SDS 383C: Homework 1

G. Paulon

September 19, 2016

Problem 1. Convergence of random variables

Assume $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(\mu, \sigma, \mu_3, \mu_4)$. The sample variance is $S_n^2 = \frac{\sum_i (X_i - \bar{X}_n)^2}{n-1}$. Prove the following statements.

(a) $S_n^2 \xrightarrow{P} \sigma^2$, that is, the WLLN for S_n^2.

$$\begin{aligned} S_n^2 =& \frac{\sum_i (X_i - \bar{X}_n)^2}{n-1} = \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 + \bar{X}_n^2 - \frac{2}{n}\bar{X}_n \sum_i X_i\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 - \bar{X}_n^2\right) = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \frac{n}{n-1}\bar{X}_n^2. \end{aligned} \tag{1}$$

The second term of the sum converges in probability to μ^2. In fact, for the WLLN, $\bar{X}_n \xrightarrow{P} \mu$ and, for the continuous mapping theorem $\bar{X}_n^2 \xrightarrow{P} \mu^2$.

Moreover, as far as the first term is concerned, it is easy to prove that $\frac{\sum_i X_i^2}{n} \xrightarrow{P} \mu_2$. In fact,

$$E\left[\frac{\sum_i X_i^2}{n}\right] = \frac{1}{n}\sum_i E[X_i^2] = \mu_2,$$

and

$$\begin{aligned} \text{Var}\left[\frac{\sum_i X_i^2}{n}\right] &= \frac{1}{n^2}\sum_i \text{Var}(X_i^2) = \frac{1}{n^2}\sum_i [EX_i^4 - (EX_i^2)^2] \\ &= \frac{1}{n}(\mu_4 - \mu_2^2). \end{aligned}$$

Therefore, using the definition of convergence in probability,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_i X_i^2 - \mu_2\right| < \epsilon\right) &= P\left(\left(\frac{1}{n}\sum_i X_i^2 - \mu_2\right)^2 < \epsilon^2\right) \\ &\leq \frac{E\left[\left(\frac{1}{n}\sum_i X_i^2 - \mu_2\right)^2\right]}{\epsilon^2} = \frac{\text{Var}\left(\frac{1}{n}\sum_i X_i^2\right)}{\epsilon^2} \\ &= \frac{\mu_4 - \mu_2^2}{n\epsilon^2} \xrightarrow[n\to+\infty]{} 0, \end{aligned}$$

i.e. $\frac{\sum_i X_i^2}{n} \xrightarrow{P} \mu_2$.

Thus, using (1) we get

$$S_n^2 = \frac{n}{n-1}\left(\frac{\sum_i X_i^2}{n} - \bar{X}_n^2\right) \xrightarrow{P} \mu_2 - \mu^2 = \sigma^2.$$

(b) $S_n \xrightarrow{P} \sigma$. The proof is straightforward: we can use the continuous mapping theorem with $g(x) = \sqrt{x}$ since the square root is a continuous function on $\mathbb{R}^+$.

(c) $\frac{\bar{X}_n}{S_n} \xrightarrow{P} \frac{\mu}{\sigma}$. The proof is straightforward: we can use the continuous mapping theorem with $g(u, v) = \frac{u}{v}$, which is continuous if $v \neq 0$ In fact, we already knew that $\bar{X}_n \xrightarrow{P} \mu$ (WLLN) and that $S_n \xrightarrow{P} \sigma$ (proved in (b)).

(d) $\sqrt{n}\frac{(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} \mathcal{N}(0, 1)$. We know that, for the Central Limit Theorem,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Moreover, we proved in part (b) that $S_n \xrightarrow{P} \sigma$, which implies that $S_n \xrightarrow{d} \sigma$, where σ is a constant. Therefore, using Slutsky's lemma, we conclude that

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} \frac{1}{\sigma}\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1).$$

Problem 2. Maximum Likelihood Estimates

(a) Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}([\theta, \theta + 1])$. We can compute the likelihood of the data

$$\begin{aligned} L(X_1, \dots, X_n; \theta) &= \prod_{i=1}^{n} P(X_i = x_i; \theta) \\ &= \prod_{i=1}^{n} \mathcal{I}_{[\theta;\theta+1]}(x_i) = \begin{cases} 1 & \text{if } \theta \leq x_i \leq \theta + 1 \quad \forall i = 1, \dots, n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore we can find an infinite number of maxima, i.e. of $\hat{\theta}$ that realize $L(X_1, \dots, X_n; \hat{\theta}) = 1$, by simply setting the two constraints

$$\begin{aligned} \hat{\theta} &\leq \min_{i=1,\dots,n}(x_i) \\ \hat{\theta} + 1 &\geq \max_{i=1,\dots,n}(x_i) \end{aligned}$$

that correspond to $\hat{\theta} \in [\max_{i=1,\dots,n}(x_i) - 1; \min_{i=1,\dots,n}(x_i)]$. The MLE exists but it is not unique.

