EPFL

Thèse n° 9493

Higher Order Asymptotics: Applications to Satellite Conjunction and Boundary Problems

Présentée le 26 mai 2023

Faculté des sciences de base
Chaire de statistique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Soumaya ELKANTASSI

Acceptée sur proposition du jury

Prof. M. Viazovska, présidente du jury
Prof. A. C. Davison, directeur de thèse
Dr M. Hejduk, rapporteur
Prof. V. Chavez, rapporteuse
Prof. M. Stensrud, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

Acknowledgements

I am deeply grateful to everyone who has supported and encouraged me throughout my Ph.D. journey.

Firstly, I would like to thank my supervisor, Prof. Anthony Davison, for his invaluable guidance, support, and mentorship. His expertise, encouragement, and optimism have been crucial in helping me navigate the complex and challenging path of doctoral research.

I am also grateful to the members of my thesis committee, Prof. Maryna Viazovska, Dr. Matthew Hejduk, Prof. Mats Stensrud and Prof. Valérie Chavez for their insightful discussions and constructive feedback. Additionally, I would like to extend my appreciation to Prof. Alessandra Brazzale for her invitation to visit the Department of Statistical Science at the University of Padova. Her valuable knowledge and experience have enriched my work.

Furthermore, I owe a great deal of gratitude to my current EPFL colleagues, Sonia, Tim, Mario, Laya, and Kartik as well as former members Jonathan, Stefano, Léo, Arwa, Thomas, Hélène, Max, Neda, and Quentin, for creating a welcoming environment and making my stay in Switzerland more enjoyable.

I would also like to express my heartfelt thanks to my parents Khadija and Mohamed, my brothers Haythem, Oussema, and Ahmed, and my friends Arbia, Yiannis, Lobna, Ismail and Zied for their unconditional support and encouragement throughout my Ph.D. journey. Their love and encouragement have been a source of strength during the most challenging times. I am also extremely grateful to have had the support of Dimitrios and my amazing Rika, without whom this journey wouldn't have been nearly as enjoyable or successful.

Finally, I am deeply humbled by the opportunity to pursue a Ph.D., and I would like to thank all my professors who have shaped and enriched my academic profile in Tunisia and KAUST, especially Prof. Raul Tempone. I am grateful to the institute of Mathematics at EPFL for providing me with the resources and opportunities to pursue my research, and for supporting my work.

Lausanne, May 18, 2023 Soumaya Elkantassi

Abstract

Higher-order asymptotics provide accurate approximations for use in parametric statistical modelling. In this thesis, we investigate using higher-order approximations in two specific settings, with a particular emphasis on the tangent exponential model.

The first chapter introduces first-order asymptotic theory and reviews key concepts such as sufficiency, significance, and exponential families. We then discuss higher-order approximations, which have been studied by many authors. The literature is rich with examples demonstrating the limitations of first-order methods when applied to models with many nuisance parameters and showcasing the increased accuracy of higher-order approximations.

The second chapter concerns collision assessment of space objects. Satellite conjunctions involving 'near misses' are becoming increasingly likely. A common approach to risk analysis involves the computation of the collision probability, but this has been regarded as having some counter-intuitive properties, and its interpretation has been debated. We formulate an approach to satellite conjunction based on a simple statistical model and discuss inference on the miss distance between the two objects, for linear and non-linear motion. We point out that the usual collision probability estimate can be badly biased, but highly accurate inference on the miss distance is possible using the tangent exponential model. The ideas are illustrated with case studies and Monte Carlo results that show its excellent performance.

In the third chapter we study statistics used to test hypotheses concerning parameters on the boundary of their domain. These often have non-standard limiting distributions, which may be poor finite-sample approximations even when the sample size is very large. We distinguish soft and hard boundary problems, discuss elementary approached to both and describe an approach to small-sample approximation based on the tangent exponential model. Numerical results show that the approach can give much improved approximations, even in small samples.

We finish the thesis with ideas for future research in the field of particle physics, including some preliminary results.

Keywords: boundary problems, higher-order asymptotics, hypothesis testing, likelihood, satellite conjunction assessment, tangent exponential model

Résumé

Les asymptotiques d'ordre supérieur fournissent des approximations précises dans le cadre du traitement statistique de modèles paramétriques. Dans cette thèse, nous étudions l'utilisation des approximations d'ordre supérieur dans deux contextes spécifiques, en mettant particulièrement l'accent sur le modèle exponentiel tangent.

Le premier chapitre présente la théorie asymptotique du premier ordre et passe en revue les concepts clés tels que la suffisance, la signification et les familles exponentielles. Nous discutons ensuite des approximations d'ordre supérieur, qui ont été étudiées par de nombreux auteurs. La littérature est riche d'exemples démontrant les limites des méthodes de premier ordre appliquées à des modèles avec plusieurs paramètres de nuisance et mettant en évidence la précision améliorée des approximations d'ordre supérieur.

Le deuxième chapitre concerne l'évaluation des collisions d'objets spatiaux. Les conjonctions de satellites impliquant des "rencontres rapprochées" sont de plus en plus probables. Une approche courante de l'analyse des risques implique le calcul de la probabilité de collision, mais celle-ci a été considérée comme ayant des propriétés contre-intuitives, et son interprétation a été débattue. Nous formulons une approche de la conjonction des satellites basée sur un modèle statistique simple et discutons l'inférence sur la distance critique entre les deux objets, pour des trajectoires linéaires et non linéaires. Nous soulignons que l'estimation habituelle de la probabilité de collision peut être fortement biaisée, mais qu'une inférence très précise de cette distance est possible en utilisant le modèle exponentiel tangent. Les idées sont illustrées par des études de cas et des résultats de Monte Carlo qui montrent d'excellentes performances.

Dans le troisième chapitre, nous étudions les statistiques utilisées pour tester les hypothèses concernant des paramètres à la limite de leur domaine. Ceux-ci ont souvent des distributions limites non standard, qui peuvent être de mauvaises approximations d'échantillon fini même lorsque la taille de l'échantillon est très grande. Nous distinguons les problèmes de limites souples et dures, discutons des approches élémentaires pour les deux et décrivons une approche de l'approximation à petit échantillon basée sur le modèle exponentiel tangent. Les résultats numériques montrent que cette approche peut donner des approximations bien meilleures, même pour de petits

échantillons.

Nous concluons avec des idées de recherches futures dans le domaine de la physique des particules, en incluant quelques résultats préliminaires.

Keywords : valeur limite de l'espace de paramètre, asymptotique d'ordre supérieur, tests d'hypothèse, vraisemblance, évaluation de la conjonction de satellites, modèle exponentiel tangent

Contents

An Introduction to Likelihood-Based Inference

"What has now appeared is that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making inferences, and that the mathematical quantity which usually appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "Likelihood" to designate this quantity; since both the words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship."

R. A. Fisher, Statistical Methods for Research Workers, 1925.

1.1 Introduction

The idea of likelihood has been around for centuries, with early suggestions dating back to Lambert in 1790. However, it was formally introduced and developed by Fisher in 1922 (Fisher, 1922). Since then, the concept of likelihood has become the basis of the most widely used method of statistical estimation and has been applied to a vast range of problems in various scientific fields. It plays a crucial role in statistical theory, methodology, and applications (Barndorff-Nielsen and Cox, 1989; Berger and Wolpert, 1988; Severini, 2000). Throughout its development, the focus of statistical inference shifted from simply summarizing the data to treating decision problems, using concepts such as hypothesis testing and point and interval estimation. To ensure the clarity and reproducibility of the mathematical formulation, it was necessary to identify optimal methods for making inferences even before the observations were collected, following the principle of repeated sampling. This approach, known as the *frequency-decision paradigm*, aims to minimize ambiguity in interpretation and ensure the reliability of the results. Many of the concepts and methods in these

developments were influenced by pioneers such as Neyman (Neyman, 1937; Neyman and Scott, 1948), Egon Pearson (Neyman and Pearson, 1928; Pearson, 1936), Wald (Wald, 1949), Lehmann (Lehmann, 1983, 1986) and others who were inspired by Fisher's earlier major contributions.

Over the past century, numerous contributions have furthered the development of likelihood-based theory. This evolution has often involved the use of asymptotic mathematics and arguments, which are common in both frequency-based and Bayesian theory. The purpose of this chapter is to introduce some basic concepts of likelihood-based inference and to review the approximations that will play a central role in the subsequent chapters. We also introduce some notation and terminology that will later be taken as prerequisites.

First order asymptotics involves linearizing the log likelihood and using the central limit theorem to obtain results related to maximum likelihood estimates, score tests, likelihood ratio tests, etc. In this context, a variety of statistics for testing a null hypothesis $\theta = \theta_0$ are available, differing by $O_p(n^{-1/2})$. In regular parametric models when there is an arbitrarily large amount of data, these statistics produce similar results, at least in theory. This is not always the case in practical use, and some statistics give qualitatively more sensible results than others. In choosing between them, one should pay special attention to the parametrization, as invariance considerations seem compelling for confidence regions (Barndorff-Nielsen and Cox, 1994, Chapter 4). First-order approximations will be discussed in Section 1.2.

In many models, the parameter vector can be divided into two parts: the *parameter of interest*, ψ, and the *nuisance parameter*, λ. The parameter of interest, ψ, is the primary focus of the study, while the nuisance parameter, λ, represents aspects of the model that are necessary for accurate modeling but not the main focus. Typically, ψ has a small number of dimensions, and often is scalar, while λ may be a high dimensional vector. In this type of setting, it is desirable to base inference on a function of the data and ψ that has properties similar to a likelihood function when there is no nuisance parameter. A natural candidate for this is the profile likelihood function, whose properties are illustrated in Severini (1998) and Pace and Salvan (1997, Chapter 4).

The profile likelihood provides a first-order approximation, but its inferential accuracy may be unsatisfactory, especially when the number of nuisance parameters is large and the sample size is small, so modifications have been proposed in an effort to improve its accuracy. One particular modification was proposed by Barndorff-Nielsen (1980, 1983, 1988), and involves the use of *higher-order asymptotic methods*. This approach is discussed in Section 1.5. Before that, in Section 1.3 , we briefly review exponential families and elaborate the idea of reduction to a marginal model and

reduction to a conditional model, given a distribution-constant statistic; these are formally unified under the factorization theorem (Jørgensen, 1994). Higher-order asymptotics are often obtained using a combination of techniques, such as Taylor series expansion of the log likelihood, asymptotic expansions for cumulants, and the Laplace expansion for evaluating integrals. These techniques and their connections to other fundamental approximations are discussed in Section 1.4.

In Section 1.6, we present a simplified version of the Barndorff-Nielsen approximation that was developed by Fraser and his colleagues in a series of articles (Fraser and Reid, 1993, 1995; Fraser et al., 1999a) called the *tangent exponential model* . This section is largely inspired by the work of Davison and Reid (2022), which provides a clear understanding of the model and illustrates the concept using straightforward examples. In addition to the refinements proposed in Cox and Reid (1987), Cox (1975), and Owen (2001), several other enhancements to the profile likelihood have been suggested in the literature and will be mentioned throughout the chapter.

1.2 General concepts

In this section, we discuss first-order asymptotic theory for likelihood-based methods, which forms the foundation for the higher-order approximations described in Section 1.5 and 1.6. These approximations hold as the sample size n becomes large; here n is used as an index of the amount of information provided by the data, and the assumption of n going to infinity allows us to obtain an approximation that can be used for finite sample size.

1.2.1 First-order theory

We consider a vector $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ of continuous responses and a statistical model for Y with joint density function $f(y;\theta)$ that depends on a parameter $\theta \in \mathbb{R}$. The likelihood function is proportional to the density evaluated at the observed data y° and regarded as a function of the unknown parameter θ, i.e.,

$$L(\theta) = L\left(\theta; y^\circ\right) = c(y^\circ) f\left(y^\circ; \theta\right),$$

with $c(\cdot)$ an arbitrary function, we use y to denote a generic response vector and y° its observed value, and $\widehat{\theta}$ is the maximum likelihood estimator and $\widehat{\theta}^\circ$ the maximum likelihood estimate computed from y°. Fisher (1956, Chapter 3) suggested using likelihood ratios to declare regions of the parameter space that are "very plausible", "somewhat plausible", and "highly implausible" according to a specified threshold. The

maximum likelihood estimator $\widehat{\theta}$ is often used as a reference point in this assessment. It is defined as the solution to the score equation, $\partial\ell(\theta)/\partial\theta = 0$, and under mild regularity conditions, which we discuss later, it has an asymptotic normal distribution, as $n \to \infty$. Under these conditions, the approximate normal distribution of $\widehat{\theta}$ is centered at the true parameter θ with variance $1/\jmath(\widehat{\theta})$, where $\jmath(\theta) = -\partial^2\ell(\theta)/\partial\theta^2$ is the observed information. This allows calibration of θ using the score statistic

$$s(\theta) = \jmath\left(\widehat{\theta}\right)^{-1/2} \frac{\partial\ell(\theta)}{\partial\theta},$$

corresponding to a linear expansion of the score function, the Wald statistic

$$w(\theta) = \jmath\left(\widehat{\theta}\right)^{1/2}\left(\widehat{\theta} - \theta\right),$$

or the likelihood root

$$r(\theta) = \text{sign}\left(\widehat{\theta} - \theta\right)\left\{2\ell(\widehat{\theta}) - 2\ell(\theta)\right\}^{1/2},$$

corresponding to a quadratic approximation to the log likelihood function.

The quantities $s(\theta)$, $r(\theta)$ and $w(\theta)$ are approximate pivots: they are functions of the data and parameter and have approximate standard normal distributions under repeated sampling if θ equals its true value. The approximations introduce so-called first-order error, of size $O(n^{-1/2})$. For more details, see, for example, Chapter 9 of Cox and Hinkley (1974). For details on using the likelihood function directly for inference see for example Royall (1997) or Burnham and Anderson (2002).

1.2.2 Several parameters

Suppose now that the density function $f(y;\theta)$ depends on an unknown d-dimensional parameter θ, which comprises a scalar interest parameter ψ and a nuisance parameter λ; ψ and λ are supposed to be variation independent. As λ is unknown, it must be replaced by an estimate, and this introduces errors. In this situation the maximum likelihood estimate $\widehat{\theta}^{\text{o}} = (\widehat{\psi}^{\text{o}}, \widehat{\lambda}^{\text{o}})$ maximises the log likelihood $\ell(\theta) = \log f(y^{\text{o}};\theta)$ with respect to θ, and the partial maximum likelihood estimate $\widehat{\theta}^{\text{o}}_{\psi} = (\psi, \widehat{\lambda}^{\text{o}}_{\psi})$ maximizes $\ell(\theta)$ with respect to λ for fixed ψ. The large-sample properties of the maximum likelihood estimator $\widehat{\theta}$ under repeated sampling are well-established (Cox and Hinkley, 1974, Chapter 9): as the sample size $n \to \infty$ and under the regularity conditions given below, $\widehat{\theta}$ has an approximate d-dimensional normal distribution with mean the true parameter θ and covariance matrix $\jmath(\widehat{\theta})^{-1}$, where $\jmath(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^{\text{T}}$ is the $d \times d$ observed information matrix and θ^{T} denotes the transpose of the $d \times 1$ vector θ.

The usual regularity conditions are

- $\mathscr{C}_1$: the true value θ^{o} of θ is interior to its parameter space Θ, which has finite dimension and is compact;

- $\mathscr{C}_2$: the densities defined by any two different values of θ are distinct;

- $\mathscr{C}_3$: there is a neighbourhood δ of θ^{o} within which the first three derivatives of the log likelihood with respect to θ exist almost surely, and for $r,s,t=1,\ldots,d$, $n^{-1}\mathrm{E}\left\{\left|\partial^3\ell(\theta)/\partial\theta_r\partial\theta_s\partial\theta_t\right|\right\}$ is uniformly bounded for $\theta\in\delta$;

- $\mathscr{C}_4$: the Fisher information matrix $i(\theta)=\mathrm{E}[J(\theta)]$ is finite and positive definite within δ, and its elements satisfy

$$i(\theta)_{rs}=\mathrm{E}\left\{\frac{\partial\ell(\theta)}{\partial\theta_r}\frac{\partial\ell(\theta)}{\partial\theta_s}\right\}=\mathrm{E}\left\{-\frac{\partial^2\ell(\theta)}{\partial\theta_r\partial\theta_s}\right\},\qquad r,s=1,\ldots,d.$$

In practice, conditions $\mathscr{C}_1-\mathscr{C}_4$ can be violated in various ways. Chapter 3 of the thesis will delve into boundary problems, which occur when condition $\mathscr{C}_1$ is not satisfied, and examine certain non-regular models in more detail. Under the above conditions, the error committed by replacing parameters in (1.4) by their estimates is $O(n^{-1/2})$, giving first-order approximations, and the same error is committed by treating

$$\text{the likelihood root}\quad r(\psi)=\mathrm{sign}(\widehat{\psi}-\psi)\left[2\left\{\ell(\widehat{\theta})-\ell(\widehat{\theta}_\psi)\right\}\right]^{1/2}, \tag{1.1}$$

$$\text{the Wald statistic}\quad w(\psi)=J_{\mathrm{p}}(\widehat{\psi})^{1/2}(\widehat{\psi}-\psi), \tag{1.2}$$

$$\text{the score statistic}\quad s(\psi)=J_{\mathrm{p}}(\widehat{\psi})^{-1/2}\frac{\partial\ell(\widehat{\theta}_\psi)}{\partial\psi}, \tag{1.3}$$

as standard normal; here J_{p} is the observed information function for the profile likelihood which can be expressed in terms of the observed information as

$$\begin{aligned}J_{\mathrm{p}}(\psi)&=J_{\psi\psi}\left(\widehat{\theta}_\psi\right)-J_{\psi\lambda}\left(\widehat{\theta}_\psi\right)J^{\lambda\lambda}\left(\widehat{\theta}_\psi\right)J_{\lambda\psi}\left(\widehat{\theta}_\psi\right)\\&=\frac{\left|J\left(\widehat{\theta}_\psi\right)\right|}{\left|J_{\lambda\lambda}\left(\widehat{\theta}_\psi\right)\right|},\end{aligned}$$

where $|\cdot|$ indicates the determinant. The last identity holds for scalar ψ using the partition of the observed information matrix and its inverse into the (ψ,λ) blocks, i.e.,

$$J(\theta)=\begin{pmatrix}J_{\psi\psi}(\theta) & J_{\psi\lambda}(\theta)\\ J_{\lambda\psi}(\theta) & J_{\lambda\lambda}(\theta)\end{pmatrix},\quad J(\theta)^{-1}=\begin{pmatrix}J^{\psi\psi}(\theta) & J^{\psi\lambda}(\theta)\\ J^{\lambda\psi}(\theta) & J^{\lambda\lambda}(\theta)\end{pmatrix}.$$

1.2.3 Significance, sufficiency and ancillarity

A key tool for statistical inference on a scalar parameter θ is the significance function

$$p^{o}(\theta) = \Pr(\widehat{\theta} \leq \widehat{\theta}^{o}; \theta), \tag{1.4}$$

which can be used for inference about θ based on observed data y^{o}. It can be used to determine the limits of a confidence interval for θ, and compare the relative strength of evidence for different hypotheses (Fraser, 2019). The pivots $w(\theta)$, $s(\theta)$ and $r(\theta)$ can be used for inference on θ since they have approximate standard normal distributions. The corresponding approximate significance functions based on observed data y^{o} are $\Phi\{w^{o}(\theta)\}$, $\Phi\{s^{o}(\theta)\}$ and $\Phi\{r^{o}(\theta)\}$. In Section 2.4, the significance function is discussed in greater detail and expressed in terms of an equivalent "evidence function", which allows us to asses the plausibility of different values of θ in light of y^{o} and make informed decisions based on the available data.

A different way of defining the significance function is through the notion of sufficiency. Let s be a sufficient statistic, i.e., a function of y such that the conditional density of Y given $S = s(Y)$ does not depend on θ for all s. That is

$$f_{Y|S}(y \mid s; \theta) = g(y, s), \quad \theta \in \Theta. \tag{1.5}$$

Then, s is "sufficient" for the parameter θ, in the sense that no other statistic that can be calculated from the data provides any additional information about θ. The definition of a sufficient statistic in (1.5) does not uniquely determine S, because it is possible to augment S with aspects of the data that are not already included. To avoid this ambiguity, we typically consider the minimal sufficient statistic, which is the lowest-dimensional statistic for which (1.5) holds.

Now, suppose that there exists a one-to-one transformation from a sufficient statistic $s(Y)$ to a pair (M, A), where the distribution of A is independent of θ and there is no further information in Y regarding θ beyond that in M. We call A an "ancillary" statistic; the term ancillary implies that A is auxiliary or supplementary in nature. Even though it does not contain any information about the parameter θ, the observed value a^{o} of A is still used to make inferences about the parameter. A more thorough and precise explanation of minimal sufficiency and the related Fisher–Neyman factorization theorem can be found in Barndorff-Nielsen (1978, section 4.2). Formulated otherwise using this decomposition, the significance function in (1.4) can be defined conditional on the observed value of $A = a^{o}$, leading to a significance function of the form

$$p^{o}(\theta) = \Pr\left(M \leq m^{o} \mid A = a^{o}; \theta\right).$$

In simpler terms, the above arguments suggest that it can be useful to focus on the distribution of the (minimal) sufficient statistic when making inferences about θ (Cox, 1958). This process of using the sufficient statistic to make inferences is called reduction by sufficiency (Barndorff-Nielsen and Cox, 1989, 1994), and is especially useful when the likelihood function is complex or the sample size is large, as it allows us to avoid the computational burden of computing the full likelihood.

Example 1

Let (Y_1, Y_2) be independent Poisson random variables with means $\mu_1 = (1-\theta)p$, and $\mu_2 = \theta p$, where p is a known constant. The log likelihood function is

$$\begin{aligned}\ell(\theta; y_1, y_2) &= \log\left\{\frac{\mu_1^{y_1}\exp(-\mu_1)}{y_1!}\frac{\mu_2^{y_2}\exp(-\mu_2)}{y_2!}\right\},\\ &= (y_1+y_2)\log p - p - \log\{(y_1+y_2)!\}\\ &\quad + \log\binom{y_1+y_2}{y_1} + y_1\log(1-\theta) + (y_2+y_1-y_1)\log\theta,\end{aligned}$$

where the minimal sufficient statistic is (Y_1, Y_2). By setting $S = Y_1$ and $A = Y_1 + Y_2$, we can write the likelihood as

$$\ell(\theta; y) = \ell_{\text{Pois}}(p; a) + \ell_{\text{B}}(\theta; s|a),$$

where ℓ_{Pois} is the log likelihood based on the random variable $A = Y_1 + Y_2$, which is distribution constant, having a Poisson distribution of mean p, and hence is ancillary, and the likelihood ℓ_{B} corresponds to the contribution of Y_1 given $A = a$, which is binomial with denominator a and parameter θ. One way to understand the role of conditioning in this example is to note that the total count, a, does not provide any information on its own to estimate θ. However, by conditioning on $A = a$, the observed information is

$$J_{\text{B}}(\theta) = \frac{s(2\theta-1) + a(1-\theta)^2}{\theta^2(1-\theta)^2}, \quad 0 < \theta < 1,$$

the same as in the original Poisson likelihood function, so we can retrieve complete information about θ. This is the case in contingency tables, where the total number of counts per row is fixed, and we use a multinomial variable to represent the corresponding row.

1.3 Exponential family models

Exponential family distributions have several desirable statistical and computational properties. These properties are often attributed to the *natural parameter space*

$$\mathscr{F} = \left\{\theta \in \mathbb{R}^d : \kappa(\theta) = \log \int e^{s(y)^{\mathrm{T}}\theta} f_0(y)\mathrm{d}y < \infty\right\},$$

constructed for a baseline density function $f_0(y)$ with support $\mathscr{Y}$, which may be either continuous or discrete. The natural observation $s(y) = \big(s_1(y), \ldots, s_d(y)\big)^{\mathrm{T}}$ consists of functions of y such that the set $\{1, s_1(y), \ldots, s_d(y)\}$ is linearly independent. In general, $\theta = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$ may depend on a parameter ϕ taking values in $\Omega \subset \mathbb{R}^d$, where $\theta(\Omega) \subseteq \mathscr{F}$. Under these assumptions, an exponential family of order d is

$$f(y;\phi) = f_0(y) \exp\left[s(y)^{\mathrm{T}}\theta(\phi) - \kappa\left\{\theta(\phi)\right\}\right], \quad y \in \mathscr{Y}, \phi \in \Omega, \tag{1.6}$$

and θ is called the *canonical parameter*. The density (1.6) is called a minimal representation and it can be demonstrated that $s(Y)$ forms a minimal sufficient statistic for the canonical parameter $\theta(\phi)$ (Fraser, 1963). If there exists a one-to-one mapping between $\mathscr{F}$ and Ω, the dependence on ϕ can be omitted, and the family is called a *natural exponential family*.

Using Hölder's inequality, it can be shown that $\mathscr{F}$ is convex and that $\kappa(\theta)$ is strictly convex on $\mathscr{F}$ (Davison, 2003, §5.2). Under (1.6), the cumulant generating function of $s(Y)$ is $K(t) = \kappa(\theta + t) - \kappa(\theta)$, so

$$\mathrm{E}_\theta\{s(Y)\} = \partial\kappa(\theta)/\partial\theta, \quad \mathrm{var}_\theta\{s(Y)\} = \partial^2\kappa(\theta)/\partial\theta\partial\theta^{\mathrm{T}}.$$

The convexity of $\kappa(\theta)$ allows an exponential family to be parameterized not only by the canonical parameter θ, but also by the mean parameter $\mu = \mathrm{E}_\theta\{s(Y)\}$. Many distributions that are typically parameterized using the mean parameterization can also be parameterized using the canonical parameterization. This property forms the basis for generalized linear models (Nelder and Wedderburn, 1972; McCulloch and Searle, 2001).

Example 2

The normal distribution with mean μ and variance σ^2 has density

$$\begin{aligned} f(y;\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y-\mu)^2/\left(2\sigma^2\right)\right\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}y^2 - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right], \end{aligned}$$

where $\phi = \left(\mu, \sigma^2\right)$. This is a two-parameter exponential family with $s_1(y) = y$, $s_2(y) = y^2$, $\theta_1(\phi) = \mu/\sigma^2$, $\theta_2(\phi) = -1/\left(2\sigma^2\right)$, $\kappa(\phi) = \mu^2/\left(2\sigma^2\right) + \log\sigma$, arising from tilting the standard normal density $f_0(y) = \exp\left(-y^2/2\right)/(2\pi)$.

Curved exponential families may arise when the set Ω is a subset of $\mathbb{R}^r$ and $r < d$. In this situation, two potential scenarios can occur:

i) The canonical parameter is a linear function of ϕ, giving an r-parameter exponential family for some $r < d$. The statistic $s(Y)$ is sufficient, but not minimal sufficient.

ii) There is no linear constraint linking θ to ϕ, giving a curved exponential family. The statistic $s(Y)$ is minimal sufficient, but may not be complete; for the notion of completeness see e.g., Moser (1996, Chapter 6).

Example 2 (ctd)

In the previous example, if the mean μ and variance σ^2 are equal, then we have a normal distribution with mean μ and variance μ, where $\mu > 0$. The canonical parameter $\theta(\phi) = \left\{1, -1/(2\mu)\right\}^{\mathrm{T}}$ satisfies the linear constraint $\theta_1 = 1$. The density is then a one-parameter exponential family with $s(y) = y^2$ as a complete sufficient statistic (Keener, 2010).

Inference in exponential families is a key setting in which the concepts of sufficiency, ancillarity, and model reduction, discussed in Section 1.2.3, can be applied. Consider an exponential family with a minimal representation of the form

$$f(y;\psi,\lambda) = f_0(y) \exp\left\{\psi s_1(y) + \lambda s_2(y) - \kappa(\psi,\lambda)\right\}, \quad y \in \mathcal{Y}.$$

Suppose that we want to make inferences about the parameter ψ. It can be shown that a partition of the natural observations $(S_1, S_2) = \{s_1(Y), s_2(Y)\}$ is sufficient for (ψ, λ) in the sense of Basu (1978). In particular, the conditional distribution of S_2 given $S_1 = s_1$ and the marginal distribution of S_2 are natural exponential families of

orders k and $d-k$, the sizes of the partition. Techniques for making inferences about the parameter of interest while eliminating the nuisance parameter are discussed in Barndorff-Nielsen and Pedersen (1968), Jørgensen and Labouriau (1995), and Pace and Salvan (1997).

1.4 Fundamental approximations

Two common density approximations used in statistical analysis are Laplace and saddlepoint approximations, typically used in Bayesian and frequentist inference. These approximations have been thoroughly studied and have well-established properties. In this section, we provide a brief overview of them in order to better understand Sections 1.5 and 1.6.

1.4.1 Laplace approximation

Suppose that $g(u)$ is a smooth convex function of u with a minimum at $u=\widehat{u}$. We denote partial derivatives of g evaluated at $\widehat{u}$ as $g_2 = d^2 g(\widehat{u})/du^2, g_3 = d^3 g(\widehat{u})/du^3$, and so forth. Using a Taylor series expansion of $g(u)$ around $\widehat{u}$, we have

$$\int_{-\infty}^{+\infty} \exp\{-ng(u)\}\,du = \left(\frac{2\pi}{ng_2}\right)^{1/2} \exp\{-ng(\widehat{u})\}\{1+O(n^{-1})\}. \tag{1.7}$$

The leading term on the right-hand side of (1.7), called the Laplace approximation, replaces the integral using the value and the second derivative of the exponent, i.e., $g(\widehat{u})$ and g_2, with a relative error of $O(n^{-1})$. The remainder term is

$$1+\frac{1}{n}\left(\frac{5}{24}\widehat{\kappa}_3^2-\frac{1}{8}\widehat{\kappa}_4\right)+O(n^{-2}),$$

where $\widehat{\kappa}_3 = g_3/g_2^{3/2}$ and $\widehat{\kappa}_4 = g_4/g_2^2$. The right-hand side of (1.7) is an asymptotic series meaning that the partial sums may not converge and the accuracy of the approximation may not be improve by adding more terms.

The Laplace approximation is often written with an additional factor $a(u)$. For instance, consider the expression

$$J_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{u_0} a(u)\exp\{-ng(u)\}\{1+O(n^{-1})\}\,du, \tag{1.8}$$

where u is a scalar, $a(u)>0$, and $g(u)$ has the same properties as in (1.7). Using two

changes of variables, we obtain

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{r_0^*}^{+\infty} \exp^{-nr^{*2}/2} \left\{1 + O\left(n^{-1}\right)\right\} dr^*, \\ &= 1 - \Phi\left(n^{1/2} r_0^*\right) + O\left(n^{-1}\right), \\ &= \Phi\left(-n^{1/2} r_0^*\right) + O\left(n^{-1}\right), \end{aligned}$$

where $r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{q_0}{r_0}\right)$, $\quad r_0 = \text{sign}(\widehat{u} - u_0) \left\{2g(u_0)\right\}^{1/2}$, and $q_0 = -\frac{g'(u_0)}{a(u_0)}$.

Further details can be found in Davison (2003, §11.3), with some slight variations, as r defined there has the opposite sign to that defined in equation (1.1).

1.4.2 Saddlepoint approximation

The saddlepoint approximation is widely used in asymptotic analysis and has close ties to the Laplace approximation. In this section, we will provide a concise summary of the former, which is used to approximate density and distribution functions and is the basis for numerous small-sample procedures.

Let $\bar{X}$ denote the average of a random sample of continuous scalar random variables $X_1, \ldots, X_n$, each having cumulant generating function $\kappa(u)$. By definition, the cumulant generating function of $n\bar{X}$ can be expressed as integral involving $f(\bar{x})$, the density of $\bar{X}$,

$$\begin{aligned} \exp\{n\kappa(u)\} &= \int_{-\infty}^{\infty} \exp\left\{un\bar{x} + \log f(\bar{x})\right\} d\bar{x} \\ &= \int_{-\infty}^{\infty} \exp\left\{-g(u,\bar{x})\right\} d\bar{x}, \end{aligned}$$

where $g(u,\bar{x}) = -un\bar{x} - \log f(\bar{x})$. Using the Laplace approximation (1.7) to the last integral yields

$$\exp\{n\kappa(u)\} = \left(\frac{2\pi}{g_2(u,\bar{x}_u)}\right)^{1/2} \exp(nu\bar{x}_u) f(\bar{x}_u) \left\{1 + O\left(n^{-1}\right)\right\},$$

where $\bar{x}_u$ minimises $g(u,\bar{x})$ over $\bar{x}$ for fixed u. It can be shown that

$$\bar{x}_u = \kappa'(u), \quad g_2(u,\bar{x}_u) = n\left\{\kappa''(u)\right\}^{-1}.$$

Hence, the saddlepoint approximation to the density of $\bar{X}$ at x is

$$f_{\bar{X}}(x) = \left\{ \frac{n}{2\pi\kappa''(u)} \right\}^{1/2} \exp\{n\kappa(u) - nux\} \left\{ 1 + O\left(n^{-1}\right) \right\}. \tag{1.9}$$

Some aspects of the derivation outlined above are not explained; for details see for example Butler (2007, Chapter 12) or Davison (2003, Section 13.2).

The saddlepoint approximation in (1.9) and the corresponding approximation to the cumulative distribution function of $\bar{X}$ involve using the cumulant generating function, $\kappa(u)$, and finding the value of $\tilde{u}$ for each x of interest. The approximation's accuracy depends on the distribution's shape and the proximity of the maximum of the cumulant generating function to the origin.

The saddlepoint approximation has been widely applied and has proven to be a powerful tool. Many techniques for accurately approximating densities and distributions using this method have been developed since its introduction in a seminal article by Daniels (1954); see Lugannani and Rice (1980), Skovgaard (1987), Srivastava and Yau (1989), and Pierce and Peters (1992). Reid provided a comprehensive review of its applications and the relevant literature in Reid (1988, 1991).

Example 3

Assume that a random variable X has a noncentral chi-squared density, which can be written as an infinite mixture of central chi-squared densities, where the weights are Poisson probabilities,

$$f(x;\theta) = \sum_{k=0}^{\infty} \frac{x^{p/2+k-1} e^{-x/2}}{\Gamma(p/2+k) 2^{p/2+k}} \frac{\theta^k e^{-\theta}}{k!}, \quad x > 0,\ \theta, p > 0;$$

here p is the degrees of freedom, θ is the noncentrality parameter, and $\Gamma(p)$ is the gamma function (Goutis and Casella, 1999). The cumulant generating function is

$$\kappa(u) = \frac{2\theta u}{1-2u} - \frac{p}{2}\log(1-2u), \quad u < 1/2.$$

The equation $\kappa'(u) - x = 0$ yields a second order polynomial in u, and the saddlepoint is

$$\tilde{u}(x) = \frac{-p + 2x - \sqrt{p^2 + 8\theta x}}{4x}, \quad x > 0.$$

The left panel of Figure 1.1 shows the saddlepoint and exact densities for $p = 7$ and $\theta = 4$. We also plot a normalized version of (1.9) in which the density is divided

by its integral, approximated numerically using Simpson's rule. The saddlepoint approximation appears to be highly accurate in this case.

If the analytic form of $\kappa(u)$ is not available or it cannot be calculated, then we can estimate $\kappa(u)$ using the estimator proposed by Davison and Hinkley (1988), i.e.,

$$\widehat{\kappa}_n(u) = \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp ux_i\right).$$

Derivative estimates of $\widehat{\kappa}_n(u)$ are then used to produce an empirical saddlepoint approximation $\widehat{f}_{\bar{X}}$. The empirical saddlepoint approximation has been shown to perform well in the central region of a distribution, as demonstrated in Feuerverger (1989) and Wang (1992). However, it has a limitation in that the solution to the equation $\widehat{\kappa}'(u) = x$ does not exist outside the convex hull of $x_1, \ldots, x_n$. This issue was addressed in Fasiolo et al. (2018) with the introduction of the extended empirical saddlepoint approximation.

Example 3 (ctd)

For noncentral chi-squared density in Example 3, we plot the empirical saddlepoint and exact densities using 10^3 samples for $p = 7$ and $\theta = 4$ in the right panel of Figure 1.1. The empirical density and its standardized version provide good approximations for the true density in the center of the range of the data, with slightly less accuracy in the right tail. This example suggests that the empirical approach may be viable for approximating densities when the cumulant generating function is unknown.

1.4.3 Bayesian asymptotics

In the Bayesian setting with parameter vector $\theta = (\psi, \lambda^{\mathrm{T}})^{\mathrm{T}}$, assume we have a prior density $\pi(\psi, \lambda)$. The posterior density of the scalar interest parameter ψ is

$$\pi(\psi \mid y) = \frac{\int f(y \mid \psi, \lambda)\pi(\psi, \lambda)d\lambda}{\int f(y \mid \psi, \lambda)\pi(\psi, \lambda)d\lambda d\psi}.$$

If both integrals are approximated using the multivariate Laplace approximation in

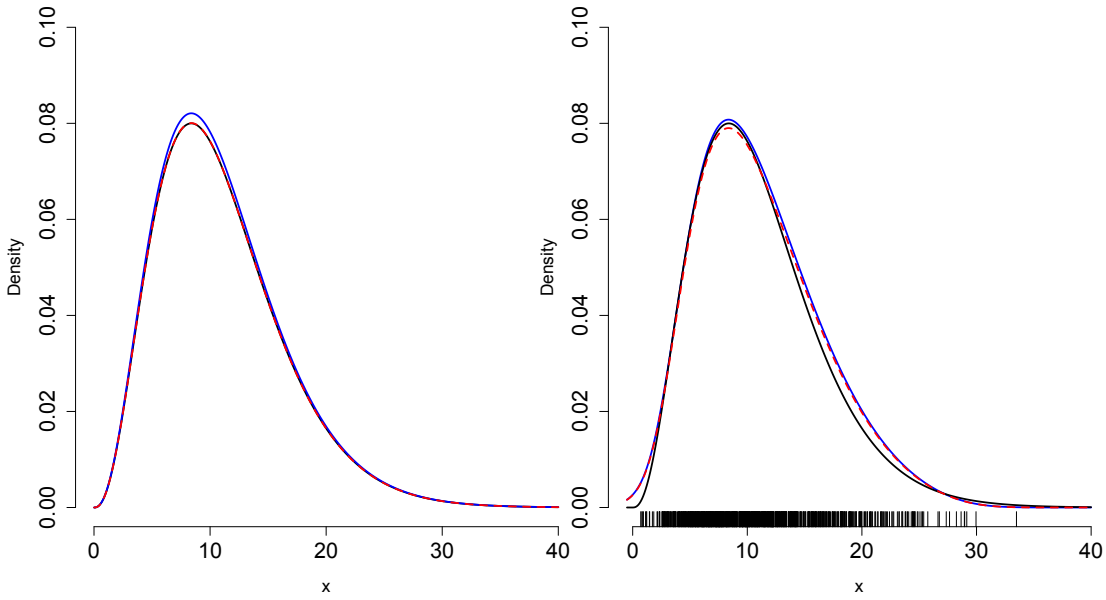


Figure 1.1 – Illustration of the saddlepoint density approximation for the non-central chi-square density with $p = 7$ and $\theta = 4$. Left panel: true density (black), saddlepoint density (blue), standardized saddlepoint (dashed red). Right panel: same densities using the empirical approximation based on 10^3 x_i; the rug of tickmarks shows the x_i themselves.

(1.7), the resulting approximation can be expressed as

$$\pi(\psi \mid y) = \left(\frac{n}{2\pi}\right)^{1/2} \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\tilde{\psi},\tilde{\lambda})}{\partial\theta\partial\theta^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\psi,\tilde{\lambda}_\psi)}{\partial\lambda\partial\lambda^{\mathrm{T}}} \right|} \right\}^{1/2} \frac{f\left(y \mid \psi, \tilde{\lambda}_\psi\right) \pi\left(\psi, \tilde{\lambda}_\psi\right)}{f(y \mid \tilde{\psi}, \tilde{\lambda}) \pi(\tilde{\psi}, \tilde{\lambda})} \left\{1 + O\left(n^{-1}\right)\right\}. \quad (1.10)$$

where $\tilde{\lambda}_\psi$ is the maximum a posteriori estimate of λ for a fixed value of ψ, and $\ell_m(\theta) = \log f(y \mid \theta) + \log \pi(\theta)$ is the log likelihood modified by the log prior. The posterior marginal cumulative distribution for ψ is obtained by integrating equation (1.10), which can be approximated using equation (1.8) by setting

$$g(\psi) = \log f(y \mid \tilde{\psi}, \tilde{\lambda}) - \log f\left(y \mid \psi, \tilde{\lambda}_\psi\right), \quad a(\psi) = \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\tilde{\psi},\tilde{\lambda})}{\partial\theta\partial\theta^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\psi,\tilde{\lambda}_\psi)}{\partial\lambda\partial\lambda^{\mathrm{T}}} \right|} \right\}^{1/2} \frac{\pi\left(\psi, \tilde{\lambda}_\psi\right)}{\pi(\tilde{\psi}, \tilde{\lambda})},$$

In this variant, the ratio of priors is included in the positive coefficient $a(u)$ of the integral instead of appearing in the exponent of (1.8). In this case, ℓ_m becomes simply the log likelihood, $\tilde{\theta}$ and $\tilde{\theta}_\psi$ are maximum likelihood estimates, the Hessians are

observed information matrices. The posterior marginal cumulative distribution for ψ is

$$\Pr(\psi \leq \psi_0 \mid y^{\mathrm{o}}) = \Phi\left\{-r_B^{*\mathrm{o}}(\psi_0)\right\}\left\{1 + O(n^{-1})\right\}, \tag{1.11}$$

where

$$\begin{aligned} r_B^{*\mathrm{o}}(\psi) &= r^{\mathrm{o}}(\psi) + \frac{1}{r^{\mathrm{o}}} \log\left\{\frac{q_B^{\mathrm{o}}(\psi)}{r^{\mathrm{o}}(\psi)}\right\}, \\ r^{\mathrm{o}}(\psi) &= \operatorname{sign}\left(\widehat{\psi} - \psi_0\right)\left[2\left\{\ell\left(\widehat{\theta}\right) - \ell\left(\widehat{\theta}_\psi\right)\right\}^{1/2}\right], \end{aligned}$$

is the likelihood root, and

$$\begin{aligned} q_B^{\mathrm{o}}(\psi) &= \frac{\mathrm{d}\ell(\psi, \widehat{\lambda}_\psi)}{\mathrm{d}\psi}\left\{\frac{\left|-\frac{\partial^2 \ell(\psi, \widehat{\lambda}_\psi)}{\partial\lambda\partial\lambda^{\mathrm{T}}}\right|}{\left|-\frac{\partial^2 \ell(\widehat{\psi}, \widehat{\lambda})}{\partial\theta\partial\theta^{\mathrm{T}}}\right|}\right\}^{1/2} \frac{\pi(\widehat{\theta})}{\pi(\widehat{\theta}_\psi)} \\ &= \frac{\mathrm{d}\ell(\widehat{\theta}_\psi)}{\mathrm{d}\psi}\left\{\frac{|j_{\lambda\lambda}(\widehat{\theta}_\psi)|}{|j(\widehat{\theta})|}\right\}^{1/2} \frac{\pi(\widehat{\theta})}{\pi(\widehat{\theta}_\psi)} \\ &= \ell'_{\mathrm{p}}(\psi)\, j_{\mathrm{p}}(\widehat{\psi})^{-1/2}\left\{\frac{\left|j_{\lambda\lambda}\left(\widehat{\theta}_\psi\right)\right|}{\left|j_{\lambda\lambda}(\widehat{\theta})\right|}\right\}^{1/2} \frac{\pi(\widehat{\theta})}{\pi\left(\widehat{\theta}_\psi\right)}. \end{aligned}$$

Each of these quantities is evaluated at the observed data $y = y^{\mathrm{o}}$ and the corresponding estimates $\widehat{\theta} = \widehat{\theta}^{\mathrm{o}}$ and $\widehat{\theta}_\psi = \widehat{\theta}^{o}_{\psi_0}$. The proof can be found in Brazzale et al. (2007, Section 8.7), although there is a typo in this reference, where q and r incorrectly have opposite signs.

Using an appropriate choice of the prior, there is a close parallel between frequentist and the Bayesian approximations. The appropriate priors are called "matching priors" and have been suggested as noninformative priors in Bayesian inference. Originally proposed by Jeffreys (1946), a constant information prior of the form

$$\pi(\theta)\mathrm{d}\theta \propto |i(\theta)|^{1/2}\ \mathrm{d}\theta$$

is invariant under reparameterization and has some special properties (Jeffreys, 1946). In the presence of nuisance parameter λ, Peers (1965), Welch and Peers (1963), and Tibshirani (1989) proposed an extension of the form

$$\pi(\psi, \lambda)\mathrm{d}\psi\mathrm{d}\lambda \propto\ i_{\psi\psi}(\psi, \lambda)^{1/2} g(\lambda)\mathrm{d}\psi\mathrm{d}\lambda,$$

where λ and ψ are orthogonal and $g(\lambda)$ is an arbitrary positive function that satisfies mild regularity conditions.

Reid et al. (2002) provide an overview of the pioneering work of Welch and Peers (1963) on matching priors, and discuss subsequent developments. Fraser and Reid (2002) introduce the concept of strong matching priors and provide insights into techniques for examining them, with a focus on location models. In Chapter 2 of the thesis, we use first-order matching priors for a curved exponential model and show the coverage of confidence intervals based on the approximate Bayesian solution is not as good as the corresponding frequentist intervals.

1.5 The p^* approximation

One of the primary higher-order asymptotic results in likelihood-based inference is the p^* approximation for the density of the maximum likelihood estimator. This transformation is exact for transformation models when properly normalized, and it coincides with the saddlepoint approximation for exponential models. In the following section, we will review it using the mathematical tools outlined in Section 1.4.

1.5.1 Definitions

Barndorff-Nielsen (1980, 1983, 1988) proposed the following approximation to the conditional density of the maximum likelihood estimator,

$$p^*(\widehat{\theta};\theta,a) = \frac{c(\theta,a)}{(2\pi)^{d/2}} \left|j(\widehat{\theta})\right|^{1/2} \exp\left\{\ell(\theta) - \ell\left(\widehat{\theta}\right)\right\}, \tag{1.12}$$

where a is an ancillary statistic , $c = c(\theta,a)$ is a renormalizing constant, $j(\widehat{\theta})$ is the observed Fisher information and $2\{\ell(\widehat{\theta}) - \ell(\theta)\}$ is the log-likelihood ratio statistic.

In the p^* approximation, the data vector is transformed into $(\widehat{\theta},a)$, which is assumed to be a one-to-one mapping. Transforming y into $(\widehat{\theta},a)$ allows us to view the right-hand side of (1.12) as a d-dimensional density for $\widehat{\theta}$ even though the sample space is n-dimensional. This dimension reduction is achieved by conditioning on the ancillary statistic a. The p^* approximation is invariant under one-to-one transformation of the data y, and is parametrization-invariant. The last property also applies to the normalization constant $c(\theta,a)$ (Barndorff-Nielsen and Cox, 1994, Section 6.2). For many models, including transformation models, the pair $(\widehat{\theta},a)$ forms a minimal sufficient statistic, and the p^* approximation is equal to the exact conditional density $p(\widehat{\theta};\theta,a)$.

In the special case when θ is a scalar, integrating the p^* approximation gives the

following approximation to the conditional cumulative distribution function of r,

$$\Phi(r^*) = \Phi\left(r + \frac{1}{r}\log\frac{q}{r}\right). \tag{1.13}$$

The pivot in (1.13), known as the *modified likelihood root*, is

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)}\log\frac{q(\theta)}{r(\theta)}. \tag{1.14}$$

It is derived, under regularity conditions, using a change of variable from $\widehat{\theta}$ to r in the exponent of equation (1.8) and the saddlepoint method discussed in Section 1.4.2. The correction term in (1.14) is

$$q = \left\{\ell_{;\widehat{\theta}}(\theta) - \ell_{;\widehat{\theta}}\left(\widehat{\theta}\right)\right\}\left\{J\left(\widehat{\theta}\right)\right\}^{-1/2} \tag{1.15}$$

and in general it depends on the sample space and mixed derivatives

$$\ell_{;\widehat{\theta}}(\theta;\widehat{\theta},a) = \frac{\partial\ell(\theta;\widehat{\theta},a)}{\partial\widehat{\theta}}, \quad \ell_{\theta;\widehat{\theta}}(\theta;\widehat{\theta},a) = \frac{\partial^2\ell(\theta;\widehat{\theta},a)}{\partial\theta\partial\widehat{\theta}^{\mathrm{T}}}.$$

Expression (1.13), known as the *Barndorff-Nielsen*'s r^* approximation, was introduced by Barndorff-Nielsen (1986) and discussed further in Barndorff-Nielsen and Cox (1994), Fraser et al. (1999a), and Reid (2003). An alternative approximation to (1.13) with the same order of asymptotic error is the *Lugannani and Rice* approximation

$$\Phi^*(r) = \Phi(r) + \left(\frac{1}{r} - \frac{1}{q}\right)\varphi(r). \tag{1.16}$$

The term $\ell_{;\widehat{\theta}}$ appears in the correction term (1.15) due to the transformation from $\widehat{\theta}$ to the likelihood root, $r(\theta)$. This term reflects the effect that changes in the data have on the likelihood function. However, these changes only occur in certain directions, namely those that maintain the value of the ancillary statistic a constant. More information about this derivation and discussion can be found in Skovgaard (1990), Fraser and Reid (1993), Fraser and Reid (1995), and Brazzale et al. (2007). In a detailed analysis, Barndorff-Nielsen and Cox (1994, Chapter 6) derived the p^* approximation using statistics other than r to measure the departure of $\widehat{\theta}$ from θ, such as the likelihood ratio statistic and Bartlett adjustments of this. They also derived asymptotic expansions of p^* that resemble Edgeworth expansions, but with coefficients that are determined by mixed derivatives of the log model function rather than cumulants.

If θ is a vector, the argument becomes more complex. In this case, we need to trans-

form θ into r and $d-1$ additional variables, say $\widehat{\theta}_\psi$. The Jacobian matrix for the transformation from $\left(r, \widehat{\theta}_\psi\right)$ to $(\widehat{\psi}, \widehat{\lambda})$ is based on likelihood differentiation. After this transformation, we integrate out the additional variables, resulting in

$$q = \left|\ell_{;\widehat{\theta}}(\widehat{\theta}) - \ell_{;\widehat{\theta}}\left(\widehat{\theta}_\psi\right) \;\; \ell_{\lambda;\widehat{\theta}}\left(\widehat{\theta}_\psi\right)\right| \left\{\left|j_{\lambda\lambda}\left(\widehat{\theta}_\psi\right)\right| \left|j(\widehat{\theta})\right|\right\}^{-1/2}. \tag{1.17}$$

The first factor on the right-hand side of (1.17) is the determinant of a $d \times d$ matrix whose first column is given by the difference of sample-space derivatives $\ell_{;\widehat{\theta}}(\widehat{\theta}) - \ell_{;\widehat{\theta}}\left(\widehat{\theta}_\psi\right)$ and whose other columns are given by the $d \times (d-1)$ matrix of mixed derivatives $\ell_{\lambda;\widehat{\theta}}\left(\widehat{\theta}_\psi\right)$. Barndorff-Nielsen and Cox (1994, Chapter 8) offer a comprehensive proof.

1.5.2 Decomposition of the correction term

Pierce and Peters (1992) examined the modified directed likelihood when the parameter of interest is a one-dimensional component of the canonical parameter in an exponential model. They suggested breaking down the adjustment term into two parts: one that addresses the potential influence of nuisance parameters and another that addresses the deviation from standard normality of the modified likelihood root. Barndorff-Nielsen and Cox (1994, Chapter 6) extend their decomposition of r^* to the general setting, obtaining

$$\begin{aligned} r^* &= r + r_{\text{NP}} + r_{\text{INF}}, \\ r_{\text{NP}} &= r^{-1} \log C, \qquad (1.18) \\ r_{\text{INF}} &= r^{-1} \log(\tilde{u}/r). \qquad (1.19) \end{aligned}$$

where

$$C = \left|\ell_{\lambda;\widehat{\lambda}}\left(\widehat{\theta}_\psi\right)\right| \left\{\left|j_{\lambda\lambda}\left(\widehat{\theta}_\psi\right)\right| \left|j_{\lambda\lambda}(\widehat{\theta})\right|\right\}^{-1/2},$$

and

$$\tilde{u} = \left|\ell_{;\widehat{\theta}}(\widehat{\theta}) - \ell_{;\widehat{\theta}}\left(\widehat{\theta}_\psi\right) \ell_{\lambda;\widehat{\theta}}\left(\widehat{\theta}_\psi\right)\right| \left\{j_{\text{p}}(\widehat{\psi})^{-1/2} \left|\ell_{\lambda;\widehat{\lambda}}\left(\widehat{\theta}_\psi\right)\right|\right\}^{-1},$$

$j_{\text{p}}(\widehat{\psi})$ is the observed profile information on evaluated at $\widehat{\psi}$, and C is a score-type statistic based on the profile log likelihood function.

Tang and Reid (2020) studied the order of the terms in equations (1.18) and (1.19) in settings with increasing numbers of nuisance parameters, focusing on linear exponential families and location-scale families.

1.5.3 Further remarks

The approximation of (1.13) by (1.16) suggests that r is approximately normally distributed. The accuracy of this approximation can be quantified using a Taylor series expansion, of the form

$$r = q + \frac{A}{\sqrt{n}}q^2 + \frac{B}{n}q^3 + O(n^{-3/2}), \tag{1.20}$$

where A and B are constants, and the expansion follows from differentiation of the normed profile likelihood $\ell\left(\widehat{\theta}_\psi\right) - \ell(\widehat{\theta})$; see in Reid (2003) and Brazzale et al. (2007, Chapter 8).

This expansion is a way to address the singularity at $\widehat{\theta}$ as both r and q approach zero when θ approaches $\widehat{\theta}$. Li (2001) and Fraser et al. (2003) developed third- and second-order bridges for the p-value around the maximum likelihood singularity for scalar and vector parameters, respectively. The resulting pivots can be viewed as Bartlett-type corrections to the likelihood ratio that are derived from the observed likelihood. In practice, the pivots are usually evaluated on a fine grid of points around $\widehat{\theta}$ and interpolated using smoothing splines (Fraser et al., 2003).

It follows from (1.20) that

$$\frac{1}{r}\log\frac{q}{r} = \frac{A}{\sqrt{n}} + \frac{B - 3A^2/2}{n}q + O(n^{-3/2}), \tag{1.21}$$

$$\frac{1}{r} - \frac{1}{q} = \frac{A}{\sqrt{n}} - \frac{B - A^2}{n}q + O(n^{-3/2}), \tag{1.22}$$

$$r^* = \frac{r - A/n^{1/2}}{\{1 + (2B - 3A^2/n)\}^{1/2}} + O\left(n^{-3/2}\right), \tag{1.23}$$

and

$$\mathrm{E}(r) = \frac{A}{\sqrt{n}} + O(n^{-3/2}), \quad \mathrm{var}(r) = 1 + \frac{1}{n}(2B - 3A^2) + O(n^{-2}).$$

The expansions in equations (1.21) and (1.22) demonstrate that both approximations in (1.16) and (1.13) are equivalent to $O(n^{-1})$. Renormalizing the density results in a normal density with an error of $O(n^{-3/2})$, as the $1/n$ term cancels out with the normalizing constant (Davison and Reid, 2022). Hence, an approximation that treats the modified likelihood root (1.23) as standard normal has a relative error of order $O(n^{-3/2})$.

The difficulty of constructing an exact or approximate ancillary statistic a for the p^* formula has limited its practical usefulness, as it is often not straightforward to

construct such a statistic, and it is uncommon to have an explicit formula for it in general models (Reid, 2003). Fraser (1990, 1991) has proposed alternative versions of the adjusted likelihood root that do not involve the transformation from y to $(\widehat{\theta}, a)$. Several other approximations have also been suggested in the literature, including those by Barndorff-Nielsen and Chamberlin (1991), DiCiccio and Martin (1993), with the approximation proposed in Skovgaard (1996, 2001) being particularly useful for both theoretical and applied purposes.

1.6 The tangent exponential model

The tangent exponential model proposed by Fraser and co-authors e.g., (Fraser and Reid, 1993, 1995; Fraser et al., 1999a) simplifies the construction of (1.14), by noting that it is not necessary to consider the full dependence of the log likelihood function on $(\widehat{\theta}, a)$. It is sufficient to examine the first derivative of the log likelihood at the observed data to understand how the log likelihood changes as $\widehat{\theta}$ changes but a is fixed. This approach is similar to the scalar parameter setting, where the full dependence of the log likelihood on the transformation is not needed. For this model, the approximation to (1.12) is

$$f_{\mathrm{TEM}}(s \mid a; \theta) = \exp\left[s^{\mathrm{T}}\varphi(\theta) + \ell\left\{\theta(\varphi); y^{\circ}\right\}\right] h(s). \tag{1.24}$$

This can be seen as a linear exponential family model with a constructed sufficient statistic $s = s(y) \in \mathbb{R}^d$ and constructed canonical parameter $\varphi(\theta) \in \mathbb{R}^d$, where $-\ell\left(\theta; y^{\circ}\right) = -\ell\left(\varphi(\theta); y^{\circ}\right)$ is the cumulant generating function. If the underlying density of y belongs to the exponential family, then $\varphi(\theta)$ is simply the canonical parameter. In more general models, the canonical parameter may depend on the data y^{o} (Davison and Reid, 2022).

1.6.1 Sufficient directions

We are interested in performing inference based on the conditional density $f(s \mid a^{\circ}; \theta)$, where a° is the observed value of the ancillary statistic A. To do so, we define the reference set $\mathscr{A}^{\circ} = \left\{y \in \mathbb{R}^n : a(y) = a^{\circ}\right\}$, as the d-dimensional manifold of the sample space on which the ancillary statistic equals its observed value a°. The reference set $\mathscr{A}^{\circ}$ can be parameterized in terms of s, at which point its tangent plane, denoted by $\mathscr{T}_s$, is determined by the columns of the $n \times d$ matrix $\partial y(s, a^{\mathrm{o}})/\partial s^{\mathrm{T}}$. This allows us to perform inference based on the conditional density $f(s \mid a^{\circ}; \theta)$ while taking into account the observed value of the ancillary statistic a°. In particular, the tangent plane $\mathscr{T}_{\circ}$ to $\mathscr{A}^{\circ}$ at y° is determined by the *sufficient directions*, i.e, the space spanned by the

columns of the matrix

$$V = \left.\frac{\partial y(s, a^\circ)}{\partial s^{\mathrm{T}}}\right|_{y=y^\circ}.$$

Constructing V does not require the knowledge of the mapping $y \mapsto (s, a)$, as

$$V = \left.\frac{\partial y}{\partial s^{\mathrm{T}}}\right|_{y=y^\circ} = \left.\frac{\partial y}{\partial \theta^{\mathrm{T}}}\right|_{y=y^{\mathrm{o}},\theta=\widehat{\theta}^{\mathrm{o}}} \times \left.\left(\frac{\partial s}{\partial \theta^{\mathrm{T}}}\right)^{-1}\right|_{y=y^{\mathrm{o}},\theta=\widehat{\theta}^{\mathrm{o}}}, \tag{1.25}$$

and we typically take V to be $\partial y/\partial\theta^{\mathrm{T}}$, evaluated at $y = y^\circ$ and $\theta = \widehat{\theta}^\circ$ since the second matrix on the right has dimension $d \times d$ and is invertible; thus the column space of V is also the column space of the first matrix on the right-hand side of (1.25) . Both span the sufficient directions, but the columns of the matrix $\partial y/\partial\theta^{\mathrm{T}}$ do not require y to be expressed in terms of (s, a).

