In this chapter we present a Bayesian approach to the parameter estimation of a hypersphere, given noisy point observations on that sphere. The Bayesian approach forces to make explicit all assumptions and beliefs about the problem at hand in the form of prior probabilities. We identify these assumptions — in our specific case isotropic Gaussian zero mean noise with given standard deviation on the point observations, as well as uniform distribution of the observations over the estimated hyperspherical surface — and give mathematical, geometrical and practical justifications for them. As a result we obtain an explicit, closed form formula for the likelihood of the parameters “location of center” and “radius” of a hypersphere in arbitrary dimensions.

In a Bayesian setting such a likelihood formula would be used as a departure for further inference or reasoning. Here we calculate the likelihood’s gradient and apply gradient descent to obtain a maximum likelihood estimate (MLE) for 3D spheres. We compare this estimator to a non-linear least squares estimator (which differs from our maximum likelihood estimator in that it implicitly minimizes the observation noise) and analyze its behavior in various setups, including varying degrees of violation of our assumptions.
5.1. Motivation

The maximum likelihood estimator (MLE) for hyperspheres that we propose here includes very classical results in 2D circle estimation but extends to spheres in 3D and beyond. In fact, by giving a closed form likelihood function, we enable extension to a full Bayesian approach to the problem. Instead of singling out a maximum, the Bayesian approach retains the likelihood function, which can be used for further reasoning. For instance, one can infer typical parameter values using probability mass.

More precisely, we assume $N$ data points in an $n$-dimensional Euclidean space. As these points may themselves be the result of some measurement or estimation process we take them to be observations degraded by noise, their true positions being unknown. Out of all possible hyperspheres on which the true points could lie, a maximum likelihood estimate selects the one that has the highest probability of having generated the observed data, given certain characteristics of the observation noise. We will assume this noise to be Gaussian and isotropic, i.e. the observations are normally distributed around the true points with no preferred direction. Furthermore, we assume the variance of the observation noise to be known (though not necessarily the same for all points). This is not an unrealistic assumption because — whether the observations are measurements by a physical sensor or the result of some precursory calculation — often, one has some idea about its magnitude.

This chapter is structured as follows. In section 5.2 we give a short historically motivated overview about the subject at hand, including a brief review of relevant literature. In section 5.3 we will try to explain what is meant here by the “Bayesian approach” and place our method within that context. The main contribution is presented in section 5.4 while in section 5.5 we discuss its results. Finally, section 5.6 contains concluding remarks and an outlook to future work. Appendices B and C at the end of this thesis provide some detailed numerical results as well as mathematical background for expressing derivations in this chapter.

5.2. Background

Maximum likelihood estimation provides an intuitive framework for estimating model parameters from observed data. However, care has to be taken when it comes to modeling the random variables involved or describing the statistical behavior of the obtained estimates.

An early step on the way to a hypersphere estimator is a circle estimator. A circle is a hypersphere in two dimensions and a number of researchers have addressed the problem of estimating it from uncertain point data. Much disagreement seems to exist, however, regarding how to set up a maximum likelihood circle estimator or how to interpret its properties. We will give a brief overview and comparison of different approaches.

Kiefer and Wolfowitz [W56] present a very systematic and general approach to maximum likelihood estimation. We will adopt their notation in this chapter. Let $X_i$, $i = 1, \ldots, N$, be vectors of random variables. Each of them has $n$ components $X_{ij}$, $j = 1, \ldots, n$. Specific values (or realizations) $x_{ij}$ of these will be called observations or data. Let their probability density
5.2 Background

Figure 5.1: A Bayesian network depicting the dependencies of the variables involved in parameter estimation. A variable at an arrow’s head is statistically dependent on the one at its tail.

function be $f(x|\theta, \alpha_i)$, where the so-called structural parameters $\theta$ are used to parametrize some model that the data adheres to, while the vectors of incidental parameters $\alpha_i$ determine the distribution of the data.

To further structure the problem, one can assume that the $\alpha_i$ are not arbitrary constants but independently and identically distributed (i.i.d.) random variables with distribution function $G(\alpha_i|\nu)$, where $\nu$ parametrizes the distribution. Formally, $G$ is a collection of random variables and is called a stochastic process. See Figure 5.1 for a Bayesian network visualizing these notions. The way one chooses to view the $\alpha_i$ — whether to regard them as unknown constants, as i.i.d. random variables or something else — determines the behavior and properties of the resulting maximum likelihood estimator.

Kiefer and Wolfowitz come up with a number of regularity assumptions on $f$, which have to hold in order for the maximum likelihood estimator to be consistent. Proving these assumptions depends on the problem at hand and properties of the space in which $G$ is assumed (or allowed) to lie.

An early step on the way to a hypersphere estimator is a circle estimator. A circle is a hypersphere in two dimensions and a number of researchers have addressed the problem of estimating it from uncertain point data. Much disagreement seems to exist, however, regarding how to set up a maximum likelihood circle estimator or how to interpret its properties. We will give a brief overview and comparison of different approaches.

Having introduced the very general notation used in [W56] we will now focus on other authors’ work with maximum likelihood circle estimation in 2D. Obviously, the model in this case is the common circle on which the true data points lie. Its parameters (i.e. radius and location of the center) are the structural parameters $\theta$. There exists any number of possible incidental parameters $\alpha$, but the most common ones are those describing the observation noise, e.g. mean and variance of the respective random variables modeling the observations.

