# Robust (Bayesian) Inference

Chris Holmes

Professor of Biostatistics, University of Oxford, Department of Statistics & Wellcome Trust Centre for Human Genetics

MLSS 2014

# Overview

- Background to the study of robust inference
  - What is "robust inference"?
  - Why is it increasingly important?
- Bayesian approaches to robust analysis
- Heavy-tailed ("outlier prone") likelihood functions
- Robustness to changes in the prior
- Local-minimax actions

# Preamble

- All of statistics, and all of statistical machine learning, is based on assumptions

  "nothing will come from nothing"

- For example, and of special interest here, underlying every probabilistic (Bayesian) machine learning approach is the explicit assumption of a joint probability distribution (the model), e.g. a likelihood function (or sampling distribution if you're Bayesian), for observations

  $$l(\theta) \propto Pr(x|\theta),$$

  where $x$ denotes data and $\theta$ denotes all parameters in the model[1], and a prior

  $$Pr(\theta)$$

[1] I will, rather loosely, use the notation $Pr(\theta)$, $p(\theta)$, $\pi(\theta)$ inter-changeably to denote either a distribution, density or mass function; making clear if the context dictates

# Preamble

- The term 'assumptions' is somewhat vague, but for our purposes it will suffice to cover statements and issues such as,
  - Normality
  - Independence and identically distributed (i.i.d.)
  - Linearity
  - Stationarity
  - Parametric form of likelihood
  - Prior (for Bayesian models)
  - Data ascertainment and stability of training data to real-world application environment

- and by 'robust' we shall take to mean

  *"....an insensitivity to small deviations from the assumptions." Huber & Ronchetti (2009)*

- In statistics, the word was coined by Box (1953) when analysing the effect of non-gaussian data when testing equality of variances
- It is important to note that robustness is *contextual* to the question being addressed
  - e.g. the Student distribution is robust to inference on location (mean) to non-Gaussian data, but is not robust to loss of power in the t-test in this scenario
  - in fact, in a Bayesian framework, I believe that decision theory is the appropriate framework to study robustness

# Preamble

- 'Robustness' is a property that statisticians (machine learners) should naturally seek out
  - somewhat vacuous, as who would actively choose to be non-robust??
- In the words of Kadane (1984a):

  *"Robustness is a fundamental issue for all statistical analyses; in fact it might be argued that robustness is the subject of statistics."*
- this viewpoint aligns well with the notion of Statistics as precision in imprecision
- Indeed, good statistical practice advocates the use of all available tools including diagnostic plots, and graphical visualisation to protect against overconfident inference and prediction

# Preamble

- In what follows, we shall see that by seeking 'robustness' we will be lead to select adopt one probability model, or class of probability models, over another one, or one procedure over another
- Following Anscombe (1960), see also Huber (1972), we can view the adoption of a robust model as a kind of insurance premium,

  > *"I am willing to pay a premium (a loss of efficiency of, say, 5 to 10% at the ideal model) to safeguard against ill effects caused by small deviations from it"*

- So by robustness we will mean a trade off, I loose a (little) bit of statistical efficiency versus an original model if true, in order to protect myself against large misjudgements if the original modelling assumptions are false

# Historical context

- Historically, robustness was studied in the context of estimates of central-location and spread
- Many eminent applied statisticians understood the danger of blindly assuming normal (Gaussian) errors in physical datasets, e.g. both Bessel (1818) and Newcomb (1886) noticed longer tails

  Newcomb speaking of the Gaussian (Normal) distribution, page 343: "As a matter of fact, however, the cases are quite exceptional in which errors are found to really follow the law. The general rule is that much more than one per cent of the errors exceed four times the probable error. In other words, it is nearly always found that some of the outstanding errors seem abnormally large."
- or the following remark from an anonymous author (Anonymous, 1821) (translated, page 189) on an early form of $\alpha$-trimmed means: "...there are certain provinces in France, where, in order to determine the average yield of an estate, it is custom to consider the yields over a period of twenty consecutive years, remove the greatest and the smallest of these numbers, and then to use the 18 remaining to calculate the mean. Those who imagined this method, must have no doubt considered that very abundant harvests and very poor harvests were exceptions from the usual course of Nature, and thus one should not make allowance for them."

- The astronomer Eddington used the mean absolute deviation, ($L_1$ norm) $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$, instead of the standard deviation ($L_2$ norm) to record variance claiming that it was a more accurate estimate of the variation in the data
  - Fisher (1920) showed that this lost 12% efficiency (as measured by asymptotic relative efficiency)
- Thus, the dangers of the "dogma of normality" (as Huber called it) were not brought to light until Tukey (1960) looked at the mixture model:

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

and showed that for values of $\epsilon$ as small as 0.002 the standard deviation is no longer better than the mean deviation.

- Huber (1964) was the first to provide a theoretical approach to robust statistical inference. Deviations from normality motivated his initial work on location and scale but this generalised to wider models, see Huber & Ronchetti (2009)
- Note: Huber's perspective is that robustness differs from *nonparametric* and *distribution-free* models. For instance, the sample mean is distribution-free but highly sensitive to outliers. Robustness relates to parametric models having the following desirable properties:
  - Efficiency: at assumed model (ideal model)
  - Stability: with respect to small perturbations to assumptions, only impairing performance slightly.
  - Breakdown protection: larger deviations should not cause catastrophe.

# What we will cover

- In these two lectures we will not spend much time on the classical field of "robust statistics", for which the interested researcher is referred to Huber & Ronchetti (2009)
  - e.g. so we shall not cover important areas of M-estimators, estimating equations, sandwich estimators, leverage plots, breakdown numbers,....
- Rather we shall focus on robust probabilistic modelling from a Bayesian perspective
- In one sense you could argue that Bayesian approaches are inherently robust as they don't condition on a single model but rather seek to accommodate all aspects of uncertainty within a joint probability model, $p(x, \theta)$, usually factorized as

$$\begin{aligned} p(x, \theta) &= p(x|\theta) \times p(\theta) \\ &= \text{Likelihood} \times \text{Prior} \end{aligned}$$

where $p(\theta)$ may cover a whole space of probability models deemed plausible *a priori*

- Inference and prediction is made through the posterior distribution, $p(\theta|x)$, accommodating the uncertainty under the support of the prior,

$$T = \int \psi(\cdot;\theta)p(\theta|x)d\theta$$

- Under a strict Bayesian position there is no issue with model robustness
  - You precisely specify your subjective beliefs through $p(x,\theta)$ 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.....

"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). In which case what does the resulting posterior distribution $p(\theta|x)$ actually represent?

- The standard solution is to first specify an operational model $p(x, \theta)$, to the best of your available time and ability, and then investigate sensitivity of inference or decisions to departures around $p(x, \theta)$, typically assuming that $p(x|\theta)$ 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), Berger (1994), Wasserman (1992), and Ruggeri *et al.* (2005), as well as the collection of papers in the edited volumes of Kadane (1984b) and Rios Insua & Ruggeri (2000).

- Robustness was one of the most active research areas in Bayesian statistics in the 1990s and earl 2000s, following which interest tailored off – principally as computational methods (MCMC) and hierarchical and nonparametric models allowed one to apply evermore complex models – so that the methods outpaced the data

- However, there are a number of recent (and long standing) developments that merit a reappraisal

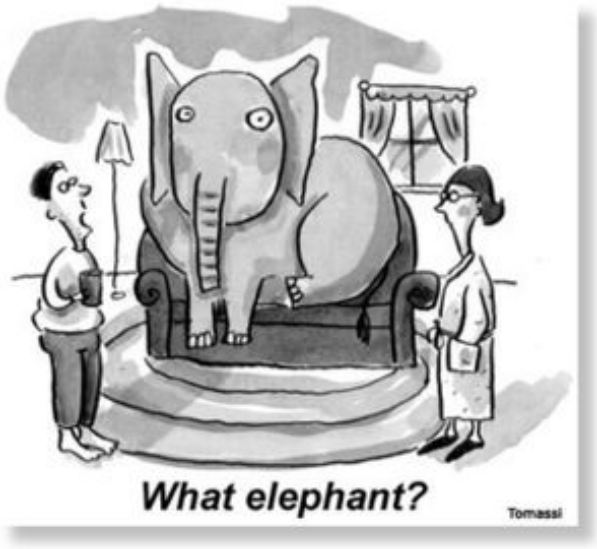
# Rise of approximate models

- In order to address the increasing complexities of modern data analysis there has been a rise in the development and use of computationally efficient approximate models
  - INLA
  - Composite Likelihoods
  - ABC
  - Variational Bayes
  - PAC likelihoods
  - Gibbs posteriors
  - Monte Carlo representations
- These models are approximate by design, but they are often the only applicable methods we have
- But what are the consequences of these approximations?
  - how robust is the inference to the inherent model misspecification?
  - if interest is in the science (rather than pure prediction) what does the posterior $p(\theta|x)$ represent?
  - what does the in hand "bag-of-samples" $\theta_i \sim p(\theta|x)$ refer to?
- In an era of big-data and big-models we can't check all assumptions

# M-closed

- Bayesian statistics is conditional on the model being true
  - the so called M-closed perspective (Bernardo & Smith (2009))
- Formally, in order to update you have to assume that the true likelihood, $p(x|\theta)$, is contained under the support of the prior
- and....

  All of Bayesian statistics is model based



# Bayesian robust priors

- Robustness to misspecification in Bayesian inference is not a new topic
- Rich literature in the 1980s and early 1990s
- Robust Bayesian analysis was one of the major research themes in the field
  - principally investigating robustness in the prior (assuming the likelihood is known)
    - as you alter the likelihood then the meaning of the prior changes
  - excellent reviews in Berger (1994; 1985 )

- The field tailored off somewhat thereafter as methods and computation outpaced the complexity of data sets
- MCMC, hierarchical models, Bayes nonparametrics all allowed for indexing of richer and richer model classes
- However, big-data has now overtaken these methods (leading to the rise of approximate models)
- This merits a reappraisal of Bayesian robustness in the modern context
  - incorporating interesting recent developments in robust control, macroeconomics and finance literature

# Let's begin with robust likelihoods....

# Robust Linear Models

- It is well known that the Normal linear model is sensitive to outliers

$$\begin{aligned} y &= X\beta + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I) \end{aligned}$$

- Conventional approach is to robustify the likelihood (sampling density) from Normal to say Student

$$\begin{aligned} y &= X\beta + \epsilon \\ \epsilon &\sim t_\nu(0, \sigma) \end{aligned}$$

- Note:
  - The standard non-Bayesian approach would be to use a robust estimator for $\beta$, such as M-estimation and Huber likelihood
  - Bayesian analysis is more complicated as you have to define a joint probability model, you're not free to just define an estimator – which is why robustness is particularly important to Bayesian analysis

# Outlier Prone Likelihoods

- In classical statistics we talk of "robustness to outliers" (unusual observations)
- In Bayesian statistics we talk of "outlier prone" sampling distributions (likelihoods) (Neyman & Scott (1971), O'Hagan (1979)); rather than "outlier resistant"
- Bayesian models are generative, hence you want a model (sampling distribution) that places reasonable probability on generating an outlying observation
- Robust Bayesian inference dates back to work of de Finetti (1961) (the founder of modern Bayesian statistics) who pointed out that outliers should not be rejected but rather probabilistically down weighted

