Investigations into the robustness of statistical decisions

James Watson

Supervised by Professor Chris Holmes

Department of Statistics

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

Corpus Christi College January 2016

To C.L.R.P. a dear random variable.

Declaration

Some work in this thesis is the product of a collaborative effort. In particular:

- Chapters 2 & 4 are joint work with my supervisor Chris Holmes, and a large part has been written up as a publication (Watson and Holmes, 2014) and accepted for publication in Statistical Science.
- Chapter 3 is joint work with Luis Nieto-Barajas and Chris Holmes, and has been written up as a publication (Watson et al., 2014) and submitted to Statistics.
- Chapter 6 is joint work with Tristan Grey-Davies, Luis Nieto-Barajas and Chris Holmes.

James Watson
January 2016

Acknowledgements

The four and a half years spent in Oxford at the Department of Statistics and at the Doctoral Training Centre have been challenging and fulfilling. I would like to thank the following people who have made the experience particularly enjoyable and fruitful.

This work was made possible by funding from the SABS-IDC and Hoffman-La Roche. At the DTC, I would especially like to thank Charlotte Deane for her support and patience during the early stages of the programme; at Roche, Alexander Manta, whose enthusiasm and engaging conversations motivated my interest in Bayesian statistics.

At the department, people have always been happy to help out and discuss problems. In particular, Tristan Grey-Davies, George Nicholson, Aimee Taylor, Pierre Jacob and Louis Aslett (aka the R guru) who have provided much advice and helpful suggestions.

Through Corpus Christi College, I've been lucky to meet some of the stranger species who study the Humanities. Some of them even decided to move in with me. Colm, Mara and Anna, thank you for putting up with my foul moods and enlightening my mind with Church politics, obscure French philosophy and rare manuscripts.

A huge thanks to Thomas, Garance, Colm, Jake, Aimee and Pierre who all have helped me out in the final stages when the money started to run dry and when I found myself homeless more than once; the bon vivant froggies, Edouard, Harold and Yacine, who have been reliable pub companions; Ollie Harriman and Ben Bussman great sporting partners.

I am very grateful to my parents, brother and sister who have always been unswervingly supportive and caring.

Finally, I am deeply indebted to my supervisor Chris Holmes, whose creativity and knowledge has been inspiring. Without his support I would certainly not have got to the end of it. I have to say, however, the work he gave me was never elementary. I am grateful for his constant guidance, at the same time as giving me the freedom to choose what to do.

Abstract

Decision theory is a cornerstone of Statistics, providing a principled framework in which to act under uncertainty. It underpins Bayesian theory via the Savage axioms, game theory via Wald's minimax, and supplies a mathematical formulation of 'rational choice'. This thesis argues that its role is of particular importance in the so-called 'big data' era. Indeed, as data have become larger, statisticians are confronted with an explosion of new methods and algorithms indexing ever more complicated statistical models. Many of these models are not only high-dimensional and highly non-linear, but are also approximate by design, e.g. deliberately making approximations for reasons such as tractability and interpretation. For Bayesian theory, and for Statistics in general, this raises many important questions, which I believe decision theory can help elucidate.

From a foundational standpoint, how does one interpret the outputs of Bayesian computations when the model is known to be approximate and misspecified? Concerns of misspecification violate the necessary assumptions for the use of the Savage axioms. Should principles such as expected loss minimisation apply in such settings? On a practical level, how can modellers assess the extent of the impact of model misspecification? How can this be integrated into the process of model construction in order to inform the user whether more work needs to be done (for example, more hours of computation, or a more accurate model)? They need to know whether the model is unreliable, or whether the conclusions of the model are robust and can be trusted. In the history of *Robust Statistics*, whose main aspects are covered in Chapter 1, there has been periodic concern with misspecification. *Robust Bayesian analysis* was a particularly active area of research through the 1980s to mid-90s, but later declined due to methodological and computational advances which overcame original concerns of misspecification.

Now, however, the complexity of datasets frequently prohibits the possibility of constructing fully specified and well-crafted models and therefore Bayesian robustness merits a reappraisal. Additionally, new methods have been developed which are characterised by their deliberately approximate and misspecified nature, such as integrated nested Laplace approximation (INLA), approximate Bayesian computation (ABC), Variational Bayes, and composite likelihoods. These all start with a premise of misspecification.

The work described in this thesis concerns the development of a comprehensive framework addressing challenges associated with imperfect models, encompassing both formal methods to assess the sensitivity of the model (Chapters 2 & 3), and diagnostic exploratory methods via graphical plots and summary statistics (Chapters 4 & 5). This framework is built on a *post hoc* sensitivity analysis of the posterior approximating model via the loss function.

Chapter 2 describes methods for estimating the sensitivity of a model with respect to the loss function by analysing the effect of local perturbations in neighbourhoods centred at the approximating model (in a Bayesian context this would be the posterior distribution). These neighbourhoods are defined using the Kullback-Leibler divergence. This approach provides a bridge between the two dominant paradigms in decision theory: Wald's minimax and Savage's expected loss criterion. Two key features of this framework are that the solution is analytical, and it unifies other well known methods in Statistics such as predictive tempering, power likelihoods and Gibbs posteriors. It also offers an interesting solution to the Ellsberg paradox. Another application of the work is in the area of computational decision theory where the statistician only has access to the model via a finite set of samples. In this context, the methods can be used at very little extra computational cost.

Chapter 3 considers nonparametric extensions to the approximating reference model. In particular, we look at the Pólya tree process, the Dirichlet process and bootstrap procedures. Again using the Kullback-Leibler divergence, it is possible to characterise random samples of these nonparametric models with respect to the base model, and therefore understand the effect of local perturbations on the distribution of loss of the approximating model.

A series of diagnostic plots and summary statistics are presented in Chapter 4, and further illustrated in Chapter 5 by means of two applications taken from the medical decision-making literature. These complete the framework of *post-hoc* assessment of model stability and allow the user to understand why the model might be sensitive to misspecification. Graphical displays are an essential part of statistical analyses, indeed the point of departure for any serious data analysis. Their use in model exploration in the context of decision theory, however, is not common. We borrow some ideas from finance and econometrics as a basis of exploratory decision-system plots. Other plots come as natural consequences of the methodology from the two previous chapters.

The final chapter examines a very specific application of statistical decision theory, notably the analysis of randomised clinical trials to assess the evidence in favour of patient heterogeneity. This problem, known as *subgroup analysis*, has traditionally been solved using predictive models which are a proxy for the real object of interest: evidence of patient heterogeneity. By formally expressing the decision problem as a hypothesis test, and working from first principles, the problem is shown to be in fact much easier than previously thought.

The method avoids issues involving counterfactuals by testing decision rules against their mirror images. It can harness the strength of well known model free tests and uses a random forest-type approach for *post-hoc* exploration of decision rules. The randomisation allows for a causal interpretation of the results.

Table of contents

Chapter 1

Introduction

> "Le doute n'est pas un état bien agréable, mais l'assurance est un état ridicule". Voltaire

> "We are not interested in falsifying our model for its own sake among other things, having built it ourselves, we know all the shortcuts taken in doing so, and can already be morally certain it is false". Gelman and Shalizi (2013)

1.1 Motivation

A mathematical model of a real phenomenon must make simplifying assumptions in order to have practical use. The model must be analytically and computationally tractable. A key question of scientific interest is how the outputs of the model are sensitive to these assumptions.

This dissertation attempts to provide a principled framework with which it is possible to quantify robustness to model assumptions in the context of probabilistic statistical models. This is an especially topical contribution given the recent explosion in data complexity due to increased collection and storage capacity, greater CPU speeds and advances in parallel architectures, all part of the so-called 'big data' era. The rise in complex data has been followed by a surge in computational methods such as MCMC that allow the specification of rich classes of statistical models. In this way, many historical modelling constraints can be avoided, for example, conjugacy in Bayesian analysis. However, many of these models are approximate by design, deliberately making assumptions known to be false, for the purposes of speed and tractability. Assessing robustness to misspecification[1] in these

[1]Here *misspecification* is left deliberately vague but can be defined as not using the algorithms (models) of choice if the statistician had infinite time and computing power. I define it at more length in section 1.3.4. From

algorithms is therefore essential in order for practitioners to be confident in their benefit as decision-informing entities.

The main argument is as follows. Model robustness is contextual and thus should be considered from a decision theoretic perspective. If it is possible to define a loss (negative utility) function, then this should be used to quantify the consequence and importance of possible misspecifications. I focus on decision theory in the Bayesian paradigm (under the Savage axioms) and look at how this can be adapted to take into account model uncertainty. This is the $\mathscr{M}_{\text{open}}$ scenario as opposed to $\mathscr{M}_{\text{closed}}$, which is the standard Bayesian framework, conditional on the model being true (see Bernardo and Smith, 1994, section 6.1.2). In this respect, many of the ideas are not new but follow in a tradition of eminent practitioners and theorists, for example Berger (1984); Box and Draper (1987); Gilboa and Schmeidler (1989); Good (1950); Hansen and Sargent (2008) to name but a few references. However the proposed framework is *ex-post*, in the sense that it is designed to assess a statistical model's capacity to inform robust decisions once the statistician's 'best guess' model has been constructed. From a decision-theoretic perspective, I shall argue that only the marginal distribution over the parameters that enter into the loss function should be used to test robustness (throwing away the 'nuisance' parameters). This is done by considering neighbourhoods around the reference model (or 'best guess' model) which could be a Bayesian posterior distribution, but in practice the methodology is agnostic as to this position. The framework provides a bridge between the two dominant and separate decision paradigms, minimaxity (Wald, 1950) and expected loss (Savage, 1954). The setting of decision-making in the context of model misspecification is denoted $\mathscr{D}_{\text{open}}$, in analogy to $\mathscr{M}_{\text{open}}$. This thesis provides a framework for working in $\mathscr{D}_{\text{open}}$ via formal methodology, diagnostic plots and summary statistics (chapters 2, 3 and 4). The final two chapters present typical case studies in medical decision-making and an alternative method for approaching a statistical decision-theory problem.

A primary interest is quantifying the robustness of statistical models where the distribution in question is only accessible via a 'bag of Monte Carlo samples'. This may seem restrictive, but is the case for many applications. The framework naturally extends to continuous distributions, under certain constraints, but the discrete perspective allows for the development of off-the-shelf tools to quantify robustness via summary statistics and graphical visualisation. Indeed, few graphical methods exist to my knowledge for visualising decision robustness as compared to those for model checking or for convergence in Monte Carlo algorithms. Andrew Gelman's opinion on model checking is (as quoted in: Michael Jordan, *What are the open problems in Bayesian statistics?* ISBA bulletin, 2011)

a philosophical standpoint, misspecification is always an issue because of the "unknown unknowns" (Donald Rumsfeld's argument regarding the minimax-style decision by the United States to invade Iraq, Department of Defense news briefing, February 2002).

> For model checking, a key open problem is developing graphical tools for understanding and comparing models. Graphics is not just for raw data; rather, complex Bayesian models give opportunity for better and more effective exploratory data analysis.

This thesis focuses on graphical visualisation as an essential tool for model introspection, and robustness validation. This is related to ideas given in Kerman et al. (2008), but directed at visualising the sensitivity of a statistical decision system, via the model and loss function.

In the remainder of this chapter, I provide a brief introduction to some of the ideas in statistical robustness, alongside the major references, in order to give a flavour of the main ideas in the field. I also introduce the two dominant decision-theoretic paradigms and provide some of the necessary background by which the work is inspired.

1.2 Robust Statistics

1.2.1 Early Developments

In a statistical context, the word *robustness* was first coined by Box (1953) when analysing the effect of non-Gaussian data in order to test the equality of variances. This differed from its original 18^{th} century meaning, when "the word 'robust' was used to refer to someone who was strong, yet boisterous, crude and vulgar" (Stigler, 1973). Indeed, robustness now plays a central role in statistical methodology as a highly desirable and necessary property, indicative of a healthy statistical approach, especially with a recent increase in the complexity of both data and accompanying models. The following quote from Kadane (1984) may not be an overstatement:

> Robustness is a fundamental issue for all statistical analyses; in fact it might be argued that robustness is the subject of statistics.

Before Box, many eminent statisticians understood the dangers of blindly following model assumptions. Historically, the most controversial of these assumptions was the Gaussian distribution for measurement errors. A fundamental task that necessitated statistical reasoning during the infancy of the field was the problem of estimating a single location parameter, given noisy data, where almost all the noise was due to measurement error. Bessel (1818) and Newcomb (1886) noticed longer tails than normality would predict in their observations. Indeed, Newcomb commented on the law of normal errors (as it was then known), page 343, saying:

> As a matter of fact, however, the cases are quite exceptional in which errors are found to really follow the law. The general rule is that much more than one per cent. [sic] of the errors exceed four times the probable error. In other words, it is nearly always found that some of the outstanding errors seem abnormally large.

Stigler (1973) argues that Newcomb (1886) gave the first principled approach to robust estimation by considering mixtures of Gaussians in order to deal with heavy tailed distributions. He also dealt with the as yet unsatisfying theory of rejection of outliers given by Pierce (1852).

However further reinforcement of the law of normal errors resulted from a now famous argument between Fisher and the astronomer Eddington. Edington (1914, page 147) advocated the use of the mean absolute deviation (L_1 norm, which is given by $\sum_{i=1}^{n} |\bar{x} - x_i|$) instead of the standard deviation (L_2 norm) as a measure of the variation in the data, claiming in a footnote: "This is contrary to the advice of most text-books but can be shown to be true". This prompted Fisher (1920) to publish a short note showing that the L_1 norm lost 12% efficiency (as measured by asymptotic relative efficiency: the limit of the ratio of the two estimators) if the true error mechanism is Gaussian, thus resolving the erroneous claim.

The full danger of approximate normality in observations was not exposed until Tukey (1960) analysed the following simple mixture model:

$$F(x) = (1-\varepsilon)\cdot\Phi(x) + \varepsilon\cdot\Phi(\frac{x}{3})$$

where Φ is a standard normal cdf. He asked the question what would happen if the true distribution deviated slightly from the assumed one. With this mixture model he showed that for values of ε as small as 0.002 the standard deviation is no longer optimal compared to the mean deviation. Up to this point, the belief that all measurement errors were Gaussian was so entrenched that Huber referred to it as "the dogma of normality". For a more complete and entertaining review of the early developments of statistical robustness, see Huber (1972). Huber blames "the dogma of normality" on a misunderstanding of the Central Limit theorem (which only implies an approximately normal distribution for the sum of independent errors) and of the Gauss-Markov theorem (stating that the sample mean is the best linear unbiased estimator). He notes: "[..] there is no reason, except mathematical convenience, to impose linearity or unbiasedness, and one might argue from sad experience that the model should also allow for a *few gross* [sic] elementary errors occurring with low probability." This said, the wisdom of hands-on experience led to the invention by practitioners of certain 'robust'

procedures, such as in the following description of the α-trimmed mean by an anonymous scholar (1821, page 189, author's translation):

> [..] there are certain provinces in France, where, in order to determine the average yield of an estate, it is customary to consider the yields over a period of twenty consecutive years, remove the greatest and the smallest of these numbers, and then to use the remaining eighteen to calculate the mean. Those who imagined this method, must have no doubt considered that very abundant harvests and very poor harvests were exceptions from the usual course of Nature, and thus one should not make allowance for them.

1.2.2 Huber and Hampel

Huber (1964) was the first to provide a theoretical framework for the robust estimation of a location parameter. He then generalised the theory to give a firm basis for robust statistical estimation (see Huber, 2009). Deviations from normality motivated his initial work. His definition of robustness is somewhat vague:

> [..] for our purposes, robustness signifies an insensitivity to small deviations from the assumptions.

However the key elements of the philosophy of statistical robustness are contained in Huber (1972), whether from a frequentist or a Bayesian perspective:

> With Anscombe (1960) I am inclined to view robustness as a kind of insurance problem: I am willing to pay a premium (a loss of efficiency of, say, 5 to 10% at the ideal model) to safeguard against ill effects caused by small deviations from it; although I am happy if the procedure performs well also under large deviations, I do not really care - inferences based upon a grossly wrong statistical model may have little significance.

He also argued that robust models are not the same as nonparametric and distribution-free models. For example, the sample mean is distribution-free but highly sensitive to outliers. Robustness is a property of parametric statistical models that have the following characteristics:

- Efficiency at the assumed model (the idealised situation)
- Stability with respect to small perturbations to the assumptions, only impairing performance slightly
- Breakdown protection: larger deviations should not cause catastrophe

M-estimators

The first general framework with which to consider robustness was constructed using the gross-error model, or ε-contamination model, introduced by Huber (1964):

$$\mathscr{P}_{F_0,\varepsilon} := \{F \quad | \quad F = (1-\varepsilon)F_0 + \varepsilon H, \quad H \in \mathscr{M}\} \tag{1.1}$$

where F_0 is a baseline distribution and $\mathscr{M}$ is a set of contamination distributions. It is important to note that this is not a topological neighbourhood. The parameter ε controls the quantity of 'contamination' allowed in the model.

M-estimation generalises procedures such as least-squares estimation or maximum likelihood, by finding an optimal $\hat{\theta}$ that satisfies:

$$\hat{\theta} = \arg\min_{\theta} \int_{\mathscr{X}} \rho(x,\theta) \mathrm{d}F(X)$$

for a general class of functions ρ ("M" stands for "maximum-likelihood type"). If ρ is differentiable, then it is said to be an M-estimator of the ψ-type. These are easier to solve for in general. Under certain conditions, it can be shown that these are consistent estimators and are asymptotically normally distributed (Chapter 4, page 113 of Staudte and Sheather, 1990).

Influence and breakdown

Hampel's approach on the other hand is composed of two key tools, the influence function and the breakdown point. The breakdown point is defined as the minimum proportion of 'contamination' (gross-errors) that can occur within a dataset in order for the statistic of interest to be arbitrarily far from the truth (meaningless). In the case of the mean for example, this is $1/n$, hence defined in the limit as zero. However for the median this is $1/2$. For a given statistic $T(x)$, the influence function $I(x)$ is the effect that a data point x has on $T(x)$. This allows an insight into the robustness of T with respect to contamination in the data (when data do not adhere to model assumptions). Again considering the sample mean, $I(x)$ is unbounded as $x \to \pm\infty$. Influence functions are particularly useful for studying the behaviour of M-estimators. Indeed, the influence function of an ψ-type M-estimator ρ is proportional to the derivative ρ'. Hence to minimise the effect of outliers it is easy to directly specify the derivative ρ' (which is then used to find the solution $\hat{\theta}$).

1.2.3 Minimax decisions

In parallel to advances in the theory of robust estimation and inference, a first approach to robust decision theory was given by Wald (1945, 1949, 1950). Wald considered the general problem of choosing a decision function $\omega \in \Omega$, mapping an observation $E \in \mathscr{X}$ to a hypothesis H_ω, where the distribution over $\mathscr{X}$ is unknown but assumed to be within a family F_θ indexed by a parameter θ. Wald considered the frequentist risk, $R[\theta, \omega] = \int L(\omega(x), \theta) F_\theta(dx)$, with $x \in \mathscr{X}$ where F_θ is the distribution function over $\mathscr{X}$ corresponding to θ. He uses the term 'weight function' to refer to the loss $L(\omega(E), x)$ His first consideration is to use a weighted average of the risk function, integrating over all possible values of θ with respect to a measure $f(\theta)$. However, he notes (Wald, 1945, page 267):

> The difficulty with this approach is that the decision function $\omega(E)$ for which [the expected loss] is a minimum will, in general, depend on the distribution function $f(\theta)$ and one can hardly justify any particular choice of $f(\theta)$. If there would exist an a priori probability distribution $g(\theta)$ of the parameter θ and if this distribution were known, one could put $f(\theta)$ equal to $g(\theta)$. However, in most of the applications not even the existence of such an a priori probability distribution of θ can be postulated, and in those few cases where the existence of an a priori distribution of θ may be assumed this distribution is usually unknown.

Thus in the absence of beliefs (model) on θ, Wald interpreted the decision problem as a zero sum two-person game, following Von Neumann and Morgenstern's work on game theory (Von Neumann and Morgenstern, 1947). To be robust the statistician protects himself against the worst possible outcome, selecting an action (here a decision rule) $\omega(E)$ according to the minimax rule[2]:

$$\hat{\omega} = \arg\inf_{\omega} \left[\sup_{\theta} R(\theta, \omega(E)) \right]$$

This is akin to the decision maker playing a two-person game with a malevolent Nature, where losses made by one agent will be gained by the other (zero sum). On selection of an action ω, Nature will select the worst possible outcome (θ^*_ω), equivalent to the assumption of a point mass distribution taken reactively to the choice of action, $\delta_{\theta^*_\omega}(\theta)$ where,

$$\theta^*_{\omega(E)} = \arg\sup_{\theta \in \Theta} R(\theta, \omega(E))$$

[2] Wald (1945) does not justify the minimax rule using a robustness argument, indeed there is no principled justification given in his paper, but robustness can be seen as one of the benefits of minimax. Another benefit comes from the computational side, as the minimax value of an action can be easier to compute than the expected loss.

In this section, the notation is adapted to fit that of Wald's original work; however, for the purposes of this thesis I do not consider decision rules but instead use the more abstract notion of a set of possible actions (alternatives) $a \in \mathscr{A}$, with a loss function L defined over $\mathscr{A} \times \Theta$. The minimax rule is thus written as:

$$\hat{a} = \arg \inf_{a \in \mathscr{A}} \left[\sup_{\theta \in \Theta} L_a(\theta) \right]$$

I am re-writing Wald's framework so that the notation is analogous to the Bayesian decision framework (see next section). Although elegant in its derivation the minimax rule has severe problems from an applied perspective. The decision maker following the minimax rule is not rational and treats all situations with extreme pessimism. Minimax assumes that Nature is reactive in selecting $\delta_{\theta^*_{\omega(E)}}(\theta)$ for your choice of $\omega \in \Omega$ irrespective of the evidence from existing information I informing on the plausible values of θ. Subsequent to Wald there has been considerable work to develop more applied procedures that protect against less extreme outcomes.

1.3 $\mathscr{M}_{\text{open}}$ and Savage decisions

1.3.1 Elements of Bayesian decision theory

The Bayesian decision paradigm (see, for example Berger, 1985, page 8) is composed of the following elements:

- Observables (data) $x \in \mathscr{X}$ and a parameter space Θ, where $\theta \in \Theta$ represents the unknown 'state of the world'.
- A joint model $\pi(x, \theta)$ over $\mathscr{X} \times \Theta$, which can be factorised into a prior $\pi(\theta)$ and a likelihood function $f(x|\theta)$, which is a valid sampling distribution for the observed data. Combined with the data x, a posterior distribution $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ is obtained.
- A set of available actions or alternatives $a \in \mathscr{A}$ and a real-valued loss function L defined over $\mathscr{A} \times \Theta$, that specifies the loss (or negative utility/reward) if a is chosen when θ is the true state of the world.
- Each action is evaluated by posterior expected loss:

$$\psi(a, L, \pi) := \int_{\Theta} L(a, \theta)\pi(\theta|x)\mathrm{d}\theta = \frac{\int_{\Theta} L(a, \theta) f(x|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta} f(x|\theta)\pi(\theta)\mathrm{d}\theta}$$

- The optimal action is defined as $a^* := \arg\inf_{a\in\mathscr{A}}\{\psi(a,L,\pi)\}$, i.e. the action minimising posterior expected loss.

Savage (1954) put together a set of postulates (known as the Savage axioms) that describe the structure of a rational agent's preferences[3] (such as transitivity for example). If these are obeyed, then a representation theorem ensues, showing that there must exist a loss function L and a distribution $\pi(\theta)$ over the unknown states of the world $\theta \in \Theta$ (Savage's small world), where the agent's ordering of preference is given by the ordering of expected loss. By preference, is meant here an ordering of all the possible consequences $c(a,\theta)$ that would arise if θ was the true state of the world and action a was chosen. Adhering to the Savage axioms implies an ordering of consequences which must induce a complete ordering of actions.

Thus, within the Bayesian framework, given a posterior distribution $\pi(\theta|x)$ and a loss function $L(a,\theta)$, posterior expected loss is the optimal criterion for ranking the alternatives a. However, this depends on $L(a,\theta)$ being the agent's true loss function and $\pi(\theta|x)$ being composed of the true prior π and true likelihood f. The existence of this posterior distribution and loss function is a consequence of the assumption that the agent obeys the Savage axioms, (Representation Theorem, Savage, 1954). The Savage axioms and thus Bayesian decision theory are prescriptive (normative) and do not pretend to describe the way people actually make decisions (this is the motivation for the Ellsberg paradox, see Ellsberg, 1961).

In reality, however, there are very few situations where the whole model is believed to be correctly specified. Historically the most controversial element of a Bayesian decision model has been the prior distribution π, seen as the most subjective component. This was understood by the founders of Bayesian theory as indicated in the following quote:

> Subjectivists should feel obligated to recognise that any opinion (so much more the initial one) is only vaguely acceptable... So it is important not only to know the exact answer for an exactly specified initial problem, but what happens changing in a reasonable neighbourhood the assumed initial opinion. De Finetti, as quoted in Dempster (1975)

The prior distribution π can never (or at least very rarely) be assumed to be true (as in exactly corresponding to the agent's prior beliefs), very often having been chosen for mathematical tractability and ease of computation. In this context, it is clear that posterior expected loss should not be blindly applied as the only optimal criterion and model 'robustness' must be verified. However, the standard Bayesian framework is a closed hypothesis space, that is to

[3] I use the imagery of an rational agent taking decisions contained within some set $\mathscr{A}$, where rational is defined as obeying the Savage axioms.

say with probability 1 Nature's true generating model is considered to be within the support of the prior, indexed by the parameter θ. In the context of model selection, this is denoted $\mathscr{M}_{\text{closed}}$ (see Bernardo and Smith, 1994, chapter 6, section 1.2). When this is not believed to be the case, the modelling happens in $\mathscr{M}_{\text{open}}$[4]. In this case additional methods are needed as the meaning of the posterior $\pi(\theta|x)$ is no longer clear. An elegant solution to the issue of $\mathscr{M}_{open}$ is given by Bissiri et al. (2013); Bissiri and Walker (2012a,b). The authors provide a framework where instead of having to specify a full joint distribution $\pi(x,\theta)$ over the observables and the parameter space, a prior $\pi(\theta)$ is specified over the parameter of interest, and this is updated with new information via a loss function.

In the context of inference, another method for dealing with $\mathscr{M}_{\text{open}}$ is to redefine the meaning of the parameter θ. Instead of postulating the existence of a $\theta^* \in \Theta$ that is the 'true' value, i.e. $\pi(x|\theta^*)$ is Nature's data generating mechanism, an alternative θ^* is proposed. This now corresponds to the value that the posterior would concentrate around given infinite data, which under certain conditions is the model $\pi(x|\theta^*)$ that is closest to the true data generating mechanism (Walker, 2013). Under this interpretation, the Bayes update of a prior distribution $\pi(\theta)$ has a well defined meaning, even though the model is 'misspecified' in the sense that π is known to not represent the statistician's beliefs about θ. In the discussion of Walker, Hoff and Wakefield (2013) point out that incorrect models can still estimate the object of interest. For example, using a normal distribution to estimate the mean of a non-Gaussian population, albeit in an inefficient manner. From this idea they show that a Bayesian sandwich estimator can be used to estimate a low dimensional parameter or a statistic of interest "whose sampling distributions are robust to model misspecification". This they call the "pseudo-true" parameter θ^* or the "pivotal quantity" $s(X,\theta)$ (where X is the data). In their words:

> If we wish to benefit from the internal consistency of subjective Bayesian inference, we need to limit our probability statements to those quantities about which we have actual information.

This can be formulated more succinctly as: 'model the object of interest'. These ideas are linked to the framework presented in this thesis, where only the misspecification that effects the loss function (what the decision makers cares about) is considered to be of importance.

In what follows, I will refer to $\{L, \pi_I, \mathscr{A}\}$ as the decision system. π_I here is the agent's best guess or approximating model over the uncertainty θ, which will usually be a Bayesian posterior distribution composed of a prior π and likelihood f, or a nonparametric prior.

[4] Bernardo and Smith also consider the $\mathscr{M}_{\text{completed}}$ case, where none of the proposed models are considered to be true, but the statistician has access to the 'true' model, with which it is possible to compare the performances of the submodels. This is developed further in section 1.3.2 and also illustrated in Chapter 5, section 5.3

Copying the notation of Bernardo & Smith, I define $\mathscr{D}_{closed}$ and $\mathscr{D}_{open}$ to describe the situations when the decision system is believed to be respectively well specified (true beliefs and true loss function) and misspecified (either the distribution π_I or the loss function or both). My perspective is that every statistical task is in a sense a decision task, as given in Bernardo and Smith (1994), page 102:

> ...in our view, the supposed dichotomy between inference and decision is illusory, since any report or communication of beliefs following the receipt of information inevitably constitutes a form of action.

It may seem curious that the data x is not included in the definition of the decision system, but this work is concerned with robustness to model assumptions, thus assuming that all the necessary pre-processing steps have been taken prior to modelling. This motivates the terminology of 'best guess' for the distribution π_I, constructed using best statistical practice, which is an area of research of its own (see for example Gelman et al., 2014). Analogous to the concept of 'shaking the data' (e.g. Efron's bootstrap), I am interested in formal methodologies and graphical visualisations which 'shake the model'.

1.3.2 Robust Bayesians

Prior robustness

For a strict Bayesian adherent, there is no issue with model robustness. The statistician specifies his beliefs via the joint distribution $\pi(x, \theta)$, updates with new data x and then chooses the optimal action using the criterion of posterior expected loss. This is completely contrary to any conventional wisdom regarding statistical modelling, as conveyed by the popular aphorism attributed to George Box: "all models are wrong, but some are useful"[5].

The more pragmatic *robust Bayes* approach has existed at least since Good (1950), and its flavour is well contained in the above quote from De Finetti. However the main advances can be attributed to work done by Berger in a series of papers: Berger (1984, 1994); Berger and Berliner (1986). The last of these papers gives a thorough and accessible review of the robust Bayesian viewpoint, alongside a lengthy discussion by other authors. I believe the most relevant points can be summarised as:

1. The model chosen should have inherent robust properties, in particular via the use of:

[5]The extended quote is: "The fact that the polynomial is an approximation does not necessarily detract from its usefulness because all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind." Box and Draper (1987), page 424.

- Flat-tailed distributions (the Student-t instead of the Gaussian for example).
- Prior distributions which are non-informative or weakly informative. This is an area which has been developed a lot since, see for example Martins et al. (2014) regarding the specification of (robust) default priors, Gelman et al. (2008) for ideas on weakly informative priors, and Berger et al. (2009) for overview on reference priors.
- Bayesian nonparametric methods (these have full support, thus alleviating concerns of M_{closed}, this is further discussed in section 1.3.2).

2. Global robustness methods, which consider a range of posterior inferences (instead of only one) under a class of models (either parametric or nonparametric). This idea is central to the work of this thesis.

Regarding this last point, and as mentioned in the previous section, the most contentious element of Bayesian models has historically been the prior distribution $\pi(\theta)$. A very natural idea is to consider a family of priors Γ instead of using the statistician's 'best guess' prior π_0. Berger and Berliner (1986) consider ε-contamination priors, defined as:

$$\Gamma = \{\pi = (1-\varepsilon)\pi_0 + \varepsilon q, q \in \mathscr{Q}\}$$

where $\mathscr{Q}$ is the class of contaminant distributions. This is analogous to Huber's gross-error model (see equation 1.1). In the same way, this is not a topological neighbourhood and for tractability it is necessary to restrict $\mathscr{Q}$ so that it is not 'too big', for instance by including only uni-modal distributions (Berger, 1994).

More generally, robust Bayesian methods are usefully classified as either *local* or *global*. Global methods take as input a family of priors Γ - such as the ε-contamination family - and then consider the range of the functional $\psi(a, L, \pi)$ (the posterior expected loss as defined in section 1.3.1) with π taken from Γ. Local methods on the other hand look at functional derivatives of the quantity $\psi(a, L, \pi)$ at some baseline point π_0. Both these methods consider perturbations with respect to the prior distribution $\pi(\theta)$, with some work looking at simultaneously changing the prior and likelihood (see Lavine, 1991). However this is more complex, as changing the likelihood will in general change the meaning of the prior. I refer the reader to Ruggeri et al. (2005) for a very complete review of robust Bayesian techniques.

Kadane and Srinivasan (1994) raise an important distinction between *decision robustness* and *loss robustness*, the former referring to situations where the optimal decision $\hat{a}$ is robust to model perturbations, and the latter when the expected loss quantity remains relatively

unchanged across a neighbourhood of models Γ. It is clear that the two criteria are not interchangeable, and they give simple examples of when the one occurs without the other. Although it would appear formally that decision robustness is more desirable, situations where the set $\mathscr{A}$ is in a sense a mathematical artefact, then loss robustness could be more important. This will be purely context-dependent and methods targeting both criteria are important.

Complex approximate models

The 1980s to mid 90s was a highly active period for research in Bayesian robustness, with the vast majority of the work focussing on sensitivity to the prior specification. However, this tailed off somewhat with the arrival of new computational methods that allowed for more flexible model specifications, alleviating the historic concern that $\pi(x,\theta)$ was indexing too restrictive a class of sub-models. However, with the explosion of computational power and increasing complexity of datasets - the so called 'big data era' - robust Bayesian methods merit a reappraisal.

Bayesian theory has often been associated with inductive inference and a mathematical formulation of rationality, as opposed to the classical school of Neyman-Pearson's hypothesis tests which are associated with the hypothetico-deductive and falsificationist view of science of Popper (1959) (see the discussion by Gelman and Shalizi, 2013). This view encourages a certain complacency with regards to model robustness. Falsifying the model is in a sense a good thing, as it reveals its strengths and limitations. In the words of Good (1992): "Instead of saying that the Newtonian theory has been refuted, it would be better to say that special relativity explains why the Newtonian theory is so good!" He suggests the (inelegant) word 'inexactified'. One suggested process for 'inexactification' is given by posterior predictive checking (Gelman et al., 1996). The Bayesian posterior predictive $P(y) = \int_\Theta f(y|\theta)\pi(\theta|x)\mathrm{d}\theta$ can be used to simulate new datasets $y_1^{rep},..,y_K^{rep}$ which can then be compared to the real data x, generally via summary statistics. This process, outside of the standard Bayesian paradigm, allows for an 'assessment' of the model.

> We use the term 'assessment' instead of 'testing' to highlight the fundamental difference between assessing the *discrepancies* between a model and data and testing the *correctness* of a model [sic]. We believe that there is a general concern that there has been too much emphasis on the latter problem and that, in the words of Tiao and Xu (1993), there is a need in practice for "... development of diagnostic tools with a greater emphasis on assessing the usefulness of an assumed model for specific purposes at hand rather than on whether the model is true."(Gelman et al., 1996, page 734)

The authors classify Bayesian model checking into three categories: "(1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking the model fits the data." Posterior predictive checks fit into the third category. The framework developed in this thesis fits into the second, but by clarifying 'reasonable' as a property pertaining to the loss function. Kerman et al. (2008) propose various graphical methods and this theme is also developed in chapter 4.

Bayesian nonparametrics

Bayesian nonparametrics has become popular due the flexibility of the models and the inherent robustness, alleviating concerns about the closed hypothesis space of traditional Bayesian modelling. Their wide support removes any worry about the 'closed hypothesis space' of Bayesian modelling. However, Bayesian nonparametric models are in general less interpretable than say mechanistic parametric models[6], where the parameters have strong interpretations, even though they might be known to be misspecified. Karabatsos (2006) provides an interesting approach to misspecification in a Bayesian framework by comparing a parametric Bayesian posterior with its nonparametric counterpart. The misspecification of the parametric model is seen as the price the statistician is willing to pay for interpretability and/or tractability. The predictive performance (with respect to the 'true' nonparametric model) is given in Kullback-Leibler divergence, which has an expected loss interpretation (KL penalty for misspecification). This gives a 'reject' action, if the parametric model is beyond a certain distance from the nonparametric model. This perspective comes under the $\mathscr{M}_{\text{complete}}$ view (Bernardo and Smith, 1994, page 365), where models are compared with respect to the user's true model.

1.3.3 Robust Control Theory

An early criticism of the Savage axioms came from researchers in Economics. They objected that the framework could not distinguish between different types of uncertainty following the distinction made by Knight (1921) between risk and uncertainty, the former being measurable the later not. This distinction appears to be respected by individuals when taking decisions, showing a 'risk averse' behaviour. Ellsberg (1961) illustrated this phenomenon in his famous paradox. Imagine two urns each containing 100 balls and every ball is either red or blue. One is told that the first urn (A) has exactly 50 red balls and 50 blue balls. No more information is

[6] 'Bayesian nonparametrics' is accepted as a misnomer, with 'infinitely parametric' being more suitable a name. The parameters in these models are hyper parameters, and do have interpretations, albeit less so than in standard parametric models.

given about the second urn (B). The person is told that they will win $100 if she picks a red ball. The question is which urn would she choose? This game is paradoxical from a Bayesian perspective as the two alternatives should be equal in expected value (under any reasonable prior), but empirically this is not the case, with a majority of participants opting for the urn A. Hence the Savage axioms do not work as a descriptive theory of rational decision making ('rational' here defining the manner in which individuals make decisions in idealised settings).

Gilboa and Schmeidler (1989) developed the theory of maxmin Expected Utility in part to solve the Ellsberg paradox which extends standard Bayesian inference to a setting with multiple priors in the form of a closed convex set Γ. An action is then scored by its expected loss under the least favourable prior within that set. Their 1989 paper formalises this and provides a solution to the Ellsberg paradox. When Γ contains only one prior, we are back again in the usual Bayesian setting. The set Γ can be seen as describing the decision maker's aversion to uncertainty. This work is closely related to Γ-*minimax* (for which the Ellsberg paradox is also used as a motivating example, see section 1 of Vidakovic (2000)).

Independent of the above developments in statistics, control theorists were investigating robustness to modelling assumptions. Control theory broadly concerns optimal intervention strategies (actions) on stochastic systems so as to maintain the process within a stable regime. Hence it is not surprising that decision stability is an important issue. When the system is linear with additive normal (white) noise the optimal intervention is well known (Whittle, 1990). Robust control theory, principally developed by Whittle, considers the case when Nature is acting against the operator through stochastic buffering by non-independent noise, see Whittle (1990). Whittle established that under a malevolent Nature with a bounded variance an optimal intervention can be calculated using standard recursive algorithms.

Again working in economics, Hansen and Sargent, in a series of influential papers (e.g. 2001a, 2001b), generalised ideas from Whittle (1990) and Gilboa and Schmeidler (1989) motivated by problems in macroeconomic time series. They define a robust action as a local-minimax act within a Kullback-Leibler (KL) neighbourhood of $\pi_I(\theta)$ through exploration of,

$$\psi_a^{\sup} := \sup_{\pi \in \Gamma_C} E_\pi[L_a(\theta)]$$

where Γ_C denotes a KL ball around π_I,

$$\Gamma_C := \{\pi : \int \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_I(\theta)}\right) d\theta \leq C\}$$

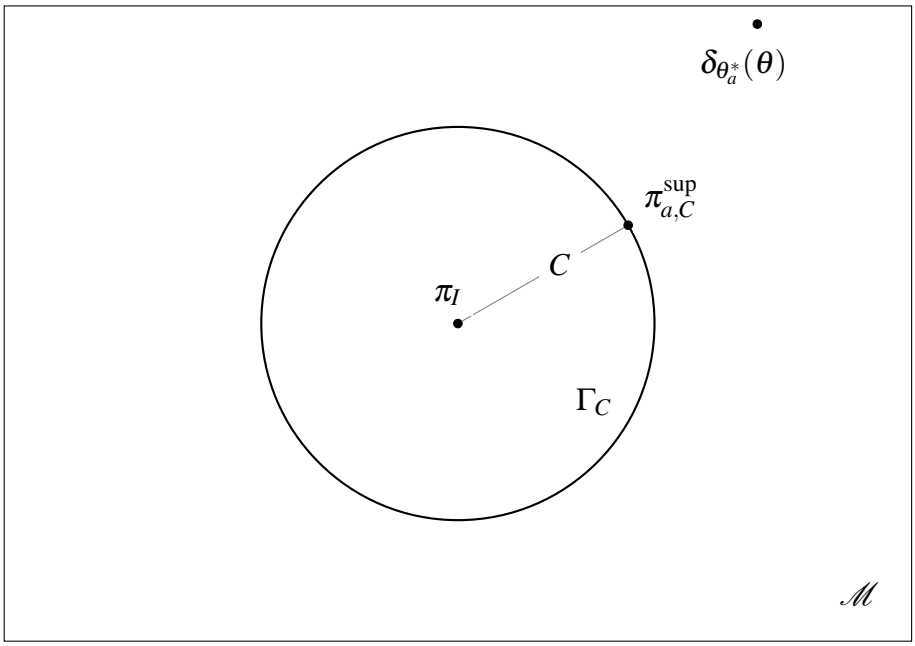

Fig. 1.1 Graphical representation of local-minimax model $\pi_{a,C}^{\sup}$ within a Kullback-Leibler ball of radius C around the reference model π_I, with global (Wald's) minimax density $\delta_{\theta_a^*}(\theta)$.

$\pi_{a,C}^{\sup}$ denotes the corresponding local-minimax distribution,

$$\pi_{a,C}^{\sup} = \arg \sup_{\pi \in \Gamma_C} E_\pi[L_a(\theta)]$$

Figure 2.1 shows a pictorial representation of this constrained minimax rule, where the reference distribution π_I is a point in the space of distributions $\mathscr{M}$ (represented by the rectangle) and the least favourable distribution $\pi_{a,C}^{\sup}$ is contained within the neighbourhood Γ_C (represented by the circle of radius C). The Wald minimax distribution is given by $\delta_{\theta_a^*}(\theta)$. Hansen and Sargent showed how $\pi_{a,C}^{\sup}$ and $\psi_a^{\sup}$ can be computed for dynamic linear systems with normal noise, see Hansen and Sargent (2008) for a thorough review and references.

Breuer and Csiszár (2013a,b), building on the work of Hansen and Sargent, derived corresponding results for arbitrary probability measures $\pi_I(\theta)$. Under mild regularity conditions, and using results from exponential families and large deviation theory, they obtain the exact form of $\pi_a^{\sup}$ for any $\pi_I(\theta)$ given the KL ball of size C, as well as an estimate for $\psi_a^{\sup}$ (see also Ahmadi-Javid, 2011, 2012).

1.3.4 What is misspecification?

I have so far kept the notion of 'model misspecification' vague. This is because I aim to tackle sensitivity issues that could be related to any step of the model building process. Inherently robust procedures - such as M-estimation - are of course desirable in model construction. However, if the expected loss criterion is used to rank decisions (based on this model) it must be understood that the *optimality* of this procedure relies on having the *true* model. Possible sources of misspecification are given in this (non-exclusive) list:

- Inputs of (Bayesian) decision-system: prior, likelihood and loss function
- Assumptions during the construction of the model (these mostly pertain to the likelihood but can have a wider impact) such as linearity, independence, homoscedasticity, stability of pattern in real world (concept drift in data-mining)
- Finite resources, in particular time and computation power. This forces the user to adapt 'approximate' methods such as approximate Bayesian computation, integrated nested Laplace approximations, variational Bayes, composite likelihoods, etc

Grünwald and van Ommen (2014) draw a distinction between *good* vs. *bad* misspecification. The authors consider high dimensional linear models which are misspecified - for example assuming homoscedasticity when in fact the variance is heteroscedastic - and analyse whether or not the posterior distribution will concentrate around some distribution $\tilde{P}$ which is nearest in Kullback-Leibler to data generating mechanism P^*. When the set of models indexed by the prior is not convex, the posterior will not concentrate around $\tilde{P}$. This is denoted bad misspecification. On the other hand, if the posterior concentrates properly around $\tilde{P}$, we have situations where the interpretation of Bayesian inference given by Walker (2013) is appropriate.

