Submitted to Statistical Science

Approximate Models and Robust Decisions

James Watson and Chris Holmes

University of Oxford

Abstract. Decisions based partly or solely on predictions from probabilistic models may be sensitive to model misspecification. Statisticians are taught from an early stage that "all models are wrong", but little formal guidance exists on how to assess the impact of model approximation on decision making, or how to proceed when optimal actions appear sensitive to model fidelity. This article presents an overview of recent developments across different disciplines to address this. We review diagnostic techniques, including graphical approaches and summary statistics, to help highlight decisions made through minimised expected loss that are sensitive to model misspecification. We then consider formal methods for decision making under model misspecification by quantifying stability of optimal actions to perturbations to the model within a neighbourhood of model space. This neighbourhood is defined in either one of two ways. Firstly, in a strong sense via an information (Kullback-Leibler) divergence around the approximating model. Secondly, using a Bayesian nonparametric model (prior) centred on the approximating model, in order to 'average out' over possible misspecifications. This is presented in the context of recent work in the robust control, macroeconomics and financial mathematics literature. We adopt a Bayesian approach throughout although the presentation is agnostic to this position.

Key words and phrases: Computational decision theory, Model misspecification, *D-open* problem, Kullback-Leibler divergence, Robustness, Bayesian nonparametrics.

Department of Statistics, University of Oxford(e-mail: cholmes@stats.ox.ac.uk)

1. INTRODUCTION

This article presents recent developments in robust decision analysis using approximate statistical models. The central theme of the paper can be summarised as follows, that the consequence of statistical model misspecification is contextual and hence should be dealt with under a decision theoretic framework e.g. (Berger, 1985; Parmigiani & Inoue, 2009). As a toy illustration consider the following: suppose that data arise from an exponential distribution, $x \sim \exp(\lambda)$, yet the statistician adopts a normal model, incorrectly assuming $x \sim N(\mu, \sigma^2)$. If interest is in the estimation of the mean $E[X]$ and the sample size is large, then there may be little consequence in the misspecification. However if the focus is on the probability of an interval event, say $X \in [a, b]$, then there might be far reaching consequences. Of course this is a toy problem and careful model checking and refinement will help in reducing misspecification, but pragmatically, especially in modern high-dimensional data settings, it seems to us inappropriate to separate the issue of model misspecification from the consequences, context, and rationale of the modelling exercise.

Statisticians are taught from an early stage that "essentially, all models are wrong, but some are useful" (Box & Draper, 1987). By "wrong" we will take to mean misspecified and by "useful" we will take to mean helpful for aiding actions (taking decisions), or rather a model is not useful if it does not aid any decision. We will refer to such situations as *D-open* problems, to highlight that Nature's true model is outside of the decision makers knowledge domain, c.f. *M-open* in Bayesian statistics which refers to problems in updating beliefs when the model is known to be wrong (Bernardo & Smith, 1994). We will adopt a Bayesian standpoint throughout although the approach we develop is generic. We will assume there is uncertainty in some aspect of the world[1], $\theta \in \Theta$, which if known would determine the loss in taking an action a as quantified through a real-valued measurable loss-function, $L_a(\theta)$. The loss will often be a joint function of states and observables, $L_a(\theta, x)$, although we shall suppress this notation for convenience. Uncertainty in θ is characterised via a probability distribution $\pi_I(\theta)$ given all available information I. Without loss of generality we will assume that θ relates to parameters of a probability model and information I is in the form of data, $x \in \mathcal{X}$, and a joint model $\pi(x, \theta)$, such that,

$$\pi_I(\theta) \equiv \pi(\theta|x) \propto f(x;\theta)\pi(\theta),$$

where $\pi(x, \theta)$ is factorised according to the sampling distribution (or likelihood) $f(x;\theta)$ and the prior $\pi(\theta)$; although more generally $\pi_I(\theta)$ simply represents the statisticians best current beliefs about the value of the unknown state θ. Following the axiomatic approach of Savage (1954) the rational coherent approach to decision making is to select an action $\hat{a}$ from the set of available actions $a \in A$ so as to minimise expected loss,

$$\hat{a} = \arg \inf_{a \in A} \mathbb{E}_{\pi_I(\theta)}[L_a(\theta)]. \tag{1}$$

This underpins the foundations of Bayesian statistics (Bernardo & Smith, 1994). The problem is that (1) assumes perfect precision in specifying $\pi(x, \theta)$. In reality

[1]Savage (1954) refers to Θ as the "small world" relevant to the decision.

the model $\pi(x, \theta)$ is misspecified, such that the decision maker acknowledges that $f(x; \theta)$ may not be Nature's true sampling distribution or $\pi(\theta)$ does not reflect all aspects of prior subjective beliefs in $f(x; \theta)$ or on the marginal $\pi(x) = \int_\theta \pi(x, \theta) d\theta$. This paper presents diagnostics and formal methods to assist in exploring the potential impact of this misspecification.

It is important to note that we will not spend much time on the area of pure inference problems such as robust estimation of summary functionals for which there is a substantial literature (Huber, 2011), or on recent work on the use of loss functions to construct posterior models (Bissiri *et al.*, 2013). We shall also pass quickly over the use of conventional prior sensitivity analysis and robust "heavy tailed" priors and likelihoods. We are principally concerned with *ex-post*[2] settings where $\pi(x, \theta)$ has been specified to the best of the modellers ability under the practical constraints of computation and time, and where concerns arise as to whether $\pi(x, \theta)$ represents the modeller's true marginal $\pi(x)$ to sufficient precision. This is particularly important when θ pertains to a high-dimensional complex model or to the value of a future predicted observation.

There is a rich literature in Bayesian statistics on model robustness, the vast majority of which relates to sensitivity to specification of the prior $\pi(\theta)$. We review the material in detail below but mention here the overviews in Berger (1994), Rios Insua & Ruggeri (2000) and Ruggeri *et al.* (2005). Bayesian robustness was a highly active area through the 1980s to mid-1990s. Interest tailored off somewhat since that time, principally due to the arrival of computational methods such as Markov chain Monte Carlo (MCMC) coupled with developments in hierarchical models, nonlinear models and nonparametric priors, see e.g. Chipman *et al.* (1998), Robert & Casella (2004), Rasmussen & Williams (2006), Denison *et al.* (2002), and Hjort *et al.* (2010). These methods allow for very flexible model specifications alleviating the historic concern that $\pi(x, \theta)$ was indexing a restrictive sub-class of models. However, a number of recent factors merit a reappraisal. In the 1990s and 2000s computational advances and hierarchical models broadly outpaced the complexity of data sets being considered by statisticians. In more recent times very high-dimensional data are becoming common, the so called "big data" era, whose size and complexities prohibit application of fully specified carefully crafted models, (e.g. National Research Council: *et al.*, 2013, Chapter 7). Relevant to this, approximate probabilistic inference techniques that are misspecified by design have emerged as important tools for applied statisticians tackling complex inference problems. For example, models involving composite likelihoods, integrated nested Laplace approximations (INLA), Variational Bayes, Approximate Bayesian Computation (ABC), all start with the premise of misspecification, see e.g. Beaumont *et al.* (2002), Fearnhead & Prangle (2012), Marjoram *et al.* (2003), Marin *et al.* (2012), Minka (2001), Ratmann *et al.* (2009), Rue *et al.* (2009), Varin *et al.* (2011), and Wainwright & Jordan (2003). Finally there have been recent developments in coherent risk measures within the macroeconomics and mathematical finance literature, building on areas of robust control, which are of importance and relevance to statisticians, as outlined in Section 2 below.

The rest is as follows. In Section 2 we review some background literature on decision robustness and quantification of expected loss under model misspecifi-

[2]Meaning here 'once the modelling has been completed'. In a Bayesian setting, this refers to dealing directly with the posterior quantities.

cation. In Section 3 we review diagnostic tools to assist applied statisticians in identifying actions which may be sensitive to model fidelity. Section 4 presents formal methods for summarising decision stability, by exploring the consequence of misspecification within local neighbourhoods around the approximate model. Section 5 contains illustrations. Conclusions are made in Section 6.

2. BACKGROUND ON DECISIONS UNDER MODEL MISSPECIFICATION

We first review some of the background literature on decisions made under model misspecification.

2.1 Minimax

The first axiomatic approach to robust statistical decision making was made by Wald (1950). In the absence of a true model, Wald interpreted the decision problem as a zero sum two-person game, following Von Neumann and Morgenstern's work on game theory (Von Neumann & Morgenstern, 1947). To be robust the statistician protects himself against the worst possible outcome, selecting an action $\hat{a}$ according to the minimax rule, which for the purposes of this paper we can consider as[3],

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

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, Nature will select the worst possible outcome, equivalent to the assumption of a point mass distribution taken reactively to your choice of action,

$$\delta_{\theta_a^*}(\theta),$$

where,

$$\theta_a^* = \arg \sup_{\theta \in \Theta} L_a(\theta).$$

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. It assumes that Nature is reactive in selecting $\delta_{\theta_a^*}(\theta)$ for your choice of $a \in A$ irrespective of the evidence from existing information I 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.

2.2 Robust Bayesian statistics

Under a strict Bayesian position there is no issue with model robustness. You precisely specify your subjective beliefs through $\pi(x, \theta)$ and condition on data to obtain posterior beliefs, taking actions according to the Savage axioms. However, even the modern founders of Bayesian statistics acknowledged issues with an approach that assumes infinite subjective precision,

[3]Wald's original work considered selection of decision functions, $\delta(x) \in A$, by non-conditional loss quantified as frequentist risk, $R[F_X, \delta(x)] = \int L(\delta, x) F(dx)$, with $x \in \mathcal{X}$ from unknown distribution F_X.

> "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)
>
> "...in practice the theory of personal probability is supposed to be an idealization of one's own standard of behaviour; that the idealization is often imperfect in such a way that an aura of vagueness is attached to many judgements of personal probability..." Savage (1954)

As Berger points out, many people somewhat distrust the Bayesian approach as "Prior distributions can never be quantified or elicited exactly (i.e. without error), especially in finite amount of time" – Assumption II in Berger (1984). This then raises the thorny issue of what does the resulting posterior distribution $\pi(\theta|x)$ actually represent?

An intuitive solution is to first specify an operational prior model $\pi_0(\theta)$, to the best of your available time and ability, and then investigate sensitivity of inference or decisions to departures around $\pi_0(\theta)$, typically assuming that $f(x;\theta)$ is known so that divergence is with respect to the prior. This idea has origins in the work of Robbins (1952) and Good (1952) with many important contributions since that time. We mention just a few pertinent areas below, referring the interested reader to the review articles of Berger (1984, 1994), Wasserman (1992), and Ruggeri *et al.* (2005), as well as the collection of papers in the edited volumes of Kadane (1984) and Rios Insua & Ruggeri (2000).

The resulting robust Bayesian methods are usefully classified as either "local" or "global". Local approaches look at functional derivatives of posterior quantities of interest with respect to perturbations around the baseline model, e.g. Ruggeri & Wasserman (1993) Sivaganesan (2000); see also Kadane & Chuang (1978) who consider asymptotic stability of decision risk. Global approaches consider variation in a posterior functional of interest, $\psi = \int h(\theta)\pi(\theta|x)d\theta$, by varying the prior π within a neighbourhood $\pi \in \Gamma$ centred around some π_0. A typical quantity would be the range $(\psi^{\inf}, \psi^{\sup})$ where $\psi^{\inf} = \inf_{\pi\in\Gamma} \psi$ and $\psi^{\sup} = \sup_{\pi\in\Gamma} \psi$. The challenge is to define the nature and size of Γ so as to capture plausible ambiguity in π_0, while taking into account factors such as ease of specification and computational tractability, Berger (1994; 1985 section 4.7). One important example being the ϵ-*contamination* neighbourhood (Berger & Berliner, 1986) formed by the mixture model,

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

where ϵ is the perceived contamination error in π_0 and $\mathcal{Q}$ is a class of contaminant distributions. It is usual to restrict $\mathcal{Q}$ so that it is not "too big", for instance by including only uni-modal distributions Berger (1994), for which it is shown that the solutions have tractable form. Other approaches consider frequentist risk, such as Γ-*minimax* that investigates the minimax Bayes (frequentist) risk of $\psi^{\sup}$ for $\pi \in \Gamma$ whereas *conditional* Γ-*minimax* procedures (Vidakovic, 2000) study the maximum expected loss across prior distributions within Γ, this being perhaps closest to the approach we develop here. Also of note is the idea of model elaboration (Carota *et al.*, 1996), whereby a parametric model is embedded into a larger class of models (for example a normal model into the family of student-t

distributions). Diagnostics concerning the sensitivity of the model can then be obtained by comparing the two alternative posterior distributions, for instance using their Kullback-Leibler divergence as a summary statistic. This is somewhat different from the approach we discuss here where we do not embed the model (posterior) in a parametric or nonparametric family but rather embed it into a neighbourhood of distributions[4].

