CODE LENGTHS FOR MODEL CLASSES WITH CONTINUOUS UNIFORM DISTRIBUTIONS

Panu Luosto

University of Helsinki
Department of Computer Science
P.O. Box 68, FI-00014 UNIVERSITY OF HELSINKI, Finland
panu.luosto@cs.helsinki.fi

ABSTRACT

Continuous uniform distributions are an important means for modelling e.g. noise and unknown portions of heterogeneous data. Even if they are simplistic models, deriving the corresponding code lengths is sometimes non-trivial. One of the obvious problems is that the set in which a uniform density gets positive values is often not known in advance in practical applications. This paper treats uniform distributions in origin-centred balls, arbitrary balls and axis-aligned boxes. We derive normalized maximum likelihood (NML) densities for the cases when the maximum likelihood parameters of the data are bounded. From the NML densities we derive code length functions that depend on the prior densities of the parameters. We generalize Rissanen's prior for positive reals for this purpose. We also suggest methods for dealing with the problems that arise from the singularities in the final code length functions.

1. INTRODUCTION

Continuous uniform distributions are as simplistic models important in the field of minimum description length (MDL) [1, 2] principle based learning. When the domain of the data is known, the uniform distribution gives the shortest worst-case code. In this paper, we assume that the domain is unknown, which makes the situation non-trivial. Our objective is to derive code length functions that are suitable for noise and entirely unknown data, or that can be used as baseline code length functions for determining the efficiency of more sophisticated models. In the same time, we avoid unnecessary assumptions about the domain. We have used our code length for the model class with uniform distributions in axis-aligned boxes in [3] where the objective is to find the best clustering with an unknown number of normally distributed clusters and one uniform cluster.

If the bounded set in which the uniform distribution gets positive values is not known in advance, even choosing its geometrical form can be a difficult design choice. We consider origin-centred balls in any dimensionality and arbitrary balls in one and two dimensions. Uniform distributions in axis-aligned boxes are simply product densities of uniform distributions in arbitrary one-dimensional balls. If the ranges of the parameters are bounded, calculating normalized maximum likelihoods (NML) [4] is straightforward for all the models mentioned above. If the parameters are unbounded, the NML density is not defined.

We derive our code lengths with unbounded parameters in all the cases according to a similar idea. To outline the method, we take as an example the simplest model, the uniform distribution in an origin-centred ball. The distribution has one parameter, the radius of the ball. Let $x^n \in (\mathbb{R}^d)^n$ be a data sequence. The maximum likelihood parameter $R(x^n)$ is equal to the distance of the farthest point in the sequence from the origin. If we restrict the data so that $R(x^n) \in [r_1,\, r_2]$, we can derive a normalized maximum likelihood $f_{\text{NML}}(x^n;\, r_1, r_2)$. But it can be difficult to give r_1 and r_2 any reasonable values before seeing the data. If we let r_1 and r_2 approach $R(x^n)$, the density grows unbounded, which means that renormalization is not possible. Also, if we fix r_1 and let $r_2 \to \infty$, or fix r_2 and let $r_1 \to 0$, the density approaches 0. Instead, we give the parameter r_1 a continuous prior density p_{r_1} and let $t > 1$. Now, we get a mixture density by integrating $f_{\text{NML}}(x^n;\, r_1, tr_1)$ over such values of r_1 that $R(x^n) \in [r_1,\, tr_1]$. That gives the density

$$f(x^n;\, p_{r_1}, t) = \int_{R(x^n)/t}^{R(x^n)} f_{\text{NML}}(x^n;\, r, tr)\, p_{r_1}(r)\, dr\,.$$

To dispose of the parameter t, we consider the limit

$$\lim_{t \to 1+} f(x^n;\, p_{r_1}, t)$$

The limiting function is a density function and a universal model [2]. We still have the problem of choosing a suitable p_{r_1}. NML encoding minimizes the worst-case excess code length compared to the maximum likelihood code length (the latter being the optimal coding method, but only with hindsight). Similarly, we should choose a flat prior which diminishes asymptotically as slow as possible in order to minimize the excess code length with all data. Section 2 introduces a generalization of Rissanen's prior for positive reals as a candidate for p_{r_1}.

With continuous distributions, it is a common practice to use the term *code length* as a synonym for the negative logarithm of the density, which corresponds to encoding

of real numbers with infinite precision. We consider just densities in this paper, taking the logarithm is left to the reader. In practical situations, data values have a finite precision and minimizing the negative logarithm of the density is not quite equal to finding the most effective way to encode the data. This does not usually cause problems, but if the density can grow unbounded in the neighbourhood of some point, the results may be surprising, especially when the data is represented with greater precision than we find trustworthy. Therefore it might be reasonable to fine-tune the model so that the density is bounded.

