The dynamical modelling of dwarf spheroidal galaxies using Gaussian-process emulation

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

AMERY GRATION

Department of Physics & Astronomy
University of Leicester

January 2020

Abstract

This thesis concerns the use of Gaussian-process emulation (Sacks et al., 1989) to build metamodels of computationally expensive dynamical models of dwarf spheriodal galaxies. These metamodels are computationally cheaper to evaluate than the models that they emulate, and hence have the potential to render tractable previously intractable problems in galactic dynamics. The first part of the thesis deals with the theoretical foundations of Gaussian-process emulation (GPE) while the second part deals with the application of GPE to the modelling of dwarf spheroidal galaxies. I give a description of the general principles of modelling and metamodelling, formally defining a physical model, and showing that the parameter spaces of such models may be made metric or pseudometric spaces. I give a formal treatment of the foundations of GPE and, building on the work of Parzen (1959), give a novel derivation of the GPE predictor and mean-squared error. I also set right some confusion and errors in the literature. In particular, I show that the GPE predictor presented by Rasmussen and Williams (2006) is biased. I quantify this bias, and discuss the circumstances under which it will be significant. In modelling dwarf spheroidal galaxies, I adopt the distribution-function approach, and use GPE to construct a metamodel of the log-likelihood. First, I construct a toy model of a dwarf spheroidal galaxy which I fit using synthetic data drawn from the same toy model. I maximize the log-likelihood using the method of efficient global optimization (Jones et al., 1998), finding that I am able to recover robust confidence regions for the parameter vector, galactic density, and velocity anisotropy with fewer than 100 model evaluations. Second, I construct a more general model. Although the resulting predictions are accurate, the metamodel fails validation, indicating that we may not trust the confidence regions associated with these predictions. We conclude that the usual simplifications made in implementing GPE render it inadequate for the task of predicting the log-likelihood in this case, and that we must consider more general methods.

Preface

This thesis concerns the use of Gaussian-process emulation (GPE) to build metamodels of computationally expensive dynamical models of dwarf spheriodal galaxies. These metamodels are computationally cheaper to evaluate than the models that they emulate, and hence have the potential to render tractable previously intractable problems in galactic dynamics. The thesis is in two parts. The first part (Chaps 1, 2, and 3) deals with the theoretical foundations of GPE, while the second part (Chaps 4 and 5) deals with the application of GPE to the modelling of dwarf spheroidal galaxies.

Chapter 1 concerns the general principles of modelling and metamodelling. I begin by formally defining a mathematical model—as an indexed family of mathematical objects (functions, differential equations, etc.) These indices are referred to as *parameters*. Typically they are tuples of real numbers, each representing a different physical quantity. The set of all such indices, when equipped with some structure forms a parameter space. However, parameter space is not formally defined in the literature. I here make a first attempt at such a definition, giving a discussion of the kind of structure we would like parameter space to have. To this end I introduce the concept of *flavour spaces* (Def. 10) and, as an example, show how they may form *flavourful vector spaces* (Def. 11). In the case of dynamical modelling we are interested in models of probability density functions (specifically probability density functions on phase space). I show that the parameter spaces of such models may be made metric or pseudometric spaces (Subsec. 1.1.2, esp. Props 18 and 19). In the remainder of the chapter I present the well-established theory of the maximum-likelihood estimation of model parameters.

Chapter 2 concerns the theoretical basis of GPE. The first use of GPE is rightly credited to Sacks et al. (1989). It is, in short, the use of the *best linear predictor* (BLP) or *best unbiased linear predictor* (BLUP) to predict the outcome of a computer experiment. Its origins, not properly acknowledged in the machine learning literature, therefore lie in the statistical literature of the 1940s and 1950s (the invention of the BLUP is normally credited to Henderson, 1950). The most natural setting for the derivation of the BLP and BLUP is the *reproducing kernel Hilbert space*. The use of reproducing kernel Hilbert spaces in statistics was pioneered by Parzen (1959), and my presentation in places closely follows his. His arguments combine mathematical rigour and geometrical insight. I add detail and clarification where I think it appropriate. Most commonly, other methods are used to derive the BLP or BLUP (see, for example, Sacks et al., 1989, who use the method of Langrange multipiers to derive the BLUP) but Parzen's methods have the advantage of greater generality.

In the first part of the chapter (Sec. 2.1) I give a thorough account of random processes. In doing so, I synthesize a large amount of material from numerous sources, in particular the graduate texts of Adler (1981), Gikhman and Skorohod (1974), and Yaglom (1962). However, in discussing stationary and isotropic random processes I introduce a new definition of *invariant* random processes (Def. 41). This generalizes the definition of stationary and isotropic random processes, which then become special cases of my new definition. In the second and third parts (Secs 2.2 and 2.3) of the chapter I outline the theory of reproducing kernel Hilbert spaces and their application to random process. In the fourth part I outline the theory of linear prediction, and in the fifth part outline the use of reproducing kernel Hilbert spaces in linear prediction. From Parzen's prediction theorem (Thm 107) I make a novel derivation of the expressions for GPE (Sec. 2.5.3, eqs 2.181, and 2.183). I finally discuss the relationship between GPE and geostatistics, where the use of the BLP and the BLUP goes by the name 'kriging'. The practical implementation of GPE relies heavily on kriging methods in a way not properly acknowledged in the literature. I identify some limitations of the kriging methodology, and a way to remedy them. Specifically, I point out that the metric structure of parameter space I have developed in Chapter 1 allows me to define new covariance functions, tailored for a particular use. I do not take advantage of this ability, but leave it for further research. Throughout this chapter I give proofs where they are useful to the development of the argument. Where I do give proofs, they are my own unless otherwise stated.

Chapter 3 concerns the practical implementation of GPE, and sets right some confusion and errors in the literature. I emphasize the importance of validating the results of GPE, a procedure that appears to have been entirely overlooked in the application of GPE to astrophysical problems thus far. The machine learning literature does not adequately distinguish between the BLP and the BLUP, and in some sources only the BLP is presented. Rasmussen and Williams (2006), for example, present only the BLP, though they do not call it such. Moreover, the particular expression they derive for the BLP is based on an unwarranted assumption.[1] It is, in general, biased. It gives biased predictions, and biased values for the confidence intervals associated with these prediction. This bias is not acknowledged by Rasmussen and Williams. I quantify it (Sec. 3.4, Eq. 3.38), and discuss the circumstances under which it will be significant.

Chapter 4 concerns the proof-of-concept application of GPE to the dynamical modelling of dwarf spheroidal galaxies. I adopt the distribution-function approach to dynamical modelling, and construct a toy model of a dwarf spheroidal galaxy which I fit using maximum-likelihood methods and synthetic data drawn from the same toy model. I assume that mass follows light and use Ossipkov and Merritt's method to construct a phase-space probability density function for a galaxy with Plummer-type density. I then use GPE to construct a metamodel of the log-likelihood, which I maximize using the method of *efficient global optimization* (Jones et al., 1998). With fewer than 100 model evaluations I am able to recover robust confidence regions for the parameter vector (Fig. 4.11), galactic density, and velocity anisotropy (Fig. 4.12). This material has already been published in *The monthly notices of the Royal Astronomical Society* (Gration and Wilkinson, 2019).

Chapter 5 concerns the application of GPE to more general models of dwarf spheroidal galaxies. I assume that the galaxy consists of a dark-matter halo and a single population of stars, each of which has a density of the generalized Hernquist type. I find the phase-space probability density function by modelling it using the method of Gerhard (1991), and then solving the density equation using the method of Cuddeford and Louis (1995). Again, I use GPE to construct a metamodel of the log-likelihood. Although the resulting predictions are accurate, the metamodel fails validation, indicating that we may not trust the confidence regions associated with these predictions. I conclude that the usual kriging methods are inadequate for the task of predicting the log-likelihood in this case, and that we must consider more general methods. I leave this for future work.

[1] To compute the BLP of a random variable Z based on a set of random variable $X_1, X_2, \ldots, X_n$ we must know the joint distribution of Z and $X_1, X_2, \ldots, X_n$, which we assume to be Gaussian in the case of GPE. Rasmussen and Williams assume the mean to be both constant and zero. They appear to claim (wrongly) that they may do this without loss of generality.

Acknowledgements

The relationship between a PhD student and his supervisor requires trust. The PhD student must trust that the problem chosen by his supervisor is interesting and tractable, while the supervisor must trust that his student has the ability to pursue the problem at hand and make it his own. I should like to thank my supervisor, Mark Wilkinson, for upholding his end of bargain so abundantly. I can only hope that I have upheld my end. I have found the subject of this thesis—Gaussian-process emulation applied to the dynamical modelling of dwarf spheroidal galaxies—to be utterly fascinating. But it was Mark who realized this first, and I should like to thank him for his foresight. For a work of physics this thesis is rather eccentric. The first three chapters almost exclusively address the mathematical setting of Gaussian-process emulation. I believe that some interesting results have arisen from the investigations I present in these chapters, and that they have direct consequences for physics. But the work was slow and involved and I should like to thank Mark for having faith in me while it was under way.

I should also like to thank the other members of the Theoretical Astrophysics Group for their help and advice over the course of my time at Leicester: Andrew King, Walter Dehnen, Sergei Nayakshin, Richard Alexander, Chris Nixon, and Graham Wynn. Last but not least I should like to thank Sylvy Anscombe for numerous discussions of the mathematical structure of Gaussian-process emulation.

Contents

List of figures

List of tables

Notation

I use an uppercase letter to represent a random variable, and a lowercase letter to represent a realization of that random variable. For example, a random variable X, may have a realization x, y, z, etc. I denote the probability density function of a random variable X by f_X. I denote the cumulative probability distribution of a random variable X by F_X.

The following symbols have reserved meanings.

$\langle \cdot, \cdot \rangle$	Inner product
$\langle \cdot, \cdot \rangle_V$	Inner product associated with a particular vector space, V
$\\| \cdot \\|$	Norm of a vector
$\mathbf{V}$	Inner-product space $(V, +, (\lambda), \langle \cdot, \cdot \rangle)$
$\mathbf{H}$	Hilbert space $(H, +, (\lambda), \langle \cdot, \cdot \rangle)$
r	Correlation function
k	Covariance function or positive-semidefinite kernel
R	Correlation matrix corresponding to correlation function r
K	Covariance matrix corresponding to covariance function k
$\mathbf{X}$	Random process, $\{X_t\}_{t \in T}$, with arbitrary index set T
$\mathbf{\Omega}$	Measurable space, $(\Omega, \mathscr{M})$ on which a random variable or random process is define
$\mathrm{E}(\cdot)$	Expectation of a random variable

$\mathrm{var}(\cdot)$	Variance of a random variable
$\mathrm{cov}(\cdot)$	Covariance of a random variable
$\mathrm{corr}(\cdot)$	Correlation of a random variable
$L(\mathbf{X})$	Linear span of a random process $\mathbf{X}$
$\mathbf{L}(\mathbf{X})$	Hilbert space spanned by the random process $\mathbf{X}$
$\mathbf{G}_k$	Reproducing kernel Hilbert space with reproducing kernel k
$\mathrm{ev}(\cdot)$	Evaluation map
$\mathbf{R}^{\Xi}$	Set of functions from Ξ to $\mathbf{R}$
$(\mathbf{H}, \Xi)$	Hilbert space of functions $f : \Xi \longrightarrow \mathbf{R}$
$L^2(\cdot)$	Square-integrable functions on a set
$\mathbf{L}^2(\cdot)$	Hilbert space of square-integrable functions on a set
ψ	Congruence between two Hilbert spaces
$\inf(\cdot)$	Infimum of a set
$\sup(\cdot)$	Supremum of a set
$\min(\cdot)$	Minumum of a set
$\max(\cdot)$	Maximum of a set
$\mathbf{N}$	Set of natural numbers, including zero
$\mathbf{Z}$	Set of integers
$\mathbf{Q}$	Set of rational numbers
$\mathbf{R}$	Set of real numbers.
$\mathbf{C}$	Set of complex numbers

‘A physical theory or world picture is a model (generally of a mathematical nature) and a set of rules that connect the elements of the model to observations.’

STEPHEN HAWKING AND LEONARD MLODINOW, *THE GRAND DESIGN*

Chapter 1

Modelling and metamodelling

The Miky Way is orbited by approximately 60 satellite galaxies (McConnachie, 2012). All but the Large Magellanic Cloud are *dwarf* galaxies, with stellar masses less than 10^9 $\mathrm{M}_\odot$. These dwarf galaxies fall into two categories: those containing gas, which therefore exhibit ongoing star formation, and those not containing gas, which therefore exhibit no such ongoing star formation. The first category of dwarfs are known as dwarf irregular (dIrr), and the second as *dwarf spheroidal* (dSph). Dwarf galaxies are diffuse and intrinsically faint, and all except the Small Magellanic Cloud (a dIrr galaxy) have been discovered since 1938, when Shapley (1938) discovered the Fornax dwarf spheroidal galaxy. The dSph galaxies that were bright enough to be discovered before the Sloan Digital Sky Survey (SDSS) are known as the *classical* dwarf spheroidal satellites. There are eight of these: Carina, Draco, Fornax, Leo I, Leo II, Sculptor, Sextans, and Ursa Minor. Those discovered since, by SDDS and subsequent surveys (such as the ATLAS survey performed on the VLT Survey Telescope, or the Dark Energy Survey), are known as *ultrafaint* dwarf spheroidal satellites.

These dwarf galaxies are some of the most dark-matter dominated stellar systems we know, with half-light mass-to-light ratios of 10 or more (Walker, Mateo, Olszewski, Peñarrubia, Evans and Gilmore, 2009). Indeed some ultrafaint dSph satellites have half-light mass-to-light ratios in excess of 1000 (Wolf et al., 2010). This makes the dwarf satellites valuable laboratories in which to study the properties of dark matter. The stellar populations of the dSph galaxies are the most direct tracers of dark matter. Of these, the classical dSph galaxies are particularly interesting since we have excellent observations of their stellar kinematics (namely stellar sky positions and line-of-sight velocites). The largest of these data sets were made by Walker, Mateo, Olszewski,

galaxy	d (kpc)	$r_{\rm h}$ (pc)	$M(r_{\rm h})$ (10^6 $\rm M_\odot$)	M_* (10^6 $\rm M_\odot$)	$M_{\rm V}$ (mag)
Leo I	254 ± 15	251 ± 27	12	5.5	-12.0
Leo II	233 ± 14	176 ± 42	4.6	0.74	-9.8
Sextans	86 ± 4	695 ± 44	25	0.44	-9.3
Ursa Minor	76 ± 3	181 ± 27	9.5	0.29	-8.8
Carina	105 ± 6	250 ± 39	6.3	0.38	-9.1
Draco	76 ± 6	221 ± 19	11	0.29	-8.8
Sculptor	86 ± 6	283 ± 45	14	2.3	-11.1
Fornax	147 ± 12	710 ± 77	56	20	-13.4

Table 1.1 Properties of the classical dwarf spheroidal galaxies: bulk distance from the Sun, d, half-light radius, $r_{\rm h}$, total mass contained within half-light radius, $M(r_{\rm h})$, stellar mass, $M_*(r_{\rm h})$, and absolute V-band magnitude, M_V (McConnachie, 2012).

Peñarrubia, Evans and Gilmore (2009) using the Michigan-MIKE Fibre Spectrograph (MMFS) of the Magellan Clay telescope at the Las Campanas Observatory, Chile. These include observatations of 775 stars in Carina, 2500 stars in Fornax, 1365 stars in Sculptor, and 440 stars in Sextans. Smaller data sets have been compiled for the remaining classical dSph galaxies by other authors, and include observations of 300 stars in Leo I (Mateo et al., 2008), 170 stars in Leo II (Koch et al., 2007), 100 stars in Draco, and 100 stars in Ursa Minor (Kleyna et al., 2002). In Table 1.1 I summarize the gross features of the classical dSph galaxies. They have stellar masses between 10^5 $\rm M_\odot$ and 10^7 $\rm M_\odot$ (Draco is the faintest, and Fornax the brightest), orbit at distances of order 100 kpc, and (Draco and Ursa Minor are the closest at 76 kpc, and Leo I the most distant at 254 kpc) have half-light radii of several hundred parsecs (Leo II is the smallest at 176 pc, and Fornax the largest at 710 pc).

That the dSph glaxies are dominated by dark matter is known from the fact that their kinematics may not explained by their stellar mass. Some additional, non-luminous, matter is required. This fact is observed in many astrophysical systems, including spiral galaxies and galaxy clusters. It is termed the *missing-mass problem*, and was first recognised by Zwicky (1933) in his observations of the Coma cluster. Later, Rubin and Ford Jr. (1970) and Rubin et al. (1978) observed it in the Andromeda Galaxy, M31. In dispersion-supported systems like dSph galaxies and galaxy clusters the missing-mass problem is most clearly manifest as too great a dispersion in the velocity of its members. It may therefore be observed using line-of-sight velocites, from which we compute the velocity dispersion using the virial theorem. The excess dispersion requires excess mass in the cluster. In disc galaxies the problem manifests itself as too great a circular speed for tracers of the galactic potential. It may therefore be observed directly in the line-of-sight velocities of gas or stars, which may be compared to Newtonian predictions. HII may be used to observe rotation within roughly the extent of the stellar disc, and HI to observe rotation beyond the stellar disc. These circular velocities are oberved to be roughly constant at large radii, contradicting the expectation that they be asymptotically Keplerian, i.e. decreasing as $v_c(r) \sim \sqrt{M/r}$. This excess circular speed, like excess velocity dispersion, necessitates excess mass within the observed radius.

It was first proposed that this missing mass might exist be in the form of massive compact halo objects (MACHOS), i.e. low-mass black holes, neutron stars, brown dwarfs, or unbound Jupiter -like exoplanets (Griest, 1991). These are now largely excluded as dark-matter candidates (Brandt, 2016), on the grounds that we observe them neither directly, in microlensing surveys, nor indirectly, by their dynamical effects (such as the disrupting of wide halo binaries, stellar streams, or the dynamical heating of their host systems). Instead, dark matter is properly understood as an intrinsic compononent of Lambda cold dark matter (ΛCDM) cosmology. In ΛCDM, the universe is modelled as spatially flat and containing a mixture of radiation, ordinary matter, nonbaryonic dark matter, and dark energy. The universe begins in a hot Big Bang and cools as it expands. In this model dark matter is essential to explain the structure that we observe in the universe (Frenk and White, 2012; Primack, 2009). Quantum fluctuations cause perturbations in the dark matter's density that are then expanded to cosmological scales by inflation at 10^{-32} s. This seeds the early universe with dark matter structures of all scales. These structures collapse under gravity, with smaller structures collapsing first to form dark matter halos. Larger structures form by the merging of these small halos. Dark matter is gravitationally dominant, and baryonic matter falls onto these halos to form galaxies. This process is known as hierarchical structure formation (Blumenthal et al., 1984).

This structure formation must be probed using simulations (Vogelsberger et al., 2014). These fall into two categories: those that simulate only the dark-matter content of the universe, such as Millennium (Springel et al., 2005), and Aquarius (Springel et al., 2008); and those that simulate both the dark-matter and baryonic content of the universe, such as Illustris (Vogelsberger et al., 2014), and EAGLE (Schaye et al., 2015). The first kind of simulations probe the large-scale distribution of dark matter, while the second also probe the structure of dark matter halos, which are affected by baryonic feedback. In this way we observe the formation of Galaxies (at separations of about a megaparsec), galaxy groups (like the Local Group, consisting of tens of galaxies), galaxy clusters (consisting of hundreds or thousands of galaxies), superclusters (like the Virgo Supercluster, consisting of hundreds of galaxy groups and clusters, or tens of thousands of galaxies), filaments, and walls, separated by enormous voids (of scales of tens to hundreds of megaparsecs).

In ΛCDM dark matter is a an as-yet unidentified particle that does not form part of the standard model, but interacts with standard-model particles by the weak force. By not interacting by the electromagnetic force, it remains invisible. Initially, the contents of the universe is in thermal equilibrium. Particles freely interact with each other, being continually created and destroyed. The interaction rate, and hence the relative abundances of species, is determined by the universe's temperature. Equilibrium is maintained by scattering, and in order to maintain equilibrium the scattering rate must exceed the rate at which the universe is expanding. Once the two rates have acheived equality for a given species that species falls out of thermal equilibrium with the others and its abundance is fixed. This phenomemon is known as *freeze out*. Once frozen out a species is said to have become *decoupled* from the rest of the universe's contents. Dark matter is said to be cold because it is nonrelativistic at this time.

According to ΛCDM the universe therefore evolves as follows. The early universe (lasting from the end of the first 10^{-12} s to the first 377 000 yr) begins as a plasma of elementary particles. By the end of the first second, quarks become bound to form hadrons (including protons and neutrons), and the temperature of the universe is 10^{10} K, at which point neutrinos freeze out and become decoupled. From this point, the mean-free path of neutrinos becomes infinite and they free-stream through the universe. Because these neutrinos have been scattered from a nonreflective, opaque surface of uniform temperature, they have the spectrum of a blackbody. These free-streaming neutrinos and their black-body spectrum may still be observed in the form of the *cosmic neutrino background*. Between 10 s and 10^3 s nucleosynthesis occurs, and protons become bound to neutrons to form atomic nuclei. Principally these nuclei are deuterium, He-3, He-4, and Li-7. They are observed, in these abundances, today (Schramm and Turner, 1998). After 377 000 yr electrons become bound to these nuclei to form neutral atoms. This process is known as *recombination*, though this is a misnomer, as electrons had not been bound to nuclei at any time before. At this time the temperature of the universe has fallen to 4000 K, and photons are frozen out and become decoupled from matter. Photons, like neutrinos now free-stream through the universe, again exhibiting a blackbody spectrum. These photons and their spectrum may be observed today in the from of the *cosmic microwave background* (Penzias and Wilson, 1965). After photon freeze-out, the early universe is said to end, and it enters a period known as the *dark ages*, during which the only new source of photons is the spin-flip transition between the two hyperfine levels of the 1s ground state of hydrogen. The dark ages come to and end at 200 Myr with the formation of the first stars, which form as hydrogen undergoes gravitational collapse. The earliest galaxies form from 380 Myr. Once stars have formed, the photons they radiate reionize the universe's hydrogen. Reionization is complete by 1 Gyr, and the universe enters its present state.

As the universe expands over the course of its history, the density of its component populations reduces. At first radiation is the most dense component, and the universe is said to be *radiation dominated*. At 47 000 yr matter becomes the most dense component, and the universe becomes *matter dominated*. In both cases the universe continues to expand but the rate of expansion decreases with time. The universe is said to *decelerate*. The era of matter domination lasts until 1 Gyr, when the density of matter falls below that of dark energy, and the universe becomes *dark-energy dominated*. This causes the rate of expansion to increase with time, and the universe to *accelerate*. This acceleration is observed in the recession of Type 1a supernovae (Riess et al., 1998; Perlmutter et al., 1999). Today the observed mass-energy density of the contents of the universe has been estimated by the Planck Collaboration (2018) to be 25.9 % dark matter 4.86 % baryonic matter, and 69.1 % dark energy.

ΛCDM therefore successfully predicts the large-scale structure of the universe, the cosmic microwave background, the abundances of the elements, and cosmic expansion. However, it faces some challenges in predicting structure at small scales (namely galactic scales of one megaparsec or less). These arise from discrepancies between the predictions of dark matter-only simulations and observations. The three best-known challenges are the *missing satellites problem* (Klypin et al., 1999; Moore et al., 1999), the *too-big-to-fail problem* Boylan-Kolchin et al. (2011), and the *core-cusp problem* (Flores and Primack, 1994; Moore, 1994).

The missing satellites problem results from the fact that Milky Way-like halos in dark matter-only cosmological simulations host thousands of satellite halos. This differs dramatically from the 60 galaxies satellite galaxies that we observe. In particular, cosmological simulations predict many more small halos than we observe small galaxies. Future surveys may find more ultra-faint satellites, but they are unlikely to find the thousands required. It may well be that dark matter halos are inefficient at forming galaxies at low mass. For example, as protogalactic gas is warmed by reionization, that gas may be prevented from collapsing onto low-mass halos. It may also be that baryonic feedback may prevent galaxy formation. In this case, low-mass halos would remain dark, and we would not observe as many low-mass galaxies as there are low-mas halos. We would expect to observe only high-mass satellites, and solve the missing satellites problem by proposing that the observed satellites galaxies reside in the highest-mass halos. A consequence of this proposed solution is that the central densities of Milky-Way satellites should be consistent with the central densities of highest-mass halos predicted by simulations (these are of order 10^{10} $M_\odot$). The too-big-to-fail problem results from the fact that the most-massive halos found in Aquarius (Springel et al., 2008) and Via Lactea II (Diemand et al., 2008). are too dense in centre. It is not clear why would galaxies not form in the highest-mass halos whilst forming forming in halos of lower mass. These galaxies should be too big to fail in this way. Recently, Ostriker et al. (2019) have suggested that this is the result of hierarchical growth itself. The Milky Way has grown by mergers, and as over its history has always been most likely to merge with the brightest of its satellites. After merger the difference in luminosity between the Milky Way and its brightest satellite is greater than it was before merger. Ostriker et al. (2019) claim that this gap in luminosity is indeed obervered in the most-recent baryonic ΛCDM simulations EAGLE and IllustrisTNG.

The core-cusp problem results from the fact dark matter-only simulations predict a universal density profile for dark-matter halos. Regardless of size, the spherically averaged density of a halo is well approximated by the Navarro-Frenk-White (NFW) formula:

$$\rho(r) = \rho_0 \left(\frac{r}{b}\right)^{-1} \left(1 + \frac{r}{b}\right)^{-2} \tag{1.1}$$

where r is distance from the galactic centre, $\rho_0 > 0$ is a normalizing factor or characteristic density, and $b > 0$ is the transition radius. This is a split-power law, in which density falls off as $\rho(r) \propto r^{-3}$ for large radii, and as $\rho(r) \propto r^{-1}$ for small radii. Such halos are said to exhibit a *cusp* in their central densities. However, observations suggest that the density of some dSph galaxies is roughly constant at small radii, in which case they are said to exhibit a *core* (de Blok, 2010; Battaglia et al., 2008; Strigari et al., 2010; Breddels and Helmi, 2013; Read and Steger, 2018). The core-cups problem may be spurious, and the cores we supposedly observe may not be statistically significant (Wolf and Bullock, 2012). Or it may be that these cores are genuine and that the problem is resolved by baryonic physics. For example, it has been shown that supernova-driven flattening may turn dark-matter cusps into cores (Navarro et al., 1996; Read and Gilmore, 2005; Mashchenko et al., 2008). In either case we require more robust dynamical modelling, which will allow us definitively identify the problem, or provide us with good targets for evolutionary simulations and help us better understand the history of these galaxies.

To date, dynamical modelling of dSphs has for the most part been based on very restrictive simplifying assumptions, namely that the dSphs are spherical and in equilibrium (see, for example, Wilkinson et al., 2002, Walker, Mateo, Olszewski, Peñarrubia, Evans and Gilmore, 2009, Strigari et al., 2010, and Read and Steger, 2017). While some authors have considered more general models (Breddels and Helmi, 2013), relaxation of these assumptions results in models that are significantly more computationally expensive, often prohibitively so. When a model is computationally expensive we may use a *metamodel*, i.e. a model of the model that is computationally cheaper. One commonly used method of metamodelling is *Gaussian-process emulation* (GPE). In the astrophysical literature it has been used to fit exoplanetary transit and secondary-eclipse light curves (Gibson et al., 2012 and Evans et al., 2015), to map interstellar extinction within the Milky Way (Sale and Magorrian, 2014, 2019), and to fit semi-analytic models of galaxy formation (Bower et al., 2010), while in the cosmological literature it has has been used to predict the non-linear matter power spectrum in the Coyote Universe simulation (Heitmann et al., 2009), and to fit gravitational-wave models (Moore et al., 2016). However, it has not been used in galactic dynamics.

Our principal interest is in what observational data can tell us about the distribution of dark matter in a dSph. Which dark-matter distributions do the data rule out? Which best account for the observations? We will adopt the distribution-function approach to the modelling of dSphs. We construct a model of the phase-space distribution function, compute the observables, and recover the parameter of this model using likelihood methods. We will be interested in using GPE to create metamodels of this likelihood function.

1.1 MODELLING

Before we embark on a discussion of metamodelling and its application to galactic dynamics, it will be worth our while to discuss some formal aspects of modelling in general. In particular it will be worth our while to discuss the mathematical structure of parameter space. Although the term 'parameter space' is ubiquitous, I know of no source that defines it. A full definition will be beyond the scope of this thesis, and we will instead be concerned with showing that, in the context of dynamical modelling, parameter space may be considered a metric or pseudometric space. Nonetheless, we will work *towards* a formal definition of parameter space, the complete statement of which must be left for future work.

We begin with a formal definition of an indexed family.

Definition 1 (indexed family). Let Θ and Ξ be nonempty sets and let f be a surjective (i.e. many-to-one) function such that

$$f\colon \Theta \longrightarrow \Xi \tag{1.2}$$

$$\theta \longmapsto f(\theta). \tag{1.3}$$

We use the notation $\xi_\theta := f(\theta)$. We call f the *indexing function*, we call the set Θ the *index set*, and we call the set Ξ the *indexed set*. The indexing function, f, is in fact a set of pairs $\{(\theta, \xi_\theta) : \theta \in \Theta\}$, which we denote $(\xi_\theta)_{\theta\in\Theta}$. We call this surjective function the *family of elements in* Ξ *indexed by* Θ (or, the *indexed family*, or the Θ-*indexed family*).

When the elements of an indexed family take on a particular significance we will refer to it as a *model*. When we view an indexed set as a model, we will refer to the indices as *parameters* and to the index set as the *set of parameters* (or *parameter set*).

Example 2 (gravitational central force). A *model of a particle's acceleration under a gravitational central force* is

$$\left(\ddot{x}(t) = -\frac{\mathrm{G}m}{|x(t)|^3}x(t)\right)_{m\in M} \tag{1.4}$$

where x is the particle's displacement from the centre of force, and the parameter m is the *mass of the source particle*. The parameter set M (denoted Θ in Def. 1) is the set of all such masses, i.e. the set of positive real numbers.

Example 3 (coupled pendulums). A *model of a pair of coupled pendulums* is

$$\left(\{m_1\ddot{x}_1 = -\frac{m_1}{l_1}x_1 - k(x_1 - x_2), m_2\ddot{x}_2 = -\frac{m_2}{l_2}x_2 - k(x_1 - x_2)\}\right)_{\theta\in\Theta} \tag{1.5}$$

where x_1 and x_2 are the displacements of the pendulums from equilibrium, and the parameter θ is a tuple of values (m_1, m_2, l_1, l_2, k).[1] The parameter set is $\Theta = \mathbf{R}^5_{>0}$.

Example 4 (statistical model). If the elements of an indexed family are probability density functions then we call that family a *statistical model*.

Two types of statistical model are of particular interest to us in astrophysics: models of measurement errors, and distribution-function models of dynamical systems.

[1] Note the parameter is the tuple (m_1, m_2, l_1, l_2, k). Its components are not themselves parameters, and we do not refer to, say, the mass m_1 as a parameter of the system.

Example 5 (Gaussian measurement errors). We represent the error of a measurement as a random variable, X, which we typically assume to have a Gaussian distribution. A *model of measurement errors* is an indexed family of Gaussian distributions

$$(f_X(\cdot;\sigma) : \mathbf{R} \longrightarrow \mathbf{R})_{\sigma \in \Sigma} \tag{1.6}$$

where $f_X(\cdot;\sigma)$ is the normal probability density function with zero mean,[2] i.e. where

$$f_X(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{x^2}{2\sigma^2} \right). \tag{1.7}$$

The parameter is the standard deviation, σ, and the parameter set is the set $\Sigma = \mathbf{R}_{>0}$.

Example 6 (distribution function model of a dynamical system). If the elements of an indexed family are probability density functions on phase-space then we call that family a *distribution function model of a dynamical system*.

Remark 7. Note that Examples 4, 5 and 6 are indexed families of functions, that Example 2 is an indexed family of differential equations, and that Example 3 is an indexed family of pairs of coupled differential equations.

A parameter set Θ is called a *parameter space* if it is endowed with some structure. For example, it might be endowed with the structure of a vector space or metric, or with a weaker structure, such as that of a topological space. (I will take the definition of these structures for granted, but for convenience include them in App. A.) What structure, then, should parameter space have?

1.1.1 Flavourful spaces

A parameter is typically a tuple consisting of physical quantities (mass, length, density, etc.) and numbers. These numbers may be real, positive, nonnegative, or integer, etc. It makes sense, for example, to add masses to masses or lengths to lengths. Similary, it makes sense to multiply masses by a constant, or to multiply lengths by a constant. It makes sense to multiply lengths by lengths, or to multiply masses by masses. But, it does not make sense to add masses to lengths, nor does it make sense to multiply masses by lengths. We may think of each element of the tuple as having a distinct *flavour*, and the parameter as being *flavourful*. We must accommodate this flavourful property of parameters in the appropriate structure of parameter space.

We may illustrate this idea by constructing a *flavourful vector space*, noting that, in general, parameter space will not be a vector space. The key is to recognize that the flavourful property

[2] Here we follow the convention that the PDF of random variable X always denoted f_X.

of parameters can be described by giving a vector space a distinguished (i.e. special) Cartesian product. Recall that a Cartesian product of vector spaces is defined by their *direct sum* (see, for example, Lang, 2004).

Definition 8 (direct sum). Let U be a vector space and let $V, W \subseteq U$ be subspaces of U. Then U is the *direct sum of* V *and* W if $U = V + W$ and $V \cap W = \{0\}$. We write $U = V \oplus W$. We call V and W *factors* of U.

A vector space may be *decomposed* into its factors.

Definition 9 (decomposition). A *decomposition* of a vector space V is a set of vector spaces $\{V_i : i = 1, 2, \ldots, n\}$ such that V is the direct sum of these vector spaces, i.e. such that

$$V = V_1 \oplus V_2 \oplus \ldots \oplus V_n. \tag{1.8}$$

We may then define a flavourful vector space as a vector space together with a given decomposition. In order to do this, we fix a finite family of vector spaces, which *represent the flavours of our parameters.*

Definition 10 (flavour space). We fix a family $\mathscr{U} = (U_i)_{i=1}^{n}$ of vector spaces. These vector spaces are called *flavour spaces.*

Each factor of the decomposition should now be isomorphic to one of these flavour spaces.

Definition 11 (flavourful vector space). Fix a finite family $\mathscr{U} = (U_i)_{i=1}^{n}$ of flavour spaces, and let V be another vector space. A *flavourful decomposition of* V *relative to* $\mathscr{U}$ is a decomposition of V into a direct sum

$$V = V_1 \oplus V_2 \oplus \ldots V_m, \tag{1.9}$$

such that for each $i \leq m$ there exists $j \leq n$ such that V_i is isomorphic to U_j. A *flavourful vector space* is a vector space equipped with a flavourful decomposition.

Each subspace is a *Cartesian factor* of the parameter space, and the parameter space itself is the *Cartesian product* of these factors. We may refer to it as a *product space*. Note that if each component of the parameter vector has a different flavour then $n = m$.

We will be able to make equivalent decompositions of a variety of mathematical structures, for example topological spaces, groups, etc. For example, we might be interested in the seven fundamental physical dimensions (length, mass, time, current, temperature, number, and luminous intensity), which may not all be represented as vector spaces. Consider, for example, the parameter space $\mathbf{R}_{\geq 0}$, representing temperature, all but the zero element of which fails to have an additive inverse. Consider, also, a compact region, $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n$, which is not closed under addition. This illustrates the fact that, in general, it will not be possible to endow parameter space with the structure of a vector space. However, we will not consider decompositions of such general spaces here.

If we are to quantify the distance between elements of a flavourful vector space we will wish to equip it with a metric. We do this by equipping it with an inner product, which then induces (i.e. is used to define) a metric. To do this we may adopt one of two approaches. We may either equip each factor with an inner product, and then extend this to the product space, or we may equip the product space with an inner product, and restrict it to each factor. By introducing an inner product, we also allow ourselves to discuss the orthogonality of our factors.

Adopting the first approach, we suppose that each flavour space is an inner product space. By abuse of notation we write, $U_i = (U_i, \langle\cdot,\cdot\rangle_i)$, where $\langle\cdot,\cdot\rangle_i : U_i \times U_i \longrightarrow \mathbf{R}_{\geq 0}$ is an inner product on the flavour space U_i. Let $V = V_1 \oplus V_2 \oplus \cdots \oplus V_m$ be a flavourful space, and let $(a_i)_{i=1}^m$ be a sequence of positive numbers, which we will call *scale constants*. We may now define an inner product, $\langle\cdot,\cdot\rangle$, on the product space, V, given by

$$\langle \sum_i u_i, \sum_i v_i \rangle = \sum_i a_i \langle u_i, v_i \rangle_i \tag{1.10}$$

for all $u_i, v_i \in V_i$. Then $(V, \langle\cdot,\cdot\rangle)$ is an inner product space. Moreover, $V_1 \oplus V_2 \oplus \cdots \oplus V_m$ is an orthogonal direct sum, i.e. the factors of the flavourful space are orthogonal to each other. (To see this, consider vectors $u \in V_i$ and $v \in V_j$ where $i \neq j$. Then $\langle u, v \rangle = a_i \langle u, 0 \rangle_i + a_j \langle 0, v \rangle_j = 0$.)

The inner product $\langle\cdot,\cdot\rangle_i$ induces a norm on the factor V_i, namely the function $\|\cdot\|_i : V_i \longrightarrow \mathbf{R}_{\geq 0}$ given by

$$\|u\|_i = \sqrt{\langle u, u \rangle_i}, \tag{1.11}$$

and in turn, a metric, namely the function $d_i : V_i \times V_i \longrightarrow \mathbf{R}_{\geq 0}$ given by

$$d_i(u, v) = \|u - v\|_i. \tag{1.12}$$

Similarly, we have the norm $\|\cdot\| : V \longrightarrow \mathbf{R}_{\geq 0}$ given by $\|u\| = \sqrt{\langle u, u \rangle}$, and metric $d : V \times V \longrightarrow \mathbf{R}_{\geq 0}$ given by $d(u, v) = \|u - v\|$. The crucial thing is as follows. Consider the unit vectors, $u_i \in U_i$ and $u_j \in U_j$, such that $\|u_i\|_i = \|u_j\|_j = 1$. Then the ratio of the norms $\|u_i\| / \|u_j\| = \sqrt{a_j / a_i}$.[3] We may think of the number $\sqrt{a_i / a_j}$ as a *relative scaling* of flavour spaces U_i and U_j.

[3] To see this, note that $\|u_i\|^2 = \langle u_i, u_i \rangle = a_i \langle u_i, u_i \rangle_i = a_i \|u_i\|_i^2$. Therefore $\|u_i\|^2 / \|u_j\|^2 = (a_i \|u_i\|_i^2)/(a_j \|u_j\|_j^2) = 1$. Hence $\|u_i\|_i / \|u_j\|_j = \sqrt{a_j / a_i}$.

Remark 12. We have extended the inner products on the factors of V to V itself in such a way that these factors become orthogonal. There are ways of extending these inner products such that the factors do not become orthogonal, but we do not pursue them here.

Adopting the second approach, we suppose that the product space, V, is an inner product space, rather than its factors. Again, by abuse of notation we write, $V = (V, \langle\cdot,\cdot\rangle)$, where $\langle\cdot,\cdot\rangle : V \times V \longrightarrow \mathbf{R}_{\geq 0}$ is an inner product on V. In this scheme, the given factors of V are not necessarily orthogonal, but we may find an orthonormal basis for V, and hence an orthogonal decomposition of V distinct from the given factors. We may find such a basis by applying the Gram–Schmidt procedure. Denote this basis $(b_i)_{i=1}^{q}$. Let L_j be the span of the vector b_j. Then L_j is a one-dimensional subspace of V. The space V has the orthogonal decomposition $L_1 \oplus L_2 \oplus \cdots \oplus L_q$.

Again, the inner product induces a norm, and in turn a metric, both of which we may restrict to each L_j. Denote the norm of factor L_i by $\|\cdot\|_{L_j}$. Then for unit vectors $v_i \in L_i$ and $v_j \in L_j$ such that $\|v_i\|_{L_i} = \|v_j\|_{L_j} = 1$ we have that $\|v_i\|/\|v_j\| = 1$. We might think of these subspaces as being natural.

1.1.2 Metric and pseudometic parameter spaces

From hereon we will be concerned specifically with those parameter spaces that index statistical models (Ex. 4), including distribution-function models of dynamical models (Ex. 6). It will not always be possible to make such spaces metric spaces, but it will always be possible to make them *pseudometric* spaces, as we will now discuss. Recall that, whereas a metric is a function of two variables that is positive-definite, symmetric, and obeys the triangle inequality, a pseudometric is nondegenerate, symmetric, and obeys the triangle inequality (see App. A). Whereas a metric is zero only if its arguments are identical, a pseudometric may be zero for nonidentical arguments.

The key here is that we may we define a metric or pseudometric on the set of probability density functions itself, and that parameter space may then inherit this metric or pseudometric. Whereas the flavourful property of parameter space makes it difficult to impose a metric on parameter space directly, there is no such problem with the set of probability density functions. They are simply nonnegative functions.

The metric or pseudometric on the set of probability density functions quantifies how different any two probability density functions are. We then take the distance between two parameters to be the distance between the probability density functions that they index. The distance between these parameters then quantifies the effect of altering the parameter.

This leaves the question of precisely which metric we should impose on the set of probability density functions. In the case that parameter space is a compact region $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n$ it is possible to endow the set of probability density functions with the structure of a Riemannian manifold. This is described by Amari (1985).[4] In this scheme, the metric tensor is given by the following definition.

Definition 13 (Fisher information metric tensor). The elements of the Fisher information metric tensor are given by

$$g_{ij}(\theta) = \int_X \partial_{\theta_i} \ln(f_X(x;\theta))\partial_{\theta_j} \ln(f_X(x;\theta))f_X(x;\theta)\,\mathrm{d}x \tag{1.13}$$

for $\theta_i, \theta_j \in \theta$.

The distance between any two points on the manifold (i.e. between any two probability density functions) is the length of the geodesic connecting those two points. We will call this the *Fisher information metric*, and denote it d. In general the Fisher information metric will not have closed form, and we will have to compute it numerically, or estimate it. However, in certain cases it is possible to find a closed-form expression. The following is one such case.

Example 14 (Riemannian manifold of Gaussian probability density functions). Consider a Riemannian manifold of univariate Gaussian probability density functions,

$$(f_X(\cdot;\mu,\sigma^2))_{(\mu,\sigma^2)\in\Theta} \tag{1.14}$$

for $\Theta = \mathbf{R} \times \mathbf{R}_{\geq 0}$. Amari (1985) shows that the Fisher information metric is given by

$$d(f_X(\cdot;\mu,\sigma^2), f_X(\cdot;\mu',\sigma^{2\prime})) = \sqrt{2}\,\mathrm{arcosh}\left(\frac{(\mu^2 - \mu'^2) + 2(\sigma^2 + \sigma^{2\prime})}{4\sigma\sigma'}\right). \tag{1.15}$$

The Fisher information metric on the set of probability density functions can be used to define a metric or pseudometric on the parameter space. In order to formalize this we will need the concept of a *pullback*. First, however, let us recall the definition of a composite function.

Definition 15 (composite function). Let $f: X \longrightarrow Y$ and $g : Y \longrightarrow Z$ be functions. The function *g composed with f*, (or *g after f*) is the function

$$g \circ f : X \longrightarrow Z \tag{1.16}$$

$$x \longmapsto g(f(x)). \tag{1.17}$$

[4] For a proof that such a such a set of probability density functions may be made into a Riemannian manifold we refer the reader to Amari directly. We will move right away to consider the metric properties of this manifold.

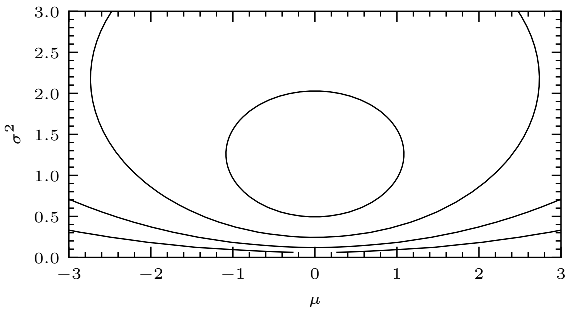

Figure 1.1 The pullback of the Fisher information metric, $d_\Theta(\theta, \theta') = 1, 2, 3$, and 4 for $\theta' = (0, 1)$. Each contour is closed and represents the set of all parameters equidistant from θ'.

We may define a pullback as follows.

Definition 16 (pullback). Let X, Y, and Z be sets and let $f : X \longrightarrow Y$ and $g : Y \longrightarrow Z$ be functions. We may *pullback* g via f to give a function $g' : X \longrightarrow Z$, defined to be the composition $g \circ f$.

Consider the indexed family $f : \Theta \longrightarrow \Xi$, and let $d : \Xi \times \Xi \longrightarrow \mathbf{R}_{\geq 0}$ be a metric or pseudometric on Ξ. Then the pullback of d via the indexing function f is the function

$$d_\Theta : \Theta \times \Theta \longrightarrow \mathbf{R}_{\geq 0} \tag{1.18}$$

$$(\theta, \theta') \longmapsto d(\xi_\theta, \xi_{\theta'}). \tag{1.19}$$

Example 17 (Riemannian manifold of Gaussian probability density functions). Consider the Riemannian manifold of univariate Gaussian probability density functions given in Example 14, which has metric d. We may pullback this metric to form the metric d_Θ given by $d_\Theta(\theta, \theta') = d(f_X(\cdot; \mu, \sigma^2), f_X(\cdot; \mu', \sigma^{2\prime}))$. I plot this in Figure 1.1 for $\theta' = (0, 1)$. Each contour represents the set of points equidistant from θ'.

Note that the pullback of a metric on the set of probability density functions is not necessarily a metric on the set of parameters. Consider the statistical model $(f_X(\cdot; \theta))_{\theta \in \Theta}$ in which two elements are identical, i.e. for which there exist $\theta, \theta' \in \Theta$ such that $\theta \neq \theta'$ and $f_X(\cdot; \theta) = f_X(\cdot; \theta')$. Thus $d_\Theta(\theta, \theta') = d(f_X(\cdot; \theta), f_X(\cdot; \theta')) = 0$, i.e. the two points θ and θ' are separated by zero distance despite being nonidentical. It is then the case that d_Θ is a pseudometric.

If there is a pseudometric on the model then the pullback of that metric is a pseudometric on the parameter space. If there is a metric on the model then the pullback of that metric is a metric on the parameter if and only if the indexing function is injective (i.e. one-to-one).

Proposition 18 (existence of pseudometric parameter spaces). Let $f : \Theta \longrightarrow \Xi$ be an indexed family and let d be a pseudometric on Ξ. Then d_Θ, the pullback of d via the indexing function f, is a pseudometric on Θ.

Proof. It is a simple matter of verification to show that d_Θ satisfies the definition of a pseudometric. □

Proposition 19 (existence of metric parameter spaces). Let $f : \Theta \longrightarrow \Xi$ be an indexed family and let d be a metric on Ξ. Then d_Θ, the pullback of d via the indexing function f, is a pseudometric on Θ if and only if f is injective.

Proof. The function d_Θ will fail to be a metric insofar as it fails to be positive semidefinite. Suppose that the indexing function is injective, and furthermore suppose that $d_\Theta(\theta, \theta') = 0$. Then $d(\xi_\theta, \xi_{\theta'}) = 0$ and hence $\xi_\theta = \xi_{\theta'}$. Thus $\theta = \theta'$ by injectivity. Conversely, suppose that the indexing function is not injective, and furthermore suppose that $\xi_\theta = \xi_{\theta'}$ and $\theta \neq \theta'$. Then $d_\Theta(\theta, \theta') = d(\xi_\theta, \xi_{\theta'}) = 0$. Thus d_Θ is not positive definite. □

We have thus shown how we may rigorously regard the parameter spaces of statistical models as metric or pseudometric spaces. In doing so, we have gone some way in providing an account of the structure of parameter spaces that is missing from the literature. Further development of this account must be left for future work, as we must now return to more practical questions of modelling.

1.2 LIKELIHOOD METHODS

In galactic dynamics, we are interested in continuous random vectors (i.e. tuples of random variables) representing the state of a stellar system, or its observable quantities. Let us denote such a random vector by X, and the model of its joint PDF by $(f_X(\cdot;\theta))_{\theta\in\Theta}$. Its true parameter, θ_0, is unknown to us. Let x be a realization of X. To recover the parameter, we will use the *principal of maximum likelihood* (Fisher, 1922), i.e. we will choose the parameter that, of all possible parameters, assigns that realization the greatest probability density.

Definition 20 (likelihood). Let $X : \mathbf{\Omega} \longrightarrow \mathbf{R}$ be a random vector with joint PDF, f_X, known to be an element of the model $(f_X(\cdot;\theta))_{\theta\in\Theta}$. Let $\mathbf{R}^{\mathbf{\Omega}}$ be the set of all real-valued random variables on the probability space $\mathbf{\Omega}$. The *likelihood* is the function

$$L : \Theta \longrightarrow \mathbf{R}^{\mathbf{\Omega}} \tag{1.20}$$

$$\theta \longmapsto f_X(X;\theta). \tag{1.21}$$

We call the value $L(\theta)$ the *likelihood of the parameter* θ. For convenience we will use the notation $L(\theta)$ and L_θ interchangeably.

