en de maximum likelihood methode.
Voor de variantie is er echter vrij weinig te zien op de plots in figuur 2. Hiervoor maakten we een extra plotje waar de punten pas vanaf $N = 1000$ werden geplot om een duidelijker beeld te krijgen van wat er gebeurd. Dit is te zien in figuur 4.



(a) *Variantie van $\hat{k}$ in functie van N_{max}*

(b) *Variantie van $\hat{\theta}$ in functie van N_{max}*

Figuur 4: *Variantie voor $\hat{k}$ en $\hat{\theta}$ voor MM en MLLH*

Deze figuren tonen duidelijk dat de variantie van beide schatters groter is als de methode van de momenten wordt toegepast. Dit betekent dat de MLLH-methode toch wat efficiënter is dan de MM.

Ten slotte zien we bij de correlaties een duidelijke scheiding. De correlaties zijn niet afhankelijk van het aantal getrokken waarden wat te verwachten is. We zien een bijna discrete scheiding tussen de correlaties. Ook is de absolute waarde van de correlatiecoëfficiënt van de schatters $\hat{k}$ en $\hat{\theta}$ kleiner bij de MLLH-methode dan bij de MM. De MLLH-schatters zijn dus minder gecorreleerd.

De MLLH-methode is efficiënter dan de MM. Er moet echter wel vermeld worden dat het bepalen van de schatters hierbij een stuk ingewikkelder was. $\hat{k}$ moest namelijk numeriek uitgerekend worden. Dit zorgde ervoor dat ook het bootstrappen bij de MLLH-methode een stuk langer duurde. Verder waren de MLLH-schatters minder gecorreleerd dan de MM-schatters. Hierdoor komen we tot dezelfde conclusie als de syllabus. De MM is een stuk eenvoudiger maar iets minder precies en dus goed voor een snelle schatting. Wanneer men preciezere resultaten wilt, is de MLLH-methode beter, maar dit is een pak ingewikkelder en duurt ook een stuk langer.

Indien we de methode slechts 1 keer moeten toepassen, is de tijd een minder groot probleem. Maar in toepassingen zoals deze, waar de methodes heel veel keer worden toegepast (in ons geval ongeveer $17,6$ miljard keer), is het verschil in tijd van veel groter belang.
Het duurde 152 seconden om de berekeningen voor de MM uit te voeren. Voor de MLLH-methode was dit 595 seconden. Wat net geen 4 keer de tijd voor de MM is.

4 Vergelijken van de gebootstrapte varianties en de schattingen van de varianties

4.1 Methode van de Momenten

	bootstrap	schatting
$V(\hat{k})$	$0.231 \cdot 10^{-3}$	$0.241 \cdot 10^{-3}$
$V(\hat{\theta})$	$0.119 \cdot 10^{-3}$	$0.127 \cdot 10^{-3}$
$\rho_{\hat{k},\hat{\theta}}$	-0.941	-0.943

Tabel 1: Vergelijking tussen de bootstrap en de schatting voor MM

Bij beide varianties en de correlatie kunnen we zien dat de waarden voor de bootstrap kleiner zijn dan die van de schatting. Het verschil is wel relatief klein, vooral bij de covariantie.

4.2 Maximum Log Likelihood Methode

	bootstrap	schatting
$V(\hat{k})$	$0.160 \cdot 10^{-3}$	$0.159 \cdot 10^{-3}$
$V(\hat{\theta})$	$88.6 \cdot 10^{-6}$	$86.8 \cdot 10^{-6}$
$\rho_{\hat{k},\hat{\theta}}$	-0.920	-0.918

Tabel 2: Vergelijking tussen de bootstrap en de schatting voor MLLH

In tegenstelling tot de methode van de momenten is de bootstrap altijd iets groter dan de schatting. Zowel voor de varianties als de correlatie. Het verschil is echter zeer minimaal. Het is ook een stuk kleiner dan het verschil dat we zien bij de methode van de momenten. De variantie voor $\hat{\theta}$ is zelfs een groteorde kleiner bij de MLLH-methode dan bij de MM. Ook dit resultaat versterkt de stelling die in deel 3 werd gemaakt. Daar stelden we dat de MLLH-methode preciezer is dan de MM.