(b) Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}([\theta, 1]), \theta \leq 1$.

i. We can compute the likelihood of the data

$$\begin{aligned} L(X_1, \dots, X_n; \theta) &= \prod_{i=1}^{n} P(X_i = x_i; \theta) \\ &= \prod_{i=1}^{n} \left\{ \frac{1}{1-\theta} \mathcal{I}_{[\theta;1]}(x_i) \right\} = \begin{cases} \frac{1}{(1-\theta)^n} & \text{if } \theta \leq x_i \leq 1 \quad \forall i = 1, \dots, n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Our goal is to maximize $\frac{1}{(1-\theta)^n}$ under the constraint $\theta \leq x_i \leq 1$. The minimum is reached for $\hat{\theta} = \min_{i=1,\dots,n}(x_i)$. Therefore the MLE exists and it is unique.

ii. To find the limit in distribution of $n(\hat{\theta}-\theta)$, we compute its cdf, i.e.

$$\begin{aligned}
P(n(\hat{\theta}-\theta) \leq t) &= P(\hat{\theta} \leq t/n + \theta) \\
&= P(\min\{x_1, \dots, x_n\} \leq t/n + \theta) \\
&= 1 - P(\min\{x_1, \dots, x_n\} > t/n + \theta) \\
&\overset{\text{ind.}}{=} 1 - \prod_{i=1}^{n} P(x_i > t/n + \theta) \\
&\overset{\text{i.d.}}{=} 1 - [P(x_i > t/n + \theta)]^n \\
&= 1 - [1 - P(x_i \leq t/n + \theta)]^n \\
&= 1 - \left[1 - \frac{t/n + \theta - \theta}{1-\theta}\right]^n \\
&= 1 - \left[1 - \frac{t}{n(1-\theta)}\right]^n \longrightarrow 1 - e^{-\frac{t}{1-\theta}},
\end{aligned}$$

which is the cdf of an exponential distribution. Therefore $n(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{E}\left(\frac{1}{1-\theta}\right)$.

iii. The MLE is not behaving the way we would expect. In fact, we know that, under mild assumptions of regularity, MLEs are asymptotically normal. However, in this case one of the hypothesis does not hold. In particular, the support of the pdf, i.e. the set $S = \{x : f(x;\theta) > 0\} = [\theta, 1]$ is not independent of θ.

(c) Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, and let $\theta := e^{\mu}$. Let, in the simulations, $\mu = 5$ and $n = 100$.

i. Use the delta method to get the variance of the estimator and a 95% confidence interval of θ.

Let us recall that, since $\bar{X}_n$ is the MLE for μ, then $e^{\bar{X}_n}$ is the MLE for e^{μ} (invariance principle). Moreover, we know (central limit theorem) that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{1} \xrightarrow{d} \mathcal{N}(0,1). \tag{2}$$

Actually, in the case when the observations are iid draws from the normal distribution, not only the asymptotic distribution but also the exact distribution of the sample mean is normal (linear combination of independent normal distributions). However, we here use the Delta method to the expression (2), obtaining

$$\frac{\sqrt{n}(e^{\bar{X}_n} - e^{\mu})}{e^{\mu}} \xrightarrow{d} \mathcal{N}(0,1),$$

which corresponds to

$$e^{\bar{X}_n} \xrightarrow{d} \mathcal{N}\left(e^{\mu}, \frac{e^{2\mu}}{n}\right).$$