A more general distinction is that here we recommend that robustness to misspecification should apply to only those states θ that enter into the loss function $L_a(\theta)$. This facilitates application to high-dimensional problems for which specification of Γ may be difficult (Sivaganesan, 1994) and helps tackle the thorny issue that changing the likelihood changes the interpretation of the prior (Ruggeri *et al.*, 2005, page 635).

2.3 Robust control, macroeconomics and finance

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 robustness 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.

In Economics one early criticism of the Savage axioms was that the framework could not distinguish between different types of uncertainty. Gilboa & Schmeidler (1989) developed a theory of maxmin Expected Utility in part to counter the famous Ellsberg paradox[5] 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).

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

[4]Section 4.3 considers an embedding of the posterior in a nonparametric model such as the Dirichlet Process, however the purpose of this is to allow for sampling of distributions within some neighbourhood of the model π_I.

[5]The standard setting for the Ellsberg paradox is as follows: imagine two urns each containing 100 balls and every ball is either red or blue. One is told that the first urn (A) has 50 red balls and 50 blue balls exactly. No more information is given about the second urn (B). Suppose you win 100$ if you pick a red ball, which urn would you choose? So there exists a set of alternatives which are equal in expected value (under any reasonable prior) but which appear to have different empirical preferences.

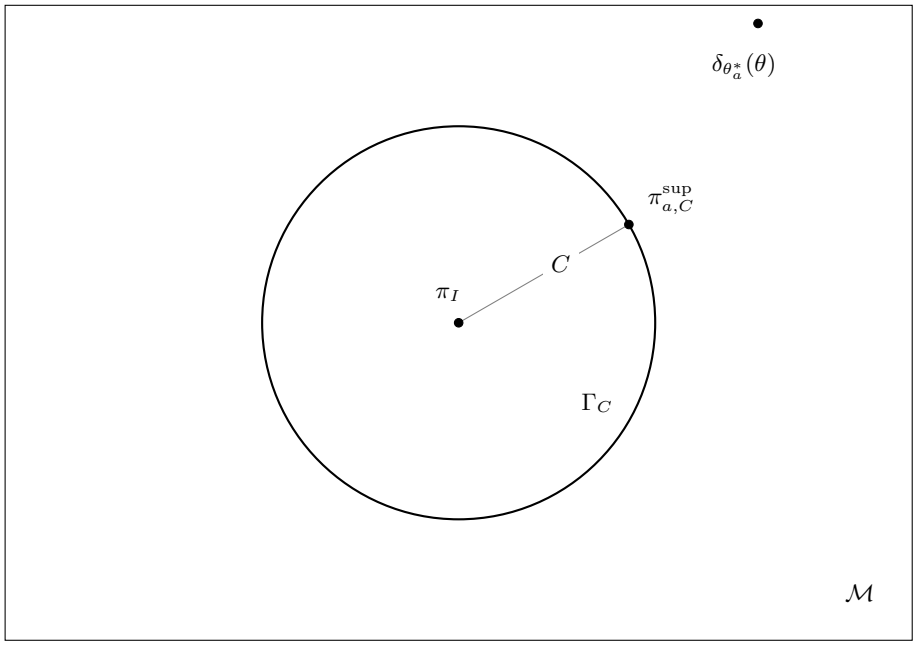

FIG 1. *Graphical representation of local-minimax model* $\pi_{a,C}^{\sup}$ *within a Kullback-Leibler ball of radius* C *around the reference model* π_I*, with global (Wald's) minimax density* $\delta_{\theta_a^*}(\theta)$*.*

neighbourhood of the posterior $\pi_I(\theta)$ through exploration of,

$$\psi_{(a)}^{\sup}(C) = \sup_{\pi \in \Gamma_C} \mathbb{E}_\pi[L_a(\theta)]$$

where Γ_C denotes a KL ball of radius C around π_I,

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

We will use $\pi_{a,C}^{\sup}$ to denote the corresponding local-minimax distribution,

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

Figure 1 shows a pictorial representation of this constrained minimax rule, where the reference distribution π_I is a point in the space $\mathcal{M}$ of all distributions on θ (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 & Sargent (2008) for a thorough review and references.

Breuer & Csiszár (2013a, 2013b), 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 $\pi_{a,C}^{\sup}$ 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). In Section 4 we derive the same result using an alternative, less general, but perhaps more intuitive proof. Before considering these formal methods we shall start with exploratory diagnostics and visualisation methods.

3. D-OPEN DIAGNOSTICS

All good statistical data analysis begins with graphical exploration and evaluation of summary statistics before formal modelling takes place. In this section we consider some graphical displays to aid understanding of when and where actions are sensitive to modelling assumptions. Section 5 further illustrates these ideas. Despite the importance of graphical statistics, there are few if any established tools for investigation of decision stability, although see Vickers & Elkin (2006) for one exception, in contrast to the multitude of methods for investigating model discrepancy and misspecification (Belsley *et al.*, 2005; Gelman, 2007; Kerman *et al.*, 2008). Here we consider graphical displays that concentrate on the relationship between the loss function $L(a, \theta)$ and 'model' or posterior $\pi_I(\theta)$ for a given a. These are examples of how a set of available actions $a \in \mathcal{A}$ can be graphically compared. They could be displayed as a preliminary step to a formal analysis of sensitivity.

3.1 Value at Risk (Quantile-Loss)

A primary tool for assessing the sensitivity with respect to misspecification is to plot 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 indicator function. We use notation $f_a(z) = F_{Z_a}(dz)$ to denote the corresponding density function and $F^{-1}_{Z_a}(q)$, for $q \in [0, 1]$, is the inverse cumulative distribution or quantile function. For robustness given a value $q \in [0, 1]$ it is possible to characterise the utility of an action a by its quantile loss or *Value at Risk* (VaR; terminology used in finance) $F^{-1}_{Z_a}(1 - q)$ rather than the expected loss. Rostek (2010) developed an axiomatic framework in which decision-makers can be uniquely characterised by a quantile $1-q$ and rational behaviour (optimal action) is defined as choosing actions $\hat{a} := \arg\min_{a \in \mathcal{A}} F^{-1}_{Z_a}(1 - q)$. For example, this method could rank them by median loss (taking $q = .5$). Or when $q = 0$ this would correspond to choosing actions using the minimax rule. 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. Although single quantiles discard much information contained in $[\pi_I(\theta), L_a(\theta)]$, plotting this function allows for immediate visualisation of how much of the tails are taken up by high loss (low utility) events. With a bag of samples simulated from the posterior marginal $\theta_1, .., \theta_m \sim \pi_I(\theta)$, this is easily approximated:

1. Sort the realised loss values, $z^{(a)}_i = L_a(\theta_i)$, from highest to lowest $\{z^{(a)}_{\upsilon_a(1)} \geq z^{(a)}_{\upsilon_a(2)} \geq \ldots \geq z^{(a)}_{\upsilon(m)}\}$, where $\upsilon_a(\cdot)$ defines the sort mapping.
2. For $q \in [0, 1]$, approximate $F^{-1}_{Z_a}(1 - q)$ by linear interpolation of the points $\left(x = k/m, y = z^{(a)}_{\upsilon_a(k)}\right)$.

We refer the reader to figure 3 from section 3.5 for an illustration of this plot.

In mathematical finance, summary statistics defined on loss distributions are known as *risk measures*. VaR is a particularly controversial risk measure as it is

widely used (written into official regulations) but is not *coherent*[6] (Basle Committee, 1996; Artzner *et al.*, 1999)[7], violating subadditivity[8]. This has motivated the use of different diagnostics such as CVaR (see next section). We note that expected loss and minimax are both coherent diagnostics. However, the use of coherence in Bayesian decision theory does not seem appropriate in general as the subadditivity axiom does not hold in many applications[9].

3.2 Conditional value at risk (upper-trimmed mean utility)

Conditional Value at Risk (CVaR, Rockafellar & Uryasev, 2000) is another popular alternative to expected loss (or utility), initially motivated by concerns of the incoherence of VaR. 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 (lowest of utility). 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 in to 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. We use the linear interpolation,

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

For a set of actions $\mathcal{A}$, 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).

[6]Note that this is a different definition of coherence from Bayesian coherence, discussed in section 4.1.2.

[7]A coherent risk measure ρ has the following properties: *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 $\rho(Z + Y)$ we mean the risk measure on the combined loss distribution of the combination of the two actions corresponding to the loss distributions Z and Y.

[8]Subadditivity corresponds to investors decreasing risk by diversifying portfolios.

[9]A trivial example is the following: Consider the two gambles, A lose £10 if a fair coin falls on Heads; B lose £10 if a fair coin falls on Tails. For the same coin, if both gambles are taken then one loses £10 with probability 1.

3.3 Cumulative Expected Loss

The *Cumulative Expected Loss* (CEL) function for action a, defined as,

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

for $q \in [0,1]$. The CEL-plot is a monotone decreasing function $J_a(q)$ and an informative graph for highlighting decision sensitivity. An action with CEL-plot that is steeply rising as $q \to 0$ is 'heavily downside' (see for example figure 9 in section 5.2), with expected-loss driven by low-probability high loss outcomes, while CEL-plot rising at 1 indicates 'heavy upside'. In particular $J_a(q)$ and its gradient 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.
- $J_a(1) = \mathbb{E}_{\pi_I(\theta)}[L_a(\theta)]$, is the expected loss of action a, and $\hat{a} = \arg\max_{a \in A} J_a(1)$ 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, e.g. Pritsker (1997).
- $J'_a(0) = \sup_{\theta \in \Theta} L_a(\theta)$, 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 $q \to 0$).

3.4 Sensitivity diagnostics

It is also worth exploring the sensitivity or leverage of the contribution of particular data observations, x_j, to the overall expected loss. In this way one might highlight "outliers" to the decision. We suggest a simple method and graphical display for assessing the sensitivity with respect to individual data points (likelihood) and/or the prior distribution. Again, let the model be represented by a bag of Monte Carlo samples $\theta_1, .., \theta_m \sim_{iid} \pi_I$, noting for a parametric model, $\pi_I(\theta_i) \propto \Pi_{j=1}^n f(x_j|\theta_i)\pi(\theta_i)$, for data x_j, likelihood f and prior π.

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 under a flat prior) has importance weights given by:

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

These weights can be used to compute the leave-one-out (LOO) estimates of the expected loss, where the prior can be considered as an 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 on expected loss for single data points can be evaluated (towards detection of points of high leverage) as can the effect of the prior, which is especially useful in small sample situations.

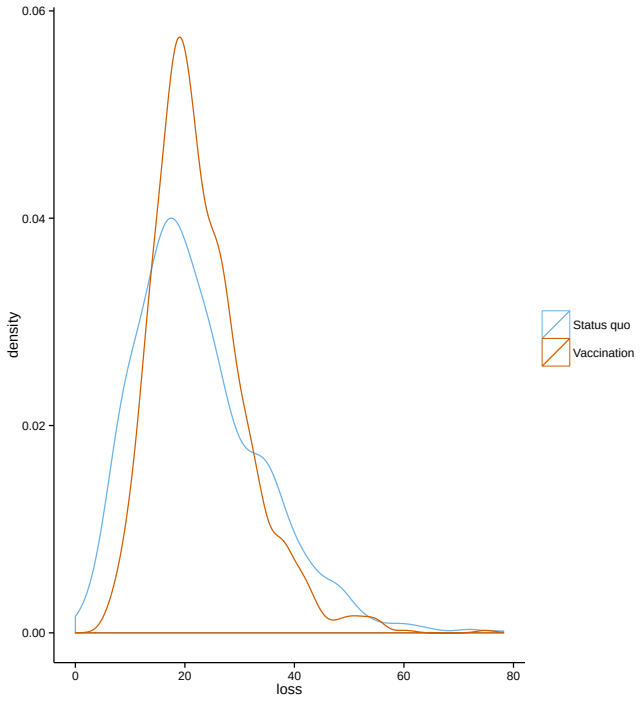

FIG 2. *Loss densities of the two actions for 'willingness to pay' $k = 19400$ (same expected loss for both actions).*

We also propose plotting the loss values $L_a(\theta_i)$ against the density estimates $\pi_I(\theta_i)$ (up to a normalising constant). This will highlight situations where the high loss samples are coming from the tails of the distribution π_I. A more general analysis could even look to model the LOO estimates as a function of x, using regression models, to highlight covariate regions of greatest leverage on the expected loss.

