# LINMA1731 – Project 2019
# Fish schools tracking

By Louis Navarre and Gilles Peiffer

**Abstract**

In this paper we propose a solution to the project for the class "Stochastic processes: Estimation and prediction" given during the Fall term of 2019 at the EPL by Luc Vandendorpe and Pierre-Antoine Absil. The average speed of each fish in a school of fish is approximated by a gamma-distributed random variable with a shape parameter $k$ and a scale parameter $s$, and various methods for estimating this quantity are given; numerical simulations are also included, using Julia and J$^{\text{u}}$MP [3]. This paper also details how to obtain these speed measurements from noisy observations of the position. Using a particle filter, one can then try to use these noisy observations in order to obtain a more accurate position for each of the fish. Finally, the effect of various parameters of the model, such as the number of particles per fish, the sampling period and the standard deviation of the observational noise is studied.

**Contents**



© 2019 Gilles Peiffer and Louis Navarre.

# Part 1. Average speed estimation

## 1. Introduction

For the purpose of this project, we assume that the speed of each fish in a school at time $i$ is a random variable $V_i$ following a Gamma distribution, as suggested in [6]. This distribution is characterized by two parameters: a shape parameter $k > 0$ and a scale parameter $s > 0$. The parameters are the same for every fish and are time invariant. The aim of this first part is to identify these two parameters using empirical observations $v_i$.

## 2. A maximum likelihood estimator for the scale parameter

Let $v_i$ be i.i.d. realisations of a random variable following a Gamma distribution $\Gamma(k, s)$ (with $i = 1, \dots, N$) [16]. We first assume that the shape parameter $k$ is known.

We start by deriving the maximum likelihood estimator of $\vartheta \coloneqq s$ based on $N$ observations [19]. Since the estimand $\vartheta$ is a deterministic quantity, we use Fisher estimation. In order to do this, let us restate the probability density function of $V_i \sim \Gamma(k, s)$:

$$f_{V_i}(v_i; k, s) = \frac{1}{\Gamma(k) s^k} v_i^{k-1} \mathrm{e}^{-\frac{v_i}{s}}, \quad i = 1, \dots, N. \tag{2.1}$$

With this in mind, we can find that the likelihood $\mathcal{L}(v_1, \dots, v_N; k, \vartheta)$ is given by

$$\mathcal{L}(v_1, \dots, v_N; k, \vartheta) = \prod_{i=1}^{N} f_{V_i}(v_i; k, \vartheta) = \prod_{i=1}^{N} \frac{1}{\Gamma(k) \vartheta^k} v_i^{k-1} \mathrm{e}^{-\frac{v_i}{\vartheta}}. \tag{2.2}$$

In order to alleviate notation, we compute instead the log-likelihood, which is generally easier to work with[1]:

$$\ell(v_1, \dots, v_N; k, \vartheta) \coloneqq \ln \mathcal{L}(v_1, \dots, v_N; k, \vartheta) \tag{2.3}$$

$$= \sum_{i=1}^{N} \ln \left( \frac{1}{\Gamma(k) \vartheta^k} v_i^{k-1} \mathrm{e}^{-\frac{v_i}{\vartheta}} \right) \tag{2.4}$$

$$= (k-1) \sum_{i=1}^{N} \ln v_i - \sum_{i=1}^{N} \frac{v_i}{\vartheta} - N\big(k \ln \vartheta + \ln \Gamma(k)\big). \tag{2.5}$$

Now, in order to obtain the maximum likelihood estimate $\hat{\vartheta}$, we must differentiate the log-likelihood with respect to the estimand $\vartheta$, and set it equal to zero:

$$\left. \frac{\partial \ell(v_1, \dots, v_N; k, \vartheta)}{\partial \vartheta} \right|_{\vartheta = \hat{\vartheta}} = -\frac{kN}{\hat{\vartheta}} + \frac{\sum_{i=1}^{N} v_i}{\hat{\vartheta}^2} = 0 \tag{2.6}$$

$$\iff \hat{\vartheta} = \frac{\sum_{i=1}^{N} v_i}{kN} = \frac{\overline{v}}{k}. \tag{2.7}$$

This then allows us to find the maximum likelihood estimator $\widehat{\Theta}$, given by

$$\widehat{\Theta} = \frac{\sum_{i=1}^{N} V_i}{kN} = \frac{\overline{V}}{k}. \tag{2.8}$$

[1] This is possible because the values of $\vartheta$ which maximize the log-likelihood also maximize the likelihood.

## 3. **Asymptotic properties of the maximum likelihood estimator**

We now wish to show some of the properties of this estimator. The definitions of these properties are given in Appendix A.

PROPERTY 3.1. *The maximum likelihood estimator derived in* (2.8) *is asymptotically unbiased, that is,*

$$\lim_{N\to+\infty} \mathbb{E}\left[g(V_1,\dots,V_N);\vartheta\right] = \vartheta\,. \tag{3.1}$$

*Proof.* We wish to prove that $\lim_{N\to+\infty} \mathbb{E}\left[\frac{\overline{V}}{k}\right] = \vartheta$. We recall that $\mathbb{E}\left[V_i\right] = k\vartheta$ for $V_i \sim \Gamma(k,\vartheta)$ and that the expected value operator is linear to obtain

$$\mathbb{E}\left[\frac{\overline{V}}{k}\right] = \frac{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N V_i\right]}{k} = \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[V_i\right]}{k} = \frac{\frac{1}{N}Nk\vartheta}{k} = \vartheta\,. \tag{3.2}$$

This proves that the maximum likelihood estimator of (2.8) is unbiased, hence it is also asymptotically unbiased. This can be seen on Figures 1a and 1b. □

PROPERTY 3.2. *The maximum likelihood estimator derived in* (2.8) *is efficient.*

*Proof.* We use the fact that the random variables are independent to simplify the computations. Since $\vartheta$ is a scalar parameter, the Fisher information matrix is a scalar, equal to

$$\mathcal{I}(\vartheta) = -N\mathbb{E}\left[\frac{\partial^2}{\partial\vartheta^2}\left((k-1)\ln v_1 - \frac{v_1}{\vartheta} - \big(k\ln\vartheta + \ln\Gamma(k)\big)\right)\right] \tag{3.3}$$

$$= N\mathbb{E}\left[\frac{\partial^2}{\partial\vartheta^2}\Big(\frac{v_1}{\vartheta} + k\ln\vartheta\Big)\right] = \frac{kN}{\vartheta^2}\,. \tag{3.4}$$