We can use this asymptotic distribution to obtain a 95% confidence interval for the unknown quantity e^{μ}. In fact

$$\begin{aligned}
&P(|e^{\bar{X}_n} - e^{\mu}| \leq t) = 1-\alpha \\
\Rightarrow &P\left(\sqrt{n}\frac{|e^{\bar{X}_n} - e^{\mu}|}{e^{\mu}} \leq \frac{\sqrt{n}}{e^{\mu}}t\right) = 1-\alpha \\
\Rightarrow &P\left(|Z| \leq \frac{\sqrt{n}}{e^{\mu}}t\right) = 1-\alpha \\
\Rightarrow &\frac{\sqrt{n}}{e^{\mu}}t = z_{1-\alpha/2} \\
\Rightarrow &t = \frac{e^{\mu} z_{1-\alpha/2}}{\sqrt{n}} \approx \frac{e^{\bar{X}_n} z_{1-\alpha/2}}{\sqrt{n}}.
\end{aligned}$$

Therefore, the 95% CI for θ is

$$\theta \in \left[e^{\bar{X}_n} - \frac{e^{\bar{X}_n} z_{1-\alpha/2}}{\sqrt{n}}; e^{\bar{X}_n} + \frac{e^{\bar{X}_n} z_{1-\alpha/2}}{\sqrt{n}}\right].$$

ii. The same confidence interval can be approximated via Bootstrap. Recall, in fact, that this method allows us to obtain the sampling distribution of the estimator $\hat{\theta}$. At each bootstrap iteration, a new dataset is obtained via sampling with replacement from the original dataset. The estimate $e^{\bar{X}_n}$ is calculated at each iteration, yielding to a sample $\theta^* = \left(e^{\bar{X}_n}\right)^{(1)}, \dots, \left(e^{\bar{X}_n}\right)^{(B)}$. Then, in order to calculate the confidence interval we can consider the 2.5% and 97.5% quantiles of this sample, i.e. $C_n = \left[\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}\right]$. Other methods are plausible, e.g. the normal interval and the pivotal intervals, but here we take the quantile approach.

In Listing A.1 the code for the Bootstrap simulation is displayed. In Listing 1 the results are shown: as one can see, the CIs are very similar in the two cases.

```
                  2.5%    97.5%
Delta Method 130.6150 194.2965
Bootstrap    136.8269 193.8129
```

Listing 1: Comparison of the CIs using the Delta method and the Bootstrap.

Problem 3. Gradient ascent

We will use a iterative algorithm to calculate the MLE of the parameters of a Dirichlet distribution. The conjugate prior to the multinomial is the Dirichlet distribution on the k-simplex, whose density is given by:

$$f(x_1, \dots, x_k|\alpha) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k+1} x_i^{\alpha_i - 1}$$

where $x_i > 0$, $\sum_i^{k+1} x_i = 1$ and $\alpha_i \geq 0$.

(a) Prove that $E[\log(x_i)] = \Psi(\alpha_i) - \Psi(\sum_{i=1}^{k+1} \alpha_i)$, where $\Psi(\alpha) = d \log \Gamma(\alpha)/d\alpha$ is the digamma function. To to that, we are going to first prove that the marginals of a Dirichlet distribution are Beta distributions.

- *Proof of the neutrality property.* We can prove a more general property, called **neutrality** of the Dirichlet distribution, which states that, if $(X_1, \dots, X_{k+1}) \sim$ Dirichlet$(\alpha_1, \dots, \alpha_{k+1})$ then the following holds:

$$X_i \perp \left(\frac{X_1}{1-X_i}, \dots, \frac{X_{i-1}}{1-X_i}, \frac{X_{i+1}}{1-X_i}, \dots, \frac{X_{k+1}}{1-X_i} \right)$$
$$X_i \sim \text{Beta}(\alpha_i, \sum_{j \neq i}^{k+1} \alpha_j)$$
$$\left(\frac{X_1}{1-X_i}, \dots, \frac{X_{i-1}}{1-X_i}, \frac{X_{i+1}}{1-X_i}, \dots, \frac{X_{k+1}}{1-X_i} \right) \sim \text{Dirichlet}\left(\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_{k+1}\right).$$

In particular, this implies that each marginal of a Dirichlet distribution is a Beta distribution. To prove this property, we simply apply the pdf transformation theorem in the multivariate case. We consider the mapping

$$\boldsymbol{q} = g(\boldsymbol{x}) = \begin{cases} q_1 = \frac{x_1}{1-x_i} \\ \dots \\ q_{i-1} = \frac{x_{i-1}}{1-x_i} \\ q_i = x_i \\ q_{i+1} = \frac{x_{i+1}}{1-x_i} \\ \dots \\ q_k = \frac{x_k}{1-x_i} \end{cases} \quad \Rightarrow \quad \boldsymbol{x} = g^{-1}(\boldsymbol{q}) = \begin{cases} x_1 = q_1(1-q_i) \\ \dots \\ x_{i-1} = q_{i-1}(1-q_i) \\ x_i = q_i \\ x_{i+1} = q_{i+1}(1-q_i) \\ \dots \\ x_k = q_k(1-q_i). \end{cases}$$