3.5 Motivating synthetic example

We use a fictitious example of a decision process taken from Baio & Dawid (2011). Consider an infectious disease for which there exists treatment medication and a new vaccine. The problem is 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. This is a standard setting for decisions made by institutions such as NICE[10] in the UK, for example. The goal is compare the two available actions: widespread 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[11], complications from the drugs which could lead to extra visits, possible hospitalisation and even death. Each of these outcomes has either a monetary cost (visit to the GP for example) or a utility measured in Quality Adjusted Life Years (QALYs). Therefore to assign a loss value to each action, it is necessary to choose

[10]National Institute for Clinical Excellence.
[11]General Practitioner

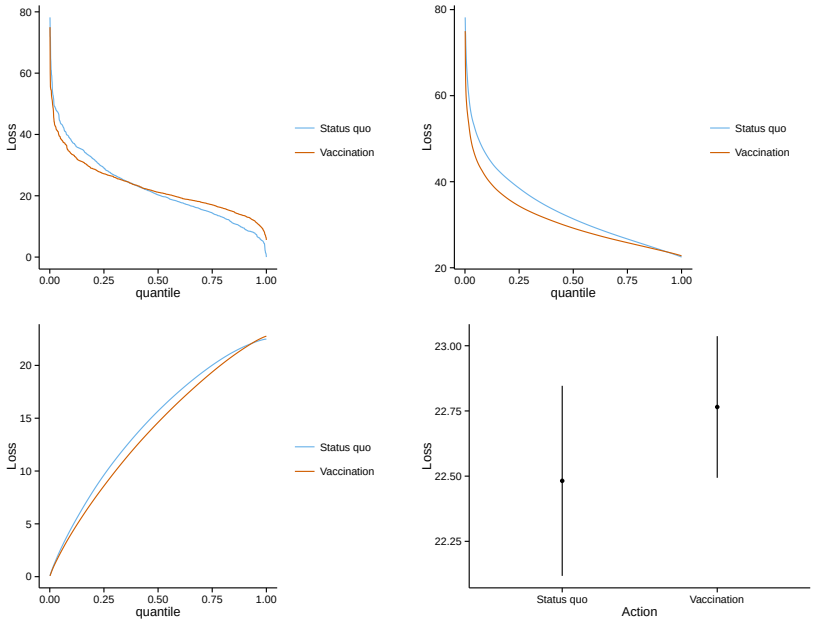

FIG 3. *Diagnostic plots for the decision system given in Baio & Dawid (2011) comparing vaccination (red) to status quo (blue). The 'willingness to pay' parameter set to 19400. From top left to bottom right: Inverse loss distribution; Conditional value at risk; Cumulative expected loss; Expected loss centred at two intervals of standard deviation.*

a conversion rate k, known as the 'willingness to pay' parameter, exchanging QALYs into pounds. 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 implements the model presented in Baio & Dawid (2011) and performs a sensitivity analysis around the parameter k. The exact details of the model are not of particular interest so we do not expose them here, our main purpose being illustrative. The model used for this setting has 28 parameters, each with informative prior distributions, these are given in table 1 of Baio & Dawid (2011). MCMC is used to estimate a posterior distribution, all the relevant details can be found in the package documentation, such as the cost function used etc. We ran the model given in BCEA using the default settings. We note that all the graphs were produced using our package *decisionSensitivityR* and all the code can be found in its documentation. In figure 2 we plot estimates of the densities of loss for the two actions for the 'willingness to pay' parameter $k = 19400$, the threshold value at which the two actions have the same expected loss. The status quo (blue) has greater variance in the distribution of loss than vaccination (red). The value $k = 19400$ is of interest because the two actions are indistinguishable with respect to expected loss. However our method allows for a principled manner of choosing between them (see solution to Ellsberg paradox, section 4.1.3).

Figure 3 illustrates the three diagnostic plots presented above for this application. We see from the inverse loss distribution (top left) that status quo (blue) has higher downside loss than vaccination (red). The CVaR plot (top right) clearer distinguishes the two actions because of this higher downside loss. In this example, the CEL is not informative, but this is context dependent, see figure 9 from

the breast cancer screening application in section 5.2. Note that the expected values (bottom right plot) should match by design, the differences being due to Monte Carlo error.

The diagnostic plots and summary statistics presented in this section allow for visualisation of dependencies between the loss function and posterior distribution, which can highlight the impact of model misspecification on decision making. We now look at formal methods via perturbations to the model (posterior distribution) in order to measure the sensitivity of the expected loss quantities.

4. D-OPEN FORMAL METHODS

In shaping the development of formal methods it may help to state some principles of robust analysis that we would hope to adhere to.

- Principle 1a: **Context**: The impact of model misspecification (approximation) is contextual and hence should be dealt with in a decision theoretic framework.
- Principle 1b: **Consequence**: Following P1a, we should be concerned with sensitivity to only those states θ that affect the expected loss – i.e. states that enter into the loss function $L_a(\cdot)$.
- Principle 2: **Coherence**: robust methods should be coherent, in that two analysts starting with the same approximate joint model (prior and likelihood) and given the same information (data), should arrive at the same robust conclusions when the same loss function is applicable.

We show that these guiding principles lead to a unique characterisation of the problem and robust solution.

An important implication of P1a is that our assessment of the robustness or sensitivity of a model will change depending on what you are doing, so that a model trained on the same data might be highly robust for one analysis and highly sensitive for another. That is, we should be investigate how expected loss $\psi^{(a)} = \int L_a(\theta)\pi_I(\theta)d\theta$ varies under model misspecification. An important implication of P1b is that we shall only be concerned with the marginal posterior model

$$\pi_I(\theta|x) = \int_\eta \pi_I(\theta, \eta|x)d\eta,$$

on the states θ that enter into the loss function $L_a(\theta)$, where η are all other parameters in the likelihood that are nuisance parameters to the loss.

In studying robustness we will consider the construction of a neighbourhood of 'close' models around the marginal $\pi_I(\theta)$. This allows for either a study of the variation of the quantity of interest (expected loss) ψ_a over all models in this neighbourhood, or can guide the construction of a nonparametric extension of the model. In this section, we explore both ideas, each providing the statistician with a different set of tools to estimate the sensitivity of the decision system.

For ease of comprehension, all the notation used throughout this paper is summarised in a glossary in Appendix B.

4.1 Kullback-Leibler neighbourhoods

To investigate formally the stability of decisions to model misspecification we suggest following an approach taken in Hansen & Sargent (2001b) and study

the variation of expected loss $\psi_{(a)}$ over models within a KL ball Γ around the *marginal* posterior density, $\pi_I(\theta)$, of the approximating model on the states that enter into the loss function. We will assume after linear transformation that the loss can be bounded, a reasonable assumption for almost all applied problems[12].

4.1.1 Properties

It is well known that the KL divergence is not symmetric, in general $\mathrm{KL}(\pi^*||\pi) \neq \mathrm{KL}(\pi||\pi^*)$ for $\pi^* \neq \pi$, and following others we consider the neighbourhood $\Gamma_C = \{\pi : \mathrm{KL}(\pi||\pi_I) \leq C\}$. This might be considered the more natural setting as here the KL divergence represents the expected self-information log-loss in using an approximate model π_I when Nature is providing outcomes according to the probability law π. The alternative neighbourhood is considered in the Appendix D[13].

Surprisingly the use of KL leads to a least favourable distribution solution with a simple form.

THEOREM 4.1. *Let* $\pi^{\sup}_{a,C} = \arg\sup_{\pi\in\Gamma_C} \mathbb{E}_\pi[L_a(\theta)]$, *with* $\Gamma_C = \{\pi : \mathrm{KL}(\pi \parallel \pi_I) \leq C\}$ *for* $C \geq 0$. *Then the solution* $\pi^{\sup}_{a,C}$ *is unique and has the following form,*

$$\pi^{\sup}_{a,C} = Z_C^{-1}\pi_I(\theta)\exp[\lambda_a(C)L_a(\theta)] \tag{2}$$

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, $\pi^{\sup}_{a,C} = \arg\sup_{\pi\in\Gamma_C} \mathbb{E}_\pi[L_a(\theta)]$, has an unconstrained Lagrange dual form, see for example Hansen *et al.* (2006) (pages 58-60),

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

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

$$\begin{aligned}
\pi^{\sup}_{a,C} &= \arg\inf_{\pi\in\mathcal{M}} \left\{\int -L_a(\theta)\pi(\theta)d\theta + \eta_a^{-1}\int \pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_I(\theta)}\right)d\theta\right\} \\
&= \arg\inf_{\pi\in\mathcal{M}} \left\{\int \pi(\theta)\log\left(\frac{\pi(\theta)}{\pi_I(\theta)\exp[\eta_a L_a(\theta)]}\right)d\theta\right\} \\
&\propto \pi_I(\theta)\exp[\eta_a L_a(\theta)] \qquad (3)
\end{aligned}$$

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

[12]In practice it is always possible to cap the loss. For instance, any model by which θ is simulated using MCMC this assumption is made. In finance, the potential losses incurred by any individual or organisation could be bounded by an arbitrarily large number, say ±GDP of the US.

[13]In the Monte Carlo setting where π_I is represented by a set of $\{\theta_i\}_{i=1}^m$ each with weight $1/m$, then any distribution π that is a reweighing of θ_i's satisfies: $\mathrm{KL}(\pi_I||\pi) \geq \mathrm{KL}(\pi||\pi_I)$ (Watson *et al.*, 2014). Because we are looking at the variation of the expected loss $\psi_{(a)}$ across the Γ_C, we want to the neighbourhood to be more exclusive for fixed values of the radius C.

By a similar argument the distribution of minimum expected loss follows:

$$\pi^{\inf}_{a,C} \propto \pi_I(\theta) \exp[-\lambda(C) L_a(\theta)]$$

Note by assuming bounded loss functions we can ensure the integrability of the densities. Breuer & Csiszár (2013a) and Ahmadi-Javid (2012) derive the same result more generally but perhaps less intuitively. Breuer & Csiszár (2013a) gives more general conditions on when the solution exists.

The Γ_C least favourable distributions, $\{\pi^{\inf}_{a,C}, \pi^{\sup}_{a,C}\}$, have an interpretable form as exponentially tilted densities, tilted toward the exponentiated loss function, with weighting $\lambda_a(C)$ a monotone function of the neighbourhood size C. For linear loss, $L_a(\theta) = A\theta$, the local least favourable $\pi^{\sup}_{a,C}$ is the well known Esscher Transform used for option pricing in actuarial science. The tilting parameter $\lambda_a(C)$ is a function of the neighbourhood size C, but we will write λ_a for convenience. λ_a and C can be thought of as interchangeable, as there is a bijective mapping between $C \geq 0$ and $\lambda_a \geq 0$, although this is not a linear mapping.

Following Theorem 4.1, the corresponding range of expected losses for each action $(\psi^{\inf}_{(a)}, \psi^{\sup}_{(a)})$ can then be plotted as a function of C for each potential action. Formally we should write $[\psi^{\inf}_{(a)}(C), \psi^{\sup}_{(a)}(C)]$ although for ease of notation we will often suppress C from the expression unless clarity dictates. The constraint $\mathrm{KL}(\pi \parallel \pi_I) \leq C$ will result in $\pi^{\sup}_{a,C}$ lodging on the boundary as the expected loss can always be increased by diverging toward $\delta_{\theta^*_a}(\theta)$ for any distribution with $\mathrm{KL}(\pi \parallel \pi_I) < C$. Substituting the solution (2) into the KL divergence function gives,

$$\begin{aligned} \mathrm{KL}(\pi^{\sup}_{a,C} \parallel \pi_I) &= \int \pi^{\sup}_{a,C}(\theta) \log\left(Z^{-1}_{\lambda_a} \exp[\lambda_a L_a(\theta)]\right) d\theta \\ &= \lambda_a \mathbb{E}_{\pi^{\sup}_{a,C}}[L_a(\theta)] - \log Z_{\lambda_a} \end{aligned}$$

So, given neighbourhood size C, the KL divergence $\mathrm{KL}(\pi^{\sup}_{a,C} \parallel \pi_I)$ is $\lambda_a(C)$ times the expected loss under $\pi^{\sup}_{a,C}$ minus the log partition function. Moreover, by Jensen's inequality,

$$\begin{aligned} \mathrm{KL}(\pi^{\sup}_{a,C} \parallel \pi_I) &= \lambda_a \mathbb{E}_{\pi^{\sup}_{a,C}}[L_a(\theta)] - \log \mathbb{E}_{\pi_I}[\exp(\lambda_a L_a(\theta)] \\ &\leq \lambda_a \left[\mathbb{E}_{\pi^{\sup}_{a,C}}[L_a(\theta)] - \mathbb{E}_{\pi_I}[L_a(\theta)]\right] \end{aligned}$$

The KL divergence is bounded above by λ_a times the difference in expected loss between the approximating and the contained minimax models.