In the following sections, there are examples of densities having singularities in which they are not defined and in the neighbourhoods of which they get arbitrarily large values. These densities are problematic to use with some data sets. We might for example have a two-point cluster the density of which grows unbounded when the distance between the points approaches zero. The small cluster could thus totally dominate the code length of a potentially complex clustering, which might lead to very unintuitive results. We give therefore some solutions how the densities can be bounded and extended to the singularities.

We use the notation $\log$ for the logarithm to the base of two and $\ln$ for the natural logarithm. The elements of a data sequence are assumed to be identically and independently distributed in all the models.

Before deriving the densities corresponding to the code length functions we introduce a very flat density function that we use as a prior for parameters.

2. A PRIOR DENSITY FOR THE REAL NUMBERS

The main criterion for priors of the parameters in our case is that they should be as uninformative as possible. In [5], Rissanen gives a density function for the reals in the interval $[1, \infty[$. We generalize it without changing its asymptotic properties by adding a parameter that defines how strongly the probability mass is concentrated in the vicinity of the origin.

We write x^y as $x \uparrow y$ for typographical reasons. Let $x \uparrow\uparrow 0 = 1$ and let $x \uparrow\uparrow y = \underbrace{x \uparrow x \uparrow \ldots \uparrow x}_{y \text{ copies of } x}$ for $x > 0$, $y \in \mathbb{N}$. Now let $b = \underbrace{2 \uparrow 2 \uparrow \ldots \uparrow 2 \uparrow}_{k-1 \text{ copies of '2}\uparrow\text{'s}} \delta$ where $k \in \mathbb{N}$ and $\delta \in [1, 2]$. For $x \in \mathbb{R}_+$, we define the density

$$f_{\mathbb{R}_+}(x;\, b) = \frac{1 - \ln 2}{(\ln 2)^k} \, \frac{1}{1 - \log \delta \, (1 - \ln 2)} \, \frac{1}{(x + b)\, h(x + b)} \tag{1}$$

where

$$h(x) = \begin{cases} 1 & \text{if } \log x \leq 1 \\ \log x \; h(\log x) & \text{otherwise.} \end{cases}$$

We verify next that $f_{\mathbb{R}_+}(\cdot;\, b)$ integrates to unity over the positive real line. Let $\log^{(k)} x = \underbrace{\log \log \ldots \log}_{k \text{ copies}} x$. Notice that

$$D_x (\ln 2)^k \, \log^{(k)} x = \frac{1}{x \, h(x)} \,,$$

if $1 \leq \log^{(k-1)} x \leq 2$ or equivalently $2 \uparrow\uparrow (k-1) \leq x \leq 2 \uparrow\uparrow k$. Thus

$$\int_{2 \uparrow\uparrow (k-1)}^{2 \uparrow\uparrow k} \frac{1}{x \, h(x)} \, dx = (\ln 2)^k$$

and

$$\int_1^\infty \frac{1}{x \, h(x)} \, dx = \sum_{k=1}^{\infty} \int_{2 \uparrow\uparrow (k-1)}^{2 \uparrow\uparrow k} \frac{1}{x \, h(x)} \, dx = \frac{\ln 2}{1 - \ln 2} \,.$$

Assuming that b is defined as above,

$$\begin{aligned} &\int_b^\infty \frac{1}{x \, h(x)} \, dx \\ &\quad = \int_1^\infty \frac{1}{x \, h(x)} \, dx - \sum_{i=1}^{k-1} \int_{2 \uparrow\uparrow (i-1)}^{2 \uparrow\uparrow i} \frac{1}{x \, h(x)} \, dx \\ &\qquad - \int_{2 \uparrow\uparrow (k-1)}^{b} \frac{1}{x \, h(x)} \, dx \\ &\quad = \frac{\ln 2}{1 - \ln 2} - \left(\frac{1 - (\ln 2)^k}{1 - \ln 2} - 1 \right) - (\ln 2)^k \, \log \delta \\ &\quad = (\ln 2)^k \left(\frac{1}{1 - \ln 2} - \log \delta \right) . \end{aligned}$$

The proof is easily completed after a variable change $x = y + b$.

The function $f_{\mathbb{R}_+}(\cdot;\, b)$ diminishes asymptotically only slightly faster than $1/(x \log x)$. When a prior for all the real numbers is needed, we simply use $(1/2) f_{\mathbb{R}_+}(|x|)$.

3. A CODE LENGTH ACCORDING TO UNIFORM DISTRIBUTIONS IN ORIGIN-CENTRED BALLS

In this section, we consider a model class consisting of uniform distributions in a sphere centred at the origin. Let

$$V_d(r) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)} r^d \,,$$

denote the volume of a d-dimensional sphere with the radius r, and let $A_{d-1}(r) = V_d(r) \; d/r$ denote the surface area of that sphere. For the distance of the farthest point in the sequence $x^n = (x_1, x_2, \ldots, x_n) \in (\mathbb{R}^d)^n$ from the origin, we use the notation $R(x^n) = \max \{ \|x_i\| \mid i \in \{1, 2, \ldots, n\} \}$.