1.4 Notation

Throughout this thesis I will refer to the approximating model as π_l. This is the statistician's 'best guess' at the uncertainty surrounding the parameter θ. In most cases this will be a Bayesian posterior distribution, but it could also be a prior (before any data has been seen), or a distribution fitted using Maximum Likelihood or any other suitable method. In this regard, the approach is agnostic as to the Bayesian paradigm. The uncertainty in the world is denoted by the parameter θ within some measurable space Θ. $\mathscr{A}$ denotes the set of 'actions' a (discrete or continuous), and the loss function will be denoted $L(a, \theta)$, abbreviated $L_a(\theta)$

Notation	Definition
Θ	Parameter space describing the uncertainty in the 'small world' of interest
$a \in \mathscr{A}$	Set of actions or alternatives
$L(a,\theta)$ or $L_a(\theta)$	Loss function defined as mapping $\mathscr{A} \times \Theta \rightarrow \mathbb{R}^+$
$L_{(a,a')}(\theta)$	Regret loss function: $L_a(\theta) - L_{a'}(\theta)$
π_I	The approximating or reference model. This could be a Bayesian posterior, or just any distribution over the parameter space Θ.
$\delta_{\theta_a^*}(\theta)$	Wald minimax distribution for action a, i.e. point mass on least favourable state θ_a^*.
C	The radius of the Kullback-Leibler ball centred at π_I
$\lambda_a(C)$	Exponential tilting parameter given in equation 2.1 for action a corresponding to least favourable distribution in KL ball of radius C
Γ_C	Set of distributions π satisfying $\mathrm{KL}(\pi\|\|\pi_I) \leq C$ (KL ball)
Γ_C^{rev}	Set of distributions π satisfying $\mathrm{KL}(\pi_I\|\|\pi) \leq C$ (reverse KL ball)
$\pi_a^{\sup}$	The least favourable distribution for action a in the KL ball of radius C
$\psi_a^{\sup}(C)$	Expected loss of action a under $\pi_{a,C}^{sup}$
$[\psi_a^{\inf}(C), \psi_a^{\sup}(C)]$	Interval of expected loss of action a in Γ_C
$\pi_{(a,a'),C}^{\sup}$	Least favourable distribution corresponding to regret loss function $L_{(a,a')}(\theta)$

Fig. 1.2 Glossary of mathematical notation

when a is considered fixed. Table 1.2 gives a comprehensive glossary of mathematical notation used in this work.

Chapter 2

Local Minimax Decisions

This chapter presents a bridging framework between the two dominant decision paradigms, minimax and expected loss. This allows for a formal assessment of the impact of model approximation. We quantify the stability of optimal actions to perturbations within an information neighbourhood defined via the Kullback-Leibler divergence. The analytical form of the local least favourable distribution is derived and it is shown how this unifies already known results and methods. In the case where the information neighbourhood is shrunk to zero, the variance of loss under the approximating model is shown to correspond to a measure of the local sensitivity of the decision system. This provides a simple solution to the well-known Ellsberg paradox of decision theory. The results lead to computationally cheap algorithms that take as input Monte Carlo representations of (Bayesian) posterior models and the corresponding user defined loss function, and output expected loss estimates under the perturbed model. This gives rise to graphical diagnostic methods which are further explored in a later chapter.

2.1 Bridging Wald and Savage

Bayesian decision theory formally ignores model misspecification. The more pragmatic *robust Bayesian* perspective has long challenged this view, but this area of research has tailed off since the advent of computational techniques such as MCMC. Such methods have removed the dependency on constrained families of models, for example, conjugate classes. However, I believe that Bayesian robustness merits a reappraisal in view of new developments in statistics concerning models that are misspecified by design. This chapter presents a formal framework for assessing the sensitivity of a decision system when assuming possible (known or unknown) misspecifications. This is done by combining the expected loss criterion with Wald's minimax, closely following ideas developed in econometrics, robust

control and finance (e.g. Breuer and Csiszár, 2013a; Gilboa and Schmeidler, 1989; Hansen and Sargent, 2008). This is done via the exploration of neighbourhoods of the posterior distribution (or approximating model) π_I, defined using the Kullback-Leibler divergence. Much of this approach builds on other work, and therefore it is important to highlight the contributions of this thesis and how the approach differs from the perspective of previous work.

Firstly, this approach is *ex-post*. That is to say we perform a sensitivity analysis after the statistician has put together his 'best guess' approximating model and loss function. The majority of robust Bayesian approaches have focused on sensitivity of the inputs of the model, the prior and/or likelihood. In contrast, we consider neighbourhoods of the *posterior* distribution over the parameter of interest θ, although this is connected.

Secondly, only the marginal distribution of π_I over the dimensions that enter into the loss function is considered. Thus the sensitivity of the decision system is defined relative to the loss function. This is different from the usual separation of loss function and model in Bayesian decision systems. We also provide a simple proof for the form of the local minimax distribution, which is less general than previous proofs but allows for more intuition behind the result. Concepts of coherence, local sensitivity and admissibility and connections to other well-known ideas are shown as well.

Last but not least, we show that this approach is well suited for the 'computational decision theory' set-up. When access to the posterior distribution is only possible via some form of Monte Carlo sampling, then the methods can be implemented at very little extra computational cost. This aspect of robust Bayesian decision is of particular interest.

Together these ideas provide a framework for D_{open}, decision making in the context of model misspecification. Because models are approximations, the criterion for defining optimal decisions is also an approximation. In the Bayesian setting this criterion is posterior expected loss and the optimal Bayesian action is therefore conditional on the posterior model.

2.2 Least favourable distributions

2.2.1 Properties

Analytical form

We consider the context where the statistician has defined a Bayesian decision-system as given in section 1.3.1. This is defined as a 4-tuple $\{\Theta, \pi_I, L, \mathscr{A}\}$, composed of the parameter space Θ, an action space $\mathscr{A}$, a (posterior) distribution π_I over Θ and a loss function L defined over $\mathscr{A} \times \Theta$. We denote by $\Theta' \subseteq \Theta$ the subspace over which L is non constant, i.e. all

the dimensions of Θ that enter into the loss $L(a,\theta)$. We assume that Θ' is a non-null set (otherwise trivial).

In order to assess the sensitivity of the decision-system to perturbations in the model π_I, we look at the variation of the expected loss as a function of its inputs:

$$\psi(a,L,\pi) = \int_\theta L(a,\theta)\pi(\theta)\mathrm{d}\theta$$

In particular, for a given loss function L, and fixing the action a, we look at the range of ψ as we vary π over a neighbourhood which is centred at the reference distribution π_I. The neighbourhood of interest is denoted Γ_C and is defined around the *marginal* of π_I over Θ' using the Kullback-Leibler divergence, where $C > 0$ is the size of the neighbourhood:

$$\Gamma_C = \{\pi : \mathrm{KL}(\pi||\pi_I) \leq C\}$$

where,

$$\mathrm{KL}(\pi||\pi_I) = \int_{\Theta'} \pi(\theta)\log\frac{\pi(\theta)}{\pi_I(\theta)}\mathrm{d}\theta$$

Within this neighbourhood Γ_C, the local least favourable distribution $\pi_a^{\sup}$ is defined as:

$$\pi_a^{\sup} = \arg\sup_{\pi\in\Gamma_C} \psi(a,L,\pi)$$

This distribution is a function of the size C of the neighbourhood Γ_C, but to alleviate notation we remove reference to C as it is clear from the context.

Figure 2.1 shows the space of all distributions $\mathscr{M}$ (rectangle), with the centring distribution π_I and the neighbourhood Γ_C (circle of radius C). The global minimax distribution $\delta_{\theta_a^{\sup}}(\theta)$ is also shown.

In order to guarantee the existence of a non-degenerate distribution $\delta_{\theta_a^*}(\theta)$ and for the following theorem we assume that the loss function is positive and is upper-bounded, with the maximum value given by $\sup_{a\in\mathscr{A}} L(a,\theta_a^*)$. Although this may seem a limiting assumption, it is one that would hold in almost every applied situation (for example in finance one could take the range $\pm$GDP of USA).

Surprisingly this situation leads to a local least favourable solution with a simple form.

Theorem 1. *Let* $\pi_a^{\sup} = \arg\sup_{\pi\in\Gamma_C}\mathbb{E}_\pi[L_a(\theta)]$*, with* $\Gamma_C = \{\pi : \mathrm{KL}(\pi \parallel \pi_I) \leq C\}$ *for* $C \geq 0$*. Then the solution* $\pi_a^{\sup}$ *is unique and has the following form,*

$$\pi_a^{\sup} = Z_C^{-1}\pi_I(\theta)\exp[\lambda_a(C)L_a(\theta)] \tag{2.1}$$

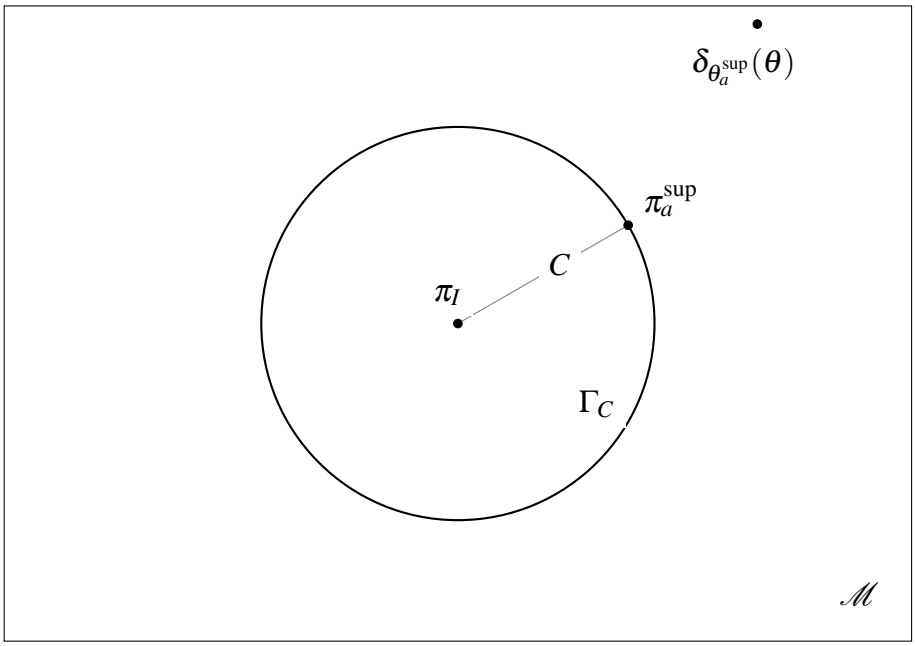

Fig. 2.1 Graphical representation of local least favourable model $\pi_a^{\sup}$ within a Kullback-Leibler ball of radius C around the reference model π_I, with global (Wald's) minimax density $\delta_{\theta_a^*}(\theta)$.

where $Z_C = \int \pi_I(\theta)\exp[\lambda_a(C)L_a(\theta)]d\theta$ *is the normalising constant or partition function, for which we assume* $Z_C < \infty$*, and* $\lambda_a(C)$ *is a non-negative real valued monotone function.*

Proof. The function minimisation problem, $\pi_a^{\sup} = \arg\max_{\pi\in\Gamma_C} \mathbb{E}_\pi[L_a(\theta)]$, has an unconstrained Lagrange dual form (see for example Hansen et al., 2006, pages 58-60) (due to strong duality),

$$\pi_a^{\sup} = \arg\inf_{\pi\in\mathscr{M}} \left\{\mathbb{E}_\pi[-L_a(\theta)] + \eta_a^{-1}\mathrm{KL}(\pi \parallel \pi_I)\right\}$$

for some $\eta_a = \eta_a(C)$ is a penalisation parameter with $\eta_a \in [0,\infty)$, and is monotone increasing in C. Hence,

$$\begin{aligned}
\pi_a^{\sup} &= \arg\inf_{\pi\in\mathscr{M}} \left\{\int -L_a(\theta)\pi(\theta)d\theta + \eta_a^{-1}\int \pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_I(\theta)}\right)d\theta\right\} \\
&= \arg\inf_{\pi\in\mathscr{M}} \left\{\int \pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_I(\theta)\exp[\eta_a L_a(\theta)]}\right)d\theta\right\} \\
&\propto \pi_I(\theta)\exp[\eta_a L_a(\theta)] \qquad (2.2)
\end{aligned}$$

The uniqueness arises from the convexity of the KL loss. The result follows, taking $\lambda_a(C) = \eta_a$. □

By a parallel argument the most favourable distribution for action a is:

$$\pi_a^{\inf} \propto \pi_I(\theta)\exp[-\lambda(C)L_a(\theta)]$$

Note that by assuming bounded loss functions we can ensure the integrability of the density $\pi_I(\theta)\exp[\lambda_a(C)L_a(\theta)]$. Breuer and Csiszár (2013b) and Ahmadi-Javid (2012) derive the same result more generally but perhaps less intuitively. Breuer and Csiszár (2013b) gives more general conditions on when the solution exists.

The parameter λ_a controls the KL divergence of the local minimax distribution $\pi_I(\theta)e^{\lambda_a L(a,\theta)}$. However the correspondence between λ_a and $\mathrm{KL}(\pi_a^{\sup}||\pi_I)$ is action specific, i.e. for a fixed $C > 0$, in general, different λ_a are needed for each action a. But if considering the action a fixed, then it is possible to think of λ as the free parameter instead of C. The relationship between λ and the KL radius C_λ is given by:

$$C_\lambda = \mathrm{KL}(\pi_a^{\sup}||\pi_I) = \int_\Theta \lambda L_a(\theta)e^{\lambda L_a(\theta)}\pi_I(\theta)\mathrm{d}\theta$$

This can be inverted (numerically) to find the corresponding λ for a fixed radius C. In this way, we see that both the standard Bayesian decision theory and Wald's minimax can be recovered, taking $\lambda_a \to 0$ and $\lambda_a \to \infty$, respectively.

This distribution $\pi_I(\theta)e^{\lambda L(a,\theta)}$ is an exponentially tilted version of the posterior π_I and is a well known object in actuarial science when the loss function is linear, i.e. $L(a,\theta) = A\theta$. The corresponding local minimax is the 'Esscher transform' (Esscher, 1932), used in option pricing (Gerber and Shiu, 1993) and to compute the so-called 'Esscher premiums' (expected loss under the Esscher transform).

Local sensitivity

Although the framework presented here fits into *global robustness* methods, it can be used to extract *local robustness* measures. In particular we consider the derivative of least favourable expected loss for a given action either as a function of the neighbourhood size C or the exponential tilting parameter λ. Firstly, differentiating w.r.t. λ we obtain the variance of loss under $\pi_a^{\sup}$:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathbb{E}_{\pi_a^{\sup}}[L_a] = \mathrm{Var}_{\pi_a^{\sup}}[L_a(\theta)]$$

The derivation is simple and can be found in Appendix A. Setting λ to 0, we see that the local sensitivity of the expected loss estimate is given by the variance of the loss under π_I.

Differentiating now w.r.t. C we need the following (applying the chain rule):

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}C_\lambda = \mathbb{E}_{\pi_a^{\sup}}[L_a(\theta)] - \mathbb{E}_{\pi_l}[L_a(\theta)]$$

For $\lambda \to 0$, this tends to zero, showing that the reciprocal derivative is infinite. This does not tell us anything particularly new but justifies using the variance of the loss as a sensitivity diagnostic. That is to say, if one were to choose between two actions that are indistinguishable in expected loss, then the variance gives a measure of the sensitivity of each action.

Solution to the Ellsberg paradox Ellsberg (1961) presents a strong argument against the use of Bayesian decision theory (or expected utility theory, as it is known in economics) as a descriptive framework for everyday decision making[1]. He considered the following thought experiment: one is confronted with a series of choices concerning bets on drawing at random either a red or a black ball from one of two urns. Both urns contain 100 balls, and in urn A it is known that there are exactly 50 red and 50 black balls. For urn B nothing more is known other than every ball must either be red or black. Empirically, the majority of people make the following selection of bets (betting on drawing a particular colour from a particular urn wins £1 if correct and zero otherwise).

- No preference between betting on drawing red or black from urn A
- No preference between betting on drawing red or black from urn B
- Prefer to bet on drawing red from urn A than red from urn B
- Prefer to bet on drawing black from urn A than black from urn B

This selection of choices displays an inherent 'risk aversion' and is incompatible with any set of probabilities assigned to each colour in each urn if the person was choosing based on

[1] Although this counterexample was popularised by Ellsberg, Keynes (1921, page 75) describes it clearly (his emphasis):

> The typical case, in which there may be a *practical* connection between weight and probable error, may be illustrated by the two cases following of balls drawn from an urn. In each case we require the probability of drawing a white ball ; in the first case we know that the urn contains black and white in equal proportions; in the second case the proportion of each colour is unknown, and each ball is as likely to be black as white. It is evident that in either case the probability of drawing a white ball is 1/2, but that the weight of the argument in favour of this conclusion is greater in the first case.

However, Ellsberg went further by showing that a particular selection of choices implies a logical inconsistency. Ellsberg's subjects included Jimmie Savage (who also did not adhere to his axioms!).

minimising expected loss. This appears to confirm the distinction made by Knight (1921) between "measurable uncertainty" and "risk"[2].

A consequence of the above result concerning the local sensitivity of actions is that it provides a method for differentiating between actions of equal expected loss. If the action is chosen to minimise the least favourable expected loss in the limit as $\lambda \to \infty$ then this corresponds to the action with lowest variance. This in turn provides a simple solution to the Ellsberg paradox.

Let $\theta_A \in [0,1]$ be the probability of drawing red from urn A, and $\theta_B \in [0,1]$ the probability of drawing red from urn B. Then $\pi(\theta_A)$ is a point mass on 1/2 (could also be seen as the limit of Beta$(\varepsilon,\varepsilon)$ as $\varepsilon \to \infty$). And $\pi(\theta_B)$ is a uniform on $[0,1]$, which is the limit of a Beta$(\varepsilon,\varepsilon)$ with $\varepsilon \to 0$. The loss of betting on drawing red from urn A can be taken as θ_A (this is expected proportion of times the bet will be successful). The same for urn B. The corresponding losses of betting on drawing black are $1-\theta_A$ and $1-\theta_B$ respectively. In this scenario, every bet has expected value 1/2, and the variance of the loss is given by the variance of the Beta distribution (zero for urn A and 1/2 for urn B). Therefore the choices presented above agree with a local minimax agent in the limit as the titling parameter $\lambda \to \infty$.

Experiment 1				Experiment 2			
Bet 1A		Bet 1B		Bet 2A		Bet 2B	
Winnings	Probability	Winnings	Probability	Winnings	Probability	Winnings	Probability
.11	1	1	.89	0	.89	0	.9
		0	.01	1	.11		
		5	.1			5	.1

Table 2.1 The Allais paradox: participants are asked to do both experiments and choose between bets A and B.

Ill-posed problems and regularisation In the standard Bayesian setting there are situations where two actions cannot be distinguished in terms of expected loss (the distinction made by Knight is meaningless for a Bayesian). However, the local minimax framework provides a principled method for discriminating between loss distributions with identical expected values. This can be seen as a regularisation mechanism for ill-posed problems in Bayesian decision theory, where 'ill-posed' is defined as the decision system not having a unique solution. Another paradox that this 'resolves' is the Allais paradox, introduced as an argument against expected utility theory in economics, which the author refers to as the "l'école américaine" (Allais, 1953). This is another example of empirical behaviour which cannot be explained by expected loss minimisation (or utility maximisation). A version of

[2]We note that he also considered a similar set-up to illustrate this distinction (see page 219 Knight, 1921).

the paradox is shown in table 2.1. A solution is assume that all bets have exactly the same loss (utility), then the principle of minimising the variance of loss would suggest choosing bets 1A and 2A (contrary to empirical choices). In this way, it is possible to see the Allais paradox as an under-determined problem, which our local minimax framework solves.

Bayesian coherent updating

Adapting results from Bissiri et al. (2013), we are are able to state the following result regarding the uniqueness of Kullback-Leibler divergence under the condition of guaranteeing coherent Bayesian updating.

Theorem 2. *Let $\pi_a^{\sup}(x,\pi_I)$ be the solution obtained by*

$$\pi_a^{\sup}(x,\pi_I) = \arg\inf_{\pi\in\mathscr{M}} \left\{ \mathbb{E}_\pi[-L_a(\theta)] + \eta_a^{-1} D(\pi \parallel \pi_I) \right\}$$

with data $x = \{x_i\}_{i=1}^n$, a centring distribution π_I, and arbitrary f-divergence measure D. Moreover, let x be partitioned as $x = \{x^{(1)}, x^{(2)}\}$, for $x^{(1)} = \{x_i\}_{i\in S}$, $x^{(2)} = \{x_j\}_{j\in\bar{S}}$, where $S, \bar{S}$ is any partition of the indices $i = 1,..,n$. For coherence we require,

$$\pi_{a,C}^{\sup}(x,\pi_I) \equiv \pi_{a,C}^{\sup}\left(x^{(2)}, \pi_{a,C}^{\sup}(x^{(1)},\pi_I)\right)$$

That is, the solution using a partial update involving $x^{(1)}$, which is subsequently updated with $x^{(2)}$, should coincide with the solution obtained using all of the data, for any partition. Then for coherence $D(\cdot||\cdot)$ is the Kullback-Leibler divergence.

The proof is given in the appendix.

Relevance of f-divergences This theorem shows that KL is the only divergence measure out of the f-divergences to provide coherent updating of the local least favourable distribution. f-divergences (confusingly, also known as g-divergences) were introduced as a general class of 'distances' between probability distributions by Ali and Silvey (1966). These are not metrics (in general they do not always obey the triangle inequality nor are symmetric), but measure how 'far apart' two distributions p_1, p_2 are, when defined over the same space X. The authors state four properties that are desirable for a measure of divergence D between distributions. The divergence D should be well defined for all pairs; its minimum value should occur when the two distributions are identical; transformations of the state space cannot increase the divergence[3]; and finally, when applied to pairs within a family of distributions

[3] It should be harder to distinguish between the two distributions if the state space is mapped to a coarser field. That means that the divergence should decrease (or stay the same), i.e. they become more similar.

indexed on the real line, the divergence should increase as the L_1 distance between the two index parameters increases. They show that expectations (and increasing functions of these expectations) of convex functions C of the ratio $\phi(x) = p_1(x)/p_2(x)$ taken w.r.t. to P_2 will have these four properties. Thus the family has a general representation as:

$$D(P_1||P_2) = f[\mathbb{E}_{P_1}\{C(\phi)\}]$$

(see Ali and Silvey, 1966, Theorem 2, page 138). This requires that P_2 is absolutely continuous with respect to P_1, C is a continuous convex function on $\mathbb{R}^+$, and f an increasing real-valued function on $\mathbb{R}$.

Notable examples of this family are total variation: $C(t) = (\sqrt{t}-1)^2$; the Hellinger distance: $C(t) = |t-1|$; and of course the Kullback-Leibler divergence, $C(t) = t\log t$. We discuss the role of Kullback-Leibler in particular in section 2.2.2.

It is clear that this is an important family of divergences and the log ratio between two probability measures is an essential feature for determining their similarity. However, in the context of computational decision theory, this means that these divergences only depend on the weights of the each atom (Monte Carlo sample). The value of the KL divergence is completely independent of the location of the atoms θ_i, see equation 2.5 in section 2.3. Using a different metric, such as the Wasserstein distance will give different benefits as it takes into account the location of the atoms as well.

Local Bayesian admissibility

In a classical setting, the notion of *admissibility* helps define a smaller class of actions that can then be further scrutinized in order to choose an optimal decision. Decision rules are *inadmissible* if there does not exist a θ such that its risk function (frequentist) is minimal (with respect to the other decisions) at θ. We note that in a Bayesian context, because the expected loss is a single quantity used to classify actions, only the action which minimizes expected loss is admissible. However if we consider the set of distributions contained within a Kullback-Leibler neighbourhood of radius C, then an analogous definition of local Bayesian admissibility can be given.

Definition 1. *An action a is Γ-dominated, or locally inadmissible in a neighbourhood Γ if,*

$$\forall \pi \in \Gamma, \quad \exists a' \in \mathscr{A}, a \neq a', \quad E_\pi[L_a] > E_\pi[L_{a'}]$$

That is to say, a is not optimal under expected loss for any distribution $\pi \in \Gamma$.

In order to be able to make such statements about the KL neighbourhoods Γ_C, it helps to first define the pairwise difference in expected losses of any two actions $(a, a') \in A$. This can be interpreted as the 'regret' loss of having chosen a instead of a':

$$L_{(a,a')}(\theta) = L_a(\theta) - L_{a'}(\theta)$$

This loss function will therefore induce the following least favourable pairwise distribution:

$$\begin{aligned} \pi^{\sup}_{(a,a')} &:= \arg\sup_{\pi \in \Gamma_C} \left\{ \mathbb{E}_\pi [L_{(a,a')}(\theta)] \right\} \\ &= Z_C^{-1} \pi_I(\theta) \exp\left(\lambda_{(a,a')} [L_a(\theta) - L_{a'}(\theta)] \right) \end{aligned}$$

with expected loss $\psi^{\sup}_{(a,a')} = \int_\theta \pi^{\sup}_{(a,a')}(\theta)[L_a(\theta) - L_{a'}(\theta)]d\theta$.

In what follows, a^* denotes the optimal action with respect to the reference distribution π_I. The size of the largest KL-neighbourhood of π_I under which all actions except a^* are Γ_C dominated is:

$$C^* := \arg\sup\{C : \psi^{\sup}_{(a,a')}(C) < 0, \ \ \forall a' \in A \setminus a\}$$

If a particular action is Γ_∞ dominated, then it is globally inadmissible (this retrieves the classical notion of admissibility).

Only plotting $\psi^{\sup}_a$ as a function of the neighbourhood size C does not give any information as to the admissibility of the actions $a \in \mathscr{A}$. In order to graphically represent admissibility (inadmissibility), it is necessary to look at least favourable distributions defined over all the pairwise regret losses. By plotting $\psi^{\sup}_{(a,a')}$ as a function of constraint radius C we can look for actions that are dominated, such that there is no $\pi \in \Gamma_C$ for which they are optimal. If the region Γ_C - defined as the largest region such that all $a \in \mathscr{A} - \{a^*\}$ are inadmissible - was quite large (by some calibration method of KL divergence) this would give trust in use of the posterior π_I to say that a^* is in fact optimal.

2.2.2 Why Kullback-Leibler?

Although the Kullback-Leibler divergence is commonly used in many applications of statistics, it is necessary to review some arguments for why its use is appropriate for defining the neighbourhood Γ_C.

As noted previously, the Kullback-Leibler divergence (Kullback, 1959; Kullback and Leibler, 1951) is an f-divergence. In particular it is asymmetric, that is to say in general $\mathrm{KL}(\pi||\pi_I) \neq \mathrm{KL}(\pi_I||\pi)$. We choose to use the KL taken with respect to π instead of π_I, as it has the natural interpretation of the expected log-likelihood when Nature is π and the

statistician uses π_I. It is also the expected loss of π_I when the loss function is negative logarithmic loss (the local proper scoring rule), plus a constant term which is the entropy of π. The logarithmic scoring rule is attractive because it turns out to be the only smooth, local proper scoring rule (see Bernardo, 1979a, Theorem 2). We discuss this further in section 2.2.3. Other desirable properties are invariance to reparametrisation of the state space, and coherence as given in section 2.2.1. Its information theoretic interpretation is the amount of information (in natural units, *nits*) that a simplified model (here π_I) does not explain about the true sampling density (π).

It is a well known result that the frequentist MLE estimate θ^* will converge in KL to the closest point in Θ to the 'true' θ_0 that describes the sampling distribution. Under certain conditions, this result also applies to Bayesian updating (see Schervish, 1995, Theorem 7.80). Hence if Nature's true θ_0 is not contained within the support of the prior then the posterior will concentrate around values of θ that minimise KL divergence to θ_0. For an example of when this concentration doesn't happen, see Grünwald and van Ommen (2014). However, when the updating does concentrate, this motivates a different interpretation for the Bayesian prior on θ when the model is thought to be misspecified, see Walker (2013).

Last but not least, defining the neighbourhood around π_I with the Kullback-Leibler divergence gives a tractable analytical solution. This means the methods can be implemented at little extra computational cost. For these reasons the KL divergence is a natural candidate for defining a neighbourhood Γ_C of the approximating model π_I.

2.2.3 Unifying statistical approaches

We show that the local minimax framework unifies some diverse well-known statistical methods.

Predictive tempering

We now consider the task of providing a predictive distribution, $\widehat{\pi(y|x)}$, for a future observation y given covariates x. The local proper scoring rule in this case is known to be the self-information logarithmic loss $L(y) = -\log \pi(y|x)$ (Bernardo, 1979a). The conventional Bayesian solution is to report your honest marginal beliefs as $\widehat{\pi(y|x)} = \pi_I(y|x)$, where given a model parametrised by θ we have $\pi_I(y|x) = \int \pi(y|x,\theta)\pi_I(\theta)d\theta$. Of course this assumes that the model is true, but moreover it assumes that the prediction problem is stable in time, as in the the prediction probability contours to not change. The latter is of particular interest in data mining, known as 'concept drift' (see below). A particular occurrence is when past observations x_{t_i} are reweighed according to how 'topical' they are - meaning the closer in

time, the more weight given. An example would be scoring a tennis player's current ability, a dampening effect would be needed for matches played further back in time.

Returning to the problem of reporting best beliefs, we can see that for a particular action - reporting a distribution $\widehat{\pi(y|x)}$ - the robust local least favourable distribution is given by,

$$\pi^{\sup}(y|x) \propto \pi(y|x)e^{-\lambda \log \pi(y|x)} = \pi(y|x)^{1-\lambda}$$

for $\lambda \in [0,1]$. This has the form of tempering the predictive distribution, taking into account additional external levels of uncertainty outside of the modelling framework. In this way, predictive annealing can be seen as a local-minimax action.

Concept drift. In data mining applications we may have access to meta-data, t_i, for the i'th observation and a belief that loss is ordered or structured by the information in t. For example, t might index time and due to 'concept drift' the analysis might hold greater loss in predicting using more historic collected observations, (e.g. Section 3.1 in Hand, 2006), though more generally t_i simply contains information relative to predictive loss. In this case the natural loss function is a weighted self-information loss, based on the empirical distribution:

$$L(\theta) = -\sum_i \Delta(t_i) \log f(y_i;\theta).$$

with $\Delta(t_i) \in (0,1)$ encapsulating the relative weight of log loss to the future predictive.

For prediction of a new observation y^* given x^* this leads to the robust solution as

$$\begin{aligned} \widehat{f_{\sup}(y^*|x^*)} &\propto \int_\theta f(y^*|x^*,\theta) \left[\prod_i f(y_i;\theta)\pi(\theta)\right] e^{-\sum_i \Delta(t_i)\log f(y_i;\theta)} d\theta \\ &\propto \int_\theta f(y^*|x^*,\theta) \left[\prod_i f(y_i;\theta)^{1-\Delta(t_i)}\pi(\theta)\right] d\theta \end{aligned}$$

that can be seen to down-weight the information in y_i used to predict y^*. For example, if t_i records the time since the current prediction time then a natural penalty is $\Delta(t_i) = \exp(-\lambda t_i)$, where λ encodes a predictive forgetting factor. For a related approach see Hastie and Tibshirani (1993).

Conditional Γ-minimax priors

Suppose that for the application at hand $\Theta' = \Theta$ (in our notation defined in section 2.2). That is to say that whole parameter space enters into the loss function (no nuisance parameters).

Then the local least favourable distribution $\pi_a^{\sup}$ can be written as:

$$\pi_a^{\sup}(\theta) \propto e^{\lambda L_a(\theta)} \Pi_{i=1}^n f(x_i|\theta)\pi(\theta)$$

where $f(\cdot|\theta)$ is the likelihood function, and $\pi(\theta)$ the prior over Θ. This can be re-written as a posterior distribution with a minimax-prior $e^{\lambda L_a(\theta)}\pi(\theta)$. The prior would therefore be action-specific. This is the setting of conditional Γ-minimax priors, see Vidakovic (2000).

However, it could be argued that this approach is highly unprincipled, for example in the discussion on loss functions and priors, Jaynes (2003, page 424) disagrees with the idea that the loss function should influence in any way the choice of prior:

> We consider it an important aspect of 'objectivity' in inference - almost a principle of morality - that we should not allow our opinions to be swayed by our desires; what we believe should be independent of what we want.

However, our framework fits into a more pragmatic approach to decision theory, where the distribution $\pi_I(\theta)e^{\lambda L(a,\theta)}$ does not pretend to reflect knowledge on the true state of θ (in a sense it is no more plausible than any other distribution in the ball Γ_C) but is a useful tool to assess stability of the decision-system as it is the worst case scenario.

Gibbs posteriors

Suppose the statistician has a well defined loss function (or risk function) L defined over data x (additive) and θ and prior beliefs π about the parameter of interest θ. In the absence of a likelihood there is not a well defined model, and it is not possible to specify a full Bayesian posterior. However, if the task at hand (action) is to provide a predictive distribution, based on the parameter θ, and conditional on the data and the prior knowledge, then in the Gibbs posterior approach constructs the distribution:

$$\pi_a^{\inf} = Z_\lambda^{-1} e^{-\lambda \sum_i L(x_i,\theta)}\pi(\theta)$$

In the literature, this is known to minimise a combination of the sample risk (loss to data or empirical loss) along with a KL penalty for deviating from the prior (see Zhang, 2006a, Proposition 5.1). This is the basis of a Gibbs posterior, or an exponentially weighted PAC-Bayesian approach (see Bissiri et al., 2013; Dalalyan and Tsybakov, 2008, 2012; Zhang, 2006a,b). An example would be median estimation, where the natural loss to data would be L_1 distance, i.e.

$$L(x_i,\theta) = |x_i - \theta|$$

In this way we can interpret Gibbs posteriors as local least favourable solutions in the absence of a known sampling distribution (Bissiri et al., 2013). In the Gibbs posterior literature, the free parameter λ is thought of as being similar to the 'temperature' used in simulated annealing.

Power likelihoods

A similar concept for dealing with model misspecification is power likelihoods. As presented by Grünwald and van Ommen (2014), Bayesian inference in M_{open} does not always behave correctly. For example, in situations where the model space is non-convex, then it is possible to have a posterior distribution which does not concentrate around π^*, the nearest distribution in KL divergence to the true data-generating π^{true}. Grünwald and van Ommen's example is Bayesian linear regression with a countably infinite covariate space and an assumption of homoscedastic variance when in fact the true variance is heteroscedastic (i.e. a misspecified model). Grünwald and van Ommen show that there exists an $\eta_{crit} < 1$ such that the posterior π_I^η defined with respect to the power likelihood for values $\eta \leq \eta_{crit}$ will concentrate around π^*:

$$\pi_I^\eta(\theta) \propto \{\Pi_{i=1}^n f(x_i|\theta)\}^\eta \pi(\theta)$$

This corresponds to the local *most* favourable distribution, with respect to the additive empirical loss function defined as the negative log-likelihood of the data:

$$L(x,\theta) = \sum_{i=1}^{n} -\log f(x_i|\theta)$$

This can be seen by writing:

$$\pi^{\inf}(\theta) \propto e^{-\lambda \Sigma_{i=1}^n -\log f(x_i|\theta)} \pi(\theta) \propto \{\Pi_{i=1}^n f(x_i|\theta)\}^\lambda \pi(\theta)$$

But in this case, λ is restricted to the interval $[0, \eta_{crit}]$ where η_{crit} is unknown.

2.3 Computational decision theory

2.3.1 'Shaking the model'

In many applied settings, the statistician would not have access to the analytical form of the posterior distribution π_I, but to an approximation in form of Monte Carlo samples $\{\theta_i\}_{i=1}^m$

(obtained using MCMC for example). We denote this approximation by $\hat{\pi}_I$, defined as:

$$\hat{\pi}_I(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\theta_i}(\theta) \tag{2.3}$$

where $\mathbb{1}_{\theta_i}$ is the indicator function taking value 1 at θ_i and zero elsewhere. Each sample θ_i has equal weight $1/m$. We note for non-degenerate functionals $g(\theta)$ of interest $\mathbb{E}_{\hat{\pi}_I}[g(\theta)]$ converges to $\mathbb{E}_{\pi_I}[g(\theta)]$, as $m \to \infty$. Thus it is possible to construct estimates of the expected loss of an action $a \in \mathscr{A}$ using the estimator

$$\hat{\psi}_a = \frac{1}{m}\sum_{i=1}^{m} L(a, \theta_i) \tag{2.4}$$

Such approximations of expected loss are discussed for example by Müller (2005) in the context of the optimal design of experiments. In the case where the action space $\mathscr{A}$ is high-dimensional, it can be a difficult task in itself to find the optimal a^* with respect to the posterior π_I. Evaluations of 2.4 are expensive, making a naive search inefficient. Müller (2005) considers ways in which this search can be carried out more efficiently, for example using model augmentation (Monte Carlo methods which solve both problems: sampling from the posterior and exploring regions of minimum loss), or simulated annealing which makes the loss surface more peaked (easier to find the minimum value). This is an important problem, but a formal consideration is beyond the scope of this work. Chapter 5, section 5.3 looks at an application where the space of all decisions is too big ($\approx 10^{30}$) and so we implement a very simple stochastic search method to identify good candidates. But in general, we are interested in the problem of comparing suitable candidates $a_1, .., a_K$ that either have been identified as minimisers of equation 2.4 or selected via some screening method (such as stochastic search, Monte Carlo). Thus we consider the set $\mathscr{A}$ to be discrete with finite cardinality K, and that K not 'too large', i.e. it is possible to compute the matrix of loss value with elements $\{L(a_j, \theta_i)\}_{j=1..K}^{i=1..m}$. We are not interested in the problem of finding the candidates - this would be done using algorithms such as in Müller (2005) - but rather how to assess whether the estimate given by 2.4 is stable in order to be used to guide choice of $a^* \in \mathscr{A}$. This stability is defined with respect to the intrinsic *ex-post* misspecification of the decision-system $\{\Theta, \pi_I, L, \mathscr{A}\}$.

The area of computational decision theory roughly deals with the following issues in the context of decision making using probabilistic models:

1. Untractable/complex $\pi_I(\theta)$ necessitating approximate methods (for example via Monte Carlo sampling).

2. Large action space $\mathscr{A}$ and costly evaluations of the loss function $L(a,\theta)$.

3. Misspecification in both $\pi_I(\theta)$ and $L(a,\theta)$.

Our methodology deals mainly with points (i) and (iii). We show that it is possible to use the theory developed in the previous sections to implement computationally cheap methods for evaluating the sensitivity of the decision-system using $\hat{\pi}_I$ as the reference (approximating) model. This can be seen as deterministic retrospective reweighing scheme for the estimator 2.4, that 'shakes' the decision system.

2.3.2 Importance sampling local least favourable distributions

Following on from section 2.2.1, the distribution of interest is the action specific local least favourable distribution $\pi_a^{\sup}(\theta) \propto \pi_I(\theta)e^{\lambda L(a,\theta)}$. It would be possible to directly target this distribution using Monte Carlo and thus compute estimates of local-minimax expected loss for $a \in \mathscr{A}$. The estimates appear amenable to Sequential Monte Carlo samplers indexed on λ, but we do not explore that here. There have been proposals in the literature for MCMC sampling from $\pi_a^{\sup}$ in specific cases, for example when the loss is defined as the sample risk, and the reference distribution is a prior used for variable selection, see Jiang and Tanner (2008, section 7). However, in general, this is not very efficient, requiring different Monte Carlo runs for each action.

An easier solution comes from importance sampling. For 'small' values of the KL radius C, it is possible to importance sample the local least favourable distribution $\pi_a^{\sup}$ using $\theta_i \sim \pi_I$. We know that the importance ratio (up to a multiplicative constant that is the normalising constant Z_C^{-1} from equation 2.1) between the two densities is simply the tilting factor:

$$w_i = \frac{\pi_a^{\sup}(\theta_i)}{\pi_I(\theta_i)} \propto e^{\lambda L(a,\theta_i)}$$

Hence we have the following importance estimate of $\pi_a^{\sup}(\theta)$, which we denote $\hat{\pi}_a^{\sup}$:

$$\hat{\pi}_a^{\sup}(\theta) = \frac{1}{\sum_i w_i}\sum_i w_i \mathbb{1}_{\theta_i} = \frac{1}{\sum_i e^{\lambda L(a,\theta_i)}}\sum_i e^{\lambda L(a,\theta_i)}\mathbb{1}_{\theta_i}$$

For large values of λ this approximation will be bad, with the variance of the un-normalised weights going to infinity. However, this is a weakness inherent to using a Monte Carlo representation of π_I in the first place. This suggests that we should explore 'small' neighbourhoods around π_I to implement the local minimax framework, where small is defined relative to the importance sampling procedure. In this case, the size of the neighbourhood C

would in part depend on the number of samples m from π_I. Measures such as the variance of the weights, or the effective sample size (see Liu, 2008, section 2.5.3) can then be used to calibrate suitable values of the KL radius C. We also look at a simpler score, the Gini coefficient on the importance weights, see section 2.3.4.

2.3.3 KL and reverse KL

The change of neighbourhood from $\mathrm{KL}(\pi \parallel \pi_I)$ to $\mathrm{KL}(\pi_I \parallel \pi)$ results in a non-analytic solution to the local least favourable distribution. However we can use numerical methods to compute the minimax optimisation. We first consider the numerical solution to $\pi_a^{\sup} = \arg\sup_{\pi \in \Gamma_C^{\mathrm{rev}}} \mathbb{E}_\pi[L_a(\theta)]$, with Γ_C^{rev} defined as $\Gamma_C^{\mathrm{rev}} = \{\pi : \mathrm{KL}(\pi_I \parallel \pi) \leq C\}$ for $C \geq 0$. We consider again the stochastic representation $\hat{\pi}_I$ of π_I given in equation 2.3. where θ_i are i.i.d. draws from π_I and $m \to \infty$. To make the solution tractable in defining a KL neighbourhood around π_I we will use the neighbourhood around $\hat{\pi}_I$. Moreover, in considering the KL divergence between π_I and an alternative model $\pi \in \Gamma_C^{\mathrm{rev}}$ we will work with a stochastic approximation to π represented as mixtures of the atoms $\{\delta_{\theta_1}, \delta_{\theta_2}, \ldots, \delta_{\theta_m}\}$:

$$\hat{\pi} = \sum_i w_i \delta_{\theta_i}(\theta)$$

for $0 \leq w_i \leq 1$, $\sum_i w_i = 1$, where the w_i's can be interpreted as importance weights $w_i \propto \pi(\theta_i)/\pi_I(\theta_i)$, so that $\mathbb{E}_{\hat{\pi}}[g(\theta)] \to \mathbb{E}_\pi[g(\theta)]$, as $m \to \infty$.

The KL divergence between π_I and π can then be approximated via the KL divergence of their stochastic representations,

$$\mathrm{KL}(\hat{\pi}_I, \hat{\pi}) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{m * w_i}. \tag{2.5}$$

From these definitions, we will now look for the probability measure maximisation

$$\hat{\pi}_a^{\sup} = \arg \sup_{\tilde{\pi} \in \hat{\Gamma}_C^{\mathrm{rev}}} \{\mathbb{E}_{\hat{\pi}}[L_a(\theta)]\} \tag{2.6}$$

Given the atomic structure of $\hat{\pi}$ the maximisation (2.6) leads to a convex optimisation in the weights,

$$\hat{\pi}_a^{\sup} = \sum_i w_i^* \delta_{\theta_i}(\theta)$$
$$w^* = \arg\sup_w \left\{ \sum_i w_i L_a(\theta_i) : -\frac{1}{m}\sum_i \log(w_i) \leq C + \log m, \sum_i w_i = 1 \right\}$$

for which standard numerical methods can be applied.

2.3.4 Calibration of Kullback-Leibler

Previous work

A few methods for interpreting and calibrating Kullback-Leibler have been proposed in the literature. We go over the main ideas in order to show that there is - to our knowledge at least - no fully satisfying solution so far. The simplest method was given by McCulloch (1989). His idea is to use a coin toss as a reference experiment in which is possible to gauge how much misspecification is acceptable. That is to say, given a biased coin (the misspecified model) where P(Heads)= p, which values of $p \in [0,1]$ could be accepted to simulate a non-biased coin ($p^* = 1/2$), the correctly specified model? The range of acceptable values $p \in [0,1]$ correspond to a range of Kullback-Leibler divergences, given by the mapping

$$p \to -\frac{1}{2}\log\{4p(1-p)\}$$

This gives a direct interpretation of KL values, for example a KL divergence of 0.83 would correspond to using a biased coin with $p = 0.95$. However, it is not clear why this should extend to continuous models of arbitrary dimension, and whether the interpretation of the KL values should remain the same.