5.2.1. Structural Model vs. Functional Model

Given a circle on which some true 2D points lie, their observations can be considered as realizations of random variables, which can be described in the established terminology by

$$X_{i1} = \theta_1 + \theta_3 \cos \alpha_{i3} + \alpha_{i1}, \quad (5.1)$$
$$X_{i2} = \theta_2 + \theta_3 \sin \alpha_{i3} + \alpha_{i2}, \quad (5.2)$$
where $\alpha_{i3}$ parametrize the distribution of the true points over the circle with center $(\theta_1, \theta_2)$ and radius $\theta_3$, while $\alpha_{i1}$ and $\alpha_{i2}$ parametrize the noise on the point coordinates (see Figure 5.2).

Berman and Culpin [C86] classify different approaches to the incidental parameters. Firstly, they assume that $\alpha_{i1}$ and $\alpha_{i2}$ are Gaussian random variables with zero mean.

Then they cite Chan [Ch65], who assumed that $\alpha_{i3}$ were fixed but unknown angles. This they call the Gaussian Cartesian functional model, since the observations are Cartesian coordinates bound by the functional relationship based on $\alpha_{i3}$. An estimation method would attempt to estimate the $\alpha_{i3}$. Instead of taking them to have fixed but unknown values, the $\alpha_{i3}$ can be regarded as independently distributed random variables. Berman and Culpin [C86] call this the Gaussian Cartesian structural model, because nothing is assumed about the distribution of the true points and only the structural parameters $\theta$ are estimated, not the $\alpha_{i3}$.

The difference between the two is the difference between a random variable and a realization. The structural model describes the true points by a probability distribution which may require any number of additional parameters. It can be implemented in different manners by constraining the possible distribution of $\alpha_{i3}$ in certain ways [An81, Be83]. In the functional model each true point is treated as a single (albeit unknown) realization of that random variable which is usually estimated along with the structural parameters.

We observe here that the Bayesian approach blurs the boundaries between the two models as both of them arise out of different assumptions about prior distributions and the decision which variables are treated as data instead of parameters and which variables are marginalized over. We will treat these issues in more detail in section 5.3.

### 5.2.2. Cartesian Model vs. Radial Model

Another way of classifying the problem statement is by assuming noise not on the Cartesian coordinates of the observations but on their polar coordinate representation. The following is called the radial model [C86].

$$X_{i1} = \theta_1 + \alpha_{i1} \cos \alpha_{i2}, \quad (5.3)$$
$$X_{i2} = \theta_2 + \alpha_{i1} \sin \alpha_{i2}, \quad (5.4)$$
5.2 Background

with various constraints on the incidental parameters. For example, Berman and Culpin claim that it is usually assumed that \( E[\alpha_i] = \theta_3 \), the circle’s radius (which is saying that the radius can be modeled as a true value \( \theta_3 \) with additive zero mean noise). If the \( \alpha_i \) are regarded as independently distributed random variables, Berman and Culpin call it the radial structural model.

Note that \( X_{i1} \) and \( X_{i2} \) are Cartesian coordinates of the observations and that Berman and Culpin chose to subsume the structural radius parameter \( \theta_3 \) with additive zero mean noise \( \alpha_{i1} \) in one single random variable \( \alpha_{i1} \). The distribution of the angular parameters \( \alpha_{i2} \) determines the distribution of the observations over the circle as well as the angular observation noise.

Even though it is not mentioned explicitly in [C86], the radial functional model could be achieved by regarding the \( \alpha_{i2} \) as fixed but unknown angles.

5.2.3. Maximum Likelihood vs. Least Squares

When it comes to fitting a model to observed data, least squares methods are very popular, because they are straightforward to implement, intuitive and often give good results. Moreover, it is a common misconception that least squares estimates are equivalent to maximum likelihood estimates under the assumption of independent Gaussian noise. This is not generally true but rather depends very much on the model used.

A maximum likelihood estimator \( \{ \hat{\theta}_{ML}, \hat{\alpha}_{ML} \} \) tries to maximize the joint likelihood of the structural and possible incidental parameters \( \theta \) and \( \alpha \) that explain all the observations.

\[
\{ \hat{\theta}_{ML}, \hat{\alpha}_{ML} \} = \arg \max_{\theta, \alpha} L_{x_1, \ldots, x_N}(\theta, \alpha),
\]

(5.5)

where the \( L_{x_1, \ldots, x_N}(\theta, \alpha) \) denotes the conditional probability \( p(x_1, \ldots, x_N|\theta, \alpha) \) and depends on the specific functional relationship between the structural and possible incidental parameters \( \theta \) and \( \alpha \), and the observations \( x_i \).

By contrast, a least squares estimator \( \{ \hat{\theta}_{LS}, \hat{\alpha}_{LS} \} \) tries to minimize the sum of squared distances between the observations \( x_i \) and the model \( f \), which depends on the structural parameters \( \theta \), but also on the assumed fixed but unknown true values of the incidental parameters \( \alpha_i \). The true values \( \alpha \), which are constrained by the structural parameters \( \theta \), are estimated along with the structural parameters.

\[
\{ \hat{\theta}_{LS}, \hat{\alpha}_{LS} \} = \arg \min_{\theta, \alpha} \sum_{i}^{N} (x_i - f(\theta, \alpha))^2
\]

(5.6)

For 2D circle estimation, the least squares estimator setting is described by the functional models above. A classical relevant example of this procedure is found in [Ch65], where a functional relationship between Cartesian coordinates \( X_{i1}(\theta, \alpha_i) = \theta_1 + \theta_3 \cos \alpha_i \) and \( X_{i2}(\theta, \alpha_i) = \theta_2 + \theta_3 \sin \alpha_i \) is used to describe a circle, whose parameters \( \theta_i \) are then estimated along with the true points (or rather their true angular parameters \( \alpha_i \)).