The term *ancillary directions* has been used to refer to the columns of V, which are obtained from the ancillary manifold $\mathscr{A}^\circ$ at $s = s^\circ$. However, Davison and Reid (2022) argued that this term is misleading because the d columns of V show how y changes in the direction of s locally at s°. Therefore, referring to them as sufficient directions is more accurate. The ancillary statistic itself varies locally at a° in the $n-d$ directions that are perpendicular to the columns of V.

In Fraser and Reid (1995), it was demonstrated that the vector V can be constructed using a vector of pivot statistics $z(y;\theta) = \{z_1(y_1,\theta),\ldots,z_n(y_n,\theta)\}^{\mathrm{T}}$. Each element $z_i(y_i,\theta)$ of this vector has a fixed distribution under the specified model. This construction of V relies on the assumption that the components of y are independent, and gives

$$V = \left.\frac{\partial y}{\partial \theta^{\mathrm{T}}}\right|_{y=y^{\mathrm{o}},\theta=\widehat{\theta}^{\mathrm{o}}} = -\left.\left(\frac{\partial z}{\partial y^{\mathrm{T}}}\right)^{-1}\right|_{y=y^{\mathrm{o}},\theta=\widehat{\theta}^{\mathrm{o}}} \times \left.\frac{\partial z}{\partial \theta^{\mathrm{T}}}\right|_{y=y^{\mathrm{o}},\theta=\widehat{\theta}^{\mathrm{o}}}, \tag{1.26}$$

these are tangent to the surface in the sample space on which the ancillary statistic is held constant.

Example 4

Suppose that Y_1/θ and $Y_2\theta$ are independent gamma variables with unit scale and shape parameter n. The joint density function of Y_1 and Y_2 is

$$f(y_1, y_2;\theta) = \frac{(y_1 y_2)^{n-1}}{\Gamma(n)^2}\exp\left(-\frac{y_1}{\theta} - y_2\theta\right), \quad y_1, y_2 > 0, \theta > 0.$$

If we set $S = (Y_1/Y_2)^{1/2}$ and $A = (Y_1 Y_2)^{1/2}$, then $Y_1 = AS$ and $Y_2 = A/S$, and A is ancillary for θ. The log likelihood for this model can be expressed using the conditional density

of S given that $A = a$,

$$\ell(\theta; s \mid a) = -a\left(\frac{s}{\theta} + \frac{\theta}{s}\right), \quad \theta > 0.$$

The maximum likelihood estimator for θ is $\widehat{\theta} = s$, and the variance of this estimator, $\text{var}(\widehat{\theta}) = j(\widehat{\theta})^{-1} = s^2/a$, decreases as the ancillary statistic a increases, leading to more precise inference for θ. As $y^{\mathrm{T}} = (y_1, y_2) = (as, a/s)$, the sufficient directions are given by

$$V = \frac{\partial y(s,a)}{\partial s} = \left(a, -a/s^2\right)^{\mathrm{T}}\Big|_{y=y^\circ}.$$

This vector represents the change in the minimal sufficient statistic (y_1, y_2) as s changes, holding the ancillary statistic a constant. In Figure 1.2, we plot the joint density $f(y_1, y_2; \theta)$ for $a^\circ = 3$, $s^\circ = 1$, and $\theta = 1$. The dashed red curve represents the reference set $\mathscr{A}^\circ$, i.e., the values of the sufficient statistics y_1 and y_2 when the ancillary statistic is fixed at $a^\circ = 3$, and the dashed blue curve represents tangent space $\mathscr{T}^\circ$ at $y^\circ = (3,3)$. Similar illustrations for various scalar parameter models can be found in Reid (2003).

1.6.2 Canonical parameter

Now, let us consider the canonical parameter $\varphi(\theta)$ defined in (1.24) using the sample space derivatives discussed earlier,

$$\varphi\left(\theta; y^{\mathrm{o}}\right) = \ell_{;V}\left(\theta; y^{\mathrm{o}}\right).$$

Previously, we introduced the $n \times d$ matrix V of sufficient directions, which is used in the sample space derivative as follows

$$\ell_{;V}\left(\theta; y^{\mathrm{o}}\right) = \frac{d}{dt}\ell\left(\theta; y^{\mathrm{o}} + Vt\right)\Big|_{t=0} = V^{\mathrm{T}}\frac{\partial \ell(\theta; y)}{\partial y}\Big|_{y=y^\circ} = \sum_{j=1}^{n} V_j^{\mathrm{T}}\frac{\partial \ell\left(\theta; y_j^{\mathrm{o}}\right)}{\partial y_j},$$

where $t = (t_1, \ldots, t_d)^{\mathrm{T}}$, and $y_j, \ldots, y_n$ have independent contributions $\ell\left(\theta; y_j\right)$ to the log likelihood function; this will be used to build the pivot $q(\psi)$.

In general, q should be based on how far $\widehat{\theta}$ differs from $\widehat{\theta}_\psi$, or, alternatively, how far $\widehat{\varphi}$ differs from $\widehat{\varphi}_\psi$. If the density has linear exponential form, q can be expressed as a modified version of the Wald pivot defined in (1.2). But in general the Wald-type

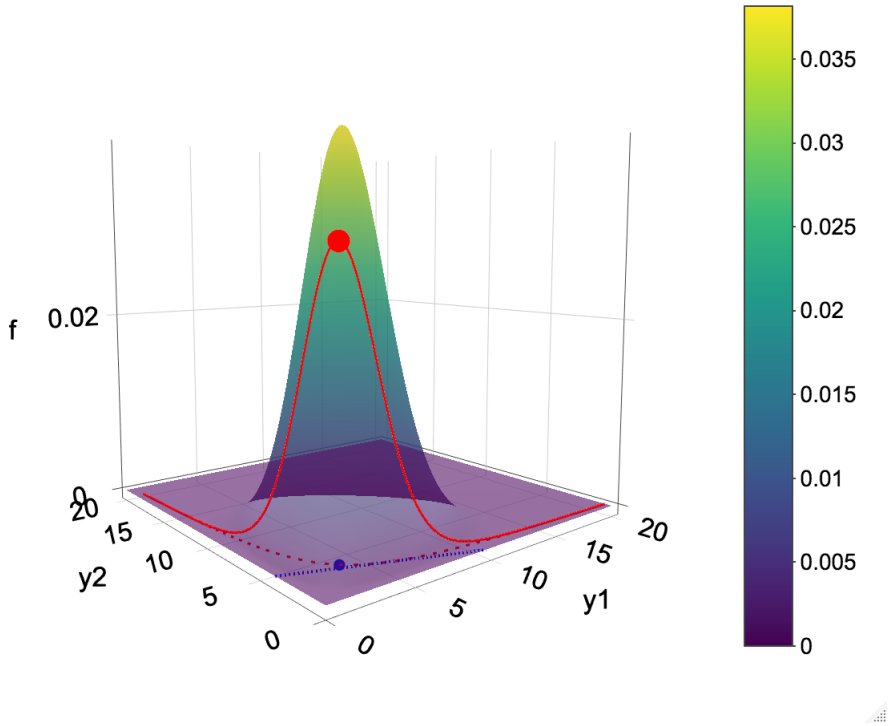


Figure 1.2 – Joint density for Example 4 with $\theta = 1$, $y^\circ = (3,3)$, giving $a^\circ = 3$ and $s^\circ = 1$. The dashed red line represents the reference set $\mathscr{A}^\circ$, and the solid red line represents the density $f(y;\theta)$ on $\mathscr{A}^\circ$. The tangent plane $\mathscr{T}^\circ$ is shown by the dashed blue line. The blue bullet represents the point y°, and the red bullet represents $f(y^\circ;\theta)$.

measure has the following form

$$q(\psi) = \text{sign}(\widehat{\psi} - \psi)\left|\chi(\widehat{\theta}) - \chi\left(\widehat{\theta}_\psi\right)\right|\left\{\frac{|j(\widehat{\varphi})|}{\left|j_{\lambda\lambda}\left(\widehat{\varphi}_\psi\right)\right|}\right\}^{1/2},$$

where the constructed parameter is given by

$$\chi(\theta) = u^{\mathrm{T}}\varphi(\theta), \quad u = \frac{\partial\psi\left(\widehat{\theta}_\psi\right)/\partial\varphi}{\left\|\partial\psi\left(\widehat{\theta}_\psi\right)/\partial\varphi\right\|},$$

i.e., the orthogonal projection of $\varphi(\theta)$ onto a unit vector u parallel to $\partial\psi\left(\widehat{\theta}_\psi\right)/\partial\varphi$. This form of q is an extension to the expression of (1.17) upon noting that partial derivatives of the log likelihood satisfy

$$\ell_\theta(\widehat{\theta};\widehat{\theta},a) = 0, \quad \ell_{\theta;\widehat{\theta}}(\widehat{\theta};\widehat{\theta},a) = j(\widehat{\theta}),$$

and computing the determinant using blocks corresponding to a partition of the

partial derivatives into (ψ, λ) components. Further details can be found in Brazzale et al. (2007, Section 8.5), Barndorff-Nielsen and Cox (1994, Chapter 6), and a guide to the literature may be found in Davison and Reid (2022). The resulting correction term can be written as

$$q(\psi) = \frac{\left|\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \;\; \varphi_\lambda(\widehat{\theta}_\psi)\right|}{\left|\varphi_\theta(\widehat{\theta})\right|} \frac{\left|j(\widehat{\theta})\right|^{1/2}}{\left|j_{\lambda\lambda}(\widehat{\theta}_\psi)\right|^{1/2}}. \tag{1.27}$$

Here $\varphi_\theta(\theta) = \partial\varphi(\theta)/\partial\theta^{\mathrm{T}}$ and $\varphi_\lambda(\theta) = \partial\varphi(\theta)/\partial\lambda^{\mathrm{T}}$ are respectively $d \times d$ and $d \times (d-1)$ matrices. The numerator of the first term of q is the determinant of a $d \times d$ matrix whose first column is $\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi)$ and whose remaining columns are $\varphi_\lambda(\widehat{\theta}_\psi)$. In (1.27), the observed information associated with the canonical parameter φ is expressed as function of $\theta = (\psi, \lambda)$, and the derivative of the parameter of interest $\partial\psi/\partial\varphi$ is a column of the inverse of φ_θ.

Example 5

To illustrate this discussion, assume that we have a single observation from the Rayleigh distribution, which arises as the length y of a bivariate normal vector whose components are independent $N(0, \psi^2)$ variables. Exact computations are possible in this case, so the quality of the approximations can be assessed. The probability density function for y is

$$f(y;\psi) = \frac{y}{\psi^2} e^{-y^2/(2\psi^2)}, \quad y > 0, \quad \psi > 0,$$

and one can readily check that $\widehat{\psi} = y/\sqrt{2}$, $j(\widehat{\psi}) = 4/\widehat{\psi}^2$ and

$$\begin{aligned} w(\psi) &= 2\left(1 - \frac{\psi}{\widehat{\psi}}\right), \\ r(\psi) &= \operatorname{sign}(\widehat{\psi} - \psi)\left[2\left\{2\log(\frac{\psi}{\widehat{\psi}}) + \left(\frac{\widehat{\psi}}{\psi}\right)^2 - 1\right\}\right]^{1/2}, \\ q(\psi) &= 1 - \left(\frac{\widehat{\psi}}{\psi}\right)^2. \end{aligned}$$

The left-hand panel of Fig 1.3 shows $\Phi\{w^{\mathrm{o}}(\psi)\}$, $\Phi\{r^{\mathrm{o}}(\psi)\}$, and $\Phi\{r^{*\mathrm{o}}(\psi)\}$ when y equals the observed value $y^{\mathrm{o}} = \sqrt{2}$, so $\widehat{\psi}$ has observed value $\widehat{\psi}^{\mathrm{o}} = 1$. The functions are decreasing in ψ and the maximum likelihood estimate and the limits of the 90% confidence interval are the values of ψ for which the functions equal 0.5 and 0.05, 0.95, respectively. The right-hand panel of Fig 1.3, which shows how $w^{\mathrm{o}}(\psi)$, $r^{\mathrm{o}}(\psi)$ and

$r^{*\mathrm{o}}(\psi)$ themselves depend on ψ, makes it easier to read off confidence intervals.

In this example $\widehat{\psi}/\psi$ has a known distribution, so the left-hand side of (1.4) provides an exact significance function. The 95% confidence interval for ψ based on r is $(0.4765, 4.186)$, but using r^* yields a confidence interval of $(0.519, 6.133)$, which is very close to the interval obtained using the exact significance function $(0.520, 6.103)$. The p-values for testing a null hypothesis $\psi = \psi_0 = 0.3$ versus an alternative $\psi > \psi_0$ using the Wald statistic and the likelihood root are $\Phi\{-w^{\mathrm{o}}(\psi_0)\} = 0.0808$, and $\Phi\{-r^{\mathrm{o}}(\psi_0)\} = 4.33 \times 10^{-5}$. So these pivots give quite different evidence about ψ_0. The exact significance probability for testing $\psi_0 = 0.3$ is 1.49×10^{-5}. The significance probability based on r^* is 1.55×10^{-5}, giving a relative error of 3.9%. The sample size here is $n = 1$, so it is no surprise that the exact function differs greatly from the large-sample approximations based on $w^{\mathrm{o}}(\psi)$ or $r^{\mathrm{o}}(\psi)$. Nevertheless, the approximation based on r^* yields near-perfect inferences: it is indistinguishable from the exact quantities.

The Rayleigh example is chosen because it allows for exact inferences to be made, demonstrating the advantage of using higher-order approximations over classical ones, which can sometimes produce poor results. In this specific example, it is straightforward to determine q, but generally, it can be more challenging, as we will see later in the thesis.

1.6.3 Discrete settings

The local canonical parameter for discrete random variables is constructed differently than for continuous distributions (Frydenberg and Jensen, 1989; Davison et al., 2006). If the random variable $(Y_1, \ldots, Y_n)$ has a discrete distribution with Y_i following a curved exponential family model of the form

$$f_i\left(y_i;\theta\right) = f_0\left(y_i\right)\exp\left\{\alpha_i(\theta)\, y_i - \kappa_i(\theta)\right\},$$

and mean $\mu_i(\theta) = \mathrm{E}\left(y_i;\theta\right)$, then the vectors V_i can be derived by considering the effect of θ on y_i through its mean $\mu_i(\theta)$ i.e.,

$$V_i = \left.\frac{\partial \mathrm{E}\left(Y_i;\theta\right)}{\partial\theta}\right|_{\widehat{\theta}^{\mathrm{o}}} = \left.\frac{\partial \mu_i(\theta)}{\partial\theta}\right|_{\widehat{\theta}^{\mathrm{o}}} \tag{1.28}$$

and the canonical parameter from the ith observation is

$$\varphi_i(\theta) = \alpha_i(\theta)\, V_i.$$

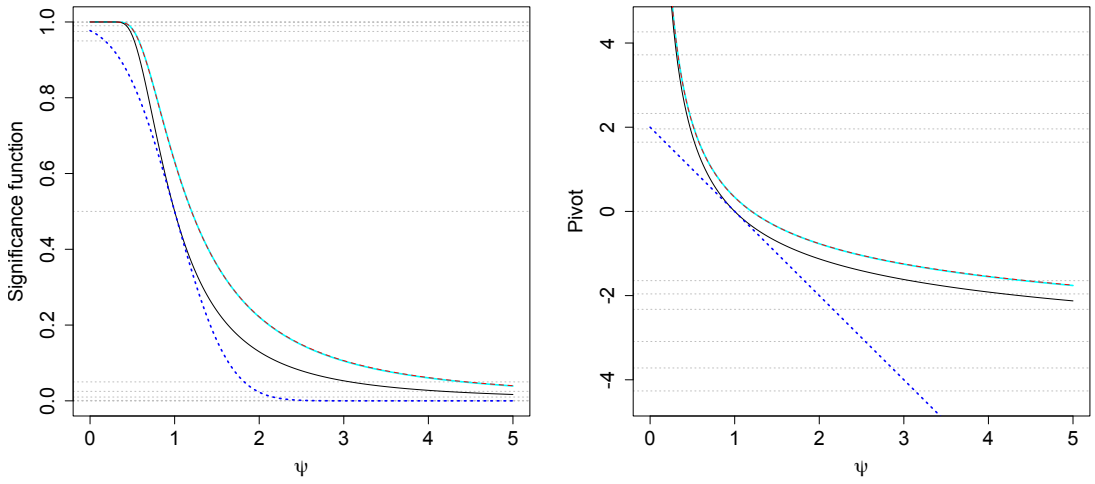


Figure 1.3 – Left: Significance functions based on likelihood root $r^{\circ}(\psi)$ (solid black), Wald statistic $w(\psi)$ (dotted blue), exact (wide cyan) and modified likelihood root $r^{*\circ}(\psi)$ (red dashes). The horizontal lines correspond to probabilities $10^{-6}, 10^{-5}$, 10^{-4}, 10^{-3}, 10^{-2}, 0.025, 0.05, 0.5, 0.95, 0.975, 0.999, 0.9999, 0.99999 and 0.999999. Right: functions from (a) transformed to the standard normal scale.

In this context, y_i is considered to be the score variable for α_i in the full exponential family model (Fraser and Reid, 2001). We can extend this approach to more general settings by replacing y_i with a locally defined score variable in the construction of $\varphi(\theta)$ and with V_i defined through

$$s_i = \frac{\partial}{\partial\theta}\ell\left(y^i;\theta\right)\Big|_{\theta=\widehat{\theta}^{\circ}}$$

and

$$V_i = \frac{\partial}{\partial\theta}\mathrm{E}\left(s_i;\theta\right)\Big|_{\theta=\widehat{\theta}^{\circ}}. \tag{1.29}$$

The score variable s_i is a $d\times 1$ vector, so V_i is a $d\times d$ matrix that represents the contribution of y_i to the expected information matrix, evaluated at $\widehat{\theta}^{\circ}$. The contribution of the i-th observation to the local reparameterization is

$$\varphi_i(\theta) = \frac{\partial\ell_i\left(\theta;y_i\right)}{\partial s_i}\Big|_{y_i^{\circ}} V_i\,.$$

Partial derivatives of the log likelihood function can be obtained as

$$\frac{\partial\ell\left(\theta;y_i\right)}{\partial s_i} = \frac{\partial\ell\left(\theta;y_i\right)}{\partial y_i}\times\left(\frac{\partial s_i}{\partial y_i}\right)^{-1}.$$

Equations (1.28–1.29) will be used in Chapter 3 for a discrete example, but similar results can also be found in Brazzale et al. (2007, Chapter 4), Davison and Sartori (2008) and Davison et al. (2006).

Other approximations proposed for use with discrete data include those of Skovgaard (1996) and Severini (1999). Severini's approximation involves moment estimators of expected values, while Skovgaard's correction term is based on cumulants of the log likelihood and produces the same approximation as in curved exponential families (Davison et al., 2006; Reid and Fraser, 2010). These approximations have a relative error that is of order $O(n^{-1})$ instead of $O(n^{-3/2})$ in continuous setting.

1.7 Summary

The literature contains many ways to approach higher-order approximations, depending on the model in question and the methods used for inference. The likelihood function is often central to these methods, allowing for higher-order adjustments to classical first-order statistics that can minimize the impact of nuisance parameters. One higher-order approximation is the p^* formula, which provides third-order refinements for likelihood-based statistics. This approximation is based on ancillary statistics and evaluated at the observed quantities. The formula is generally accurate to order $O(n^{-3/2})$, and it is, in fact, exact for many important models. The modified likelihood root r^* is another important higher-order asymptotic quantity for likelihood-based inference. It has better inferential properties than the ordinary likelihood root, and has been extensively studied, revised, and applied to many models since its introduction by Barndorff-Nielsen in 1986.

An approach that unfolded later to simplify the p^* formula is the tangent exponential model, which improves asymptotic normal approximation with the same relative error of $O(n^{-3/2})$. This is based on similar concepts: conditioning, sufficiency, and ancillarity, but only requires computation of the observed likelihood and its first sample-space derivative. The tangent exponential model has contributed greatly to the literature on higher-order asymptotics by making it more accessible and applicable to biology, sociology, and other applied areas, due to its relatively simple and familiar quantities. A variety of model classes have been studied by Brazzale et al. (2007), and a more detailed version of similar computations can be found in Fraser et al. (1999b). A literature guide can be found in Davison and Reid (2022).

In Chapter 2 of the thesis, we extend the list of applications of higher order approximation to include satellite conjunction assessment, where highly accurate estimation of confidence limits and tail probabilities is of crucial importance. In Chapter 3, we study

the performance of first-order approximations when the parameter of interest is on the boundary of the parameter space and discuss techniques such as bias correction for the profile score and Edgeworth expansion for its distribution. We also examine the improvement of third-order pivots for such irregular problems.

Statistical Formulation of Conjunction Assessment

2.1 Introduction

The expansion of the aerospace industry and the increasing number of space objects, especially in Low Earth Orbit (LEO), where most spacecraft operate, means that risk assessment and collision avoidance manoeuvres are vital to ensure their safety. According to the European Space Agency (ESA) (ESA, 2022), in more than 60 years of space activity, more than 6250 launches have resulted in 13630 satellites being placed into Earth Orbit. Of these satellites, 65% are still in orbit, but only 48% are functional. The active satellites face navigating through tons of accumulated space debris recorded in the US Space Surveillance Network (SSN) (Chatters et al., 2009, Chapter 19). The SSN only tracks objects larger than 5–10 cm in low Earth orbit (LEO) and 30 cm to 1 m in geostationary (GEO) orbit. However, it is estimated that there are more than 132 million fragments of debris ranging in size from 1 mm to 10 cm that are not included in the SSN's catalogue (ESA, 2022). These smaller pieces can still pose a threat to active satellites.

The increasing rate at which new objects are added to space justifies the concerns that specialists raise about the overall safety of existing spacecraft and the long-term sustainability of space activities (Union of Concerned Scientistis, 2022). According to the Satellite Industry Association (SIA)(SIA, 2022), the number of commercial satellites launched during 2021 and the first semester of 2022 increased by more than 40% as compared to 2020. The current race between private space companies and governments will add to this.

Although these numbers are alarming, there is an emerging effort to raise awareness of space safety and global debris mitigation in parallel to a growing industry working on sustainable satellite activities (Virgili, 2016; Letizia et al., 2019; International

Standards Organisation, 2016; Lewis, 2020). Among many other initiatives (Braun et al., 2013; Inter-Agency Space Debris Coordination Committee, 2007; International Standards Organisation, 2016, 2019), the United Nations Committee on the Peaceful Uses of Outer Space (UNCOPUOS) released guidelines to provide an overview of space activities and to quantify internationally endorsed mitigation measures (Committee on the Peaceful Uses of Outer Space, 2019). While the development of such voluntary guidelines is promising, the lack of both communication and collaboration on an international level and the absence of legislation make space sustainability out of reach.

Before addressing conjunction assessment, one must formally define the notion of conjunction. According to ESA (2022), a conjunction is a close geometric approach between two objects, irrespective of their activity status, triggering an operator analysis but not necessarily an avoidance manoeuvre or implying a collision. The motion between these two bodies is described based on relative motion theory in contrast to absolute motion, two controversial interpretations of motion philosophically debated since antiquity (Armstrong, 1963; Vallado, 2013). The use of relative motion is fundamental in space missions such as spacecraft formation flying (Inalhan et al., 2002), space rendezvous and proximity (Curtis, 2010, Chapter 7) (Burnett and Schaub, 2022), relative orbital navigation problems (Alfriend et al., 2010, Chapter 12), and space object surveillance (Terui and ichiro Nishida, 2007; Kenneth, 2015).

Relative motion models are divided into two categories: algebraic and geometrical. Both have been successfully applied to many space missions (Klinkrad, 2006). Algebraic models for relative errors were proposed by Hill (1878) and Clohessy and Wiltshire (1960) and extensively explored in Yamanaka and Ankersen (2002). These models are based on the dynamical equations describing each object's position and velocity vectors expressed in a relative frame. Geometrical models, on the other hand, illustrate the evolution of an equivalent 6-dimensional vector that contains the orbital elements; the scalar magnitude, and the angular representations of the two orbits (Schaub and Alfriend, 2002; Schaub, 2004). Either model specifies the two-body orbit and provides a complete set of initial conditions for solving an initial value problem for a set of differential equations. Depending on the model's complexity and the orbit type, the solution of this system, when evaluated at the time of the closest approach, provides a prediction for the relative state vector and the associated covariance matrix.

Conjunction assessment for orbiting objects is generally done by representing the two objects as ellipsoids and attempting to estimate the probability that they will collide, as in Vallado (2013, Section 11.7) or Chen et al. (2017, Chapter 5). This is calculated at the time of the closest approach using the estimated position and velocity vectors

for the two objects and the associated error covariances. In short-term conjunction, it can be expressed as an integral of a two-dimensional Gaussian probability density function over the collision cross-sectional area. Although unavailable in explicit form, this integral can readily be evaluated semi-analytically using different approaches with comparable accuracy (Foster and Estes, 1992; Chan, 1997; Alfriend et al., 1999; Alfano, 2005a, 2006b; Patera, 2005; Garcia-Pelayo and Hernando-Ayuso, 2016).

The simplified calculation of the collision probability offered in short-term encounters is not valid in long-term conjunction: the motion is not linear, the relative velocity is not constant, and its uncertainty is not negligible. In this case, one needs to compute the integral of the probability density through the volume swept out by the combined hard body sphere, which is generally expressed as an integral over time of a time-dependent collision probability. Although this integration is complicated due to the changing direction of the hard body sphere and the combined position-error ellipsoid throughout the encounter, most of the methods proposed for the two-dimensional integral have been revised and extended to cover nonlinear motion (Patera, 2003, 2006; Alfano, 2006a; Kenneth, 2015). Breaking the collision tube into sufficiently small cylinders makes the motion nearly linear in each section. The total integral is then the sum over individual two-dimensional integrals in each section. Hall (2021) summarises the literature on probability calculation for nonlinear and repeated conjunctions.

The probability of collision, if computed as described above, has a perplexing behavior in which both more precise and less precise measurements typically reduce the collision probability, a 'dilution' property that has been seen as paradoxical, and its interpretation is not seen as clear-cut (Balch, 2016; Balch et al., 2019). Hejduk et al. (2019) gives an insight into this phenomenon by graphically illustrating what happens in both the dilution and robust regions and explaining how risk assessment analysts proceed when presented with suspiciously low collision probability values. Similar works have also been pursued by other authors to discuss how to use the probability of collision operationally (Alfano, 2005a; Alfano and Oltrogge, 2018; Hejduk and Snow, 2019; Siminski et al., 2021). However, the mere fact that such a dilution effect exists makes the use of collision probability as a conjunction assessment metric less straightforward.

Another criterion for risk assessment is the closest approach, or miss distance, as a miss distance that is likely to be lower than a specified safety threshold indicates a situation that requires close inspection. To have an idea about both criteria, we use conjunction data for 1100 pairs of space objects tracked over one week in October 2022. The data is published by Space-Track.org as part of Conjunction Data Messages

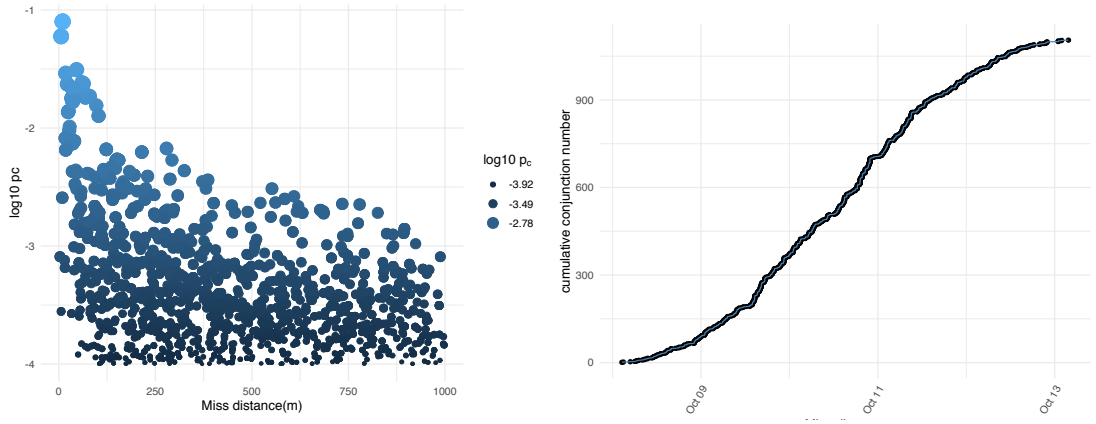


Figure 2.1 – Conjunction probabilities and miss distances from CDMs published by Space-Track.org during the second week of October 2022.

(CDMs). In the left panel of Figure 2.1, we plot conjunction probabilities versus miss distances. In general, as the miss distance increases, the conjunction probability decreases. However, this depends on the uncertainty of the observed relative state vector. In the plot shown, the miss distances are less than 1 km and the conjunction probabilities are greater than 10^{-4}. The likelihood of a collision is relatively high for such conjunctions . The right panel of Figure 2.1 shows the accumulated number of expected conjunctions over this period and highlights the frequency of such events. The miss distance is usually estimated repeatedly over the approach to conjunction, and the collision probability is updated accordingly. There are many algorithms to compute the miss distance, often as the root of a polynomial equation (Alfano and Negron, 1993; Gronchi, 2005; Armellin et al., 2010) that depends on the relative path of the objects, which may be highly nonlinear when they are far apart.

Most early studies on conjunction assessment, as well as current ones, focus on improving the use of the probability of collision. Over time, an extensive literature has developed covering a wide range of conjunction configurations, but little is established on the use of the miss distance. Chan (2011) derived the distribution of the squared miss distance, which has a generalized non-central chi-square distribution under a standard probability model. A recent addition to the literature uses characteristic function inversion to evaluate the distribution of the squared miss distance and to estimate collision probabilities (Bernstein et al., 2021). There has been relatively little research on the connection between the miss distance and the collision probability, despite the fact that these two metrics are the main tools used in practice to assess the risk of a collision. While Modenini et al. (2022) showed that the Mahalanobis miss distance (McLachlan, 1999), is well connected to the estimating the collision probability, more work is needed to fully understand the connection and to develop

more accurate and reliable methods for assessing the risk of a collision.

In this work, we formulate a statistical model for conjunction assessment that resolves the apparent difficulties with the collision probability and suggest that the miss distance is a more appropriate focus of interest. We discuss inference for this distance based on standard likelihood theory (Cox and Hinkley, 1974, chapter 9), and show that improved theory, presented in Chapter 1, is both highly accurate and should give results similar to a Bayesian formulation (Brazzale et al., 2007). Our approach is based on significance functions (Cunen et al., 2020) and provides both point and interval estimates for the miss distance, with the intervals containing the true miss distance with a specified probability under the model. We can also test whether the true distance is higher than the safety threshold in order to make decisions about avoidance manoeuvres (Neyman, 1937).

This chapter is organized as follows. In Section 2.3 we formulate conjunction assessment in statistical terms and discuss the choice of the parametrization. We also discuss the relationship between the collision probability and miss distance. After the brief review of methods for computing the probability of collision in Section 2.2, we point out that despite the success of this metric, it still suffers from some issues. We explain its downward bias and elucidate the 'dilution paradox' in Section 2.5. In Section 2.4, we discuss decision-making in the context of conjunction assessment, and we show in Section 2.6 how significance functions provide calibrated frequentist inference and link likelihood inference to the Bayesian approach. Then, we apply these ideas to satellite conjunction, and in Section 2.7, study their numerical properties based on four case studies and comment on limitations. In Appendix 2.9, we present possible extensions of the approach and provide details of our numerical implementation.

2.2 Probability of collision

The precise position of the two space objects is usually unavailable due to inevitable errors in orbit determination. Uncertainty in observing the relative position vector is then described using a three-dimensional probability density function with a covariance matrix Ω^{-1}. The covariance of the position of the objects is often dominated by systematic errors, such as errors in the orbit determination process or errors in the models used to predict the motion of the objects. It is common for practitioners to assume that Ω^{-1} is known, and below we follow this assumption. A collision probability is defined as the three-dimensional integral of this density over a collision region that we call $\mathscr{C}$. The core of most methods for computing collision probabilities is to perform this integration.

In practice, random errors in tracking data are typically assumed to follow a Gaussian distribution (Alfano, 2005a), but in reality this assumption may not hold. Errors in the model used to track the data can be difficult to quantify and may not necessarily follow a Gaussian distribution (Alfriend et al., 1999). Under the normality assumption, the collision probability is conveniently represented as

$$p_c = \frac{1}{\sqrt{(2\pi)^3 \det(\Omega^{-1})}} \int_{\mathscr{C}} \exp\left\{-\tfrac{1}{2}\left(x-\mu\right)^{\mathrm{T}} \Omega \left(x-\mu\right)\right\} \mathrm{d}x, \tag{2.1}$$

where μ denotes coordinates of the nominal position vector, x is the relative position vector, and Ω^{-1} is the associated 3×3 covariance matrix.

The direction of the axes depends on the choice of the coordinate system used to describe the relative motion. The RSW (radial, along-track, and cross-track) and NTW (tangential, normal, and cross-track) systems are two of the most common types of satellite-based coordinates; see Vallado (2013) for an exhaustive description of the existing coordinate systems. In short-term encounters, the relative motion satisfies the following assumptions:

(i) the trajectory of the second object relative to the primary object is linear;

(ii) the relative position vector has a Gaussian distribution; and

(iii) the relative velocity is sufficiently large that its uncertainty can be ignored.

The last assumption ensures a brief encounter time and a constant covariance, and the probability of collision can be reduced to two-dimensional integral, as we shall see below.

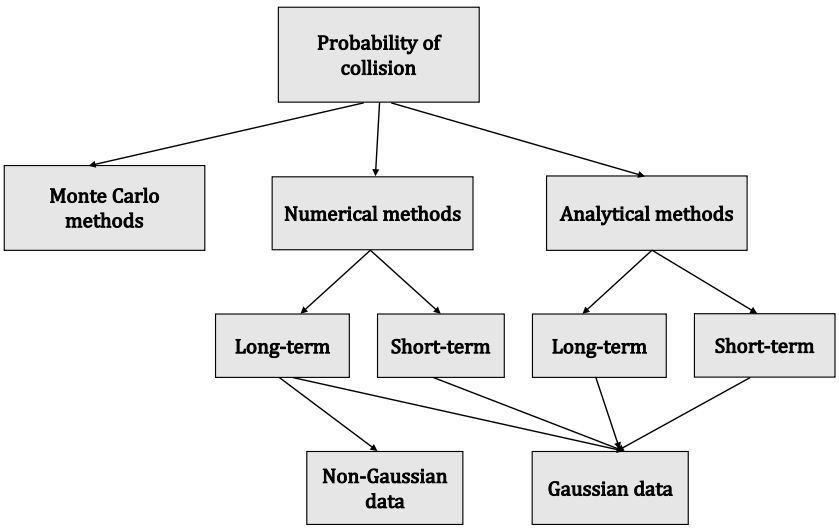


Figure 2.2 – Diagram summarising different collision probability methods.

In the process of reviewing the methods that have been developed for calculating the probability of a collision, we have organized these methods according to the diagram shown in Figure 2.2. We will briefly cover the highlighted categories in the following sections.

2.2.1 Monte Carlo methods

A wide variety of Monte Carlo (MC)-based methods are available to compute the probability of collision (Alfano, 2006b). In such methods, the state vector of each object is first sampled according to its distribution at the initial time t_0. Then, for each sampled state, the relative position and velocity are propagated to the time at which the distance between the two objects is at a minimum, called the Time of Closest Approach (TCA). There are various models for orbit propagation (Chen et al., 2017), each with its strengths and limitations, and the choice of which to use depends on the specific requirements of the conjunction assessment. At the point of the closest approach, a collision is deemed to occur if the relative distance is less than a small threshold. An estimate of the probability of collision is

$$\widehat{p}_c = \frac{N_\mathrm{c}}{N_\mathrm{t}},$$

where N_c is the number of counted collision samples out of a total of N_t samples .

This method is not well-suited to dealing with small probabilities, since the number of samples required to guarantee a good estimate is inversely proportional to the actual probability. For example, to attain a 1% relative accuracy for a true probability of 10^{-4} with a 95% confidence level, at least 10^9 independent random simulations are required (Dagum et al., 2000). The effort spent generating these trajectories is wasted since most of the trajectories will not contribute to the probability estimate. Another shortcoming of the MC method is that it fails to take advantage of possible dimension reduction by projecting quantities into the encounter plan; instead, it uses the 6-dimensional state vectors and the associated covariance matrix.

Advanced Monte Carlo techniques, such as Importance Sampling (IS)(Pastel, 2011), Subset Simulation (SS), Line Sampling (LS), and Brute Force Monte Carlo (BFMC) have been developed, aiming at increasing precision for a given computational effort and concentrating on the most valuable parts of the sample space. These techniques have been explored within the collision assessment community, and different MC-based methods have been proposed (Losacco et al., 2019; Hall, 2021; Hall et al., 2018; Li et al., 2022).

2.2.2 Numerical methods

In the last three decades, several numerical methods for computing (2.1) have been proposed. These methods rotate around the ideas of (i) ignoring the marginal component of the density in the direction of the relative velocity, (ii) reducing the two-dimensional integral to one dimension, and (iii) approximating it numerically. The main challenge in approximating (2.1) numerically is to propose an approach with acceptable precision and a realistic computational cost. Below, we present one of the earliest methods introduced by Foster and Estes (1992) and currently used by the National Aeronautics and Space Administration (NASA) as one of the methods to assess on-orbit risk.

This method is based on rotating the relative quantities to the UVW frame, defined as

$$U = \left(v_p \times v_s\right) / \left\|v_p \times v_s\right\|, \quad V = (v_s - v_p) / \left\|v_s - v_p\right\|, \quad W = U \times V.$$

where μ_p, μ_s, v_p and v_s are the current positions and velocities of the primary and secondary objects; all are 3×1 vectors. The displacement vector in the encounter plane, which is spanned by $\{U, W\}$, has a bivariate normal distribution with mean (μ_U, μ_W) and variance (σ_U^2, σ_W^2). To obtain these uncertainties, one needs to transform

each covariance, Ω_i^{-1} $(i = 1,2)$, from the UVW frame of the corresponding object to the encounter plane using a rotation matrix of the form

$$T(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}.$$

The projected position vector has a covariance matrix of the form $\Omega_j^{-1} = T(\theta_r)\tilde{\Omega}_j^{-1}T(\theta_r)^{\mathrm{T}}$, where $\tilde{\Omega}_j^{-1} = \mathrm{diag}(\sigma_{j1}^2, \sigma_{j3}^2, \sigma_{j3}^2)$ for $j = \{s, p\}$ and θ_r is the angle between the v axis and the primary object. The combined covariance matrix of the relative position is $\Omega_{\mathrm{UVW}}^{-1} = \Omega_s^{-1} + \Omega_p^{-1} = \mathrm{diag}(\sigma_U^2, \sigma_V^2, \sigma_W^2)$, where

$$\sigma_U^2 = \sigma_{s1}^2 + \sigma_{p1}^2, \quad \sigma_W^2 = \sigma_{s2}^2 \sin^2\theta_r + \sigma_{s3}^2 \cos^2\theta_r + \sigma_{p2}^2 \sin^2\theta_r + \sigma_{s3}^2 \cos^2\theta_r.$$

After integrating out the component along the V axis, the density function is

$$f(u,w) = \frac{1}{2\pi\sigma_U\sigma_V} \exp\left[-\frac{1}{2}\left\{\frac{(u-\mu_U)^2}{\sigma_U^2} + \frac{(w-\mu_W)^2}{\sigma_W^2}\right\}\right].$$

Expressing u, w, μ_U, and μ_W in polar coordinates, we obtain

$$u = r\sin\theta, \quad w = r\cos\theta, \quad \mu_U = \psi\sin\phi, \quad \mu_W = \psi\cos\phi,$$

where

$$r = \sqrt{u^2 + w^2}, \quad \psi = \sqrt{\mu_U^2 + \mu_W^2}.$$

The probability of collision is then defined as

$$p_c = \frac{1}{2\pi\sigma_U\sigma_W} \int_{r=0}^{\mathrm{HBR}} \int_{\theta=0}^{2\pi} \exp\left[-\frac{1}{2}\left\{\frac{\left(r\cos\theta - \psi\cos\phi\right)^2}{\sigma_W^2} + \frac{\left(\sin\theta - \psi\sin\phi\right)^2}{\sigma_U^2}\right\}\right] r\,\mathrm{d}r\,\mathrm{d}\theta,$$

where the combined hard-body radius, denoted by HBR, is a minimum safety threshold.

This method is implemented using a numerical discretization scheme, sufficiently accurate, but slower than other available algorithms if small steps of r and θ are used (Alfano, 2007).

In the same line of work, we find the method developed by Alfano (2005a), where (2.1) is expressed as series combinations of error functions and an exponential term. This series is then divided into odd and even components and approximated using Simpson's one-third rule. Patera (2001, 2005) simplifies the probability of collision to

a one-dimensional closed path integral about the perimeter of the hard-body circle. Patera's method does not require the space objects to be spherically shaped, so objects of an irregular shape can be taken into account. Comparisons of these methods and their performance can be found in Chan (2008), Alfano (2007) and Serra et al. (2016). In Section 2.7, we use the methods of Patera (2001) and Foster and Estes (1992). These give very similar results for the scenarios we study, with the difference between them being noticeable only in the 6th decimal place.

For cases where the uncertainty in the relative state vector is not Gaussian, most of the methods available for computing the probability of collision rely on numerical solutions. The techniques proposed in such studies vary from the use of Gaussian mixture models (GMM) as in Zhang et al. (2020), the reconstruction probability density function through the principle of maximum entropy (Adurthi and Singla, 2015) or MC-based techniques as in Jones and Doostan (2013). A brief comparison of numerical methods for computing the collision probability in short and long-term encounters is presented in Li et al. (2022).

2.2.3 Analytical methods

The methods previously described are based on numerical discretization of the integral rather than on analytical formulae. Chan (2008) was the first successful attempt to provide a closed-form expression for p_c under a set of simplifications. Chan's formulation is based on a scale transformation that reshapes the circular collision cross-section centered at (μ_x, μ_y) to an equivalent elliptic one and transforms the two-dimensional Gaussian density to an isotropic Gaussian density function with symmetrized position standard deviation σ. Again expressing $x = (x_1, x_2)$ in terms of the polar coordinate system, we have

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta, \quad \mu_1 = \psi\cos\phi, \quad \mu_2 = \psi\sin\phi$$

Expression (2.1) becomes

$$p_c = \frac{1}{2\pi\sigma^2}\int_0^{\mathrm{HBR}} r\exp\left(-\frac{r^2+\psi^2}{2\sigma^2}\right)\left\{\int_0^{2\pi}\exp\left(\frac{r\psi\cos\left(\theta-\phi\right)}{\sigma^2}\right)\mathrm{d}\theta\right\}\mathrm{d}r.$$

The inner integrand can be expressed using Taylor series expansion as

$$\exp\left(\frac{\psi r\cos\left(\theta-\phi\right)}{\sigma^2}\right) = \sum_{k=0}^{\infty}\frac{1}{k!}\frac{\psi^k r^k}{\sigma^{2k}}\cos^k\left(\theta-\phi\right),$$

and integrating this function with respect to θ gives

$$\int_0^{2\pi} \exp\left\{\frac{\psi r \cos\left(\theta-\phi\right)}{\sigma^2}\right)\mathrm{d}\theta = \sum_{k=0}^{\infty}\frac{1}{k!}\frac{\psi^k r^k}{\sigma^{2k}}\int_0^{2\pi}\cos^k\left(\theta-\phi\right)\mathrm{d}\theta. \tag{2.2}$$

As

$$\int_0^{2\pi}\cos^k\left(\theta-\phi\right)\mathrm{d}\theta = \begin{cases} 2\pi\dfrac{(k-1)!!}{k!!}, & \text{for even } k,\\ 0, & \text{for odd } k,\end{cases}$$

where $k!!$ denotes the double factorial of k, (2.2) becomes

$$\int_0^{2\pi}\exp\left(\frac{\psi r\cos\left(\theta-\phi\right)}{\sigma^2}\right)\mathrm{d}\theta = 2\pi\sum_{k=0}^{\infty}\frac{1}{(k!)^2}\left(\frac{\psi r}{2\sigma^2}\right)^{2k} = 2\pi I_0\left(\frac{\psi r}{\sigma^2}\right),$$

where

$$I_0(x) = \sum_{k=0}^{\infty}\frac{1}{(k!)^2}\left(\frac{x}{2}\right)^{2k},$$

is the zero-order modified Bessel function of the first kind.

Plugging this one-dimensional integral into the expression for the collision probability, we have

$$\begin{aligned} p_c &= \int_0^{\mathrm{HBR}}\frac{r}{\sigma^2}\exp\left(-\frac{r^2+\psi^2}{2\sigma^2}\right)I_0\left(\frac{\psi r}{\sigma^2}\right)\mathrm{d}r\\ &= \int_0^{\mathrm{HBR}}\frac{r}{\sigma^2}\exp\left(-\frac{r^2+\psi^2}{2\sigma^2}\right)\left(\sum_{k=0}^{\infty}I_0^{(k)}\right)\mathrm{d}r\\ &= \sum_{k=0}^{\infty}\int_0^{\mathrm{HBR}}\frac{r}{\sigma^2}\exp\left(-\frac{r^2+\psi^2}{2\sigma^2}\right)I_0^{(k)}\mathrm{d}r, \end{aligned}$$

where $I_0^{(k)} = \dfrac{1}{(k!)^2}\left(\dfrac{\psi r}{2\sigma^2}\right)^{2k}$. The probability of collision is then given by the infinite sum

$$p_c = \sum_{k=0}^{\infty}p_k,$$

with $p_k = \displaystyle\int_0^{\mathrm{HBR}}\frac{r}{\sigma^2}\exp\left(-\frac{r^2+\psi^2}{2\sigma^2}\right)I_0^{(k)}\mathrm{d}r$. The first term of this infinite series can be shown to be

$$p_0 = \mathrm{e}^{-v}\left(1-\mathrm{e}^{-u}\right),$$

where $u = \frac{1}{2}\left(\mathrm{HBR}/\sigma\right)^2$ and $v = \frac{1}{2}\left(\psi/\sigma\right)^2$ are dimensionless quantities, and p_k satisfies the recursive relationship

$$p_k = a_k p_{k-1} - b_k, \quad k = 1,2,\ldots$$

where

$$a_k = \frac{v}{k}, \quad b_k = \frac{u^k v^k}{k!k!} \mathrm{e}^{-(v+u)}.$$

This method gives an upper bound for the truncation error $S_n = \sum_{i=n+1}^{\infty} p_i$, viz

$$S_n < \frac{1}{n!(n+1)!} u^{n+1} v^n e^{-v} e^{uv}.$$

This allows users to choose the number of terms to include in the expansion based on the specific geometry of the conjunction and by comparing the results with other methods. For example, retaining the first term in this sum yields a one-dimensional Rice density function, which is easy to evaluate. The relative error of this one-term approximation is of order 10^{-5}, generally considered to be negligible for most encounters between space objects.

Another analytical method for satellite conjunction assessment proposed by Chan (2008) involves inverting the moment-generating function of the squared-miss distance. Serra et al. (2016) used the Laplace transform and D-finite functions to provide an analytic expression for the integral in (2.1). The formula is a product of an exponential term and a convergent power series with positive coefficients. Analytic bounds on the truncation error are also derived and are used to obtain accurate results. A closely-related method based on characteristic function inversion (CFI) is proposed by Bernstein et al. (2021). The approach uses CFI to evaluate the generalized chi-square distribution of the squared miss distance.

2.3 Statistical modeling of conjunction assessment

2.3.1 Statistical formulation of space conjunction

A parametric statistical model treats the available data y as the realisation of a random variable whose probability density function $f(y;\vartheta)$ is determined by unknown parameters ϑ, and the usual goal is to use y for inference about the value of a scalar parameter $\psi = \psi(\vartheta)$. Below, we follow Elkantassi and Davison (2022), and take ψ to be the miss distance and suppose, as before, that a collision occurs if the two objects come closer than a minimum distance $\psi_{\min}$ = HBR. We assume that the conjunction is sufficiently close that the relative motion can be taken to be linear. Suppose initially that the current positions μ_p and μ_s and velocities ν_p and ν_s of the two objects are known exactly. Define $\mu = \mu_s - \mu_p$ and $\nu = \nu_s - \nu_p$, so the second object is considered relative to an origin at the primary object. In this frame of reference and under relative

linear motion, the second object traverses the line $\mu + t\nu$, where $t \in \mathbb{R}$. Its distance from the origin, $(\mu + t\nu)^{\mathrm{T}}(\mu + t\nu)$, is minimised by choosing $t = -\nu^{\mathrm{T}}\mu/\nu^{\mathrm{T}}\nu$, at which point the minimum squared distance is $\psi^2 = \mu^{\mathrm{T}}\mu - \left(\mu^{\mathrm{T}}\nu\right)^2/\nu^{\mathrm{T}}\nu$. In terms of spherical polar coordinates, we have

$$\mu = \|\mu\| \left(\sin\theta_1 \cos\phi_1, \sin\theta_1 \sin\phi_1, \cos\theta_1\right)^{\mathrm{T}}, \tag{2.3}$$

$$\nu = \|\nu\| \left(\sin\theta_2 \cos\phi_2, \sin\theta_2 \sin\phi_2, \cos\theta_2\right)^{\mathrm{T}}, \tag{2.4}$$

where $\|\cdot\|$ denotes the Euclidean norm and $0 \le \theta_1, \theta_2 \le \pi$ and $0 \le \phi_1, \phi_2 < 2\pi$ are the polar and azimuthal angles for μ, ν. The minimum distance between the two objects, the miss distance, is

$$\psi = \|\mu\|(1 - \cos^2\beta)^{1/2} = \|\mu\||\sin\beta|, \tag{2.5}$$

where β, the angle between the location and velocity vectors μ and ν, satisfies

$$\cos\beta = \sin\theta_1 \cos\phi_1 \sin\theta_2 \cos\phi_2 + \sin\theta_1 \sin\phi_1 \sin\theta_2 \sin\phi_2 + \cos\theta_1 \cos\theta_2. \tag{2.6}$$

When $\beta = 0$, we distinguish two cases: if $\mu^{\mathrm{T}}\nu < 0$ the second object will pass through the origin, leading to a collision, whereas if $\mu^{\mathrm{T}}\nu > 0$ the second object is heading away from the origin, so its current position is the closest it will come to the first object. More generally, $\psi < \|\mu\|$ only if $\cos\beta < 0$, i.e., $\pi/2 < \beta < 3\pi/2$. For μ and ν to be collinear but pointing in opposite directions we need $\phi_2 = \pi + \phi_1$ and $\theta_2 = \pi - \theta_1$, and then $\cos\beta = \cos(\theta_1 + \theta_2) = -1$, so $\beta = \pi$ and hence $\psi = 0$, as expected. To lighten the notation below we write $\vartheta = (\psi, \lambda)$ where

$$\lambda = (\theta_1, \phi_1, \|\nu\|, \theta_2, \phi_2). \tag{2.7}$$

In parametrizing the relative state vector as a six-dimensional parameter vector (ψ, λ), it is important to ensure that the transformation preserves the parameter space of ϑ. In particular, when $\Theta = \Psi \times \Lambda$, where Ψ and Λ are the parameter spaces for ψ and λ, we say that the parametrization is variation independent. If the allowable values of ψ were to depend on λ, this would introduce irregularities into the model, and the usual asymptotic theory might not apply. We show below that this is the case when the minimum distance ψ is expressed in Cartesian coordinates.

Assume that the relative position and relative velocity vectors in the Cartesian coordinate system are $(\mu, \nu) = (\mu_1, \mu_2, \mu_3, \nu_1, \nu_2, \nu_3)$. We would like to express one component,

for example, μ_1, in terms of ψ and the other parameters. We proceed as follows

$$\begin{aligned}
\psi^2 &= \mu^{\mathrm{T}}\mu - \left(\mu^{\mathrm{T}}\nu\right)^2 / \nu^{\mathrm{T}}\nu, \\
&= \mu_1^2 + \mu_2^2 + \mu_3^2 - \left(\mu_1\nu_1 + \mu_2\nu_2 + \mu_3\nu_3\right)^2 / \left(\nu_1^2 + \nu_2^2 + \nu_3^2\right), \\
&= \mu_1^2\left(\frac{\nu_2^2+\nu_3^2}{\nu_1^2+\nu_2^2+\nu_3^2}\right) + \mu_2^2\left(\frac{\nu_1^2+\nu_3^2}{\nu_1+\nu_2^2+\nu_3^2}\right) + \mu_3^2\left(\frac{\nu_1^2+\nu_2^2}{\nu_1^2+\nu_2^2+\nu_3^2}\right) \\
&\quad - 2\mu_1\frac{\nu_1(\nu_2\mu_2+\nu_3\mu_3)}{\nu_1^2+\nu_2^2+\nu_3^2} - \frac{2\mu_2\nu_2\mu_3\nu_3}{\nu_1^2+\nu_2^2+\nu_3^2}, \\
&= A\mu_1^2 - 2B\mu_1 + C,
\end{aligned}$$

where

$$A = \frac{\nu_2^2+\nu_3^2}{\nu_1^2+\nu_2^2+\nu_3^2}, \quad B = \frac{\nu_1(\nu_2\mu_2+\nu_3\mu_3)}{\nu_1^2+\nu_2^2+\nu_3^2}, \quad C = \frac{\nu_1^2(\mu_2^2+\mu_3^2)+(\mu_2\nu_3-\mu_3\nu_2)^2}{\nu_1^2+\nu_2^2+\nu_3^2}.$$

This implies that μ_1 solves the quadratic equation $A\mu_1^2 + 2B\mu_1 + (C-\psi^2) = 0$, i.e.,

$$\mu_1 = \frac{B \pm \sqrt{D}}{A}, \quad D = B^2 + A(\psi^2 - C),$$

which can be simplified to

$$\mu_1 = \frac{\nu_1(\nu_2\mu_2+\nu_3\mu_3) \pm \sqrt{\Delta}}{\nu_2^2+\nu_3^2},$$

where

$$\begin{aligned}
\Delta &= \nu_1^2(\nu_2\mu_2+\nu_3\mu_3)^2 + [\psi^2(\nu_1^2+\nu_2^2+\nu_3^2) - \{\nu_1^2(\mu_2^2+\mu_3^2) + (\mu_2\nu_3-\mu_3\nu_2)^2\}](\nu_2^2+\nu_3^2) \\
&= \nu^{\mathrm{T}}\nu\left\{\psi^2(\nu_2^2+\nu_3^2) - (\mu_2\nu_3-\mu_3\nu_2)^2\right\},
\end{aligned}$$

which will give an expression for μ_1 in terms of the interest parameter ψ and the nuisance parameter $\lambda = (\mu_2, \mu_3, \nu_1, \nu_2, \nu_3)$. Note, however, that we require that $\Delta > 0$, which implies that we must have $\psi^2 > (\mu_2\nu_3 - \mu_3\nu_2)^2/(\nu_2^2+\nu_3^2)$, and this is only zero if (μ_2, μ_3) and (ν_2, ν_3) are collinear, in which case the satellite can pass through the origin. To see this another way, $(\mu_2\nu_3 - \mu_3\nu_2)/(\nu_2^2+\nu_3^2)^{1/2}$ is the projection of (μ_2, μ_3) onto a unit vector $(-\nu_3, \nu_2)/(\nu_2^2+\nu_3^2)^{1/2}$ orthogonal to (ν_2, ν_3).

The above expressions give two possible values for μ_1, but the geometry implies that we must choose the solution for which $\mu^{\mathrm{T}}\nu < 0$, as otherwise, the satellite is heading away from the origin. For any specified distance ψ greater than the shortest distance, for which $\Delta = 0$, it is clear from the geometry that if $\nu_1 > 0$, then we should take the

root with the minus sign, giving

$$\mu_1(\psi,\lambda) = \frac{\nu_1(\nu_2\mu_2+\nu_3\mu_3) - \left[\nu^{\mathrm{T}}\nu\left\{\psi^2(\nu_2^2+\nu_3^2) - (\mu_2\nu_3-\mu_3\nu_2)^2\right\}\right]^{1/2}}{\nu_2^2+\nu_3^2}.$$

The argument is that if $\nu_1 > 0$ and we denote the two roots by μ_1^- and $\mu_1^+ = \mu_1^- + 2\Delta'$, where $\Delta' > 0$, then

$$\mu_1^+\nu_1 + \mu_2\nu_2 + \mu_3\nu_3 = \mu_1^-\nu_1 + \mu_2\nu_2 + \mu_3\nu_3 + 2\nu_1\Delta' > \mu_1^-\nu_1 + \mu_2\nu_2 + \mu_3\nu_3.$$

This implies that the root for $\mu^{\mathrm{T}}\nu < 0$ must be given by μ_1^-. Likewise, we should choose μ_1^+ if $\nu_1 < 0$.

The constraint on ψ implies that it is not variation independent of λ. This might cause problems when eliminating λ from the log-likelihood function, which is why we prefer the polar-based parametrization given in (2.5-2.7).

In the above deterministic setting, the collision probability $p \equiv p(\vartheta)$ takes two possible values,

$$p(\vartheta) = \begin{cases} 0, & \psi > \psi_{\min}, \\ 1, & 0 \le \psi \le \psi_{\min}, \end{cases} \tag{2.8}$$

so a decision-maker can make an ideal decision. In reality, of course, both μ and ν are observed with error. In the next section we follow the literature and assume that the available observations on the positions and velocities of the two objects have a multivariate normal distribution with a known covariance matrix.

2.3.2 Short-term encounters

In short-term encounters, a simplified version of the problem treats the relative velocity vector ν as known. In this case, the last three components of y and $\eta(\vartheta)$ corresponding to the relative velocity can be dropped, and only the 3×3 corner of Ω^{-1} pertaining to the relative position need be retained. The conjunction can then be visualised in the encounter plane, which is normal to ν and passes through the origin.

To understand how the projection is performed, assume for now the random data consist of a 3×1 vector y containing the estimated position of the second object relative to the first, and it is assumed that $y \sim \mathcal{N}_3(\mu,\Omega^{-1})$, where μ and Ω are of respective dimensions 3×1 and 3×3. Let A be a 3×3 orthogonal matrix whose final column is $\nu/\|\nu\|$ and whose other columns are chosen so that the upper left 2×2 corner of $A^{\mathrm{T}}\Omega^{-1}A$ is diagonal, say, $D = \mathrm{diag}(d_1^2, d_2^2)$. If so, then $x' = A^{\mathrm{T}}y \sim \mathcal{N}_3(A^{\mathrm{T}}\mu, A^{\mathrm{T}}\Omega^{-1}A)$,

and the first two components of x' are independent, with distribution $\mathcal{N}_2(\xi, D)$, say. Hence ξ is the projection of the true position of the second object along its velocity vector onto the encounter plane, and thus $\psi = \|\xi\|$ is the miss distance. In terms of the coordinates in the encounter plane defined by A we can write $\xi = (\psi\cos\lambda, \psi\sin\lambda)$ and $x = (x_1, x_2)$, where $\psi > 0$ is the parameter of interest and $\lambda \in [0, 2\pi)$ is the nuisance parameter. We describe how to obtain A later in this section.