Hansen and Sargent approach the problem from a different angle. When calibrating the radius C of a neighbourhood Γ_C centred at π_I, they ask whether the least favourable distribution $\pi_a^{\sup}$ is 'indistinguishable' from π_I if the statistician had a (small) finite sample of size N (see chapter 9, Hansen and Sargent, 2008). One can think of this as a hypothesis testing scenario, with two possible models. The authors compute detection error probabilities, using model selection techniques, based on likelihood ratio tests. This allows the user to determine a plausible probability (function of the radius C) of selecting the wrong model given the available data, which can be inverted to find C (by simulation). The argument is that the user should worry about statistically indistinguishable models π which change the

expected loss of an action a. This approach is principled but in many cases even the detection error probabilities could be difficult to compute.

Finally, the KL provides a useful upper bound on the loss distribution Z_a of an action a under the reference distribution π_I (see Ahmadi-Javid, 2012; Breuer and Csiszár, 2013a). Let $\alpha := e^{-C}$, where C is again the radius of the KL neighbourhood. Then:

$$P\{Z_a \geq \mathbb{E}_{\pi_a^{\sup}}[L_a(\theta)]\} \leq \alpha$$

It allows the user to upper bound the tail of the loss distribution Z_a, which is useful but it is unclear how tight this bound is.

Ad hoc calibration

In the computational decision theory set-up, the calibration should be done with the representation of the decision system in mind. In particular, when a distribution $\pi(\theta)$ is represented as a 'bag of samples' $\{\theta_i\}_{i=1}^m$, the weights w_i used to approximate $\pi^{\sup}$ are important for understanding the accuracy of the approximation. As an example, this motivates the re-sampling step used in sequential Monte Carlo methods (which kills off particles with too low weight). In a similar manner, when we approximate the least favourable distribution $\pi_a^{\sup}(\theta)$, the quality of the importance sampling decreases as the KL radius increases, which means we are constrained to a maximum ball size. If the weights corresponding to the distribution π_I are all $1/m$, then for a sequence of distributions $\hat{\pi}_a^{\sup,C_j}$ at increasing divergences C_j, we can use the same methods as those used for the calibration of importance sampling itself. The most widely used 'rule of thumb' method is the *effective sample size* (ESS), which here would be defined as:

$$ESS(\hat{\pi}_a^{\sup}, m) = \frac{m}{1 + \mathrm{Var}_{\pi_I}(\mathbf{w})}$$

where $\mathbf{w}$ is the vector of importance weights. In order to avoid having to compute the normalising constant of $\pi_a^{\sup}$, the variance term can be estimated by the coefficient of variation of the unnormalised weights (Liu, 2008, section 2.5.3):

$$cv^2(\mathbf{w}) = \frac{\sum_i (w_i - \bar{w})^2}{(m-1)\bar{w}^2}$$

where $\bar{w}$ is the sample average of the weights w_i.

A simple alternative to using the ESS is the Gini coefficient on the normalised weights. This measures the 'inequality' of the weight distribution, with the situation when all weights

are $1/n$ corresponding to perfect equality (coefficient of zero) and the global least favourable distribution (one sample has all the weight) corresponding to a coefficient of 1.

Another method comes from the link with the power priors from Ibrahim and Chen (2000). These stem from the idea of tempering priors that are computed from historical data. This is very similar to the posterior tempering given in section 2.2.3. The tempering coefficient directly corresponds to a Kullback-Leibler divergence. The calibration of this tempering parameter is discussed in the power prior literature, for an example its application see Brian (2010). This tempering can also be interpreted as a way of reducing the data size - similar to the concept of the 'Safe Bayes' (Grünwald and van Ommen, 2014) where learning is deliberately 'slowed down' (power likelihood: suboptimal learning from data) in order to achieve correct behaviour.

A final suggestion is to use the data themselves for calibration process, in the form of posterior predictive checks (see, for example Gelman et al., 1996). This can be extended in a simple way for the least favourable distribution π_a^{sup} by using reweighing (importance sample weights) to generate replicate datasets.

2.3.5 Implementation - R package

The framework discussed in this chapter is implemented in an R package *decisionSensitivityR*. This takes three inputs: a set of m Monte Carlo samples $\theta_i \sim \pi_I$, a discrete set of actions, and a loss function $L(a,\theta)$ defined over $\mathscr{A} \times \Theta$. The software executes the following steps:

- Compute a loss matrix, with entry $l_{ij} = L(a_j, \theta_i)$. Select the top k actions for analysis (k is a user specified parameter).
- For each action, compute the expected losses under exponential tilting. Invert the KL-λ relationship to find the correct values of λ.
- Same procedure for the 'regret-loss' between the optimal Bayes action and the $k-1$ others.

The output of the main function *preliminaryAnalysis* is a series of plots that allow the user to diagnose the following:

- What are suitable candidate values for the KL radius C?
- Is the optimal action sensitive to model perturbations within a reasonable neighbourhood Γ_C?
- Sensitivity of the decision system under D_{open}

The third point is developed in chapter 4 where we look at a series of plots that explore the sensitivity of a decision system

2.4 Discussion

Decision making in D_{open} cannot only depend on expected loss estimates to score actions. It must take into account the knowledge that the posterior distribution (or reference model) π_I is an approximation of the statistician's 'true' posterior beliefs, or Nature 'true' data generating mechanism. It is therefore necessary to assess the sensitivity of the expected loss quantities. This chapter considers the exploration of neighbourhoods of 'close' models centred at the approximating model π_I. In particular we highlight the following points for why this is an attractive methodology for D_{open} decisions:

- It bridges the two dominant decision paradigms: minimax and expected loss.
- Provides a principled approach for perturbing a statistical model in a decision-theoretic context.
- Computationally cheap to implement when the model is represented by Monte Carlo samples.

In chapter 4 we explore further the computational element of this framework via diagnostic plots that aim to visually represent potential weaknesses in the decision system. This gives a partial solution to the issue of calibration of the Kullback-Leibler divergence. However, interpretation of Kullback-Leibler and its calibration remains in our view an open problem.

The formal framework lacks the following aspects in order to be complete. Firstly, it does not give the statistician any recommendations for how sensitivity to misspecification should be reduced. A possible solution to this is presented in chapter 4, discussing diagnostics. This can allow the user to distinguish between sensitivity coming from the choice of prior or the data, i.e. the choice of likelihood. Another idea is to formalise a 'reconsider' action. A formulation of this is suggested in Karabatsos (2006). Formally this must specify the cost or added expected loss of more modelling, which in practice may be hard. Karabatsos looks at the information lost when using a restricted but more interpretable model versus a fully nonparametric but unwieldy model (that he considers is the 'true' model because of its wide support). Our opinion is that graphical representations of the model and its restrictions can be the most useful for practitioners when the loss functions are difficult to define.

Chapter 3

Ex-post nonparametric model extensions

The previous chapter explored the use of *local least favourable distributions* to assess the stability of a decision-system. Results concerning the analytical form of these distributions, the Bayesian coherency of the approach, and the simplicity of implementation, all make this framework attractive. However, it inherits the same drawbacks as Wald's minimax theory, notably that in many situations, the statistician does not believe in a 'reactive' Nature, and thus the local minimax framework is too conservative a paradigm. For these reasons, it may be preferable to explore a whole neighbourhood Γ of the reference model π_I, and analyse the loss distribution induced by the set of models $\pi \in \Gamma$. This necessitates defining a measure over all the distributions within a neighbourhood of π_I. This is a standard problem in Bayesian nonparametrics and in this chapter we consider two well known nonparametric processes, the Pólya Tree Process (PT) and the Dirichlet Process (DP). We present results concerning the distance of random draws from these processes with respect to the baseline measure, characterised in Kullback-Leibler divergence. This enables sampling within a KL neighbourhood of the posterior distribution π_I. Whereas chapter 2 (section 2.3) considers deterministic reweighing of these samples, this chapter explores stochastic reweighing using these nonparametric model extensions. Bootstrap procedures for reweighing are also considered and analogous results are given characterising their divergence in Kullback-Leibler.

3.1 Pólya trees

If the model π_I is believed to be misspecified, then a natural approach from a Bayesian perspective is to look at distributions over models around π_I. This would be instead of only considering the least favourable distribution within a neighbourhood of π_I, as was done in Chapter 2. Bayesian nonparametrics provides the necessary tools for such an approach. In

this chapter we look at two popular Bayesian nonparametric constructions, the Pólya tree process and the Dirichlet process. Both of these can be centred at π_I, and random draws can then be characterised in terms of the Kullback-Leibler divergence from the baseline distribution for the Pólya tree, and in terms of L_1 distance for the loss distribution in the case of the Dirichlet process. Using nonparametric model extensions in this way allows us to characterise the effect of local perturbations to the approximating model π_I. We also look at the connections between the Dirichlet process and a generalised version of the Bayesian bootstrap which also provides a method for sampling distributions 'close to' π_I. We first discuss the Pólya tree process and derive some results concerning the Kullback-Leibler divergence of draws from this process with respect to the baseline measure.

3.1.1 Introduction

The Pólya tree process was first introduced by Lavine (1992) as an intermediate class of distributions between the Dirichlet process (Ferguson, 1973) and tailfree processes (Fabius et al., 1964; Freedman, 1963). Although the Pólya tree process is an elegant mathematical construction providing tractable inference and sampling, its dependence on a sample space partition impacts inference and is badly suited to high dimensions. However, we are interested in using Pólya trees as ex-post model extensions over the subspace of Θ that enters into the loss function $L(a,\theta)$ (see section 1.3.1 for notation outline). For a fixed loss function $L_a(\cdot)$, this is equivalent to centring the Pólya tree on the distribution of loss $F_a(Z)$ (where Z is the loss), which is a 1-dimensional object.

Random processes such as the Pólya tree are usually characterized by their mean and variance around a centring distribution π_0. For instance, if we have a random distribution π with a Dirichlet process law, denoted $\pi \sim \mathrm{DP}(\alpha,\pi_0)$, where π_0 is the centring distribution, then $\mathbb{E}(\pi)=\pi_0$, and α is a precision parameter that controls the dispersion of π from π_0. Similarly, if π is a random distribution with law governed by a Pólya tree process (using notation from Hanson, 2006), $\pi \sim \mathrm{PT}(\alpha,\rho,\pi_0)$, where ρ denotes the concentration function, the choice of α and the partition structure Π defining the tree, will imply that the draws are centred around π_0, where α is again the precision parameter. Moreover, the concentration function $\rho(\cdot)$ controls the speed at which the variance of the branching probabilities defining the Pólya Tree increase or decrease. In this section we highlight the connection between the parametrisation of the Pólya Tree process (PT) and the distance in Kullback-Leibler divergence between random draws and the baseline measure. As an aside, these results give insight into the parametrisation of the Pólya tree process and how it can be used to sample distributions within a KL neighbourhood of the reference distribution π_I using rejection sampling. The results presented in this chapter allows for a principled method for choosing

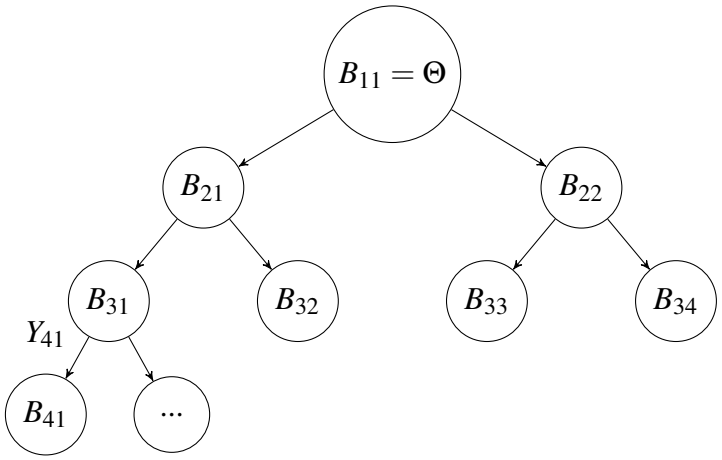

Fig. 3.1 Illustration of the partition tree and associated probabilities in a Pólya tree process. The partition tree is fixed but the associated probabilities are random.

both the truncation level M. These results link up to the previous chapter and allow for a principled construction of a nonparametric extension to the posterior model π_I.

We first give an overview of the Pólya tree construction and the notation used in this chapter.

3.1.2 Notation and construction

The construction of a Pólya tree process relies on a partition of the sample space Θ. For our purposes, the space on which we will define the Pólya tree will in fact be $\mathbb{R}$, so we consider $(\mathbb{R}, \mathscr{B})$ as the measurable space of interest, with $\mathbb{R}$ the real line and $\mathscr{B}$ the Borel sigma algebra of subsets of $\mathbb{R}$.

To define the binary partition tree, we use the notation from Nieto-Barajas and Müller (2012): $\Pi = \{B_{mj} : m \in \mathbb{N}, j = 1,..,2^m\}$, where the index m denotes the level in the tree and j the location of the partitioning subset at that level. The sets at level 1 are denoted by (B_{11}, B_{12}); the partitioning subsets of B_{11} are (B_{21}, B_{22}), and $B_{12} = B_{23} \cup B_{24}$, such that $(B_{21}, B_{22}, B_{23}, B_{24})$ denote the sets at level 2. In general, at level m, the set B_{mj} is split into two disjoint sets $(B_{m+1,2j-1}, B_{m+1,2j})$, where $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$ and $B_{m+1,2j-1} \cup B_{m+1,2j} = B_{mj}$.

Every set B_{mj} has an associated random branching probability Y_{mj}. We define $Y_{m+1,2j-1} = F(B_{m+1,2j-1} \mid B_{mj})$, and $Y_{m+1,2j} = 1 - Y_{m+1,2j-1} = F(B_{m+1,2j} \mid B_{mj})$. We denote by $\mathscr{Y} = \{Y_{mj}\}$ the set of random branching probabilities associated with the partition tree Π. This recursive splitting is shown in figure 3.1, each node corresponding to a subset of Θ, and each arrow given a branching probability Y_{mj}.

Definition 2. *(Lavine, 1992). Let $\mathscr{A}_m = \{\alpha_{mj},\, j = 1,\ldots,2^m\}$ be non-negative real numbers, $m = 1,2,\ldots$, and let $\mathscr{A} = \bigcup \mathscr{A}_m$. A random probability measure π on $(\mathbb{R},\mathscr{B})$ is said to have a Polya tree prior with parameters $(\Pi,\mathscr{A})$, if for $m = 1,2,\ldots$ there exist random variables $\mathbb{Y}_m = \{Y_{m,2j-1}\}$ for $j = 1,\ldots,2^{m-1}$, such that the following hold:*

1. *All the random variables in $\mathscr{Y} = \cup_m\{\mathscr{Y}_m\}$ are independent.*
2. *$\forall m \in \mathbb{N}$ and every $j = 1,\ldots,2^{m-1}$, $Y_{m,2j-1} \sim \text{Beta}(\alpha_{m,2j-1},\alpha_{m,2j})$.*
3. *$\forall m \in \mathbb{N}$ and every $j = 1,\ldots,2^m$*

$$\pi(B_{mj}) = \prod_{k=1}^{m} Y_{m-k+1,j_{m-k+1}^{(m,j)}},$$

where $j_{k-1}^{(m,j)} = \lceil j_k^{(m,j)}/2 \rceil$ is a recursive decreasing formula, whose initial value is $j_m^{(m,j)} = j$, that locates the set B_{mj} with its ancestors upwards in the tree. $\lceil\cdot\rceil$ denotes the ceiling function, and $Y_{m,2j} = 1 - Y_{m,2j-1}$ for $j = 1,\ldots,2^{m-1}$.

There are several ways of centring the process around a parametric probability measure π_0. The simplest and most used method (Hanson and Johnson, 2002) consists of matching the partition with the dyadic quantiles of the desired centring measure and keeping α_{mj} constant within each level m. More explicitly, at each level m we take

$$B_{mj} = \left[\pi_0^{-1}\left(\frac{j-1}{2^m}\right), \pi_0^{-1}\left(\frac{j}{2^m}\right)\right], \tag{3.1}$$

for $j = 1,\ldots,2^m$, with $\pi_0^{-1}(0) = -\infty$ and $\pi_0^{-1}(1) = \infty$. If we further take $\alpha_{mj} = \alpha_m$ for $j = 1,\ldots,2^m$ we get $\mathbb{E}[\pi(B_{mj})] = \pi_0(B_{mj})$ where π is a random draw from $\text{PT}(\alpha,\rho,\pi_0)$.

In particular, we take $\alpha_{mj} = \alpha\rho(m)$, so that the parameter α can be interpreted as a precision parameter of the Polya tree (Walker and Mallick, 1997), and the function ρ controls the speed at which the variance of the branching probabilities moves down in the tree. In this chapter, we refer to α as the concentration parameter and $\rho(m)$ the concentration function. According to Ferguson (1974), $\rho(m) = 1/2^m$ defines an a.s. discrete measure that coincides with the Dirichlet process (Ferguson, 1973), and $\rho(m) = 1$ defines a continuous singular measure. Moreover, if ρ is such that $\sum_{m=1}^{\infty}\rho(m)^{-1} < \infty$ it guarantees that π is absolutely continuous (Kraft, 1964), e.g., $\rho(m) = m^2, m^3, 2^m, 4^m$.

We can truncate the tree at a finite level M to define a finite PT process. At the lowest level M, we assign the probability within each set B_{Mj} according to π_0. In this case the

random probability measure defined will have a density of the form:

$$\pi(\theta) = \prod_{m=1}^{M} Y_{m,j_m^{(x)}} 2^M \pi_0(\theta), \tag{3.2}$$

for $\theta \in \mathbb{R}$, and with $j_m^{(\theta)}$ identifying the set at level m that contains θ. This maintains the condition $\mathbb{E}[\pi] = \pi_0$. We denote a finite Polya tree process as $\mathrm{PT}_M(\alpha, \rho, \pi_0)$. When $M \to \infty$ the finite tree converges to a Polya tree process.

3.1.3 Properties of random draws

In this section we look at the distance in Kullback-Leibler divergence of random draws from a Pólya tree process. We recall this divergence between two probability measures π and π' is:

$$\mathrm{KL}(\pi||\pi') = \mathbb{E}_\pi \left[\log \left\{ \frac{\pi(\theta)}{\pi'(\theta)} \right\} \right] = \int \log \left\{ \frac{\pi(\theta)}{\pi'(\theta)} \right\} \pi(\theta) \mathrm{d}\theta. \tag{3.3}$$

where π is absolutely continuous with respect to π'.

In what follows, π_0 will refer to the centring distribution of the Pólya tree, and π a random draw, $\pi \sim \mathrm{PT}_M(\alpha, \rho, \pi_0)$. It is not difficult to show that the KL divergence between π_0 and π is a random variable that does not depend on π_0, and is given by:

$$\mathrm{KL}(\pi_0||\pi) = -\sum_{m=1}^{M} \sum_{j=1}^{2^m} (\log Y_{mj}) \frac{1}{2^m} - M \log 2. \tag{3.4}$$

Since the KL divergence measure is asymmetric, we can reverse the role of π and π_0. In this case the reverse KL divergence becomes:

$$\mathrm{KL}(\pi||\pi_0) = \sum_{m=1}^{M} \sum_{j=1}^{2^m} (\log Y_{mj}) \prod_{k=1}^{m} Y_{m-k+1, j_{m-k+1}^{(m,j)}} + M \log 2. \tag{3.5}$$

We now present results that characterize the first two moments of these divergences.

Proposition 1. *Let $\pi \sim \mathrm{PT}_M(\alpha, \rho, \pi_0)$. Then the Kullback-Leibler divergence between π_0 and π, defined in* (3.4)*, has mean and variance given by*

$$\mathbb{E}\{\mathrm{KL}(\pi_0||\pi)\} = \sum_{m=1}^{M} \{\psi_0(2\alpha\rho(m)) - \psi_0(\alpha\rho(m)) - \log 2\} \tag{3.6}$$

and

$$Var\{\mathrm{KL}(\pi_0||\pi)\} = \sum_{m=1}^{M} \frac{1}{2^m} \{\psi_1(\alpha\rho(m)) - 2\psi_1(2\alpha\rho(m))\} \tag{3.7}$$

where $\psi_0(\cdot)$ and $\psi_1(\cdot)$ denote the digamma and trigamma functions respectively[1].

Proof. The expected value is a consequence of the fact that the geometric mean of a Beta random variable is: $\mathbb{E}(\log Y_{mj}) = \psi_0[2\alpha\rho(m)] - \psi_0[\alpha\rho(m)]$. For the variance, we note that the random variables Y_{mj} are independent across m, and for the same m, Y_{mj} and Y_{mk} are independent for $|k-j| > 1$. Noting that $\mathrm{Var}(\log Y_{mj}) = \psi_1[\alpha\rho(m)] - \psi_1[2\alpha\rho(m)]$ and since $Y_{m,2j} = 1 - Y_{m,2j-1}$, for $j = 1,\ldots,2^{m-1}$, with $\mathrm{Cov}[\log Y_{m,2j-1}, \log(1-Y_{m,2j})] = -\psi_1[2\alpha\rho(m)]$, the result follows. □

We now concentrate on the limiting behaviour of the expected KL value as a function of the finite tree level M. For some cases of the function $\rho(\cdot)$ this limit is finite. In particular we consider the following concentration functions:

$$\rho_1(m) = 1/2^m, \;\; \rho_2(m) = 1, \;\; \rho_3(m) = m^\delta, \;\; \text{and} \;\; \rho_4(\mathrm{m}) = \delta^{\mathrm{m}}, \tag{3.8}$$

where $\delta > 1$, to define discrete, singular and two absolutely continuous measures, respectively.

Corollary 1. *Let $\mathscr{E}_M := \mathbb{E}\{\mathrm{KL}(\pi_0||\pi)\}$, given in Proposition 1, to make explicit the dependence on the maximum level M. For the families $\rho_3(m)$ and $\rho_4(m)$ in expression (3.8), the limit of the expected KL divergence, as $M \to \infty$, is bounded respectively by:*

$$\lim_{M\to\infty} \mathscr{E}_M \leq \frac{1}{4\alpha}\zeta(\delta) + \frac{1}{\alpha^2}\zeta(\delta^2) \tag{3.9}$$

$$\lim_{M\to\infty} \mathscr{E}_M \leq \frac{\alpha(\delta+1)+4}{4\alpha^2(\delta^2-1)} \tag{3.10}$$

where $\delta > 1$, and $\zeta(\delta) = \sum_{n=1}^{\infty} n^{-\delta}$ is the Riemann zeta function.

Proof. The digamma function can be expanded as: $\psi_0(x) = \log x - (1/2)x^{-1} - \mathscr{O}(x^{-2})$, from which these inequalities follow. □

It is possible to derive analogous results in the case of the reverse KL given in 3.5.

[1] The digamma function is defined as the logarithmic derivative of the gamma function, i.e. $\psi_0(x) = \frac{\mathrm{d}}{\mathrm{dx}} \log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. In similar fashion, the trigamma function is defined as the second derivative.

Proposition 2. *Let* $\pi \sim \mathrm{PT}_M(\alpha, \rho, \pi_0)$. *Then the Kullback-Leibler divergence between* π *and* π_0, *defined in* (3.5), *has mean and variance given by*

$$\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\} = \sum_{m=1}^{M} \{\psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+1) + \log 2\}$$

and

$$Var\{\mathrm{KL}(\pi||\pi_0)\} = A + B,$$

where

$$\begin{aligned}
A &= \sum_{m=1}^{M} \left[\left\{ \prod_{k=1}^{m} \left(\frac{\alpha\rho(k)+1}{2\alpha\rho(k)+1} \right) \right\} \lambda_5(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right] \\
B &= \sum_{m=1}^{M} \left(\left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k)+1}{2\alpha\rho(k)+1} \right) \right\} \lambda_6(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right. \\
&\quad + \sum_{j=1}^{m-1} \left[\left(\frac{\alpha\rho(j)}{2\alpha\rho(j)+1} \right) \left\{ \prod_{k=1}^{j-1} \left(\frac{\alpha\rho(k)+1}{2\alpha\rho(k)+1} \right) \right\} \lambda_2^2(m) - \left(\frac{1}{2} \right)^j \lambda_2^2(m) \right] \\
&\quad + 2 \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k)+1}{2\alpha\rho(k)+1} \right) \right\} \sum_{j=m+1}^{M} \left\{ \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_3(m)\lambda_2(j) \right. \\
&\quad \left. \left. + \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_4(m)\lambda_2(j) - \lambda_2(m)\lambda_2(j) \right\} \right)
\end{aligned}$$

with

$$\begin{aligned}
\lambda_2(m) &= \psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+1) \\
\lambda_3(m) &= \psi_0(\alpha\rho(m)+2) - \psi_0(2\alpha\rho(m)+2) \\
\lambda_4(m) &= \psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+2) \\
\lambda_5(m) &= \psi_1(\alpha\rho(m)+2) - \psi_1(2\alpha\rho(m)+2) + \{\psi_0(\alpha\rho(m)+2) - \psi_0(2\alpha\rho(m)+2)\}^2 \\
\lambda_6(m) &= \{\psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+2)\}^2 - \psi_1(2\alpha\rho(m)+2)
\end{aligned}$$

Proof. The expected value follows by using independence properties and by noting that $\mathbb{E}\{(\log Y_{mj})Y_{mj}\} = \lambda_2(m)/2$. For the variance, we first bring the variance operator within the sum by splitting it into the sum of variances of each element plus the sum of covariances. The variance of each element is defined in terms of first and second moments and relies on

independence properties. Noting that:

$$\begin{aligned}
\mathbb{E}\{(\log Y_{mj})Y_{mj}\} &= \lambda_2(m)/2 \\
\mathbb{E}\{(\log Y_{mj})Y_{mj}^2\} &= \frac{1}{2}\left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1}\right)\lambda_3(m) \\
\mathbb{E}\{(\log Y_{mj})Y_{mj}(1-Y_{mj})\} &= \frac{1}{2}\left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1}\right)\lambda_4(m) \\
\mathbb{E}\{(\log Y_{mj})^2 Y_{mj}^2\} &= \frac{1}{2}\left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1}\right)\lambda_5(m) \\
\mathbb{E}\{(\log Y_{mj})\log(1-Y_{mj})Y_{mj}(1-Y_{mj})\} &= \frac{1}{2}\left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1}\right)\lambda_6(m)
\end{aligned}$$

the result is obtained. □

Figures 3.2 and 3.3 illustrate the behaviour of the mean and standard deviation, respectively, as a function of the truncation level M for the two KL divergences (3.4) (centred at π_0, empty dots) and (3.5) (centred at π, solid dots). The four panels in each figure correspond to choices of $\rho(m) = 1/2^m, 1, m^\delta, \delta^m$, as given in (3.8). In all cases we use $\alpha = 1$, and $\delta = 2$ (the so-called canonical choice). The plots show that $\mathbb{E}\{\mathrm{KL}(\pi_0||\pi\} \geq \mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\}$ for all M and for all functions ρ. Apart from the singular continuous case, $\rho_2(m) = 1$, the variances of $\mathrm{KL}(\pi_0||\pi)$ are also larger that those of $\mathrm{KL}(\pi||\pi_0)$.

In the case of the Dirichlet process where $\rho_1(m) = 1/2^m$, the mean value of the KL and the reverse KL diverge to infinity as $M \to \infty$[2]. The KL (3.4) increases at an exponential rate whereas for the reverse KL (3.5) the growth rate is constant. As for the standard deviations, that of the KL also diverges as $M \to \infty$, however, that of the reverse KL converges.

The precision function $\rho_2(m) = 1$, which defines a singular continuous random distribution (Ferguson, 1974), has asymptotic constant expected values for both KL and reverse KL in the limit of M. The variance of the KL converges to a finite value when $M \to \infty$, but for the reverse KL the variance increases at a constant rate as a function of M. In the case of the two continuous processes, obtained with precision functions ρ_3 and ρ_4, the expected values for KL and the reverse KL converge in the limit, as given by the upper bounds in Corollary 1. Interestingly, the variances for the two KL divergences are asymptotically constant.

3.1.4 Parametrisation using KL divergence

One consequence of these results concerns the parametrisation of Pólya tree processes in practical applications, for example in the context of Bayesian inference. The user must

[2]Figure 3.2 appears to show that $\mathbb{E}\{KL(\pi||\pi_0)\}$ remains constant, but this is an artefact due to the scale.

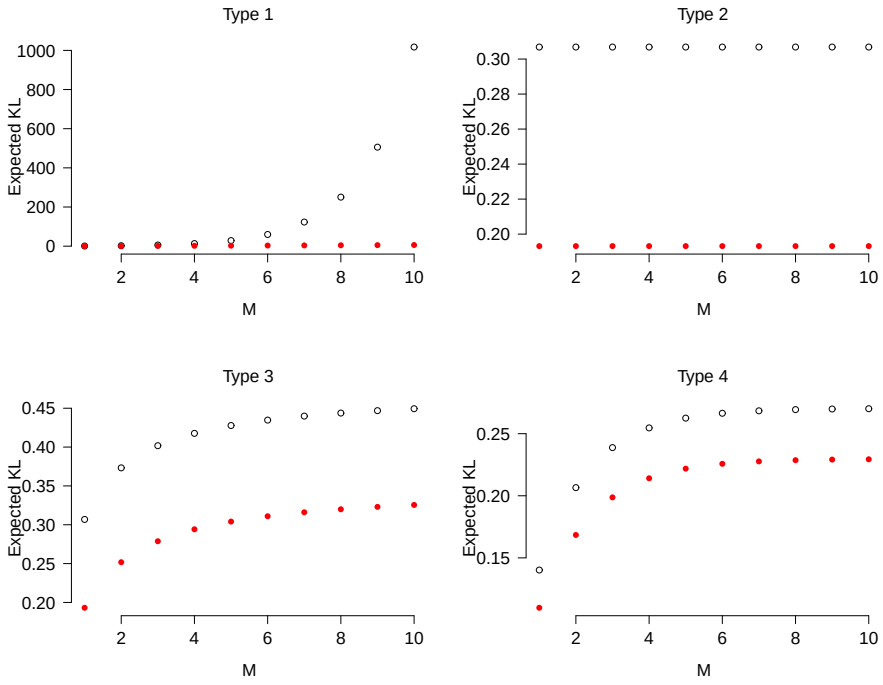

Fig. 3.2 Comparison between expected values of KL for different values of M. $\mathbb{E}\{\mathrm{KL}(\pi_0||\pi)\}$ (empty dots) and $\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\}$ (solid dots). Type 1 to 4 denote the different ρ functions given in (3.8).

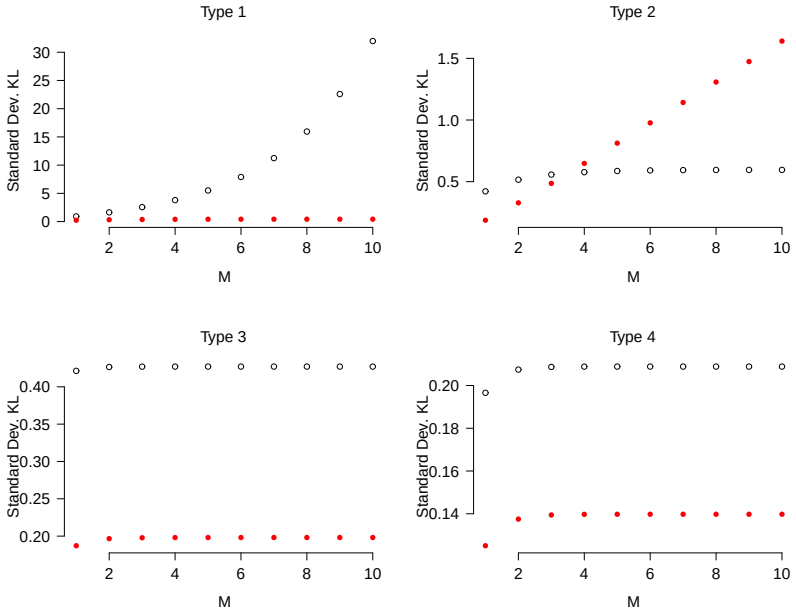

Fig. 3.3 Comparison between standard deviations of KL for different values of M. $\sqrt{\text{Var}\{\text{KL}(\pi_0||\pi)\}}$ (empty dots) and $\sqrt{\text{Var}\{\text{KL}(\pi||\pi_0)\}}$ (solid dots). Type 1 to 4 denote the different ρ functions given in (3.8).

specify three parameters: the truncation level M, the concentration parameter α, and the form of the function $\rho(m)$. This function is usually constructed in its continuous version, i.e. the precision function ρ satisfies the continuity property, for example ρ_3 and ρ_4 as given in (3.8). Lavine (1992) recommends $\rho_3(m) = m^2$ as a "sensible canonical choice", which as been adopted as the standard choice in the vast majority of applications (see for example Hanson and Johnson, 2002; Karabatsos, 2006; Muliere and Walker, 1997; Walker et al., 1999; Walker and Mallick, 1997). The choice of M has usually been done with a rule of thumb (e.g. Hanson, 2006), say $M = \log_2(n)$ with n being the data sample size. The author notes 'a law of diminishing returns' when increasing the truncation level from $M \to M+1$. Our study confirms this by plotting the diversity of draws as measured in KL against M, which suggest that a Pólya tree prior with as a low as $M = 4$ and $\rho(m) = 2^m$ can produce random draws that are equally far from the centring distribution as with a larger M (see two bottom panels in Figures 3.2 and 3.3). We argue that in order to make full use of the finite nature of the tree, the various combinations of the parameters should be considered. We have shown that it is possible to fully characterise draws from a truncated Pólya tree in terms of these three parameters. Careless choices can lead to a prior that overly concentrated around the baseline π_0. These results provide a principled method for choosing the parametrisation. In figure 3.4 we show that using $\rho_3(m)$ with a choice of $\delta = 1.01$ (empty dots) gives greater gains

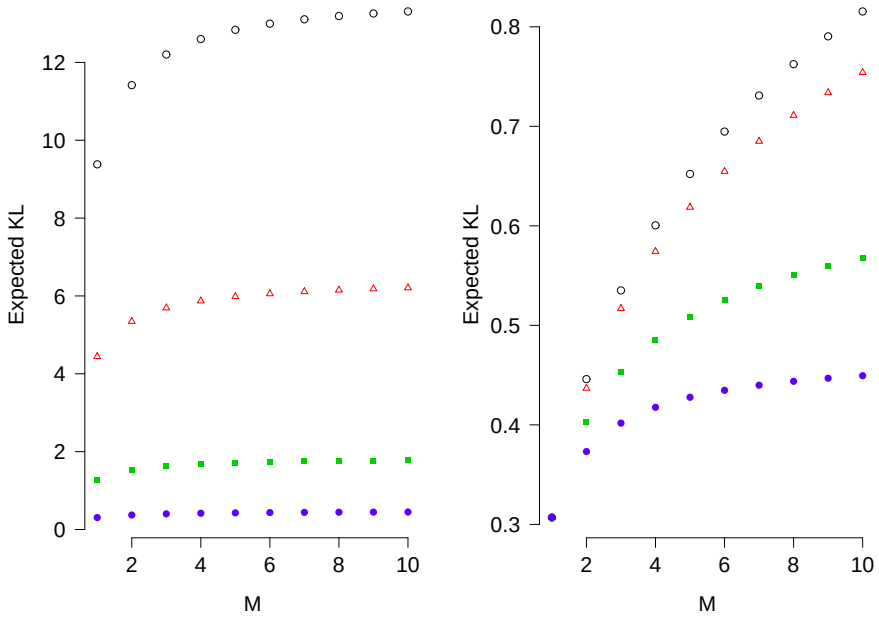

Fig. 3.4 Expected KL$(\pi_0||\pi)$ for different values of α (left panel) and δ (right panel) for the concentration function $\rho(m) = m^{\delta}$, as a function of m. Left: $\alpha = 0.05$ (empty dots); $\alpha = 0.1$ (triangles); $\alpha = 0.3$ (squares); $\alpha = 1$ (solid dots). Right: $\delta = 1.01$ (empty dots); $\delta = 1.1$ (triangles); $\delta = 1.5$ (squares); $\delta = 2$ (solid dots)

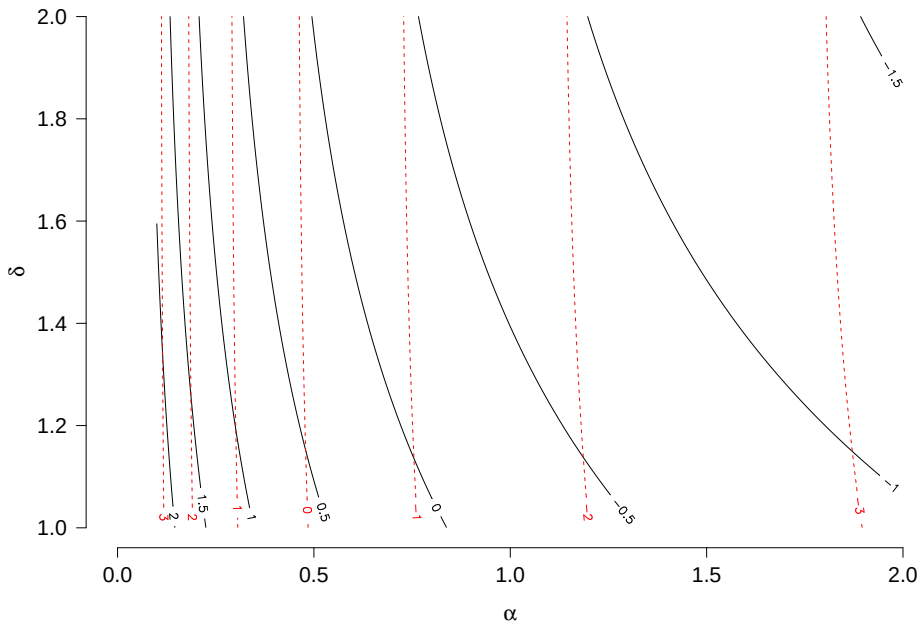

Fig. 3.5 Overlaid contour plots of the log-expected KL (black lines) and the log-variance of the KL (dashed red lines) as functions of the two parameters (α, δ) at regular intervals of 1/2 and 1, respectively, of draws from a Pólya tree processes with truncation level $M = 10$.

in expected KL as M is increased as compared to those obtained for the standard choice of $\delta = 2$ and decreasing the parameter α. The concentration around the baseline measure is highly sensitive to this choice of exponent, thus questioning the "sensible canonical choice" of $\delta = 2$ given by Lavine (1992). Figure 3.5 shows overlaid contour plots of the expected KL and its variance as a function of (α, δ) when taking $\rho(m) = m^{\delta}$ as concentration function. This highlights the role both parameters play in specifying the location of random draws from the process, and shows that changing the concentration parameter α has a greater effect than the exponent δ.

3.1.5 Retrospective stochastic reweighing

Returning to the context of chapter 2, section 2.3, we now consider π_0 as the empirical density obtained from n samples from a continuous density π_I. Having explored deterministic reweighing of these samples, targeting the least favourable distribution $\pi_a^{\sup}$, we can use the results from the previous section to explore stochastic reweighing of the samples θ_i.

Without loss of generality, suppose that $n = 2^k$, for some $k > 1$, and that $\theta_1, .., \theta_n \sim \pi_I$. As in equation 2.3, we define $\hat{\pi}_I := \sum_{i=1}^{m} 1/m \mathbb{1}_{\theta_i}(\theta)$, the Monte Carlo representation of π_I.

We define the following finite partition tree Π: $B_{11} = \Theta$; $\{B_{21}, \bar{B}_{22}\}$ is a partition of Θ such that $\theta_1, .., \theta_{2^{k-1}} \in B_{21}$ and $\theta_{2^{k-1}+1}, .., \theta_n \in B_{22}$. At level $m \leq k+1$ of the tree, B_{mj} contains 2^{k-m+1} samples exactly. At level $k+1$ we have that $\theta_j \in B_{(k+1)j}$ for $j \in 1..2^k$. Because of the finite representation $\hat{\pi}_I$, we do not define the partition tree any further. The expected loss taken with respect to $\hat{\pi}_I$ is:

$$\begin{aligned}\psi^{(a)}_{\hat{\pi}_I} &= \mathbb{E}_{\hat{\pi}_I}[L_a(\theta)] = \int_\Theta L_a(\theta)\hat{\pi}_I(\theta)\mathrm{d}\theta \\ &= \sum_{i=1}^{2^k} \int_{B_{(k+1)i}} L_a(\theta)\hat{\pi}_I(\theta)\mathrm{d}\theta \\ &= \sum_{i=1}^{2^k} 2^{-k} L_a(\theta_i)\end{aligned}$$

Because the finite Pólya tree is defined on the observed samples θ_i, we only need to sample branching probabilities up to level $k+1$ in order to have an approximation of the expected loss taken with respect to a random draw $\hat{\pi} \sim \mathrm{PT}(\alpha, \rho, \hat{\pi}_I)$.

Results from section 3.1.3 give us the exact expected KL divergence of a random draw $\hat{\pi}$, and the variance of this KL divergence. Analogous to the diagram given in figure 2.1, the Pólya tree allows for sampling roughly contained within an 'ε-doughnut' neighbourhood around $\hat{\pi}_I$. A graphical representation of this is shown in figure 3.6.

Thus we have a procedure for sampling estimates of expected loss with respect to random draws from this KL ε-*doughnut* neighbourhood centred at π_I:

- Sample $\theta_i \sim \pi_I$
- For $m = 1, .., k$ and $j = 1, .., 2^{m-1}$:
  - Sample $Y_{m,2j-1} \sim \mathrm{Beta}[\alpha\rho(m), \alpha\rho(m)]$
  - $Y_{m,2j} = 1 - Y_{m,2j-1}$
- Set weights $w_i = \Pi_{m=1}^k Y_{m,j^{(\theta_i)}}$
- Accept if $\sum_i w_i \log(n w_i) \in [C-\varepsilon, C+\varepsilon]$ (within the KL ε-doughnut). Reject otherwise.
  - The expected loss is given by: $\psi_a^{\hat{\pi}} = \sum_{i=1}^n w_i l_a(\theta_i)$

We illustrate this method with $n = 2^{10}$ draws from a standard normal distribution with expected KL divergence of 1. Figure 3.7 shows the resulting densities for the four concentration functions $\rho(m)$ given in (3.8). For $\rho_1(m) = \delta^{-m}, \rho_3(m) = m^\delta$ and $\rho_4(m) = \delta^m$, we set $\delta = 1.2$ ($\rho_3(m) = 1$ is not dependent on δ). The concentration parameter α is

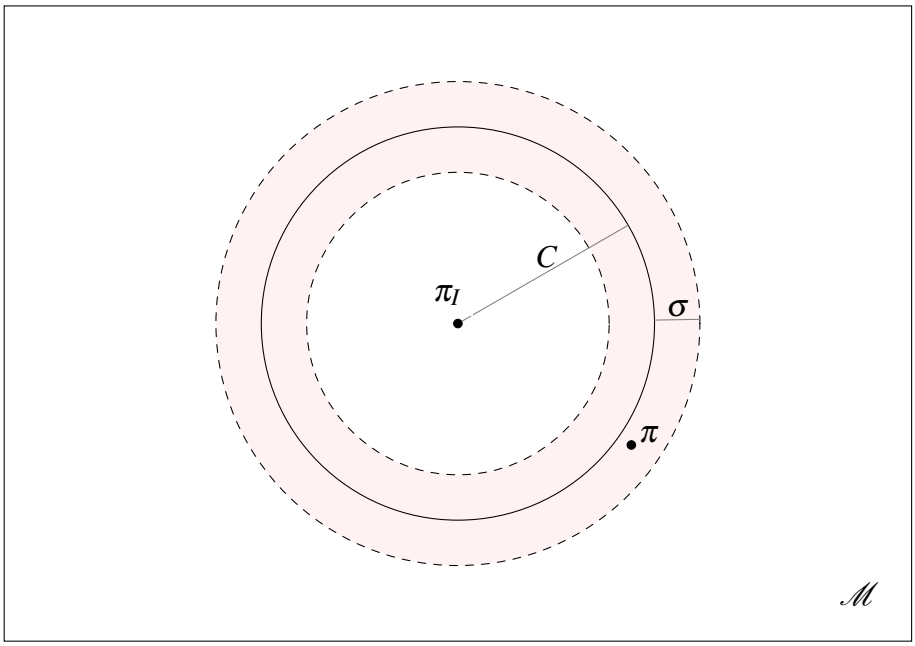

Fig. 3.6 Graphical illustration of the *doughnut KL neighbourhood* centred at π_I with an ε range. The radius C is the expected KL divergence of a random draw π from a $\text{PT}(\alpha, \rho, \pi_I)$, given in equation 3.6. $\mathscr{M}$ is the space of all measures on Θ.

then taken as the solution $\mathbb{E}[\text{KL}(\hat{\pi}_I||\hat{\pi})] = 1$ for $\hat{\pi} \sim \text{PT}_n(\alpha, \delta, \hat{\pi}_I)$. This gives values of $\alpha \approx 6.9, 2.5, 0.7, 1.1$, respectively. The plot shows all the draws (there is no rejection step).