- de Finetti (1961) considered the following sampling model for the location parameter, $\mu$, of the Normal

$$y_i \sim N(\mu, \lambda_i \sigma^2)$$

where each observation has its own variance parameter, distributed Inverse-Gamma,

$$\lambda_i \sim IG(\nu/2, \nu/2)$$

"each observation is taken using an instrument with normal error, but each time chosen at random from a collection of instruments of different precisions, the distribution of the precisions being that indicated (by the mixing distribution)" (de Finetti, 1961)

- We can clearly generate samples according to the above joint distribution through the following simple algorithm
  1. Draw $\lambda_i \sim IG(\nu/2, \nu/2)$
  2. Draw $y_i|\lambda_i \sim N(\mu, \lambda_i\sigma^2)$

- We can plot out the joint samples $\{y_i, \lambda_i\}$, for various values of $\nu$, say $\nu \in \{2, 10, 100\}$ with $\mu = 0$ and $\sigma^2 = 1$

$\nu = 2$

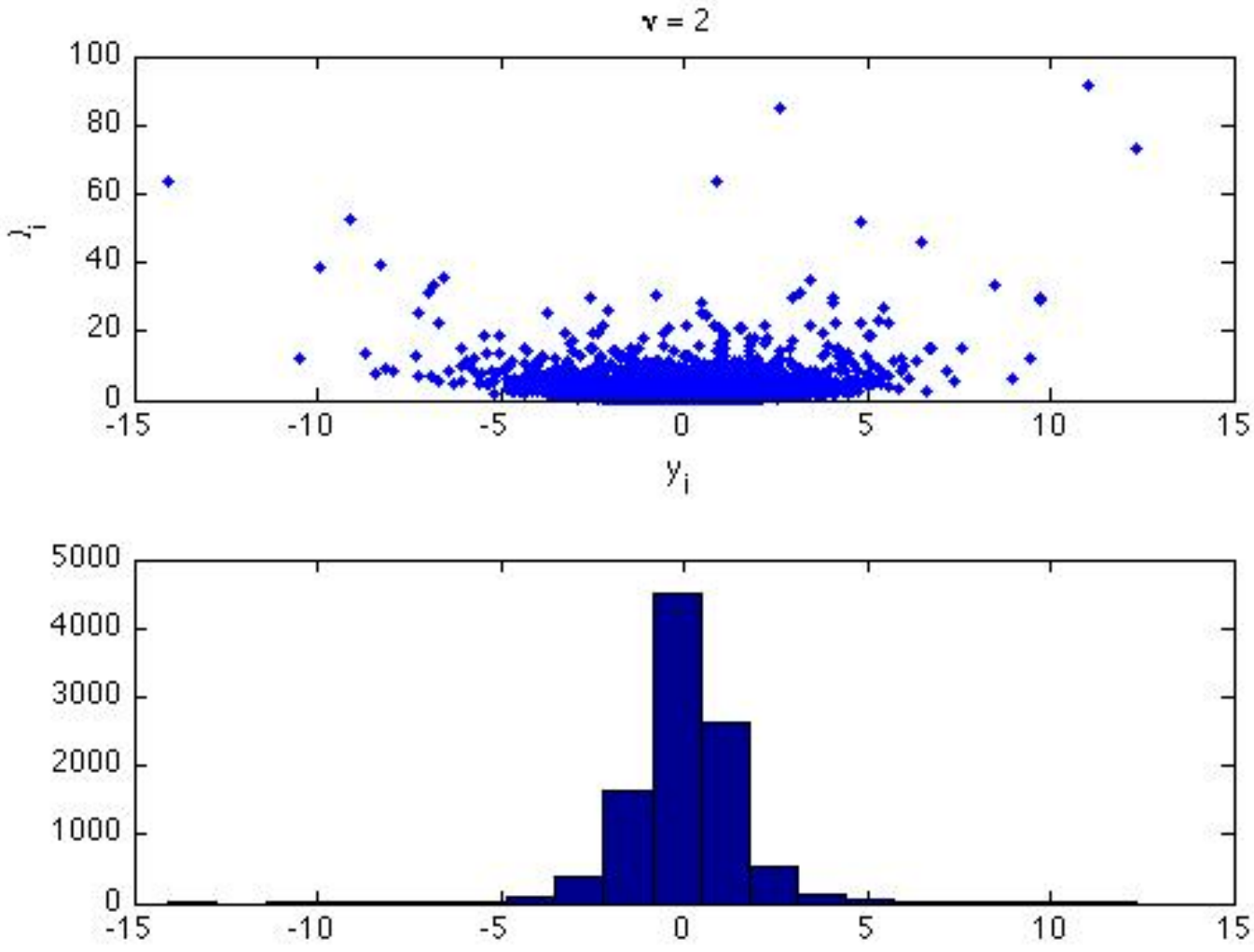



Figure: Sample of $\{y_i, \lambda_i\}$ for $\nu = 2$, $\lambda_i \sim IG(1, 1)$

# $\nu = 10$

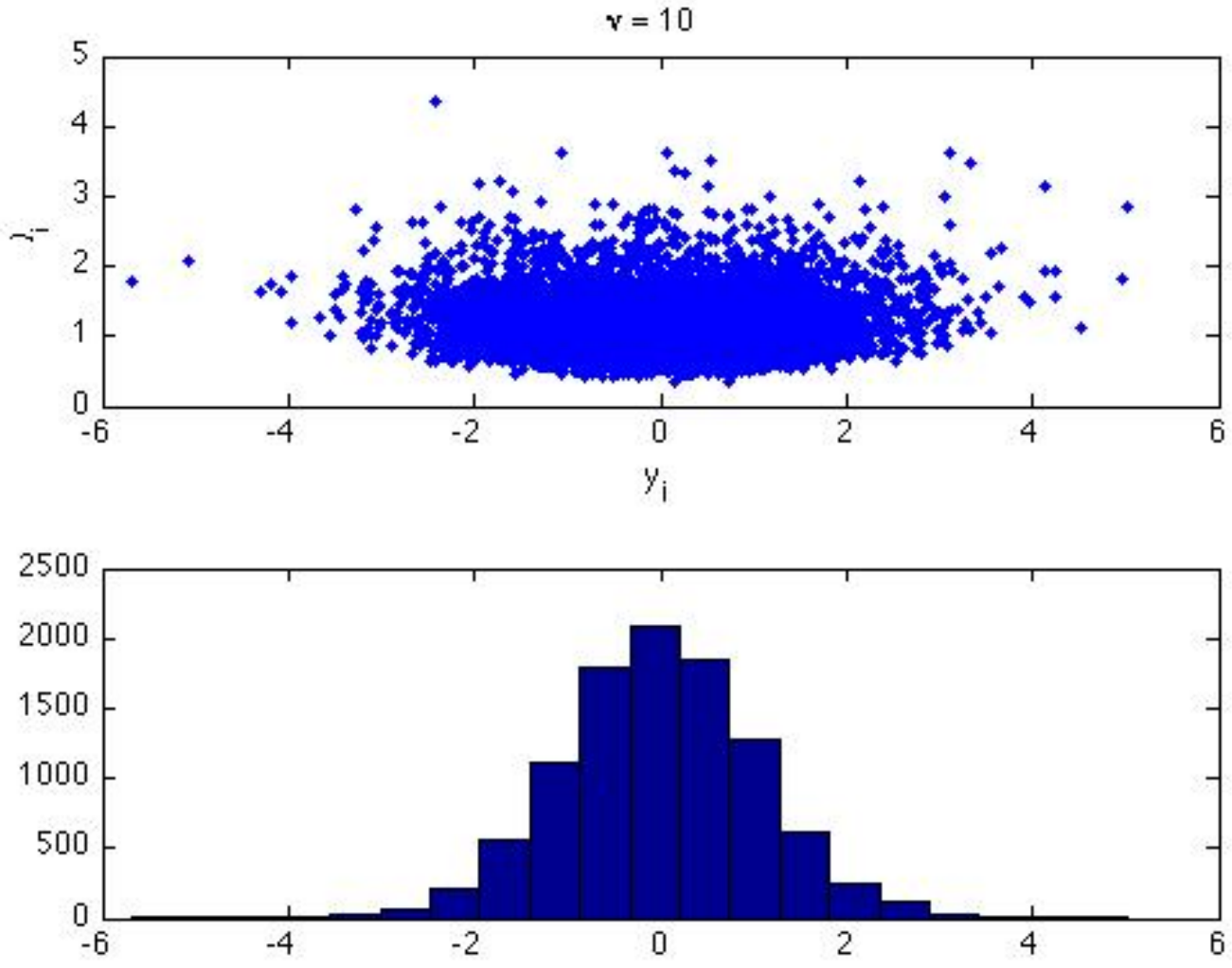



Figure: Sample of $\{y_i, \lambda_i\}$ for $\nu = 10$, $\lambda_i \sim IG(5, 5)$

# $\nu = 100$

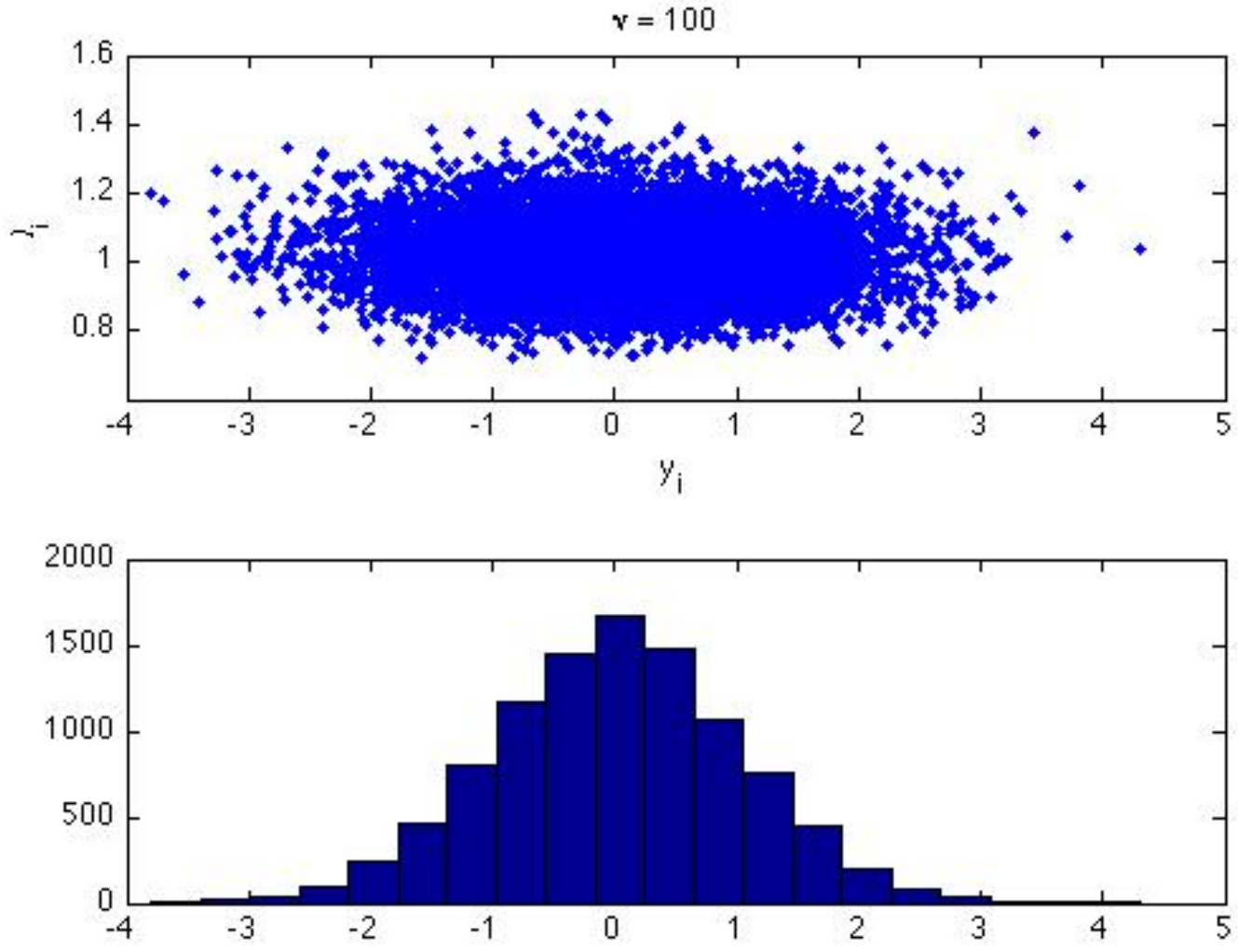



Figure: Sample of $\{y_i, \lambda_i\}$ for $\nu = 100$, $\lambda_i \sim IG(50, 50)$

# qq-plot of $\{y_i^{(\nu=10)}, y_i^{(\nu=100)}\}$

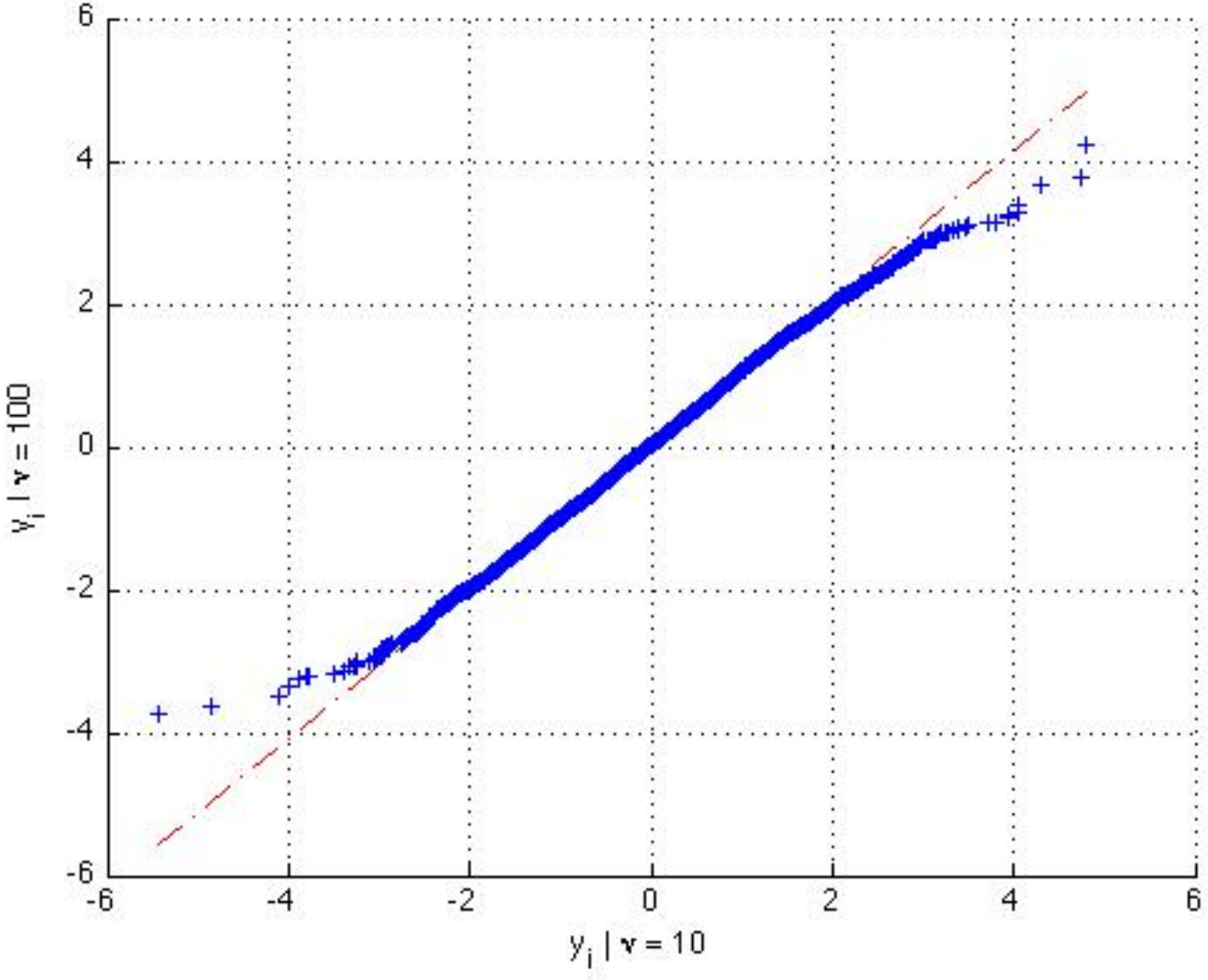



Figure: qq-plot of (sorted) samples from $y_i|\nu = 10$ versus $y_i|\nu = 100$

In fact it is well known that the hierarchical model,

$$\begin{aligned} y_i &\sim N(\mu, \lambda_i \sigma^2) \\ \lambda_i &\sim IG(\nu/2, \nu/2) \end{aligned}$$

defines a marginal distribution $\pi(y_i) = \int_\lambda \pi(y_i|\lambda)\pi(\lambda)d\lambda$,

$$y_i \sim St_\nu(\mu, \sigma^2)$$

a Student density with $\nu$ degrees of freedom

- So that the "robust" Student density naturally arises by thinking of each observation as having some additional individual component of variance say due to individual experimental heterogeneity, de Finetti (1961)

# Characterising Influence via Score Functions

- More formally, score functions provide a really useful method to examine model sensitivity to observations
- Suppose that our current beliefs about parameters $\theta$ are summarized by $p(\theta)$