The transformation from y to x' is invertible and depends only on the known quantities v and Ω, so no statistical information is lost in using x' instead of y. As discussed before, in satellite conjunction analysis, it is customary to ignore the coordinate x_3 of y in the direction orthogonal to the encounter plane, which in statistical terms amounts to basing inference on the joint density of x_1 and x_2, leading to the log likelihood (2.11).

To obtain a suitable 3×3 projection matrix A, note that if we define

$$B = (b_1, b_2, v) = \begin{pmatrix} 0 & v_2^2 + v_3^2 & v_1 \\ v_3 & -v_1 v_2 & v_2 \\ -v_2 & -v_1 v_3 & v_3 \end{pmatrix}, \quad N = \begin{pmatrix} \|b_1\|^{-1} & 0 & 0 \\ 0 & \|b_2\|^{-1} & 0 \\ 0 & 0 & \|v\|^{-1} \end{pmatrix},$$

then the first two columns of the orthonormal matrix BN span the encounter plane. If C denotes the 3×2 matrix containing these columns, then $C^{\mathrm{T}} y$ is the orthogonal projection of y onto the encounter plane. If VDV^{T} is the spectral decomposition of $C^{\mathrm{T}}\Omega^{-1}C$, then $(CV)^{\mathrm{T}}\Omega^{-1}(CV) = V^{\mathrm{T}}C^{\mathrm{T}}\Omega^{-1}CV = D = \mathrm{diag}(d_1^2, d_2^2)$. Thus $(CV)^{\mathrm{T}} y$ is bivariate normal with mean $\xi = (CV)^{\mathrm{T}}\mu$ and covariance matrix D. Hence if we let $A = (CV, v/\|v\|)$, then the first two elements of $x' = A^{\mathrm{T}} y$ are independent and satisfy $x_1 \sim \mathcal{N}(\xi_1, d_1^2)$ and $x_2 \sim \mathcal{N}(\xi_2, d_2^2)$.

2.3.3 Inference for the miss distance

In Sections 2.3 and 2.3.2, we developed a statistical model and identified a suitable set of parameters for the conjunction assessment problem. The next step is to make inferences about the miss distance. However, the reader may wonder why we have chosen to focus on the miss distance rather than the probability of collision presented in Section 2.2. This is an important point, as we believe that the miss distance is a more appropriate metric for satellite conjunction assessment for specific reasons, which we will discuss in detail in Section 2.5. Before doing so, it is practical to introduce some basic tools and clarify the notation while discussing inference for the miss distance.

Assume that vector y containing the observed position and velocity of the second

object relative to the first has a six-dimensional normal distribution with mean vector $\eta(\vartheta) = \{\mu(\psi,\lambda)^{\mathrm{T}}, \nu(\lambda)^{\mathrm{T}}\}^{\mathrm{T}}$ given by equations (2.3)-(2.6) and known 6×6 covariance matrix Ω^{-1}, whose inverse Ω is known as the dispersion matrix.

After dropping irrelevant additive constants, the log-likelihood function is

$$\ell(\vartheta) = -\tfrac{1}{2}\{y-\eta(\vartheta)\}^{\mathrm{T}}\Omega\{y-\eta(\vartheta)\}, \tag{2.9}$$

where $\vartheta = (\psi,\lambda)$ contains the miss distance ψ for which inference is required and the five-dimensional nuisance parameter vector $\lambda = (\theta_1,\phi_1,\|\nu\|,\theta_2,\phi_2)$. The maximum likelihood estimator $\widehat{\vartheta}$ based on the observed relative distance and velocity contained in the 6×1 vector y satisfies $\eta(\widehat{\vartheta}) = y$, and the observed information matrix is

$$J(\widehat{\vartheta}) = \left.\frac{\partial\eta^{\mathrm{T}}(\vartheta)}{\partial\vartheta}\Omega\frac{\partial\eta(\vartheta)}{\partial\vartheta^{\mathrm{T}}}\right|_{\vartheta=\widehat{\vartheta}}, \tag{2.10}$$

where $\partial\eta^{\mathrm{T}}(\vartheta)/\partial\vartheta$ is a 6×6 matrix. Hence $\ell(\widehat{\vartheta}) = 0$, $\widehat{\vartheta}$ is simply a transformation of y, and $\widehat{\vartheta}_\psi = (\psi,\widehat{\lambda}_\psi)$ minimises the weighted sum of squares in (2.9) for fixed ψ. These quantities allow inference on ψ based on the likelihood root (1.1) and the Wald statistic (1.3), while the more accurate modified likelihood root (1.14) also requires (1.27).

Here there are six parameters and a single six-dimensional observation y, so the sample size is $n = 1$, and it appears that we cannot expect large-sample approximations to apply. However, the covariance matrix for an average of n independent observations would be $(n\Omega)^{-1}$, so a large sample size n is equivalent to a small variance for the observations or equivalently large Ω, which is the correct gauge of accuracy; c.f. Secetion 2.5.1

As we saw in Section 2.3.2, the model simplifies greatly when the velocity vector ν is known. In this case, y and $\eta(\vartheta)$ are replaced by the projections of the observed position vector and the true position of the second object into the encounter plane. To do so, the first two elements of y are scaled by C, and rotated by V to obtain a vector with diagonal covariance matrix D. The projected three-dimensional vector x' has a Gaussian distribution $x' = A^T y \sim N_3(\xi',\Sigma)$ where

$$\xi' = A^T\mu = \begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix}, \quad \Sigma = A^T\Omega^{-1}A = \begin{pmatrix} D & \dfrac{(CV)^T\Omega^{-1}\nu}{||\nu||} \\ \dfrac{\nu^T\Omega^{-1}(CV)}{||\nu||} & \dfrac{\nu^T\Omega^{-1}\nu}{||\nu||^2}\end{pmatrix}.$$

The structure of Σ shows that (x_1,x_2), and x_3 are correlated since the local frame naturally depends on the direction ν. In practice, uncertainty in the direction of ν

can be integrated out by simply dropping $x_3 \sim N(\xi_3, \nu^T \Omega^{-1} \nu / ||\nu||^2)$. However, we can instead base inference about ϑ using the full likelihood of the 3-dimensional state vector. This can be written as

$$\ell_F(\vartheta) = \ell(\vartheta) + \ell_\perp(\vartheta),$$

where $\ell(\vartheta)$ is the likelihood for (x_1, x_2) in the encounter plane described in (2.11) and $\ell_\perp(\vartheta)$ is the contribution of $x_3|(x_1, x_2) = (a_1, a_2)$, which follows a Gaussian distribution $N(\bar{\xi}, \bar{\sigma}^2)$, with

$$\bar{\xi} = \xi_3 + \frac{\nu^{\mathrm{T}} \Omega^{-1} (CV)}{||\nu||} \Omega^{-1} D^{-1} \begin{pmatrix} a_1 - \xi_1 \\ a_2 - \xi_2 \end{pmatrix}, \quad \bar{\sigma}^2 = \frac{\nu^{\mathrm{T}} \Omega^{-1} \nu^{\mathrm{T}}}{||\nu||} - \frac{\nu^{\mathrm{T}} \Omega^{-1} (CV)}{||\nu||} D^{-1} \frac{(CV)^{\mathrm{T}} \Omega^{-1} \nu}{||\nu||}.$$

The observed information using the full likelihood is $j_{\mathrm{F}}(\vartheta) = j(\vartheta) + j_\perp(\vartheta)$. By comparing the information contained in the full likelihood function with the marginal likelihood, and determine whether the knowledge of x_3 provides additional information about the parameters ϑ.

Consider now the log likelihood based on the joint density of $x = (x_1, x_2)$ with mean $\xi = (\xi_1, \xi_2) = (\psi \cos\lambda, \psi \sin\lambda)$, and a diagonal covariance matrix D,

$$\ell(\psi, \lambda) = -\frac{1}{2} \left\{ \frac{(x_1 - \psi\cos\lambda)^2}{d_1^2} + \frac{(x_2 - \psi\sin\lambda)^2}{d_2^2} \right\}, \quad \psi > 0, 0 \le \lambda < 2\pi. \tag{2.11}$$

The detailed calculations for inference based on this model may be found in Section 2.6.1, where insight into the general problem can be gained by considering the case $d_1 = d_2$.

The approach advocated by Carpenter (2019) is related to the discussion above, but the confidence statements therein are based on the marginal quantiles of the miss distance distribution rather than on likelihood theory. That approach does not allow for uncertainty about the nuisance parameter and appears to be equivalent to basing inference on the Wald statistic, which can perform very poorly in nonlinear settings.

2.4 Calibration, decisions and evidence

Statistical inference involves statements about the properties of a probability distribution that is assumed to have given rise to observed data y^{o}. In the simplest setting the distribution depends only on a scalar parameter ψ, which is the focus of interest, and the likelihood function $L(\psi) = f(y^{\mathrm{o}}; \psi)$ is used to compare the plausibility of different values of ψ as explanations for y^{o}. The best-fitting model is provided by the maximum

likelihood estimate based on y^o, $\widehat{\psi}^o$, and the relative likelihood function $L(\psi)/L(\widehat{\psi}^o)$, which has maximum value 1, allows values of ψ to be compared. A 'pure likelihood' approach (e.g., Edwards, 1972, Chapter 3) treats any ψ for which $L(\psi) \geq c\,L(\widehat{\psi}^o)$ as plausible, but with c chosen essentially arbitrarily. In practice further information is typically used to choose c.

Bayesian inference treats ψ as a random variable and is based on a density $\pi(\psi)$ that weights values of ψ according to their plausibility prior to seeing the data. This is updated in light of the observed data y^o using Bayes' formula, resulting in the posterior distribution function

$$\Pr(\psi \leq \psi_0 \mid y^o) = \frac{\displaystyle\int_{-\infty}^{\psi_0} L(\psi)\pi(\psi)\mathrm{d}\psi}{\displaystyle\int_{-\infty}^{\infty} L(\psi)\pi(\psi)\mathrm{d}\psi}. \tag{2.12}$$

Clearly this calculation depends on the prior density $\pi(\psi)$; if this is badly chosen then (2.12) may have poor properties when used repeatedly. The most obvious choice of prior in the satellite conjunction setting is uniform on the position of the secondary space object, but this has the undesirable properties mentioned in Section 2.5.1.

When the losses due to possible evasive actions can be specified, the data can be used to choose the action that minimises the expected posterior loss. For the simplest possible formulation in the collision avoidance context, suppose that ψ represents the unknown miss distance, that the two actions $a = 0$ and $a = 1$ correspond to 'do nothing' and 'take evasive action', and that the loss l_{ae} when action a is taken and event e occurs is as given in Table 2.1; the loss in doing nothing in the case of no collision is zero. If $\psi \leq \psi_0$ results in a collision, then the posterior expected loss due to taking action $a \in \{0, 1\}$ is

$$l_{a0}\Pr(\psi > \psi_0 \mid y^o) + l_{a1}\Pr(\psi \leq \psi_0 \mid y^o) = l_{a0}(1-p) + l_{a1}p,$$

say, which is minimised by doing nothing if $l_{01}p < l_{10}(1-p) + l_{11}p$. Equivalently, evasive action should be taken if $\Pr(\psi \leq \psi_0 \mid y^o) \geq l_{10}/(l_{10} + l_{01} - l_{11})$. Thus if the losses are known explicitly, they provide a threshold for action, and the use of a decision rule such as "take evasive action if the posterior probability of collision exceeds 10^{-4}" corresponds to an implicit ratio of losses. This decision setup could be made more realistic, and in any case would only be regarded as guidance in a practical setting; our point is that it provides a rational basis for considering action when (2.12) exceeds a threshold, and explicitly links that threshold to potential losses.

Table 2.1 – Basic decision analysis for satellite conjunction, with losses l_{ae} corresponding to action a and event e.

Action	Event	
	No collision	Collision
Do nothing, $a = 0$	$l_{00} = 0$	l_{01}
Evasive action, $a = 1$	l_{10}	l_{11}

Bayesian inference has some appeal, but nevertheless other approaches to calibrating the likelihood are often preferred. Repeated sampling inference invokes hypothetical repetition of the random experiment that is presumed to have led to the observed data (Fisher, 1973, pp. 33–38). In the simplest case $\widehat{\psi}^{\mathrm{o}}$ is regarded as a realization of a random variable $\widehat{\psi}$ that has a normal distribution, $\mathcal{N}(\psi, \lambda^2)$, under repeated sampling, with λ known. This implies that

$$\Pr(\widehat{\psi} \leq \widehat{\psi}^{o}; \psi) = \Phi\left\{(\widehat{\psi}^{\mathrm{o}} - \psi)/\lambda\right\}, \tag{2.13}$$

where Φ denotes the standard normal cumulative distribution function. The significance function (2.13), also called the confidence distribution or P-value function (Fraser, 2017, 2019; Schweder and Hjort, 2016), is then used for inference on ψ, as discussed in Section 1.2.3 of Chapter 1. Our later discussion simplifies if framed in terms of the equivalent 'evidence function'

$$p^{\mathrm{o}}(\psi) = \Pr(\widehat{\psi} \geq \widehat{\psi}^{\mathrm{o}}; \psi) = 1 - \Phi\left\{(\widehat{\psi}^{\mathrm{o}} - \psi)/\lambda\right\} = \Phi\left\{(\psi - \widehat{\psi}^{\mathrm{o}})/\lambda\right\}, \tag{2.14}$$

and we use this henceforth. For example, the null hypothesis that $\psi = \psi_0$ can be tested against the alternative that $\psi > \psi_0$ by computing the significance probability

$$p_{\mathrm{obs}} = p^{\mathrm{o}}(\psi_0), \tag{2.15}$$

small values of which are regarded as evidence against the null hypothesis in favour of the alternative. Likewise a two-sided $(1-2\alpha) \times 100\%$ confidence interval $\mathscr{I}_{1-2\alpha}$ for the value of ψ underlying the data, the so-called 'true value', has as its lower and upper limits L_α and U_α the solutions to the equations $p^{o}(L_\alpha) = \alpha$ and $p^{\mathrm{o}}(U_\alpha) = 1 - \alpha$, and this yields $\mathscr{I}_{1-2\alpha} = (L_\alpha, U_\alpha) = (\widehat{\psi}^{\mathrm{o}} - \lambda z_{1-\alpha}, \widehat{\psi}^{\mathrm{o}} - \lambda z_\alpha)$, where z_p, the p quantile of the standard normal distribution, satisfies $\Phi(z_p) = p$ and $0 < p < 1$. The limits of $\mathscr{I}_{1-2\alpha}$ simplify to the familiar $\widehat{\psi}^{\mathrm{o}} \pm \lambda z_{1-\alpha}$ on recalling that $z_{1-\alpha} = -z_\alpha$. In this ideal case the inferences are perfectly calibrated: under repeated sampling with $\psi = \psi_0$ the significance probability p_{obs} has an exact uniform distribution and $\mathscr{I}_{1-2\alpha}$ contains

ψ_0 with probability exactly $1-2\alpha$, for any α in the interval $[0, 0.5]$. When $\psi = \psi_0$, therefore, there is a probability p_{obs} that a decision to reject this hypothesis in favor of the alternative based on a significance probability p_{obs} will be incorrect.

The evidence function $p^{\text{o}}(\psi)$ is increasing in ψ and has the properties of a cumulative distribution function, so its derivative

$$\frac{\partial p^{\text{o}}(\psi)}{\partial \psi}, \tag{2.16}$$

has the formal properties of a probability density function for ψ. It can be regarded as a frequentist summary of the information about ψ based on the observed data; see Schweder and Hjort (2016), where it is called a confidence density. Unlike the posterior density obtained from (2.12), no prior information is involved, and despite the use of the word 'density' for (2.16), ψ is regarded as an unknown constant rather than as the value of a random variable.

2.5 Conjunction probability or miss distance ?

2.5.1 Limitations of collision probability

As discussed in Section 2.3.2, a simplified version of the problem treats the relative velocity vector v as known in short-term encounters. In this case, the observed position y of the second object, its actual position μ, and the density of y, can be projected orthogonally in the direction of v into the encounter plane, leading to the projected observed position x having a bivariate normal density $f(x;\xi)$ that is centred at the projected true position ξ with known diagonal covariance matrix $D = \text{diag}(d_1^2, d_2^2)$.

As shown in the left-hand panel of Figure 2.3, ξ and x are two-dimensional vectors: the unknown point at which the second object will actually pass through the encounter plane, ξ, is at a distance $\psi = \|\xi\|$ from the origin. Using this notation, the collision probability is zero unless $\psi \le \psi_{\min}$, and the observed x is a realisation of a bivariate normal variable with mean ξ. The right-hand panel of Figure 2.3 shows how the collision probability is being computed: the density is assumed to be centered at x and the estimator $\widehat{p}_c$ is the integral of this density over the disk of radius $\psi_{\min}$ around the origin. Thus we can write $\widehat{p}_c = p_c(x)$, where

$$p_c(\xi) = \int_{\{x' : \|x'\| \le \psi_{\min}\}} f(x';\xi)\,\mathrm{d}x'. \tag{2.17}$$

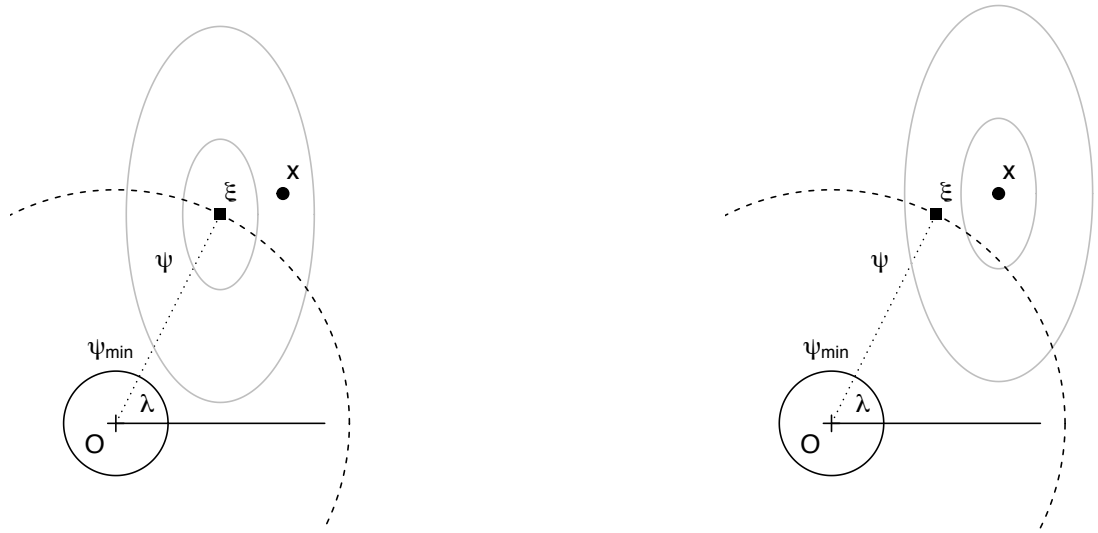


Figure 2.3 – Statistical formulation of satellite conjunction in the encounter plane. The primary object is at the origin O and the solid circle around it has radius $\psi_{\min}$, the combined hard-body radius. The true position at which the second object will cross the encounter plane, ξ, can be expressed in terms of the polar coordinates ψ and λ. The noisy observation of the second object will cross the encounter plane at x. Left panel: the ellipses indicate the true density of x, with mean at ξ. Right panel: the ellipses indicate the assumed density of x when computing the collision probability estimate $\widehat{p}_c$.

One might regard $p_c(x)$ as an estimate of $p_c(\xi)$, since the unknown ξ is replaced with the known x in computing the integral. Seen through the lens of the statistical model, however, it is not immediately obvious why $p_c(\xi)$ is of interest, as it is the probability that the noisy observation x will appear to pass within the hard-body radius, rather than the probability (2.17) that the second object itself will do so. Moreover the form of the contours of $f(x;\xi)$ implies that there is more probability outside the circle of radius ψ than inside it, so x will tend to be further away from the origin than ξ. Another way to see this is to note that the Euclidean norm of x, $\|x\|$, satisfies

$$\mathrm{E}(\|x\|^2) = \mathrm{E}(x_1^2 + x_2^2) = \xi_1^2 + \xi_2^2 + d_1^2 + d_2^2 = \psi^2 \left\{1 + (d_1^2 + d_2^2)/\psi^2\right\}, \tag{2.18}$$

and hence $\mathrm{E}\{\|x\|\} \approx \psi \left\{1 + (d_1^2 + d_2^2)/\psi^2\right\}^{1/2}$ for large ψ, i.e., the mean length of x exceeds ψ by an amount that depends on ξ and D. Hence $\widehat{p}_c = p_c(x)$ will tend to be smaller than $p_c(\xi)$. Figure 2.4 shows the effect of this in a test case described thoroughly in Section 2.7.4, in which $\xi = (11.84, -1.36)^{\mathrm{T}}$m, so $\psi = 11.92$m. Each panel compares the values of $p_c(\xi)$ with boxplots of the values of $p_c(x)$ for 20,000 values of x generated as bivariate normal, $\mathcal{N}_2\{\xi, c^2\mathrm{diag}(d_1^2, d_2^2)\}$, with $d_1 = 25.1$m, $d_2 = 11.61$m and various values of c^2. The collision probability is computed using a contour inte-

gral transformation (Patera, 2005), which is then approximated numerically using the trapezoidal rule. The values of $p_c(x)$ vary over many orders of magnitude and can be much lower than $p_c(\xi)$. For example, the boxplot for $c^2 = 10^{-2}$ and hard body radius 5m shows that although $p_c(\xi) \approx 10^{-3}$, around 25% of the values of $p_c(x)$ are below 10^{-4}. The panels of Figure 2.5, which show the probabilities $p_c(x)$ for a subset of the simulated observations x contributing to Figure 2.4, also illustrate the potential for the occurrence of extremely small values of $p_c(x)$, both when collision is certain and when it will not occur.

Despite these comments, one reason to use $\widehat{p}_c$ is that it is a Bayesian estimator. If a prior density $\pi(\xi)$ for ξ is placed on the encounter plane, then the posterior probability that ξ lies within the hard-body radius is given in terms of the posterior density of ξ, i.e.,

$$f(\xi \mid x) = \frac{f(x;\xi)\pi(\xi)}{\int f(x;\xi)\pi(\xi)\,\mathrm{d}\xi},$$

by

$$\Pr(\psi \le \psi_{\min} \mid x) = \Pr(\|\xi\| \le \psi_{\min} \mid x) = \int_{\{\xi:\|\xi\|\le\psi_{\min}\}} f(\xi \mid x)\,\mathrm{d}\xi, \tag{2.19}$$

and if $\pi(\xi)$ is constant and $f(x;\xi)$ is bivariate normal, then $f(\xi \mid x) = f(x;\xi)$ and (2.19) equals $\widehat{p}_c = p_c(x)$. This explains why $\widehat{p}_c$ is a plausible estimator of (2.8), but does not alter its downward bias. Moreover a constant prior for ξ is improper, with the undesirable property that the ratio of the probability inside any disk around the origin is zero relative to the probability outside that disk, thus expressing a prior belief that ξ is infinitely far from the origin, i.e., the second object will traverse the encounter plane infinitely far from the first. This provides an alternative explanation of the behaviour illustrated in Figure 2.4.

Another practical problem that arises using the probability of collision as a risk assessment tool is a likely sense of false security. In general, the lower the quality of the data, the less confidence one has in the findings. However, using (2.17), we notice that the more uncertain we are about the true location of the two space objects, the less likely they will collide. This counter-intuitive phenomenon is referred to as "probability dilution" (Balch et al., 2019). Take for example the test case we described before, in which $\xi = (11.84, -1.36)^{\mathrm{T}}$m, so $\psi = 11.92$m. In left panel of Figure 2.6, we choose HBR = 5m, and for ease of illustration, we set $d = d_1 = d_2$ to vary from 0.01 to 200. Reading this panel from left to right, we see that the probability of collision increases as the relative uncertainty grows. However, at $d = 1.6$ the probability of collision eventually reaches a maximum $p_{c_{\max}} = 6.485 \times 10^{-2}$, and past that point, as relative uncertainty rises, the probability of collision declines. The decreasing region corresponds to the probability

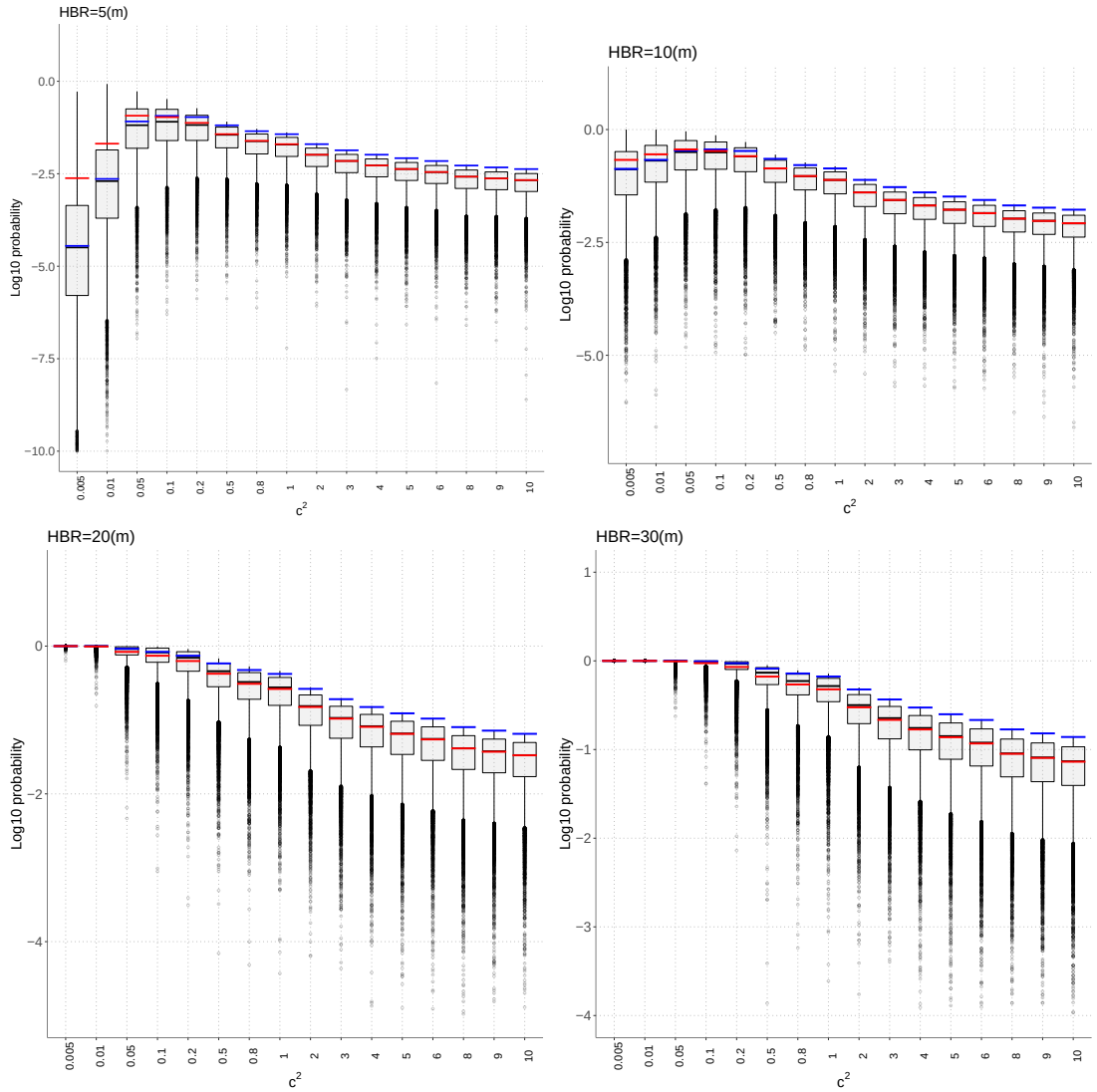


Figure 2.4 – Behaviour of $\widehat{p}_c = p_c(x)$ for hard-body radius $\psi_{\min} = 5, 10, 20$ and 30m in Case Study C, with $\psi = 11.92$m, as a function of the variance of $x \sim \mathcal{N}_2\{\xi, c^2\text{diag}(d_1^2, d_2^2)\}$. For each c^2 in $0.005, 0.01, \ldots, 10$ we computed $p_c(\xi)$ (blue segments) and $p_c(x)$ (boxplots) for 20,000 simulated values of x. The average values of $p_c(x)$ are shown by the red segments. In the top panels $\psi > \psi_{\min}$, so $p_c(\xi) \to 0$ as $c \to 0$, whereas in the bottom two panels $\psi < \psi_{\min}$, then $p_c(\xi) \to 1$ as $c \to 0$.

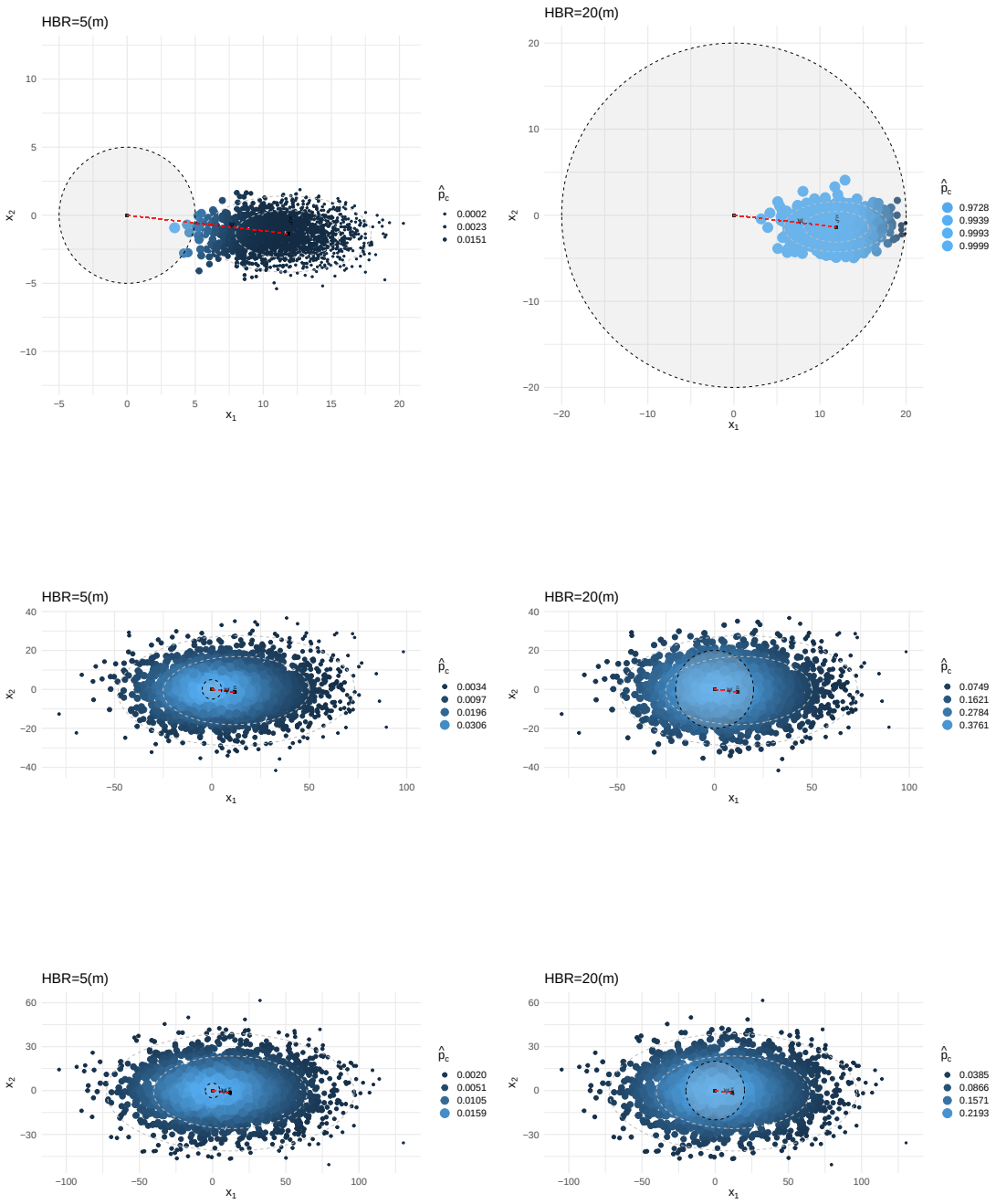


Figure 2.5 – Encounter plane for Case Study C, with hard body radius (HBR, 5m in the left-hand panels and 20m in the right-hand panels) shown as the dashed circle, and the true point ξ at which a second object will traverse the plane (at the end of the dashed red line). The point ξ lies inside the HBR in the right-hand panels, resulting in a collision, but outside the HBR in the left-hand panels. In each panel the blue points show 1000 simulated points x, where $x \sim \mathcal{N}_2(\xi, c^2 D)$, with $c^2 = 10^{-2}, 1$, and 2 (top to bottom). The dashed ellipses show the shape of the matrix $D = \text{diag}(d_1^2, d_2^2)$. The estimated collision probability $p_c(x)$ corresponding to each x is indicated by the size and shade of its blue point.

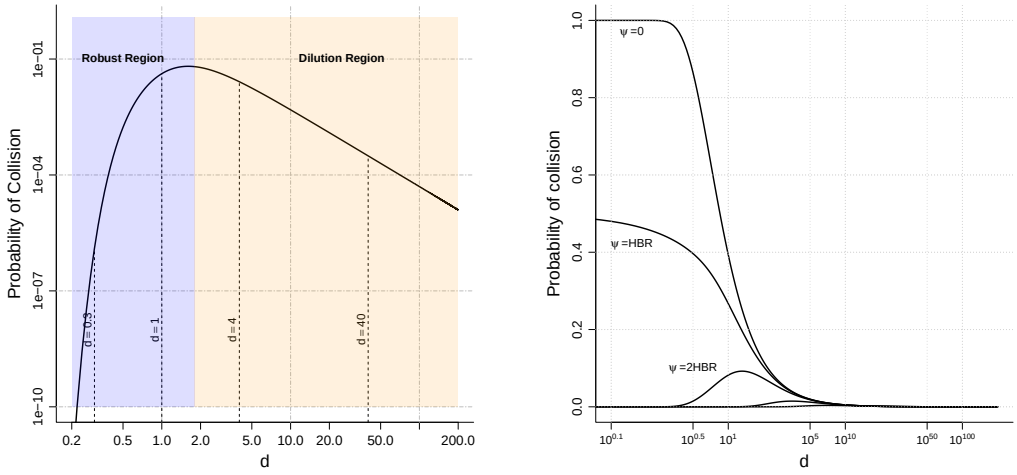


Figure 2.6 – Left panel: Collision probability as a function of the relative uncertainty for case study C with HBR = 5m and $\psi = 11.93$m. Right panel: Collision probability as a function of the relative uncertainty (semi-log scale) when $\psi = k \times \text{HBR}$, for $k = 0, 1, 2, 3$.

dilution, in which we have a false assurance of safe conjunction. In the right panel of Figure 2.6, we show that the overall maximum conjunction probability and the dilution region depend on the true miss distance. $\psi = 0$ yields the overall maximum computable collision probability, then the peak per curve will decrease as the ratio ψ gets bigger since the satellite is deemed safer, and the dilution region is shifted to the right.

This behavior of the probability of collision has been discussed in previous works such as Balch et al. (2019) and Hejduk et al. (2019). The “dilution effect", as it is sometimes referred to, results from a shrinkage of the integration region as the relative uncertainty increases. The maximum probability of collision, denoted by $p_{c_{\max}}$, corresponds to the maximum projected area of the covariance within the disk of radius HBR. If the uncertainty is increased or decreased from this point, the probability of collision will decrease.

In practice, the Conjunction Assessment Risk Analysis (CARA) group based at NASA assumes that all conjunctions with a probability of collision greater than a threshold, $\varepsilon \approx 10^{-4}$, are of high risk. Alfano (2005b) proposed a technique that does not addresses the fundamental problem of probability dilution, but assesses the maximum probability for various satellite sizes, encounter geometries, and covariance sizes and shapes. This value is compared with ε to decide if a mitigation is needed. Depending on the size of this maximum probability and the updated states, the remediation action needed might be fairly large or urgent. In Balch (2016), an alternative approach is proposed for addressing the limitations of the probability of collision when the un-

certainty is large. This approach involves using a“random-to-fuzzy transformation", as proposed by Zadeh (1965), to represent the collision risk in a more appropriate way. According to Balch, this transformation is particularly useful for dealing with large uncertainty. The confidence distribution approach, which has been discussed by Schweder and Hjort (2002), Balch (2012), Singh et al. (2007) and Cunen et al. (2020), is another method suggested for addressing the dilution effect. This approach involves expressing the probability of collision in terms of a confidence interval, rather than a point estimate. Both of these proposals aim to provide a more accurate representation of the collision risk in cases where the uncertainty is large.

2.5.2 Why miss distance?

Two further points are immediately clear from the discussion in Section 2.5.1 and the statistical model for satellite conjunction in Section 2.3.

First, the conventional target of inference, the true collision probability (2.8), depends on the unknown relative position and velocity of the two objects, but not on the covariance matrix Ω^{-1} or the data y. The collision probability is generally estimated by $\widehat{p}_c$, which depends on both Ω^{-1} and y. The 'paradoxical' behaviour whereby $\widehat{p}_c$ is very tiny when the data have a very large variance and then increases when that variance decreases is the behaviour of an estimator, not of a parameter of the model. The estimator $\widehat{p}_c$ depends on Ω^{-1}, and as the variance decreases we expect that either $\widehat{p}_c \to 1$, if a collision will occur, or $\widehat{p}_c \to 0$, otherwise; in both cases $\widehat{p}_c \to p(\vartheta)$, as we should expect when the data become noiseless. Thus probability dilution is the natural behaviour of an estimator in response to changes in the variability of the underlying data, not a probability paradox.

Second, the fact that $p(\vartheta)$ takes just two values, whereas the miss distance ψ takes values in a continuum suggests that ψ is a preferable target of inference. For example, the maximum likelihood estimator of $p(\vartheta)$ is $I(\widehat{\psi} \leq \psi_{\min})$, where $I(\cdot)$ and $\widehat{\psi}$ denote the indicator function and the maximum likelihood estimator of ψ, and clearly $\widehat{\psi}$ is more informative. Moreover, in a Bayesian framework, the unknown parameters are regarded as random variables and the posterior probability of collision given the data is obtained by integrating over the posterior density of ψ given y; c.f. (2.19) when ν is known. In this setting, the collision probability is also based on the available knowledge about the miss distance ψ, which is the more fundamental of the two. In Section 2.6.1, we show how inference on ψ provides a significance probability with an interpretation akin to that of $\widehat{p}_c$ and in Section 2.6.2, we discuss Bayesian inference in more detail.

Motivated by these considerations, we turn to inference on ψ based on the observed value of y. If the data suggest that ψ is lower than a safety threshold ψ_0, possibly with $\psi_0 > \psi_{\min}$ for a safety margin, then action to avert a collision should be considered. Our goal below is therefore inference on the unknown miss distance ψ, allowing for the fact that λ is also unknown and must be estimated; in the six-dimensional case we can write the relative distance vector using (2.3) but with $\|\mu\|$ replaced by $\psi/|\sin\beta|$, for $\beta \neq \pi$. In statistical terms the scalar ψ is the primary object of inference, the so-called interest parameter, whereas the 5×1 vector λ of nuisance parameters, while essential for realistic modelling, is of only secondary concern. In the special case with known velocity vector ν and considering only the encounter plane, the miss distance ψ remains the parameter of interest and its interpretation is unchanged, but the nuisance parameter λ is scalar. This simplifies the problem, but the statistical issue remains the same.

2.6 Improved inference for the miss distance

2.6.1 Tangent exponential model

The normal model in (2.9) is a curved exponential family (Davison, 2003, Section 5.2) in terms of ϑ, so we can take $\varphi(\vartheta) = \eta(\vartheta)$, and the computation of $r^*(\psi)$ only involves $\eta(\vartheta)$ and its derivatives. In order to obtain the matrix V when the velocity vector ν is unknown, we define the vector of pivots as $z(y,\vartheta) = \Omega^{1/2}\{y - \eta(\vartheta)\}$; these are independent and standard normal under the model. Partial differentiation yields

$$V = \frac{\partial \eta(\vartheta)}{\partial \vartheta^{\mathrm{T}}} = \eta_\vartheta(\vartheta),$$

evaluated at the maximum likelihood estimate $\widehat{\vartheta}^{\mathrm{o}}$ corresponding to y^{o}. The log-likelihood (2.9) has derivative $\Omega(\eta - y)$ with respect to y, so the the canonical parameter of the tangent exponential model may be written in the form $\varphi(\vartheta) = G\eta(\vartheta) + a$, where $G = -\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta}^{\mathrm{o}})\Omega$ and $a = \eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta}^{\mathrm{o}})\Omega y^{\mathrm{o}}$ are both constant with respect to ϑ and G is full-rank. Any canonical parameter that is an affine transformation of η gives the same expression for $q(\psi)$, because

$$\frac{\left|\varphi(\widehat{\vartheta}) - \varphi\left(\widehat{\vartheta}_\psi\right) \quad \varphi_\lambda\left(\widehat{\vartheta}_\psi\right)\right|}{\left|\varphi_\vartheta(\widehat{\vartheta})\right|} = \frac{\left|G\left\{\eta(\widehat{\vartheta}) - \eta(\widehat{\vartheta}_\psi)\right\} \quad G\eta_\lambda(\widehat{\vartheta}_\psi)\right|}{\left|G\eta_\vartheta(\widehat{\vartheta})\right|} = \frac{\left|\eta(\widehat{\vartheta}) - \eta(\widehat{\vartheta}_\psi) \quad \eta_\lambda(\widehat{\vartheta}_\psi)\right|}{\left|\eta_\vartheta(\widehat{\vartheta})\right|}, \tag{2.20}$$

where $\eta_\lambda(\vartheta) = \partial\eta(\vartheta)/\partial\lambda^{\mathrm{T}}$. Hence we can take the constructed parameter φ to be the state vector, $\varphi(\vartheta) = \eta(\vartheta)$. To compute (2.20), we need the 6×6 Jacobian η_ϑ and the

second derivatives $\eta_{\vartheta\vartheta}$, a $6\times6\times6$ tensor containing the second derivatives of η, needed to compute $j_{\lambda\lambda}(\widehat{\vartheta}_\psi)$. The score equation and the observed Fisher information can respectively be given as

$$\frac{\partial\eta^{\mathrm{T}}(\vartheta)}{\partial\vartheta}\Omega\{y-\eta(\vartheta)\}=0,\quad \frac{\partial\eta^{\mathrm{T}}(\vartheta)}{\partial\vartheta}\Omega\frac{\partial\eta(\vartheta)}{\partial\vartheta_r}-\frac{\partial^2\eta^{\mathrm{T}}(\vartheta)}{\partial\vartheta\partial\vartheta_r}\Omega\{y-\eta(\vartheta)\},\quad r=1,\ldots,6. \tag{2.21}$$

The observed information matrix evaluated at the maximum likelihood estimate for the full model equals $j(\widehat{\vartheta})=\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta})\Omega\eta_\vartheta(\widehat{\vartheta})$, because the score equation for the full model implies that $\eta(\widehat{\vartheta})=y$. A fuller version of a similar computation is given by Fraser et al. (1999b).

In the simplified problem with known ν there are just two parameters, $\psi>0$ and $\lambda\in[0,2\pi)$, the log likelihood reduces to (2.11) and $\varphi(\psi,\lambda)=\xi=(\psi\cos\lambda,\psi\sin\lambda)^{\mathrm{T}}$. But $\widehat{\xi}=(x_1,x_2)^{\mathrm{T}}$, so

$$\widehat{\psi}=(x_1^2+x_2^2)^{1/2},\quad \widehat{\lambda}=\arctan(x_2/x_1). \tag{2.22}$$

If ψ is fixed, then $\widehat{\lambda}_\psi$ is readily found as the unique minimum of the sum of squares in (2.11); a good starting value should be $\widehat{\lambda}$. Then the likelihood root reduces to

$$r(\psi)=\operatorname{sign}(\widehat{\psi}-\psi)\frac{1}{d_1d_2}\left\{d_2^2\left(x_1-\psi\cos\widehat{\lambda}_\psi\right)^2+d_1^2\left(x_2-\psi\sin\widehat{\lambda}_\psi\right)^2\right\}^{1/2}.$$

To obtain q, we use (2.10) to simplify the ratio

$$\frac{|j(\widehat{\vartheta})|^{1/2}}{|\varphi_\vartheta(\widehat{\vartheta})|}=\frac{\left|\eta_\vartheta(\widehat{\vartheta})\Omega\eta_\vartheta(\widehat{\vartheta})\right|^{1/2}}{|\eta_\vartheta(\widehat{\vartheta})|}=|\Omega|^{1/2}=\frac{1}{d_1d_2}.$$

The determinant involving the canonical parameter and its derivatives is

$$\left|\varphi(\widehat{\theta})-\varphi(\widehat{\theta}_\psi)\ \varphi(\widehat{\theta}_\lambda)\right|=\begin{vmatrix}x_1-\psi\cos\widehat{\lambda}_\psi & -\psi\sin\widehat{\lambda}_\psi\\ x_2-\psi\sin\widehat{\lambda}_\psi & \psi\cos\widehat{\lambda}_\psi\end{vmatrix}=\psi\left(x_1\cos\widehat{\lambda}_\psi+x_2\sin\widehat{\lambda}_\psi\right)-\psi^2.$$

The information component associated with the nuisance parameter is

$$j_{\lambda\lambda}=\frac{1}{d_1^2d_2^2}\left\{d_2^2\psi\sin\lambda\left(x_1-\psi\cos\lambda\right)+d_1^2\psi\cos\lambda\left(x_2-\psi\sin\lambda\right)+\psi^2\left(d_2^2\sin^2\lambda+d_1^2\cos^2\right)\right\},$$

so

$$q(\psi) = \frac{\left|\varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi)\,\varphi(\widehat{\theta}_\lambda)\right|}{|j_{\lambda\lambda}(\widehat{\vartheta})|}\,|\Omega|^{1/2},$$
$$= \psi^{1/2}\frac{x_1\cos\widehat{\lambda}_\psi + x_2\sin\widehat{\lambda}_\psi - \psi}{\left\{d_2^2(x_1\cos\widehat{\lambda}_\psi - \psi\cos 2\widehat{\lambda}_\psi) + d_1^2(x_2\sin\widehat{\lambda}_\psi + \psi\cos 2\widehat{\lambda}_\psi)\right\}^{1/2}}.$$

When $d_1 = d_2 = d$, say, then $\widehat{\lambda}_\psi \equiv \widehat{\lambda}$ does not depend on ψ and after simplification we obtain

$$\begin{aligned} r(\psi) &= w(\psi) = (\widehat{\psi} - \psi)/d, \\ q(\psi) &= r(\psi)(\psi/\widehat{\psi})^{1/2}, \\ r^*(\psi) &= \frac{\widehat{\psi} - \psi}{d} + \frac{d}{2(\widehat{\psi} - \psi)}\log(\psi/\widehat{\psi}), \quad \psi > 0. \end{aligned} \tag{2.23}$$

Note that $r^*(\psi) \to -d/(2\widehat{\psi})$ when $\psi \to \widehat{\psi}$. The fact that $r^*(\psi) < r(\psi)$ for all ψ in this setting implies that confidence intervals for ψ based on $r^*(\psi)$ will be closer to the origin than those based on $r(\psi)$, and significance levels for fixed ψ will be higher, leading to more conservative inferences; this is illustrated in Figure 2.7. We should consider evasive action based on $r(\psi)$ when $\Phi\{-r^{\mathrm{o}}(\psi_0)\} > \varepsilon$, i.e., when $\widehat{\psi}^{\mathrm{o}} - dz_\varepsilon$ is smaller than the hard-body radius ψ_0; if $\varepsilon = 10^{-4}$ for example, then $z_\varepsilon = -3.72$. Notice that this rule relates the observed distance of the second object from the origin, $\widehat{\psi}^{\mathrm{o}}$, to the measurement uncertainty, d. Use of $r^{*\mathrm{o}}(\psi)$ will lead to very similar conclusions in most cases, though it is not monotone in ψ when $\psi \to 0$.

2.6.2 Bayesian approximation

As discussed in Section 1.4.3, the Bayesian analog of q is

$$q_B = \ell'_{\mathrm{p}}(\psi)\, j_{\mathrm{p}}(\widehat{\psi})^{-1/2}\left\{\frac{\left|j_{\lambda\lambda}\left(\widehat{\vartheta}_\psi\right)\right|}{\left|j_{\lambda\lambda}(\widehat{\vartheta})\right|}\right\}^{1/2}\frac{\pi(\widehat{\vartheta})}{\pi\left(\widehat{\vartheta}_\psi\right)}. \tag{2.24}$$

For $\widehat{\vartheta}^{\mathrm{T}} = (\psi, \widehat{\lambda}_\psi^{\mathrm{T}})$, we write

$$\frac{\mathrm{d}\ell(\widehat{\vartheta}_\psi)}{\mathrm{d}\psi} = \ell_\psi(\widehat{\vartheta}_\psi) + \frac{\partial\widehat{\lambda}_\psi^{\mathrm{T}}}{\partial\psi}\ell_\lambda(\widehat{\vartheta}_\psi) = \ell_\psi(\widehat{\vartheta}_\psi) - \ell_{\psi\lambda}(\widehat{\vartheta}_\psi)\ell_{\lambda\lambda}^{-1}(\widehat{\vartheta}_\psi)\ell_\lambda(\widehat{\vartheta}_\psi),$$

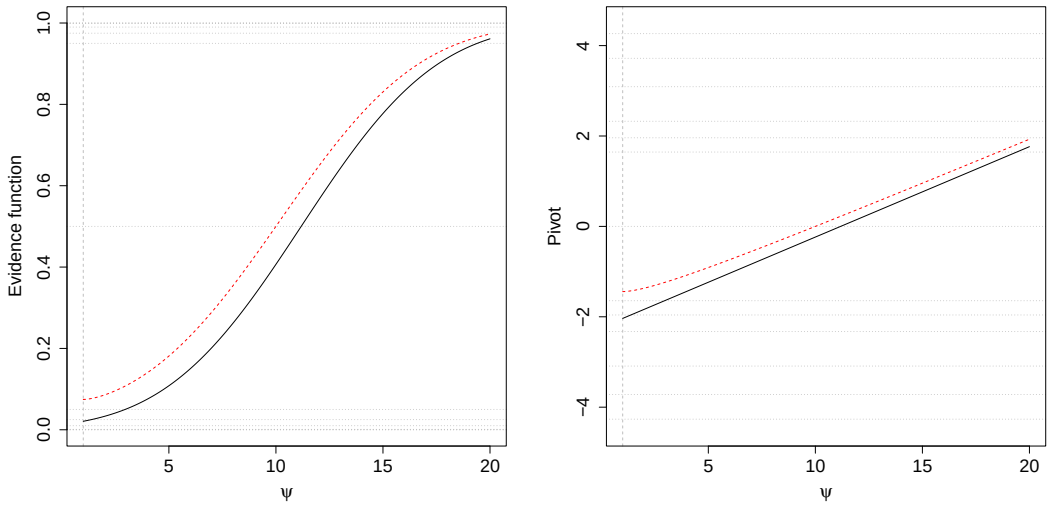


Figure 2.7 – Illustration of the two-dimensional setting described in (2.23) for $x_1 = 10$, $x_2 = 5$, and $d = 5$. Left panel: Evidence function based on likelihood root $r^{\mathrm{o}}(\psi)$ (solid black), and modified likelihood root $r^{*\mathrm{o}}(\psi)$ (red dashes). The horizontal lines correspond to probabilities $10^{-6}, 10^{-5}$, 10^{-4}, 10^{-3}, 10^{-2}, 0.025, 0.05, 0.5, 0.95, 0.975, 0.999, 0.9999, 0.99999 and 0.999999. Right panel: the same quantities transformed to the standard normal scale.

where the second equality follows from noting that differentiating the equation $\ell_\lambda(\widehat{\vartheta}_\psi) = 0$ that defines $\widehat{\lambda}_\psi$ yields

$$\ell_{\psi\lambda}(\widehat{\vartheta}_\psi) + \frac{\partial \widehat{\lambda}_\psi^{\mathrm{T}}}{\partial \psi}\ell_{\lambda\lambda}(\widehat{\vartheta}_\psi) = 0,$$

with the matrix $\ell_{\lambda\lambda}(\widehat{\vartheta}_\psi)$ invertible because it is the Hessian corresponding to the maximum of ℓ in the λ direction for fixed ψ. A standard identity for the determinant of a partitioned matrix gives

$$\left|\ell_\vartheta(\widehat{\vartheta}_\psi) \quad \ell_{\vartheta\lambda}(\widehat{\vartheta}_\psi)\right| = \begin{vmatrix} \ell_\psi(\widehat{\vartheta}_\psi) & \ell_{\psi\lambda}(\widehat{\vartheta}_\psi) \\ \ell_\lambda(\widehat{\vartheta}_\psi) & \ell_{\lambda\lambda}(\widehat{\vartheta}_\psi) \end{vmatrix} = \{\ell_\psi(\widehat{\vartheta}_\psi) - \ell_{\psi\lambda}(\widehat{\vartheta}_\psi)\ell_{\lambda\lambda}^{-1}(\widehat{\vartheta}_\psi)\ell_\lambda(\widehat{\vartheta}_\psi)\} \times \left|\ell_{\lambda\lambda}(\widehat{\vartheta}_\psi)\right|, \tag{2.25}$$

and the expression for the observed information in (2.21) implies that we can write

$$-\ell_{\vartheta\lambda_r}(\vartheta) = \eta_\vartheta^{\mathrm{T}}(\vartheta)\Omega\eta_{\lambda_r}(\vartheta) + A_r(\vartheta)\{y - \eta(\vartheta)\}, \quad r = 1, \ldots, d-1, \tag{2.26}$$

where $A(\vartheta)$ involves second derivatives of η. Now $y = \eta(\widehat{\vartheta})$, so if $\widehat{\psi} - \psi$ is of order $n^{-1/2}$, then $y - \eta(\widehat{\vartheta}_\psi) = \eta(\widehat{\vartheta}) - \eta(\widehat{\vartheta}_\psi)$ is also $O(n^{-1/2})$, and so too is the second term on the

right of (2.26). Thus (2.25) equals

$$\begin{aligned}
\left|\ell_\vartheta(\widehat{\vartheta}_\psi) \quad \ell_{\vartheta\lambda}(\widehat{\vartheta}_\psi)\right| &= (-1)^{d-1}\left|\ell_\vartheta(\widehat{\vartheta}_\psi) \quad -\ell_{\vartheta\lambda}(\widehat{\vartheta}_\psi)\right| \\
&= (-1)^{d-1}\left|\eta_\vartheta(\widehat{\vartheta}_\psi)\Omega\{y-\eta(\widehat{\vartheta}_\psi)\} \quad \eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta}_\psi)\Omega\eta_\lambda(\widehat{\vartheta}_\psi)\right| + O(n^{-1/2}) \\
&= (-1)^{d-1}\left|\eta_\vartheta(\widehat{\vartheta}_\psi)\right| \times |\Omega| \times \left|\eta(\widehat{\vartheta})-\eta(\widehat{\vartheta}_\psi) \quad \eta_\lambda(\widehat{\vartheta}_\psi)\right| + O(n^{-1/2}).
\end{aligned}$$

As ϑ is of dimension d and $|J(\widehat{\vartheta})| = \left|\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta})\right|^2 |\Omega|$, equation (2.24) equals

$$\begin{aligned}
q_B(\psi) &= \frac{\left|\ell_\vartheta(\widehat{\vartheta}_\psi) \quad \ell_{\vartheta\lambda}(\widehat{\vartheta}_\psi)\right|}{|\ell_{\lambda\lambda}(\widehat{\vartheta}_\psi)|} \times \frac{|J_{\lambda\lambda}(\widehat{\vartheta}_\psi)|^{1/2}}{|J(\widehat{\vartheta})|^{1/2}} \frac{\pi(\widehat{\vartheta})}{\pi(\widehat{\vartheta}_\psi)} \\
&= \frac{\left|\ell_\vartheta(\widehat{\vartheta}_\psi) \quad -\ell_{\vartheta\lambda}(\widehat{\vartheta}_\psi)\right|}{|J_{\lambda\lambda}(\widehat{\vartheta}_\psi)|^{1/2}|J(\widehat{\vartheta})|^{1/2}} \times \frac{\pi(\widehat{\vartheta})}{\pi(\widehat{\vartheta}_\psi)} \\
&= \frac{\left|\eta_\vartheta(\widehat{\vartheta}_\psi)\right||\Omega|\left|\eta(\widehat{\vartheta})-\eta(\widehat{\vartheta}_\psi) \quad \eta_\lambda(\widehat{\vartheta}_\psi)\right|}{|J_{\lambda\lambda}(\widehat{\vartheta}_\psi)|^{1/2}\left|\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta})\right||\Omega|^{1/2}} \times \frac{\pi(\widehat{\vartheta})}{\pi(\widehat{\vartheta}_\psi)} + O(n^{-1/2}) \\
&= q(\psi) \times \frac{\left|\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta}_\psi)\right|}{\left|\eta_\vartheta^{\mathrm{T}}(\widehat{\vartheta})\right|} \times \frac{\pi(\widehat{\vartheta})}{\pi(\widehat{\vartheta}_\psi)} + O(n^{-1/2}), \qquad (2.27)
\end{aligned}$$

where $q(\psi)$ is given in equation (1.27).

The Jeffreys prior is the root of the determinant of the Fisher information matrix,

$$\pi(\vartheta) \propto \left|\eta_\vartheta^{\mathrm{T}}(\vartheta)\Omega\eta_\vartheta(\vartheta)\right|^{1/2} \propto \left|\eta_\vartheta(\vartheta)\right|,$$

as the constant $|\Omega|$ can be ignored. If this prior is used then (2.27) simplifies to equation (2.20), plus a term of $O(n^{-1/2})$. Hence

$$r_B(\psi) = -r^*(\psi) + O(n^{-1}),$$

and inferences from both pivots will be the same to this order of error. Hence Bayesian and frequentist confidence intervals for ψ_0 differ by only $O\left(n^{-1}\right)$.

The Jeffreys prior gives inferences invariant to $1-1$ transformation of ϑ and thus is often regarded as a natural choice, though it is criticized by Fraser et al. (2016a). In the conjunction problem, this prior has the undesirable property mentioned at the end of Section 2.5.1 of attributing zero probability to any sphere around the origin, relative to outside that region. In a similar context Davison and Sartori (2008) compare the performances of $r_B^{*o}(\psi)$ and $r^{*o}(\psi)$ and find that the former performs rather worse. This same behavior is also observed in the present setting, owing to the downward bias

of $\widehat{p}_c$; see case study C in Section 2.7. The Bayesian confidence intervals are slightly shorter and tend to contain the true parameter less often. Bayesian approximations are not our primary focus in this work; rather, we find it reassuring that Bayesian and frequentist inferences can be approached in the same way. If reliable prior information was available, then a Bayesian approach would be appropriate.