By plotting out the interval $[\psi^{\inf}_{(a)}(C), \psi^{\sup}_{(a)}(C)]$ for each action as a function of KL divergence constraint $C : \mathrm{KL}(\pi \parallel \pi_I) \leq C$ we can look for crossing points between the supremum loss $\psi^{\sup}_{(\hat{a})}$ under the optimal action $\hat{a}$ chosen by the approximating model and the infimum loss under all other actions, $\psi^{\inf} := \inf_{a \in A \setminus \hat{a}} \{\psi^{\inf}_{(a)}\}$

4.1.2 Coherence

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 4.2. *Let $\pi^{\sup}_{a,C}(x,\pi_I)$ be the solution obtained by*

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

with data $x = \{x_i\}_{i=1}^n$, a centring distribution π_I, and arbitrary g-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,

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

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 $\{x^{(1)}, x^{(2)}\}$, for any partition. Then for coherence the divergence $D(\cdot||\cdot)$ is the Kullback-Leibler divergence.

The proof is given in Appendix A. □

This theorem shows that KL is the only divergence measure to provide coherent updating of the local least favourable distribution.

4.1.3 Local sensitivity and regularisation of ill-posed decisions.

Although the framework presented here fits into *global robustness* methods, it is also possible to extract *local robustness* measures and show how they can restore well-posedness to certain decision problems. Consider an ill-posed decision such as the Ellsberg paradox mentioned earlier. The non-uniqueness of the optimal action can be solved via the introduction of a regularization term, c.f. Tikhonov regularization and ridge-regression. In our context this amounts to exploring robust actions in the limit as the neighbourhood size $C \to 0$.

In Appendix C we show that the derivative at zero of the least favourable expected loss w.r.t. λ (exponential tilting parameter) is exactly the variance of the loss distribution.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}|_{\lambda=0} \quad \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a] = \mathrm{Var}_{\pi_I}[L_a(\theta)]$$

This justifies the use of the variance of loss as a local regularization term which 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 $\lambda \to 0$ then this corresponds to the action with lowest variance. For the Ellsberg paradox the robust solution is to select the urn with exactly 50 red and 50 black balls.

4.1.4 Local Bayesian admissibility.

In a classical setting, the notion of *admissibility* defines a subclass of actions that can then be scrutinized in order to choose an optimal decision. A decision is denoted inadmissible if there is no $\theta \in \Theta$ 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 the single quantity used to classify actions, only the action a^* which minimizes expected loss (with respect to π_I) is admissible. However, if we consider the set of posterior distributions contained within a Kullback-Leibler neighbourhood of radius C, then an analogous definition of admissibility can be given. An action a is said to be admissible if there exists a

distribution $\pi \in \Gamma_C$ such that a minimises expected loss in the set $\mathcal{A}$ as calculated with respect to π. The previous results tell us about the expected loss under the least favourable distribution in Γ_C for each action a, but this does not in general tell us whether an action a is admissible in Γ_C. The optimal Bayes action a^* is always admissible for all $C \geq 0$ (because it is optimal under $\pi_I \in \Gamma_C$). However for an action $a \neq a^*$ to be admissible, there must exist a distribution π such that a is better than a^* under expected loss with respect to some $\pi \in \Gamma$. The existence of this distribution can be deduced from the previous results if we look at the regret loss function between the two actions:

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

and the corresponding least favourable pairwise distribution:

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

with expected loss $\psi^{\sup}_{(a,a')}(C) = \int_\theta \pi^{\sup}_{(a,a'),C}(\theta)[L_a(\theta) - L_{a'}(\theta)]d\theta$. The sign of this expected loss indicates whether a is admissible with respect to a^* (if it is positive).

This can be extended to every other action in the set $\mathcal{A}$, thus giving the following definition:

DEFINITION 1. *We say that an action a is C^*-dominated, or locally-inadmissible up to level C^* when,*

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

If a is ∞-dominated then it is globally inadmissible (this retrieves the classical notion of admissibility).

4.1.5 Calibration of neighbourhood size.

In most scenarios the local least favourable distribution $Z_\lambda^{-1}\pi_I(\theta)\exp[\lambda_a L_a(\theta)]$ will not have closed form. When $\pi_I(\theta)$ is represented as a Monte Carlo bag of samples $\{\theta_i\}_{i=1}^m \sim_{iid} \pi_I$ the distribution can be approximated by using π_I as an importance sampler leading to,

$$\begin{aligned} \tilde{\pi}^{\sup}_{a,C} &= \frac{1}{Z_{\lambda_a}} \sum_i w_i \delta_{\theta_i}(\theta) \\ w_i &= \exp[\lambda_a L_a(\theta_i)], \quad Z_{\lambda_a} = \sum_i w_i \end{aligned}$$

We can then use $\tilde{\pi}^{\sup}_{a,C}$ to calculate $(\psi^{\inf}_{(a)}, \psi^{\sup}_{(a)})$. For a small neighbourhood size and hence small λ_a relative to $L_a(\theta)$ this approximation should be accurate. In general if π_I is thought to be a useful model to the truth then the neighbourhood size should be kept small. However as λ increases, the variance of the un-normalised importance weights will grow exponentially and the approximation error with it. In this situation the problem appears amenable to sequential Monte Carlo samplers (Del Moral *et al.*, 2006) taking $\lambda_a \geq 0$ as the "time index" although here we do not explore this any further.

This points to the much wider and important open issue of how to choose the size of the neighbourhood Γ. In the Monte Carlo setting of this problem, the statistician might explore candidate KL values using one or more of the following ideas:

- Explore the distribution of the importance weights and deciding whether this is 'degenerate', for example by looking at the variance of the weights that increases with C.
- Simular to the above, define an *inequality* constraint for the distribution of importance weights, for example consider KL divergences up to the point when 99% of the mass of importance weights is assigned to 1% of the importance samples.
- Calibrate to the distribution of KL values, $KL[\pi(\theta|x_{-d})||\pi(\theta)]$, obtained in a partial update of the prior $\pi(\theta)$ to the posterior given some % subset of the full data set x_{-d}. For example, we could examine a KL neighbourhood size in Γ equating to the average KL divergence of the posterior to the prior using 10% of the data.

Overall, we consider that the calibration of the Kullback-Leibler divergence remains an open problem, even though this divergence is used in many applications. McCulloch (1989) proposes a general solution using a Bernoulli distribution, but it is not obvious that this translates well into a method for the calibration of any continuous distribution. Further options are presented in the Discussion section below.

4.2 Illustrative Solutions

In this section we illustrate the form of some canonical solutions found using the robust decision approach of Section 4.1.

4.2.1 Information annealing.

Under model misspecification there is greater uncertainty in the statistical analysis than supposed by a conventional Bayesian update, whereby the joint model is assumed true. To put it another way, there is less information in the experiment than is conditioned on. In such situations it can be interesting to anneal the information in the following manner.

Likelihood annealing. We consider robustness to the use of the likelihood function, $f(y;\theta)$, which is usually assumed known in a conventional Bayesian update. In this case, under a local proper scoring rule, the natural loss function is the self-information loss $L(y) = -\log f(y|x)$ Bernardo & Smith (1994).

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

This leads to the robust Bayesian update as

$$\tilde{\pi}(\theta|x) \propto [f(y;\theta)]^{1-\lambda}\pi(\theta)$$

for $\lambda \in [0,1]$. This has the form of an annealing of the information in the likelihood function reducing the influence of the data relative to the prior due to concerns on the likelihood specification. This down weighting of the likelihood has previously been considered by others, without the formal justification of appealing to the principles P1–P2 above, e.g. Grünwald & van Ommen (2014); Miller & Dunson (2015); Walker & Hjort (2001).

Concept Drift. More generally suppose there is meta-information, u_i, recorded on each observation, x_i and a belief that robustness to the task at hand may associate with u_i. For example, in data-mining applications when the task is to provide a predictive model, u may record the time at which an observation was collected and there is concern that the underlying system being modeled is not dynamically stable. This is known as "concept drift" (e.g. Section 3.1 in Hand, 2006), although more generally u simply contains information relative to predictive loss. The natural loss function is now a weighted self-information loss, based on the empirical distribution:

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

with $\Delta(u_i) \in (0,1)$ 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

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

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

Predictive annealing. Suppose the task is to provide a marginal predictive distribution, $\widehat{f(y|x)}$, for a future observation y given covariates x, without knowledge or respect to where or how robustness may affect the model. The local proper scoring rule in this case is the self-information logarithmic loss $L(y) = -\log f(y|x)$. The conventional Bayesian solution is to report your honest marginal beliefs as $\widehat{f(y|x)} = f_I(y|x)$, where given a model parametrised by θ we have $f_I(y|x) = \int f(y|x,\theta)\pi_I(\theta)d\theta$. However this assumes that the model is true and stable in time, both of these assumptions being potentially dubious. The solution above protects against misspecification and leads to

$$\widehat{f^*_{\sup}(y|x)} \propto f_I(y|x)^{1-\lambda},$$

for $\lambda \in [0,1]$. This has the form of annealing the predictive distribution, taking into account additional external levels of uncertainty outside of the modelling framework.

4.2.2 General Bayesian updates, Gibbs Posteriors and PAC-Bayes.

Suppose you hold prior beliefs about a set of parameters θ but don't know how to specify the likelihood $f(x|\theta)$, and hence lack a model $\pi(x,\theta)$. For example, suppose θ refers to the median of F_X with unknown distribution. Suppose the task (action) is to provide your best subjective beliefs $\pi(\theta|\cdot)$ conditional on information in the data and prior knowledge. We don't have a likelihood but we could have

a well defined prior hence $\pi_I(\theta) = \pi(\theta)$. In this situation there may be a well defined loss function on the data that we would wish to *maximise* utility against for specifying beliefs, e.g. for the median we should take

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

The distribution that *minimises* the expected loss within a certain KL divergence of the prior is given by the local-maximin distribution,

$$\pi^{\inf}_{a,C} = Z^{-1}_{\lambda_a} e^{-\lambda_a \sum_i |x_i - \theta|} \pi(\theta)$$

This has the form of a Gibbs Posterior or an exponentially weighted PAC-Bayesian approach (Zhang, 2006a,b; Bissiri *et al.*, 2013; Dalalyan & Tsybakov, 2008, 2012). In this way we can interpret Gibbs posteriors as local-maximin solutions in the absence of a known sampling distribution (Bissiri *et al.*, 2013).

4.2.3 Conditional Γ*-minimax priors*

There is a direct relationship between the solution under Section 4.1. and Γ-minimax priors (Vidakovic, 2000) when the $L_a(\theta)$ involves all the parameters in a parametric model so that the posterior is

$$\pi_I(\theta) \propto \Pi^n_{j=1} f(x_j|\theta)\pi(\theta)$$

with likelihood $f(\cdot|\theta)$ and prior $\pi(\theta)$.

Thus the posterior least favourable distribution

$$\pi^{\sup}_a(\theta) \propto e^{\lambda_a L_a(\theta)} \Pi^n_{j=1} f(x_j|\theta)\pi(\theta)$$

can be considered a Bayes update using the minimax prior $\left[e^{\lambda_a L_a(\theta)}\pi(\theta)\right]$ (dropping the normalisation constant). This is an action specific prior.

Note that the KL divergence is the only divergence to ensure this coherency, and also that the "prior" $\pi^{\sup}_a(\theta)$ is data dependent if the loss function uses the empirical risk, i.e. is of the form $L_a(\theta, X)$.