We note this is different from the usual Pólya Tree construction. When truncated at a level M, this has a density proportional to the baseline density π_I (but reweighed by w_i in the i^{th} partition). In our case, the density is atomic - the plots in figure 3.7 show the estimated density from the reweighed atoms.

3.2 Bootstrap procedures

3.2.1 Divergence between discrete distributions

We consider the general setting where π_0 is a discrete measure with n atoms $\{\theta_1, ..., \theta_n\}$, i.e., the density is given by $\pi_0(\theta) = \sum_{i=1}^n p_i \delta_{\theta_i}(\theta)$, with $p_i > 0$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^n p_i = 1$. Let $\mathbf{w} = (w_1, \ldots, w_n)$ be random weights such that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$ almost surely. Let π be a random distribution defined as a reweighing of the atoms of π_0 with the random weights $\mathbf{w}$. In notation, $\pi(\theta) = \sum_{i=1}^n w_i \delta_{\theta_i}(\theta)$.

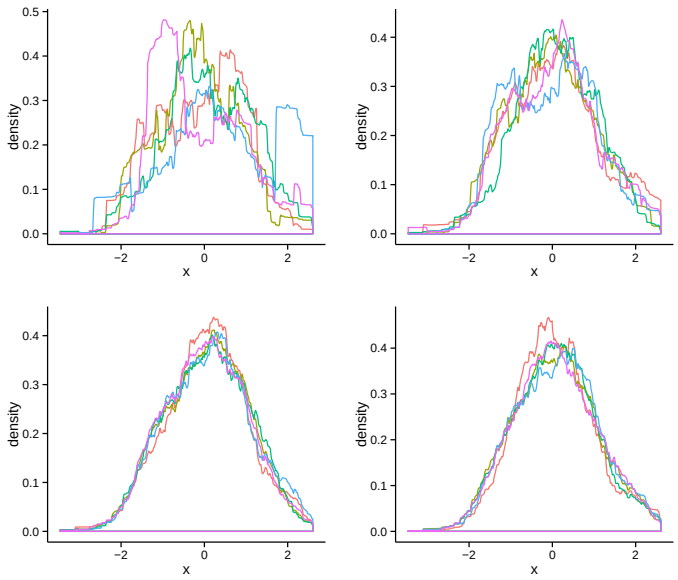

Fig. 3.7 Pólya tree draws with expected KL divergence of 1 from a standard normal distribution (no rejection step). Top left to bottom right: draws for each type of concentration function given in 3.8 with $\delta = 1.2$. The values of α are respectively: $\approx 6.9, 2.5, 0.7, 1.1$. With a rejection step, the acceptance rates are approximately 0.6, 0.4, 0.1 and 0.2 respectively (simulated by a 1000 draws).

The Kullback-Leibler divergence between π_0 and π does not depend on the atoms locations and is given by:

$$KL(\pi_0||\pi) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{w_i}\right), \tag{3.11}$$

and the reverse Kullback-Leibler has the form

$$KL(\pi||\pi_0) = \sum_{i=1}^{n} w_i \log\left(\frac{w_i}{p_i}\right). \tag{3.12}$$

The special case where $p_i = 1/n$ for $i = 1,..,n$ is of particular interest as it coincides with the setting where the distribution of interest π_0 can only be accessed via a 'bag of Monte Carlo samples' as in sections 2.3 and 3.1.5. In this case where π_0 is a uniform density, we first highlight an important property in the relationship between the divergences (3.11) and (3.12).

Proposition 3. *Consider the KL divergences* (3.11) *and* (3.12). *If* $p_i = 1/n$ *for* $i = 1,\ldots,n$, *then for any given re-weighing vector* $\mathbf{w}$ *taken from the simplex* $\mathscr{Q}_n := \{w : w_i \geq 0, \sum_{i=1}^{n} w_i =$

$1\}$ *we have that*

$$\mathrm{KL}(\pi_0||\pi) \geq \mathrm{KL}(\pi||\pi_0).$$

Proof. Let $h(\mathbf{w}) := \mathrm{KL}(\pi_0||\pi) - \mathrm{KL}(\pi||\pi_0)$. Using expressions (3.11) and (3.12), $h(w)$ becomes $h(\mathbf{w}) = -\sum(1/n + w_i)\log(w_i)$. We note that $h(\mathbf{w}) = 0$ at $\mathbf{w}^* = (1/n, \ldots, 1/n)$ and is infinite on all the simplex boundaries. Moreover, h is convex and by straightforward differentiation we see that $h''(\mathbf{w}^*)$ is positive. The result follows. □

This inequality between the two divergences provides a reason for choosing one of the two divergences, depending on whether it is preferable to have a smaller or larger neighbourhood, i.e. for a fixed radius C, (3.12) defines a smaller neighbourhood than (3.11). With this in mind, we now look at characterising both the frequentist and Bayesian bootstrap methods in terms of the Kullback-Leibler divergence, in order to use them for sampling in a neighbourhood of an approximating model.

3.2.2 Characterising the frequentist bootstrap

Using the notation from the previous section, we take $n\mathbf{w} \sim \mathrm{Mult}(n, \mathbf{p})$[3], a multinomial distribution with n trials and n categories with probability of success $\mathbf{p} = (p_1, \ldots, p_n)$. The random distribution π defined by $\mathbf{w}$ will have expectation π_0, where again π_0 is defined by the vector $\mathbf{p}$. Note that if $p_i = 1/n$ for $i = 1, \ldots, n$ this choice of distribution for the weights $\mathbf{w}$ coincides with the frequentist bootstrap (Efron, 1979) for which the atoms $\{\theta_i\}$ are replaced by i.i.d. random variables $\{X_i\}$.

We note that the KL divergence (3.11) will in general not be defined, as w_i can be zero. In fact, for large n and for $p_i = 1/n$ in the previous multinomial choice, approximately one third of the weights will be zero. However, $0\log 0$ is defined by convention as 0, so the reverse KL (3.12) is well defined.

Proposition 4. *The expected value of the Kullback-Leibler divergence between a "bootstrap" draw π, with $n\mathbf{w} \sim Mult(n, \mathbf{p})$, and its centring distribution π_0, defined in* (3.12)*, has the following upper bound:*

$$\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\} \leq \sum_{i=1}^{n} p_i \log\left(p_i + \frac{1-p_i}{n}\right) - H(\mathbf{p}) \tag{3.13}$$

where $H(\mathbf{p}) = \sum_{i=1}^{n} p_i \log p_i$, the entropy of the vector $\mathbf{p}$. For the special case when $p_i = 1/n$, we have $\mathbb{E}\{\mathrm{KL}(f||f_0)\} \leq \log(2 - 1/n) \leq \log 2$

[3] We use this notation to emphasise the fact that $\mathbf{w}$ represents a random probability mass function, but taking values on the set $\{0, 1/n, 2/n, .., 1\}$. A factor of n is needed for the vector to be distributed according to a multinomial distribution.

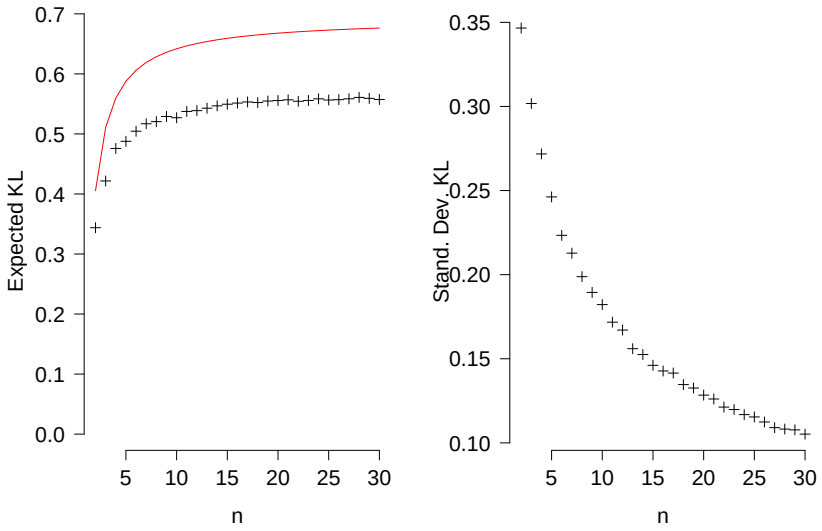

Fig. 3.8 Estimated expected value and standard deviation of KL divergence (3.12) as a function of n for the frequentist bootstrap. Left: estimated mean (black crosses) and upper bound given in (3.13), red line. Right: estimated standard deviation.

Proof.

$$\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\} = \sum_{i=1}^{n} \mathbb{E}\{w_i \log w_i\} - \sum_{i=1}^{n} \mathbb{E}\{w_i\} \log p_i.$$

Working on the individual expected values,

$$\mathbb{E}\{w_i \log w_i\} = \sum_{k=1}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k} \left(\frac{k}{n}\right) \log\left(\frac{k}{n}\right).$$

From which we get $\mathbb{E}(w_i \log w_i) = (1/n)\mathbb{E}\{\log((v_i+1)/n)\}$, with $v_i \sim \mathrm{Bin}(n-1, 1/n)$. Using Jensen's inequality we get $\mathbb{E}\{w_i \log w_i\} \leq p_i \log(p_i + (1-p_i)/n)$. Substituting this into the original sum and using $\mathbb{E}\{w_i\} = p_i$ gives the result. □

Although we do not have an analytical expression for the expected value and the variance of the KL divergence (3.12) as a function of n, we can still study its behaviour via simulation. In Figure 3.8 we show estimates of these quantities based on a sample of size 1000. We use $p_i = 1/n$ and values of n ranging from 2 to 30. For the expected value we also include the upper bound obtained in Proposition 4 as a dashed line in the left panel. From the figure we observe that the expected value of the KL tends to converge to a finite number when the number of atoms n increases (upper bounded by $\log 2$). On the other hand, the variance decreases to zero as n grows.

Thus the frequentist bootstrap provides a simple nonparametric procedure for sampling random measures within a KL neighbourhood (again a 'doughnut' neighbourhood as given in figure 3.6). We now consider the Bayesian analogue ad derive similar results.

3.2.3 Generalised Bayesian bootstrap

An alternative way for the random π to be centred around π_0 is by sampling weights $\mathbf{w}$ from a Dirichlet distribution with parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ such that $\beta_i = \alpha_n p_i$, $i = 1, \ldots, n$, with $\alpha_n > 0$ a parameter changing as a function of the number of atoms. This is denoted $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. It is straightforward to prove that $\mathbb{E}(\pi) = \pi_0$, and that the form of α_n parametrises the precision, in the same way as the Pólya tree case. If we take $\alpha_n = n$, $p_i = 1/n$ and replace the atoms $\{\theta_i\}$ by i.i.d. random variables $\{X_i\}$, we obtain the original Bayesian bootstrap proposed by Rubin (1981)[4]. Sampling from a Dirichlet with parameter vector $\alpha_n \mathbf{p}$ gives a generalised version of this bootstrap procedure.

KL divergence properties

In this setting, both $\text{KL}(\pi_0 || \pi)$ and reverse $\text{KL}(\pi || \pi_0)$, given in (3.11) and (3.12) respectively, are well defined since $w_i \neq 0$ almost surely. Their expected values and variances can be obtained in closed form as functions of α_n and $\mathbf{p}$.

Proposition 5. *Let π be a "generalised Bayesian bootstrap" draw around π_0 with weights $\mathbf{w} \sim Dir(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in* (3.11) *has mean and variance:*

$$\mathbb{E}\{\text{KL}(\pi_0 || \pi)\} = H(\mathbf{p}) - \sum_{i=1}^{n} p_i \{\psi_0(\alpha_n p_i) - \psi_0(\alpha_n)\}$$

$$Var\{\text{KL}(\pi_0 || \pi)\} = \sum_{i=1}^{n} p_i^2 \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$$

where ψ_0 and ψ_1 are the digamma and trigamma functions.

Proof. This result follows from $\mathbb{E}(\log w_i) = \psi_0(\alpha_n p_i) - \psi_0(\alpha_n)$ and linearity of expectation. The variance follows from $\text{Var}(\log w_i) = \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$, and $\text{Cov}(\log w_i, \log w_j) = \psi_1(\alpha_n p_i)\delta_{ij} - \psi_1(\alpha_n)$, where δ_{ij} is the Kronecker delta function taking value 1 when $i = j$ and 0 otherwise. □

The limiting behaviour of this expected KL and its variance, as n tends to infinity, can more easily be studied for the special case of $p_i = 1/n$, $i = 1, \ldots, n$. When $\alpha_n = \alpha$,

[4] It is interesting to note that in his original work he only considers this special case.

i.e. constant, they both diverge to infinity. In the limit, this is a well known construction of the Dirichlet process, when the atoms are sampled i.i.d. from a baseline measure π_0. However, if we make α_n grow linearly with n, say $\alpha_n = \alpha n$, then $\lim_{n\to\infty} \mathbb{E}\{\mathrm{KL}(\pi_0||\pi)\} = \log(\alpha) - \psi_0(\alpha)$ and $\lim_{n\to\infty} Var\{\mathrm{KL}(\pi_0||\pi)\} = 0$. These values are obtained by noting that $\psi_0(n)$ behaves like $\log(n)$ for large n. Finally, if we increase the rate at which α_n grows with n, say $\alpha_n = \alpha n^2$, both mean and variance of the KL converge to zero as $n \to \infty$.

Proposition 6. *Let π be a "generalised Bayesian bootstrap" draw around π_0 with weights $\mathbf{w} \sim Dir(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in* (3.12) *has mean:*

$$\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\} = \sum_{i=1}^{n} p_i \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\} - H(\mathbf{p}) \tag{3.14}$$

where $H(\mathbf{p}) := \sum_{i=1}^{n} p_i \log p_i$ the entropy of the vector $\mathbf{p}$, and the variance given by

$$\begin{aligned} Var\left(\mathrm{KL}(\pi||\pi_0)\right) = &\sum_{i=1}^{n} \left\{ Var(w_i \log w_i) + (\log p_i)^2 Var(w_i) - 2(\log p_i) Cov(w_i \log w_i, w_i) \right\} \\ +2 &\sum_{i<j} \left\{ Cov(w_i \log w_i, w_j \log w_j) + (\log p_i)(\log p_j) Cov(w_i, w_j) - 2(\log p_j) Cov(w_i \log w_i, w_j) \right\} \end{aligned} \tag{3.15}$$

where each of the elements are given in the footnote[5].

Proof. Note that each $w_i \sim \mathrm{Beta}\{\alpha_n p_i, \alpha_n(1-p_i)\}$ and thus we have that $\mathbb{E}(w_i \log w_i) = p_i\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}$. Using linearity of expectation and substituting this expression we obtain the mean. Using properties of the variance and covariance of sums we get the second part of the result. □

Similarly to the previous case, if we take $p_i = 1/n$ and $\alpha_n = \alpha n$, when $n \to \infty$ then $\mathbb{E}\{\mathrm{KL}(\pi||\pi_0)\} \to \psi_0(\alpha + 1) - \log(\alpha)$. It is possible to show analytically that each term in (3.15) goes to zero as $n \to \infty$, but this can also be seen using the relation between the two KL's given in Proposition **??**, and noting that the variance involves a monotonic transformation, hence we have that $\mathrm{Var}\{\mathrm{KL}(\pi_0||\pi)\} \geq \mathrm{Var}\{\mathrm{KL}(\pi||\pi_0)\}$. From the previous result it follows that $\lim_{n\to\infty} \mathrm{Var}\{\mathrm{KL}(\pi||\pi_0)\} = 0$ for these choices of p_i and α_n. In Figure 3.9 we compare

[5] $\mathrm{Var}(w_i) = p_i(1-p_i)/(\alpha_n+1)$, $\mathrm{Cov}(w_i, w_j) = -p_i p_j/(\alpha_n+1)$, $\mathrm{Var}(w_i \log w_i) = p_i(\alpha_n p_i+1)/(\alpha_n+1)\{\psi_1(\alpha_n p_i+2) - \psi_1(\alpha_n+2) + [\psi_0(\alpha_n p_i+2) - \psi_0(\alpha_n+2)]^2\} - p_i^2\{\psi_0(\alpha_n p_i+1) - \psi_0(\alpha_n+1)\}^2$, $\mathrm{Cov}(w_i \log w_i, w_i) = p_i(\alpha_n p_i+1)/(\alpha_n+1)\{\psi_0(\alpha_n p_i+2) - \psi_0(\alpha_n+2)\} - p_i^2\{\psi_0(\alpha_n p_i+1) - \psi_0(\alpha_n+1)\}$, $\mathrm{Cov}(w_i \log w_i, w_j) = p_i p_j\{-\psi_0(\alpha_n p_i+1)/(\alpha_n+1) + \psi_0(\alpha_n+1) - \alpha_n \psi_0(\alpha_n+2)/(\alpha_n+1)\}$, $\mathrm{Cov}(w_i \log w_i, w_j \log w_j) = \alpha_n p_i p_j/(\alpha_n+1)[\{\psi_0(\alpha_n p_i+1) - \psi_0(\alpha_n+2)\}\{\psi_0(\alpha_n p_j+1) - \psi_0(\alpha_n+2)\} - \psi_1(\alpha_n+2)] - p_i p_j\{\psi_0(\alpha_n p_i+1) - \psi_0(\alpha_n+1)\}\{\psi_0(\alpha_n p_j+1) - \psi_0(\alpha_n+1)\}$.

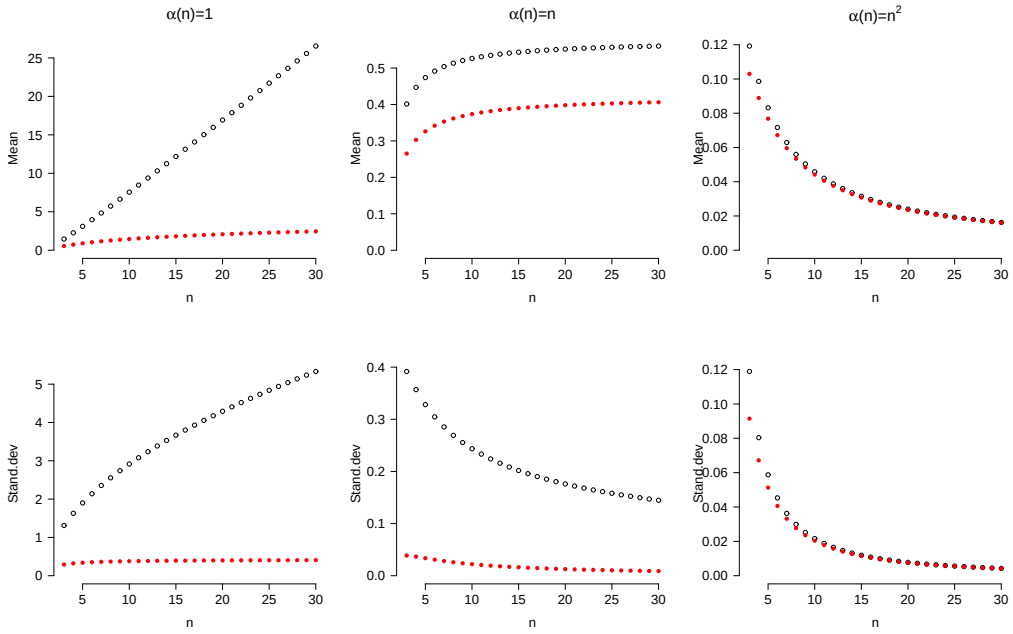

Fig. 3.9 Expected value (top row) and standard deviation (bottom row) of $\mathrm{KL}(\pi_I||\pi)$ (black empty dots) and $\mathrm{KL}(\pi||\pi_I)$ (red solid dots) as a function of the number of atoms n. The columns from left to right correspond to values of α_n of $1, n$ and n^2.

the expected value and variance of both KL and reverse KL for $p_i = 1/n$ and different values of α_n as a function of n. The first column corresponds to $\alpha_n = 1$, the second column to $\alpha_n = n$ and the third to $\alpha_n = n^2$, which induce high, moderate and small variance in the $\mathbf{w}$ respectively. As shown in proposition 3, the expected value and variance of $\mathrm{KL}(\pi_0||\pi)$ are larger than those of $\mathrm{KL}(\pi||\pi_0)$. Their limiting behaviours can also be assessed from the graphs.

3.2.4 KL divergence and weak convergence

In our construction, atoms θ_i are sampled i.i.d. from the baseline measure π_I and then stochastically reweighed by the vector $\mathbf{w} \sim \mathrm{Dir}(\alpha_\mathbf{n}\mathbf{p})$. Therefore π represents a random process centred around the empirical distribution of π_I. Ishwaran and Zarepour (2002) considered exactly this random probability process and derived results for the limiting behaviour for a variety of choices of α_n (see Theorem 3, page 948). When $\alpha_n = \alpha$, then π is distributed according to a Dirichlet process $\mathrm{DP}(\alpha, \pi_I)$, in the limit as $n \to \infty$. If $\alpha_n = \alpha n$, then we have almost sure weak convergence of π to π_I, as $n \to \infty$. For the third case considered here, $\alpha_n = \alpha n^2$, π converges in probability to π_I, as $n \to \infty$.

The case where $\alpha_n = \alpha n$ is of particular interest. Although we have weak convergence of $\pi \to \pi_I$, the random distribution does not converge in KL divergence. In other words, although functionals of π tend to the functionals of π_I, the KL divergence between the two densities remains finite. This becomes apparent when considering the random quantity nw_i, which comes into the equation (3.11), whose variance becomes asymptotically $1/\alpha$, as $n \to \infty$. To the unwary observer this comes as a surprising result. It shows to what extent convergence in Kullback-Leibler is a strong statement.

3.3 Dirichlet process

3.3.1 Characterising perturbations to loss distributions

In the final section of this chapter, we focus on the Dirichlet process, a special case of the Pólya tree process when the concentration function $\rho(m) = 1/2^m$. This is by far one of the most widely used nonparametric processes, and we consider a different construction, which is a particular setting of the generalised Bayesian bootstrap.

As stressed at the start of this chapter, it is more natural from a Bayesian standpoint to characterise the variation in expected loss arising over all models in some neighbourhood Γ, rather than performing minimax optimisation within the neighbourhood. In order to quantify this uncertainty and take expectations over distributions in the neighbourhood of π_I, we require a probability distribution on a set of probability measures centred on π_I. The Pólya tree process almost gives this - sampling within a 'doughnut' shaped neighbourhood rather than an KL ball. However, in a decision-theoretic context, the statistician is interested in estimates of functionals of the distributions $\pi \in \Gamma$. In particular the functionals $\psi_a : \pi \to \mathbb{E}_\pi[L_a(\theta)]$ for $a \in \mathscr{A}$ (expected loss). As shown in the apparent paradox of section 3.2.4, two sequences of distributions π_n, π_n^* can be infinitely divergent in Kullback-Leibler, or can remain at a finite distance in total variation metric, but weakly converge, i.e. their functionals converge. Thus, if we set a nonparametric distribution Π over measures π, that is centred at π_I: instead of studying the 'distance' between draws $\pi \sim \Pi$ and the reference distribution π_I, we can study the distance between the induced distributions $F_{a,\pi}(z)$ and $F_{a,\pi_I}(z)$, the (cumulative) distributions of loss for action a. A suitable candidate distribution Π should have wide support (to overcome the possible misspecification) and it should be possible to characterise the distance of the induced distributions $F_{a,\pi}$. The Dirichlet Process (DP) allows for exactly such a construction.

Definition 3. Dirichlet Process: *Given a state space $\mathscr{X}$ we say that a random measure P is a Dirichlet Process on $\mathscr{X}$, $P \sim \mathrm{DP}(\alpha, P_0)$, with concentration parameter α and baseline*

measure P_0 if for every finite measurable partition $\{B_1,\ldots,B_k\}$ of $\mathscr{X}$, the joint distribution of $\{P(B_1),\ldots,P(B_k)\}$ is a k-dimensional Dirichlet distribution $\mathrm{Dir_k}\{\alpha \mathrm{P_0(B_1)},\ldots,\alpha \mathrm{P_0(B_k)}\}$.

Using this definition we can then sample from distributions in the neighbourhood of π_I according to $\pi \sim \mathrm{DP}(\alpha,\pi_I)$, for some $\alpha > 0$. In practice we can consider a draw from the DP via a constructive definition,

$$\begin{aligned} \{\theta_i\}_{i=1}^m &\sim \pi_I \\ \underline{w} &\sim \mathrm{Dir}_m(\alpha/m,\ldots,\alpha/m), \\ \hat{\pi}(\theta) &:= \sum_{i=1}^m w_i \delta_{\theta_i}(\theta) \end{aligned} \tag{3.16}$$

where the θ_i's are i.i.d. from π_I and independent of the Dirichlet weights. As $m \to \infty$, $\tilde{\pi}$ tends to a draw $\pi \sim \mathrm{DP}(\alpha,\pi_I)$. This construction fits well with the Monte Carlo context, where π_I is represented by a bag of samples $\{\theta_i\}_{i=1}^m$. If we draw multiple vectors $\underline{w}^{(1)},..,\underline{w}^{(k)} \sim \mathrm{Dir}_m$, then in the limit $m \to \infty$, each corresponds to an independent draw from the $\mathrm{DP}(\alpha,\pi_I)$, conditional on the atoms θ_i. Ideally, one would want to resample a set $\{\theta_i\}_{i=1}^m$ at each step. But this would not be feasible in practice and would defeat the purpose of constructing an *ex-post* methodology for analysing sensitivity. Therefore, this construction of the Dirichlet Process is more adapted than say the stick-breaking representation. This also comes as a special case of the generalised Bayesian bootstrap presented in section 3.2.3. The expected KL of a draw is infinite, thus characterising draws in terms of KL is not appropriate. However we look at characterising the draws in terms of the loss distribution.

For an action a, the expected loss under the re-weighed draw $\hat{\pi}$ is given by:

$$\psi_a^{\tilde{\pi}} = \sum_i w_i L_a(\theta_i) \tag{3.17}$$

and the loss distribution by:

$$F_{a,\hat{\pi}}(z) = \sum_i w_i \mathbb{1}_{z \le L_a(\theta_i)}(z)$$

In what follows, without loss of generality, we fix a and consider the θ_i to be ordered by loss, i.e. $L_a(\theta_1) \le ... \le L_a(\theta_m)$. Let $v_i = \sum_{j=1}^i w_i$, the cumulative summed weights, and $x_i := i/m$ for $i = 1,..,m$. We also consider that the loss function $L(a,\theta)$ has undergone the following linear transformation (which does not alter the ranking of actions under expected loss):

$$L(a,\theta) \to \frac{L(a,\theta) - \min_{a,\theta} L(a,\theta)}{\max_{a,\theta} L(a,\theta) - \min_{a,\theta} L(a,\theta)} \tag{3.18}$$

This means each loss cdf takes values between [0,1]. We can study the L_1 distance between the empirical distribution[6] $F_{a,\tilde{\pi}_I}$ and the reweighed version $F_{a,\tilde{\pi}}$ which is given by:

$$\sum_{i=1}^{m} |v_i - x_i| \cdot [L_a(\theta_i) - L_a(\theta_{i-1})]$$

For a fixed sample $\{\theta_i\}_{i=1}^m$, the increments $L_a(\theta_i) - L_a(\theta_{i-1})$ are also fixed, and it is possible to compute the expected difference $|v_i - x_i|$ by noting that $v_i \sim \text{Beta}(x_i\alpha, (1-x_i)\alpha)$. This is given by:

$$\mathbb{E}_v\{|v_i - x_i|\} = \frac{2}{\alpha} \frac{[x_i^{x_i}(1-x_i)^{(1-x_i)}]^\alpha}{\text{Beta}(x_i\alpha, (1-x_i)\alpha)}$$

As a consequence of the linear transformation given in (3.18), this L_1 difference is bounded by 1/2. $\mathbb{E}_{\underline{w}}\{|F_{a,\hat{\pi}} - F_{a,\hat{\pi}_I}|\}$ is dependent on the concentration parameter α which controls how close the draws $F_{a,\tilde{\pi}}$ are from the reference loss distribution; increasing α shrinks the L_1 distance. However, it is important the note that this distance will also be dependent on the form of the loss function, i.e the increments $L_a(\theta_i) - L_a(\theta_{i-1})$.

3.3.2 Probability of optimality

From properties of the Dirichlet Process, we know that $\mathbb{E}_\Pi[L_a(\theta)] = \mathbb{E}_{\pi_I}[L_a(\theta)]$, where Π is the nonparametric measure defined in equation (3.16). Thus if an action a is optimal under the criterion of posterior expected loss (taken with respect to π_I), it will remain optimal under expected loss taken with respect to Π. Instead of looking at expected loss we consider the probability that a particular action will be optimal when drawing a random $\pi \sim \text{DP}(\pi_I, \alpha)$ (and computing expected loss with respect to this random π). That is to say, each random draw π will induce a distribution of loss $F_{a,\pi}$ for action a. The probability that a is optimal will depend on the concentration parameter α. As the concentration parameter $\alpha \to \infty$, the random loss distribution $F_{a,\pi}$ tends to F_{a,π_I} in probability under the L_1 norm, thus giving back the optimality mapping induced by π_I. This gives rise to a diagnostic graph, where the probability of optimality of each action is plotted against the parameter α. The probability of optimality is non-analytical in the general case, and dependent on the form of the loss function $L(a, \theta)$. However, given a Monte Carlo representation of π_I and thus a matrix of loss values (number of samples θ_i times number of actions) it is easy to approximate via successive draws $\underline{w} \sim \text{Dir}(\alpha/\text{n}, .., \alpha/\text{n})$ and using the construction given in (3.17).

[6] empirical in the sense that it corresponds to π_I through i.i.d. sampling.

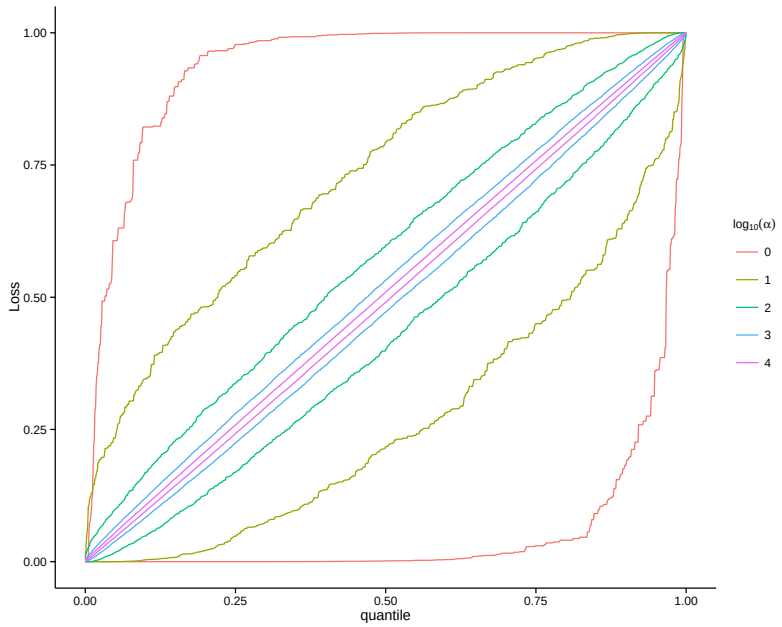

Fig. 3.10 95% confidence intervals around the reference action a^* for different values of the concentration parameter α. From widest to tightest: $\log_{10}(\alpha)$ values are 0,1,2,3,4.

3.3.3 Calibration of the Dirichlet Process extension

Contrary to the methodology proposed in chapter 2, using a Dirichlet Process as a nonparametric model extension to test robustness gives a framework which is not action specific. However, it also relies on a free parameter α which needs to be calibrated in a principled manner. In order to do this, we define a reference action a^*, such that the loss distribution is given by $F_{a^*,\pi}(z) = z$, for $z \in [0,1]$[7].

With the construction given in (3.17), if we draw weights $\underline{w} \sim \text{Dir}_m(\alpha/m, \ldots, \alpha/m)$ and use uniform loss intervals, i.e. reweighing the distribution $F_{a^*,\pi}(z)$, it is then possible to compute 95% confidence intervals for draws from a Dirichlet Process for a specific value of the parameter α. Figure 3.10 shows an approximation of these confidence intervals for a series of values α whose $\log_{10}$ values are integers from 0 to 4. It is clear that low values of α (between 0 and 10 for example) imply a very low trust in the model, as the draws can vary hugely from the reference distribution. However the statistician might want the decisions to be robust to higher values of α where the draws are much tighter. In chapter 5 we illustrate the use of this method.

[7] After the linear transformation given in (3.18).

3.4 Summary

A nonparametric model extension for assessing the robustness of an approximating model π_I is a natural choice in a Bayesian framework. Instead of only considering one distribution within a certain neighbourhood, it averages out over all 'close' distributions, where in the case of the Pólya tree and the bootstrap this is KL divergence, and for the Dirichlet process the L_1 distance defined over loss distributions. Concerning the distinction made by Kadane and Srinivasan (1994) between decision robustness and loss robustness, this framework is helpful for assessing decision robustness, as illustrated by the optimality plots of section 3.3.2. The next chapter focusses on assessing robustness via graphical diagnostics, some of those motivation stem from the material in this chapter and chapter 2.

Chapter 4

Qualitative diagnostic methods

This chapter explores ideas for assessing the stability of a decision-system in practice, via diagnostic plots and summary statistics, based on the theoretical concepts presented in Chapters 2 and 3. I want to stress the importance of graphical diagnostic plots to visualise a (Bayesian) decision-system. This allows the user to understand possible weaknesses by exploring the relationship between the loss function $L(a, \theta)$ and the posterior distribution or approximating reference model $\pi_I(\theta)$. These methods for assessing model sensitivity are an essential part of statistical modelling and are recommended as Many of the diagnostic plots are inspired by ideas from the mathematical finance and actuarial literature. I illustrate these plots and summary statistics with a fictional application from the medical decision making literature.

4.1 Motivation

Every good statistical analysis must commence with a graphical exploration of the data, however complicated the underlying patterns. The reasons for visualising the data range from simple outlier detection and sanity checks, to a pragmatic Bayesian perspective where the data provide information regarding plausible prior beliefs. Thus data visualisation is recommended as a primary tool for any statistical investigation. Despite the importance of graphical statistics, there are few if any established tools for the investigation of decision stability, in contrast to the multitude of methods for investigating model discrepancy and misspecification (Belsley et al., 2005; Gelman, 2007; Kerman et al., 2008). In the same way that diagnostic tools, such as trace plots for MCMC chains, allow the user to verify the inner workings of a model or algorithm, I think that analogous tools would be useful in the context of decision-theoretic problems. Indeed, if the application at hand is suitable for a decision-theoretic approach, then it is essential to understand the relationship between the

model π_I, which characterises the uncertainty, and the loss function L, which characterises the objectives of the modelling exercise.

Because of meta-uncertainty, i.e. uncertainty regarding the specification of the model and the loss, is is important to understand how misspecification in one effects the contribution of the other. Visual diagnostic plots can play a key role in investigating robustness properties with regards to misspecification in this context. Visual representations are especially suited when the model π_I is represented by a finite collection of Monte Carlo samples θ_i. Throughout this chapter it is considered that the statistician has access to π_I in the form of m samples $\theta_i \sim \pi_I$. The diagnostic plots and summary statistics that I present can be given qualitative interpretations and are intended for comparative purposes. That is to say, when the user has narrowed down the action set to a small selection, these plots can highlight different robustness properties of each action/alternative.

The chapter is structured as follows. Firstly, the diagnostic plots that display the relationship between the model and the loss function. These are inspired by tools used in financial mathematics and econometrics. The notion of 'coherence' is discussed in relation to these diagnostic tools and its role in Bayesian decision theory. Secondly, plots motivated by the formal framework of Chapters 2 and 3. All these plots and summary statistics are illustrated throughout with a synthetic example taken from the medical decision making literature.

4.1.1 Illustrative example

Consider a certain infectious disease for which there exists treatment medication and a new vaccine. One is interested in knowing whether the vaccination should be publicly funded, or whether the status-quo should remain in place, whereby patients visit their doctor and are prescribed over-the-counter drugs (OTC). This is a standard setting for decisions made by institutions such as NICE[1] in the UK, for example. I use a fictitious example of such a decision process taken from Baio and Dawid (2011), itself inspired by a real model given in Turner et al. (2006). The goal is to compare the two available actions: population wide vaccination or status quo. The modelling must take into account the uncertainty with regards to the efficacy of the vaccine and its coverage were it to be implemented. With regards the status quo action, the modelling has to consider the number of visits to the GP[2], complications from the drugs, possibly leading to extra visits, hospitalisation and even death. Each of these outcomes has either a monetary cost (visit to the GP for example) and/or a loss (negative utility) measured in Quality Adjusted Life Years, known as QALYs, (Loomes and McKenzie, 1989; Zeckhauser and Shepard, 1976). Therefore, in order to assign a loss value

[1] National Institute for Clinical Excellence.
[2] General Practitioner (doctor).

to each action, it is necessary to choose a conversion rate k, known as the 'willingness to pay' parameter, exchanging QALYs into pounds. Thus the loss of an action a is calculated as:

$$L_a(\theta) = \text{Cost}(a,\theta) - k * \text{QALY}(a,\theta) \tag{4.1}$$

Most of the decision literature focusses on the sensitivity of the decision system with respect to the specification of k. The R package *BCEA* (Bayesian Cost-Effectiveness Analysis) developed by Gianluca Baio has a statistical model for the *vaccination* vs *status quo* application. This example is given in detail in Baio and Dawid (2011). The *BCEA* package illustrates how to perform a sensitivity analysis on the parameter k.

I have used their model to illustrate the methods as it is good example of a sufficiently complex model (28 parameters) with a discrete set of actions (*vaccination* and *status quo*) and a well defined loss function, of the form given in equation 4.1.

The main components of the decision-system are the following:

- Parameters governing clinical outcomes in a population of N individuals: influenza infection rate; vaccination coverage; GP visits; minor/major complications; hospitalization; death; side effects of vaccine. For some of these outcomes indexed by j, there will be a baseline rate β_j and a proportional reduction due to vaccination ρ_j in the people who received the vaccination. All the binary outcome variables are modelled with binomials with informative hyperparameters.

- The number ω_j of QALYs lost for each outcome j. These are modelled using lognormal distributions, again with expert knowledge to set hyperparameters.

- The fixed costs (in pounds) for certain outcomes, i.e. cost of medication, GP visits, vaccination etc.

The exact form of the informative prior distributions can be found in table 1 of Baio and Dawid (2011). MCMC is used to draw samples from the model (there is no data, so this is sampling directly from the prior), and all the relevant details can be found in the package documentation, such as the exact settings of the cost function and hyperparameters. I ran the model given in *BCEA* using the default settings. Note that all the graphs were produced using my package *decisionSensitivityR* and the relevant code can be found in its documentation.

4.2 Evaluating 'risk'

4.2.1 Motivation

The series of plots and summary statistics presented in this section explores the relationship between the loss function $L(a,\theta)$ and the posterior distribution $\pi_I(\theta)$. Bayesian decision theory relies on the correct specification of these two objects whose construction is independent from one another. As noted by Jaynes (2003, Chapter 13, page 1320), it is tempting to think that maybe one only needs the product of the two, thus defining a new decision theory framework. However he rejects this for the following reasons:

> (1) priors and loss functions have very different - almost opposite - roles to play, both in the mathematical theory and in 'real life', and (2) the theory of inference involving priors is more fundamental than that of loss functions.

In Bayesian decision theory these two elements are constructed independently of one another. However, their combination defines the ranking of the actions $a \in \mathscr{A}$ and thus the Savage action, the optimal a^*. If there is a concern of misspecification in π_I, then the relationship between the model and loss is fundamental to knowing whether this misspecification matters. For example, would misspecification in the 'tails' of π_I be an issue? This is to say, would small misspecifications in areas of low probability have a high impact on the overall expected loss of an action?

Such questions motivate the study of the distribution of loss, which is a standard approach in the financial and actuarial literature where the loss distribution of a particular action is known as the risk of a position. In this chapter this is denoted Z_a, and via graphical tools, I attempt to show how it is possible to make qualitative statements regarding the robustness of available actions. The tools are similar to those used in finance and actuarial science and are particularly useful in the setting of Bayesian decision theory when the model π_I is not trusted.

4.2.2 Value-at-Risk

A primary tool for assessing the sensitivity with respect to misspecification of a functional of interest, is the distribution of loss, where Z_a denotes the random loss variable under $\pi_I(\theta)$:

$$F_{Z_a}(z) = Pr(Z_a \leq z) = \int_{\theta\in\Theta} I[L_a(\theta) \leq z]\pi_I(d\theta),$$

where $I[\cdot]$ is the identity function. I use notation $f_a(z) = F_{Z_a}(dz)$ to denote the corresponding density function and $F_{Z_a}^{-1}(q)$, for $q \in [0,1]$, is the inverse cumulative distribution or quantile

function. For a given $q \in [0,1]$ it is possible to characterise the value of an action a by its quantile loss or *Value at Risk*: $F_{Z_a}^{-1}(1-q)$ (VaR; terminology used in finance). VaR is a controversial diagnostic tool in finance because of non-coherence. The relevance of this concept is discussed later on in section 4.2.5.

Rostek (2010) developed an axiomatic framework in which decision-makers are uniquely characterised by a quantile $1-q$ and rational behaviour (the choice of optimal action) is defined as choosing $\hat{a} := \arg\inf_{a \in \mathscr{A}} F_{Z_a}^{-1}(1-q)$. Note that this incorporates the minimax rule by taking $q = 0$. The author argues that quantile maximisation is attractive to practitioners as its key characteristic is robustness, specifically to misspecification in the tails of the loss distribution. The values at particular quantiles provide informative summary statistics, which are alternatives to the mean value of loss. The work by Rostek justifies their use in a decision theoretic context. However, single quantiles discard much information contained in $[\pi_I(\theta), L_a(\theta)]$, plotting the whole function allows for immediate visualisation of the effect of the tail event on the estimate of expected loss. Indeed, it is easy to approximate the VaR plot with a 'bag of samples' $\{\theta_i\}_{i=1}^m \sim \pi_I(\theta)$ using the following method:

1. Sort the realised loss values, $z_i^{(a)} = L_a(\theta_i)$, from highest to lowest $\{z_{\upsilon_a(1)}^{(a)} \geq z_{\upsilon_a(2)}^{(a)} \geq \ldots \geq z_{\upsilon(m)}^{(a)}\}$, where $\upsilon_a(\cdot)$ defines the sort mapping.

2. For $q \in [0,1]$, approximate $F_{Z_a}^{-1}(1-q)$ by linear interpolation of the points $\left(x = k/m, y = z_{\upsilon_a(k)}^{(a)}\right)$.

From the VaR plot, it is immediately possible to observe whether there are crossing points between the actions. This allows the user to see which action is optimal for higher/lower quantiles of loss. This can be used in the context of the work done by Rostek.