2.6.3 Evidence and decisions

In Section 2.4 we argued that functions such as (2.15) allow inference on the true miss distance ψ, either by constructing confidence intervals or as an assessment of the evidence that ψ equals some particular value ψ_0. In the present context ψ_0 might be a safety threshold, and then one approach to inference on ψ is to test the null hypothesis $H_0 : \psi = \psi_0$ against the alternative hypothesis $H_+ : \psi > \psi_0$, with evasive action to be considered if (c.f. equation (2.15))

$$p_{\text{obs}} = p^{\text{o}}(\psi_0) = \Phi\left\{-r^{*\text{o}}(\psi_0)\right\} > \varepsilon,$$

i.e., H_0 cannot be rejected at level ε. If the significance probability is correctly calibrated and the true miss distance is ψ_0, then the false positive probability, that of considering action unnecessarily, would be $1-\varepsilon$, whatever ε is chosen. In practice ε is often taken to be 10^{-4}. The choice of H_+ as alternative hypothesis ensures that p_{obs} is small when the estimated collision probability $\widehat{p}_c$ and the Bayesian posterior probability $\Pr(\psi \le \psi_0 \mid y)$ would also be small, despite their different interpretations and properties. Hejduk et al. (2019) argue that the null hypothesis $\psi \le \psi_0$ is unnatural, since it implies that the 'null' situation is to anticipate a collision, but it appears more important to us to ensure that small values of p_{obs} correspond to small collision probabilities. Our approach is supported by regarding hypothesis testing as attempting 'proof by stochastic contradiction': the null hypothesis represents an assumption that is regarded as absurd (disproved) when the corresponding significance level is sufficiently small. If so, it makes sense to take $\psi = \psi_0$ as the null hypothesis, as we hope that this will be contradicted by the data and no evasive action will need be considered.

The interpretation of the threshold ε in terms of an elementary decision analysis was described in Section 2.4, and since conjunction analysis is intended to assist decision-making, consideration of losses seems an appropriate basis for choosing ε. Although our earlier discussion suggested considering evasive action when the posterior probability $\Pr(\psi \le \psi_0 \mid y)$ exceeds ε, it seems better to replace it by a significance probability p_{obs} computed as $\Phi\left\{-r^{\text{o}}(\psi_0)\right\}$ or $\Phi\left\{-r^{*\text{o}}(\psi_0)\right\}$, which are approximately uniformly distributed under H_0.

The abuse of hypothesis tests has been much discussed (e.g., Carpenter et al., 2017), and it is often suggested that they be systematically replaced by confidence intervals. Plotting $\Phi\{-r^{*\mathrm{o}}(\psi)\}$ as a function of ψ, as in Figure 2.9, allows the construction of confidence intervals for ψ, the form of which expresses the uncertainty and suggests what power is available: a narrow interval corresponds to more precise estimation and hence higher power for rejecting hypotheses such as H_0 above. Unlike a two-sided $(1-2\alpha)$ confidence interval $[L_\alpha, U_\alpha]$, one-sided intervals such as $[L_\alpha, +\infty)$ or $[0, U_\alpha)$ give no information about the accuracy of the estimate, so a two-sided interval provides a better basis for risk assessment. A one-sided confidence interval (L_ε, ∞) that does not contain ψ_0 leads to the same decision as observing $p_{\mathrm{obs}} < \varepsilon$, and it seems partly a matter of taste which is preferred: computing p_{obs} alone is quicker but is less informative than a plot of $\Phi\{-r^{*\mathrm{o}}(\psi)\}$. In navigating the literature, it is enlightening to recognize that hypothesis testing has a variety of distinct uses (Cox, 2020), one of which is to flag situations that merit more detailed scrutiny. In conjunction analysis, one might therefore plot evidence functions only when $p_{\mathrm{obs}} > \varepsilon$, thereby focusing on those conjunctions requiring careful consideration.

2.7 Numerical results

2.7.1 General setup

Below we investigate the accuracy of the normal approximations to the Wald statistic, the likelihood root and the modified likelihood root in four case studies. We do so in terms of one-sided error rates for confidence intervals (L_α, U_α) for the true miss distance ψ_0 and use $\Pr_0$ to indicate probability computed when $\psi = \psi_0$. An ideal two-sided equi-tailed confidence interval with coverage probability $1-2\alpha$ should satisfy $\Pr_0(L_\alpha \le \psi_0 \le U_\alpha) = 1-2\alpha$ and have one-sided left-tail and right-tail error rates $\Pr_0(\psi_0 < L_\alpha)$ and $\Pr_0(U_\alpha < \psi_0)$ both equal to α, for any $\alpha \in (0, 0.5)$. Departures from this will indicate deficiencies of the confidence intervals and the corresponding tests, whereas close agreement will indicate that the inferences are well-calibrated. As $w(\psi)$, $r(\psi)$ and $r^*(\psi)$ are decreasing in ψ and should ideally have standard normal distributions, inaccurate left-tail error rates correspond to departures from normality in the upper tail $w(\psi)$, $r(\psi)$ and $r^*(\psi)$. In the present setting, accurate left-tail error rates are most important, since they correspond to well-calibrated significance probabilities and confidence intervals of form (L_α, ∞).

As mentioned in Section 2.2, the form of the covariance matrix depends on the type of the conjunction (Chen et al., 2017, Chapter 5). In short-term conjunctions, uncertainty on the velocity is negligible compared to uncertainty on the position. In long-term

conjunctions, the motion is nonlinear and the computations are more involved. In both cases, the quality of risk assessment depends heavily on the covariance matrix, which is usually intentionally inflated to improve the fidelity of the error modeling. Below we suppose that the error covariance matrix for the relative distance and velocity of the second satellite relative to the first is given by

$$\Omega^{-1} = \begin{bmatrix} P_1 & P_{12} \\ P_{12} & P_2 \end{bmatrix},$$

where P_1, P_2, and P_{12} are the position, the velocity and the cross-correlation covariance matrices, of units km^2, km^2 s^{-2} and km^2 s^{-1} respectively. The six eigenvalues of Ω^{-1} are difficult to interpret physically, and can vary greatly.

In our first two case studies, we assume that $P_{12} = 0_{3\times 3}$, and choose $P_1 = \tau\sigma^2 I_3$ and $P_2 = \sigma^2 I_3$. This choice implies that the standard deviation of position errors along each axis direction is $\sqrt{\tau}\sigma$ (km) and the standard deviation of velocity errors is σ (km/s). Uncertainty on the position is typically larger than that on the velocity, and then $\tau > 1$. In the last two case studies, we consider quantities projected into the encounter plane, so the covariance matrix is a two-dimensional diagonal matrix with standard deviation of position errors in km.

2.7.2 Case study A: Simulated data

The relative quantities and spherical coordinates of two satellites in this case study are given in column A of Table 2.2. The relative distance and speed are around 102 km and around 11.7 km/s, the value of σ^2 varies from 10^{-3} km^2 to 2 km^2, and that of τ varies from 1 to 3.

Table 2.4 shows the error rates for the Wald statistic, the likelihood root and the modified likelihood root based on 10^4 datasets simulated for various combinations of values of σ and τ. For very small σ^2, i.e., high-precision measurement of position and velocity, all three sets of error rates are close to the nominal values. However, problems with the Wald statistic and, to a lesser extent, the likelihood root start to appear when $\sigma^2 \geq 0.1$, with the left-tail error systematically too high and the right-tail error systematically too low. The modified root behaves much better overall, though its right-tail error also rises as σ^2 increases. As mentioned in Section 2.7.1, tests of H_+ require accuracy in the left tail, so right-tail error is less important.

These remarks are confirmed by the Gaussian QQ-plots of simulated values of the three quantities in Figure 2.8. If the distribution is exactly Gaussian, then the confidence

Table 2.2 – Conjunction geometry of case studies: A, a simulated example; B, the U.S. and Russian satellite collision event; C, an event with high probability of conjunction; and D an event with minimum miss distance.

Variable	Case study			
	A	B	C	D
Miss distance (m)	$10^3 \times 35.267$	698.011	11.917	3.345
ΔX (m)	$-10^3 \times 100$	−258.909	−7.678	2.875
ΔY (m)	$-10^3 \times 20$	−635.813	−9.152	−2.382
ΔZ (m)	0	126.229	0.564	−1.074
ΔV_x (km/s)	10	10.580	9.926	−1.099
ΔV_y (km/s)	6	−3.733	−9.653	−11.840
ΔV_z (km/s)	1	3.126	−4.110	1.313
θ_1	1.570	1.389	1.618	1.850
θ_2	1.485	1.299	1.860	-0.691
ϕ_1	−2.944	1.957	−2.269	1.460
ϕ_2	−2.944	−0.339	− 0.772	-1.663

Table 2.3 – Position and velocity coordinates of the primary (O_1) and the secondary (O_2) objects for case studies B, C and D.

	B		C		D	
	O_1	O_2	O_1	O_2	O_1	O_2
X (km)	−1457.273	−1457.532	−1818.269	−1818.277	1935.852	1935.849
Y (km)	1589.568	1588.932	1040.564	1040.555	562.737	562.740
Z (km)	6814.189	6814.316	−6772.707	−6772.708	6779.432	6779.433
V_x (km/s)	−7.001	3.578	-3.610	6.317	−4.907	−3.808
V_y (km/s)	−2.439	−6.172	6.269	−3.384	−5.371	6.469
V_z (km/s)	−0.9262	2.200	1.933	−2.177	1.843	0.530

intervals are exactly calibrated, so a departure from the line of unit slope through the origin implies a lack of calibration. For small σ^2, all three statistics have standard normal distributions and give comparable results, but for larger σ^2, the Wald statistic and the likelihood root are shifted to the right and right-skewed, more strikingly for larger values of τ. This reflects the upward bias of the estimated distance, discussed in Section 2.5.1, and explains the asymmetric error rates in Table 2.4, with lower probabilities for the right than for the left. The asymmetry increases with larger uncertainties on the relative distance and velocity and with smaller nominal error rates.

Figure 2.8 shows that the modified likelihood root r^* corrects the departure from normality in the upper tail even for $\sigma^2 = 5$, and its error rates are closer to the nominal

rates in all cases considered. For $\sigma^2 > 2$ and for 1% nominal levels, the Wald statistic and the likelihood root show extreme overcoverage on the right and undercoverage on the left; although the modified likelihood root provides a considerable improvement, its right-tail error is somewhat smaller than the nominal value.

2.7.3 Case study B: US and Russian collision event

Our second example is the 10 February 2009 collision of the U.S. operational communications satellite Iridium 33 and the decommissioned Russian communications satellite Cosmos 2251, whose relative configuration is given in Table 2.2 and absolute coordinates in Earth-centered inertial (ECI) coordinates are given in Table 2.3. Figure 2.9 shows the evidence functions for this conjunction for $(\sigma,\tau) = (10^{-1},5)$, which suggest that evasive action would be essential if the safety threshold is 20m, i.e., $\psi_0 = 0.02$km (the dashed vertical line). On the other hand, the probability of collision, 1.14687×10^{-5}, is lower than the conventional $\varepsilon = 10^{-4}$, implying that such action is unnecessary. The Wald statistic and the likelihood root are indistinguishable, but the modified likelihood root is shifted slightly to the left, increasing the evidence that the true miss distance is below ψ_0. The significance probabilities for testing $\psi = \psi_0 = 20$m against $H_+ : \psi > \psi_0$ are 1.2×10^{-3} for the Wald statistic and the likelihood root, and 7.2×10^{-3} for the modified likelihood root, so the same conclusions would be drawn using any of these quantities if $\varepsilon = 10^{-4}$. Clearly, however, the conclusions might disagree in other circumstances. The posterior probability $\Pr(\psi \leq \psi_0 \mid y^{\mathrm{o}})$ can be expected to be similar to $\Phi\{-r^{*\mathrm{o}}(\psi_0)\}$.

Table 2.5 gives left and right error rates for different values of σ^2 and τ for this conjunction. Its first row corresponds to the case where the standard deviation of the position error along each axis is 10m and the standard deviation of the velocity errors is 10m/s, giving $(\sigma,\tau) = (10^{-2},1)$. In the lower rows we first increase uncertainty on the position while keeping that on the velocity fixed by increasing τ, and then increase both velocity and position errors by increasing σ^2. For these simulations the true miss distance and relative velocity are $\psi = 698$m and $\|v\| = 11.648 \times 10^3$m/s.

The error rates for the Wald statistic and the likelihood root are almost identical, implying that the corresponding pivots are indistinguishable. For small velocity errors, with $\sigma^2 < 10^{-3}$, there is no significant difference in the error rates for the three statistics. For larger velocity variance, the overall error rates found by summing the left and the right error rates equal the nominal values, but left-tail error dominates the sum. In these cases the modified likelihood root is more symmetric and shows fewer extreme values, especially for large τ, so interval estimates based on r^* are more reliable.

Table 2.4 – Empirical left- and right-tail error rates (%) at nominal levels 10%, 5%, and 1% for case study A, estimated from 10^4 Monte Carlo samples. The standard errors (SE) appear in the last line.

		Left tail (%)			Right tail (%)		
Uncertainty	Statistic	5	2.5	0.5	5	2.5	0.5
$(\sigma^2,\tau)=(10^{-3},1)$	w	5.22	2.58	0.54	5.03	2.49	0.49
	r	5.20	2.56	0.53	5.04	2.51	0.50
	r^*	5.15	2.54	0.53	5.13	2.54	0.51
$(\sigma^2,\tau)=(10^{-3},2)$	w	5.43	2.61	0.43	4.96	2.61	0.59
	r	5.43	2.61	0.43	4.98	2.61	0.61
	r^*	5.40	2.58	0.43	5.02	2.62	0.62
$(\sigma^2,\tau)=(10^{-3},3)$	w	4.93	2.48	0.53	4.83	2.46	0.54
	r	4.91	2.47	0.52	4.88	2.46	0.55
	r^*	4.85	2.40	0.50	4.90	2.49	0.55
$(\sigma^2,\tau)=(10^{-1},1)$	w	5.45	2.88	0.64	4.15	2.09	0.37
	r	5.26	2.78	0.59	4.27	2.15	0.37
	r^*	4.89	2.53	0.54	4.65	2.37	0.42
$(\sigma^2,\tau)=(10^{-1},2)$	w	5.62	2.98	0.67	4.60	2.28	0.50
	r	5.48	2.75	0.61	4.70	2.32	0.54
	r^*	5.06	2.48	0.54	5.06	2.56	0.64
$(\sigma^2,\tau)=(10^{-1},3)$	w	5.66	2.86	0.65	4.49	2.26	0.49
	r	5.54	2.70	0.63	4.58	2.35	0.50
	r^*	5.06	2.50	0.59	4.92	2.57	0.53
$(\sigma^2,\tau)=(1,1)$	w	7.21	4.10	1.04	3.18	1.37	0.16
	r	6.45	3.36	0.68	3.35	1.53	0.21
	r^*	5.39	2.64	0.55	4.56	2.37	0.31
$(\sigma^2,\tau)=(1,2)$	w	6.58	3.72	0.92	3.64	1.50	0.24
	r	5.92	3.16	0.57	3.80	1.62	0.29
	r^*	4.81	2.40	0.41	5.32	2.59	0.48
$(\sigma^2,\tau)=(1,3)$	w	6.40	3.50	0.96	3.73	1.64	0.25
	r	5.74	3.03	0.64	3.94	1.76	0.25
	r^*	4.74	2.49	0.54	5.19	2.72	0.46
$(\sigma^2,\tau)=(2,1)$	w	7.78	4.57	1.41	2.12	0.69	0.03
	r	6.52	3.56	0.78	2.27	0.75	0.03
	r^*	5.19	2.64	0.58	4.46	1.85	0.10
$(\sigma^2,\tau)=(2,2)$	w	7.94	4.59	1.33	2.30	0.71	0.05
	r	6.77	3.49	0.71	2.43	0.88	0.05
	r^*	5.18	2.60	0.54	4.62	2.09	0.18
$(\sigma^2,\tau)=(2,3)$	w	7.67	4.07	1.15	2.19	0.73	0
	r	6.49	3.24	0.65	2.39	0.86	0
	r^*	4.71	2.34	0.49	4.48	1.98	0.11
SE		0.22	0.16	0.07	0.22	0.16	0.07

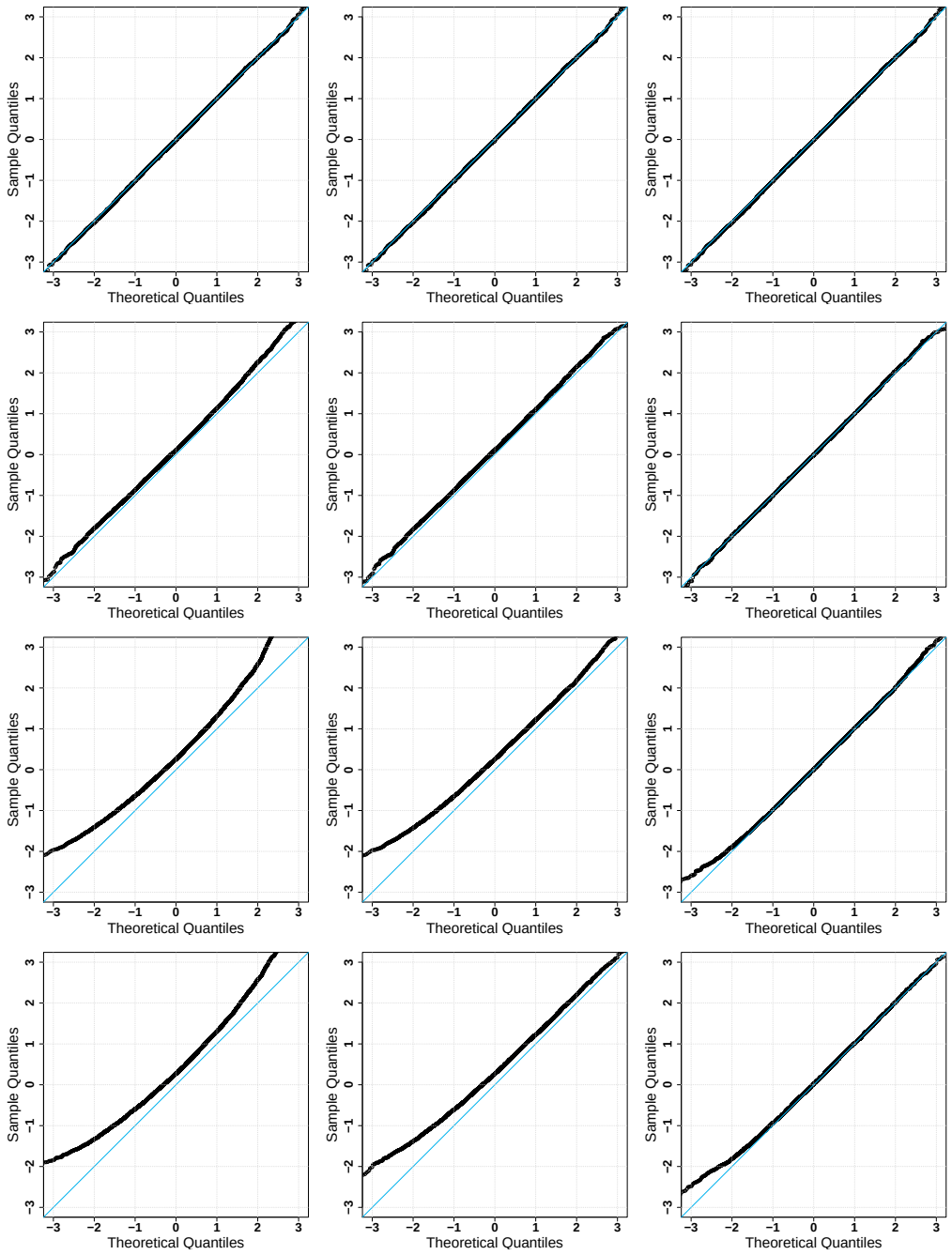


Figure 2.8 – Case study A: normal QQ-plots of $w(\psi^o)$ (left), $r(\psi^o)$ (middle) and $r^*(\psi^o)$ (right) based on 10^4 Monte Carlo sample quantiles, with (σ^2, τ) equal to $(10^{-3}, 1), (1,1), (5,2), (5,5)$ (top to bottom).

Table 2.5 – Empirical left- and right-tail error rates (%) at nominal levels 10%, 5%, and 1% confidence intervals for the parameter ψ for case study B, based on 10^4 Monte Carlo replications. The standard errors (SE) appear in the last line.

		Left tail (%)			Right tail (%)		
Uncertainty	Statistic	5	2.5	0.5	5	2.5	0.5
$(\sigma^2,\tau)=(10^{-4},1)$	w	5.06	2.54	0.42	5.00	2.53	0.64
	r	5.06	2.54	0.42	5.00	2.53	0.64
	r^*	4.98	2.46	0.41	5.05	2.62	0.65
$(\sigma^2,\tau)=(10^{-4},2)$	w	5.27	2.73	0.60	4.66	2.29	0.52
	r	5.27	2.73	0.60	4.66	2.290	0.52
	r^*	5.17	2.67	0.56	4.78	2.33	0.53
$(\sigma^2,\tau)=(10^{-4},4)$	w	4.67	2.29	0.49	4.53	2.34	0.43
	r	4.67	2.29	0.49	4.53	2.34	0.43
	r^*	4.58	2.21	0.47	4.67	2.40	0.47
$(\sigma^2,\tau)=(10^{-4},10^2)$	w	5.89	2.98	0.74	4.33	2.05	0.47
	r	5.89	2.98	0.74	4.33	2.05	0.47
	r^*	5.18	2.73	0.65	5.14	2.59	0.61
$(\sigma^2,\tau)=(10^{-3},1)$	w	5.17	2.66	0.50	4.60	2.16	0.40
	r	5.17	2.66	0.50	4.60	2.16	0.40
	r^*	4.96	2.55	0.45	4.88	2.37	0.44
$(\sigma^2,\tau)=(10^{-3},2)$	w	5.00	2.50	0.48	4.93	2.35	0.37
	r	5.00	2.50	0.48	4.93	2.35	0.37
	r^*	4.76	2.34	0.45	5.31	2.54	0.42
$(\sigma^2,\tau)=(10^{-3},4)$	w	5.25	2.73	0.61	4.54	2.23	0.38
	r	5.25	2.73	0.61	4.54	2.23	0.38
	r^*	4.74	2.52	0.52	5.00	2.45	0.43
$(\sigma^2,\tau)=(10^{-2},1)$	w	5.75	2.96	0.54	4.53	2.06	0.31
	r	5.75	2.96	0.54	4.54	2.07	0.32
	r^*	5.09	2.56	0.44	5.33	2.70	0.43
$(\sigma^2,\tau)=(10^{-2},2)$	w	5.96	2.94	0.51	3.69	1.92	0.30
	r	5.96	2.94	0.51	3.69	1.92	0.30
	$r*$	4.89	2.39	0.38	4.87	2.53	0.49
$(\sigma^2,\tau)=(10^{-2},4)$	w	6.20	3.26	0.57	2.73	1.14	0
	r	6.20	3.26	0.57	2.73	1.14	0
	$r*$	4.78	2.45	0.41	4.96	2.30	0.31
SE		0.22	0.16	0.07	0.22	0.16	0.07

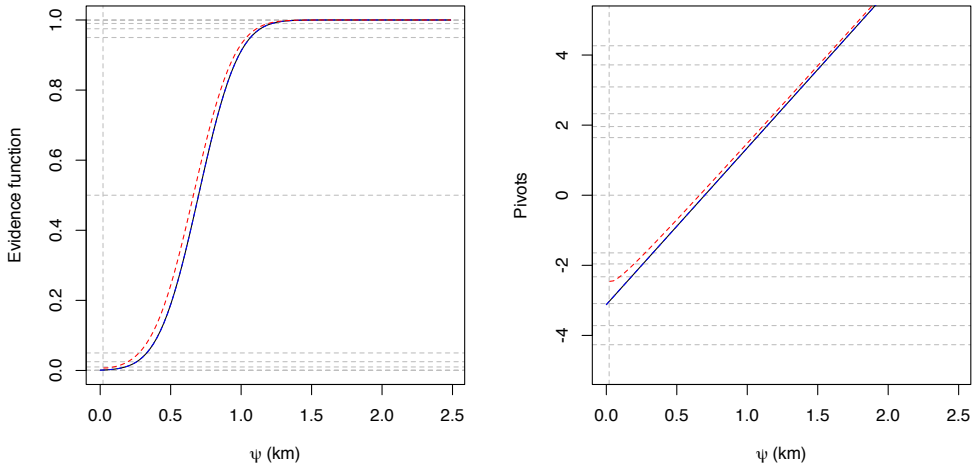


Figure 2.9 – Summaries for case study B. Left: evidence functions based on likelihood root $r^{\mathrm{o}}(\psi)$ (solid black), Wald statistic $w(\psi)$ (dotted blue), and modified likelihood root $r^{*\mathrm{o}}(\psi)$ (red dashes). Right: the same quantities transformed to the standard normal scale. The likelihood root and Wald statistic are almost indistinguishable.

2.7.4 Case study C: High p_c event

Our third case study is based on a NASA test case[1] as part of their publicly-released Conjunction Assessment Risk Analysis (CARA) tool. Each published test case contains data that can be readily converted to the ECI reference frame for both objects. The test cases also include state covariance data, originally expressed in the UVW reference frame. The UVW covariance matrices can be transformed to the 6×6 ECI state covariance matrices Ω_p^{-1} and Ω_s^{-1} for the two objects. These transformed quantities are then used to define the relative ECI state at close approach, η, and the associated combined covariance, $\Omega^{-1} = \Omega_p^{-1} + \Omega_s^{-1}$.

Column C of Table 2.2 and Table 2.3 show the conjunction elements for this example. Since the motion is linear and uncertainty on the velocity can be ignored, we use the reduced two-dimensional model representing conjunction in the encounter plane. The projected quantities in this plane, as described in Section 2.3.2 are given by

$$A = (CV, v/\|v\|) = \begin{bmatrix} 0.72 & 0.11 & -0.69 \\ 0.68 & -0.30 & 0.67 \\ 0.13 & 0.95 & 0.28 \end{bmatrix},$$

[1] https://github.com/nasa/CARA_Analysis_Tools/tree/master/two-dimension_Pc/UnitTest/InputFiles

$$x^{\mathrm{o}} = A^{\mathrm{T}} y^{\mathrm{o}} = \begin{bmatrix} 11.84 \\ -1.36 \end{bmatrix}, \quad D = \mathrm{diag}\left(d_1^2, d_2^2\right) = \begin{bmatrix} 25.1^2 & 0.00 \\ 0.00 & 11.61^2 \end{bmatrix},$$

so $\widehat{\psi}^o = \|x^{\mathrm{o}}\| = 11.92$m. The evidence functions plotted in Figure 2.10 show that with $\psi_0 = 10$ m the significance probabilities p_{obs} are all of order 0.5, though that for the modified likelihood root is closer to 0.6; clearly evasive action would be essential in this case.

Table 2.6 shows empirical error rates for the three approximate pivots based on 10^6 values of x simulated from the bivariate normal distribution with $\xi = x^{\mathrm{o}}$ and covariance matrix $c^2 D$, with c^2 varying from 0.005 to 4, as in Figure 2.4. In order to study the error rates for very rare events, we took α from 2.5% down to 0.005%, the latter corresponding to two-sided 99.99% confidence intervals; the one-sided error rates for $\alpha = 0.05\%$ and $\alpha = 0.005\%$ span the level $\varepsilon = 0.01\%$ above which evasive action might be considered. Under the real conditions of this event, $c^2 = 1$, the left-tail error is systematically high and the right-tail error is exactly zero. This is unsurprising, because in this setting the pivots can be close to zero, and with increasing uncertainty the upper confidence limit is almost invariably larger than ψ_0. As c^2 decreases, the uncertainty becomes unrealistically small and the properties of all three pivots improve. When c^2 increases, the error rates for r and w remain poor, but r^* behaves better, particularly in the left tail, which is of the most interest. Although the errors for r^* are closer to the nominal level, it should be jointly interpreted with the other pivots, especially when the uncertainties are large relative to the miss distance.

The coverage properties of the noninformative Bayesian version of r^*, r_B^* are different and not as satisfactory as those of r^* itself. This is evident from the simulation results reported in Table 2.7, which show that for large uncertainties, the right-tail error is smaller than the nominal value, while the more critical left-tail error is significantly larger than the nominal value. This suggests that confidence intervals based on r_B^* are shifted to the right and may give a false indication of a safe conjunction. This is certainly due to the undesirable properties of the noninformative prior, which suffers from the same“dilution effect" as the probability of collision. In other words, using a noninformative prior may lead to confidence intervals that are shifted upward, and therefore may not accurately reflect the actual risk of a collision.

In a second experiment, we fixed D to the real uncertainty matrix but increased ξ to $c'\xi$, for $c' > 1$ to give a situation in which we expect $\mathrm{E}(\|x\|) \approx \psi_0$ for large c', hence reducing the bias of the collision probability $\widehat{p}_c$. Table 2.8 shows the resulting error rates, again based on 10^6 simulated values of x. All the error rates approach their nominal values as c' increases, but r^* again performs best overall, particularly in the left tail.

Table 2.6 – Empirical left- and right-tail error rates (%) at nominal levels α = 2.5%, 0.5%, 0.05% and 0.005% for case study C, with variance matrix $c^2 D$, based on 10^6 Monte Carlo samples. The standard errors (SE) appear in the last line.

		Left tail (%)				Right tail (%)			
c^2	Statistic	2.5	0.5	0.05	0.005	2.5	0.5	0.05	0.005
	w	2.5986	0.5229	0.0504	0.0062	2.4616	0.4985	0.0542	0.0062
0.005	r	2.6115	0.5275	0.0507	0.0062	2.4274	0.4817	0.0494	0.0047
	r^*	2.5206	0.5059	0.0490	0.0060	2.5256	0.5031	0.0524	0.0048
	w	2.6242	0.5278	0.0531	0.0054	2.4530	0.5194	0.0665	0.0117
0.01	r	2.6411	0.5347	0.0548	0.0057	2.3701	0.4644	0.0433	0.0039
	r^*	2.5064	0.5044	0.0500	0.0053	2.5087	0.5008	0.0476	0.0042
	w	2.7484	0.5611	0.0538	0.0051	3.1974	1.3743	0.4281	0.0616
0.05	r	2.8128	0.5804	0.0570	0.0055	0.8226	0.0000	0.0000	0.0000
	r^*	2.5127	0.5070	0.0474	0.0047	1.3042	0.0000	0.0000	0.0000
	w	2.7974	0.5681	0.0556	0.0057	2.0243	0.3167	0.0000	0.0000
0.1	r	2.9171	0.6048	0.0623	0.0062	0.0000	0.0000	0.0000	0.0000
	r^*	2.4808	0.5020	0.0481	0.0048	0.0041	0.0004	0.0001	0.0000
	w	2.8968	0.5855	0.0602	0.0061	0.1192	0.0000	0.0000	0.0000
0.2	r	3.1142	0.6499	0.0680	0.0071	0.0000	0.0000	0.0000	0.0000
	r^*	2.4780	0.4964	0.0497	0.0050	0.0414	0.0079	0.0013	0.0004
	w	3.1192	0.6139	0.0650	0.0082	0.0000	0.0000	0.0000	0.0000
0.5	r	3.6289	0.7466	0.0807	0.0104	0.0000	0.0000	0.0000	0.0000
	r^*	2.4578	0.4864	0.0510	0.0070	0.3016	0.0862	0.0212	0.0080
	w	3.3729	0.6352	0.0666	0.0088	0.0000	0.0000	0.0000	0.0000
0.8	r	4.1558	0.8421	0.0891	0.0110	0.0000	0.0000	0.0000	0.0000
	r^*	2.4974	0.4877	0.0508	0.0066	0.6050	0.2070	0.0618	0.0264
	w	3.5192	0.6687	0.0619	0.0050	0.0000	0.0000	0.0000	0.0000
1	r	4.4715	0.9130	0.0920	0.0075	0.0000	0.0000	0.0000	0.0000
	r^*	2.5252	0.5021	0.0475	0.0036	0.7748	0.2829	0.0899	0.0380
	w	4.1574	0.7724	0.0737	0.0081	0.0000	0.0000	0.0000	0.0000
2	r	5.7200	1.1922	0.1211	0.0129	0.0000	0.0000	0.0000	0.0000
	r^*	2.6649	0.5265	0.0512	0.0052	1.5817	0.6505	0.2498	0.1222
	w	4.6041	0.8318	0.0761	0.0070	0.0000	0.0000	0.0000	0.0000
3	r	6.5660	1.3730	0.1428	0.0131	0.0000	0.0000	0.0000	0.0000
	r^*	2.7466	0.5273	0.0516	0.0052	2.2104	0.9820	0.4040	0.2065
	w	4.9401	0.8807	0.0772	0.0070	0.0000	0.0000	0.0000	0.0000
4	r	7.2084	1.5009	0.1495	0.0147	0.0000	0.0000	0.0000	0.0000
	r^*	2.8042	0.5438	0.0520	0.0057	2.8894	1.3420	0.5913	0.3171

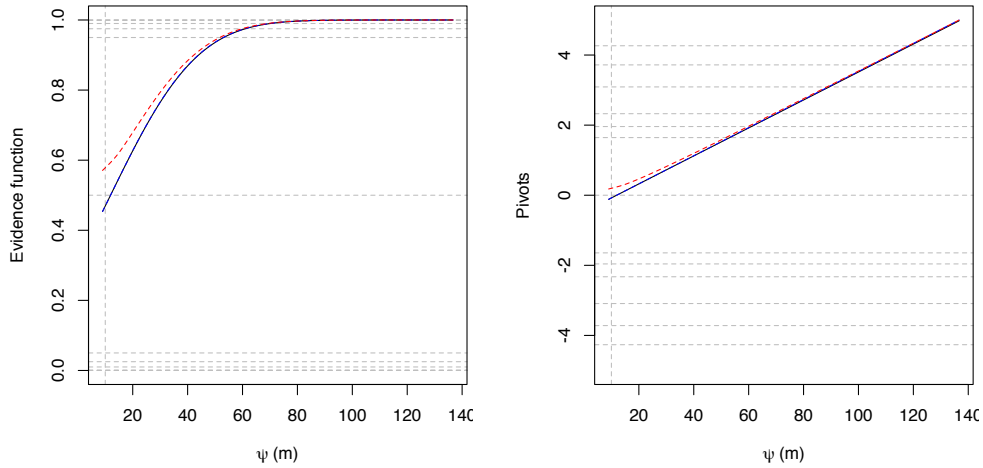


Figure 2.10 – Summaries for case study C. Left: evidence functions based on likelihood root $r^{\mathrm{o}}(\psi)$ (solid black), Wald statistic $w(\psi)$ (dotted blue), and modified likelihood root $r^{*\mathrm{o}}(\psi)$ (red dashes), Right: the same quantities transformed to the standard normal scale. The likelihood root and Wald statistic are almost indistinguishable.

Table 2.7 – Empirical left- and right-tail error rates (%) at nominal levels $\alpha = 2.5\%, 0.5\%, 0.05\%$ and 0.005% of r_B^* for case study C, with variance matrix $c^2 D$, based on 10^6 Monte Carlo samples.

	Left tail (%)				Right tail (%)			
c^2	2.5	0.5	0.05	0.005	2.5	0.5	0.05	0.005
0.005	2.3931	0.4768	0.0454	0.0049	2.5914	0.5330	0.0526	0.0051
0.01	2.3400	0.4635	0.0460	0.0040	2.6760	0.5473	0.0532	0.0056
0.05	2.1672	0.4256	0.0448	0.0046	2.9490	0.5843	0.0547	0.0057
0.1	2.0850	0.3938	0.0377	0.0046	2.8930	0.5094	0.0151	0.0034
0.2	1.8713	0.3509	0.0306	0.0028	2.7698	0.4252	0.1253	0.0522
0.5	1.6067	0.3066	0.0284	0.0031	4.8841	2.1380	0.8695	0.4446
0.8	1.4546	0.2765	0.0266	0.0020	6.8234	3.4632	1.6459	0.9044
1	1.3936	0.2628	0.0265	0.0024	7.8495	4.1776	2.0854	1.2062
2	1.2198	0.2325	0.0227	0.0024	12.8030	7.4818	4.1114	2.5450
3	1.1145	0.2027	0.0187	0.0023	15.6118	9.5435	5.4722	3.5047
4	1.0595	0.1863	0.0176	0.0017	17.7259	11.0861	6.5343	4.2717

Table 2.8 – Empirical left- and right-tail error rates (%) at nominal levels α = 2.5%, 0.5%, 0.05% and 0.005% for case study C, with true position vector $c'\xi$ in the encounter plane, based on 10^6 Monte Carlo samples. The standard errors (SE) appear in the last line.

		Left tail (%)				Right tail (%)			
c'	Statistic	2.5	0.5	0.05	0.005	2.5	0.5	0.05	0.005
	w	2.9354	0.5837	0.0570	0.0066	0.0045	0.0000	0.0000	0.0000
2	r	3.2085	0.6631	0.0677	0.0079	0.0000	0.0000	0.0000	0.0000
	r^*	2.4625	0.4821	0.0478	0.0055	0.0749	0.0157	0.0027	0.0007
	w	2.8360	0.5736	0.0620	0.0063	1.6491	0.1615	0.0000	0.0000
3	r	2.9632	0.6124	0.0675	0.0070	0.0000	0.0000	0.0000	0.0000
	r^*	2.5088	0.5057	0.0530	0.0048	0.0052	0.0005	0.0000	0.0000
	w	2.7575	0.5586	0.0580	0.0053	3.0591	1.2037	0.1975	0.0077
4	r	2.8315	0.5846	0.0615	0.0058	0.0000	0.0000	0.0000	0.0000
	r^*	2.4990	0.4999	0.0506	0.0046	0.0052	0.0000	0.0000	0.0000
	w	2.7127	0.5559	0.0523	0.0048	3.1485	1.3268	0.5772	0.2385
5	r	2.7627	0.5703	0.0554	0.0052	2.1960	0.4215	0.0380	0.0027
	r^*	2.5079	0.5045	0.0475	0.0040	2.5059	0.4962	0.0480	0.0037
	w	2.6819	0.5358	0.0523	0.0047	2.7548	0.9488	0.4112	0.2252
6	r	2.7249	0.5494	0.0537	0.0049	2.2379	0.4364	0.0428	0.0051
	r^*	2.4966	0.4969	0.0483	0.0044	2.4794	0.4978	0.0508	0.0060

2.7.5 Case study D: Minimum miss distance event

The motivation behind trying different examples is to expand the evaluation of our approach against real or nearly real conjunction data, using values for risk assessment thresholds that are much closer to what is employed operationally. In particular, we are interested in testing how the proposed approach performs overall for conjunctions with small miss distances and similarly small HBR thresholds. For that, we consider another example from the limited set of conjunction data released by the CARA group of NASA described in column D of Table 2.2 and Table 2.3. The other released data are the subject of discussion in Alfano (2006b, 2007). These scenarios are not entirely appropriate to give our approach a reasonable exercise, because they are meant to stress the two-dimensional p_c, test the linearity of the relative motion, deal with situations in which the covariance matrix in not positive definite, or situations in which only one covariance of the relative motion is available.

After converting coordinates to the ECI reference frame and defining the relative state vector for the chosen case study, we have $\psi = 3.34$m and

$$x^{\mathrm{o}} = A^{\mathrm{T}} y^{\mathrm{o}} = \begin{bmatrix} 3.31 \\ 0.46 \end{bmatrix}, \quad D = \mathrm{diag}\left(d_1^2, d_2^2\right) = \begin{bmatrix} 2245.55^2 & 0.00 \\ 0.00 & 51.34^2 \end{bmatrix},$$

where

$$A = (CV, v/\|v\|) = \begin{bmatrix} 0.96 & -0.26 & -0.09 \\ -0.12 & -0.08 & -0.99 \\ -0.25 & -0.96 & 0.11 \end{bmatrix},$$

The probability of collision for this test case using different uncertainties $c^2 D$ where $c^2 = 10^{-3}, 10^{-2}, 10^{-1}, 1, 5$, and an HBR between 1 and 30m is given in the left panel of Figure 2.11. The large diagonal elements of D indicate little certainty about the observed state vector and the corresponding miss distance. Using the real covariance, $c^2 = 1$, and an HBR that is shorter than 3m, we obtain collision probabilities smaller than 3.90×10^{-5}, values that indicate the satellite is safe. Shrinking the uncertainty, i.e., $c^2 < 1$, increases the collision risk, and the probabilities become larger than the operational threshold (horizontal dashed line) for most HBR values; this implies that the situation requires a close inspection and remediation might be necessary. Upon consideration, the low value of p_c lies in the dilution region of the probability of collision, and so do the low values obtained by growing the uncertainty for $c^2 > 1$.

In this example, we study the coverage of one-sided right-tail confidence intervals of the form $[L_\alpha, +\infty)$, and we estimate the empirical left-tail error for α from 5% to 0.01%. This choice is not only driven by the fact that calibrated left-tail errors are more crucial

for conjunction assessment but also due to the unusually short miss distance and the large uncertainties, which result in point estimates of ψ bigger than ψ_0 for most values of c^2 considered. The right panel of Figure 2.11 shows histograms of $\widehat{\psi}$ based on 10^6 sampled data used in Table 2.9. This plot shows that even for small values of c^2, the estimated miss distance is larger than the observed value of 3.14m, and when using the real covariance, $\widehat{\psi}$ can reach 6km. Since point estimates of ψ are always bigger than ψ_0, the likelihood root and the Wald statistics, under the null hypothesis, are consistently positive. Therefore, we focus on the postive quantiles in the QQ-plots shown in Figure 2.13.

Unlike in the case study C, the diagonal elements of the covariance matrix D and the scaled covariance c^2D have different magnitudes; the condition number of D is 1912. This implies that the bivariate normal distribution has a covariance error ellipse with an eccentricity that is almost one, i.e., a high degree of ovalness. So, shrinking the covariance, although it results in a smaller ellipse, preserves the asymmetry in the uncertainty along the two axes defining the encounter plane.

While the right-tail error is systematically zero because the upper confidence bound is consistently greater than $\psi_0 = 3.14$m, the left-tail error is bigger than the nominal values for most of the scenarios examined in Table 2.9. We consider values of c^2 varying from 10^{-6} to 0.2, then d_1 varies from 2.245m to 1.004km, and d_2 from 0.051m to 22.961m. As we can see, for small uncertainty, all statistics have similar error rates, but as c^2 increases, the Wald statistic behaves better than the likelihood root, even though both are far from the nominal values. The distribution of the modified likelihood root under the null hypothesis is closer to standard normal and produces left-tail errors that are much closer to the nominal value; see Table 2.9 and the third column of Figure 2.13.

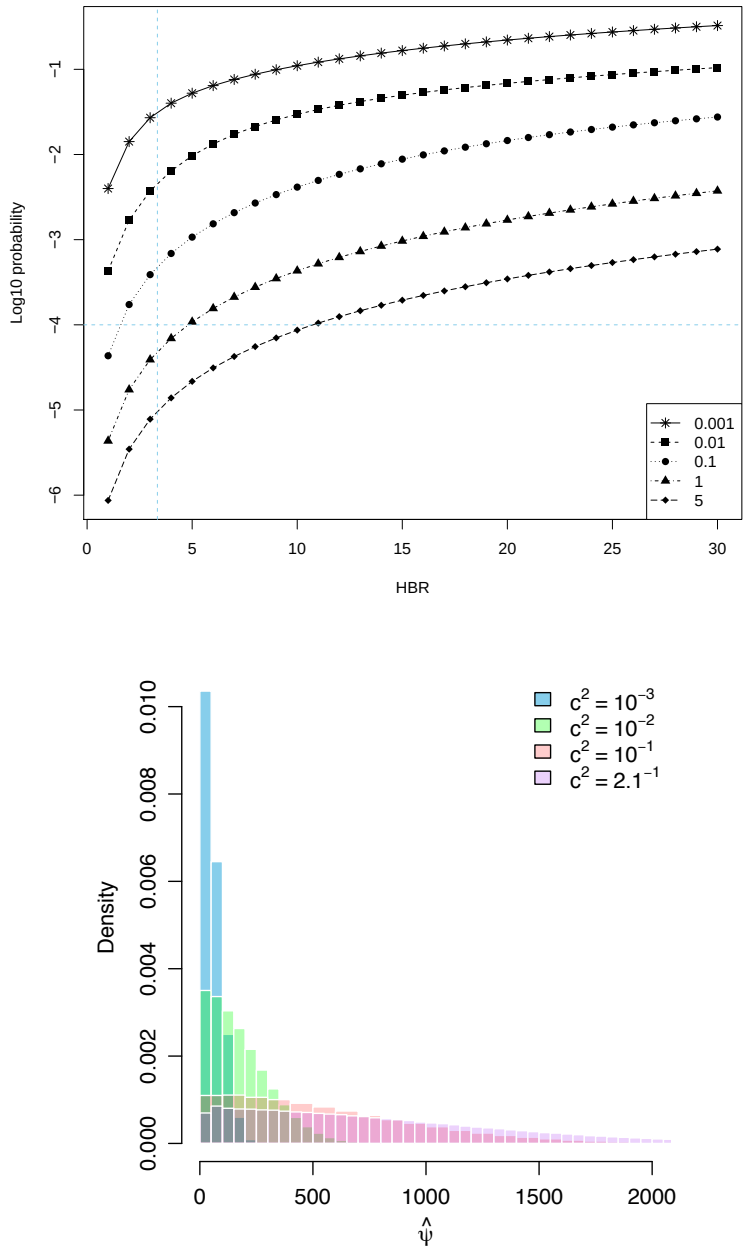


Figure 2.11 – Case study D: Top: probability of collision with uncertainty in the encounter plane set to c^2D and HBR(m) varying from 1m to 30m, the dashed blue vertical line is $\psi^o = 3.34$m, and the dashed blue horizontal line is the operational threshold 10^{-4}. Bottom: histograms of the estimated miss distance based on 10^6 data sampled from the bivariate normal distribution with variance c^2D.

Table 2.9 – Empirical left- and right-tail error rates (%) at nominal levels α = 5%, 1%, 0.1% and 0.01% for case study D based on 10^6 Monte Carlo samples. The standard errors (SE) appear in the last line.

		Left tail (%)			
c^2	Statistic	5.00	1	0.10	0.01
10^{-6}	w	4.9817	0.9956	0.0961	0.0084
	r	5.0218	1.0100	0.0978	0.0087
	r^*	5.0203	1.0098	0.0976	0.0087
$2\cdot10^{-6}$	w	5.0057	0.9984	0.1014	0.0108
	r	5.0457	1.0120	0.1041	0.0109
	r^*	5.0429	1.0115	0.1040	0.0109
10^{-4}	w	7.6229	1.4427	0.1383	0.0125
	r	7.6684	1.4535	0.1391	0.0126
	r^*	7.6386	1.4461	0.1383	0.0125
$2\cdot10^{-4}$	w	8.1552	1.5569	0.1437	0.0173
	r	8.2105	1.5732	0.1453	0.0175
	r^*	8.1608	1.5592	0.1436	0.0172
10^{-3}	w	9.079	1.767	0.169	0.020
	r	9.358	1.827	0.176	0.021
	r^*	8.740	1.700	0.162	0.020
$2\cdot10^{-3}$	w	9.3951	1.8701	0.1829	0.0158
	r	10.2440	2.0462	0.2036	0.0171
	r^*	8.2992	1.6536	0.1622	0.0148
10^{-2}	w	9.715	1.911	0.194	0.017
	r	14.406	3.021	0.314	0.028
	r^*	5.732	1.135	0.118	0.010
$2\cdot10^{-2}$	w	9.715	1.911	0.194	0.017
	r	14.406	3.021	0.314	0.028
	r^*	5.732	1.135	0.118	0.010
10^{-1}	w	9.8856	1.9698	0.1993	0.0215
	r	20.9553	4.9731	0.5711	0.0632
	r^*	2.9182	0.5928	0.0643	0.0062
$2\cdot10^{-1}$	w	9.950	1.988	0.203	0.021
	r	22.307	5.425	0.641	0.069
	r^*	2.446	0.501	0.055	0.006
SE ($\times10^{-3}$)		21.79	9.94	3.16	0.99

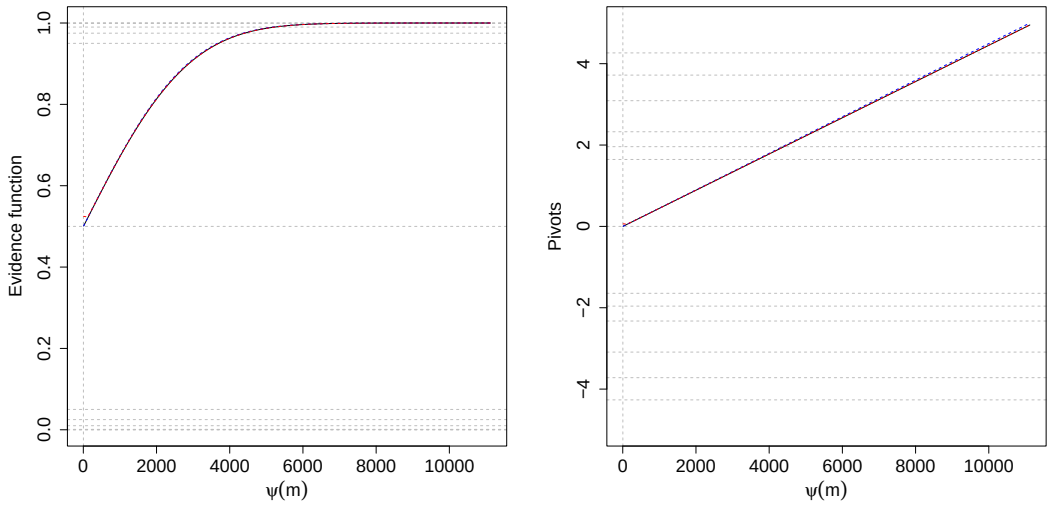


Figure 2.12 – Summaries for case study D. Left: evidence functions based on likelihood root $r^{\mathrm{o}}(\psi)$ (solid black), Wald statistic $w(\psi)$ (dotted blue), and modified likelihood root $r^{*\mathrm{o}}(\psi)$ (red dashes), Right: the same quantities transformed to the standard normal scale. All pivots are almost indistinguishable.

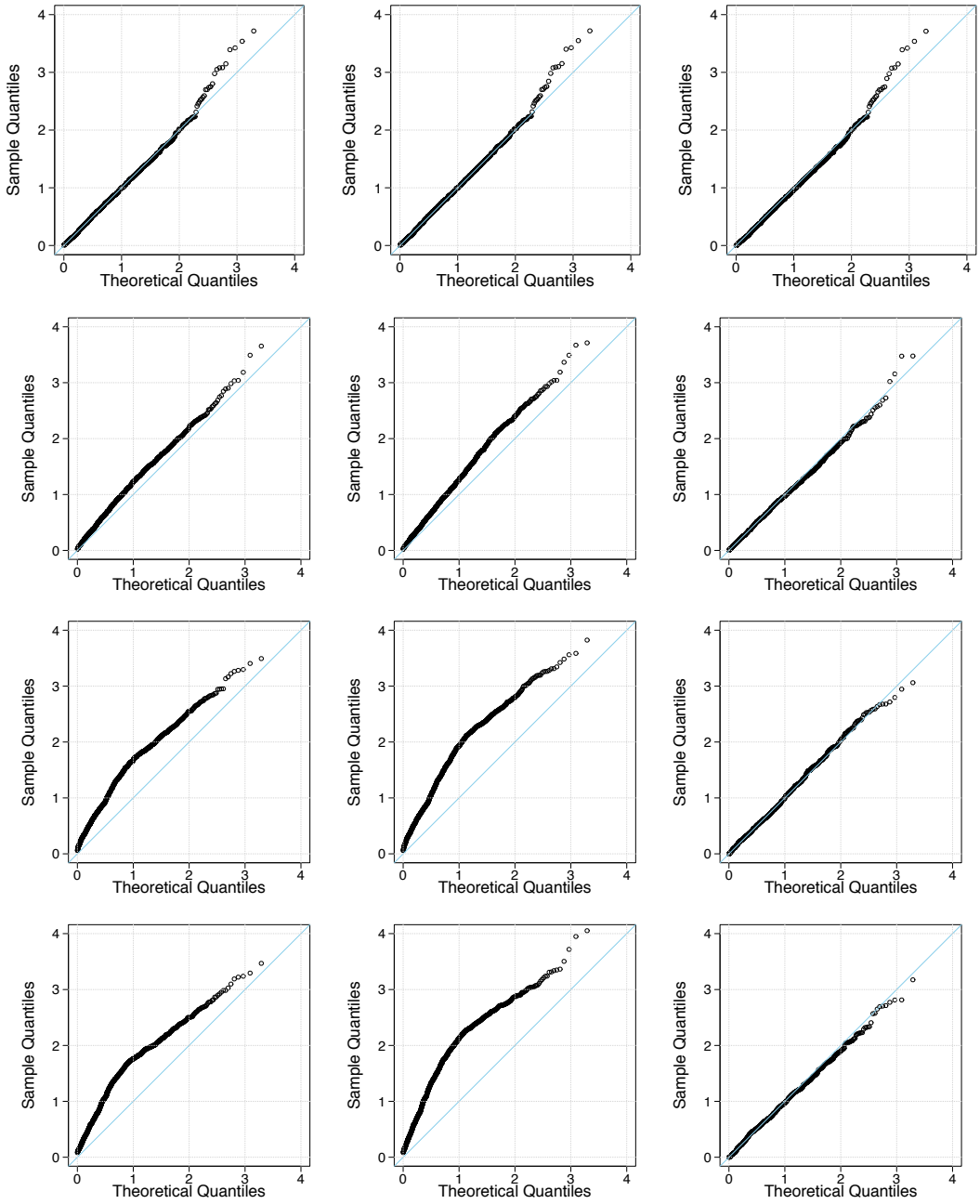


Figure 2.13 – Case study D: normal QQ-plots of $w(\psi^o)$ (left), $r(\psi^o)$ (middle) and $r^*(\psi^o)$ (right) based on 10^6 Monte Carlo sample quantiles with variance $c^2 D$ where c^2 equals $10^{-3}, 10^{-2}, 10^{-1}, 2 \cdot 10^{-1}$ (top to bottom).

2.8 Conclusion

In this chapter, we reviewed the main methods for determining the collision probability of space objects in relative linear motion. We formulated conjunction assessment in statistical terms and discussed likelihood inference on the miss distance, both when the relative velocity can be taken as known and when its uncertainty must be taken into account. If a constant prior density is placed on the encounter plane, the posterior probability that the second object lies within the hard-body radius equals the usual collision probability. This immediate interpretation of the probability of collision as a Bayesian formulation is not further explored since any other plausible prior would need to reflect the underlying risk of collision due to orbital crowding. As currently practiced, no such prior is available in conjunction analysis.

We studied the repeated-sampling properties of likelihood statistics in this setting. Viewed in our framework, the collision probability estimator has a downward bias that seems not to have been noticed previously, and the so-called probability dilution paradox vanishes, since it refers to the properties of an estimator of the collision probability rather than to the probability itself.

Examples illustrate inference on the miss distance, suggest that standard likelihood confidence intervals may need improvement when uncertainty on the relative distance and velocity is large. The numerical results show that an improved approximation gives appreciably better inferences. However, the previously highlighted Bayesian interpretation of the collision probability does not alter its downward bias and the use of noninformative priors such as the Jeffreys prior does not improve coverage properties of our approach, so one should not be optimistic about the effect of Bayesian correction in the context of conjunction assessment. In our setup, the estimated collision probability is replaced by a significance probability for testing whether the true miss distance is larger than a safety threshold. If the model is correctly specified then this probability is calibrated in a repeated-sampling sense and thus provides a statistically well-founded basis for avoidance decisions.

2.9 Appendices of Chapter 2

Appendix A: Implementation details

The main steps for the numerical implementation of higher-order quantities are given in Table 2.11. Numerical experiments were performed with an Intel® Xeon(R) CPU E5-1650 v3 @3.50GHz × 12 processor with 64GB ram. To obtain coverage of the confidence intervals for a specific case study, the R code is designed to use parallel loops for speed-up purposes. It takes roughly 8 seconds to run 10^3 simulations in the two-dimensional settings and 15 seconds for six-dimensional model.

Appendix B: Successive observations

In some cases successive six-dimensional observation vectors $y_1, \ldots, y_n$ and corresponding 6×6 variance matrices $\Omega_1^{-1}, \ldots, \Omega_n^{-1}$ are available, with the variance matrices increasingly concentrated as information accrues on a conjunction. If the observations can be regarded as independent, then the corresponding log likelihood is

$$\ell(\psi, \lambda_1, \ldots, \lambda_n) = -\frac{1}{2} \sum_{j=1}^{n} \{y_j - \eta(\vartheta_j)\}^{\mathrm{T}} \Omega_j \{y_j - \eta(\vartheta_j)\},$$

where $\vartheta_j = (\psi, \lambda_j)$, with ψ representing the miss distance common to all the observations and $\lambda_1, \ldots, \lambda_n$ representing 5×1 vectors of nuisance parameters corresponding to $y_1, \ldots, y_n$. The more precise y_j are automatically given higher weight, since the corresponding dispersion matrices Ω_j are larger. In this case the overall parameter vector is $\vartheta = (\psi, \lambda_1, \ldots, \lambda_n)$, and the approach of Section 2.6 can be applied with minor changes. A similar but more complex generalisation should be feasible when the observations are dependent due to batch updates.

More complicated geometric discussion leading to a different form for $\eta(\vartheta_j)$ would be needed if the relative motions could not be considered to be rectilinear.

Likelihood methods rest on distributional assumptions, and one might query whether it is appropriate to assume normality of the observation vector. Provided suitable data are available, standard methods could be used to check model adequacy and used to modify the model if this was found to be necessary. For example, the multivariate Student t or Laplace distributions might be used, though the numerical details would be more complex.

In a Bayesian set-up, the incorporation of reliable prior information would be valuable,

Numerical details

Input: 6×1 state vector y containing the observed 3×1 relative position and velocity vectors $\widehat{\mu}$ and $\widehat{v}$, 6×6 covariance matrix Ω^{-1} for y, and a nominal confidence level α.

1. Compute the overall maximum likelihood estimates $\widehat{\theta}$: (i) in the six-dimensional setting, $\widehat{\psi}$ and $\widehat{\lambda} = (\widehat{\theta}_1, \widehat{\phi}_1, \widehat{\|v\|}, \widehat{\theta}_2, \widehat{\phi}_2)$ are obtained from $(\widehat{\mu}, \widehat{v})$ using equations (2.3)–(2.6); and (ii) in the simplified two-dimensional setting, we first need to project relative quantities into the encounter plane using the matrix A (see Section 2.3.2), then $\widehat{\psi}$ and $\widehat{\lambda}$ are given in equation (2.22).