4.3 Characterising variation of expected loss within the neighbourhood Γ

From a Bayesian standpoint rather than consider the "worst case" least favourable distribution in Γ_C it is more natural to characterise the distribution in expected loss arising over all models in the neighbourhood Γ. 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. This is classically a problem in Bayesian nonparametrics, see for example Hjort *et al.* (2010). However, in a decision-theoretic context, only the functionals of the distributions $\pi \in \Gamma$ are of importance. In particular the functionals $\psi_a : \pi \to \mathbb{E}_\pi[L_a(\theta)]$ for $a \in \mathcal{A}$ (expected loss). It is important here to note that 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, see Watson *et al.* (2014) for an example and a further discussion of this. 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.

4.3.1 Dirichlet Processes for functional neighbourhoods

DEFINITION 2. **Dirichlet Process:** *Given a state space $\mathcal{X}$ we say that a random measure P is a Dirichlet Process on $\mathcal{X}$, $P \sim DP(\alpha, P_0)$, with concentration parameter α and baseline measure P_0 if for every finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathcal{X}$, the joint distribution of $\{P(B_1), \ldots, P(B_k)\}$ is a k-dimensional Dirichlet distribution $Dir_k\{\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\}$.*

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

$$\begin{aligned} \{\theta_i\}_{i=1}^m &\sim \pi_I \\ \underline{w} &\sim \mathrm{Dir}_m(\alpha/m, \ldots, \alpha/m), \\ \tilde{\pi}(\theta) &:= \sum_{i=1}^m w_i \delta_{\theta_i}(\theta) \end{aligned} \tag{4}$$

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 $\pi \sim DP(\alpha, \pi_I)$. This construction fits well with the Monte Carlo context, where π_I is represented by a set 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 $DP(\alpha, \pi_I)$, conditional on the atoms θ_i. Ideally, we 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 our 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.

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

$$\psi_a^{\tilde{\pi}} = \sum_i w_i L_a(\theta_i) \tag{5}$$

and the loss distribution by:

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

In what follows, without loss of generality, we fix a and consider the θ_i to be ordered by loss, i.e. $L_a(\theta_1) \leq ... \leq L_a(\theta_m)$. 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):

$$L(a, \theta) \to \frac{L(a, \theta) - \min_{a,\theta} L(a, \theta)}{\max_{a,\theta} L(a, \theta) - \min_{a,\theta} L(a, \theta)} \tag{6}$$

This means each loss cdf takes values between [0,1]. We can study the L_1 distance between the original empirical distribution[14] $F_{a,\tilde{\pi}_I}$ with weights $w_i = 1/m$ 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:

$$\mathbb{E}_v\{|v_i - x_i|\} = \frac{2}{\alpha} \frac{[x_i^{x_i}(1-x_i)^{(1-x_i)}]^\alpha}{\text{Beta}(x_i\alpha, (1-x_i)\alpha)} \tag{7}$$

As a consequence of the linear transformation given in (6), this L_1 difference is bounded by 1/2. $\mathbb{E}_{\underline{w}}\{|F_{a,\tilde{\pi}} - F_{a,\tilde{\pi_I}}|\}$ 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 $L_a(\theta_i) - L_a(\theta_{i-1})$. Using this result we can set the parameter α in order to sample distributions at a certain L_1 distance thus allowing for the calibration of the diagnostic plot presented in the following section.

4.3.2 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 (4). 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 $\pi \sim DP(\pi_I, \alpha)$ (and computing expected loss with respect to this random π). That is to say, each random draw π will induce a ranking of the actions as given by expected loss with respect to π. 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 as a function of the expected L_1 distance of the DP random draws. This 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 $\underline{w} \sim Dir(\alpha/n, .., \alpha/n)$ and using the construction given in (5). We can then use equation 7 to compute the corresponding L_1 distance for a given α value.

5. APPLICATIONS

5.1 Synthetic Example, continued

To illustrate the ideas from sections 4.1 and 4.3, we continue with the vaccination vs. status quo decision problem given in section 3.5. Figure 4 plots the

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

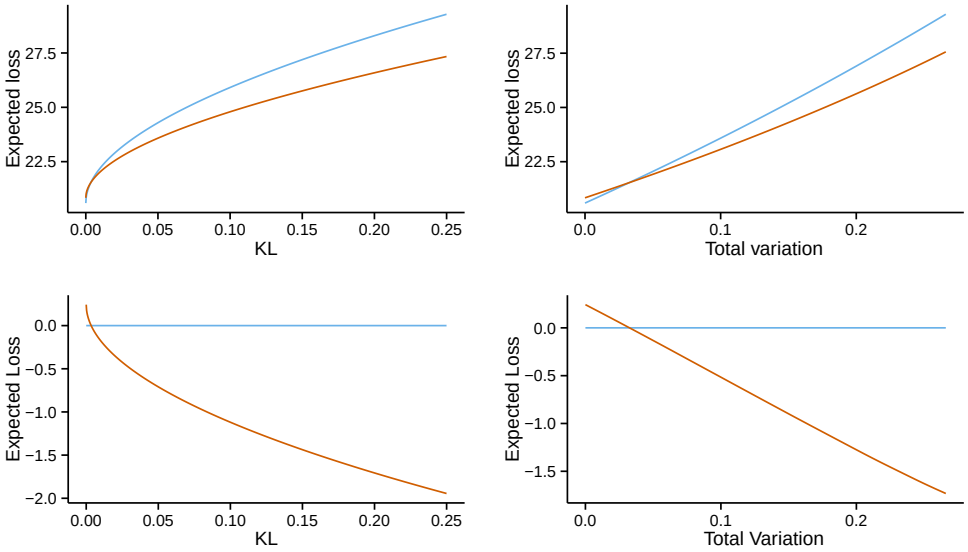

FIG 4. *Diagnostic plots for the local least favourable distribution, comparing vaccination (red, dark line) to status quo (blue, light line). Top row: local least favourable expected loss as a function of KL radius (left) and total variation (right); Bottom row: As for top row but plotting differences in the expected loss between non optimal and optimal actions again measured in KL divergence (left) and total variation (right).*

expected loss both as a function of the ball size as measured in KL divergence (left column) and also as a function of the distance between π_I and π_a^{sup} as measured in total variation (L_1 distance - right column). Showing these plots together provides a method for interpreting the values of KL divergence. We note that the L_1 distance and the KL divergence are related by Pinsker's inequality. These plots show that for relatively small values of KL (and of L_1 distance) the *status quo* action is no longer optimal. This can be explained by the greater variance of the *vaccination* action loss distribution.

Figure 5 shows the probability of optimality for the *status quo* action as a function of the L_1 distance of the draws from the Dirichlet Process (as explained in Section 4.3.1). This shows that for similar L_1 distances, the optimal Bayes action remains the most probable optimal action as defined in Section 4.3.1. This highlights that in fact, the system is decision robust under small perturbations as measured in L_1 distance (see Section 6 for a discussion on decision robustness vs. loss robustness) even though it is not loss robust.

The *vaccination* vs. *status quo* decision problem is synthetic but it allows us the illustrate the diagnostic plots based on the formal methodology presented in section 4. We see that the two diagnostic methods respectively based on KL balls and on Dirichlet process extensions highlight different ways in which the optimal action can be sensitive to model misspecification. The local least favourable distribution concentrates on the high loss values of each action, thus making the vaccination action preferable for very small KL values (see figure 4, top right). However using a Dirichlet model extension, the decision system is robust to symmetrical perturbations around π_I. This is shown in figure 5 where even for small perturbations as measured in expected L_1 distance, the *vaccination* action (blue)

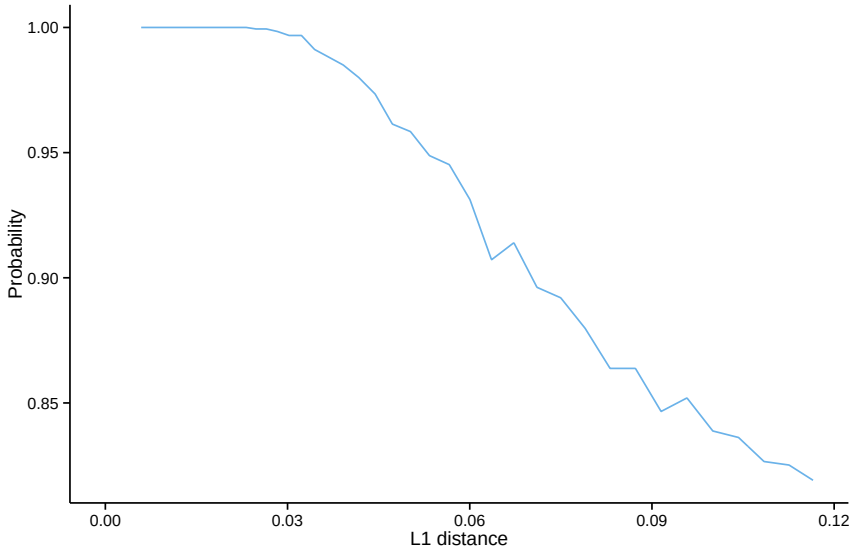

FIG 5. *Probability of optimality of the status quo action under a nonparametric extended model using a* $DP(\pi_I, \alpha)$. *This is plotted as a function of the expected* L_1 *distance of the draws as given in section 4.3.1.*

is not more probably optimal than the *status quo* (red). This shows strong stability to these symmetrical perturbations. Taking a decision as to whether to trust the model or not would be context dependent.

5.2 Example: Optimal Screening Design for Breast Cancer

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 precisely 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. This is of course the gold standard which permits causal inference. However 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 using a statistical model of disease progression throughout a population. Parmigiani (1993) proposed using a semi-Markov process consisting of four states 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, transition time denoted t_D) or contract the disease, modeled by a transition to state B (denoted t_B), the pre-clinical stage. This is followed by a transition to either the clinical stage of the disease (transition time t_C) or death. It was assumed that each transition happens after

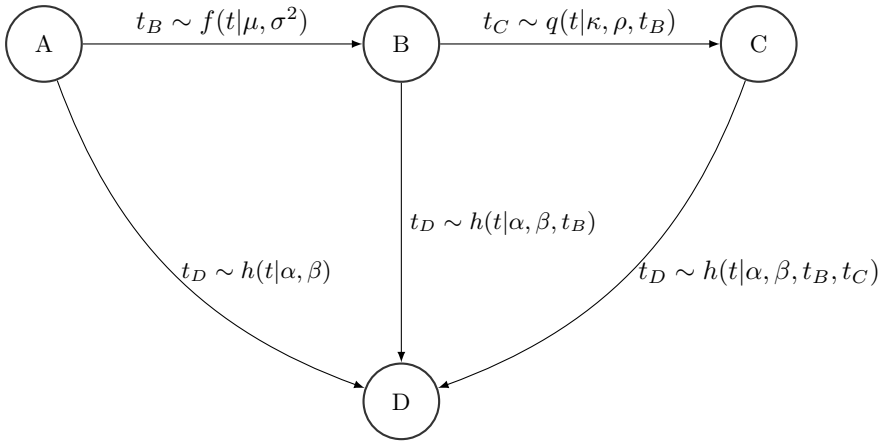

FIG 6. *Graphical model of the transitions and transition densities between states.*

a time t with the following densities:

$$\begin{aligned} t_D \sim h(t|\alpha, \beta) &= \text{Weibull}(\alpha, \beta) \\ t_B \sim f(t|\mu, \sigma^2) &= \text{LogNormal}(\mu, \sigma^2) \\ t_C \sim q(t|\kappa, \rho) &= \text{LogLogistic}(\kappa, \rho) \end{aligned} \tag{8}$$

Figure 6 shows a graphical model of the four state semi-Markov process with 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 $t_D < t_B + t_C$ (assuming that all individuals will contract the disease if they lived long enough). For a screening schedule $a = (t_0, \delta)$ the loss function is defined as follows (a function of the times $t = \{t_B, t_C, t_D\}$):

$$L(a, t) = r \cdot n_a(t) + \mathbb{1}_C \tag{9}$$

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. Each screen has an age-dependent false-negative rate modeled 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. To simulate transition times for individuals from this model, we used 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). This is based on data from the HIP study Shapiro *et al.* (1988). Figure 7 shows the estimated marginal densities for 10^4 sampled times for each transition event[15].