Note that it is $L(\theta)$ that is the random variable, not L. If $X = x$ (i.e. if the random variable X takes the realized value x) then $L(\theta) = f_X(X = x;\theta)$ (i.e. the random variable $L(\theta)$ takes the realized value $f_X(X = x;\theta)$), and if $X = x'$ then $L(\theta) = f_X(X = x';\theta)$. Similarly , if $X = x$ then $L(\theta) = f_X(X = x;\theta)$ and $L(\theta') = f_X(X = x;\theta')$.[5] We will be interested in the relative likelihood of two parameters and hence define the *likelihood ratio*.

Definition 21 (likelihood ratio). The *likelihood ratio* is the function

$$\Lambda : \Theta \times \Theta \longrightarrow \mathbf{R} \tag{1.22}$$

$$(\eta,\theta) \longmapsto L(\eta)/L(\theta). \tag{1.23}$$

We will say that η is *more likely* than θ if $\Lambda(\eta,\theta) > 1$. Note that the likelihood ratio is invariant under arbitrary scalings of the likelihood, $L(\theta) \longmapsto aL(\theta)$ for some real a. Therefore, we are not interested in absolute values of L. It will often be convenient to work with the natural logarithm of the likelihood, which we call the 'support'.

Definition 22 (support). Let $\bar{\mathbf{R}} = \mathbf{R} \cup \{-\infty\}$, and let $\bar{\mathbf{R}}^{\mathbf{\Omega}}$ be the set of all $\mathbf{R}$-valued functions on the probability space $\mathbf{\Omega}$. The *support*[6] is the function

$$S : \Theta \longrightarrow \bar{\mathbf{R}}^{\mathbf{\Omega}} \tag{1.24}$$

$$\theta \longmapsto \ln(L(\theta)) \tag{1.25}$$

where we understand that $\ln(0) = -\infty$. We call the value $S(\theta)$ the *support for the parameter* θ. For convenience we will use the notation $S(\theta)$ and S_θ interchangeably.

[5] Once we have defined random processes, in Chapter 2, we will be able to see the likelihood as the random process $L = (L_\theta)_{\theta\in\Theta}$.

[6] Some authors (for example, Wasserman, 2004) use the term 'log-likelihood of θ' instead of 'support for θ', which they denote $l(\theta)$. We will follow Edwards (1972) in using the latter term. This has the happy consequence that we may respect our convention that random variables are denoted by a capital letter, in this case S_θ.

Note that, just as $L(\theta)$ is the random variable not L, so is $S(\theta)$ the random variable, not S.

Definition 23 (maximum-likelihood estimator). A *maximum-likelihood estimator* (MLE) of the parameter is any element

$$\hat{\Theta} \in \operatorname*{argmax}_{\theta \in \Theta}(L(\theta)). \tag{1.26}$$

Note that the MLE need neither exist not be unique. There may be no maximum of the set $\{L(\theta)\}_{\theta\in\Theta}$ or there may be multiple maxima. Because the logarithm is a monotonic function, the maximum of the support function coincides with the maximum of the likelihood function, i.e. $\operatorname{argmax}_{\theta\in\Theta}(L(\theta)) = \operatorname{argmax}_{\theta\in\Theta}(S(\theta))$.

1.2.1 Asymptotic behaviour of the MLE

The maximum-likelihood estimator is a random variable. Its value is dependent on the value of X. What happens as as we observe an increasingly large number of realizations of X? Let $(X_i)_{i=1}^n$ be indendent and identically distributed random vectors, each with joint PDF $f_X(\cdot, \theta_0)$ known to be an element of the model $((f_X(\cdot, \theta))_{\theta\in\Theta}$. The joint PDF of these random vectors is

$$f_{(X_1,X_2,\ldots,X_n)}(\cdot;\theta) = \prod_{i=1}^{n} f_X(\cdot,\theta). \tag{1.27}$$

We denote the likelihood of parameter θ by $L_n(\theta)$, its support by $S_n(\theta)$, and the MLE by $\hat{\Theta}_n$ where the subscript n emphasizes the dependence of these quantities on the number of observations. What happens as $n \longrightarrow \infty$? In our case X is a random vector representing a single stellar observation, and $(X_i)_{i=1}^n$ is a set of observations for a stellar system. What, then, happens as we observe increasingly may many stars?

Under certain regularity conditions we find that the MLE is *consistent* (i.e. it converges on the true parameter) and *asymptotically normal* (i.e. its distribution converges on a normal distribution with some given variance matrix). These regularity conditions amount to requiring that the indexing function is injective, that the parameter space is compact, and that the PDF is sufficiently smooth (for a discussion, see Wasserman, 2004, p. 126). If a model isn't regular we may consider a regular subset of that model, in which case we will say that a model is *locally regular*.

Recall that we say a sequence of random variables $X_1, X_2, \ldots, X_n$ *converges in probability* towards the random variable X (which may be a constant, i.e. a trivial random variable) if for all positive ε it is the case that $\lim_{n\longrightarrow\infty} P(|X_n - X| > \varepsilon) = 0$, and that we write $X_n \xrightarrow{p} X$ to denote this.

Proposition 24 (consistency). The maximum-likelihood estimate of a parameter is consistent, i.e. $\hat{\Theta}_n \xrightarrow{p} \theta_0$.

Proof. A proof is given by Wasserman (2004). □

To characterize the asymptotic normality of the MLE we will require the Fisher information.

Definition 25 (Fisher information). The *Fisher information* is the matrix-valued function

$$I_n : \Theta \longrightarrow \mathbf{R}^{n \times n} \tag{1.28}$$

$$\theta \longmapsto (\mathrm{E}_\theta(\partial_{\theta_i} S_n(\theta) \partial_{\theta_j} S_n(\theta)))_{ij}. \tag{1.29}$$

Note that $I_n = nI_1$, and that $I_1 = \mathrm{E}_\theta(\partial_{\theta_i} S_1(\theta)\partial_{\theta_j} S_1(\theta)) = -\mathrm{E}_\theta(\partial_{\theta_i \theta_j} S_1(\theta))$, i.e. the matrix I_1 is the negative of the expectation of the Hessian of the support. Note that it is positive semidefinite. Note also that $\mathrm{cov}(\partial_{\theta_i} S(\theta), \partial_{\theta_j} S(\theta)) = I_{ij}$.

Remark 26. We have already encountered the Fisher information, I_1, in our discussion of the structure of parameter space (Def. 13).

Recall that we say a sequence of random variables $X_1, X_2, \ldots, X_n$, with cumulative distribution functions $F_{X_1}, F_{X_2}, \ldots, F_{X_n}$ *converges in distribution* towards a random variable X, with cumulative distibution function F_X if $\lim_{n \longrightarrow \infty} F_{X_n} = F_X$, and that we write $X_n \xrightarrow{d} F_X$ to denote this.

Proposition 27 (asymptotic normality). The maximum-likelihood estimate converges in distribution,

$$n^{1/2} I_1^{1/2} (\hat{\Theta}_n - \theta_0) \xrightarrow{d} N(0, I)) \tag{1.30}$$

where I is the identity matrix, and $I_1^{1/2}$ is the square root of the Fisher information matrix I_1 (this is well defined as the Fisher information matrix is positive semidefinite).

Proof. A proof is given by Wasserman (2004). □

Remark 28. Note that in the limit, $\mathrm{cov}(\hat{\Theta}_{n,i}, \hat{\Theta}_{n,j}) \longrightarrow I^{-1}_{n,ij}$, where $I^{-1}_{n,ij}$ is the ij-th element of the inverse of the Fisher information, I_n.

A more suggestive notation for convergence in distribution is

$$\hat{\Theta}_n \approx N(\theta_0, I_n^{-1}(\theta_0)). \tag{1.31}$$

We may think of $\hat{\Theta}_n$ as having a distribution that is approximately normal with mean θ_0 and variance $I_n^{-1}(\theta_0)$. In fact, it may be shown (Wasserman, 2004, p. 129) that

$$\hat{\Theta}_n \approx N(\theta_0, I_n^{-1}(\hat{\Theta}_n)), \tag{1.32}$$

i.e. instead of evaluating the Fisher information at θ_0, we may evaluate it at the realization of $\hat{\Theta}_n$. This allows us to determine the confidence region, C_n, defined by the boundary that is the solution to the equation

$$(\theta - \hat{\Theta}_n)^{\mathrm{t}} I_n^{-1}(\theta - \hat{\Theta}_n) = \chi_q^2(1-\alpha) \tag{1.33}$$

where χ_q^2 is the quantile function (i.e. the inverse of the cumulative distribution function) for the chi-squared distribution for q degrees of freedom and α is the critical value. The probability $P(\theta \in C_n) \longrightarrow \alpha$ as $n \longrightarrow \infty$. We say that C_n *traps* θ_0 *with probability* α.

We have noted that the Fisher information I_1 is the negative of the expectation of the Hessian of the support. The Fisher information, $I_1(\theta)$ therefore quantifies the curvature of the support, $S_1(\theta)$. Let us consider a parameter containing a single element, meaning that the support is a function of one variable. Evaluated at θ_0, it quantifies the breadth of the peak in the support. A narrow peak (i.e. large curvature and large Fisher information) indicates that the maximum is well constrained. A broad peak (i.e. small curvature and small Fisher information) indicates that the maximum is poorly constrained.

Proposition 29 (equivariance). Let $g : \Theta \longrightarrow \mathbf{R}$ be a function, and let $\hat{\Theta}_n$ be the MLE of of θ_0. Then $g(\hat{\Theta}_n)$ is the MLE of $g(\theta)$.

Proof. A proof is given by Wasserman (2004). □

By the asymptotic normality of maximum-likelihood estimates (Prop. 27), the distribution of $g(\hat{\Theta}_n)$ must itself be asymptotically normal. Its distribution may be computed explicitly using the following proposition.

Proposition 30 (delta method). Let $\nabla = (\partial_{\theta_i})_i$, and let $g : \Theta \longrightarrow \mathbf{R}$ be a differentiable function such that $\nabla g \neq 0$. Then

$$g(\hat{\Theta}_n) \xrightarrow{d} N(g(\theta_0), \nabla^{\mathrm{t}} g(\theta_0) I_n^{-1} \nabla g(\theta_0)). \tag{1.34}$$

Proof. A proof is given by Wasserman (2004). □

This allows us to determine the confidence region

$$C_n = (g(\hat{\Theta}_n) - \chi_1^2(1-\alpha)\sqrt{\nabla^t g I_n^{-1} \nabla g}, g(\hat{\Theta}_n) + \chi_1^2(1-\alpha)\sqrt{\nabla^t g I_n^{-1} \nabla g}), \tag{1.35}$$

which traps θ_0 with probability α.

It is not always possible to compute the Fisher information. In particualar, we may be unable to compute the expectated value.

Definition 31 (observed Fisher information). The *observed Fisher information* is the matrix-valued function

$$J_n : \Theta \longrightarrow \mathbf{R}^{n\times n} \tag{1.36}$$

$$\theta \longmapsto (-\partial_{\theta_i \theta_j} S_n(\theta))_{ij}. \tag{1.37}$$

Consider an element of the matrix,

$$-\partial_{\theta_i\theta_j} S_n(\theta) = -\partial_{\theta_i\theta_j} \ln\left(\prod_{k=1}^{n} f(X_i;\theta\right) \tag{1.38}$$

$$= -\sum_{k=1} \partial_{\theta_i\theta_j} \ln(f(X_i;\theta)). \tag{1.39}$$

The summands of this expression are independent and identically distributed. By the law of large numbers, therefore, the element's average converges on the expected value of any single term, i.e.

$$\frac{1}{n} J_{n,ij}(\theta) \xrightarrow{p} I_{ij}(\theta). \tag{1.40}$$

More suggestively we may write $J_n \approx nI(\theta) = I_n(\theta)$. In fact, it may be shown that $J_n \approx I_n(\hat{\Theta}_n)$. Hence we may rewrite expressions 1.32 and 1.33 with J_n substituted for I_n. We may then make the same substitution in our statement of the delta method (Prop. 30).

1.3 METAMODELLING

Distribution function models of dwarf spheroidal galaxies are typically expensive to evaluate, and it is impractical to maximize the likelihood of the parameters. We are therefore interested in constructing metamodels of them. In our case, the model is a family of probability density functions for the observable quantities, namely the sky positions and line-of-sight velocities of stars observed in a dSph (a full discussion will come in Ch. 4). We create a metamodel by evaluating the model for some small number of distinct parameters, and then using the results of these evaluations to *predict* the value of the model for some arbitrary parameter. We do this without needing to make an additional evaluation of the model. In just the same way that a *computer simulation* is a means of evaluating a model, an *emulator* is a means of evaluating a metamodel.[7] And in just the same way that a simulation *run* is a is an evaluation of a model for a single parameter, an emulator run is an evaluation of a metamodel for a single parameter.

Although the value of an element of the model, $f(x;\theta)$, is deterministic, we treat it as if it were random, i.e. we treat it as if it were the realization of a random variable. The metamodel is a family of such random variables, one representing the value of the model for each parameter. Metamodels allow us to compute a best guess for the value of the given element of the model, together with a measure of confidence in that best guess.

We will use GPE to construct our metamodels. To give some idea of the potential of this method, let us look at the results of the emulation of the Forrester function (Forrester et al., 2008). The Forrester function is the function $f\colon [0,1] \longrightarrow \mathbf{R}$ given by

$$f(x) = (6x-2)^2 \sin(12x-4). \tag{1.41}$$

It has no particular physical significance, but is rather used as a test function for optimization methods, because it has multiple minima and an inflexion point. The example is artificial, as we would never have any need to emulate this function. We may always evaluate the Forrester function directly. Suppose, however, that it were expensive to evaluate and that we wished to make as few evaluations as possible. (In this sense $f(x)$ stands in for the likelihood of a parameter $L(\theta)$.) With GPE we may evaluate the function for some small number of distinct arguments (in this case 10), and use these values to predict the values of the function for its entire domain. GPE does this with no knowledge of the formula for the Forrester function. The results are shown in Figure 1.2. The predicted values are very close to the true values everywhere, and the true values always within the five-sigma confidence intervals of the predicted values. These confidence intervals are themselves small.

Having hinted at the potential of GPE, let us now turn our attention to its theoretical foundations.

[7] Presumably the term 'emulation' originates with the fact that one computer programme is being used to *emulate* another. This is somewhat analagous to the use of software (*emulators*) to allow one operating system to run programmes written for another.

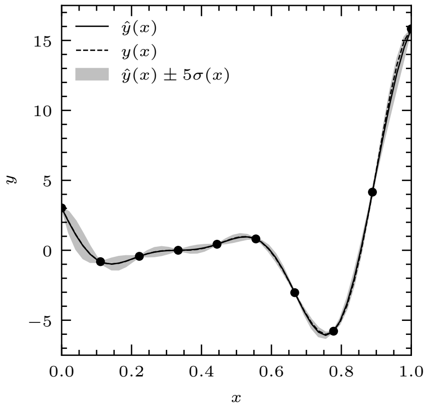

Figure 1.2 The Forrester function (dashed line) and its predicted values (solid line) computed using GPE based on the 10 samples shown (filled circles). The five-sigma confidence interval for these predictions is also shown (grey band).

Chapter 2

The mathematical structure of Gaussian-process emulation

Consider the following problem of inference. Given a family of random variables, $X_1, X_2, \ldots, X_n$, what can we say about some other random variable, Z? If Z is independent of this family we can of course say nothing. But if it is dependent on it we might want to do one of two things. We might want to provide some best guess for its realized value. Or we might want to provide a *prediction* for that random variable based on this family, namely some other random variable that appoximates it. Now consider an extension to this problem. Given the same family of random random variables, $X_1, X_2, \ldots, X_n$, what can we say about some other *family* of random variables, $Z_1, Z_2, \ldots, Z_m$. Again, we might want to do one of two things. We might want to provide some best guess for the realized values of its elements. Or we might want to provide a *prediction* for that second family based on the first family, namely some other family of random variables that approximates it. Now consider a further extension to the problem. Rather than families of random variables indexed by the sets $\{1, 2, \ldots, n\}$ and $\{1, 2, \ldots, m\}$, consider instead families of random variables indexed by arbitrary index sets, T and S, namely $\{X_t\}_{t \in T}$ and $\{Z_s\}_{s \in S}$. Such families of random variables are known a *random processes*. Given a random process, $\mathbf{X} = \{X_t\}_{t \in T}$, what can we say about some other random process, $\mathbf{Z} = \{Z_s\}_{s \in S}$? We may want to provide best guesses for the realized values of $\mathbf{Z}$ based on $\mathbf{X}$. Or we might want to provide a prediction for $\mathbf{Z}$ based on $\mathbf{X}$. Each process, $\mathbf{X}$ and $\mathbf{Z}$, may have one element, finitely many elements, or infinitely many elements.[1] The general case of predicting every element of $\mathbf{Z}$ seems not to have a name. Let

[1] In the case of that $\mathbf{X}$ is a finite subset of $\mathbf{Z}$, some authors also use the terms 'forecast', 'extrapolation' or 'interpolation' instead of 'prediction'(for example Yaglom, 1961; Gikhman and Skorohod, 1974).

us call it *replication*.[2] We may think of a replication as being a family of predictions of random variables. Or, we may think of a prediction as being a single value of a replication of a random process. In talking about prediction and replication we are really talking about the same thing.

In *linear prediction* (Parzen, 1959) we seek a predictor for Z that is a linear combination of the elements of $\mathbf{X}$. Let us denote this prediction by $\hat{Z}$. Then we have that

$$\hat{Z} = \sum_{t \in T} a_t X_t. \tag{2.1}$$

We choose the elements $(a_t)_{t\in T}$ so as minimize the *mean-squared error*, $\mathrm{E}((Z - \hat{Z})^2)$. We call this prediction the *best linear predictor* (Parzen, 1959). We may further impose the constraint that the expected values of Z and $\hat{Z}$ are equal, i.e. that $\mathrm{E}(\hat{Z}) = \mathrm{E}(Z)$. We call this the *best unbiased linear predictor* (Parzen, 1959). In the general case, these linear combinations will not be finite sums, and will be interpreted as convergent series in a Hilbert space of random variables. Hilbert spaces are already familiar to physicists as the proper setting for the formulation of quantum mechanics (Böhm, 1978). In this way we may view prediction geometrically. In fact, we will see that we may find the best linear predictor and the best unbiased linear predictor by projecting the random variable Z onto the appropriate subspace.

In particular we use *reproducing kernel Hilbert spaces*, which may be used to represent arbirtrary Hilbert spaces, and in which our calculations become tractable. A reproducing kernel Hilbert space is a Hilbert space of functions that has the following property: if two functions are close in norm then they are also close point-wise. Or, given two functions $f, g : \Xi \longrightarrow \mathbf{R}$, if $\|f - g\|$ is small then so is $|f(\xi) - g(\xi)|$. (We will rigorously define the words 'close' and 'small' in due course.)[3] The general theory of reproducing kernel Hilbert spaces was developed by Aronszajn (1950). It was first applied to the problem of prediction by Parzen (1959), the principal result of his work being an expression for the *best unbiased linear predictor* for Z, which I will call the 'Parzen prediction theorem' (Thm 107). A special case of this theorem allows us to compute explicit predictions for Z when $\mathbf{X}$ is finite. This special case may be used to create metamodels and emulators. The first such use was by Sacks et al. (1989). We will here summarize this work, first establishing the theory of random processes, then the theory of reproducing kernel Hilbert spaces, then the theory of prediction, and finally the theory of prediction in the setting of reproducing kernel Hilbert spaces.

[2] I have Sylvy Anscombe to thank for coining this term.

[3] It is worth pointing out even at this early stage that L^2, the Hilbert space of square-integrable functions, cannot be endowed with the structure of a reproducing kernel Hilbert space.

2.1 RANDOM PROCESSES

We will fundamentally be interested in random processes, namely indexed families of random variables. (A summary of the theory of random variables is given in App. A.)

Definition 32 (random process). A *random process* (also *stochastic process*, *random function*, or just *process*) is an indexed family of random variables on a probablity space $\mathbf{\Omega}$, which we denote $\mathbf{X} = \{X_t : \mathbf{\Omega} \longrightarrow \mathbf{R}\}_{t \in T}$. We call T the *index set*, and any element $t \in T$ an *index*.

Remark 33. The index set is arbitrary. I will generally be interested in the case that $T \subset \mathbf{R}^n$. Some authors reserve the term 'random process' for the case that T is one-dimensional, i.e. a subset of $\mathbf{R}$, and use the term 'random field' in the case that T is multidimensional, i.e. a subset of $\mathbf{R}^n$ for some $n > 1$. In this case, the term 'time series' is also used for a (one-dimensional) random process. Furthermore, if T is discrete (respectively continuous) then the random process is said to be a *discrete-time random process* (respectively *continuous-time random process*). I will not make this distinction, but will use the term 'random process' regardless of the dimension of T.

If the index set of a random process is finite we will call that random process a *random vector*. In particular, a finite subset of a random process is a random vector. We denote the cumulative distribution function (CDF) of a random vector $\mathbf{X}$ by $F_{\mathbf{X}}$, and its probability density function (PDF) by $f_{\mathbf{X}}$. Kolmogorov's extension theorem (App. A) tells us that any random process is determined by the set of all finite-dimensional joint distributions of its elements.

Definition 34 (realization of a random process). Given a fixed element of the probability space, $\omega \in \mathbf{\Omega}$, the *realization of* $\mathbf{X}$ *at* ω (also, *sample function of* $\mathbf{X}$ *at* ω, or *sample path of* $\mathbf{X}$ *at* ω) is the function

$$f \colon T \longrightarrow \mathbf{R} \tag{2.2}$$

$$t \longmapsto X_t(\omega). \tag{2.3}$$

(See Adler, 1981, p. 14.)

A realization of a random process is a function, $T \longmapsto \mathbf{R}$, in the same way that a realization of a random variable is a real number. We may think of ourselves as drawing a function from a random process in the same way that we may think of ourselves as drawing a real number from a random variable.

Definition 35 (order of a random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. Then $\mathbf{X}$ is of *order* n (or n*-th order*) if the n-th moment of X_t exists for all $t \in T$.

Note that if a random process is of order n then it is also of order $n-1$.

Definition 36 (mean, covariance, and correlation function). Let $\mathbf{X} = \{X_t\}_{t\in T}$ be a random process. If $\mathbf{X}$ is first-order then the *mean function for* $\mathbf{X}$ is the function

$$m : T \longrightarrow \mathbf{R} \tag{2.4}$$

$$t \longmapsto \mathrm{E}(X_t) \tag{2.5}$$

where $\mathrm{E}(X_t)$ is the expected value of X_t. If $\mathbf{X}$ is second-order then the *covariance function for* $\mathbf{X}$ (also, *autocovariance function for* $\mathbf{X}$) is the function

$$k : T \times T \longrightarrow \mathbf{R} \tag{2.6}$$

$$(s,t) \longmapsto \mathrm{cov}(X_s, X_t), \tag{2.7}$$

where $\mathrm{cov}(X_s, X_t)$ is the covariance of random variables X_s and X_t, and the *correlation function* is

$$r : T \times T \longrightarrow \mathbf{R} \tag{2.8}$$

$$(s,t) \longmapsto \mathrm{corr}(X_s, X_t) \tag{2.9}$$

where $\mathrm{corr}(X_s, X_t)$ is the correlation of random variables X_s and X_t. (Definitions of expected value, covariance and correlation are given in App. A.)

If $\mathrm{E}(X_t) = 0$ for all $t \in T$ we say that a random process is *centred.* Note that the variance of an element $X_t \in \mathbf{X}$ is $k(t,t)$.[4] Note also that if a random process has covariance function k then it has correlation function given by

$$r(s,t) = \frac{k(s,t)}{\sqrt{k(s,s)k(t,t)}} \tag{2.10}$$

(since $r(s,t) = \mathrm{corr}(X_s, X_t) = \mathrm{var}(X_s, X_t)/\sqrt{\mathrm{var}(X_s)\mathrm{var}(X_t)} = k(s,t)/\sqrt{k(s,s)k(t,t)}$).

Definition 37 (positive-semidefinite kernel). Let Ξ be an arbitrary set. A symmetric function $k : \Xi \times \Xi \longrightarrow \mathbf{R}$ is a *positive semi-definite kernel* (respectively, a *positive definite kernel*) if the matrix

$$(k(\xi_i, \xi_j))_{ij} \tag{2.11}$$

is positive semi definite (respectively, positive definite), for all distinct $\xi_1, \xi_2, \dots, \xi_n \in \Xi$.

[4] Some authors refer to the variance of a random process as the *average power of* $\mathbf{X}$ *at* t.

Equivalently, a symmetric function k is a positive-semidefinite kernel (respectively, a positive-definite kernel) if

$$\sum_{i,j} k(\xi_i, \xi_j) u_i u_j \tag{2.12}$$

is nonnegative (respectively, positive) for all distinct $\xi_1, \xi_2, \ldots, \xi_n \in \Xi$ and all $u_1, u_2, \ldots, u_n \in \mathbf{R}$. This is simply the definition of a positive-semidefinite matrix (respectively, positive-definite matrix) applied to the matrix $(k(\xi_i, \xi_j))_{ij}$.

Example 38 (dot product). The dot product,

$$f\colon \mathbf{R}^n \times \mathbf{R}^n \longrightarrow \mathbf{R} \tag{2.13}$$

$$(v, \xi) \longmapsto v^{\mathrm{t}}\xi \tag{2.14}$$

is a positive-definite kernel. First, note that f is symmetric, and consider the set of elements $(x_i)_{i=1}^n$ where $x_i \in \mathbf{R}^n$. By definition, the function f is positive definite if the matrix $K = (x_i^{\mathrm{t}} x_j)_{ij}$ is positive definite, i.e. if condition 2.12 holds. Recall that a matrix is positve definite if its eigenvalues are positive. As the eigenvalues of K are indeed positive, so K is positive definite.

Remark 39. The word 'kernel' has at least three meanings in the mathematical literature. There is the *kernel* in the algebraic sense, namely the preimage of zero under a homomorphism between algebraic structures, such as groups, rings, vector spaces, etc. (Bourbaki, 1989, Ch. I Sec. 4, No 5 Def. 8, and Ch. II Sec. 1, No 3). There is the *integral kernel*, namely the two-variable function used in an integral transform (Courant and Hilbert, 1989, Ch. III). And there is is the *positive-semidefinite kernel*, as defined above. I will only use the last of these three definitions. Nonetheless, for the sake of clarity, I will always use the term 'positive-semidefinite kernel', and never the term 'kernel'. Positive-semidefinite kernels and integral kernels are related concepts , but positive-semidefinite kernels and algebraic kernels are unrelated. Aronszajn (1950) discusses the relationship between positive-semidefinite kernels and integral kernels.)

Theorem 40 (characterization of covariance functions). *A function, k, is the covariance function of a random process if and only if it is a positive-semidefinite kernel.*

Proof. A proof is given by Parzen (1959, p. 15). □

This theorem justifies my use of the symbol k for both covariance functions and positive-semidefinite functions. The non-trivial direction of this theorem (the *only if* part of the statement) is the *existence* of a random process that has a given positive-semidefinite kernel as its covariance function.

2.1.1 Stationary and isotropic random processes

We now wish to define *stationary* and *isotropic* random processes. These are random processes for which the mean and covariance are invariant under translations and rotations of the index set, T. Of course it must be possible to define such translations and rotations. This is not possible on all index sets. The most general case is that of a group $\mathbf{G}$ acting on the index set T (see App. A). For example, T might be real three-dimensional space and $\mathbf{G}$ the group of rotations. I therefore introduce a more general definition of *invariant random processes*. Stationary and isotropic random processes then become special cases of this definition.

Definition 41 (invariant and weakly invariant random processes). Let $\mathbf{X} = \{X_t\}_{t\in T}$ be a random process, and suppose that $\mathbf{G} = (G, \cdot)$ is a group acting on T (on the left, say). We say that $\mathbf{X}$ is $\mathbf{G}$-*invariant* if

$$F_{(X_{t_1},\dots,X_{t_n})} = F_{(X_{gt_1},\cdots,X_{gt_n})}, \tag{2.15}$$

for all $g \in G$, $t_1, \dots, t_n \in T$. We say that $\mathbf{X}$ is *weakly* $\mathbf{G}$-*invariant* if

$$\mathrm{E}(X_t) = \mathrm{E}(X_{gt}),$$

for all $g \in G$, $t \in T$ and

$$\mathrm{cov}(X_{t_1}, X_{t_2}) = \mathrm{cov}(X_{gt_1}, X_{gt_2}),$$

for all $g \in G$, $t_1, t_2 \in T$.

Remark 42. We might think of the group $(\mathbf{R}, +)$ acting on $\mathbf{R}$ (translations), or of the group $SO(n)$ acting on $\mathbf{R}^n$ (rotations).

In other words, a process is $\mathbf{G}$-invariant if the joint cumulative distribution functions of all finite subsets of $\mathbf{X}$ is invariant under the action of $\mathbf{G}$ on T. A process is weakly stationary if the mean and variance of the random process both exist and are invariant under under the action of $\mathbf{G}$ on T. A $\mathbf{G}$-invariant process fails to be weakly $\mathbf{G}$-invariant if either the mean or covariance do not exist. (Note that the definition of $\mathbf{G}$-invariance does not require the mean or covariance of the random process process to exist.) A weakly $\mathbf{G}$-invariant process fails to be $\mathbf{G}$-invariant if the cumulative distribution functions are not invariant under the action of $\mathbf{G}$ on T despite the mean and covariance being invariant under the action of $\mathbf{G}$ on T.

Given this new definition of invariance, we may now view the normal definition of stationarity as a special case. Specifically , we may now define stationary random processes as follows.

Definition 43 (stationary and weakly stationary random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. Suppose that T itself admits an Abelian group operation, $+$, and consider the regular action of $(T, +)$ on T (i.e. $s \in T$ acts by $t \longmapsto s + t$). We say that $\mathbf{X}$ is *stationary*[5] if $\mathbf{X}$ is $(T, +)$-invariant. Similarly, we say that $\mathbf{X}$ is *weakly stationary*[6] if $\mathbf{X}$ is weakly $(T, +)$-invariant.

Remark 44. If $\mathbf{X}$ is a stationary random process, its distributions are invariant under translations of T. For all sequences $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and for all $u \in T$ it is the case that

$$F_{(X_{t_1}, X_{t_2}, \ldots, X_{t_n})} = F_{(X_{u+t_1}, X_{u+t_2}, \ldots, X_{u+t_n})}. \tag{2.16}$$

If $\mathbf{X}$ is a weakly stationary random process then for all $s, t, u \in T$ it is the case that $\mathrm{E}(X_s) = \mathrm{E}(X_{u+s})$ and $\mathrm{cov}(X_s, X_{u+s}) = \mathrm{cov}(X_t, X_{u+t})$.

If a random process is weakly stationary then its mean and variance are constant. Denote the variance by σ^2. Then $k(t, t) = \sigma^2$ for all $t \in T$. Furthermore, it is the case that

$$k(s, t) = \sigma^2 r(s, t) \tag{2.17}$$

for all $s, t \in T$. If a random process is neither stationary nor weakly stationary, it is *nonstationary* (also, *evolutionary*).

The covariance of two elements of a stationary random process depends only on the separation of their indices in parameter space. Hence, the covariance function may be expressed as a function of this separation. We will see examples of such functions later (Exs 58–55).

Definition 45 (metacovariance and metacorrelation function). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a weakly stationary random process with covariance function k, and correlation function r. The *metacovariance function*[7] *for* $\mathbf{X}$ is the function

$$\kappa : T \longrightarrow \mathbf{R} \tag{2.18}$$

$$t \longmapsto k(s, s + t), \tag{2.19}$$

for some (equivalently, all) $s \in T$. The *metacorrelation function for* $\mathbf{X}$ is the function

$$\rho : T \longrightarrow \mathbf{R} \tag{2.20}$$

$$t \longmapsto r(s, s + t), \tag{2.21}$$

[5] Some authors use the terms 'stongly stationary', 'strictly stationary', and 'homogeneous' as synonyms for 'stationary'.

[6] Some authors use the terms 'second-order stationary', 'stationary in the wide-sense', 'covariance stationary', or 'weakly homogeneous' as synonyms for 'weakly stationary'.

[7] Some authors use the term 'covariance function', despite this term already being used for k. I have coined the term 'metacovariance function' to avoid this ambiguity.

for some (equivalently, all) $s \in T$. These are well defined since $\mathbf{X}$ is weakly stationary.

Definition 46 (positive definite function). Let G be a group. A function $\kappa : G \longrightarrow \mathbf{R}$ is *positive semi-definite* (respectively, *positive definite*) if the function

$$f : G \times G \longrightarrow \mathbf{R} \tag{2.22}$$

$$(s, t) \longmapsto \kappa(s - t) \tag{2.23}$$

is a positive semi-definite kernel (respectively, positive definite kernel).

Proposition 47 (existence of metacovariance functions). A function, κ, is the metacovariance of a weakly stationary random process if and only if it is a positive-semidefinite function.

Proof. To show the *if* part of the claim, observe that if $\kappa : G \longrightarrow \mathbf{R}$ is a metacovariance function then by definition (of the metacovariance function) there exists a covariance function $k : G \times G \longrightarrow \mathbf{R}$, which must be a positive-semidefinite kernel (by Thm 40). Hence κ is a positive-semidefinite function. To show the *only if* part of the claim, observe that if $\kappa : G \longrightarrow \mathbf{R}$ is a positive-semidefinite function then by definition (of the positive-semidefinite function) there exists a function $f : G \times G \longrightarrow \mathbf{R}$ that is a positive-semidefinite kernel and hence a covariance function (again, by Thm 40). Hence κ is a metacovariance function. □

This proposition justifies my use of the symbol κ for both metacovariance functions and positive-definite functions. Just as stationarity is a special case of invariance, so is isotropy. In fact, we may define isotropic random processes as follows.

Definition 48 (isotropic and weakly isotropic random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. Suppose that T is the set $\mathbf{R}^n$, and that G is the special orthogonal group $\mathrm{SO}(n)$, which acts on $T = \mathbf{R}^n$ in the usual way. Then $\mathbf{X}$ is *isotropic*[8] if $\mathbf{X}$ is G-invariant. Accordingly, we say that $\mathbf{X}$ is *weakly isotropic*[9] if $\mathbf{X}$ is weakly G-invariant.

Remark 49. If $\mathbf{X}$ is an isotropic random process, its distributions are invariant under rotations of T. For all sequences $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ and for all rotations g it is the case that

$$F_{(x_{t_1}, x_{t_2}, \ldots, x_{t_n})} = F_{(x_{gt_1}, x_{gt_2}, \ldots, x_{gt_n})}. \tag{2.24}$$

If $\mathbf{X}$ is a weakly isotropic random process then for all $s, t, u \in T$ and all $g \in G$ it is the case that $\mathrm{E}(X_t) = \mathrm{E}(X_{gt})$ and $\mathrm{cov}(X_s, X_t) = \mathrm{cov}(X_{gs}, X_{gt})$. (See Ivanov and Leonenko, 1989, p. 11.)

[8] Some authors use the terms 'strictly isotropic' and 'strongly isotropic'.

[9] Some authors use the terms 'isotropic in the wide-sense'.

If a random process, $\mathbf{X}$, is both weakly isotropic and weakly stationary then it is the case that the metacovariance $\kappa(r) = \kappa(\|r\|)$, and that the metacorrelation $\rho(r) = \rho(\|r\|)$.

2.1.1.1 Increments of a random process

Two final definitions concern the *increment* of a random process.

Definition 50 (increment of random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. An *increment of* $\mathbf{X}$ is the difference of any two of its elements, $X_s - X_t$.

Definition 51 (intrinsically stationary random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. Suppose again that T admits an Abelian group operation, +. Then $\mathbf{X}$ is *instrinsically stationary* if

$$\mathrm{E}(X_s) = \mathrm{E}(X_t) \tag{2.25}$$

and

$$\mathrm{var}(X_s - X_{u+s}) = \mathrm{var}(X_t - X_{u+t}), \tag{2.26}$$

for all $s, t, u \in T$.

The class of random processes with stationary increments includes the class of weakly-stationary random processes, as we see in the following theorem.

Proposition 52. If a random process is weakly stationary then it has stationary increments.

Proof. Let $\boldsymbol{X}_{t \in T}$ be a weakly-stationary random process. Then for all $u, s, t \in T$ it is the case that $\mathrm{E}(X_s) = \mathrm{E}(X_t)$, that $\mathrm{var}(X_s) = \mathrm{var}(X_t)$, and that$\mathrm{cov}(X_s, X_{u+s}) = \mathrm{cov}(X_t, X_{u+t}) =$. Hence

$$\mathrm{var}(X_s - X_{u+s}) = \mathrm{var}(X_s) + \mathrm{var}(X_{u+s}) + 2\mathrm{cov}(X_s, X_{u+s}) \tag{2.27}$$
$$= \mathrm{var}(X_t) + \mathrm{var}(X_{u+t}) + 2\mathrm{cov}(X_t, X_{u+t}) \tag{2.28}$$
$$= \mathrm{var}(X_t - X_{u+t}). \tag{2.29}$$

Hence, $\boldsymbol{X}_{t \in T}$ has stationary increments. □

2.1.1.2 Examples of random processes

Random processes are used to model numerous physical phenomena. Perhaps the two most important random processes in physics are the Wiener process and the Poisson process. The Wiener process is used to model Brownian motion (Wiener, 1923). The Poisson process was first used to model α-particle detections (Rutherford et al., 1910), and calls received at a telephone exchange (Erlang, 1909).[10] It may be used to model numerous other phenomena, including particle collisions in an ideal gas. The most important of all random processes, however, is the Gaussian random process.

Definition 53 (Gaussian random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. Let $T \subseteq \mathbf{R}^n$. Then $\mathbf{X}$ is *Gaussian* (also, *normal*) if any finite subset of $\mathbf{X}$ is normally distributed.

Recall that a q-dimensional Gaussian random vector, $\mathbf{X} \sim N(\mu, \Sigma)$, has joint PDF given by

$$f_{\mathbf{X}}(x) = \frac{1}{\sqrt{(2\pi)^q|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{t}}\Sigma^{-1}(x-\mu)\right) \tag{2.30}$$

where $\mu \in \mathbf{R}^n$ and $\Sigma \in \mathbf{R}^{n \times n}$ is a positive-definite matrix. In the case that $n = 1$, $\mu = 0$, and $\Sigma = 1$, $\mathbf{X}$ is said to have the *standard* normal distribution.

Remark 54. Note that the distribution of a Gaussian process is determined by its mean and covariance function (Kallenberg, 1997, p. 200, Lemma 11.1). Furthermore, note that a Gaussian process is stationary if and only if it is weakly stationary. To see this recall the discussion following our definition of invariant random processes (Def. 43, Rem. 44). For a stationary random process to be weakly stationary it is sufficient that the mean and covariance exist, which they do. For a weakly-stationary random process to be stationary it is sufficient that the mean and covariance of the random process completely define the joint CDF of all finite subsets of that random process, which they also do.

We now discuss some common examples of covariance functions used to define Gaussian processes on the parameter space $T = \mathbf{R}^n$, equipped with a norm $\|\cdot\|$. These are found throughout the literature (e.g. Sacks et al., 1989; Rasmussen and Williams, 2006). They are necessarily positive definite. Assume that we have chosen a basis and that in this basis the norm is given by

$$\|t\| = \sqrt{t^{\mathrm{t}}Mt} \tag{2.31}$$

where M is a positive-definite matrix, which we call the 'metric matrix'. If M is the identity matrix, then this is the usual Euclidean norm. If it is some other positive-definite matrix, we will refer to it as the *generalized Euclidean norm*. In a vector space equipped with the Euclidean norm, the set of all vectors with unit norm defines a n-dimensional sphere. In a vector space equipped with the generalized Euclidean norm, the set of all vectors with unit norm defines an n-dimensional ellipsoid.

[10] A history of the Poisson process is given by Stirzaker (2000).

Example 55 (white-noise covariance function). A covariance function, k, is a *white-noise covariance function* if there exists $\sigma^2 \in \mathbf{R}_{>0}$ such that

$$k(s,t) = \sigma^2 \delta(s,t) \tag{2.32}$$

for all $s, t \in T$, where δ is the Kronecker delta. If the random process has white-noise covariance function and constant mean then it is stationary and isotropic. The metacorrelation function is then given by

$$\rho(r) = \begin{cases} 1 & \text{if } r = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

(To see this note that by definition $\rho(r) = r(t, t+r)$. Hence $\rho(r) = k(t, t+r)/\sigma^2 = \delta(t, t+r)$.)

Example 56 (Matérn covariance function). A covariance function, k, is a *Matérn covariance function* (Matérn, 1986, p. 18)[11] if there exist $\sigma^2, a, \nu \in \mathbf{R}_{>0}$ such that

$$k(s,t) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (a\|s-t\|)^\nu K_\nu(a\|s-t\|)) \tag{2.34}$$

where K_ν is the modified Bessel function of the second kind, and Γ is the Gamma function. The constant a is called the *scale constant.* If the random process has Matérn covariance function and constant mean then it is stationary and isotropic. The metacorrelation function is then given by

$$\rho(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} (ar)^\nu K_\nu (ar) . \tag{2.35}$$

Suppose that $\nu = p + 1/2$ for $p \in \mathbf{N}$, i.e. suppose that that ν is half integer. In this case the metacorrelation function is given by

$$\rho(r) = \exp(-ar) \frac{p!}{2p!} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} (2ar)^{p-i}. \tag{2.36}$$

We call a Matérn covariance function with half-integer ν the $(p+1)$*-th order autoregressive covariance function.* Note that the first-order autoregressive metacorrelation function is given by

$$\rho(r) = \exp(-ar). \tag{2.37}$$

We also call this the *exponential metacorrelation function.*

Example 57 (Ornstein-Uhlenbeck covariance function). In the case that the parameter space has dimension $n = 1$, and $\nu = 1/2$ the Matérn covariance function is known as the *Ornstein-Uhlenbeck*

[11] A history of the Matérn covariance function is given by Guttorp and Gneiting (2006).

covariance function. In this case, the covariance function is given by

$$k(s,t) = \sigma^2 \exp(-a\|s-t\|) \tag{2.38}$$

for all $s,t \in T$. The metacorrelation function is given by

$$\rho(r) = \exp(-ar). \tag{2.39}$$

This is the exponential metacorrelation function (Eq. 2.37) for the special case of one-dimesensional parameter space.

Example 58 (p-exponential covariance function). A covariance function, k, is a *p-exponential covariance function* if there exist $\sigma^2, a \in \mathbf{R}_{>0}$ and $p \in \mathbf{N} \setminus \{0\}$ such that

$$k(s,t) = \sigma^2 \exp(-a\|s-t\|^p). \tag{2.40}$$

for all $s,t \in T$. If the random process has p-exponential covariance function and constant mean then it is stationary and isotropic. The metacorrelation function is then given by

$$\rho(r) = \exp(-ar^p). \tag{2.41}$$

The 1-exponential covariance function is normally called the 'exponential covariance function'. The 2-exponential covariance function is normally called the 'squared-exponential covariance function'. A constant-mean Gaussian process with p-exponential covariance is stationary. In Figures 2.1 and 2.2 I plot realizations of a centred Gaussian process for dimension $n = 1$ and $n = 2$, with covariance given by the squared-exponential covariance function. Note that these realizations are continuous.[12]

The above examples are of covariance functions defined on the real index set $T = \mathbf{R}^n$. Typically we imagine that T is time ($n = 1$) or space ($n = 1$, 2, or 3). In the case of constant mean, all of these examples define stationary and isotropic random processes. Stationary and isotropic random processes are of use to us if we believe that the mean and variance of the phenomenon we are modelling is invariant under translations and rotations of T.

2.2 GENERAL THEORY OF REPRODUCING KERNEL HILBERT SPACES

We have said that we are interested in linear prediction, and that the proper setting for linear prediction is the Hilbert space and the reproducing kernel Hilbert space (Aronszajn, 1950; Parzen, 1959). Let us now consider these subjects formally.

[12] This is a consequence of the Kolmogorov continuity theorem, which we will not consider here.

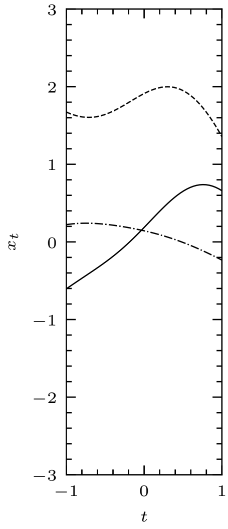

Figure 2.1 Three realizations of a centred Gaussian process defined on the real interval $[-1, 1]$ with squared-exponential covariance function (Ex. 58) where $a = 1/2$ and $\|\cdot\|$ is the usual Euclidean norm, given by $\|v\| = |v|$. Note that the realizations are continuous.

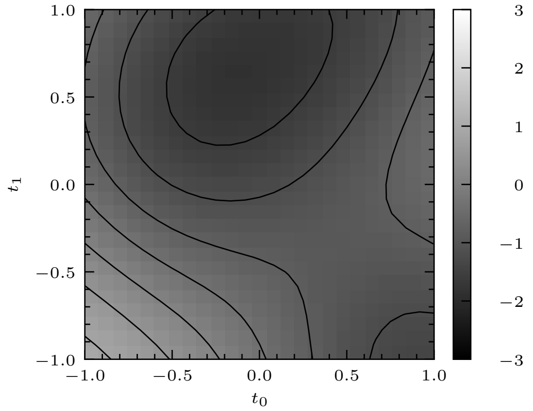

Figure 2.2 A realization of a centred Gaussian process defined on the real region $[-1, 1] \times [-1, 1]$ with squared-exponential covariance function (Ex. 58) where $a = 1/2$ and $\|\cdot\|$ is the usual Euclidean norm, given by $\|v\| = \sqrt{v^{t}v}$. Note that the realization is continuous.

2.2.1 Hilbert spaces

Definition 59 (Hilbert space). Let $\mathbf{V} = (V, \langle\cdot,\cdot\rangle)$ be a real inner product space. Let $\|\cdot\|$ be the norm induced by the inner product, and let d be the metric induced by this norm. Then $\mathbf{V}$ is a *Hilbert space* if the metric space (V, d) is complete (Def. 117, App. A).

Let $\mathbf{V} = (V, \langle\cdot,\cdot\rangle)$ be a Hilbert space. We denote by V^* the dual space of V, namely the real vector space consisting of linear functionals on $\mathbf{V}$, together with the usual addition and scalar multiplication of real-valued functions. Note that a Hilbert space is in particular a topological vector space: the norm, $\|\cdot\|$, induces a topology, $\mathscr{T}_{\mathrm{s}}$, which we call the *strong topology*. (For the reader unfamiliar with general topology, please consult the App. A.)

We denote by V' the subspace of V^* consisting of those linear functionals on $\mathbf{V}$ that are continuous with respect to the strong topology on V. Note that such a linear functional is continuous with respect to $\mathscr{T}_s$ if and only if it is bounded (see, for example, Rudin, 1991). We may endow V' with the norm

$$\|\varphi\|_{V'} := \sup\{|\varphi(f)| : \|f\| \leq 1, f \in \mathbf{V}\}, \tag{2.42}$$

for $\varphi \in V'$. It can be verified that $\|\cdot\|_{V'}$ satisfies the parallelogram law, and that it thus defines an inner product,

$$\langle\varphi,\psi\rangle_{V'} := \frac{1}{2}\big(\|\varphi+\psi\|^2_{V'} - \|\varphi\|^2_{V'} - \|\psi\|^2_{V'}\big), \tag{2.43}$$

for $\varphi, \psi \in V'$.

Definition 60 (continuous dual of a Hilbert space). The *continuous dual space* of $\mathbf{V}$ is the real inner product space $\mathbf{V}' := (V', \langle\cdot,\cdot\rangle_{V'})$.[13]

Recall that a set $A \subseteq V$ of vectors in an inner product space $\mathbf{V}$ is *orthogonal* if

$$\langle u, v\rangle = 0, \tag{2.44}$$

for all distinct $u, v \in A$. Such a set A is *orthonormal* if furthermore

$$\langle v, v\rangle = 1, \tag{2.45}$$

for all $v \in A$. The *span* of a set A, which we denote by $\operatorname{span}(A)$, is the closure of the linear span of A in $\mathbf{V}$. We say that A *spans* $\mathbf{V}$ (or that A is a *spanning set* of $\mathbf{V}$) if $\operatorname{span}(A) = \mathbf{V}$. Finally, a set A is an *orthonormal basis* of $\mathbf{V}$ if it is both orthonormal and a spanning set of $\mathbf{V}$.

[13] We use the word 'continuous' to distinguish this space from the dual of a vector space (Def. 134, App. A). Some authors use the term 'algebraic dual' instead of 'dual', and 'dual' instead of 'continuous dual'.

Theorem 61. *Every Hilbert space* $\mathbf{H}$ *has an orthonormal basis, and any two orthonormal bases of* $\mathbf{H}$ *have the same cardinality (i.e. number of elements).*

Proof. A proof is given by Bourbaki (1981). □

Definition 62 (dimension of a Hilbert space). The *dimension* of $\mathbf{H}$, as a Hilbert space, is the cardinality of some orthonormal basis (equivalently, of all orthonormal bases) of $\mathbf{H}$.