Under the assumption of Gaussian isotropic noise, the above makes for equivalence of max-
imum likelihood and least squares methods in a certain sense as the sum of squared distances between observations \( x_i \) and estimated true points \((X_{i1}, X_{i2})\) is minimized. However, the least squares estimator also *implicitly* minimizes the amount of noise in the observations. This is behavior that the user of a least squares estimator may not have asked for. If the amount of noise can be known, this behavior is even detrimental to the estimate. Additionally, as each observation \( x_i \) adds a parameter (or degree of freedom) \( \alpha_i \) to the least squares estimate, it becomes more vulnerable to overfitting the data.

### 5.2.4. Consistency of MLE in General and in Particular

A very common quality measure for an estimator is *consistency*. That means that the expectation of the estimator tends towards the true value of the estimated quantity as the amount of data \( N \) tends towards infinity. As mentioned above, Kiefer and Wolfowitz [W56] propose a number of regularity assumptions under which the maximum likelihood estimator is consistent. Their proof pertains to maximum likelihood estimation in general and is not limited to circle or hypersphere estimation.

Chan [Ch65] points out that for the Gaussian Cartesian functional model of circle estimation, where \( X_{i1} \) and \( X_{i2} \) are normally distributed, a necessary and sufficient condition for the maximum likelihood estimator to be consistent is that both \( 1/N \sum \cos \alpha_{i3} \) and \( 1/N \sum \sin \alpha_{i3} \) tend to zero as \( N \to \infty \). Geometrically, this assumption is fulfilled if the data is uniformly distributed over the whole circle. Anderson [An81] remarks that attempts to establish a consistent estimator for the incidental parameters \( \alpha_{i3} \) introduces instability of the maximum likelihood estimator if the support of \( \alpha_{i3} \) is less than half the circle. She, too, reverts to the assumption that the observations are approximately uniformly distributed over the circle.

### 5.2.5. Efficiency of the MLE

Also, Anderson [An81] establishes formulas for the variance of the estimator for the circle’s radius and for the variance of the noise in the observations. Moreover, she brings up the *asymptotic relative efficiency* (ARE) of the maximum likelihood estimator for the circle’s center. This ratio compares the capability of two estimators to establish the same confidence in the estimate. We will omit the mathematical details here and only report on Anderson’s results.

Under the assumption of uniform distribution an alternative consistent estimator for the circle’s center is the arithmetic mean of the observations’ locations. Both estimators, the MLE estimator and the arithmetic mean, would eventually (i.e. given an infinite number of observations) find the true value. But as long as the amount of data is not infinite they are both afflicted by uncertainty which depends on the amount of data \( N < \infty \) seen so far.

Comparing the uncertainty of these two estimators, Anderson comes to the conclusion that the ARE of the maximum likelihood estimator depends on the ratio of observation noise \( \sigma^2 \) to the circle’s radius \( \rho^2 \). Anderson reports an ARE of 0.991 at \( \sigma^2/\rho^2 = 1.0 \), dropping to 0.810 at \( \sigma^2/\rho^2 = 0.3 \) then to 0.358 at \( \sigma^2/\rho^2 = 0.1 \) and down to 0.039 at \( \sigma^2/\rho^2 = 0.01 \). By definition of
the ARE, this means that the maximum likelihood estimator needs only 99.1% (respectively 81%, 35.8% and 3.9%) of the data to achieve the same level of uncertainty/confidence as the estimator derived from the observation mean.

5.3. A Bayesian Approach

We now assume a Bayesian standpoint. It should have become clear that, from a Bayesian point of view, there are a number of decisions that every author of a statistical parameter estimator makes, explicitly or — more often — implicitly. In this section we will introduce the general Bayesian approach while pointing out the decisions we make along the way for our specific hypersphere estimator and explain why we make them. Finally we will arrive at a likelihood formula which can be computed explicitly, yielding an estimate for the hypersphere parameters.

Formally, the likelihood maps a set of structural and incidental parameters to the conditional probability density of the observations given these parameters, i.e.

$$ (\theta, \alpha) \mapsto L_x(\theta, \alpha) = p(x|\theta, \alpha). $$

(5.7)

We would like to stress the fact that the likelihood is a function of the parameters $\theta$ and $\alpha$, not of the data $x$ — whereas $p(x|\theta, \alpha)$ denotes the probability of observing $x$ given the parameters $\theta$ and $\alpha$. Here we will use the notation of $L_x(\theta, \alpha)$ (read: the likelihood of the parameters $\theta$ and $\alpha$, given the data $x$) instead of $L(\theta, \alpha)$. Sometimes — especially in the context of maximum likelihood estimation — the likelihood function is only of interest up to a multiplicative constant.