[15]We calibrate the Weibull distribution with values $\alpha = 7.233, \beta = 82.651$ which are the values used in Parmigiani (1993)

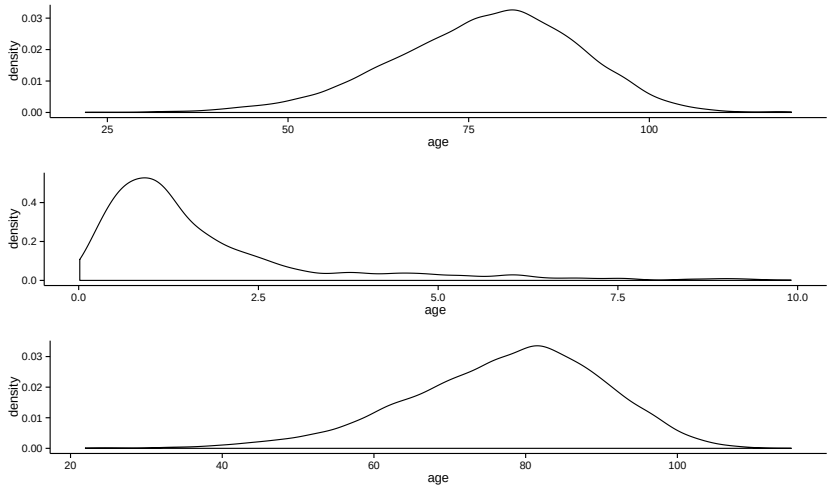

FIG 7. *Probabilistic model of transition times: (from top to bottom) marginal densities of transition times to preclinical stage, transition from preclinical to clinical stage, and death times.*

There is a heated debate as to whether breast cancer screening is in fact beneficial at all due to high false positive rates and screening related anxiety, see for instance Løberg *et al.* (2015). In the UK, the NHS invites women for screening biennially, starting from 50-53 years of age. Some authors argue for earlier screening starting times, from age 40, see Moss *et al.* (2015) and counter-argument Autier (2015). With this in mind, we carried out an ex-post analysis of the screening model and considered 54 alternative schedules, consisting of all combinations of starting ages taken from the set $\{40, 42, 44, 46, 48, 50\}$ (years) and screening frequencies of $\{1, 1.25, 1.5, .., 2.75, 3\}$ (years). This choice of screening schedules is mainly illustrative for our purposes: the optimal schedule will heavily depend on the choice of r (trade-off ratio in equation 9) which we do not attempt to justify (the value 10^{-3} was taken from the section 4.5 of Ruggeri *et al.* (2005) where the authors also considered this application). In order for the plots to be legible, we selected the top 6 schedules[16] (as ordered by a Monte Carlo approximation of the expected loss under the reference model) for analysis. However, there is no reason not to analyse a greater number of schedules other than for clarity in plotting. The top left plot in figure 8 shows the loss density plot of the optimal action corresponding to the schedule $a = (t_0 = 48, \delta = 1.5)$ (units in years) and a trade-off parameter $r = 10^{-3}$. The other three plots show the corresponding loss density for the minimax distributions at KL values equivalent to reassigning 2,5 and 10% of the mass, respectively. The effect can be see as transferring the mass from left to right, i.e. from low loss to high loss. The losses incurred for a particular schedule $a = (t_0, \delta)$ can be seen to be highly bimodal. Most of the population do not contract the disease and therefore contribute a loss of $r \cdot n_a$ (cost of screen times number of total screens during lifetime). The loss contributed by those who do contract the clinical stage of the disease is of magnitude $1/r$ greater by definition.

Figure 9 gives four diagnostic plots for the loss distributions: inverse loss dis-

[16]Given in order of increasing expected loss these are: (48,1.5), (45,1.5), (42,1.5), (50,1.75), (50,2.5) and (48,3).

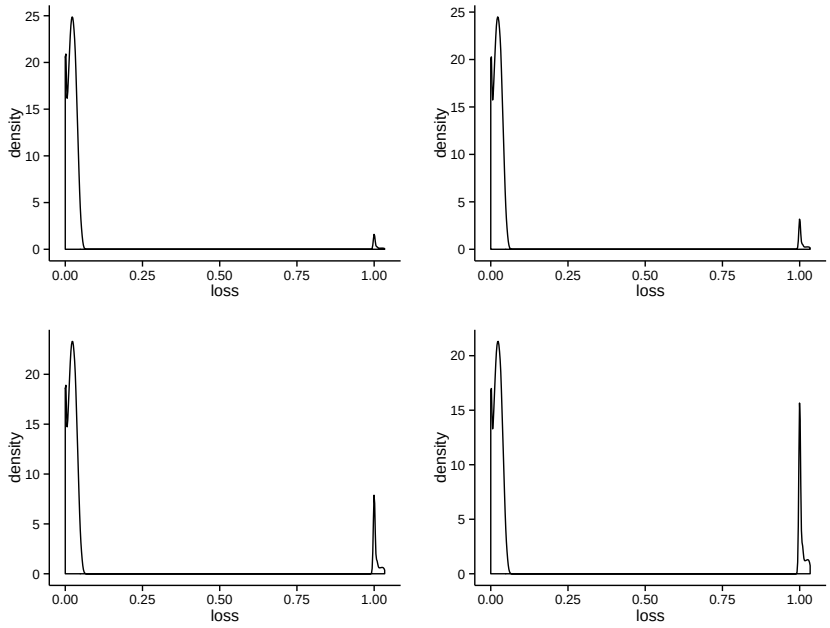

FIG 8. *Top left: loss density for the optimal action (start at 48, frequency 18 months) under the approximating model π_I given in (8). Going from top right to bottom right: loss densities, for the same action, under the local least favourable distribution at KL divergences equivalent to a reassignment of mass of 2%, 5% and 10%, respectively. These are approximately: 0.008, 0.08, 0.3. The effect can be seen as increasing the mass put onto high loss events.*

tributions; the Value at Risk; the Conditional Value at Risk; and the Conditional Expected Loss. These are defined in section 3 and are shown here with the schedules (decisions) aforementioned. Both the loss distributions and the upper-trimmed mean losses (CVaR) are almost indistinguishable. However, the Conditional Expected Loss plot very clearly shows that the expected loss values are driven by low probability events (around 10% of the mass).

The diagnostic plot in figure 10 which based on the theory given in section 4.1 confirms that the decision system is sensitive to small changes in the model. The difference in expected loss under the local least favourable distributions between action (45,1.5) and the optimal action is negative for small KL values (bottom plot, figure 10). Hence small perturbations (in KL divergence) changes the optimality of the actions. This is also apparent from figure 11, where we plot the local pairwise admissibility of the optimal action under π_I (see section 4.1.4). For very small neighbourhoods in KL, alternative actions to that of the optimal Bayes action (at $C = 0$) are no longer inadmissible and should be given serious consideration.

As a final diagnostic plot we look at the probability of optimality under the Dirichlet extension model (section 4.3). Figure 12 shows the probability of each action being optimal with respect to a random DP reweighting of π_I as a function of the expected L_1 distance of the draw (see Section 4.3.1). We see that the probability of optimality of the Bayes action (48, 1.5) rapidly decreases for small increments of L_1 and has probability of optimality less than 1/2 for an L_1 distance greater than 0.01. For larger values of L_1 distance, the action (48,3) becomes most

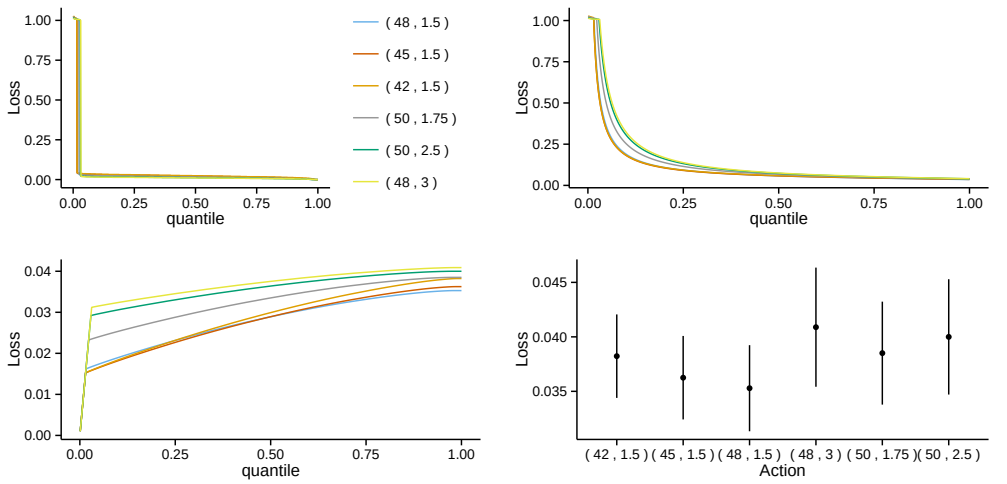

FIG 9. *Model diagnostic plots. From top left to bottom right: inverse loss distributions of the 6 actions (all very close in shape); Upper trimmed mean loss which differentiates the actions by showing the higher downside in some schedules; conditional expected loss; estimates of expected loss centred inside intervals of two standard deviations. We see that the expected loss* ψ_{a,π_I} *for all actions is driven by low probability, high loss events (shape of CEL plot).*

optimal, with all other actions having an almost zero probability of optimality. This shows the lack of decision robustness in this problem, mainly due to flatness of the loss surface.

This application highlights an interesting distinction that must be made when considering model misspecification in a decision-theoretic context. The loss surface is very flat for changes in screening schedule. That is to say, there is little relative difference in expected loss between similar screening schedules. This is also noted by Ruggeri *et al.* (2005) in their analysis of the problem. 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 an action a^*. We discuss this issue further in section 6.

6. CONCLUSIONS

The goal of this article is to assist decision making by providing statistical methods for exploring sensitivity to model misspecification. We hope this will generate further debate and research in this field. The increase in complex high-dimensional statistical analysis problems has driven a corresponding rise in the use of approximate probabilistic techniques. This merits a reappraisal of existing diagnostics and formal methods for characterising the stability of inference to approximate modelling.

The three principles, (P1a, P1b, P2), underpinning the formal methods presented in Section 4 advocate that the neighbourhood should be defined with respect to the marginal distribution on only those elements that enter into the loss function. We showed that to achieve coherence (P2), the Kullback-Leibler divergence is the only measure to use for Γ. Further motivation for using KL is given in Chapters 1 and 9 in Hansen & Sargent (2008). Other advantages stated in the literature for using KL as a divergence measure include

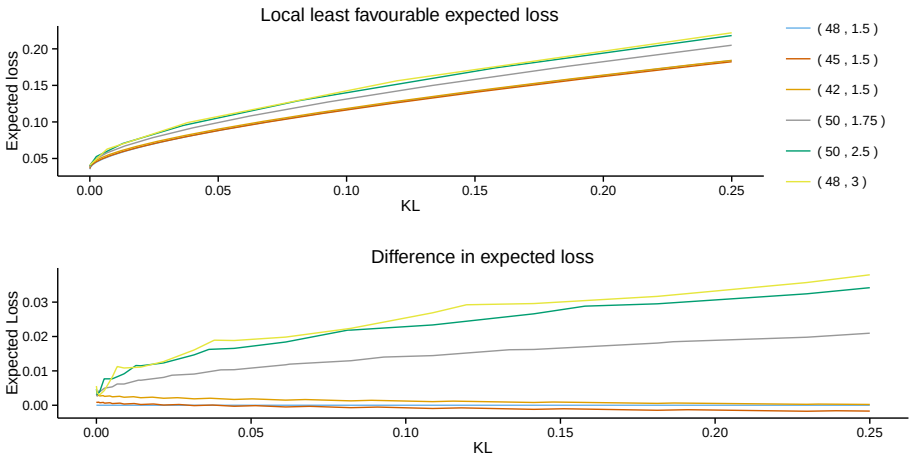

FIG 10. *Diagnostic plots for local least favourable distribution. Top: plot of supremum expected loss versus the size C of the KL neighbourhood; Bottom: difference between the supremum expected loss of each action and that of the optimal action a^*.*

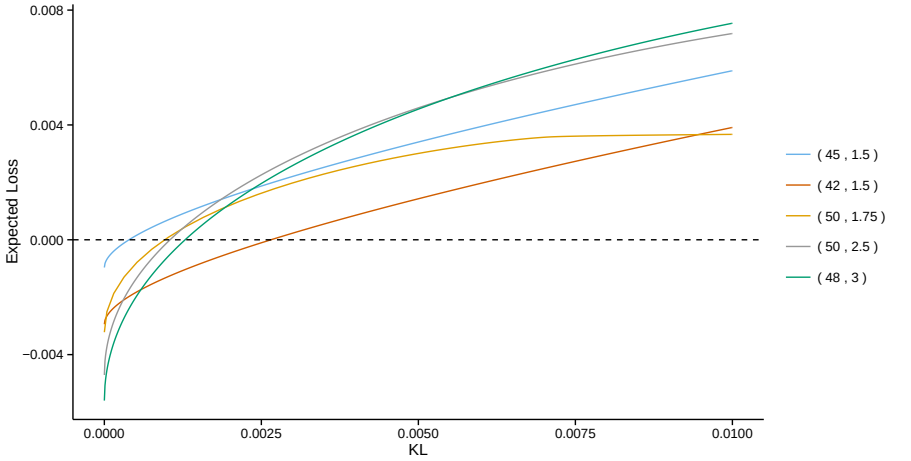

FIG 11. *Local Bayesian admissibility plot.*

- its invariance to re-parametrisation
- its information theoretic representation as the number of bits of information needed to recover p from model q
- its decision theoretic representation as the expected log-loss in using q to approximate p when using proper local scoring rules (Bernardo & Smith, 1994)
- KL bounds the L1 divergence $\mathrm{KL}(p \parallel q) \leq \parallel p - q \parallel_1$.