The Jacobian matrix of this transformation is the $k \times k$ matrix

$$J = \begin{pmatrix} 1-q_i & \dots & 0 & -q_1 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & & 1-q_i & -q_{i-1} & \vdots & & \vdots \\ \vdots & & 0 & 1 & 0 & & \vdots \\ \vdots & & \vdots & -q_{i+1} & 1-q_i & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & -q_k & 0 & \dots & 1-q_i \end{pmatrix}$$

whose determinant is $||J|| = (1-q_i)^{k-1}$. Therefore,

$$
\begin{aligned}
f_Q(\boldsymbol{q}) &= f_X(g^{-1}(\boldsymbol{q})) \cdot ||J|| \\
&= \frac{\Gamma(\sum_{i=1}^{k+1}\alpha_i)}{\prod_{i=1}^{k+1}\Gamma(\alpha_i)} q_i^{\alpha_i - 1} \prod_{j\neq i}^{k} (q_j(1-q_i))^{\alpha_j - 1} \left[1 - q_i - \sum_{j\neq i}^{k}(1-q_i)q_j\right]^{\alpha_{k+1}-1} (1-q_i)^{k-1} \\
&= \frac{\Gamma(\sum_{i=1}^{k+1}\alpha_i)}{\prod_{i=1}^{k+1}\Gamma(\alpha_i)} q_i^{\alpha_i - 1}(1-q_i)^{\sum_{j\neq i}^{k}(\alpha_j - 1)} \prod_{j\neq i}^{k} q_j^{\alpha_j-1}(1-q_i)^{\alpha_{k+1}-1} \left[1 - \sum_{j\neq i}^{k} q_j\right]^{\alpha_{k+1}-1} (1-q_i)^{k-1} \\
&= \frac{\Gamma(\alpha_i + \sum_{j\neq i}^{k+1}\alpha_j)}{\Gamma(\alpha_i)\Gamma(\sum_{j\neq i}^{k+1}\alpha_j)} q_i^{\alpha_i-1}(1-q_i)^{\sum_{j\neq i}^{k+1}\alpha_j - 1} \\
&\quad \cdot \frac{\Gamma(\sum_{j\neq i}^{k+1}\alpha_j)}{\prod_{j\neq i}^{k+1}\Gamma(\alpha_j)} \prod_{j\neq i}^{k} q_j^{\alpha_j - 1}\left[1 - \sum_{j\neq i}^{k} q_j\right]^{\alpha_{k+1}-1},
\end{aligned}
$$

which is the thesis. In fact we can see that the density is factorized in two independent terms: the first one is the density Beta$(\alpha_i, \sum_{j\neq i}^{k+1}\alpha_j)$; the second one the density Dirichlet$(\boldsymbol{\alpha}^{(-i)})$, where $\boldsymbol{\alpha}^{(-i)} = (\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_{k+1})$. $\square$

- *Proof marginalizing out the variables.* As an alternative proof, we can just find any marginal distribution by integrating out all the other variables. In our case, given $(X_1, \dots, X_{k+1}) \sim$ Dirichlet$(\alpha_1, \dots, \alpha_{k+1})$, to find the marginal X_i we can firstly integrate out x_1, that is

$$
\begin{aligned}
\mathcal{L}(X_2, \dots, X_{k+1}) &= \int_{\mathcal{X}_1} \mathcal{L}(X_1, \dots, X_{k+1}) dx_1 \\
&= \int_0^{1-\sum_{j=2}^{k}x_j} \frac{\Gamma(\alpha_1 + \dots + \alpha_{k+1})}{\Gamma(\alpha_1)\dots\Gamma(\alpha_{k+1})} x_1^{\alpha_1 - 1} \dots x_k^{\alpha_k - 1}(1 - \sum_{j=1}^{k} x_j)^{\alpha_{k+1}-1} dx_1 \\
&= \frac{\Gamma(\alpha_1 + \dots + \alpha_{k+1})}{\Gamma(\alpha_1)\dots\Gamma(\alpha_{k+1})} x_2^{\alpha_2 - 1} \dots x_k^{\alpha_k - 1} \int_0^{1-\sum_{j=2}^{k}x_j} x_1^{\alpha_1-1}(1 - x_1 - \sum_{j=2}^{k} x_j)^{\alpha_{k+1}-1} dx_1.
\end{aligned}
$$