Now we regard the incidental parameters $\alpha$ not as arbitrary given constants but as random variables governed by some (possibly unknown) stochastic process $G(\alpha|\nu)$ parametrized by $\nu$. In the terminology above this means that we are assuming a structural model, where the true values can only be described by a random variable of which the observations are realizations. Taking this into account the likelihood is described by

$$ L_{x,\alpha}(\theta, \nu) = p(x|\theta, \alpha)G(\alpha|\nu) = p(x, \alpha|\theta, \nu). $$

(5.8)

We assume a structural model because treating the incidental parameters as fixed values which need to be estimated would introduce additional degrees of freedom to our estimator which then would become more prone to overfitting the data. Note that at this point (5.8) we have abandoned the idea of estimating the true values $\alpha_i$. Instead we estimate the parameters $\nu$ of the process governing their probability distribution.

Anderson [An81] differentiates four cases of estimation problems with respect to the process $G$. Firstly, the process $G$ could be (assumed to be) a uniform distribution. Secondly, $G$ could be known except for a fixed number of parameters $\nu$, for instance $G$ could be a normal distribution with unknown mean and variance. Thirdly, $G$ could be unknown, in which case it
A Bayesian Approach to Hypersphere Estimation

lies in a function space parametrized by $\nu$ and can be estimated consistently \cite{W56}. In practice, however, it may be infeasible to estimate $\nu$ — and thereby $G$ — without further constraining the space in which $G$ is allowed to lie. Finally, if nothing is known about the incidental parameters, one could try to eliminate them from the functional relationship and end up with a likelihood that does not involve a $G$. For the functional circular model, Anderson says that it is provably impossible to eliminate the incidental parameters. Therefore, this approach is also not an option for our hypersphere estimator. Alternatively, in order to avoid introducing $G$ altogether, one could treat the incidental parameters as unknown constants to be estimated along with the structural parameters, which brings us back to the functional model and to the initial formulation of the likelihood (5.7).

In what we would like to call the “full Bayesian approach” it is even possible to introduce prior probabilities on the possible distributions $G$. This is very similar to the third case quoted above. Instead of strictly constraining the function space, these prior probabilities are used to guide the search for $G$. If one is interested in estimating only the structural parameters, though, one will marginalize the prior distribution out of the likelihood formula.

$$L_{x, \alpha}(\theta) = \int p(x|\theta, \alpha) G(\alpha|\nu) p(\nu) d\nu$$

$$= \int p(x|\theta, \alpha) G(\alpha, \nu) d\nu$$

$$= p(x|\theta, \alpha) G(\alpha)$$

(5.9)

where, for brevity’s sake, we extended the notation of probability densities to that of the stochastic process in a straightforward way.

With these remarks we are now finally in a position where we can formally specify the model we use for our hypersphere estimator. Instead of estimating the distribution of the incidental parameters or even marginalizing it out of our likelihood function, we will assume a certain distribution $G$.

Our goal is to fit a hypersphere to point data. We assume independent, isotropic Gaussian noise on the Cartesian coordinates of $N$ point observations. In many real world applications, a Gaussian distribution is a realistic — or at least very common — assumption, validated by the central limit theorem. The means and variances of these Gaussian distributions make up the incidental parameters of our estimator. We assume the variances to be known. As said before, in many applications this is the case, and in our setting we only want to estimate the structural (hypersphere) parameters. With the noise characteristics known we have the variance(s) $\sigma_i^2$, $i = 1, \ldots, N$ given. The incidental parameters that are left are the means of the point observations. Without any further information we assume that they are uniformly distributed over the surface of the hypersphere-to-be-estimated; we do not have an a priori reason to assume a concentration around some point. Because we are not interested in estimating the incidental parameters, we will marginalize them out of our likelihood formula.

For a single point observation $x_i$, given the hypersphere $s(\theta_c, \theta_r)$ with structural parameters
radius $\theta_r$ and location of center $\theta_c$, the true point $\alpha_i$ could lie anywhere along $s$ with equal probability. The observation is a realization of a normally distributed random variable with mean $\alpha_i$ and known variance $\sigma_i^2$. Then (5.10) becomes

$$L_X(\theta) = \int p(X, \alpha|\theta) \, d\alpha.$$  \hfill (5.11)

In our specific case, we start out from (5.9) and marginalize out the incidental parameters.

$$L_X(\theta) = \int L_{x_i, \alpha_i, \sigma}(\theta_c, \theta_r) \, d\alpha_i \, d\sigma$$

$$= \int p(x_i|\theta_c, \theta_r, \alpha_i, \sigma) \, G(\alpha_i, \sigma) \, d\alpha_i \, d\sigma$$

$$= \int p(x_i|\theta_c, \theta_r, \alpha_i, \sigma) \, U_s(\theta_c, \theta_r)(\alpha_i) \, \delta(\sigma - \sigma_i) \, d\alpha_i \, d\sigma$$

$$= \oint_{s(\theta_c, \theta_r)} N(x_i|\alpha_i, \sigma_i^2) \, d\alpha_i.$$  \hfill (5.12)

Note that one fundamental assumption we make is that there is a common hypersphere to be estimated. This means, given a set of structural parameters $\theta$, the probability of the true points lying anywhere but on the indicated hypersphere is zero. We use $U_s(\theta_c, \theta_r)(\alpha_i)$ to denote the uniform distribution of the $\alpha_i$ restricted to the hypersphere $s$ parametrized by $\theta_c$ and $\theta_r$. Therefore, the integral over the whole observation space becomes the integral over that hypersphere. We use the Dirac delta $\delta(\sigma - \sigma_i)$ as a probability distribution that is zero for all $\sigma \neq \sigma_i$ to model the fact that we know the value of $\sigma_i$. We will explicitly evaluate the integral (5.12) in the next section. But first, having placed our estimator in the context of Bayesian reasoning, some remarks are in order.