However, none of this provides a constructive approach for choosing the KL radius C. In chapter 9 of Hansen & Sargent (2008), the authors suggest using detection error probabilities to calibrate the size of the neighbourhood Γ. This stems from the concept of statistically indistinguishable models given a finite data sample of size N. Using model selection principles based on likelihood ratio tests, the user determines a plausible probability (a function of the radius C) of selecting

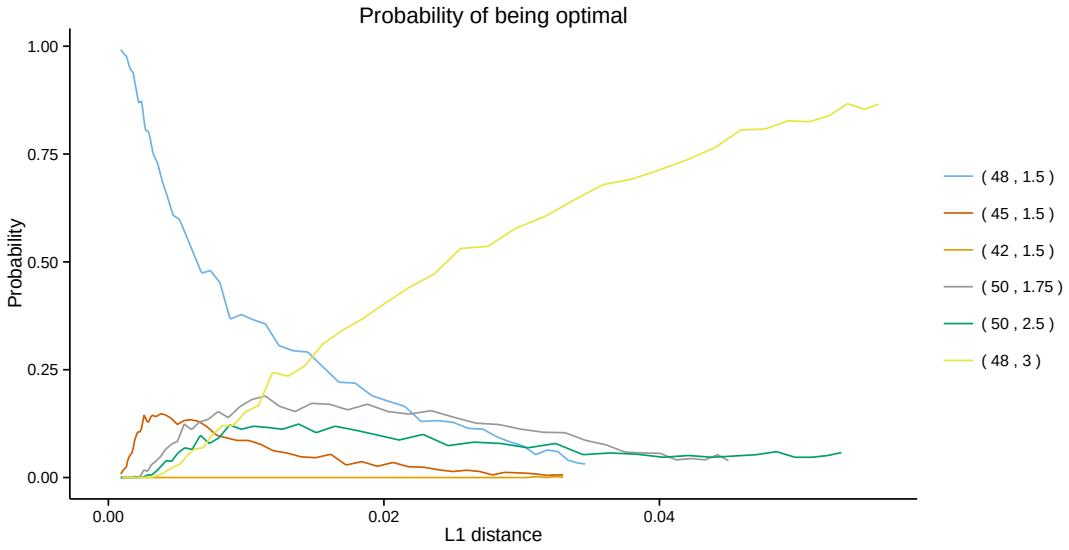

FIG 12. *Probability of being optimal for the top six actions selected by expected loss as a function of the expected L_1 distance of the re-weighted draws. The legend gives the actions ordered by increasing expected loss. We observe the optimal action (under π_I) ceases to be most likely optimal action when the L_1 distance is greater than ≈ 0.015. For draws further that 0.02, the most probable optimal action is (48,3)*

the wrong model given the available data, and then invert this value to find C (by simulation). Although this is a principled method, in many cases even the detection error probability could be difficult to calibrate.

In terms of implementation, we showed that the approaches in Section 4 have simple numerical solutions via re-weighted Monte Carlo samples drawn from the approximating model, using exponentially tilted weights for the local-minimax solution and stochastic Dirichlet weights for the marginal loss distribution. This has the advantage that robustness can be explored *a posteriori* using standard outputs from conventional Bayesian analysis.

Finally, it is important to note the distinction between "decision robustness" and "loss robustness" as discussed in Kadane & Srinivasan (1994). A system is said to be decision robust if perturbations to the model do not effect the optimality of an action $\hat{a}$. On the other hand, it is said to be loss robust, if those perturbations do not effect the overall expected loss of the action $\hat{a}$ (in a relative sense). It is clear that a decision system can have one property without the other. Which is more desirable will be highly context dependent. Throughout the article we have taken the loss function to be known. However, it is clear that loss misspecification is also an important element of robust decision making. Further work is needed to develop a unified approach for dealing with this. Our framework does not take into account misspecification in the loss function. Certain loss functions are often chosen for computational ease or because they posses other desirable properties such as convexity. Also, elicitation of the true loss function can be difficult (for an example see the application discussed in section 5.2). Hence for completeness, a robustness analysis of a decision system should take this into account.

ACKNOWLEDGEMENTS

We would like to thank George Nicholson and Tristan Gray-Davies for many helpful discussions and suggestions; Luis Nieto-Barajas in particular for checking all the results and providing new insights. Watson was supported by an EPRSC grant from the Industrial Doctorate Centre (SABS-IDC) at Oxford University and by Hoffman-La Roche. Holmes gratefully acknowledges support for this research from the Oxford-Man Institute, the EPSRC, the i-Like program grant, the Medical Research Council and the Wellcome Trust.

APPENDIX A: PROOF OF THEOREM 4.2 IN SECTION 4.

Reproduced and amended from (Bissiri *et al.*, 2013).

Assume that Θ contains at least two distinct points, say θ_1 and θ_2. Otherwise, π is degenerate and the thesis is trivially satisfied. To prove this theorem, it is sufficient to consider the case $n = 2$ and a very specific choice for π, taking $\pi = p_0\delta_{\theta_1} + (1 - p_0)\delta_{\theta_2}$, where $0 < p_0 < 1$. Any probability measure absolutely continuous with respect to π has to be equal to $p\delta_{\theta_1} + (1 - p)\delta_{\theta_2}$, for some $0 \leq p \leq 1$. Therefore, in this specific situation, the cost function, $l(\cdot) = \left\{\mathbb{E}_\pi[-L(\theta)] + \lambda^{-1}g(\pi \parallel \pi_I)\right\}$, to be minimised becomes:

$$l(p, p_0, L_I) := p\, L_I(\theta_1) \,+\, (1-p)\, L_I(\theta_2) \\ +\, p\, g\left(\frac{p}{p_0}\right) \,+\, (1-p)\, g\left(\frac{1-p}{1-p_0}\right),$$

where g is a divergence measure, $L_I(\theta_i) = L(\theta_i, I_1) + L(\theta_i, I_2)$ for data $I = (I_1, I_2)$ and $L_I(\theta_i) = L_1(\theta_i, I_j)$ for $I = I_j$, $i, j = 1, 2$. Denote by p_1 the probability $\pi_{I_1}(\{\theta_1\})$, i.e. the minimum point of $l(p, p_1, L_{(I_1,I_2)})$ as a function of p, and by p_2 the probability $\pi_{(I_1,I_2))}(\{\theta_1\})$. By hypotheses, p_2 is the unique minimum point of both loss functions $l(p, p_1, L_{I_2})$ and $l(p, p_0, L_{(I_1,I_2)})$. Again by hypothesis, we shall consider only those functions L_{I_1} and L_{I_2} such that each one of the functions $l(p, p_0, L_{I_1})$, $l(p, p_1, L_{I_2})$, and $l(p, p_0, L_{(I_1,I_2)})$, as a function of p, has a unique minimum point, which is p_1 for the first one and p_2 for the second and third one. The values p_1 and p_2 have to be strictly bigger than zero and strictly smaller than one: this was proved by Bissiri and Walker (2012) in their Lemma 2. Hence, p_1 has to be a stationary point of $l(p, p_0, h_{I_1})$ and p_2 of both the functions $l(p, p_1, L_{I_2})$ and $l(p, p_0, L_{(I_1,I_2)})$. Therefore,

$$g'\left(\frac{p_1}{p_0}\right) - g'\left(\frac{1-p_1}{1-p_0}\right) = L_{I_1}(\theta_2) - L_{I_1}(\theta_1), \tag{10}$$

$$g'\left(\frac{p_2}{p_0}\right) - g'\left(\frac{1-p_2}{1-p_0}\right) = L_{(I_1,I_2)}(\theta_2) - L_{(I_1,I_2)}(\theta_1), \tag{11}$$

$$g'\left(\frac{p_2}{p_1}\right) - g'\left(\frac{1-p_2}{1-p_1}\right) = L_{I_2}(\theta_2) - L_{I_2}(\theta_1). \tag{12}$$

Recall that $L_{(I_1,I_2)} = L_{I_2} + L_{I_1}$. Therefore, summing up term by term (10) and (12), and considering (11), one obtains:

$$g'\left(\frac{p_2}{p_0}\right) - g'\left(\frac{1-p_2}{1-p_0}\right) \\ = g'\left(\frac{p_1}{p_0}\right) - g'\left(\frac{1-p_1}{1-p_0}\right) + g'\left(\frac{p_2}{p_1}\right) - g'\left(\frac{1-p_2}{1-p_1}\right). \tag{13}$$

Recall that by hypothesis (10)–(12) need to hold for every two functions L_{I_1} and L_{I_2} arbitrarily chosen with the only requirement that p_1 and p_2 uniquely exist. Hence, (13) needs to hold for every (p_0, p_1, p_2) in $(0, 1)^3$. By substituting $t = p_0$, $x = p_1/p_0$ and $y = p_2/p_1$, (13) becomes

$$g'(xy) - g'\left(\frac{1-txy}{1-t}\right) \\ = g'(x) - g'\left(\frac{1-tx}{1-t}\right) + g'(y) - g'\left(\frac{1-txy}{1-tx}\right), \tag{14}$$

which holds for every $0 < t < 1$, and every $x, y > 0$ such that $x < 1/t$ and $y < 1/(xt)$. Being g convex and differentiable, its derivative g' is continuous. Therefore, letting t go to zero, (14) implies that

$$g'(xy) = g'(x) + g'(y) - g'(1) \tag{15}$$

holds true for every $x, y > 0$. Define the function $\varphi(\cdot) = g'(\cdot) - g'(1)$. This function is continuous, being g' such, and by (15), $\varphi(xy) = \varphi(x) + \varphi(y)$ holds for every $x, y > 0$. Hence, $\varphi(\cdot)$ is $k \ln(\cdot)$ for some k, and therefore

$$g'(x) = k \ln(x) + g'(1), \tag{16}$$

where $k = (g'(2) - g'(1))/\ln(2)$. Being g convex, g' is not decreasing and therefore $k \geq 0$. If $k = 0$, then g' is constant, which is impossible, otherwise, for any h_I, p_1 satisfying (10) either would not exist or would not be unique. Therefore, k must be positive. Being $g(1) = 0$ by assumption, (16) implies that $g(x) = k\,x \ln(x) + (g'(1) - k)(x - 1)$. Hence,

$$g(\pi_1, \pi_2) = k \int \ln\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right) \mathrm{d}\pi_1$$

holds true for some $k > 0$ and for every couple of measures (π_1, π_2) on Θ such that π_1 is absolutely continuous with respect to π_2.