Definition 63 (congruence of Hilbert space). Let $\mathbf{H}_1 = (H_1, \langle\cdot,\cdot\rangle_1)$ and $\mathbf{H}_2 = (H_2, \langle\cdot,\cdot\rangle_2)$ be two Hilbert spaces. A map $\psi : \mathbf{H}_1 \longrightarrow \mathbf{H}_2$ is an *isometry* if

$$\langle v_1, v_2\rangle_1 = \langle \psi(v_1), \psi(v_2)\rangle_2. \tag{2.46}$$

for all $v_1, v_2 \in \mathbf{H}_1$. If ψ is both an isometry and a linear isomorphism then it is called a *congruence*. We say that $\mathbf{H}_1$ and $\mathbf{H}_2$ are *congruent* if there exists a congruence $\psi : \mathbf{H}_1 \longrightarrow \mathbf{H}_2$, and write $\mathbf{H}_1 \cong \mathbf{H}_2$.

We might instead call a congruence an 'isometric isomorphism'. It is the correct notion of isomorphism for the study of Hilbert spaces.

Remark 64. A congruence maps limits to limits, i.e. if ψ is a congruence then

$$u = \lim(u_n)_n \iff \psi(u) = \lim(\psi(u_n))_n, \tag{2.47}$$

for all vectors u and sequences of vectors $(u_n)_n$.

Theorem 65 (congruence theorem). *Let* $\mathbf{H}_1 = (H_1, \langle\cdot,\cdot\rangle_1)$ *and* $\mathbf{H}_2 = (H_2, \langle\cdot,\cdot\rangle_2)$ *be two Hilbert spaces. Let* $U = \{u_t \in H_1 : t \in T\}$ *and* $V = \{v_t \in H_2 : t \in T\}$ *be two sets of vectors, both indexed by a set* T*, such that* U *spans* $\mathbf{H}_1$*,* V *spans* $\mathbf{H}_2$*, and*

$$\langle u_s, u_t\rangle_1 = \langle v_s, v_t\rangle_2 \tag{2.48}$$

for all $s, t \in T$*. Then there is a congruence* $\psi : \mathbf{H}_1 \longrightarrow \mathbf{H}_2$ *such that*

$$\psi(u_t) = v_t \tag{2.49}$$

for all $t \in T$*.*

Proof. A proof is given by Parzen (1959, p. 11). □

Corollary 66. *Two Hilbert spaces are congruent if and only if they are of the same dimension.*

Proof. A proof is given by Parzen (1959, p. 13). □

Theorem 67 (Riesz representation theorem). *Let* $\mathbf{H}$ *be a Hilbert space. The map*

$$\Phi : \mathbf{H} \longrightarrow \mathbf{H}' \tag{2.50}$$

$$f \longmapsto [\Phi_f : g \longmapsto \langle f, g \rangle] \tag{2.51}$$

is a congruence.

Proof. A proof is given by Bourbaki (1981, Ch. V Sec. 1 No 7 Thm 3). □

Corollary 68. *If* $\mathbf{H}$ *is a Hilbert space then the continuous dual space* $\mathbf{H}'$ *is a Hilbert space.*

2.2.2 Hilbert spaces of functions

A *vector space of (real-valued) functions* is a pair (V, Ξ) where V is a vector space and Ξ is a set. Moreover we suppose that V is a set of functions on Ξ (i.e. V is a subset of $\mathbf{R}^{\Xi}$) and that addition and scalar multiplication in V are addition and scalar multiplication of functions, i.e.

$$(f+g)(\xi) = f(\xi) + g(\xi) \tag{2.52}$$

$$(\lambda f)(\xi) = \lambda f(\xi), \tag{2.53}$$

for $f, g \in V$, $\lambda \in \mathbf{R}$, and $\xi \in \Xi$. An *inner product space* (respectively, *Hilbert space*) *of functions* is a pair $(\mathbf{V}, \Xi)$, where $\mathbf{V}$ is an inner product space (respectively, Hilbert space) with underlying vector space V such that (V, Ξ) is a vector space of functions. We do not assume, unless otherwise stated, that the inner product satisfies any additional properties involving Ξ.

An inner product space of functions posseses not only a strong topology, $\mathscr{T}_{\mathrm{s}}$, induced by the norm but also a *weak topology*, $\mathscr{T}_{\mathrm{w}}$, induced by the *topology of pointwise convergence* on $\mathbf{R}^{\Xi}$: that is, a sequence $(f_n)_n$ in V converges in $\mathscr{T}_{\mathrm{w}}$ to f if and only if for all $\xi \in \Xi$ we have $f_n(\xi) \longrightarrow f(\xi)$. Accordingly, we say that a sequence in an inner product space of functions *weakly converges* if it converges in the weak topology.

Definition 69 (evaluation map). Given $(\mathbf{V}, \Xi)$, the *evaluation map* is the function

$$\mathrm{ev} : \Xi \longrightarrow \mathbf{V}^* \tag{2.54}$$

$$\xi \longmapsto [\mathrm{ev}_\xi : f \longmapsto f(\xi)]. \tag{2.55}$$

We will often be concerned with the case that the image $\mathrm{ev}(\Xi)$ is contained in the continuous dual $\mathbf{V}'$, which is not always so, as the following example illustrates.

Example 70. Consider $C([0,1])$, the space of continuous functions $[0,1] \longrightarrow \mathbf{R}$ endowed with the L^2-norm, given by

$$\|f\| = \left(\int_0^1 |f(x)|^2 \, \mathrm{d}x \right)^{1/2}.$$

The map ev_0 evaluates each function at 0. For each $n > 0$, consider the piecewise linear function $f_n : [0,1] \longrightarrow \mathbf{R}$ given by

$$f_n(x) = \begin{cases} 1 - xn & \text{for } x \leq 1/n \\ 0 & \text{for } x > 1/n. \end{cases}$$

Then $f_n \longrightarrow c_0$, the constantly zero function. Nevertheless, $\mathrm{ev}_0(f_n) \not\longrightarrow \mathrm{ev}_0(c_0)$, and so ev_0 is not continuous.

Example 71. Every inner product space is congruent to a space of functions: the map

$$\Phi : \mathbf{V} \longrightarrow \mathbf{V}' \tag{2.56}$$

$$x \longmapsto [\Phi_x : y \longmapsto \langle x, y \rangle] \tag{2.57}$$

is a non-singular linear transformation and an isometry.

2.2.2.1 Functional completion

Any metric space (X, d) admits a completion, which is unique up to isometry over (X, d). Similarly, any real inner product space $\mathbf{V} = (V, \langle \cdot, \cdot \rangle)$ admits a completion, which is unique up to congruence over $\mathbf{V}$.

Definition 72 (functional completion). Let $(\mathbf{V}, \Xi)$ be an inner product space of functions. A *functional completion* of $(\mathbf{V}, \Xi)$ is an extension, $(\mathbf{V}, \Xi) \subseteq (\bar{\mathbf{V}}, \Xi)$, of inner product spaces of functions such that

(a) $\bar{\mathbf{V}}$ is complete,

(b) $\mathbf{V}$ is dense in $\bar{\mathbf{V}}$, and

(c) the image of Ξ under the evaluation map $\bar{\mathrm{ev}}$ is contained in $\bar{\mathbf{V}}'$.

We say $(\mathbf{V}, \Xi)$ is *functionally complete* if it is equal to its own functional completion.

The third condition (c) is equivalent to requiring that $\bar{\mathrm{ev}}(\xi) : \bar{\mathbf{V}} \longrightarrow \mathbf{R}$ is continuous with respect to the strong topology for each $\xi \in \Xi$. Note that the functional completion of an inner product space of functions is unique, if it exits.

Theorem 73 (existence of functional completions). *Let* $(\mathbf{V}, \Xi)$ *be an inner product space of functions. The following are equivalent.*

(i) *The inner product space of functions* $(\mathbf{V}, \Xi)$ *admits a functional completion.*

(ii) *We have both:*

(a) $\operatorname{ev}(\Xi) \subseteq \mathbf{V}'$, *and*

(b) *every Cauchy sequence in* $\mathbf{V}$ *which weakly converges to* 0 *must strongly converge to* 0.

Proof. A proof is given by Aronszajn (1950). □

2.2.3 Reproducing kernels and reproducing kernel Hilbert spaces

Having introduced Hilbert spaces, I now introduce the reproducing kernel, and reproducing kernel Hilbert spaces. In this section $(\mathbf{H}, \Xi)$, with $\mathbf{H} = (H, \langle \cdot, \cdot \rangle)$, will denote a Hilbert space of functions. The general theory of reproducing kernel Hilbert spaces was developed by Aronszajn (1950).

Definition 74 (reproducing kernel). A function $k : \Xi \times \Xi \longrightarrow \mathbf{R}$ is a *reproducing kernel* for $(\mathbf{H}, \Xi)$ if

(i) for all $x \in \Xi$, the map

$$\begin{aligned} k(x, \cdot) : \Xi &\longrightarrow \mathbf{R} \\ y &\longmapsto k(x, y) \end{aligned}$$

is in $\mathbf{H}$, and

(ii) for all $x \in \Xi$ and $f \in \mathbf{H}$, we have

$$\langle k(x, \cdot), f \rangle = f(x).$$

If a function meets the second of these conditions it is said to have the *reproducing property*. We say $(\mathbf{H}, \Xi)$ is a *reproducing kernel Hilbert space* if there exists a reproducing kernel for it.

Example 75. A finite dimensional inner product space of functions is a reproducing kernel Hilbert space. To see this, let $(\mathbf{V}, \Xi)$ be an inner product space of functions of dimension n. Immediately,

$\mathbf{V}$ is a Hilbert space, because it is complete under the norm induced by the inner product. Moreover, there is an orthonormal basis for $\mathbf{V}$, which we denote by $\{e_1, \ldots, e_n\}$. Indeed, $\{e_1, \ldots, e_n\}$ is also a linear basis for V. Define the function

$$k : \Xi \times \Xi \longrightarrow \mathbf{R} \tag{2.58}$$

$$(x, y) \longmapsto \sum_{i=1}^{n} e_i(x) e_i(y). \tag{2.59}$$

For each $x \in \Xi$, the function $k(x, \cdot)$ is an element of $\mathbf{V}$ since it is the finite sum

$$k(x, \cdot) = \sum_{i=1}^{n} e_i(x) e_i. \tag{2.60}$$

Moreover, k possesses the reproducing property: given $x \in \Xi$ and $f = \sum_i \alpha_i e_i \in \mathbf{V}$, we have

$$\langle k(x, \cdot), f \rangle = \langle \sum_i e_i(x) e_i, \sum_j \alpha_j e_j \rangle \tag{2.61}$$

$$= \sum_i \alpha_i e_i(x) \tag{2.62}$$

$$= f(x). \tag{2.63}$$

Thus $(\mathbf{V}, \Xi)$ is a reproducing kernel Hilbert space.

Lemma 76 (representation of a reproducing kernel). *If k is a reproducing kernel for $(\mathbf{H}, \Xi)$, then k is symmetric and*

$$\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y) \tag{2.64}$$

for all $x, y \in \Xi$.

Proof. By (i), the function $k(y, \cdot)$ is in $\mathbf{H}$. Applying (ii) to $f = k(y, \cdot)$, we have

$$\langle k(x, \cdot), k(y, \cdot) \rangle = k(y, x). \tag{2.65}$$

The result now follows from the symmetry of the inner product. □

Lemma 77 (Aronszajn inequality). *If k is a reproducing kernel for $(\mathbf{H}, \Xi)$, then we have*

$$|f(x) - g(x)| \leq \sqrt{k(x, x)} \|f - g\| \tag{2.66}$$

for $f, g \in \mathbf{H}$ and $x \in \Xi$.

Proof. Using the reproducing property of k, the Cauchy-Schwarz inequality, the definition of the norm, and Lemma 76 (in that order) we see that

$$\begin{aligned} |f(x) - g(x)| &= |\langle k(x,\cdot), f - g\rangle| && (2.67)\\ &\leq \|k(x,\cdot)\| \|f - g\| && (2.68)\\ &= \sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle} \|f - g\| && (2.69)\\ &= \sqrt{k(x,x)} \|f - g\| && (2.70) \end{aligned}$$

as required. □

We name the above inequality after Aronszajn. It is useful in proving the following theorems.

Theorem 78 (uniqueness of the reproducing kernel). *A Hilbert space of functions* $(\mathbf{H}, \Xi)$ *admits at most one reproducing kernel.*

Proof. Let $k_1, k_2 : \Xi \times \Xi \longrightarrow \mathbf{R}$ be two reproducing kernels for $(\mathbf{H}, \Xi)$, and let $x \in \Xi$. Then, for all $f \in H$, it is the case

$$\langle k_1(x,\cdot), f\rangle = f(x) = \langle k_2(x,\cdot), f\rangle, \tag{2.71}$$

by the reproducing property. In particular, taking $f = k_1(x,\cdot) - k_2(x,\cdot)$, we have

$$\langle k_1(x,\cdot) - k_2(x,\cdot), k_1(x,\cdot) - k_2(x,\cdot)\rangle = 0, \tag{2.72}$$

by linearity. By positive-definiteness, we have $k_1(x,\cdot) = k_2(x,\cdot)$. Since this in turn holds for all $x \in \Xi$, we have $k_1 = k_2$, as required. □

Theorem 79 (existence of reproducing kernel Hilbert spaces). *For a Hilbert space of functions* $(\mathbf{H}, \Xi)$*, the following are equivalent:*

(i) $(\mathbf{H}, \Xi)$ *is a reproducing kernel Hilbert space,*

(ii) the image $\mathrm{ev}(\Xi)$ *is contained in* $\mathbf{H}'$.

Proof. Our proof follows that of Aronszajn (1950). We begin by showing that the first condition implies the second. Suppose that $(\mathbf{H}, \Xi)$ is a reproducing kernel Hilbert space with reproducing kernel $k : \Xi \times \Xi \longrightarrow \mathbf{R}$. Let $x \in \Xi$. We must show that $\mathrm{ev}_x : \mathbf{H} \longrightarrow \mathbf{R}$ is continuous with respect to $\mathscr{T}_s$. By the Aronszajn inequality, we have

$$|\mathrm{ev}_x(f) - \mathrm{ev}_x(g)| = |f(x) - g(x)| \leq \sqrt{k(x,x)} \|f - g\| \tag{2.73}$$

for all $f, g \in \mathbf{H}$. Since $\sqrt{k(x,x)}$ does not depend on f or g, this shows that ev_x is continuous. Thus the image $\mathrm{ev}(\Xi)$ is contained in $\mathbf{H}'$.

We now show that the second condition implies the first. Conversely, we suppose that $\mathrm{ev}(\Xi)$ is contained in $\mathbf{H}'$. Let $x \in \Xi$. We consider the functional $\mathrm{ev}_x \in \mathbf{H}'$. By Theorem 67, $\Phi : \mathbf{H} \longrightarrow \mathbf{H}'$ is a congruence which maps $f \in \mathbf{H}$ to the function $\Phi_f : g \longmapsto \langle f, g \rangle$. In particular, Φ is surjective, so there exists $f_x \in \mathbf{H}$ such that $\Phi_{f_x} = \mathrm{ev}_x$. By writing $k(x, \cdot) = f_x$, we have defined a function $k : \Xi \times \Xi \longrightarrow \mathbf{R}$. It remains to verify that k is a reproducing kernel. Condition (i) of the definition is automatically satisfied since $k(x, \cdot) = f_x \in \mathbf{H}$. For condition (ii) of the definition, let $g \in \mathbf{H}$. Then

$$\langle k(x, \cdot), g \rangle = \Phi_{k(x,\cdot)}(g) = \mathrm{ev}_x(g) = g(x), \tag{2.74}$$

as required. □

Theorem 80. *Let $\mathbf{H}_k$ be a reproducing kernel Hilbert space with reproducing kernel k. Then the set $\{k(t, \cdot) : t \in T\}$ spans $\mathbf{H}_k$.*

Proof. A proof is given by Parzen (1959, p. 49, Theorem 5B) □

The following proposition is what makes reproducing kernel Hilbert spaces so interesting and useful.

Proposition 81. If $(\mathbf{H}, \Xi)$ is a reproducing kernel Hilbert space then convergence in the strong topology implies convergence in the weak topology.

Proof. This follows from the Aronszajn inequality. □

Remark 82. Let $(\mathbf{H}, \Xi)$ be a Hilbert space of functions. Then $(\mathbf{H}, \Xi)$ is functionally complete if and only if it is a reproducing kernel Hilbert space, by the existence theorem.

2.2.4 The Moore–Aronszajn theorem

The following theorem is given by Aronszajn (1950) who attributes it to E. H. Moore. It provides a characterization of reproducing kernels without explicit reference to their corresponding reproducing kernel Hilbert space. It can be thought of as the converse of the uniqueness of reproducing kernels (Thm 78).

Theorem 83 (Moore–Aronszajn theorem). *Let $k : \Xi \times \Xi \longrightarrow \mathbf{R}$ be a function. The following are equivalent.*

(i) The function k is a positive semi-definite kernel.

(ii) *The function k is a reproducing kernel for a unique Hilbert space of functions* $(\mathbf{H}, \Xi)$.

Proof. A proof is given by Aronszajn (1950). □

Remark 84. As before, one would perhaps expect the phrase 'up to congruence' to appear in the above theorem. However, by definition, such a Hilbert space must be a subspace of the space of functions from Ξ to $\mathbf{R}$, namely $\mathbf{R}^{\Xi}$.

2.2.5 Reproducing kernel Hilbert space representations of Hilbert spaces

An arbitrary Hilbert space (not necessarily a Hilbert space of functions) may be identified with a reproducing kernel Hilbert space by means of a congruence, i.e. we may always use a congruence to map an arbitrary Hilbert space into a reproducing kernel Hilbert space. Such a reproducing kernel Hilbert space is said to *represent* this arbitrary Hilbert space. The following congruence allows a natural representation of this kind (Parzen, 1959). It will be of crucial importance later, where we will represent Hilbert spaces of random variables by reproducing kernel Hilbert spaces in order to exploit their additional structure.

Definition 85 (canonical congruence). Let $\mathbf{H} = (H, \langle\cdot,\cdot\rangle_H)$ be the Hilbert space with a basis $\{v_t : t \in T\}$, and let $\mathbf{G}_k = (G, \langle\cdot,\cdot\rangle_G, k)$ be the reproducing kernel Hilbert space with reproducing kernel

$$k : T \times T \longrightarrow \mathbf{R} \tag{2.75}$$

$$(s,t) \longmapsto \langle v_s, v_t\rangle_H. \tag{2.76}$$

The *canonical congruence* between $\mathbf{H}$ and $\mathbf{G}_k$ is the congruence $\psi : \mathbf{H} \longrightarrow \mathbf{G}_k$ such that

$$v_t \longmapsto k(t,\cdot). \tag{2.77}$$

The canonical congruence exists by the congruence theorem (Thm 65). A vector $v \in \mathbf{H}$ may be repesented as $v = \sum_i a_i v_{t_i}$ (I leave the height of the sum ambiguous). Then there is a unique function $f \in \mathbf{G}_k$ such that $f = \psi(v)$. In fact:

$$f = \psi(v) \tag{2.78}$$

$$= \psi(\sum_i a_i v_{t_i}) \tag{2.79}$$

$$= \sum_i a_i \psi(v_{t_i}) \tag{2.80}$$

$$= \sum_i a_i k(t_i,\cdot). \tag{2.81}$$

For convenience I list these relationships in Table 2.1.

$\mathbf{H}$	$\mathbf{G}_k$
v_t	$k(t,\cdot)$
$\sum_t a_t v_t$	$\sum_t a_t k(t,\cdot)$

Table 2.1 Equivalent objects in $\mathbf{H}$ (a Hilbert space with basis $\{v_t : t \in T\}$), and $\mathbf{G}_k$ (the reproducing kernel Hilbert space with reproducing kernel given by $k(s,t) = \langle v_s, v_t \rangle_H$) under the canonical congruence $\psi : \mathbf{H} \longrightarrow \mathbf{G}_k$ (Def. 85). The space $\mathbf{H}$ *is not* necessarily a Hilbert space of functions. The space $\mathbf{G}_k$ *is* necessarily a Hilbert space of functions (specifically, functions on T). We make use of the congruence between $\mathbf{H}$ and $\mathbf{G}_k$ to exploit the structure of reproducing kernel Hilbert spaces.

Example 86 (finite index set). Given a function $f \in \mathbf{H}_k$, we may wish to find the coefficients $\{a_i\}_i$ explicitly. Suppose that $T = \{t_1, t_2, \dots, t_n\}$ is finite. Then we have the n simultaneous equations

$$f(t_j) = \sum_i a_i k(t_i, t_j) \tag{2.82}$$

for all $t_j \in T$. Because $\{v_t\}_{t \in T}$ is linearly independent, the matrix $K = (k(s_i, t_j))_{ij}$ is nonsingular, meaning that the inverse, K^{-1}, exists. We denote the ij-th element of K^{-1} by K_{ij}^{-1}. The above simultaneous equations then have solutions

$$a_i = \sum_j K_{ji}^{-1} f(t_j) \tag{2.83}$$

for all $i \in \{1, 2, \dots, n\}$. Note that this gives an explicit expression for the inverse congruence,

$$\psi^{-1}(f) = \sum_{i,j} K_{ij}^{-1} f(t_i) v_{t_j}. \tag{2.84}$$

2.3 REPRODUCING KERNEL HILBERT SPACE REPRESENTATIONS OF RANDOM PROCESSES

We are now in a position to give an account of random processes in the setting of reproducing kernel Hilbert spaces, and to use this formalism to give a geometric account of prediction. In is this account that was devoped by Parzen in the 1950s. First, let us fix in advance a probability space $\mathbf{\Omega} = (\Omega, \mathscr{M}, P)$, where Ω is an arbitrary set, $\mathscr{M}$ is a σ-algebra on Ω, and P is a probability measure on $(\Omega, \mathscr{M})$. (A brief account of the basics of probability theory is included in App. A.) We will necessarily be interested in second-order random processes. Let $L^2(\mathbf{\Omega})$ denote the set of all second-order random variables with domain $\mathbf{\Omega}$. That is, the domain of $L^2(\mathbf{\Omega})$ is the set of Lebesgue- and square-integrable functions $\mathbf{\Omega} \longrightarrow \mathbf{R}$ modulo the equivalence relation of equality almost everywhere.[14] The set $L^2(\mathbf{\Omega})$ admits the structure of a vector space. The function

$$\langle \cdot, \cdot \rangle : L^2(\mathbf{\Omega}) \times L^2(\mathbf{\Omega}) \longrightarrow \mathbf{R} \tag{2.85}$$

$$(X, Y) \longmapsto \mathrm{E}(XY) \tag{2.86}$$

is an inner product on $L^2(\mathbf{\Omega})$. In fact, the space $\mathbf{L}^2(\mathbf{\Omega}) = (L^2(\mathbf{\Omega}), \langle \cdot, \cdot \rangle)$ is complete with respect to the norm induced by this inner product, and is therefore a Hilbert space. Note, however, that elements of $\mathbf{L}^2(\mathbf{\Omega})$ are *not* functions on $\mathbf{\Omega}$, so *a fortiori* $\mathbf{L}^2(\mathbf{\Omega})$ is not a Hilbert space of functions.

Now consider the following subset of these random variables.

Definition 87 (linear span of a random process). Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process. The *linear span of* $\mathbf{X}$ is the set of all linear combinations of elements of $\mathbf{X}$, i.e. the set[15]

$$L(\mathbf{X}) = \{\sum_{i=1}^{n} a_i X_{t_i} : n \in \mathbf{N}, X_{t_i} \in \mathbf{X}, a_i \in \mathbf{R}\}. \tag{2.87}$$

Definition 88 (Hilbert space spanned by a random process). Let $L(\mathbf{X})$ be the linear span of a second-order random process $\mathbf{X}$. This linear span admits the structure of a vector space, which in turn admits a completion under the norm induced by the inner product. Denote this complete linear span by $\bar{L}(\mathbf{X})$. Then $\mathbf{L}(\mathbf{X}) := (\bar{L}(\mathbf{X}), \langle \cdot, \cdot \rangle)$ is a Hilbert space. We say that $\mathbf{L}(\mathbf{X})$ is *the Hilbert space spanned by the random process* $\mathbf{X}$.

For the sake of completeness we note that the norm is given by

$$\|X\| = \sqrt{\langle X, X \rangle} \tag{2.88}$$

$$= \sqrt{\mathrm{E}(X^2)}. \tag{2.89}$$

This in turn induces a metric, given by

$$d(X, Y) = \|X - Y\| \tag{2.90}$$

$$= \sqrt{\mathrm{E}((X - Y)^2)}. \tag{2.91}$$

As ever, a set $\{Y_i : i \in I\}$ of random variables is *orthogonal* if

$$\langle Y_i, Y_j \rangle = \mathrm{E}(Y_i Y_j) \tag{2.92}$$

$$= 0 \tag{2.93}$$

[14] Taking the equivalence classes modulo this equivalence relation ensures that $\langle \cdot, \cdot \rangle$ is positive definite.

[15] Some authors (for example, Parzen, 1959) call the linear span of a random process the 'the linear manifold spanned by a random process'.

for $i \neq j$, and is *orthonormal* if moreover

$$\langle Y_i, Y_i \rangle = \mathrm{E}(Y_i Y_i) \tag{2.94}$$

$$= 1, \tag{2.95}$$

for all i. Note that

$$\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{2.96}$$

$$= \langle X, Y \rangle - \mathrm{E}(X)\mathrm{E}(Y). \tag{2.97}$$

If a set of random variables is centred, i.e. if all its elements have zero mean, then it is orthogonal if all its elements are uncorrellated, and orthonormal if furthermore its elements have unit variance.

Having defined the Hilbert space spanned by a random process, we may now define the reproducing kernel Hilbert space representation of that Hilbert space, following Definition 85. There is then a canonical congruence between the Hilbert space spanned by a random process and its reproducing kernel Hilbert space representation. We may formalize this with the following theorem.

Theorem 89 (Loève representation theorem). *Let* $\mathbf{X}$ *be a second-order random process with covariance function* k*. Then* $\mathbf{L}(\mathbf{X})$ *is congruent with the reproducing kernel Hilbert space* $\mathbf{H}_k$.

Proof. The theorem is a restatement of the existence of canonical congruences. □

2.4 PREDICTION

In what follows $\mathbf{X} = \{X_t\}_{t \in T}$ is a random process on a probability space $\mathbf{\Omega} = (\Omega, \mathscr{M}, p)$. We wish to *predict* a random variable, $Z : \Omega \longrightarrow \mathbf{R}$ defined on the same probability space. As usual, $\mathbf{L}(\mathbf{X})$ is the Hilbert space spanned by $\mathbf{X}$.

Definition 90 (linear predictor). A *linear predictor based on* $\mathbf{X}$ is an element of $\mathbf{L}(\mathbf{X})$. (See Parzen, 1959, p. 53, and Berlinet and Thomas-Agnan, 2004, p. 77.)

Definition 91 (best linear predictor of a random variable). A linear predictor, $Y \in \mathbf{L}(\mathbf{X})$, of Z based on $\mathbf{X}$, is the *best linear predictor* (BLP) of Z if it minimizes the mean-squared error

$$\mathrm{MSE} : \mathbf{L}(\mathbf{X}) \longrightarrow \mathbf{R} \tag{2.98}$$

$$Y \longmapsto \mathrm{E}((Z - Y)^2). \tag{2.99}$$

(See Parzen, 1959, p. 125.)

Definition 92 (unbiased linear predictor of a random variable). A linear predictor, $Y \in \mathbf{L}(\mathbf{X})$ of Z based on $\mathbf{X}$, is said to be *unbiased* if

$$\mathrm{E}(Y) = \mathrm{E}(Z). \tag{2.100}$$

(See Parzen, 1959, p. 126.)

Definition 93 (best unbiased linear predictor of a random variable). An unbiased linear predictor, $Y \in \mathbf{L}(\mathbf{X})$, of Z based on $\mathbf{X}$, is the *best unbiased linear predictor*[16] (BLUP) of Z if it minimizes the mean-squared error

$$\mathrm{MSE} : \mathbf{L}(\mathbf{X}) \longrightarrow \mathbf{R} \tag{2.101}$$

$$Y \longmapsto \mathrm{E}((Z - Y)^2). \tag{2.102}$$

(Parzen, 1959, p. 125.)

We will denote the both the BLP and BLUP of a random variable Z by $\hat{Z}$.

Remark 94. Note that

$$\mathrm{MSE}(Y) = \mathrm{E}((Z - Y)^2) \tag{2.103}$$

$$= \mathrm{var}(Z - Y) + (\mathrm{E}(Z - Y))^2 \tag{2.104}$$

$$= \mathrm{var}(Z - Y) + \mathrm{bias}(Y)^2 \tag{2.105}$$

where $\mathrm{bias}(Y) = \mathrm{E}(Z - Y)$. For an unbiased linear predictor, $\mathrm{bias}(Y) = 0$. Therefore, the mean-squared error of an unbiased linear predictor of Z is

$$\mathrm{MSE}(Y) = \mathrm{var}(Z - Y). \tag{2.106}$$

Proposition 95. The BLP of Z based on $\mathbf{X}$ is the conditional expectation $\mathrm{E}(Z \mid \mathbf{X})$, and its MSE is $\mathrm{E}(\mathrm{var}(Z \mid \mathbf{X}))$.

Proof. Let $Y \in \mathbf{L}(\mathbf{X})$ be a linear predictor of Z based on $\mathbf{X}$. The mean-squared error of Y is

$$\mathrm{MSE}(Y) = \mathrm{E}((Z - Y)^2)) \tag{2.107}$$

$$= \mathrm{E}((Z - \mathrm{E}(Z \mid X) + \mathrm{E}(Z \mid X) - Y)^2)) \tag{2.108}$$

$$= \mathrm{E}((Z - \mathrm{E}(Z \mid X))^2 + 2(Z - \mathrm{E}(Z \mid X))(\mathrm{E}(Z \mid X) - Y) + (\mathrm{E}(Z \mid X) - Y)^2) \tag{2.109}$$

$$= \mathrm{E}(\mathrm{E}((Z - \mathrm{E}(Z \mid X))^2 \mid X)) + \mathrm{E}(2(\mathrm{E}(Z \mid X) - Y)\mathrm{E}((Z - \mathrm{E}(Z \mid X)) \mid X))) + \tag{2.110}$$

$$\mathrm{E}((\mathrm{E}(Z \mid X) - Y)^2) \tag{2.111}$$

$$= \mathrm{E}(\mathrm{var}(Z \mid X)) + \mathrm{E}((\mathrm{E}(Z \mid X) - Y)^2). \tag{2.112}$$

[16] Also called by some authors ‘the best linear unbiased predictor’ (for example, Sacks et al., 1989).

This expression is a minimum when the second term vanishes, i.e when $Y = \mathrm{E}(Z \mid X)$. When this is the case, we have that $\mathrm{MSE}(Y) = \mathrm{E}(\mathrm{var}(Z \mid X))$. □

Remark 96. Recall that the conditional expectation $\mathrm{E}(Z \mid \mathbf{X})$ is a random variable, not a real number.

Example 97 (known joint normal distribution). In the case that the joint distribution of $\mathbf{X}$ and Z is known, we may use the above theorem to calculate the BLP explicitly. Suppose that $\mathbf{X} = (X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ is a finite random vector and that we know Z and $\mathbf{X}$ to have a joint normal distribution. We may form the augmented random vector $\mathbf{X}' = (Z, \mathbf{X})$. By indexing from zero we may write $\mathbf{X}' = (X_{t_0}, X_{t_1}, \ldots, X_{t_n})$, where $X_{t_0} = Z$. Suppose that the mean of this augmented random vector is $m' = (m_{X_{t_0}}, m_{\mathbf{X}})$ and that its covariance is $K' = (k(t_i, t_j))_{ij}$. This is the block matrix

$$K' = \begin{bmatrix} (k(t_0, t_0)) & (k(t_0, t_i))_i^{\mathrm{t}} \\ (k(t_i, t_0))_i & (k(t_i, t_j))_{ij} \end{bmatrix}. \tag{2.113}$$

Then, using the properties of normal distributions (see, for example, Rasmussen and Williams, 2006, p. 200), we find that

$$Z \mid (\mathbf{X} = \boldsymbol{x}) \sim N(m(\boldsymbol{x}), \sigma^2) \tag{2.114}$$

where

$$m(\boldsymbol{x}) = m_{t_0} + (k(t_0, t_i))_i^{\mathrm{t}} (k(t_i, t_j))_{ij}^{-1} (\boldsymbol{x} - m_{\mathbf{X}}) \text{ and} \tag{2.115}$$

$$\sigma^2 = k(t_0, t_0) - (k(t_0, t_i))_i^{\mathrm{t}} (k(t_i, t_j))_{ij}^{-1} (k(t_i, t_0))_i. \tag{2.116}$$

By Proposition 95 the BLP is

$$\hat{Z} = \mathrm{E}(Z \mid \mathbf{X}) \tag{2.117}$$

$$= m_{t_0} + (k(t_0, t_i))_i^{\mathrm{t}} (k(t_i, t_j))_{ij}^{-1} (\boldsymbol{X} - m_{\mathbf{X}}) \tag{2.118}$$

with MSE

$$\mathrm{MSE}(\hat{Z}) = \mathrm{E}(\mathrm{var}(Z \mid \mathbf{X}) \tag{2.119}$$

$$= \sigma^2. \tag{2.120}$$

The BLP is a linear combination of Gaussian random variables, and hence a Gaussian random variable itself. Note that, in this case, $\hat{Z}$ is unbiased because $\mathrm{E}(\hat{Z}) = m_{t_0} = \mathrm{E}(Z)$.

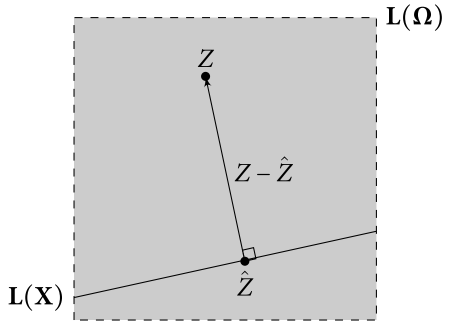

Figure 2.3 Let $\mathbf{L}(\boldsymbol{\Omega})$ be the Hilbert space of all random variables on a probability space $\boldsymbol{\Omega}$. Let $Z \in \mathbf{L}(\boldsymbol{\Omega})$ be a such a random variable. Let $\mathbf{X}$ be a random process and let $\mathbf{L}(\mathbf{X})$ be the Hilbert space spanned by $\mathbf{X}$. Then $\mathbf{L}(\mathbf{X}) \subset \mathbf{L}(\boldsymbol{\Omega})$. The least squares predictor for Z based on $\mathbf{X}$, denoted $\hat{Z}$, is the orthogonal projection of Z onto $\mathbf{L}(\mathbf{X})$. The mean-squared error is the norm of the vector $Z - \hat{Z}$.

2.5 PREDICTION IN THE SETTING OF REPRODUCING KERNEL HILBERT SPACES

Having considered the theory of random processes, reproducing kernel Hilbert spaces, and prediction, we are now in a position to give an account of prediction in the setting of reproducing kernel Hilbert spaces. The key insight is that prediction may be cast as a minimization problem in the appropriate Hilbert space. We have a random variable $Z \in \mathbf{L}(\boldsymbol{\Omega})$, and a subspace $\mathbf{L}(\mathbf{X}) \subset \mathbf{L}(\boldsymbol{\Omega})$. The norm on $\mathbf{L}(\boldsymbol{\Omega})$ is given by $\|U\| = \sqrt{\mathrm{E}(U^2)}$. Therefore the MSE of a linear predictor of Z is $\mathrm{MSE}(U) = \mathrm{E}((Z - U)^2) = \|Z - U\|^2$. To find the BLP of Z we must therefore find the random variable $\hat{Z} \in \mathbf{L}(\mathbf{X})$ that minimizes this norm. That such a minimum exists and is unique is a consequence of the following theorem.

Theorem 98. *Let* $\mathbf{H}$ *be a Hilbert space and let* $\mathbf{G} \subseteq \mathbf{H}$ *be a Hilbert subspace. For all* $v \in \mathbf{H}$ *there is a unique point* $\hat{v} \in \mathbf{G}$ *such that* $\|v - \hat{v}\| = \inf\{\|v - u\| : u \in \mathbf{G}\}$.

Proof. A proof is given by Bourbaki (1981, Ch. V Sec. 1 No 5 Thm 1). □

As a corollary of this theorem, the BLP of a random variable exists and is unique.

Theorem 99. *Given* $\mathbf{H}$, $\mathbf{G}$, v, *and* $\hat{v}$ *as in Theorem 98, it is the case that* $\hat{v}$ *is the orthogonal projection of* $v \in \mathbf{H}$ *onto* $\mathbf{G}$.

Proof. A proof is given by Bourbaki (1981, Ch. V Sec. 1 No 6 Thm 2). □

Corollary 100. *In the setting of Definition 91, the BLP of Z based on* $\mathbf{X}$ *is the orthogonal projection of Z onto* $\mathbf{L}(\mathbf{X})$.

Example 101 (polynomial projection of a square-integrable function). To illustrate the value of projection, let us project a square-integrable function on the unit interval onto a polynomial subspace of degree n, $\mathbf{P}^n([0,1])$. Denote the function by f and its projection by $\hat{f}$. The space of square-integrable functions on the unit interval is a Hilbert space. We may write $\mathbf{P}_n([0,1]) \subseteq \mathbf{L}^2([0,1])$. The subspace $\mathbf{P}_n([0,1])$ admits a finite orthonormal basis. One possible choice of such a basis is the first $n+1$ normalized shifted Legendre polynomials, $\{\tilde{p}_n\}_{i=0}^n$, where (Courant and Hilbert, 1989)

$$\tilde{p}_n(x) = \frac{2n+1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - x)^n. \tag{2.121}$$

It is therefore the case (see Proposition 132) that

$$\hat{f}(x) = \sum_{i=0}^{n} \langle f, \tilde{p}_i \rangle \tilde{p}_i(x) \tag{2.122}$$

where

$$\langle f, \tilde{p}_i \rangle = \int_0^1 f(x)\tilde{p}_i(x)\,\mathrm{d}x. \tag{2.123}$$

Suppose, for the sake of concreteness, that f is the exponential function on the unit interval, and $n = 2$. The first three normalized shifted Legendre polynomials are

$$\tilde{p}_0(x) = 1, \tag{2.124}$$

$$\tilde{p}_1(x) = 3(2x-1), \text{and} \tag{2.125}$$

$$\tilde{p}_2(x) = 5(6x^2 + 6x + 1). \tag{2.126}$$

Hence

$$\hat{f}(x) = (\mathrm{e}-1) + 3(3-\mathrm{e})(2x-1) + 5(7\mathrm{e}-19)(6x^2+6x+1). \tag{2.127}$$

This is the quadratic projection of the exponential function on the unit interval. We plot it in Figure 2.4.

2.5.1 Generalized integral equations

In the remainder of this section we work with a Hilbert space $\mathbf{H}$, an indexed subset $\{v_t : t \in T\}$, and an element $v \in \mathbf{H}$. By analogy with the notation $\mathbf{L}(\mathbf{X})$ (Def. 88) we denote by $\mathbf{L}(\{v_t : t \in T\})$

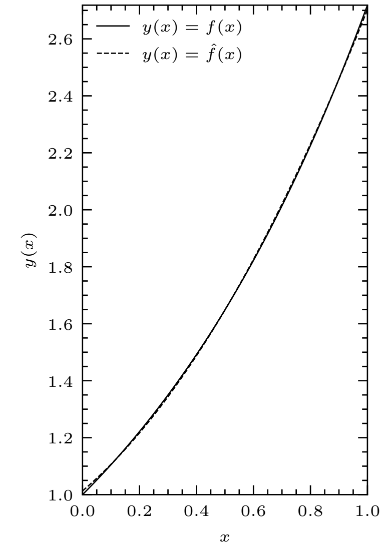

Figure 2.4 The exponential function on the unit interval, f, and its least-squares second-degree polynomial estimate, $\hat{f}$, given by expression 2.127.

the completion of the subspace of $\mathbf{H}$ spanned by $\{v_t : t \in T\}$. As in Theorems 98 and 99, we let $\hat{v}$ denote the orthogonal projection of v onto $\mathbf{L}(\{v_t : t \in T\})$. As a consequence, we have

$$\langle v - \hat{v}, u \rangle = 0, \tag{2.128}$$

for all $u \in \mathbf{L}(\{v_t : t \in T\})$. We view this as a family of equations for $\hat{v}$, with v given and u ranging over $\mathbf{L}(\{v_t : t \in T\})$. Equivalently, we seek $\hat{v} \in \mathbf{H}$ of minimum norm such that

$$\langle \hat{v}, u \rangle = \langle v, u \rangle \tag{2.129}$$

for all $u \in \mathbf{L}(\{v_t : t \in T\})$ (see Figure 2.3). This equation is a *generalized integral equation* (see Parzen, 1959, p. 58). Its solution, if it exists, may be found using the theory of reproducing kernel Hilbert spaces.

Definition 102 (generalized integral equation). Let $\mathbf{H}$ be a Hilbert space, let $\{v_t : t \in T\} \subseteq H$ be a set of vectors indexed by a set T, let $f : T \longrightarrow \mathbf{R}$, and let v be a variable element[17] of $\mathbf{L}(\{v_t : t \in T\})$. A system of equations of the form

$$\langle v, v_t \rangle = f(t) \text{ for all } t \in T \tag{2.130}$$

[17] By 'variable element' we mean an unknown or undetermined element, not a random variable.

is called a *generalized integral equation*. Such an element $v \in \mathbf{L}(\{v_t : t \in T\})$ solving these equations is called a *solution* to the generalized integral equation.

Example 103 (integral equation). Consider the Hilbert space of square-integrable functions, $\mathbf{L}^2([a,b])$. Let $\{k(\cdot,t) : t \in [a,b]\} \subseteq \mathbf{L}^2([a,b])$ be a set of functions. Let $x \in \mathbf{L}(\{k(\cdot,t) : t \in [a,b]\})$ be an element of the linear span of $\{k(\cdot,t) : t \in [a,b]\}$, and let $f : [a,b] \longrightarrow \mathbf{R}$ be a function. The generalized integral equation

$$\langle x, k(\cdot,t)\rangle = f(t) \tag{2.131}$$

is an integral equation of the form

$$\int_a^b x(s)k(s,t)\,\mathrm{d}s = f(t). \tag{2.132}$$

The fact that Definition 102 generalizes this equality is the origin of the name 'generalized integral equation'.

Note that generalized integral equations do not necessarily have a solution. The following theorem tells us when a solution does exist, and allows us to find the solution when it exists.

Theorem 104 (generalized integral equation theorem). *Let* $\mathbf{H}$ *be a Hilbert space, let* $\{v_t : t \in T\} \subseteq H$ *be a set of vectors indexed by a set* T*, and let* $f : T \longrightarrow \mathbf{R}$*. The following are equivalent.*

(a) There exists a unique minimum-norm solution $v_0 \in \mathbf{L}(\{v_t : t \in T\})$ *to the generalized integral equation*

$$\langle v_0, v_t\rangle = f(t),$$

for all $t \in T$*.*

(b) The function f *is an element of the reproducing kernel Hilbert space* $\mathbf{G}_k$ *with kernel* k *defined by*

$$k(s,t) = \langle v_s, v_t\rangle,$$

for all $s, t \in T$*.*

Furthermore, in this case,

$$v_0 = \psi^{-1}(f) \tag{2.133}$$

where ψ *is the* canonical congruence $\psi : \mathbf{L}(\{v_t : t \in T\}) \longrightarrow \mathbf{G}_k$ *(Def. 85). The norm of* v_0 *is*

$$\|v_0\| = \|f\|_G. \tag{2.134}$$

Proof. The Hilbert space $\mathbf{L}(\{v_t : t \in T\})$ and reproducing kernel Hilbert space $\mathbf{G}_k$ are congruent. In fact the canonical congruence $\psi : \mathbf{L}(\{v_t : t \in T\}) \longrightarrow \mathbf{G}_k$ is such that $\psi(v_t) = k(t, \cdot)$ (Def. 85). Therefore, if $f \in \mathbf{G}_k$ then $v = \psi^{-1}(f)$ is a solution, because

$$\langle v, v_t \rangle = \langle \psi(v), k(t, \cdot) \rangle_G \tag{2.135}$$

$$= \langle f, k(t, \cdot) \rangle_G \tag{2.136}$$

$$= f(t) \tag{2.137}$$

(where we have used the properties of congruences, Def. 63, the canonical congruence, Def. 85, and the reproducing property, Def. 74, in that order). Conversely, if there exists a solution $v \in \mathbf{L}(\{v_t : t \in T\})$ then $\psi(v) = f \in \mathbf{G}_k$. To show that v is unique, when also we demand that it has minimum norm, we suppose that there exists $f' \in \mathbf{G}_k$ such that $v = \psi^{-1}(f')$. Then

$$f'(t) = \langle f', k(t, \cdot) \rangle \tag{2.138}$$

$$= \langle v, v_t \rangle \tag{2.139}$$

$$= f(t). \tag{2.140}$$

By the properties of congruences, we have that $\|v\| = \langle v, v \rangle = \langle f, f \rangle_G = \|f\|_G$. □

The theorem states that the unique minimum-norm solution of a generalized integral equation exists if f is an element of the reproducing kernel Hilbert space spanned by $\{v_t : t \in T\}$ with reproducing kernel k. In this case, the existence of the canonical congruence between the two spaces allows us to find the solution. In fact, the solution of the equation is the preimage of f under the canonical congruence. With this theorem now in place, we are in a position to solve the projection equation (eq. 2.129). I show a schematic representation of this solution method in Figure 2.5. We may use then use the generalized integral equation theorem to find the BLP.

Theorem 105 (Parzen's BLP theorem). *Let $\mathbf{X} = \{X_t\}_{t \in T}$ be a random process, and let $Z : \Omega \longrightarrow \mathbf{R}$ be a random variable defined on the same probability space as $\mathbf{X}$. Let k be the function given by $k(s,t) = \mathrm{E}(X_s X_t)$, and let $\mathbf{G}_k$ be the reproducing kernel Hilbert space with reproducing kernel k. Let ψ be the canonical congruence from $\mathbf{L}(\mathbf{X})$ to $\mathbf{G}_k$. Let $\rho_Z : T \longrightarrow \mathbf{R}$ be the function given by $\rho_Z(t) = \mathrm{E}(Z X_t)$. Then the BLP of Z based on $\mathbf{X}$ is*

$$\hat{Z} = \psi^{-1}(\rho_Z), \tag{2.141}$$

which has mean-squared error

$$\mathrm{MSE}(\hat{Z}) = \mathrm{E}(Z^2) - \|\rho_Z\|^2_{\mathbf{G}_k}. \tag{2.142}$$

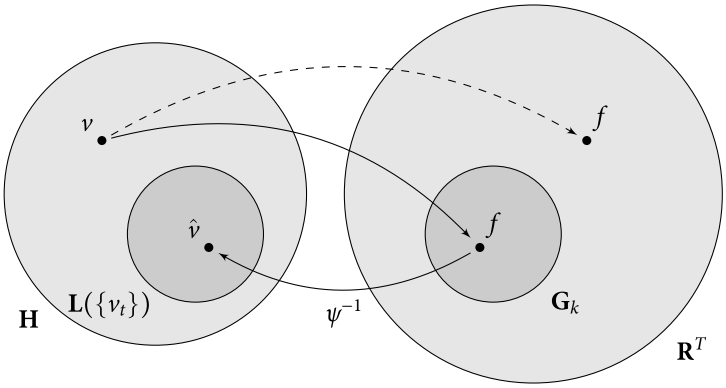

Figure 2.5 Finding the orthogonal projection of a vector, $v \in \mathbf{H}$, onto a subspace, $\mathbf{L}(\{v_t : t \in T\})$. We denote this orthogonal projection $\hat{v}$. In order to find $\hat{v}$ we must solve the equation $\langle \hat{v}, v_t \rangle = \langle v, v_t \rangle$ for all $t \in T$ (eq. 2.129). This is a generalized integral equation of the form $\langle \hat{v}, v_t \rangle = f(t)$ where the function $f : T \longrightarrow \mathbf{R}$ is given by $\langle v, v_t \rangle$. Thus $f \in \mathbf{R}^T$. The reproducing kernel Hilbert space $\mathbf{G}_k$, where k is such that $k(s,t) = \langle v_s, v_t \rangle$, is a subset of $\mathbf{R}^T$. Under the canonical congruence, $\psi : \mathbf{L}(\{v_t : t \in T\}) \longrightarrow \mathbf{G}_k$, it is the case that $\hat{v}$ exists if and only if $f \in \mathbf{G}_k$, whereupon, $\hat{v} = \psi^{-1}(f)$ (Thm 104).