We consider first a special case where the radius of the smallest enclosing sphere belongs to a certain interval. Let $r_1, r_2 > 0$ and assume that $r_1 < r_2$. Let $A = \{x^n \in (\mathbb{R}^d)^n \mid R(x^n) \in [r_1, r_2]\}$. The maximum likelihood for $x^n \in A$ is $g_{\text{ML}}(x^n;\, r_1, r_2) = V_d(R(x^n))^{-n}$.

The normalization integral for the NML density is

$$\begin{aligned}
&\int_{x^n \in A} g_{\mathrm{ML}}(x^n;\, r_1, r_2)\, dx^n \\
&\quad = n \int_{x_1 \in \overline{B}(\overline{0}, r_2) \setminus B(\overline{0}, r_1)} \int_{x_2 \in \overline{B}(\overline{0}, \|x_1\|)} \cdots \\
&\qquad\qquad \cdots \int_{x_n \in \overline{B}(\overline{0}, \|x_1\|)} \frac{dx_n\, dx_{n-1} \ldots dx_1}{V_d(\|x_1\|)^n} \\
&\quad = n \int_{x_1 \in \overline{B}(\overline{0}, r_2) \setminus B(\overline{0}, r_1)} \frac{1}{V_d(\|x_1\|)}\, dx_1 \\
&\quad = n \int_{r_1}^{r_2} \int_{y \in \partial B(\overline{0}, r)} \frac{1}{V_d(r)}\, dy\, dr \\
&\quad = n \int_{r_1}^{r_2} \frac{A_{d-1}(r)}{V_d(r)}\, dr \\
&\quad = n \int_{r_1}^{r_2} \frac{d}{r}\, dr \\
&\quad = nd \ln \frac{r_2}{r_1}
\end{aligned}$$

which yields the NML density function

$$f_{\mathrm{NML}}(x^n;\, r_1, r_2) = \frac{1}{V_d(R(x^n))^n} \frac{1}{nd} \frac{1}{\ln(r_2/r_1)} \qquad (2)$$

if $x^n \in A$. The parameters r_1 and r_2 are something we would like to get rid of, because we can seldom give them reasonable values before looking at the data. Setting the parameters to their maximum likelihood values $r_1 = r_2 = R(x^n)$ results in an infinite density, which implies that a renormalization schema is not possible.

Instead, we give r_1 a continuous prior density p_{r_1}, and make r_2 a function of r_1, $r_2(r_1) = tr_1$, where $t > 1$. Integrating the coefficient $1/\ln(r_2/r_1) = 1/\ln t$ over the values of r_1 such that $R(x^n) \in [r_1, r_2] = [r_1, tr_1]$ yields

$$\int_{R(x^n)/t}^{R(x^n)} \frac{1}{\ln t}\, p_{r_1}(r)\, dr\,.$$

When t approaches 1 from above, let the limiting function be

$$\begin{aligned}
u(x^n,\, p_{r_1}) &= \lim_{t \to 1+} \int_{R(x^n)/t}^{R(x^n)} \frac{1}{\ln t}\, p_{r_1}(r)\, dr \\
&= \lim_{t \to 1+} \left(R(x^n) - \frac{R(x^n)}{t} \right) \frac{1}{\ln t}\, p_{r_1}(R(x^n)) \\
&= R(x^n)\, p_{r_1}(R(x^n))\,.
\end{aligned}$$

Replacing the coefficient $1/\ln(r_2/r_1)$ with $u(x^n,\, p_{r_1})$ in (2) yields the function

$$f(x^n;\, p_{r_1}) = \frac{R(x^n)}{V_d(R(x^n))^n} \frac{1}{nd}\, p_{r_1}(R(x^n)) \qquad (3)$$

if $x^n \in (\mathbb{R}^d)^n$ and $R(x^n) > 0$. It is easy to check that (3) is a valid density by integrating over $\{x^n \in (\mathbb{R}^d)^n \mid R(x^n) > 0\}$.

We compare the result briefly with a more straightforward solution. Assume that $R(x^n) \geq a > 0$ and let $p(r) = \epsilon a^\epsilon r^{-1-\epsilon}$ where $\epsilon > 0$. Consider the mixture density

$$\begin{aligned}
f_0(x^n) &= \int_{R(x^n)}^{\infty} V_d(r)^{-n}\, p(r)\, dr \\
&= \frac{1}{V_d(R(x^n))^n} \frac{1}{R(x^n)^\epsilon} \frac{\epsilon a^\epsilon}{nd + \epsilon}\,.
\end{aligned}$$

Let $p_{r_1} = f_{\mathbb{R}_+}(\cdot;\, b)$ as defined in (1). Then the ratio

$$\frac{f_0(x^n)}{f(x^n;\, p_{r_1})} \propto \frac{1}{R(x^n)^{1+\epsilon}\, p_{r_1}(R(x^n))}$$

approaches zero when $R(x^n) \to \infty$.