In general I believe that the inverse loss distribution is a better diagnostic tool than the loss distribution as it puts the loss on the y-axis. This focusses the user's attention on the values of loss in the upper and lower tails rather than the quantiles themselves. The shape of the curve can be used as a qualitative tool to assess stability. For example a steep curve near $q = 0$ or $q = 1$ implies extreme values (relative to other loss values) in the tails that have a large effect on the expected loss estimate. If these tails are slightly misspecified, this could lead to large deviations in the estimate of expected loss.

To illustrate these points, I point to the reader to figures 4.1 and 4.2. The first plots the loss distribution of the two actions in the medical decision making example (section 4.1.1) when the 'willingness to pay' parameter (from equation 4.1) is set at $k = 15000$. The *status quo* is given in grey and the *vaccination* alternative in yellow. One can see that the *status quo* is preferable to *vaccination* with a slightly lower value of expected loss. However, its distribution of loss has greater variance than the *vaccination* alternative, which would suggest that it is less robust (see section 2.2.1). The top left plot of figure 4.2 shows the corresponding

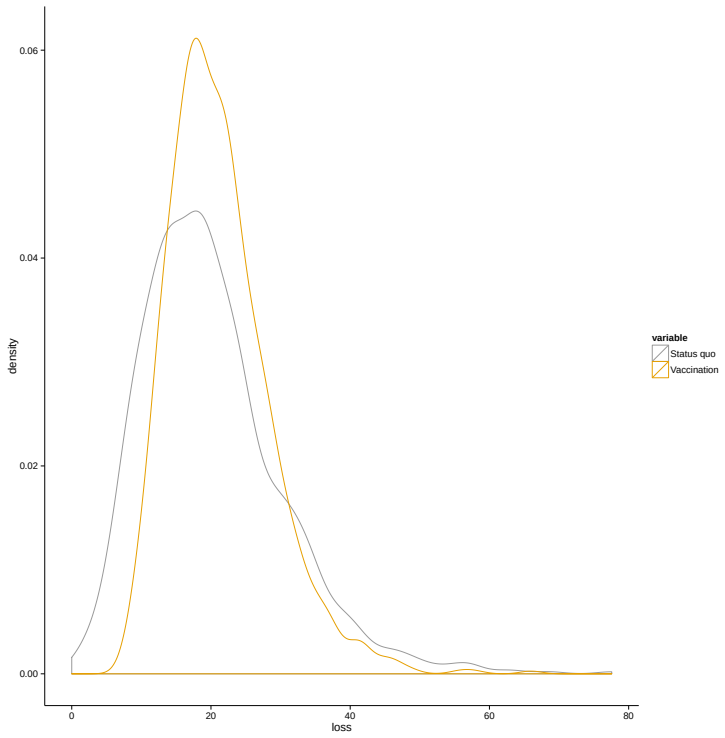

Fig. 4.1 Loss distributions of *vaccination* and *status quo* actions for a ‘willingness to pay’ parameter $k = 15000$. The expected loss value are respectively $\approx 26.9, 25.7$

inverse loss distributions for the two actions with the same value of k. This highlights that the *status quo* action (grey) is more unstable than the vaccination alternative, with the curve being steeper around both 0 and 1. The flatter profile of the *vaccination* action is indicative of a 'safer' loss distribution, i.e. tighter around the mode, which would probably suffer less from misspecification in the tails. However, of note is the crossing point at $q \approx 0.25$. For smaller quantiles, the *status quo* incurs higher loss. Thus to prefer *vaccination* over *status quo*, one would have to be strongly risk averse.

I stress that this graph has primarily a comparative purpose. If a discrete set of possible alternatives $\mathscr{A}$ has been identified as 'plausible', then this plot - and the following plots - allow for direct comparison between the alternatives when the expected loss estimate is not trusted (due to suspected or known misspecification). I also think that the inverse loss distributions are easier to compare than the loss density plots, although both are informative.

4.2.3 CVaR and CEL

Conditional Value at Risk In finance the *Conditional Value at Risk* (CVaR, Rockafellar and Uryasev, 2000, 2002) is another popular alternative to expected loss (or utility). To a statistician it represents the lower trimmed mean of loss (or upper trimmed mean of utility),

$$G_a(q) = \frac{1}{q}\int_{F_{Z_a}^{-1}(1-q)}^{\infty} z f_a(z) dz.$$

This gives the expected value of an action conditional on the event (θ) occurring above a quantile of loss. q can be seen as regulating the amount of pessimism towards Nature, with $\lim_{q\to 0}\sup_a G_a(q)$ corresponding to the minimax rule.

Another strategy for taking into account model misspecification is by considering the two-sided *trimmed expected loss*, defined as:

$$H(q) = \frac{1}{1-q}\int_{F_{Z_a}^{-1}(q/2)}^{F_{Z_a}^{-1}(1-q/2)} z f_a(z) dz$$

This is a robust measure of expected loss formed by discarding events with highest and lowest loss. Both these statistics are easily approximated using the bag of samples $\{\theta_i\}_{i=1}^n$ and the sort mapping υ defined previously. I use the linear interpolation,

$$\hat{G}_a(\frac{k}{m}) = \frac{1}{k}\sum_{i=1}^{k} L_a(\theta_{\upsilon(i)}), \quad k = 0,..,m$$

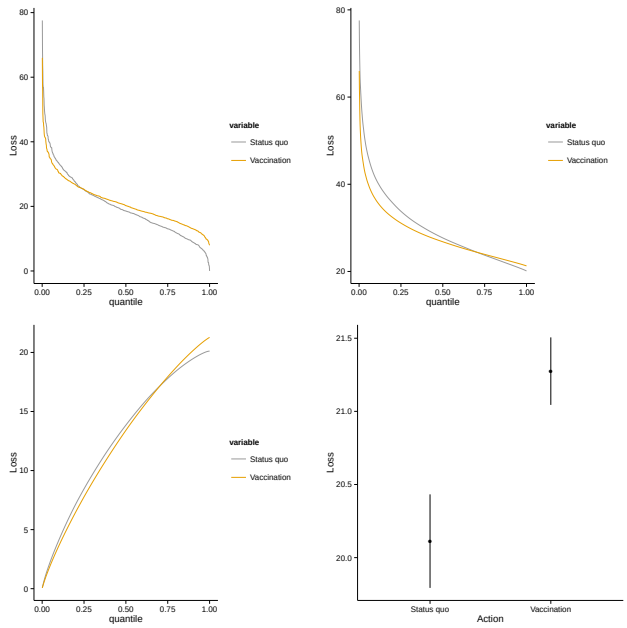

Fig. 4.2 Diagnostic plots for the decision system given in Baio and Dawid (2011) comparing vaccination (yellow) to status quo (grey) with the 'willingness to pay' parameter set to 15000. From top left to bottom right: Inverse loss distribution; Conditional value at risk; Cumulative expected loss; Expected loss centred at two intervals of standard deviation.

For a set of actions $\mathscr{A}$, it is possible to quantify the stability of the optimal action a^* evaluated under expected loss, by observing the first CVaR crossing point. That is to say the first value $q \in [0,1]$ such that a^* is no longer optimal, evaluated under CVaR(q).

Cumulative Expected Loss By the simple transformation $J_a(q) = qG_a(q)$ one obtains a monotone increasing function (if all the loss values are positive). This allows for a direct comparison of the contribution to the overall expected loss from high and loss quantiles of loss. Thus the *Cumulative Expected Loss* (CEL) function for action a is defined as,

$$J_a(q) = \int_{F_{Z_a}^{-1}(1-q)}^{\infty} z f_a(z) dz.$$

for $q \in [0,1]$. The overall shape of $J_a(q)$ provides a qualitative description of decision sensitivity. An action with CEL-plot that is steeply rising as $q \to 0$ is 'heavily downside' with expected-loss driven by low-probability high loss outcomes, while CEL-plot rising at 1 indicates 'heavy upside'. In particular $J_a(q)$ has a number of useful features:

- $J_a(q)$ quantifies the contribution to the expected loss of action a, from the $100 \times (1 - q)\%$ set of outcomes with greatest loss.

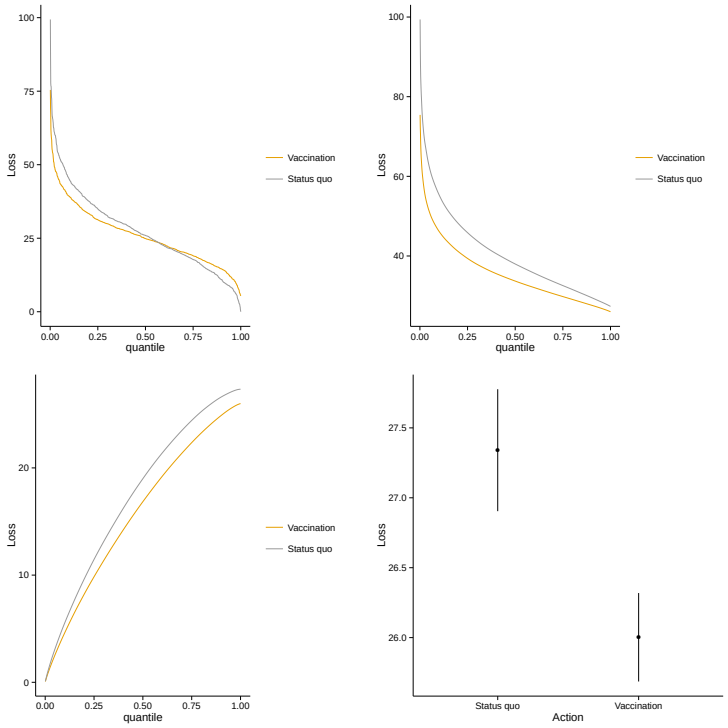

Fig. 4.3 Diagnostic plots for the decision system given in Baio and Dawid (2011) comparing vaccination (yellow) to status quo (grey) with the 'willingness to pay' parameter set to 25000. From top left to bottom right: Inverse loss distribution; Conditional value at risk; Cumulative expected loss; Expected loss centred at two intervals of standard deviation.

- $J_a(1) = \mathbb{E}_{\pi_I(\theta)}[L_a(\theta)]$, is the expected loss of action a, and $a^* = \arg\max_{a\in A} J_a(1)$ is the optimal Savage action.
- $J'_a(q) = \inf_{z^*\in\mathbb{R}^+}\{Pr(Z_a \leq z^*) = 1-q\}$, the gradient of the curve at $J_a(q)$ gives the threshold loss value z^*, such that under action a we can expect with probability $(1-q)$ the outcome to have loss less than or equal to z^*. This is the 'Value-at-Risk' of action a outlined above.
- $J'_a(0) = \sup_{\theta\in\Theta} L_a(\theta)$, and hence the Wald minimax action is given by: $\tilde{a} = \arg\min_{a\in A} J'_a(0)$ (the action with steepest gradient as $q \to 0$).

As noted previously, these graphs have a comparative purpose, and depending on the application they will highlight differences in the loss distributions or not.

The top right and bottom left plots in figure 4.2 show the CVaR and CEL plots respectively for a 'willingness-to-pay' parameter of $k = 15000$. In this application, the CEL curves have the same profile so are not useful in distinguishing the two actions. However, the CVaR clearly shows that the vaccination alternative is safer, as the blue curve is dominated by the

red for most values of q except very close to 1 (i.e. the expected loss estimate). In chapter 5 there are examples of the opposite situation, i.e. the CVAR plot not showing any difference but the CEL plot discriminating between actions.

It is interesting to contrast the diagnostic plots in figures 4.2 and 4.3. The former shows that for $k = 15000$, the difference between the two actions is very slight, and in fact the action with higher expected loss is more robust to misspecification in the model. For the latter, however, when $k = 25000$, there are no crossing points in the CVaR nor the CEL plots, with the *status quo* action very clearly the optimal choice.

4.2.4 Loss-density plot

As a final diagnostic plot linking the posterior and the loss function, I look at the relationship between the density estimates $\pi_I(\theta_i)$ and their corresponding losses $L(a,\theta_i)$ for the collection of Monte Carlo samples θ_i and a set of actions $a \in \mathscr{A}$. I note that it might not be possible in all applications to directly access these estimates $\pi_I(\theta_i)$, but in this case kernel density estimation methods could be used. This is of interest because it allows the user to see whether the expected loss estimate is driven by events which have low probability of occurring.

The sensitivity of the estimates $1/m\sum_i L(a,\theta_i)$ for $a \in \mathscr{A}$ could be qualitatively determined by a process similar to outlier detection, where outlier status would be determined by how extreme the value of $\pi_I(\theta_i)$ is (lower values being more extreme). The range of values $\pi_I(\theta_i)$ can be transformed so as to be contained within (0,1] by dividing by $\max_i \pi_I(\theta_i)$. Here are three alternatives methods for constructing the loss-density plot.

- Directly construct a scatterplot of the pairs $\{\pi_I(\theta_i), L(a,\theta_i)\}$.
- Bag the values $\pi_I(\theta_i)$ into K bins of equal size or with equal number of elements.
- Plot the values $\pi_I(\theta_i)$ on a logarithmic scale.

This diagnostic plot can highlight under-explored areas of loss, which correspond to tails of the distribution $\pi_I(\theta)$. The previous plots given in sections 4.2.2 and 4.2.3 (VaR, CVaR and CEL) do not tell us whether the losses $L(a,\theta_i)$ are coming from the tails of π_I but rather whether they are coming from the tail of F_{Z_a} (loss distribution).

Figure 4.4 plots the loss-density relationship for 4 selected parameters: ϕ, the vaccine coverage probability; ρ, the vaccine effectiveness probability; π, the probability of infection in the non-vaccinated group; and 'death', the probability of dying in the non-vaccinated group (this is based on previous states which are discrete, which explains the finite number of possible densities in the bottom right plot). As ϕ does not effect the distribution of loss under the action *status quo*, the loss densities are identical (top left plot). In the same way, ρ

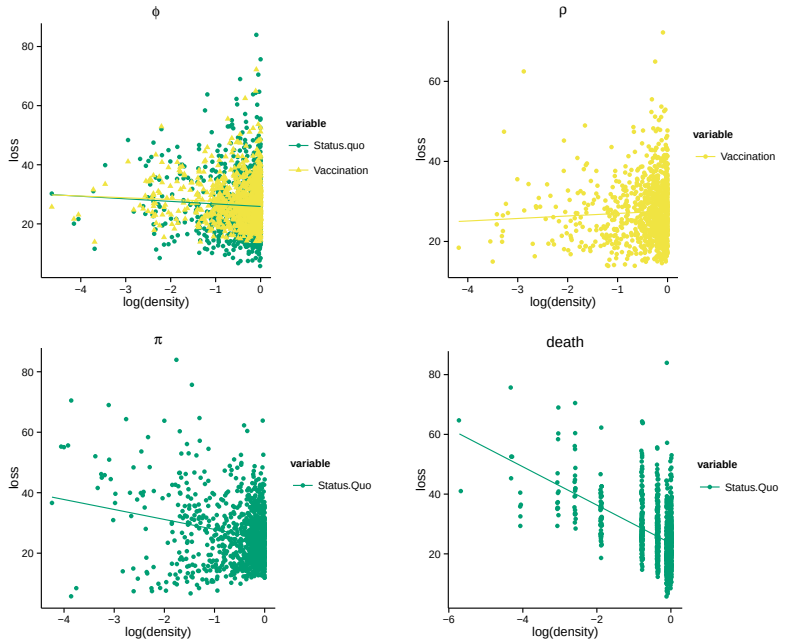

Fig. 4.4 Loss log-density plot for 4 informative parameters of the BCEA model; willingness-to-pay: $k = 15000$. For the parameters ρ, π and 'death' the loss function is constant for one of the two actions, so I only plot the action of interest. The lines represent a linear fit to the points (using default linear fit in *ggplot2*).

does not effect the loss for the *status quo* group so I only plot the loss-density relationship for the *vaccination* action. Same for π and 'death', which only effect the *status quo* action. I also fit for each action a linear regression curve (using the default method from the *R* plotting package *ggplot2*). Unsurprisingly, 'death' comes as the most informative parameter, the rare events being correlated to much higher losses.

Outlier detection - sensitivity to the prior As a final idea, which is not implemented in this work because the BCEA model is synthetic and has no data, it is possible to look at the effect on the posterior distribution of removing one (or more) datapoint x_j. I propose a simple importance sampling method for evaluating $\pi_{I-\{x_j\}}$ and $\pi_{I-\pi}$, denoting respectively the posterior without the datum x_j and the posterior without the prior π (the posterior with a flat prior). The prior can be considered as one datum in the computation of the posterior. The importance weights are given by:

$$w_i^{-x_j} = \frac{1}{f(x_j|\theta_i)}, \quad w_i^{-\pi} = \frac{1}{\pi(\theta_i)}$$

These weights give leave-one-out (LOO) estimates of the expected loss, where the prior can be considered as one extra data point:

$$\psi_a^{-x_j} = \frac{1}{\sum_i w_i^{-x_j}} \sum_i w_i^{-x_j} L_a(\theta_i)$$

$$\psi_a^{-\pi} = \frac{1}{\sum_i w_i^{-\pi}} \sum_i w_i^{-\pi} L_a(\theta_i)$$

Thus the effect of single data points can be evaluated (detection of outliers) as can the effect of the prior, which is especially useful in small sample situations.

4.2.5 'Coherent' diagnostics

In finance and actuarial science, diagnostic statistics on 'risky positions' are called *risk measures*. Note that these are not probability measures, but rather functionals defined on loss distributions for a given $a \in \mathscr{A}$ (in my notation). These are denoted $\rho : Z_a \to \mathbb{R}$. An important notion in this literature is that of a *coherent risk measure* (Artzner et al., 1999), which is defined as a function ρ satisfying four axioms that financial positions[3] (in our framework these are actions) are thought to obey. These are: *translational invariance*: $\rho(Z+c) = \rho(Z) + c$, where c is a constant; *monotonicity*: if $Z$is stochastically dominated by Y, then $\rho(Z) \leq \rho(Y)$; *positive homogeneity*: $\rho(\lambda Z) = \lambda\rho(Z)$, for $\lambda \geq 0$; *subadditivity*: $\rho(Z+Y) \leq \rho(Z) + \rho(Y)$. By $\rho(Z+Y)$ is meant the risk measure on the combined loss distribution of the combination of the two actions corresponding to the loss distributions Z and Y. Because of the importance of coherency (which is not at all related to the notion of Bayesian coherency outlined in section 2.2.1) in the financial literature I briefly discuss its position for some diagnostic summary statistics.

A simple example of a risk measure is the α-scaled standard deviation: $\rho(Z) = \alpha\sqrt{Var(Z)}$. This is positive homogeneous and sub-additive but not translation invariant or monotone. A more interesting case is that of the Value-at-Risk for a given quantile q: $\text{VaR}_q(Z) = F_Z^{-1}(1-q)$. This fails the coherency test as it is not sub-additive (see for example Artzner et al., 1999). This has made VaR particularly controversial as it is written into official regulations (Basle Committee, 1996) (defining how risky positions are allowed to be). In financial applications the sub-additivity axiom is regarded as highly important as it implies that an investor reduces his overall risk (expected loss) by diversifying his portfolio. This is one of the reasons that VaR has been rejected in finance as a reliable measure of a portfolio's risk. The *coherent* alternative that has been proposed for VaR is the Conditional Value-at-Risk

[3]For example buying one stock, which is a position on the market.

(also known as expected shortfall in finance). Note that expected loss and expected loss under the local least favourable distribution are both coherent risk measures. Indeed, Theorem 3.2 from Ahmadi-Javid (2012) states that any risk measure represented by the supremum of a family of distributions will be coherent. However, in statistical applications it is not clear why this axiom should be desirable in general[4], therefore not eliminating VaR_q as a useful statistic and basis for a diagnostic plot.

4.3 Exploring Kullback-Leibler neighbourhoods

The plots discussed in the previous sections look at the loss distribution under the reference model π_I. Following the ideas in chapter 2, it is natural to plot the change in this quantity within a neighbourhood Γ_C, and more importantly as a function of the neighbourhood size.

4.3.1 Stability of Bayesian optimal actions

The result from section 2.2.1 (Chapter 2, local sensitivity) justifies using the variance of the loss distribution to measure how sensitive the expected loss estimate is to perturbations with respect to the least favourable distribution for very small neighbourhoods. But larger neighbourhoods may be of interest, and Chapter 2 shows how to importance sample the least favourable distribution π_a^{sup}. Thus it is natural to graphically compare a set of actions by plotting their expected loss under the local least favourable distribution, as a function of the neighbourhood size C. This allows the user (decision maker) to see how sensitive the expected loss estimates are to changes in neighbourhood size, and which actions are more sensitive than others. At $C = 0$ the actions are ranked by expected loss under the reference model π_I, and as C increases, this ranking may be altered. In the limit, $C \to \infty$, the ranking will correspond to that of minimax. Thus we can observe change points, where the optimal (Bayesian) action is no longer optimal. As an example, in figure 4.5 (top plot) I plot the expected least favourable loss as a function of C for *vaccination* vs. *status quo*. In this setting we see that the optimal Bayes action (*status quo* - grey line) becomes sub-optimal for KL values of $C \approx 0.075$ (the calibration of these values is considered in the next section). This is because the variance of *status quo* action is much larger than that of the *vaccination*, as noted in the previous section. Thus as C becomes larger, more weight is placed onto the high loss events, which are higher for the *status quo* action. The bottom plot shows the difference

[4] A trivial example of where it does not hold is given by the following: consider two gambles, A, lose £10 if a fair coin falls on Heads; B, lose £10 if the fair coin falls on Tails. For the same coin, if both gambles are taken then one loses £10 with probability 1.

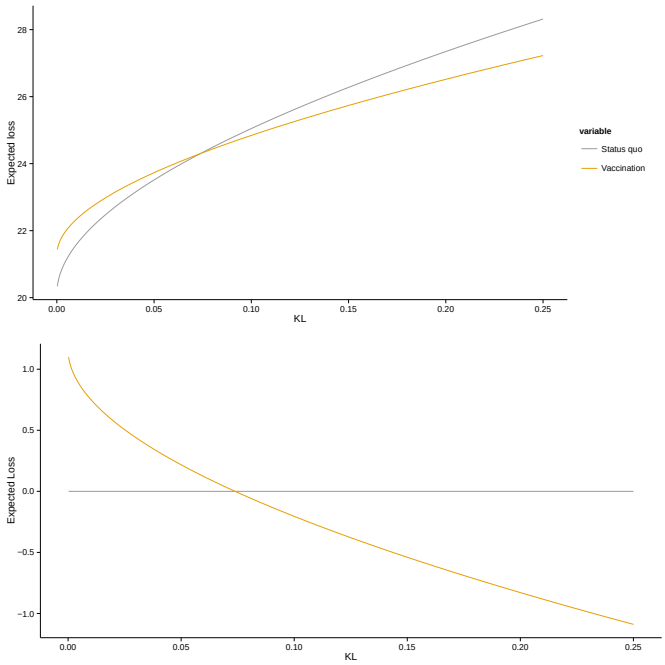

Fig. 4.5 Expected loss under local least favourable distribution within a KL ball as function of the KL radius C (top). Difference in expected loss between each action and the Bayes optimal action (*status quo* - bottom plot).

in expected loss between the two actions, also as a function of the radius C. The crossing point can be observed at $C \approx 0.075$.

Chapter 2 looked at *local admissibility* of actions. This is defined with respect to a neighbourhood Γ_C, and is computed using the 'regret' loss function $L_{a,a'}(\theta) = L_a(\theta) - L_{a'}(\theta)$. If this expectation is negative for the least favourable distribution in Γ_C, then it is also negative for all distributions in Γ_C, which means that there is no distribution $\pi \in \Gamma_C$ such that $a > a'$. This makes a' locally admissible in the region Γ_C. By plotting the least favourable expected loss of the regret loss between the optimal Bayes action a^* and a set of others, it is possible to compute the largest region $\Gamma_{C_{\max}}$ such that a^* is locally admissible. In figure 4.6 I plot the expected loss of the regret loss between *status quo* and *vaccination*. This identifies a KL ball of approximately the same size as in figure 4.5 with regards to the crossing point at zero (note this will not be the case in general, as it is a stronger condition).

These two plots give an idea of the stability of the optimal Bayes action by characterising the largest KL neighbourhood such that the action is optimal, either under a local minimax expected loss, or with respect to every distribution within that neighbourhood. It is also of importance to understand how exactly the least favourable distribution is changing π_I.

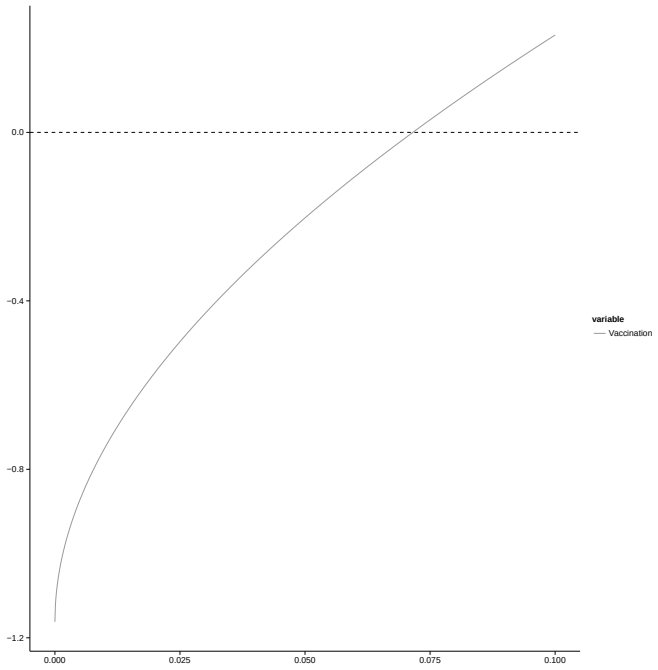

Fig. 4.6 Local admissibility: expected loss under the local least favourable distribution of the 'regret' loss function $L_{SQ}(\theta) - L_{vac}(\theta)$. We see that *status quo* is no longer optimal for KL neighbourhoods of radii greater than 0.015.

4.3.2 Calibration of Kullback-Leibler

Distribution of importance weights

Section 2.3.4 presented some methods for the calibration of the Kullback-Leibler divergence. In the computational decision theory set-up where π_I is approximated by a bag of samples θ_i, it is natural to use the discrete nature of the approximation to help calibrate feasible values. In particular I proposed using the distribution of the (normalised) importance weights, $w_i = e^{\lambda L_a(\theta_i)}$, to identify a plausible range of KL values. This can be done using the Gini coefficient for example, which is defined as the twice the difference in area between the line $y = x$ and the cumulative distribution of (sorted) weights, for $x \in [0,1]$. An example is given in figure 4.7 using the importance weights for the *vaccination* alternative corresponding to a Gini coefficient of 0.3. The curved red line corresponds to the cumulative distribution of the importance weights, the Gini coefficient is given by twice the area between the black and red lines. The black line is known as the line of 'perfect equality', i.e. every $w_i = 1/n$.

The main attraction of using a score such as the Gini coefficient is that it is defined on the interval $[0,1]$ (makes it easier to set default values) and is well-known. This allows a user to narrow down the choice of possible values of the KL divergence, whilst at the same

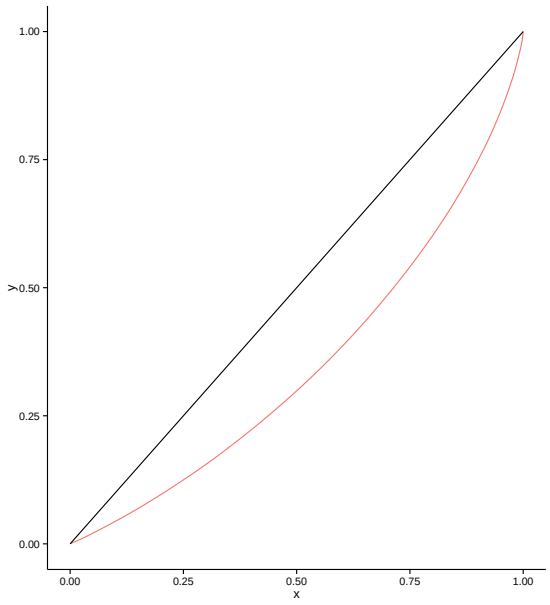

Fig. 4.7 Cumulative distribution of importance weights (red line) for the *vaccination* alternative that corresponds to a Gini coefficient of 0.3 (twice the area between the black and red lines).

time choosing KL values which are small enough so that the importance sampling is a valid method for approximating π_a^{sup}.

How plausible is π_a^{sup}?

In order to interpret the meaning of a particular KL divergence, a first idea is to simply plot the corresponding loss distribution. The loss distribution can in some cases (see for example section 5.2, chapter 5) be directly interpreted as plausible or not (for example the minimax loss distribution, a point mass on $\theta_a^* = \arg\max_{\theta\in\Theta} L_a(\theta)$, would often be easy to reject as too pessimistic). Figure 4.8 shows the loss distribution of the two actions (*status quo* - grey; *vaccination* - yellow) under the reference distribution π_I (top left) and then under local least favourable distributions π_a^{sup} for KL divergences corresponding to Gini coefficients of 0.1, 0.2, and 0.3 respectively (top right to bottom right). We note a shift of mass from low to high values of loss, where the undulating pattern (very visible in the bottom right plot, Gini coefficient of 0.3) is an artefact of the unequal weight distribution between samples. This would suggest that the a Gini coefficient of 0.3 is too high in this particular example. Chapter 5 gives two particular applications which illustrate how these methods of calibration can be used to choose plausible values of KL. It also looks at plotting marginals of the tilted distribution. Which marginals should be plotted will be context dependent.

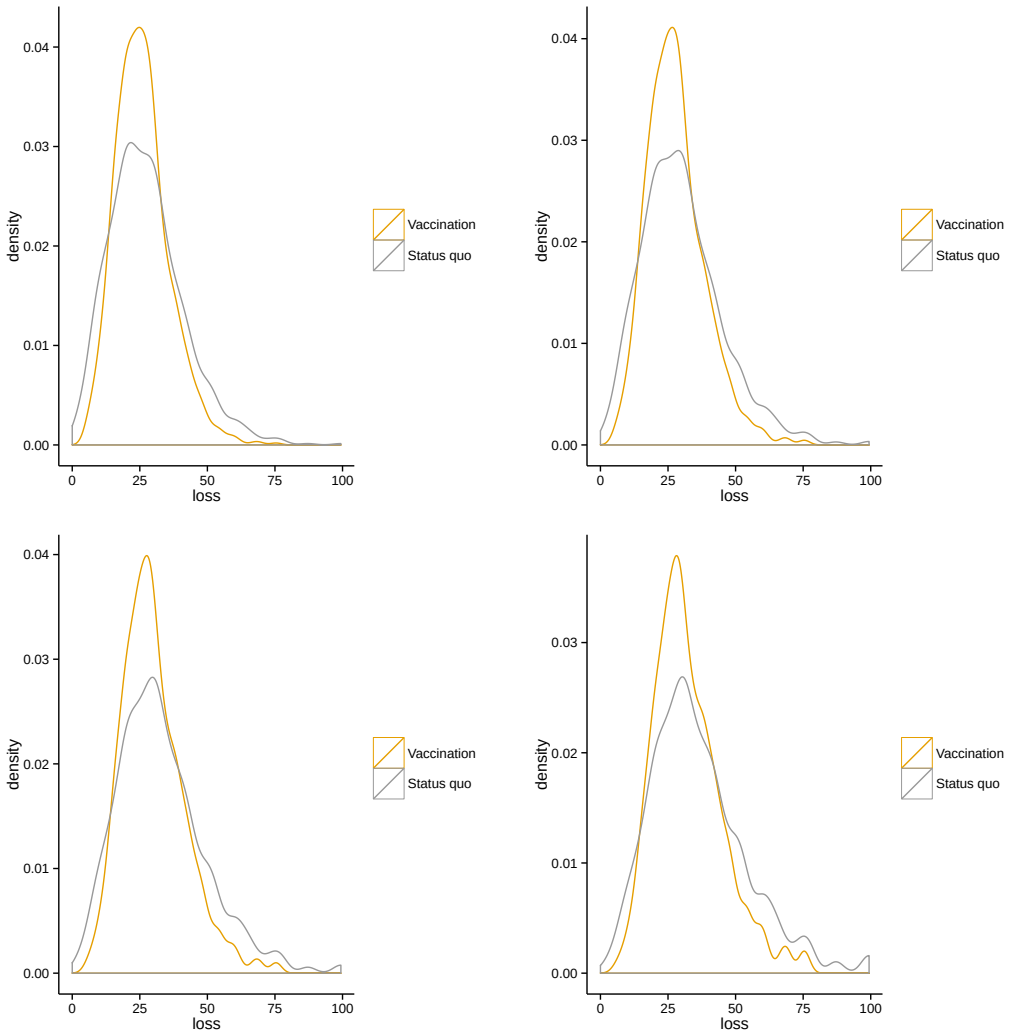

Fig. 4.8 Local least favourable loss distribution. Top left: loss distributions under π_I. Top right to bottom left: loss distributions at KL divergences corresponding a Gini coefficient of the re-weighing vector of 0.1, 0.2 and 0.3 respectively.

4.4 Discussion

This chapter is motivated by the lack of available graphical tools to diagnose the robustness of formal decision systems. It presents a series of methods which can be used together to qualitatively assess the stability of a formal decision theoretic system. When concerns of model misspecification are present, this is a vital part of the construction and validation of the decision system. Although it does not highlight where exactly in the model misspecification may occur, it does allow the decision maker to qualify the importance of possible misspecifications. The final section of Chapter 2 noted that the framework lacked a convincing method for the calibration of the Kullback-Leibler divergence. This is indeed true, and this chapter gives only qualitative methods for this calibration which partially solve the problem. However, even though this work does not fully answer how to interpret units of KL, in the context of these plots (such as figure 4.5), a main element is the behaviour of the expected loss values of the actions as the KL ball is enlarged. The local minimax framework provides a principled method for considering the effect of model misspecification and these plots show how to implement these ideas in practice. This is further developed in the next chapter, where I consider two applications from medical decision making. A key part of the methodology is the computational aspect, whereby these plots can be produced at little to no extra computational cost when Monte Carlo is used to estimate the expected loss values. Part of my contribution has been to write an *R* package which automatically produces the series of plots presented in this chapter, given inputs consisting of Monte Carlo samples, a loss function and a set of actions (decisions).

Chapter 5

Case studies in medical decision making

This chapter considers the practical implementation of a formal decision-theoretic framework. I first discuss some difficulties regarding the use of decision theory in practice. I then present two case studies, both in medical decision making, which necessitate the use of a probabilistic model. In both cases the loss function is easy to elicit, albeit with one free parameter.

The use of the formal sensitivity analysis framework is demonstrated on these two case studies by means of diagnostic plots and summary statistics, all implemented using an R package developed for these purposes.

5.1 Difficulties with the expected loss paradigm

Chapter 1 gave a number of well-known objections to the Savage axioms which motivate a D_{open} approach to Bayesian decision theory. Very rarely the decision maker can assume they are working in D_{closed} because of model misspecification, either known or suspected. There are, however, many fixes for this problem, ranging from practical adjustments such as robust Bayesian analysis (looking at the sensitivity of the posterior inferences with respect to the prior distribution), methodological changes such as the use of nonparametric Bayesian models (which have full support so that in theory the statistician is in M_{closed}), to even a re-interpretation of the prior distribution as given by Walker (2013). Chapters 2-4 of this thesis considered an *ex post* framework for assessing the robustness of a formal decision system. This is motivated by exploring 'neighbourhoods' of the approximating model and producing diagnostic plots that graphically depict model robustness. The purpose of these methods is the development of automated tools for checking the robustness of a statistical decision system. I believe that this is of use to the statistical community. However, it is necessary to address an important practical impediment to the implementation of Bayesian decision theory. This is the specification of the loss function itself.

It can be argued that loss functions have less of a foundational background than prior distributions (see Chapter 13, page 423 of Jaynes, 2003) which are based on principles such as, for example, optimal information processing rules (Zellner, 1988), maximising entropy (Jaynes, 1957) or decision-theoretic principles (Savage, 1954). This is not true in general, certain loss functions are motivated by foundational principles, such as the negative logarithmic loss, $L(\pi, y) = -\log \pi(y|x)$ for a predictive model π given data x, discussed in section 2.2.3. This loss function forces 'honesty' when reporting beliefs $\pi(y|x)$. But it is true that there is more literature concerning prior elicitation than loss elicitation in situations where expert knowledge does not apply, whether it is by means of objective priors (Bernardo, 1979b; Ghosh, 2011), weakly informative priors (Gelman et al., 2008), or default priors (Martins et al., 2014). The difficulty of eliciting loss functions is alleviated somewhat by the idea of 'strategic equivalence' (Ruggeri et al., 2005), whereby for a given problem, there will be a class of loss functions for which the ranking of actions under expected loss remains the same. Therefore it is only necessary, for example, to specify the loss function up to a linear transformation (with a positive first order coefficient).

On the other hand, much of the area of *Robust Statistics* focusses on the loss function as the object of primary interest. For example, M-estimation (as described in Chapter 1, section 1.2.2) considers minimising general loss functions of the form $\rho(x_i; \theta)$ (instead of $-f(x_i; \theta)$, which leads to maximum likelihood estimation) to estimate location parameters. This is part of the framework denoted as M_{free} by Bissiri et al. (2013) for *model free*, as opposed to M_{open} and M_{closed} (see Chapter 1, section 1.3.1). Bissiri et al. extend this to a general framework in which it is possible to update beliefs using loss functions. In this way, the loss function replaces the role played by the likelihood in Bayesian statistics. The update involves both a loss to the data and a loss to the prior, the later of which can be derived as the Kullback-Leibler divergence (this is the unique solution in order to obtain coherent updating, see Theorem 1, Bissiri et al., 2013). The loss to the data would be problem specific however. However, every application needs a tailor made loss function, which is often not the case, with choices of loss function often made for analytical and tractability reasons (for example the zero-one loss, squared error loss). There are some examples of the contrary (for example, specific loss functions for Bayesian imaging, Rue, 1995), but this is not the usual practice.

The general problem of loss elicitation is beyond the scope of this work, but here are some ideas of how it may be approached in robust way. The decision problems considered in this chapter are medical decision problems, where the loss function is of the form:

$$L(a, \theta) = Cost(a, \theta) - r \cdot Benefit(a, \theta)$$

where r is a trade-off parameter between the loss incurred by action a (this will often be monetary cost) and the utility of a (benefit of treatment for example, could be in units such as Quality Adjusted Life Years - QALYs, see Loomes and McKenzie, 1989; Zeckhauser and Shepard, 1976). This trade-off can be difficult to specify as it can seem arbitrary. This has motivated a sensitivity analysis dedicated to this one parameter, known as the 'willingness-to-pay' parameter, see Baio and Dawid (2011).

A solution to having to make such trade-offs is proposed by Müller (2005), which he calls 'stylised Bayes'. Instead of performing a formal decision analysis, the decision maker specifies desirable frequentist properties that the posterior distribution should possess. The problem can be approached from the opposite angle as well, instead of constructing a prescriptive loss function, the decision maker looks at previous decisions which can then determine plausible ranges for parameters such as r. This has been done for institutions such as NICE (Devlin and Parkin, 2004). Even though medical decision problems encounter such difficulties, it would seem that formal decision-theoretic analyses have been successful in this area. The next two sections concentrate on two examples. Chapter 6 goes further by considering an alternative to specifying a full model, whereby it is possible to directly target the loss distribution.

5.2 Breast cancer screening

5.2.1 Motivation and model design

Public health policy is an area where the application of statistical modelling can be used to optimally allocate resources. Breast cancer screening for healthy women over a certain age to detect asymptomatic tumours is a hotly debated and controversial issue for which it is difficult to fully quantify the benefits. A recent independent review (Marmot et al., 2012), commissioned by Cancer Research UK and the Department of Health (England) concludes that only a randomised clinical trial would fully resolve this issue. A primary issue is determining the optimal screening schedule, consisting of a starting time t_0 (age of first screen), and a frequency δ for subsequent screens. It is of course sharply infeasible to trial all combinations of schedules (t_0, δ). An optimal trial design however can be constructed with the help of a statistical model of the disease progression within a population. Parmigiani (1993) proposed using a four-state semi-Markov process, which generalises to any chronic disease characterised by an asymptomatic stage. All individuals start in state A, disease-free. They then transition either to the absorbing state D (death) or contract the disease, modelled by a transition to state B, the pre-clinical stage. This is followed by a transition to either the

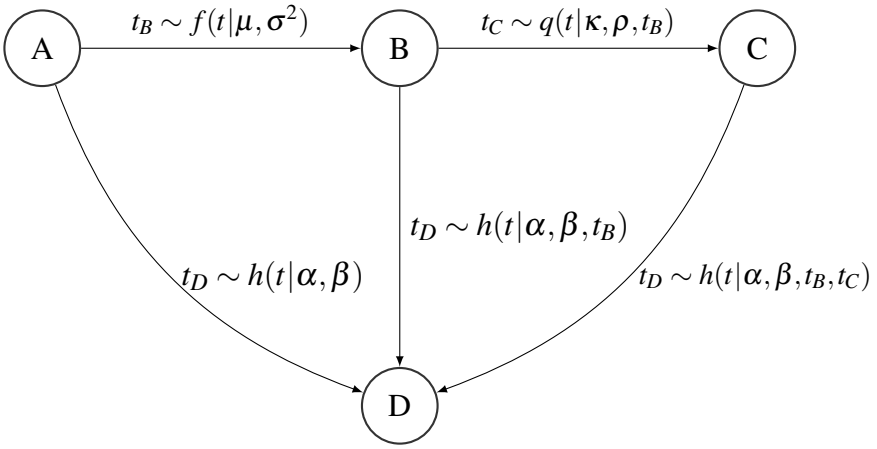

Fig. 5.1 Graphical model of the transitions and transition densities between states.

clinical stage of the disease or death. It was assumed that each transition happens after a time t with the following densities:

$$\begin{aligned} t_D &\sim h(t|\alpha,\beta) = \text{Weibull}(\alpha,\beta) \\ t_B &\sim f(t|\mu,\sigma^2) = \text{LogNormal}(\mu,\sigma^2) \\ t_C &\sim q(t|\kappa,\rho) = \text{LogLogistic}(\kappa,\rho) \end{aligned} \tag{5.1}$$

Figure 5.1 shows a graphical model of the four-state semi-Markov process with the associated transition densities. An individual is characterised by the triple $t = (t_B, t_C, t_D)$ where the symptomatic stage of the disease is contracted only when $t_D > t_B + t_C$ (this assumes that all individuals will contract the disease if they live long enough). For a screening schedule $a = (t_0, \delta)$ the loss function is defined as follows (a function of the times $t = (t_B, t_C, t_D)$):

$$L(a,t) = r \cdot n_a(t) + \mathbb{1}_C \tag{5.2}$$

where n_a is the number of screening schedules an individual will receive during their lifetime, until they die or enter into the symptomatic stage of the disease. $\mathbb{1}_C$ is the indicator function, taking value 1 for the event that the pre-clinical tumour is not detected by screening or occurs before t_0, and zero otherwise. r trades off the cost of one screen against the cost incurred by the onset of the clinical disease. Note that this is a different free parameter from the loss function in Chapter 4 (section 4.1.1), which converts units of 'life' (QALYs) into monetary units. In this application, both components of the loss function can be elicited in monetary units.

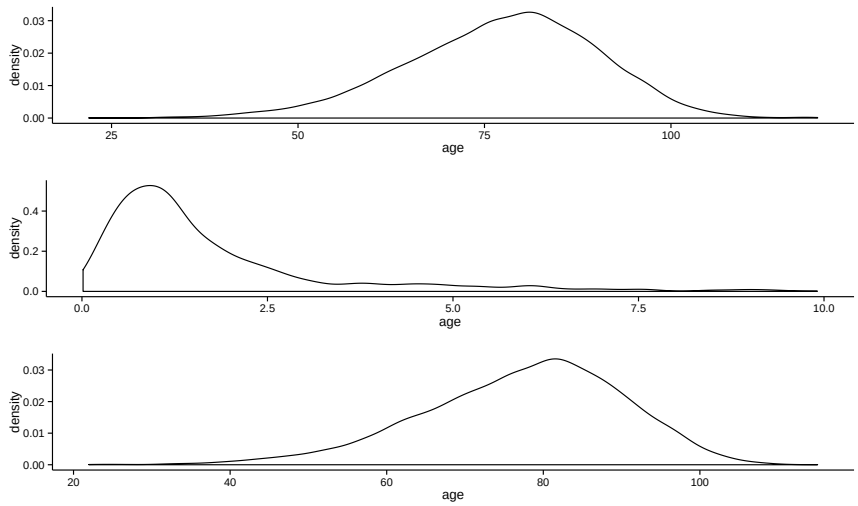

Fig. 5.2 Probabilistic model of transition times: (from top to bottom) marginal densities of transition times to preclinical stage, transition from preclinical to clinical stage, and death times. Note the time scales are not the same for the three plots.