Proof. The BLP of Z based on $\mathbf{X}$ is the orthogonal projection of Z onto $\mathbf{L}(\mathbf{X})$, by Corollary 100. Hence it is the case that $\langle \hat{Z}, U \rangle = \langle Z, U \rangle$, for all $U \in \mathbf{L}(\mathbf{X})$. By the generalized integral equation theorem, Theorem 104, we have that $\hat{Z} = \psi^{-1}(\rho_Z)$. □

2.5.2 Finding the BLUP

Instead of finding the BLP of a random variable Z, we will now find its BLUP. We will do this under the assumption the mean of Z is an element of some class of possible mean functions. As before, we have a random variable $Z \in \mathbf{L}(\boldsymbol{\Omega})$, and a random process $\mathbf{X}$ defined on the same probability space. We have $\mathbf{L}(\mathbf{X})$, the Hilbert space spanned by the random process $\mathbf{X} = \{X_t\}_{t \in T}$. We have $\mathbf{G}_k$, the reproducing kernel Hilbert space with reproducing kernel k given by $k(s,t) = \langle X_s, X_t \rangle_{L(\mathbf{X})}$, and the canonical congruence $\psi : \mathbf{L}(\mathbf{X}) \longrightarrow \mathbf{G}_k$. We will assume that the true mean of $\mathbf{X}$, denoted m_0, is an unknown element of a known subspace of the reproducing kernel Hilbert space, namely $\mathbf{M} \subseteq \mathbf{G}_k$.

Any $m \in \mathbf{M}$ is a function $T \longrightarrow \mathbf{R}$. By extending linearly, each m induces a function $\mathrm{E}_m : \mathbf{L}(\mathbf{X}) \longrightarrow \mathbf{R}$. If we know an explicit formula for the elements m of M, then we may extend the domain of E_m to $\mathbf{L}(\boldsymbol{\Omega})$. Thus, when we are explicitly given M, it makes sense to write $\mathrm{E}_m(Z)$, for all $m \in \mathbf{M}$. For the sake of clarity, we will even write the expectation $\mathrm{E}(X)$ as $\mathrm{E}_{m_0}(X)$, with which it is anyway equal, for all $X \in \mathbf{L}(\boldsymbol{\Omega})$. The mean-squared error of a random variable Y is $\mathrm{MSE}(Y) = \mathrm{E}((Z-Y)^2) = \mathrm{E}_{m_0}((Z-Y)^2)$, so it seems that to compute the mean-squared error we must be able to compute the expectation, which we do not know *a priori*. But this is not so. Recall that in the unbiased case (eq. 2.106), when

$$\mathrm{E}_{m_0}(Z) = \mathrm{E}_{m_0}(Y) \tag{2.143}$$

we have that

$$\begin{aligned} \mathrm{MSE}(Y) &= \mathrm{var}(Z-Y) && (2.144) \\ &= \mathrm{var}(Z) + \mathrm{var}(Y) - 2\mathrm{cov}(Z,Y) && (2.145) \end{aligned}$$

where we have used the fact that $\mathrm{var}(aZ+bY) = a^2\mathrm{var}(Z)+b^2\mathrm{var}(Y)+2ab\mathrm{cov}(Z,Y)$. If we know these three terms, $\mathrm{var}(Z)$, $\mathrm{var}(Y)$, and $2\mathrm{cov}(Z,Y)$, then we may compute the mean-squared error without knowing the true mean, m_0. Rather that assume that we know $\mathrm{cov}(Z,Y)$ directly, we will instead assume that we know the function ρ_Z given by

$$\rho_Z(t) = \mathrm{cov}(Z, X_t). \tag{2.146}$$

This then allows us to compute $\mathrm{cov}(Z,Y)$.

We will say that Z is *predictable* if there exists $Y \in \mathbf{L}(\mathbf{X})$ such that

$$\mathrm{E}_m(Z) = \mathrm{E}_m(Y), \tag{2.147}$$

for all $m \in \mathbf{M}$. Such a Y is called a *uniformly unbiased* linear predictor for Z. (Parzen, 1959, p. 122). Our aim is to find such a Y with minimum mean-squared error, which we couch as a minimization problem in the reproducing kernel Hilbert space $\mathbf{G}_k$. To this end we note the following.

Lemma 106 (representation of the mean and covariance). *Let $g, h \in \mathbf{G}_k$, and write $U = \psi^{-1}(g)$ and $V = \psi^{-1}(h)$. It is the case that*

$$\begin{aligned} \mathrm{E}_m(U) &= \langle g, m\rangle && (2.148) \\ \mathrm{cov}(U,V) &= \langle g, h\rangle, \textit{ and} && (2.149) \\ \mathrm{var}(U) &= \|g\|^2. && (2.150) \end{aligned}$$

Proof. It is the case that $g = \sum_i a_i k(t_i, \cdot)$. Using this fact, the linearity of the inner product, and the reproducing property of k, we have that

$$\langle g, m\rangle = \langle m, \sum_i a_i k(t_i, \cdot)\rangle \tag{2.151}$$

$$= \sum_i a_i \langle m, k(t_i, \cdot)\rangle \tag{2.152}$$

$$= \sum_i a_i m(t_i) \tag{2.153}$$

$$= \mathrm{E}_m(u). \tag{2.154}$$

We also have that

$$\langle g, h\rangle = \langle \sum_i a_i k(t_i, \cdot), \sum_j a_j k(t_j, \cdot)\rangle \tag{2.155}$$

$$= \langle \sum_i a_i k(t_i, \cdot), \sum_j a_j k(t_j, \cdot)\rangle \tag{2.156}$$

$$= \sum_{i,j} a_i a_j \langle k(t_i, \cdot), k(t_j, \cdot)\rangle \tag{2.157}$$

$$= \sum_{i,j} a_i a_j k(t_i, t_j) \tag{2.158}$$

$$= \mathrm{cov}(u, v). \tag{2.159}$$

We therefore have that

$$\mathrm{var}(u) = \mathrm{cov}(u, u) \tag{2.160}$$

$$= \langle g, g\rangle \tag{2.161}$$

$$= \|g\|^2, \tag{2.162}$$

as required. □

Given these results, we find that the MSE of a random variable $Y = \psi^{-1}(g)$ is

$$\mathrm{MSE}(Y) = \mathrm{var}(Z) + \mathrm{var}(Y) - 2\mathrm{cov}(Z, Y) \tag{2.163}$$

$$= \mathrm{var}(Z) + \langle g, g\rangle - 2\langle \rho_Z, g\rangle \tag{2.164}$$

$$= \mathrm{var}(Z) + \langle g, g\rangle - \langle \rho_Z, g\rangle - \langle g, \rho_Z\rangle + \langle \rho_Z, \rho_Z\rangle - \langle \rho_Z, \rho_Z\rangle \tag{2.165}$$

$$= \mathrm{var}(Z) + \langle g, g - \rho_Z\rangle - \langle \rho_Z, g - \rho_Z\rangle - \langle \rho_Z, \rho_Z\rangle \tag{2.166}$$

$$= \mathrm{var}Z + \langle g - \rho_Z, g - \rho_Z\rangle - \langle \rho_Z, \rho_Z\rangle \tag{2.167}$$

$$= \mathrm{var}(Z) - \|\rho_Z\|^2 + \|g - \rho_Z\|^2. \tag{2.168}$$

Thus, to find the BLUP of Z we must find g such that $\psi^{-1}(g)$ is uniformly unbiased, subject to which g minimizes the norm $\|g - \rho_Z\|$. (The term $\|\rho_Z\|$ is a constant.)

We may formalize this argument in the following theorem, due to Parzen (1959, p. 123, Thm 11A). This is the principle result of the reproducing kernel Hilbert space approach to prediction.

Theorem 107 (Parzen prediction theorem). *Let* $\mathbf{X} = \{X_t\}_{t \in T}$ *be a second-order random process with covariance function* k *and mean function* $m \in \mathbf{M} \subset \mathbf{G}_k$. *Let* Z *be a predictable random variable. The BLUP of* Z *is*

$$\hat{Z} = \psi^{-1}(\hat{\rho}_Z) \tag{2.169}$$

where $\hat{\rho}_Z$ *is the minimizer of* $\|\rho_Z - f\|$, *for all* $f \in \mathbf{G}_k$, *subject to the constraint that* $\langle \hat{\rho}_Z, m \rangle = \mathrm{E}_m(Z)$, *for all* $m \in \mathbf{M}$. *Furthermore, the mean-squared error of* $\hat{Z}$ *is*

$$\mathrm{MSE}(\hat{Z}) = \mathrm{var}(Z) - \|\rho_Z\|^2 + \|\rho_Z - \hat{\rho}_Z\|^2. \tag{2.170}$$

Proof. We first define the set of functions which correspond to uniformly unbiased linear predictors:

$$U := \{g \in G_k : \forall m \in M \, \langle g, m \rangle = E_m(Z)\}. \tag{2.171}$$

Thus $Y \in \mathbf{L}(\mathbf{X})$ is a uniformly unbiased linear predictor for Z if and only if $\psi(Y) \in U$. Such a predictor Y is *best* (subject to uniform unbiasedness) if and only if $g = \psi(Y)$ minimizes

$$\|g - \rho_Z\| \tag{2.172}$$

among $g \in U$. By the projection theorem (Thm 99), this amounts to choosing g to be the projection of ρ_Z onto U. Denote this projection by $\hat{\rho}_Z$ and write $\hat{Z} := \psi^{-1}(\hat{\rho}_Z)$. Then $\hat{Z}$ is the required BLUP. For the final claim, we compute the mean-squared error:

$$\mathrm{MSE}(\hat{Z}) = \mathrm{E}_{m_0}((Z - \hat{Z})^2) \tag{2.173}$$

$$= \mathrm{var}(Z) - \|\rho_Z\|^2 + \|\rho_Z - \hat{\rho}_Z\|^2, \tag{2.174}$$

by Lemma 106. □

2.5.3 Finite-dimensional space of mean functions

Having established the Parzen prediction theorem, we may consider the special case of finite-dimensional $\mathbf{M}$. This will yield a readily computably expression for the BLUP that we may use to construct metamodels. Let us suppose, then, that $\mathbf{M}$ is of finite dimension, q, with known basis $(\varphi_j)_{j=1}^q$. Recall that ρ_Z is given and our aim is the compute $\hat{\rho}_Z$. For convenience, we introduce a function h defined by $h := \hat{\rho}_Z - \rho_Z$. Our task may now be re-expressed in terms of h. We seek h of minimum norm subject to the constraint

$$\langle h, m \rangle = \langle \hat{\rho}_Z, m \rangle - \langle \rho_Z, m \rangle, \tag{2.175}$$

for all $m \in \mathbf{M}$. The orthogonal projection of such an h onto $\mathbf{M}$ must still satisfy this constraint, and thus we equivalently seek $h \in \mathbf{M}$ satisfying Equation 2.175, for all $m \in \mathbf{M}$. Since $(\varphi_j)_{j=1}^q$ spans $\mathbf{M}$, it even suffices to find $h \in \mathbf{M}$ satisfying $\langle h, \varphi_j \rangle = d_j$, where $d_j := \langle \hat{\rho}_Z, \varphi_j \rangle - \langle \rho_Z, \varphi_j \rangle$, for all $j \in \{1, 2, \ldots, q\}$. If we write $h = \sum_{i=1}^q a_i \varphi_i$ then $\langle h, \varphi_j \rangle = \sum_{i=1}^q a_i \langle \varphi_i, \varphi_j \rangle$. Therefore, it suffices that $d_j = \sum_{i=1}^q a_i \langle \varphi_i, \varphi_j \rangle$ for all j. Let us define the Gram matrix, $M = (\langle \varphi_i, \varphi_j \rangle)_{ij}$. We may rewrite the last expression as $d_j = \sum_{i=1}^q a_i M_{ij}$. This is a family of simultaneous equations, which we may solve to find that $a_i = \sum_{j=1}^q M_{ji}^{-1} d_j$. Therefore $h = \sum_{i,j} d_j M_{ji}^{-1} \varphi_i$. Furthermore,

$$\hat{Z} = \psi^{-1}(\hat{\rho}_Z) \tag{2.176}$$
$$= \psi^{-1}(\rho_Z) + \psi^{-1}(h) \tag{2.177}$$
$$= \psi^{-1}(\rho_Z) + \sum_{i,j} d_j M_{ji}^{-1} \psi^{-1}(\varphi_i) \tag{2.178}$$

with mean-squared error

$$\operatorname{MSE}(\hat{Z}) = \operatorname{var}(Z) - \|\rho_Z\|^2 + \|h\|^2 \tag{2.179}$$
$$= \operatorname{var}(Z) - \|\rho_Z\|^2 + \sum_{i,j} d_j M_{ji}^{-1} d_i. \tag{2.180}$$

This is an agreement with Parzen (1959, p. 124), who gives the same expressions without proof.

It is convenient to put this in matrix notation. Let $\varphi = (\varphi_i(t_0))_i$, let $A = (\varphi_j(t_i))_{ij}$, let $k = (k(t_0, t_i))_i$, and let $\hat{\mathrm{B}} = (A^{\mathrm{t}} K^{-1} A)^{-1} A^{\mathrm{t}} K^{-1} X$. Then the above expressions are equivalent to the following:

$$\hat{Z} = \varphi^{\mathrm{t}} (A^{\mathrm{t}} K^{-1} A)^{-1} A^{\mathrm{t}} K^{-1} X + k^{\mathrm{t}} K^{-1} (X - A(A^{\mathrm{t}} K^{-1} A)^{-1} A^{\mathrm{t}} K^{-1} X) \tag{2.181}$$
$$= \varphi^{\mathrm{t}} \hat{\mathrm{B}} + k^{\mathrm{t}} K^{-1} (X - A\hat{\mathrm{B}}), \tag{2.182}$$

and

$$\operatorname{MSE}(\hat{Z}) = k(t_0, t_0) - k^{\mathrm{t}} K^{-1} k + (\varphi^{\mathrm{t}} - k^{\mathrm{t}} K^{-1} A)(A^{\mathrm{t}} K^{-1} A)^{-1} (\varphi - A^{\mathrm{t}} K^{-1} k). \tag{2.183}$$

These expressions are the same as those found by Sacks et al. (1989, eq. 7, p. 413) using different means. Howeve, nowhere in the literature is an explicit connection made between Parzen's prediction theorem and these expressions. We have therefore derived them in a novel manner.

Note that the expression for the BLUP, $\hat{Z}$, is a sum of two terms. The first is an estimate of the mean, $m(t_0)$. Recall that $m(t_0) = \sum_{i=1}^{q} \beta_i \varphi_i(t_0)$, or in matrix notation $m(t_0) = \varphi^{\mathrm{t}}\mathrm{B}$. By comparing this expression with the first term of the expression 2.182, namely $\varphi^{\mathrm{t}}\mathrm{B}$, we identify $\hat{\mathrm{B}}$ as the estimator of B, i.e. as an estimator of the unknown coefficients $(\beta_i)_{i=1}^{q}$. This is referred to as the *generalized least-squares estimator* of B (Seber and Lee, 2003). The second term is a weighted sum of the residuals $(X - A\hat{\mathrm{B}})$. Note that $A\mathrm{B}$ is just the tuple of mean values $(m(t))_{t\in T}$, and that $A\hat{\mathrm{B}}$ is the generalized least-squares estimator of this tuple. Hence $(X - A\hat{\mathrm{B}})$ is indeed a tuple of residuals for this estimator. The tuple $k^{\mathrm{t}}K^{-1}$ may be thought of a tuple of *weights*, whereupon the second term of expression 2.182, namely $k^{\mathrm{t}}K^{-1}(X - A\hat{\mathrm{B}})$, becomes a weighted sum of the residuals.

2.5.4 Gaussian process emulation

The BLUP, as given in expression 2.182, is used in computer science to create emulators for computationally expensive simulations. Recall that we think of a simulation as a means of evaluating a parameterized function, $(f(\cdot;t))_{t\in T}$, where T is the parameter space of that function. We create a metamodel consisting of a random process, $\mathbf{Z} = \{Z_t\}_{t\in T}$, and then select a subset of that process $\mathbf{X} \subset \mathbf{Z}$, which we use to generate the BLUP of an arbitrary element $Z \in \mathbf{Z}$, along with its associated MSE. A realization of $\mathbf{X}$ is generated, which gives a realization of $\hat{Z}$. This is our best guess for the true value of Z. To create confidence regions for this best guess, we must know something about the distribution of our process, $\mathbf{Z}$. It is common to assume that $\mathbf{Z}$ is Gaussian. In this case prediction is known as 'Gaussian process emulation' (Rasmussen and Williams, 2006).[18] We may then compute a confidence region for our best guess in the usual way. This assumption of Gaussianity is useful but not necessary, and we might do well to refer to prediction in the context of computer experiments as 'random process emulation' to emphasize this fact. It is also worth mentioning the use of the BLUP in spatial statistics and, in particular, geostatistics. In this context, prediction is known as *kriging* (Krige, 1951). Typically, the predicted quantity is the density of a mineral deposit within a potential mining field. The density of the deposit is represented as the realization of a random process $\mathbf{Z} = \{Z_t\}_{t\in T}$ where $T = \mathbf{R}^3$ represents the three-dimensional volume of mining field in question. Again, a subset of that process is chosen, $\mathbf{X} \subset \mathbf{Z}$, a realization of $\mathbf{X}$ is generated by practical experiment, i.e. by sinking bore-holes to the correct positions. Kriging comes in three

[18] Presumably the term 'emulation' originates with the fact that one computer programme is being used to *emulate* another. This is somewhat analagous to the use of software (*emulators*) to allow one operating system to run programmes written for another.

flavours: *ordinary kriging*, *simple kriging*, and *universal kriging*. In ordinary kriging (Matheron, 1963; Cressie, 1988) we compute the BLUP under the assumption that the random process $\mathbf{X}$ is intrinsically stationary (Def. 51). In simple kriging (Cressie, 1990) we compute the BLP (Def. 91) under the assumption that the mean is a known constant. In universal kriging (Cressie, 1986) we compute the BLUP under the assumption that the mean, m, is a polynomial of given order, and the condition that the random process $\mathbf{Z} - m$ is intrinsically stationary.

Much of the work in random process emulation has been borrowed from kriging. Indeed Sacks et al. (1989) state that the method of emulation has been 'adopted from kriging in the spatial statistics literature'. But there is a fundamental difference that has been ignored by the literature. In spatial statistics the parameter space is $\mathbf{R}^3$. In random process emulation, the parameter space is far more general, as we have discussed in Chapter 1. Concepts of stationarity and isotropy that apply in the case of $\mathbf{R}^3$ (or, indeed, $\mathbf{R}^n$) do not necessarily apply in the case of such parameter spaces. To this end, the literature implicitly assumes that parameter space is either embedded in $\mathbf{R}^n$, or mapped to $\mathbf{R}^n$ under some continous transformation. This is deeply unsatisfactory and should be remedied by the development of new covariance functions based on the instrinsic structure of parameter space. I plan to pursue this goal in future work.

Bower et al. (2010) do an OLS fit first. They fix the correlation length by eye, using knowledge of the function they

Chapter 3

Mean and covariance specification in Gaussian-process emulation

In finding the BLUP we have assumed that the covariance function, k, is known and that the mean belongs to some known family of means. In practice we know neither the covariance nor the family to which the mean belongs. Similarly, in finding the BLP we have assumed that the joint distribution of the process $\mathbf{X}$ and the random variable Z is known. In the Gaussian case (Ex. 97) this amounts to assuming that the mean and covariance of the distribution are known. In practice, we do not know this joint distribution. What then, should we do? In the case of the BLUP we must approximate the covariance and the family of means. And in the case of the BLP we must approximate the joint distribution.

For the BLUP we will use the method of ordinary kriging, assuming that the random process, $\mathbf{X}$, is stationary, and hence has constant mean. For the BLP we will use the method of simple kriging, assuming that the random process, $\mathbf{X}$, has constant mean. Moreover, we will assume in both cases that the random vector, $\mathbf{X}$ and the predictable random variable Z are both drawn from the same stationary Gaussian random process. Further, we will assume that the covariance function itself is an element of some family of covariance functions. The assumption of Gaussianity then allows us to optimize the parameter of this covariance function using the principle of maximum likelihood (Santner et al., 2003). Given a covariance function, and a family of mean functions in this way, we are then in a position to compute the BLUP directly. To to compute BLP, however, we must assume some value for the mean function. There is no good way to do this, and it is frequently assumed, arbitrarily, to be zero (Rasmussen and Williams, 2006). We will avoid making this this assumption, but will instead attempt to quantify the bias of the resulting predictor.

3.1 CHOOSING THE COVARIANCE FUNCTION

Following common practice (Sacks et al., 1989), we will assume that the covariance function, k, is an element of some parameterized family of covariance functions, say one of the families given in Examples 56–58. In other words, we will assume the covariance function to be an element of the model $(k_\theta)_{\theta\in\Theta}$. The tuple θ is the parameter of the covariance function, and the set Θ is the set of all possible parameters of the covariance function.

Under the assumption of known covariance function, we had a single predictor for the BLUP or BLP, which we denoted $\hat{Z}$ in each case. Under the assumption that the covariance function is one element of a family of covariance functions, we now have multiple predictors, one for each element of the family of covariance functions. We may denote each such predictor $\hat{Z}_\theta$. In the case of the BLP (Ex. 97, Eq. 2.118) we may write

$$\hat{Z}_\theta = m_Z + k_\theta^{\mathrm{t}} K_\theta^{-1}(X - m_X) \tag{3.1}$$

where m_Z is the mean of Z, m_X is the mean of X, $k_\theta = (k_\theta(t_0, t_i))_i$ and $K_\theta = (k_\theta(t_i, t_j))_{ij}$. And in the case of the BLUP we may write

$$\hat{Z}_\theta = \varphi^{\mathrm{t}}\hat{\mathrm{B}}_\theta + k_\theta^{\mathrm{t}} K_\theta^{-1}(X - A\hat{\mathrm{B}}_\theta) \tag{3.2}$$

where $\varphi = (\varphi_i(t_0))_i$ is a tuple of basis functions for the space of mean functions, $A = (\varphi_j(t_i))_{ij}$, and $\hat{\mathrm{B}}_\theta = (A^{\mathrm{t}} K_\theta^{-1} A)^{-1} A^{\mathrm{t}} K_\theta^{-1} X$. However, the predictor should not be understood as a parameterized function. It has been derived on the *assumption* that the covariance function is known. To emphasize this fact we will refer to θ not as a parameter of the predictor but rather as a *hyperparameter* of the predictor, Z.

We may optimize our choice of hyperparameter using the principle of maximum likelihood. Assume that we have a realization of the random vector $\mathbf{X}$, giving data $((t_i, x_i))_{i=1}^n$. We choose the parameter of the covariance function that makes the data most probable. By assumption, the random process from which our data are drawn is Gaussian. The PDF for the random vector $\mathbf{X}$ is therefore

$$f_{\mathbf{X}}(x; m_X, \theta) = \frac{1}{\sqrt{(2\pi)^n |K_\theta|}} \exp\left(-\frac{1}{2}(x - m_X)^{\mathrm{t}} K_\theta^{-1}(x - m_X)\right). \tag{3.3}$$

where $x = (x_1, x_2, \ldots, x_n)$, $m_X = (m_{t_1}, m_{X_{t_2}}, \ldots, m_{X_{t_n}})$. By definition, the likelihood of θ is

$$L_\Theta(\theta; m_X, x) = f_X(x; m_X, \theta) \tag{3.4}$$

and hence the support for θ is

$$S_\Theta(\theta; m_X, x) = -\frac{1}{2}(x - m_X)^{\mathrm{t}} K_\theta^{-1}(x - m_X) - \frac{1}{2}\ln(|K_\theta|) - \frac{n}{2}\ln(2\pi). \tag{3.5}$$

If we make the further assumption that the random process has uniform variance then there exists a correlation function r such that $k_\theta = \sigma^2 r_\theta$ (Def. 36).[1] Let us define the *correlation matrix*, $R_\theta := (r_\theta(t_i, t_j))_{ij} = K_\theta/\sigma^2$. Then

$$S_\Theta(\theta; m_X, x) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln(|R_\theta|) - \frac{1}{2\sigma^2}(x - m_X)^{\mathrm{t}} R_\theta^{-1}(x - m_X) - \frac{n}{2}\ln(2\pi). \tag{3.6}$$

Assuming that the support is differentiable, we find that

$$\frac{\partial S}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}(x - m_X)^{\mathrm{t}} R_\theta^{-1}(x - m_X). \tag{3.7}$$

This has a unique root in σ^2, namely

$$\sigma^2 = \frac{1}{n}(x - m_X)^{\mathrm{t}} R_\theta^{-1}(x - m_X). \tag{3.8}$$

Hence (by substituting into expression 3.6) we find that

$$S_\Theta(\theta; m_X, x) = -\frac{n}{2}\ln\left(\frac{1}{n}(x - m_X)^{\mathrm{t}} R_\theta^{-1}(x - m_X)\right) - \frac{1}{2}\ln(|R_\theta|) - \frac{n}{2} - \frac{n}{2}\ln(2\pi). \tag{3.9}$$

If we are computing the BLP, in which case we know the mean, we may maximize this directly. If we are computing the BLUP, we do not know the mean vector m_X, but do have an estimator for it, $A\hat{\mathrm{B}}_\theta$ (see eq. 2.181 and the subsequent disccussion). Note also that $\hat{\mathrm{B}}_\theta = (A^{\mathrm{t}} K_\theta^{-1} A)^{-1} A^{\mathrm{t}} K_\theta^{-1} x = (A^{\mathrm{t}} R_\theta^{-1} A)^{-1} A^{\mathrm{t}} R_\theta^{-1} x$. We may therefore form the *concentrated support*,

$$S_\Theta(\theta; x) = -\frac{n}{2}\ln\left(\frac{1}{n}(x - A\hat{\mathrm{B}}_\theta)^{\mathrm{t}} R_\theta^{-1}(x - A\hat{\mathrm{B}}_\theta)\right) - \frac{1}{2}\ln(|R_\theta|) - \frac{n}{2} - \frac{n}{2}\ln(2\pi). \tag{3.10}$$

This depends only on the parameters of the correlation function, and not the variance.

To find the maximum-likelihood estimate of θ, namely $\hat{\theta}$, we may maximize this function subject to the constraint that R_θ is positive-semidefinite (positive-semidefiniteness being a necessary property of covariance matrices). Given $\hat{\theta}$ we may then compute

$$\hat{\sigma}^2 = \frac{1}{n}(x - A\hat{\mathrm{B}}_{\hat{\theta}})^{\mathrm{t}} R_{\hat{\theta}}^{-1}(x - A\hat{\mathrm{B}}_{\hat{\theta}}). \tag{3.11}$$

Expression 3.9 consists of three terms. Rasmussen and Williams (2006) point out that the first is a measure of fit quality, the second a complexity penalty, and the third a normalization constant. The function S_Θ will in general have multiple maxima, each maximum giving a different tradeoff

[1] This is necessarily the case if the random process is stationary (Rem. 44 and Eq. 2.16).

between fit quality and complexity. Note that the complexity penalty is a function of R_θ only, and quantifies the complexity of our attempted fit independent of the data. For a complicated fit, the covariance of any two points is low. Hence the determinant of the covariance matrix R_θ is small, and $\ln(|R_\theta|)$ diverges with complexity (i.e. as $|R_\theta| \longrightarrow 0$ so $\ln(|R_\theta|) \longrightarrow -\infty$). This strongly penalizes complex models.

3.2 KRIGING METHODS

By far the most common methods used in the construction of meta models are those of simple and ordinary kriging Cressie (1986, 1990). Despite the fact that they ignore the intrinsic statistical structure of the random process in question, we will follow the crowd and adopt these methods also. Recall that in ordinary kriging we compute the BLUP under the assumption of stationarity. In this case the mean is necessarily constant. Recall also, that in simple kriging compute the BLP under the assumption of constant mean. In this case it also common to further impose the assumption of stationarity. We thus have two cases: stationary Gaussian random process with known mean (the BLP) and unknown mean (the BLUP).

We must choose a family of covariance functions that ensures stationarity. By far the most commonly used covariance function is the squared-exponential (Def. 58) for $a = 1/2$, given by

$$k(s,t) = \sigma^2 \exp\left(-\frac{1}{2}\|s-t\|^2\right). \tag{3.12}$$

where

$$\|s-t\| = \sqrt{(s-t)^{\mathrm{t}} M (s-t)}. \tag{3.13}$$

(In fact this covariance function ensures that the random process is isotropic as well as stationary.) Furthermore, it is common to assume that the metric matrix is diagonal, $M = \operatorname{diag}(m_1, m_2, \ldots, m_D)$. The hyperparameter of the predictor is therefore the tuple $\theta = (\sigma^2, m_1, m_2, \ldots, m_D)$. The parameter of the covariance function is the tuple θ, and the parameter of the correlation function is the tuple m.

We may show (see, for example, Loeppky et al., 2009) that the mean squared gradient is

$$\mathrm{E}\left(\frac{\partial X(t)}{\partial t_i}\right)^2 = \sigma^2 m_i, \tag{3.14}$$

and therefore call m_i the *sensitivity* of the model to the i-th parameter. Because M is positive-semidefinite it has a unique positive-semidefinite inverse, which in turn has a unique square root, i.e. there exists a unique matrix L (not to be confused with the likelihood, L) such that $M = L^{-2}$. We may rewrite the squared exponential as

$$k(s,t) = \sigma^2 \exp\left(-\frac{1}{2}(s-t)L^{-2}(s-t)\right). \tag{3.15}$$

If M is diagonal then so is L and $L = \mathrm{diag}(l_1, l_2, \dots, l_D)$ where $m_i = l_i^{-2}$ for all i. This is a Gaussian function (not playing the role of a PDF) with root-mean square widths l_1, l_2, …, l_D. We call the value l_i the *correlation length* (also *scale length*) for the i-th component of the parameter. It is the characteristic scale length of the correlation function for the i-th component of the parameter.

To compute the BLUP, we assume constant but unknown mean, such that $m(t) = \mu$ for all $t \in T$. Then we have a single basis function, $\varphi = 1$. Recalling the definitions associated with Equations 2.181 and 2.183, we now have that

$$\varphi = (\varphi_i(t_0))_i = 1 \tag{3.16}$$

and

$$A = (\varphi_j(t_i))_{ij} = 1_n \tag{3.17}$$

where 1_n is the ones vector of length n. Then we require that the generalized least-squares estimate of the mean is $\varphi^{\mathrm{t}}\hat{\mathrm{B}} = \hat{\mu}$, and find that $\hat{\mathrm{B}} = \hat{\mu}$. We then have that

$$\hat{\mu} = (A^{\mathrm{t}}K^{-1}A)^{-1}A^{\mathrm{t}}K^{-1}x \tag{3.18}$$

$$= (1_n^{\mathrm{t}}K^{-1}1_n)^{-1}1_n^{\mathrm{t}}K^{-1}x \tag{3.19}$$

$$= \frac{1_n^{\mathrm{t}}K^{-1}x}{1_n^{\mathrm{t}}K^{-1}1_n}. \tag{3.20}$$

By subsituting the expressions $\varphi = 1$, $A = 1_n$, and $\hat{\mathrm{B}} = \hat{\mu}$ into Equation 2.181 we find the ordinary kriging predictor:

$$\hat{Z} = \hat{\mu} + k^{\mathrm{t}}K^{-1}(X - 1_n^{\mathrm{t}}\hat{\mu}). \tag{3.21}$$

Similarly (by subsituting the same expressions into Eq. 2.183) we find the MSE of the ordinary kriging predictor:

$$\mathrm{MSE}(\hat{Z}) = k(t_0, t_0) - k^{\mathrm{t}}K^{-1}k + \frac{(1 - k^{\mathrm{t}}K^{-1}1_n)^2}{1_n^{\mathrm{t}}K^{-1}1_n}. \tag{3.22}$$

These expression are consitent with those found by Cressie (1986) using different means. Note that if the random process has constant variance then the mean, $\hat{\mu}$, and the predictor, $\hat{Z}$, are independent of this variance. The variance is required only to compute the MSE.

Example 108 (Forrester function). Let us return to the example of the Forrester function (Eq. 1.41). We may now see in detail how we create a metamodel of this function. We will use ordinary kriging to compute the BLUP for an arbitrary value of the function, which we denote $\hat{y}(x)$. We assume the squared-exponential covariance function with diagonal metric matrix. First we generate a sample of the Forrester function, size $n = 10$. Second, we optimize the parameters of the covariance function using the maximum-likelihood method (expressions 3.10 and 3.11). The results of this optimization are shown in Table 3.1. We plot the support for the covariance function parameters in Figure 3.2. Third, we compute the predictor for the desired value of the function, along with the associated MSE. Because, by assumption, the random process is Gaussian this MSE is equal to the variance of the function and we may hence compute a confidence interval for our prediction directly. We plot the predictions for 100 such values in Figure 3.1, along with their 5σ confidence intervals. Note the values of the Forrester function are everywhere within 5σ of their predicted values.

It is enlightening to compute predictions for the values of the function given a parameter of the covariance function that differs from its MLE. According to Table 3.1 the MLE of the correlation hyperparameter is $m = 39.2857$. The associated correlation length is $l = 39.2857^{-1/2} = 0.1595$. Let us suppose that we choose a correlation length of factor four either side of this value, i.e. suppose that $l = 0.1595/4 = 0.03989$ (whereupon $m = 628.5712$) or that $l = 0.1595 \times 4 = 0.6382$ (whereupon $m = 2.4554$).[2] We retain the MLE of $\sigma^2 = 58.2386$. Using these values we recompute the predictions shown shown in Figure 3.1. These are shown in Figure 3.3.

In the first case the predictions exhibit overfitting: predictions are biased towards the mean, $\hat{mu}$, with high MSE. In the second case, the predictions exhibit underfitting: they have low MSE. In the first case, the length scale is in fact shorter than the separation of samples, meaning that even adjacent samples are poorly correlated, when in fact they are well correlated. In the second case, the length scale is comprable to that of the width of the domain, meaning that all samples are well correlated, when in fact only adjacent samples are well correlated. In the first case the MSE is so large as to make the prediction useless. We can have very little confidence in the predictions. In the second case, the MSE is so low that the values of the Forrester function are not everywhere within 5σ of their predicted values.

[2] We would choose a more natural factor of 10 but a length scale of $10l$ results in numerical instability, and prevents us from computing predictions. We address the issue of numerical instability later in the chapter.

σ^2	m	l	$\hat{\mu}$
58.2386	39.2857	0.1595	4.0956

Table 3.1 Maximum-likelihood estimate for the hyperparameter elements, σ^2 and m, of the squared-exponential covariance function used in emulating the Forrester function. Also shown are the length scale, $l = m^{-1/2}$, and the the maximum-likelihood estimate of the random process mean, $\hat{\mu} = (1_n^t K^{-1} x)/(1_n^t K^{-1} 1_n)$ (Eq. 3.20).

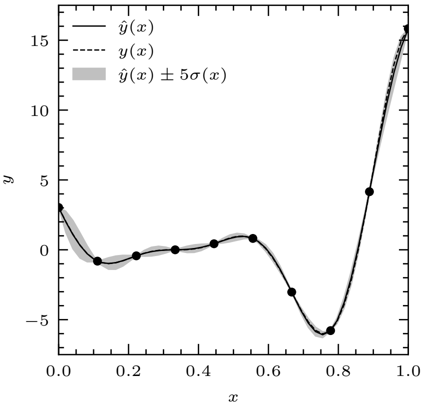

Figure 3.1 The Forrester function (Ex. 108), and its predicted values computed using GPE and the sample shown (filled black circles). The five-sigma confidence intervals for the predictions are also shown.

3.3 VALIDATING THE EMULATOR

We would like to know how good our predictions are. In the example of the Forrester function given above we were able to compare our predictions with the function's values. This will in general not be possible. Our principal tool in assessing the quality of our predictions is a procedure known as *validation* (Wasserman, 2007). Having constructing a predictor, we would ideally test it by evaluating the model at some number of parameters $t \in T$, not used in the construction of the predictor. In the case that the model is expensive to evaluate, this imposes an impractical computational burden. Instead we use *leave-one-out cross-validation* (LOOCV). We omit the i-th pair, (x_i, t_i), from our data to give the *reduced data*, $\{(x_j, t_j)\}_{j \neq i}$. Using these reduced data we then compute a prediction for x_i. We will call this the *i-th LOOCV prediction*, and denote it $\hat{x}_{-i}$.

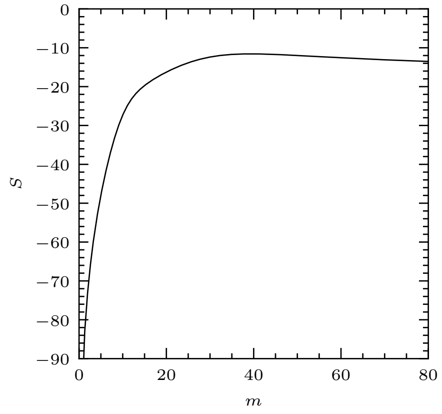

Figure 3.2 Support for the correlation hyperparameters of the squared-exponential covariance function used in emulating the Forrester function.

The omission of a single datum should not significantly affect the performance of the predictor, and we therefore expect this prediction to be close to the value x_i. We may quantify this closeness using the *i-th LOOCV residual* (Wasserman, 2007), i.e.

$$r_{-i} = x_i - \hat{x}_{-i}. \tag{3.23}$$

By assumption, the random process is Gaussian, meaning the variance of the predictor is equal to the MSE. The standard deviation of the predictor is the square root of its MSE (Eq. 3.22), which we denote σ_{-i}. We may therefore form the *i-th standardized LOOCV residual* (Jones et al., 1998), i.e.

$$e_{-i} = \frac{r_{-i}}{\sigma_{-i}}. \tag{3.24}$$

A further useful statistic is the *LOOCV score* (Wasserman, 2007), i.e. the mean-squared LOOCV residuals

$$R^2 = \frac{1}{n}\sum_{i}^{n} r_{-i}^2. \tag{3.25}$$

We expect the LOOCV residuals to be normally distributed with zero mean and unit variance. In particular, according to three-sigma rule, we should not expect to observe values outside the interval $[-3, 3]$. Moreover we expect the LOOCV residuals to exhibit no trend in parameter space, T. The LOOCV score tends to zero in the infinite-training data limit. Typically, we require it to be less than 10 % of the range of the sample, $\text{range}(x) := \max(x) - \min(x)$ (Jones et al., 1998).

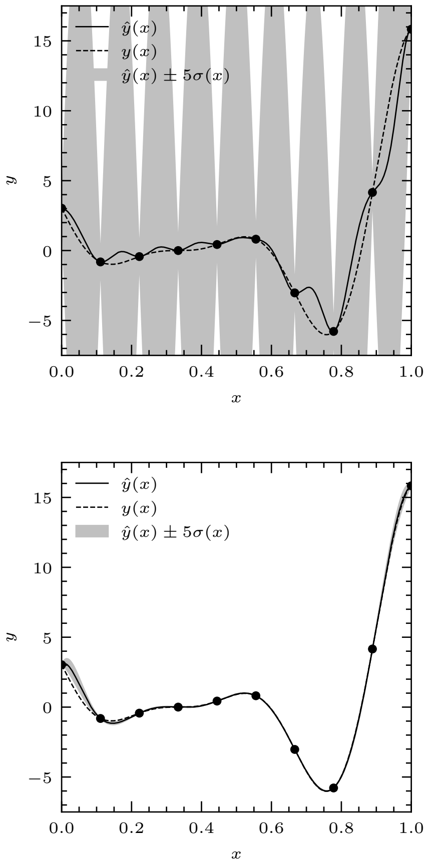

Figure 3.3 When the length scale is too small (top) the metamodel exhibits overfitting: predictions are biased towards the mean, and the MSE is large. When the length scale is too large (bottom) the metamodel exhibits underfitting: the MSE is small.

We may use these statistics to validate the predictor, since we should observe these properties in our results. A failure to observe these properties is diagnostic of poor performance of the predictor.

One way to compare two distributions is by means of a *quantile-quantile plot*, in which we plot the quantiles of one distribution against the quantiles of the other (Wilk and Gnanadesikan, 1968). If the two distributions are the same, then the points will lie on the diagonal. If either distribution is empirical we may substitute its ordered observed values for its quantiles. To check that the standardized LOOCV residuals are distributed as required, we therefore plot the *n* ordered LOOCV residuals of our fit against the *n*-th quantiles of the normal distribution. We may check for trends in the LOOCV residuals by plotting them against their coordinates in T.

In summary then, the validation of our predictor consists of computing the cross-validation score and making the following plots:

(a) LOOCV predictions against equivalent values,

(b) LOOCV residuals against coordinates in T, and

(c) quantile-quantile plot showing ordered LOOCV residuals against equivalent quantiles of the standard Gaussian distribution.

Example 109. We may perform validation for the example of the Forrester function (Ex. 108). The plots are shown in Figure 3.4. Looking at the plot of the LOOCV predictions against equivalent values, we see that the points lie on the diagonal, indicating good prediction accuracy. Looking at the plot of the cross-validation residuals against their coordinates in T, we see that the points exhibit no trend parameter space, T. Looking at the quantile-quantile plot we see that the points depart somewhat from the diagonal, indicating in this case, that the distribution of the cross-validation residuals is lighter in the tails than is the normal distribution. However, this plot is of limited use for such a small sample size of $n = 10$, so this fact should not overly concern us. We compute the LOOCV score to be $R = 1.5693$. We say that our predictor for the Forrester function has not failed validation.

We expect poor performance of our predictor when neighbouring points are poorly correlated, i.e. when the scale of features in our function is approximately equal to or less than the point separation. We also expect the accuracy to be poor at the boundary of parameter space, where the model is constrained by data on one side only.

3.4 AVOIDING MEAN MISSPECIFICATION

In computing the BLP we must know the mean of the joint distribution of Z and $\mathbf{X}$. This leaves us needing to know the mean of our random process. Rasmussen and Williams (2006) compute the BLP under the assumption that this mean is zero. They thus derive the expressions

$$\hat{Z} = k^{\mathrm{t}} K^{-1} X. \tag{3.26}$$

and

$$\mathrm{MSE}(\hat{Z}) = k(t,t) - k^{\mathrm{t}} K^{\mathrm{t}} k. \tag{3.27}$$

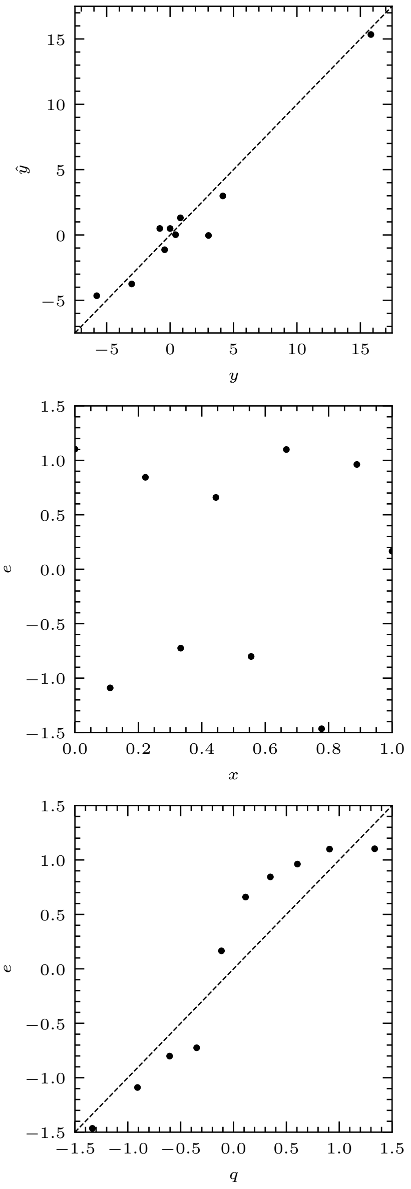

Figure 3.4 Validation plots for the emulation of the Forrester function. LOOCV predicted values against equivalent true values. LOOCV residuals. Quantile-quantile plot showing ordered LOOCV residuals against equivalent quantiles of the standard Gaussian distribution.

(These are Eqs 2.118 and 2.120 for $m_{t_0} = 0$ and $m_{\mathbf{X}} = 0_n$, where 0_n is the zero vector of length n.) They state that this may be done without loss of generality, arguing that the use of zero mean is only a notational convenience.[3] Regrettably, however, they are not able to make such a simplification for notational convenience. In doing so, we are assuming that the mean is constant, and that this constant is zero. These are both strong assumptions. If the random process does not satisfy them, the results of our emulation will be biased. The biasedness of the BLP is, to the best of my knowledge, acknowledged nowhere in the literature on GPE.

It is worthwhile asking how the BLP computed under the assumption of zero mean will differ from the BLUP computed under the assumption of constant but unknown mean. In other words, what is the bias of the BLP computed under assumption of zero mean? To this end we note that the mean-squared error may be decomposed into a variance and a bias term, as follows:

$$\mathrm{MSE}(\hat{Z}) = \mathrm{E}((\hat{Z} - Z)^2) \tag{3.28}$$

$$= \mathrm{var}(\hat{Z} - Z) + (\mathrm{E}(\hat{Z} - Z))^2 \tag{3.29}$$

$$= \mathrm{var}(\hat{Z} - Z) + \mathrm{bias}(\hat{Z})^2 \tag{3.30}$$

where $\mathrm{bias}(\hat{Z}) := \mathrm{E}(\hat{Z} - Z)$. If, as Rasmussen and Williams assume, $\hat{Z} = k^{\mathrm{t}} K^{-1} X$ then we find that

$$\mathrm{var}(\hat{Z} - Z) = \mathrm{var}(Z) + \mathrm{var}(\hat{Z}) - 2\mathrm{cov}(\hat{Z}, Z) \tag{3.31}$$

$$= \mathrm{var}(Z) + \mathrm{var}(k^{\mathrm{t}} K^{-1} X) - 2\mathrm{cov}(k^{\mathrm{t}} K^{-1} X, Z) \tag{3.32}$$

$$= k(x, x) + \mathrm{var}(k^{\mathrm{t}} K^{-1} X) - 2\mathrm{cov}(k^{\mathrm{t}} K^{-1} X, Z) \tag{3.33}$$

$$= k(x, x) + k^{\mathrm{t}} K^{-1} K K^{-1} k - 2k^{\mathrm{t}} K^{-1} k \tag{3.34}$$

$$= k(x, x) + k^{\mathrm{t}} K^{-1} k - 2k^{\mathrm{t}} K^{-1} k \tag{3.35}$$

$$= k(x, x) - k^{\mathrm{t}} K^{-1} k \tag{3.36}$$

where we have used the fact that $\mathrm{var}(a^{\mathrm{t}} X) = a^{\mathrm{t}} K a$, and that $\mathrm{cov}(a^{\mathrm{t}} X, Z) = a^{\mathrm{t}} k$. We also find that

$$\mathrm{bias}(\hat{Z}) = \mathrm{E}(k^{\mathrm{t}} K^{-1} X - Z) \tag{3.37}$$

$$= k^{\mathrm{t}} K^{-1} \mathrm{E}(X) - m(t) \tag{3.38}$$

where $\mathrm{E}(X) = (\mathrm{E}(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_n)) = (m(t_1), m(t_2), \ldots, m(t_n))$. The bias is dependent on the mean and the covariance of the random process, X. We can, in general, say no more about it. The zero-mean BLP is therefore unbiased only in the case that $k^{\mathrm{t}} K^{-1} \mathrm{E}(X) = m(t)$.

[3] They state that 'for notational simplicity we will take the mean function to be zero, although this need not be done' (Rasmussen and Williams, 2006, p. 13), referring to fact that non-zero mean may used, as we have in deriving the BLUP.

Suppose that the mean of the random process is in fact constant, but not necessarily zero, i.e. suppose that $m(t) = \mu$ for all $t \in T$. Then the bias is

$$\operatorname{bias}(\hat{Z}) = \mu k^{\mathrm{t}} K^{-1} 1 - \mu \tag{3.39}$$

$$= \mu(k^{\mathrm{t}} K^{-1} 1 - 1). \tag{3.40}$$

In this case the zero-mean BLP is unbiased only in the cases that $\mu = 0$ or $k^{\mathrm{t}} K^{-1} 1 = 1$. This latter requirement is equivalent to the elements of the tuple $K^{-1} k$ summing to 1.

It is also worthwhile asking when the BLUP reduces to the zero-mean BLP. For this to be the case we require that the generalized least-squares estimate of the mean, $\hat{\mu}$, is zero. Recall that

$$\hat{\mu} = \frac{1_n^{\mathrm{t}} K^{-1} X}{1_n^{\mathrm{t}} K^{-1} 1_n}. \tag{3.41}$$

Therefore, $\hat{\mu} = 0$ if

$$1^{\mathrm{t}} K^{-1} X = 0. \tag{3.42}$$

Note that this is the requirement that the realized value of $1^{\mathrm{t}} K^{-1} X$ be zero. In other words we require that $1^{\mathrm{t}} K^{-1} x = 0$. This is equivalent to the requirement that the elements of the tuple $K^{-1} x$ sum to zero.

Both the BLUP and the BLP are in use in machine learning literature. Use of the BLP results in both biased predictions (eq. 2.118, cp. 2.182) and biased confidence intervals for those predictions (eq. 2.120, cp. 2.183). However, the biasedness of the BLP is never acknowledged. In particular is not acknowledged in the textbook by Rasmussen and Williams (2006). The BLP is used exclusively in the astrophysical literature. The consequences of this are not clear. In the case that the assumed mean function (be it zero or nonzero) is well-motivated then the biasedness is presumably not too severe. For example, Gibson et al. (2012) use GPE to fit exoplanetary transit light curves. They assume the mean function to be a transit function of the kind proposed by Mandel and Agol (2002) and determine the parameters of both the mean and covariance functions using maximum likelihood (eq. 3.9). In such a case, the use of the BLP rather than the BLUP is entirely appropriate. When the mean is not well-motivated the consequences are not so predictable. For example, Bower et al. (2010) use GPE to create metamodels of the luminosity functions that are the output of galactic evolution model GALFORM (Bower et al., 2006). Specifically, they use GPE to model the residuals of an ordinary least squares (OLS) regression analysis. They first fit a cubic polynomial to a sample of GALFORM's output using OLS, and then fit the resulting residuals using GPE. If the residuals are normally distributed then GPE is suitable for modelling them. But this is not necessarily the case. Moreover, this two-stage process fails to take advantage of the BLUP's abiliity to predict both mean and residuals at once (eq. 2.182 and following discussion).