We must also compute the variance of the ML estimator $\widehat{\Theta}$, which is given by

$$\mathbb{V}\left[\widehat{\Theta}\right] = \mathbb{V}\left[\frac{\overline{V}}{k}\right] = \frac{\vartheta^2}{kN}\,. \tag{3.5}$$

The Cramér–Rao lower bound is thus reached for all values of $\vartheta$, which concludes the proof. This is shown on Figure 2. □

PROPERTY 3.3. *The maximum likelihood estimator of* (2.8) *is best asymptotically normal.*

*Proof.* In our case, we can show using the Cramér–Rao lower bound that $\Sigma$ is minimal if it is equal to $\mathcal{I}^{-1}(\vartheta)$. To alleviate notations, we will write $\ell(\vartheta)$ instead of $\ell(v_1,\dots,v_N;k,\vartheta)$. By definition, since $\hat{\vartheta} = \arg\max_\vartheta \ell(\vartheta)$, we know that $\ell'(\hat{\vartheta}) = 0$. Let $\vartheta_0$ be the true value of the parameter $\vartheta$. We can then use Taylor expansion on $\ell'(\hat{\vartheta})$ around $\hat{\vartheta} = \vartheta_0$ to obtain

$$\ell'(\hat{\vartheta}) = \ell'(\vartheta_0) + \frac{\ell''(\vartheta_0)}{1!}(\hat{\vartheta}-\vartheta_0) + \mathcal{O}\left((\hat{\vartheta}-\vartheta_0)^2\right)\,. \tag{3.6}$$

We know the expression on the left is zero, hence

$$\ell'(\vartheta_0) = -\ell''(\vartheta_0)(\hat{\vartheta}-\vartheta_0) + \mathcal{O}\left((\hat{\vartheta}-\vartheta_0)^2\right)\,. \tag{3.7}$$

Rearranging and multiplying by $\sqrt{n}$, we get

$$\sqrt{n}(\hat{\vartheta}-\vartheta_0) = \frac{\ell'(\vartheta_0)/\sqrt{n}}{-\ell''(\vartheta_0)/n + \mathcal{O}\left((\hat{\vartheta}-\vartheta_0)/n\right)}\,. \tag{3.8}$$

Next, we need to show that $\frac{1}{\sqrt{n}}\ell'(\vartheta_0) \sim \mathcal{N}\big(0, \mathcal{I}(\vartheta_0)\big)$. This is done using the Lindeberg–Lévy central limit theorem, in Appendix B [14]. We know that $\frac{1}{N}\ell''(\vartheta_0) = \mathcal{I}(\vartheta_0)$. Finally, we can rewrite

$$\sqrt{N}(\hat{\vartheta} - \vartheta_0) \sim \frac{\mathcal{N}\big(0, \mathcal{I}(\vartheta_0)\big)}{\mathcal{I}(\vartheta_0)} = \mathcal{N}\big(0, \mathcal{I}^{-1}(\vartheta_0)\big)\,, \tag{3.9}$$

where we didn't take into account the remainder of the Taylor series, which goes to zero. This proves that the ML estimator is best asymptotically normal. □

PROPERTY 3.4. *The maximum likelihood estimator of* (2.8) *is consistent.*

*Proof.* We have shown that the estimator is unbiased, hence its MSE is equal to its variance. Since the estimator is efficient by Property 3.2, we know that its variance is equal to the Cramér–Rao lower bound, $\operatorname{cov}\widehat{\Theta} = \mathcal{I}^{-1}(\vartheta)$. We found this lower bound to be equal to $\frac{\vartheta^2}{kN}$ in (3.4). We have

$$\lim_{N\to+\infty} \operatorname{cov}\widehat{\Theta} = \lim_{N\to+\infty} \frac{\vartheta^2}{kN} = 0\,. \tag{3.10}$$

This proves that the variance (and hence the mean square error) of the estimator goes to zero as $N$ goes to infinity, hence the estimator is consistent. □

## 4. **Joint maximum likelihood estimation**

We now consider $V_i \sim \Gamma(k, s)$ (for $i = 1, \dots, N$) with both $k$ and $s$ unknown. Before, we assumed $k$ was known, so we could maximize the log-likelihood function with respect to $s$. Now, we have to maximize this function with respect to $s$ and $k$ at the same time. We know the maximum likelihood estimator of $s$, $\hat{s} = f(k)$. Therefore, in the log-likelihood function, we can replace all the occurrences of $s$ by the estimator we found, $\hat{s}$. One then gets a function of $k$ only, which can be differentiated and its derivative set to zero. Solving this, one can find the maximum likelihood estimator of $k$. We abusively write $\ell(\vartheta)$ instead of $\ell(v_1, \dots, v_N; \vartheta)$. Substituting in the estimator $\hat{s}$ instead of $\vartheta$ in the log-likelihood given in (2.5), one finds

$$\ell(\vartheta) = (k-1)\sum_{i=1}^{N} \ln v_i - \sum_{i=1}^{N} \frac{k v_i}{\overline{v}} - Nk \ln \overline{v} + Nk \ln k - N \ln \Gamma(k)\big)\,. \tag{4.1}$$

Taking the derivative of this function with respect to $k$, we get

$$\left.\frac{\partial \ell(\vartheta)}{\partial k}\right|_{k=\hat{k}} = \sum_{i=1}^{N} \ln v_i - N - N \ln \overline{v} + N \ln \hat{k} + \frac{N\hat{k}}{\hat{k}} - N\frac{\Gamma(\hat{k})}{\Gamma(\hat{k})}\psi_0(\hat{k}) \tag{4.2}$$

$$= \sum_{i=1}^{N} \ln v_i - N \ln \sum_{i=1}^{N} v_i + N \ln \hat{k} + N \ln N - N\psi_0(\hat{k})\,, \tag{4.3}$$

where $\psi_0$ is the digamma function, i.e. the logarithmic derivative of the gamma function [4]. One must now look for a root of this equation:

$$\ln \hat{k} - \psi_0(\hat{k}) = \ln\left(\sum_{i=1}^{N} v_i\right) - \ln N - \frac{\sum_{i=1}^{N} \ln v_i}{N} \tag{4.4}$$

$$\iff \ln \hat{k} - \psi_0(\hat{k}) = \ln\left(\frac{\sum_{i=1}^{N} v_i}{N}\right) - \frac{\sum_{i=1}^{N} \ln v_i}{N}\,. \tag{4.5}$$

This equation has no closed-form solution for $\hat{k}$, but can be approximated using numerical methods since the function is very well-behaved.