- The influence of a new observation, drawn from $p(x \mid \theta)$, on beliefs about a particular component $\theta_j$ can be usefully measured through the posterior score function [West (1984), Haro-Lopez & Smith (1999)],

$$\frac{\partial}{\partial \theta_j} \log p(\theta \mid x) = \frac{\partial}{\partial \theta_j} \log p(\theta) + \frac{\partial}{\partial \theta_j} \log p(x \mid \theta).$$

- The dependence of the posterior score function on the data $x$ is through the *efficient score function* $\frac{\partial}{\partial \theta_j} \log p(x \mid \theta)$

- Clearly

$$\frac{\partial}{\partial \theta_j} \log p(\theta \mid x) = \sum_i \frac{\partial}{\partial \theta_j} \log p(\theta \mid x_i)$$

  which highlights the additive contribution of the $i$'th observation, $x_i$, to the $j$'th partial derivative of the log-likelihood function evaluated at $\theta$

- The key point to note is that at the posterior mode, maximum a posteriori (MAP), we have

$$\sum_i \frac{\partial}{\partial \theta_j} \log p(\widehat{\theta} \mid x_i) = 0$$

- That is, at the MAP, $\widehat{\theta} = \arg\max_\theta p(\theta|x)$, the sum of the individual scores must equal 0

# Efficient score functions

- Intuitively we can see that if one of the scores is large, then this will influence the MAP estimate, which must move (so to speak) in order to position itself such that, by definition, $\sum_i \frac{\partial}{\partial \theta_j} \log p(\widehat{\theta} \mid x_i) = 0$
- The MAP is at the balancing point of the score function realisations
- So by examining the form of the score functions for different sampling distributions (models) we can see how outlier prone a distribution is
  - Remember – "outlier prone" is a good thing here

# Score Function examples

Consider a standardized observation $z$, which is mean zero'd and standard deviation 1, $z(x, \mu, \sigma) := (x - \mu)/\sigma$

We can look at the relative influence on the central location (mean) and scale (variance) of an observation under the Gaussian and Student distributions

- Gaussian, $N(x|\mu, \sigma^2)$:
  - location efficient score

$$\frac{\partial}{\partial \mu} \log \left[ N(x \mid \mu, \sigma^2) \right] = \frac{z}{\sigma}$$

  - scale efficient score

$$\frac{\partial}{\partial \sigma} \log \left[ N(x \mid \mu, \sigma^2) \right] = \frac{z^2 - 1}{\sigma}$$