Depending on the choice of p_{r_1}, the density $f(x^n; p_{r_1})$ can grow unbounded when $R(x^n)$ approaches zero. A simple solution to get a bounded density is to make p_{r_1} a function of n and d, and to let $p_{r_1}(R)$ grow relative to R^{nd-1} in an interval $[0,\, \epsilon]$, which keeps $f(x^n; p_{r_1})$ constant when $R(x^n) \in [0,\, \epsilon]$. As a concrete example, let $b = 2 \uparrow\uparrow (k-1)$, where $k \in \mathbb{N}$ and we have used the notation explained in Section 2. Let also $\epsilon = \underbrace{2 \uparrow 2 \uparrow \ldots \uparrow 2 \uparrow}_{k-1 \text{ copies of '2}\uparrow\text{'s}} \alpha - b$ where $\alpha \in]1,\, 2]$. A continuous density fulfilling the previous requirements is

$$p_{r_1}(R) = \begin{cases} c\, f_{\mathbb{R}_+}(\epsilon;\, b)\, \epsilon^{1-nd}\, R^{nd-1} & \text{if } R \in [0,\, \epsilon[\\ c\, f_{\mathbb{R}_+}(R;\, b) & \text{if } R \geq \epsilon\,, \end{cases}$$

where $f_{\mathbb{R}_+}$ is a density defined in (1) and c is a constant for normalization. Because

$$\int_0^\epsilon f_{\mathbb{R}_+}(R;\, b)\, dR = (1 - \ln 2) \log \alpha$$

and

$$\int_0^\epsilon \frac{f_{\mathbb{R}_+}(\epsilon;\, b)}{\epsilon^{nd-1}} R^{nd-1}\, dR = \frac{f_{\mathbb{R}_+}(\epsilon;\, b)}{nd} \epsilon\,,$$

we get

$$c = \left(1 - (1 - \ln 2) \log \alpha + \frac{f_{\mathbb{R}_+}(\epsilon;\, b)}{nd} \epsilon \right)^{-1}\,.$$

4. A CODE LENGTH ACCORDING TO UNIFORM DISTRIBUTIONS IN ARBITRARY BALLS

We consider here modelling a data sequence according to a uniform distribution in an arbitrary ball, first in one and then in two dimensions. The one-dimensional case is important because a uniform distribution in an axis-aligned box is the product of the densities of the coordinates according to uniform distributions in one-dimensional balls.

In the first subsection, we assume that the minimum and maximum values of the one-dimensional sequence are unequal. In the second subsection, we bound the density not by choosing a special prior but by altering the models slightly. The third and fourth subsections examine the two-dimensional case.

Let $c(x^n)$ denote the centre of the smallest enclosing ball of $x^n \in (\mathbb{R}^d)^n$, and let $r(x^n)$ denote the radius of that ball.

4.1. One dimension, $\min(x^n) \neq \max(x^n)$

We restrict the data with the maximum likelihood parameters first. Let $c_0 \in \mathbb{R}$ and let $\delta, r_1, r_2 > 0$ Assume that $r_1 < r_2$. Let the set of sequences to be considered be $A = \{x^n \in \mathbb{R}^n \mid c(x^n) \in [c_0 - \delta,\, c_0 + \delta],\, r(x^n) \in [r_1,\, r_2]\}$. The maximum likelihood of $x^n \in A$ is

$$g_{\mathrm{ML}}(x^n) = \frac{1}{(2r(x^n))^n}\,,$$

and the corresponding normalizing integral is

$$\begin{aligned}
&\mathcal{C}(c_0, \delta, r_1, r_2) \qquad (4)\\
&\quad = \int_{x^n \in A} g_{\mathrm{ML}}(x^n)\, dx^n\\
&\quad = n(n-1) \iint_{\substack{x_1, x_2 \in \mathbb{R}:\\ (x_1+x_2)/2 \,\in [c_0-\delta, c_0+\delta],\\ (x_2-x_1)/2 \,\in [r_1, r_2]}} \int_{x_1}^{x_2} \cdots\\
&\qquad\qquad \cdots \int_{x_1}^{x_2} \frac{1}{(x_2 - x_1)^n} dx_n\, dx_{n-1} \ldots dx_2\, dx_1\\
&\quad = n(n-1) \iint_{\substack{x_1, x_2 \in \mathbb{R}:\\ (x_1+x_2)/2 \in [c_0-\delta, c_0+\delta],\\ (x_2-x_1)/2 \,\in [r_1, r_2]}} \frac{dx_2\, dx_1}{(x_2 - x_1)^2}\\
&\quad = n(n-1) \int_{c_0-\delta}^{c_0+\delta} \int_{r_1}^{r_2} \frac{2}{(2r)^2}\, dr\, dc \qquad (5)\\
&\quad = n(n-1)\, \delta \left(\frac{1}{r_1} - \frac{1}{r_2}\right).
\end{aligned}$$