Given a closed form formula for the likelihood of the structural (and incidental) parameters, one can employ standard methods — such as gradient descent or the Newton-Raphson method — to single out its maximum. However, the Bayesian approach has several advantages. First of all, depending on the further data processing chain, it is not always necessary to come up with a single estimate for the involved parameters. Rather the likelihood formula provides a means to assign a probability value to each possible set of parameters. Additionally, in many cases the maximum likelihood estimate may not represent a typical outcome. In this respect probability mass is more indicative than the magnitude of the likelihood (which may have a very high, yet very narrow peak), and it can only be determined by integrating the likelihood over a certain range. Furthermore, even if the main interest lies in a specific estimate, one can compute the variance of the estimate, which gives a clue about the confidence in it. This can be very useful in order to base further decisions on the result or in keeping track of the error buildup in processing uncertain data. Also, given a closed form of the likelihood, it is possible to marginalize out certain parameters that may be uninteresting in specific applications (known as nuisance parameters) or to incorporate prior knowledge about the distribution of parameters.
5.4. Evaluation of the Closed Form Likelihood

Before we write out the integral (5.12) we shall give some geometrical intuition regarding the derivation of the result. See Figure 5.3 for an example in three dimensions. Imagine a straight line from the hypersphere’s center $\theta_c$ through the observed point $x_i$. The intersection point of that line with the hypersphere $s$ is one of many candidates for the true point that generated the observation. From $\theta_c$, a candidate $\alpha_i$ and the observed point $x_i$ are viewed under an angle $\tau_i$. Now, because the assumed observation noise is isotropic, we can express the probability of $\alpha_i$ having generated $x_i$ purely in terms of their mutual distance $d$, which only depends on $\tau_i$. This means that a rotation of the configuration about the imagined line joining $\theta_c$ and $x_i$ does not change the probability. More specifically, every point on a hypersphere with one dimension less and a radius that only depends on $\tau_i$ contributes the same probability. Therefore, as we let the viewed angle $\tau_i$ run from 0 to $\pi$ we integrate the probabilities over the whole hypersphere as in (5.12).

Indeed, this observation, which greatly simplifies calculations, is the main reason why we assume both isotropic noise on the observations, as well as uniform distribution of the true points around the hypersphere. With anisotropic noise the probability of an $\alpha_i$ having generated $x_i$ would not only depend on their distance, but also on their direction. While — even under isotropic noise — without uniform distribution of $\alpha_i$ a rotation about the imagined line would change that probability.

The squared distance, $d(\tau_i)^2$, between the observation $x_i$ and $\alpha_i$ is

$$d(\tau_i)^2 = (\|\theta_c - x_i\| - \theta_r \cos \tau_i)^2 + (\theta_r \sin \tau_i)^2 = \|\theta_c - x_i\|^2 + \theta_r^2 - 2 \|\theta_c - x_i\| \theta_r \cos \tau_i$$

(5.13)

Let $n$ be the dimensionality of the space of observed points. The hypersphere we are trying to estimate will have dimension $n - 1$, whereas the hypersphere of true points which could have generated the observation for fixed $\tau_i$ has dimension $n - 2$ and radius $(\theta_r \sin \tau_i)$. The surface
area of a hypersphere of co-dimension 2 is given by the following expression.

\[
A(n - 2, \theta_r \sin \tau_i) = (n - 1) \frac{\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} (\theta_r \sin \tau_i)^{n-2}
\]

\[
= \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} (\theta_r \sin \tau_i)^{n-2},
\]

(5.14)

where \(\Gamma\) denotes the *gamma function* and we used the fact that \(\frac{\Gamma\left(\frac{n+1}{2}\right)}{(n - 1)} = \frac{1}{2} \Gamma\left(\frac{n-1}{2}\right)\).

Since we assume the observation noise to be Gaussian, our integral (5.12) evaluates to

\[
L_{x_i, \sigma_i}(\theta_c, \theta_r) = \int_0^\pi \frac{1}{\sqrt{(2\pi \sigma_i^2)^n}} \exp\left(-\frac{d(\tau_i)^2}{2\sigma_i^2}\right) A(n - 2, \theta_r \sin \tau_i) \, d\tau_i
\]

\[
= \int_0^\pi \frac{1}{\sqrt{(2\pi \sigma_i^2)^n}} \cdot \exp\left(-\frac{||\theta_c - x_i||^2 + \theta_r^2}{2\sigma_i^2} - 2\sqrt{||\theta_c - x_i||^2 \theta_r \cos \tau_i}\right)
\]

\[
\cdot \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} (\theta_r \sin \tau_i)^{n-2} \, d\tau_i
\]

\[
= \frac{1}{\sqrt{(2\pi \sigma_i^2)^n}} \exp\left(-\frac{||\theta_c - x_i||^2 + \theta_r^2}{2\sigma_i^2}\right) \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \theta_r^{n-2}
\]