# Gaussian efficient score functions

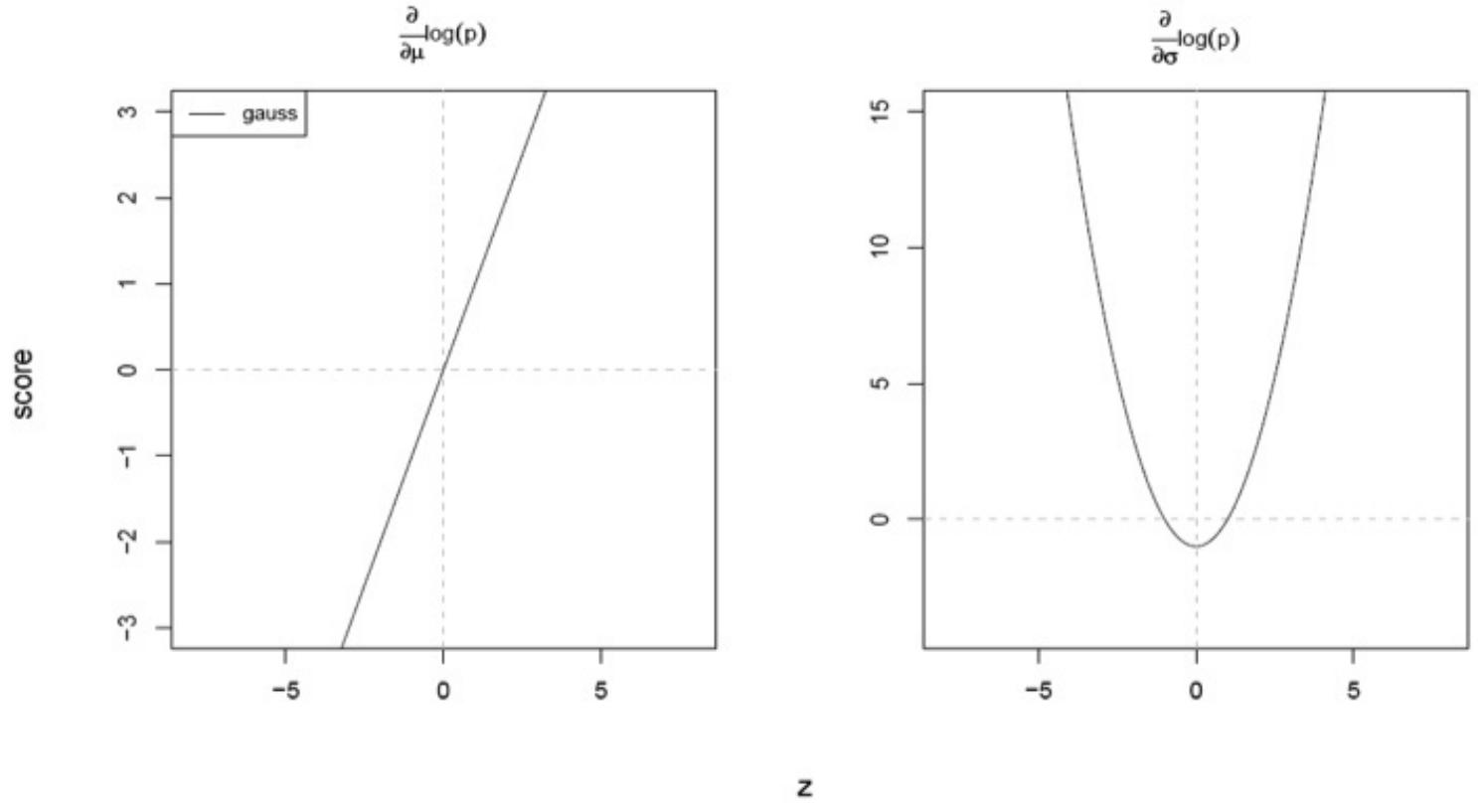
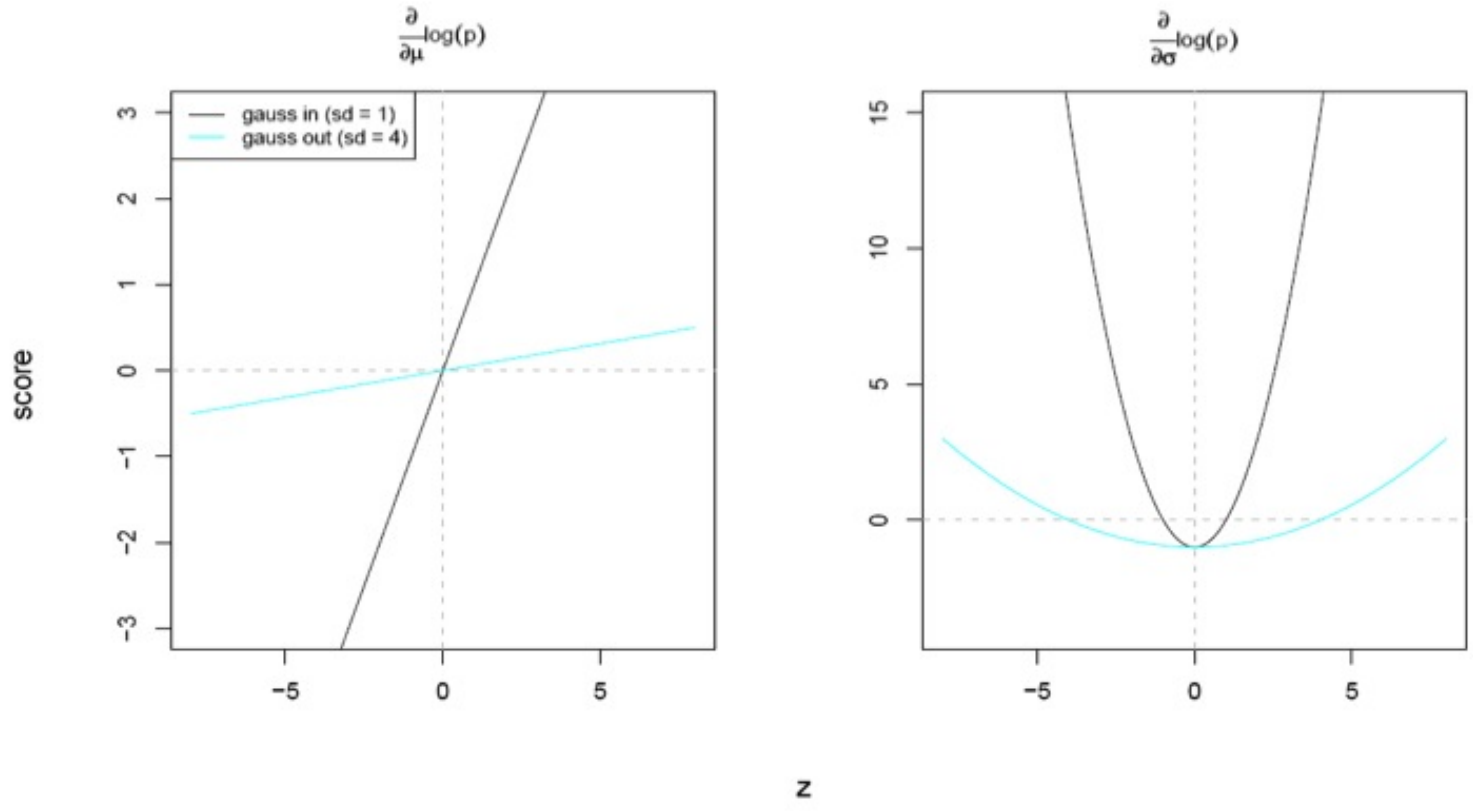



Figure: efficient score functions for location and scale of a Gaussian

# Gaussian efficient score functions for $\sigma = 1$ and $\sigma = 4$




Note:

- the influence (slope) on the location is $\sigma^{-1}$
- observations, $|x| \leq \sigma$, have a negative score on $\frac{\partial}{\partial \sigma} \log p(\sigma|x)$, and those $|x| > \sigma$ a positive

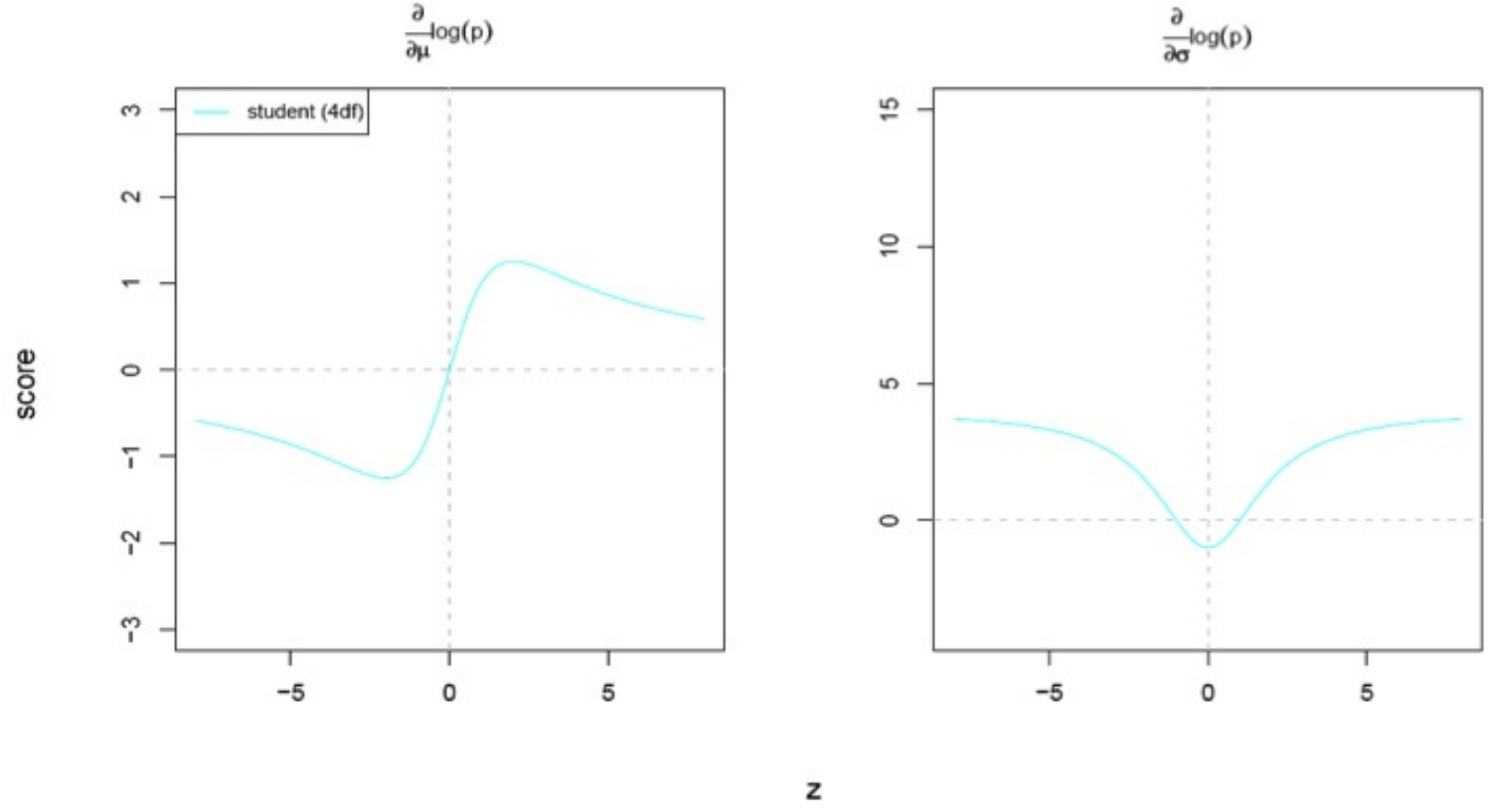
- Student, $t_\nu(x \mid \mu, \sigma^2)$:
  - location efficient score

$$\frac{\partial}{\partial \mu} \log \left[ t_\nu(x \mid \mu, \sigma^2) \right] = \frac{1}{\sigma} \left( \frac{(\nu+1)z}{\nu + z^2} \right)$$

  - scale efficient score

$$\frac{\partial}{\partial \sigma} \log \left[ t_\nu(x \mid \mu, \sigma^2) \right] = \frac{1}{\sigma} \left( \frac{(\nu+1)z^2}{\nu + z^2} - 1 \right)$$

# Student df$= 4$, $\sigma = 1$, efficient score functions




For the Student we note that

- $\frac{\partial}{\partial \mu} \log \left[ t_\nu(x \mid \mu, \sigma^2) \right] \to \frac{1}{\sigma} \left( \frac{\nu+1}{z} \right)$ as $z \to \infty$
- $\frac{\partial}{\partial \sigma} \log \left( t_\nu(x \mid \mu, \sigma^2) \right) \to \frac{\nu}{\sigma}$ as $|z| \to \infty$.