There was a coordinate change $(x_1, x_2) = (c - r,\, c + r)$ at (5) in the previous integration. Dividing the maximum likelihood by the normalizing integral yields the NML density function

$$f_{\mathrm{NML}}(x^n;\, c_0, \delta, r_1, r_2) = \frac{1}{(2r(x^n))^n} \frac{1}{n(n-1)} \frac{1}{\delta} \frac{r_1 r_2}{r_2 - r_1} \qquad (6)$$

if $x^n \in A$.

The next step is to replace c_0, δ, r_1 and r_2 with more general parameters that allow us to define a non-zero density for all $x^n \in \mathbb{R}$ having $r(x^n) > 0$. Our solution is similar to the one in Section 3.

We assume that r_1 is independent of δ and c_0. Consider the parameters r_1 and r_2 first. Let again $t > 1$ and $r_2(r_1) = tr_1$. Requiring that $r(x^n) \in [r_1,\, r_2] = [r_1,\, tr_1]$, we replace the coefficient $(r_1 r_2)/(r_2 - r_1) = (tr_1)/(t-1)$ in (6) with the integral

$$\int_{r(x^n)/t}^{r(x^n)} \frac{tr}{t-1}\, p_{r_1}(r)\, dr\,,$$

where p_{r_1} is a continuous prior of the parameter r_1. Letting t approach 1 from above, we get

$$\begin{aligned}
&\lim_{t \to 1+} \int_{r(x^n)/t}^{r(x^n)} \frac{tr}{t-1}\, p_{r_1}(r)\, dr\\
&\quad = \lim_{t \to 1+} \left(r(x^n) - \frac{r(x^n)}{t}\right) \frac{t\, r(x^n)}{t-1}\, p_{r_1}(r(x^n))\\
&\quad = r(x^n)^2\, p_{r_1}(r(x^n))\,.
\end{aligned}$$

Next, we get rid of the coefficient $1/\delta$ and the dependence on c_0 in (6). Let $\delta > 0$ and let p_{c_0} be a continuous prior density function of the parameter c_0. The integration goes over all such values of c_0 that $c(x^n) \in [c_0 - \delta,\, c_0 + \delta]$. In a similar fashion as above, we substitute $1/\delta$ with the limiting function

$$\lim_{\delta \to 0+} \int_{c(x^n)-\delta}^{c(x^n)+\delta} \frac{1}{\delta}\, p_{c_0}(c)\, dc = 2 p_{c_0}(c(x^n))\,. \qquad (7)$$

The final density function is thus

$$f(x^n;\, p_{c_0}, p_{r_1}) = \frac{1}{(2r(x^n))^{n-2}} \frac{p_{c_0}(c(x^n))\, p_{r_1}(r(x^n))}{2n(n-1)} \qquad (8)$$

if $x^n \in \mathbb{R}^n$ and $r(x^n) > 0$.

The sequences consisting of equal points are problematic singularities this time. We can bound the density and extend it to the singularities as at the end of Section 3, choosing a special prior p_{r_1} that keeps $f(x^n;\, p_{c_0},\, p_{r_1})$ constant when $r(x^n) \in [0, \epsilon]$. For the case $n = 1$ we let $f((x_1);\, p_{c_0}) = p_{c_0}(x)$.

A naive solution for bounding the density is to add one extra point to the beginning of the sequence in order to ensure that the difference between the maximum and minimum values in the sequence is greater than some positive ϵ. By decoding, this point is simply discarded. But then if $r(x^n) > \epsilon$, the number of extra bits needed compared to (8) is not a constant any more but $\log r(x^n) + \log(n+1) - \log(n-1) + 1$.

In the next subsection, we provide yet another solution how to bound the density.

4.2. One dimension, bounded maximum likelihood

We restrict the largest possible density by bounding the radius parameter of the models from below. Assume for the time being that $n \in \{2, 3, \ldots\}$. We shall see later that the result applies for $n = 1$ as well. Let $x^n \in \mathbb{R}^n$ and let the smallest radius that can be used for encoding to be $\epsilon > 0$. The maximum likelihood is thus

$$g_{\mathrm{ML}}(x^n;\, \epsilon) = \begin{cases} (2r(x^n))^{-n} & \text{if } r(x^n) \geq \epsilon\\ (2\epsilon)^{-n} & \text{otherwise}\,. \end{cases}$$