2. Compute the observed information matrix $J(\widehat{\vartheta})$ given by (2.10) (a 6×6 or 2×2 matrix), the estimated variance of each $\widehat{\vartheta}_r$ is the (r,r) element of $J(\widehat{\vartheta})^{-1}$. Let $\mathrm{se}(\widehat{\psi})$ denote the square root of the estimated variance for $\widehat{\psi}$.

3. Define a grid $\mathcal{G} = \{\psi'_1, \ldots, \psi'_{n_\psi}\}$ of values of ψ that includes the maximum likelihood estimate $\widehat{\psi}$. We use a non-uniform grid in the interval $\left[\max\left(\widehat{\psi} - z_{1-\alpha}\mathrm{se}(\widehat{\psi}), 0\right), \widehat{\psi} + z_\alpha \mathrm{se}(\widehat{\psi})\right]$, such that the mesh is finer for small ψ and coarser for larger ψ.

4. For each $\psi \in \mathcal{G}$, obtain the constrained estimates $\widehat{\lambda}_\psi$ subject to $0 \le \theta_1, \theta_2 \le \pi$, $-\pi \le \phi_1, \phi_2 \le \pi$ and $0 < ||v||$, by minimising the sum of squares in (2.9), and store the corresponding values of $\widehat{\vartheta}_\psi = (\psi, \widehat{\lambda}_\psi)$. We have found that

   - it may help to transform the components of ϑ to take values in the real line, for example removing the restrictions by maximizing in terms of $\log||v||, \log\{\theta_i/(\theta_i - \pi)\}$, and $\tan(\phi_i/2)$, for $i = 1,2$, and
   - constrained optimisation using the 'Rvmmin' or 'nlminb' solvers in the optimx() function of the R package optimx leads to an algorithm that is overall robust, fast, and generally insensitive to perturbations in initial values.

5. Use partial derivatives of the log-likelihood and the mean vector η to evaluate $\ell(\widehat{\vartheta}_\psi)$, $J_{\lambda\lambda}(\widehat{\vartheta}_\psi)$, $\eta(\widehat{\vartheta}_\psi)$ and $\eta_\lambda(\widehat{\vartheta})$ and then use expressions (1.1), (1.3) and (1.27) to obtain the values of $r(\psi)$, $w(\psi)$, $q(\psi)$ and $r^*(\psi)$ on $\mathcal{G}$.

6. Interpolate $r(\psi)$ and $r^*(\psi)$ on $\mathcal{G}$ by (for example) a cubic smoothing spline in which the values of ψ are treated as functions of those of $r(\psi)$ and $r^*(\psi)$. Very large values of r^* arising in a few cases due to numerical instabilities when $|r| < 0.1$ are excluded.

7. If required, obtain the point estimate $\widehat{\psi}^*$ of ψ by using the interpolating spline for $r^*(\psi) = 0$. This is not needed for the Wald or the likelihood root, as $\widehat{\psi}$ is already known.

8. Obtain a $(1-2\alpha)$ confidence interval $(\psi_\alpha, \psi_{1-\alpha})$ based on $r(\psi)$ as the solutions of $r(\psi) = \pm z_\alpha$. If the equation $r(\psi) = -z_\alpha$ cannot be solved, then the lower limit of the confidence interval is $\psi_\alpha = 0$. Confidence intervals based on $r^*(\psi)$ are obtained likewise.

Output: Point estimates $\widehat{\psi}$ and $\widehat{\psi}^*$ of the miss distance and corresponding two-sided $(1-2\alpha)$ confidence intervals.

Table 2.11 – Algorithm for likelihood inference on the miss distance

and might for example be based on the output of a filtering approach to tracking. One might then treat conjunction analysis as a prediction problem rather than an estimation problem, and then the Bayesian formalism would be attractive.

For the satellite problem, we use a Bayesian formulation and let η_0 and y_0 respectively denote the state vector and its observed value at time $t = 0$, where

$$\eta_0 \sim \mathcal{N}\left(\eta_0', \Sigma_0^{-1}\right), \quad y_0 \mid \eta_0 \sim \mathcal{N}\left(\eta_0, \Omega_0^{-1}\right)$$

Here η_0' and Σ_0^{-1} represent prior knowledge about η_0. The posterior distribution of η_0 given y_0 is then

$$\eta_0 \mid y_0 \sim \mathcal{N}\left\{(\Omega_0 + \Sigma_0)^{-1}\left(\Omega_0 y_0 + \Sigma_0 \eta_0'\right), (\Omega_0 + \Sigma_0)^{-1}\right\}$$

which reduces to

$$\eta_0 \mid y_0 \sim \mathcal{N}\left(y_0, \Omega_0^{-1}\right)$$

when $\Sigma_0 \to 0$, reflecting prior ignorance of the initial state vector. This posterior distribution is then updated in later steps of the Kalman filter when subsequent observations $y_1, \ldots, y_n$ are seen at respective times $t_1, \ldots, t_n$. Let $\mathcal{H}_j = \{y_0, \ldots, y_j\}$ denote the observations seen up to and including time t_j, write $\eta_j = \eta_{t_j}$, and let m_j and S_j respectively denote the posterior mean and variance matrix for η_j conditional on $\mathcal{H}_j$. Note that m_j is a linear function of $y_0, \ldots, y_j$. We suppose that

$$\eta_{j+1} = \Phi_{j+1}\eta_j + u_{j+1},$$

where Φ_{j+1} is a known transition matrix for the evolution of the state vector between times t_j and t_{j+1}, and $u_{j+1} \sim \mathcal{N}\left(0, \Sigma_{j+1}^{-1}\right)$ represents random influences on the state vector between these times. This implies that

$$\eta_{j+1} \Big| \mathcal{H}_j \sim \mathcal{N}\left(\Phi_{j+1} m_j, \Phi_{j+1} S_j \Phi_{j+1}^{\mathrm{T}} + \Sigma_{j+1}^{-1}\right), \quad y_{j+1} \Big| \eta_{j+1} \sim \mathcal{N}\left(\eta_{j+1}, \Omega_{j+1}^{-1}\right),$$

and standard calculations with the normal distribution then show that

$$\eta_j \mid \mathcal{H}_{j-1}, y_j \sim \mathcal{N}\left(m_{j+1}, S_{j+1}\right)$$

where $\{\mathscr{H}_{j-1}, y_j\} \equiv \mathscr{H}_{j+1}$ and, setting $A_{j+1} = \Phi_{j+1} S_j \Phi_{j+1}^{\mathrm{T}} + \Sigma_{j+1}^{-1}$,

$$\begin{aligned} m_{j+1} &= \Phi_{j+1} m_j + A_{j+1} \left(A_{j+1} + \Omega_{j+1}^{-1} \right)^{-1} \left(y_{j+1} - \Phi_{j+1} m_j \right) \\ &= \left(A_{j+1} + \Omega_{j+1}^{-1} \right)^{-1} \left(A_{j+1} y_{j+1} + \Omega_{j+1}^{-1} \Phi_{j+1} m_j \right), \\ S_{j+1} &= A_{j+1} - A_{j+1} \left(A_{j+1} + \Omega_{j+1}^{-1} \right)^{-1} A_{j+1}. \end{aligned}$$

Both of these expressions make sense: the posterior mean for η_{j+1} is a weighted average between the new estimate y_{j+1} and the updated previous estimate, $\Phi_{j+1} m_j$; and if $\Omega_{j+1}^{-1} = 0$ then y_{j+1} provides precise information about the state at time t_{j+1}, so $S_{j+1} = 0$, whereas if $\Omega_{j+1}^{-1} = \infty$, then no additional information is provided by y_{j+1} and then $S_{j+1} = A_{j+1}$ is entirely based on $\mathscr{H}_j$ plus information about the subsequent evolution of the state, but not the next observation. Hence, if observations are available at times $0, t_1, \ldots, t_n$, leading to conditional mean and variance matrix m_n and S_n at time t_n, Bayesian inference can then be performed, approximating the posterior distribution for the minimal distance ψ between the two space objects using the previous computations.

Accurate Inference in Boundary Problems

3.1 Introduction

Inference procedures based on likelihood theory form the backbone of many statistical methods owing to the appeal of likelihood as a measure of plausibility, the generality of the likelihood paradigm, the flexibility with which new problems can be addressed, and close links to Bayesian ideas. The original notion now encompasses a wide range of related ideas, including conditional and marginal likelihoods, partial likelihoods, empirical likelihoods, quasi- and pseudo-likelihoods, and composite likelihoods; see for example Pawitan (2001). An appealing aspect is that the standard theory leads to a few well-understood, simple and widely applicable approximations for inference. These typically rely on normal and chi-squared distributions and have been easy to apply since well before the computer age. Over the past few decades this classical theory has developed further, and as we saw in earlier chapters in its modern form it can yield highly accurate inferences based on parametric models even for very small samples.

Much less attention has been paid to so-called non-regular cases, under which the standard conditions for validity of these classical approximations do not hold. These conditions, which are typically of Cramér type (Cramér, 1946, §33.3), include differentiability of the underlying joint probability or density function up to a suitable order and finiteness of the Fisher information matrix. Unfortunately they fail for models that are commonly used in applications of much practical interest in genetics, reliability, econometrics and many other fields. One example is so-called endpoint problems, in which the support of the observations must be estimated; in this case the shape of the density function at the limits of its support determines the accuracy with which the endpoint, and possibly other parameters, can be estimated (Smith, 1985). Non-regularity can arise in many other ways. A highly cited review of nonregular

problems is Smith (1989); see also Cheng and Traylor (1995). Further examples can be found in Barndorff-Nielsen and Cox (1994, §3.8), Davison (2003, §4.6) and Cox (2006, Chapter 7). Brazzale and Mameli (2022) group non-regular problems into three broad classes. One broad class consists of change-point problems, and a related class is situations in which one or more components of the parameter vanishes when another component is set to a particular value. In both cases approximate distributions for likelihood-based statistics can be complex and their usefulness may be limited in realistic settings. A third class of non-regular problems comprises so-called boundary cases, where it is desired to test the hypothesis that some interest parameter ψ equals a null value ψ_0 against the alternative that $\psi > \psi_0$, and ψ_0 lies on the boundary of its domain. Informally, the methodological difficulties in likelihood-based inference occur because the maximum likelihood estimate can only fall on the 'right-hand' side of ψ_0. If the maximum occurs on the boundary, ψ_0, the score function need not be zero and the distributions of related likelihood statistics will not converge to the typical normal or chi-squared distributions. Because of the difficulties inherent to the derivation of the limiting distribution of the likelihood ratio statistic, practitioners tend to ignore the boundary problem and to proceed as if ψ_0 was an interior point of its parameter space. This naïve approach may lead to highly inaccurate inferences especially for complex models.

In this work, we study finite-sample approximations for certain boundary problems and show how they may be greatly improved using higher-order likelihood procedures. As discussed in Chapter 1, an extensive literature on higher-order likelihood inference for regular models, in both classical and Bayesian frameworks, is available (Brazzale et al., 2007). Severini (2000) and Barndorff-Nielsen and Cox (1994) show how highly accurate approximations to the distributions of test statistics and pivots may be obtained for a variety of parametric statistical models. However, the only precursor paper for higher-order inference for boundary problems of which we are aware is Castillo and López-Ratera (2006), who demonstrate the validity of an improved signed likelihood ratio statistic when testing a boundary hypothesis on a scalar parameter in an exponential family.

One problem we tackle is testing for a null component of variance in a mixed-effects model. This includes as a special case comparison of parametric regression models with semiparametric alternatives, under the now-standard formulation of spline regression as a linear mixed model (Laird and Ware, 1982; Ruppert et al., 2003; McCulloch and Searle, 2001; Wood, 2017). The asymptotic distribution of the likelihood ratio statistic in such cases is typically a $\bar{\chi}^2$, that is, a mixture of chi-squared variables with known probabilities and degrees of freedom (Self and Liang, 1987), though other limiting distributions are found as well (Sinha et al., 2012). The second problem we address

is infinite mixtures, which embrace models such as the Student t with continuous $\psi^{-1} = \nu$ degrees of freedom, the negative binomial with overdispersion parameter $\psi^{-1} = \nu$, and the generalized Pareto distribution with shape parameter ψ. These distributions become identical to the Normal, Poisson and exponential, respectively, if $\psi \to 0$. Here, the limiting distribution of the likelihood ratio statistic puts mass $\frac{1}{2}$ at $\psi = 0$, with the remaining probability spread as a χ^2_1 distribution, though we shall see that finite-sample results are unreliable even with large sample sizes.

As we will discuss in Section 3.2, this is because this type of problem places a *hard* boundary, which cannot be crossed, on the domain of ψ. However, this is not the case for a *soft* boundary, where $\psi = \psi_0$ lies on the edge of the 'statistical' parameter space, but can be an interior point of the 'mathematical' parameter space for which the density function is well-defined. Because of the existence of this 'enlarged' parameter space, no difficulties with the existence of derivatives arise on the statistical boundary, which may justify the better, though still unsatisfactory performance, of large-sample likelihood pivots. A distinction hence needs be made according to whether the boundary is soft or hard.

As we shall see in Sections 3.4 and 3.5, no difficulties with the standard higher-order methods will be observed for the former. Research for the latter case is incomplete as hard boundary problems are more difficult. Likelihood pivots on the boundary need be calculated as left limits for $\psi \to \psi_0$, but, the required derivatives may not exist on the boundary or may be numerically unstable. Depending on the shape of the log likelihood function at the boundary, the score function may be heavily skewed, and the limiting distribution far from normal or chi-squared, even for very large sample sizes. They may not even have the classical $n^{1/2}$ asymptotic order at the boundary but for example, $\left(n \log n\right)^{1/2}$, as seen in Ledford and Tawn (1996).

This chapter is organized as follows: Section 3.2 reviews the literature on boundary problems and motivates our work with two examples illustrating the poor performance of first-order pivots regardless of the boundary type. Section 3.3 presents direct methods for improving finite-sample approximation of the mixing probabilities, notably using the profile score and Edgeworth expansion for its distribution. We also propose a rough-and-ready remedy using a "shadow" estimator. However, this approximation can only be applied when boundary probabilities can be computed or approximated and presupposes that the shadow estimator has an approximately normal distribution. In Section 3.4, we examine the direct improvements presented in Section 3.3 for some soft and hard boundary examples. Numerical results for the tangent exponential model using Monte Carlo simulations and real datasets are shown in Sections 3.5 and 3.6.

3.2 Boundary problems

3.2.1 Background

Let $\ell(\psi, \lambda)$ denote the log likelihood for a parametric statistical model for data with sample size n, possibly notional, with scalar parameter of interest ψ and nuisance parameter λ, where $\psi \in \Psi = [\psi_0, \infty)$ and the value of λ that generated the data is interior to an open set Λ. Suppose we wish to test the boundary null hypothesis $H_0 : \psi = \psi_0$. If ψ_0 were interior to Ψ, then under mild regularity conditions the signed likelihood ratio statistic

$$r(\psi_0) = \text{sign}(\widehat{\psi} - \psi_0) \left[2 \left\{ \ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi_0, \widehat{\lambda}_0) \right\} \right]^{1/2}, \tag{3.1}$$

would have a standard normal distribution in large samples, and its square $r(\psi_0)^2$ would follow a chi-squared distribution with 1 degree of freedom. Here $\widehat{\lambda}_0$ denotes the maximum likelihood estimator of λ when $\psi = \psi_0$. In the present setting, $\psi = \psi_0$ under the null hypothesis, a boundary problem arises.

Research on boundary problems was initiated by Chernoff (1954) who derived the asymptotic null distribution of the likelihood ratio statistic for testing whether ψ lies on one or the other side of a smooth $(d-1)$-dimensional surface in a d-dimensional space when the true parameter value lies on the surface. Using geometrical arguments, Chernoff established that this distribution is equivalent to the distribution of the likelihood ratio statistic for testing suitable restrictions on the mean of a multivariate normal distribution with covariance matrix given by the inverse of the Fisher information matrix using a single observation. A later cornerstone contribution which inspired many researchers and fuelled an enormous literature is the highly-cited article by Self and Liang (1987). This time, ψ_0 no longer need be an interior point of the parameter space, but can fall onto the boundary. However, the parameter space must be regular enough to be asymptotically approximated by a cone with vertex at ψ_0. A further major step forward in likelihood asymptotics for boundary problems was marked by Kopylev and Sinha (2011) and Sinha et al. (2012), who derived the null distribution of the likelihood ratio statistic by algebraic arguments. From a technical point of view, the derivation of a closed-form expression for the limiting distribution of the likelihood ratio becomes more difficult the more nuisance parameters lie on the boundary of the parameter space. All these contributions are summarized in Chow et al. (2012), with some interesting examples and an account of the areas of interest in genetics and biology. In general terms, the asymptotic distribution turns out to be a

chi-bar squared distribution (Kudo, 1963) with cumulative distribution function

$$\Pr(\bar{\chi}^2 \leq c) = \sum_{\nu=0}^{d} \omega_\nu \Pr(\chi^2_\nu \leq c),$$

that is, a mixture of chi-squared distributions with degrees of freedom ν, and probabilities ω_ν, where ν varies from 0 to d, and χ^2_0 is a point mass in zero. This mixture depends on the number and type of parameters, and on the geometry of the tangent cone at the null hypothesis.

Later we consider testing for a null component of variance in a mixed effects model. As mathematically the likelihood function will be defined also on a suitable extension of the parameter space, we call this type of problem a soft boundary problem. Crainiceanu et al. (2002) and Crainiceanu and Ruppert (2004a) derived the finite sample distributions of the likelihood ratio and the restricted likelihood ratio tests if there is a single variance component. They showed that asymptotic results give very poor approximations in analysis of variance and penalized spline models. One reason for this is that the asymptotic mixing proportions may be wildly inaccurate even in relatively large samples. Susko (2013) showed that likelihood ratio tests using data-dependent degrees of freedom give conservative asymptotic type I error.

A further common example of a boundary problem is a mixture model, in which the data y follow the density $\psi f(y;\xi) + (1-\psi) g(y;\lambda)$, and $\psi_0 = 0$ represents the possibility that the data originate from the density $g(y;\lambda)$. The parameters ξ disappear under the null hypothesis (Davies, 1987; Ritz and Skovgaard, 2005), and the distribution of the likelihood ratio statistic for testing $\psi = \psi_0$ may typically be approximated by the supremum of some function of a Gaussian process. We do not address this situation in this work, which is concerned with the simpler setting in which all nuisance parameters are present under the null hypothesis, but the limiting distribution of the likelihood ratio statistic is still not chi-squared. When ψ cannot go below ψ_0, we call this a hard boundary problem, under which mathematical and numerical difficulties arise in obtaining the maximum likelihood estimate, i.e.,in obtaining a solution that satisfies the score equations and has nonsingular information matrix.

Ross (1990) called models with soft boundaries pseudomodels and listed some interesting cases, such as the double-exponential regression models. This interpretation of boundary problems is inspired by the work of Chant (1974). Another example of soft boundaries is the generalization of the Weibull model to include negative powers and its extended form, the Generalized Extreme Value (GEV) distribution. This example is discussed in Hosking (1984),Smith (1985), and Cheng and Traylor (1995), and connected to the Generalized Pareto (GP) distribution we study in Section 3.4. A limited

number of hard boundary examples are well-studied. Dannemann and Holzmann (2008) examined the finite-sample distribution of the likelihood ratio when testing for zero entries of the transition matrix in hidden Markov models. This is related to testing the parameters of the stationary distribution of the underlying Markov chain. Bartolucci (2006) also studied boundary problems for latent Markov models, for a combination of joint boundary null hypotheses. Further examples include testing for the mixture properties in finite mixtures when the distributions belong to several families (Chen and Kalbfleisch, 2005). To handle models with hard boundaries, some researchers have proposed using techniques such as reparameterization or penalization of the original model to transform the model into one with a non-singular information matrix (Lee, 1993; Rotnitzky et al., 2000). In this work, we offer alternative approaches for dealing with these types of irregular models.

3.2.2 Soft boundaries

Consider the one-way random effects model

$$Y_{ij} = \mu + b_i + \varepsilon_{ij}, \quad i = 1,\ldots,k, j = 1,\ldots,m, \tag{3.2}$$

where μ is the overall mean and the b_i and the ε_{ij} are mutually independent normal random variates having zero means and variances σ_b^2 and σ^2. This corresponds to a sample of independent observations divided into k groups each of size m, so $n = mk$. Let $\psi = \sigma_b^2/\sigma^2$ and set $\lambda = (\mu, \sigma^2)$, so that a test of the boundary hypothesis $H_0 : \psi = \psi_0 = 0$ corresponds to testing $b_1 = \cdots = b_k = 0$. In this case, an exact test is available using the F distribution of the ratio of between- and within-group mean squares, but it is instructive to apply the large-sample approximation nonetheless.

The work of Chernoff (1956) implies that the likelihood root has large-sample distribution

$$\Pr_0\{r(\psi_0) \le x\} = \tfrac{1}{2} H(x) + \tfrac{1}{2}\{2\Phi(x) - 1\} I(x > 0), \quad x \in \mathbb{R}, \tag{3.3}$$

where $\Pr_0(\cdot)$ denotes a probability computed under the null hypothesis H_0, $H(x)$ denotes the Heaviside function, $\Phi(x)$ is the standard normal distribution function, and $I(\cdot)$ is the indicator function. The p-value for a test of H_0 is $p_{\text{obs}} = \Pr_0\{r(\psi_0) \ge r_{\text{obs}}\}$, where r_{obs} is the observed value of $r(\psi_0)$. If $\widehat{\psi} = \psi_0$, i.e., the maximum likelihood estimate of ψ lies on the boundary, then $r_{\text{obs}} = 0$ and (3.3) yields $p_{\text{obs}} = 1$, whereas if $r_{\text{obs}} > 0$, i.e., $\widehat{\psi} > \psi_0$, then (3.3) yields $p_{\text{obs}} = \Phi(-r_{\text{obs}})$.

Apart from additive constants the log likelihood may be written as

$$\ell(\mu,\sigma^2,\psi) = -\tfrac{1}{2}\left\{mk\log\sigma^2 + \frac{C_2}{\sigma^2} + k\log(1+m\psi) + \frac{C_1}{\sigma^2(1+m\psi)} + \frac{km(\bar{y}_{\cdot\cdot}-\mu)^2}{\sigma^2(1+m\psi)}\right\}, \tag{3.4}$$

where the grand mean $\bar{y}_{\cdot\cdot}$ and the sums of squares C_1 and C_2 between groups and within groups are mutually independent and satisfy

$$C_1 = m\sum_{i=1}^{k}\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2 \sim \sigma^2(1+m\psi)\chi^2_{k-1}, \quad C_2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2 \sim \sigma^2\chi^2_{k(m-1)},$$

where $\bar{y}_{\cdot\cdot}$ and $\bar{y}_{i\cdot}$ are respectively the overall average and that for the ith group. It is easily checked that the profile log likelihood for ψ may be written as

$$\ell_{\mathrm{p}}(\psi) \equiv -\frac{k}{2}\left\{m\log\left(C_2 + \frac{C_1}{1+m\psi}\right) + \log(1+m\psi)\right\}, \quad \psi \geq 0.$$

The log likelihood function in (3.4) is defined for values of $\psi \geq 0$. However, we can extend the range of ψ to include values down to $-1/m$, so $\psi_0 = 0$ is a soft boundary.

Using the likelihood root defined in (3.1) we have $\Pr_0\{r(\psi_0) > 0\} = \Pr(B > m^{-1})$, where B has the beta distribution with parameters $(k-1)/2$ and $k(m-1)/2$. There are two aspects to the asymptotic approximation (3.3), namely the limiting probabilities and the half-normal approximation that applies for positive x. In some cases the probabilities are very far from $1/2$, even in very large samples. Indeed, Table 3.1 shows that this probability is far from the asymptotic value of $1/2$ for any values of k and m likely to arise in practice. In this case, the limiting probability of $1/2$ appears as $k \to \infty$, and the asymptotic approximation degrades slightly as m increases for fixed k.

It is natural to try and improve the approximation by using the restricted log likelihood (Harville, 1977), which amounts to dropping the last term in (3.4) and replacing the coefficients mk and k for the logarithmic terms by $mk-1$ and $k-1$ respectively. The resulting probability of a positive gradient, $\Pr\{B > (k-1)/(km-1)\}$, is also shown in Table 3.1. This does provide an improvement, but the the power loss remains large.

3.2.3 Hard boundaries

We now consider an example discussed in Wasserman et al. (2020) and Tse and Davison (2022), which concerns testing for the mixing proportion in a two-component Gaussian mixture model. Suppose we have independent random variables $(Y_1,\ldots,Y_n)$,

Table 3.1 – Probability (%) of positive maximum likelihood estimator of the variance ratio in a variance components model (3.2).

k	$m=5$		$m=10$		$m=20$		$m=30$	
	Usual	REML	Usual	REML	Usual	REML	Usual	REML
5	32.2	43.0	30.3	41.7	29.5	41.1	29.2	40.9
10	37.7	45.5	36.3	44.6	35.6	44.1	35.4	44.0
20	41.4	47.0	40.4	46.3	39.9	45.9	39.7	45.8
30	43.0	47.6	42.1	47.0	41.7	46.7	41.6	46.6
50	44.6	48.1	43.9	47.7	43.6	47.5	43.5	47.4
100	46.2	48.7	45.7	48.4	45.5	48.2	45.4	48.2

where

$$Y_i \sim \tfrac{1}{2}\mathcal{N}_p\left(\mu_1, I_p\right) + \tfrac{1}{2}\mathcal{N}_p\left(\mu_2, I_p\right), \tag{3.5}$$

$\mu_1 = \lambda - \psi 1_p$, $\mu_2 = \lambda + \psi 1_p$, λ is a p-dimensional vector of nuisance parameters and ψ is the scalar parameter of interest. Homogeneity can be imposed on the model by setting $\mu_1 = \mu_2$, or equivalently $\psi = 0$. The squared Euclidean distance between the two component means is

$$\left|\left|\mu_2 - \mu_1\right|\right|_2^2 = \left|\left|2\psi 1_p\right|\right|_2^2 = 4p\psi^2.$$

To account for the increased distance between samples in high-dimensional spaces, we consider the standardized distance $\psi = \delta/(2\sqrt{p})$. This standardized distance helps to balance the effect of increasing dimensionality on the distance between samples; see the two-dimensional illustration in Figure 3.1.

Gaussian mixture models are useful for model-based clustering, as they can effectively capture complex patterns in the data. For a comprehensive overview of Gaussian mixture models and their applications, see Lindsay (1995), McLachlan and Peel (2000), Fraley and Raftery (2002), and Hennig (2010). Testing for homogeneity against a two-component Gaussian mixture has been addressed in many works concentrating on the distribution of the likelihood ratio statistic. Much of the early work centers around the univariate normal mixture. See McLachlan (1987) and Thode et al. (1988) for results on unknown but common variances. For a mixture with different means and variances, results are in McLachlan (1987), Feng and McCulloch (1996), Hathaway (1985). These rely on Monte Carlo simulations and concern finite-sample distributions. Ghosh and Sen (1984) was the first successful attempt to develop an asymptotic distribution of the likelihood ratio for a two-component mixture of arbitrary densities. Hartigan (1985) and Bickel and Chernoff (1993) refined the original proof and showed that the mean parameters for Gaussian mixtures need to be bounded, as otherwise the

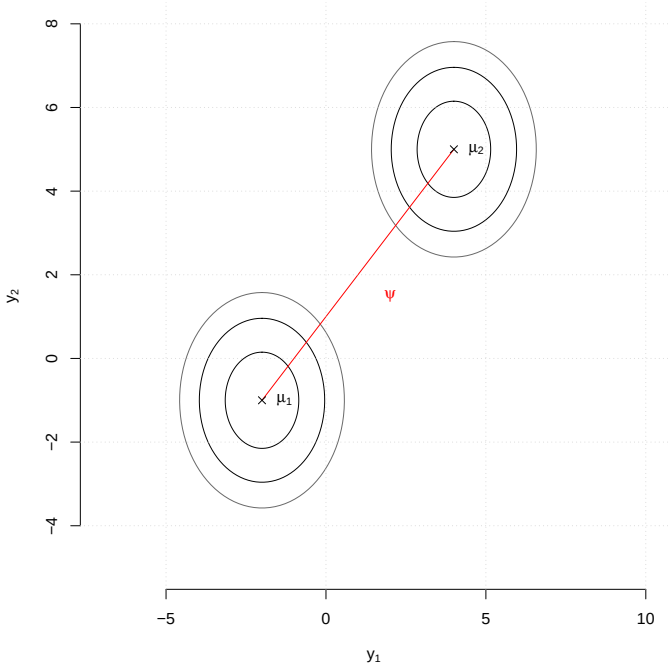


Figure 3.1 – Level plots for two-component Gaussian mixture when $\lambda = (0,1)^{\mathrm{T}}$ and $\psi = 3/2$.

likelihood ratio diverges to infinity.

In Equation (3.5), we consider a multivariate example of a Gaussian mixture model. The key parameter of the mixture model is the scalar ψ, which simplifies the asymptotic distributions under the null hypothesis. However, if the mixture model were to depend on both the means and variances of the populations, the asymptotic distribution of the likelihood ratio would be related to a Gaussian random field, making the computation more complex (Donoho and Jin, 2004).

For the mixture in (3.5), under the null hypothesis $\psi = 0$, the likelihood ratio is

$$\{r(0)\}^2 = 2\left\{\ell\left(y;\widehat{\theta}\right) - \ell\left(y;\widehat{\theta}_{\psi_0}\right)\right\},$$

where $\widehat{\theta}$ is the maximum likelihood estimator of θ, usually obtained via the EM algorithm, and $\widehat{\theta}_{\psi_0} = \left(0, \bar{Y}_0\right)$ is the restricted maximum likelihood estimator of λ under H_0. The probability of a positive estimate of ψ for combinations of p, the dimension of the nuisance parameter and the sample size n, is given in Table 3.2; they are smaller than $1/2$ even for n tending to ∞. So wrong conclusion might be drawn if we test the null hypothesis H_0 using the asymptotic distribution of the likelihood root in (3.3).

To accurately compute the boundary probability in small and moderate-sized samples,

Table 3.2 – Probability (%) of positive maximum likelihood estimate of the distance between the means of two-component Gaussian mixture.

n	$p=2$	$p=5$	$p=10$	$p=20$
30	41.23	42.02	41.60	41.05
50	44.33	43.38	43.67	43.46
100	45.22	45.30	45.92	45.55
200	46.82	46.76	47.76	47.86
500	48.92	47.75	48.05	49.18

for models with soft or hard boundaries, it may be necessary to use an alternative method. The examples provided illustrate some of the difficulties that need to be addressed and solutions will be sought in the following sections.

3.3 Direct improvement on first-order approximations

The profile likelihood is a primary tool for inference on the parameter of interest ψ. However, treating $\ell_{\mathrm{p}}(\psi)$ as an ordinary log likelihood can give poor results, especially if the dimension of the nuisance parameter λ is high and the sample size n is small. The modified likelihood root presented in Chapter 1 alleviates some of these problems. Below, we investigate other methods for improving the finite-sample approximation to the distribution of the likelihood root in (3.3).

3.3.1 Simple solution

Feng and McCulloch (1992) suggest enlarging the parameter space to guarantee that the likelihood ratio maintains the common limiting distribution, but this approach works only when the null hypothesis is uniquely identified. A counterexample for finite mixtures is given by Böhning et al. (1994) in their discussion of Cheng and Traylor (1995). Furthermore, this approach is only applicable to soft boundary problems, and is not suitable for addressing hard boundary problems, where the limit of the statistic must be computed as ψ approaches ψ_0.

Another natural way to improve the approximation is to visualise a “shadow" maximum likelihood estimator $\tilde{\psi}$ that equals the true estimator $\widehat{\psi}$ when the latter is positive, but can take negative values; imagine that $\tilde{\psi} \dot{\sim} \mathcal{N}(\delta, \tau^2)$ and $\widehat{\psi} = \max(\tilde{\psi}, 0)$. Maximum likelihood estimators of on boundaries are typically downwardly biased, so $\delta < 0$ for a shadow estimator, and thus $\Pr(\widehat{\psi} > 0) = \Pr(\tilde{\psi} > 0) = 1 - \Phi(\delta/\tau) < 1/2$, and

correspondingly the score $\partial \ell_{\mathrm{p}}(\psi)/\partial\psi$ on the boundary has a negative bias.

This argument suggests that the p-value $\Phi(-r_{\mathrm{obs}})$ that is computed directly from (3.3) will be too large, thus leading to a loss of power for testing the boundary hypothesis, but also suggests a rough-and-ready remedy when the shadow estimator is approximately normal and its variance is known. This is the case for the shadow likelihood root $\tilde{r}(\psi_0)$, for which $\tau = 1$ to first order. If $\Pr\{\tilde{r}(\psi_0) > 0\} \approx p_+$, then $\delta \approx \Phi^{-1}(p_+)$, and an improved p-value when r_{obs} is positive equals

$$\begin{aligned} \Pr_0\{r(\psi_0) > r_{\mathrm{obs}}\} &= \Pr_0\{\tilde{r}(\psi_0) > r_{\mathrm{obs}}\} \\ &\approx 1 - \Phi(r_{\mathrm{obs}} - \delta) \\ &= \Phi(\delta - r_{\mathrm{obs}}) \\ &\approx \Phi\left\{\Phi^{-1}(p_+) - r_{\mathrm{obs}}\right\}. \end{aligned} \tag{3.6}$$

This approximation will typically be smaller than $\Phi(-r_{\mathrm{obs}})$, because $\Phi^{-1}(p_+) < 0$, but it can only be applied when p_+ can be computed or approximated and presupposes that the shadow estimator has an approximately a normal distribution.

3.3.2 Profile score

Below, we explore two ways of improving the finite-sample approximation of (3.3) by using the distribution of the profile score under the null hypothesis. McCullagh and Tibshirani (1990) suggest using an adjusted profile likelihood to address some of the inherent issues of the profile likelihood, such as its bias and overly optimistic variance estimates.

Under mild conditions, the log likelihood will have a local maximum on the boundary if the profile log likelihood $\ell_{\mathrm{p}}(\psi)$ for ψ has a negative gradient there; put another way, $\widehat{\psi} > \psi_0$ if and only if $\partial \ell_{\mathrm{p}}(\psi)/\partial\psi > 0$ when evaluated at $\psi = \psi_0$, see Figure 3.2. This suggests that the profile score is a natural starting point for the finite-sample approximation

$$p_+ = \Pr_0(\widehat{\psi} > \psi_0) = \Pr_0\left\{ \left.\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right|_{\psi=\psi_0} > 0 \right\}. \tag{3.7}$$

Unlike the distribution of the maximum likelihood estimator, that of the profile score statistic is continuous, without a point mass at the origin. McCullagh and Tibshirani (1990) studied the behavior of the profile score statistic in finite samples, and proposed an adjustment for it at each parameter value. This correction was later further studied and extended to semiparametric models (Kauermann, 2002; Bellio et al., 2008). Below,

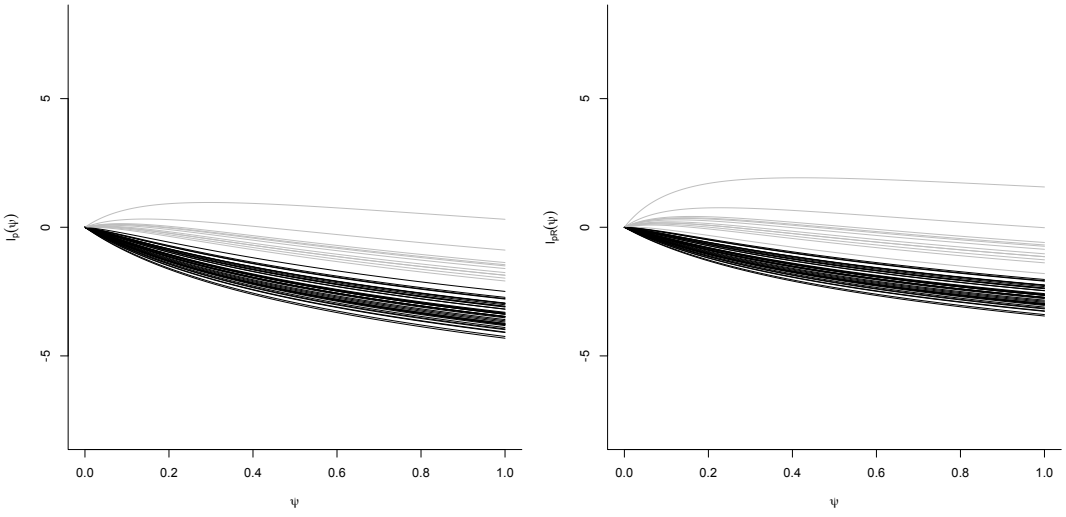


Figure 3.2 – Profile log likelihoods for data simulated from a one-way classification model (3.2) where $k = 5$ and $m = 5$; Left panel: the ordinary profile log likelihood. Right panel: the restricted profile log likelihood. Grey curves have a positive gradient at the origin, i.e., $\widehat{\psi} > 0$, and black curves correspond to likelihood functions maximised at the origin.

we describe the original version in more detail.

Let us write the derivatives of the log likelihood ℓ with respect to ψ and the components $\lambda_i, \lambda_j, \ldots$, of λ as

$$U_\psi = \frac{\partial \ell}{\partial \psi}, \quad U_i = \frac{\partial \ell}{\partial \lambda_i}, \quad U_{ij} = \frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_j}, \quad U_{\psi i} = \frac{\partial^2 \ell}{\partial \psi \partial \lambda_i},$$

and so fourth. Note that U_ψ and U_i are $O_\mathrm{p}(n^{1/2})$, whereas $U_{\psi i}$ and U_{ij} are in general $O_\mathrm{p}(n)$. Using Taylor series expansion of the profile score, McCullagh and Tibshirani (1990) show that

$$\begin{aligned}\frac{\partial \ell_\mathrm{p}}{\partial \psi} = U_\psi + U_{\psi i}\left\{\kappa^{i,j}U_j + \kappa^{i,j}\kappa^{k,l}\left(U_{jk} - \kappa_{jk}\right)U_l + \frac{1}{2}\kappa^{ijk}U_kU_l\right\} \\ + \frac{1}{2}U_{\psi ij}\kappa^{i,k}\kappa^{j,l}U_kU_l + O_\mathrm{p}\left(n^{-1/2}\right),\end{aligned} \tag{3.8}$$

where Einstein's summation convention implies that we sum over repeated sub- and superscripts (McCullagh, 1987), the indices i, j, k, l refer to components of the nuisance parameter λ, $\kappa_{i,j}$ is the (i, j) component of the Fisher information matrix for λ, $\kappa^{i,j}$ is a component of the corresponding inverse matrix, and $\kappa_{\psi,i}$ is a component of the Fisher information corresponding to ψ and λ_i and $\kappa^{ijk} = \kappa^{i,i'}\kappa^{j,j'}\kappa^{k,k'}\kappa_{i'j'k'}$.

The expectation of (3.8) is

$$\mathrm{E}\left\{\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right\} = -\tfrac{1}{2}(\kappa_{\psi,i,j} - \kappa_{\psi,k}\kappa^{k,l}\kappa_{l,i,j})\kappa^{i,j} - \tfrac{1}{2}(\kappa_{\psi,ij} - \kappa_{\psi,k}\kappa^{k,l}\kappa_{l,ij})\kappa^{i,j} + O(n^{-1/2}), \tag{3.9}$$

where

$$\kappa_{\psi,i,j} = \mathrm{E}\{U_\psi U_i U_j\}, \quad \kappa_{\psi,ij} = \mathrm{E}\{U_\psi U_{ij}\}.$$

The terms $\kappa_{i,j}$, $\kappa_{\psi,i,j}$ and so on are $O(n)$, and $\kappa^{i,j}$ is $O(n^{-1})$, so the expected profile score is $O(1)$. Moreover a standard computation gives

$$\mathrm{var}\left\{\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right\} = \kappa_{\psi,\psi} - \kappa_{\psi,r}\kappa_{s,\psi}\kappa^{r,s} + O(n^{1/2}), \tag{3.10}$$

The approximate moments can be used in a normal approximation of (3.7), i.e., without adjustment we have

$$\mathrm{Pr}_0\left\{\left.\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right|_{\psi=\psi_0} > 0\right\} \doteq 1 - \Phi(0) = \frac{1}{2}.$$

Bias adjustment gives

$$\begin{aligned} p_+^{\mathrm{a}} =& \mathrm{Pr}_0\left\{\left.\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right|_{\psi=\psi_0} > 0\right\} \\ \doteq& 1 - \Phi\left(-\frac{\mu_{\mathrm{p}}}{\sigma_{\mathrm{p}}}\right) \\ =& \Phi\left(\frac{\mu_{\mathrm{p}}}{\sigma_{\mathrm{p}}}\right), \end{aligned}$$

where

$$\mu_{\mathrm{p}} \doteq \mathrm{E}_0\left(\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right), \quad \sigma_{\mathrm{p}}^2 \doteq \mathrm{var}_0\left(\frac{\partial \ell_{\mathrm{p}}(\psi)}{\partial \psi}\right).$$

It is common for the profile score to have a negative bias on the boundary. This suggests that μ_{p} is likely to be negative and p_+^{a} is expected to be smaller than the asymptotic value of 1/2. Later, we will compare p_+^{a} to the probability obtained via numerical simulations to see how well the two agree.

3.3.3 Edgeworth expansion

The asymptotic normal distribution resulting from the central limit theorem is the foundation of many statistical approximations. Edgeworth expansion is another

classical technique that provides an expansion of the distribution of standardized sum offering corrections, usually of order $O(n^{-1/2})$ and $O(n^{-1})$. Unlike the saddlepoint approximation, the Edgeworth series requires only the first few cumulants and these can often be computed without knowing the generating function.

Let U_n^* denote the standardized version of $U_n = \sum_i^n U_i$, where $U_1, \ldots, U_n$ are independent replicates of a continuous random variable with finite cumulants κ_r, and standardized cumulants $\rho_r = \kappa_r / \kappa_2^{r/2}$. The Edgeworth expansion for the distribution of $U_n^* = \sum_{i=1}^{n} (U_i - \kappa_1)/\kappa_2^{1/2}$ is

$$F_n^*(u) = \Phi(u) - \varphi(u) \left[\frac{\rho_3}{6n^{1/2}} H_2(u) + \frac{1}{n} \left\{ \frac{\rho_4}{24} H_3(u) + \frac{\rho_3^2}{72} H_5(u) \right\} + O\left(n^{-3/2}\right) \right], \quad (3.11)$$

where $\varphi(\cdot)$ is the standard normal density and H_r denotes the rth-order Hermite polynomial, given by

$$H_1(u) = u, \quad H_2(u) = u^2 - 1, \quad H_3(u) = u^3 - 3u,$$
$$H_4(u) = u^4 - 6u^2 + 3, \quad H_5(u) = u^5 - 10u^3 + 15u.$$

The leading term of the Edgeworth expansion (3.11) gives the standard normal approximation for U_n^*. Terms beyond the normal approximation can be expressed as the product of the normal density, Hermite polynomials in u, and the skewness and kurtosis of U_n^* (Pace and Salvan, 1997).

The Edgeworth expansion has been extensively used for theoretical work, for example by Hall (1987, 1991, 1992), Barndorff-Nielsen and Cox (1979), and van der Vaart (1998). For a discussion of Edgeworth series in the discrete case, see Esseen (1945) and Kolassa and McCullagh (1990). Applications are widely reported in Ferrari et al. (1997, 2001), and Gerlovina et al. (2017).

Below, we consider an expansion for the distribution of the profile score with a correction that involves the skewness. The truncation after ρ_3 provides an approximate distribution F_n^* with a remainder of order $O(n^{-1})$. To obtain the third cumulant, we start by noting that the profile score as written in (3.8) is the sum of

$$\frac{\partial \ell_{\mathrm{p}}}{\partial \psi} = A_n + B_n + O_{\mathrm{p}}\left(n^{-1/2}\right),$$

where

$$A_n = U_\psi + U_{\psi i}\kappa^{i,j}U_j = O_{\mathrm{p}}\left(n^{1/2}\right),$$
$$B_n = U_{\psi i}\left\{\kappa^{i,j}\kappa^{k,l}\left(U_{jk}-\kappa_{jk}\right)U_l + \tfrac{1}{2}\kappa^{ijk}U_jU_k\right\} + \tfrac{1}{2}U_{\psi ij}\kappa^{i,k}\kappa^{j,l}U_kU_1 = O_{\mathrm{p}}(1).$$

Rewriting A_n and B_n in terms of the centred variables $U_{ij}-\kappa_{ij}$, and $U_{ijk}-\kappa_{ijk}$ gives

$$\begin{aligned}A_n &\doteq U_\psi + \kappa_{\psi i}\kappa^{i,j}U_j,\\ B_n &\doteq \left(U_{\psi i}-\kappa_{\psi i}\right)k^{i,j}U_j + \kappa_{\psi i}\left\{\kappa^{i,j}\kappa^{k,l}\left(U_{jk}-\kappa_{jk}\right)U_l + \tfrac{1}{2}\kappa^{ijk}U_jU_k\right\}\\ &\quad + \tfrac{1}{2}\kappa_{\psi ij}\kappa^{i,k}\kappa^{j,l}U_kU_l.\end{aligned}$$

We then have

$$\left(\frac{\partial \ell_{\mathrm{p}}}{\partial \psi}\right)^3 = A_n^3 + 3A_n^2B_n + O_{\mathrm{p}}\left(n^{1/2}\right). \tag{3.12}$$

Expanding each term in (3.12) and applying the expectation, we have

$$\mathrm{E}\left(A_n^3\right) = \kappa_{\psi,\psi,\psi} + 3\kappa_{\psi i}\kappa_{\psi,\psi,j}\kappa^{i,j} + 3\kappa_{\psi i}\kappa_{\psi k}\kappa^{i,j}\kappa^{k,l}\kappa_{\psi,j,l} + \kappa_{\psi i}\kappa_{\psi j}\kappa_{\psi k}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{r,s,t},$$

and

$$\begin{aligned}\mathrm{E}\left(A_2^2B_n\right) =& \kappa_{\psi,\psi,\psi k,l}\kappa^{k,l} + \kappa_{\psi,\psi,rn,p}\kappa_{\psi m}\kappa^{m,r}\kappa^{n,p}\\ &+ \tfrac{1}{2}\kappa_{\psi,\psi,r,n}\kappa_{\psi m}\kappa^{mrn} + \tfrac{1}{2}\kappa_{\psi,\psi,m,r}\kappa_{\psi kl}\kappa^{k,m}\kappa^{l,r}\\ &+ 2\kappa_{\psi,\psi k,j,l}\kappa_{\psi i}\kappa^{i,j}\kappa^{k,l} + 2\kappa_{\psi,j,rn,p}\kappa_{\psi i}\kappa_{\psi m}\kappa^{i,j}\kappa^{m,r}\kappa^{n,p}\\ &+ \kappa_{\psi,j,r,n}\kappa_{\psi i}\kappa_{\psi m}\kappa^{i,j}\kappa^{mrn} + \kappa_{\psi,j,m,r}\kappa_{\psi i}\kappa_{\psi kl}\kappa^{i,j}\kappa^{k,m}\kappa^{l,r}\\ &+ \kappa_{\psi k,n,z,l}\kappa_{\psi i}\kappa_{\psi s}\kappa^{i,n}\kappa^{s,z}\kappa^{k,l} + \kappa_{n,z,rt,p}\kappa_{\psi i}\kappa_{\psi s}\kappa_{\psi m}\kappa^{i,n}\kappa^{s,z}\kappa^{m,r}\kappa^{t,p}\\ &+ \tfrac{1}{2}\kappa_{n,z,r,p}\kappa_{\psi i}\kappa_{\psi s}\kappa_{\psi m}\kappa^{i,n}\kappa^{s,z}\kappa^{mrp} + \tfrac{1}{2}\kappa_{n,z,m,r}\kappa_{\psi i}\kappa_{\psi s}\kappa_{\psi kl}\kappa^{i,n}\kappa^{s,z}\kappa^{k,m}\kappa^{l,r}.\end{aligned}$$

Terms appearing in the third-order moment of the score are functions of x centered variables of orders $O(n^{x/2})$ if x is even and $O(n^{(x-1)/2})$ if x is odd, so $\mathrm{E}\left(A_n^3\right)$ is $O(n)$ whereas $\mathrm{E}\left(A_2^2B_n\right) = O(1)$; see Pace and Salvan (1997, Chapter 3). We then use the following relations between central moments and cumulants, where we ignore the contributions of order $O(n)$.

$$\begin{aligned}\mathrm{E}\left(U_iU_jU_kH_{uv}\right) &= \kappa_{i,uv}\kappa_{j,k}[3] + O(n),\\ \mathrm{E}\left(U_iU_jU_kU_l\right) &= \kappa_{i,j}\kappa_{k,l}[3] + O(n).\end{aligned}$$

[3] refers to the sum over permutations of the indices i, j and k, and the moments

satisfy the second Bartlett identity (Bartlett, 1953).

The third-order cumulant of the profile score is

$$\begin{aligned}\kappa_3 &\approx \mathrm{E}\left\{\left(U_\mathrm{p}\right)^3\right\} - 3\,\mathrm{E}\left(U_\mathrm{p}\right)\mathrm{E}\left\{\left(U_\mathrm{p}\right)^2\right\} \\ &= \mathrm{E}\left\{\left(U_\mathrm{p}\right)^3\right\} - 3\,\mathrm{E}(U_\mathrm{p})\left[\mathrm{var}\left(U_\mathrm{p}\right) + \mathrm{E}\left\{\left(U_\mathrm{p}\right)^2\right\}\right],\end{aligned}$$

Moments in this expression are given in (3.9), (3.10) and the expectation of (3.12). κ_3 is then standardized by $\kappa_2^{3/2}$, and serves as a correction term of order $O(n^{-1/2})$ in (3.11).

The probability of a positive gradient at the origin obtained using the Edgeworth expansion is for the distribution of the profile score is then

$$\begin{aligned}p_+^\mathrm{e} =& \mathrm{Pr}_0\left\{\left.\frac{\partial \ell_\mathrm{p}(\psi)}{\partial \psi}\right|_{\psi=\psi_0} > 0\right\} \\ \doteq& 1 - F_n^*\left(-\frac{\mu_\mathrm{p}}{\sigma_\mathrm{p}}\right),\end{aligned}$$

where F_n^* is in (3.11).

To summarize, we defined two probabilities: p_+^a and p_+^e using the distribution of the profile score and Edgeworth expansion for its cumulative distribution function. These probabilities are used to approximate (3.7) and will be compared to simulation-based probabilities and the asymptotic value of 1/2 in the next section.

3.4 Applications

In Section 3.2, we distinguished between soft and hard boundaries. This section illustrates the limitations of standard methods through examples falling under these two varieties of boundary problems, and explores the improved approximations proposed in Section 3.3.

3.4.1 Soft boundaries

Mixed effects models

The variance components example presented in Section 3.2.2 is paradigmatic of several widely-used models that can be brought under the single umbrella of linear

mixed models. These are models of the form

$$y = X\beta + Zb + \varepsilon, \tag{3.13}$$

where the response vector y is n-dimensional, β is a p-dimensional vector of fixed-effect parameters, b is a k-dimensional vector of random-effect coefficients with a known symmetric positive definite variance matrix $\sigma_b^2\Sigma$ and ε is an n-dimensional vector of uncorrelated random errors with variance σ^2. The $n\times p$ matrix X and the $n\times k$ matrix Z indicate how y depends on the fixed parameters β and the random variables b. If (b,ε) has a normal distribution with mean zero and the given covariance matrices, the marginal distribution of Y is normal with $\mathrm{E}(Y) = X\beta$ and $\mathrm{cov}(Y) = \sigma^2\Delta(\psi)^{-1}$, where $\psi = \sigma_b^2/\sigma^2$, and $\Delta(\psi)^{-1} = \psi ZZ^{\mathrm{T}} + I_n$. For this model, a test on the boundary $H_0 : \psi = \psi_0 = 0$ corresponds to testing for constant means $b_1 = \cdots = b_k = 0$.

The log likelihood based on the marginal distribution of y is

$$\ell(\psi,\beta,\sigma^2) \equiv -\frac{1}{2}\left\{n\log\sigma^2 - \log|\Delta(\psi)| + \frac{1}{\sigma^2}(y-X\beta)^{\mathrm{T}}\Delta(\psi)(y-X\beta)\right\}, \quad \psi \geq 0. \tag{3.14}$$

For fixed ψ, the maximum likelihood estimates of σ^2 and β are

$$\widehat{\sigma}_\psi^2 = \frac{1}{n}\left(y - X\widehat{\beta}_\psi\right)^{\mathrm{T}}\Delta(\psi)\left(y - X\widehat{\beta}_\psi\right), \quad \widehat{\beta}_\psi = \left\{X^{\mathrm{T}}\Delta(\psi)X\right\}^{-1}X^{\mathrm{T}}\Delta(\psi)y, \tag{3.15}$$

and the profile log likelihood for the variance ratio ψ is

$$\ell_{\mathrm{p}}(\psi) \equiv -\frac{1}{2}\left\{n\log\widehat{\sigma}_\psi^2 - \log|\Delta(\psi)|\right\}, \quad \psi \geq 0\,. \tag{3.16}$$

Under the null hypothesis H_0, the probability that $\widehat{\psi} > 0$ corresponds to a positive log likelihood gradient at $\psi = \psi_0$, which is

$$\Pr_0\left\{\left.\frac{\partial\ell_{\mathrm{p}}(\psi)}{\partial\psi}\right|_{\psi=\psi_0} > 0\right\} = \Pr_0\left(e^{\mathrm{T}}Qe > 0\right), \tag{3.17}$$

where e has a standard multivariate normal distribution, $Q = n(I-H)ZZ^{\mathrm{T}}(I-H) - \mathrm{tr}\left(Z^{\mathrm{T}}Z\right)(I-H)$, and $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. If $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of the matrix Q, (3.17) equals

$$\Pr_0\left(\sum_{i=1}^{n}\lambda_i e_i^2 > 0\right).$$

For the linear mixed effects model described in (3.13), the log restricted likelihood

(Harville, 1977) is

$$\ell_{\mathrm{R}}(\psi,\sigma^2) \equiv -\frac{1}{2}\left\{\ell(\psi,\widehat{\beta}_\psi,\sigma^2) - p\log\sigma^2 + \log|X^{\mathrm{T}}\Delta(\psi)X|\right\}. \tag{3.18}$$

where $\widehat{\beta}_\psi$ is defined in (3.15). The probability of positive estimates using the profile log restricted likelihood is

$$\Pr_0\left\{\left.\frac{\partial\ell_{\mathrm{R}}\left(\psi,\widehat{\sigma}^2_{\psi,\mathrm{R}}\right)}{\partial\psi}\right|_{\psi=\psi_0} > 0\right\} = \Pr_0\left(e^{\mathrm{T}}Q_{\mathrm{R}}e > 0\right), \tag{3.19}$$

where $\widehat{\sigma}^2_{\psi,\mathrm{R}} = (y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Delta(\psi)(y - X\widehat{\beta}_\psi)/(n-p)$ is the restricted likelihood estimator of σ^2, and $Q_{\mathrm{R}} = (n-p)(I-H)ZZ^{\mathrm{T}}(I-H) - \mathrm{tr}\left\{ZZ^{\mathrm{T}}(I-H)\right\}$. Details of the partial log likelihood derivatives in (3.17) and (3.19) are provided in Appendix 3.8.1.

One way to compute (3.17) is to use the saddlepoint approximation (Barndorff-Nielsen and Cox, 1979)

$$\Pr\left(e^{\mathrm{T}}Qe > q\right) = 1 - F(q) \simeq \begin{cases} 1 - \Phi\left\{w + \dfrac{1}{w}\log\left(\dfrac{v}{w}\right)\right\}, & q \neq \mathrm{E}\left\{Q(e)\right\}, \\ \dfrac{1}{2} - \dfrac{\kappa'''(0)}{6\sqrt{2\pi}\kappa''(0)^{3/2}}, & q = E\left\{Q(e)\right\}, \end{cases} \tag{3.20}$$

where $\kappa(\cdot)$ is the cumulant generating function of $Q(e) = e^{\mathrm{T}}Qe$, i.e.,

$$\kappa(\xi) = -\frac{1}{2}\sum_{i=1}^{n}\log(1-2\xi\lambda_i), \quad \xi < \frac{1}{2\min\limits_{i=1,\ldots,n} 2\lambda_i},$$

and

$$w = \mathrm{sign}(\widehat{\xi})\left[2\left\{\widehat{\xi}q - \kappa\left(\widehat{\xi}\right)\right\}\right]^{1/2}, \quad v = \widehat{\xi}\left\{\kappa''(\widehat{\xi})\right\}^{1/2};$$

where $\widehat{\xi}$ is the saddlepoint.

The saddlepoint approximation, when applied to the models in (3.13), provides a efficient way to calculate the boundary probability. Kuonen (1999) demonstrated that using (3.20) is comparable in speed to exact methods, almost as accurate, and much easier to implement. Another approximation for the distribution of quadratic forms in normal variates was first suggested by Pearson (1959) and later by Imhof (1961)

$$\mathrm{pr}\left(e^{\mathrm{T}}Qe > q\right) \simeq \mathrm{pr}\left(\chi^2_b > r\right),$$

where χ^2_b denotes a chi-squared variable with $b = c_2^3/c_3^2$ degrees of freedom, $r = q - c_1(b/c_2)^{1/2} + b$, and $c_s = \mathrm{tr}(Q^s)$, for $s = 1,2,3$. If $e^{\mathrm{T}}Qe$ is non-positive, the same

approximation holds but one must assume $c_3 > 0$.

Penalized splines

Nonparametric regression using natural cubic splines is equivalent to a particular linear mixed effects model, so testing for a linear regression versus a general smooth alternative can be viewed as testing for a zero variance component.