## 5. Numerical experiments with various estimators

For the numerical simulation, $N$ random variables were generated from a distribution with parameters $\Gamma(1, 2)$, for different values of $N$ (`10:50:1000`). For each value of $N$, the experiment was repeated $M = 500$ times.

In order to use method of moments estimation for the Gamma distribution given in (2.1), one first needs to know its characteristic function: $\varphi_{V_i}(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{j}tV_i}\right] = (1 - \mathrm{j}st)^{-k}$. The $n$th moment is given by $\mu_n = \mathbb{E}\left[V_i^n\right] = \mathrm{j}^{-n}\varphi_{V_i}^{(n)}(0)$ [15]. Since the parameter vector has dimension two, we need to compute the first two moments. These are given by $\mu_1 = ks$ and $\mu_2 = k(k+1)s^2$. One can use the sample moments $\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} v_i$ and $\hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N} v_i^2$ to estimate $\mu_1$ and $\mu_2$. Using some simple algebra, one then finds

$$\hat{k}_{\mathrm{MOM}} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}\,, \quad \hat{s}_{\mathrm{MOM}} = \frac{\hat{\mu}_2}{\hat{\mu}_1} - \hat{\mu}_1\,. \tag{5.1}$$

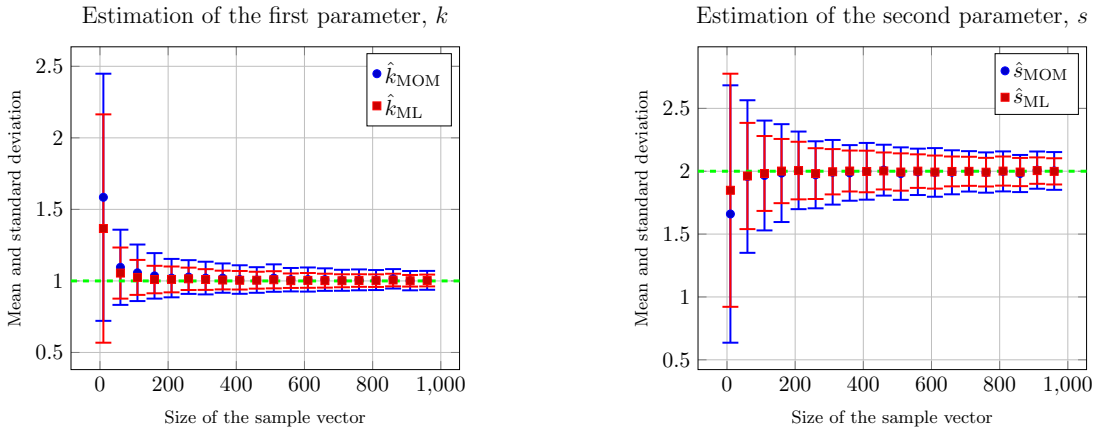
The maximum likelihood estimators are computed using the formulas in (2.8) and (4.5), for the given sample. As mentioned in Section 4, the maximum likelihood estimator for $k$ has no closed-form solution, but can be approximated using numerical methods which require an initial guess. One such first guess could be provided by the method of moments estimator for the parameter, that is

$$\hat{k}_{\mathrm{ML}}^{(0)} = \hat{k}_{\mathrm{MOM}} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}\,. \tag{5.2}$$

Another possible choice for the first guess is

$$\hat{k}_{\mathrm{ML}}^{(0)} = \frac{3 - \xi + \sqrt{(\xi - 3)^2 + 24\xi}}{12\xi}\,, \quad \text{where } \xi = \ln \overline{v} + \frac{1}{N}\sum_{i=1}^{N} \ln v_i\,. \tag{5.3}$$

This guess can be shown to be within 1.5 % of the actual value [11]. The estimator for $s$ can then be found from (2.8), using $\hat{k}_{\mathrm{ML}}$ instead of $k$.




(a) (b)

Figure 1. Both the method of moments estimator and the maximum likelihood estimator get increasingly accurate as the sample size goes up. The green line is the true value of the parameter, while the dots and squares indicate the mean of the estimators for a given value of $N$. The error bars are determined by the standard deviation of the estimators.

On Figures 1a and 1b, the mean and standard deviation are shown for the $M$ values of both the method of moments estimator and the maximum likelihood estimator, for different values of the size

of the sample, $N$. For both parameters, both estimators are unbiased, but the ML estimator has a lower variance. This should not come as a surprise: since it is efficient, every other estimator must have a greater or equal asymptotic variance. As is shown numerically in Section 7, this variance asymptotically goes to $\mathcal{I}^{-1}(\vartheta)$, the Cramér–Rao lower bound. This is also expected, in light of Property 3.2.[2]

## 6. An analytical derivation of the Fisher information matrix

One can analytically derive the Fisher information matrix. Since there are two estimators, the dimensions of the matrix are $2 \times 2$. Using the definition of the Fisher information matrix given in Theorem A.1, the Fisher matrix is given by

$$\mathcal{I}(\vartheta) = \begin{pmatrix} -\mathbb{E}\left[\dfrac{\partial^2 \ell(v;k,s)}{\partial s^2}\right] & -\mathbb{E}\left[\dfrac{\partial^2 \ell(v;k,s)}{\partial k \partial s}\right] \\ -\mathbb{E}\left[\dfrac{\partial^2 \ell(v;k,s)}{\partial s \partial k}\right] & -\mathbb{E}\left[\dfrac{\partial^2 \ell(v;k,s)}{\partial k^2}\right] \end{pmatrix}, \tag{6.1}$$

where $\ell(v;k,s)$ is used as a shorthand for $\ell(v_1,\dots,v_N;k,s)$, given by

$$\ell(v_1,\dots,v_N;k,s) = (k-1)\sum_{i=1}^{N} \ln v_i - \sum_{i=1}^{N} \frac{v_i}{s} - N\big(k \ln s + \ln \Gamma(k)\big)\,. \tag{6.2}$$

Computing the entries of the matrix is fairly tedious, and the details are in Appendix C. The Fisher information matrix is then

$$\mathcal{I}(\vartheta) = N \begin{pmatrix} \frac{k}{s^2} & \frac{1}{s} \\ \frac{1}{s} & \psi_1(k) \end{pmatrix}, \tag{6.3}$$

and the Cramér–Rao lower bound, given by the inverse of the Fisher information matrix, is

$$\mathcal{I}^{-1}(\vartheta) = \frac{1}{N} \begin{pmatrix} \psi_1(k) & -\frac{1}{s} \\ -\frac{1}{s} & \frac{k}{s^2} \end{pmatrix} \frac{s^2}{k\psi_1(k) - 1}\,, \tag{6.4}$$

where $\psi_1$ is the trigamma function, i.e. the second derivative of the logarithm of the gamma function [5].