Let $\delta, r_2 > 0$ and let $c_0 \in \mathbb{R}$. We calculate the normalizing integral in the bounded set $\{x^n \in \mathbb{R}^n \mid c(x^n) \in [c_o - \delta,\, c_0 + \delta],\, r(x^n) \leq r_2\}$ first. The integral consist of two parts:

$$\begin{aligned}
\mathcal{C}(c_0, \delta, r_2) = &\int_{\substack{x^n \in \mathbb{R}^n:\\ c(x^n) \in [c_0-\delta,\, c_0+\delta],\\ r(x^n) \in [0,\, \epsilon[}} \frac{1}{(2\epsilon)^n}\, dx^n\\
&+ \int_{\substack{x^n \in \mathbb{R}^n:\\ c(x^n) \in [c_0-\delta,\, c_0+\delta],\\ r(x^n) \in [\epsilon,\, r_2]}} \frac{1}{(2r(x^n))^n}\, dx^n\,.
\end{aligned}$$

The second term is $\mathcal{C}(c, \delta, \epsilon, r_2)$ as in (4). The first term is equal to

$$\begin{aligned} & n(n-1) \int_{\substack{x_1, x_2 \in \mathbb{R}: \\ (x_1+x_2)/2 \in [c-\delta,\, c+\delta] \\ (x_2-x_1)/2 \in [0, \epsilon[}} \frac{(x_2 - x_1)^{n-2}}{(2\epsilon)^n} \, dx_2 \, dx_1 \\ & = n(n-1) \int_{c-\delta}^{c+\delta} \int_0^{\epsilon} \frac{(2r)^{n-2}}{(2\epsilon)^n} 2 \, dr \, dc \\ & = \frac{n\delta}{\epsilon} \, . \end{aligned}$$

Putting these together yields

$$\mathcal{C}(c_0, \delta, r_2) = n(n-1)\delta \left(\frac{1}{\epsilon} - \frac{1}{r_2} \right) + \frac{n\delta}{\epsilon} \, .$$

When r_2 approaches infinity, $\mathcal{C}(c, \delta, r_2) \to (n^2\delta)/\epsilon$. We normalize the maximum likelihood by this limit and use a prior for the parameter c_0 in a similar way as in (7), getting the density

$$f_\epsilon(x^n;\, p_{c_0}) = \frac{1}{(2r(x^n))^n} \frac{2\epsilon}{n^2} \, p_{c_0}(c(x^n)) \qquad (9)$$

if $r(x^n) \geq \epsilon$, and

$$f_\epsilon(x^n;\, p_{c_0}) = \frac{1}{(2\epsilon)^{n-1}} \frac{p_{c_0}(c(x^n))}{n^2} \qquad (10)$$

if $r(x^n) \in [0, \epsilon[$. Letting $n = 1$ in (10) yields $f_\epsilon((x_1);\, p_{c_0}) = p_{c_o}(x_1)$, which is a valid density.

4.3. Two dimensions, $r(x^n) > 0$

Next, we consider the arbitrary ball model in two dimensions. Assume first that $n \in \{3, 4, \dots\}$. The final result is also valid for $n = 2$, which we shall see after the calculation of the normalizing integral at (11).

Let $(C_1, C_2) \in \mathbb{R}^2$ and let $\delta, r_1, r_2 > 0$, $r_1 < r_2$. We assume first that the centre of the smallest enclosing ball of the point sequence belongs to the set $D = [C_1 - \delta,\, C_1 + \delta] \times [C_2 - \delta,\, C_2 + \delta]$ and that the radius of that ball is in the interval $[r_1,\, r_2]$. Let the set of sequences fulfilling these conditions be $A = \{x^n \in (\mathbb{R}^2)^n \mid c(x^n) \in D,\, r(x^n) \in [r_1,\, r_2]\}$. Denote the maximum likelihood in this set $g_{\mathrm{ML}}(x^n) = (\pi r(x^n)^2)^{-n}$.

There must be at least two different points x_i, x_j in the sequence x^n such that $x_i, x_j \in \partial B(c(x^n), r(x^n))$. If x_i and x_j are the only points in x^n belonging to the border of the smallest enclosing ball, then $c(x^n) = (x_i + x_j)/2$. If at least three points of x^n belong to $\partial B(c(x^n), r(x^n))$, there are three different indices $i, j, k \in \{1, 2, \dots, n\}$ such that $x_i, x_j,\, x_k \in \partial B(c(x^n), r(x^n))$ and $c(x^n) \in \mathrm{conv}\{x_i, x_j, x_k\}$, where $\mathrm{conv}\{x_i, x_j, x_k\}$ is the convex hull of the set $\{x_i,\, x_j, x_k\}$.