\[
\cdot \int_0^\pi \exp\left(\frac{\sqrt{||\theta_c - x_i||^2 \theta_r \cos \tau_i}}{\sigma_i^2} \sin \tau_i\right) (\sin \tau_i)^{n-2} \, d\tau_i
\]

\[
= \frac{1}{\sqrt{(2\pi \sigma_i^2)^n}} \exp\left(-\frac{||\theta_c - x_i||^2 + \theta_r^2}{2\sigma_i^2}\right) \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \theta_r^{n-2}
\]

\[
\cdot \sqrt{n} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} {}_0F_1\left(\frac{n}{2}, \frac{||\theta_c - x_i||^2 \theta_r^2}{4\sigma_i^2}\right)
\]

\[
= \frac{2\theta_r^{n-2}}{\sqrt{(2\pi \sigma_i^2)^n} \Gamma\left(\frac{n}{2}\right)} \exp\left(-\frac{||\theta_c - x_i||^2 + \theta_r^2}{2\sigma_i^2}\right)
\]

\[
\cdot {}_0F_1\left(\frac{n}{2}, \frac{||\theta_c - x_i||^2 \theta_r^2}{4\sigma_i^2}\right),
\]

(5.15)

where \(\theta_0\) denotes the *confluent hypergeometric limit function* (see Appendix C).

We will assume the observations \(x_i\) to be statistically independent. In order to investigate a simplified case of practical relevance in detail, we will also assume the noise level for each observation to be the same, i.e. \(\sigma_i = \sigma, i = 1, \ldots, N\). The joint likelihood for \(N\) point
observations is therefore given by

\[
L_{x, \sigma}(\theta_c, \theta_r) = \prod_{i=1}^{N} L_{x_i}(\theta_c, \theta_r) = \left( \frac{2\theta_r^{n-2}}{\sqrt{(2\sigma^2)^n \Gamma(n/2)}} \right)^N \exp \left( -\frac{N\theta_r^2 + \sum_{i=1}^{N} \|\theta_c - x_i\|^2}{2\sigma^2} \right) 
\cdot \prod_{i=1}^{N} F_1 \left( \frac{n}{2}; \frac{\|\theta_c - x_i\|^2 \theta_r^2}{4\sigma^4} \right).
\] (5.16)

Maximum likelihood estimation in this case is obviously highly non-linear in the structural parameters to be estimated. In order to employ, say, a gradient descent method to search for the maximum, one needs to know the gradient of the likelihood function. We therefore calculate the first derivatives with respect to \(\theta_c\) and \(\theta_r\), respectively.

\[
\nabla_{\theta_c} L_{x, \sigma}(\theta_c, \theta_r) = L_{x, \sigma}(\theta_c, \theta_r) \sum_{i=1}^{N} \left( \frac{\theta_c - x_i}{\sigma^2} + \frac{(\theta_c - x_i)\theta_r^2}{n\sigma^4} F_1 \left( 1 + \frac{n}{2}, \frac{\|\theta_c - x_i\|^2 \theta_r^2}{4\sigma^4} \right) \right),
\]
(5.17)

\[
\nabla_{\theta_r} L_{x, \sigma}(\theta_c, \theta_r) = L_{x, \sigma}(\theta_c, \theta_r) \left( \frac{N(n-2)}{\theta_r^2} - \frac{N\theta_r}{\sigma^2} \right)
+ \sum_{i=1}^{N} \frac{\theta_r\|\theta_c - x_i\|^2}{n\sigma^4} F_1 \left( 1 + \frac{n}{2}, \frac{\|\theta_c - x_i\|^2 \theta_r^2}{4\sigma^4} \right),
\]
(5.18)

where we have used \((\partial/\partial x) F_1(n/2, x) = 2 F_1(1 + n/2, x)/n\).

The fact that the gradient is a multiple of the joint likelihood works to our advantage. Firstly, when looking for zeros of (5.17) and (5.18), we are looking for parameters for which either the sum term or \(L_{x, \sigma}(\theta_c, \theta_r)\) is zero. Since the likelihood is essentially a probability it will not be zero at its maximum, so we only need to find parameters for which the sum term is zero. From a computational point of view, focusing only on the sum yields a numerically much more well behaved gradient. Additionally, the joint likelihood is \textit{very} small, such that multiplying by it could introduce numerical instability. Usually this problem is tackled by taking the logarithm of the joint likelihood (yielding the log-likelihood), such that, instead of having to evaluate a large product as in (5.16), one evaluates a large sum.

5.5. Results

The first test that our formula has to pass is the following. Does it agree with the previously known likelihood formula for 2D circle estimation? If we substitute \(n = 2\) into (5.16) we get

\[
L_{x, \sigma}(\theta_c, \theta_r) = \left( \frac{1}{2\sigma^2} \right)^2 \exp \left( -\frac{N\theta_r^2 + \sum_{i=1}^{N} \|\theta_c - x_i\|^2}{2\sigma^2} \right) \prod_{i=1}^{N} I_0 \left( \frac{\sqrt{\|\theta_c - x_i\|^2 \theta_r^2}}{\sigma^2} \right).
\] (5.19)
where \( I_0 \) denotes the modified Bessel function of the first kind (see also Appendix C). This agrees — up to a multiplicative constant — with the likelihood formula cited by Anderson, who gives the log-likelihood for the circle’s center in [An81].