Each screen has an age-dependent false-negative rate modelled with a logistic function:

$$\beta(t) = \frac{1}{1 + e^{-b_o - b_1(t-\tilde{t})}}$$

where $\tilde{t}$ is the average age at entry in the study group.

5.2.2 Privacy and reproducible research

For this particular application, I did not have access to a dataset in order to construct a posterior distribution for the model described in the previous section. However, it so happens that previous researchers, Wu et al. (2007), have made posterior samples of their model freely available online. These are from their analysis of the HIP study (Shapiro et al., 1988). The robustness analysis I present only depends on posterior samples $\theta_j = (\mu, \sigma^2, \kappa, \rho, b_0, b_1)$ of the parameters in the model given in 5.1 and the loss function given in 5.2.

The fact that it is possible to proceed with the analysis in this way highlights an interesting aspect of the methods developed. Many datasets, such as the HIP study on breast cancer screening, are confidential and cannot be made freely available online. This is true even after anonymisation, as there can be enough information left in the data to identify patients using techniques such as cross-referencing other public datasets (see for example Narayanan and Shmatikov, 2008). However, if a formal Bayesian statistical analysis, using a

model with known[1] prior $\pi(\theta)$ and known likelihood $f(x|\theta)$ was performed on an unknown dataset $\{x_i\}_{i=1}^n$ with $n \geq 2$, then I posit that posterior samples θ_j would not give *enough*[2] information regarding the individual data points. Summary statistics concerning x_i can of course be derived from the θ_j's. Thus a researcher can make the analysis public by revealing $\left\{\pi, f, \{\theta_j\}_{j=1}^K\right\}$.

A formal description of this problem would require some thought as to what information concerning the data must be kept secret, along with other considerations as to how much prior knowledge there may be already available concerning x_i. I believe this is an interesting problem and deserves further development. In this way, analyses on confidential datasets can be made public, a small step to help move research towards more transparency and greater reproducibility.

5.2.3 *Ex-post* analysis

For the purposes of the analysis I simulated transition times for individuals in the population at risk. This is done by using the 2000 posterior parameter samples for $\theta = (\mu, \sigma^2, \kappa, \rho, b_0, b_1)$ given in the supplementary materials of Wu et al. (2007), based on data from the HIP study (Shapiro et al., 1988). Figure 5.2 shows the estimated marginal densities for 10^4 sampled times for each transition event[3].

To make the action space tractable, I first selected 32 alternative schedules, consisting of all combinations with:

$$t_0 \in \{55, 57, 59, 61, 63, 65, 67, 69\}$$
$$\delta \in \{9, 12, 15, 18, 24\}$$

where t_0 is in years, and δ is in months.

This choice of screening schedules is mainly illustrative. The optimal schedule will heavily depend on the choice of r (trade-off ratio in equation 5.2) which I do not attempt to justify (the value 10^{-3} was taken from the section 4.5 of Ruggeri et al. (2005) where the authors considered this same application). In order for the plots to be legible, the top 6 schedules were selected[4] (as ordered by expected loss under the reference model) for analysis (the only reason not to analyse a greater number of schedules is for clarity in plotting).

[1] By *known* and *unknown* I mean publicly available or not.

[2] What exactly is/is not *enough* is contextual.

[3] The Weibull distribution has values $\alpha = 7.233, \beta = 82.651$ which are the values used in Parmigiani (1993).

[4] Given in order of increasing expected loss these are: (59,15), (55,15), (57,15), (61,15), (55,18) and (59,18).

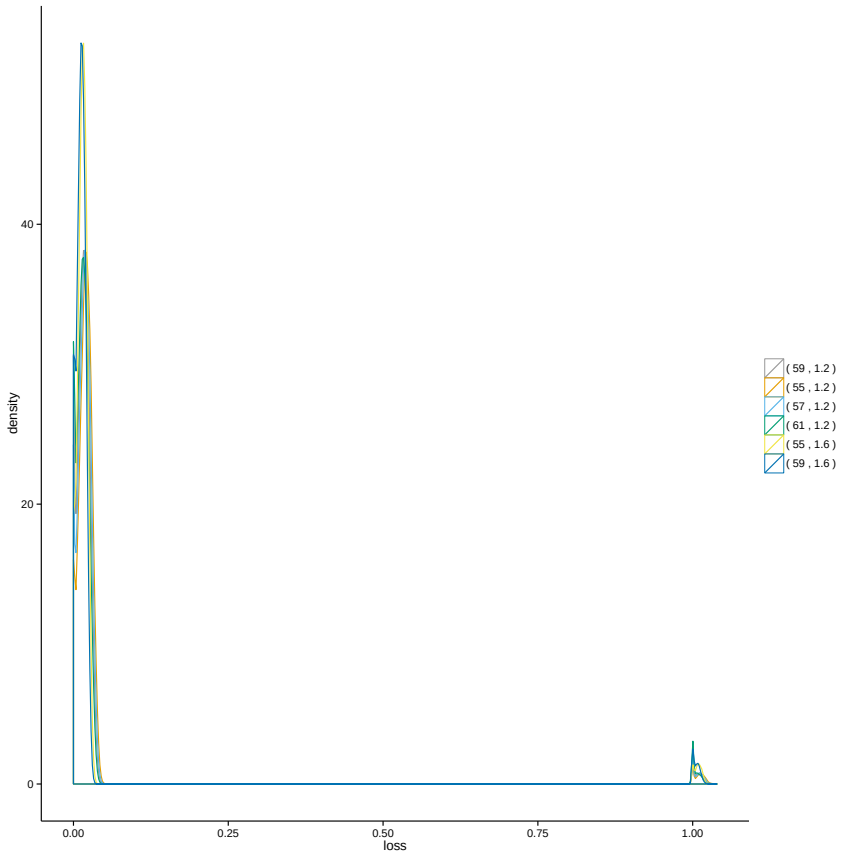

Fig. 5.3 Loss distributions of top 6 schedules out of the 32 considered.

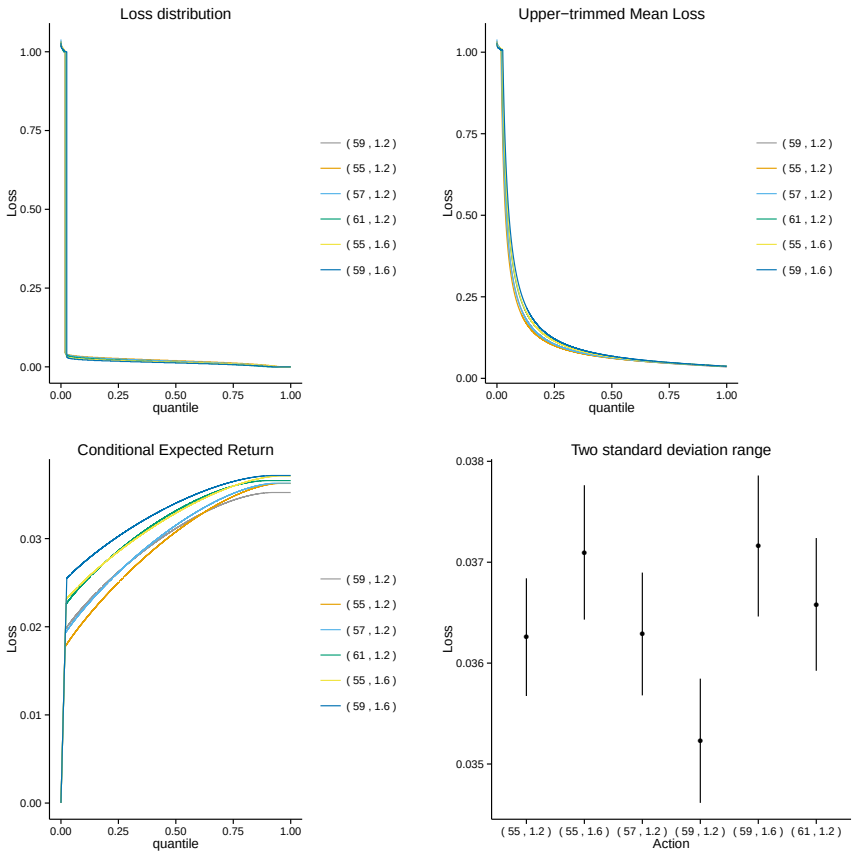

Fig. 5.4 Model diagnostic plots. From top left to bottom right: inverse loss distributions of the 6 actions (all very close in shape); upper trimmed mean loss (CVaR) which differentiates the actions by showing the higher downside in some schedules; conditional expected loss (CEL); estimates of expected loss centred inside intervals of two standard deviations.

The losses incurred for a particular schedule $a = (t_0, \delta)$ can be seen to be highly bimodal (the loss distributions of the top 6 actions scored by expected loss under π_I are shown in figure 5.3). Most of the population do not contract the disease and therefore contribute a loss of $r \cdot n_a$ (cost of each screen multiplied by the number of total screens during lifetime). The loss contributed by those who contract the clinical stage of the disease is therefore of magnitude $1/r$ greater.

Figure 5.4 shows the four diagnostic plots for the loss distribution presented in Chapter 4: the inverse loss distribution, the Value at Risk, the Conditional Value at Risk and the Conditional Expected Loss. The inverse loss distribution shows the large contribution of very rare events to the overall expected loss. The CVaR plot does not differentiate the 6 schedules, but the CEL plot has a branching out structure, where the actions with a larger contribution from tail events to the overall expected loss branch out higher up than the others. We observe a crossing point between the schedule $(59, 12)$ and $(55, 12)$ at approximately the 70^{th} quantile, which would suggest the $(55, 12)$ schedule is more robust (fewer high loss events). This plot clearly shows that the expected loss values are driven by low probability events (less than 10% of the mass).

Looking at the sensitivity of the 6 schedules with respect to Kullback-Leibler neighbourhoods around π_I, we see that for KL values larger than ≈ 0.01, the optimal Bayes action is no longer optimal under local least favourable expected loss, with the schedule $(55, 12)$ showing less sensitivity. This is given in figure 5.5, with the top plot showing the local least favourable expected loss as a function of the KL radius, and the bottom plot showing the difference in expected loss between each action and the optimal Bayes action (schedule $(59, 12)$). This confirms that the decision-system is sensitive to small changes in the model. This is also apparent from figure 5.6, where I plot the local admissibility of the optimal action under π_I (see Chapter 2, section 2.2.1). For very small neighbourhoods in KL divergence, the optimal Bayes action is no longer locally admissible.

To make sense of the values of KL, one can look at the local least favourable loss distributions for varying KL values, calculated using the Gini coefficient on the importance sampling weights (see Chapter 2, section 2.3.4). Figure 5.7 shows these loss density plots for the 6 top actions under π_I and then $\pi_a^{\sup}$ for $\mathrm{KL} = 0.1, 0.3, 0.6$ respectively, going from top left to bottom right. The effect can be see as transferring the mass from left to right, i.e. from low loss to high loss.

As a final diagnostic plot I look at the probability of optimality under the Dirichlet extension model from Chapter 3, section 3.3.2. Figure 5.8 shows this probability as a function of the concentration parameter α for each action in the top 6 selected. We can see that for large values of α (greater than 10^4) the optimal action under π_I is recovered.

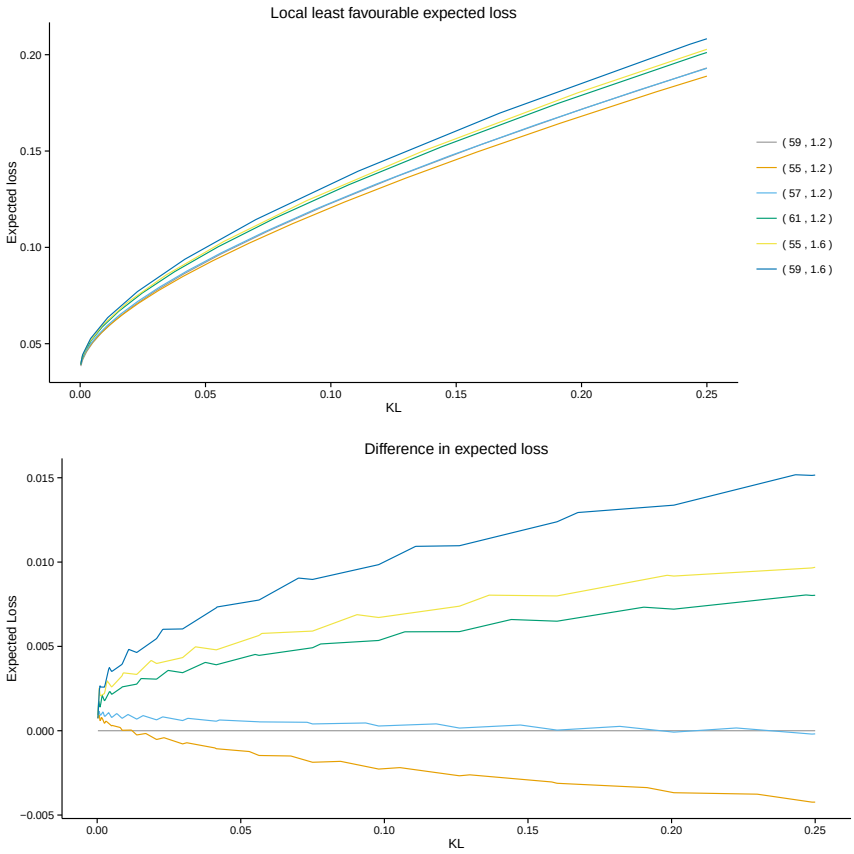

Fig. 5.5 Diagnostic plots for local least favourable distribution. Top: local least favourable expected loss plotted against the size C of the KL neighbourhood; bottom: difference between the minimax expected loss of each action and that of the optimal action a^*.

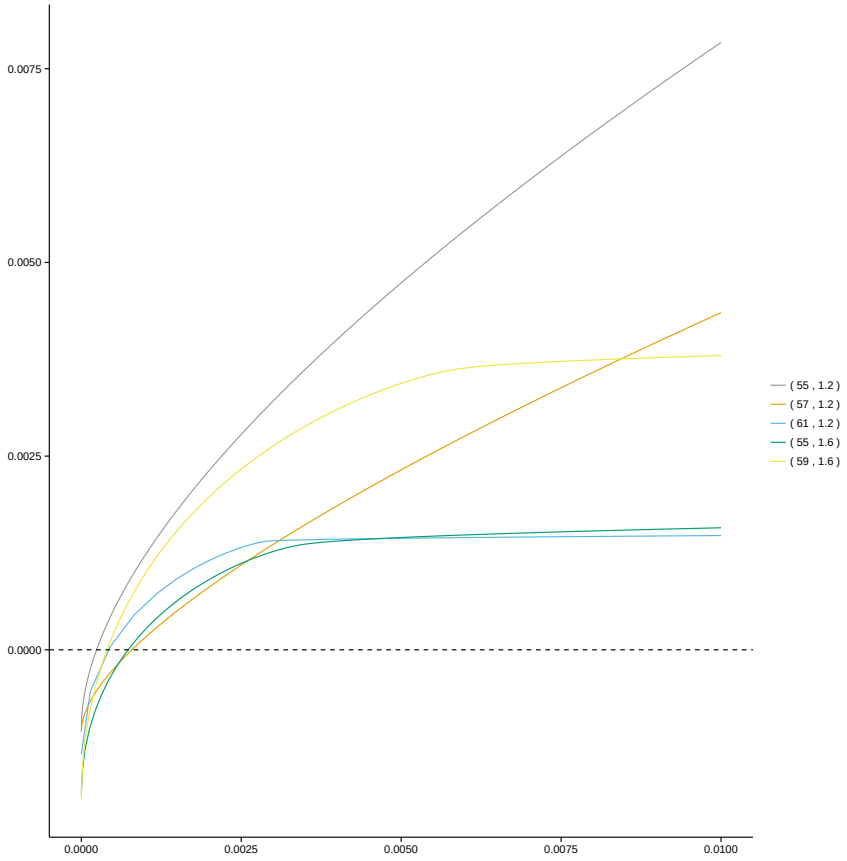

Fig. 5.6 Local Bayesian admissibility: local least favourable 'regret' loss between each action and the optimal Bayes action (59,15). This is a function of the KL divergence C (x-axis).

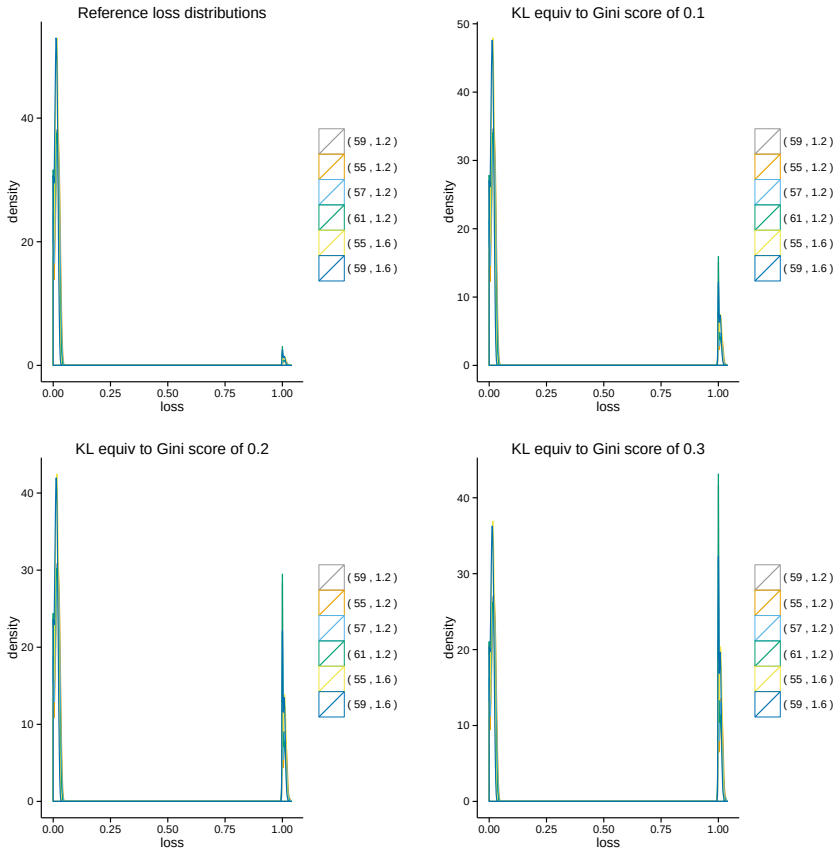

Fig. 5.7 Loss distribution under local least favourable distributions for three values of KL radius C. These are respectively: 0.1, 0.3 and 0.6.

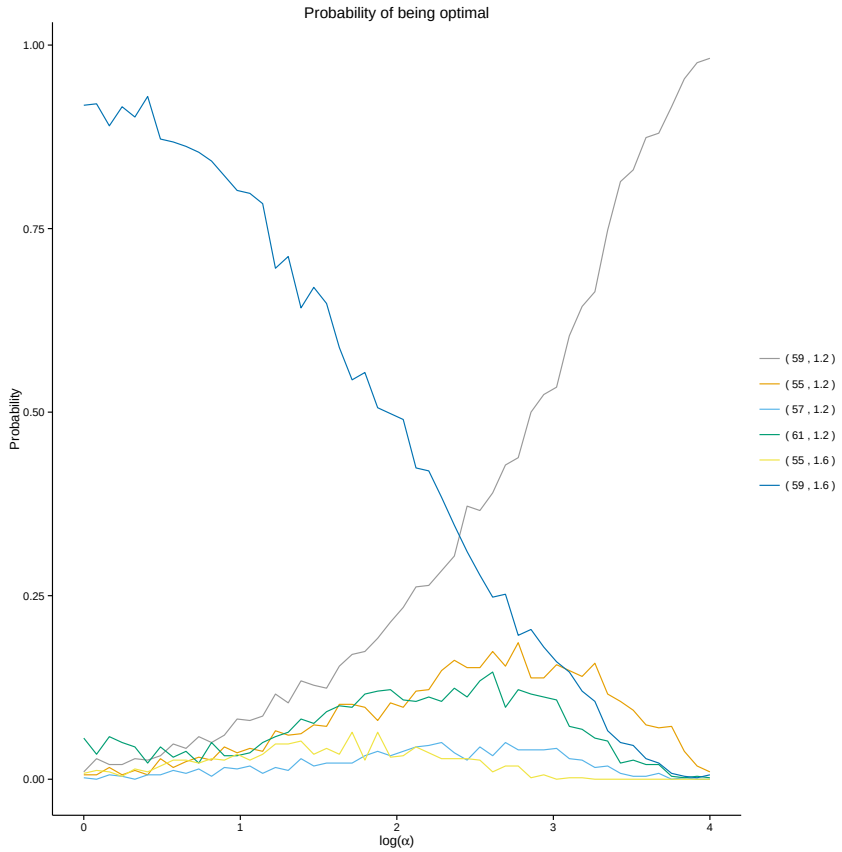

Fig. 5.8 Estimated probability of being optimal under expected loss in the Dirichlet extension model as a function of α for each action in the top 6 schedules selected. The α-values are plotted on a $\log_{10}$ scale.

However, for smaller values, the least optimal under π_I (of the 6 selected) has a higher probability of being optimal. Using figure 3.10 from section 3.3.3 we can see that these are fairly large values of $\log \alpha$. This shows the lack of decision robustness in this problem, mainly due to flatness of the loss surface.

5.2.4 Discussion

This application is based on a fairly simple model π_I but serves well to illustrate the use of the methods developed in earlier chapters. This chapter illustrates how the plots can be interpreted and how information about the robustness of the decision system can thus be extracted. It also highlights an interesting point regarding privacy issues in the context of reproducible research. From this analysis it is possible to conclude the following. Firstly, the loss surface is very flat for changes in screening schedules. This has also been noted by Ruggeri et al. (2005) in their analysis of the problem. This means that the optimal schedule is

highly sensitive to perturbations in the model. However, these perturbations do not drastically change the overall expected loss. This highlights an interesting distinction made by Kadane and Srinivasan (1994) when considering model misspecification in a decision-theoretic context between decision robustness and loss robustness. These are two separate properties of a decision system, one not implying the other. This particular application is robust to changes in the model (in an expected loss sense) but not however decision robust. I.e. small perturbations to the model will change the optimality of the Bayes action a^*.

5.3 Bayesian Variable Selection with Cost

5.3.1 Motivation and notation

For high-dimensional parameter spaces, one goal of statistical inference is to select an optimal subspace that reduces dimensionality but permits informative modelling. Regression is a standard example where the number p of potential covariates for the prediction of a response y can be large. In particular, I consider the case when data collection comes at a cost, i.e. it is an increasing function of the number of covariates included. Therefore it is necessary to find a minimal set of maximally informative variables. A decision-theoretic approach means this problem well defined.

Let $\mathscr{X} = \{x_j\}_{j=1}^p$ be the set of possible covariates. Then an action a is defined as a subset $\gamma^a = \{\gamma_i^a\}_{j=1}^p$ where $\gamma_j^a = 1$ if the j^{th} covariate is included and zero otherwise. γ^a is a possible model, and the decision task is one of model selection. The loss function defined over a model γ^a must trade accuracy of prediction against total cost of data collection:

$$L(\gamma^a) = L_p(\gamma^a) + C\sum_{j=1}^{p} c_j \cdot \gamma_j^a \tag{5.3}$$

This is the setting considered by Brown et al. (1999). Here, L_p is the loss incurred when predicting $\pi(y|X,\gamma^a)$, and c_j is the cost of including the j^{th} variable in the regression. This makes the assumption that the cost is constant with respect to the inclusion of each variable. C is a constant that represents the conversion of units of cost into units of expected predictive loss. The calibration of this parameter is an immediate issue, but this is context dependent. An additional sensitivity analysis should be performed with respect to the choice of this value.

5.3.2 RAND Quality of Hospital Care data

To illustrate model robustness under misspecification in this particular context, I looked at data collected from a large study by the RAND corporation on quality of hospital care (QoHC) in the US from the 1980s, Draper et al. (1990); Kahn et al. (1990). It is typically very expensive to directly assess the quality of care for a given hospital. However in theory it should be possible to approximate it by considering each hospital as a black box whose performance can be measured by looking at the inputs and outputs. For example, comparing mortality rates of different hospitals (outputs) based on sickness at admission (inputs). The problem is therefore dependent on selecting the most reliable measure of sickness at admission (disease specific) whilst also taking into account their relative costs of measurement. For example, recording the age of a new patient will be less expensive (cost in time for instance) than taking DNA samples or performing an ECG test. This is a process that should only really work for diseases where there is a strong predictive link between the state of the patient and their outcome. Therefore it is possible to compare expected mortality rates and observed mortality rates.

Previous work and formulation of problem

The RAND dataset consists of 83 covariates and a univariate response $y \in \{0,1\}$, which is 1 if the patient dies within 30 days of admission and zero otherwise. The sample is of size $n = 2532$, all elderly American patients hospitalised with pneumonia in the period between 1980-1986. Because of high redundancy in many of the predictors, the aim of the study was to select a minimal set of variables that would best predict the outcome y. The original analysis of the study by Kahn et al. (1990) used standard classical backward-selection methods to choose a parsimonious set of $p = 14$ variables with good predictive accuracy. But it did not take into account the difference in the cost of collection of the 83 variables, which is a key part of the problem (they included the most expensive variable in their analysis, the Total Apache II score). The cost is estimated in minutes (time needed on average to extract the corresponding measure from medical records) and varies from 0.5 to 10 with median cost 1.

A further analysis was done by Fouskakis and Draper (2008), who looked at principled methods for a trade-off between the cost of variable inclusion against loss in predictive accuracy. The authors use a classification loss for which it is necessary to specify a cut-off value for the predictive model and the loss values for each misclassification and correct classification. A cross-validation sample was then used to estimate the expected loss of the model which was constructed using Bayesian logistic regression. The paper's purpose was to compare the performance of a set of stochastic search methods, in particular, simulated

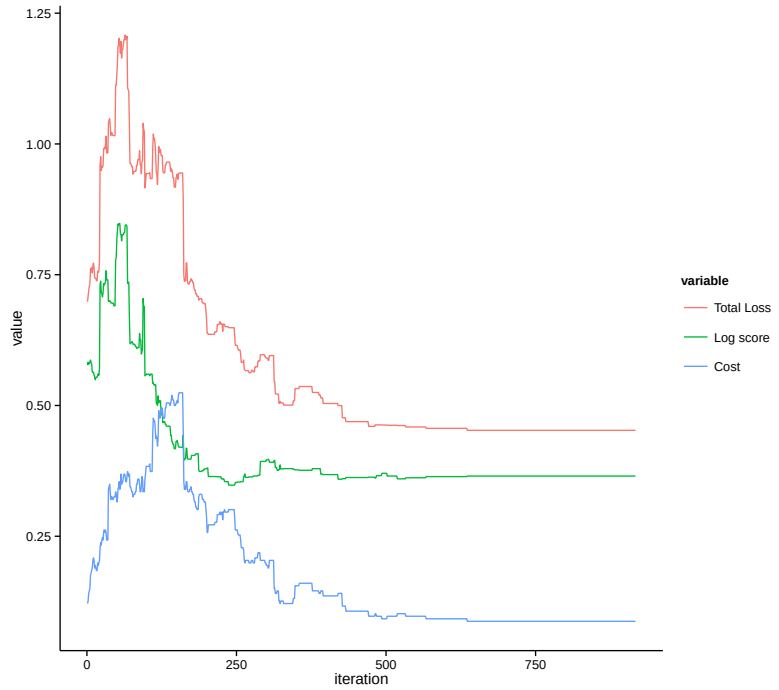

Fig. 5.9 Trace of expected loss (red) during a typical run of the vanilla simulated annealing search algorithm. The blue and green lines correspond to the two components of expected loss, respectively the negative logarithmic score and the fixed cost from variable inclusion (rescaled by $C = 1$ from eq 5.3), respectively.

annealing, genetic algorithms and Tabu search. In a further paper, Fouskakis et al. (2009) then looked at the use of modified priors that take into the account the varying costs. The subsequent posterior model odds then lead to a cost-adjusted version of the BIC method for variable selection. Interestingly they found much agreement between the decision-theoretic version and this cost-adjusted BIC.

Inspired by this previous work, I also take a decision-theoretic approach but using the loss function given in 5.3 which explicitly trades-off the benefit (predictive accuracy) and the cost (sum of the costs of each variable included in the model). Bayesian logistic regression was used to compute an approximate model $\pi_I(\theta)$. The predictive expected loss is chosen to be the negative logarithmic scoring rule for a given model γ_a, defined as:

$$L_p(r,i) = -\ln r_i$$

where r is a the vector of probabilities reported, when the i^{th} event happens. The true predictive distribution over outcomes given covariates x_i is unknown, but it is possible to use the full model $\pi_I(y|X)$ (built on all available covariates) as the best approximation to Nature.

This is the marginal of the logistic regression parameters, i.e.:

$$\pi_I(y|X) = \int_\beta \pi(y|\beta, X)\pi(\beta)\mathrm{d}\beta \tag{5.4}$$

where the integral is approximated using Monte Carlo sampling. This was done using the R package *BayesLogit* developed by Polson et al. (2013)[5]. Therefore the expected loss from diminished predictive accuracy using the reduced model $\pi_{\gamma_a}(y|X)$ is:

$$\begin{aligned} E_{\pi_I}[L_p(\gamma_a)] = &- \pi_I(y=1|X)\ln\pi_{\gamma_a}(y=1|X) &(5.5)\\ &- \pi_I(y=0|X)\ln\pi_{\gamma_a}(y=0|X)] &(5.6)\end{aligned}$$

Given a choice of model γ_a, under a negative logarithmic scoring rule, the expected loss under the least favourable distribution value of the decision is:

$$\max_{y'=\{0,1\}}\{-\ln\pi_{\gamma_a}(y'|X)\} = -\ln\left(\min_{y'=\{0,1\}}\{\pi_{\gamma_a}(y'|X)\}\right)$$

That is to say, the logarithmic score when the more unlikely event occurs. An intermediate least favourable distribution π^*_λ, using the notation from earlier chapters, for $\lambda \geq 0$ is:

$$\begin{aligned}\pi^{\sup}_a(y'|X) &\propto \pi_I(y'|X)e^{\lambda L_p(\gamma_a,y')}\\ &= \pi_I(y'|X)e^{-\lambda\ln(\pi_{\gamma_a}(y'|X))}\\ &= \pi_I(y'|X)\pi_{\gamma_a}(y'|X)^{-\lambda}\end{aligned}$$

It is not possible to evaluate this over future outcomes, but it can be approximated by bootstrapping X values from the empirical $\hat{F_X}$ and y's from the predictive $\pi_I(y|X)$. I.e. for $i = 1,..,n$, computing the value of $\pi^{\sup}_a(y_i|X_i)$ for a specific value of λ. The approximate Kullback-Leibler divergence of this distribution $\pi^{\sup}_a$ is then given by:

$$KL(\pi^{\sup}_a||\pi_I) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{y'=0,1}\frac{\pi_I(y'|X_i)\pi_{\gamma_a}(y'|X_i)}{Z}\log\frac{\pi_{\gamma_a}(y'|X_i)^{-\lambda}}{Z}\right)$$

where Z is the normalization constant for the distribution $\pi^{\sup}_a$ at (y_i, X_i). Note that it is not necessary to consider the loss incurred by cost of variable inclusion when computing the local least favourable distribution, as this is a constant value which is independent of the predictive model π_I. However it is used to compute the overall expected loss of the model γ_a.

[5] I used the default settings of the method *logit*, but drawing 1000 samples (samp=1000) and burn-in of 10000 samples (burn=10000).

Stochastic search in decision space

The space of possible models γ is of size $2^{83} \approx 9.7 \times 10^{24}$, which with modern computational resources is too large for brute force methods. In order to select a set of 'good' models, I implemented a vanilla simulated annealing search algorithm over the space $\{0,1\}^{83}$, starting with 10 random positions initialised at 1[6]. New proposals are generated by flipping the value of a random position with the acceptance rate for higher expected loss models following a geometric cooling regime. My main interest is to find a set of plausible solutions on which to test robustness properties under assumptions of model misspecification. I ran the algorithm 100 times, each time with a cooling schedule of 0.99, and for 1000 iterations. The starting vector was random with 10 non-zero entries. Figure 5.9 shows the trace of the expected loss of actions searched during a typical run using this implementation on the RAND dataset. The expected loss value (shown in red) is broken down into its two components, the contribution of the negative logarithmic score (shown in green) and the contribution from the fixed variable cost (shown in blue). The cost vector (units given in minutes) was normalised so that $\sum_{j=1}^{83} c_j = 1$. The constant C in equation (5.3) was chosen to be 1.

Figure 5.10 presents the results of these 100 experiments, alongside the variables selected by the methods in Fouskakis and Draper (2008) and original RAND paper Kahn et al. (1990). Note that for the 100 runs, there is a lot of crossover between the models selected, which would suggest that the algorithm is finding the actions nearby to the global maximum. In fact, only 44 unique models were discovered in these 100 runs. There is also some crossover between the variables selected here and those selected by previous work. This is shown in figure 5.10, where the first two columns correspond to variables selected by Kahn et al. (1990) and Fouskakis and Draper (2008) respectively.

5.3.3 *Ex post* analysis

Having found 44 candidate models (actions) from the 100 stochastic search experiments, it is possible to apply the machinery developed in Chapters 2-4. In order for the plots to be legible, I selected the top 8 models for the robustness analysis. These actions are given in figure 5.11. Each model has between 9 and 11 variables included. The term 'model' and 'action' are used interchangeably as each action is a model, unless there is confusion as to the whether it refers to the action γ^a or to the probabilistic model π_l.

[6] I chose to initialise 10 variables for the starting vector, as the results from Fouskakis and Draper (2008) indicate a peak in expected utility as a function of the number of predictors retained between 6-10 variables. This is reproduced in the results, with the all best models containing between 6 and 13 covariates, with the majority having between 9 and 11.

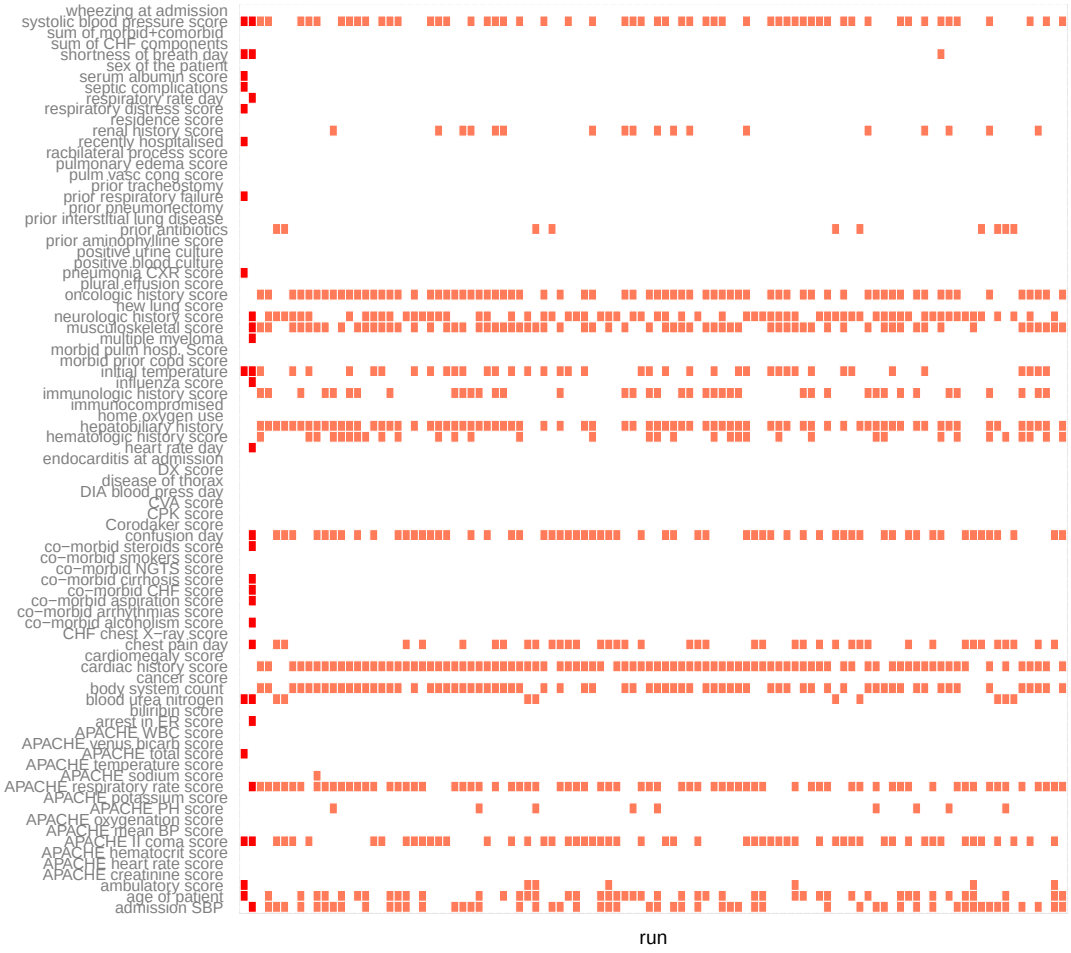

Fig. 5.10 Heatmap comparing results from 100 runs (columns 3 to 100) of the stochastic search algorithm and the variables included by the original RAND paper (Kahn et al., 1990, 1st column) and Fouskakis and Draper's decision theoretic algorithm (2008, 2nd column). The rows correspond to all the available variables. Each red square represents a variable (row) being included in the model (column). The colours in the first two columns are brighter for ease of comparison between our method and previous work.

Fig. 5.11 The variables included by the top 8 models (actions) from the 100 runs of stochastic search algorithm. The model indices correspond to their indices from the 100 stochastic search experiments, which I also use to name the models in the legends of figures 5.12-5.17

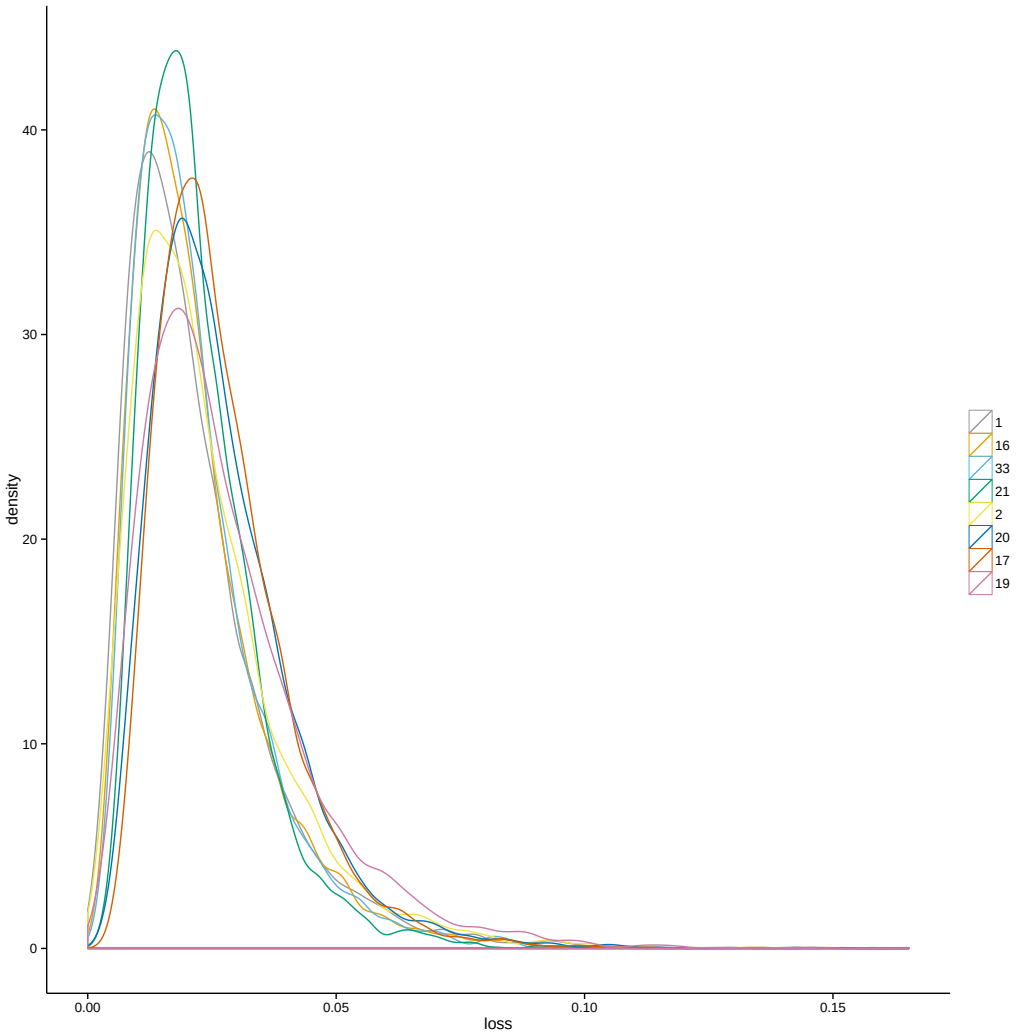

Fig. 5.12 Loss distributions of the top 8 models selected by the stochastic search algorithm. The legend gives their indices from the 100 runs of the algorithm with random starting points.

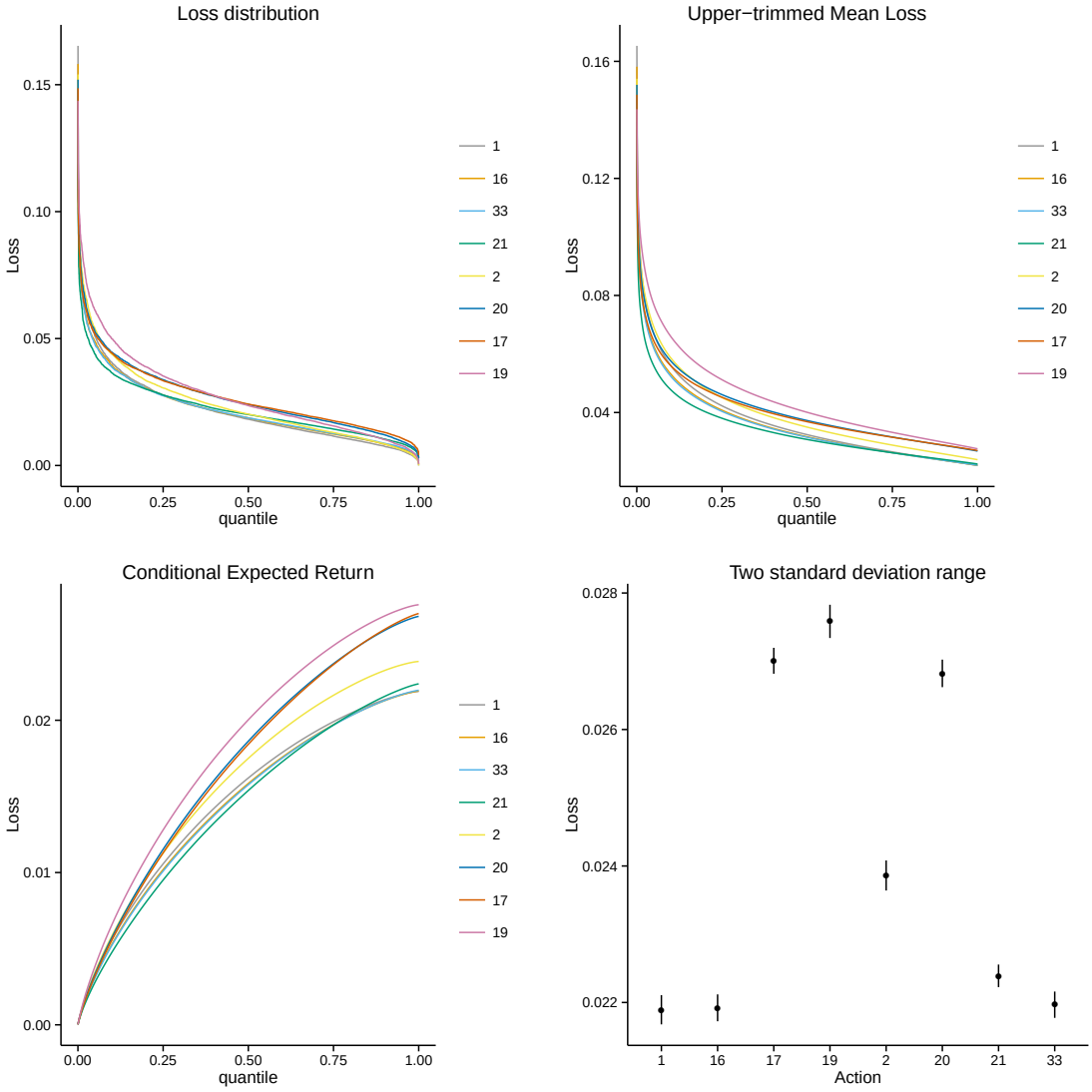

Fig. 5.13 Model diagnostic plots for top 8 actions.