We derive then integral $\int_{x^n \in A} g_{\mathrm{ML}}(x^n) \, dx^n$ by dividing the integrating space into two parts whose intersection is a null set. First, consider a situation where the points x_1 and x_2 determine the minimal enclosing ball. Let $x_i(j)$ denote the jth coordinate of x_i. We change the coordinate system using the function

$$\begin{aligned} w(c_1, c_2, r, \theta) &= (c_1 + r\cos\theta, \;\; c_2 + r\sin\theta, \\ &\qquad c_1 - r\cos\theta, \;\; c_2 - r\sin\theta) \\ &= (x_1(1), x_1(2), x_2(1), x_2(2)) \, . \end{aligned}$$

Hence, $|\det w'(c_1, c_2, r, \theta)| = 4r$. The integral with x_1 and x_2 as outermost points is

$$\begin{aligned} & I_2(D, r_1, r_2) \\ & = \int_{\substack{x_1, x_2 \in \partial B(c(x^n), r(x^n)), \\ c(x^n) = (x_1+x_2)/2}} \; \int_{\substack{x^n \in A: \\ }} g_{\mathrm{ML}}(x^n) \, dx^n \\ & = \iint_{\substack{x_1, x_2 \in \mathbb{R}^2: \\ (x_1+x_2)/2 \in D, \\ r_1 < \|x_1 - x_2\|/2 < r_2}} \; \int\cdots\int_{\substack{x_3, \dots, x_n \in \\ \overline{B}((x_1+x_2)/2, \|x_1-x_2\|/2)}} \\ & \qquad \left(\pi \left(\frac{\|x_1 - x_2\|}{2} \right)^2 \right)^{-n} dx_n \, dx_{n-1} \dots dx_1 \\ & = \iint_{\substack{x_1, x_2 \in \mathbb{R}^2: \\ (x_1+x_2)/2 \in D, \\ r_1 < \|x_1 - x_2\|/2 < r_2}} \left(\pi \left(\frac{\|x_1 - x_2\|}{2} \right)^2 \right)^{-2} dx_2 dx_1 \\ & = \iint_{(c_1, c_2) \in D} \int_0^{2\pi} \int_{r_1}^{r_2} \frac{4r}{(\pi r^2)^2} \, dr \, d\theta \, dc_2 \, dc_1 \\ & = (4\delta^2)(2\pi) \frac{4}{\pi^2} \int_{r_1}^{r_2} \frac{1}{r^3} dr \\ & = \frac{16\delta^2}{\pi} \left(\frac{1}{r_1^2} - \frac{1}{r_2^2} \right) . \end{aligned}$$

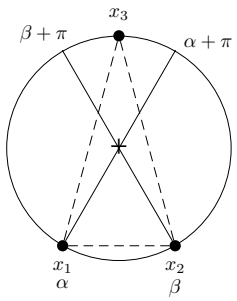

Figure 1. Three points and their smallest enclosing ball. If and only if the angular coordinate of x_3 is between $\alpha + \pi$ and $\beta + \pi$, the centre of the smallest enclosing ball is the point marked with a cross.

Next, consider the situation where the points x_1, x_2 and x_3 determine the smallest enclosing ball (figure 1). The function for the change of coordinates is now

$$\begin{aligned} w(c_1, c_2, r, \alpha, \beta, \gamma) = (& c_1 + r\cos\alpha, \;\; c_2 + r\sin\alpha, \\ & c_1 + r\cos\beta, \;\; c_2 + r\sin\beta, \\ & c_1 + r\cos\gamma, \;\; c_2 + r\sin\gamma) \, , \end{aligned}$$

and $|\det w'(c_1, c_2, \alpha, \beta, \gamma)| = \big|\sin(\alpha - \beta) + \sin(\gamma - \alpha) + \sin(\beta - \gamma)\big| \, r^3$.

The integral with x_1, x_2 and x_3 as outermost points without any fixed ordering is

$$\begin{aligned}
&I_3(D, r_1, r_2) \\
&\quad= \int_{\substack{x^n \in A: \\ x_1,x_2,x_3 \in \partial B(c(x^n), r(x^n)), \\ c(x^n) \in \operatorname{conv}\{x_1,x_2,x_3\}}} g_{\mathrm{ML}}(x^n)\, dx^n \\
&\quad= \iiint_{\substack{x_1,x_2,x_3 \in \mathbb{R}^2: \\ c((x_1,x_2,x_3)) \in D, \\ r_1 < r((x_1,x_2,x_3)) < r_2}} \int \cdots \int_{\substack{x_4,\ldots,x_n \in \\ B(c((x_1,x_2,x_3)), r((x_1,x_2,x_3)))}} \\
&\qquad\qquad (\pi r((x_1,x_2,x_3))^2)^{-n}\, dx_n\, dx_{n-1} \ldots dx_1 \\
&\quad= \iiint_{\substack{x_1,x_2,x_3 \in \mathbb{R}^2: \\ c((x_1,x_2,x_3)) \in D, \\ r_1 < r((x_1,x_2,x_3)) < r_2}} \frac{dx_3\, dx_2\, dx_1}{(\pi r((x_1,x_2,x_3))^2)^3} \\
&\quad= \iint_{(c_1,c_2) \in D} \int_{r_1}^{r_2} \int_0^{2\pi} 2 \int_{\alpha}^{\alpha+\pi} \int_{\alpha+\pi}^{\beta+\pi} \\
&\qquad \left| \sin(\alpha-\beta) + \sin(\gamma-\alpha) + \sin(\beta-\gamma) \right| \frac{r^3}{(\pi r^2)^3} \\
&\qquad d\gamma\, d\beta\, d\alpha\, dr\, dc_2\, dc_1 \\
&\quad= \frac{2}{\pi^3}(4\delta^2)\frac{1}{2}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right)(2\pi)(3\pi) \\
&\quad= \frac{24\delta^2}{\pi}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right).
\end{aligned}$$