Remark that, for the sake of simplicity, we are omitting the indicator function over the k-dimensional simplex. Using the substitution $u = \frac{x_1}{1-\sum_{j=2}^{k}x_j}$, which corresponds to normalizing the argument of the integral, we obtain

$$\begin{aligned}\mathcal{L}(X_2,\dots,X_{k+1}) =& \frac{\Gamma(\sum_{j=1}^{k+1}\alpha_j)}{\prod_{j=1}^{k+1}\Gamma(\alpha_j)} x_2^{\alpha_2-1}\dots x_k^{\alpha_k-1} \\ & \cdot \int_0^1 u^{\alpha_1-1}(1-\sum_{j=2}^{k}x_j)^{\alpha_1-1}[(1-u)(1-\sum_{j=2}^{k}x_j)]^{\alpha_{k+1}-1}(1-\sum_{j=2}^{k}x_j)du \\ =& \frac{\Gamma(\sum_{j=1}^{k+1}\alpha_j)}{\prod_{j=1}^{k+1}\Gamma(\alpha_j)} x_2^{\alpha_2-1}\dots x_k^{\alpha_k-1}(1-\sum_{j=2}^{k}x_j)^{\alpha_1+\alpha_{k+1}-1}\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_{k+1}-1}du \\ =& \frac{\Gamma(\sum_{j=1}^{k+1}\alpha_j)}{\prod_{j=1}^{k+1}\Gamma(\alpha_j)} x_2^{\alpha_2-1}\dots x_k^{\alpha_k-1}(1-\sum_{j=2}^{k}x_j)^{\alpha_1+\alpha_{k+1}-1}\frac{\Gamma(\alpha_1)\Gamma(\alpha_{k+1})}{\Gamma(\alpha_1+\alpha_{k+1})} \\ =& \frac{\Gamma(\alpha_2+\dots+(\alpha_1+\alpha_{k+1}))}{\Gamma(\alpha_2)\dots\Gamma(\alpha_k)\Gamma(\alpha_1+\alpha_{k+1})} x_2^{\alpha_2-1}\dots x_k^{\alpha_k-1}(1-\sum_{j=2}^{k}x_j)^{\alpha_1+\alpha_{k+1}-1}\end{aligned}$$

and we recognize that $(X_2,\dots,X_k,X_{k+1}) \sim$ Dirichlet$(\alpha_2,\dots,\alpha_k,\alpha_1+\alpha_{k+1})$. Integrating iteratively over all the other variables but x_i, we get

$$(X_i, X_{k+1}) \sim \text{Dirichlet}(\alpha_i, \sum_{j\neq i}^{k+1}\alpha_j),$$

which is equivalent to $X_i \sim$ Beta$(\alpha_i, \sum_{j\neq i}^{k+1}\alpha_j)$. □

Now that we proved that $X_i \sim$ Beta$(\alpha_i, \sum_{j\neq i}^{k+1}\alpha_j)$, using the hint we conclude

$$E[\log(X_i)] = \Psi(\alpha_i) - \Psi\left(\sum_{j=1}^{k+1}\alpha_j\right).$$

(b) Let us suppose now we have n data points $\{x^{(i)}, i=1,\dots,n\}$ generated from $f(x|\alpha)$. The log-likelihood of the i-th data point is then

$$l(x^{(i)};\alpha) = \log\Gamma(\sum_{j=1}^{k+1}\alpha_j) - \sum_{j=1}^{k+1}(\log\Gamma(\alpha_j)) + \sum_{j=1}^{k+1}(\alpha_j-1)\log x_j^{(i)}.$$

By deriving the total log-likelihood of the sample we obtain

$$\begin{aligned}\frac{\partial\sum_{i=1}^{n}l(x^{(i)};\alpha)}{\partial\alpha_h} &= \sum_{i=1}^{n}\left\{\Psi(\sum_{j=1}^{k+1}\alpha_j) - \Psi(\alpha_h) + \log x_h^{(i)}\right\} \\ &= N\Psi(\sum_{j=1}^{k+1}\alpha_j) - N\Psi(\alpha_h) + N\overline{\log x_h},\end{aligned}$$

and, by setting the expression above to be equal to 0, we get to the following equation for the MLE $\hat{\alpha}$

$$\Psi(\sum_{j=1}^{k+1}\hat{\alpha}_j) = \Psi(\hat{\alpha}_h) - \overline{\log x_h}, \tag{3}$$

where $\overline{\log x_h} = \frac{1}{n}\sum_{i=1}^{n} \log x_h^{(i)}$ is the average computed from data.