Consider a set of points $(x_1, y_1), \ldots, (x_n, y_n)$, where $a < x_1 < \cdots < x_n < b$ and assume that

$$y_i = \mu(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{3.21}$$

where μ is differentiable on $[a, b]$ with absolutely continuous first derivative μ'. To choose μ to balance fidelity to the data and smoothness, we take μ to minimize the penalized sum of squares

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 + \frac{1}{\psi} \int_a^b \mu''(x)^2 \, \mathrm{d}x,$$

where $\psi > 0$ is a dimensionless parameter, usually denoted $1/\lambda$ in the literature. The integral is a roughness penalty on μ''. When $\psi \to \infty$, no penalty is applied, and there are n degrees of freedom, corresponding to the unconstrained variation of each element of the vector μ. As $\psi \to 0$, the penalty becomes so large that $\mu(x)$ is forced to become a straight line, i.e., a curve with two degrees of freedom. Intermediate values of ψ give curves lying between these two extremes. Green and Silverman (1994) show that the resulting μ is a natural cubic spline and that the penalty can be written as

$$\int \mu''(x)^2 \, \mathrm{d}x = \mu^{\mathrm{T}} K \mu,$$

where $\mu^{\mathrm{T}} = (\mu(x_1), \ldots, \mu(x_n))$ and K is an $n \times n$ matrix of rank $n-2$. The kernel of K is spanned by $\{1_n, x\}$ and the penalty matrix can be written as $K = ADA^{\mathrm{T}}$, where $A = (a_1, \ldots, a_n)$ is an orthogonal matrix, the columns of which are the eigenvectors of K, while D is a diagonal matrix of the corresponding eigenvalues $d_1 = d_2 = 0 < d_3 \leq \cdots \leq d_n$. The eigenvalue decomposition of K implies that

$$K = \sum_{j=1}^{n} d_j a_j a_j^{\mathrm{T}}, \quad K_+^{-1} = \sum_{j=3}^{n} d_j^{-1} a_j a_j^{\mathrm{T}}, \quad K K_+^{-1} = \operatorname{diag}(0, 0, I_{n-2}).$$

Then we can write (3.21) as

$$y = \mu + \varepsilon = X\beta + Zb + \varepsilon, \tag{3.22}$$

where $X_{n\times 2} = (a_1, a_2)$, $Z_{n\times(n-2)} = (d_3^{-1/2}a_3, \ldots, d_n^{-1/2}a_n)$, $b \sim \mathcal{N}_{n-2}(0, \sigma_b^2 I)$ independent of $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. The distribution of μ is $\mathcal{N}_n(X\beta, \psi\sigma^2 ZZ^{\mathrm{T}})$, where $\psi = \sigma_b^2/\sigma^2$. The penalty equals

$$\mu^{\mathrm{T}} K \mu = (X\beta + Zb)^{\mathrm{T}} K (X\beta + Zb) = b^{\mathrm{T}} \begin{pmatrix} 0_{2\times(n-2)} \\ D_+^{-1/2} \end{pmatrix}^{\mathrm{T}} A^{\mathrm{T}} A D A^{\mathrm{T}} A \begin{pmatrix} 0_{2\times(n-2)} \\ D_+^{-1/2} \end{pmatrix} b = b^{\mathrm{T}} b.$$

The usual and the restricted log likelihoods for this model are given in equations (3.14) and (3.18) respectively. So testing for a linear fit against a smooth curve is equivalent to testing $H_0 : \psi = 0$ ($\sigma_b^2 = 0$) against $H_A : \psi > 0$ ($\sigma_b^2 > 0$).

Generalized Pareto distribution

The generalized Pareto (GP) distribution function with scale $\sigma > 0$ and shape $\xi \in \mathbb{R}$ is

$$G(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x}{\sigma}\right)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right), & \xi = 0, \end{cases} \tag{3.23}$$

where $x_+ = \max(0, x)$. A generalized Pareto random variable has support $[0, -\sigma/\xi)$ if $\xi < 0$ and $\mathbb{R}_+$ otherwise. The distribution in (3.23) has three basic shapes, corresponding to a limiting distribution of exceedances from different classes of underlying distributions (Embrechts et al., 1997).

Consider a null hypothesis $H_0 : \xi = \xi_0 = 0$, under which we test whether the data have an exponential distribution. The distribution under the null can be regarded as an infinite mixture of exponential variables. Assume that a random variable X is exponentially distributed with rate $\lambda \sim \Gamma(\nu, 1/s)$. Then, the marginal distribution of X is GP$\{\xi = 1/\nu, \sigma = 1/(\nu s)\}$, because

$$\begin{aligned} \Pr(X > x) &= \int_0^{+\infty} \exp(-\lambda x) \frac{\lambda^{\nu-1} \exp\left(\frac{-\lambda}{s}\right)}{\Gamma(\nu) s^\nu} \mathrm{d}\lambda \\ &= \frac{1}{\Gamma(\nu) s^\nu} \int_0^{+\infty} \lambda^{\nu-1} \exp\left\{-\lambda\left(x + \frac{1}{s}\right)\right\} \mathrm{d}\lambda \\ &= (1 + sx)^{-\nu}, \quad x > 0, \quad \sigma, \nu > 0. \end{aligned}$$

Table 3.3 shows probabilities of positive estimator for the shape parameter based on 10^4 samples of exponential variables for different sample sizes. The simulations suggest that the probability of positive estimates, denoted p_+^n, is smaller than 1/2.

Table 3.3 – Probability (%) of positive maximum likelihood estimator of the shape parameter in GP($\xi = 0, \sigma$) described in (3.23). The probabilities p_+^n, p_+^{a} and p_+^{e} are obtained using (i) 10^4 simulations from GP($\xi = 0, \sigma$), (ii) the distribution of the profile score, and (iii) Edgeworth expansion for the distribution of the profile score, respectively.

prob	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$	$n = 200$
p_+^n	32.9	36.0	38.3	40.5	40.5	43.1
p_+^{a}	41.1	43.7	44.8	45.5	46.0	47.2
p_+^{e}	41.0	43.6	44.8	45.5	46.0	47.2

Similar work was pursued by Hosking (1984), who considered whether the shape parameter is zero in the generalized extreme value distribution. This is equivalent to testing whether the data follow a Gumbel distribution rather than a type II or III generalized extreme value distribution. Results therein show poor agreement between finite-sample and asymptotic distributions and authors do not recommend using the likelihood ratio or Wald statistics for hypothesis testing.

The moments of the profile score for the GP distribution are $\mu_{\mathrm{p}} = -1$, $\sigma_{\mathrm{p}}^2 = n$ and $\kappa_3 = 13n$; see Appendix 3.8.2 for details. Table 3.3 shows that the resulting adjusted probability p_+^a is closer to the simulated 'true' probability p_+^n than to the asymptotic value of 1/2, though it is still appreciably larger than p_+^n. The further adjustment provided by the Edgeworth expansion goes in the right direction but p_+^e is essentially equal to p_+^a.

3.4.2 Hard boundaries: Mixture models

Infinite mixtures

One model for heavy-tailed data is the Student t distribution, which we write as in Davison (2003, §4.6) , as

$$f(y;\psi,\mu,\sigma) = \frac{\Gamma\{(\psi^{-1}+1)/2\}\psi^{1/2}}{(\sigma^2\pi)^{1/2}\Gamma\{1/(2\psi)\}}\left\{1+\psi(y-\mu)^2/\sigma^2\right\}^{-(\psi^{-1}+1)/2}, \tag{3.24}$$

where $\psi, \sigma > 0$ and $-\infty < \mu, y < \infty$. This generalizes the Student t density with $\psi^{-1} = \nu$ degrees of freedom to continuous ψ, with $\psi = \psi_0 = 0$ corresponding to the normal density. It furthermore allows us to interpret the Student t distribution as an infinite mixture of normal variates. Let $X \sim N(0, s^2)$ be a centered Gaussian variable with

Table 3.4 – Probability (%) of positive maximum likelihood estimator of the reciprocal of the degrees of freedom in a Student t density described in (3.24). The probabilities p_+^n, p_+^{a} and p_+^{e} are obtained using (i) 10^4 standard normal samples, (ii) the distribution of the profile score, and (iii) Edgeworth expansion for the distribution of the profile score, respectively.

prob	$n=10$	$n=20$	$n=50$	$n=100$	$n=200$	$n=500$	$n=1000$
p_+^n	21.1	26.3	33.5	36.3	39.7	42.7	44.0
p_+^{a}	34.9	39.2	43.1	45.1	46.5	47.8	48.4
p_+^{e}	33.6	38.8	43.0	45.1	46.5	47.8	48.4

variance s^2. If $s^{-2} \sim \Gamma(\nu/2, \nu/2)$, then the marginal distribution of X is Student t with ν degrees of freedom. So testing for Gaussianity of the data can also be viewed as testing for a zero variance component.

Simulations based on 10^4 normal samples show that the probability of positive $r(\psi_0)$ for a sample of size n is smaller than 1/2, as shown in Table 3.4. For $n = 10$, only 21% of the estimates are positive. Large-sample results are still unreliable since the probability of a point mass at zero is more than 50% even for $n > 200$. These results suggest that the null tail probability is substantially over-estimated, and the likelihood root rejects H_0 too rarely especially for small n.

In this example, tedious calculations involving moments of the Gaussian distribution (see Appendix 3.8.2), yield $\mu_{\mathrm{p}} = -3/2$, $\sigma^2_{\mathrm{p}} = 3n/2$ and $\kappa_3 = 369n/8$ for the profile score. The corrected probabilities are overall closer to p_+^n. Although the correction goes in the right direction, it is not wholly efficacious.

Testing for overdispersion

Suppose that count data Y follow a Poisson distribution but the rate has a gamma distribution with mean ξ and shape parameter ν. The resulting density is the negative binomial,

$$f(y;\theta) = \frac{\Gamma(\nu+y)}{\Gamma(\nu)\, y!} \frac{\nu^\nu \xi^y}{(\nu+\xi)^{\nu+y}}, \quad y = 0, 1, \ldots, \xi, \nu > 0. \tag{3.25}$$

The variance of a negative binomial random variable is $\xi + \xi^2/\nu$, so ξ^2/ν is the additional variance compared to that of a Poisson variable with mean ξ. The overdispersion is controlled by $1/\nu$, scaled by the square of the mean ξ^2. Under the above model, we consider the null hypothesis $H_0 : \psi = 1/\nu = 0$ against the alternative that $\psi > 0$, to test

Table 3.5 – Probability (%) of positive maximum likelihood estimator of the dispersion parameter in a negative binomial distribution defined in (3.25). The probabilities p_+^n, p_+^{a} and p_+^{e} are obtained using (i) 10^4 simulations from Poisson variables, (ii) the distribution of the profile score, and (iii) Edgeworth expansion for the distribution of the profile score, respectively.

prob	$n=10$	$n=20$	$n=50$	$n=100$	$n=200$
p_+^n	33.5	37.1	41.5	44.3	45.9
p_+^{a}	41.1	43.7	46.0	47.2	48.0
p_+^{e}	40.9	43.6	46.0	47.2	48.0

whether the data are consistent with a Poisson distribution. Table 3.5 gives results based on 10^4 Poisson samples with rate ξ and size n. The probabilities are consistent with the previous findings, as all frequencies are smaller than $1/2$.

The profile score in this example has the following moments $\mu_{\mathrm{p}} = -\xi/2$, and $\sigma_{\mathrm{p}}^2 = n\xi^2/2$; see Appendix 3.8.2. The corrected probabilities are close to p_+^n but are not by any means perfect. Given the extra effort required to compute the third-order cumulant for the profile score, computing p_+^{e} does not seem worthwhile.

3.5 Results for the tangent exponential model

The tangent exponential model derivation for the examples discussed in Section 3.4 is outlined in detail in Appendix 3.8.3. Additional computational information relevant to these examples can be found in Appendices 3.8.4 and 3.8.5. While the technical details of the derivations have been included in the appendices to allow for a greater focus on the interpretation of the results, it is important to note that these derivations were the tools we used to obtain the results below.

3.5.1 Example: Variance components

For the one-way classification example in (3.2), th ordinary profile likelihood has $\mu_{\mathrm{p}} = -(m-1)/2$ and $\sigma_{\mathrm{p}}^2 = m(m-1)k/2$, see Appendix 3.8.2 for details of the usual and restricted profile scores. We have

$$p_+^{\mathrm{a}} = \Phi\left(-\sqrt{\frac{m-1}{2mk}}\right), \quad m,k>1.$$

Table 3.6 – Probability (%) of non-zero estimates in variance components model (3.2) using the distribution of the profile score, and Edgeworth expansion for the distribution of the profile score, denoted as p_+^{a} and p_+^{e}, respectively.

k	$m=5$		$m=10$		$m=20$		$m=30$	
	p_+^{a}	p_+^{e}	p_+^{a}	p_+^{e}	p_+^{a}	p_+^{e}	p_+^{a}	p_+^{e}
5	38.86	38.85	38.29	38.20	37.90	37.89	37.80	37.79
10	42.07	42.07	41.60	41.60	41.37	41.37	41.30	41.30
20	44.37	44.37	44.03	44.03	43.87	43.87	43.82	43.82
30	45.40	45.40	45.12	45.12	45.00	45.00	44.95	44.95
50	46.43	46.43	46.22	46.22	46.11	46.11	46.08	46.08
100	47.47	47.47	47.32	47.32	47.25	47.25	47.22	47.22

which converges to the asymptotic value $1/2$ for large k, $(m-1)/m \approx 1$ unless m is small. The corresponding probabilities, shown in Table 3.6, show an improvement compared to the column “Usual" of Table 3.1 obtained using the beta distribution due to the bias correction. As previously noted, the Edgeworth expansion does not offer a significant improvement relative to the amount of effort required. While adding the kurtosis term to the expansion may improve the probability, it is more practical to use simulations for these examples.

The one-way classification model is a soft boundary problem. We consider ψ such that $\psi > -1/m$ as an extended parameter space for numerical purposes but restrict the interpretation of the results to $\psi \geq 0$. For this particular example, closed-form expressions are available for the maximum likelihood estimates $\widehat{\theta}, \widehat{\theta}_\psi$ and the pivots $r(\psi)$, $q(\psi)$. So to obtain the pivot $r^*(\psi)$, no numerical optimization or differentiation is needed; see expressions (3.29–3.30) and further details in Appendix 3.8.3. Table 3.7 gives probabilities of positive $\widehat{\psi}$ and $\widehat{\psi}^*$ based on 10^4 simulated observations using the ordinary and the restricted likelihoods. Results for r^* are much better than those for r, as all probabilities are closer to the asymptotic value of $1/2$, even for small values of k and m.

In addition to the incorrect use of the asymptotic mixing probability of $1/2$, another problem from using the asymptotic distribution of the likelihood root is the assumption that the non-zero part is standard normal. Figure 3.3 shows Gaussian QQ-plots of simulated values of the ordinary and the restricted likelihood roots, and the corresponding modified pivots. For small k, the likelihood root tends to be smaller than standard normal quantiles, especially for fixed k and larger m. Sample quantiles of the restricted likelihood root, displayed in the right panels of Figure 3.3, show a minor departure from unit slope, implying that the corresponding distribution is closer to standard normal than is that of the ordinary-based likelihood root. The modified

Table 3.7 – Probability (%) of non-zero estimates in variance components model (3.2) using 10^4 simulations of r and r^* for the usual likelihood "Usual" and the restricted likelihood "REML". The dataset consists of k groups each of size m. Figures in bold equal 50% up to simulation error.

		$m=5$		$m=10$		$m=20$		$m=30$	
k	pivot	Usual	REML	Usual	REML	Usual	REML	Usual	REML
5	r	32.8	43.5	29.8	41.1	30.2	41.2	29.6	41.3
	r^*	48.7	**50.3**	47.6	**49.0**	48.5	**49.7**	**49.2**	**50.5**
10	r	37.1	44.9	36.2	44.5	36.7	45.0	35.6	44.0
	r^*	**49.3**	**49.7**	**49.5**	**49.8**	**50.3**	**50.6**	**49.5**	**49.9**
20	r	40.6	46.1	39.9	45.8	39.9	46.1	40.2	46.5
	r^*	**49.0**	**49.1**	**49.3**	**49.5**	**49.7**	**49.9**	**50.5**	**50.6**
30	r	42.0	46.7	42.8	47.5	42.4	47.7	41.6	46.4
	r^*	**49.3**	**49.4**	**50.5**	**50.6**	**51.0**	**51.0**	**49.5**	**49.7**
50	r	44.7	48.2	43.2	47.5	43.7	47.4	43.5	47.5
	r^*	**50.2**	**50.3**	**49.5**	**49.5**	**49.7**	**49.7**	**49.7**	**49.8**
100	r	46.3	**49.0**	44.9	47.7	45.3	48.3	45.7	48.4
	r^*	**50.3**	**50.3**	**49.2**	**49.2**	**50.0**	**50.0**	**50.2**	**50.2**

likelihood root efficiently corrects the departure from normality and produces better results overall, except for small k where both pivots have light upper tails. In this setting, the corrected likelihood root produces almost identical results to the REML-based version, as shown in Figure 3.4. The correction terms, $\log(q/r)/r$, are of different magnitudes. However, in unbalanced settings and with additional fixed effects, REML-based solutions may perform better (Chatterjee and Das, 1983; Corbeil and Searle, 1976; Brown and Kempton, 1994).

In Table 3.8, we study the coverage of one-sided confidence intervals of the form $[L_\alpha, +\infty)$ for $\alpha < 0.4$. Lower-bound intervals correspond to the upper tail of the pivots in the QQ-plots of Figure 3.3, and are better suited for testing $\psi = 0$ against $\psi > 0$. Table 3.8 shows that empirical coverage based on r^* is excellent, as the left-tail error is equal to the nominal value for all values of α and considered sample sizes.

A natural extension of the previous model is the balanced two-way crossed random model with one observation per cell

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}, \quad i = 1, \ldots, k, j = 1, \ldots, m, \tag{3.26}$$

where a_i, b_j and ε_{ij} are independently normally distributed with mean zero and variances σ_a^2, σ_b^2 and σ^2 respectively. This model can be written as $y = X\beta + Z_1 b_1 + Z_2 b_2 + \varepsilon$, where X is an $mk \times 1$ vector of ones, b_1 contains the a_i's, b_2 the b_j's, and

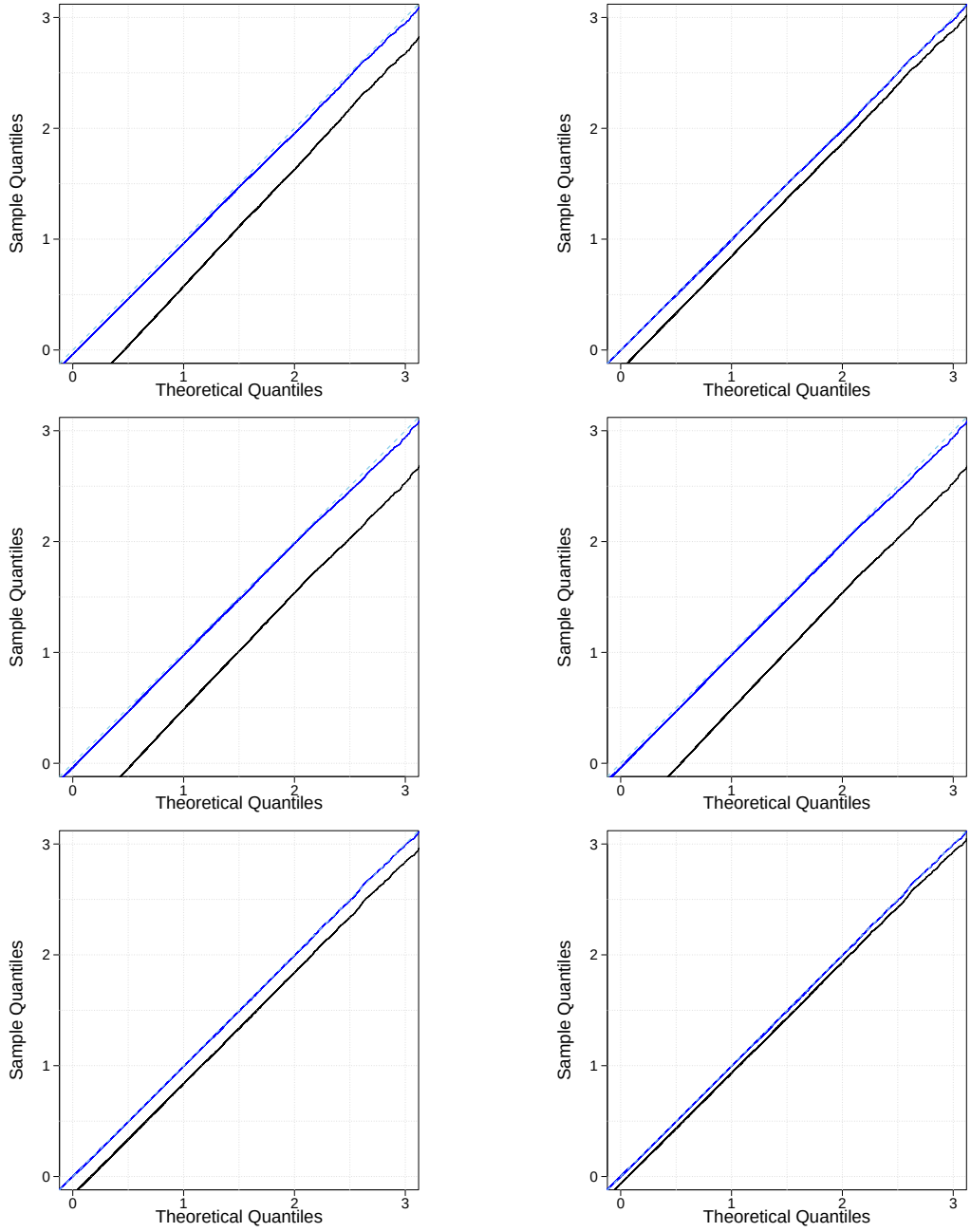


Figure 3.3 – Left panels: Gaussian QQ-plots based on 10^4 Monte Carlo samples of $r(\psi_0)$ (black) and $r^*(\psi_0)$ (blue) in one-way classification model with $(k,m) = (5,5)$, $(5,30)$, $(50,30)$ (top to bottom). Right panels: the corresponding REML solutions.

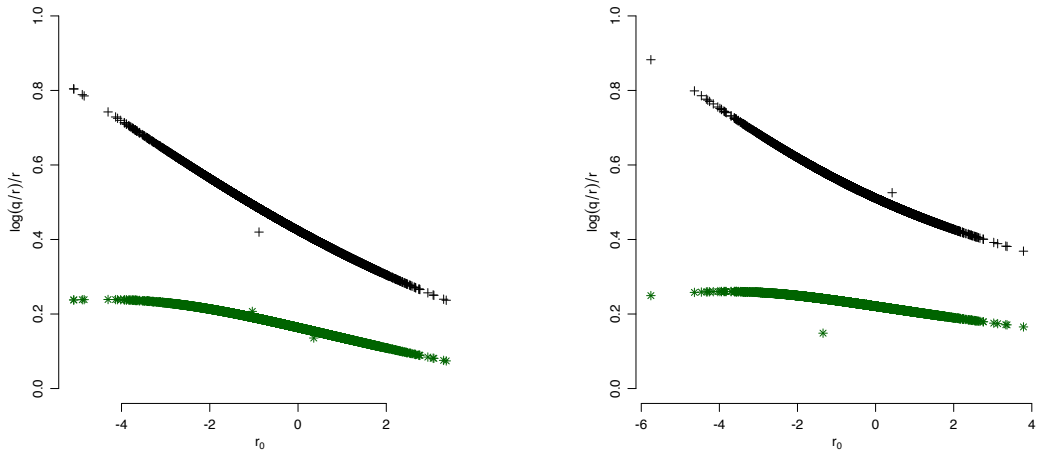


Figure 3.4 – Correction term for the usual likelihood (black +) and restricted likelihood (green *) plotted against the likelihood root for 10^4 simulated data with $(k,m) = (5,5)$ (left) and $(k,m) = (5,30)$ (right).

ε the ε_{ij}'s. If we add an interaction term c_{ij}, for experiments with more than one observation per cell, then the model is a two-way crossed effects with interaction, and the Z's may be written using the Kronecker product of matrices and vectors of ones of appropriate dimension as in Miller (1977) or the squared sum representation as in Sahai and Ojeda (2004, Chapter 4).

Define $\psi_1 = \sigma_a^2/\sigma^2$, $\psi_2 = \sigma_b^2/\sigma^2$, and consider the null hypothesis $H_0 : \psi_1 = 0$, under which we test the significance of the first factor. Under the null hypothesis, the true value of ψ_2 is large enough to ensure that that second factor is consistently positive even for small values of k and m; Figure 3.5 shows the proportion of positive variance components in such a setting. Simulations in Susko (2013) show that the proportions of positive variance components in a model where one parameter is zero depend on how far the parameters are from their null hypothesis values, and suggest that a conditional chi-square test is more powerful than the classical chi-bar test.

In our example, the probabilities of a positive estimate for different m and k when $\psi_2 = 1$ and $\sigma^2 = 1$ using 10^4 simulations of r and r^* based on the restricted likelihood are given in Table 3.9. It is interesting to note the similarities between the two variance component examples: most probabilities based on the likelihood root are not very far from $1/2$, but improvement is possible. When computed based on the modified likelihood root, these probabilities equal $1/2$, and the distribution of the corresponding sample quantiles is closer to the standard normal, as shown in Figure 3.6. These results support the conclusion of Stein et al. (2014), who studied the effectiveness of the standard likelihood ratio test in mixed linear models with small sample

Table 3.8 – Empirical coverage probabilities of the right-tail confidence intervals in the one-way classification model based on 10^4 simulations for k groups each of size m. Figures in bold equal the nominal values up to simulation error.

	$(k,m)=(5,5)$		$(k,m)=(5,30)$		$(k,m)=(50,30)$	
α	r	r^*	r	r^*	r	r^*
0.005	0.001	**0.004**	0.001	**0.004**	**0.004**	**0.004**
0.010	0.003	**0.008**	0.003	**0.009**	0.006	**0.009**
0.025	0.009	**0.022**	0.008	**0.024**	0.017	**0.024**
0.050	0.022	**0.046**	0.018	**0.048**	0.036	**0.050**
0.100	0.047	**0.098**	0.039	**0.097**	0.076	**0.100**
0.200	0.110	**0.195**	0.092	**0.196**	0.160	**0.198**
0.300	0.175	**0.293**	0.153	**0.292**	0.247	**0.292**
0.400	0.247	**0.391**	0.220	**0.391**	0.335	**0.393**

sizes. They demonstrated that other methods, such as the bootstrap-based test, the Bartlett-corrected usual test, and the adjusted profile likelihood ratio test, produce better results for two specific mixed linear models.

Example: Penalized splines

Testing for polynomial regression versus a non-parametric alternative has often been addressed using spline fits. The choice of the basis function and the penalty usually depends on the complexity of the problem at hand and the structure of the underlying data (de Boor, 1978; Green and Silverman, 1994; Hastie and Tibshirani, 1990; Wahba, 1990; Wood, 2017). Under the null hypothesis, this is a soft boundary problem as

$$\psi > -\frac{1}{\max\{\text{eigen}(ZZ^{\mathrm{T}})\}},$$

where the lower bound in this inequality is negative.

A convenient property of smoothing splines in (3.4.1) is that the penalty is written as $\mu^{\mathrm{T}}K\mu$ with a suitably defined penalty matrix K. Crainiceanu and Ruppert (2004b) and Berry et al. (2002) considered non-parametric regression using a truncated power basis of order p

$$\mu(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{n} (x-\tau_k)_+^p, \tag{3.27}$$

where $\tau_1,\ldots,\tau_n$ are the knots. They investigated the use of the likelihood ratio as a statistic for hypothesis testing when $p=0$, corresponding to a constant mean β_0,

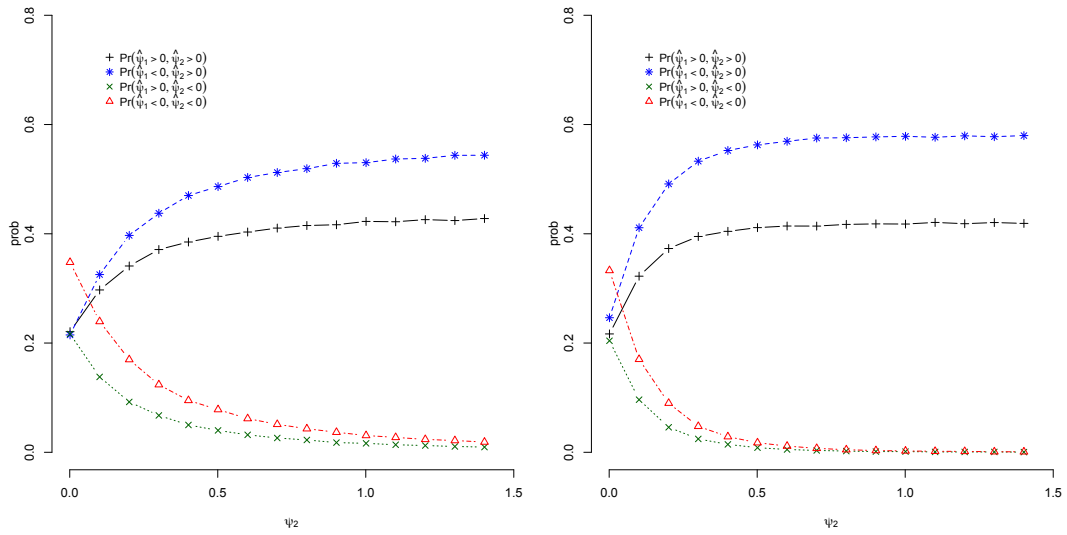


Figure 3.5 – Proportion of positive variance components in model (3.26), where $(k,m) = (5,5)$ (left), and $(k,m) = (5,10)$ (right). $\Pr(\widehat{\psi}_1 > 0, \widehat{\psi}_2 > 0)$ (black "+"), $\Pr(\widehat{\psi}_1 < 0, \widehat{\psi}_2 > 0)$ (blue"$*$"), $\Pr(\widehat{\psi}_1 > 0, \widehat{\psi}_2 > 0)$ (green "$\times$"), $\Pr(\widehat{\psi}_1 < 0, \widehat{\psi}_2 < 0)$ (red"$\triangle$").

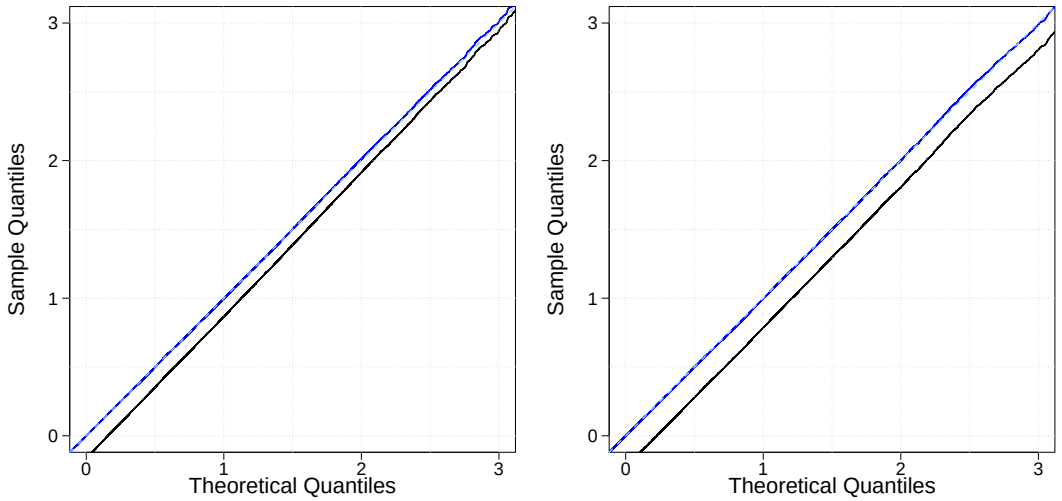


Figure 3.6 – Gaussian QQ-plots based on 10^4 Monte Carlo samples of $r(\psi_0)$ (black) and $r^*(\psi_0)$ (blue) in two-way classification model (3.26), with $(k,m)(5,5);(5,30);(50,5)$ (left to right).

Table 3.9 – Probability (%) of non-zero estimates of the variance of factor I in a two-way classification model (3.26) using 10^4 simulations of r and r^* for k groups each of size m. Figures in bold equal 50% up to simulation error.

k		$m=5$	$m=10$	$m=20$	$m=30$
5	r	44.0	42.5	41.9	41.0
	r^*	**50.1**	**50.4**	**50.5**	**49.7**
10	r	45.7	44.5	43.3	43.0
	r^*	**50.1**	**49.9**	**49.5**	**49.0**
20	r	47.4	46.9	46.0	46.7
	$r*$	**50.4**	**50.4**	**49.8**	**50.6**
30	r	47.4	46.4	46.9	46.9
	r^*	**49.8**	**49.6**	**50.3**	**50.4**
50	r	**49.0**	48.1	47.0	47.8
	r^*	**50.7**	**50.4**	**49.3**	**50.3**

and $p = 1$, corresponding to a linear polynomial. The percentage of zero variance components in these cases, for moderate numbers of equally spaced knots, was found to be greater than 90%. The restricted likelihood ratio was found to be more effective, for example using $n = 20$ knots the asymptotic probability mass at zero is 0.65 for restricted likelihood ratio and 0.95 for the usual likelihood. So the usual likelihood ratio is essentially useless for hypothesis testing and also for higher-order correction.

Natural cubic splines generally have the desirable properties of being very stable and numerically efficient. However, for other choices of regression functions, for instance B-splines, the penalty matrix can be obtained using an invertible change of basis (Ruppert et al., 2003). In addition to their numerical appeal, we favor natural cubic splines because of the structure of the covariance matrix under the alternative hypothesis. For the truncated power basis in the example of Crainiceanu and Ruppert (2004b), when $\psi > 0$, the variance of the data increases as x increases. This implies that the variance increases in a way that is often statistically unnatural, as a broadly constant variance is typically observed in most applications, see the panels of Figure 3.7. This may explain the large point mass at zero obtained in Ruppert et al. (2003) where the spread of all data points suggest that H_0 is more plausible than H_1.

Consider testing for a linear regression versus a general alternative modeled as natural cubic splines. We follow Ruppert et al. (2003), and take x to be equally spaced on $[0,1]$, and n equally-spaced knots. Table 3.10 presents the probability of a positive gradient at the origin based on the saddlepoint approximation (3.20), and 10^4 samples of r and r^* from the equivalent linear mixed effects model described in (3.22) . In more than two-thirds of the simulations, the likelihood is maximized at $\psi = 0$. On

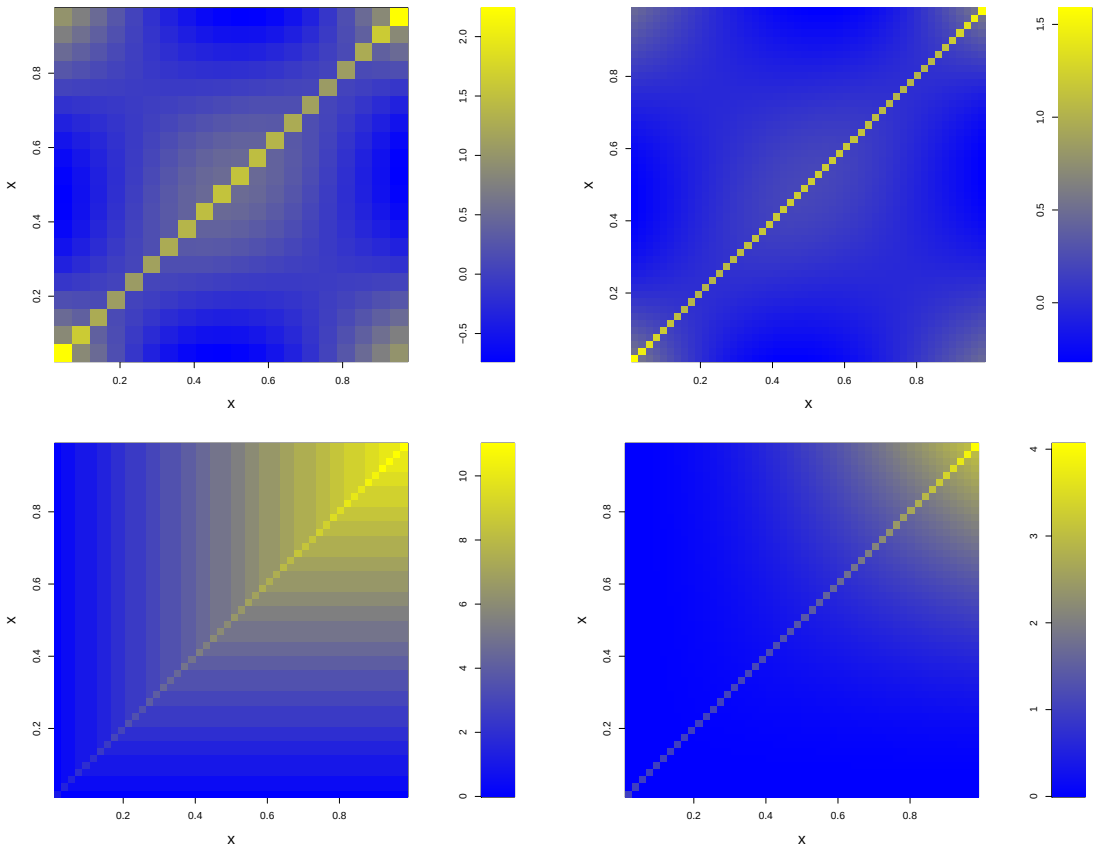


Figure 3.7 – Covariance matrix $I_n + \psi ZZ^{\mathrm{T}}$ for $\psi = 0.5$ in penalized splines model. Top panels: natural cubic splines with $n = 20$ (left) and $n = 50$ (right). Bottom panels: constant and linear truncated power basis with 20 knots for $n = 50$ (left to right).

the other hand, the probabilities based on r^* are closer to $1/2$ even for $n = 10$ knots. QQ-plots in Figure 3.8 compare positive sample quantiles of the two pivots to those of the standard normal distribution. The likelihood root has a negative intercept in both plots and is particularly right-skewed for small n, in alignment with the proportion of boundary estimates shown in Table 3.10. The modified likelihood root corrects the departure from normality.

Example: Generalized Pareto

Testing for a zero-shape parameter in generalized Pareto distribution can be regarded as a soft boundary problem since ξ can be negative, but we limit our interpretation to the positive values. When the parameter $\xi < 0$, the upper endpoint of the distribution depends on ξ, and the model is irregular when $\xi < -1/2$.

Table 3.10 – Probability (%) of positive estimates in a penalized spline model with n knots using the saddlepoint approximation “sp" described in (3.20), and the pivots r and r^*, based on 10^4 replicate samples. Figures in bold equal 50% up to simulation error.

	$n=10$	$n=20$	$n=30$	$n=50$	$n=100$	$n=200$	$n=300$
sp	42.73	36.94	35.28	34.03	33.14	33.14	32.72
r	42.79	39.03	36.67	35.42	34.55	34.35	34.02
r^*	**49.01**	**50.93**	**50.77**	**50.41**	**50.55**	**50.41**	**50.80**

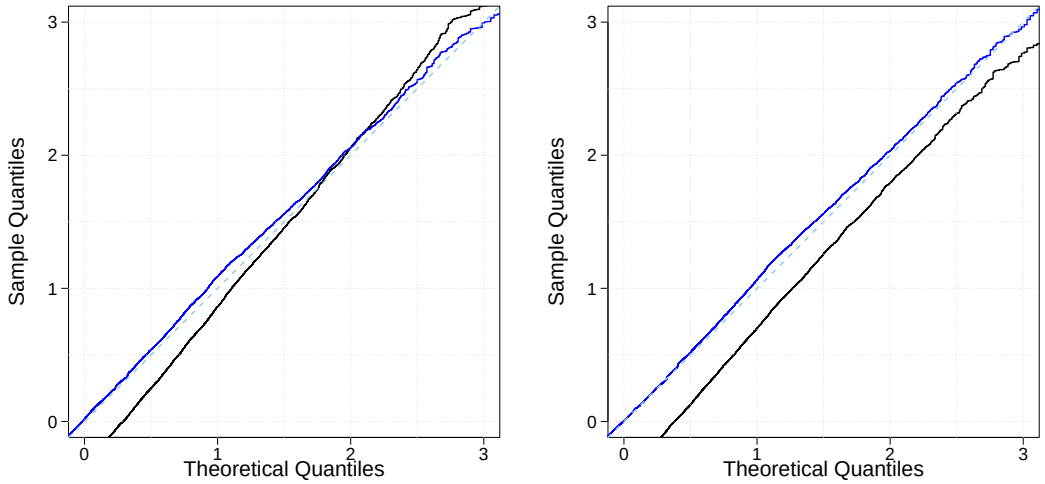


Figure 3.8 – Gaussian QQ-plots of the non-zero components of the likelihood root (black) and modified likelihood root (blue) in penalized splines for $n=20$ (left) and $n=50$ (right).

The log likelihood for a random sample $x=(x_1,\ldots,x_n)$ from $\mathrm{GP}(\sigma,\xi)$ is

$$\ell(\sigma,\xi)=-n\log(\sigma)-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1+\frac{\xi x_i}{\sigma}\right),\quad \xi\neq 0,\sigma>\max(0,-\xi x_{(n)}). \tag{3.28}$$

Under the full model, there is no closed-form expression for the maximum likelihood estimators of (3.28). Grimshaw (1993) reduced the two-dimensional numerical search for the zeros of the log likelihood gradient vector to a one-dimensional numerical search in a transformed parametrization as in Davison (1984). In this example, numerical optimization of the generalized Pareto distribution is carried out using the function “fit.gpd" from the R “mev" package (Belzile et al., 2022).

Table 3.11 – Probability (%) of positive maximum likelihood estimator of the shape parameter in a generalized Pareto distribution (3.23) based on 10^4 Monte Carlo samples. Figures in bold equal 50% up to simulation error.

n	40	60	80	100	200	400
r	36.01	38.26	40.52	40.51	43.10	46.04
r^*	**49.96**	**50.21**	**50.83**	**50.37**	**49.66**	**50.76**

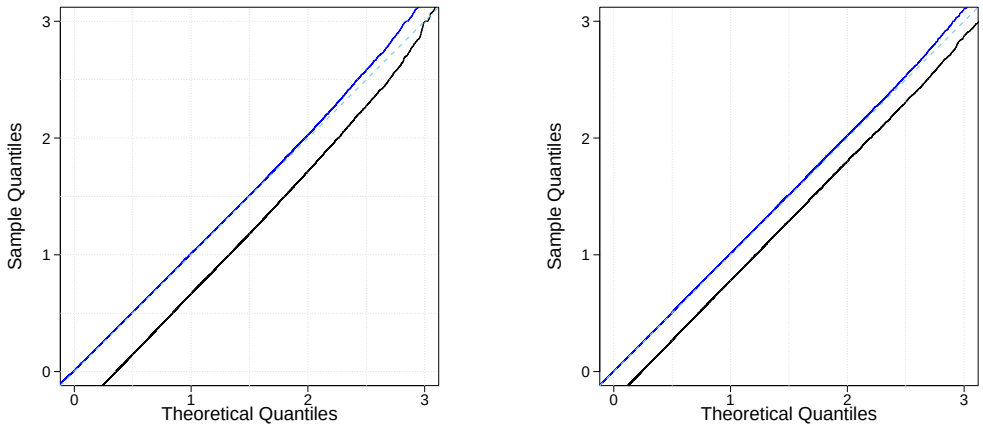


Figure 3.9 – Gaussian QQ-plots of the non-zero components of the likelihood root (black) and modified likelihood root (blue) in generalized Pareto distribution with shape $\xi = 0$ based on 10^4 simulations for $n = 40$ (left) and $n = 60$ (right).

Fitting the generalized Pareto distribution to 10^4 samples of exponential variables with size n gives the probabilities in Table 3.11. The finite-sample distribution of r^* appears to be almost identical to the asymptotic distribution for all chosen n: the proportion of positive estimates of the shape parameter is 50%, and the distribution of the positive components of the corrected likelihood root follows a standard normal distribution as shown in Figure 3.9.

In Table 3.12, we show the empirical coverage of the right-tail intervals for $n = 40, 60, 100$. For small nominal values, intervals based on r tend to contain ψ_0 less often, especially for small sample sizes. This is due to the left-shifted distribution of the pivots, but r^* provides better coverage with left-tail errors equal to the nominal values. The results of this study are consistent with those reported by Hosking (1984), who examined the two-sided alternative $\xi \neq 0$ in order to determine the direction of any deviation from the null hypothesis $\xi = 0$.

Table 3.12 – Empirical coverage probabilities for the generalized Pareto distribution where $\xi = 0$, based on 10^4 simulations with $n = 40, 60$ and 100. Figures in bold equal the nominal values up to simulation error.

	$n = 40$		$n = 60$		$n = 100$	
α	r	r^*	r	r^*	r	r^*
0.005	0.002	**0.005**	0.003	**0.006**	0.003	**0.005**
0.010	0.004	**0.009**	0.005	**0.012**	0.006	**0.011**
0.025	0.012	**0.026**	0.015	**0.028**	0.014	**0.024**
0.050	0.026	**0.050**	0.031	**0.053**	0.030	**0.048**
0.100	0.053	**0.098**	0.063	**0.105**	0.065	**0.098**
0.200	0.117	**0.199**	0.134	**0.205**	0.143	**0.203**
0.300	0.191	**0.302**	0.212	**0.306**	0.227	**0.306**
0.400	0.273	**0.400**	0.295	**0.404**	0.318	**0.403**

3.5.2 Example: Student t

The usual calculations in Davison (2003, Chapter 7) show that in conventional asymptotics, the power for testing $H_0 : \psi = \psi_0$ when in fact $\psi = \psi_1$ depends on

$$\tilde{\delta} = n^{1/2}\frac{\psi_1 - \psi_0}{b},$$

so if $\psi_1 = \psi_0 + c/n^a$, then the power is asymptotically zero if $a > 1/2$. Hence ψ_0 and ψ_1 are statistically indistinguishable, but as $\psi_1 > 0$, calculations of r and q can be performed at ψ_1, though not at ψ_0. Of course, one has to choose a and c so that the value of ψ_1 is close to the true null. This supports a possible framework for examining how higher-order approximation performs for hard-boundary problems such as with the Student t distribution and other examples.

For hard-boundary problems, we only seek to improve the p-values for H_0 by correcting the distribution of the non-zero part of the likelihood root. We consider a null hypothesis H_0 under which we test that the reciprocal of the degrees of freedom $\psi_0 = 1/\nu_0 = b/n^{(1+\varepsilon)/2}$, for $c = 1$, and $\varepsilon = 0.2, 0.4, 1$. The smaller ψ_0, the higher the degrees of freedom, and then the Student t distribution approximates the normal distribution for both small and large sample sizes, as illustrated in Figure 3.10. In principle, smaller values of ε and c could be used, but simulations near the boundary show that the observed information matrix is often not positive definite, and higher-order correction might be more prone to failure. Maximum likelihood estimates under the full model are obtained using the EM algorithm (Liu and Rubin, 1995);(Davison, 2003, Chapter 5); details are in Appendix 3.8.4.

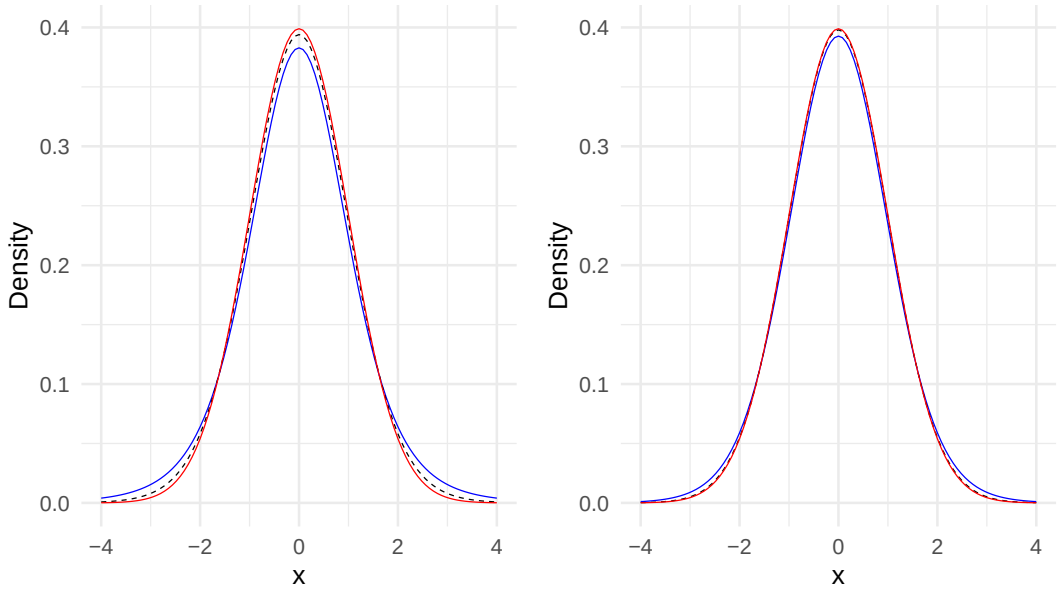


Figure 3.10 – Standard normal density (red) vs Student t with $\nu = n^{(1+\varepsilon)/2}$ degrees of freedom for $\varepsilon = 0.2$ (blue) and $\varepsilon = 1$ (black) and sample size $n = 20$ (left) and 100 (right).

Figure 3.11 shows QQ-plots for r and r^* under H_0, in which r^* has sample quantiles closer to unit slope and the flat segment around the origin corresponds to the extra point mass at zero. The shadow likelihood root $\tilde{r}(\psi_0)$ discussed in Section 3.3.1 is a quick fix to the finite sample-distribution of the likelihood root, and in this case, shifts the sample quantiles to the right, but more is needed.

Table 3.13 shows that the modified likelihood root performs better in terms of empirical coverage for small sample sizes. But both pivots tend to undercover the true value, particularly for large nominal values, as the estimated lower bounds of the confidence intervals approach the origin. This is especially evident for $\varepsilon > 0.4$, $n \geq 100$ and $\alpha > 0.2$, where the modified likelihood root, r^*, does not perform well but it is still better than r. Figure 3.12 shows that p-values using the modified root r^* have a smaller relative error than those for r. This confirms that first-order pivots, such as the likelihood root, produce large p-values that fail to reject the null hypothesis as often as needed. The corrected p-values are less conservative, with a relative error closer to zero.

Table 3.13 – Empirical coverage probabilities of the right-tail confidence intervals based on 10^4 simulations from the Student t distribution $\nu = n^{-(1+\varepsilon)/2}$ degrees of freedom, for $n = 20, 50$ and 100 and $\varepsilon = 0.2, 0.4$ and 1. Figures in bold equal the nominal values up to simulation error.

		$n = 20$		$n = 50$		$n = 100$	
	α	r	r^*	r	r^*	r	r^*
$\varepsilon = 0.2$	0.005	0.002	**0.004**	0.003	**0.005**	0.002	**0.004**
	0.010	0.005	**0.009**	0.007	**0.010**	0.006	**0.009**
	0.025	0.015	**0.024**	0.015	**0.024**	0.015	**0.023**
	0.050	0.031	**0.050**	0.033	**0.050**	0.032	**0.049**
	0.100	0.064	**0.103**	0.067	**0.101**	0.067	**0.099**
	0.200	0.129	**0.202**	0.140	**0.204**	0.147	**0.200**
	0.300	0.194	**0.299**	0.212	**0.298**	0.222	**0.303**
	0.400	0.262	0.304	0.288	0.338	0.306	0.352
$\varepsilon = 0.4$	0.005	0.003	**0.006**	0.002	**0.005**	0.003	0.007
	0.010	**0.008**	**0.011**	0.005	**0.010**	0.006	**0.011**
	0.025	0.017	**0.026**	0.014	**0.023**	0.016	**0.028**
	0.050	0.032	**0.051**	0.029	**0.048**	0.033	**0.051**
	0.100	0.060	**0.100**	0.061	**0.101**	0.068	**0.102**
	0.200	0.121	**0.208**	0.127	**0.195**	0.142	0.213
	0.300	0.182	0.286	0.196	**0.303**	0.219	0.313
	0.400	0.251	0.286	0.274	0.323	0.298	0.339
$\varepsilon = 1$	0.005	0.003	**0.005**	0.003	**0.006**	**0.004**	0.009
	0.010	0.005	**0.010**	0.005	**0.010**	0.007	0.014
	0.025	0.013	**0.024**	0.012	**0.026**	0.014	**0.028**
	0.050	0.025	**0.049**	0.026	**0.053**	0.030	**0.054**
	0.100	0.050	**0.105**	0.056	**0.104**	0.060	0.113
	0.200	0.110	0.221	0.120	0.220	0.133	0.225
	0.300	0.169	0.260	0.192	0.288	0.201	0.320
	0.400	0.226	0.262	0.260	0.299	0.283	0.329

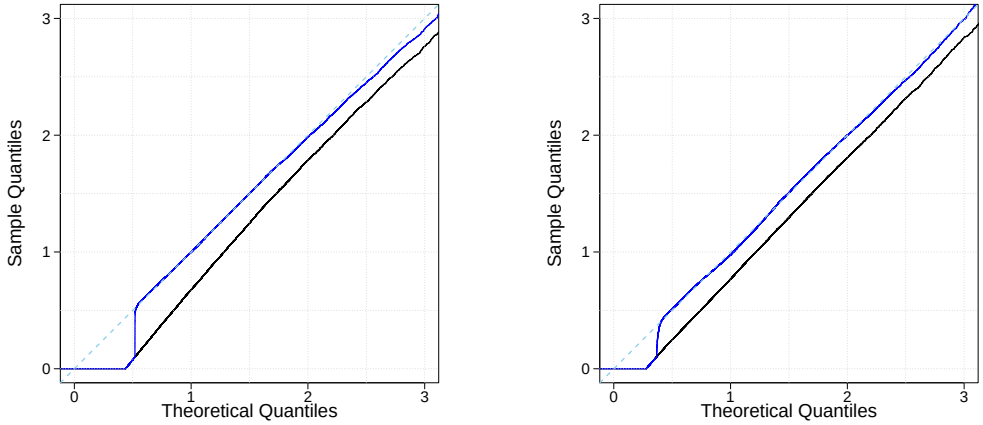


Figure 3.11 – Gaussian QQ-plots of the non-zero components of the likelihood root (black) and modified likelihood root (blue) based on 10^4 simulations in the Student t with $\nu = n^{(1+\varepsilon)/2}$ degrees of freedom for $\varepsilon = 0.2$ where $n = 20$ (left) and $n = 100$ (right).

3.5.3 Example: Negative binomial

The log likelihood function based on a single observation from the negative binomial density in (3.25) is

$$\ell(\xi, \nu) = A(y+\nu) - A(\nu) + \nu\log(\nu) + y\log(\xi) - (\nu+y)\log(\nu+\xi), \quad \xi, \nu > 0,$$

where $A(\nu) = \log\Gamma(\nu)$ is the log-gamma function, with derivative the digamma function.

Results based on 10^4 simulated data from the Poisson distribution with $\xi = 2$ show that $r(\psi_0)$ is far from its asymptotic standard normal distribution. For this example, two versions of r^* are implemented; the second one is based on Skovgard's approximation to the sample space derivative (Skovgaard, 1996). Both versions yield similar results, so we report those computed using the tangent exponential model formulation.

The discrepancy from the standard normal distribution is corrected by r^*, which does a satisfactory job but tends to over-correct large quantiles, resulting in a right-skewed tail for large n. The length of the flat segment around the origin shrinks as the point mass at zero converges to 1/2 for larger sample sizes. As shown in Figure 3.14, the relative error of the p-values based on the likelihood root decreases by half when the sample size n increases from 20 to 50. However, the p-values obtained from the modified likelihood root are less conservative. If the sample size is large (e.g., $n > 200$), both relative errors tend to be small, and the likelihood root may be preferred.

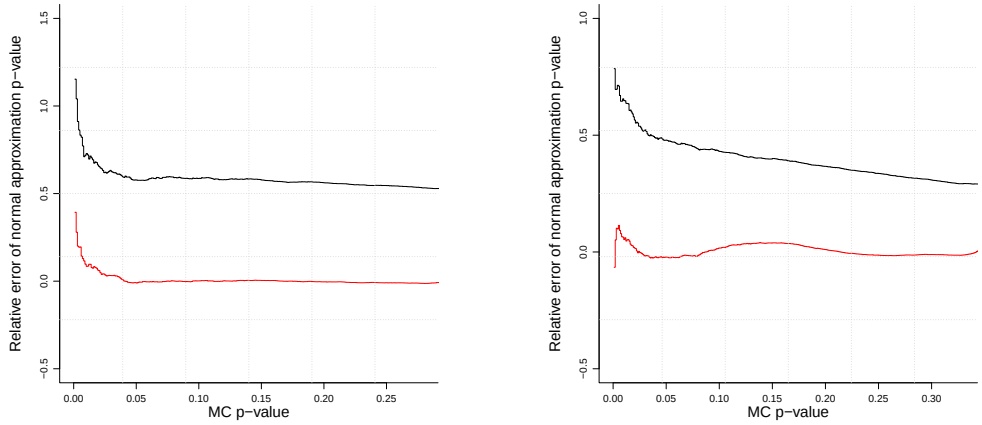


Figure 3.12 – Relative error of the normal approximation to the p-value using r (black) and r^* (red) as a function of $\Pr_0(r > r_{\text{obs}})$ approximated using 10^5 Monte Carlo samples for the Student t distribution with $\nu = n^{(1+\varepsilon)/2}$ degrees of freedom for $\varepsilon = 0.2$, $n = 20$ (right) and $n = 100$ (left).

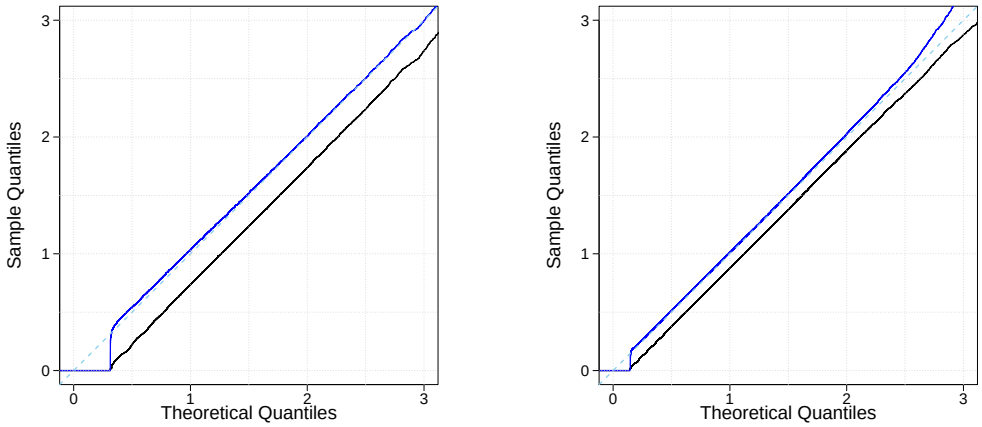


Figure 3.13 – Gaussian QQ-plots of the non-zero components of the likelihood root (black) and modified likelihood root (blue) based on 10^4 simulations in the negative binomial example for $n = 20$ (left) and $n = 100$ (right).

3.5.4 Example: Gaussian mixture

In our last example, we revisit the hard boundary problem mentioned in the motivational section. Under the null hypothesis $H_0 : \psi = 0$, we are testing whether the data is from a single normal distribution $\mathcal{N}_p(\lambda, I_p)$, rather than a Gaussian mixture. We set $\lambda = (1, \ldots, p)^{\mathrm{T}}$, and the parameters are estimated using the EM algorithm (see details in Appendix 3.8.4).

Figure 3.15 shows sample quantiles of $r(\psi_0)$ under the null hypothesis for two different dimensions of the nuisance parameter vector, $p = 5$ and $p = 10$. The sample quantiles

Table 3.14 – Empirical coverage probabilities of the right-tail confidence intervals based on 10^4 simulations for the negative binomial distribution for $n = 20,\ 50$ and 100. Figures in bold equal the nominal values up to simulation error.