# Gaussian and Student efficient score functions compared

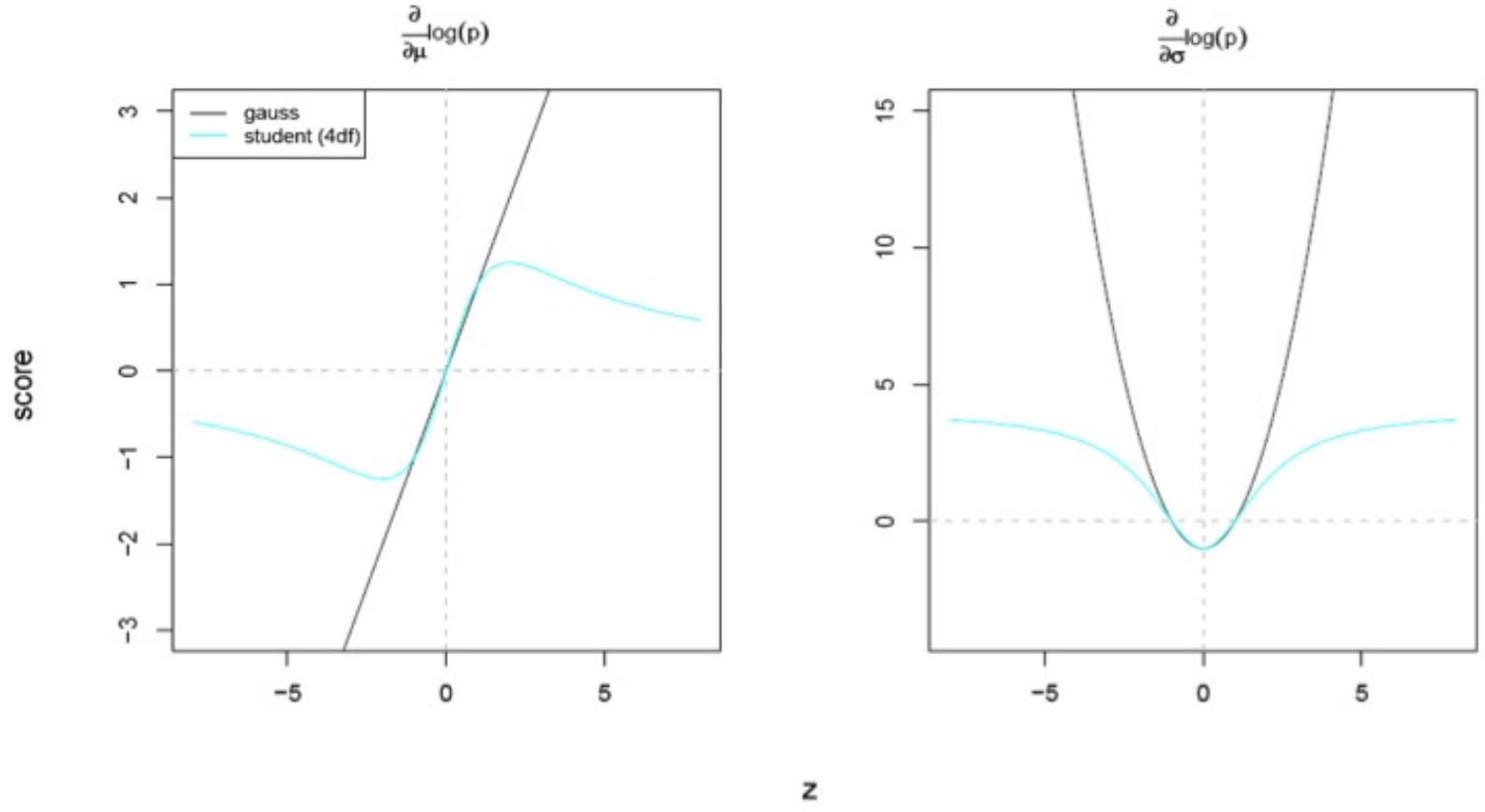



Figure: efficient score functions overlaid for location and scale of a Normal and Student-4

# Robust Bayesian Analysis (to the prior)

- In his excellent review Berger (1994) defines

  *"Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs."*

  - for other good overviews of robust Bayesian analysis see Rios Insua & Ruggeri (2000) and Ruggeri *et al.* (2005)

- Bayesian analysis involves the joint specification of $p(x|\theta)$ and the prior $p(\theta)$
- The methods discussed above gives us tools to think about robust outlier prone sampling distributions (likelihoods) so what about tools for robust prior specification?

# Robust prior specification

- The majority of papers dealing with formal methods for robust Bayesian analysis are concerned with prior specification
- This is not so surprising as
  - in many situations it may be felt that priors are more uncertain than the sampling distribution (likelihood function)
  - the major objection/contrast to a non-Bayesian approach is in the use of the prior
  - the prior is defined with reference to the parameters in the model (likelihood), so changing the likelihood would necessitate a change in the prior anyhow
- One way to deal with sensitivity in the prior is to construct default (automated) specifications that have minimal influence (in some sense) on the answers, e.g. so called noninformative, reference, and maximum entropy priors
- A difficulty here is that what is a noninformative prior for one question might be highly influential for another

# Sensitivity to your operational prior specification

- Perhaps the clearest way, as mentioned in the introduction, is to progress with an operational prior, $p_0(\theta)$, specified to the best of your time and effort constraints
- Then explore sensitivity of "answers" to perturbations around the prior via a class of distribution functions
- Note:
  - most "answers" or posterior quantities of interest can be written as functionals, $\psi$, which are expectations with respect to the posterior model

    $$\psi = \int g(\theta) p(\theta|x) d\theta$$

    e.g. posterior mean; credible intervals; predictions; quantiles;.....
- 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
- We shall focus on global approaches.....

# Global prior robustness

- Global approaches consider variation in a posterior functional of interest, $\psi$, within a neighbourhood

$$p(\theta) \in \Gamma$$

where $\Gamma$ is a ball (or class) of distribution functions around the operational prior model $p_0$

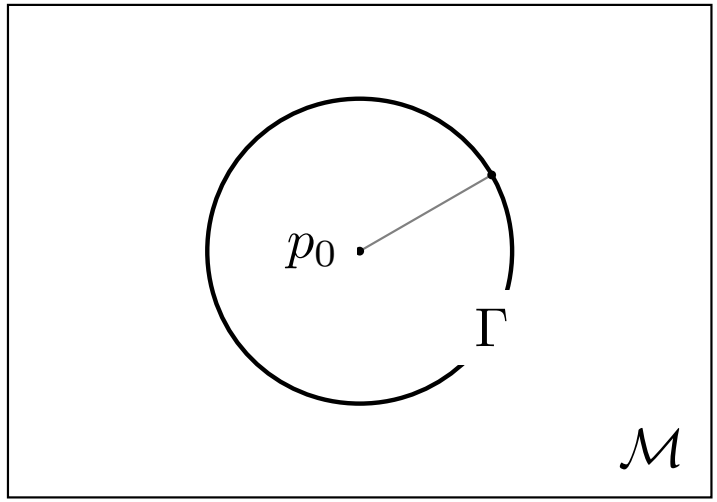



Figure: Graphical representation of neighbourhood of alternative priors constructed around the operational prior $p_0$

- A typical quantity for sensitivity analysis would be the range $(\psi_{\inf}, \psi_{\sup})$ where

$$\psi_{\inf} = \inf_{\pi \in \Gamma} \int g(\theta) p(x|\theta) \pi(\theta) d\theta$$

and

$$\psi_{\sup} = \sup_{\pi \in \Gamma} \int g(\theta) p(x|\theta) \pi(\theta) d\theta$$

- The challenge is to define the class of priors (the nature and size of $\Gamma$) so as to capture plausible ambiguity in $p_0$, while taking into account factors such as ease of specification and computational tractability, Berger (1994; 1985 section 4.7)

# Choosing the neighbourhood class of priors, $\Gamma$

- The art and science of choosing good classes of priors includes
  - the class should be easy to elicit and interpret
  - easy to handle computationally
  - be wide enough to cover reasonable uncertainty, but not so wide as to be implausible
  - adaptable to high dimensions and possible constraints
- One important example is the $\epsilon$-*contamination* neighbourhood (Berger & Berliner, 1986) formed by the mixture model,

$$\Gamma = \{\pi = (1-\epsilon)p_0 + \epsilon q, q \in \mathcal{Q}\},$$

where $\epsilon$ is the perceived contamination error in $p_0$ 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
- A nice feature of the $\epsilon$-contaminated class of priors is that the posterior has a mixture form

$$p_\Gamma(\theta|x) = w(x)p_0(\theta|x) + [1 - w(x)]q(\theta|x)$$

with weights

$$w(x) = 1 + \frac{\epsilon m(x|q)}{(1-\epsilon)m(x|p_0)}$$

where $m(x|\cdot)$ denotes the marginal (integrated) likelihood; prior predictive, or evidence

- Other approaches consider frequentist risk, such as $\Gamma$-*minimax* that investigates the minimax Bayes (frequentist) risk of $\psi_{\text{sup}}$ for $\pi \in \Gamma$ whereas *conditional* $\Gamma$-*minimax* procedures (Vidakovic, 2000) study the maximum expected loss across prior distributions within $\Gamma$

# Robust control and econometrics

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

- 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; see Hansen & Sargent (2008) for a thorough review and references.

- H&S defined a robust action as a local-minimax act within a Kullback-Leibler (KL) neighbourhood of the posterior $\pi_I \equiv \pi(\theta|x)$ through exploration of,

$$\psi_{\sup}^{(a)} = \sup_{\pi \in \Gamma} \int L(\theta)\pi(\theta|x)d\theta$$

  where $\Gamma$ denotes a KL ball around $\pi(\theta|x)$,

$$\Gamma = \{\pi : \int \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi(\theta|x)}\right) d\theta \leq C\}.$$

  and $L(\theta)$ is a real-valued loss function
  - e.g. $L(\theta)$ could be the weighted prediction error, although more formally it is the loss in taking the action if the unknown state happened to be $\theta$

# Ex-post robustness

- So consider $\pi(\theta|x)$ as a useful, but approximate model
- Investigate robustness via the minimax (worst possible outcome) distribution, $\pi_a^{(\sup)} \equiv \pi_a^*$, constrained to be within some neighbourhood, $\Gamma$, around $\pi(\theta|x)$