3.5 CONDITIONING THE MODEL

We have assumed that the random process, **X** is Gaussian. This assumption may be unwarranted. In this case, the predictor will perfom poorly, and may fail validation. If this is the case, however, we may transform the random process to make it Gaussian. To simplify matters, we will continue to assume that the random process is stationary, and hence has constant variance. The fact that the variance is constant means that elements of the sample, $x_1, x_2, ..., x_n$, are drawn from identical normal distributions. Considered together, we expect them to be distributed normally with mean μ and variance σ^2. If we do not observe this distribution in our training data we may transform it to ensure that we do. Such a transformation is said to be *variance stabilizing* (Bartlett, 1947). Note, however, that althought the elements of the sample $(x_i)_i$ should be drawn from identical Gaussian distibutions, we should not expect to see a full Gaussian distribution in our data. There is no reason the sample spans the mean of this distribution, let alone symmetrically, and may be preferentially drawn from some region of the distribution. This is especially true of unimodal functions, for which the data are dawn entirely from one side of the mean of the random process.

Variance-stabilizing transformations are normally chosen by practical experiment from a number of a number of standard transformations. One such transformation is the *Box–Cox transformation* (Box and Cox, 1964). Let $(x_i)_{i=1}^n$ be a sample of the random process X_i. Then the Box–Cox transformation is the function g such that

$$g(x_i; \lambda_1, \lambda_2) = \frac{(x_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} \tag{3.43}$$

for some real λ_1, λ_2 such that $\lambda_2 > -x_i$ for all i. Note that this expression is just a scaled power law, with the scaling chosen such that $\lim_{\lambda_1 \longrightarrow 0} g(x_i) = \ln(x_i + \lambda_2)$. For $\lambda > 1$ (respectively $\lambda < 1$) the Box–Cox transformation has the effect of compressing (respectively expanding) relatively large values, and expanding (respectively compressing) relatively small values. It is common to round the value of λ_1 to the nearest half-integer, e.g. to use one of the values $2, 1, 1/2, 0, -1/2, -1$, or -2 (the square, identity, square root, logarithm, reciprocal square root, reciprocal, or reciprocal square) which give the transformation a ready interpretation. Frequently, a Box–Cox transformation is powerful enough that a predictor that failed validation for a given data set, may no longer fail validation once that data set has been transformed. In the case of ordinary least-squares regression there is a more rigorous way of choosing the parameter of the Box–Cox transformation, but it is of limited use in the case of GPE.[4]

3.6 COMPUTATIONAL PRACTICALITIES

The expressions for the BLP and BLUP are simple matrix expressions, and very well-suited for evaluation by computer. The only complication in their evaluation is in the inversion of the covariance matrix K. This inverted matrix, K^{-1}, appears once in the expression for the BLP (Eq. 3.1), in the term $k^{\mathrm{t}}K^{-1}(X - m_X)$, and three times in the expression for the ordinary-kriging BLUP (Eqs 3.21), in the terms $k^{\mathrm{t}}K^{-1}(X - \hat{\mu}1_n)$, $1_n^{\mathrm{t}}K^{-1}X$, and $1_n^{\mathrm{t}}K^{-1}1_n$. The explicit inversion of K can be numerically inaccurate, and unduly expensive. Instead of inverting K^{-1} explicitly we therefore introduce the variables α and β, and solve the expressions $K\alpha = 1_n$ and $K\beta = x$. The solutions give values for $K^{-1}1_n$ and $K^{-1}x$ without our having to invert K explicitly. This method is preferrable even when the inverted matrix is being used multiple times, as it is in the case of the BLUP.

One efficient method for solving the system $K\xi = \psi$ (for $\xi = \alpha$ and $\psi = 1_n$ or $\xi = \beta$ and $\psi = x$) is by means of the Cholesky decomposition (Serre, 2002, Thm 8.2.1). Because K is positive-semidefinite there exists a lower-triangular matrix L such that $K = LL^{\mathrm{t}}$. (The same statement is true of positive-definite matrices, in which case L is furthermore unique.) Thus we may write $LL^{\mathrm{t}}\xi = y$. We may then solve for $L^{\mathrm{t}}\xi$ by forwards substitution, and in turn for ξ by backwards substitution. Whereas the decomposition of K has computational complexity order $O(n^3)$, the forward and backward substitutions are of order $O(n^2)$. The factorization must be performed only once of course. We may also use the Cholesky decomposition to compute the determinant of the covariance matrix, which is

$$|K| = |LL^{\mathrm{t}}| \tag{3.44}$$

$$= |L||L^{\mathrm{t}}| \tag{3.45}$$

$$= \left(\prod_{i=1}^{n} L_{ii}\right)^2 \tag{3.46}$$

(where we have used the fact that the determinant of a triangular matrix is equal to the product of its diagonal elements), and has logarithm,

$$\ln(|K|) = 2\sum_{i=1}^{n} \ln(L_{ii}). \tag{3.47}$$

[4] See the paper by Box and Cox (1964) for details.

This latter expression is required for the maximum-likelihood estimate of the parameters of the covariance function (eq. 3.5).

Nonethless, it may still be the case that the matrix K is nearly singular, and this will result in numerical instability in the computation of α and β. Recall that a matrix is singular when its rows are linearly dependent. This occurs in the case of K in the following case. Let $X_j \in X$ be a random variable. Its covariance with all random variables $(X_i)_{i=1}^n$ forms the j-th row of K: $(k(t_j, t_i))_{i=1}^n$. Now let $X_l \in X$, where $m \neq j$, be a different random variable. Its covariance with all random variables $(X_i)_{i=1}^n$ forms the l-th row of K: $(k(t_l, t_i))_{i=1}^n$. If $(k(t_j, t_i))_{i=1}^n$ and $(k(t_l, t_i))_{i=1}^n$ are nearly equal then K is nearly singular. The random variables X_j and X_l may be thought of as being effectively identical. One provides no more information than the other. This occurs when two points are very close together, in the sense that their separation is very much less than the scale length. If the length scale is very large then all points are equally well correlated with all points. We should therefore expect K to be nearly singular, and our predictor to be numerically unstable in the case of very large length scales, and should be alert to this possibility in practice.

Chapter 4

Gaussian-process emulation and galactic modelling

In constructing a distribution function model of a stellar system (Ex. 6), we treat stellar positions, $\boldsymbol{X}$, and velocities, $\boldsymbol{V}$, as random vectors, meaning that the state of a star is represented by the random vector, $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{V})$. The phase-space probability density function for a single star is then denoted $f_{\boldsymbol{W}}$. We assume that the PDF is an element of the model $(f(\cdot\,; \boldsymbol{a}))_{\boldsymbol{a}\in\boldsymbol{A}}$, where the parameter $\boldsymbol{a}$ is a d-dimensional real vector, the elements of which represent the total galactic mass, galactic scale length, velocity anisotropy, etc.

From the phase-space PDF we may calculate the observable properties of the system. For a dSph these observables are typically the projected stellar positions, (represented by the random variables X and Y) and the line-of-sight velocity (represented by the random variable V_z), which we represent by the random vector $\boldsymbol{W}_\mathrm{p} = (X, Y, V_z)$. The PDF for these observables is given by the marginalization of the phase-space PDF:

$$f_{\boldsymbol{W}_\mathrm{p}}(\boldsymbol{w}_\mathrm{p}; \boldsymbol{a}) = \int_{\mathbf{R}^3} f_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{a})\, \mathrm{d}z\, \mathrm{d}v_x\, \mathrm{d}v_y. \tag{4.1}$$

Of course, we do not observe realizations of $\boldsymbol{W}_\mathrm{p}$ directly because there are errors associated with our measurements. Let us represent these errors by the random variables E_X, E_Y, and E_{V_z}, which we assume to be normally distributed with known variances.[1] We may then form the random vector $\boldsymbol{E}_{\boldsymbol{W}_\mathrm{p}} = (E_X, E_Y, E_{V_z})$, which has a joint probability density function (PDF) denoted $f_{\boldsymbol{E}_{\boldsymbol{W}_\mathrm{p}}}$. We therefore observe realizations of the random vector $\boldsymbol{W}'_\mathrm{p} = \boldsymbol{W}_\mathrm{p} + \boldsymbol{E}_{\boldsymbol{W}_\mathrm{p}}$, the PDF of which

[1] Following convention, we have used italic type for the variable, E, and roman type for the expectation operator, E.

is given by

$$f_{W'_{\mathrm{p}}}(\boldsymbol{w}'_{\mathrm{p}}; \boldsymbol{a}) = (f_{W_{\mathrm{p}}} * f_{E_{W_{\mathrm{p}}}})(\boldsymbol{w}'_{\mathrm{p}}; \boldsymbol{a}) \tag{4.2}$$

$$= \int_{\mathbf{R}^3} f_{W_{\mathrm{p}}}(\boldsymbol{w}_{\mathrm{p}}; \boldsymbol{a}) f_{E_{W_{\mathrm{p}}}}(\boldsymbol{w}'_{\mathrm{p}} - \boldsymbol{w}_{\mathrm{p}})\, \mathrm{d}\boldsymbol{w}_{\mathrm{p}}. \tag{4.3}$$

We can reasonably assume that the states of stars are independent and identically distributed. If the errors are also independent and identically distributed then the joint marginalized PDF for N stars is given by

$$f_{(W'_{\mathrm{p},1},\ldots,W'_{\mathrm{p},N})}(\boldsymbol{w}'_{\mathrm{p},1}, \ldots, \boldsymbol{w}'_{\mathrm{p},N}; \boldsymbol{a}) = \prod_{i=1}^{N} f_{W'_{\mathrm{p}}}(\boldsymbol{w}'_{\mathrm{p},i}; \boldsymbol{a}). \tag{4.4}$$

We may then optimize the parameter of the phase-space PDF using the maximum likelihood method. By definition, the likelihood of model parameter $\boldsymbol{a}$ is (Def. 20)

$$L(\boldsymbol{a}; \boldsymbol{w}'_{\mathrm{p},1}, \ldots, \boldsymbol{w}'_{\mathrm{p},N}) = f_{(W'_{\mathrm{p},1},\ldots,W'_{\mathrm{p},N})}(\boldsymbol{w}'_{\mathrm{p},1}, \ldots, \boldsymbol{w}'_{\mathrm{p},N}; \boldsymbol{a}). \tag{4.5}$$

We recover the parameter by maximizing this function for given data, namely the observed values of $\boldsymbol{w}'_{\mathrm{p},1}, \ldots, \boldsymbol{w}'_{\mathrm{p},N}$, aware that the function may have multiple maxima. Even the simplest physically interesting phase-space PDF will fail to have closed-form integrals of the kind required (Eqs 4.1 and 4.3). These integrals will have to be computed numerically, and can be computationally expensive. It is in these cases that GPE can be used to reduce the computational burden. In this chapter, we will illustrate the use of GPE by emulating the likelihood of a Plummer model of a dSph galaxy, which we fit to synthetic data generated using the true PDF.

4.1 EMULATING THE LIKELIHOOD FUNCTIONS

We are interested in proof-of-concept, and will therefore adopt the most straightforward method—simple kriging. We will compute the BLP assuming a Gaussian random process with zero mean, and the squared-exponential covariance function, given by

$$k(\boldsymbol{a}, \boldsymbol{a}') = \sigma_{\mathrm{SE}}^2 \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{a}')^{\mathrm{t}} M (\boldsymbol{a} - \boldsymbol{a}')\right) \tag{4.6}$$

where we assume M is diagonal and positive definite. To accommodate the assumption of a Gaussian random process with zero mean, we first make a Box–Cox transformation of the likelihood and subtract its mean. To accommodate the assumption of stationarity and isotropy that is implicit in using the squared-exponential covariance function, we make a suitable reparameterization of the model. Let us denote the Box–Cox transformation of the likelihood of a parameter, $L(\boldsymbol{a})$, by $Y(\boldsymbol{a})$. We evaluate this quantity for n distinct values of $\boldsymbol{a}$, giving us a vector

$\boldsymbol{y} = (y(\boldsymbol{a}_1), y(\boldsymbol{a}_2), \ldots, y(\boldsymbol{a}_n))$. According to this prescription, the BLP (Ex. 97, Eq. 2.118) of the likelihood of a parameter $\boldsymbol{a}^*$ is

$$\hat{y}(\boldsymbol{a}^*) = k^{\mathrm{t}}(\boldsymbol{a}^*)K^{-1}\boldsymbol{y}, \tag{4.7}$$

with MSE (or variance)

$$\sigma^2(\boldsymbol{a}^*) = \sigma^2_{\mathrm{SE}} - k^{\mathrm{t}}(\boldsymbol{a}^*)K^{-1}k(\boldsymbol{a}^*). \tag{4.8}$$

We call the set $\{(\boldsymbol{a}_i, y(\boldsymbol{a}_i))\}_{i=1}^n$ the *data*, we call the set $(\boldsymbol{a}_i)_{i=1}^n$ the *design* and we call the set $(y(\boldsymbol{a}_i))_{i=1}^n$ the *sample*. Assuming we are able to choose some region of parameter space for which we wish emulate a function, we must chose a *design* (i.e. an arrangement of points in parameter space at which we will compute our sample). If *a priori* we know nothing about our function we wish the design to be *space-filling*, i.e. to have uniform density throughout parameter space. We also wish all projections of the design onto lower-dimensional subspaces to be space-filling, as the model may have low sensitivity to some parameter components. Lattices are a poor choice for such a design, as their size grows exponentially with the dimension of the parameter space. The most commonly used designs satisfying the above space-filling requirements are *Latin hypercubes* (McKay et al., 1979). In Latin hypercube sampling (LHS), a d-dimensional parameter space is partitioned into a d-dimensional hypergrid of n^d cells and n points are placed in these cells such that there is only one point in any hyper-row or hyper-column of cells.

We may optimize the space-filling property of the design by maximizing the minimum separation of pairs of points in all projections of the design onto lower-dimensional subspaces (Santner et al., 2003). LHS designs are restricted to rectangular regions. It is possible to form space-filling designs on nonrectangular regions (e.g. Draguljić et al., 2012) but we do not consider these here.

We choose the size of our design, n, so that the GPE model has acceptable accuracy. This size depends on the difficulty of the problem, i.e. the complexity of the function we are emulating. The more difficult the problem, the greater n will need to be. The question obviously arises: how do we choose an appropriate value of n for a particular problem?

We note that n is satisfactory if the MSE (Eq. 4.8) is small, and that the MSE is a function of $\boldsymbol{M}$ (through the covariance function, k), n (through the size of the matrix $\boldsymbol{K}$), and d (the dimensional of the parameter space). We wish to understand the relationship between these quantities. To this end, Loeppky *et al.* (2008) introduce the *total sensitivity*,

$$\tau := \sum_{i=1}^{d} m_i, \tag{4.9}$$

and the *sparsity*,

$$\psi := \sum_{i=1}^{d} m_i^2, \tag{4.10}$$

where $(m_i)_{i=1}^d$ is the set of elements of the metric matrix. Recall that the length scale l_i is defined such that $m_i = l_i^{-2}$. Consider the squared separation of a pair of design points, $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$, namely $d^2(\boldsymbol{a}_i, \boldsymbol{a}_j)$. For a random LHS design, this separation is the realization of a random variable, H. Loeppky *et al.* show that for such a design H is distributed with mean $\mathrm{E}(H) = \mu(n)\tau$ and variance $\mathrm{var}(H) = \nu(n)\psi$ where μ and ν are weak and strictly decreasing functions of n that converge to a positive constant. The accuracy of our emulator will be good when $\mathrm{E}(H)$ is small (i.e. when the mean separation of sample points is small, and hence mean correlation is good) and when $\mathrm{var}(H)$ is large (i.e. when many pairs of points have separations smaller than the mean and are hence even better correlated).

If we minimize ψ whilst keeping τ constant (i.e. if we minimize ψ subject to the constraint $\sum_{i=1}^d m_i = c$ for some real c) we find that $m_j = c/d$ for all j, i.e. we find that ψ is a minimum (and hence the accuracy poor) when the parameter components are equally active, and hence $\psi = c^2/d$. On the other hand, if we maximize ψ whilst keeping τ constant we find that $m_j = c$ for some j and $m_i = 0$ for $i \neq j$, i.e. we find that ψ is a maximum (and hence the accuracy good) when only one parameter is active, and hence $\psi = c$. For fixed τ, therefore, ψ quantifies the sparsity of the matrix $\boldsymbol{M}$. In the case that all parameter components are equally active it is the case that $\mathrm{E}(H) = dc$ and that $\psi = dc^2$, i.e. that both the mean and the variance of the separation are proportional to the number of parameter components, and that for a sufficiently large number, the accuracy will be poor.

The accuracy of our predictor depends on both the total sensitivity and the sparsity. It does not depend on the total number of parameter components but rather on the number of active parameter components. Suppose that the parameter space has been mapped to a hypercube of side h. Motivated by practical experiment, Loeppky *et al.* propose that if $\tau h^2 = 3$ then the problem is "easy", and if $\tau h^2 = 40$ the problem is "very difficult". If $\tau h^2 = 10$ the the problem will be tractable if ψ is small but intractable if ψ is large. Moreover, for easy problems the convergence of the LOOCV score, R^2 (Eq. 3.25), to zero is fast whereas for difficult problems the convergence is slow. As a rule of thumb, easy problems will have good accuracy for $n = 10d$, whereas difficult problems will require significantly greater n.

Training is therefore best done iteratively. We first take a sample of size $10d$ and validate the emulator. If the model accuracy is poor the covariance function is misspecified, in which case we must use a different covariance function, or an unduly large number of training data. Due to the slow convergence of the MSE in this case (i.e. the case where a sample size of $10d$ is too small), we will need to resample the function in a smaller region of parameter space. If the model accuracy is good, we augment our data. To do this we require some figure of merit for choosing new design points. If we wish to emulate the function faithfully across the region, we might resample at points of maximum variance. If we wish to maximize the function, as we do here, we might use the *expected improvement*. In this case the procedure is known as *efficient global optimization* (Jones et al., 1998).

We follow the presentation of Schonlau and Welch (1996), which we reproduce here in our own notation, for clarity. In the mathematical literature, optimization problems are couched in terms of minimization rather than maximization. We adopt this convention here, understanding of course that we may maximize a function by minimizing its negative.

Suppose that we are performing GPE, and using training data $\{(\boldsymbol{a}_i, y(\boldsymbol{a}_i))\}_{i=1}^n$. The minimum of our sample is $y_{\min} = \min(y(\boldsymbol{a}_i))_{i=1}^n$. We would like to know where to sample in order to improve the accuracy of this minimum. To this end we define the *improvement in the minimum*,

$$I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a})) := \max(y_{\min} - Y(\boldsymbol{a}), 0). \tag{4.11}$$

This is a random variable, the PDF of which is

$$f_{I_Y}(y(\boldsymbol{a})) = \max(y_{\min} - y(\boldsymbol{a}), 0) \tag{4.12}$$

$$= \begin{cases} y_{\min} - y(\boldsymbol{a}) & \text{if } y(\boldsymbol{a}) < y_{\min}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

By definition the expected improvement is

$$\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a}))) = \int_R I_{Y(\boldsymbol{a})}(y(\boldsymbol{a})) f_{Y(\boldsymbol{a})}(y(\boldsymbol{a}))\,\mathrm{d}y(\boldsymbol{a}), \tag{4.14}$$

where $f_{Y(\boldsymbol{a})}$ is the PDF of $Y(\boldsymbol{a})$. In the case of GPE we know that $Y(\boldsymbol{a}) \sim N(\hat{y}(\boldsymbol{a}), \hat{\sigma}^2(\boldsymbol{a}))$, i.e. $f_{Y(\boldsymbol{a})}$ is the normal (i.e. Gaussian) PDF $\varphi(y(\boldsymbol{a}); \hat{y}(\boldsymbol{a}), \hat{\sigma}^2(\boldsymbol{a}))$ (see Eqs 4.7 and 4.8). If $\hat{\sigma}^2(\boldsymbol{a}) = 0$ then the value $y(\boldsymbol{a})$ is known with certainty and we cannot expect any improvement, hence $\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a}))) = 0$. If $\hat{\sigma}^2(\boldsymbol{a}) > 0$ then we may make the change of variables from $y(\boldsymbol{a})$ to $u'(\boldsymbol{a}) = (y(\boldsymbol{a}) - \hat{y}(\boldsymbol{a}))/\hat{\sigma}(\boldsymbol{a})$ to find that the expected improvement is

$$\begin{aligned} &\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a}))) \\ &\quad = \hat{\sigma}(\boldsymbol{a}) \int_{-\infty}^{u(\boldsymbol{a})} (u(\boldsymbol{a}) - u'(\boldsymbol{a}))\varphi(u'(\boldsymbol{a}); 0, 1)\,\mathrm{d}u'(\boldsymbol{a}) \end{aligned} \tag{4.15}$$

Algorithm 1 Efficient global optimization

```
Require: objective function, $y$, sample of objective function, $D := \{(\boldsymbol{a}_i, y(\boldsymbol{a}_i))\}_{i=1}^n$, and stopping threshold, $\varepsilon$.
Ensure: global minimum of objective function.
  $E_{\max} \leftarrow \max(\{\mathrm{E}(I_{Y(\boldsymbol{a})}(y(\boldsymbol{a})) \mid \boldsymbol{a} \in \boldsymbol{A}\}))$
  $\boldsymbol{a}_{\max} \leftarrow \operatorname{argmax}(\{\mathrm{E}(I_{Y(\boldsymbol{a})}(y(\boldsymbol{a})) \mid \boldsymbol{a} \in \boldsymbol{A}\}))$
  while $E_{\max} > \varepsilon$: do
    $D \leftarrow D \cup (\boldsymbol{a}_{\max}, y(\boldsymbol{a}_{\max}))$
    $E_{\max} \leftarrow \max(\{\mathrm{E}(I_{Y(\boldsymbol{a})}(y(\boldsymbol{a})) \mid \boldsymbol{a} \in \boldsymbol{A}\}))$
  end while
  return argmin$(\{y(\boldsymbol{a}_i)\}_{i=1}^n)$.
```

where $u(\boldsymbol{a}) := (y_{\min} - \hat{y}(\boldsymbol{a}))/\hat{\sigma}(\boldsymbol{a})$. Thus,

$$\begin{aligned}&\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a})))\\ &\quad= \begin{cases}\hat{\sigma}(\boldsymbol{a})(u(\boldsymbol{a})\Phi(u(\boldsymbol{a});0,1) + \varphi(u(\boldsymbol{a});0,1)) & \text{if } 0 < \hat{\sigma}(\boldsymbol{a}),\\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{4.16}$$

where $\Phi(u(\boldsymbol{a});0,1)$ is the normal cumulative distribution function. We augment our training data, with the pair $(\boldsymbol{a}_{n+1}, y(\boldsymbol{a}_{n+1}))$ where $\boldsymbol{a}_{n+1} = \operatorname{argmax}(\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a}))))$, and then iterate this procedure until $\mathrm{E}(I_{Y(\boldsymbol{a})}(Y(\boldsymbol{a})))$ is smaller than some threshold, ε. Efficient global optimization (EGO) is implemented by Algorithm 1.

The expected improvement for $0 < \hat{\sigma}^2(\boldsymbol{a})$ is the sum of two terms in $u(\boldsymbol{a})$. The first term dominates if $u(\boldsymbol{a})$ is large, while the second term dominates if $u(\boldsymbol{a})$ is small. For given $\hat{y}(\boldsymbol{a})$ it is the case that $u(\boldsymbol{a})$ is large if $\hat{\sigma}(\boldsymbol{a})$ is small (which will be the case close to design points, including the current minimum) and $u(\boldsymbol{a})$ is small if $\hat{\sigma}(\boldsymbol{x})$ is large (which will be the case away from design points, including the current minimum). The expected improvement is therefore a tradeoff between probable small improvements (near to the current minimum) and improbable large improvements (remote from the current minimum), or between local and global search. The fact that the expected improvement is a trade off between local and global search in this way makes multistart optimization a sensible choice. We may use gradient-based methods as the gradient of the expected improvement has closed form.

The efficient global optimization algorithm reduces the problem of the prohibitively expensive optimization of y to the cheap optimization of the expected improvement. There is a convergence theorem (Vazquez and Bect, 2010) that guarantees that the expected improvement produces a sequence of points that is dense in the parameter space under mild assumptions about the covariance function, so that the result is guaranteed to be a global minimum in the infinite-sample limit. However, we do not know of any theorems concerning the rate of convergence.

We illustrate EGO by reproducing an example given by Jones et al. (1998), namely the minimization of the *Branin function*, a real-valued function of two variables used as a test for optimization. For the sake of completeness, we also produce figures equivalent to theirs. The Branin function is defined by the formula

$$y(a_1, a_2) = \alpha(a_2 - \beta^2 + \gamma a_1 - \delta)^2 + \zeta(1 - \eta)\cos a_1 + \eta, \tag{4.17}$$

where $\alpha = 1$, $\beta = 5.1/(4\pi^2)$, $\gamma = 5/\pi$, $\delta = 6$, $\zeta = 10$, $\eta = 1/(8\pi)$. It has three global minima, at $(a_1, a_2) = (-\pi, 12.275)$, $(\pi, 2.275)$, and $(9.425, 2.475)$ where the function takes the value 0.398. It is evaluated on the domain $a_1 \in [-5, 10]$, $a_2 \in [0, 15]$. We treat the function as a realization of a random process, $\{Y_{\boldsymbol{a}}\}_{\boldsymbol{a} \in \boldsymbol{A}}$, where the parameter space, $\boldsymbol{A} = [-5, 10] \times [0, 15]$. We create a LHS design for the parameter space, namely the set $(\boldsymbol{a}_i)_{i=1}^{n}$, of size $n = 10d = 20$ and evaluate the function, y, at these points, giving the data $\{(\boldsymbol{a}_i, y(\boldsymbol{a}_i))\}_{i=1}^{n}$. The function and the design are plotted in Figure 4.1. We assume a squared-exponential covariance function, $k_{\mathrm{SE}}(\boldsymbol{a}, \boldsymbol{a}') = \sigma_{\mathrm{SE}}^2 \exp(-(\boldsymbol{a} - \boldsymbol{a}')^{\mathrm{t}} \boldsymbol{M}(\boldsymbol{a} - \boldsymbol{a}')/2)$ where $\boldsymbol{M} = \mathrm{diag}(m_1, m_2)$, and then optimize its hyperparameter vector, $(\sigma_{\mathrm{SE}}^2, m_1, m_2)$, using the maximum likelihood method, finding that $\sigma^2 = 36\,500$, $m_1 = 0.0633$, and $m_2 = 0.00580$, or equivalently that $l_1 = 3.98$, $l_2 = 13.1$. For the sake of illustration, we plot the likelihood of the hyperparameter vector in Figure 4.2 as well as the mean and variance of the Gaussian-process predictor for the entire domain in Figure 4.3. We also compute the LOOCV score finding that $\sqrt{R^2} = 2.36$, and the maximum standardized predicted error, finding that the most extreme value of the standardized residuals is e_{-i} is 2.63 (Eq. 3.24). Diagnostic plots are shown in Figure 4.4. We see that the standardized LOOCV residuals are distributed normally and exhibit no trend across the parameter space. The total sensitivity, $\tau h^2 = 14.0$, is moderate, explaining this success. Satisfied that our model is accurate, we then iteratively augment the data using the maximum expected improvement and a stopping criterion of $\varepsilon = 0.001$. For illustration we plot the expected improvement for the first iteration (Figure 4.5) and see that it has three maxima close to the three minima of the Branin function. The algorithm requires a total of eight iterations to find a minimum to an accuracy of 1.3 %. We plot the augmented design in Figure 4.5.

As with any global optimization method, it is sensible to polish the result, which may lack precision. We may do so by resampling the function in the neighbourhood of this result, and again

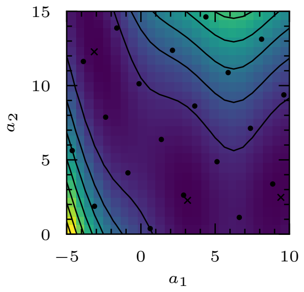

Figure 4.1 The Branin function (Eq. 4.17) and the Latin square design (marked with filled circles) used in its emulation. Arbitrary, equally-spaced contours are shown. Following Jones et al. (1998), we use it to illustrate the methods of Gaussian process emulation (Sacks et al., 1989) and efficient global optimization (Jones et al., 1998). It has three global minima (marked with crosses).

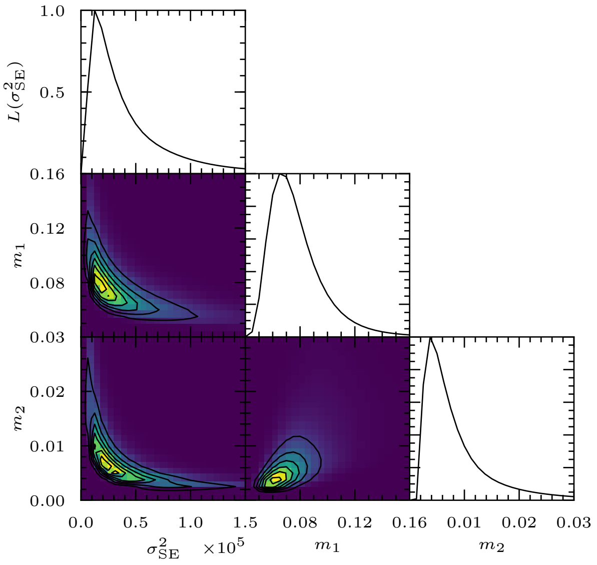

Figure 4.2 The likelihood (Eq. 3.9) of the hyperparameter of the squared-exponential covariance function (Eq. 4.6), used in the emulation of the Branin function (Eq. 4.17). In each panel the marginal likelihood is shown (i.e. the likelihood has been integrated over the parameter components not shown), scaled to the unit interval. Arbitrary, equally-spaced contours are shown. The maximum likelihood is found at $(\sigma^2_{\mathrm{SE}}, m_1, m_2) = (36\,500, 0.0633, 0.00580,)$, i.e. for length scales, $l_1 = 3.98$, and $l_2 = 13.1$.

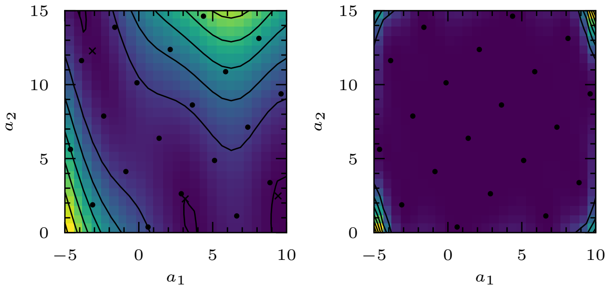

Figure 4.3 The mean (left) and variance (right) of GPE estimate of the Branin function, computed using the squared-exponential covariance function and the design shown (filled circles). In the left panel contours are drawn at the same levels as in Figure 4.1. The variance is high in regions of parameter space that have been poorly sampled, or near the boundary of parameter space, where the predictor is constrained by fewer data than elsewhere.

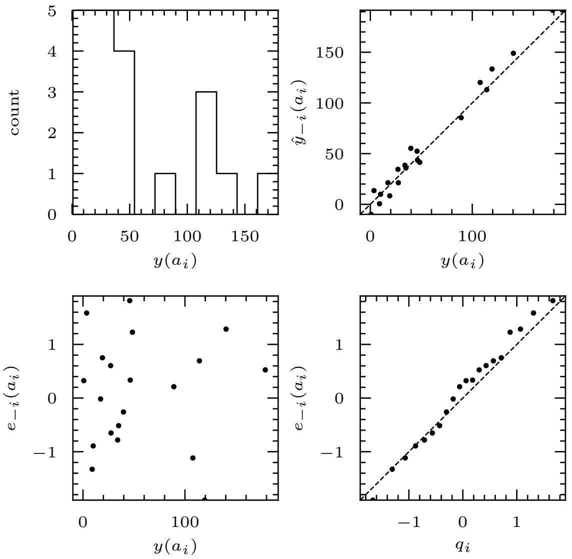

Figure 4.4 Diagnostic plots for the emulation of the Branin function. Top-left: distribution of function values. Top-right: predicted values from a LOOCV analysis against true values. Bottom-left: standardized LOOCV residuals from a LOOCV analysis (Eq. 3.24) against true values. Bottom-right: quantile-quantile plot showing the ordered standardized LOOCV residuals from a LOOCV analysis against the equivalent quantiles of the normal distribution.

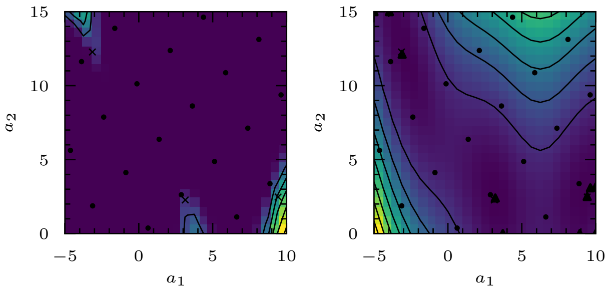

Figure 4.5 The expected improvement in the minimum (left) of our sample of the Branin function, computed using the design shown (filled circles). There are three maxima (marked by crosses), each close to a minimum of the Branin function. The Branin function and the augmented design (marked with filled triangles) determined by the EGO algorithm (right). The new data cluster about the function's three minima. Once the initial design has been computed only eight additional design points are required to find a global minimum with an accuracy of 1.3 %.

performing GPE. The length scales of the GPE fit provides a guide to the size of this neighbourhood. If $\boldsymbol{a}_0 = (a_{0,1}, a_{0,2}, a_{0,3})$ is the result of the EGO algorithm then we resample the function in the region $\boldsymbol{A}' = [a_{0,1} - \delta l_1, a_{0,1} + \delta l_1] \times [a_{0,2} - \delta l_2, a_{0,2} + \delta l_2] \times [a_{0,3} - \delta l_3, a_{0,3} + \delta l_3]$, where (l_0, l_1, l_2) is the vector of length scales found in the final iteration, and δ is some positive real number less than one. Our polished minimum is then the minimum of the GPE predictor. We can find this minimum using gradient-based optimization, or, as the GPE predictor is cheap, by brute force i.e. by searching over a fine lattice of test points covering the whole region $\boldsymbol{A}'$. (This brute force method returns more than the minimum, of course. It maps out the function over the entirety of $\boldsymbol{A}'$. In general the cheapness of the GPE predictor will allow us to do just this. When working in very high-dimensional parameter spaces the inversion the covariance matrix, $\boldsymbol{K}$, may not be so cheap, and we may wish to map out the region using, for example, Markov chain Monte Carlo methods.) There are two additional benefits to this polishing step. First, we may use a high termination threshold, ε, which reduces the number of iterations required by the EGO algorithm. Second, the Hessian of the GPE estimate is available in closed form, and provides an estimate of the Hessian of the function. If the function in question is a likelihood, this allows us to compute an estimate of the Fisher information matrix. The derivative of a Gaussian process is itself a Gaussian process (Adler, 2010):

$$\frac{\partial Y}{\partial \boldsymbol{a}} \sim \left(\frac{\partial \hat{y}}{\partial \boldsymbol{a}}, \frac{\partial^2 k(\boldsymbol{a}, \boldsymbol{a}')}{\partial \boldsymbol{a} \partial \boldsymbol{a}'^{\mathrm{t}}} \right). \tag{4.18}$$

If $H = \partial^2 y / \partial \boldsymbol{a} \partial \boldsymbol{a}^{\mathrm{t}}$ is the Hessian of the function y, then an estimate for the Hessian is

$$\hat{H} = \frac{\partial^2 \hat{y}}{\partial \boldsymbol{a} \partial \boldsymbol{a}^{\mathrm{t}}}. \tag{4.19}$$

We compute the Hessian for the case of the squared-exponential covariance function in Appendix B.

4.1.1 Computational expense

The computational complexity of Equations 4.7 and 4.8 is dominated by the inversion of the matrix $\boldsymbol{K}$, which is of order $O(n^3)$ or better. This inversion may be done indirectly by solving the systems $\boldsymbol{K\alpha} = \boldsymbol{y}$ and $\boldsymbol{K\beta} = \boldsymbol{k}(\boldsymbol{a})$ for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. Moreover, it must be performed only once in each case, regardless of the number of test evaluations required. Once the inversion has been performed, evaluation of the mean involves one matrix-vector multiplication, with complexity $O(n^2)$, followed by one vector-vector multiplication for each evaluation, with complexity $O(n)$. Each evaluation of the variance involves one matrix-vector multiplication and one vector-vector multiplication. Thus the computational complexity is $O(n^3)$. The covariance matrix $\boldsymbol{K}$ must also be inverted for every step in the optimization of the hyperparameter, which must be done at each iteration of the EGO algorithm. Furthermore, it must be computed explicitly for the validation step. Nonetheless, the total computational expense of these inversions is negligible compared with the expense of any interesting astrophysical simulation. The expense of the method is largely in the computation of the training data, and their augmentation when using the EGO method. The initial sampling is trivially parallel, and linear in the dimension of parameter space, but the EGO method is necessarily sequential. (A batch-sequential extension of EGO is available, which makes it possible to perform up to 10 function evaluations at each iteration.) In general we cannot estimate in advance the number of iterations required for EGO without knowing the rate of convergence of the EGO algorithm, and we do not explore this further here.

4.2 PLUMMER MODEL OF A DWARF SPHEROIDAL GALAXY

4.2.1 The anisotropic Plummer sphere

We illustrate the use of GPE for stellar dynamical modelling using the toy model of an anisotropic Plummer sphere of the Osipkov-Merritt type. The relative potential of the Plummer sphere is given by the formula

$$\Psi(r) = \frac{GM}{\sqrt{r^2 + b^2}}, \tag{4.20}$$

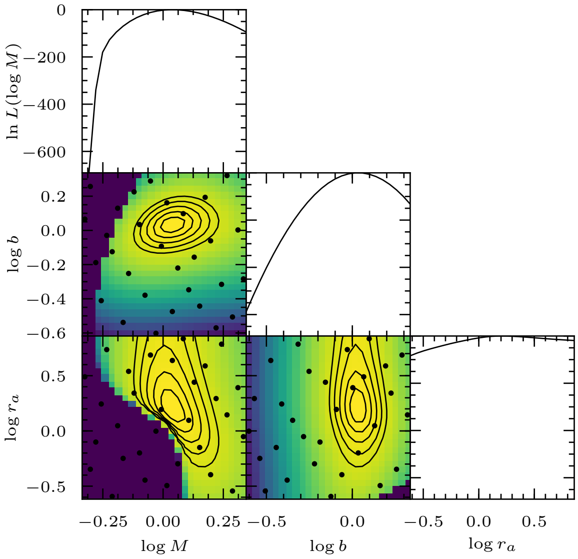

Figure 4.6 The log-marginalized likelihood for the anistropic Plummer model with parameter $(\log(M), \log(b), \log(r_\text{a})) = (0, 0, \log(2)) = (0, 0, 0.301)$ computed using data for 1000 stars generated by the same model (Eq. 4.30). In each panel the likelihood has been marginalized over the unshown parameter components, scaled to the unit interval, and its natural logarithm plotted. The peak therefore has a value of zero. Contours are shown at $-40.650, -27.342, -16.729, -8.740, -3.247$ corresponding to a Gaussian approximation of the likelihood at the 68.268 %, 95.449 %, 99.730 %, 99.993 %, and 99.999 % levels. The darkest regions have likelihoods of zero. The maximum likelihood is found at $\log M = 0.0327$, $\log b = -0.0213$, $\log r_\text{a} = 0.305$. We do not emulate this function directly, but rather its Box-Cox transformation $\ln(L + \lambda)$, where λ is an arbitrary small constant, here taken to be the smallest non-zero element of our sample of L. The design used in the emulation of L is shown with filled circles.

and its density given by the formula

$$\rho(r) = \frac{3M}{4\pi b^3}\left(1 + \frac{r^2}{b^2}\right)^{-5/2} \tag{4.21}$$

where M is the galactic mass, r the radius, and b the galactic scale length (Plummer, 1911). Let us consider an isotropic system described by this potential-denisty pair. The phase-space PDF may be expressed as a function of relative energy only Binney and Tremaine (2008).

Following Osipkov (1979) and Merritt (1985), we define the variable $Q = \mathscr{E} - L^2/2r_\mathrm{a}^2$ where $\mathscr{E}$ is the relative energy, L is the magnitude of the angular momentum, and r_a is the anisotropy radius. By use of Eddington's inversion formula we then find that the phase-space DF for a star may be expressed as a function of Q:

$$f_Q(Q) = \frac{3Mb^2}{\pi^3\sqrt{2}r_\mathrm{a}^2}\left(\frac{16(r_\mathrm{a}^2 - b^2)}{7}Q^{7/2} + (GM)^2Q^{3/2}\right). \tag{4.22}$$

The marginalized DF is given by the integral of f_Q with respect to the line-of-sight position and proper-motion velocities. If we define the parameter vector $\boldsymbol{a} = (M, b, r_\mathrm{a})$ and work in cylindrical coordinates with the z-axis parallel to the line of sight, this PDF is

$$f_{(R_\mathrm{p},V_z)}(r_\mathrm{p}, v_z; \boldsymbol{a}) = 2\pi\int_R\int_R\int_R f_Q(Q)\,\mathrm{d}v_{r_\mathrm{p}}v_\varphi\,\mathrm{d}z. \tag{4.23}$$

The inner double integral may be computed analytically using the method given by Carollo et al. (1995):

$$\int_R\int_R f_Q(Q)\,\mathrm{d}v_{r_\mathrm{p}}\,\mathrm{d}v_\varphi = \begin{cases} 2\pi g(r, r_\mathrm{p})F(Q_\mathrm{max}) & \text{if } 0 < Q_\mathrm{max}, \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

where

$$g(r_{r,\mathrm{p}}) := \frac{a^2}{\sqrt{(r_\mathrm{a}^2 + r^2)(r_\mathrm{a}^2 + r^2 - r_\mathrm{p}^2)}}, \tag{4.25}$$

$$F(Q) := \frac{6b^2}{\pi^3\sqrt{2}r_\mathrm{a}^2(GM)^5}\left(\frac{16(r_\mathrm{a}^2 - b^2)}{63}Q^{9/2} + \frac{(GM)^2}{5}Q^{5/2}\right), \tag{4.26}$$

and

$$Q_\mathrm{max}(r_\mathrm{p}, z, v_z) = \Psi(r) - \left(\frac{r_\mathrm{a}^2 + r^2}{r_\mathrm{a}^2 + r^2 - r_\mathrm{p}^2}\right)\frac{v_z^2}{2}. \tag{4.27}$$

However, the outer integral in Equation 4.23 must be computed numerically.

We may account for observational errors using Equation 4.3:

$$f_{(R'_\mathrm{p},V'_z)}(r'_\mathrm{p}, v'_z; \boldsymbol{a}) = (f_{(R_\mathrm{p},V_z)} * f_{(E_{R_\mathrm{p}},E_{V_z})})(\boldsymbol{w}'_\mathrm{p}; \boldsymbol{a}), \tag{4.28}$$

where $f_{(E_{R_\mathrm{p}},E_{V_z})}$ is the joint PDF for the errors on our measurements. Using Equation 4.4 we may form the joint marginalized PDF for a galaxy of N stars:

$$\begin{aligned} & f_{((R_{\mathrm{p},1},V_{z,1}),\ldots,(R_{\mathrm{p},N},V_{z,N})}(r_{\mathrm{p},1},v_{z,1},\ldots,r_{\mathrm{p},N},v_{z,N};\boldsymbol{a}) \\ & \quad = \prod_{i=1}^{N} f_{(R_\mathrm{p},V_z)}(r_{\mathrm{p},i},v_{z,i};\boldsymbol{a}). \end{aligned} \tag{4.29}$$

We recover the parameter, $\boldsymbol{a}$, by maximizing the likelihood

$$\begin{aligned} & L(\boldsymbol{a};r_{\mathrm{p},1},v_{z,2},\ldots,r_{\mathrm{p},N},v_{z,N}) \\ & \quad = f_{((R_{\mathrm{p},1},V_{z,1}),\ldots,(R_{\mathrm{p},N},V_{z,N}))}(r_{\mathrm{p},1},v_{z,1},\ldots,r_{\mathrm{p},N},v_{z,N};\boldsymbol{a}). \end{aligned} \tag{4.30}$$

We will assume that the error is zero, as this presents the maximally difficult case for GPE for the following reason. If the error is zero then there exist regions of parameter space for which the likelihood is also zero (these are the regions of parameter space for which the line-of-sight velocity exceeds the local escape velocity). In this case it impossible to transform the sample so that is normally distributed (Section 3.5). The squared-exponential is therefore necessarily misspecified and the emulator necessarily underperforms. Despite this misspecifation, it is possible to successfully emulate the likelihood. The inclusion of errors would only improve the emulator's performance by reducing the mean square error.

4.2.2 Optimization of the likelihood

We work in mass units of $10^9\mathrm{M}_\odot$, and distance units of kpc (meaning that the gravitational constant is $G = 4.302 \times 10^3 \mathrm{kpc}\mathrm{M}_\odot^{-1}\mathrm{km}^2\mathrm{s}^{-2}$). We use synthetic data generated using the same model and parameter $(M, b, r_\mathrm{a}) = (1, 1, 2)$. The data consist of sky positions and line-of-sight velocities for 1000 stars, each with zero error. In the case of the anistropic Plummer model the likelihood is cheaply computed, and is shown in Figure 4.6.

It happens to be unimodal and approximately Gaussian, but note that in general this is not the case. Suppose that the likelihood were not cheaply computed. In this case we would proceed as follows. First we choose the region of parameter space on which we wish to emulate. By the virial theorem we know that $3\langle v_z^2\rangle = GM_\mathrm{virial}/r_\mathrm{g}$ where $\langle v_z^2\rangle$ is the line-of-sight velocity dispersion, M_virial is the virial mass, and r_g is the gravitational radius (Binney and Tremaine, 2008). We may approximate the gravitational radius by $r_{1/2}/0.45$ (Binney and Tremaine, 2008), where $r_{1/2}$ is the half-light radius, and note that for the Plummer sphere, $r_{1/2} = b/\sqrt{2^{2/3} - 1}$. We might estimate the true value of M to be within a factor of three either side of M_virial. Similarly, we might estimate

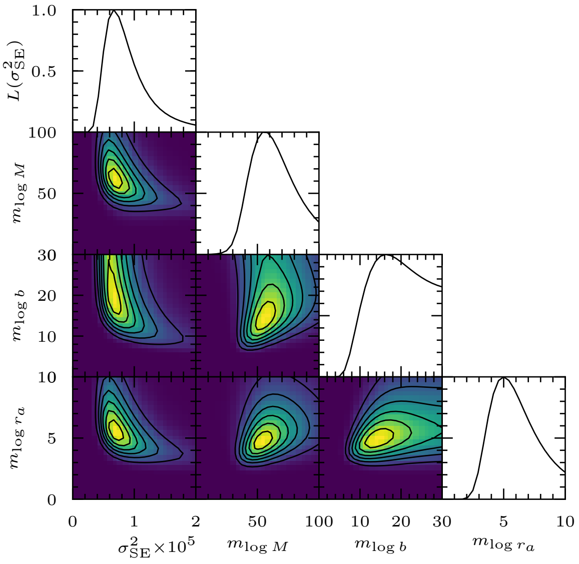

Figure 4.7 The likelihood (Eq. 3.9) of the hyperparameter of the squared-exponential covariance function (Eq. 4.6), used in the emulation of the Plummer-model likelihood (Eq. 4.30). In each panel the marginal likelihood is shown (i.e. the likelihood has been integrated over the parameter components not shown), scaled to the unit interval. Arbitrary, equally-spaced contours in linear space are shown. The maximum likelihood is found at $(\sigma^2_{\mathrm{SE}}, m_{\log M}, m_{\log b}, m_{\log r_\mathrm{a}}) = (81\,000, 58.0, 14.5, 5.10)$, i.e. for length scales, $l_{\log M} = 0.131$, $l_{\log b} = 0.262$, and $l_{\log r_\mathrm{a}} = 0.443$.