	$n = 20$		$n = 50$		$n = 100$	
α	r	r^*	r	r^*	r	r^*
0.005	0.003	**0.005**	**0.004**	**0.005**	0.003	**0.005**
0.010	0.005	**0.010**	0.006	**0.011**	0.006	**0.010**
0.025	0.012	**0.025**	0.018	**0.027**	0.018	**0.027**
0.050	0.028	**0.054**	0.036	**0.054**	**0.041**	**0.053**
0.100	0.064	**0.104**	0.076	**0.103**	0.081	**0.102**
0.200	0.132	0.210	0.158	0.211	0.164	**0.196**
0.300	0.212	0.314	0.246	0.312	0.249	**0.297**
0.400	0.298	0.375	0.338	0.414	0.349	**0.401**

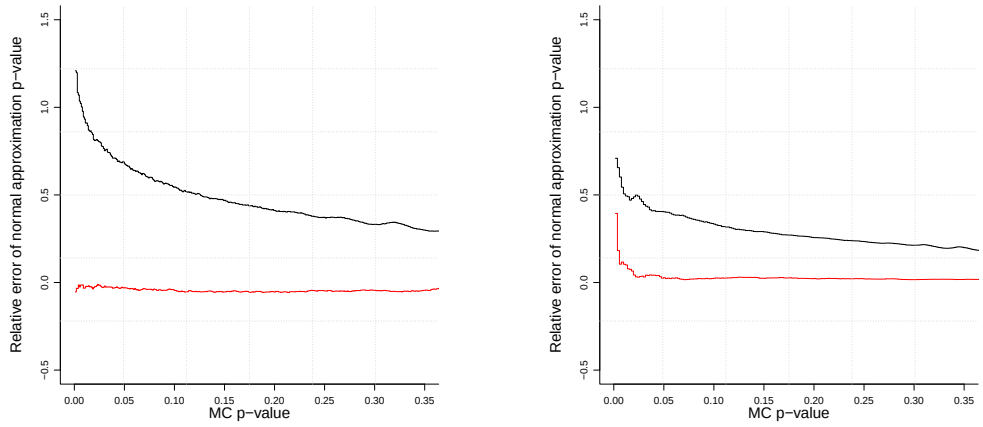


Figure 3.14 – Relative error of the normal approximation to the p-value using r (black) and r^* (red) as a function of $\Pr_0(r > r_{\text{obs}})$ approximated using 10^5 Monte Carlo samples for the negative binomial with $\xi = 2$ and sample size $n = 20$ (right) and $n = 50$ (left).

of the likelihood root are smaller than the standard normal quantiles. The QQ-plots of r^* are closer to the asymptotic distribution, with a slight deviation in the upper tail for fixed sample size n and large p. Table 3.15 shows the empirical coverage of the confidence intervals and demonstrates that r^* is vastly preferable to r for testing H_0, in line with previous results on the use of r^* in hard boundary problems.

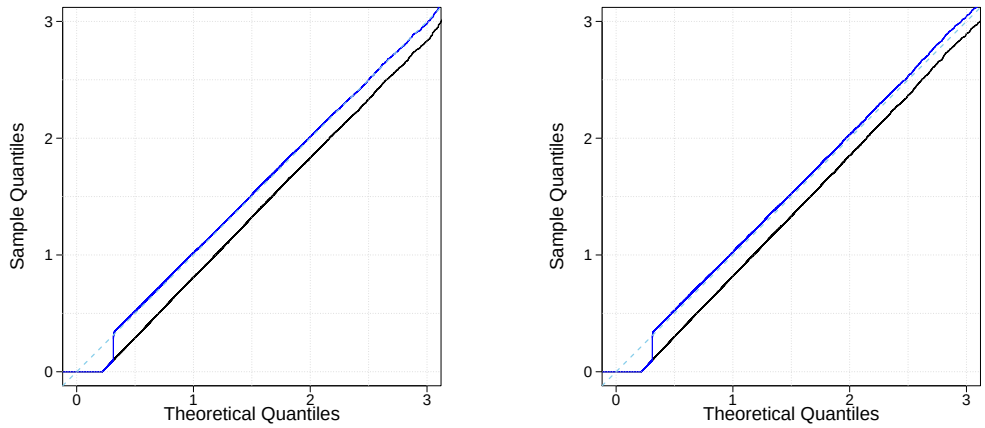


Figure 3.15 – Gaussian QQ-plots of the non-zero components of the likelihood root (black) and modified likelihood root (blue) based on 10^4 simulations in two-Gaussian mixture for $n = 30$, $p = 5$ (left) and $p = 10$ (right).

3.6 Data illustrations

Below we illustrate the use of the tangent exponential model for three datasets denoted $\mathscr{D}_i$, $i = 1,2,3$. These data are assumed to be observations from distributions discussed in Section 3.4, namely the variance components model (3.26), the penalized splines model (3.21), and the generalized Pareto distribution model (3.23). For these examples, third-order results are different from the first-order ones, and show that higher-order approximations can confirm that we have a boundary estimate for the parameter of interest or lead to a different conclusion. A good number of data illustrations in the literature of third-order approximations for non-boundary problems can be found in Severini (2000, Chapter 7), Davison (2003), Brazzale et al. (2007), and Butler (2007, Chapter 1,2).

Chimpanzee data

Consider a dataset that we denote $\mathscr{D}_1$, which represents the times (min) for four chimpanzees to learn each of ten words shown in Table 3.16. The data are discussed in Davison (2003, Chapter 10) and taken from Brown and Hollander (1977). A possible model for log time is the variance components model in (3.26) where y_{ij} describes the linguistic capacity of chimpanzee i to learn the word j, for $i = 1,\ldots,4$, $j = 1,\ldots,10$. Under these assumptions, we assume that the response is sampled from a larger population whose variation is of interest. However, we can consider a variance component model where one factor is treated as a fixed effect and design further tests for a particular word or chimpanzee.

Table 3.15 – Empirical coverage probabilities of the right-tail confidence intervals based on 10^4 simulations for the two-Gaussian mixture for $n = 30$, and $p = 5, 10, 20$. Figures in bold equal the nominal values up to simulation error.

	$p=5$		$p=10$		$p=20$	
α	r	r^*	r	r^*	r	r^*
0.005	0.002	**0.004**	0.003	**0.005**	0.003	**0.005**
0.010	0.005	**0.009**	0.006	**0.009**	0.006	**0.009**
0.025	0.015	**0.025**	0.016	**0.023**	0.016	**0.025**
0.050	0.035	**0.050**	0.034	**0.052**	0.035	**0.051**
0.100	0.069	**0.099**	0.072	**0.102**	0.073	**0.103**
0.200	0.151	**0.208**	0.145	**0.200**	0.152	**0.207**
0.300	0.238	**0.309**	0.228	**0.302**	0.238	**0.307**
0.400	0.324	0.381	0.318	0.376	0.330	0.382

Table 3.16 – Time (min) for four chimpanzees to learn each of ten words.

Chimpanzee	Word									
	1	2	3	4	5	6	7	8	9	10
1	178	60	177	36	225	345	40	2	287	14
2	78	14	80	15	10	115	10	12	129	80
3	99	18	20	25	15	54	25	10	476	55
4	297	20	195	18	24	420	40	15	372	190

Analysis of variance gives the sums of squares $C_c = 5.33$, $C_w = 45.69$, and $C_e = 17.65$, where the degrees of freedom for chimps is $c - 1 = 3$, for words $w - 1 = 9$, and for the residual $(c-1)(w-1)$. Estimates of the variance components and their standard deviations are $\widehat{\sigma}_c^2 = 0.112$ (0.335), $\widehat{\sigma}_w^2 = 1.105$ (1.051), $\widehat{\sigma}^2 = 0.654$ (0.808).

One aspect of interest is testing if chimp should be treated as a random effect since the corresponding variance is rather small. The null hypothesis, in this case, is $\psi = \sigma_c^2/\sigma^2 = 0$. The maximum likelihood estimate of ψ, the 95% confidence intervals using the Wald statistic, the likelihood root, and the analogous modified likelihood root are given in column $\mathscr{D}_1$ of Table 3.17. The lower bounds of the confidence intervals include negative values for ψ, although only the positive values are meaningful for these examples. The third-order approximation shifts the pivots to the right and lengthens the confidence intervals based on the first-order pivot, especially the inappropriate short interval based on the Wald statistic; see Figure 3.16 for a full summary. For small degrees of freedom, such as $c = 4$, representing chimp as a random effect seems unnecessary; the higher-order approximation confirms this conclusion.

Table 3.17 – Summaries of the chimpanzee data $\mathscr{D}_1$, the Venice sea-level data $\mathscr{D}_2$, and the precipitation data $\mathscr{D}_3$.

ψ		$\mathscr{D}_1$ Ratio of variances σ_c^2/σ^2	$\mathscr{D}_2$ Ratio of variances σ_b^2/σ^2	$\mathscr{D}_3$ shape parameter of GP
Point estimate	w	0.171	0.270	-0.052
	r	0.171	0.271	-0.052
	r^*	0.235	0.516	0.002
95% CI	w	$(-0.286, 0.630)$	$(-0.729, 1.272)$	$(-0.339, 0.233)$
	r	$(-0.033, 2.385)$	$(-0.020, 4.612)$	$(-0.294, 0.336)$
	r^*	$(-0.025, 3.752)$	$(-0.016, 6.790)$	$(-0.246, 0.418)$
pivots(H_0)	w^{o}	0.734	0.531	-0.361
	r^{o}	1.325	1.065	-0.336
	$r^{*\text{o}}$	1.514	1.469	0.0134

Venice sea level

In a second example we test whether a natural cubic spline is needed to fit the average of Venice's first 10-largest sea levels (cm) from 1887 to 2019. The available data $\mathscr{D}_2$ consist of the ten highest observations per year except for 1935, for which only the largest six observations are available (Davison, 2003, Chapter 10). In Figure 3.17, we plot the ten maximum sea levels and their average; the highest level ever recorded was 198 cm in the historic floods of 1966.

The null hypothesis being tested is $\psi = s_b^2/\sigma^2 = 0$, and further details can be found in Section 3.4.1. The fit summary in column $\mathscr{D}_2$ of Table 3.17 shows that the p-values using r and r^* are 0.14 and 0.07, respectively. As expected, the inference based on r^* is more stringent, though it does not alter the conclusion in this case. However, this example illustrates that higher-order approximations can lead to different conclusion when the exact p-value is close to a critical value such as $\alpha = 0.1$. When the alternative hypothesis is considered, i.e., using smoothing splines in the mixed model representation, the fitted model plotted in blue in Figure 3.17 is different from the linear fit shown in black dashes.

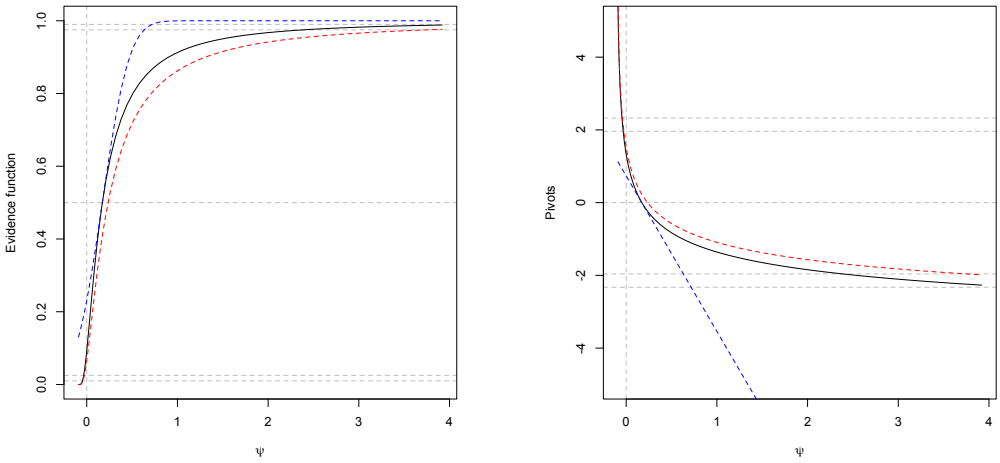


Figure 3.16 – Summaries for the chimpanzee data $\mathscr{D}_1$. Left: evidence function based on likelihood root $r(\psi)$ (black), Wald statistic $w(\psi)$ (dashed blue), and modified likelihood root $r^*(\psi)$ (dashed red). Right: pivots on standard normal scale.

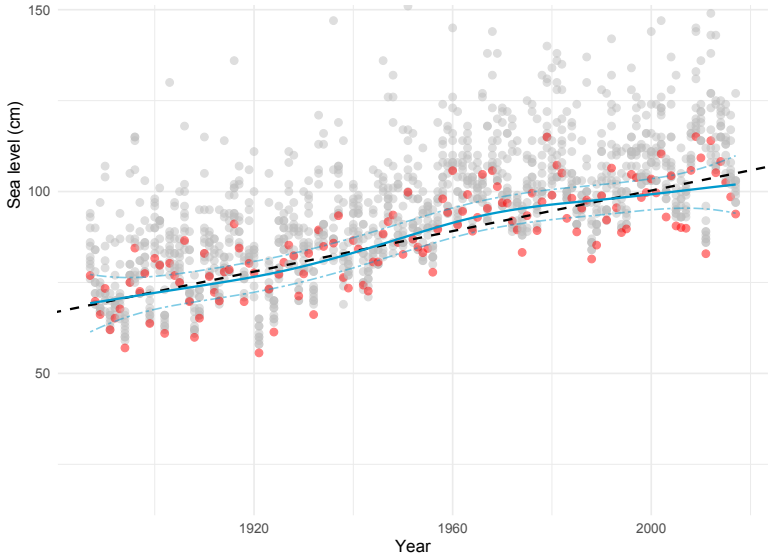


Figure 3.17 – Largest 10 annual sea-levels (cm) in Venice for each year from 1887 to 2019 (grey bullets), their average (red bullets), linear fit (dashed black) and spline fit (blue).

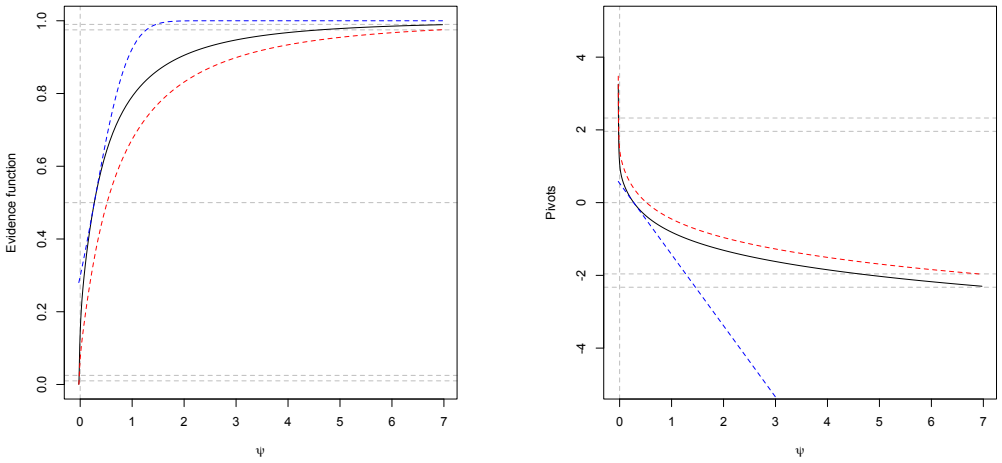


Figure 3.18 – Summaries for the Venice sea level data $\mathscr{D}_2$. Left: evidence function based on likelihood root $r(\psi)$ (black), Wald statistic $w(\psi)$ (blue), and modified likelihood root $r^*(\psi)$ (red dashed). Right: pivots on standard normal scale.

Precipitation data

Consider a third dataset $\mathscr{D}_3$, which consists of monthly precipitation (mm) data in the Czech Republic from 1981 to 2020. The data are publicly available in the R package “pRecipe". A detailed record of this dataset and other variables describing the global atmosphere, land surface and ocean waves based on atmospheric reanalysis is discussed in Hersbach et al. (2020). In the left panel of Figure 3.19, we plot the precipitation levels and, in the right panel, the corresponding seasonal boxplots.

Out of a total of 491 observations, 42 observations had values above a threshold of 115 mm. The maximum likelihood estimates of the generalized Pareto distribution, fitted to the standardized data, are $\widehat{\xi} = -0.052$ (0.1502) and $\sigma = 0.693$ (0.1493). The threshold was chosen using a mean residual life plot, but less subjective methods have been proposed (Dupuis, 1998; Northrop and Coleman, 2014). The right panel of Figure 3.20 gives the profile plot for the shape parameter. Point estimates and the corresponding 95% confidence intervals are given in the column $\mathscr{D}_3$ of Table 3.17. Confidence intervals based on the Wald statistic are likely to be biased and too short, as in the previous two examples. The modified likelihood root offers a considerable correction to the first-order pivot, especially for the upper bound of the confidence interval and the point estimate of the shape, which becomes positive. Nevertheless, p-values based on the three pivots are too large to reject the null hypothesis $H : \xi = 0$.

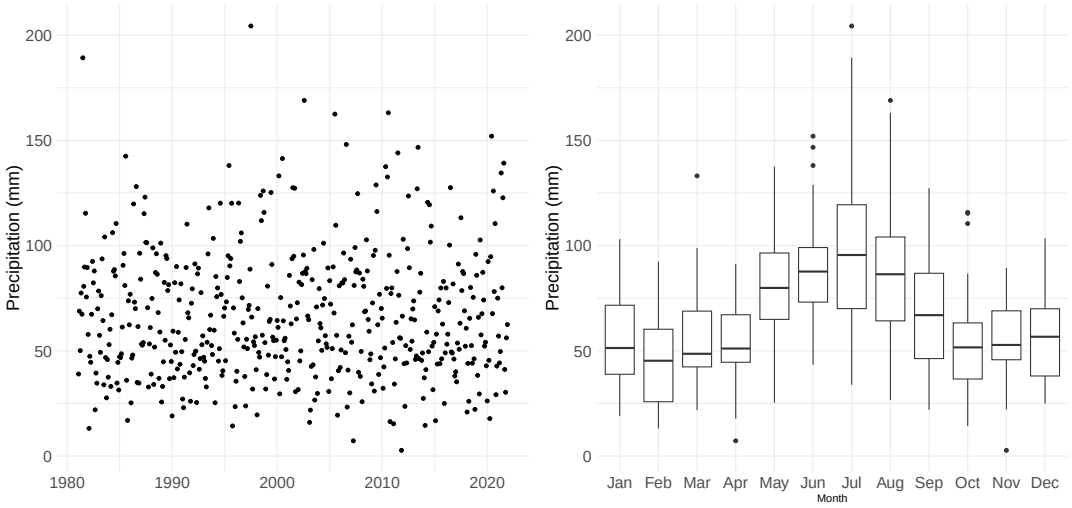


Figure 3.19 – Left: Monthly precipitation data (mm) in the Czech Republic from 1981 to 2020. Right: Seasonal boxplots of the precipitation levels.

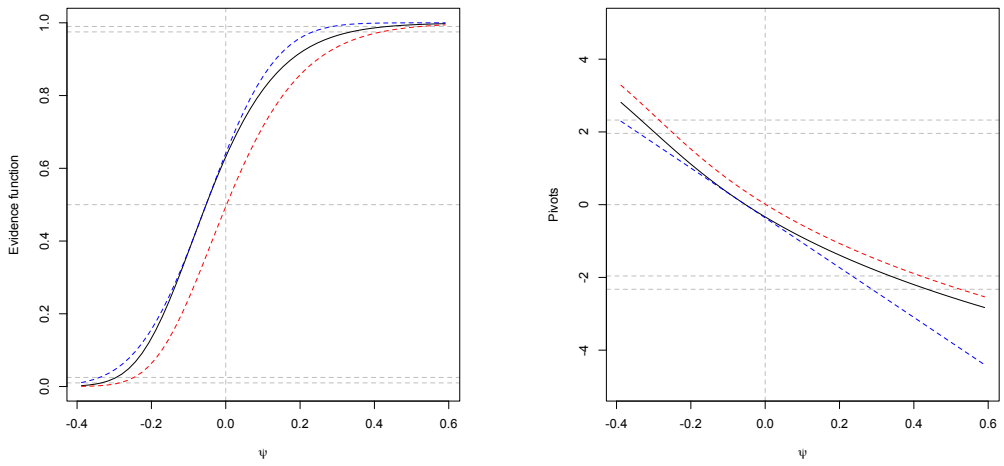


Figure 3.20 – Summaries for the precipitation data $\mathscr{D}_3$. Left: evidence function based on likelihood root $r(\psi)$ (black), Wald statistic $w(\psi)$ (dashed blue), and modified likelihood root $r^*(\psi)$ (dashed red). Right: pivots on standard normal scale.

3.7 Conclusion

In this chapter we discussed the finite-sample distribution of the likelihood root statistic for testing a null hypothesis when the parameter of interest is on the boundary of its domain. In dealing with boundary problems, we distinguished between soft and hard boundaries, where the latter are, as their name implies, more challenging to handle. Our numerical results suggest that first-order approximations perform poorly even when the sample size is very large. This is consistent with conclusions drawn from previous works such as Hosking (1984), Davison (2003), and Crainiceanu and Ruppert (2004a).

Then we considered approximating the distribution of the profile score under the null hypothesis. The reason for deriving analytic expressions for its moments is that these expressions can often reveal which aspects of the model have the greatest impact on the poor reliability of the zeroth-order approximation. These corrections give probabilities close to simulation-based results for the examples we considered. However, the Edgeworth expansion is not fully effective in this context.

Under such nonstandard settings, we applied the tangent exponential model, which is known to produce good results even in small samples (Guedes et al., 2020; Brazzale et al., 2007; Severini, 2000; Fraser, 2017). The results demonstrate the adequacy of the approach, especially for soft boundary problems such as testing for a zero variance component in a normal model. This is equivalent to testing whether a spline expansion is needed in semi-parametric regression using the mixed-model representation. In these cases, the proportion of positive estimates is closer to 1/2, the finite-sample approximation to the distribution of the non-zero part of r^* is closer to the standard normal distribution, and the error of one-sided confidence intervals is close to the nominal value. While the extension from one to multiple-classification models seems straightforward for independent random effects, it is more complex with random coefficients. For example, testing a null random slope would imply setting two parameters at zero, and the mixture distribution of the likelihood root becomes more complex. In this case, it is adequate to use a higher order approximation for a vector-valued parameter of interest such as Bartlet or Skovgaard corrections to the likelihood ratio statistic (Skovgaard, 2001) which have been studied for different classes of models; see Skovgaard (1996), Ferrari and Cysneiros (2008) and Melo et al. (2009). Exploring directional inference as in Fraser et al. (2016b) for boundary problems can be a future extension of this work. Generalized linear mixed models seem out of reach at the moment for multiple reasons (Breslow and Clayton, 1993; Bonat and Ribeiro Jr, 2016). First, the calculation of higher-order solutions requires numerical differentiation of the pivotal quantities and relies heavily on the

eigenvalues of the design matrices; the output is not always stable. Also, this can be time-consuming if the model contains several parameters and the matrices are large. The choice of the link function and its effect on the distribution of interest still need to be addressed.

When the parameter space is not expandable, the tangent exponential model still improves first-order results. As $\psi \to \psi_0$, the distribution of the non-zero part is closer to the standard normal for all studied examples. However, the inflated point mass at ψ_0 is a limitation, as r^* has a singularity at zero. In this case, we seek to improve the finite -sample approximations using different approaches.

Studying higher-order adjustments for boundary problems for various models with different fields of applications is a step toward better understanding a broad class of irregular models. One such application is testing that the miss distance equals a safety threshold for the conjunction assessment presented in Chapter 2. It was shown that this is a boundary problem for large relative uncertainties. Further work to improve these new developments will be very valuable.

3.8 Appendices of Chapter 3

3.8.1 Appendix A: Score of linear mixed model

Partial derivative of the restricted log likelihood defined in (3.18) with respect to ψ is

$$\frac{\partial \ell_{\mathrm{R}}\left(\psi, \widehat{\sigma}^2_{\psi,\mathrm{R}}\right)}{\partial \psi} = -\frac{1}{2}\left\{(n-p)\frac{\partial \log \widehat{\sigma}^2_{\psi,\mathrm{R}}}{\partial \psi} - \frac{\partial \log|\Delta(\psi)|}{\partial \psi} + \frac{\partial \log|X^{\mathrm{T}}\Delta(\psi)X|}{\partial \psi}\right\}.$$

Using matrix derivatives that involve the log determinant, we have

$$\begin{aligned}
\frac{\partial \log \widehat{\sigma}^2_{\psi,\mathrm{R}}}{\partial \psi} &= -\frac{1}{\widehat{\sigma}^2_{\psi,\mathrm{R}}}(y - X\widehat{\beta}_{\psi})^{\mathrm{T}} M(\psi)(y - X\widehat{\beta}_{\psi}), \\
\frac{\partial \log|\Delta(\psi)|}{\partial \psi} &= -\sum_{j=1}^{n}\frac{\lambda_j}{1+\psi\lambda_j}, \\
\frac{\partial \log|X^{\mathrm{T}}\Delta(\psi)X|}{\partial \psi} &= -\operatorname{tr}\left\{(X^{\mathrm{T}}\Delta(\psi)X)^{-1}X^{\mathrm{T}}\Delta(\psi)ZZ^{\mathrm{T}}\Delta(\psi)X\right\},
\end{aligned}$$

where $\widehat{\sigma}^2_{\psi,\mathrm{R}} = (y - X\widehat{\beta}_{\psi})^{\mathrm{T}}\Delta(\psi)(y - X\widehat{\beta}_{\psi})/(n-p)$, λ_j is the j-th eigenvalue of the matrix ZZ^{T}, and $M(\psi) = \left\{-2\Delta(\psi)X\left(X^{\mathrm{T}}\Delta(\psi)X\right)^{-1}X^{\mathrm{T}} + I_n\right\}\Delta(\psi)ZZ^{\mathrm{T}}\Delta(\psi)$.

The score of the profile restricted log likelihood evaluated at $\psi = 0$ is

$$\begin{aligned}
\left.\frac{\partial \ell_{\mathrm{R}}\left(\psi, \widehat{\sigma}^2_{\psi,\mathrm{R}}\right)}{\partial \psi}\right|_{\psi=0} &= -\frac{1}{2}\left\{-(n-p)\frac{y^{\mathrm{T}}(I-H)^{\mathrm{T}}(-2H+I_n)ZZ^{\mathrm{T}}(I-H)y}{y^{\mathrm{T}}(I-H)y} + \operatorname{tr}\left\{(I-H)ZZ^{\mathrm{T}}\right\}\right\} \\
&= -\frac{1}{2}\left[-(n-p)\frac{y^{\mathrm{T}}(I-H)ZZ^{\mathrm{T}}(I-H)y}{y^{\mathrm{T}}(I-H)y} + \operatorname{tr}\left\{(I-H)ZZ^{\mathrm{T}}\right\}\right],
\end{aligned}$$

where $\widehat{\beta}_0 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ and $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is usual hat matrix in a linear fit. Similar calculations are performed when we consider the score based on the ordinary log likelihood.

3.8.2 Appendix B: Moments of the profile score

Below, we give calculations for the score in the examples presented in Section 3.4.

Variance components

For the one-way classification model, using the ordinary likelihood in (3.4), we take $\theta = (\lambda^{\mathrm{T}}, \psi)^{\mathrm{T}}$ where $\lambda = (\mu, \sigma^2)$. First and second-order derivatives of the ordinary log likelihood are

$$\ell_\mu = \frac{km(\bar{y}_{..} - \mu)}{\sigma^2(1+m\psi)}, \quad \ell_{\sigma^2} = -\frac{1}{2}\left\{\frac{mk}{\sigma^2} - \frac{C_2}{\sigma^4} - \frac{C_1}{\sigma^4(1+m\psi)} - \frac{mk(\bar{y}_{..}-\mu)^2}{\sigma^4(1+m\psi)}\right\},$$

$$\ell_\psi = -\frac{1}{2}\left\{\frac{mk}{1+m\psi} - \frac{C_1 m}{\sigma^2(1+m\psi)^2} - \frac{m^2 k(\bar{y}_{..}-\mu)^2}{\sigma^2(1+m\psi)^2}\right\},$$

$$\ell_{\mu\mu} = -\frac{mk}{\sigma^2(1+m\psi)}, \quad \ell_{\mu\sigma^2} = -\frac{mk(\bar{y}_{..}-\mu)}{\sigma^4(1+m\psi)}, \quad \ell_{\mu\psi} = -\frac{m^2 k(\bar{y}_{..}-\mu)}{\sigma^2(1+m\psi)^2},$$

$$\ell_{\sigma^2\sigma^2} = \frac{mk}{2\sigma^4} - \frac{C_2}{\sigma^6} - \frac{C_1}{\sigma^6(1+m\psi)} - \frac{mk(\bar{y}_{..}-\mu)^2}{\sigma^6(1+m\psi)},$$

$$\ell_{\sigma^2\psi} = -\tfrac{1}{2}\left\{\frac{mC_1}{\sigma^4(1+m\psi)^2} + \frac{m^2 k(\bar{y}_{..}-\mu)^2}{\sigma^4(1+m\psi)^2}\right\},$$

$$\ell_{\psi\psi} = \frac{m^2 k}{2(1+m\psi)^2} - \frac{m^2 C_1}{\sigma^2(1+m\psi)^3} - \frac{m^3 k(\bar{y}_{..}-\mu)^2}{\sigma^2(1+m\psi)^3}.$$

The expected information matrix is

$$i(\theta) = \frac{1}{2\sigma^4(1+m\psi)^2}\begin{bmatrix} 2mk\sigma^2(1+m\psi) & 0 & 0 \\ 0 & mk(1+m\psi)^2 & mk\sigma^2(1+m\psi) \\ 0 & mk\sigma^2(1+m\psi) & m^2k\sigma^4 \end{bmatrix}.$$

The inverse of the (λ, λ) block is

$$i_\lambda(\theta)^{-1} = \begin{bmatrix} \dfrac{\sigma^2(1+m\psi)}{mk} & 0 \\ 0 & \dfrac{2\sigma^4}{mk} \end{bmatrix}.$$

The moments of the profile score $U_{\mathrm{p}} = \partial\ell_{\mathrm{p}}/\partial\psi$ are

$$\mathrm{E}(U_{\mathrm{p}}) = -\frac{m-1}{2}, \quad \mathrm{var}(U_{\mathrm{p}}) = \frac{mk(m-1)}{2},$$

$$\mathrm{E}\left\{(U_{\mathrm{p}})^3\right\} = \frac{mk(m^2-3m+2)}{4}, \quad \rho_3 = \frac{m^3(8k+3) - m^2(18k+9) + m(10k+9) - 3}{2\sqrt{2}\sqrt{m^3k^3(m-1)}}.$$

In the restricted model, $\lambda = \psi$, and partial derivatives of the restricted log likelihood $\theta = (\sigma^2, \psi)^{\mathrm{T}}$ are

$$\begin{aligned}
\ell_{\sigma^2} &= -\frac{mk-1}{2\sigma^2} + \frac{C_2}{2\sigma^4} + \frac{C_1}{2\sigma^4\left(1+m\psi\right)}, \\
\ell_{\psi} &= -\frac{(k-1)\,m}{2(1+m\psi)} + \frac{C_1\,m}{2\sigma^2\left(1+m\psi\right)^2}, \\
\ell_{\sigma^2\sigma^2} &= \frac{mk-1}{2\sigma^4} - \frac{C_2}{\sigma^6} - \frac{C_1}{\sigma^6\left(1+m\psi\right)}, \\
\ell_{\sigma^2\psi} &= -\frac{C_1\,m}{2\sigma^4\left(1+m\psi\right)^2}, \\
\ell_{\psi\psi} &= \frac{(k-1)\,m^2}{2\left(1+m\psi\right)^2} - \frac{C_1\,m^2}{\sigma^2\left(1+m\psi\right)^3}.
\end{aligned}$$

The expected information matrix is

$$i(\theta) = \frac{1}{2\sigma^4(1+m\psi)^2}\begin{bmatrix} (mk-1)\left(1+m\psi\right)^2 & m\,(k-1)\,\sigma^2(1+m\psi) \\ m\,(k-1)\,\sigma^2(1+m\psi) & m^2(k-1)\sigma^4 \end{bmatrix},$$

from which we obtain

$$\kappa^{\sigma^2,\sigma^2} = 2\,\frac{\sigma^4}{mk-1}.$$

Moments of the profile score under the null hypothesis

$$\begin{aligned}
\mathrm{E}\left(U_{\mathrm{p}}\right) &= 0, \quad \mathrm{var}\left(U_{\mathrm{p}}\right) = \frac{(k-1)\,m^2\,\{k\,(m-2)+1\}}{k(m-1)}, \\
\mathrm{E}\left\{\left(U_{\mathrm{p}}\right)^3\right\} &= \frac{m^3(k-1)\left(4k^2m^2-27k^2m-19k^2+19mk+65k-42\right)}{4(mk-1)^2}, \\
\rho_3 &= \frac{\left(4m^2-27m-19\right)k^2+(19m+65)k-42}{2\sqrt{k^3(m-1)^3(mk-1)(k-1)}}.
\end{aligned}$$

Generalized Pareto distribution

Using an expansion of the log likelihood function (3.28) around small ξ, we have

$$\begin{aligned}
\ell_{\xi} &= \frac{y(y-2\sigma)}{2\sigma^2}, \quad \ell_{\sigma} = \frac{y-\sigma}{\sigma^2}, \\
\ell_{\xi\xi} &= \frac{y^2(3\sigma-2y)}{3\sigma^3}, \quad \ell_{\sigma\sigma} = \frac{y(\sigma-y)}{\sigma^3}, \quad \ell_{\xi\sigma} = \frac{y(\sigma-y)}{\sigma^3}.
\end{aligned}$$

The expected information matrix for $\theta = (\xi, \sigma)$ at $(0, \sigma)$ is

$$i(\theta) = \begin{bmatrix} 2 & \sigma^{-1} \\ \sigma^{-1} & \sigma^{-2} \end{bmatrix}.$$

Under the null hypothesis, the profile score has the following moments

$$\mathrm{E}\big(U_\mathrm{p}\big) = -1, \quad \mathrm{var}\big(U_\mathrm{p}\big) = n, \quad \mathrm{E}\Big\{\big(U_\mathrm{p}\big)^3\Big\} = 7n, \quad \rho_3 = \frac{13}{\sqrt{n}}.$$

Student t distribution

Partial derivatives of the log likelihood using a Taylor series expansion of $\log f(\theta)$ defined in (3.24) for $\theta = (\mu, \sigma^2, \psi)$ around $\psi = 0$ gives

$$\begin{aligned} \ell_\mu &= \frac{z}{\sigma}, \quad \ell_{\sigma^2} = \frac{z^2-1}{2\sigma^2}, \quad \ell_\psi = \frac{z^4 - 2z^2 - 1}{4}, \\ \ell_{\mu\mu} &= -\frac{1}{\sigma^2}, \quad \ell_{\sigma^2\sigma^2} = -\frac{2z^2-1}{2\sigma^4}, \quad \ell_{\psi\psi} = \frac{z^4}{2} - \frac{z^6}{3}, \\ \ell_{\mu\psi} &= \frac{z - z^3}{\sigma}, \quad \ell_{\sigma^2\psi} = \frac{z^2 - z^4}{2\sigma^2}, \quad \ell_{\mu\sigma^2} = -\frac{z}{\sigma^3}. \end{aligned}$$

where $z = (y-\mu)/\sigma \sim \mathcal{N}(0,1)$. The m-th moment of a standard normal variable is

$$\mu_m = \mathrm{E}\big(Z^m\big) = \begin{cases} 0, & m \text{ odd}, \\ 2^{-m/2}\dfrac{m!}{(m/2)!}, & m \text{ even}, \end{cases}$$

giving

$$\mu_2 = 1, \quad \mu_4 = 3, \quad \mu_6 = 15, \quad \mu_8 = 105, \quad \mu_{10} = 945, \quad \mu_{12} = 10395.$$

The expected information matrix and its inverse at $\psi = 0$ are

$$i(\theta) = \begin{bmatrix} \dfrac{1}{\sigma^2} & 0 & 0 \\ 0 & \dfrac{1}{2\sigma^4} & \dfrac{1}{\sigma^2} \\ 0 & \dfrac{1}{\sigma^2} & \dfrac{7}{2} \end{bmatrix}, \quad i_\lambda(\theta)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

Under the null hypothesis, the profile score satisfies

$$\mathrm{E}\big(U_\mathrm{p}\big) = -\frac{3}{2}, \quad \mathrm{var}\big(U_\mathrm{p}\big) = \frac{3}{2}n, \quad \mathrm{E}\Big\{\big(U_\mathrm{p}\big)^3\Big\} = \frac{117n}{4}, \quad \rho_3 = \frac{41}{4}\sqrt{\frac{6}{n}}.$$

Negative binomial distribution

A Taylor series expansion of the negative binomial density in (3.25) around the origin $\psi = 1/\nu$ gives

$$\begin{aligned}\ell_\psi &= \tfrac{1}{2}\left(\xi^2 - 2\xi y + (-1+y)y\right), \quad \ell_\xi = -1 + \frac{y}{\xi},\\ \ell_{\psi\psi} &= \frac{1}{6}\left(-4\xi^3 + 6\xi^2 y + y\left(-1 + 3y - 2y^2\right)\right),\\ \ell_{\xi\xi} &= -\frac{y}{\xi^2}, \quad \ell_{\psi\xi} = \xi - \mathrm{y}.\end{aligned}$$

The expected information matrix for $\theta = (\psi, \xi)$ is

$$i(\theta) = \begin{bmatrix} \dfrac{\xi^2}{2} & 0 \\ 0 & \dfrac{\xi^2\psi^2 - \xi\psi + 1}{\xi} \end{bmatrix},$$

The profile score has the following moments

$$\mathrm{E}\left(U_\mathrm{p}\right) = -\frac{\xi}{2}, \quad \mathrm{var}\left(U_\mathrm{p}\right) = \frac{n\xi^2}{2}, \quad \mathrm{E}\left\{\left(U_\mathrm{p}\right)^3\right\} = \frac{1}{4}n\xi^2(\xi+2), \quad \rho_3 = \frac{11\xi + 4}{2\sqrt{2n}\xi}.$$

3.8.3 Appendix C: Components of the tangent exponential model

Variance components

For the one-way classification model, using the restricted likelihood, the parameter vector is $\theta = (\psi, \sigma^2)$, the data is $y = (y_1, y_2) = (C_1, C_2)$. We consider the following pivots to define the sufficient directions

$$z_1(\theta, y) = \frac{C_1}{\sigma^2(1+m\psi)}, \quad z_2(\theta, y) = \frac{C_2}{\sigma^2},$$

which gives

$$v_\psi = \left\{\frac{m}{m-1}\frac{k-1}{k}C_2, k(m-1)\frac{C_1}{C_2}\right\}, \quad v_{\sigma^2} = \{0, k(m-1)\}.$$

The likelihood derivatives with respect to the data are

$$\frac{\partial \ell_\mathrm{R}}{\partial y_1} = -\frac{1}{2\sigma^2(1+m\psi)}, \quad \frac{\partial \ell_\mathrm{R}}{\partial y_2} = -\frac{1}{2\sigma^2},$$

which are weighted by the columns of V in order to produce the canonical parameter $\varphi(\theta) = (\varphi_1(\theta), \varphi_2(\theta))^T$, where

$$\varphi_1(\theta) = -\frac{1}{2\sigma^2}\frac{m}{m-1}\frac{k-1}{k}\frac{C_2}{1+m\psi},$$
$$\varphi_2(\theta) = -\frac{1}{2\sigma^2}k(m-1)\left(1+\frac{1}{1+m\psi}\frac{C_1}{C_2}\right).$$

The observed information matrix evaluated at the MLE is

$$j(\widehat{\theta}) = \frac{1}{2}\begin{bmatrix} \frac{m^2}{(m-1)^2}\frac{(k-1)^3}{k^2}\left(\frac{C_2}{C_1}\right)^2 & \frac{m(k-1)^2}{C_1} \\ \frac{m(k-1)^2}{C_1} & \frac{k^2(m-1)^2(km-1)}{C_2^2} \end{bmatrix}.$$

The other quantity needed to compute to compute $q(\psi)$ is the matrix

$$\varphi_\theta(\theta) = \frac{1}{2\sigma^2}\begin{bmatrix} \frac{m^2}{m-1}\frac{k-1}{k}\frac{C_2}{(1+m\psi)^2} & \frac{1}{\sigma^2}\frac{m}{m-1}\frac{k-1}{k}\frac{C_2}{(1+m\psi)} \\ \frac{mk(m-1)}{(1+m\psi)^2}\frac{C_1}{C_2} & \frac{k(m-1)}{\sigma^2}\left(1+\frac{C_1}{C_2(1+m\psi)}\right) \end{bmatrix},$$

the second column of which contains $\varphi_\lambda(\theta)$ and should be evaluated at

$$\widehat{\sigma}^2_\psi = \frac{1}{km-1}\left(C_2 + \frac{C_1}{1+m\psi}\right).$$

Plugging in all expressions, we obtain

$$r(\psi) \equiv \text{sign}(\widehat{\psi}-\psi)\left[(km-1)\log\left\{C_1+C_2\left(1+m\psi\right)\right\} - k(m-1)\log(1+m\psi)\right]^{1/2},$$
$$q(\psi) = \frac{k(m-1)C_1-(k-1)(1+m\psi)C_2}{C_1+C_2(1+m\psi)}\left\{\frac{1}{2}\frac{(km-1)}{k(k-1)(m-1)}\right\}^{1/2}. \tag{3.29}$$

If instead, we consider the ordinary likelihood, we use the extra pivot

$$z_3(\theta, y) = km\frac{(\bar{y}_{..}-\mu)^2}{\sigma^2(1+m\psi)}.$$

In this case, we obtain

$$r(\psi) \equiv \text{sign}(\widehat{\psi}-\psi)\left[mk\log\left\{C_1+C_2(1+m\psi)\right\} - k(m-1)\log(1+m\psi)\right]^{1/2}, \tag{3.30}$$
$$q(\psi) = m\left\{(m-1)\,C_1 - C_2\left(1+m\psi\right)\right\}\left[\frac{k}{2(m-1)}\frac{C_1}{\left\{C_1+C_2(1+m\psi)\right\}^3}\right]^{1/2},$$

so, we have closed-form expressions for the quantities needed to compute $r^*(\psi)$ when using the usual or the restricted likelihoods.

A straightforward extension of the previous model is a two-way classification model with no interaction for which $\theta = (\psi_1, \psi_2, \sigma^2)^{\mathrm{T}}$ and $y = (C_1, C_2, C_3)$. We consider the following three pivots,

$$z_1(\theta, y) = \frac{C_1}{\sigma^2(1+m\psi_1)}, \quad z_2(\theta, y) = \frac{C_2}{\sigma^2(1+k\psi_2)}, \quad z_3(\theta, y) = \frac{C_3}{\sigma^2}.$$

This gives

$$\begin{aligned} v_{\psi_1} &= \left\{ \frac{m}{m-1} C_3, 0, (k-1)(m-1)\frac{C_1}{C_3} \right\}, \\ v_{\psi_2} &= \left\{ 0, \frac{k}{k-1} C_2, (k-1)(m-1)\frac{C_2}{C_3} \right\}, \\ v_{\sigma^2} &= \{0, 0, (k-1)(m-1)\} \ . \end{aligned}$$

The corresponding log likelihood derivatives with respect to the data are

$$\frac{\partial \ell_{\mathrm{R}}}{\partial y_1} = -\frac{1}{2\sigma^2(1+m\psi_1)}, \quad \frac{\partial \ell_{\mathrm{R}}}{\partial y_2} = -\frac{1}{2\sigma^2(1+k\psi_2)}, \quad \frac{\partial \ell_{\mathrm{R}}}{\partial y_3} = -\frac{1}{2\sigma^2}.$$

Components of the three-dimensional canonical parameter are

$$\begin{aligned} \varphi_1(\theta) &= -\frac{1}{2}\frac{m}{m-1}\frac{C_3}{\sigma^2(1+m\psi_1)}, \\ \varphi_2(\theta) &= -\frac{1}{2}\frac{k}{k-1}\frac{C_3}{\sigma^2(1+k\psi_2)}, \\ \varphi_3(\theta) &= -\frac{1}{2}\frac{(m-1)(k-1)}{\sigma^2}\left\{ \frac{C_1}{1+m\psi_1} + \frac{C_2}{1+k\psi_2} + 1 \right\}. \end{aligned}$$

These expressions give

$$\begin{aligned} r(\psi_1) &\equiv \mathrm{sign}(\widehat{\psi}_1 - \psi_1)\left[m(k-1)\log\left\{ \left(1+m\psi_1\right) C_3 + C_1 + \right\} - (k-1)(m-1)\log\left(m\psi_1 + 1\right) \right]^{1/2}, \\ q(\psi_1) &= \left\{ \frac{m(k-1)}{2(m-1)} \right\}^{1/2} \frac{(m-1)C_1 - (1+m\psi_1)C_3}{C_3(1+m\psi_1) + C_1}. \end{aligned} \tag{3.31}$$

Linear mixed models

In the previous section, we showed that for one- and two-way classification models, closed-form expressions of $r(\theta), q(\psi)$, and $r^*(\psi)$ are available. In a more general context, as in model (3.13), for $\theta = (\psi, \beta, \sigma^2)$, we take the pivots to be the elements of

the $n \times 1$ vector of scaled martingale differences $z(y;\theta) = \Delta(\psi)^{1/2}(y - X\beta)/\sigma$, which have independent standard normal distributions. Derivatives of the pivotal quantities give

$$v_\beta = X\,, \quad v_{\sigma^2} = \frac{(y - X\beta)}{2\sigma^2}\bigg|_{(y^o,\widehat{\theta}^o)}.$$

To obtain the column vector associated with ψ, we consider the eigenvalue decomposition of $\Delta(\psi)^{-1} = P\Lambda(\psi)P^{\mathrm{T}}$, where P is independent of ψ, and $\Lambda(\psi)$ is diagonal matrix with elements $1 + \psi\lambda_j$. Using this decomposition, we have

$$\frac{\partial z}{\partial \psi} = P\tilde{\Lambda}(\psi)P^{\mathrm{T}},$$

where $\tilde{\Lambda}(\psi)$ is a diagonal with elements $\lambda_j/\{2(1+\psi\lambda)^{3/2}\}$. This gives

$$v_\psi = P\Lambda(\psi)^{1/2}\tilde{\Lambda}(\psi)P(y - X\beta)\big|_{(y^o,\widehat{\theta}^o)}.$$

The partial derivative of the ordinary log likelihood with respect to the data is

$$\frac{\partial \ell(\theta)}{\partial y}\bigg|_{(y=y^o)} = -\frac{1}{2\sigma^2}(y^o - X\beta)^{\mathrm{T}}\Delta(\psi),$$

We then have all that is needed to obtain a local parametrization $\varphi(\theta)$.

For the restricted likelihood model described in (3.18), we can take the pivotal quantities to be $\sigma^{-1}\Sigma(\psi)^{-1/2}e \sim \mathcal{N}_n(0, I_n)$, where e is the residual vector $e = y - X\widehat{B} = H(\psi)y$, $H(\psi) = I_n - X\{X^{\mathrm{T}}\Delta(\psi)X\}^{-1}X^{\mathrm{T}}\Delta(\psi)$, and $\Sigma(\psi) = H(\psi)\Delta(\psi)^{-1}H(\psi)^{\mathrm{T}}$. We then proceed as under the ordinary likelihood.

Generalized Pareto distribution

Suppose that $\{y_i\}_{i=1}^n$ are samples of GP(ξ,σ). For $\theta = (\psi,\lambda)$ where $\psi = \xi$ and $\lambda = \sigma$, let the CDF of y_i be the pivot $z_i(\theta) = GP(y_i;\theta)$, then we have

$$\mathrm{v}_\xi = \frac{y_i}{\sigma}\bigg|_{(y^o,\widehat{\theta}^o)}, \quad \mathrm{v}_\sigma = \frac{1}{\xi}\frac{y_i}{\sigma + \xi y_i} - \frac{1}{\xi^2}\log\Big(1 + \xi\frac{y_i}{\sigma}\Big)\bigg|_{(y^o,\widehat{\theta}^o)}.$$

while the derivative of the log likelihood with respect to y_i is

$$\frac{\partial \ell}{\partial y_i}\bigg|_{(y_i=y_i^o)} = \frac{1+\xi}{\sigma + \xi y_i^o}.$$

These expressions can be substituted into $\varphi(\theta)$ to obtain the modified likelihood root.

Student t distribution

For the Student t example, we use the cumulative distribution function $F(y_i,\theta)$ as pivotal statistics for y_i, where $\theta^T=(\psi,\lambda)$ and $\lambda=(\mu,\sigma^2)$. The CDF can be written as

$$\begin{aligned}F(y;\psi,\mu,\sigma^2)=&\frac{\Gamma\left(\frac{\psi+1}{2\psi}\right)\sqrt{\psi}}{\Gamma\left(\frac{1}{2\psi}\right)\sqrt{\pi}}\int_{-\infty}^{\frac{y-\mu}{\sigma}}\left(1+\psi x\right)^{-\frac{\psi+1}{2\psi}}\mathrm{d}x\\=&\frac{\sqrt{\psi}}{B(\frac{1}{2},\frac{1}{2\psi})}\int_{-\infty}^{\frac{y-\mu}{\sigma}}\left(1+\psi x\right)^{-\frac{\psi+1}{2\psi}}\mathrm{d}x\\=&A(\psi)C(z,\psi,\mu,\sigma),\end{aligned}$$

where $z=(y-\mu)/\sigma$, $A(\psi)$ is a normalization constant depending only on ψ, and $C(y,\psi,\mu,\sigma)$ is the integral term. For small ψ, $1/\psi\to\infty$, we use the Stirling approximate to give an asymptotic formula for the Beta function. Then

$$A(\psi)=\frac{\sqrt{\psi}}{B(\frac{1}{2},\frac{1}{2\psi})}\doteq\frac{\sqrt{\psi}}{\Gamma\left(\frac{1}{2}\right)\sqrt{2\psi}}=\frac{1}{\sqrt{2\pi}}.$$

On the other hand

$$\begin{aligned}C(z,\theta)=&\int_{-\infty}^{z}\left(1+\psi x\right)^{-\frac{\psi+1}{2\psi}}\mathrm{d}x\\=&\int_{-\infty}^{z}\exp\left\{-\frac{1+\psi}{2\psi}\log(1+\psi x^2)\right\}\mathrm{d}x\\=&\int_{-\infty}^{z}\exp\left[-\frac{1+\psi}{2\psi}\left\{\psi x^2-\frac{\psi^2x^4}{2}+o\left(\psi^2\right)\right\}\right]\mathrm{d}x\\=&\int_{-\infty}^{z}\exp\left\{-\frac{x^2}{2}+\frac{\psi}{2}\left(\frac{x^4}{2}-x^2\right)+o\left(\psi^2\right)\right\}\mathrm{d}x\\\doteq&\sqrt{2\pi}\Phi\left(\frac{y-\mu}{\sigma}\right)+\frac{\sqrt{2\pi}\psi}{2}\int_{-\infty}^{z}\phi(x)\left(\frac{x^4}{2}-x^2\right)\mathrm{d}x\\=&\sqrt{2\pi}\Phi\left(z\right)+\frac{\sqrt{2\pi}\psi}{4}\left\{-z^3\phi(z)-z\phi(z)+\Phi(z)\right\}.\end{aligned}$$

Using the approximation for the density and its integral around the boundary, partial derivatives of the pivots are

$$\begin{aligned}\frac{\partial F}{\partial \psi} &= \frac{1}{4}\left\{-z^3\phi(z) - z\phi(z) + \Phi(z)\right\},\\ \frac{\partial F}{\partial \mu} &= -f(y;\psi,\mu,\sigma^2),\\ \frac{\partial F}{\partial \sigma^2} &= -\frac{y-\mu}{2\sigma^2}f(y;\psi,\mu,\sigma^2).\end{aligned}$$

This yields the following columns for the matrix V,

$$v_\psi = \frac{1}{4}\left\{z^3 + z - \frac{\Phi(z)}{\phi(z)}\right\}\Bigg|_{(z^\mathrm{o},\widehat{\theta}^\mathrm{o})}, \quad v_\mu = 1_n, \quad v_{\sigma^2} = \frac{z}{2\sigma}\Big|_{(z^\mathrm{o},\widehat{\theta}^\mathrm{o})}.$$

For larger values of ψ, i.e., far from the boundary, we directly differentiate the CDF with respect to the parameter of interest, and the expression for v_ψ becomes

$$\begin{aligned}v_\psi = -\sigma^{-\frac{1+\psi}{\psi}}(y-\mu)\Bigg[&\psi\sigma^{\frac{1+\psi}{\psi}} + \left\{\sigma^2 + \psi(y-\mu)^2\right\}^{\frac{1+\psi}{2\psi}}\\ &\left\{{}_2F_1\left[\begin{matrix}\frac{1}{2}, \frac{1+\psi}{2\psi}\\ \frac{3}{2}\end{matrix}; -\frac{\psi(y-\mu)^2}{\sigma^2}\right]\left\{\zeta\left(\frac{1}{2\psi}\right) - \zeta\left(\frac{1+\psi}{\psi}\right)\right\} - {}_2F_1^b\left[\begin{matrix}\frac{1}{2}, \frac{1+\psi}{2\psi}\\ \frac{3}{2}\end{matrix}; -\frac{\psi(y-\mu)^2}{\sigma^2}\right]\right\}\Bigg]\Bigg|_{(y^\mathrm{o},\widehat{\theta}^\mathrm{o})},\end{aligned}$$

where $\zeta(\cdot)$ is the digamma function, ${}_2F_1\left[\begin{smallmatrix}a & b\\ & b\end{smallmatrix}; x\right]$ and ${}_2F_1^b\left[\begin{smallmatrix}a & b\\ & c\end{smallmatrix}; x\right]$ denote respectively the hypergeometric function with parameter (a,b,c) and its derivative with respect to b, both evaluated at x. To define the canonical parameter $\varphi(\theta)$, we also require the derivative of log likelihood with respect to the data.

$$\frac{\partial \ell}{\partial y_i}\Big|_{y_i = y_i^\mathrm{o}} = \frac{(1+\psi)(\mu - y_i^\mathrm{o})}{\psi(y-\mu)^2 + \sigma^2}.$$

Negative binomial distribution

A complete account of higher-order approximations for discrete data, particularly count and contingency table data is given in Davison et al. (2006). Recall that the log likelihood function for a single observation of a negative binomial variable is

$$\ell(\theta) = A(y+\nu) - A(\nu) + \nu\log(\nu) + y\log(\xi) - (\nu+y)\log(\nu+\xi),$$

where $A(\nu) = \log\{\Gamma(\nu)\}$ is the log-gamma function. Differentiating the log likelihood with respect to ξ and ν at $\left(\widehat{\xi}^{\mathrm{o}}, \widehat{\nu}^{\mathrm{o}}\right)$ gives $s = (s_1, s_2)$ where

$$s^1 = \left.\frac{\partial l}{\partial \xi}\right|_{\widehat{\theta}^{\mathrm{o}}} = \frac{y}{\widehat{\xi}^{\mathrm{o}}} - \frac{\left(\widehat{\nu}^{\mathrm{o}} + y\right)}{\left(\widehat{\nu}^{\mathrm{o}} + \widehat{\xi}^{\mathrm{o}}\right)},$$
$$s^2 = \left.\frac{\partial l}{\partial \nu}\right|_{\widehat{\theta}^{\mathrm{o}}} = \zeta\left(\widehat{\nu}^{\mathrm{o}} + y\right) - \zeta\left(\widehat{\nu}^{\mathrm{o}}\right) + 1 + \log\left(\widehat{\nu}^{\mathrm{o}}\right) - \log\left(\widehat{\nu}^{\mathrm{o}} + \widehat{\xi}^{\mathrm{o}}\right) - \frac{\left(\widehat{\nu}^{\mathrm{o}} + y\right)}{\left(\widehat{\nu}^{\mathrm{o}} + \widehat{\xi}^{\mathrm{o}}\right)}.$$

The s^i is affinely equivalent to the simpler form $\left(y_i,\ \zeta\left(\widehat{\nu}^{\mathrm{o}} + y_i\right) - \dfrac{y_i + \widehat{\nu}^{\mathrm{o}}}{\widehat{\xi}^{\mathrm{o}} + \widehat{\nu}^{\mathrm{o}}}\right)^{\mathrm{T}}$, where $\zeta(\nu)$ is the digamma function.

For the i-th observation in a sample of size n, the columns of V_i are

$$v_\xi^i = \left\{\frac{\widehat{\nu}^{\mathrm{o}}}{\widehat{\xi}^{\mathrm{o}}\left(\widehat{\xi}^{\mathrm{o}} + \widehat{\nu}^{\mathrm{o}}\right)},\ 0\right\}, \quad v_\nu^i = \left[0,\ \zeta'\left(\widehat{\nu}^{\mathrm{o}}\right) - \tfrac{1}{\widehat{\nu}^{\mathrm{o}}} - E\left\{\zeta'\left(\widehat{\nu}^{\mathrm{o}} + y_i\right)\right\} + \frac{1}{\widehat{\nu}^{\mathrm{o}} + \widehat{\xi}^{\mathrm{o}}}\right].$$

The matrix V^i is diagonal since the parameters ν and ξ are orthogonal. Partial derivatives of the log likelihood are

$$\frac{\partial \ell(\theta)}{\partial s_i^j} = \left\{\frac{\partial \ell(\theta)}{\partial y_i}\right\}\left(\frac{\partial s_i^j}{\partial y_i}\right)^{-1}, \quad j = 1, 2.$$

The two-dimensional canonical parameter $\varphi(\theta)$ is

$$\varphi_1(\theta) = \sum_{i=1}^{n}\left\{\zeta\left(\nu + y_i^{\mathrm{o}}\right) + \log\left(\frac{\xi}{\nu + \xi}\right)\right\},$$
$$\varphi_2(\theta) = \sum_{i=1}^{n}\frac{\zeta\left(\nu + y_i^{\mathrm{o}}\right) + \log\{\xi/(\nu + \xi)\}}{\zeta'\left(\widehat{\nu}^{\mathrm{o}} + y_i^{\mathrm{o}}\right) - 1/\left(\widehat{\nu}^{\mathrm{o}} + \widehat{\xi}^{\mathrm{o}}\right)} v_\nu^i.$$

Two-component Gaussian mixture

Assume the pivot for y_i is

$$z_i(\theta, y_i) = \tfrac{1}{2}\varphi_n\left(y_i - \psi 1_p + \lambda\right) + \tfrac{1}{2}\varphi_n\left(y_i - \psi 1_p - \lambda\right).$$

Then V^i is a $p \times (p+1)$ matrix with columns

$$v^i_\psi = \frac{\left(y_i - \psi 1_p - \lambda\right)^T 1_p \varphi\left(y_i - \psi 1_p + \lambda\right) - \left(y_i + \psi 1_p - \lambda\right)^T 1_p \varphi\left(y_i + \psi 1_p + \lambda\right)}{\varphi\left(y_i - \psi 1_p + \lambda\right) + \varphi\left(y_i - \psi 1_p - \lambda\right)},$$
$$v^i_\lambda = I_p.$$

To obtain the $(p+1)$-dimensional canonical parameter $\varphi(\theta)$, we need partial derivatives of the the log likelihood $\ell(\theta) = \sum_{i=1}^{n} \log\left\{\frac{1}{2}\varphi\left(y_i - \psi 1_p + \lambda\right) + \frac{1}{2}\varphi\left(y_i - \psi 1_p - \lambda\right)\right\}$ given by

$$\frac{\partial \ell}{\partial y_i} = -\frac{\left(y_i + \psi 1_p - \lambda\right)^T I_p \varphi\left(y_i + \psi 1_p + \lambda\right) + \left(y_i - \psi 1_p - \lambda\right)^T I_p \varphi\left(y_i - \psi 1_p + \lambda\right)}{\varphi\left(y_i - \psi 1_p + \lambda\right) + \varphi\left(y_i - \psi 1_p - \lambda\right)}.$$

3.8.4 Appendix D: EM algorithm

EM algorithm for Student t

Assume that $Y \sim \mathcal{N}(\mu, \sigma^2/\eta)$, where η is the unobserved variable which follows $\Gamma(\nu/2, \nu/2)$. Then the marginal distribution of Y can be viewed as an infinite mixture of gamma variables as briefly discussed in Section 3.4.2. For $\theta = (\nu, \mu, \sigma)$, the complete-data log likelihood based on a random sample $(y_1, \eta_1), \ldots, (y_n, \eta_n)$ is

$$\sum_{i=1}^{n} \log f(y_i, \eta_i, \theta) = \sum_{i=1}^{n} -\frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log\eta_i - \frac{\eta_i}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2 - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\frac{\nu}{2} + \left(\frac{\nu}{2} - 1\right)\log\eta_i - \frac{\nu}{2}\eta_i.$$

The conditional density $f(\eta|y,\theta)$ follows upon noting that the gamma distribution is the conjugate prior to a normal distribution with shape and scale parameters given respectively by

$$\alpha = \frac{\nu + 1}{2}, \quad \beta = \frac{\nu}{2} + \left(\frac{y - \mu}{\sigma}\right)^2,$$

So, taking the expectation of the complete log likelihood with respect to $f(\eta \mid y, \theta_0)$ gives

$$Q(\theta, \theta_0) = -\frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left(\log\eta_i\right) - \tfrac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2 \mathrm{E}\left(\eta_i\right) - n\log\Gamma\left(\frac{\nu}{2}\right) + \tfrac{n\nu}{2}\log\tfrac{\nu}{2} + \left(\tfrac{\nu}{2} - 1\right)\sum_{i=1}^{n} \mathrm{E}\left(\log\eta_i\right) - \tfrac{\nu}{2}\sum_{i=1}^{n} \mathrm{E}\left(\eta_i\right)$$

where $\mathrm{E}(\eta_i) = \alpha_i^0/\beta_i^0$, and $\mathrm{E}(\log\eta_i) = \zeta(\alpha_i^0) - \log\beta_i^0$, ζ is the digamma function. The superscript $(\cdot)^0$ indicates that the expectations are evaluated with respect to θ_0.