## 7. Numerical evidence of convergence to the Cramér–Rao bound

Figure 2 gives the spectral norm as defined in [18] of

$$\mathscr{R} = \operatorname{cov} \widehat{\Theta} \oslash \mathcal{I}(\vartheta) - \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{7.1}$$

where $\oslash$ denotes Hadamard division as defined in [17] and $\widehat{\Theta}$ is $(\hat{k}_{\mathrm{ML}}, \hat{s}_{\mathrm{ML}})$, for different values of $N$. If this estimator is efficient, we know by Definition A.2 that its covariance should asymptotically go to the inverse of the Fisher information matrix.

Let the covariance matrix be defined as in [13], that is

$$\operatorname{cov} \widehat{\Theta} = \frac{1}{M-1} \begin{pmatrix} \sum_{i=1}^{M} \hat{s}_{i,\mathrm{c}}^2 & \sum_{i=1}^{M} \hat{k}_{i,\mathrm{c}} \hat{s}_{i,\mathrm{c}} \\ \sum_{i=1}^{M} \hat{s}_{i,\mathrm{c}} \hat{k}_{i,\mathrm{c}} & \sum_{i=1}^{M} \hat{k}_{i,\mathrm{c}}^2 \end{pmatrix}. \tag{7.2}$$

where $\hat{k}_i$ and $\hat{s}_i$ are the maximum likelihood estimators for the sample in the $i$th repetition of the experiment, and $\hat{\varrho}_{i,\mathrm{c}} = \hat{\varrho}_i - \tilde{\varrho}$, with $\tilde{\varrho}$ denoting the mean over all experiments of $\hat{\varrho}$ for a given value of $N$, for $\varrho = k, s$.

[2]While Property 3.2 is only proved for the estimator in (2.8), it also holds for the estimator found in (4.5), as it is a general property of maximum likelihood estimators.

If one takes the Hadamard division of the covariance matrix by the inverse of the Fisher information matrix, the result should converge to a matrix of ones [17]. In order to visualize this convergence, the matrix must be "centered", by removing one in every position[3].

For any norm[4] of the resulting matrix $\mathscr{R}$, one can then observe that $\lim_{N\to+\infty} \|\mathscr{R}\| = 0$. This is shown for the spectral norm on Figure 2. To generate this picture, estimators were computed for samples of size $N \in \{\,10, 50, 150, 3000\,\}$. This experiment was repeated $M = 10000$ times. The ratio matrix was then computed as the element-wise division between the empirical covariance of the estimators for a given $N$ and the inverse of the Fisher information matrix, given by (6.4). Finally, the matrix was centered, and its spectral norm was computed.

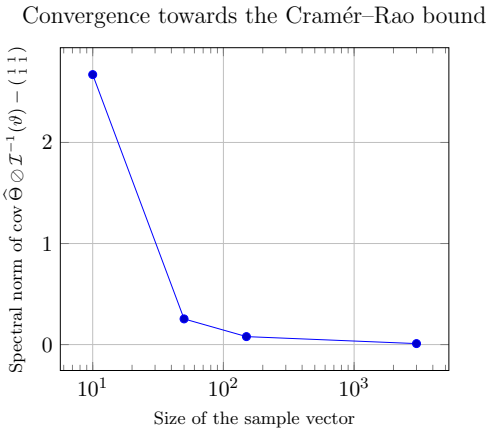



Figure 2. Using an adequately-centered spectral norm, one can visualize to which extent the Cramér–Rao lower bound is reached. The symbol "$\oslash$" denotes Hadamard division [17].

## Part 2. Particle filtering

### 8. Introduction

For the second part of the project, the task is to use noisy observations of the position of a fish in order to compute the parameters of the Gamma distribution governing its speed. With these parameters, a simulation is then completed which models the reaction and movement of the fish to a nearby predator. More precisely, a sequential Monte Carlo method is used to extract the real positions from the noisy measurements. In order to do this, one can play with various parameters, such as the number of fish, the number of particles per fish in the simulation, the time step between observations or the standard deviation of the noise which is added to the position to obtain the observations.

### 9. Estimation of the parameters from noisy measurements

Recall that the speed of the fish follows a Gamma distribution ($V_i \sim \Gamma(k, s)$). Here, one gets noisy observations of the position of a fish at given instants in time. It is rather straightforward to transform a noisy observation of the speed from these positions measurement, knowing the sampling period $t_s$. Once we have extracted a noisy approximation of the speed from this set of data, we can

[3]If the matrix is not centered so as to make its asymptotic value equal to the zero matrix, then the norm does not suffice to prove convergence to a matrix of ones.

[4]For this paper, the spectral norm is used, but any norm would give similar results.

repeat what we explained earlier in the document, in Section 5, to estimate the parameters of the distribution.

We begin by defining the formula for the speed more rigorously: from $x_p(i)$, the noisy measurement of the position of fish $p$ at step $i$, one can get $v_p(i)$, the noisy observation of the speed of that fish at that step, by

$$\boldsymbol{v}_p(i) = \frac{\left\|\boldsymbol{x}_p(i+1) + \boldsymbol{n}_p(i+1) - \big(\boldsymbol{x}_p(i) + \boldsymbol{n}_p(i)\big)\right\|}{t_\mathrm{s}}\,, \quad \forall p\,, \tag{9.1}$$

with $\boldsymbol{n}_p(i) \sim \mathcal{N}(0, \sigma^2_{\mathrm{obs}})$ an additive white noise. Using the estimation techniques of Section 4, one can get estimators of the two parameters of the Gamma distribution of the speed, $\hat{k}$ and $\hat{s}$, based on the observed speeds. The values found with this method are $\hat{k} = 3.9541$ and $\hat{s} = 0.3028$.