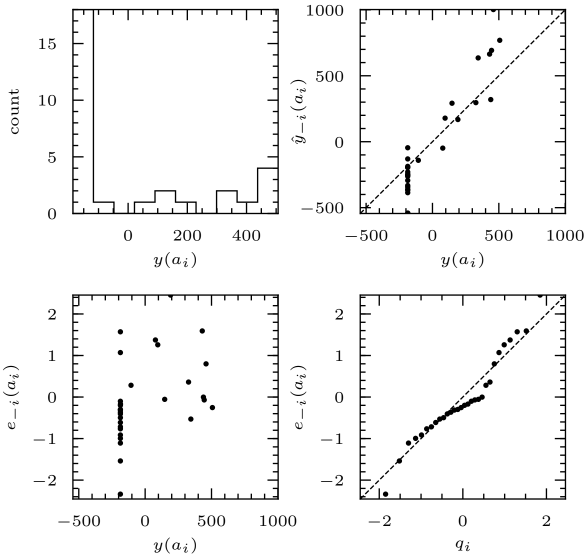

Figure 4.8 Diagnostic plots for the emulation of the transformed Plummer model likelihood. Top-left: distribution of likelihood values. Top-right: predicted values from a LOOCV analysis against true values. Bottom-left: standardized LOOCV residuals from a LOOCV analysis (Eq. 3.24) against true values. Bottom-right: quantile-quantile plot showing the ordered standardized LOOCV residuals from a LOOCV analysis against the equivalent quantiles of the normal distribution.

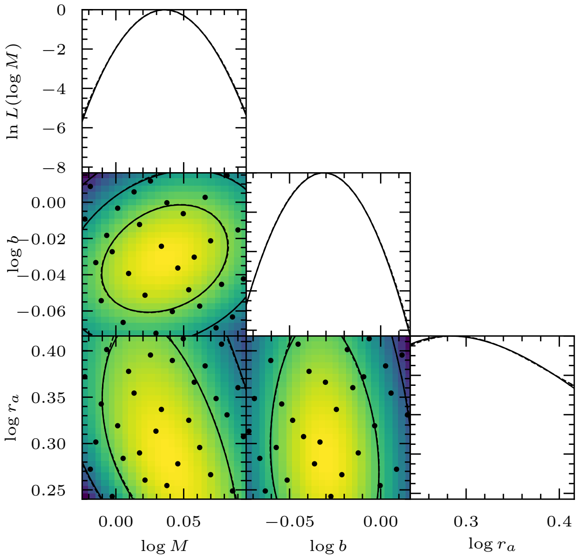

Figure 4.9 The log-marginalized likelihood for the anisotropic Plummer model evaluated on the region of parameter space $\boldsymbol{X}' = [-0.0249, 0.0961] \times [-0.0738, 0.0162] \times [0.240, 0.416]$ (dashed line), and the mean of the GPE estimate (solid line). The likelihood is resampled on this region in order to polish the maximum-likelihood estimate of the parameter vector. Although both the true log-marginalized likelihood and its GPE estimate are plotted they are barely distinguishable.

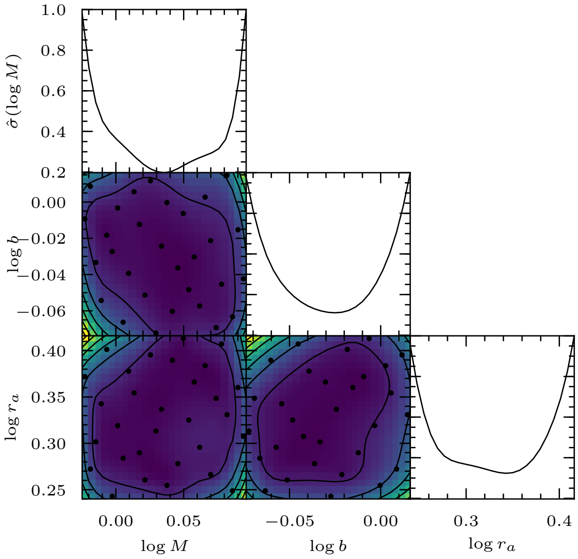

Figure 4.10 The marginalized standard deviation of the GPE estimate of the log-likelihood shown in Figure 4.9. In each panel the likelihood has been marginalized over the unshown parameter components and scaled to the unit interval. Arbitrary, equally-spaced contours are shown.

the true value of b to be a factor of three either side of its estimate, and that the true value of r_a to be within an order of magnitude either side of its estimate.

However, in the case that the observed data have no error, the feasible region is bounded below by the curve $Q_\mathrm{max} = 0$ (the minimum value Q can take). For a given projected radius, the maximum line-of-sight velocity is therefore set by the condition

$$v_z^2 = \frac{2\Psi(r)(r_\mathrm{a}^2 + r^2 - r_\mathrm{p}^2)}{r_\mathrm{a}^2 + r^2} \tag{4.31}$$

where $r_\mathrm{p} \leq r$. We must maximize this expression. The maximum value, $(v_z)_\mathrm{max}$, occurs at a radius determined by the equation

$$\left.\frac{\mathrm{d}v_z}{\mathrm{d}r}\right|_{r=r_\mathrm{max}} = 0 \tag{4.32}$$

which, if it exists, is unique, or (if this equation has no solution on account of the constraint $r \geq r_\mathrm{p}$) at a radius $r = r_\mathrm{p}$.

In the isotropic limit, $r_\mathrm{a} = \infty$, we have $v_z = v_{z,\mathrm{max}}$ at $r = r_\mathrm{p}$, and therefore

$$\frac{M^2}{(v_z^2 r_\mathrm{p}/2G)^2} - \frac{b^2}{r_\mathrm{p}^2} = 1, \tag{4.33}$$

a hyperbola in M and b. Each pair (r_p, v_z) defines such a hyperbola. In the anisotropic case, Equation 4.32 gives

$$r_\mathrm{max}^2 = \frac{(3r_\mathrm{p}^2 - 2r_\mathrm{a}^2) - r_\mathrm{p}\sqrt{9r_\mathrm{p}^2 - 8r_\mathrm{a}^2 + 8b^2}}{2} \tag{4.34}$$

if the discriminant and numerator are real and nonnegative, i.e. if

$$r_\mathrm{p} \geq \frac{r_\mathrm{a}^2}{\sqrt{r_\mathrm{a}^2 + 2b^2}}, \text{ and} \tag{4.35}$$

$$r_\mathrm{a} \geq b. \tag{4.36}$$

Otherwise, $r_\mathrm{max} = r_\mathrm{p}$. In the point-mass limit, $b = 0$, and upon substituting Equation 4.34 into 4.31 we find an equation in M and r_a. Again, each pair (r_p, v_z) defines such an equation. A given parameter vector is forbidden if the observed line-of-sight velocity of any star is greater than this maximum allowed velocity.

For our data we find that $\langle v_z^2 \rangle = 504$ km^2 s^{-2}, and $r_{1/2} = 0.944$ kpc. Thus $b = 0.723$ kpc and $M_\mathrm{virial} = 0.737 \times 10^9$ M$_\odot$. The total mass, M, is bounded below by the maximum value of $v_z^2 r_\mathrm{p}/2G$, namely 0.461×10^9 M$_\odot$. We therefore choose to emulate on the region of parameter space $\boldsymbol{A} = [0.461, 2.21] \times [0.241, 2.17] \times [0.241, 7.23]$. We make a logarithmic transformation of the parameter space (according to the prescription given in section 3.5), mapping a parameter vector $\boldsymbol{a} = (M, b, r_\mathrm{a})$ to $\boldsymbol{x} = (\log M, \log b, \log r_\mathrm{a})$. The transformed parameter space is $\boldsymbol{X} = [-0.336, 0.345] \times [-0.618, 0.337] \times [-0.618, 0.860]$. We also transform the likelihood (again according to the prescription given in section 3.5) from $L(\boldsymbol{a})$ to $\ln(L_{\boldsymbol{X}}(\boldsymbol{x}) + \varepsilon)$ where $\varepsilon = \min(L_{\boldsymbol{X}}(\boldsymbol{x}_i))_{i=1}^{n}$. We sample this transformed likelihood using a LHS design of size $n = 10d = 30$ giving the training data $\{(\boldsymbol{x}_i, \ln(L_{\boldsymbol{X}}(\boldsymbol{x}_i) + \varepsilon))\}_{i=1}^{n}$. We then optimize the model hyperparameter vector, $(\sigma_\mathrm{SE}^2, m_{\log M}, m_{\log b}, m_{\log r_\mathrm{a}})$, using the maximum-likelihood method (Ch. 3 Sec. 3.1), finding that $\sigma_\mathrm{SE}^2 = 81\,000$, $m_{\log M} = 58.0$, $m_{\log b} = 14.5$, $m_{\log r_\mathrm{a}} = 5.10$, or equivalently that $l_{\log M} = 0.131$, $l_{\log b} = 0.262$, and $l_{\log r_\mathrm{a}} = 0.443$. We find that the LOOCV score is $\sqrt{R^2} = 69.3$, and that the extreme value of the standardized LOOCV residuals is 2.33. Diagnostic plots (Figure 4.8) show that the standardized LOOCV residuals are distributed normally and show no trend across the parameter space. The results of the validation are acceptable, meaning that we may proceed to maximize the transformed likelihood using EGO. Using a stopping threshold of $\varepsilon = 0.001$, the EGO algorithm requires 33 iterations to find the maximum at $\log M = 0.0356$, $\log b = -0.0288$, and $\log r_\mathrm{a} = 0.328$. At the last iteration the maximum LOO-likelihood estimate of the hyperparameter vector is $\boldsymbol{\theta} = (203\,000, 17.0, 30.8, 8.10)$, meaning that the length scales are $l_{\log M} = 0.242$, $l_{\log b} = 0.180$, and $l_{\log r_\mathrm{a}} = 0.351$. We then polish this result by resampling the likelihood in its neighbourhood, and again performing GPE. We choose the region that is within one quarter of a length scale in each element of the parameter vector, namely $\boldsymbol{X}' = [-0.0249, 0.0961] \times [-0.0738, 0.0162] \times [0.240, 0.416]$. We again transform our sample of the likelihood, finding that the most-appropriate transformation is to $\ln L(\boldsymbol{x})$, where no offset is required as the likelihood is everywhere nonzero in this new region of parameter space. The maximum LOO-likelihood estimate of the hyperparameter vector is $\hat{\boldsymbol{\theta}} = (436, 24.1, 28.3, 1.42)$, meaning that the length scales are $l_{\log M} = 0.204$, $l_{\log b} = 0.188$, and $l_{\log r_\mathrm{a}} = 0.839$. We find the maximum at $\log M = 0.0327$, $\log b = -0.0213$, and $\log r_\mathrm{a} = 0.305$. In Figure 4.9 we plot the log-likelihood and its GPE estimate for the region $\boldsymbol{X}'$, and in Figure 4.10 we plot the variance of this estimate.

We note that for this three-dimensional model we have recovered the MLE with approximately 100 evaluations of the likelihood. The first and last sets of 30 evaluations may each be made in parallel, effectively reducing this number to many fewer than 100. Batch-sequential EGO, would reduce the effective number of runs still further.

The total sensitivity in the initial step of emulation is $\tau = 47.9$, indicating that this problem is very difficult. In explaining this we note that the likelihood is very sharply peaked. Another way of putting this is to say that it has multiple length scales (the function is highly sensitive to changes in the parameter vector around its maximum, but insensitive to such changes away from its maximum). The squared-exponential covariance function, which assumes a single set of length scales is thus grossly misspecified. The sharpness of this peak is due to several factors: (1) that our data are drawn from the same model we are fitting, (2) that the dimension of our parameter space is small, and (3) that there is no error associated with our synthetic observations. The dynamical model is well-specified and its parameter tightly constrained by the data. The problem of multiple length-scales persists even in the transformed data, to which we see an approximation in Figure 4.6. In this case there is a sharp cliff on the boundary of the permitted and forbidden regions of parameter space. Such forbidden regions exist only for data with zero errors. We thus expect the task of fitting this perfectly specified, low-dimensional toy model to perfect data to be the maximally difficult case for emulation. We expect it to be considerably harder than the task of fitting more sophisticated models to imperfect data, the likelihoods of which will be less sharply peaked, and for which forbidden regions of parameter space do not exist.

4.2.3 Confidence region

The Hessian of the log-likelihood is available to us as a consequence of the polishing step (Eq. 4.19). Hence, we may compute an estimate of the Fisher information matrix, $\boldsymbol{I}$ (Def. 25), without further evaluation of the dynamical model. However, our predictor $\hat{y}$ is for the log-likelihood, $\ln L(\boldsymbol{x})$, expressed as a function of the transformed parameters, $\boldsymbol{x}$. Thus, the first derivative is

$$\frac{\partial \ln L}{\partial a_j} = \frac{\partial \hat{y}}{\partial a_j} \tag{4.37}$$

$$= \frac{\partial \hat{y}}{\partial x_j}\frac{\partial x_j}{\partial a_j}, \tag{4.38}$$

and the second derivative is

$$\frac{\partial^2 \ln L}{\partial a_i \partial a_j} = \frac{\partial^2 \hat{y}}{\partial a_i \partial a_j} \tag{4.39}$$

$$= \frac{\partial^2 \hat{y}}{\partial x_i \partial x_j}\frac{\partial x_i}{\partial a_i}\frac{\partial x_j}{\partial a_j} + \frac{\partial \hat{y}}{\partial x_j}\frac{\partial^2 x_j}{\partial a_i \partial a_j}, \tag{4.40}$$

where the second term vanishes at the maximum. Given that $x_i = \log a_i$ we have that

$$\frac{\partial x_i}{\partial a_i} = \frac{1}{a_i \ln 10}, \text{ and} \tag{4.41}$$

$$\frac{\partial^2 x_i}{\partial a_i \partial a_j} = -\frac{1}{a_i^2 \ln 10}\delta_{ij}. \tag{4.42}$$

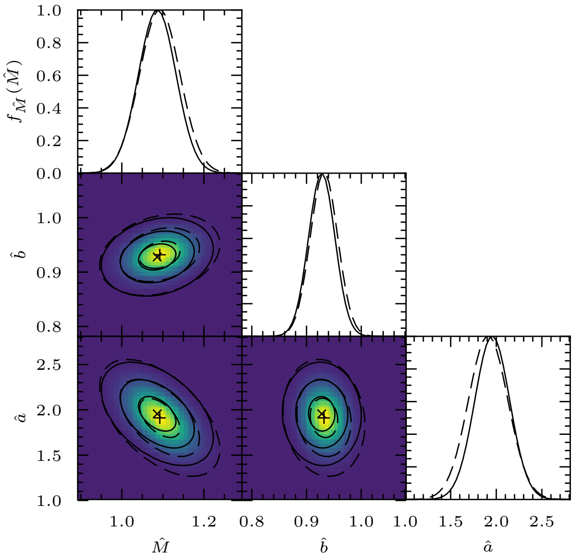

Figure 4.11 The 68 %, 95 % and 99.7 % confidence regions for the maximum-likelihood estimate of the Plummer-model parameter, (M, b, r_a), computed using the Fisher information matrix (dashed line), and the mean of its GPE estimate (solid line).

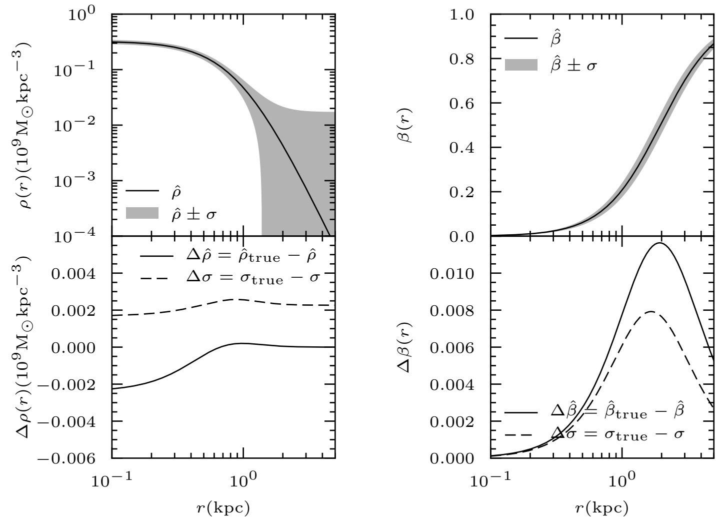

Figure 4.12 The maximum-likelihood estimates of the Plummer-model density and anisotropy parameter computed using GPE (top panels), together with the deviation of these estimates from their true values (bottom panels).

In Figure 4.11 we plot the confidence regions for the maximum-likelihood estimate of the parameter. The galactic mass, *M*, and scale length, *b*, are well constrained but the anistropy parameter, r_{a}, is less so. This is as we would expect. For a self-consistent model of this kind the mass and extent of a galaxy are functions of one another through Poisson's equation. All of the observational data therefore contain information about the *M* and *b*. In the anisotropic case, however, there is an additional length scale, r_{a}, which we must determine using data at radii greater than this value. Stars at smaller radii do not constrain the length scale, meaning that only a subset of our data contain information about it.

Given the distribution of the MLE for the parameter vector we may also compute the distribution of the MLE for the density and for Binney's anisotropy parameter using Proposition 30. The density is given by Equation 4.49 and hence the MLE for the density at radius *r* is

$$\hat{P}(r) \sim N(\rho(r; \hat{\boldsymbol{a}}), \sigma_\rho^2), \tag{4.43}$$

where

$$\sigma_\rho^2 = \left(\frac{\partial\rho(r; \hat{\boldsymbol{a}})}{\partial\boldsymbol{a}}\right)^{\mathrm{t}} \boldsymbol{I}^{-1}(\hat{\boldsymbol{a}}) \frac{\partial\rho(r; \hat{\boldsymbol{a}})}{\partial\boldsymbol{a}}. \tag{4.44}$$

For an Ossipkov-Merritt model, Binney's anisotropy parameter (Binney and Tremaine, 2008),

$$\beta(r) = \frac{1}{1 + r_\mathrm{a}^2/r^2}. \tag{4.45}$$

Hence, the MLE for Binney's anisotropy parameter,

$$\hat{B}(r) \sim N(\beta(r; \hat{\boldsymbol{a}}), \sigma_\beta^2), \tag{4.46}$$

where

$$\sigma_\beta^2 = \left(\frac{\partial\beta(r; \hat{\boldsymbol{a}})}{\partial\boldsymbol{a}}\right)^\mathrm{t} \boldsymbol{I}^{-1}(\hat{\boldsymbol{a}}) \frac{\partial\beta(r; \hat{\boldsymbol{a}})}{\partial\boldsymbol{a}}. \tag{4.47}$$

We plot the distributions of these quantities in Figure 4.12. These are the principal results of our work.

In the upper-left panel of these plots we plot the GPE predictions for the density, $\hat{\rho}$, and the GPE prediction for its one-sigma confidence region, σ_ρ. In the lower-left panel we plot the difference of the GPE predictions for the density, and the true maximum-likelihood values, which we denote $\Delta\hat{\rho} = \rho_\mathrm{true} - \hat{\rho}$. In this panel we also plot the difference of the GPE predictions for the confidence region and the true maximum-likelihood confidence region, which we denote $\Delta\sigma_\rho = \sigma_{\rho,\mathrm{true}} - \sigma_\rho$. In the upper-right panel of these plots we plot the GPE predictions for the velocity anisotropy, $\hat{\beta}$, and the GPE prediction for its one-sigma confidence region, σ_β. In the lower-right panel we plot the difference of the GPE predictions for the velocity anisotropy, and the true maximum-likelihood values, which we denote $\Delta\hat{\beta} = \beta_\mathrm{true} - \hat{\beta}$. Again, we also plot the difference of the GPE predictions for the confidence region and the true maximum-likelihood confidence region, which we denote $\Delta\sigma_\beta = \sigma_{\beta_\mathrm{true}} - \sigma_\beta$.

Both the GPE predictions for density and velocity anisotropy as well as the GPE predictions for their confidence regions are excellent. In the case of density, the error in the predicted values $\Delta\hat{\rho}$ is at greatest less than 0.3 %. The error in the predicted confidence region $\Delta\sigma_\rho$ is similarly at greatest less than 0.3 %. In the case of velocity anisotropy, the error in the predicted values $\Delta\hat{\beta}$ is at greatest less than 1.2 %. The error in the predicted confidence region $\Delta\sigma_\rho$ is at greatest 0.8 %.

4.2.4 Contamination of data

We have assumed throughout that our model of the data is perfect: that every star in our sample is known to be a bound member of the galaxy in question, and that this galaxy is well described by a Plummer model. In practice we do not know that our model is perfect. It may fail in two significant ways: our model of the galaxy may be misspecified, or our observations may include

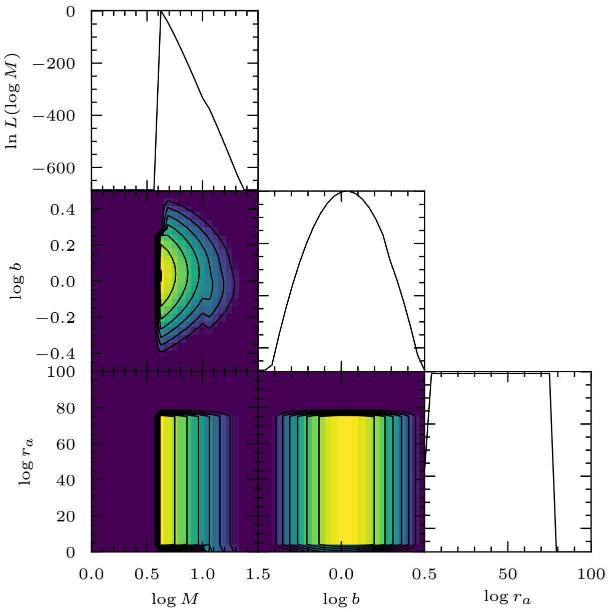

Figure 4.13 The log-marginalized likelihood for the anistropic Plummer model with parameter $(\log(M), \log(b), \log(r_\mathrm{a})) = (0, 0, \log(2)) = (0, 0, 0.301)$ computed using data for 1000 stars generated by the same model (Eq. 4.30) and an additional 10 % contaminating stars. Contours are the same as those shown in Fig. 4.6.

stars that are not bound members of the galaxy. Our data may be contaminated by foreground and background stars. In fact our data will *necessarily* be contaminated in this way. This can bias the results of our maximum-likelihood anaylsis, and we must be alert to this fact.

To help understand the problem of contamination let us recall the way in which stellar catalogues of dSphs are created. First, we identify red giants that are candidates for belonging to the dSph in question. This is done using stellar photometry to produce a color-magnitude diagram for the appropriate portion of the sky. Stars from the red giant branch are identified, and their spectra then taken. This allows us to compute their line-of-sight velocities. The resulting sample is contaminated by foreground and background stars belonging principally to the Milky Way's stellar halo. The stars of the halo and the dwarf are kinematically distinct, forming two distinct populations, each with a different mean velocity and velocity dispersion. Typically the halo stars have a small mean velocity and broad velocity dispersion, while the dSph stars have high mean velocity and narrow velocity dispersion. This reflects the fact that the dSphs are tightly bound systems, undergoing mean orbital motion around the Milky Way. Typically, the majority of these contaminating stars are eliminated from our sample kinematically, using *sigma clipping* (e.g. Kleyna et al., 2004): a Gaussian is fitted to dSph velocity dispersion, and all stars with velicities greater than, say, 3σ from the mean are rejected. The remaining stars form our catalogue. The sigma-clipping method is crude, and alternatives are available (Walker, Mateo, Olszewski, Sen and Woodroofe, 2009 use a more sophisticated statistical method that takes advantage of both the kinematic and metallicity information provided by the stellar spectra). However, no method will completely eliminate contamination of our catalogue. We can only reduce it to an acceptable level.

Contamination modifies both the velocity distribution and spatial distribution of stars. The resulting sample is no longer a likely realization of our model and as a consequence the maximum-likelihood estimate of the parameter will be biased. In particular the contamination increases the velocity dispersion of our sample, inflating the wings of the velocity distribution. In particular, the maximum velocity of stars in our sample increases. In order to allow for such increased velocities the maximum-likelihood estimate of the dSph mass will have to be greater than its true value. We have noted that, in the case that our data have zero errors, regions of parameter space are forbidden due to any single star having a line-of-sight velocity greater than the local escape velocity for the parameters in these regions. Contaminating stars, being unbound, may exceed the local escape velocity for a given parameter, causing that parameter incorrectly to have zero likelihood. Thus, contaminating stars also modify the forbidden region of parameter space. In general we expect the forbidden region to be more extensive than it in fact is. Additional regions of parameter space are incorrectly deemed impossible.

In the Osipkov-Merritt model, stars are isotropic in the inner part of the dSph (i.e. inside the anisotropy radius, r_a), and radially anisotropic in the outer parts (i.e. outside the anisotropy radius, r_a). This means that line-of-sight velocities decrease with radius across the face of the galaxy. (Of course the speed of stars must also decrease with radius as their potential increases and they become decreasingly well-bound.) However, the line-of-sight velocities of the contaminating stars do not vary with radius in this way. Instead, their distribution is independent of position. Contamination has the effect of reducing the velocity anisotropy in the outer parts. We therefore expect the maximum-likelihood estimate of the anisotropy radius to also increase.

The effect of contamination on the maximum-likelihood estimate of the length-scale of the galaxy will be less dramatic. We have increased the surface density by a constant uniformly across the extent of the galaxy. The fall-off in density, which the scale-length quantifies, is therefore unaffected. Thus we expect the maximum-likelihood estimate of the length-scale to increase, but only marginally.

To illustrate these effects, we add 100 contaminating stars to our synthetic data set of 1000 stars drawn from the Plummer model. We then recompute the likelihood. We choose our contaminating stars to be drawn from uniform distributions in both space and velocity. We assume that stars are distributed uniformly within a radius of 5 kpc, and that line-of-sight velocities are distributed uniformly between -100 km s^{-1} and 100 km s^{-1}. This choice is consistent with the maximum observed speeds of halo stars (Helmi, 2008). The resulting sample is akin to that of the Fornax dSph, which has mean line-of-sight velocities approximately equal to those of the contaminating population (Walker, Mateo, Olszewski, Sen and Woodroofe, 2009). The new log-likelihood is plotted in Figure 4.13.

As we expect, the maximum-likelihood estimate of the galactic mass is significantly increased: from $\log M = 0.0327$ to $\log M = 0.625$. This is also true of the velocity anisotropy radius, which increases from $\log r_\mathrm{a} = 0.305$ to $\log r_\mathrm{a} = 54.1667$. This very large anisotropy radius is effectively infinite, corresponding to an isotropic rather than anisotropic system. The scale-length, however, changes only marginally, from $\log b = -0.0213$ to $\log b = 0.04167$. The addition of contamination has therefore resulted in the system being incorrectly identified as isotropic, and as having a mass nearly four times greater than its true value. It is worth noting that the log-likelihood is very far from being quadratic (Prop. 27), meaning that we are not in the asymptotic limit where likelihood methods are reliable . This itself alerts us to the fact that either we have insufficient

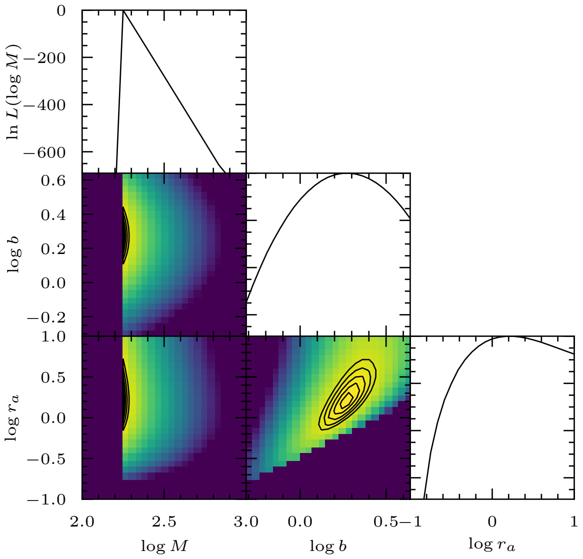

Figure 4.14 The log-marginalized likelihood for the anistropic Plummer model computed using data for 1000 stars generated by the isotropic Hernquist model (Eq. 4.50). The maximum likelihood is found at $\log M = 2.25$, $\log b = 0.2825$, $\log r_\mathrm{a} = 0.25$. Contours are the same as those shown in Fig. 4.6.

data, or our model is unsuitable for the data. In practice contaminating stars are likely to account for less than 10 % of our observations (a few percent is more realistic.) We have introduced severe contamination of our data to illustrate the dominanting effect it may have on our analysis, and to emphasize the importance of clean kinematic data or appropriate modelling.

4.2.4.1 Model misspecification

It is also worth briefly considering the effect of misspecification of the galaxy model itself, independently of the question of data contamination. Suppose, for the sake of illustration, that we fit our Osipkov-Merritt Plummer model to data drawn from an altogether different galaxy model. For the illustration we choose to generate data from an isotropic Hernquist model (Hernquist, 1990). The potential and density of the Hernquist model are given by

$$\Psi(r) = \frac{GM}{r+b}, \tag{4.48}$$

and

$$\rho(r) = \frac{M}{2\pi b^3}\left(\frac{r}{b}\right)^{-1}\left(1+\left(\frac{r}{b}\right)\right)^{-3} \tag{4.49}$$

where again, M is the galactic mass, r the radius, and b the galactic scale length. By use of Eddington's inversion formula we may find a phase-space PDF with isotropic velocity distribution. This is given by (Hernquist, 1990; Baes and Dejonghe, 2002)

$$f_{\mathscr{E}}(\mathscr{E}) = \frac{1}{8\sqrt{2}\pi^3}\left(\frac{\sqrt{\mathscr{E}}(1-2\mathscr{E})(8\mathscr{E}^2-8\mathscr{E}-3)}{(1-\mathscr{E})^2} + \frac{2\arcsin(\sqrt{\mathscr{E}})}{(1-\mathscr{E})^{5/2}}\right). \tag{4.50}$$

Using this PDF, we generate sky-positions and line-of-sight velocities for 1000 stars using a galactic mass of $M = 10^9$ and length-scale $b = 1$ k pc.[2] Again, we assume zero errors on our data.

We note that in the inner part of the galaxy, the Hernquist density is $\rho(r) \sim r^{-1}$, and that in the outer part of the galaxy, the Hernquist density is $\rho(r) \sim r^{-4}$. The differs from the Plummer density, which is $\rho(r) =$ const. in the inner part, and is $\rho(r) \sim r^{-5}$ in the outer part. We also note that the Hernquist model results in line-of-sight velocity distributions with lower dispersion than for the Plummer model. We might therefore expect the maximum-likelihood estimate of the mass to decrease. However, this effect is dominated by the fact that the central density of a Hernquist model is higher than that of Plummer model. By fitting a Plummer model to this data, we require an inflated *total mass* to satisfy this excess central density. The maximum-likelihood estimate of the galactic mass will therefore again be greater than the true value. Again, the forbidden region of parameter space will be too great.

In an isotropic system, this increased mass would increase the velocity dispersion, by allowing for greater escape velocities. However, as we have noted, the velocity dispersion of the Hernquist model is smaller than that of the Plummer model. By reducing the isotropy, however, we may reduce the velocity dispersion. By making the system anisotropic, we assume that the measured line-of-sight velocity at that large radii is only a small component of the total velocity. By overestimating the total velocity of stars in the outer part of the galaxy in this way, we reduce the velocity dispersion of the galaxy as a whole.

We plot the likelihood of the Plummer-model parameters using data from the Hernquist model in Figure 4.14. As we expect, the maximum-likelihood estimate of the galactic mass is too great. In this case, $\log M = 2.25$, meaning that the mass is overestimated by nearly two orders of magnitude. The scale-length, $\log b = 0.2825$, is nearly twice as great as its true value. The velocity anisotropy radius, $\log r_{\mathrm{a}} = 0.25$, is much too small, meaning that the galaxy has been incorrectly

[2] In fact these data have kindly been generated by Walter Dehnen, using his code MKsphere.

identified as anisotropic rather than isotropic. Again, it is worth noting that the log-likelihood is far from being quadratic, indicating that likelihood methods have failed in this instance.

The toy model we have used in this chapter is deliberately simple. Moreover, our assumption of perfect data drawn from the same model has been driven by by the fact that this perfection results in sharply peaked likelihoods, with multiple length-scales, and that this is the maximally difficult case for GPE. We wished to know that, as our data and modelling improve the method of GPE will not fail. The question of model misspecification is distinct from the question of GPE's usefulness. The method of GPE is agnostic about the quality of the model data it is trained on. Indeed it is the need for more-sophisticated modelling, with its additional computational expense, that drives the need for increased computational efficiency.

Chapter 5

Gaussian-process emulation and the generalized Hernquist model

We would like to consider more general distribution-function models of dwarf spheroidal galaxies than the Plummer model we have considered so far. To construct distribution functions of this kind we may specify a potential-density pair and then solve the integral equation

$$\rho(\boldsymbol{x};\boldsymbol{a}) = \int_{\mathbf{R}^3} f(\boldsymbol{x},\boldsymbol{v};\boldsymbol{a})\,\mathrm{d}\boldsymbol{v} \tag{5.1}$$

for $f(\boldsymbol{x},\boldsymbol{v};\boldsymbol{a})$. If the density is spherically symmetric then the phase-space PDF, f, depends only on the energy and angular momentum (Binney and Tremaine, 2008). Moreover, if the system is spherically symmetric in all its properties then the phase-space PDF depends only on the energy and magnitude of the angular momentum. We write this PDF $f_{(E,L)}$. To preserve the distinction between random variables and their realized values, we write E and L for the random variables representing energy and the magnitude of angular momentum, and e and l for their realized values. Thus, $f_{(E,L)}$ takes values $f_{(E,L)}(e,l)$ for all $e,l \geq 0$. For convenience, we work with relative potential, $\Psi := -\Phi + \Phi_0$, and relative energy, $\varepsilon := e + \Phi_0$, where Φ_0 is chosen such that the relative energy is always positive. We denote the random variable associated with relative energy by $\mathscr{E}$, and the PDF $f_{(\mathscr{E},L)}$.

5.1 FACTORIZING THE PHASE-SPACE PROBABILITY DENSITY FUNCTION

By definition of conditional probability it is the case that

$$f_{(\mathscr{E},L)}(\varepsilon,l) = f_{L\,|\,\mathscr{E}}(l\,|\,\varepsilon) f_{\mathscr{E}}(\varepsilon). \tag{5.2}$$

where the function $f_{\mathscr{E}}$ is the PDF for the relative energy, and the function $f_{L \mid \mathscr{E}}$ is the conditional PDF for the magnitude of the angular momentum at a given relative energy. If the distribution of the magnitude of angular momentum is constant then the distribution function depends only on the energy. In this case we may use Eddington's method to solve the density integral (Eq. 5.1). If this is not the case, then a number of methods are available for performing this inversion. We will do this by constructing a model for the PDF $f_{L \mid \mathscr{E}}$, and then numerically solving the density integral using the method of Cuddeford and Louis (1995). To construct our model of $f_{L \mid \mathscr{E}}$ we follow Gerhard (1991) by introducing the function x given by

$$x(l, \varepsilon) = \frac{l}{l_0 + l_\mathrm{c}(\varepsilon)} \tag{5.3}$$

where $l_0 \in \mathbf{R}_{\geq 0}$ is the *angular momentum constant*, and where $l_\mathrm{c}(\varepsilon)$ is the angular momentum of a circular orbit of relative energy ε. We then introduce the *circularity function h* which gives the function $g = h \circ x$. We assume that $f_{L \mid \mathscr{E}}$ is an element of the family of such functions.

Note that $l \in [0, l_\mathrm{c}(\varepsilon)]$ meaning that $x(l, \varepsilon) \in [0, l_\mathrm{c}(\varepsilon)/(l_0 + l_\mathrm{c}(\varepsilon))]$, with x able to take the maximum value of one only if $l_0 = 0$. If $x = 0$ for a given star, then that star is on a plunge orbit, and if $x = x(l_\mathrm{c}(\varepsilon), \varepsilon)$ then that star is on a circular orbit. If x takes a value between 0 and 1 for a given star then it is on an elliptical orbits. The circularity function therefore suppresses or enhances orbits of a particular type, thus controlling the velocity anisotropy at a given radius. Decreasing circularity functions suppress circular orbits, while increasing circularity functions suppress plunge orbits. The constant circularity function gives an isotropic distribution of velocities.

We may think of l_0 as a the angular momentum of the equivalent circular orbit of some characteristic energy, which in turn defines a characteristic radius, which we call the 'anisotropy radius'. The most tightly bound stars have energies approximately equal to the potential at the centre of the system. Therefore $l < l_\mathrm{c}(e) \ll l_0$, and $x \approx 0$. Hence, the velocity distribution is isotropic in the core. The most loosely bound stars, have energies approximately equal to zero. Therefore $l \approx l_\mathrm{c}(e) \gg l_0$. Hence, the velocity distribution is anisotropic outside the anisotropy radius.

5.1.1 Solving the density integral

We must now specify the potential-density pair, for which we use the generalized Hernquist model (or α-β-γ model), proposed by Zhao (1996) as a model for spherical galaxies and galactic bulges. The density is given by

$$\rho(r; \rho_0, \alpha, \beta, \gamma) = \rho_0 \left(\frac{r}{b}\right)^{-\gamma} \left(1 + \left(\frac{r}{b}\right)^{\alpha}\right)^{(\gamma-\beta)/\alpha} \tag{5.4}$$

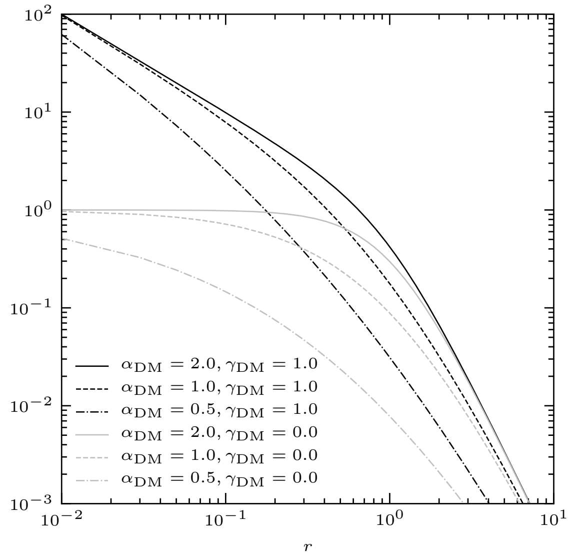

Figure 5.1 The unnormalized generalized Hernquist density, ρ/ρ_0, for outer log-slope $\beta = 3$, and inner log-slope $\gamma = 0$, and 1, and sharpness $\alpha = 0.5, 1$, and 2.

where $\alpha, b > 0$ and $\beta, \gamma \geq 0$, and the normalizing constant is

$$\rho_0 = \frac{M}{4\pi\alpha \mathrm{B}(\alpha(3-\gamma), \alpha(\beta-3), 1)}. \tag{5.5}$$

Here, B is the *incomplete Beta function*, given by

$$\mathrm{B}(a, b, x) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t. \tag{5.6}$$

The constant b is called the *scale length.* It is a split-power law, with gradients $-\gamma$ and $-\beta$ in the small- and large-radius limits. The parameter α determines the width of the transition between these two regimes. The greater α, the greater the width of the transition. We therefore call γ the *inner log-slope*, β the *outer log-slope*, and α the *sharpness.* We plot the density for a range of parameter values in Figure 5.1. If the inner log-slope is zero, then the density is approximately constant at small radii, and the galaxy is said to exhibit a *core.* If, however, the inner log-slope is greater than one, then the density diverges at small radii, and the galaxy is said to exhibit a *cusp.* Our principal interest is in distinguishing between cored and cusped dSphs, and hence in resolving the core-cusp problem (Ch. 1).

Zhao (1996) shows that the relative potential associated with the generalized Hernquist density model is given by

$$\Psi(r;\rho_0,b,\alpha,\beta,\gamma) = 4\pi\rho(r;\rho_0,b,\alpha,\beta,\gamma)f_{0,0}(r;\alpha,\beta,\gamma)r^2, \tag{5.7}$$

where

$$\begin{aligned} f_{0,0}(r;\alpha,\beta,\gamma) = {} & \frac{\alpha\mathrm{B}(\alpha(3-\gamma),\alpha(\beta-3),(1+r^\alpha)^{-1})}{(1+r^\alpha)^{-\alpha(3-\gamma)}(1-(1+r^\alpha)^{-1})^{\alpha(\beta-3)}} + \\ & \frac{\alpha\mathrm{B}(\alpha(\beta-2),\alpha(-\gamma+2),1-(1+r^\alpha)^{-1})}{(1-(1+r^\alpha)^{-1})^{\alpha(\beta-2)}(1+r^\alpha)^{-\alpha(-\gamma+2)}}. \end{aligned} \tag{5.8}$$

Using the generalized Hernquist density, we construct a model of a dwarf spheroidal galaxy that consists of a dark-matter halo and a single stellar population, which makes no contribution to the galactic potential. The stars act as a tracers of the dark-matter potential. The dark-matter halo and stellar population both have generalized Hernquist density profiles, given by

$$\rho_i(r) = \rho(r;\rho_{0,i},\alpha_i,\beta_i,\gamma_i) \tag{5.9}$$

for $i = \mathrm{DM}, *$. The galactic relative potential is therefore given by

$$\Psi(r) = \Psi_{\mathrm{DM}}(r) \tag{5.10}$$

$$= \Psi(r;\rho_{0,\mathrm{DM}},b_{\mathrm{DM}},\alpha_{\mathrm{DM}},\beta_{\mathrm{DM}},\gamma_{\mathrm{DM}}). \tag{5.11}$$

We will write $\rho_{0,\mathrm{DM}}$ as ρ_0 for convenience. The stellar density is therefore given by

$$\rho_*(r) = \int_{\mathbf{R}^3} f_{(\mathscr{E},L)}(\varepsilon,l)\,\mathrm{d}\boldsymbol{v} \tag{5.12}$$

where the relative energy is

$$\varepsilon = \Psi_{\mathrm{DM}} - \frac{1}{2}v^2. \tag{5.13}$$

This integral equation may be solved using the method described by Cuddeford and Louis (1995), as follows. First, we perform a change of variables to find that

$$\rho_*(\Psi) = 4\sqrt{2}\pi\int_0^\Psi\int_0^{\pi/2}\sqrt{\Psi-\varepsilon}f_{\mathscr{E}}(\varepsilon)f_{L\,|\,\mathscr{E}}(l_{\max}\sin(\theta),\varepsilon)\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\varepsilon \tag{5.14}$$

where $l_{\max} = \sqrt{2(\Psi-\varepsilon)}$.We then make the the definition

$$K(\Psi,\varepsilon) := \int_0^{\pi/2} f_{L|\mathscr{E}}(l_{\max}\sin(\theta),\varepsilon)\sin(\theta)\,\mathrm{d}\theta \tag{5.15}$$

and rewrite the density integral as

$$\rho_*(\Psi) = 4\sqrt{2}\pi\int_0^\Psi\sqrt{\Psi-\varepsilon}K(\Psi,\varepsilon)f_{\mathscr{E}}(\varepsilon)\,\mathrm{d}\varepsilon. \tag{5.16}$$

We also rewrite the interval $[0, \Psi)$ as the union of n intervals of width h:

$$\bigcup_{i=1}^{n} [(i-1)h, ih). \tag{5.17}$$

The density at $\Psi = jh$ is then given by

$$\rho_*(jh) = 4\sqrt{2}\pi \sum_{i=1}^{j} \int_{(i-1)h}^{jh} \sqrt{jh - \varepsilon} K(jh, \varepsilon) f_{\mathscr{E}}(\varepsilon) \, \mathrm{d}\varepsilon. \tag{5.18}$$

This expression is still exact. We have only manipulated expression 5.12. However, we continue to follow Cuddeford and Louis (1995) by now making some simplifying assumptions. Specifically, we assume that f and K are slowly varying over any subinterval, and may be replaced by their midpoint values. Let f_i and K_i be these midpoint values, i.e. let

$$f_i := f((1 - 1/2)h) \tag{5.19}$$

$$K_{ij} := K(jh, (1 - 1/2)h). \tag{5.20}$$

Furthermore, we make the definition

$$I_{ij} := \int_{(i-1)h}^{ih} \sqrt{jh - \varepsilon} \, \mathrm{d}\varepsilon. \tag{5.21}$$

Under these assumptions, the density at $\Psi = jh$ is now given by

$$\rho_*(jh) = 4\sqrt{2}\pi \sum_{i=1}^{j} K_{ij} f_{\mathscr{E},i} I_{ij}. \tag{5.22}$$

We may now write $f_{\mathscr{E},i}$ as a recurrence relation by observing that

$$\rho_*(jh) = 4\sqrt{2}\pi \left(K_{jj} f_{\mathscr{E},j} I_{jj} + \sum_{i=1}^{j-1} K_{ij} f_{\mathscr{E},i} I_{ij} \right) \tag{5.23}$$

and that hence

$$f_{\mathscr{E},j} = \frac{\rho_*(jh)/(4\sqrt{2}\pi) - \sum_{i=1}^{j-1} K_{ij} f_{\mathscr{E},i} I_{ij}}{K_{jj} I_{jj}}. \tag{5.24}$$

The term $f_{\mathscr{E},n}$ is then the solution to the density equation (eq. 5.16). To find this explicitly we observe that the first term of the recurrence relation is

$$f_{\mathscr{E},1} = \frac{\rho_*(h)}{4\sqrt{2}\pi K(h, h/2) I_{11}}, \tag{5.25}$$

and that

$$I_{ij} = -\frac{2}{3}(jh)^{3/2}((1 - i/j)^{3/2} - (1 - (i-1)/j)^{3/2}), \tag{5.26}$$

which gives us that

$$I_{11} = \frac{2}{3}h^{3/2}. \tag{5.27}$$

Having inverted the integral in this way we are left needing to compute the observable quantities. The joint PDF for a single star is given by

$$f_{R_\mathrm{p},V} = \int_\mathbf{R}\int_\mathbf{R}\int_\mathbf{R} f_{\mathscr{E}\,|\,L}(\varepsilon, l) f_{\mathscr{E}}(\varepsilon)\,\mathrm{d}z\,\mathrm{d}v_x\,\mathrm{d}v_y \tag{5.28}$$

where $l(\boldsymbol{r},\boldsymbol{v}) = |\boldsymbol{r}\times\boldsymbol{v}|$, and $\varepsilon(\boldsymbol{r},\boldsymbol{v}) = \Psi(\boldsymbol{r}) - \frac{1}{2}\boldsymbol{v}^2$ where $\boldsymbol{r} = (x, y, z)$. As we have discussed in Chapter 4 (eq. 4.3) this PDF must be convolved with the PDF for velocity errors, which we assume to be Gaussian with known variance. The resulting integral is then performed numerically, using Gaussian quadrature. Given data consisting of sky positions and line-of-sight velocities for n stars, $D = \left((r_{\mathrm{p},i}, v_{z,i})\right)_{i=1}^{n}$, we may compute the support of a given parameter, $S(\boldsymbol{\alpha}) := \ln(L(\boldsymbol{\alpha}))$ by taking the product of these integrals. This is implemented by Mark Wilkinson's C programme, HernDF.

We would like to recover the the model parameter by maximizing this likelihood. However, in this high-dimensional parameter space direct maximization of the likelihood is entirely impractical. Even maximization of an MCMC sample, which we might expect to require several hundred thousand sequential evaluations of the likelihood, is impractical. Instead, we explore the potential of GPE for the maximization of the likelihood. We test our method on synthetic data, leaving the use of observational data, specifically that for the Fornax dwarf, for future work.

5.2 GAIA CHALLENGE

The Gaia challenge test suite is a database of synthetic kinematic data released to accompany the 'Gaia challenge' workshop held at the University of Surrey, 19–23 August 2014, with attendants 'invited to apply their favourite methods to these mock data to recover the underlying gravitational potential and/or phase space distribution function.'

The Gaia Challenge data is drawn from a model that consists of a dark-matter halo and a single stellar population, which does not contribute to the galactic potential. Like our model, the density of both halo and stars are is modelled by generalized Hernquist model. The PDF for each stellar component is constructed according to the method of Osipkov (1979) and Merritt (1985), resulting in a model parameterized by a tuple of 10 real numbers,

$$\boldsymbol{\alpha} = (\rho_{0,\mathrm{DM}}, b_\mathrm{DM}, \alpha_\mathrm{DM}, \beta_\mathrm{DM}, \gamma_\mathrm{DM}, b_*, \alpha_*, \beta_*, \gamma_*, r_{\mathrm{a},*}), \tag{5.29}$$

where $r_{\mathrm{a},*}$ is the Osipkov–Merritt anisotropy radius. Sky positions and line-of-sight velocities are given for 32 distinct parameter values, as follows. The dark-matter parameter components are set to $\alpha_{\mathrm{DM}} = 1$, $\beta_{\mathrm{DM}} = 3$, and $\gamma_{\mathrm{DM}} = 0$ or 1, with $b_{\mathrm{DM}} = 1$ k pc and $\rho_{0,\mathrm{DM}} = 0.064\ \mathrm{M}_\odot\,\mathrm{pc}^{-3}$ in the case that $\gamma_{\mathrm{DM}} = 1$ and $\rho_{0,\mathrm{DM}} = 0.4\ \mathrm{M}_\odot\,\mathrm{pc}^{-3}$ in the case that $\gamma_{\mathrm{DM}} = 0$. The stellar parameter components are set to $\alpha_* = 2$, $\beta_* = 5$, and $\gamma_* = 0.1$ or 1, with $b_*/b_{\mathrm{DM}} = 0.1, 0.25, 0.5$, or 1, and $r_{\mathrm{a},*} = 1$ or ∞ k pc. (An infinite Osipkov–Merritt anisotropy radius, $r_{\mathrm{a},*} = \infty$ k pc, corresponds to an everywhere isotropic velocity distribution.) We consider the case of $\rho_{0,\mathrm{DM}} = 0.064\ \mathrm{M}_\odot\,\mathrm{pc}^{-3}$, $b_{\mathrm{DM}} = 1$ k pc, $\alpha_{\mathrm{DM}} = 2$, $\beta_{\mathrm{DM}} = 3.5$, $\gamma_{\mathrm{DM}} = 1$, $b_* = 0.25$ k pc, $\alpha_* = 1$, $\beta_* = 5$, $\gamma_* = 0.1$, $r_{\mathrm{a},*} = \infty$ k pc. This describes a Plummer-type stellar population with an isotropic velocity distribution, embedded within a cusped dark-matter halo.