(c) From now on, we work with a dataset for a Dirichlet over the 2-dimensional simplex.

i. The scatterplot of the data is displayed in Figure 1.

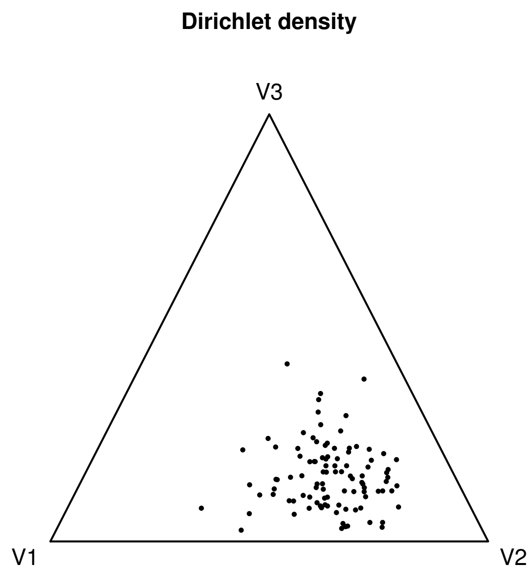

Figure 1: Scatterplot of the data represented on the simplex. Data points are represented in black.

ii. The log-likelihood is convex in α since the Dirichlet distribution is in the exponential family. Therefore, a simple algorithm can be obtained by setting the gradient equal to 0. A fixed-point iteration for maximizing the likelihood can be derived from (3), yielding

$$\Psi(\alpha_h^{new}) = \Psi(\sum_{j=1}^{k+1} \alpha_j^{old}) + \overline{\log x_h}.$$

In this iterative algorithm a convergence criterion has to be chosen. In our setting, we used the relative increment of the log-likelihood. In other terms, we defined the quantity $\delta = \frac{l(\alpha^{(k+1)}) - l(\alpha^{(k)})}{|l(\alpha^{(k)}) + \epsilon|}$ and the algorithm stops when δ is smaller than a certain threshold (in our case set to 10^{-10}). The constant $\epsilon = 10^{-3}$ is only needed in order to ensure numerical stability to the computation of the error at each step.

Running the gradient ascent method from the starting point $\alpha_0 = (1, 1, 1)$ with a tolerance error equal to 10^{-10} leads to convergence after 215 iterations. The result is the MLE for α, i.e. $\hat{\alpha} = (6.3878, 12.6291, 3.4034)$.

The log-likelihood as a function of iterations is shown in Figure 2.

iii. The scatter plot of the data together with a contour plot of the Dirichlet distribution with optimal parameters $\hat{\alpha}$ is displayed in Figure 3.

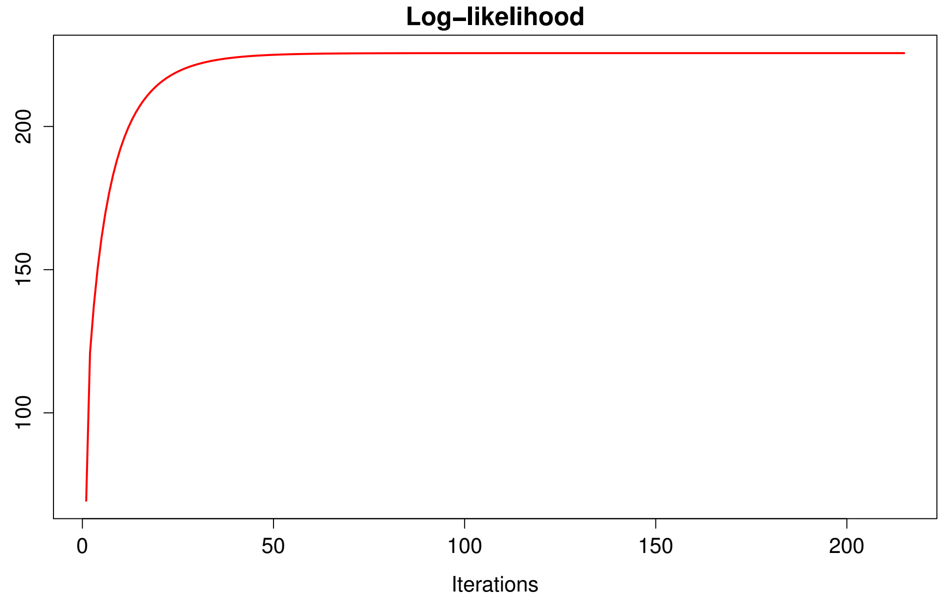

Figure 2: Plot of the log-likelihood as a function of iterations.