We can the that the zero-mean white noise has no effect on our estimators. Indeed, as shown in (9.1), the white-noise from time $i+1$ is "cancelled" with the white-noise from time $i$. We hence subtract the bias (which follows the same distribution) of two biased measurements of the position, and it is then "cancelled".

## 10. **Sequential Monte Carlo method**

We now would like to track the position of the fish and the predator, based on noisy measurements generated by `GenerateObservations`. In this case, one must exlude a linear evolution, hence using a Kalman filter would give bad results; we can however use a sequential Monte Carlo (SMC) method. At each time step, we must generate a large number of samples from the posterior distribution, and use those samples to generate the next state, as explained in [7, 9]. The SMC method does not require us to compute those samples at each time step. The following notations will be used: $f(x_t \mid y_1, \dots, y_t)$ is the posterior density function at time $t$, and $x_t^i$ is a sample from this posterior density. We also define the sample set $\mathcal{S}_t = \{x_t^i\}$, for $i = 1, 2, \dots, N_\mathrm{p}$. Our goal is to use the observations in $\mathcal{S}_t$ and the noisy measurements of the position, $y_{t+1}^p$ to compute the sample set $\mathcal{S}_{t+1}$. After doing so, we first create the new sample set then correct its values using the noisy measurements.

*Step* 1 (Initialization). When one first calls the function, the initial sample set $\mathcal{S}_1$ is empty. It is then filled using noisy observations $y_1^p$, together with an additive zero-mean white-noise for each particle. This intial guess allows us to start from the only approximation of the position we know (that is, its observed value). To compute the initial orientation of the $p$th fish, we use again the noisy measurements from $y_2^p$ and $y_1^p$. Throughout the computation, we keep in memory all the sample sets, for each time step, and for each fish and predator. Since we work in a subset of $\mathbb{R}^2$, each sample $x_t^i$ is a 2-dimensional vector, that is: $x_t^i = \begin{pmatrix}\xi & \upsilon\end{pmatrix}^T$.

*Step* 2 (Prediction). To predict the samples at step $t+1$, we can use the posterior density samples at step $t$, using the prediction density $f(x_{t+1} \mid y_1, \dots, y_t)$. We can compute the prediction set $\tilde{\mathcal{S}}_{t+1}$ of samples $\tilde{x}_t^i$. One can show that these last samples are samples from $f(x_{t+1} \mid y_1, \dots, y_t)$. [8]

In our case, we use the `StateUpdate` function to generate the samples at time $t+1$, which gives the state model update of the position.

*Step* 3 (Correction). We use the prediction set computed earlier, $\tilde{\mathcal{S}}_{t+1}$, to generate posterior samples from the posterior density function. The goal here is to compute weights $\tilde{w}_{t+1}^i$ for each particle of this set. Those weights allow to give more importance to particles which have a higher probability to be close to the true value of the position at time $t+1$. Hence, the particles which are further from the true value will have a lower weight. These weights are then normalized, which gives $\tilde{w}_{t+1}^i$. One can show that if we resample from the set $\tilde{\mathcal{S}}_{t+1}$ with probabilities given by

$\{\tilde{w}_{t+1}^1, \ldots, \tilde{w}_{t+1}^{N_\mathrm{p}}\}$, then the resulting values are samples from the posterior. In this case, because the system is very well-modeled, one should be wary of using weights blindly. In order to avoid having an all-zero weight vector due to floating point inaccuracies, one should check for such errors and if needed, replace all weights with the same value, say one. A more intelligent technique for this is explained in Section 12. Generating $N_\mathrm{p}$ resamples, which we call $x_{t+1}^i$, gives the sample set

$$\mathcal{S}_{t+1} = \{x_{t+1}^1, x_{t+1}^2, \ldots, x_{t+1}^{N_\mathrm{p}}\}\,, \tag{10.1}$$

which is the corrected sample set at time $t+1$.

*Step* 4 (Termination). For each time step, we have computed the posterior sample set of the position for each fish and for the predator. However, the true aim of the particle filter is to give a valid approximation of the noiseless value of the position of each fish. One can get these approximations, for time $t$ and fish $p$, by computing the sample mean of the sample set $\mathcal{S}_t$, relative to the $p$th fish.

## 11. **Mean squared error analysis**

Using the particle filter of Section 10, one can perform several experiments in order to determine its robustness and dependance on various parameters of the model. In order to quantify this, we use the mean square error, defined as

$$E_\mathrm{MSE} = \frac{1}{NP}\sum_{p=1}^{P}\sum_{i=1}^{N} \|\boldsymbol{x}_p(i) - \hat{\boldsymbol{x}}_p(i)\|_2^2\,. \tag{11.1}$$

Experiments were conducted for different values of three parameters. The default values were $w = 20$, $P = 3$, $N = 100$, $N_\mathrm{p} = 100$, $t_\mathrm{s} = 0.1$ and $\sigma_\mathrm{obs} = 0.2$. All experiments were repeated $M = 15$ times, in order to get a decent amount of data.

First, the effect of varying the number of particles per fish was studied. Simulations were run with $N_\mathrm{p} = 1, 2, 5, 10, 20, 50, 100, 150, 200, 250$. The results are shown on Figure 3. On this figure,

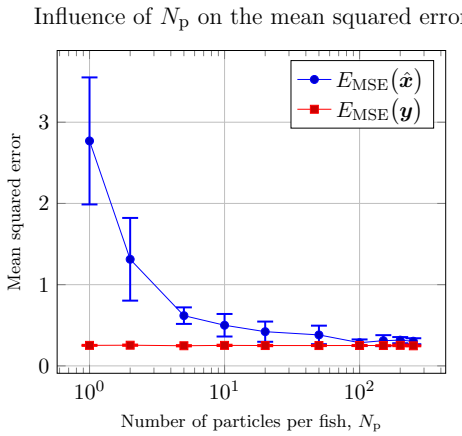



Figure 3. Mean squared error resulting from the filtered position estimates, for varying values of $N_\mathrm{p}$. Values were chosen so as to make sure the asymptotic decrease in mean squared error is clearly visible.

we see that the mean squared error decreases as the number of particles gets bigger. The filter also becomes more confident about the spread of the value, as seen by the increasingly shorter bars. This is what one would expect, when thinking about the underlying mechanism at work: an increase in the number of particles means a higher chance of one of those particles having a high likelihood

function, hence being a good estimate for the position. This can also explain the lower standard deviation: more particles are likely to have a high likelihood. We can also note that the value seems to converge towards a value of approximately 0.3. Another interesting observation is the fact that the value found by simply returning the observed positions, that is, the $y$ vector, one gets a better result than with the filter.