Figure 5.12 shows the loss densities of these top 8 actions under the reference distribution π_I (which I recall is the full model with all variables included, given in equation 5.4). For this plot and all subsequent plots in this section, the action labels correspond to the experiment index in which the model was discovered. All the densities are fairly similar, some with more extreme losses in the tails. As a preliminary check one can look at the model diagnostic plots from section 5.4, Chapter 4. These are given in figure 5.13. The inverse loss distributions are very similar, as are the CVaR (upper-trimmed mean loss), but the CEL differentiates the actions from losses incurred the 50^{th} quantile and above. This is a consequence of the fact that the peaks in the loss densities happen at different values (see figure 5.12). Note crossing points between the top four actions around the 80^{th} quantile of the CEL plot, with action 21 optimal for $q \leq .8$.

It is interesting to see how the least favourable loss densities are distorted as the radius of the Kullback-Leibler ball is increased. This is shown in figure 5.14 and seems to rule out the KL radius of 0.26 that corresponds to a Gini coefficient of 0.3 on distribution of weights for the least favourable distribution.

Figure 5.15 (top plot) shows the expected loss under the local least favourable distribution of the top 8 actions found by the vanilla stochastic search algorithm as a function of the radius of the Kullback Leibler ball. The bottom plot shows the difference in this expected loss quantity between each action and the optimal Bayes action under π_I. This allows for easier comparison between the actions and observation of the crossing points. We see that actions '16' and '33' are highly sensitive to very small changes in the KL radius ('1' is suboptimal for very small values of the KL divergence). The model '33' is less sensitive to small changes but its local least favourable expected loss is more robust than the other actions for larger values of the KL radius. Indeed, it is the optimal action under the local minimax loss at a KL divergence of ≈ 0.02 (this corresponds to the loss densities shown in the bottom left plot of figure 5.14, for which the weight distribution of the importance sampled local minimax has a Gini coefficient of 0.2). This robustness confirms that the CEL diagnostic plot, which also showed '21' as optimal for the upper quantiles of loss less than 0.8 (see figure 5.13).

The final two plots, showing the regret loss (5.16) between each action and the optimal Bayes action '1', and the probability of optimality under the Dirichlet extension model (5.17) are included for completeness. The local admissibility plot in figure 5.16 confirms the sensitivity of the action '1' in relation to actions '16' and '33'. Figure 5.17 shows that the decision system does seem robust to symmetrical perturbations to the model.

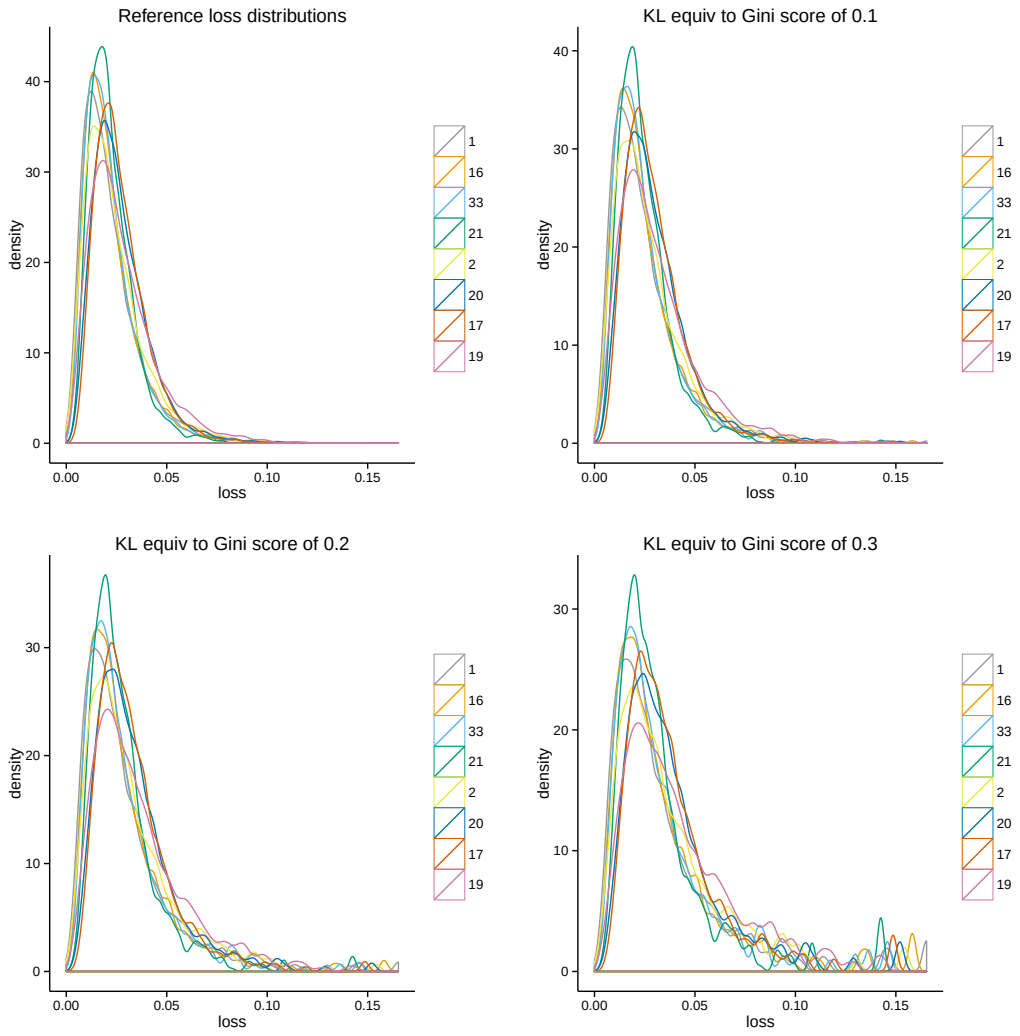

Fig. 5.14 Loss distributions of actions under the reference model and then under KL divergences of 0.02, .1 and 0.26. These correspond to Gini coefficients of .1, .2 and .3 respectively.

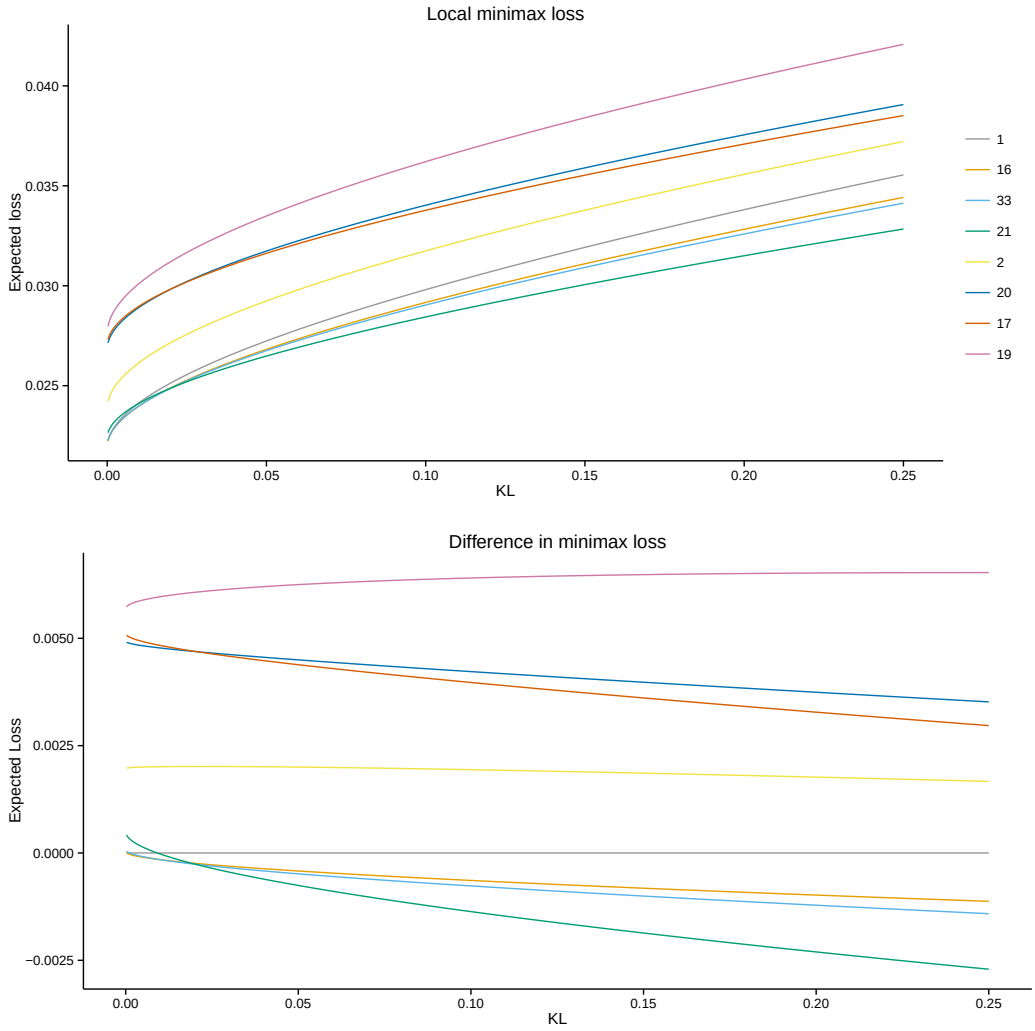

Fig. 5.15 Top: expected loss under the local least favourable distribution as a function of the KL radius C. Bottom: difference in expected loss under the local least favourable distribution between Bayes action (model 1) and the other 7 actions. This highlights the crossing points.

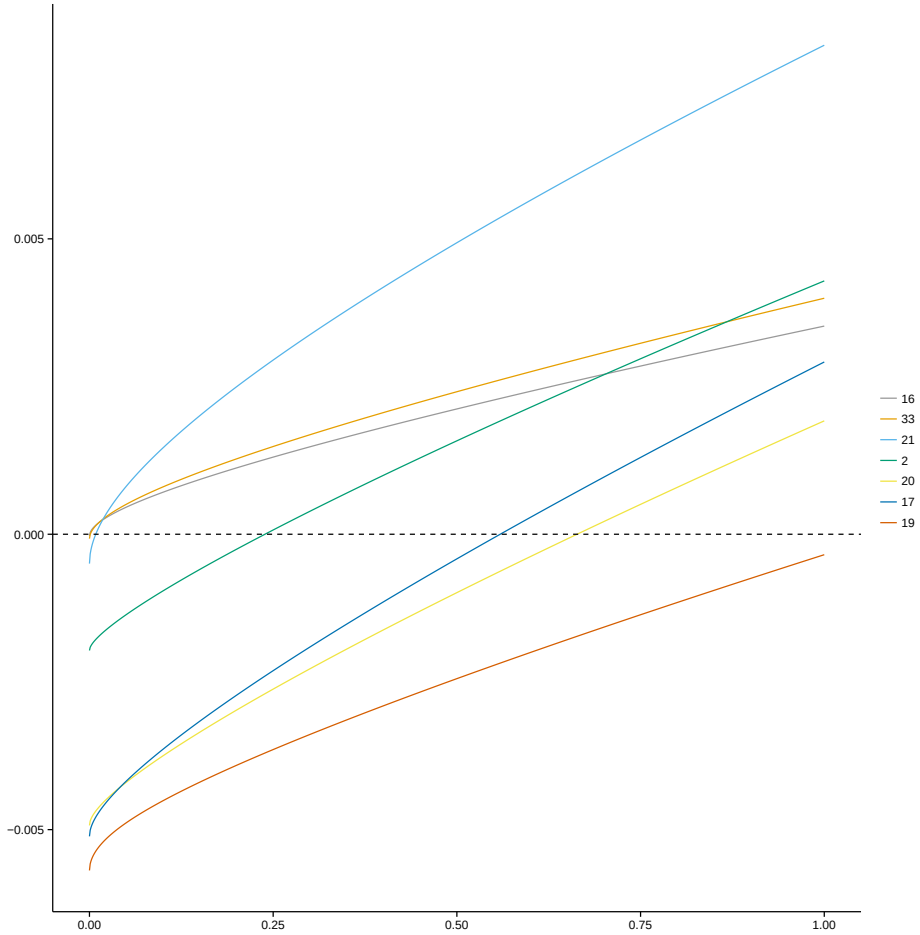

Fig. 5.16 Regret loss between actions and the optimal Bayes action (index 1).

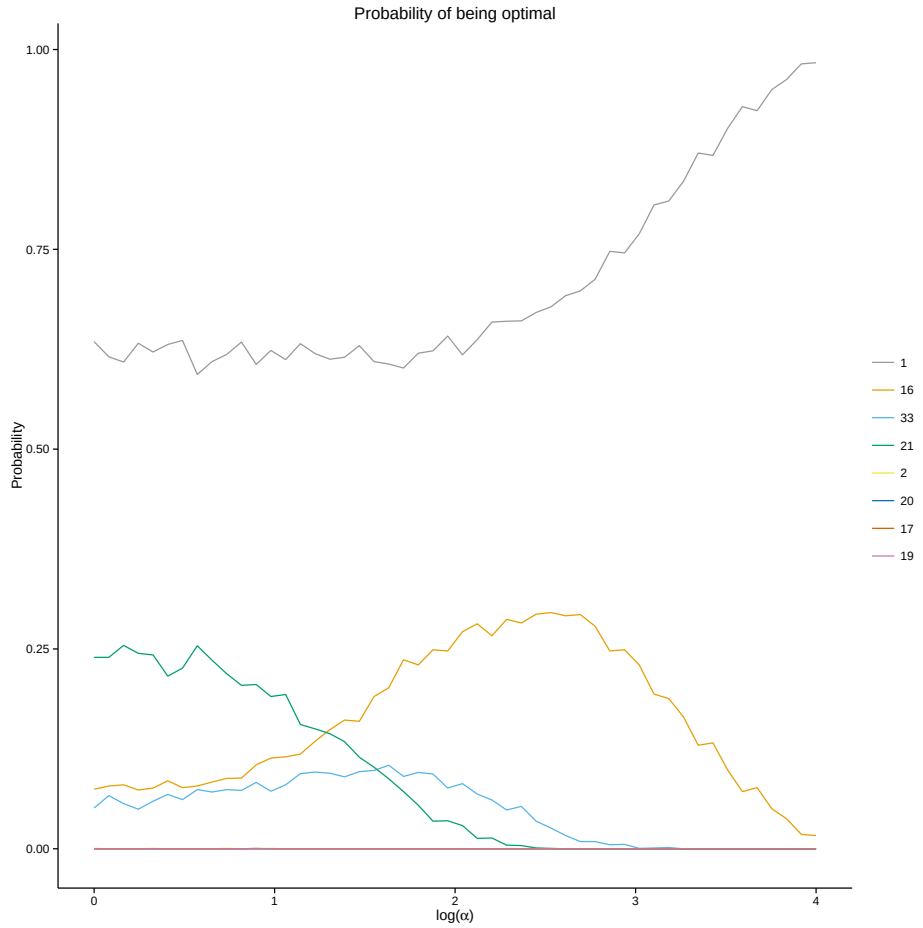

Fig. 5.17 Probability of optimality under a Dirichlet extension model as a function of the concentration parameter α.

5.3.4 Discussion of results

In this case study I considered how to formally analyse a variable selection problem with the added complication of variable inclusion cost. In an idealised situation, the statistician would use the full predictive model $\pi_I(y|x)$. However, because of practical considerations regarding the use of the model (in this case, the desire to minimise the cost of implementing a scheme for the evaluation of hospital quality of care) this is not possible. But the full model can be used here to test the performance of a predictive sub-model π_{γ_a}, in a very similar flavour to the work of Karabatsos (2006), whereby a 'correct' nonparametric Bayesian model is used to evaluate the accuracy of a 'wrong' parametric model. Within this $\mathscr{M}_{\text{completed}}$ framework, the methods from Chapters 2-4 can be illustrated to show how they can be used to analyse the sensitivity of a set of actions (here sub-models π_{γ_a}). This exercise shows that the optimal Bayes action found by the stochastic search algorithm is in fact not the most robust sub-model. The action '21' is slightly sub-optimal under π_I but performs much better than the other actions when one looks at robustness within KL neighbourhoods.

Chapter 6

Decision rules in randomised trials

6.1 Motivation

> "Can we identify the individual patient for whom one or the other of the treatments is the right answer?" Hill (1966)

6.1.1 Bypassing models

Statistical problems are usually broken down into three categories. Problems of pure inference, problems of prediction, and decision problems. As stated in the Introduction, all of these can in fact be posed as decision problems, for example, point estimation is a decision problem regarding which point to report, prediction is a decision problem regarding which predictive distribution to use, etc. However, for some statistical applications, the ultimate decision is more central to the issue at hand than for others. When it is difficult to posit a reasonable model π_I, it could be justified to bypass the model construction altogether and directly seek an 'optimal' (with respect to some chosen objective function) decision or decision rule. An interest in this idea was initially motivated by a seemingly paradoxical aspect of the standard Bayesian procedure (build a model π_I, choose a loss function L and then optimise to find the decision a^* minimising expected loss under π_I). If the ultimate decision a^* is the only element of interest - and we note that this is generally not the case, as the posterior inferences from π_I will often be scientifically meaningful, and more importantly falsifiable - then why waste time and effort in specifying all the uncertainty in π_I, to then deterministically choose a^* (there is no uncertainty attached to the choice of the Bayes action). Of course, there are deep foundational reasons for doing this, as given by Savage (1954), or Robert (2007) for a more recent exposition. However, we argue that there are cases

where difficulties in the construction of a model π_I make a more direct approach attractive[1]. The statement "model the object of interest" seems banal, but there are cases where a simpler target not only avoids concerns of model misspecification but also is in fact the actual target of interest. We focus on a problem in stratified medicine, in the context of the analysis of randomised clinical trials (RCTs). By simplifying the problem, working from first principles and directly targeting the object of interest, we present a powerful and robust method for the detection of patient heterogeneity in RCTs.

Up till now, this work has considered a set of actions $a \in \mathscr{A}$, which is combined with a loss function $L(\theta, a)$ and a distribution $\pi_I(\theta)$ to obtain a ranking of the actions by their expected loss $\int_\Theta L(\theta, a)\pi_I(\theta)\mathrm{d}\theta$. This chapter, however, considers decision rules, which map patients (i.e. covariate space) to treatments (action space).

6.1.2 Stratified medicine

In this chapter, we focus on the specific example of randomised (clinical) trials, where two or more interventions are compared and the allocation of the intervention is made at random. The main application of randomised trials is in clinical medicine where they represent the gold standard for reporting causal inferences, for example, whether or not a specific drug helps to cure a disease. Subgroup analysis, or the classification of patients into different treatment groups by baseline characteristics, is a major area of research in the analysis of randomised trials. In an era where more and more genetic data is recorded and analysed, there is much hope that treatment allocation to each patients will become more and more personalised. That is to say, individual covariates, such as age, sex, genetic markers etc, will determine the choice of treatment. This is already the case for certain diseases, such as some cancers, whereby patients are stratified by genetic markers (see for example van't Veer and Bernards, 2008).

This work is motivated by the following argument. The development of stratified medicine should happen in two stages. First, identifying whether for a particular drug/treatment there is evidence of patient heterogeneity. This is in the context of the analysis of clinical trials which were not designed to test a priori subgroup hypotheses. Second, the development of clinical trials and statistical methods which allow for the exact characterisation of subgroups. We argue that the first stage is a problem of *hypothesis testing*, the second stage is a problem of *design* and *prediction*. The existing statistical methodologies (which we review in part in section 6.2) which deal with subgroup analysis focus on the second stage, i.e. the development of methods for clinical trial design and predictive models which determine optimal treatment

[1] These could be computational difficulties, or concerns of misspecification because of the lack of understanding of the underlying processes being modelled.

subgroups. However, for the analysis of RCTs which have not been specifically designed to test suspected subgroups, this is the wrong approach. By posing the analysis as a testing problem instead of a prediction problem, we attempt to solve an easier problem but whose conclusions are more robust. The robustness stems from the fact that we need to make fewer assumptions regarding the data as compared to a predictive model. These ideas can be used at an earlier stage of research into patient heterogeneity.

6.2 Literature review

6.2.1 Notation

Before reviewing the existing methods from the literature, we first layout our notation that we will use for the rest of this chapter. This allows for a more rigorous examination of some of the methods using a consistent notation throughout.

The setting we consider is as follows. The data are in the form of a triple (X_i, Y_i, T_i). $X_i \in \mathbb{R}^p$ is the i$^{\text{th}}$ patient's covariates; $Y_i \in \mathbb{R}$ is a real-valued observed outcome or response (we assume this is continuous and higher values are more desirable); and there are two randomised treatment arms, $T_i \in \{-1, 1\}$, with $\pi(T_i)$ the probability of receiving treatment T_i, this is identical for each patient. T is a random variable, which is independent of X and Y. In what follows, we will denote $R_i \in \{1, .., n\}$ the rank of the observed response Y_i (using the convention that the greatest value of Y_i has rank 1). We define a decision rule as a function $D : \mathbb{R}^p \to \{-1, 1\}$, a mapping from the patient's covariates to a treatment assignment.

In our notation, Y and T represent the outcome and the treatment assignment under randomisation (i.e. T is a random variable, and Y is a random variable dependent on T). Y^D and T^D represent the outcome and treatment assignment under a decision rule D, hence T^D is no longer a random variable but a deterministic function of X.
Y is in fact a function of T and X, so we could write $Y(T, X)$, which under the decision rule D, is $Y^D(T^D, X)$. For ease of exposition we drop these dependencies.

6.2.2 Difficulties of subgroup analysis

From a clinical perspective, subgroup analysis is a controversial topic. Sun et al. (2012) did a systematic review of all randomised clinical trials published during the year 2007 (a total of 407). About half of these trials (207) reported subgroup effects and very few met the criteria of 'credibility' set out by the authors[2]. They set out 10 criteria to gauge the credibility of

[2]Of 64 making claims about the primary outcome, 54 met four or fewer of the 10 criteria.

subgroup claims. It is important to note that these subgroups analyses are usually based on a regression model which tests whether there is an interaction between the treatment and the covariates. The criteria are intended to reduce the false positive rate and render the analyses more credible. They are as follows (reproduced from Sun et al.)

- Design:
  - Was the subgroup variable a baseline characteristic?
  - Was the subgroup variable a stratification factor at randomisation?
  - Was the subgroup hypothesis specified a priori?
  - Was the subgroup analysis one of a small number of subgroup hypotheses tested (≤ 5)?
- Analysis
  - Was the test of interaction significant (interaction $P < 0.05$)?
  - Was the significant interaction effect independent, if there were multiple significant interactions?
- Context
  - Was the direction of subgroup effect correctly prespecified?
  - Was the subgroup effect consistent with evidence from previous related studies?
  - Was the subgroup effect consistent across related outcomes?
  - Was there any indirect evidence to support the apparent subgroup effect - for example, biological rationale, laboratory tests, animal studies?

These are in part inspired by Rothwell (2005), a comprehensive study of how subgroup analyses should be interpreted, and when such recommendations should be taken into the clinic and be translated into the standard of care. Sun et al. (2014) gives a more hands-on guide as to the interpretation of subgroup analyses, along with many examples of previous mistakes. Many studies report differences within groups (i.e. one group has a stronger effect than the other), but do not test the strength of the interaction between the groups and the treatment. It is much more likely to see a subgroup effect by only looking at the difference of effect within groups. Exaggerated claims of the presence of clinically significant subgroups has lead to well known mistakes, such as aspirin being ineffective in secondary prevention of stroke in women (Anastos et al., 1991). A major problem concerns the correct control of the false positive rate. Vaguely pre-defined hypotheses lead to what Andrew Gelman calls “the

garden of forking paths" (see Gelman and Loken, 2013)[3]. This is a problem of not properly controlling multiple testing.

The consensus agreement is that for a subgroup analysis to be credible, it should be pre-defined before the study, along with its direction and preferably its strength. The main problem with subgroup analyses in much of the literature are that they are usually underpowered (the study was not designed to detect the subgroup effect), thus carrying a high false positive rate. Also, many trial populations are unrepresentative of the overall population of interest. For example, women, children, and elderly people are often under-represented. This makes any generalisation of the findings difficult. Rothwell (2005) recommends stratification of the randomisation at the level of the suspected subgroups when they are posited a priori (this ensures equal representation between different treatment groups). The fact that a post hoc finding can be justified by a biological rationale does not help much in confirming the finding either. In the words of Douglas Altman, "[..] biological plausibility is the weakest reason, as doctors seem able to find a biologically plausible explanation for any finding" (Altman, 1998)[4].

For these reasons, subgroup analysis and the discovery of optimal treatment allocation rules are controversial. However, there are definite cases where subgroups of patients exist. In the words of Peter Rothwell (Rothwell, 2005):

> Differences between groups of patients in underlying pathology, biology, or genetics can lead to clinically important heterogeneity of treatment effects. Examples will probably be identified more frequently as our understanding of the molecular mechanisms of disease is enhanced.

The increased availability of genetic data has given this problem greater importance, making the promise of personalised medicine very real. However, it is clear that the problem should be approached with caution. In this chapter, we look at ways to discover whether there is evidence for patient heterogeneity in a randomised clinical trial. Its purpose is not to design decision rules which can directly be brought into the clinic (for reasons discussed above and by Rothwell), but instead to serve as a hypothesis generation procedure which can then be used to inform the development of further trials.

[3]Gelman and Loken are not referring to cases of deliberate dishonesty where the p-values are 'hacked' by under-reporting multiple testing, but rather to situations where vague pre-defined hypotheses lead to multiple testing, and thus the methods are contingent on the data. This a problem of statistical rigour, not malpractice.

[4]This suggests a process similar to that used in the justice system, whereby the suspect is lined-up alongside known innocents. A decision rule (or subgroup) discovered post hoc from the data could be 'lined-up' with 9 other decision rules discovered from the permuted data (permute the response to covariates) which are thus known to be 'innocent'. If it was deemed most guilty of being a plausible suspect then this would give it more credibility. One would need to see whether clinicians would embrace this idea of a 'hypothesis line-up'.

6.2.3 Predictive methods

We briefly review some current methods for exploratory subgroup analyses which fall into a predictive framework. The standard approach would be as follows. Construct a model $\pi(Y|X,T)$, where Y is the response, X are the covariates, and T is the (randomised) treatment. Subgroups can then be formed by maximising the predicted response over possible treatments T. This approach necessitates the construction of a model with an interaction term between the treatment assignment and the covariates X. This can also be framed as a hypothesis test, where the coefficient of the interaction term is zero. The immediate limitations of this approach is that the form of the model may be hard to justify because of limited understanding of the relationship between response, treatment and covariates. Also the interaction terms make the model high dimensional, which in a hypothesis test framework means there will be a difficulty in controlling the false discovery rate. However, our main argument against this approach is that the predictive model is a *proxy* tool for the test of evidence in favour of subgroups.

One solution to this problem is given by Qian and Murphy (2011). They consider *individualized treatment rules*, which are decision rules mapping patient covariates to treatment assignments thus defining patient subgroups. The problem of high dimensionality and variable selection is solved via a two step procedure. First, they estimate the conditional mean response for each treatment using a linear model with L_1-penalised least squares. The optimal decision rule is then derived from this conditional mean model. This two step procedure is more robust to misspecification. They also derive upper bounds on the difference between the true optimal decision rule and the estimated optimal decision rule.

Another popular method (at the time of writing this has had 50 citations in 3 years) comes from Foster et al. (2011). They look at predictive models of the form (using their notation):

$$P_{ij} = \mathbb{P}(Y_i = 1|T_i = j, X_i)$$

for $j = 0,1$ and a binary response variable Y (i indexes the patients). The novelty of this approach is to use random forests to estimate these predictive quantities. This is done by inputing to the random forest the quantities $\{X_i, T_i, X_i\mathbb{1}(T_i = 1), X_i\mathbb{1}(T_i = 0)\}$ and outputting an estimate of $\mathbb{P}(Y_i = 1)$. This is then used to predict or classify for new observations the quantity: $Z_i = P_{i1} - P_{i0}$. This quantity is estimated from the data, as for each patient, we only observe one of $\{P_{i1}, P_{i0}\}$ (the other is a counterfactual). For reasons which will be elaborated later on, it is important to note that this method uses a random forest algorithm in an entirely conventional fashion, where the decision trees are the backbone to the predictive model. One limitation is that it does not easily extend to non-binary response data. However, it is

an interesting model which although may be robust (via the random forest construction) is aimed a fitting a predictive model to estimate subgroups within the data.

Foster et al. (2011); Ruberg et al. (2010) both acknowledge the need for pre-defined subgroups in order for a subgroup analysis to be worth translating into clinical guidelines. However, Foster et al. justify a post hoc exploratory analysis by arguing that the statistical method should be pre-defined instead of the subgroups:

> An alternative strategy to predefining the subgroups is to predefine the statistical approach that is going to be used to find subgroups.

Although we do see the benefit of having easy-to-replicate analyses, this comment is beside the point. Any method that attempts a post hoc analysis must be able to explicitly define what the false discovery rate will be, or the what the evidence in favour of the subgroups is, for example using permutation tests which can take into account multiple testing or overfitting. Foster et al. go on to say:

> Although in principle the methods we present could be viewed as providing definitive evidence for a subgroup, in practice the methods are more useful for giving leads and suggestions for future work and better than what one could achieve by simply looking for interactions. In reality, an actual new trial would be required for the results to be confirmed and accepted.

We agree with this point, and our method is very clearly designed to be used in an exploratory manner. As argued above, we believe stratified medicine should go through two stages, and our method tackles the initial stage of testing whether there is evidence for patient heterogeneity to treatment. For this reason, we believe that predictive models such as those given by Foster et al. (2011); Matsouaka et al. (2014); Murphy (2003); Qian and Murphy (2011); Ruberg et al. (2010) may be good for second stage evaluation of pre-defined hypotheses, but are not well suited for early stage hypothesis generation. By attempting to construct a predictive model $\pi(Y|X,T)$, we argue that in general, these methods are attempting to solve a harder problem than is actually necessary.

6.2.4 Optimising the decision surface

Zhao et al. (2012) take a different approach to the problem. They do not attempt to construct a predictive model $\pi(Y|X,T)$ but instead directly attempt to optimise a function $f(X,T)$, which minimises the loss (or maximises utility) at every X. This function f labels the optimal action, and is thus a proxy for a decision rule D_f. In this way, they bypass the construction of a predictive model and directly optimise the decision surface, i.e. find the optimal subgroups

in the data. This method, called outcome weighted learning (OWL), has similar features to ours, so we outline the ideas in some detail.

It is motivated by a concern of misspecification, in particular the difficulty of constructing a parametric or semiparametric model of the response. Instead of treating the problem as a regression or prediction problem, they look at it as a classification problem (with two possible treatment choices). This comes from writing the expected response with respect to the joint probability distribution induced by the decision rule D. If $\mathbb{P}$ is the joint probability distribution over $\{Y,X,T\}$, and the $\mathbb{P}_D$ the joint over $\{Y,X,T^D\}$ (induced by the decision rule D), then it is easy to see that:

$$E_{\mathbb{P}_D}[Y] = \int Y \mathrm{d}\mathbb{P}_D = \int Y \frac{\mathrm{d}\mathbb{P}_D}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{P} = E_{\mathbb{P}}\left[\frac{\mathbb{1}(T = D(X))}{\pi(T)} Y\right] \tag{6.1}$$

where $\pi(T)$ is the probability of the random assignment T being either 1 or -1.

In this way the expectation of the response under the decision rule D can be estimated using the empirical distribution of $\mathbb{P}$. Finding the optimal decision rule D^* (i.e. maximising 6.1) is shown to be equivalent to minimising the following using the empirical distribution of X (equation 2.1 in Zhao et al.):

$$D^* = \arg\min_D \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{\pi(T_i)} \mathbb{1}(T_i \neq D(X_i)) \tag{6.2}$$

To simplify the problem, $D(X_i)$ is expressed as the sign of a smooth decision function $f(X_i)$ chosen from some family of decision functions. The function f is a mapping $f : X \to \mathbb{R}$, where $f(X) > 0$ means that $D(X) = 1$, and $f(X) < 0$ means that $D(X) = -1$. f would be chosen from a family of mappings $\mathscr{F}$ over which one would solve the optimisation problem of equation 6.2. Because of the non-convexity of 6.2, it is easier to introduce a convex surrogate loss, for example the hinge loss, along with a complexity penalty on f. This gives the following optimisation problem over functions f (equation 2.2 in Zhao et al.):

$$f^* = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{\pi(T_i)} [1 - T_i f(X_i)]^+ + \lambda_n ||f||^2$$

where $x^+ = \max(x, 0)$, and $||\cdot||$ is some norm for f. λ_n is a penalty parameter on the complexity of the decision rule f. This is similar to a weighted support vector machine (see Bishop, 2006, Chapter 7).

This method therefore only uses (on average) half the data, i.e. the data that do not agree with the decision function f. It also uses a potentially different dataset to score each decision function $f \in \mathscr{F}$. From our perspective, this method suffers a few disadvantages.

Firstly, it is not intuitive as to why the algorithm is only using half the data to find an optimal D^*. Secondly, there is a necessity to properly tune the method to avoid overfitting (a trivial solution to 6.2 is to assign every X_i the opposite treatment from the observed T_i). Last of all, there is no intuition as to what the randomisation adds in the discovery of a correct decision rule D (at least it is not explicitly stated in the paper). We also note that the main novelty of this approach is to take the quantity given in equation 6.1, and instead of maximising it using a sparse linear model as have done Qian and Murphy (2011) previously, they directly optimise it in a nonparametric way (which leads to a weighted SVM).

Building on this method, we look at a more intuitive approach that clearly makes use of the randomisation and that also has the advantages of being model-free.

6.3 Testing decision rules

As stated above, a preliminary subgroup analysis should not be approached from a prediction perspective unless there is a strong understanding of the relationship between the response, the covariates and treatment. On the other hand, we advocate thinking of the problem in a testing context. Instead of attempting to optimise some objective function, such as 6.2, over a family of decision rules, we posit the following null hypothesis: the decision rule performs (on average) no better than random allocation of treatment. We show that this is equivalent to testing the decision rule against its mirror decision rule. In this way it is possible to leverage well known tests, suitable for the data at hand.

6.3.1 Decision rule versus mirror

We recall that Y is the outcome under randomisation, and Y^D the outcome under the decision rule D. The *probability of treatment benefit* under D is:

$$\begin{aligned}\alpha_D &:= P(Y^D > Y) = \int_X P(Y^D > Y|X)\pi(\mathrm{d}X) \\ &\approx \sum_{i=1}^{n} \frac{1}{n} P(Y_i^D > Y_i|X_i)\end{aligned}$$

estimated using the empirical distribution of X. This of course cannot directly be estimated from the data, because we do not have the values Y_i^D when $D(X_i) \neq T_i$. If Ω represents the set of decision rules, the goal is to find:

$$D^*(X) := \arg\max_{D\in\Omega} [\alpha_D]$$

For a particular decision rule D, the null hypothesis can be stated in the following manner:

$$H_0 := P(Y^D > Y) \leq 0.5$$

That is to say, half the time, patient response to treatment which was assigned using the decision rule D is no better than if the treatment had been assigned at random.

In the same way, we recall that the treatment under randomisation is denoted T. Conditioning on the decision rule being the same or different than randomisation:

$$\begin{aligned} P(Y^D > Y) =& P\{Y^D > Y | T = D(X)\} P\{T = D(X)\} + \\ & P\{Y^D > Y | T \neq D(X)\} P\{T \neq D(X)\} \\ =& \frac{1}{4} + \frac{1}{2} P\{Y^D > Y^{\tilde{D}}\} \end{aligned}$$

where $\tilde{D}$ is the *mirror decision rule* of D (i.e. always prescribes the opposite treatment).

Thus, in order to test H_0, we only need to test $D(X)$ versus its mirror decision rule $\tilde{D}(X)$. This motivates the use of a testing procedure (model free or not) that compares D and $\tilde{D}$. For example, under H_0, the sum S_D of the ranks of the data points conforming to $D(X)$ are distributed according the Mann-Whitney-U statistic with parameters (n, n_D).

$$S_D = \sum_{i=1}^{n} R_i \mathbb{1}\{T_i = D(X_i)\} \tag{6.3}$$

$$n_D = \sum_{i=1}^{n} \mathbb{1}\{T_i = D(X_i)\} \tag{6.4}$$

where R_i is the rank of the response Y_i of the i^{th} patient.

It is possible to derive an analogous result using expectations, where the null hypothesis is

$$H_0 := E[Y^D - Y] = 0$$

In the same way:

$$\begin{aligned} E[Y^D - Y] =& E[Y^D - Y | T = D(X)] P\{T = D(X)\} + \\ & E[Y^D - Y | T \neq D(X)] P\{T \neq D(X)\} \\ =& \frac{1}{2} E[Y^D - Y^{\tilde{D}}] \end{aligned}$$

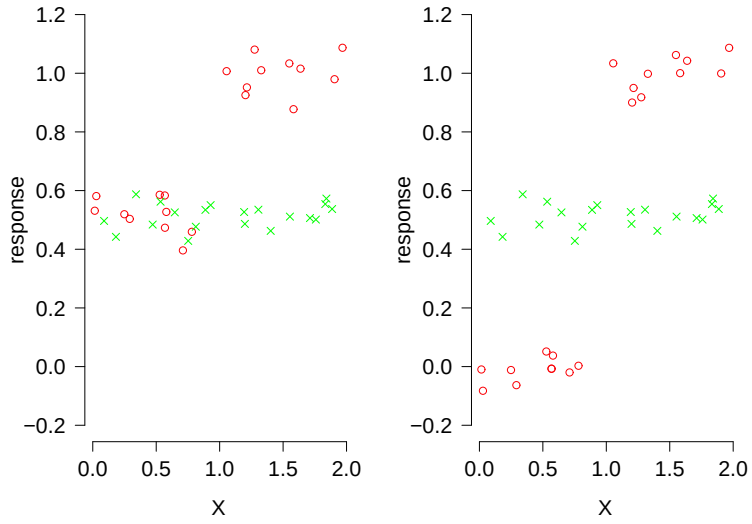

Fig. 6.1 Toy illustration: two scenarios where patient covariates are univariate, drawn from Uniform$[0,2]$. Left-hand side: $T=-1$ (red circles) is better than $T=1$ (green crosses) for $X \in [1,2]$ and equal for $X \in [0,1]$. Right-hand side: $T=-1$ is worse than $T=1$ for $X \in [0,1]$ and better for $X \in [1,2]$.

Therefore, if the assumption of normality is acceptable, or the dataset is large enough for the Central Limit Theorem to apply, then the student t-test is applicable.

Extension to survival data In many clinical trials, the outcome (response Y) is a survival time, which is very often left-truncated and/or right censored. For example, the entry into the trial could be contingent on a previous operation (left truncation), and the outcome could be measured by survival up to a pre-determined period (right censoring). In this case, the response would contain the truncation time, the censoring indicator and survival time. However, our approach very naturally extends to data of this type, by selecting a suitable test. For example, the using the log-rank test.

Non unique decision rules We briefly illustrate the ideas above with an example that highlights a non-intuitive aspect of our methodology. This can appear to be a problem with the method but is in fact a particularity of the setting. Consider the two simple examples shown in figure 6.1. This shows two scenarios where the response is a function of a random treatment drawn from $\{-1,1\}$ and a covariate which is a uniform random variable between 0 and 2. The response is normally distributed, with its mean dependent on X and T. Each point on the plot is a simulated patient and red circles correspond to those assigned $T=-1$

and green crosses those assigned $T = 1$. For the first scenario (left-hand plot), any decision rule which assigns $T = -1$, when $X > 1$, is optimal regardless which treatment is assigned for $X < 1$. On the other hand, for the second scenario, only the decision rule which assigns treatment -1 for $X > 1$ and treatment 1 for $X < 1$ is optimal. This difference between these two example points to an important aspect of the testing approach that we develop here.

As further illustration, figure 6.2 shows a proposed split at $X = 1/2$. This provides a visualisation of the splitting and testing procedure. The decision rule corresponding to this split would assign treatment -1 for $X \leq 1/2$ and treatment 1 otherwise. The test is performed on the observed X_i whose corresponding T_i agree with $D(X_i)$ (blue circles) versus those which disagree (black circles). A split at $X = 1/2$ does not completely separate out the blue and black circles in a way that one is always at least as good as the other. Indeed, we see that only a split at $X = 1$ would completely separate the blue and black points, thus maximising the test statistic.

The purpose of these two examples is to show that optimal decision rules are not unique. If the model fitted to $Y|X,T$ was looking for an interaction between components of X and the treatment assignment, then in the case of the first scenario (figure 6.1, left plot), the interaction would be detected for $X > 1$, thus providing a decision rule which splits at 1. However, in our testing framework, any split for $X < 1$ will give exactly the same test statistic. In this way, it is possible to determine subgroups where there is (or not) a difference in treatment effect.

6.3.2 Random forests for testing

The testing approach described above provides a method for estimating the evidence from data in favour of a particular decision rule. However, finding decision rules in a post hoc exploratory context necessitates searching over a tractable family of decision rules. The random forest algorithm is particularly well suited for this purpose.

Random forests (Breiman, 2001) are a popular ensemble learning method for both *classification* and *regression* tasks. They proceed by constructing a set of M decision trees and then averaging over the vote of each tree in the set to classify or predict a new observation. This use of a committee of trees is a form of *bagging* (short for bootstrap aggregation), whereby the trees in the committee have low pairwise correlation and each tree has low bias. Low pairwise correlation between trees is achieved by both bootstrapping the data and subsampling the set of predictors at each tree node. Thus, overall, the forest has low bias, but also low variance. Bootstrap aggregation procedures in general, and random forests in particular, perform well when each individual component has high variance and low bias. The aggregation allows the overall variance to be low, all the while maintaining the low bias.

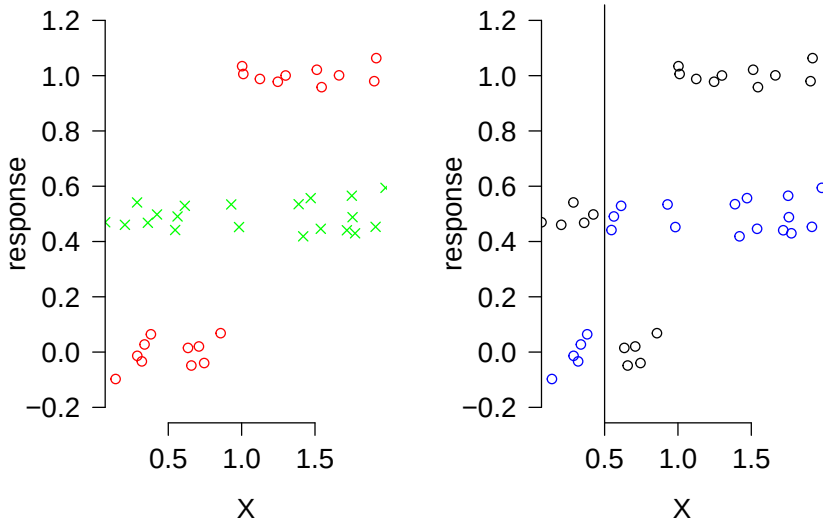

Fig. 6.2 Testing the decision rule against its mirror on a proposed split at $X = 1/2$. Left: responses under treatment -1 (green crosses) and 1 (red circles); Right: the decision rule gives the assignments in blue, its mirror the assignments in black.