We use a sample size of $n = 10\,000$, as recommended by the Gaia Challenge, although sample sizes of order 1000 are more realistic. The Gaia Challenge data consist of stellar radii with zero errors, and line-of-sight velocities with a median error of approximately 2 km s^{-1}. The error associated with these line-of-sight velocities is consistent with real data (Walker, Mateo, Olszewski, Sen and Woodroofe, 2009). However, in practice we do know the stellar radii exactly. Although sky-positions have very small errors compared to those of velocity, and may be neglected, there is non-neglible error associated with stellar radii due to the fact that we must estimate the centre of the galaxy. The Gaia Challenge ignores this fact, and we will too.

We here restrict ourselves to exploring the likelihood for the dark-matter components of the parameter, and fix the stellar components of the parameter to their true values. We also assume that we know the dSph to be istropic, and hence use a circularity function that is constantly one. For convenience, we transform the parameter components ρ_0, b_{DM}, and b_* to $\log(\rho_0/10^9\mathrm{M}_\odot)$, $\log(b_{\mathrm{DM}}/1\ \mathrm{kpc})$, and $\log(b_*/1\ \mathrm{kpc})$. The true values of these transformed parameter components are then $\log(\rho_0/10^9\mathrm{M}_\odot) = 7.8062$, $\log(b_{\mathrm{DM}}/1\ \mathrm{kpc}) = 0$, and $\log(b_*/1\ \mathrm{kpc}) = -0.6021$. Our model of the phase-space PDF is parameterized by a tuple of five real numbers,

$$\boldsymbol{\alpha} = (\rho_{0,\mathrm{DM}}, b_{\mathrm{DM}}, \alpha_{\mathrm{DM}}, \beta_{\mathrm{DM}}, \gamma_{\mathrm{DM}}). \tag{5.30}$$

We choose to emulate the support within the bounds shown in Table 5.1. Simple mass estimators may always be used to constrain the galactic mass (and hence the normalizing density) to within a factor three, as we have discussed in Chapter 4 (Sec. 4.2.2). In fact our range spans an order of one magnitude, centred on the true normalizing density, $\log(\rho_0/10^9 M_\odot) = 7.8$. The outer-slope, β_{DM}, takes a minimum value of 3.5 (determined by the computational necessities of HernDF), and includes both the Hernquist and Plummer density models (corresponding to

component	interval	
	min.	max.
$\log(\rho_0/10^9 M_\odot)$	7.5	8.5
b_{DM}	0.6	5
α_{DM}	0.5	2
β_{DM}	3.5	5
γ_{DM}	0	1.5

Table 5.1 Bounds on the dark-matter parameter components used in the emulation of the support.

$\beta_{\mathrm{DM}} = 4$ and 5 respectively). The inner log-slope ranges between 0 and 1.5, thus spanning the interesting cases of 0 and 1. The sharpness, α, ranges over values that allow for very steep ($\alpha = 2$) and very gradual ($\alpha = 0.5$) transitions between the inner and outer density profiles. HernDF takes 6.5 s to evaluate the support for a parameter given a stellar sample size of $n = 10\,000$, meaning that it is just feasible to evaluate the support for all five dark-matter components using a coarse lattice of 10^5 points.

We perform GPE of the support by sampling it using a LHS design and squared-exponential covariance function, and generate predictions using the same lattice of parameter values used to evaluate the true support. We require that the results of GPE pass LOOCV cross validation (Ch. 3, Sec. 3.3) with good precision. Typically, we might require that $\sqrt{R^2}/\,\mathrm{range}(\boldsymbol{y}) < 0.1$, but for our purposes would like that $\sqrt{R^2}/\,\mathrm{range}(\boldsymbol{y}) < 0.01$. Morever, we require that nowhere is the absolute value of the residual unduly large. In this case, the fact that we have a lattice of support values allows us to also perform out-of-sample validation of our GPE predictions. We have seen (Ch. 4, Sec. 4.1) that the rule of thumb is that a sample size of $n = 10d$ should produce acceptable results. However, there is no reason that this rule should hold in our case. The likelihood, which might be unimodal and sharply peaked, is somewhat different from the functions usually considered in the GPE literature. We therefore perform GPE for a range of sample sizes. In fact, we evaluate and emulate three-, and four-dimensional subspaces also. In each case, we have fixed the remaining parameter components to their true values.

Three dimensions. Figure 5.2 shows the log-likelihood for the inner log-slope, normalizing

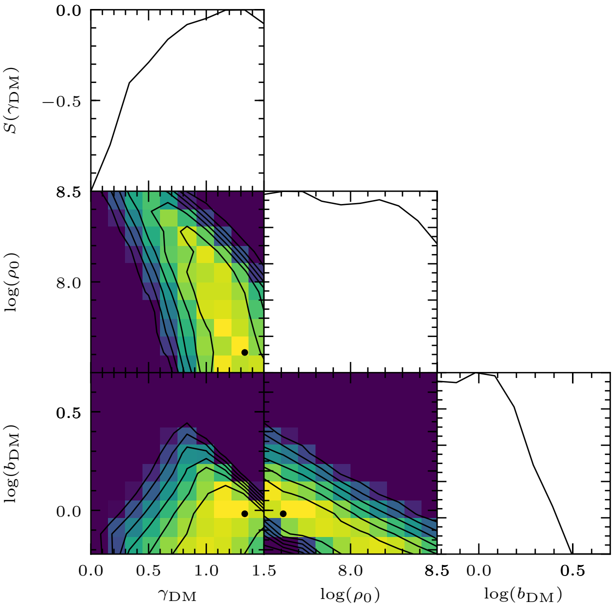

Figure 5.2 Support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, and $\log(b_{\rm DM})$ of the parameter of the generalized Hernquist model. A total of 10^3 evaluations of the likelihood have been required. The unshown parameter components have been set to their true values.

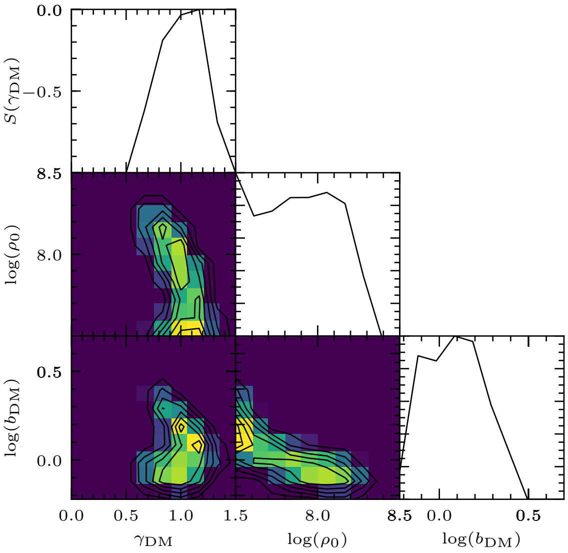

Figure 5.3 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, and $\log(b_{\rm DM})$ of the parameter of the generalized Hernquist model (Fig. 5.2), sample size $n = 10d = 30$.

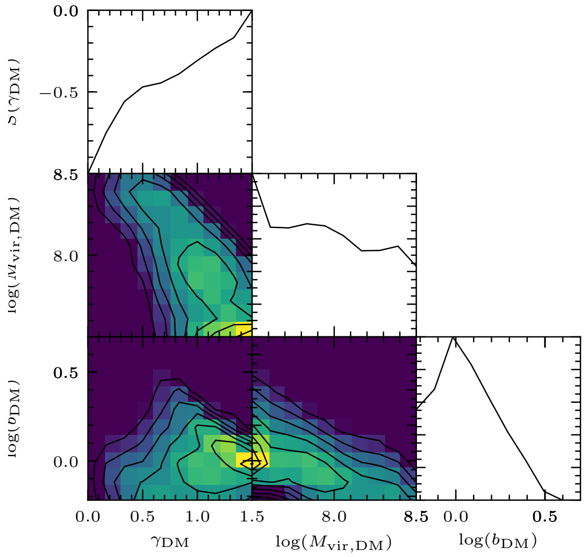

Figure 5.4 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, and $\log(b_{\rm DM})$ of the parameter of the generalized Hernquist model (Fig. 5.2), sample size $n = 20d = 60$.

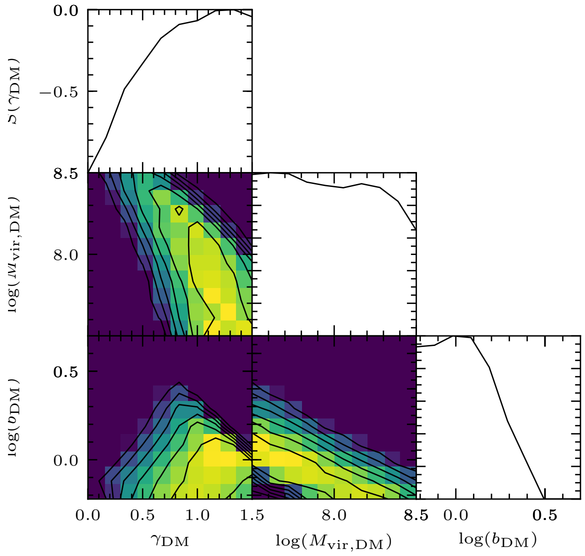

Figure 5.5 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, and $\log(b_{\rm DM})$ of the parameter of the generalized Hernquist model (Fig. 5.2), sample size $n = 60d = 180$.

sample size	μ	σ^2	$m_{\gamma_{\mathrm{DM}}}$	$m_{\log(\rho_0/10^9 M_\odot)}$	$m_{\log(b_{\mathrm{DM}})}$
30	1.1065	0.5310	1.1232	1.1213	3.2601
60	1.4368	1.5649	0.4135	2.2385	3.7679
120	1.5320	1.1768	0.6456	1.3324	6.0596
180	2.2466	2.1951	0.5447	1.8445	7.5362

Table 5.2 Hyperparameters of the random process for the GPE of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$) triples.

sample size	R^2	$\sqrt{R^2}/\mathrm{range}(\boldsymbol{y})$	$\max(\lvert r/\ \mathrm{range}(\boldsymbol{y})\rvert)$
30	0.001265	0.03557	0.09538
60	0.0001981	0.01408	0.05862
120	0.000003094	0.001759	0.01016
180	0.0000001633	0.000404	0.002200

Table 5.3 LOOCV score, R^2, root-LOOCV score, $\sqrt{R^2}/\mathrm{range}(\boldsymbol{y})$, and greatest absolute residual, $\max(\lvert r/\ \mathrm{range}(\boldsymbol{y})\rvert)$ for the GPE of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$) triples.

α_i	MLE	argmax($\hat{L}$)			
		30	60	120	180
γ_{DM}	1.3333	1.1667	1.5	1.5	1.3333
$\log(\rho_0/10^9 \mathrm{M}_\odot)$	7.6111	7.5	7.5	7.5	7.6111
b_{DM}	-0.01722	0.08509	-0.01722	-0.01722	-0.01722

Table 5.4 MLE and maximum of the GPE predictions of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$) triples.

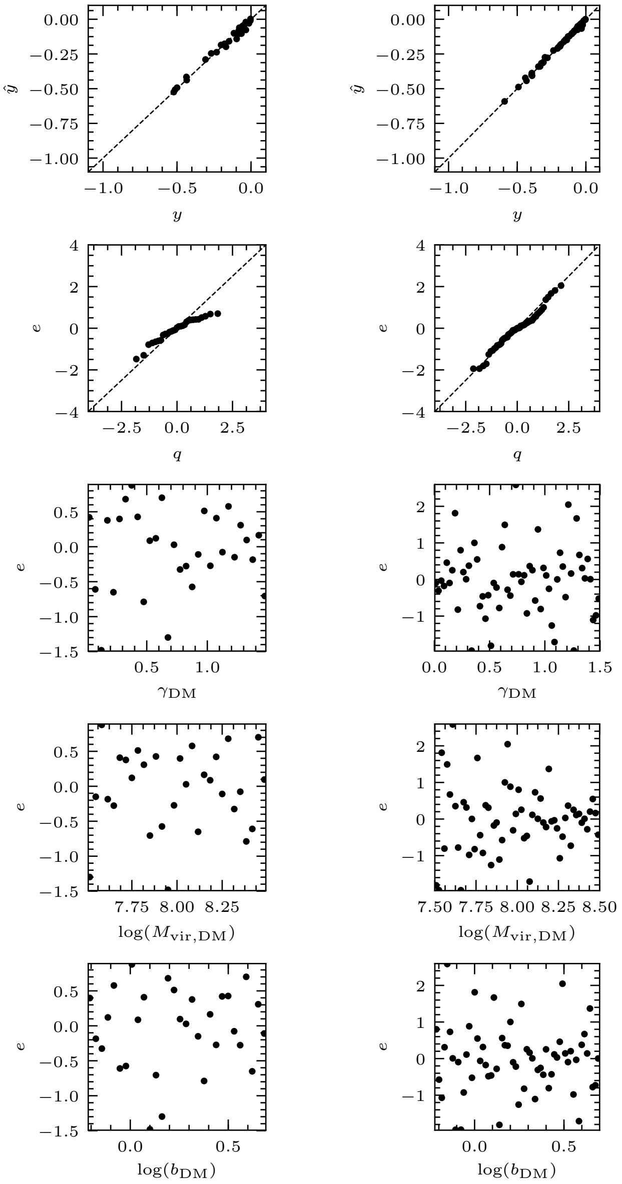

Figure 5.6 Validation plots for the GPE predictions of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$) triples for sample sizes n = 30 (left) and n = 60 (right).

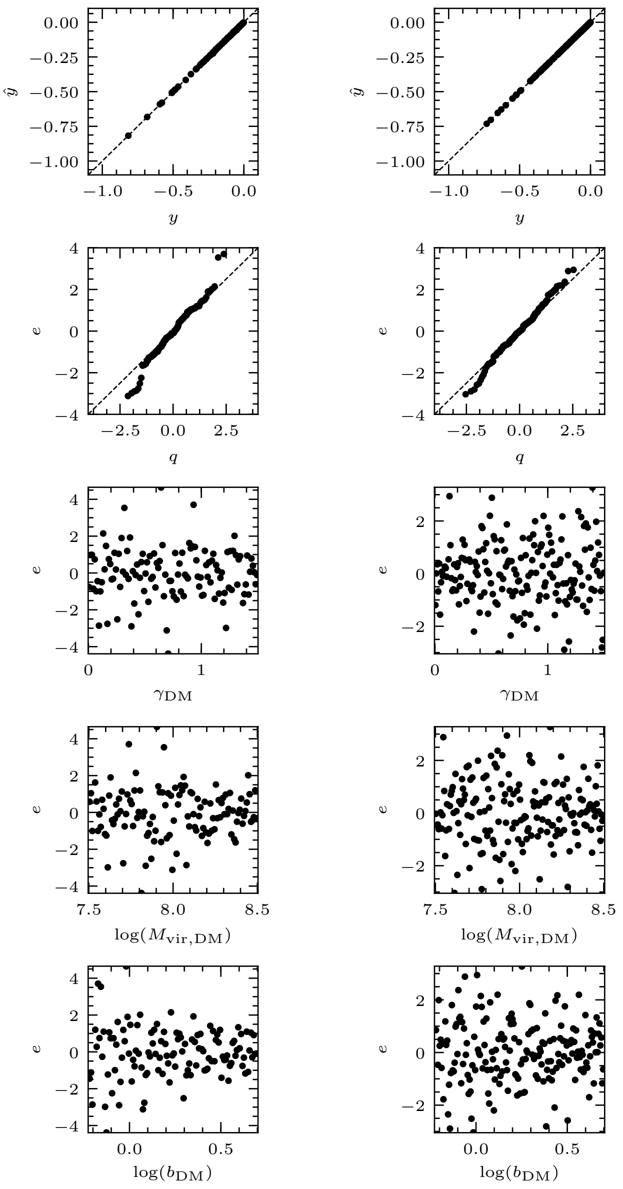

Figure 5.7 Validation plots for the GPE predictions of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$) triples for sample sizes n = 120 (left) and n = 180 (right).

density and length-scale together, under the assumption that the remaining parameter components (outer log-slope, and sharpness) have been fixed at their true values. The results of GPE are shown for $n = 10d = 30$ in Figure 5.3, for $n = 20d = 60$ in Figure 5.4, and for $n = 60d = 180$ in Figure 5.5. In Table 5.2 we show the GPE hyperparameters, computed using the maximum-likelihood method (Ch. 3, Sec. 3.1). The MLE of the tuple $(\gamma_{\text{DM}}, \log(\rho_0), \log(b_{\text{DM}}))$ is $(1.3333 \pm 0.1667, 7.6111 \pm 0.1111, -0.01722 \pm 0.1023)$, where the error is equal to the pixel size we have used to the evaluate likelihood. Note that the true value of the MLE is not equal to true value of the parmameter, but rather the parameter value found by maximizing the support, not its emulation. The true value of the MLE and the true value of the parameter are not necessarily equal due the finite size of our sample, and the existence of observational errors. The maximum in the GPE predictions is shown, for each value of n, in Table 5.4. These maxima are consistent with their true values for all values of n. Validation plots are shown for $n = 10d = 30$ and $n = 20d = 60$ in Figure 5.6, and for $n = 40d = 120$ and $n = 60d = 180$ in Figure 5.7. LOOCV statistics are shown in Tables 5.3.

For $n = 30$ we see (Fig. 5.6, left-hand side, top panel) that the LOOCV predictions show significant scatter when compared to their equivalent true values. Moreover, the Q-Q plots indicates (left-hand side, panel second from top) that the LOOCV residuals are significantly non-Gaussian, although they are all well within the interval $[-3, -3]$. Furthermore, the LOOCV residuals when plotted against each parameter component (left-hand side, lower three panels) show significant bias. In all cases, the LOOCV residuals are negatively biased. For the component γ_{DM} there is also significant boundary bias. The residuals are much larger for small values of γ_{DM} than they are for large γ_{DM}. The normalized LOOCV score, $\sqrt{R^2}/\,\text{range}(\boldsymbol{y}) = 0.03557$ (Tab. 5.3) is well below the useful threshold of 0.1, but not at the level of 0.01 we would like. GPE therefore fails validation for the case of $n = 30$.

We can compare the GPE predictions with the true log-likelihood directly by inspecting Figures 5.3 and 5.2). We see clearly that the predictor is failing, and that in particular it is failing at the boundaries. The support takes on dramatically negative values in the region of the boundary, and these are not adequately represented in the sample $\boldsymbol{y}$. The least sample value is -11146.9, which occurs at $(\gamma_{\text{DM}}, \log(\rho_0/10^9 M_\odot), \log(b_{\text{DM}})) = (1.1250, 8.3500, 0.5302)$. Compare this with the least support value of -21357.0, which occurs at $(\gamma_{\text{DM}}, \log(\rho_0/10^9 M_\odot), \log(b_{\text{DM}})) = (1.5000, 8.500, 0.6990)$. (Note that this is a vertex of the sampled parameter space.) There is a difference of almost a factor two between these support values, despite their arguments being close in parameter space. The GPE prediction for the least value is $\log(\hat{L}(1.5000, 8.500, 0.6990)) =$

-16807.1339, and hence the residual of this prediction is $-78.7\,\%$. This is vastly different from the normalized LOOCV score $\sqrt{R^2}/\operatorname{range}(\boldsymbol{y})$.

The problem persists as we increase the sample size. For $n = 60$ (Fig. 5.4), the validaton plots (Fig. 5.6, right-hand side) show there is improvement in the scatter of the LOOCV predictions plotted against their equivalent true values. Moreover, the Q-Q plot indicates the LOOCV residuals are consistent with being Gaussian. However, the LOOCV residuals are again biased. This time they show a small positive bias, although they all remain in the interval $[-3, 3]$. They also show significant boundary bias for the parameter component $\rho_{0,\mathrm{DM}}$. The residuals are much larger for small values of $\rho_{0,\mathrm{DM}}$ than they are for large $\rho_{0,\mathrm{DM}}$. The normalized LOOCV score, $\sqrt{R^2}/\operatorname{range}(\boldsymbol{y}) = 0.01408$ is again well below the useful threshold of 0.1, but still greater than 0.01. GPE therefore fails validation for $n = 60$.

For $n = 180$, we see that GPE still fails validation. Although the LOOCV predictions are accurate, the LOOCV standardized residuals now clearly non-Gaussian. The Q-Q plot tells us that their distribution is more heavily tailed than that of a Gaussian distribution. Indeed they no longer fall within the range $[-3, 3]$. The LOOCV residuals are now biased for each of the parameter components: they increase with γ_{DM}, and decrease with both $\log(\rho_{0,\mathrm{DM}}$ and $\log(b_{\mathrm{DM}}$. This failure of validation occurs despite the fact that the normalized LOOCV score is $\sqrt{R^2}/\operatorname{range}(\boldsymbol{y}) = 0.002200$, which is significantly smaller than we require. What is clear from the validation plots is the existence of boundary bias, and in particular one-sided boundary bias. By comparing Figures 5.2 and 5.5 We see clearly that the predictor is failing at the boundaries. We have noted that this is were the support takes on dramatically negative values, and that these are not adequately represented in our sample, $\boldsymbol{y}$. In this case, the least element of the sample is -15629.6, and occurs at $(\gamma_{\mathrm{DM}}, \log(\rho_0/10^9 M_\odot), \log(b_{\mathrm{DM}})) = (1.2792, 8.4361, 0.6146)$. In this case, the LOOCV residual is $-35.50\,\%$.

The failure of validation suggests that the assumption of normality required for GPE is violated. The support function is not well modelled by a Gaussian random process with the mean or squared-exponential covariance functions we have specified. We note that the log-likelihood has multiple lenth-scales, and changes rapidly near the large-γ_{DM} and small $\log(\rho_0)$ boundaries. We propose that this failure of validation is typical of functions which rapidly change value near the boundary, noting that a similar problem is observed by Jones et al. (1998) in emulating the Goldstein–Price function (Goldstein and Price, 1971).

Four dimensions. Figure 5.8 shows the log-likelihood for the inner log-slope, normalizing density, length-scale and outer log-slope together. Just as GPE failed validation in three dimen-

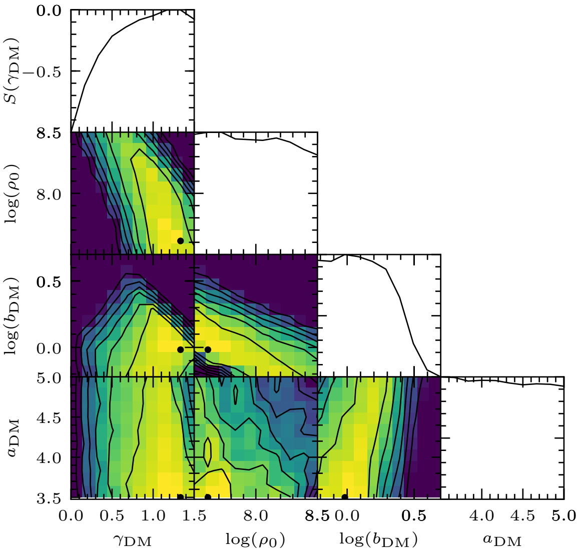

Figure 5.8 Support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\rm DM})$, and $\beta_{\rm DM}$ of the parameter of the generalized Hernquist model. A total of 10^4 evaluations of the likelihood have been required. The unshown parameter components have been set to their true values.

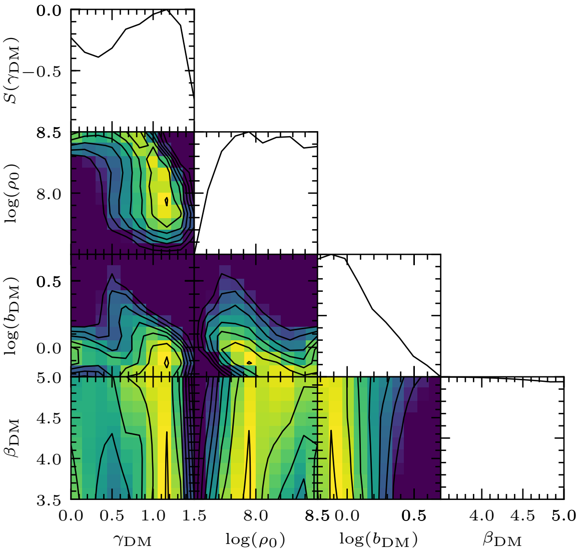

Figure 5.9 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\rm DM})$, and $\beta_{\rm DM}$ of the parameter of the generalized Hernquist model (Fig. 5.8), sample size $n = 10d = 40$.

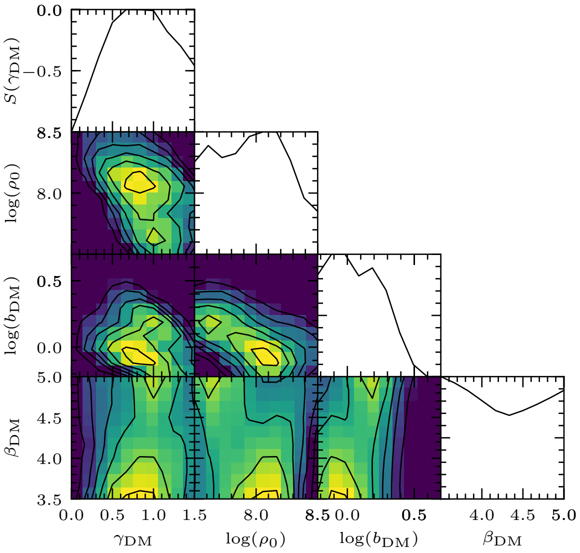

Figure 5.10 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\rm DM})$, and $\beta_{\rm DM}$ of the parameter of the generalized Hernquist model (Fig. 5.8), sample size $n = 20d = 80$.

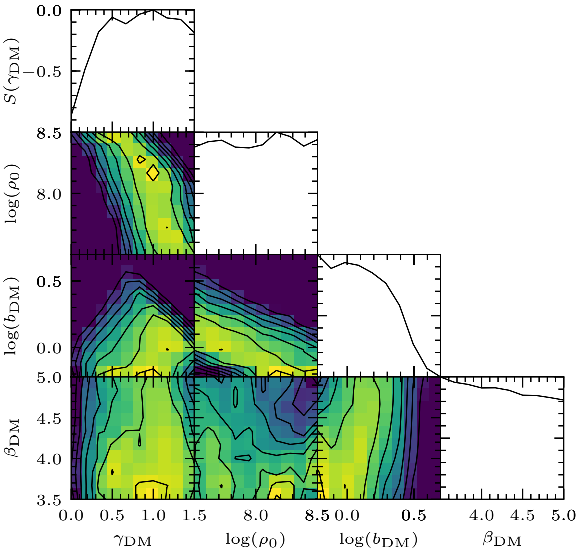

Figure 5.11 GPE predictions for the support for the components $\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\rm DM})$, and $\beta_{\rm DM}$ of the parameter of the generalized Hernquist model (Fig. 5.8), sample size $n = 60d = 240$.

sample size	μ	σ^2	$m_{\gamma_{\mathrm{DM}}}$	$m_{\log(\rho_0/10^9 M_\odot)}$	$m_{\log(b_{\mathrm{DM}})}$	$m_{\beta_{\mathrm{DM}}}$
40	0.6819	0.2197	0.9390	1.3023	1.4915	0.003208
80	0.8347	0.2920	0.5528	1.4235	4.4425	0.04491
160	0.9140	0.3453	0.9789	1.6861	4.8925	0.07708
240	0.9945	0.7102	0.5437	1.4950	7.8548	0.09595

Table 5.5 Hyperparameters of the random process for the GPE of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$, β_{DM}) tuples.

sample size	R^2	$\sqrt{R^2}/\mathrm{range}(\boldsymbol{y})$	$\max(\lvert r/\,\mathrm{range}(\boldsymbol{y})\rvert)$
40	0.0002209	0.01486	0.04818
80	0.0001898	0.01378	0.06867
160	0.00002410	0.004909	0.02661
240	0.000002616	0.001618	0.006743

Table 5.6 LOOCV score, R^2, root-LOOCV score, $\sqrt{R^2}/\mathrm{range}(\boldsymbol{y})$, and greatest absolute residual, $\max(\lvert r/\,\mathrm{range}(\boldsymbol{y})\rvert)$ for the GPE of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$, β_{DM}) tuples.

α_i	MLE	$\mathrm{argmax}(\hat{L})$			
		40	80	160	240
γ_{DM}	1.3333	1.5	1.0000	1.0000	1.3333
$\log(\rho_0/10^9 M_\odot)$	7.6111	7.6111	8.1667	7.8333	7.6111
b_{DM}	-0.01722	-0.2218	-0.2218	-0.01722	-0.01722
β_{DM}	3.5000	3.5000	3.5000	3.6667	3.5000

Table 5.7 MLE and maximum of the GPE predictions of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$, β_{DM}) tuples.

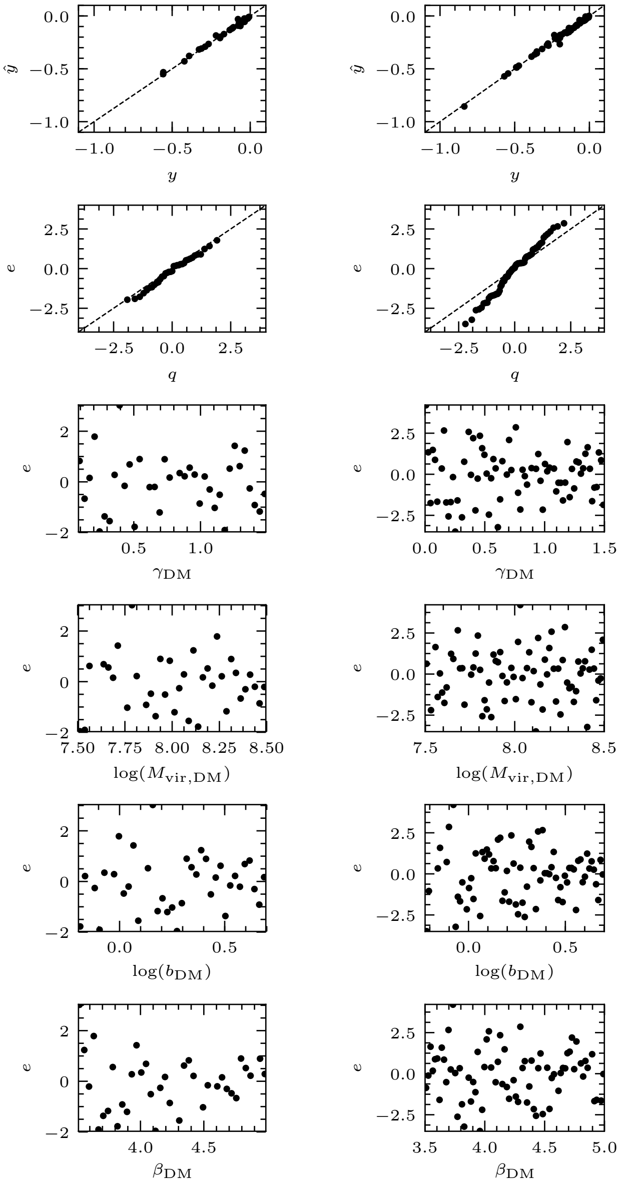

Figure 5.12 Validation plots for the GPE predictions of the support of ($\gamma_{\rm DM}$, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\rm DM})$, $\beta_{\rm DM}$) tuples for sample sizes n = 30 (left) and n = 60 (right).

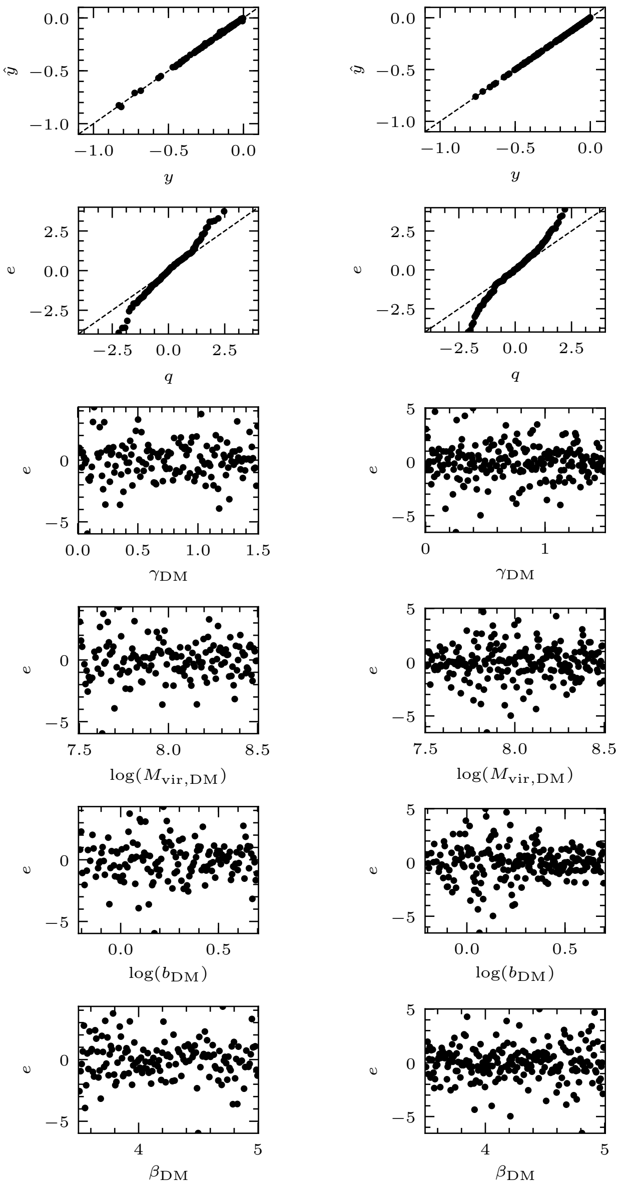

Figure 5.13 Validation plots for the GPE predictions of the support of (γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$, β_{DM}) tuples for sample sizes n = 120 (left) and n = 180 (right).

sions, so it fails in four dimensions. The results of this GPE are shown for $n = 10d = 40$ in Figure 5.9, for $n = 20d = 80$ in Figure 5.10, and for $n = 60d = 240$ in Figure 5.11. In Table 5.5 we show the GPE hyperparameters. The MLE of the tuple $(\gamma_{\mathrm{DM}}, \log(\rho_0), \log(b_{\mathrm{DM}}), \beta_{\mathrm{DM}})$ is $(1.3333 \pm 0.1667, 7.6111 \pm 0.1111, -0.01722 \pm 0.1023, 3.5 \pm 0.1667)$, which may be compared with the maxima of the GPE predictions shown, for each value of n, in Table 5.7. Again, these maxima of the GPE predictions for $(\gamma_{\mathrm{DM}}, \log(\rho_0/10^9 M_\odot), b_{\mathrm{DM}})$ are consistent with their true values for all values of n. Validation plots are shown for $n = 10d = 40$ and $n = 20d = 80$ in Figure 5.12, and for $n = 40d = 180$ and $n = 60d = 240$ in Figure 5.13. LOOCV statistics are shown in Tables 5.6,

For $n = 10d = 40$, the the LOOCV predictions show significant scatter when compared to their equivalent true values. Although the Q-Q plot shows that the distribution of LOOCV residuals is consistent with being Gaussian, they are not uniformly distributed over parameter space. Again, they are display a small negative bais, and are greater for small values of $\rho_{0,\mathrm{DM}}$ than for large values of $\rho_{0,\mathrm{DM}}$. The same is true for $\log(b_{\mathrm{DM}})$, and β_{DM}. The normalized LOOCV score is $\sqrt{R^2}/\operatorname{range}(\boldsymbol{y}) = 0.01486$, which is again well below the useful threshold of 0.1, but too great for our requirements. We can again compare the GPE predictions with the true log-likelihood directly in Figures 5.8 and 5.9. We see clearly that the predictor is failing, in particular at the boundaries. The support takes on dramatically negative values in the region of the boundary, and In this case, the least sample value is -11905.5, which occurs at $(\gamma_{\mathrm{DM}}, \log(\rho_0/10^9 M_\odot), \log(b_{\mathrm{DM}})) = (1.0688, 8.3875, 0.6414, 4.6813)$. Compare this with the least support value of -21357.0, which occurs at $(\gamma_{\mathrm{DM}}, \log(\rho_0/10^9 M_\odot), \log(b_{\mathrm{DM}})) = (1.5000, 8.5000, 0.6990, 3.5000)$. (Again, this is vertex of sampled parameter space.) The prediction for the least value is $\log(\hat{L}(1.5000, 8.5000, 0.6990, 3.5000)) = 21078.9476$. The residual of the prediction is $-1.302\,\%$, which is good. However, the worst residual occurs for the predicted value for $L(0.0000, 7.5000, -2.2185, 5.0000)$, which is $51.560\,\%$. Again, this occurs at a vertex, and again it is vastly different from both LOOCV quantities $\sqrt{R^2}/\operatorname{range}(\boldsymbol{y})$ and $\max(|r/\operatorname{range}(\boldsymbol{y})|)$.

Again, the problem persists as we increase the sample size. Let us consider $n = 60d = 240$. (The story is very much the same for $n = 80$ and $n = 180$.) Althought the LOOCV predictions are good, the Q-Q plot shows the distribution of the LOOCV residuals to be dramatically non-Gaussian, with very heavy tails. Moreover, these residuals no longer fall within the interval $[-3, 3]$, with some being greater that five. Again, this is due to boundary bias, due to the GPE failing to precisely predict the very negative values of the log-likelihood at the boundary. The least sampled value is -16314.2, which occurs at $(\gamma_{\rm DM}, \log(\rho_0/10^9 M_\odot), \log(b_{\rm DM})) = (1.4969, 8.4646, 0.4553, 3.5656)$. In this case, the residual of prediction is $-1.889\,\%$. The worst residual, however, occurs for the predicted value for $\log(\hat{L}(0.0000, 7.5000, -2.2185, 5.0000))$, which is $278.41\,\%$. It is clear that in four dimensions, as in three dimension, the assumption of normality required for GPE is violated. The support is not a likely realization of a random process with constant mean and squared-exponential covariance.

The usual solution to failed validation is to transform the function we are emulating, in the transform function is better suited for emulation (Jones et al., 1998). However, we find that the standard transformations—$\sqrt{-S}$, $\operatorname{arcsinh}(S)$, etc. (Bartlett, 1947)—do not significantly improve the validation statistics or diagnostic plots. We might, alternatively, preferentially sample the support at the the boundary. But the support is in principle arbitrarily negative at the boundary, and may in practice be undefined there due to numerical overflow in its computation. This points to a more significant problem with attempting to emulate the support: we may not be able to choose the parameter bounds to satisfy the requirements of GPE. In emulating the support, we may require unduly tight bounds on the model parameter. These results emphasizes the importance of the validation step, which has been entirely neglected in the application of GPE to astrophysics thus far in the literature (for example, Bower et al., 2010; Gibson et al., 2012; Sale and Magorrian, 2019). It should be regarded as a strength of the method that we may recognize when it underperforms.

Satisfied that GPE is failing to adequately predict the support for a given parameter, we do not attempt GPE in five dimensions. Nonetheless, we do plot the support for five dimensions in Figure 5.14. It is clear from this that HernDF itself is not working properly. It indicates that the MLE of the model parameter occurs outside the sampled parameter space. This is too great a difference from the true parameter value to be due to stellar sample size or observational error. It must be a problem with the implemenation of our method. The problem appears to be less severe when the component α has been marginalized out of the support, suggesting that the problem may be associated with the component α. It is also the case that the marginalizations $(\log(\beta_{\rm DM}), \alpha_{\rm DM})$ and $(\log(\rho_0/10^0 \mathrm{M}_\odot), \log(\beta_{\rm DM}))$ appear unexpectedly noisy. This problem with HernDF is independent of the failure of GPE. GPE has failed neither because of the location of the support's maximum nor the support's unexpected noisiness, but because of its very negative values near the boundary of parameter space.

We conclude that ordinary kriging with a squared-exponential covariance function is inadequate

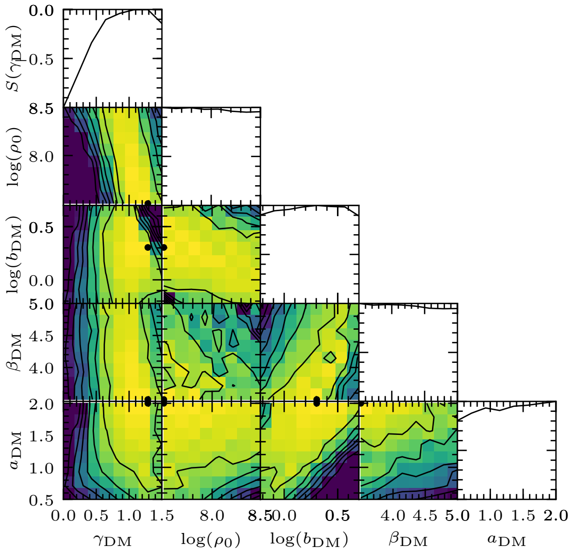

Figure 5.14 Support for the components γ_{DM}, $\log(\rho_0/10^9 M_\odot)$, $\log(b_{\mathrm{DM}})$, β_{DM}, and α_{DM} of the parameter of the generalized Hernquist model. A total of 10^5 evaluations of the likelihood have been required.

for constructing metamodels of the support. The natural next step is to change the coviance function, or to consider universal kriging (Ch. 2 Sec. 4.1). In universal kriging we use polynomial mean rather than constant mean. What degree polynomial should we consider? We leave this for future work, but note that in certain circumstances the likelihood is asymptotically normal (Thm 27), meaning that the support is therefore asymptotically quadratic. If we believe that these circumstances hold, we might therefore consider a polynomial basis of degree two: $(x_1^i x_2^j \ldots x_1^k)_{i+j+\cdots+k \leq 2}$. For a high-dimensional parameter space, the size of this basis is large. But note that the number of basis functions does not affect the number of hyperparameter components.[1] Looking again at Equation 2.181 we may think of the term $\varphi^{\mathrm{t}}\hat{\mathrm{B}}$ as accounting for the quadratic trend, and the term $k^{\mathrm{t}}K^{-1}(X - A\hat{\mathrm{B}})$ as accounting for departures from this trend.

[1] The number of basis functions does affect the size of the matrix A in Eq. 2.181, however, and hence the complexity of compututing A^{t} and $(A^{\mathrm{t}}K^{-1}A)^{-1}$.

Conclusion

In this thesis I have shown how we may construct dynamical metamodels of isolated dSph galaxies. This method is a novel combination of distribution-function modelling, maximum-likelihood parameter recovery, and GPE. The essentials of the method were presented in Chapter 4: we construct a model of the distribution-function for a given dSph galaxy, and then marginalize this model to form the PDF of the observable quantities, namely the sky positions and line-of-sight velocities of its constituent stars. We then identify a feasible region of parameter space, and compute the likelihoods for a sample of parameters within this region. With this information we construct a metamodel of the likelihood, which allows us to predict the likelihood of a given parameter without having to again evaluate the likelihood directly. In the case that the likelihood is unimodal, we may then use EGO to maximize the emulated likelihood and hence recover the model parameters and their associated confidence regions. In the case the likelihood is not unimodal, we may use the emulator to reconstruct the likelihood for the entirety of the feasible region of parameter space.

The advantage of this method is in the enormous reduction in computational complexity, and hence in the quantity of computing time we need. Likelihoods are expensive to evaluate. Even for simple models the likelihood has no closed-form expression, and can require of order 10 s of computing time to evaluate numerically. In high-dimensional parameter spaces this is prohibitive. The parameter space is simply too large for us to use conventional maximization schemes, or even to evaluate the likelihood on a coarse lattice. GPE therefore makes possible the use of sophisticated dynamical models that have previously been impossible. In Chapter 4 I presented a toy implementation of my method, in which I constructed a metamodel of a Plummer sphere with a distribution function of the Osipkov–Merritt type. I showed that it was possible to recover the model parameter, in this case a tuple of length three, using fewer than 100 evaluations of the likelihood. This is two orders of mangitude less than the number of evaluations required by a very coarse lattice search of 10^3 evaluations.

The methods of GPE and EGO are widely used. However there are some particular issues in their application to astrophysics. Most significantly, we must be careful to distinguish between the BLP (eq. 2.181) and the BLUP (eq. 2.118). The BLP is a biased predictor, a fact that is never commented upon. In many applications it seems that this is not problematic. But it may well be problematic in the case of unimodal functions. Hence we must be particularly wary when using it to construct metamodels of likelihoods. In this case it is by far preferable to use the BLUP for our metamodels.

It is also the case that the use of GPE in the prediction of computation experiments relies heavily on the methods developed for geostatistics, where GPE goes by the name 'kriging'. The BLP goes by the name 'ordinary kriging', and the BLUP by the names 'ordinary kriging' (if the mean is taken to be constant) and 'universal kriging' (if the mean is taken to be polynomial). The most significant difference between the two fields is in the strucuture of the input space. In geostatistics the input space is Euclidean, $\mathbf{R}^n$, and covariance functions are untroublingly taken to be functions of the distance between two points in input space. In computer experimentation the input space is not necessarily Euclidean, and we are not free to equip these input spaces with the Euclidean metric. We may frequently be able to embed these spaces in, or transform them to, $\mathbf{R}^n$, but it is not clear what effect this has on the covariance of the resulting Gaussian process. The rigorous approach here is to define a proper metric on the input space, and construct covariance functions using this metric. I have shown how this may be done in Chapter 1. This opens the door to the construction of new covariance functions, each tailored to the physical problem at hand. I leave the construction of such covariance functions for future work.

In the presence of such issues, I have emphasized the importance of validating the results of GPE. Validation appears to a limited extent in the machine learning literature (for example, it is considered briefly by Jones et al., 1998), but not at all in the astrophysical literature. For example is does not appear in any of the papers by Gibson et al. (2012), Evans et al. (2015), Sale and Magorrian (2014), Sale and Magorrian (2019), Bower et al. (2010), or Heitmann et al. (2009), that I listed in Chapter 1. We do not know *a priori* that the function we are seeking to emulate is a likely realization of a Gaussian random process with the mean and variance we have chosen. If it is not then GPE will give unreliable predictions and, moreover, unreliable confidence intervals for those predictions. The importance of validation was made manifest in Chapter 5, in which I have constructed a more realistic metamodel of a dSph galaxy than the one presented in Chapter 4. This assumed that the dSph in question consisted of a single stellar population tracing the potential of a dark-matter halo. The density of both stellar population and halo were assumed to be described by the generalized Hernquist model, and the stellar distribution function was constructed using the methods of Gerhard (1991), Cuddeford and Louis (1995). Here, ordinary kriging, in which we construct a metamodel under the assumption of intrinsic stationarity (Def. 51) results in predictions that fail validation. We should consider it a virtue of the method of GPE that we can tell when it fails, and we should be wary of results that are not validated in this way.

Throughout, my interest has been in what an observational data set can tell us about the stellar system from which it is drawn. In particular, for a dSph, I would like to know which dark-matter distributions a kinematic data set rules out, and which dark-matter distributions best account for the data. Once we have constructed a metamodel of the likelihood of our model parameters we may easily construct metamodels for other physically interesting quantities. In Chapter 4 I used the results of GPE to constructed maximum-likelihood estimates for the galactic density profile and for Binney's anosotopy parameter. These were in excellent agreement with the true quantities. Both predictions and the confidence intervals for those predictions were accurate to the order of 1 % or less.

5.3 FUTURE WORK

The most pressing task at hand is the resolution of the issues outlined in Chapter 5. Here, the ordinary kriging predictions failed validation. This was not due to the increase in the dimension of parameter space, but rather to the presence of multiple length scales in the likelihood. The assumptions of ordinary kriging, and in particular the adoption of the standard squared-exponential covariance function are unsuitable for such a problem. I expect that the the use of universal kriging or the adoption of a better-specified covariance function will resolve this issue. The natural next step is then to use GPE to construct very general metamodels of dSph galaxes, which include multiple stellar populations and non-spherical density distributions.

I also plan to develop an information-theoretic analysis of the existing data. The information content of this data may be quantified by the Fisher information (Def. 25), which determines the confidence region for the maximum-likelihood estimate of the parameter vector (eq. 1.33). Again, the expense of dynamical modelling makes it difficult to evaluate the Fisher information using standard methods, as we must maximize the support (i.e. the log-likelihood) for the model's parameter tuple. However, the problem should yield to the application of EGO. This will allow me to answer the question, 'What can the kinematic data tell us?'. Can they satisfactorily constrain the model parameter vector? Do they allow us to discriminate between competing models? If not, which data are needed? In the same spirit, it makes sense to investigate the principal components of the parameter space using the emulated likelihood. This is equivalent to finding the orthogonal decomposition of the random process we use to construct its metamodel (a fact that goes by the name of the Karhunen-Loéve theorem, Parzen, 1959). By finding the principal components of the parameter space we are finding the natural parameterization of the system in hand. We might, for example, expect this to be the random variable representing the actions and angles of the system.