Next, we studied the influence of the sampling period $t_s$. Simulations were run with $t_s = 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5$. The results are shown on Figure 4. As one would expect, lowering the

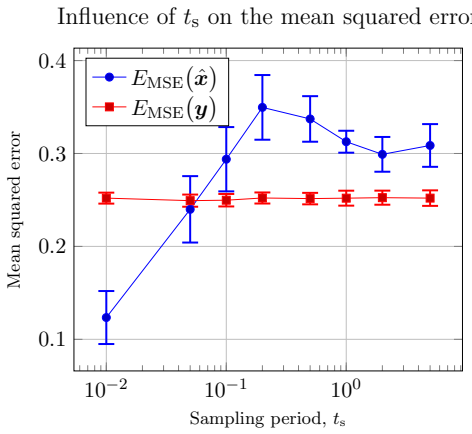



Figure 4. Mean squared error resulting from the filtered position estimates, for varying values of $t_s$. Values were chosen so as to both illustrate the strong increase in performance as the sampling period goes down, while also showing the apparent randomness once it reaches "ridiculously high" values, on the order of multiple seconds.

sampling period leads to more accurate results. To visualize why this must be true, one can simply imagine the probability density function's shape: as the sampling period goes down, the number of positions that can be reached within that duration also goes down. This means that the shape tends to that of a Dirac distribution, i.e. a deterministic value, as $t_s$ goes to zero. However, one must be careful to use enough particles: if the probable positions are very close, there is a non-negligible chance of not having any particles fall in the region, leading to bad results. For very large values of $t_s$, e.g. multiple seconds, we observe an almost constant evolution of the mean squared error. This also makes sense, with a similar argument to the one used to justify the good performance for small sampling periods. As the sampling period goes up, the range of distances that could be reached also widens. This means that a large range of positions are almost equiprobable, meaning the random number generator decides, for a big part, which estimates are chosen. This analysis was tested by running the same simulation multiple times, with the expected results.

Finally, the last parameter whose influence was examined is the standard deviation on the observed values, $\sigma_{obs}$. Simulations were run with $\sigma_{obs} = 0.001, 0.01, 0.1, 0.2, 0.5, 1, 2, 5$. The results are shown on Figure 5. With this last plot, we can observe that when the added observation noise function tends to a centered Dirac distribution, the filtered estimates, as well as the observations, are almost perfectly equal to the true value of the position, with a very small spread. However, when the noise's standard deviation is increased, the spread increases and the mean squared error increases. To explain this phenomenon, one has to remember that the filter will use this value of the standard deviation to place its particles, and that a very low variance means these particles will all be clustered around the true value. When the variance goes up, the filter is much better than simply taking the raw observations, as it eliminates part of the noise by using the Monte Carlo technique.

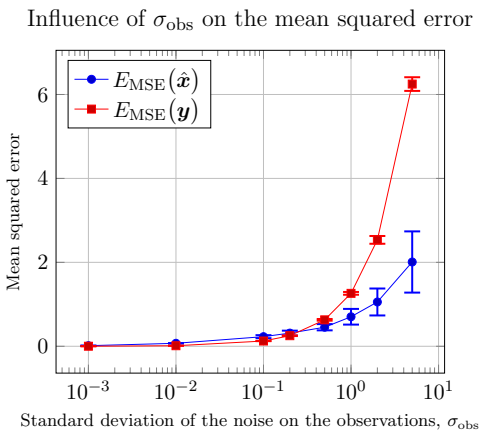



Figure 5. Mean squared error resulting from the filtered position estimates, for varying values of $\sigma_{\text{obs}}$. Values were chosen so as to visualize the expected low value near $\sigma_{\text{obs}} = 0$ and the gain over simply using observations as estimates when the standard deviation increases.

## 12. **Theoretical improvements for the filter**

As we observed in the previous section, the filter is by no means perfect already. Some simple ways to improve its performance exist. We give a (non-exhaustive) list:

- Simulations are only as good as the data/observations they work with, hence if possible, computing some dimensions of the filter analytically should always be preferred. This will not only provide a considerable improvement in the precision of the results, but will also drastically reduce the computation time. In the literature, this is referred to as Rao–Blackwellization. [1]
- When working with increasingly narrow density functions, one needs to be careful to use a sufficient amount of particles. If not, there is a non-negligible chance that no particles will fall in the probable region. This is the same phenomenon as the one causing issues with weighted sampling in Section 10. To handle this better, one could compute weights using the $c$th power of the likelihood returned by the observation model, with $0 < c < 1$. A good value of $c$ in most cases is $\frac{1}{M}$, where $M$ is the dimensionality of the observation model (i.e. $M = 2$, in this case). Other solutions include using a low-pass filter, or convoluting with a Gaussian, in order to "blur" the sharpness of the likelihood. [2]
- Repeated resampling in the absence of any actual sensory observations could lead to loss of diversity. In order to avoid this, one should only resample when necessary (for example, if some weights are too far apart), and if one does decide to resample, use low variance resampling. [10]

## 13. **Conclusion**

In Section 2, we looked at maximum likelihood estimation for the scale parameter of a Gamma distribution, $s$. We found the result to be $\hat{s}_{\text{ML}} = \frac{\overline{V}}{k}$ in (2.8). We proved some asymptotic properties of this estimator in Section 3, showing that it is asymptotically unbiased, efficient, best asymptotically normal and consistent. After establishing this, in Section 4, we looked into estimating both parameters of the Gamma distribution jointly, using our previous estimator. This led us to find that $\hat{k}_{\text{ML}}$ is the solution of (4.5). This estimator has no closed form, hence in Section 5 we turned to numerical methods to estimate it; comparing the maximum likelihood estimators with the method

of moments estimators showed that both are unbiased (for both parameters), with the former generally having lower variance, while the latter can be useful as an initial guess for the maximum likelihood estimator. Section 6, together with Appendix C, details an analytical derivation of the bound given by the Cramér–Rao inequality of Theorem A.1. The simulation in Section 7 showed empirical evidence of the efficiency of the maximum likelihood estimator.