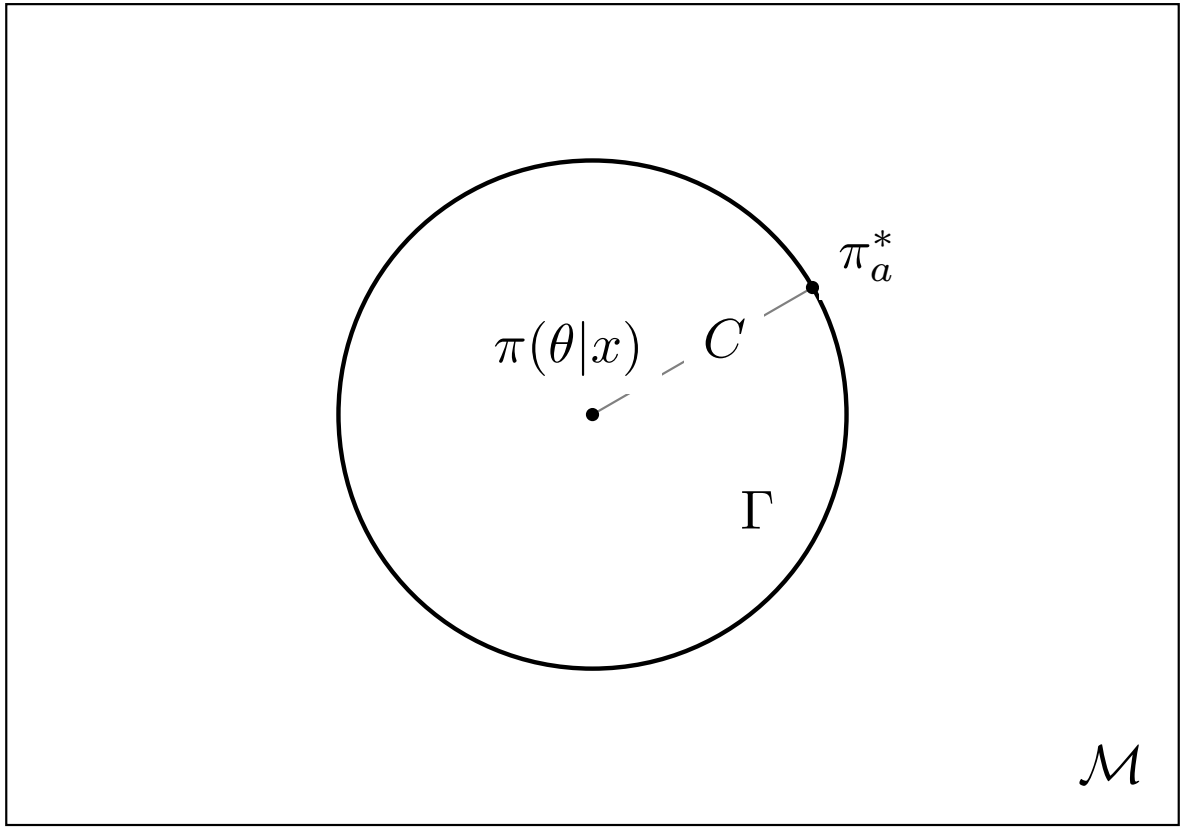



Figure: Graphical representation of local-minimax distribution $\pi^*$ within a ball of radius $C$ around the reference model $\pi_I$

- H&S considered the Gaussian dynamic state-space model (Kalman filter). Breuer & Csiszár (2013a, 2013b), building on the work of Hansen and Sargent, derived corresponding results for arbitrary probability measures. Under mild regularity conditions, and using results from exponential families and large deviation theory they obtain the exact form of $\pi_a^*$ for any $\pi(\theta|x)$ given the KL ball of size $C$, as well as an estimate for $\psi_{\mathrm{sup}}^{(a)}$, see also Ahmadi-Javid (2011, 2012)
- In fact there is a fairly simple form (and proof) for the worst possible local model

# Decision robustness

- The local-minimax distribution is defined as

$$\pi_a^{(\sup)} = \arg\sup_{\pi \in \Gamma} E_\pi[L_a(\theta)]$$

  with expected loss $E_{\pi_a^*}[L_a(\theta)]$, where $L_a(\theta)$ quantifies the loss (negative reward) you will receive if the true state is $\theta$
- In particular (for many good reasons) we take the Kullback-Leibler (KL) neighbourhood $\Gamma$ around $\pi(\theta|x)$
  - where $\Gamma \subset \mathcal{M}$ is the KL ball of distributions around $\pi(\theta|x)$

$$\Gamma = \{\pi : \int \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi(\theta|x)}\right) d\theta \leq C\}.$$

- Somewhat remarkably, under mild regularity conditions, and using results from exponential families and large deviation theory, Breuer & Csiszár obtained the exact form of $\pi_a^*$ for any $\pi(\theta|x)$

# Form of local-minimax distribution

In fact there is a less general but simple(r) proof for the form of the local-minimax distribution

## Theorem

*Let*

$$\pi_a^* = \arg\sup_{\pi \in \Gamma} E_\pi[L_a(\theta)]$$

*be the local-minimax distribution for* $\Gamma = \{\pi : KL(\pi \parallel \pi_I) \leq C\}$

*Then* $\pi_a^*$ *has the following form,*

$$\pi_a^*(\theta) \quad = \quad \frac{e^{\lambda L_a(\theta)} \pi_I(\theta)}{\int e^{\lambda L_a(\theta)} \pi_I(\theta) d\theta}$$

*where* $\lambda$ *a non-negative real valued scalar monotone in* $C$*.*

*Proof.*

The function minimisation problem, $\pi_a^* = \arg\max_{\pi \in \Gamma} E_\pi[L_a(\theta)]$, has a Lagrange dual form, (think of this as penalized nonparametric likelihood functions, c.f. the Lasso)

$$\pi_a^* = \arg \inf_{\pi \in \mathcal{M}} \left\{ E_\pi[-L_a(\theta)] + \lambda^{-1} KL(\pi \parallel \pi_I) \right\}$$

for some $\lambda \in [0, \infty)$, with $C < C' \implies \lambda < \lambda'$.

So the task is to select a distribution function that minimises the rhs, from the space of all probability measures,

Hence,

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

The result follows. Watson & Holmes (2014) □

See Ahmadi-Javid (2012), Breuer & Csiszár (2013a) who arrive at the same result through different means.

# Comments on the local-minimax $\pi_a^*$

- The $\Gamma$-local minimax distribution $\frac{e^{\lambda L_a(\theta)}\pi_I(\theta)}{\int e^{\lambda L_a(\theta)}\pi_I(\theta)d\theta}$ looks like an exponentially tilted density
- For linear loss, $L_a(\theta) = A\theta$, the local-minimax $\pi_a^*$ is the well known Esscher Transform used for option pricing in actuarial science
- Also, clearly,
  - $\lambda \to 0$, $\pi^* \to \pi_I$ leads back to Savage
  - $\lambda \to \infty$, $\pi^* \to \delta_{\theta^*}(\theta)$ leads to Wald's minimax solution

  where $\lambda$ defines the neighbourhood (KL ball) around $\pi_I$
- Note:
  - $\pi_a^*$ will lie on the boundary of the ball, so $KL(\pi_a^* \parallel \pi_I) = C$
  - $\pi_a^*$ is unique by strict convexity of the KL divergence
  - The most favourable distribution is, by symmetry,

$$\pi_a^{(\inf)} \propto e^{-\lambda L_a(\theta)}\pi_I(\theta)$$

- Substituting $\pi_a^*$ back into the KL divergence we see,

$$\begin{aligned} C &= KL(\pi_a^* \parallel \pi_I) \\ &= \int \pi_a^*(\theta) \log \left( \mathcal{Z}_{\pi^*}^{-1} \exp[\lambda L_a(\theta)] \right) d\theta \\ &= \lambda E_{\pi_a^*}[L_a(\theta)] - \log \mathcal{Z}_{\pi_a^*} \end{aligned}$$

  where $\mathcal{Z}_\pi$ denotes the normalising constant (prior predictive, marginal likelihood) or "partition function" of $\pi$

- Moreover, by Jensen's inequality

$$\begin{aligned} KL(\pi_a^* \parallel \pi_I) &= \lambda E_{\pi_a^*}[L_a(\theta)] - \log E_{\pi_I}[\exp(\lambda L_a(\theta)] \\ &\leq \lambda \left[ E_{\pi_a^*}[L_a(\theta)] - E_{\pi_I}[L_a(\theta)] \right] \end{aligned}$$

  The KL divergence is bounded above by $\lambda$ times the difference in expected losses under the two distributions.

# KL and coherence

- In fact the KL divergence is the only coherent divergence criteria one can use (for proof see Watson & Holmes (2014) taken from Appendix of Bissiri *et al.* (2013))
- That is, given a data set $x = \{x_1, \ldots, x_n\}$, prior model $\pi(x, \theta)$, and loss function $L(\theta)$. For coherence we require,

$$\pi_a^{(\text{sup})}(\theta|x_{1,\ldots,m}) f(x_{m+1,\ldots,n}|\theta) = \pi_a^{(sup)}(\theta|x_{1,\ldots,n})$$

  where $f(x|\theta)$ is the likelihood, for all $m = 0, \ldots, n$
- In words, you should obtain the same inference for the same information. The KL is the only divergence to achieve this.
- Conceptually this means we can consider specifying the operator *a priori* to obtain a local-minimax robust (data dependent) prior

$$\pi^*(\theta) \propto \pi(\theta) e^{\lambda L(\theta)}$$

  and update this with the standard likelihood

# Example I – Prediction

- Consider providing a predictive distribution for a future observation $\widehat{\pi(y|x)}$, for response variable of interest $y$ given covariates $x$
- The local proper scoring rule suggests the log-loss $L(y) = -\log \pi(y|x)$ (Bernardo & Smith, 1994)
  - Note: scoring rules are designed to keep Bayesian's honest in their model specifications
- Clearly the standard Bayesian solution is report your honest beliefs as $\widehat{\pi_I(y|x)} = \pi(y|x)$, where $\pi_I(y|x) = \int \pi(y|x,\theta)\pi_I(\theta)d\theta$
- However this assumes that the model is correct and moreover that the predictive contours do not change with time
  - c.f. "concept drift" in data mining

- The local-minimax solution above however leads to

$$\begin{aligned}\widehat{\pi(y|x)} &\propto e^{-\lambda \log \pi(y|x)} \pi(y|x) \\ &\propto \pi_I(y|x)^{1-\lambda}\end{aligned}$$

  for $\lambda \in (0, 1)$

- This leads to a tempered distribution that fattens the tails and smooths the modes of $\pi(y|x)$, with temperature given by your trust in your model

- We see that predictive tempering arises as the decision theoretic (local-minimax) solution to prediction under model misspecification
  - you should only report $\pi(y|x)$ if you have complete faith in your model, equating to $\lambda = 0$
    - c.f. Hjort and Walker (2001) on consistency

# Example II – unknown likelihood (and PAC-Bayesian)

- Suppose you hold prior beliefs about a parameter $\theta$ but don't know how to specify $p(x|\theta)$, and hence lack a model $p(x, \theta)$
- For example, consider $\theta$ as the median of $F_X$ with unknown distribution
- We don't have a likelihood but we could have a well defined prior hence
- and a well defined loss function that we would wish to maximise utility against for specifying beliefs, e.g. for the median we should take

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

- This leads to the local-minimax distribution as

$$\pi_a^{(\inf)} = Z^{-1} e^{-\lambda \sum_i |x_i - \theta|} p(\theta)$$

  where $\lambda$ trades off fidelity to the data against fidelity to the prior
- This has the form of a Gibbs Posterior or PAC-Bayesian approach (Zhang (2006a,b); Bissiri *et al.* (2013); Dalalyan & Tsybakov (2008, 2012); Langford & Schapire (2005))
- So we can interpret PAC-Bayesian or Gibbs posteriors as decision theoretic local-maximin solutions in the absence of a known sampling distribution