The M-step involves maximizing $Q(\theta,\theta_0)$ where we solve the following equations for θ,

$$\begin{aligned}
\frac{\partial Q}{\partial \mu} &= \sum_{i=1}^{n} \frac{y_i-\mu}{\sigma^2}\,\mathrm{E}\left(\eta_i\right),\\
\frac{\partial Q}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2 \mathrm{E}\left(\eta_i\right),\\
\frac{\partial Q}{\partial \nu} &= -\frac{n}{2}\zeta\left(\frac{\nu}{2}\right) + \frac{n}{2}\log\frac{\nu}{2} + \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left(\log\eta_i\right) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left(\eta_i\right).
\end{aligned}$$

This yields the closed-form solutions

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} y_i\,\mathrm{E}\left(\eta_i\right)}{\sum_{i=1}^{n}\mathrm{E}\left(\eta_i\right)}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\mathrm{E}\left(\eta_i\right),$$

where we require update of $\widehat{\mu}$ in order to update $\widehat{\sigma}^2$, and $\widehat{\nu}$ has to be found numerically by solving

$$\zeta\left(\frac{\nu}{2}\right) - \log\frac{\nu}{2} = 1 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\log\eta_i\right) - \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\eta_i\right).$$

Given values of $\theta_0 = \left(\nu_0,\mu_0,\sigma_0\right)$, the EM algorithm finds ν numerically, and updates to $\theta^\dagger = \left(\nu^\dagger,\mu^\dagger,\sigma^\dagger\right)$. In each iteration, we check for convergence using $|\theta^\dagger - \theta_0|$. If convergence is not yet attained, θ_0 is replaced by $\theta^\dagger$ and the cycle is repeated.

EM algorithm for the two-Gaussian mixture

The derivation of the EM algorithm for the two-Gaussian mixture is a special case of the mixture models discussed in Davison (2003, Example 5.36) and has been derived in Tse and Davison (2022). We have

$$Q\left(\theta;\theta_0\right) = \sum_{r=1}^{c}\left\{\sum_{j=1}^{n} w_r\left(y_j;\theta_0\right)\right\}\log\pi_r + \sum_{r=1}^{c}\sum_{j=1}^{n} w_r\left(y_j;\theta_0\right)\log f_r\left(y_j;\theta\right),$$

where $\theta = (\psi,\lambda)$, c is the number of components, π_r is the probability that a random variable Y comes from the r-th population with a density $f_r(y;\theta)$, and $\mathrm{w_r}\left(\mathrm{y};\theta_0\right) = \pi_\mathrm{r}'\mathrm{f_r}\left(\mathrm{y};\theta_0\right)/\sum_{\mathrm{s}=1}^{\mathrm{c}}\pi_\mathrm{s}'\mathrm{f_r}\left(\mathrm{y};\theta_0\right)$. In our case, we have a mixture of two-Gaussian compo-

nents and known probabilities $\pi_{\mathrm{r}} \equiv 1/2$, so

$$Q(\theta;\theta_0) = -\frac{1}{2}\sum_{j=1}^{n}\left\{ w_1(y_j;\theta_0)\sum_{k=1}^{p}\left[y_{j,k}-(\lambda_k-\psi)\right]^2 + w_2(y_j;\theta_0)\sum_{k=1}^{p}\left[y_{j,k}-(\lambda_k+\psi)\right]^2\right\}.$$

Differentiation yields

$$\frac{\partial}{\partial\psi}Q(\theta;\theta_0) = \sum_{j=1}^{n} w_1(y_j;\theta_0)\left(-\sum_{k=1}^{p} y_{j,k} + \sum_{k=1}^{p}\lambda_k - p\psi\right) + \sum_{j=1}^{n} w_2(y_j;\theta_0)\left(\sum_{k=1}^{p} y_{j,k} - \sum_{k=1}^{p}\lambda_k - p\psi\right),$$

$$\begin{aligned}\frac{\partial}{\partial\lambda_i}Q(\theta;\theta_0) = &\sum_{j=1}^{n} w_1(y_j;\theta_0)\,y_{j,i} - \lambda_i\sum_{j=1}^{n} w_1(y_j;\theta_0) + \psi\sum_{j=1}^{n} w_1(y_j;\theta_0)\\ &+\sum_{j=1}^{n} w_2(y_j;\theta_0)\,y_{j,i} - \lambda_i\sum_{j=1}^{n} w_2(y_j;\theta_0) - \psi\sum_{j=1}^{n} w_2(y_j;\theta_0).\end{aligned}$$

For a compact solution, we write

$$\begin{aligned}\widehat{\psi} &= \frac{1}{2p}\left(\frac{B}{D}-\frac{A}{C}\right),\\ \widehat{\lambda}_i &= \frac{E_i+F_i}{C+D} + \frac{BC-AD}{2pD(C+D)} + \frac{AD-BC}{2pC(C+D)}, \quad i=1,\ldots,p,\end{aligned}$$

where

$$A = \sum_{k=1}^{p}\sum_{j=1}^{n} w_1(y_j;\theta_0)\,y_{j,k}, \quad B = \sum_{k=1}^{p}\sum_{j=1}^{n} w_2(y_j;\theta_0)\,y_{j,k}, \quad C = \sum_{j=1}^{n} w_1(y_j;\theta_0),$$

$$D := \sum_{j=1}^{n} w_2(y_j;\theta_0), E_i = \sum_{j=1}^{n} w_1(y_j;\theta_0)\,y_{j,i}, \quad F_i = \sum_{j=1}^{n} w_2(y_j;\theta_0)\,y_{j,i}$$

Given values of θ_0, the EM algorithm simply involves computing the weights $w_{\mathrm{r}}(\mathrm{y_j},\theta_0)$ for these values, updating to $\theta^{\dagger} = (\psi^{\dagger},\lambda^{\dagger})$, and checking for convergence at each step.

3.8.5 Appendix E: Details of computations

Algorithm 1: Probability of positive estimate

1. **Input:** R replications of the data which is sampled from a specific model.
2. For $i = 1$ to R:
   (a) Find the maximum likelihood estimate $\widehat{\theta}$, the restricted estimates $\widehat{\theta}_{\psi}$, and evaluate $j(\widehat{\theta})^{-1}$.
   (b) Define a grid of points for the parameter of interest $\mathscr{G} = \{l_{\mathrm{b}}, \ldots, u_{\mathrm{b}}\}$, where $l_{\mathrm{b}} \leq 0$. We use a transformation of the grid $\mathscr{G}$ that is finer around the boundary and coarser on the right of the boundary of the parameter space, $\mathscr{G}' = \mathscr{G} + \frac{c}{\pi}\sin(\pi\mathscr{G} + \pi)$, where $0 \leq c \leq 1$ is a tuning parameter
   (c) If closed-form expressions are unavailable, parameter estimation is performed numerically by maximizing the log likelihood function using either a quasi-Newton optimization algorithm with analytic first derivatives and second derivatives or the EM algorithm (linear convergence but more stable).
   (d) At each point of $\mathscr{G}'$ evaluate φ, q and then r^*.
   (e) Estimate $\widehat{\psi}$ and $\widehat{\psi}^*$ as solutions of $r(\psi) = 0$ and $r^*(\psi) = 0$.
   (f) Evaluate the pivots r_0 and r_0^* under the null hypothesis to obtain p-values for the null hypothesis H_0. .
3. **Output:** Estimated probabilities of positive estimates are $\frac{1}{R}\sum_{i=1}^{R} I(\widehat{\psi}_i > 0)$ and $\frac{1}{R}\sum_{i=1}^{R} I(\widehat{\psi}_i^* > 0)$.

Future work: Signal detection

4.1 Motivation and preliminary results

High-energy physics experiments, such as those conducted in the Large Hadron Collider at CERN, involve nonnegative parameters that are small, maybe zero, to detect small signals in the presence of background noise. The main challenge in such experiments is to decide, given a particular observation, whether it originated from the background noise alone or from a noisy signal.

We follow the model in Davison and Sartori (2008), and assume that for a single channel, the available data y_1, y_2, y_3 are assumed to be realizations of independent Poisson random variables with means $\gamma\psi + \beta$, βt and γu. The detectability of the signal in this model depends, amongst other factors, on the background rate at which the event occurs, $\beta > 0$, the efficiency of the measurement device, $1 \geq \gamma > 0$, and the length of the subsidiary experiments to estimate these parameters, $t > 0$ and $u > 0$.

The goal is to summarize the evidence concerning ψ, large estimates of which suggest the presence of the signal. However, for small ψ, which is the case when the particle mass is either equal to zero or comparable to the experimental precision, the parameter of interest is on the boundary of its parameter space. The distribution of the maximum likelihood estimator for the signal has a mixture distribution as emphasized in Mandelkern (2002), with similar interpretations to the examples discussed in Chapter 3. In principle, the nuisance parameters are positive and $\psi \geq 0$, but it is mathematically reasonable to consider negative values for ψ, provided $\psi > -\beta/\gamma$. Testing for $H_0 : \psi = 0$, using this extended parameter space, is a soft boundary problem, though we restrict the interpretation of the results to the physically meaningful values $\psi \geq 0$.

The statistical model described above was discussed in Davison and Sartori (2008)

with a focus on inference for the signal using higher-order approximations in both frequentist and noninformative Bayesian setups. Mandelkern (2002) was the first to bring attention to the central issues in high-energy physics. This led to further discussions within the statistics community (Fraser et al., 2004; Ermini Leaf and Liu, 2012; Martin et al., 2012; Plante, 2020; Bickel, 2022). Analogous representations of the statistical model using a Gaussian model with lower bounds for the mean parameter or an equivalent multinomial model are used for signal detection (Davison and Sartori, 2008; Mandelkern, 2002).

The model easily extends to multiple channels, where the nuisance parameters β_k and γ_k are channel-specific. We now consider K realizations $y_k = (y_{1k}, y_{2k}, y_{3k})$, where the three components are assumed to be independent Poisson variables with respective means $(\gamma_k\psi + \beta_k, \beta_k t_k, \gamma_k u_k)$. The log likelihood function for $\theta = (\psi, \gamma_1, \beta_1, \ldots, \gamma_K, \beta_K)$ is

$$\ell(\theta) = \sum_{k=1}^{K} y_{1k}\log(\gamma_k\psi+\beta_k) - (\gamma_k\psi+\beta_k) + y_{2k}\log(\beta_k t_k) - (\beta_k t_k) + y_{3k}\log(\gamma_k u_k) - (\gamma_k u_k). \tag{4.1}$$

The MLEs satisfy the equations

$$\sum_{k}^{K} \widehat{\gamma}_k c_k = 0, \quad c_k = t_k - \tfrac{y_{2k}}{\widehat{\beta}_k}, \quad \widehat{\psi} c_k = u_k - \tfrac{y_{3k}}{\widehat{\gamma}_k}, \qquad k = 1, \ldots, K,$$

where $c_k = \left(\dfrac{y_{1k}}{\widehat{\gamma}_k\widehat{\psi} + \widehat{\beta}_k} - 1 \right)$.

The expected information matrix is

$$i(\theta) = \begin{pmatrix} A & d_1 & e_1 & d_2 & \cdots & d_N & e_N \\ d_1 & a_1 & c_1 & 0 & \cdots & 0 & 0 \\ e_1 & c_1 & b_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ d_N & 0 & 0 & 0 & \cdots & a_N & c_N \\ e_N & 0 & 0 & 0 & \cdots & c_N & b_N \end{pmatrix},$$

where the non-zero coefficients are

$$
\begin{aligned}
A &= -E\left[\frac{\partial^2 \ell(\theta)}{\partial \psi^2}\right] = \sum_{k=1}^{K} \frac{\gamma_k^2}{\gamma_k \psi + \beta_k}, \\
a_k &= E\left[-\frac{\partial^2 \ell(\theta)}{\partial \gamma_k^2}\right] = \frac{\psi^2}{(\gamma_k \psi + \beta_k)} + \frac{u_k}{\gamma_k}, \\
b_k &= E\left[-\frac{\partial^2 \ell(\theta)}{\partial \beta_k^2}\right] = \frac{1}{(\gamma_k \psi + \beta_k)} + \frac{t_k}{\beta_k}, \\
c_k &= E\left[-\frac{\partial^2 \ell(\theta)}{\partial \beta_k \partial \gamma_k}\right] = \frac{\psi}{(\gamma_k \psi + \beta_k)}, \\
d_k &= E\left[-\frac{\partial^2 \ell(\theta)}{\partial \psi \partial \gamma_k}\right] = 1 - \frac{\beta_k}{(\gamma_k \psi + \beta_k)}, \\
e_k &= E\left[-\frac{\partial^2 \ell(\theta)}{\partial \psi \partial \beta_k}\right] = \frac{\gamma_k}{(\gamma_k \psi + \beta_k)}.
\end{aligned}
$$

Tedious calculation gives the following approximation for the variance of the MLE of the signal ψ

$$
\mathrm{var}(\widehat{\psi}) \doteq \left\{ \sum_{k}^{K} \frac{\gamma_k^2 u_k t_k}{\psi^2 t_k \gamma_k + \beta_k u_k + u_k t_k (\gamma_k \psi + \beta_k)} \right\}^{-1}.
$$

The question of interest is whether inference on ψ is best improved by increasing the u_k and the t_k, or the number of channels K. These are two different schemes since increasing u_k and t_k increases the information on the individual nuisance parameters while increasing K also increases the number of nuisance parameters.

Define the precision function $f(\psi) = \mathrm{var}(\widehat{\psi})^{-1}$. First-order Taylor series expansion of f for small ψ gives

$$
f(\psi) \approx \sum_{k=1}^{K} \frac{\gamma_k^2}{\beta_k} \frac{t_k}{1+t_k} \left(1 - \psi \frac{\gamma_k}{\beta_k} \frac{t_k}{1+t_k}\right),
$$

where

$$
f(0) = \sum_{k=1}^{K} \frac{\gamma_k^2}{\beta_k} \frac{t_k}{1+t_k}, \quad f'(0) = -\sum_{k=1}^{K} \frac{\gamma_k^3}{\beta_k^2} \left(\frac{t_k}{1+t_k}\right)^2,
$$

In general, the t_k's have little effect since $t_k/(1+t_k)$ are bounded. If we consider a second-order expansion

$$
f''(0) = 2 \sum_{k=1}^{K} \left(\frac{\gamma_k}{\beta_k}\right)^3 \left(\frac{t_k}{1+t_k}\right)^2 \left\{ \left(\frac{\gamma_k t_k}{1+t_k}\right) - \frac{\beta_k}{u_k} \right\},
$$

then both the u_k's and t_k's affect the precision function, although the γ_k's are bounded. Assume the β_k, γ_k, u_k and t_k's are the same for all $k = 1, \ldots, K$. Then the asymptotic precision approximation is

$$f(\widehat{\psi}) = \frac{K\gamma^2 ut}{\{\psi^2\gamma t + \beta u + ut(\gamma\psi + \beta)\}},$$

which converges to $\dfrac{K\gamma^2}{\gamma\psi + \beta}$ for $u, t \to \infty$, that is, to the amount of information from K replicates of one channel with rate $\mu = \psi\gamma + \beta$.

To sum up, under these simplifications, the variance reduces if K increases (more channels), γ approaches one (efficiency of the measurements increases) or β decreases (background noise drops). In Figure 4.1, we plot the standard deviation for $\psi = 1$ and increasing number of channels K as a function of β, γ, u and t.

4.2 Constrained problem

In practice increasing the number of channels might be constrained to some maximum allocated cost $\mathscr{M}$. Consider the following constrained optimization problem in which we maximize the precision function, i.e., minimize the variance, with a constraint on the total cost

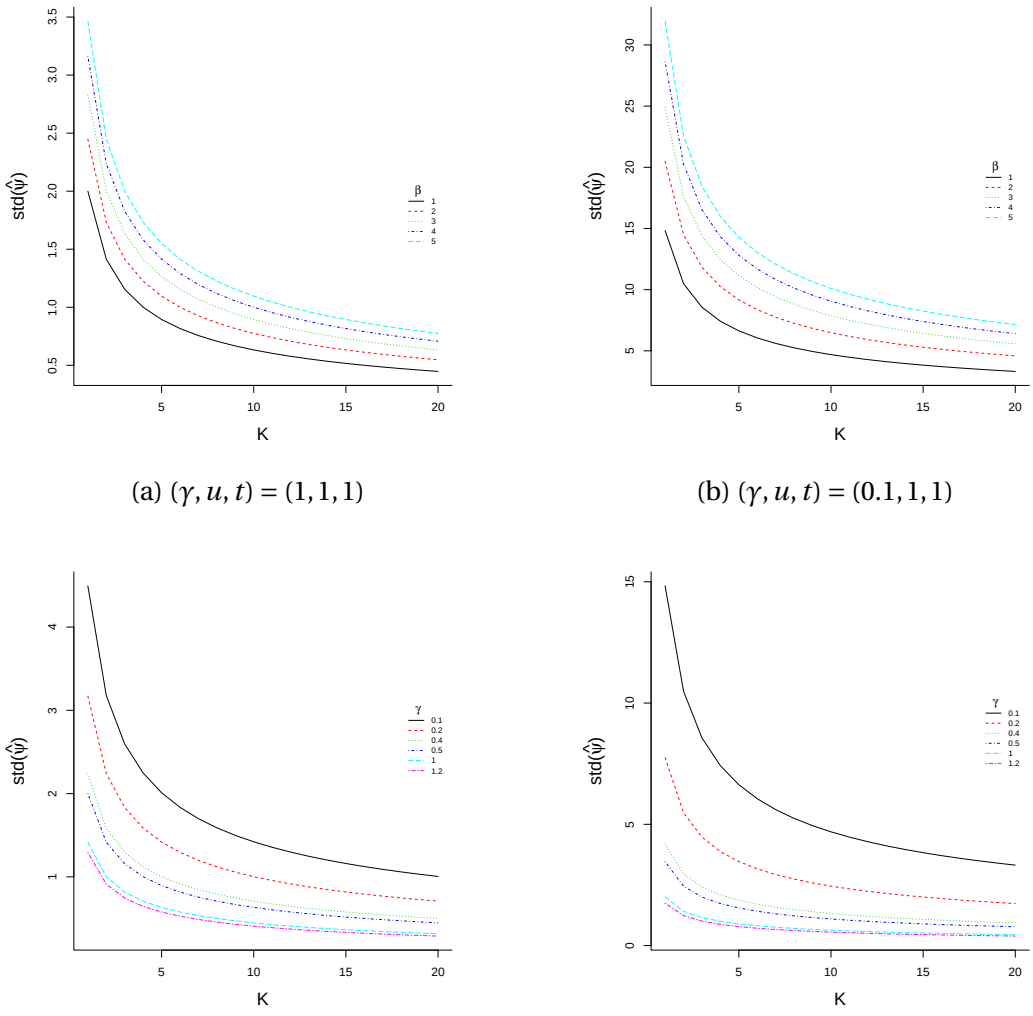
$$\begin{aligned} \max_{K,u,t} \quad & f(\widehat{\psi}) \\ \text{s.t.} \quad & K(c_1 + c_2 u + c_3 t) \leq \mathscr{M}, \end{aligned} \tag{4.2}$$

where c_1 is the cost of adding a channel, c_2 and c_3 are the costs of observing a channel for a single unit of time while estimating β and γ, respectively. We refer to Nocedal and Wright (2006) for solutions of constrained optimization problems.

The Lagrangian function corresponding to the constrained optimization problem is

$$\mathscr{L}(K, u, t, \alpha) = \text{var}(\widehat{\psi}) + \alpha\{K(c_1 + c_2 u + c_3 t) - \mathscr{M}\}.$$

If there exists a local solution of the constrained problem, then there exists a Lagrange multiplier α^* such that Karush–Kuhn–Tucker (KKT) conditions are satisfied



(a) $(\gamma, u, t) = (1, 1, 1)$

(b) $(\gamma, u, t) = (0.1, 1, 1)$

(c) $(\beta, u, t) = (10^{-3}, 1, 1)$

(d) $(\beta, u, t) = (1, 1, 1)$

Figure 4.1 – Standard deviation of the signal for $\psi = 1$ and increasing number of channels K as a function of β, γ, u and t.

for $(K^*, u^*, t^*, \alpha^*)$,

$$\frac{\partial \mathscr{L}}{\partial K} = 0 \Leftrightarrow -\frac{\left\{\psi^2 \frac{\gamma}{u} + \frac{\beta}{t} + (\gamma\psi + \beta)\right\}}{K^2\gamma^2} + \alpha\left(c_1 + c_2 u + c_3 t\right) = 0, \tag{4.3}$$

$$\frac{\partial \mathscr{L}}{\partial u} = 0 \Leftrightarrow -\frac{\psi^2}{K\gamma u^2} + \alpha K c_2 = 0, \tag{4.4}$$

$$\frac{\partial \mathscr{L}}{\partial t} = 0 \Leftrightarrow -\frac{\beta}{K\gamma^2 t^2} + \alpha K c_3 = 0, \tag{4.5}$$

where $\alpha \geq 0$, and $\alpha\{K(c_1 + c_2 u + c_3 t) - \mathscr{M}\} = 0$ for a feasible solution. Equating the α's obtained from (4.4) and (4.5) yields

$$\frac{\psi^2 \gamma}{c_2 u^2} = \frac{\beta}{c_3 t^2},$$

which implies that

$$\frac{t}{u} = \sqrt{\frac{\beta}{\gamma}\frac{c_2}{c_3}}\psi^{-1}.$$

Multiplying equations (4.4) and (4.5) respectively by u, and t, and substituting the corresponding quantities in equation (4.3), we obtain the Lagrange multiplier

$$\alpha^* = \frac{\gamma\psi + \beta}{K\gamma^2(M - Kc_2 u - Kc_3 t)} = \frac{\gamma\psi + \beta}{K^2 c_1 \gamma^2} = \frac{\gamma\psi + \beta}{K\gamma^2}\frac{1}{Kc_1}.$$

The optimal values of u and t are obtained by plugging α^* into (4.4) and (4.5), and are

$$u^* = \psi\sqrt{\frac{c_1}{c_2}}\sqrt{\frac{\gamma}{\psi\gamma + \beta}}, \tag{4.6}$$

$$t^* = \sqrt{\frac{c_1}{c_3}\frac{\beta}{\psi\gamma + \beta}}. \tag{4.7}$$

The optimal number of channels is, thus,

$$K^* = \frac{\mathscr{M}}{c_1 + c_2 u^* + c_3 t^*} = \frac{\mathscr{M}\sqrt{\psi\gamma + \beta}}{\sqrt{c_1 c_1}\sqrt{\psi\gamma + \beta} + \sqrt{c_1 c_2 \gamma}\psi + \sqrt{c_1 c_3 \beta}}. \tag{4.8}$$

The optimal trade-off between K, u and t based on the relative costs c_1, c_2, c_3 and the total cost $\mathscr{M}$ is given in equations (4.6), (4.7) and (4.8), respectively. First-order asymptotics suggests that to improve the precision of the estimated ψ, one should primarily increase the number of channels K. Since the ratio of t^*/u^* is constant, one is not interested in increasing the time allocated to estimate β and γ. If $\psi = 0$, then $u^* = 0$, which makes sense as there is no signal to be detected.

4.3 Improved inference for the signal

A future extension of this work is intended to explore what values for u and t are admissible from a physical point of view. We may also include the maximum number of installed channels $K_{\max}$ within the facility to the constrained optimization problem in (4.2). Another point that may change the perception of this problem is which of

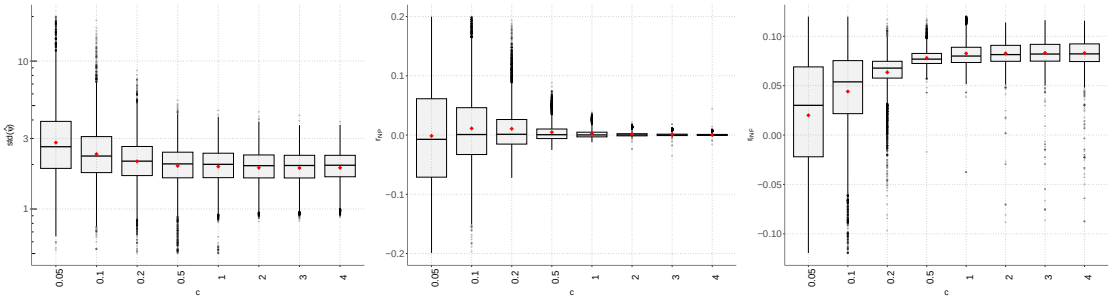


Figure 4.2 – Boxplots of std($\widehat{\psi}$) (left), r_{NP} (middle), and r_{INF} (right) in a one-channel where where $\psi = 1$, $\gamma = 1$,$\beta = \exp(1.1)$, for $c = 0.05, \ldots, 4$. The average is in red bullets.

the parameters in λ or the data (u_k, t_k) can be easily changed during an experiment, especially when the background noise sources produce events indistinguishable from signal events.

Figure 4.2 and 4.3 show boxplots of the standard deviation of the signal and the components of the modified likelihood root r_{NP} and r_{INF} in one- and multi-channel experiments. Since the variables have Poisson distributions, considering $c \times (u_k, t_k)$ in instead of (u_k, t_k) for every channel where $c > 0$ is effectively the same as changing the corresponding γ_k and β_k. However, it makes more sense to think of the parameters as being fixed, and the observation scheme is enhanced by increasing the allocated times t_k and u_k. In a one-channel experiment, for large c, r_{INF} has fairly more weight than r_{NP} as we only have two nuisance parameters, but in a multi-channel setup, r_{NP} increases with increasing K and the magnitude of the correction decreases with increasing allocated time c. These results align with the study of the precision function in (4.1).

Future directions of this work include writing the model in (4.1) to satisfy the requirements by Sartori (2003) and Tang and Reid (2020), which applies to models for stratified data in a two-index asymptotics setting. In particular, we could explore under which asymptotic setting the modified likelihood root gives improvements to the usual asymptotic distribution in a multi-channel experiment. As shown in the previous section, first-order approximation using the inverse of the Fisher information matrix suggests increasing the number of channels. However, this implies that for (4.1), a curved exponential model with $3K$ observations, the $2K$-dimensional nuisance parameter inevitably increases as well. These results not only have the potential to give improved inference for the signal parameter, as shown in Sartori (2003) but also increase the precision of the detected signal, which is critical for experiments with a weak signal.

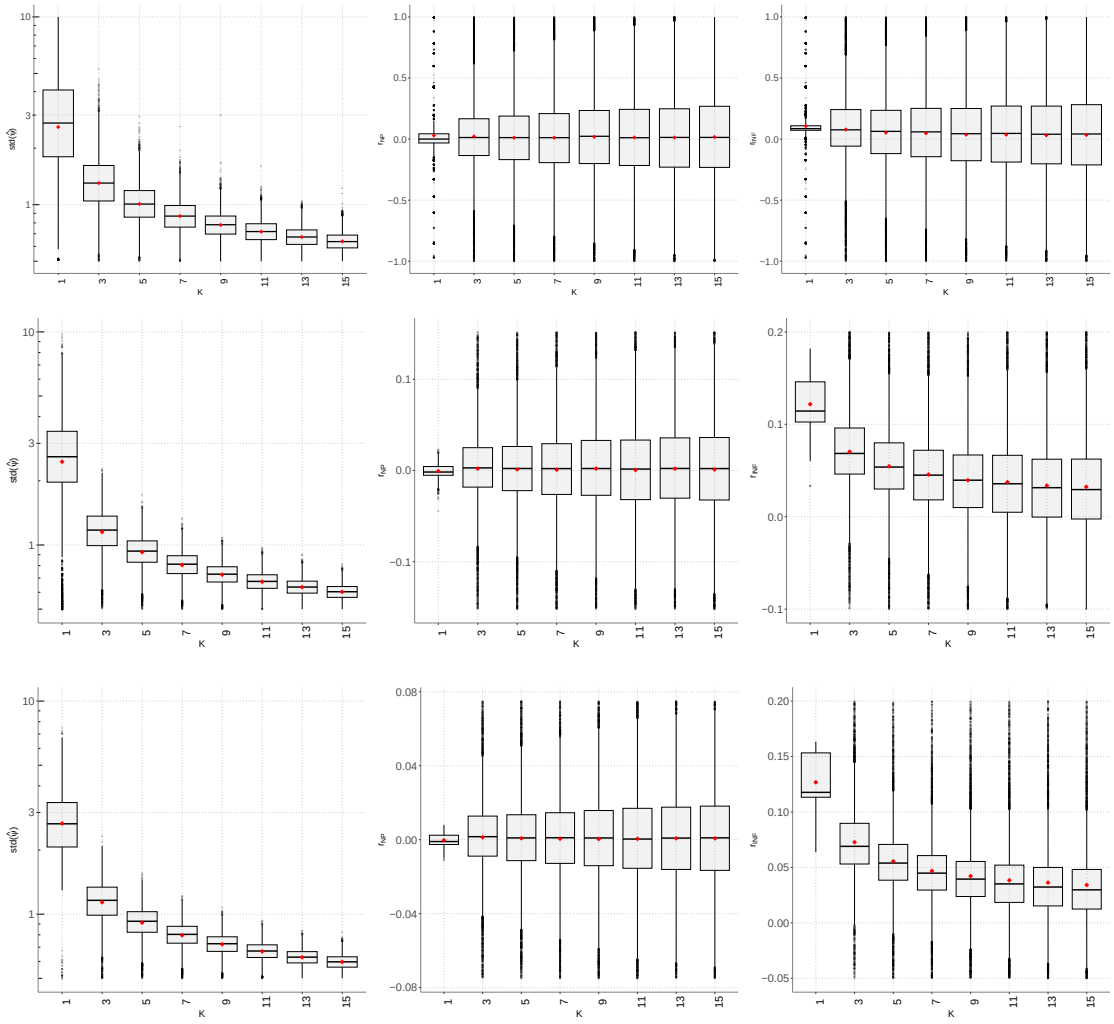


Figure 4.3 – Boxplots of the standard deviation std($\widehat{\psi}$) (left), r_{NP} (middle), and r_{INF} (right) in a multiple-channel experiment where $\gamma_k = 0.5,\ldots,0.9$, $\beta_k = 0.5+0.1k$, $t_k = c(15+2k)$, $u_k = c(50+2k)$, for $k = 1,\ldots,K$, and $c = 0.2,1,2$ (top to bottom). The average is in red bullets

Bibliography

Adurthi, N. and Singla, P. (2015). Conjugate unscented transformation-based approach for accurate conjunction analysis. *Journal of Guidance, Control, and Dynamics*, 38(9):1642–1658.

Alfano, S. (2005a). A numerical implementation of spherical object collision probability. *The Journal of the Astronautical Sciences*, 53(1):103–109.

Alfano, S. (2005b). Relating position uncertainty to maximum conjunction probability. *The Journal of the Astronautical Sciences*, 53(2):193–205.

Alfano, S. (2006a). Addressing nonlinear relative motion for spacecraft collision probability. In *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*, Keystone, CO.

Alfano, S. (2006b). Satellite collision probability enhancements. *Journal of Guidance Control and Dynamics*, 29(3):588–592.

Alfano, S. (2007). Review of conjunction probability methods for short term encounters. *Advances in the Astronautical Sciences*, 127(1):719–746.

Alfano, S. and Negron, D. J. (1993). Determining satellite close approaches. *Journal of the Astronautical Sciences*, 41(2):217–225.

Alfano, S. and Oltrogge, D. (2018). Probability of collision: Valuation, variability, visualization, and validity. *Acta Astronautica*, 148:301–316.

Alfriend, K. T., Akella, M. R., Frisbee, J., Foster, J. L., Lee, D. J., and Wilkins, M. (1999). Probability of collision error analysis. *Space Debris*, 1(1):21–35.

Alfriend, K. T., Vadali, S. R., Gurfil, P., How, J. P., and Breger, L. S. (2010). *Spacecraft Formation Flying*. Butterworth-Heinemann, Oxford.

Armellin, R., Lizia, P. D., Berz, M., and Makino, K. (2010). Computing the critical points of the distance function between two keplerian orbits via rigorous global optimization. *Celestial Mechanics and Dynamical Astronomy*, 107(3):377–395.

Armstrong, D. M. (1963). Absolute and relative motion. *Mind*, 72(286):209–223.

Balch, M. S. (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning*, 53(7):1003–1019.

Balch, M. S. (2016). A corrector for probability dilution in satellite conjunction analysis. In *18th AIAA Non-Deterministic Approaches Conference*, San Diego. The American Institute of Aeronautics and Astronautics.

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):20180565.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, 67(2):293–310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73:307–322.

Barndorff-Nielsen, O. E. (1988). *Parametric Statistical Models and Likelihood*. Springer: New York.

Barndorff-Nielsen, O. E. and Chamberlin, S. R. (1991). An ancillary invariant modification of the signed log likelihood ratio. *Scandinavian Journal of Statistics*, 18(4):341–352.

Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications (with Discussion). *Journal of the Royal Statistical Society series B*, 41(3):279–312.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall, London.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

Barndorff-Nielsen, O. E. and Pedersen, K. (1968). Sufficient data reduction and exponential families. *Mathematica Scandinavica*, 22(1):197–202.

Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika*, 40(1/2):12–19.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(2):155–178.

Basu, D. (1978). On partial sufficiency: A review. *Journal of Statistical Planning and Inference*, 2(1):1–13.

Bellio, R., Greco, L., and Ventura, L. (2008). Modified quasi-profile likelihoods from estimating functions. *Journal of Statistical Planning and Inference*, 138(10):3059–3068.

Belzile, L. et al. (2022). *mev: Modelling Extreme Values*. R package version 1.14.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, California.

Bernstein, J., Schlafly, E., Schneider, M., and Miller, C. (2021). Evaluating space object conjunction probabilities using characteristic function inversion. Technical Report LLNL-TR-827434, Laurence Livermore National Laboratory.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

Bickel, D. R. (2022). Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support. *Communications in Statistics - Theory and Methods*, 51(10):3142–3163.

Bickel, P. J. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In Ghosh, J. K., Mitra, S. K., Parthasarathy, K. R., and Prakasa Rao, B. L. S., editors, *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, pages 83–96, New York Wiley.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388.

Bonat, W. H. and Ribeiro Jr, P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, 27(2):83–89.

Braun, V., Gelhaus, J., Kebschull, C., Sánchez-Ortiz, N., Oliveira, J., Domínguez-González, R., Wiedemann, C., Krag, H., and Vörsmann, P. (2013). Drama 2.0 - ESA's space debris risk assessment and mitigation analysis tool suite. In *64th International Astronautical Congress*, Beijing.

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics*. Cambridge University Press.

Brazzale, A. R. and Mameli, V. (2022). Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio. arXiv preprint arXiv:2206.15178.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brown, B. W. and Hollander, M. (1977). *Statistics: A Biomedical Introduction*. Wiley, New York.

Brown, H. K. and Kempton, R. A. (1994). The application of REML in clinical trials. *Statistics in Medicine*, 13(16):1601–1617.

Burnett, E. R. and Schaub, H. (2022). Spacecraft relative motion dynamics and control using fundamental solution constants. In *AIAA SCITECH 2022 Forum*, San Diego, CA & Virtual.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. Springer: New York.

Butler, R. W. (2007). *Saddlepoint Approximations with Applications*. Cambridge University Press.

Carpenter, J. (2019). Covariance realism is not enough. *AAS Advances in the Astronautical Sciences*.

Carpenter, J., Alfano, S., Hall, D. T., Hejduk, M., and Gaebler, J. (2017). Relevance of the American Statistical Asoociation's warning on p-values for conjunction assessment. In *AAS/AIAA Astrodynamics Specialist Conference*, Stevenson WA.

Castillo, J. D. and López-Ratera, A. (2006). Saddlepoint approximation in exponential models with boundary points. *Bernoulli*, 12(3):491–500.

Chan, K. (1997). Collision probability analyses for earth orbiting satellites. In *Pacific Basin Societies, Advances in the Astronautical Sciences, 7th International Space Conference*, volume 96, pages 1033–1048.

Chan, K. (2008). *Spacecraft Collision Probability*. Aerospace Press, El Segundo, CA.

Chan, K. (2011). Miss distance – generalized variance non-central chi distribution. In *AAS/AIAA Space Flight Mechanics Meeting*, pages 11–175, Girdwood AK,.

Chant, D. (1974). On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*, 61(2):291–298.

Chatterjee, S. K. and Das, K. (1983). Estimation of variance components in an unbalanced one-way classification. *Journal of Statistical Planning and Inference*, 8(1):27–41.

Chatters, E. P., Crothers, B. J., Command, A., College, S., and Seminars, S. R. E. (2009). AU-18 space primer. Technical report, Air University Press.

Chen, J. and Kalbfleisch, J. D. (2005). Modified likelihood ratio test in finite mixture models with a structural parameter. *Journal of Statistical Planning and Inference*, 129(1):93–107.

Chen, L., Bai, X. Z., Liang, Y. G., and Li, K. B. (2017). *Orbital Data Applications for Space Objects: Conjunction Assessment and Situation Analysis*. Springer, Singapore.

Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems (with Discussion). *Journal of the Royal Statistical Society series B*, 57(1):3–44.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573 – 578.

Chernoff, H. (1956). Large-sample theory: Parametric case. *Annals of Mathematical Statistics*, 27(1):1–22.

Chow, J., Crezee, J., and Kopylev, L. (2012). Constrained parameters in applications: Review of issues and approaches. *International Scholarly Research Notices Biomathematics*, 2012(Article ID 8729).

Clohessy, W. H. and Wiltshire, R. S. (1960). Terminal guidance system for satellite rendezvous. *Journal of the Aerospace Sciences*, 27(9):653–658.

Committee on the Peaceful Uses of Outer Space (2019). *Guidelines for the Long-Term Sustainability of Outer Space Activities*. United Nations.

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–76.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Cox, D. R. (2020). Statistical significance. *Annual Review of Statistics and Its Application*, 7(1):1–10.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *Journal of the Royal Statistical Society series B*, 49(1):1–39.

Crainiceanu, C. M. and Ruppert, D. (2004a). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, series B*, 66(1):165–185.

Crainiceanu, C. M. and Ruppert, D. (2004b). Restricted likelihood ratio tests in nonparametric longitudinal models. *Statistica Sinica*, 14(3):713–729.

Crainiceanu, C. M., Ruppert, D., and Vogelsang, T. J. (2002). Probability that the MLE of a variance component is zero with applications to likelihood ratio tests. *Available at www.orie.cornell.edu/davidr/papers*.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Cunen, C., Hjort, N. L., and Schweder, T. (2020). Confidence in confidence distributions! *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2237):20190781.

Curtis, H. D. (2010). Relative motion and rendezvous. In *Orbital Mechanics for Engineering Students (Second Edition)*, Aerospace Engineering, pages 391–427. Butterworth-Heinemann, Boston.

Dagum, P., Karp, R., Luby, M., and Ross, S. (2000). An optimal algorithm for Monte Carlo estimation. *SIAM Journal on Computing*, 29(5):1484–1496.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25(4):631–650.

Dannemann, J. and Holzmann, H. (2008). Likelihood ratio testing for hidden Markov models under non-standard conditions. *Scandinavian Journal of Statistics*, 35(2):309–321.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43.

Davison, A. C. (1984). Modelling excesses over high thresholds, with an application. In de Oliveira, J. T., editor, *Statistical Extremes and Applications*, pages 461–482. Springer Netherlands.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press.

Davison, A. C., Fraser, D. A. S., and Reid, N. (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society series B*, 68(3):495–508.

Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, 75(3):417–431.

Davison, A. C. and Reid, N. (2022). The tangent exponential model. In Berger, J. O., Meng, X.-L., Reid, N., and Xie, M., editors, *Handbook of Bayesian, Fiducial and Frequentist Inference*, page to appear. Chapman & Hall/CRC, Boca Raton, FL.

Davison, A. C. and Sartori, N. (2008). The Banff Challenge: Statistical detection of a noisy signal. *Statistical Science*, 23(3):354–364.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer: New York.

DiCiccio, T. J. and Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *Journal of the Royal Statistical Society series B*, 55(1):305–316.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.

Dupuis, D. J. (1998). Exceedances over high thresholds: A guide to threshold selection. *Extremes*, 1(3):251–261.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Elkantassi, S. and Davison, A. C. (2022). Space oddity? A statistical formulation of conjunction assessment. *Journal of Guidance, Control, and Dynamics*, 45(12):2258–2274.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer: Berlin.

Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *International Journal of Approximate Reasoning*, 53(5):709–727.

ESA (2022). ESA's annual space environment report. Technical report, ESA Space Debris Office, Darmstadt, Germany.

Esseen, C.-G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica*, 77(1):1 – 125.

Fasiolo, M., Wood, S. N., Hartig, F., and Bravington, M. V. (2018). An extended empirical saddlepoint approximation for intractable likelihoods. *Electronic Journal of Statistics*, 12(1):1544 – 1578.

Feng, Z. D. and McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, 13(4):325–332.

Feng, Z. D. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):609–617.

Ferrari, S. L. P., Cordeiro, G. M., and Cribari-Neto, F. (2001). Higher-order asymptotic refinements for score tests in proper dispersion models. *Journal of Statistical Planning and Inference*, 97(1):177–190.

Ferrari, S. L. P. and Cysneiros, A. H. (2008). Skovgaard's adjustment to likelihood ratio tests in exponential family nonlinear models. *Statistics & Probability Letters*, 78(17):3047–3055.

Ferrari, S. L. P., Uribe-Opazo, M. A., and Cribari-Neto, F. (1997). Second order asymptotics for score tests in exponential family nonlinear models. *Journal of Statistical Computation and Simulation*, 59(2):179–194.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika*, 76(3):457–464.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, series A*, 222:309–368.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd Edinburgh.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. Collier–Macmillan, New York.

Foster, J. L. and Estes, H. S. (1992). *A Parametric Analysis of Orbital Debris Collision Probability and Maneuver Rate for Space Vehicles*. NASA, National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Fraser, D. A. S. (1963). On sufficiency and the exponential family. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(1):115–123.

Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77(1):65–76.

Fraser, D. A. S. (1991). Statistical inference: Likelihood to significance. *Journal of the American Statistical Association*, 86(414):258–265.

Fraser, D. A. S. (2017). p-values: The insight to modern statistical inference. *Annual Review of Statistics and its Application*, 4(1):1–14.

Fraser, D. A. S. (2019). The p-value function and statistical inference. *American Statistician*, 73(1):135–147.

Fraser, D. A. S., Bédard, M., Wong, A., Lin, W., and Fraser, A. M. (2016a). Bayes, reproducibility and the quest for truth. *Statistical Science*, 31(4):578–590.

Fraser, D. A. S. and Reid, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations to distribution functions. *Statistica Sinica*, 3(1):67–82.

Fraser, D. A. S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica*, 7(2):33–53.

Fraser, D. A. S. and Reid, N. (2001). Ancillary information for statistical inference. In Ahmed, S. E. and Reid, N., editors, *Empirical Bayes and Likelihood Inference*, pages 185–207. Springer, New York.

Fraser, D. A. S. and Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *Journal of Statistical Planning and Inference*, 103(1):263–285.

Fraser, D. A. S., Reid, N., Li, R., and Wong, A. (2003). p-value formulas from likelihood asymptotics: bridging the singularities. *Journal of Statistical Research*, 37(1):1–15.

Fraser, D. A. S., Reid, N., and Sartori, N. (2016b). Accurate directional inference for vector parameters. *Biometrika*, 103(3):625–639.

Fraser, D. A. S., Reid, N., and Wong, A. C. M. (2004). Inference for bounded parameters: A different perspective. *Physical Review, D*, 69(3):033002.

Fraser, D. A. S., Reid, N., and Wu, J. (1999a). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, 86(2):249–264.

Fraser, D. A. S., Wong, A., and Wu, J. (1999b). Regression analysis, nonlinear or nonnormal: Simple and accurate calculation of p-values from likelihood analysis. *Journal of the American Statistical Association*, 94(448):1265–1295.

Frydenberg, M. and Jensen, J. L. (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika*, 76(4):655–661.

Garcia-Pelayo, R. and Hernando-Ayuso, J. (2016). Series for collision probability in short-encounter model. *Journal of Guidance Control and Dynamics*, 39(8):1904–1912.

Gerlovina, I., van der Laan, M. J., and Hubbard, A. (2017). Edgeworth expansions provide a cautionary tale. *The International Journal of Biostatistics*, 13(1):20170012.

Ghosh, J. K. and Sen, P. K. (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. North Carolina State University. Dept. of Statistics.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Neneralized Linear Models: a Roughness Penalty Approach*. Chapman & Hall, London.

Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics*, 35(2):185–191.

Gronchi, G. F. (2005). An algebraic method to compute the critical points of the distance function between two Keplerian orbits. *Celestial Mechanics and Dynamical Astronomy*, 93(1):295–329.

Guedes, A. C., Cribari-Neto, F., and Espinheira, P. L. (2020). Modified likelihood ratio tests for unit gamma regressions. *Journal of Applied Statistics*, 47(9):1562–1586.

Hall, D. T. (2021). Expected collision rates for tracked satellites. *Journal of Spacecraft and Rockets*, 58(3):715–728.

Hall, D. T., Casali, S. J., Johnson, L. C., Skrehart, B. B., and Baars, L. G. (2018). High fidelity collision probabilities estimated using brute force Monte Carlo simulations. In *AIAA/AAS Astrodynamics Specialist Conference*.

Hall, P. (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. *The Annals of Probability*, 15(3):920–931.

Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics*, 22(2):215–232.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, pages 807–810, Monterey, Wadsworth.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):20–338.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795 – 800.

Hejduk, M. D. and Snow, D. E. (2019). Satellite conjunction "probability," "possibility," and "plausibility": a categorization of competing conjunction assessment risk assessment paradigms. In *AAS/AIAA Astrodynamics Specialist Conference*, number GSFC-E-DAA-TN71111-1.

Hejduk, M. D., Snow, D. E., and Newman, L. K. (2019). Satellite conjunction assessment risk analysis for "dilution region" events: Issues and operational approaches. volume 28, Austin, TX.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1):3–34.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hill, G. W. (1878). Researches in the lunar theory. *American Journal of Mathematics*, 1(1):5–26.

Hosking, J. R. M. (1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika*, 71(2):367–374.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal varibles. *Biometrika*, 48(3):352–363.

Inalhan, G., Tillerson, M., and How, J. P. (2002). Relative dynamics and control of spacecraft formations in eccentric orbits. *Journal of Guidance, Control, and Dynamics*, 25(1):48–59.

Inter-Agency Space Debris Coordination Committee (2007). *Space Debris Mitigation Guidelines*. IADC.

International Standards Organisation (2016). Space systems: Estimation of orbit lifetime. Technical report, ISO.

International Standards Organisation (2019). Space systems: Space debris mitigation requirements. Technical report, ISO.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jørgensen, B. (1994). The rules of conditional inference: Is there a universal definition of nonformation? *Journal of the Italian Statistical Society*, 3(3):355–384.

Jørgensen, B. and Labouriau, R. (1995). Exponential families and theoretical inference. Lecture Notes.

Jones, B. A. and Doostan, A. (2013). Satellite collision probability estimation using polynomial chaos expansions. *Advances in Space Research*, 52(11):1860–1875.

Kauermann, G. (2002). On a small sample adjustment for the profile score function in semiparametric smoothing models. *Journal of Multivariate Analysis*, 82(2):471–485.

Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer: New York.

Kenneth, C. F. (2015). Formulation of collision probability with time-dependent probability distribution functions. In *AAS/AIAA Space Flight Mechanics Meeting*, number 15-233. American Astronautical Society.

Klinkrad, H. (2006). *Space Debris: Models and Risk Analysis*. Astronautical Engineering. Springer, Berlin, Heidelberg.

Kolassa, J. E. and McCullagh, P. (1990). Edgeworth series for lattice distributions. *The Annals of Statistics*, 18(2):981 – 985.

Kopylev, L. and Sinha, B. (2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. *Sankhyā: The Indian Journal of Statistics, Series B*, 73(1):20–41.

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 50(3-4):403–418.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Lee, L. F. (1993). Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. *Econometric Theory*, 9(3):413–430.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

Letizia, F., Lemmens, S., Bastida Virgili, B., and Krag, H. (2019). Application of a debris index for global evaluation of mitigation strategies. *Acta Astronautica*, 161:348–362.

Lewis, H. (2020). Evaluation of debris mitigation options for a large constellation. *Journal of Space Safety Engineering*, 7(3):192–197.

Li, J.-S., Yang, Z., and Luo, Y.-Z. (2022). A review of space-object collision probability computation methods. *Astrodynamics*, 6(2):95–120.

Li, R. (2001). *On Asymptotic Likelihood Inference, Removing p-value Singularities*. PhD thesis, University of Toronto.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Number 5 in NSF-CBMS Regional Conference Series in Probability and Statistics. Institute for Mathematical Statistics, Hayward, CA.

Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM AND ECME. *Statistica Sinica*, 5(1):19–39.

Losacco, M., M. Romano, a. P. D. L., Colombo, C., Armellin, R., Morselli, A., and Pérez, J. (2019). Advanced Monte Carlo sampling techniques for orbital conjunctions analysis and near earth objects impact probability computation. In *1st NEO and Debris Detection Conference*. ESA Space Safety Programme Office.

Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12(2):475–490.

Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statistical Science*, 17(2):149–159.

Martin, R., Leaf, D. E., and Liu, C. (2012). Optimal inferential models for a Poisson mean. arXiv preprint arXiv:1207.0105.

McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London.

McCullagh, P. and Tibshirani, R. J. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society series B*, 52(2):325–344.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324.

McLachlan, G. J. (1999). Mahalanobis distance. *Resonance*, 4(6):20–26.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Melo, T. F., Vasconcellos, K. L., and Lemonte, A. J. (2009). Some restriction tests in a new class of regression models for proportions. *Computational Statistics & Data Analysis*, 53(12):3972–3979.

Miller, J. J. (1977). Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance. *The Annals of Statistics*, 5(4):746 – 762.

Modenini, D., Curzi, G., and Locarini, A. (2022). Relations between collision probability, Mahalanobis distance, and confidence intervals for conjunction assessment. *Journal of Spacecraft and Rockets*, 59(4):1125–1134.

Moser, B. K. (1996). *Linear Models*. Elsevier, San Diego.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society series A*, 135(3):370–384.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, 20A(1/2):175–240.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer: New York.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*. World Scientific, Singapore.

Pastel, R. (2011). Estimating satellite versus debris collision probabilities via the adaptive splitting technique. In *Proc. 3rd Int. Conf. Comput. Modeling Simulation*, pages 1–6.

Patera, R. (2001). General method for calculating satellite collision probability. *Journal of Guidance, Control, and Dynamics*, 24(4):716–722.

Patera, R. (2005). Calculating collision probability for arbitrary space vehicle shapes via numerical quadrature. *Journal of Guidance Control and Dynamics*, 28(6):1326–1328.

Patera, R. P. (2003). Satellite collision probability for nonlinear relative motion. *Journal of Guidance, Control, and Dynamics*, 26(5):728–733.

Patera, R. P. (2006). Collision probability for larger bodies having nonlinear relative motion. *Journal of Guidance Control and Dynamics*, 29:1468–1472.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pearson, E. S. (1959). Note on the distribution of non-central χ^2. *Biometrika*, 46:364.

Pearson, K. (1936). Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2):34–59.

Peers, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society series B*, 27(1):9–16.

Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):701–737.

Plante, A. (2020). A Gaussian alternative to using improper confidence intervals. *Canadian Journal of Statistics*, 48(4):773–801.

Reid, N. (1988). Saddlepoint methods and statistical inference (with Discussion). *Statistical Science*, 3(2):213–238.

Reid, N. (1991). Approximations and asymptotics. In Hinkley, D. V., Reid, N., and Snell, E. J., editors, *Statistical Theory and Modelling: In Honour of Sir David Cox, FRS*, pages 287–305. Chapman & Hall, London.

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics*, 31(6):1695–1731.

Reid, N. and Fraser, D. A. S. (2010). Mean loglikelihood and higher order approximations. *Biometrika*, 97(1):159–170.

Reid, N., Mukerjee, R., and Fraser, D. A. S. (2002). Some aspects of matching priors. In Moore, M., Froda, S., and Léger, C., editors, *Mathematical Statistics and Applications: Festschrift for Constance van Eeden*, volume 42 of *Lecture Notes — Monograph Series*, pages 31–44. Institute of Mathematical Statistics, Hayward, California.

Ritz, C. and Skovgaard, I. M. (2005). Likelihood ratio tests in curved exponential families with nuisance parameters present only under the alternative. *Biometrika*, 92:507–17.

Ross, G. (1990). *Nonlinear Estimation*. Springer: New York.

Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. M. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2):243–284.

Royall, R. M. (1997). *Statistical Evidence: a Likelihood Paradigm*. Chapman & Hall/CRC, London.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Sahai, H. and Ojeda, M. M. (2004). *Analysis of Variance for Random Models*. Birkhäuser Boston, MA.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, 90(3):533–549.

Schaub, H. (2004). Relative orbit geometry through classical orbit element differences. *Journal of Guidance, Control, and Dynamics*, 27(5):839–848.

Schaub, H. and Alfriend, K. T. (2002). Hybrid Cartesian and orbit element feedback law for formation flying spacecraft. *Journal of Guidance, Control, and Dynamics*, 25(2):387–393.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Serra, R., Arzelier, D., Joldes, M., Lasserre, J.-B., Rondepierre, A., and Salvy, B. (2016). Fast and accurate computation of orbital collision probability for short-term encounters. *Journal of Guidance, Control, and Dynamics*, 39(5):1009–1021.

Severini, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, 85(3):507–522.

Severini, T. A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika*, 86(2):235–247.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Clarendon Press, Oxford.

SIA (2022). State of the satellite industry report. Technical report, Satellite Industry Association.

Siminski, J., Merz, K., Virgili, B. B., Braun, V., Flegel, S., Flohrer, T., Funke, Q., Horstmann, A., Lemmens, S., Letizia, F., Mclean, F., Sanvido, S., and Schaus, V. (2021). ESA's collision avoidance service: current status and special cases. In *8th European Conference on Space Debris*. ESA Space Debris Office.

Singh, K., Xie, M., and Strawderman, W. (2007). Confidence distribution (CD) – distribution estimator of a parameter. *Lecture Notes Monograph Series*, 54:132–150.

Sinha, B. K., Kopylev, L., and Fox, J. (2012). Some new aspects of statistical inference for multistage dose-response models with applications. *Pakistan Journal of Statistics and Operation Research*, 8(3):441–478.

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability*, 24(4):875–887.

Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *The Annals of Statistics*, 18(2):779–789.

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, 2(2):145–166.

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics*, 28(1):3–32.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1989). A survey of nonregular problems. *Bulletin of the International Statistical Institute*, 53:353–372.

Srivastava, M. S. and Yau, W. K. (1989). Saddlepoint method for obtaining tail probability of Wilks' likelihood ratio test. *Journal of Multivariate Analysis*, 31(1):117–126.

Stein, M. C., da Silva, M. F., and Duczmal, L. H. (2014). Alternatives to the usual likelihood ratio test in mixed linear models. *Computational Statistics & Data Analysis*, 69:184–197.

Susko, E. (2013). Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika*, 100(4):1019–1023.

Tang, Y. and Reid, N. (2020). Modified likelihood root in high dimensions. *Journal of the Royal Statistical Society Series B*, 82(5):1349–1369.

Terui, F. and ichiro Nishida, S. (2007). Relative motion estimation and control to a failed satellite by machine vision. *IFAC Proceedings Volumes*, 40(7):639–644. 17th IFAC Symposium on Automatic Control in Aerospace.

Thode, Jr, H. C., Finch, S. J., and Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics*, 44(4):1195–1201.

Tibshirani, R. J. (1989). Noninformative priors for one parameter of many. *Biometrika*, 76(3):604–608.

Tse, T. and Davison, A. C. (2022). A note on universal inference. *STAT*, 11(1):e501.

Union of Concerned Scientistis (2022). Satellite database. https://www.ucsusa.org/resources/satellite-database.

Vallado, D. A. (2013). *Fundamentals of Astrodynamics and Applications*. Microcosm Press, Hawthorne, CA, fourth edition.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Virgili, B. (2016). Delta (debris environment long-term analysis). In *6th International Conference on Astrodynamics Tools and Techniques (ICATT)*, Darmstadt, Germany.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wang, S. (1992). General saddlepoint approximations in the bootstrap. *Statistics & Probability Letters*, 13(1):61–66.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society series B*, 25(2):318–329.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton.

Yamanaka, K. and Ankersen, F. (2002). New state transition matrix for relative motion on an arbitrary elliptical orbit. *Journal of Guidance, Control, and Dynamics*, 25(1):60–66.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zhang, S., Fu, T., Chen, D., and Cao, H. (2020). Satellite instantaneous collision probability computation using equivalent volume cuboids. *Journal of Guidance, Control, and Dynamics*, 43(9):1757–1763.