In Section 9, we established how to extract speed measurements from noisy observations of a fish's position, then applied the estimation techniques defined previously to parameterize this speed as a gamma-distributed random variable. Section 10 explains the theoretical base for particle filters, and links this to our practical implementation for this particular project. Using this particular implementation, we were able to provide a simulation using both Matlab and Julia, whose results are summarized, interpreted and discussed in Section 11. Finally, in Section 12, we reflected on the shortcomings of the current implementation as pointed out in the previous section, and suggested an improved version of the SMC method.

## Appendix A. Definitions of properties

*Definition* A.1 (Unbiased estimator). The Fisher estimator $\widehat{\Theta} = g(Z)$ of $\vartheta$ is *unbiased* if

$$m_{\widehat{\Theta};\vartheta} \coloneqq \mathbb{E}\left[g(Z);\vartheta\right] = \vartheta\,, \quad \text{for all } \vartheta\,. \tag{A.1}$$

Theorem A.1 (Cramér–Rao inequality). *If $Z = (Z_1, \dots, Z_N)^T$ with i.i.d. random variables $Z_k$ and if its probability density function given by $f_Z(z;\vartheta) = \prod_{k=1}^{N} f_{Z_k}(z_k;\vartheta)$ satisfies certain regularity conditions, then the covariance of any unbiased estimator $\widehat{\Theta}$ satisfies the* Cramér–Rao inequality

$$\operatorname{cov}\widehat{\Theta} \succeq \mathcal{I}^{-1}(\vartheta)\,, \tag{A.2}$$

*where $\mathcal{I}(\vartheta)$ is the $p \times p$* Fisher information matrix, *defined by*

$$\left[\mathcal{I}(\vartheta)\right]_{i,j} \coloneqq -\mathbb{E}\left[\frac{\partial^2 \ln f_Z(z;\vartheta)}{\partial\vartheta_i \partial\vartheta_j}\right]\,. \tag{A.3}$$

*Definition* A.2 (Efficient estimator). An unbiased estimator is said to be *efficient* if it reaches the Cramér–Rao bound for all values of $\vartheta$, that is,

$$\operatorname{cov}\widehat{\Theta} = \mathcal{I}^{-1}(\vartheta)\,, \quad \forall\vartheta\,. \tag{A.4}$$

*Definition* A.3 (Best asymptotically normal estimator). A sequence of consistent estimators $\{\widehat{\Theta}_N(Z)\}_{N\in\mathbb{N}}$ of $\vartheta$ is called *best asymptotically normal* if

$$\sqrt{N}\left(\widehat{\Theta}_N(Z) - \vartheta\right) \xrightarrow[N\to+\infty]{\mathcal{D}} \mathcal{N}(0,\Sigma)\,, \tag{A.5}$$

for some minimal positive definite matrix $\Sigma$.

*Definition* A.4 (Consistent estimator). A sequence $\{\widehat{\Theta}_N(Z)\}_{N\in\mathbb{N}}$ of estimators of $\vartheta$ is called *consistent* if

$$\operatorname*{plim}_{N\to+\infty} \widehat{\Theta}_N(Z) = \vartheta\,. \tag{A.6}$$

Theorem 9.1 of [12] proves that this is equivalent to the statement that the MSE of the estimator converges to zero as $N$ goes to infinity.

## Appendix B. Omitted proofs

The following is a proof of asymptotic normality for the estimator in (2.8), using the Lindeberg–Lévy formulation of the central limit theorem.

*Proof.* We want to prove that $\frac{1}{\sqrt{N}}\ell'(\vartheta_0) \sim \mathcal{N}\big(0, \mathcal{I}(\vartheta_0)\big)$. We simplify notation by writing $f(v_i)$ instead of $f_{V_i}(v_i;k,\vartheta)$ and using Euler's notation for partial derivatives [20]. First, we show that

the expected value of $\frac{1}{\sqrt{N}}\ell'(\vartheta_0)$ is zero.

$$\mathbb{E}\left[\frac{\ell'(\vartheta_0)}{\sqrt{N}}\right] = \int_{-\infty}^{+\infty} \partial_{\vartheta_0}\left(\sum_{i=1}^{N}\frac{\ln f(v_i)}{\sqrt{N}}\right) f(v_i)\,\mathrm{d}v_i \tag{B.1}$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\frac{\partial_\vartheta f(v_i)}{f(v_i)} f(v_i)\,\mathrm{d}v_i \tag{B.2}$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\frac{\partial_\vartheta f(v_i)}{f(v_i)} f(v_i)\,\mathrm{d}v_i \tag{B.3}$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\partial_\vartheta f(v_i)\,\mathrm{d}v_i \tag{B.4}$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\partial_\vartheta\int_{-\infty}^{+\infty} f(v_i)\,\mathrm{d}v_i \tag{B.5}$$

$$= 0\,. \tag{B.6}$$

Next, we compute the variance of $\frac{1}{N}\ell'(\vartheta_0)$ (for $i = 1, \ldots, N$)

$$\mathbb{E}\left[\left(\partial_{\vartheta_0}\ln f(v_i)\right)^2\right] = \int_{-\infty}^{+\infty}\partial_{\vartheta_0}\ln f(v_i)\frac{\partial_{\vartheta_0} f(v_i)}{f(v_i)} f(v_i)\,\mathrm{d}v_i \tag{B.7}$$

$$= \int_{-\infty}^{+\infty}\partial_{\vartheta_0}\ln f(v_i)\partial_{\vartheta_0} f(v_i)\,\mathrm{d}v_i\,. \tag{B.8}$$