# Computational decision theory

- Conventional computational decision theory via Monte Carlo computes the expected loss of an action $a$ given partial information $x$ and a model $p(\theta|x)$ via

$$\begin{aligned} \widehat{U}_a &= \frac{1}{N}\sum_{i=1}^{N} L_a(\theta_i) \\ \theta_i &\sim \pi(\theta|x) \end{aligned}$$

  where, as stated above, $L_a(\cdot)$ is a real-valued loss (negative utility, or negative reward) function and $\pi_I(\theta) \equiv \pi(\theta|x)$ denotes Your subjective posterior beliefs, given all available information

- We can write this as a weighted average with uniform weights on each MC sample drawn from the posterior

$$\begin{aligned} \widehat{U}_a &= \sum_{i=1}^{N} w_i L_a(\theta_i) \\ \theta_i &\sim \pi(\theta|x) \\ w_i &= \frac{1}{N} \end{aligned}$$

# Robust computational decision theory

- Robust methods use retrospective re-sampling or re-setting of the weights,

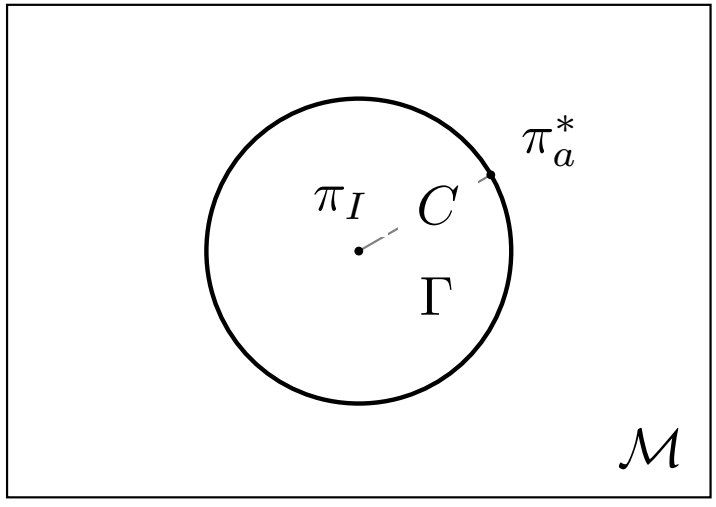
$$\widetilde{w}_i \neq \frac{1}{N}$$

  either deterministically (as above with $w_i \propto \exp[\lambda L(\theta_i)]$) or stochastically (as below), we are able to make formal statements about decision robustness under model approximation

- Informally you can think of this as shaking your model to explore robustness, by retrospectively perturbing the Monte Carlo sample weights in a very particular way
  - there are close connections to Bayes nonparametrics and the Bayesian bootstrap

# Exploring variation in expected losses within $\Gamma$

- The use of the minimax outcomes gives us one (extreme) solution, $\pi_a^{(\sup)}$ or $\pi_a^{(\inf)}$, in the KL ball
- Looking at the distribution of expected loss within the ball might be more informative, and more natural from a Bayesian perspective




- However, with the neighbourhood $\Gamma$ defined via $\mathrm{KL}(\pi \parallel \pi_I)$ calculating this distribution over the set $\Gamma = \{\pi : \mathrm{KL}(\pi \parallel \pi_I) \leq C\}$ is challenging

# Change of neighbourhood to $\Gamma = \{\pi : \mathrm{KL}(\pi_I \parallel \pi) < C\}$

- We now consider a change of neighbourhood, to the divergence of Nature from $\pi_I$, $\mathrm{KL}(\pi_I \parallel \pi)$, so now we define

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

- Interestingly we no longer have an analytic form for the local-minimax distribution

$$\pi_a^{(\mathrm{sup})} \neq Z^{-1} e^{\lambda L(\theta)} \pi_I(\theta)$$

- However we do have Monte Carlo methods for calculating both the minimax *and* the variation in loss over $\Gamma$

# Taking expectations over $\Gamma$

- To take expectations over distributions in the neighbourhood of $\pi_I$ we require a probability distribution on probability measures within $\Gamma$ centred on $\pi_I$
- This is classically a problem in Bayesian nonparametrics
- But here we are considering nonparamateric priors on posterior distributions to explore the ex-post sensitivity of misspecification
- The Dirichlet Process (Ferguson 1973) is the cornerstone of Bayes NP and a natural place to start

## Definition

**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 $\alpha$ 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 $Dir(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.

# Bayesian Nonparametric prior

- Using a DP we can sample from distributions in the neighbourhood of $\pi_I$ by setting the baseline measure of the DP to be $\pi_I$,

$$\pi \sim DP(\alpha, \pi_I)$$

where $\alpha$ defines the neighbourhood size: larger $\alpha$ means a tighter (smaller) neighbourhood concentrated around the baseline $\pi_I$

In practice the DP has a constructive definition,

$$\begin{aligned} \tilde{\pi} &= \sum_{i=1}^{m} w_i \delta_{\theta_i}(\theta) \\ \theta_i &\sim \pi_I \\ \underline{w} &\sim \mathrm{Dir}_m(\alpha/m, \ldots, \alpha/m), \\ m &\to \infty \end{aligned} \tag{2}$$

where the $\theta_i$'s are iid from $\pi_I$ and independent of the Dirichlet weights

- This highlights one difficulty with the use of the DP. The discrete atomic structure means that two draws do not have the same support
- Hence the Kullback-Leibler divergence $\mathrm{KL}(\pi_I \parallel \tilde{\pi})$ is not defined
- In order to circumvent this we introduce the notion of coupled-DPs.

**Coupled Dirichlet Process.** We say that two or more Dirichlet process samples $\{\pi^{(i)}, \pi^{(j)}\}$ are coupled,

$$\pi^{(i)}, \pi^{(j)} \sim CDP(\alpha_i, \alpha_j, \pi_0)$$

if they share a common set of baseline atoms, $\{\delta_{\theta_s}(\theta)\}_{s=1}^m$ drawn from the baseline measure, $\pi_0$, but with independent Dirichlet weights,

$$\begin{aligned}
\tilde{\pi}^{(i)} &= \sum_{s=1}^{m} w_s^{(i)} \delta_{\theta_s}(\theta) \\
\tilde{\pi}^{(j)} &= \sum_{s=1}^{m} w_s^{(j)} \delta_{\theta_s}(\theta) \\
\theta_s &\sim \pi_0 \\
\underline{w}^{(i)} &\sim \mathrm{Dir}_m(\alpha_i/m, \ldots, \alpha_i/m), \\
\underline{w}^{(j)} &\sim \mathrm{Dir}_m(\alpha_j/m, \ldots, \alpha_j/m), \\
m &\to \infty
\end{aligned}$$

- Now, given two realisations we find the KL divergence as,

$$KL(\tilde{\pi}^{(i)} \parallel \tilde{\pi}^{(j)}) = \sum_{s=1}^{m} w_s^{(i)} \log\left(\frac{w_s^{(i)}}{w_s^{(j)}}\right)$$

- Now consider the Monte Carlo representation of the baseline measure $\tilde{\pi} \sim DP(\infty, \pi_I)$,

$$\begin{aligned} \tilde{\pi}_I &= \frac{1}{m} \sum_{i=1}^{m} \delta_{\theta_i}(\theta) \\ \theta_i &\sim \pi_I \\ m &\to \infty \end{aligned} \tag{3}$$

- Then we can compute a measure of divergence of $\pi^* \sim DP(\alpha, \pi_I)$ to its baseline measure via the coupled Dirichlet process representation, $\tilde{\pi}^*, \tilde{\pi}_I \sim CDP(\alpha, \infty, \pi_I)$, for which,

$$\begin{aligned} KL(\tilde{\pi}_I \parallel \tilde{\pi}^*) &= \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{1}{m w_i}\right) \\ \underline{w} &\sim \mathrm{Dir}_m(\alpha/m, \ldots, \alpha/m), \end{aligned}$$

- This KL is itself a random variable and it is interesting to investigate its distribution from the perspective of the neighbourhood size around $\pi_I$

- It will be helpful to consider the representation of the Dirichlet as a normalised sum of Gamma random variables,

$$\begin{aligned} G_i &\sim \mathrm{Gam}(\alpha, 1) \\ w_i &= \frac{G_i}{G_1 + \cdots + G_m} \end{aligned}$$

  then $\underline{w} \sim \mathrm{Dir}_m(\alpha/m, \ldots, \alpha/m)$

- Now let $m \to \infty$ and let $G = \sum_i G_i$, so that $G \sim \mathrm{Gam}(m\alpha, 1)$

- Under this representation we have for $\tilde{\pi}, \tilde{\pi}_I \sim CDP(\alpha, \infty, \pi_I)$,

$$\begin{aligned} \mathrm{KL}(\tilde{\pi}_I \parallel \tilde{\pi}) &= \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{1}{m w_i}\right) \\ &= \log(G) - \frac{1}{m} \sum_i \log(G_i) - \log m. \end{aligned}$$

# Distribution of $\mathrm{KL}(\tilde{\pi}_I \parallel \tilde{\pi})$

**Proposition:** Let $\tilde{\pi}, \tilde{\pi}_I \sim CDP(\alpha, \infty, \pi_I)$ then $\tilde{\pi}$ is distributed on the shell of a KL ball, $\tilde{\Gamma}$, centered on $\tilde{\pi_I}$ of radius $C$ with,

$$C = \log \alpha - \psi_0(\alpha)$$

where $\psi_0(\cdot)$ denotes the digamma function

*Proof:* from the properties of the log-Gamma distribution □

- That is, when you sample from a DP you only sample in a very small region of model space at a KL divergence of $\log \alpha - \psi_0(\alpha)$ from the baseline measure
- This is analogous to the well known result that for the multivariate Gaussian, $X \sim N(0, \sigma^2 I_p)$, for large $p$ we have

$$Pr(|\parallel X \parallel^2 - \sigma^2 p| > \epsilon \sigma^2 p) \leq 2e^{-p\epsilon^2/24}$$

almost the entire measure is contained in a thin-shell of radius $\sigma\sqrt{p}$