Alongside a continued exploration of likelihood methods, it would also be interesting to consider the emulation of other quantities of physical interest. For example, we might use GPE to construct metamodels of the phase-space distribution function directly, rather than metamodels of the likelihood. There are two potential advantages of this approach: the distribution function may be more suitable for emulation than the likelihood function, and, having found a GPE estimate for the distribution function, the integrals that give the galaxy's observable quantities are cheaply computed using stochastic calculus.

Throughout this thesis I have have been interested in static dynamical model of a dSph galaxies. These assume that the galaxy is in dynamical equilibrium. This is a good assumption for the isolated dSph galaxies, like Fornax. Others, such as the Sagittarius dSph galaxy, are clearly not in equilibrium, and are instead stongly interacting with the Milky Way. In the process of this interaction they are stripped of their outer envelope of stars, which form tidal tails. GPE, however, is suitable for constructing metamodels of all kinds of model. I therefore plan to use it to fit N-body models of these galaxies to observations of tidally disturbed dSphs. This work is a natural extension of that by Ural et al. (2015), who constructed an N-body evolutionary model of Carina that included both its internal dynamics and its bulk motion within the Galactic potential to determine its preinfall mass. Ural et al. used Markov-chain Monte Carlo (MCMC) methods to sample the likelihood of the parameters of these N-body models. This required a total of 19 000 model evaluations. Whereas MCMC methods require the sequential evaluation of these N-body models, GPE is trivially parallelizable, and hence provides a clear computational advantage. Ural et al.'s work has shown that the existing observational data can tell us a lot more about the history of the Milky Way satellites than we might have expected, provided we have the modelling tools and computational resources. I would be able to fit models to all the dSph galaxies for which we have good enough data. This would allow me to determine the masses of the Milky Way satellites both now and in the past, when they first fell onto the Milky Way. We might emulate the likelihood obtained by matching the final simulation snapshot to the observed data, or we might emulate the snapshots themselves. It is not clear which approach would gives us the greater insight, or which is most computationally efficient. Both approaches are worth exploring.

It is also the case that I have considered only the application of GPE to scalar-valued multivariable functions. But GPE may also be applied to vector- or matrix-valued multivariable functions, in which case it is known as *multitask GPE* (Bonilla et al., 2008). We may thus use GPE when the output of our simulation is an image, or full probability distribution function. Whereas GPE has been applied to a variety of astrophysical and cosmological tasks, multitask GPE is, to the best of my knowledge, new to these fields. Using this method I plan to characterize the dependence of tidal tails on their pre-infall properties. Like scalar-valued GPE, multitask GPE has broad applicability and may be used for fields outside galactic dynamics. Some possible applications include the emulation of: the distribution of exoplanetary orbits after migration through their exosolar systems and the distribution of exoplanetary orbits in stellar clusters. In both cases multitask GPE could be used to predict a tuple quantifying the PDF in question (say, the mean, variance, and skewness of that PDF).

Ultimately I would like to develop robust methods for modelling the complete evolutionary histories of dSph galaxies. This would involve the development of hybrid *N*-body and hydrodynamical models that contain both dynamics and chemistry. Again, GPE would allow me to explore greater regions of parameter space than has been possible, allowing me to vary the orbit, pre-infall mass, initial gas distribution, and supernova history. These models would then be fitted to observations of dwarf-spheroidal bulk motions and detailed chemical histories known from Gaia and high-resolution spectroscopic data respectively. Can GPE be made robust for problems of this kind? If it can then it will provide a powerful tool for the study of galactic dynamics.

Appendix A

Basic definitions

For the sake of convenience I here recall some basic definitions concerning topological vector spaces (Sec. A.1) and probability theory (Sec. A.2).

A.1 SPACES

A *space* is a set endowed with some structure. In the following definitions, X is an arbitrary set. The definitive text on the subject is that by Bourbaki (1981).

Definition 110 (topological space). We say that a set $\mathscr{T}$ of subsets of X is a *topology on X* if

(a) $\varnothing, X \in \mathscr{T}$,

(b) $\mathscr{T}$ is closed under arbitrary unions, i.e. for all $\mathscr{U} \subseteq \mathscr{T}$ we have $\bigcup_{U \in \mathscr{U}} U \in \mathscr{T}$, and

(c) $\mathscr{T}$ is closed under finite intersections, i.e. for all $U_1, \ldots, U_n \in \mathscr{T}$ we have $\bigcap_{i=1}^{n} U_i \in \mathscr{T}$.

In this case we say that $(X, \mathscr{T})$ is a *topological space*.

Definition 111 (Hausdorff space). A topological space $(X, \mathscr{T})$ is *Hausdorff* if for all $a, b \in X$ with $a \neq b$ there exist $U, V \in \mathscr{T}$ with $a \in U$, $b \in V$, and $U \cap V = \varnothing$.

Definition 112 (metric). We say that a function $d : X \times X \longrightarrow \mathbf{R}_{\geq 0}$ is a *metric on X* if

(a) it is positive definite, i.e. if for all $x, y \in X$

$$d(x, y) = 0 \iff x = y,$$

(b) it is symmetric, i.e. if for all $x, y \in X$

$$d(x,y) = d(y,x),$$

(c) it obeys the triangle inequality, i.e. if for all $x, y, z \in X$

$$d(x,z) \leq d(x,y) + d(y,z).$$

In this case we say that (X, d) is a *metric space*.

Definition 113 (pseudometric). We say that $d : X \times X \longrightarrow \mathbf{R}_{\geq 0}$ is a *pseudometric on X* if it is symmetric and it obeys the triangle inequality. In this case we say that (X, d) is a *pseudometric space*.

Remark 114. A pseudometric d induces on X a topology, which we will denote by $\mathscr{T}_d$. This applies in particular to metrics d. A topology induced by a metric will be Hausdorff, but a topology induced by a pseudometric need not be Hausdorff in general.

Example 115 (pseudometric). Consider the function $d : \mathbf{Z} \times \mathbf{Z} \longrightarrow \mathbf{R}_{\geq 0}$ given by

$$d(m,n) := \big||m| - |n|\big|,$$

for $m, n \in \mathbf{Z}$. Then d is a symmetric positive semi-definite function, which satisfies the triangle inequality. To see this, without loss of generality suppose that $m \geq n \geq 0$ and $l \geq 0$, then

$$\begin{aligned} \big||m| - |n|\big| &= m - n && \text{(A.1)} \\ &= m - l + l - n && \text{(A.2)} \\ &\leq |m - l| + |l - n| && \text{(A.3)} \\ &= \big||m| - |l|\big| + \big||n| - |l|\big|. && \text{(A.4)} \end{aligned}$$

Therefore $(\mathbf{Z}, d)$ is a pseudometric space but not a metric space. Nevertheless, since for example $d(1, -1) = 0$, d is not a metric.

Definition 116 (Cauchy sequence). Let (X, d) be a metric space. A sequence $(a_n)_n$ in X is *Cauchy* if for all $\varepsilon > 0$ there exists $N \in \mathbf{N}$ such that for all $m, n > N$ we have

$$|a_n - a_m| < \varepsilon. \tag{A.5}$$

Definition 117 (completeness of a metric space). A metric space (X, d) is *complete* if every Cauchy sequence has a limit in X.

Remark 118 (completeness of a pseudometric space). Just as for a metric space, we say that a pseudometric space (X, d) is *complete* if every Cauchy sequence is convergent. However, since the topology induced by a pseudometric need not in general be Hausdorff, limits of sequences need not be unique.

A.1.1 Vector spaces

We consider only vector spaces over the field $\mathbf{R}$.

Definition 119 (vector space). A *vector space* is a tuple $\mathbf{V} = (V; +, -, 0, \mu)$, where $(V; +, -, 0)$ is an Abelian group and $\mu : \mathbf{R} \times V \longrightarrow V$ is called 'scalar multiplication', and written $\alpha u = \mu(\alpha, u)$, which together satisfy the axioms below.

The axioms are as follows. For all $u, v \in V$ and $\alpha, \beta \in \mathbf{R}$ we have:

(a) $\alpha(u + v) = \alpha u + \alpha v$,

(b) $(\alpha + \beta)u = \alpha u + \beta v$,

(c) $(\alpha\beta)u = \alpha(\beta u)$, and

(d) $1u = u$.

Definition 120 (linear independence, linear spanning, linear bases). Let V be a vector space. A set $A \subseteq V$ of vectors is *linearly independent* if

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \implies \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0,$$

for all pairwise distinct $v_1, v_2, \ldots, v_n \in A$, and all $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbf{R}$. The *linear span* of A, which we denote by $\operatorname{span}(A)$, is the set of elements

$$\sum_{i=1}^{n} \alpha_i v_i, \tag{A.6}$$

for $v_1, v_2, \ldots, v_n \in A$ and $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbf{R}$. We say that *$A$ linearly spans V* if $\operatorname{span}(A) = V$. Finally, A is a *linear basis* if it is linearly independent and linearly spans V.

Fact 121. *Every vector space V has a linear basis, and any two linear bases of V have the same cardinality. Moreover, for a set A the following are equivalent:*

(a) A is a linear basis,

(b) A is a minimal linear spanning set,

(c) *A is a maximal linearly independent set.*

Here, A is maximal *linearly independent if every proper superset $B \supset A$ is not linearly independent. Similarly, A is* minimal *spanning if every proper subset $B \subset A$ is not linear spanning.*

Definition 122. The *linear dimension* of V is the cardinality of some linear basis (equivalently, of all linear bases) of V.

Definition 123 (normed vector space). Let V be a vector space and let $\|\cdot\|$ be a nonnegative function

$$\|\cdot\| : V \longrightarrow \mathbf{R}_{\geq 0} \tag{A.7}$$

$$v \longmapsto \|v\|. \tag{A.8}$$

The function $\|\cdot\|$ is a *norm on V* if

(a) it is positive-definite, i.e. if for all $v \in V$

$$\|v\| = 0 \Longleftrightarrow v = 0,$$

(b) it is absolutely scalable, i.e. if for all $k \in K$ and for all $v \in V$

$$\|kv\| = k\|v\|,$$

(c) it obeys the triangle inequality, i.e. for all $u, v \in V$,

$$\|u + v\| \leq \|u\| + \|v\|.$$

A pair $(V, \|\cdot\|)$ is called a *normed vector space*.

Definition 124 (induced metric). Let $(V, \|\cdot\|)$ be a normed vector space. The function

$$d : V \times V \longrightarrow \mathbf{R}_{\geq 0} \tag{A.9}$$

$$(v_1, v_2) \longmapsto \sqrt{\|v_1 - v_2\|} \tag{A.10}$$

is a metric on V. Such a metric is said to have been *induced* by the norm.

Definition 125 (inner product). Let V be a vector space, and let $\langle\cdot,\cdot\rangle$ be a function

$$\langle\cdot,\cdot\rangle : V \times V \longrightarrow \mathbf{R} \tag{A.11}$$

$$(v, w) \longmapsto \langle v, w\rangle. \tag{A.12}$$

The function $\langle\cdot,\cdot\rangle$ is an *inner product* on V if

(a) it is positive definite, i.e. if for all $v \in V$

$$\langle v, v \rangle = 0 \Longleftrightarrow v = 0,$$

(b) it is symmetric, i.e. if for all $v, w \in V$

$$\langle v, w \rangle = \langle w, v \rangle,$$

(c) it is linear in the first argument, i.e. if for all $k \in \mathbf{R}$ and for all $u, v, w \in V$,

$$\langle k(u + v), w \rangle = k(\langle u, w \rangle + \langle v, w \rangle).$$

A pair $(V, \langle , \rangle)$ is called an *inner product space*.

Remark 126. Because every inner product induces a norm, and every norm induces a metric, we may naturally view every inner product space and every normed vector space as a metric space. Thus is makes sense to speak of complete inner product spaces and complete normed vector spaces.

Definition 127 (Banach space). A complete normed vector space is called a *Banach space*.

Definition 128 (Hilbert spaces). A complete inner product space is called a *Hilbert space*.

Remark 129. As we have seen, each inner product induces a norm, and this gives a function from the set of inner products to the set of norms, which is not surjective in general.

Definition 130 (projection of a vector). Let U, V and W be vector spaces such that $U = V \oplus W$. An endomorphism $\pi : U \longrightarrow U$ is a *projection* if

$$\pi : U \longrightarrow U \tag{A.13}$$

$$v + w \longmapsto v, \tag{A.14}$$

for all $v \in V$ and $w \in W$.

Proposition 131. Let V be a vector space and let $\pi : V \longrightarrow V$ be a projection. Then π is *idempotent*, i.e. $\pi \circ \pi = \pi$.

Proof. Clearly the map given in Definition 130 is idempotent. □

Proposition 132. Let V be a Hilbert space. Let $\{e_i\}_{i\in I}$ be an orthonormal set, and let $W \leq V$ be the closure of the span of $\{e_i\}_{i\in I}$. Write $V = W \oplus W^{\perp}$, and form the orthogonal projection π_W such that $\pi_W(w + x) = w$ for all $w \in W$ and $x \in W^{\perp}$. Then for all $v \in V$,

$$\pi_W(v) = \sum_{i\in I} \langle v, e_i \rangle e_i, \tag{A.15}$$

of which only countably many terms are non-zero.

Proof. A proof is given by Bourbaki (1981, Ch. V Sec. 2 No 3 Prop. 4). □

A.1.2 Duals of vector spaces

Having defined a vector space, we may now define its *dual.*

Definition 133 (linear functional). Let V be a vector space. A *linear functional* (also *linear form*) on V is a linear map $f\colon V \longrightarrow \mathbf{R}$.

Definition 134 (dual vector space). Let V be a vector space. Its *dual vector space*, V^*, is the set of all linear functionals on V, endowed with the structure of a vector space.

Theorem 135. *Let V be a vector space of finite dimension, and let V^* be its dual vector space. Then* $\dim V^* = \dim V$.

Proof. A proof is given by Bourbaki (1989, Ch. II Sec. 7 No 5 Thm. 4). □

A.2 PROBABILITY

An excellent text on the foundations of probability is that by Kallenberg (1997).

A.2.1 Measure theory

If Ω is a set, then we denote its power set (i.e. the set of all subsets of Ω) by $\mathscr{P}(\Omega)$.

Definition 136 (σ-algebra). A set $\mathscr{M} \subseteq \mathscr{P}(\Omega)$ is a *σ-algebra* on Ω if

(a) $\varnothing, \Omega \in \mathscr{M}$,

(b) $\mathscr{M}$ is closed under complements, i.e. if $A \in \mathscr{M}$ then $\Omega \setminus A \in \mathscr{M}$, and

(c) $\mathscr{M}$ is closed under countable unions, i.e. if $A_i \in \mathscr{M}$ for all i in a countable index set I then $\bigcup_{i\in I} A_i \in \mathscr{M}$.

The pair $(\Omega, \mathscr{M})$ is then called a *measurable space*.

Note that a σ-algebra is closed under countable intersections, by De Morgan's Laws.

Definition 137 (measure). A *measure* on $(\Omega, \mathscr{M})$ is a function $\mu : \mathscr{M} \longrightarrow \mathbf{R}_{\geq 0} \cup \{\infty\}$ such that

(a) $\mu(\varnothing) = 0$ and

(b) μ is countably additive, i.e. for $(A_i)_{i \in I}$ a countable family of pairwise disjoint elements of $\mathscr{M}$ we have

$$\mu\Big(\bigcup_{i \in I} A_i\Big) = \sum_{i \in I} \mu(A_n),$$

where the sum is interpreted as ∞ if it diverges.

The triple $(\Omega, \mathscr{M}, \mu)$ is then called a *measure space*.

Definition 138 (measurable function). Given two measurable spaces $(\Omega_1, \mathscr{M}_1)$ and $(\Omega_2, \mathscr{M}_2)$, a function $f : \Omega_1 \longrightarrow \Omega_2$ is *measurable* if

$$f^{-1}(A) \in \mathscr{M}_1,$$

for all $A \in \mathscr{M}_2$. (Note that $f^{-1}(A)$ denotes the preimage of A.) We write $f : (\Omega_1, \mathscr{M}_1) \longrightarrow (\Omega_2, \mathscr{M}_2)$ to indicated that f is measurable.

Definition 139 (push-forward measure). Let $(\Omega_1, \mathscr{M}_1, \mu)$ be a measure space, let $(\Omega_2, \mathscr{M}_2)$ be a measurable space, and let $f : (\Omega_1, \mathscr{M}_1) \longrightarrow (\Omega_2, \mathscr{M}_2)$ be a measurable function. We define the *push-forward measure* to be

$$\mu_f : \mathscr{M}_2 \longrightarrow \mathbf{R}_{\geq 0} \cup \{\infty\} \tag{A.16}$$

$$A \longmapsto \mu(f^{-1}(A)), \tag{A.17}$$

for $A \in \mathscr{M}_2$. Some authors prefer the notation $f_* \mu$ for the push-forward measure μ_f.

Lemma 140. *The push-forward measure μ_f, as defined above, really is a measure on $(\Omega_2, \mathscr{M}_2)$.*

Proof. The complement of a pre-image is the pre-image of the complement, and the union of pre-images is the pre-image of the union. □

Definition 141. Let $(\Omega_1, \mathscr{M}_1, \mu)$ be a measure space, and let $f : (\Omega_1, \mathscr{M}_1) \longrightarrow (\Omega_2, \mathscr{M}_2)$ be a measurable function. The σ-algebra *generated* by f, which we denote by $\sigma(f)$, is by definition the smallest σ-algebra on Ω_1 with respect to which f is measurable. In fact $\sigma(f)$ is the pre-image of $\mathscr{M}_2$ under f, i.e. the set of pre-images $f^{-1}(B)$, for $B \in \mathscr{M}_2$.

A.2.2 Lebesgue measure

If $I \subseteq \mathbf{R}$ is an interval, then it is equal to either (a,b), $(a,b]$, $[a,b)$, or $[a,b]$, for some $a,b \in \mathbf{R} \cup \{\pm\infty\}$. The *length* of I is $l(I) := b - a$, with appropriate meaning if either a or b is ∞ or $-\infty$.

Definition 142 (Outer measure). We define the *outer measure* on $\mathbf{R}$ to be the function

$$\mu_{\text{out}} : \mathscr{P}(\mathbf{R}) \longrightarrow \mathbf{R}_{\geq 0} \cup \{\infty\} \tag{A.18}$$

$$A \longmapsto \inf\left\{ \sum_{n=0}^{\infty} l(I_n) \;\middle|\; A \subseteq \bigcup_{n=0}^{\infty} I_n, (I_n)_n \text{ a countable family of intervals}\right\}. \tag{A.19}$$

Importantly, we note that μ_{out} is not *a priori* a measure! We say $A \in \mathscr{P}(\mathbf{R})$ is *Lebesgue measurable* if

$$\mu_{\text{out}}(X) = \mu_{\text{out}}(X \cap A) + \mu_{\text{out}}(X \setminus A),$$

for all $X \in \mathscr{P}(\mathbf{R})$.

Definition 143. We denote by $\mathscr{M}_{\text{Leb}}$ the set of Lebesgue-measurable subsets of $\mathbf{R}$.

In fact $\mathscr{M}_{\text{Leb}}$ is a σ-algebra.

Definition 144 (Lebesgue measure). The *Lebesgue measure*, which we denote by μ_{Leb}, is the restriction of μ_{out} to $\mathscr{M}_{\text{Leb}}$.

Then $(\mathbf{R}, \mathscr{M}_{\text{Leb}}, \mu_{\text{Leb}})$ is a measure space.

Definition 145. The *Borel σ-algebra*, denoted $\mathscr{M}_{\text{Bor}}$, is the σ-algebra generated by open subsets of $\mathbf{R}$.

Theorem 146 (Radon-Nikodym theorem). *Let $(X, \mathscr{M})$ be a measurable space. Let μ and ν be σ-finite measures on $(X, \mathscr{M})$. If ν is absolutely continuous with respect to μ then there exists a measurable function $f : X \longrightarrow \mathbf{R}_{\geq 0}$ such that*

$$\nu(A) = \int_A f \, \mathrm{d}\mu \tag{A.20}$$

for all $A \in \mathscr{M}$; and f is unique up to equality μ-almost everywhere.

Proof. A proof is given by Kallenberg (1997, Thm A1.3, p. 456). □

The function f is called the *Radon-Nikodym* derivative and is denoted

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}, \tag{A.21}$$

and may be called the *density* of ν with respect to μ.

A.2.3 Probability spaces

Definition 147 (probability measure). A measure P on $(\Omega, \mathscr{M})$ is a *probability measure* if $P(\Omega) = 1$. Then the measure space $(\Omega, \mathscr{M}, P)$ is called a *probability space*.

If P is a probability measure, then $P(A) \in [0,1]$, for all $A \in \mathscr{M}$. As is customary, when dealing with a probability measure P, we say 'P-almost sure' instead of 'P-almost everywhere'.

Definition 148 (random variable). A *random element* of $(\Sigma, \mathscr{N})$ is a measurable function $X : (\Omega, \mathscr{M}) \longrightarrow (\Sigma, \mathscr{N})$, for some measurable space $(\Omega, \mathscr{M})$. A *random variable* is simply a random element of $(\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$. A *random vector* is a random element of $(\mathbf{R}^n, \mathscr{M}^n_{\mathrm{Bor}})$, where $\mathscr{M}^n_{\mathrm{Bor}}$ is the σ-algebra of Borel subsets of $\mathbf{R}^n$.

Remark 149. Let $(\Omega, \mathscr{M}, P)$ be a probability space and let $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ be a random variable. Recall that P_X is the push-forward of P along X. Let $B \in \mathscr{M}_{\mathrm{Bor}}$ be any Borel set. To recover the usual notion of the 'probability that X takes its value in B', we may introduce the following notation

$$p(X \in B) := P_X(B) \tag{A.22}$$

$$P(\{\omega \in \Omega : X(\omega) \in B\}). \tag{A.23}$$

Recall the definition of a random process from Chapter 2.

Definition 150 (random process). A *random process* (also *stochastic process*, *random function*, or just *process*) is an indexed family of random variables on a probablity space $\mathbf{\Omega}$, which we denote $\mathbf{X} = \{X_t : \mathbf{\Omega} \longrightarrow \mathbf{R}\}_{t \in T}$. We call T the *index set*, and any element $t \in T$ an *index*.

Following (Kallenberg, 1997, p. 92), we let $\hat{T}$ denote the set of finite subsets of T. For $A, B \in \hat{T}$, with $A \subseteq B$, there is a projection

$$\begin{aligned} \pi_{B,A} : \mathbf{R}^B &\longrightarrow \mathbf{R}^A \\ (x_t)_{t \in B} &\longmapsto (x_t)_{t \in A}. \end{aligned}$$

A family $\mathbf{P} = (P_A : A \in \hat{T})$, where P_A is a probability measure on $\mathbf{R}^A$, indexed by $\hat{T}$, is *projective* if $P_B \circ \pi^{-1}_{B,A} = P_A$, for all $A \subseteq B \in \hat{T}$. A foundational result in the theory of random processes is the following, due to Kolmogorov.

Theorem 151 (Kolmogorov's extension theorem). *Let* $\mathbf{P} = (P_A : A \in \hat{T})$ *be a family of probability measures, as above. Then* $\mathbf{P}$ *is projective if and only if there is a random process* $\mathbf{X} = (X_t : t \in T)$ *such that* P_A *is the joint probability distribution of* $(X_t)_{t \in A}$*, for all* $A \in \hat{T}$.

Proof. A proof is given by Kallenberg (1997, p. 92, Theorem 5.16). □

A.2.4 Expectation and conditional expectation

For the rest of this chapter we fix a probability space $\mathbf{\Omega} := (\Omega, \mathscr{M}, P)$. Given a random variable $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$, and a set $A \in \mathscr{M}$, we write

$$P^X(A) := \int_A X \, \mathrm{d}P, \tag{A.24}$$

when the integral converges, when it does not we leave $P^X(A)$ undefined.

Definition 152 (expectation). Let $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ be a random variable which is absolutely P-integrable. The *expectation* (also *expected value*) of X is

$$\mathrm{E}(X) := P^X(\Omega) = \int_\Omega X \, \mathrm{d}P. \tag{A.25}$$

If X is not absolutely P-integrable then we leave the expectation of X undefined. We will say that a random variable has *finite expectation* to mean that it is absolutely P-integrable.

Definition 153 (conditional expectation). Let $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ be a random variable with finite expectation, and let $\mathscr{M}_0 \subseteq \mathscr{M}$ be a σ-algebra. A *conditional expectation* of X given $\mathscr{M}_0$ is a random variable $X_0 : (\Omega, \mathscr{M}_0) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ such that

$$\int_A X_0 \, \mathrm{d}P = \int_A X \, \mathrm{d}P, \text{ for all } A \in \mathscr{M}_0. \tag{A.26}$$

In other words, X_0 is such that $P^{X_0} = P^X|_{\mathscr{M}_0}$. We denote such a random variable X_0 by $\mathrm{E}(X \mid \mathscr{M}_0)$.

If $Y : (\Omega, \mathscr{M}) \longrightarrow (\Sigma, \mathscr{L})$ is another random variable, the *conditional expectation* of X given Y is $E(X \mid Y) := E(X \mid \sigma(Y))$.[1]

We will justify the existence and uniqueness (up to P-almost sure equality) of $E(X \mid \mathscr{M}_0)$ in Proposition 156.

Remark 154. We may think of $E(X \mid \mathscr{M}_0)$ as an approximation of X which is constrained by the requirement of being a measurable function $(\Omega, \mathscr{M}_0) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$. In particular, if $\mathscr{M}_\varnothing = \{\varnothing, \Omega\}$ denotes the most trivial σ-algebra on Ω, then $E(X \mid \mathscr{M}_\varnothing)$ is the constant function with value $E(X)$. To see this, note that any measurable function $(\Omega, \mathscr{M}_\varnothing) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ is constant.

Lemma 155. *Let $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ be a random variable with finite expectation. Then $P^X \ll P$.*

[1] Recall from Definition 141 that $\sigma(Y)$ is the σ-algebra generated by Y.

Proof. Let $A \in \mathscr{M}$ and suppose that $P(A) = 0$. Then any integral on A with respect to P will be zero, so in particular $P^X(A) = 0$. □

Proposition 156 (existence and uniqueness of conditional expectation). Let $X : (\Omega, \mathscr{M}) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}})$ be a random variable with finite expectation, and let $\mathscr{M}_0 \subseteq \mathscr{M}$ be a σ-algebra. The conditional expectation $E(X \mid \mathscr{M}_0)$ exists and is unique up to P-almost sure equality.

Proof. We have already argued in Lemma 155 that $P^X \ll P$. Therefore $P^X|_{\mathscr{M}_0} \ll P|_{\mathscr{M}_0}$. Moreover, both $P^X|_{\mathscr{M}_0}$ and $P|_{\mathscr{M}_0}$ are finite: in the former case by assumption that X has finite expectation, and in the latter case by assumption that P is a probability measure. Therefore, we may apply the Radon–Nikodym Theorem (for signed measures) to obtain a measurable function

$$X_0 := \frac{\mathrm{d}P^X|_{\mathscr{M}_0}}{\mathrm{d}P|_{\mathscr{M}_0}} : (\Omega, \mathscr{M}_0) \longrightarrow (\mathbf{R}, \mathscr{M}_{\mathrm{Bor}}), \tag{A.27}$$

which is absolutely $P|_{\mathscr{M}_0}$-integrable such that

$$P^X|_{\mathscr{M}_0}(A) = \int_A X_0 \,\mathrm{d}P|_{\mathscr{M}_0}, \text{ for all } A \in \mathscr{M}_0. \tag{A.28}$$

Note that X_0 is unique up to P-almost sure equality. Unpacking the equation in (A.28), we have

$$\int_A X \,\mathrm{d}P = \int_A X_0 \,\mathrm{d}P|_{\mathscr{M}_0}, \text{ for all } A \in \mathscr{M}_0. \tag{A.29}$$

Finally, since integration is ‘insensitive to refinement’ (see (Tao, 2011, Exercise 1.4.40(v))), X_0 is absolutely P-integrable and

$$\int_A X \,\mathrm{d}P = \int_A X_0 \,\mathrm{d}P, \text{ for all } A \in \mathscr{M}_0, \tag{A.30}$$

as required. □

Appendix B

Derivatives of the squared-exponential covariance function

The *squared-exponential* covariance function (Ch. 2, Ex. 58) is given by the formula

$$k(s,t) := \sigma^2 \exp\left(-\frac{1}{2}(s-t)^{\mathrm{t}}M(s-t)\right). \tag{B.1}$$

Its first- and second-order partial derivatives are given by the formulas

$$\frac{\partial k(s,t)}{\partial s} = -k(s,t)M(s-t), \tag{B.2}$$

$$\frac{\partial k(s,t)}{\partial t} = -\frac{\partial k(s,t)}{\partial s}, \tag{B.3}$$

$$\frac{\partial^2 k(s,t)}{\partial s \partial s^{\mathrm{t}}} = k(s,t)M\left((s-t)(s-t)^{\mathrm{t}}M - I\right), \tag{B.4}$$

$$\frac{\partial^2 k(s,t)}{\partial s \partial s^{\mathrm{t}}} = \frac{\partial^2 k(s,t)}{\partial s \partial s^{\mathrm{t}}}, \text{ and} \tag{B.5}$$

$$\frac{\partial^2 k(s,t)}{\partial s \partial s^{\mathrm{t}}} = -\frac{\partial^2 k(s,t)}{\partial s \partial s^{\mathrm{t}}}. \tag{B.6}$$

Bibliography

Adler, R. J. (1981), *The geometry of random fields*, SIAM, Philadelphia.

Adler, R. J. (2010), *The geometry of random fields*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Amari, S.-I. (1985), *Differential-geometrical methods in statistics*, Springer-Verlag, New York.

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Transactions of the American Mathematical Society* **68**(3), 337–404.

Baes, M. and Dejonghe, H. (2002), 'The Hernquist model revisited: completely analytical anisotropic dynamical models', **393**, 485–497.

Bartlett, M. S. (1947), 'The use of transformations', *Biometrics* **3**(1), 39–52.

Battaglia, G., Helmi, A., Tolstoy, E., Irwin, M., Hill, V. and Jablonka, P. (2008), 'The kinematic status and mass content of the Sculptor dwarf spheroidal galaxy', *The astrophysical journal* **681**, L13–L16.

Berlinet, A. and Thomas-Agnan, C. (2004), *Reproducing kernel Hilbert spaces in probability and statistics*, Kluwer Academic Publisher, Boston.

Binney, J. and Tremaine, S. (2008), *Galactic dynamics*, second edition edn, Princeton University Press, Princeton.

Blumenthal, G. R., Faber, S. M., Primack, J. R. and Rees, M. J. (1984), 'Formation of galaxies and large-scale structure with cold dark matter', *Nature* **311**, 517–25.

Böhm, A. (1978), *The rigged Hilbert space and quantum mechanics: lectures in mathematical physics at the University of Texas*, Lecture notes in physics, Springer-Verlag, Berlin.

Bonilla, E. V., Chai, K. M. and Williams, C. (2008), Multi-task gaussian process prediction, *in* J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds, 'Advances in Neural information processing systems 20', Curran Associates, Inc., pp. 153–160.

Bourbaki, N. (1981), *Topological vector spaces: chapters 1–5*, Springer, Berlin.

Bourbaki, N. (1989), *Algebra I: chapters 1–3*, Springer, Berlin.

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S. and Lacey, C. G. (2006), 'Breaking the hierarchy of galaxy formation', *Monthly notices of the Royal Astronomical Society* **370**(2), 645–655.

Bower, R. G., Vernon, I., Goldstein, M., Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S. and Frenk, C. S. (2010), 'The parameter space of galaxy formation', *Monthly notices of the Royal Astronomical Society* **407**(4), 2017–2045.

Box, G. E. P. and Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2), 211–252.

Boylan-Kolchin, M., Bullock, J. S. and Kaplinghat, M. (2011), 'Too big to fail? the puzzling darkness of massive milky way subhaloes', *Monthly Notices of the Royal Astronomical Society: Letters* **415**(1), L40–L44.

Brandt, T. D. (2016), 'Constraints on macho dark matter from compact stellar systems in ultra-faint dwarf galaxies', *The astrophysical journal* **824**, L31.

Breddels, M. A. and Helmi, A. (2013), 'Model comparison of the dark matter profiles of Fornax, Sculptor, Carina and Sextans', **558**, A35.

Carollo, C. M., de Zeeuw, P. T. and van der Marel, R. P. (1995), 'Velocity profiles of Osipkov-Merritt models', *Monthly notices of the Royal Astronomical Society* **276**.

Courant, R. and Hilbert, D. (1989), *Methods of mathematical physics*, Vol. 1, John Wiley & Sons, New York.

Cressie, N. (1986), 'Kriging nonstationary data', *Journal of the American Statistical Association* **81**(395), 625–634.

Cressie, N. (1988), 'Spatial prediction and ordinary kriging', *Mathematical geology* **20**(4), 405–21.

Cressie, N. (1990), 'The origins of kriging', *Mathematical geology* **22**(3), 239–52.

Cuddeford, P. and Louis, P. (1995), 'Spherical galaxian distribution functions with adjustable anisotropy', *Monthly notices of the Royal Astronomical Society* **275**(4), 1017–27.

de Blok, W. J. G. (2010), 'The core-cusp problem', *Advances in astronomy* **2010**, 1–14.

Diemand, J., Kuhlen, M., Madau, P., Zemp, M., Moore, B., Potter, D. and Stadel, J. (2008), 'Clumps and streams in the local dark matter distribution', *Nature* **454**(7205), 735–738.

Draguljić, D., Santner, T. J. and Dean, A. M. (2012), 'Noncollapsing space-filling designs for bounded nonrectangular regions', *Technometrics* **54**(2), 169–178.

Edwards, A. W. F. (1972), *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference*, Cambridge University Press, London.

Erlang, A. K. (1909), 'The theory of probabilities and telephone conversations', *Nyt tidsskrift for matematik B* **20**, 33–9.

Evans, T. M., Aigrain, S., Gibson, N., Barstow, J. K., Amundsen, D. S., Tremblin, P. and Mourier, P. (2015), 'A uniform analysis of HD209458b Spitzer/IRAC lightcurves with Gaussian process models', *Monthly notices of the Royal Astronomical Society* **451**(1), 680–694.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–68.

Flores, R. A. and Primack, J. R. (1994), 'Observational and theoretical constraints on singular dark matter halos', *The astrophysical journal* **427**, L1.

Forrester, A., Sobester, A. and Keane, A. (2008), *Engineering design via surrogate modelling: a practical guide*, Wiley, Chichester.

Frenk, C. and White, S. (2012), 'Dark matter and cosmic structure', *Annalen der Physik* **524**(9–10), 507–534.

Gerhard, O. E. (1991), 'A new family of distribution functions for spherical galaxies', *Monthly notices of the Royal Astronomical Society* **250**(4), 812–830.

Gibson, N. P., Aigrain, S., Roberts, S., Evans, T. M., Osborne, M. and Pont, F. (2012), 'A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy', *Monthly notices of the Royal Astronomical Society* **419**(3), 2683–2694.

Gikhman, I. I. and Skorohod, A. V. (1974), *The theory of stochastic processes*, Springer-Verlag, Berlin.

Goldstein, A. A. and Price, J. F. (1971), 'On descent from local minima', *Mathematics of computation* **25**(115), 569–74.

Gration, A. L. and Wilkinson, M. I. (2019), 'Dynamical modelling of dwarf spheroidal galaxies using Gaussian-process emulation', *Monthly notices of the Royal Astronomical Society* **485**(4), 4878–92.

Griest, K. (1991), 'Galactic microlensing as a method of detecting massive compact halo objects', *The astrophysical journal* **366**, 412.

Guttorp, P. and Gneiting, T. (2006), 'Studies in the history of probability and statistics: XLIX. On the Matérn correlation family', *Biometrika* **93**(4), 989–995.

Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B. J. and Wagner, C. (2009), 'The Coyote Universe II: cosmological models and precision emulation of the nonlinear matter power spectrum', *The astrophysical journal* **705**, 156–174.

Helmi, A. (2008), 'The stellar halo of the Galaxy', *The Astronomy and Astrophysics Review* **15**(3), 145–188.

Henderson, C. R. (1950), 'Estimation of genetic parameters', *Annals of mathematical statistics* (21), 309–310.

Hernquist, L. (1990), 'An analytical model for spherical galaxies and bulges', *The astrophysical journal* **356**, 359–64.

Ivanov, A. A. and Leonenko, N. (1989), *Statistical analysis of random fields*, Kluwer Academic Publisher, Dordrecht.

Jones, D. R., Schonlau, M. and Welch, W. J. (1998), 'Efficient global optimization of expensive black-box functions', *Journal of global optimization* **13**(4), 455–492.

Kallenberg, O. (1997), *Foundations of modern probability*, Springer, New York.

Kleyna, J. T., Wilkinson, M. I., Evans, N. W. and Gilmore, G. (2004), 'A photometrically and kinematically distinct core in the Sextans dwarf spheroidal galaxy', *Monthly notices of the Royal Astronomical Society* **354**(4), L66–L72.

Kleyna, J., Wilkinson, M. I., Evans, N. W., Gilmore, G. and Frayn, C. (2002), 'Dark matter in dwarf spheroidals—ii. observations and modelling of draco', *Monthly notices of the Royal Astronomical Society* **330**(4), 792–806.

Klypin, A., Kravtsov, A. V., Valenzuela, O. and Prada, F. (1999), 'Where are the missing galactic satellites?', *The astrophysical journal* **522**(1), 82–92.

Koch, A., Kleyna, J. T., Wilkinson, M. I., Grebel, E. K., Gilmore, G. F., Evans, N. W., Wyse, R. F. G. and Harbeck, D. R. (2007), 'Stellar kinematics in the remote leo ii dwarf spheroidal galaxy—another brick in the wall', *The astronomical journal* **134**(2), 566––578.

Krige, D. G. (1951), A statistical approach to some mine valuation and allied problems on the Witwatersrand, Master's thesis, University of the Witwatersrand.

Lang, S. (2004), *Linear algebra*, Springer, New York.

Loeppky, J. L., Sacks, J. and Welch, W. J. (2009), 'Choosing the sample size of a computer experiment: a practical guide', *Technometrics* **51**(4), 366–376.

Mandel, K. and Agol, E. (2002), 'Analytic light curves for planetary transit searches', *The astrophysical journal* **580**(2), L171–L175.

Mashchenko, S., Wadsley, J. and Couchman, H. M. P. (2008), 'Stellar feedback in dwarf galaxy formation', *Science* **319**, 174–177.

Mateo, M., Olszewski, E. W. and Walker, M. G. (2008), 'the velocity dispersion profile of the remote dwarf spheroidal galaxy leo i: a tidal hit and run?', *The astrophysical journal* **675**, 201––233.

Matérn, B. (1986), *Spatial Variation*, Lecture Notes in Statistics, second edn, Springer-Verlag, New York.

Matheron, G. (1963), 'Principles of geostatistics', *Economic geology* **58**(8), 1246–1266.

McConnachie, A. W. (2012), 'The observed properties of dwarf galaxies in and around the Local Group', *The astronomical journal* **144**(1), 4.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979), 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics* **21**(2), 239–245.

Merritt, D. (1985), 'Spherical stellar systems with spheroidal velocity distributions', *The astronomical journal* **90**, 1027–1037.

Moore, B. (1994), 'Evidence against dissipation-less dark matter from observations of galaxy haloes', *Nature* **370**(6491), 629–631.

Moore, B., Ghigna, S., Governato, F., Lake, G., Quinn, T., Stadel, J. and Tozzi, P. (1999), 'Dark matter substructure within galactic halos', *The astrophysical journal* **524**(1), L19–L22.

Moore, C. J., Berry, C. P. L., Chua, A. J. K. and Gair, J. R. (2016), 'Improving gravitational-wave parameter estimation using Gaussian process regression', *Physical Review D* **93**(6).

Navarro, J. F., Eke, V. R. and Frenk, C. S. (1996), 'The cores of dwarf galaxy haloes', *Monthly notices of the Royal Astronomical Society* **283**(3), L72–L78.

Osipkov, L. P. (1979), 'Spherical systems of gravitating bodies with an ellipsoidal velocity distribution', *Pisma v astronomicheskii zhurnal* **5**, 77–80.

Ostriker, J. P., Choi, E., Chow, A. and Guha, K. (2019), 'Mind the gap: is the too big to fail problem resolved?', available at `https://arxiv.org/abs/1904.10471`.

Parzen, E. (1959), Statistical inference on time series by Hilbert space methods I, Technical Report 23, Department of Statistics, Stanford University.

Penzias, A. A. and Wilson, R. W. (1965), 'A measurement of excess antenna temperature at 4080 Mc/s.', *The astrophysical journal* **142**, 419–421.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., Groom, D. E., Hook, I. M., Kim, A. G., Kim, M. Y., Lee, J. C., Nunes, N. J., Pain, R., Pennypacker, C. R., Quimby, R., Lidman, C., Ellis, R. S., Irwin, M., McMahon, R. G., Ruiz-Lapuente, P., Walton, N., Schaefer, B., Boyle, B. J., Filippenko, A. V., Matheson, T., Fruchter, A. S., Panagia, N., Newberg, H. J. M., Couch, W. J. and Project, T. S. C. (1999), 'Measurements of Ω and Λ from 42 High-Redshift Supernovae', *The astrophysical journal* **517**(2), 565–586.

Planck Collaboration (2018), 'Planck 2018 results. VI. Cosmological parameters', available at `https://ui.adsabs.harvard.edu/abs/2018arXiv180706209P`.

Plummer, H. C. (1911), 'On the problem of distribution in globular star clusters', *Monthly notices of the Royal Astronomical Society* **71**, 460–70.

Primack, J. R. (2009), Dark matter and galaxy formation, *in* F. Roig, D. Lopes, R. de La Reza and V. Ortega, eds, 'American Institute of Physics Conference Series', Vol. 1192 of *American Institute of Physics Conference Series*, pp. 101–37.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian processes for machine learning*, MIT Press, Cambridge.

Read, J. I. and Gilmore, G. (2005), 'Mass loss from dwarf spheroidal galaxies: the origins of shallow dark matter cores and exponential surface brightness profiles', *Monthly notices of the Royal Astronomical Society* **356**(1), 107–124.

Read, J. I. and Steger, P. (2017), 'How to break the density-anisotropy degeneracy in spherical stellar systems', *Monthly notices of the Royal Astronomical Society* **471**(4), 4541–4558.

Read, J. I. and Steger, P. (2018), 'The case for a cold dark matter cusp in Draco', *Monthly notices of the Royal Astronomical Society* **481**(1), 860–877.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B. and Tonry, J. (1998), 'Observational evidence from supernovae for an accelerating universe and a cosmological constant', *The astronomical journal* **116**(3), 1009–1038.

Rubin, V. C. and Ford Jr., W. K. (1970), 'Rotation of the Andromeda nebula from a spectroscopic survey of emission regions', *The astrophysical journal* **159**, 379–403.

Rubin, V. C., Ford Jr, W. K. and Thonnard, N. (1978), 'Extended rotation curves of high-luminosity spiral galaxies. IV. Systematic dynamical properties, Sa $\longrightarrow$ Sc', *The astrophysical journal* **225**, 107–11.

Rudin, W. (1991), *Functional analysis*, second edn, McGraw-Hill, New York.

Rutherford, E., Geiger, H. and Bateman, H. (1910), 'The probability variations in the distribution of α-particles', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **20**(118), 698–707.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989), 'Design and analysis of computer experiments', *Statistical science* **4**(4), 409–23.

Sale, S. and Magorrian, J. (2014), 'Three-dimensional extinction mapping using Gaussian random fields', *Monthly notices of the Royal Astronomical Society* **445**(1), 256–269.

Sale, S. and Magorrian, J. (2019), 'Large-scale three-dimensional gaussian process extinction mapping', *Monthly notices of the Royal Astronomical Society* **481**(1), 494–508.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003), *The design and analysis of computer experiments*, Springer, New York.

Schaye, J., Crain, R. A., Bower, R. G., Furlong, M., Schaller, M., Theuns, T., Dalla Vecchia, C., Frenk, C. S., McCarthy, I. G., Helly, J. C., Jenkins, A., Rosas-Guevara, Y. M., White, S. D. M., Baes, M., Booth, C. M., Camps, P., Navarro, J. F., Qu, Y., Rahmati, A., Sawala, T., Thomas, P. A. and Trayford, J. (2015), 'The EAGLE project: simulating the evolution and assembly of galaxies and their environments', *Monthly notices of the Royal Astronomical Society* **446**(1), 521–554.

Schonlau, M. and Welch, W. J. (1996), Global optimization with nonparametric function fitting, *in* 'Proceedings of the ASA', American Statistical Association, pp. 183–186.

Schramm, D. N. and Turner, M. S. (1998), 'Big-bang nucleosynthesis enters the precision era', *Rev. Mod. Phys.* **70**, 303–318.

Seber, G. A. F. and Lee, A. J. (2003), *Linear regression analysis*, Wiley, Hoboken.

Serre, D. (2002), *Matrices: theory and applications*, Graduate texts in mathematics, Springer, New York.

Shapley, H. (1938), 'Two stellar systems of a new kind', *Nature* **142**(3598), 715–716.

Springel, V., Wang, J., Vogelsberger, M., Ludlow, A., Jenkins, A., Helmi, A., Navarro, J. F., Frenk, C. S. and White, S. D. M. (2008), 'The Aquarius Project: the subhaloes of galactic haloes', *Monthly notices of the Royal Astronomical Society* **391**(4), 1685–1711.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J. and Pearce, F. (2005), 'Simulations of the formation, evolution and clustering of galaxies and quasars', **435**(7042), 629–636.

Stirzaker, D. (2000), 'Advice to hedgehogs, or, constants can vary', *The mathematical gazette* **84**(500), 197–210.

Strigari, L. E., Frenk, C. S. and White, S. D. M. (2010), ‘Kinematics of Milky Way satellites in a Lambda cold dark matter universe’, *Monthly notices of the Royal Astronomical Society* **408**(4), 2364–2372.

Tao, T. (2011), *An introduction to measure theory*, American Mathematical Society, Providence, Rhode Island.

Ural, U., Wilkinson, M. I., Read, J. I. and Walker, M. G. (2015), ‘A low pre-infall mass for the Carina dwarf galaxy from disequilibrium modelling’, *Nature Communications* **6**(1).

Vazquez, E. and Bect, J. (2010), ‘Convergence properties of the expected improvement algorithm with fixed mean and covariance functions’, *Journal of Statistical Planning and Inference* **140**(11), 3088–3095.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Bird, S., Nelson, D. and Hernquist, L. (2014), ‘Properties of galaxies reproduced by a hydrodynamic simulation’, *Nature* **509**(7499), 177–182.

Walker, M. G., Mateo, M., Olszewski, E. W., Peñarrubia, J., Evans, N. W. and Gilmore, G. (2009), ‘A universal mass profile for dwarf spheroidal galaxies?’, *The astrophysical journal* **704**(2), 1274.

Walker, M. G., Mateo, M., Olszewski, E. W., Sen, B. and Woodroofe, M. (2009), ‘Clean kinematic samples in dwarf spheroidals: an algorithm for evaluating membership and estimating distribution parameters when contamination is present’, *The Astronomical Journal* **137**(2), 3109–3138.

Wasserman, L. (2004), *All of statistics: a concise course in statistical inference*, Springer, New York.

Wasserman, L. (2007), *All of nonparametric statistics: a concise course in nonparametric statistical inference*, Springer, New York.

Wiener, N. (1923), ‘Differential space’, *Journal of mathematics and physics* **2**(1–4), 131–74.

Wilk, M. B. and Gnanadesikan, R. (1968), ‘Probability plotting methods for the analysis of data’, *Biometrika* **55**(1), 1–17.

Wilkinson, M. I., Kleyna, J., Evans, N. W. and Gilmore, G. (2002), ‘Dark matter in dwarf spheroidals—I. Models’, *Monthly notices of the Royal Astronomical Society* **330**, 778–791.

Wolf, J. and Bullock, J. S. (2012), 'Dark matter concentrations and a search for cores in Milky Way dwarf satellites', available at `https://arxiv.org/abs/1203.4240`.

Wolf, J., Martinez, G. D., Bullock, J. S., Kaplinghat, M., Geha, M., Munoz, R. R., Simon, J. D. and Avedo, F. F. (2010), 'Accurate masses for dispersion-supported galaxies', *Monthly notices of the Royal Astronomical Society* .

Yaglom, A. M. (1961), Second-order homogeneous random fields, *in* 'Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. 2—contributions to probability theory', University of California Press, Berkeley, pp. 593–622.

Yaglom, A. M. (1962), *An introduction to the theory of stationary random functions*, Prentice-Hall, Englewood Cliffs.

Zhao, H. (1996), 'Analytical models for galactic nuclei', *Monthly notices of the Royal Astronomical Society* **278**, 488–96.

Zwicky, F. (1933), 'Die Rotverschiebung von extragalaktischen Nebeln', *Helvetica Physica Acta* **6**, 110–27. Reprint, Zwicky, F. (2009), 'The redshift of extragalactic nebulae', *General relativity and gravity* **41**, 207–24.