Using the product rule, we can then find

$$= -\int_{-\infty}^{+\infty}\partial_{\vartheta_0\vartheta_0}\ln f(v_i) f(v_i)\,\mathrm{d}v_i + \int_{-\infty}^{+\infty}\partial_{\vartheta_0}\left(\partial_{\vartheta_0}\ln f(v_i) f(v_i)\right)\mathrm{d}v_i \tag{B.9}$$

$$= -\mathbb{E}\left[\partial_{\vartheta_0\vartheta_0}\ln f(v_i)\right] + \partial_{\vartheta_0}\int_{-\infty}^{\infty}\frac{\partial_{\vartheta_0} f(v_i)}{f(v_i)} f(v_i)\,\mathrm{d}v_i \tag{B.10}$$

$$= \mathcal{I}(\vartheta)\,, \tag{B.11}$$

where the last expression can be shown to be zero by a similar argument as the one used above for the expected value. Knowing this, one easily finds

$$\mathbb{V}\left[\frac{\ell'(\vartheta_0)}{\sqrt{N}}\right] = \frac{1}{N}\mathbb{V}\left[\sum_{i=1}^{N}\partial_{\vartheta_0}\ln f(v_i)\right] = \mathcal{I}(\vartheta_0)\,, \tag{B.12}$$

since the random variables are i.i.d.. Using the Lindeberg–Lévy CLT, we thus have

$$\frac{\ell'(\vartheta_0)}{\sqrt{N}} \sim \mathcal{N}\big(0, \mathcal{I}(\vartheta_0)\big)\,. \tag{B.13}$$

□

## Appendix C. Computation of the Fisher information matrix

In order to derive an analytical expression for the Fisher information matrix, we start by computing various derivatives of the log-likelihood which will be needed later (writing $\ell(v; k, s)$

instead of $\ell(v_1, \dots, v_N; k, s)$):

$$\frac{\partial \ell(v;k,s)}{\partial s} = \frac{1}{s^2}\sum_{i=1}^{N} v_i - \frac{Nk}{s}\,, \tag{C.1}$$

$$\frac{\partial \ell(v;k,s)}{\partial k} = \sum_{i=1}^{N} \ln v_i - N\ln s - N\psi_0(k)\,, \tag{C.2}$$

$$\frac{\partial^2 \ell(v;k,s)}{\partial s^2} = -\frac{2}{s^3}\sum_{i=1}^{N} v_i + \frac{Nk}{s^2}\,, \tag{C.3}$$

$$\frac{\partial^2 \ell(v;k,s)}{\partial k^2} = -N\psi_1(k)\,, \tag{C.4}$$

$$\frac{\partial^2 \ell(v;k,s)}{\partial k \partial s} = \frac{\partial^2 \ell(v;k,s)}{\partial s \partial k} = -\frac{N}{s}\,. \tag{C.5}$$

In order to find the entries of the Fisher matrix, one must take the expectation of the previous derivatives:

$$\Big[\mathcal{I}(\vartheta)\Big]_{0,0} = -\mathbb{E}\left[\frac{\partial^2 \ell(v;k,s)}{\partial s^2}\right] = -\mathbb{E}\left[-\frac{2}{s^3}\sum_{i=1}^{N} v_i + \frac{Nk}{s^2}\right] \tag{C.6}$$

$$= \frac{2}{s^3}Nks - \frac{Nk}{s^2} = \frac{Nk}{s^2}\,, \tag{C.7}$$

$$\Big[\mathcal{I}(\vartheta)\Big]_{0,1} = \Big[\mathcal{I}(\vartheta)\Big]_{1,0} = -\mathbb{E}\left[\frac{\partial^2 \ell(v;k,s)}{\partial k \partial s}\right] = \frac{N}{s}\,, \tag{C.8}$$

$$\Big[\mathcal{I}(\vartheta)\Big]_{1,1} = -\mathbb{E}\left[\frac{\partial^2 \ell(v;k,s)}{\partial k^2}\right] = -N\psi_1(k)\,. \tag{C.9}$$

## References


[1] A. Doucet, Rao–Blackwellised Particle Filtering for Dynamic Bayesian Networks, *Uncertainty in Artificial Intelligence Proceedings*, 176–183.

[2] Drew Bagnell, Statistical techniques in robotics. Available at http://www.cs.cmu.edu/~16831-f14/notes/F11/16831_lecture04_tianyul.pdf.

[3] I. Dunning, J. Huchette, and M. Lubin, JuMP: A Modeling Language for Mathematical Optimization, *SIAM Review* **59** no. 2 (2017), 295–320. http://dx.doi.org/10.1137/15M1020575.

[4] Eric W. Weisstein, Digamma Function, From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/DigammaFunction.html, [Online; accessed 9-April-2019].

[5] Eric W. Weisstein, Trigamma Function, From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/TrigammaFunction.html, [Online; accessed 10-April-2019].

[6] A. Huth and C. Wissel, The Simulation of the Movement of Fish Schools, *Journal of Theoretical Biology* **156** no. 1 (1992), 365–385.

[7] J. S. Liu, *Monte Carlo strategies in Scientific Computing*, Springer.

[8] M. J. Maron, *Numerical Analysis*: *A Practical Approach*, Macmillan Publishing Company, 1982.

[9] A. Srivastava, Computational Methods in Statistics, 2009.

[10] C. Stachniss, *Robotic Mapping and Exploration*, Springer.

[11] Thomas P. Minka, Estimating a Gamma distribution, https://tminka.github.io/papers/minka-gamma.pdf, 2002, [Online; accessed 6-April-2019].

[12] D. D. Wackerly, W. Mendenhall iii, and R. L. Scheaffer, *Mathematical Statistics with Applications*, 7 ed., Brooks/Cole, Cengage Learning, 20 Davis Drive, Belmont, CA 94002, USA, 2008.

[13] Wikipedia contributors, Sample mean and covariance — Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Sample_mean_and_covariance#Sample_covariance, 2018, [Online; accessed 6-April-2019].

[14] Wikipedia contributors, Central limit theorem — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Central_limit_theorem, 2019, [Online; accessed 12-April-2019].
[15] Wikipedia contributors, Characteristic function (probability theory) — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Characteristic_function_(probability_theory), 2019, [Online; accessed 12-April-2019].
[16] Wikipedia contributors, Gamma distribution — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Gamma_distribution, 2019, [Online; accessed 11-April-2019].
[17] Wikipedia contributors, Hadamard product (matrices) — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Hadamard_product_(matrices)#Analogous_operations, 2019, [Online; accessed 17-April-2019].
[18] Wikipedia contributors, Matrix norm — Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Matrix_norm#Spectral_norm, 2019, [Online; accessed 17-April-2019].
[19] Wikipedia contributors, Maximum likelihood estimation — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Maximum_likelihood_estimation, 2019, [Online; accessed 8-April-2019].
[20] Wikipedia contributors, Notation for differentiation — Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Notation_for_differentiation#Euler's_notation, 2019, [Online; accessed 12-April-2019].

Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
*E-mail*: navarre.louis@student.uclouvain.be

Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
*E-mail*: gilles.peiffer@student.uclouvain.be