Of special interest to us is the case of \( n = 3 \), because it is of geometric importance in many applications and the highest dimension which is still easily visualizable. The explicit likelihood formula becomes

\[
L_{x, \sigma}(\theta_c, \theta_r) = \left( \frac{4\theta_r}{\sqrt{2\pi \sigma^2}} \right)^N \exp \left( - \frac{N\theta_r^2 + \sum_{i=1}^N \| \theta_c - x_i \|^2}{2\sigma^2} \right) \prod_{i=1}^N \frac{\sigma^2 \sinh \left( \frac{\| \theta_c - x_i \|^2 \theta_r}{\sigma^2} \right)}{\| \theta_c - x_i \|^2}.
\]

Using (5.20) we implemented a maximum likelihood estimator for the sphere parameters in order to investigate its accuracy and its efficiency under varying conditions. Among other things we are interested in the behavior of the maximum likelihood estimator when the assumption of uniform distribution which we used to derive (5.20) is not reflected by observations on a true sphere. To this end we have run a number of experiments with simulated data. The simulated true points were either uniformly distributed over the surface of the sphere or they were distributed according to a normal distribution with varying standard deviations. In detail, we set up our experiments as follows.

We assume a true 3D sphere with radius 10 centered at the origin. On this sphere we sample a number of points by introducing a reference plane which can rotate around the \( y \)-axis through an angle that we randomly sample from a uniform distribution between \(-\pi/2\) and \(\pi/2\). From this plane we determine an azimuth angle sampled from a normal distribution with zero mean and a standard deviation which we call the spread parameter \( q \). Should this latter sample lie outside the interval \([-\pi, \pi]\), we just discard it. The fact that the resulting points are not samples of a normal distribution does not matter. Our main focus is on obtaining some concentration of the sampled points, not on their specific distribution. If it falls inside the interval we accept the resulting point and add isotropic Gaussian noise with zero mean and variance \( \sigma^2 \) to its three Cartesian coordinates. See Figure 5.5 for typical sample distributions.

We repeat this process \( N \) times and take the \( N \) resulting points as input for the likelihood formula (5.20). We find a maximum by gradient descent using the gradient formulas (5.17) and (5.18).

In addition, we perform an experiment where the reference plane is rotated through a random angle uniformly distributed between \(-\pi\) and \(\pi\) and the azimuth angle distributed according to \( \cos \varphi \) with \( \varphi \) between 0 and \( \pi/2 \). This results in a distribution of sample points over the
A solid sphere that is imaged under (approximately) parallel projection. Points on the half sphere turned away from the sensor are not visible due to self occlusion. Points at the “apex”, closest to the sensor, have a higher chance to be observed than points on the “equator”. More precisely the probability of a point being observed is distributed according to \( \cos \varphi \) on the interval of \( \varphi \in [0, \pi/2] \). By rotating this figure out of the drawing plane we obtain the setup in 3D described in the text.

surface of half the sphere with higher chance of occurring near the “apex” and lower chance of occurring near the equator. The setup emulates a situation that occurs in practice when a solid sphere is imaged under (approximately) parallel projection, for example, by a camera, a laser range finder or a similar device. In these cases half the sphere — namely the part turned away from the sensor — is not visible and it is more likely to observe points on the sphere surface where the sphere normal points in the direction of the sensor. See Figure 5.4 for a visualization of this setup.

For our experiments we vary \( N \in \{4, 10, 20, 100\} \), \( q \in \{0.1, 0.2, 0.5, 1\} \) and \( \sigma \in \{0.1, 0.5, 1, 2, 5\} \). For each of the resulting \( 4 \times 5 \times 5 \) cases we perform 5000 independent runs. Additionally, we perform the same experiment with sample points uniformly distributed around the true sphere, varying \( N \) and \( \sigma \).

As a measure of the quality of our estimator we calculate its bias, i.e. the difference between the expectation of the estimated parameters \( \hat{\theta} \) and their true value \( \theta_0 \).

\[
\text{bias}[\hat{\theta}] = E[\hat{\theta}] - \theta_0
\]

Since we are estimating multiple parameters, this will be a vector. We will calculate the magnitude, \( b \), of that vector using the Euclidean metric. Other measures of quality are the trace, \( \text{tr} \), as well as the determinant, \( \det \), of the covariance matrix \( \Sigma(\hat{\theta}) = \text{cov}(\hat{\theta}) \). Instead of calculating them analytically, we approximate all these values by the sample mean and the sample covariance of the 5000 runs.

For comparison we implemented a non-linear least squares estimator for the sphere parameters \( \theta \), which uses gradient descent to minimize the sum of squared orthogonal distances between the observed points and the estimated sphere. We will denote results pertaining to the maximum likelihood estimator (respectively, the non-linear least squares estimator) by a subscript MLE (respectively, subscript NLS). For detailed numerical results see Appendix B.
5.5 Results

In the tables summarizing the results, we find several things worth remarking on. Firstly, we see that both estimators exhibit a bias that is more severe the more concentrated the sample points are on the sphere. As the spread parameter increases, the bias reduces significantly with an increasing number of points. We would like to point out that this is not an artifact of the type of estimators we used (i.e. maximum likelihood vs. least squares estimator), but of the assumptions that we made. More precisely, we assumed that the sample points are uniformly distributed around the sphere and therefore our estimator will try to estimate a sphere around

\textbf{Figure 5.5.:} 20 points sampled from a sphere at the origin with radius 10. The points’ spread around the point \((0, 10, 0)\) on the sphere varies. The sampled points have added Gaussian noise with standard deviation \(\sigma = 0.1\) on their Cartesian coordinates.
A Bayesian Approach to Hypersphere Estimation

Figure 5.6: Some estimates. The blue sphere is the true sphere of radius 10 from which the observed points were sampled. The red sphere indicates the non-linear least squares estimate, while the green one shows the maximum likelihood estimate. Images are scaled to fit the respective plot.

which the points are uniformly distributed. In Figure 5.6 on the right we see the effect of this assumption.
5.5 Results

![Figure 5.7](image_url)

**Figure 5.7:** The estimators’ bias as a function of the point spread parameter for 20 points around a sphere at the origin with radius 10. The bias has been calculated using the sample mean of 5000 independent estimations.