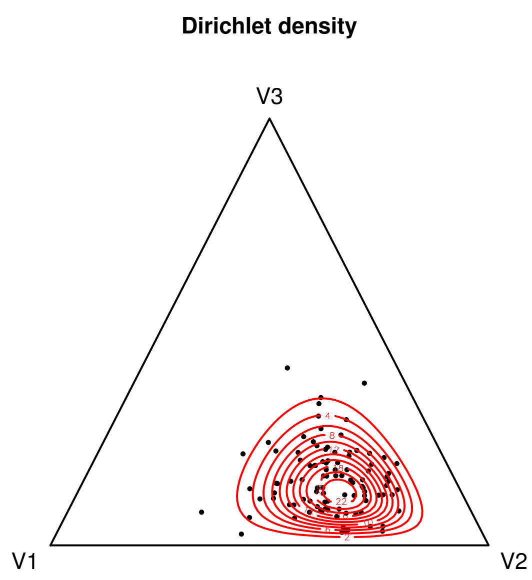

Figure 3: Scatterplot of the data represented on the simplex. Data points are represented in black; the contour plot of the density of the Dirichlet distribution with optimal parameters $\hat{\alpha}$ is overlapped in red.

Appendix A

R code

```
n <- 100
mu <- 5

set.seed(123)
# We generate the "first" dataset: X1, ..., Xn Normal (mu, 1)
X <- rnorm(n, mu, 1)

# We compute the MLE for the mean, that is, Xbar
Xbar <- mean(X)

# BOOTSTRAP
B <- 10000
mu_hat <- array(NA,dim=B)
for (b in 1:B){
  X_sampled <- sample(X, n, replace = T)
  mu_hat[b] <- mean(X_sampled)
}
theta_hat <- exp(mu_hat)
var(theta_hat)
exp(2*Xbar)/n

# Set the confidence level
alpha <- 0.05

# Bootstrap CI
quantile(exp(mu_hat), probs = c(alpha/2,1-alpha/2))

# Theoretical CI
c(exp(Xbar)*(1 - qnorm(1 - alpha/2)/sqrt(n)),exp(Xbar)*(1 + qnorm(1 - alpha/2)/sqrt(n)))
```

Listing A.1: Bootstrap simulation.

```
# Compute the log-likelihood of the Dirichlet distribution
log_lik <- function(alpha, y){
  N <- dim(y)[1]
  ll <- N*(lgamma(sum(alpha)) - sum(lgamma(alpha))) + sum((alpha-1)*t(log(y)))
  return(ll)
}

# Function for the gradient ascent of the Dirichlet model
```

```
gradient_ascent <- function(y, alpha0, maxiter, tol){

  alphas <- array(NA, dim=c(maxiter, length(alpha0)))
  alphas[1,] <- alpha0 # Initial guess
  ll <- array(NA, dim = maxiter)
  ll[1] <- log_lik(alphas[1,], y)

  for (iter in 2:maxiter){
    psialpha_new <- apply(log(y), 2, mean) + digamma(sum(alphas[iter-1,]))
    alphas[iter,] <- newton_roots(psialpha_new)
    ll[iter] <- log_lik(alphas[iter,], y)

    # Convergence check
    if (abs(ll[iter-1] - ll[iter])/abs(ll[iter-1] + 1E-3) < tol){
      cat('Algorithm has converged after', iter, 'iterations')
      ll <- ll[1:iter]
      alphas <- alphas[1:iter,]
      break;
    }
    else if (iter == maxiter & abs(ll[iter-1] - ll[iter])/abs(ll[iter-1] + 1E-3) >= tol){
      print('WARNING: algorithm has not converged')
      break;
    }
  }
  return(list("ll" = ll, "alpha" = alphas[iter,]))
}
```

Listing A.2: Gradient ascent.