Using symmetry among the points, we get the normalizing integral

$$\begin{aligned}
&\int_{x^n \in A} g_{\mathrm{ML}}(x^n)\, dx^n \qquad (11) \\
&= \binom{n}{2} I_2(D, r_1, r_2) + \binom{n}{3} I_3(D, r_1, r_2) \\
&= \frac{n(n-1)}{2}\frac{16\delta^2}{\pi}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right) \\
&\quad + \frac{n(n-1)(n-2)}{6}\frac{24\delta^2}{\pi}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right) \\
&= 4n^2(n-1)\frac{\delta^2}{\pi}\left(\frac{1}{r_1^2} - \frac{1}{r_2^2}\right).
\end{aligned}$$

The calculation above is valid also for $n = 2$ if we define $\binom{2}{3} \equiv 0$. The corresponding normalized density is

$$f_{\mathrm{NML}}(x^n; D, r_1, r_2) = \frac{1}{(\pi r(x^n)^2)^n} \frac{1}{4n^2(n-1)} \frac{\pi}{\delta^2} \frac{r_1^2 r_2^2}{r_2^2 - r_1^2}$$

if $x^n \in A$ and $n \in \{2, 3, \ldots\}$.

When we give the centre of the smallest enclosing ball a prior density p_c and the radius r_1 a prior p_{r_1}, we can derive the final density as before. Let $c_i(x^n)$ denote the ith coordinate of $c(x^n)$. The limits are

$$\begin{aligned}
&\lim_{\delta \to 0+} \int_{c_1(x^n)-\delta}^{c_1(x^n)+\delta} \int_{c_2(x^n)-\delta}^{c_2(x^n)+\delta} \frac{1}{\delta^2}\, p_c((c_1, c_2))\, dc_2\, dc_1 \\
&\quad = 4\, p_c((c_1(x^n), c_2(x^n)))
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{t \to 1+} \int_{r(x^n)/t}^{r(x^n)} \frac{r^2 (tr)^2}{(tr)^2 - r^2}\, p_{r_1}(r)\, dr \\
&\quad = \lim_{t \to 1+} \left(r(x^n) - \frac{r(x^n)}{t} \right) r(x^n)^2 \frac{t^2}{t^2 - 1}\, p_{r_1}(r(x^n)) \\
&\quad = \frac{1}{2}\, r(x^n)^3\, p_{r_1}(r(x^n)).
\end{aligned}$$

The final density is thus

$$\begin{aligned}
&f(x^n;\, p_c, p_{r_1}) \qquad (12) \\
&\quad = \frac{1}{\pi^{n-1} r(x^n)^{2n-3}} \frac{p_c(c(x^n))\, p_{r_1}(r(x^n))}{2n^2(n-1)}.
\end{aligned}$$

if $x^n \in \{y^n \in (\mathbb{R}^2)^n \mid r(y^n) > 0\}$, where $n \in \{2, 3, \ldots\}$.

4.4. Two dimensions, bounded maximum likelihood

We omit the calculations here, because they are essentially similar to those in the one-dimensional case in Subsection 4.2. Let $n \in \{1, 2, 3, \ldots\}$. The final density is

$$f_\epsilon(x^n;\, p_c) = \frac{1}{\pi^{n-1}\, r(x^n)^{2n}} \frac{\epsilon^2}{n^3}\, p_c(c(x^n))$$

if $r(x^n) \geq \epsilon$, and $f_\epsilon(x^n; p_c) = p_c(c(x^n))/(\pi^{n-1}\epsilon^{2n-2}n^3)$ if $r(x^n) \in [0, \epsilon[$.

5. REFERENCES

[1] Jorma Rissanen, *Information and Complexity in Statistical Modeling*, Springer Verlag, New York, 2007.

[2] Peter D. Grünwald, *The Minimum Description Length Principle*, The MIT Press, 2007.

[3] Panu Luosto, Jyrki Kivinen, and Heikki Mannila, "Gaussian clusters and noise: an approach based on the minimum description length principle," in *Discovery Science*, 2010, to appear.

[4] Jorma Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[5] Jorma Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, Jun. 1983.