Empirically, random forests do very well at prediction and classification tasks (at the time of writing the original paper has more than 16000 citations on google scholar). Other advantages of the method include: few tuning parameters, fast computation and interpretable results via measures such as variable importance and out-of-bag prediction (using the non-bootstrapped samples to cross-validate each decision tree in the forest). For a more in-depth overview of the random forest algorithm with precise details of its implementation, I refer the reader to Chapter 15 of Hastie et al. (2009).

Decision trees, and by extension ensembles of decision trees, constitute a tractable family of decision rules. However, it is not possible to apply the conventional random forest algorithm in this context, as our problem does not fall into a classification framework (there are no labels as we do not know the counterfactual response $\bar{Y}_i$: the response under the treatment $\bar{T}_i$) nor a regression framework. Instead we apply a random forest-type algorithm in a *testing* context. In particular, testing a decision rule D versus its mirror decision rule. To emphasise this point, it is not possible to use the *R* package *randomForest* as it is not designed to run our algorithm. We note that the method given by Foster et al. (2011) also uses random forests. As discussed in section 6.2.3, however, this is in the conventional setting, where they are used for regression or classification with either T_i or $X_i * T_i$ included in the set of predictor variables.

The use of random forests for testing is not novel, indeed it has already been extensively developed in the context of nonparametric tests for survival analysis by Ishwaran et al. (2008) under the name *survival forests*. The log-rank test underpins their methodology for choosing optimal splits on the data. Our approach differs from Ishwaran et al. as it applies specifically to multi-arm randomised data and is designed to incorporate any test, not only survival tests. Testing a split on the data is different to testing a decision rule against its mirror via a split on the data. It is a simple idea, but powerful when applied to randomised trial data in order to detect potential patient heterogeneity. Another use of decision trees is in Su et al. (2009), who construct them by scoring splits using an interaction test between treatment variable and the two subgroups. This again falls into a prediction framework.

Algorithm

The algorithm follows the standard construction of the random forest algorithm, differing only in how the optimal split at each node in each decision tree is chosen. In order to reduce the pairwise correlation between trees, we bootstrap the data with replacement for the construction of each tree. The remaining data can be then used to cross-validate the decision tree. At each level of the tree, the algorithm selects a random subset of covariates (the number of these is discussed in the next section). For each of these covariates, every 'allowed' split is proposed. For robustness purposes, and to avoid overfitting, the allowed splits are defined as the quintiles of the covariate (this is for continuous covariates, for discrete ones, we assume they are ordinal and propose splits at each level). Each split defines a decision rule which assigns treatment -1 to the left of the split, and treatment 1 to the right of the split. This decision rule has a mirror rule, which just swaps these assignments. Thus it is possible to test the decision rule (split) against its mirror using the data that conform to the decision rule and those which conform to its mirror. This, on average, will be half the data for each. The right plot of figure 6.2 illustrates this procedure.

The split that maximises the test statistic (for example using the Mann-Whitney-U test, log-rank test, or the student-t test) is then chosen to define that node of the tree. The data to the left of the split are passed onto the left child, and those to the right are passed to the right child. The procedure is iterated for each child, up to a maximum tree depth (a user-defined parameter). In order for the optimal split to be chosen, it must satisfy certain constraints, such as a p-value below some pre-defined threshold, and with a minimum number of the elements, both for the test and for each child.

When a forest $F = \{DT_j\}_{j=1}^{J}$ has been constructed, where DT_j is the j^{th} decision tree, the ultimate decision rule is defined as the consensus vote over the forest. That is to say, for a

new datum X', every tree assigns a treatment $T = DT_j(X')$ and the majority assignment is chosen. This is the same as for the conventional random forest procedure.

Diagnostics Using canonical tests such as the Mann-Whitney-U, Student-t or log-rank, we can construct decision trees where splits are determined by maximising the test statistic. If the data are bootstrapped, and only random subset of the covariates are used at each step, then this mimics the random forest procedure. However, it is also necessary to measure the overall performance of the random forest. We can obtain analogous metrics to the conventional random forest methodology:

- Compute for each covariate a measure of variable importance. The naive way of doing this would be by simply counting the occurrences of each variable in the forest. A more interesting way of computing the variable importance is by using the out-of-bag samples. For each variable in each tree, one can compare the p-value of that tree against the p-value under permutation of the variable (in the out-of-bag data).
- Assign each tree a p-value calculated using the out-of-bag samples. This is an automatic cross-validation procedure. Under the null hypothesis, this should be uniformly distributed.
- Compute the set of variable importance scores and tree p-values under the null by permuting the dataset. We look at this in more depth in section 6.3.3.

Tuning parameters The main parameters that effect the performance of the algorithm are given in the following list. There are also more minor parameters which are needed for the tree construction such as $\min_{split}$: the minimum number of data points which must be present in every leaf of the tree; $\min_{test}$: the minimum number of data points that can be used to test a split. This is the number of data points which conform to decision rule and to its mirror. These are less important for the overall performance, and it is fairly easy to set default values. However, the following parameters effect the predictive performance and the computational cost of the algorithm:

- D, the maximum depth of each tree. Greater depth reduces the bias of each tree but increases the cost of constructing each tree. From simulations studies (see next section) the performance of the algorithm is fairly sensitive to this parameter, with deeper trees ($D > 6$) tending to over-fit.
- P, the number of covariates to subset at each split. Smaller values of P reduces the correlation between trees (also reduced by bootstrapping the data). Lower correlation

between trees reduces the overall variance of the forest (see page 588 Hastie et al., 2009), however if the number of covariates which have a true interaction with the treatment is low, then a lower P will increase overfitting, i.e. selecting covariates which have relation to treatment response.

- J, the number of trees in the forest. Larger values of J reduces the variance of the forest, of course at higher computational cost. Rule-of-thumb values such as 500 (the default value for the *R* package *randomForest*) seems to perform well (not much benefit from growing more trees).

- p_{cutoff}, the p-value at which to select a split. At each node of the tree, the best proposed split is only accepted if the p-value corresponding to the test statistic at that split is below p_{cutoff}. This parameter would need tuning, but from our simulations, the forest construction is not highly sensitive to the choice of cutoff value.

It can be argued that random forests is a black box algorithm which is difficult to interpret and does not provide insight into the system being modelled. It's main strength comes from the ensemble learning which reduces the overall variance, but it is not possible, for example, to visualise 500 decision trees. However, it is possible to plot the overall decision rule (obtained from the consensus vote) on low dimensional sets of covariates. Even though the data may be high-dimensional, evidence for patient heterogeneity would come from a low-dimensional subspace, which is much easier to plot. We illustrate these types of plots in the next section.

6.3.3 Characterising over-fitting

One issue with this method is quantifying whether the random forest is over-fitting the data. It is of course possible to use standard methods such as ten-fold cross-validation to estimate the over-fitting but this implies using less of the data to train the forest of decision trees. We propose instead the following simple idea:

- Fit the random forest using the whole dataset $\{Y,T,X\}$ with J trees. This gives an overall decision rule D.

- Test the decision rule D against its mirror $\tilde{D}$, again using the data $\{Y,T,X\}$, to give a p-value p^{fit} (note this p-value is *not* uniformly distributed as the data are used twice with multiple testing).

- For K iterations: permute the response Y to the covariates X and treatment T, and repeat the above procedure to obtain analogous p-values $\{p_1^{\text{perm}},..,p_K^{\text{perm}}\}$.

- The rank of p^{fit} in the set $\{p_1^{\text{perm}},..,p_K^{\text{perm}}\}$ is uniformly distributed under the null hypothesis (no patient heterogeneity). This accounts for the double use of the data and the multiple testing in the procedure.

This method gives a Monte Carlo estimate of the true p-value corresponding to the fitted random forest. We denote this the *evidence p-value*, which is a corrected version of the p-value computed using the data twice. This of course implies running a full analysis $K+1$ times, but all these computations can be done in parallel. A major attraction of the method is that is gives a single measure of evidence of heterogeneity in the data. This procedure is illustrated with an example in the next section.

6.4 Comparison study with OWL

To finish this chapter I present a brief simulation study comparing the performance of our algorithm with that of OWL (Zhao et al., 2012). The authors kindly let us use their software which implements linear OWL (the kernel function f in equation 6.2.4 is a linear function of the covariates X). We look at three synthetic examples taken from Zhao et al., section 4. It is important to note that although none of these examples are in my opinion biologically plausible patterns (see figure 6.3), they are of interest because of the difficult for decision trees to accurately determine such structures.

All the examples consist of a 50-dimensional covariate space X, with every $X_i \sim \text{Unif}[-1,1]$. The treatment T is uniformly sampled from $\{-1,1\}$. The response R^a is normally distributed with mean: $Q_a(X,T) = 1+2X_1+X_2+0.5X_3+f_a(X,T)$ and standard deviation 1, where, for $a=1,2,3$, $f_a(X,T)$ is given by:

1. $f_1(X,T) = 0.442(1-X_1-X_2)T$ (linear decision rule)
2. $f_2(X,T) = (X_2-0.25X_1^2-1)T$ (parabolic decision rule)
3. $f_3(X,T) = (0.5-X_1^2-X_2^2)(X_1^2+X_2^2-0.3)T$ (ring decision rule)

Thus the response is only determined by the first two covariates X_1 and X_2, the other 48 being just noise. The optimal decision surface as a function of these first two covariates is given in figure 6.3, where red corresponds to treatment 1 being optimal, and white to treatment -1 being optimal. We run the linear OWL code provided by the authors and then compare the two methods by predicting on a 50×50 grid of points over $[0,1]^2$ (values of covariates X_1 and X_2) and a random vector for values of $X_3,..,X_{50}$.

The random forest algorithm used the Mann-Whitney-U test and was run with the following parametrisation: $D=4$ (depth of each tree); $P=25$ (number of covariates selected

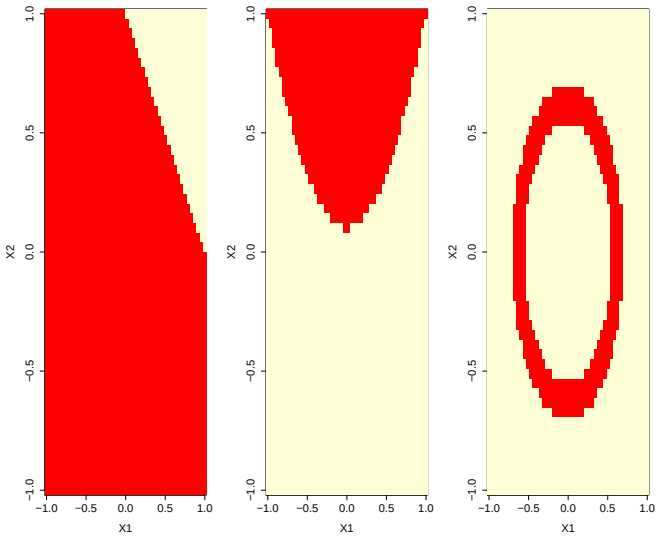

Fig. 6.3 Three synthetic examples taken from Zhao et al. (2012), from left to right corresponding to the mean response function Q_a defining a linear, parabolic and ring decision rule, respectively. Red corresponds to treatment 1 being optimal, and white to treatment -1 being optimal.

at each step); $J = 500$ (number of trees); $p_{\text{cutoff}} = 0.1$ (cutoff p-value needed to make a split). To fit the method, we used $n = 250$ datapoints were generated where $R^a \sim \mathcal{N}(Q_a(X,T),1)$.

Figure 6.4 shows the results for both methods. We do not compare the two methods using a measure such as mean squared error as this is can be deceptive as to the true performance of each method. In these three simple examples it is possible to visually compare the performance. The left column of figure 6.4 shows the predictive contours for our algorithm, and the right column those of OWL. These can then be compared to the true optimal decision surface given in figure 6.3. For a mean response generated as $Q_1(X,T)$ (linear rule), the linear OWL picks up the decision contour almost perfectly as it is designed to find exactly such decision rules. Our algorithm doesn't do badly, clearly finding a difference between the top right hand corner of the grid and the rest. Both of the methods performs similarly for data generated with mean response $Q_2(X,T)$ (parabolic rule). Our algorithm does not reproduce the curved decision rule. For $Q_3(X,T)$ (ring decision rule), the linear OWL cannot fit such a structure, and our algorithm finds the central part of the decision surface but doesn't pick up the ring structure[5].

We use this last example to test out our proposed method for estimating the evidence in the data of patient heterogeneity (evidence p-values: see section 6.3.3). The corresponding

[5]This is mainly a problem of the depth required to find such a complicated structure. Other simulations show that if the structure is a closed ring then the algorithm fits it well.

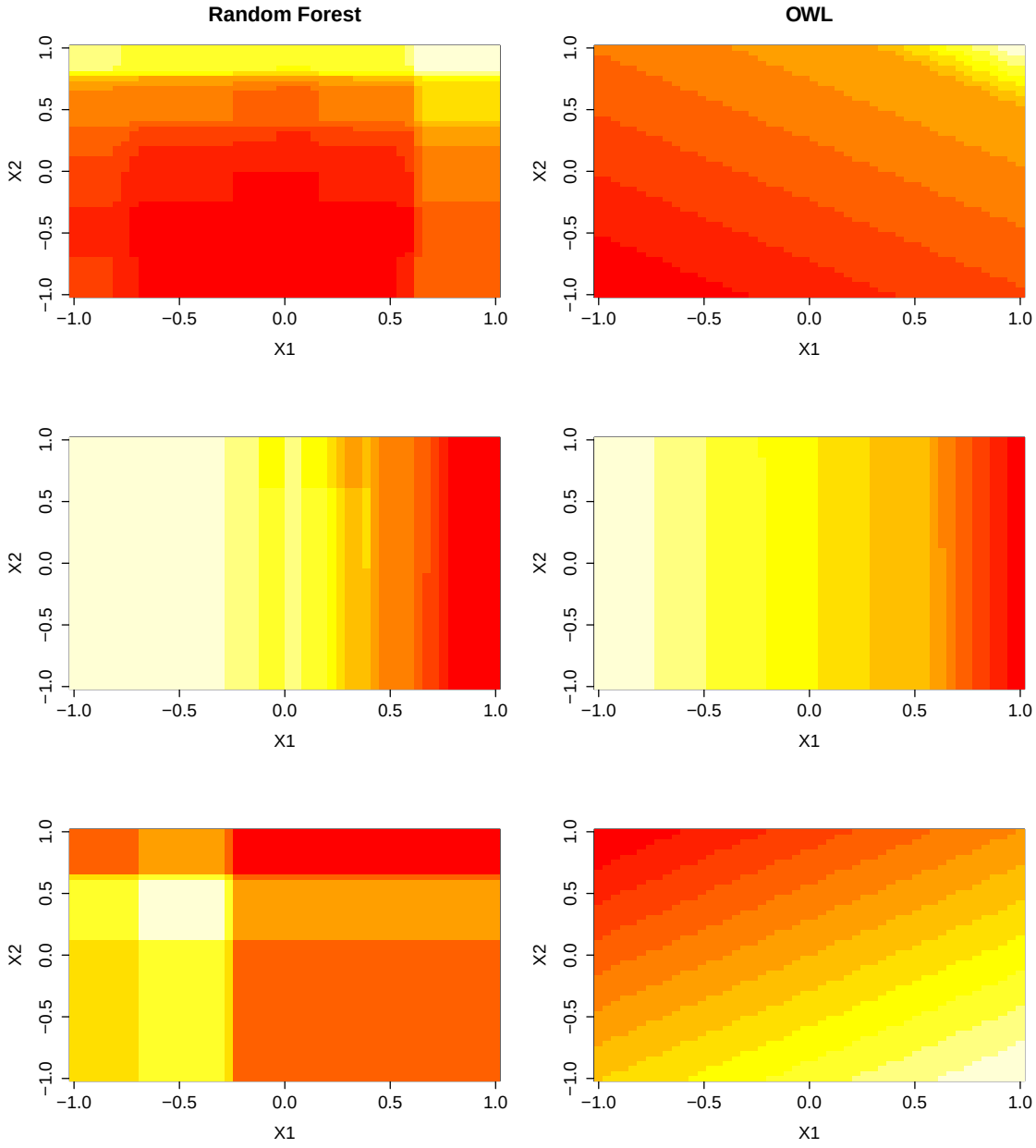

Fig. 6.4 Comparison of performance between the random forest method (left column) and OWL (right column) for synthetic examples 1-3 (respectively rows from top to bottom). Each plot is a heatmap of the predictive decision surface for the first two covariates, with a randomly generated vector of remaining covariates. Red hues are where treatment 1 is optimal, white where treatment -1 is optimal.

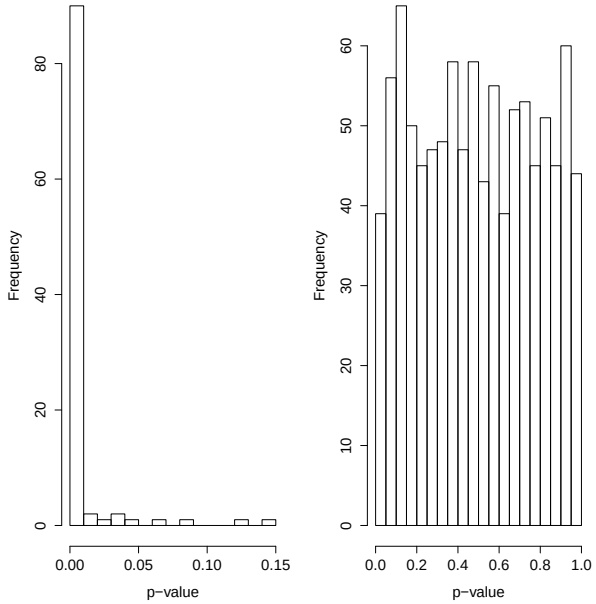

Fig. 6.5 Left: histogram of evidence p-values computed for 100 simulated datasets for the ring decision rule with $n = 150$; Right: histogram of evidence p-values for one simulated dataset ($n = 150$), permuted 1000 times.

p-value for the decision rule defined by the fitted random forest tested against its mirror on the data is ≈ 0.01. We re-run this analysis on 100 new datasets to estimate the distribution of this p-value, the histogram of which is given in the left hand plot of figure 6.5. In general we see that these p-values are very small (less than 0.01). For the dataset that gave an p-value of ≈ 0.01 (there was a bias in the selection of this particular dataset in order to get a number not too small!), we ran 1000 times the algorithm with the response permuted to treatment and covariates. The histogram of the p-values is given in the right hand plot of figure 6.5. Using the Monte Carlo estimate method described in section 6.3.3, we obtain a corrected evidence p-value of 0.018 (its rank was 18^{th} in the 1000 simulations). This very clearly shows that the method is indeed picking up the signal for patient heterogeneity in the non-permuted data, and is not overfitting the permuted dataset. Indeed the histogram of p-values for the permuted datasets appears very close to a uniform distribution.

To summarise, other than in the scenario where the true decision rule is linear, our method seems to perform qualitatively at least as well as OWL with the added benefit of being able to produce a single measure of evidence in favour of underlying patient heterogeneity.

6.5 Discussion

The method we present in this chapter falls into the category of exploratory, post hoc data analysis, specifically designed for testing decision rules using randomised trial data. Its purpose is to see whether it is possible to reject a null hypothesis, defined as no patient heterogeneity to treatment. In view of the importance of stratified medicine in pharmaceutical research and drug development, this is a key problem. The accepted consensus is, in our opinion, given by papers such as Sun et al. (2014), and dismisses the use of 'data mining' methods for the determination of clinically relevant subgroups. We agree with this statement, however, in order to make full use of all the information contained in data from large randomised clinical trials, specially designed methods are needed. By framing the problem as a hypothesis test, where the null is truly the object of interest (no patient heterogeneity), and working from first principles, we see that this reduces to simply testing decision rules versus their mirror decision rules. It is then possible to choose whichever test (model free or not) is most suitable to the data at hand. It is important to note that this method will never be as powerful as a fully parametric approach when the data really satisfy strong model assumptions, for example, linearity, Gaussian noise, etc. If the use of a parametric or semiparametric model is justified, then our method would not apply and statisticians should make use of one of many of the plethora of predictive models available in the literature. The strength of this method can be fully appreciated in situations when valid model assumptions are scarce and a more robust method is needed. Added to that we have a Monte Carlo method for approximating a single measure of evidence of heterogeneity in the data. This is done by simple permutation tests in order to approximate the distribution of p-values under the null distribution of the algorithm.

Another aspect of the method developed in this chapter concerns the use of random forests for testing. In itself, this is not novel and has already been studied by Ishwaran et al. (2008) in the context of survival analysis. We argue that the use of random forest-type algorithms for hypothesis testing extends to any testing mechanism, for example the Mann-Whitney-U, student-t, etc. Ensemble learning, and in particular random forests, is fast and easily applicable to high dimensional data. We harness the well known advantages of the random forest approach, which can identify complex interaction patterns, to explore possible decision rules.

Chapter 7

Further work

7.1 Assessing the impact of model misspecification

Adapting the principles of decision theory to complex and imperfect models is a challenge for statistical research. The work done in Chapters 2-5 proposes a framework in which the principles of Bayesian decision theory can be unified with Wald's minimax to guard against known and unknown misspecification in the model. This formal framework is completed with diagnostic plots and summary statistics which allow the statistician to qualitatively explore the sensitivity of the decision system. However, much remains to be done. I briefly expose some ideas for further work and possible improvements.

7.1.1 Interpreting the Kullback-Leibler divergence

Chapters 2 & 3 presented a formal framework with which it is possible to assess the sensitivity of a decision system when the underlying model π_I is believed misspecified. An essential component is the construction of an information neighbourhood centred at the approximating model π_I. The 'distance' used to define this neighbourhood is the Kullback-Leibler divergence, the choice of which is justified by theoretical principles (for example, the logarithmic scoring rule being the unique proper local scoring rule), and is appealing because of properties such as an analytical solution for the least favourable distribution. But the correct calibration of the size C of this neighbourhood is of course key to understanding how sensitive the decision system is to misspecification. This requires an interpretation of units of Kullback-Leibler, so that a level of confidence in the model can be translated into a KL radius. Chapter 2, section 2.3.4, overviews suggestions from the literature and some new ideas of how the KL radius can be calibrated and converted into more meaningful units. However, to my knowledge, there is no entirely satisfactory method which would allow practitioners

to use this framework in a fully automated manner and it remains an open problem. My work gives some partial solutions mainly based on qualitative assessments of interpretable parameters (either of the model or pertaining to the finite nature of the representation of the model itself, e.g. in the computational decision theory setup).

Principled calibration of the Kullback-Leibler divergence is also important for the work done in Chapter 3 regarding the Pólya Tree process. The results can be applied to two different contexts. Firstly, the parametrisation of a Pólya Tree process centred at a model π_I defines the first two moments of the (random) distance in Kullback-Leibler of random draws from the process with respect to π_I. This in turn gives a computationally cheap method for sampling distributions at a given KL divergence of the approximating model, which can be used to assess sensitivity of the decision system via this nonparametric extension of the model. Secondly, an important aspect of this work relates to the use of the Pólya Tree process in the context of Bayesian inference as a nonparametric prior. Instead of using a default parameterisation ("sensible canonical choice", Lavine, 1992), a more principled approach would be for the statistician to specify his prior knowledge via beliefs concerning the divergence of the 'true' model from the posited baseline model. Both of these necessitate a valid interpretation of the units of divergence.

7.1.2 Computation decision theory

What I refer to as 'computational decision theory' is the use of Bayesian decision theory in situations where the implementation of the idealised decision procedure has to make short cuts because of the complexity of the analysis.

These short cuts can pose conceptual issues (because of concerns of misspecification), and are driven by computational issues demanding approximate methods (tractable optimisation, computation of loss etc) and practical issues (for example, data privacy, see Chapter 5, section 5.2.2). This requires new ways of approaching decision theoretic problems and I believe some of the work related to this aspect is of particular interest and merits further exploration. For example, the framework of Chapter 2 is appealing because it can be implemented at very little extra computational cost when the distribution π_I has a 'bag of samples' representation. The Monte Carlo samples can also be used as a basis for diagnostic plots, visually representing the structure of the model π_I in relation to the loss function (Chapter 4). But this work only explores very small part of what can be done. Some as yet undeveloped ideas are as follows.

In order to estimate the least favourable distribution $\pi_a^{\sup}$, an importance sampling method was proposed (Chapter 2, section 2.3), because the likelihood ratio is known to be $e^{\lambda L(a,\theta)}$. The accuracy of this method however depends entirely on the form of the loss function $L(a,\theta)$. In situations where the desired radius C (and therefore the tilting parameter λ) is too

large for a given loss function, other methods may be required. One idea could be to instead use a Sequential Monte Carlo sampler to approximate $\pi_a^{\sup}$ with λ as the time parameter.

Another problem is to how to sample all distributions exactly within a given KL ball centred at π_I. The Pólya Tree model extension (Chapter 3) samples within a 'donut neighbourhood' and the Dirichlet process model extension samples at an infinite KL divergence. In the case where π_I is represented by a finite number of n samples, the problem is reduced to sampling within a KL constrained n-simplex. This may well be suited for a Hamiltonian Monte Carlo approach. Rejection sampling with a Dirichlet distribution would have too low an acceptance rate for small KL radii.

As mentioned in Chapter 5, for many applied problems it can be difficult to estimate the optimal Bayes action a^* because the integration required for the expected loss estimate at each action a is computationally expensive (Müller, 2005). This could be when the action space $\mathscr{A}$ is continuous and high-dimensional, making the optimisation problem intractable. In the context of a concern of model misspecification, it might be possible to simultaneously search for an optimal point a^* whilst at the same time assessing the robustness. This is to say, both $\int_\Theta L(a,\theta)\pi_I(\theta)\mathrm{d}\theta$ and $\int_\Theta L(a,\theta)\pi_a^{\sup}(\theta)\mathrm{d}\theta$ would be used to score each action a.

7.1.3 Loss functions and misspecification

As mentioned at the start of Chapter 5, a real impediment to the use of Bayesian decision theory in practice is the elicitation of the loss function $L(a,\theta)$. Misspecification in the loss function is as important as misspecification in the model $\pi_I(\theta)$ (although the loss functions are not unique - see the notion of strategic equivalence, Chapter 5, section 5.1). However, there are very few methods for assessing the sensitivity of the loss as compared to sensitivity the model. Some ideas are given in the review of Bayesian robustness by Ruggeri et al. (2005). Indeed, an straightforward approach would be to extend ideas concerning model neighbourhoods to the loss function and consider a class of loss functions instead of a single loss function. Ruggeri et al. (2005) give some examples of these classes. In the same way, the ε contamination neighbourhood for probability measures (see Chapter 1, section 1.2.2) can be extended to loss functions, where a baseline loss function L_0 can have ε additive contamination from a member L from a given class of losses.

In the same way that this work looks at the range of expected loss within a KL neighbourhood, it is then possible to look at the corresponding range over the classes of loss functions. There is still no obviously principled way of choosing the class in the first place, with notions of neighbourhoods not easily extended to loss functions.

As shown in Ruggeri et al. (2005), a decision system can be insensitive to changes in the loss or changes in model, but sensitive to changes in both. This means that methods for jointly estimating the robustness of a decision system are needed.

7.2 Decision rules for randomised trials

7.2.1 Methodological Improvements

The work presented in Chapter 6 on assessing the evidence of patient heterogeneity to treatment is ongoing and still at an early stage. A main objective of the work is to provide a universal methodology which clinicians and medical researchers can use as an exploratory tool in the context of the analysis of randomised clinical trials. I argue that the strength of the algorithm presented in Chapter 6 is twofold.

Firstly, it is simple to understand and relatively fast to run on large datasets (having the same performance as the conventional random forest method[1]). The simplicity of the method is key for its widespread use. Secondly, its minimal set of assumptions makes it widely applicable and highly robust. Indeed, much more so than a method relying on the construction of a predictive model. However, improvements are still needed in order for it to be adopted by researchers. Here are ideas which I hope will be added to the methodology.

Although the method itself relies on a very simple idea (testing decision rules against their mirror images), the random forest algorithm used to search the space of decision rules is in my opinion a 'black box' type algorithm. Its efficiency has been empirically proven by its widespread use and the fact that it has the status of a state-of-the-art method. Its main strength is for prediction. But the fitted model itself (T decision trees) is complex to understand, even with measures such as variable importance and out-of-bag p-values. Decision trees themselves, however, are simple to understand and visually represent. If it were possible to construct a *consensus decision tree* which represented the whole forest of decision trees, then this would make the method more interpretable. The construction of this consensus decision tree would be an optimisation problem using a misclassification loss for example. Its purpose would not be for prediction but for a visual representation of the fitted model. I note that this is different from, say, selecting the best decision tree in the forest (which would not include all the variables and would still have high variance).

Another idea to improve the algorithm could be called *deforestation*. The out-of-bag p-values are accurate measures of the strength of each tree in the forest. Using them to prune

[1] I would not claim that random forests are a 'big data' method (large n). However, they apply well to high-dimensional datasets (large p), for example SNP data.

the forest of decision trees could help increase the accuracy of the forest. This could help when only a few of the predictor variables define the subgroups and the number of predictor variables sub-selected at each step is small.

7.2.2 Empirical performance

Although the method is motivated as a *post hoc* exploratory tool, it is simple to extend it to subgroup analysis which conforms to the criteria defined by Sun et al. (2012, see section 6.2.2, Chapter 6). Suppose that $D(X)$ is a well defined subgroup, posited before the randomised trial. Then the p-value which corresponds to testing $D(X)$ against its mirror $D(\tilde{X})$ using the data from the trial is uniformly distributed under the null hypothesis. This gives a single measure of the evidence supporting the heterogeneity posited by the decision rule $D(X)$. More importantly however, it makes no assumptions regarding the relation between the response, treatment and covariates, other than the assumptions made for the test. In this manner, it is a completely general method for analysing patient heterogeneity, making the minimal set of assumptions.

I would like to know how this method would then perform when used to retrospectively analyse trials for which there were reported subgroups effects that were subsequently not replicated in further trials. Rothwell (2005) provides a list of such trials. By applying the method to these datasets, testing the supposed decision rules against mirror image, it would be possible to see how much evidence our method gives to the possibility of patient heterogeneity in cases where we now know there is none. Such empirical evidence could help reassure researchers that the method is not overconfident and not overfitting.

References

Ahmadi-Javid, A. (2011). An information-theoretic approach to constructing coherent risk measures. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2125–2127. IEEE.

Ahmadi-Javid, A. (2012). Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546.

Altman, D. G. (1998). Within trial variation? A false trail? *Journal of Clinical Epidemiology*, 51(4):301–303.

Anastos, K., Charney, P., Charon, R., et al. (1991). Hypertension in women: what is really known? *Ann Intern Med*, 115:287–293.

Anonymous (1821). Dissertation sur la recherche du milieu le plus probable. *Ann. Math. Pures et Appl.*, 12:181–204.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, 2(2):123–146.

Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

Baio, G. and Dawid, A. P. (2011). Probabilistic sensitivity analysis in health economics. *Statistical methods in medical research*.

Basle Committee (1996). Amendment to the capital accord to incorporate market risks. *Basle Committee on Banking Supervision*.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Berger, J. O. (1984). The robust Bayesian viewpoint (with discussion). *Robustness in Bayesian Statistics (J. Kadane, ed.)*, pages 63–124.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer, second edition.

Berger, J. O. (1994). An overview of robust Bayesian analysis – with discussion. *Test*, 3(1):5–124.

Berger, J. O. and Berliner, L. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *The Annals of Statistics*, 14(1):461–486.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, pages 905–938.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.*, 7(3):686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Bessel, F. W. (1818). *Fundamenta Astronomiae*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer New York.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2013). A general framework for updating belief distributions. Technical report.

Bissiri, P. G. and Walker, S. G. (2012a). Converting information into probability measures with the Kullback-Leibler divergence. *Annals of the Institute of Statistical Mathematics*, 64(6):1139–1160.

Bissiri, P. G. and Walker, S. G. (2012b). On Bayesian learning via loss functions. *Journal of statistical planning and inference*, 142(12):3167–3173.

Box, G. and Draper, N. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40(3-4):318–335.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breuer, T. and Csiszár, I. (2013a). Measuring distribution model risk. *arXiv preprint*.

Breuer, T. and Csiszár, I. (2013b). Systematic stress tests with entropic plausibility constraints. *Journal of Banking & Finance*, 37(5):1552–1559.

Brian, N. (2010). Bayesian analysis using power priors with application to paediatric quality of care. *Journal of Biometrics & Biostatistics*.

Brown, B. J., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, 86(3):635–648.

Dalalyan, A. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61.

Dalalyan, A. and Tsybakov, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78:1423–1443.

Dempster, A. P. (1975). A subjectivist look at robustness. *Bull. Internat. Statist. Inst*, 46:349–374.

Devlin, N. and Parkin, D. (2004). Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health economics*, 13(5):437–452.

Draper, D., Kahn, K. L., Reinisch, E. J., Sherwood, M. J., Carney, M. F., Kosecoff, J., Keeler, E. B., Rogers, W. H., Savitt, H., and Allen, H. (1990). Studying the effects of the DRG-based prospective payment system on quality of care: design, sampling, and fieldwork. *The Journal of the American Medical Association*, 264(15):1956–1961.

Edington, A. S. (1914). *Stellar movements and the structure of the universe*.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, pages 643–669.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*.

Fabius, J. et al. (1964). Asymptotic behavior of Bayes' estimates. *The Annals of Mathematical Statistics*, 35(2):846–856.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880.

Fouskakis, D. and Draper, D. (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *Journal of the American Statistical Association*, 103(484):1367–1381.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *The Annals of Applied Statistics*, 3(2):pp. 663–690.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403.

Gelman, A. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, third edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. Technical report.

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 6(4):733–760.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Gerber, H. U. and Shiu, E. S. W. (1993). Option pricing by Esscher transforms. *Transactions of the Society of Actuaries*, 46:99–191.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.*, 26(2):187–202.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153.

Good, I. (1950). *Probability and the weighing of evidence*. C. Griffin.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606.

Grünwald, P. and van Ommen, T. (2014). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Technical report.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14.

Hansen, L. P. and Sargent, T. J. (2001a). Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics*, 4(3):519–535.

Hansen, L. P. and Sargent, T. J. (2001b). Robust control and model uncertainty. *The American Economic Review*, 91(2):60–66.

Hansen, L. P. and Sargent, T. J. (2008). *Robustness*. Princeton University Press.

Hansen, L. P., Sargent, T. J., Turmuhambetova, G., and Williams, N. (2006). Robust control and model misspecification. *Journal of Economic Theory*, 128(1):45–90.

Hanson, T. (2006). Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*, 101(476).

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*, 97(460).

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, second edition.

Hill, B. A. (1966). Reflections on controlled trials. *Annals of the rheumatic diseases*, 25(2):107.

Hoff, P. and Wakefield, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters: A discussion of Bayesian inference with misspecified models by Stephen Walker. *Journal of Statistical Planning and Inference*, 143(10):1638–1642.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1972). The 1972 Wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, pages 1041–1067.

Huber, P. J. (2009). *Robust statistics*. Wiley, second edition.

Ibrahim, J. G. and Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, pages 841–860.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12(3):941–963.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.*, 36(5):2207–2231.

Kadane, J. B., editor (1984). *Robustness of Bayesian analyses*. Studies in Bayesian econometrics. North-Holland.

Kadane, J. B. and Srinivasan, C. (1994). Discussion of Berger, J. O., An overview of robust Bayesian analysis – with discussion. *Test*, 3(1):116–120.

Kahn, K. L., Rubenstein, L. V., Draper, D., Kosecoff, J., Rogers, W. H., Keeler, E. B., and Brook, R. H. (1990). The effects of the DRG-based prospective payment system on quality of care for hospitalized Medicare patients. *Journal of the American Medical Association*, 264:1953–1955.

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, 50(2):123–148.

Kerman, J., Gelman, A., Zheng, T., and Ding, Y. (2008). Visualization in Bayesian Data Analysis. In *Handbook of Data Visualization*, Springer Handbooks Comp.Statistics, pages 709–724. Springer Berlin Heidelberg.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan and Co.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co.

Kraft, C. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):385–388.

Kullback, S. (1959). *Information theory and statistics*. Gloucester MA: Peter Smith.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

Lavine, M. (1991). Sensitivity in Bayesian statistics: the prior and the likelihood. *Journal of the American Statistical Association*, 86(414):pp. 396–399.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, pages 1222–1235.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer series in statistics. Springer.

Loomes, G. and McKenzie, L. (1989). The use of QALYs in health care decision making. *Social science & medicine*, 28(4):299–308.

Marmot, M. et al. (2012). The benefits and harms of breast cancer screening: an independent review. *Lancet*, 380:1778–1786.

Martins, T. G., Simpson, D. P., Riebler, A., Rue, H., and Sorbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint*.

Matsouaka, R. A., Li, J., and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics*, 70(3):489–499.

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association*, 84(406):473–478.

Muliere, P. and Walker, S. (1997). A Bayesian non-parametric approach to survival analysis using Pólya trees. *Scandinavian Journal of Statistics*, 24(3):331–340.

Müller, P. (2005). Simulation based optimal design. In *Bayesian Thinking Modeling and Computation*, volume 25 of *Handbook of Statistics*, pages 509 – 518. Elsevier.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, pages 343–366.

Nieto-Barajas, L. E. and Müller, P. (2012). The rubbery Pólya tree. *Scandinavian Journal of Statistics*, 39(1):166–184.

Parmigiani, G. (1993). On optimal screening ages. *Journal of the American Statistical Association*, 88(422):622–628.

Pierce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal*, 2:161–163.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. Technical report.

Popper, K. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2):1180–1210.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science.

Rockafellar, R. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471.

Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1):339–371.

Rothwell, P. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

Ruberg, S. J., Chen, L., and Wang, Y. (2010). The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical trials*.

Rubin, D. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Rue, H. (1995). New loss functions in bayesian imaging. *Journal of the American Statistical Association*, 90(431):900–908.

Ruggeri, F., Ríos Insua, D., and Martín, J. (2005). Robust Bayesian analysis. *Handbook of statistics*, 25:623–667.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schervish, M. J. (1995). *Theory of Statistics*. Springer series in statistics. Springer.

Shapiro, S., Venet, W., Strax, P., and Venet, L. (1988). Periodic screening for breast cancer: the Health Insurance Plan project and its sequelae, 1963-1986. *The John Hopkins University Press*.

Staudte, R. G. and Sheather, S. J. (1990). *Location-Dispersion Estimation*, pages 95–147. John Wiley & Sons, Inc.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation, 1885-1920. *Journal of the American Statistical Association*, 68(344):pp. 872–879.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10:141–158.

Sun, X., Briel, M., Busse, J. W., et al. (2012). Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*, 344.

Sun, X., Ioannidis, J. P. A., Agoritsas, T., Alba, A. C., and Guyatt, G. (2014). How to use a subgroup analysis: user's guide to the medical literature. *JAMA*, 311(4):405–411.

Tiao, G. C. and Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, 80(3):623–641.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press.

Turner, D. A., Wailoo, A. J., Cooper, N. J., Sutton, A. J., Abrams, K. R., and Nicholson, K. G. (2006). The cost-effectiveness of influenza vaccination of healthy adults 50-64 years of age. *Vaccine*, 24(7):1035 – 1043.

van't Veer, L. J. and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–570.

Vidakovic, B. (2000). Γ-minimax: a paradigm for conservative robust Bayesians. In Rios Insua, D. and Ruggeri, F., editors, *Robust Bayesian analysis*, pages 241–259. Springer.

Von Neumann, J. and Morgenstern, O. (1947). *The theory of games and economic behavior*. Princeton University Press.

Wald, A. (1945). Statistical decision functions which minimize the maximum risk. *Annals of Mathematic Statistics*, 46(2):265–280.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205.

Wald, A. (1950). *Statistical decision functions.* Wiley.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633.

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):pp. 485–527.

Walker, S. G. and Mallick, B. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):845–860.

Watson, J. and Holmes, C. C. (2014). Approximate models and robust decisions. *ArXiv: preprint*.

Watson, J., Nieto-Barajas, L., and Holmes, C. (2014). Characterising variation of nonparametric random probability measures using the kullback-leibler divergence. *arXiv preprint arXiv:1411.6578*.

Whittle, P. (1990). *Risk-sensitive Optimal Control*. Wiley.

Wu, D., Rosner, G. L., and Broemeling, L. D. (2007). Bayesian inference for the lead time in periodic cancer screening. *Biometrics*, 63(3):873–880.

Zeckhauser, R. and Shepard, D. (1976). Where now for saving lives? *Law and contemporary problems*, pages 5–45.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):pp. 278–280.

Zhang, T. (2006a). From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Annals of Statistics*, 34:2180–2210.

Zhang, T. (2006b). Information theoretical upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52:1307–1321.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Appendix A

Local Sensitivity Analysis

Proof. We define $\psi(\lambda) = \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)] = \int_\Theta L_a(\theta)\pi_I(\theta)e^{\lambda L_a(\theta)}Z_\lambda^{-1}\mathrm{d}\theta$
where $Z_\lambda = \int_\Theta \pi_I(\theta)e^{\lambda L_a(\theta)}\mathrm{d}\theta$ (normalising constant).

$$\frac{\mathrm{d}\psi}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\int_\Theta L_a(\theta)\pi^{\sup}_{a,C(\lambda)}(\theta)\mathrm{d}\theta = \int_\Theta L_a(\theta)\pi_I(\theta)\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(e^{\lambda L_a(\theta)}Z_\lambda^{-1}\right)\mathrm{d}\theta$$

$$= \int_\Theta L_a(\theta)\pi_I(\theta)\left(\frac{L_a(\theta)e^{\lambda L_a(\theta)}Z_\lambda - e^{\lambda L_a(\theta)}\frac{\mathrm{d}Z_\lambda}{\mathrm{d}\lambda}}{Z_\lambda^2}\right)\mathrm{d}\theta$$

$$= \int_\Theta L_a(\theta)^2\pi_I(\theta)e^{\lambda L_a(\theta)}Z_\lambda^{-1}\mathrm{d}\theta - \int_\Theta L_a(\theta)\pi_I(\theta)e^{\lambda L_a(\theta)}Z_\lambda^{-1}\left(\int_\Theta L_a(\theta)\pi_I(\theta)e^{\lambda L_a(\theta)}Z_\lambda^{-1}\mathrm{d}\theta\right)\mathrm{d}\theta$$

$$= \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)^2] - \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)]^2 = \mathrm{Var}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)]$$

For $\lambda > 0$, define the corresponding KL divergence C_λ:

$$C_\lambda := K(\lambda) := \int_\Theta \pi_I(\theta)\log\frac{\pi_I(\theta)Z_\lambda}{\pi_I(\theta)e^{\lambda L_a(\theta)}}\mathrm{d}\theta$$

Hence:

$$\frac{\mathrm{d}K}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\int_\Theta \pi_I(\theta)\left(\log Z_\lambda - \lambda L_a(\theta)\right)\mathrm{d}\theta = \frac{\mathrm{d}}{\mathrm{d}\lambda}\log Z_\lambda - \int_\Theta \frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda\pi_I(\theta)L_a(\theta)\mathrm{d}\theta$$

$$= Z_\lambda^{-1}\int_\Theta L_a(\theta)\pi_I(\theta)e^{\lambda L_a(\theta)}\mathrm{d}\theta - \int_\Theta \pi_I(\theta)L_a(\theta)\mathrm{d}\theta = \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)] - \mathbb{E}_{\pi_I}[L_a(\theta)]$$

So the reciprocal derivative is:

$$\frac{\mathrm{d}}{\mathrm{d}C_\lambda}(K^{-1}) = \frac{1}{\frac{\mathrm{d}K}{\mathrm{d}\lambda}\left(K^{-1}(C_\lambda)\right)}$$

□