The images on the left of Figure 5.6 do not exhibit this behavior that drastically, because we have very low noise. The observed points provide stronger evidence for the true hypersphere they were sampled from. The amount of noise is another assumption that we make, and in Figure 5.6 we can clearly see those two assumptions compete. On the left the low noise assumption dominates the estimate, while on the right the assumption of uniform distribution takes over.

Note that while the bias of the two estimators can differ by as much as a factor of five (Table B.1, spread 0.1, 100 points, $\sigma = 0.5$) they are usually of the same order of magnitude. In general, the maximum likelihood estimator seems to lead to a slightly higher bias. This trend reverses (and is particularly pronounced when the number of point observations is small), when the points are indeed uniformly distributed around the sphere (compare the b-values in individual cells of Table B.5), which serves to show the influence of the violated assumption in the other cases. Figure 5.7 shows the bias as a function of the degree to which the assumption is violated. The bias has been calculated empirically, using the sample mean of 5000 runs.

The determinant of the covariance matrix is a measure for the hypervolume of the covariance ellipsoid. Therefore we expect it to be growing with increasing dimensionality, in general. The least extreme difference between the two compared estimators we found is a factor of about 1.17 in favor of the maximum likelihood estimator (Table B.4, spread 1.0, 4 points, $\sigma = 0.5$), but usually the two differ by many orders of magnitude, which can be as extreme as $10^{28}$ (Table B.1, spread 0.1, 100 points, $\sigma = 0.5$).

Numerical results for the experiment with points distributed according to a cosine function (recall Figure 5.4) can be found in table B.6. As one might expect, in practice as well as in our simulation experiments this particular distribution is similar to our artificial distribution described earlier, but with a spread parameter between $q = 0.5$ and $q = 1.0$. The numerical results support this intuition.

The empirical results seem to indicate that the more extreme the concentration of sample points the larger the difference between the determinants of the covariance matrices becomes. But this explanation is a little too simplistic. Since — due to the assumption of a uniform
5.6. Conclusion

In this chapter we have applied the Bayesian approach towards inference problems to the estimation of hyperspheres from point data. We have derived an explicit likelihood formula for the hypersphere parameters “radius” and “location of center” in \( n \) dimensions and calculated the likelihood gradient, which we used to obtain a maximum likelihood estimate for 3D spheres.

The Bayesian approach has several advantages, one of which is to force the author to make explicit all assumptions in the form of prior probabilities. Our assumptions about the esti-
5.6 Conclusion

Figure 5.9.: The ratio of determinants $\sqrt{\text{det}_{\text{NLS}}/\text{det}_{\text{MLE}}}$, ordered by ascending quotient $\sigma^2/\rho^2$, with $\rho$ denoting the mean radius of the estimated sphere. Note, that the horizontal axis is not linear. It ranges from $10^{-4} \times 1.377$ to $10^4 \times 7.053$ with about half the values $\leq 1$.

Information problem are that observed points are degraded by zero mean isotropic Gaussian noise whose variance is known (and therefore does not have to be estimated) as well as uniform distribution of the observed points around the hypersphere to be estimated. We have shown that it is these assumptions — rather than the nature of the problem or the type of estimator — that are responsible for the observed bias of the estimator, if points are not uniformly distributed around the hypersphere.

These assumptions were made because either they are justifiable in many practical applications — e.g. Gaussian noise as a consequence of the central limit theorem — or because they greatly simplify the derivation of the formulas — e.g. uniform distribution and isotropic noise simplify the involved integrals.

Our empirical results indicate that, given the same amount of data, a maximum likelihood estimator has much lower uncertainty about the estimated hypersphere parameters than a non-linear least squares estimator. Among other factors, this is due to the fact that our estimator treats the amount of observation noise as given, whereas a least squares estimator tries to minimize that as part of the estimation process. Alternatively, a maximum likelihood estimator can reach the same level of uncertainty as a least squares estimator with a lower amount of data, i.e. it is more efficient.

In order to harvest the benefits of the full Bayesian approach in a specific application some modifications may have to be made. Firstly, it should be carefully determined which parameters are to be estimated and which parameters are incidental or should be regarded as nuisance parameters. Secondly, all prior knowledge (or belief) about the distribution of the parameters should be incorporated in the (likelihood) formulas. Note, that in some settings this can change “online” as more data is observed.

Finally, in this chapter we analyzed the behavior of a maximum likelihood estimator for hyperspheres, which results in a single output. However, if further decisions or estimates should be based on this output it is better not to home in on a single estimate prematurely.
Preferably one should retain the full likelihood function and propagate it properly through the estimation process.