APPENDIX B: GLOSSARY OF TERMS

Notation	Definition
Θ	Parameter space describing the uncertainty in the 'small world' of interest.
$a \in \mathcal{A}$	Set of actions or alternatives.
$L(a, \theta)$ or $L_a(\theta)$	Loss function defined as mapping $\mathcal{A} \times \Theta \to \mathbb{R}^+$
$L_{(a,a')}(\theta)$	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 uncertainty Θ.
C	The radius of the Kullback-Leibler ball centred at π_I
$\lambda_a(C)$	Exponential tilting parameter given in equation 2 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)
$\tilde{\pi}$, $\tilde{\pi}_I$	The distributions π, π_I approximated by a bag of Monte Carlo samples
$\pi_{a,C}^{\sup}$	The least favourable distribution for action a in the KL ball of radius C centred at π_I
$\psi_a^{\sup}(C)$	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 $L_{(a,a')}(\theta)$

APPENDIX C: LOCAL SENSITIVITY ANALYSIS

We can look at the derivative of least favourable expected loss for a given action either as a function the ball size C or the tilting parameter λ. Firstly, differentiating wrt λ gives:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbb{E}_{\pi_{a,c(\lambda)}^{\sup}}[L_a] = \mathrm{Var}_{\pi_{a,C(\lambda)}^{\sup}}[L_a(\theta)]$$

Setting λ to 0, we see that the 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):

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}C_\lambda = \mathbb{E}_{\pi^{\sup}_{a,C(\lambda)}}[L_a(\theta)] - \mathbb{E}_{\pi_I}[L_a(\theta)]$$

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:

$$\frac{\mathrm{d}}{\mathrm{d}C_\lambda}(K^{-1}) = \frac{1}{\frac{\mathrm{d}K}{\mathrm{d}\lambda}\left(K^{-1}(C_\lambda)\right)}$$

□

APPENDIX D: REVERSE KL NEIGHBOURHOOD: $\mathrm{KL}(\pi_I||\pi)$

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-minimax and maximin distributions. However we can use numerical methods to compute the minimax optimisation. We consider the numerical solution to $\pi^{\sup}_{a,C} = \arg\sup_{\pi\in\Gamma^{\mathrm{rev}}_C}\mathbb{E}_\pi[L_a(\theta)]$, with Γ^{rev}_C now defined herein as $\Gamma^{\mathrm{rev}}_C = \{\pi : \mathrm{KL}(\pi_I \parallel \pi) \leq C\}$ for $C \geq 0$.

Numerical approximation: Consider a stochastic representation of π_I via

$$\begin{aligned} \tilde{\pi}_I &\equiv \frac{1}{m}\sum_{i=1}^{m}\delta_{\theta_i}(\theta) \\ \theta_i &\sim \pi_I(\theta) \end{aligned} \tag{17}$$

where θ_i are iid draws from π_I and $m \to \infty$. In practice this is often the actual model that statisticians work with, via a "bag of samples" Monte Carlo representation of π_I. We note for non-degenerate functionals $g(\theta)$ of interest $\mathbb{E}_{\tilde{\pi}_I}[g(\theta)]$ converges to $\mathbb{E}_{\pi_I}[g(\theta)]$, as $m \to \infty$. To make the solution tractable in defining a KL neighbourhood around π_I we will use the neighbourhood around $\tilde{\pi}_I$. Moreover, in considering the KL divergence between π_I and an alternative model $\pi \in \Gamma_C^{\text{rev}}$ we will work with a stochastic approximation to π represented as mixtures of the atoms $\{\delta_{\theta_1}, \delta_{\theta_2}, \ldots, \delta_{\theta_m}\}$ in (17),

$$\tilde{\pi} = \sum_i w_i \delta_{\theta_i}(\theta) \tag{18}$$

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

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

$$\text{KL}(\tilde{\pi}_I, \tilde{\pi}) = \frac{1}{m}\sum_{i=1}^{m}\log\frac{1}{m * w_i}.$$

with KL ball $\tilde{\Gamma}_C^{\text{rev}}$ defined similarly, $\tilde{\Gamma}_C^{\text{rev}} = \{\tilde{\pi} : \text{KL}(\tilde{\pi}_I, \tilde{\pi}) \leq C\}$.

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

$$\tilde{\pi}_{a,C}^{\sup} = \arg \sup_{\tilde{\pi} \in \tilde{\Gamma}_C^{\text{rev}}} \{\mathbb{E}_{\tilde{\pi}}[L_a(\theta)]\} \tag{19}$$

Given the atomic structure of $\tilde{\pi}$ the maximisation (19) leads to a convex optimisation in the weights,

$$\begin{aligned} \tilde{\pi}_{a,C}^{\sup} &= \sum_i w_i^* \delta_{\theta_i}(\theta) \\ w^* &= \arg\sup_w \left\{ \sum_i w_i L_a(\theta_i) : -\frac{1}{m}\sum_i \log(w_i) \leq C + \log m, \sum_i w_i = 1 \right\} \end{aligned}$$

for which standard numerical methods can be applied.

REFERENCES

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

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

Artzner, Philippe, Delbaen, Freddy, Eber, Jean-Marc, & Heath, David. 1999. Coherent Measures of Risk. *Mathematical Finance*, **9**(3), 203–228.

Autier, Philippe. 2015. Breast cancer: Doubtful health benefit of screening from 40 years of age. *Nature Reviews Clinical Oncology*, **12**(10), 570–572.

Baio, Gianluca, & Dawid, A Philip. 2011. Probabilistic sensitivity analysis in health economics. *Statistical methods in medical research*, 0962280211419832.

Basle Committee. 1996. Amendment to the Capital Accord to Incorporate Market Risks. *Basle Committee on Banking Supervision*.

Beaumont, M.A, Zhang, W., & Balding, D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

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

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

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*. Springer.

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

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

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

Bissiri, P.G., Holmes, C.C., & Walker, S.G. 2013. *A General Framework for Updating Belief Distributions*. http://arxiv.org/abs/1306.6430 *to appear* Journal of the Royal Statistical Society, Series B.

Box, George E.P., & Draper, N.R. 1987. *Empirical model-building and response surfaces*. John Wiley & Sons.

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

Breuer, Thomas, & Csiszár, Imre. 2013b. Measuring Distribution Model Risk. *arXiv preprint arXiv:1301.4832*.

Carota, Cinzia, Parmigiani, Giovanni, & Polson, Nicholas G. 1996. Diagnostic measures for model criticism. *Journal of the American Statistical Association*, **91**(434), 753–762.

Chipman, H.A., George, E.I., & McCulloch, R.E. 1998. Bayesian CART model search. *Journal of the American Statistical Association*, **93**(443), 935–948.

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

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

Del Moral, Pierre, Doucet, Arnaud, & Jasra, Ajay. 2006. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(3), 411–436.

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

Denison, D., Holmes, C.C., Mallick, B., & A.F.M., Smith. 2002. *Bayesian methods for nonlinear classification and regression*. Wiley.

Fearnhead, P., & Prangle, D. 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(3), 419–474.

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

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

Good, I.J. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 107–114.

Grünwald, Peter, & van Ommen, Thijs. 2014. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *arXiv preprint arXiv:1412.3730*.

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

Hansen, Lars Peter, Sargent, Thomas J, Turmuhambetova, Gauhar, & Williams, Noah. 2006. Robust control and model misspecification. *Journal of Economic Theory*, **128**(1), 45–90.

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

Hansen, L.P., & Sargent, T.J. 2001b. Robust control and model uncertainty. *The American Economic Review*, **91**(2), 60–66.

Hansen, L.P., & Sargent, T.J. 2008. *Robustness.* Princeton university press.

Hastie, Trevor, & Tibshirani, Robert. 1993. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hjort, N.L., Holmes, C.C., Muller, P., & Walker, S.G. 2010. *Bayesian nonparametrics.* Cambridge University Press.

Huber, Peter J. 2011. *Robust statistics.* Springer.

Kadane, J.B. (ed). 1984. *Robustness of Bayesian analyses.* Vol. 4. North Holland.

Kadane, J.B., & Chuang, D.T. 1978. Stable decision problems. *The Annals of Statistics*, 1095–1110.

Kadane, J.B., & Srinivasan, C. 1994. Discussion of Berger, J.O., An overview of robust Bayesian analysis – with discussion. *Test*, **3**(1), 116–120.

Kerman, Jouni, Gelman, Andrew, Zheng, Tian, & Ding, Yuejing. 2008. Visualization in Bayesian Data Analysis. *Pages 709–724 of: Handbook of Data Visualization.* Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg.

Løberg, Magnus, Lousdal, Mette L, Bretthauer, Michael, & Kalager, Mette. 2015. Benefits and harms of mammography screening. *Breast Cancer Research*, **17**(1), 63.

Marin, JM., Pudlo, P., Robert, C.P., & Ryder, R.J. 2012. Approximate Bayesian computational methods. *Statistics and Computing*, **22**(6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100**(26), 15324–15328.

Marmot, M.G., *et al.* 2012. The benefits and harms of breast cancer screening: an independent review. *Lancet*, **380**, 1778–1786.

McCulloch, Robert E. 1989. Local model influence. *Journal of the American Statistical Association*, **84**(406), 473–478.

Miller, Jeffrey W, & Dunson, David B. 2015. Robust Bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101.*

Minka, Thomas P. 2001. Expectation propagation for approximate Bayesian inference. *Pages 362–369 of: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc.

Moss, Sue M, Wale, Christopher, Smith, Robert, Evans, Andrew, Cuckle, Howard, & Duffy, Stephen W. 2015. Effect of mammographic screening from age 40 years on breast cancer mortality in the {UK} Age trial at 17 years' follow-up: a randomised controlled trial. *The Lancet Oncology*, **16**(9), 1123 – 1132.

National Research Council:, Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, & Division on Engineering and Physical Sciences. 2013. *Frontiers in Massive Data Analysis.* The National Academies Press.

Parmigiani, Giovanni. 1993. On optimal screening ages. *Journal of the American Statistical Association*, **88**(422), 622–628.

Parmigiani, Giovanni, & Inoue, Lurdes Y. T. 2009. *Decision Theory.* John Wiley & Sons, Ltd.

Pritsker, M. 1997. Evaluating value at risk methodologies: accuracy versus computational time. *Journal of Financial Services Research*, **12**(2-3), 201–242.

Rasmussen, C.E., & Williams, C.K.I. 2006. Gaussian processes for machine learning.

Ratmann, O., Andrieu, C., Wiuf, C., & Richardson, S. 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, **106**(26), 10576–10581.

Rios Insua, D., & Ruggeri, F. (eds). 2000. *Robust Bayesian Analysis.* Springer.

Robbins, H. 1952. Asymptotically Sub-Minimax Solutions of the Compound Decision Problem'in J. *Page 13 of: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.*

Robert, C.P, & Casella, G. 2004. *Monte Carlo statistical methods.* Vol. 319. Citeseer.

Rockafellar, R.T, & Uryasev, S. 2000. Optimization of conditional value-at-risk. *Journal of risk*, **2**, 21–42.

Rostek, M. 2010. Quantile Maximization in Decision Theory. *The Review of Economic Studies*, **77**(1), 339–371.

Rue, Håvard, Martino, Sara, & Chopin, Nicolas. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2), 319–392.

Ruggeri, F., & Wasserman, L. 1993. Infinitesimal sensitivity of posterior distributions. *Canadian*

Journal of Statistics, **21**(2), 195–203.
Ruggeri, F., Ríos Insua, D., & Martín, J. 2005. Robust Bayesian Analysis. *Handbook of statistics*, **25**, 623–667.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Shapiro, S, Venet, W, Strax, P, & Venet, L. 1988. Periodic screening for breast cancer: the Health Insurance Plan project and its sequelae, 1963-1986. *Baltimore, Maryland: The John Hopkins University Press*.
Sivaganesan, S. 1994. Discussion of Berger, J.O., An overview of robust Bayesian analysis – with discussion. *Test*, **3**(1), 116–120.
Sivaganesan, S. 2000. Global and local robustness approaches: uses and limitations. *Pages 89–108 of:* Rios Insua, D., & Ruggeri, F. (eds), *Robust Bayesian Analysis*. Springer.
Varin, Cristiano, Reid, Nancy, & Firth, David. 2011. An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5–42.
Vickers, Andrew J, & Elkin, Elena B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, **26**(6), 565–574.
Vidakovic, B. 2000. Γ-minimax: a paradigm for conservative robust Bayesians. *Pages 241–259 of:* Rios Insua, D., & Ruggeri, F. (eds), *Robust bayesian analysis*. Springer.
Von Neumann, J., & Morgenstern, O. 1947. *The theory of games and economic behavior*. Princeton university press.
Wainwright, M., & Jordan, M.I. 2003. 2003. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1–305.
Wald, A. 1950. *Statistical decision functions*. Wiley.
Walker, Stephen, & Hjort, Nils Lid. 2001. On bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(4), 811–821.
Wasserman, L. 1992. Recent methodological advances in robust Bayesian inference. **4**, 483–502.
Watson, J., Nieto-Barajas, L., & Holmes, C. 2014. *Characterising variation of nonparametric random probability models using the Kullback-Leibler divergence*. Tech. rept.
Whittle, P. 1990. *Risk-sensitive Optimal Control*. Wiley.
Wu, Dongfeng, Rosner, Gary L, & Broemeling, Lyle D. 2007. Bayesian inference for the lead time in periodic cancer screening. *Biometrics*, **63**(3), 873–880.
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.