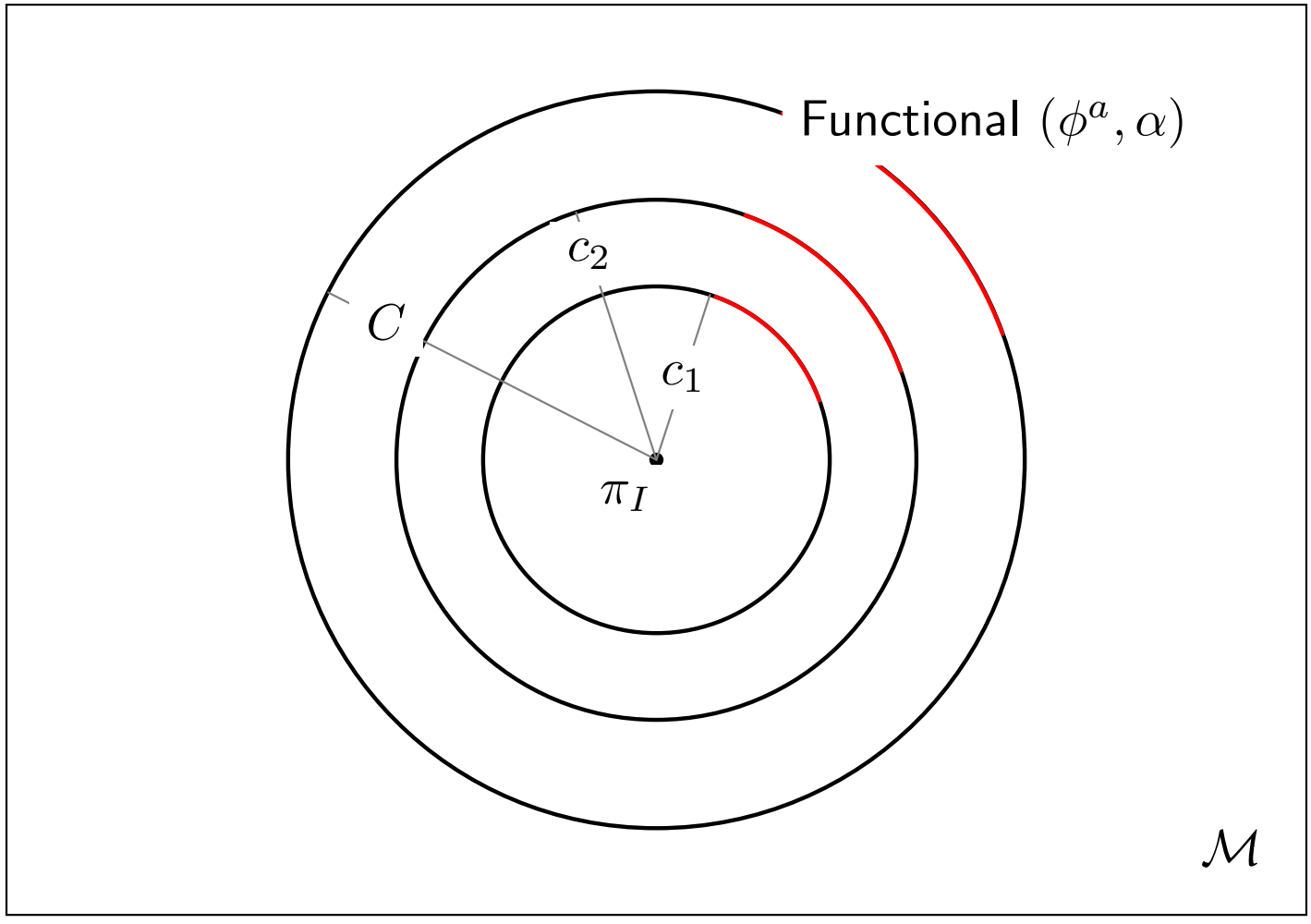



Figure: Sample KL radii $c \sim U[0, C]$ and then sample $\pi \sim DP(\alpha(c), \pi_I)$ and calculate distribution on a functional of interest

# To recap....

We can retrospectively use MC samples from $\pi_I(\theta)$ to explore sensitivity to model misspecification and robustness of decisions (and expected loss) by using local re-weighted representations

$$\begin{aligned} \tilde{\pi} &= \frac{1}{\sum w_i} \sum_i w_i \delta_{\theta_i}(\theta) \\ \theta_i &\sim p(\theta|x) \end{aligned}$$

Where:

- Local-minimax, $\mathrm{KL}(\pi \parallel \pi(\theta|\mathrm{x})) \leq \mathrm{C}$, deterministic re-weights

$$w_i = e^{\lambda L(\theta_i)}$$

- Coupled-DP, $\mathrm{KL}(\pi(\theta|\mathrm{x}) \parallel \pi) \leq \mathrm{C}$, stochastic re-weights

$$\begin{aligned} w_i &\sim \mathrm{Ga}(\alpha^{-1}[c], 1) \\ c &\sim U(0, C) \end{aligned}$$

# Case Study I: Regression Variable Selection with Costs

- We consider the problem of variable selection under costs
- Let $\mathcal{X} = \{x_j\}_{j=1}^{p}$ denote the set of possible regressors with response variable of interest $Y$
- Let $c_j$ denote the cost in measuring the $j$th variable
  - e.g. costs may relate to genotyping, sequencing or collection of phenotype data
- An action $a$ is defined as a subset of covariates $\gamma_a = \{\gamma_i\}_{j=1}^{p}$ where $\gamma_j = 1$ if the $j^{th}$ regressor is included and zero otherwise
- $\gamma_a$ can be seen as a possible model, and the decision task is one of model selection
- The loss function defined over a model $\gamma_a$ trades accuracy of prediction against total cost of data collection:

$$L_a(\gamma) = L(y) + \sum_{j=1}^{p} c_j * \gamma_{aj}$$

# Prediction of hospital mortality rate

- We consider a data set of hospital admissions analysed in Fouskakis & Draper (2008)
- $n = 2,532$ observations of 83 regressors and a univariate response $y \in \{0, 1\}$ which is 1 if the patient dies within 30 days of admission and zero otherwise
- Collection costs for variables vary from 0.5 to 10 with a median cost of 1
- Loss to data is the log-information

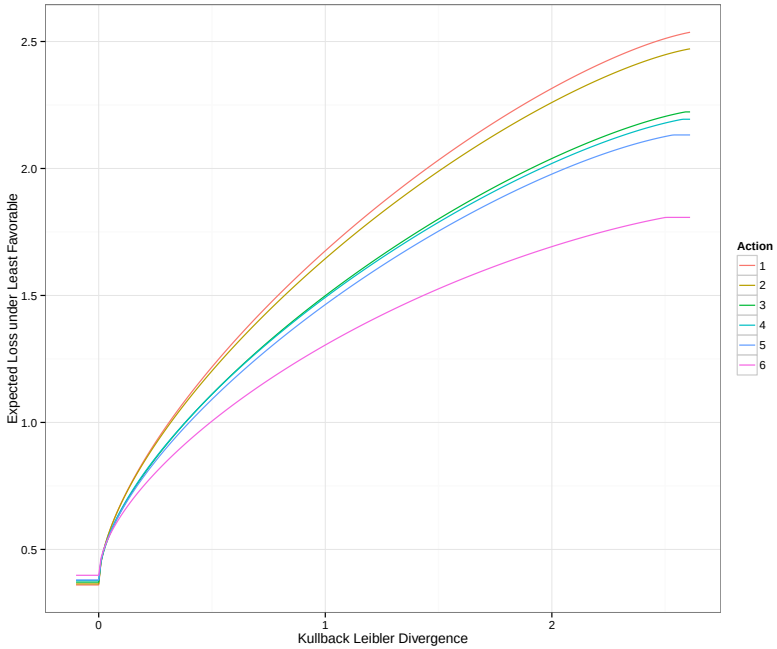
$$L(y) = \sum_i -\log \pi(y_i|x_i)$$

- We use simulated annealing to generate six models with highest expected utility
- Explore robustness of the "best" model

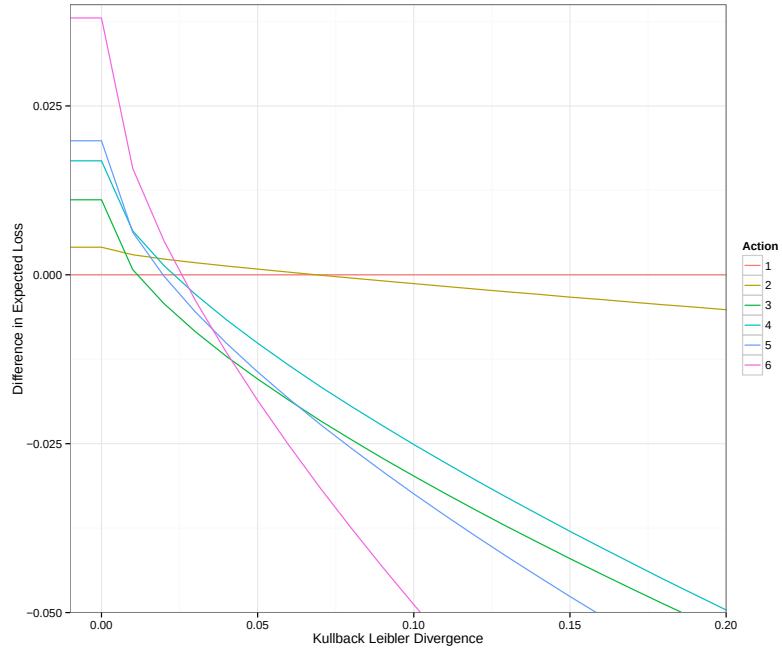
# Summary of top models found by simulated annealing

| Variable | | | Model | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Action 1 | Action 2 | Action 3 | Action 4 | Action5 | Action 6 |
| Index | Name | Cost | | | | | | |
| 1 | Systolic BP score | 0.5 | | * | * | * | * | |
| 2 | Age | 0.5 | | | * | * | * | * |
| 3 | Blood urea nitrogen | 1.5 | | | | | | * |
| 4 | APACHE II coma score | 2.5 | * | * | | | | * |
| 12 | Initial temperature | 0.5 | * | * | * | | | * |
| 14 | Chest pain day 1? | 0.5 | | | | | * | * |
| 35 | Hepatobiliary history? | 0.5 | * | * | * | * | * | |
| 36 | Renal history score | 1.0 | | | | | * | |
| 37 | APACHE respiratory rate score | 1.0 | * | | * | * | | * |
| 51 | Confusion day 1 | 0.5 | * | * | | | | * |
| 54 | Pulmonary edema score | 0.5 | | | | | | * |
| 61 | Wheezing at admission? | 0.5 | | | | * | | |
| 62 | Blood system count | 2.5 | * | * | * | * | * | |
| 68 | Co-morbid smoking score | 0.5 | * | | | | | |
| 70 | APACHE pH score | 1.0 | | | | | * | |
| 74 | Cardiac history score | 0.5 | * | * | * | * | * | * |
| 75 | Neurologic history score | 0.5 | * | * | * | * | * | * |
| 76 | Oncologic history score | 0.5 | * | * | * | * | * | |
| 77 | Immunologic history score | 0.5 | | | * | * | * | |
| 78 | Musculoskeletal score | 0.5 | * | * | * | * | | |

# Local-minimax loss from six models (actions) $\gamma_a$




(a) Expected Loss under least favorable distribution, six models




(b) Difference in expected loss near origin: showing crossing point.

Figure: Comparison of optimality under minimax expected loss as a function of KL radius, for the top 6 actions selected by simulated annealing search

- Note that the "optimal" action assuming a true model, $\pi_I(\theta)$, becomes the least optimal for KL radii greater than 0.1 (crossing points approximately between 0.0125 and 0.075)

# Case Study II – Screening design

- Public health policy is an area where the application of statistical modeling can be used to optimally allocate resources
- Many countries operate a breast cancer screening policy (a hotly debated and controversial issue) for healthy women over a threshold age to detect asymptomatic tumors
- A primary issue is determining the optimal screening schedule (action), consisting of $a = (t_0, \delta)$,
  - a starting time $t_0$ (age of first screen), and,
  - a frequency $\delta$ for subsequent screens
- Parmigiani (1993) proposed using a semi-Markov process consisting of four states which generalizes to any chronic disease characterized by an asymptomatic stage.

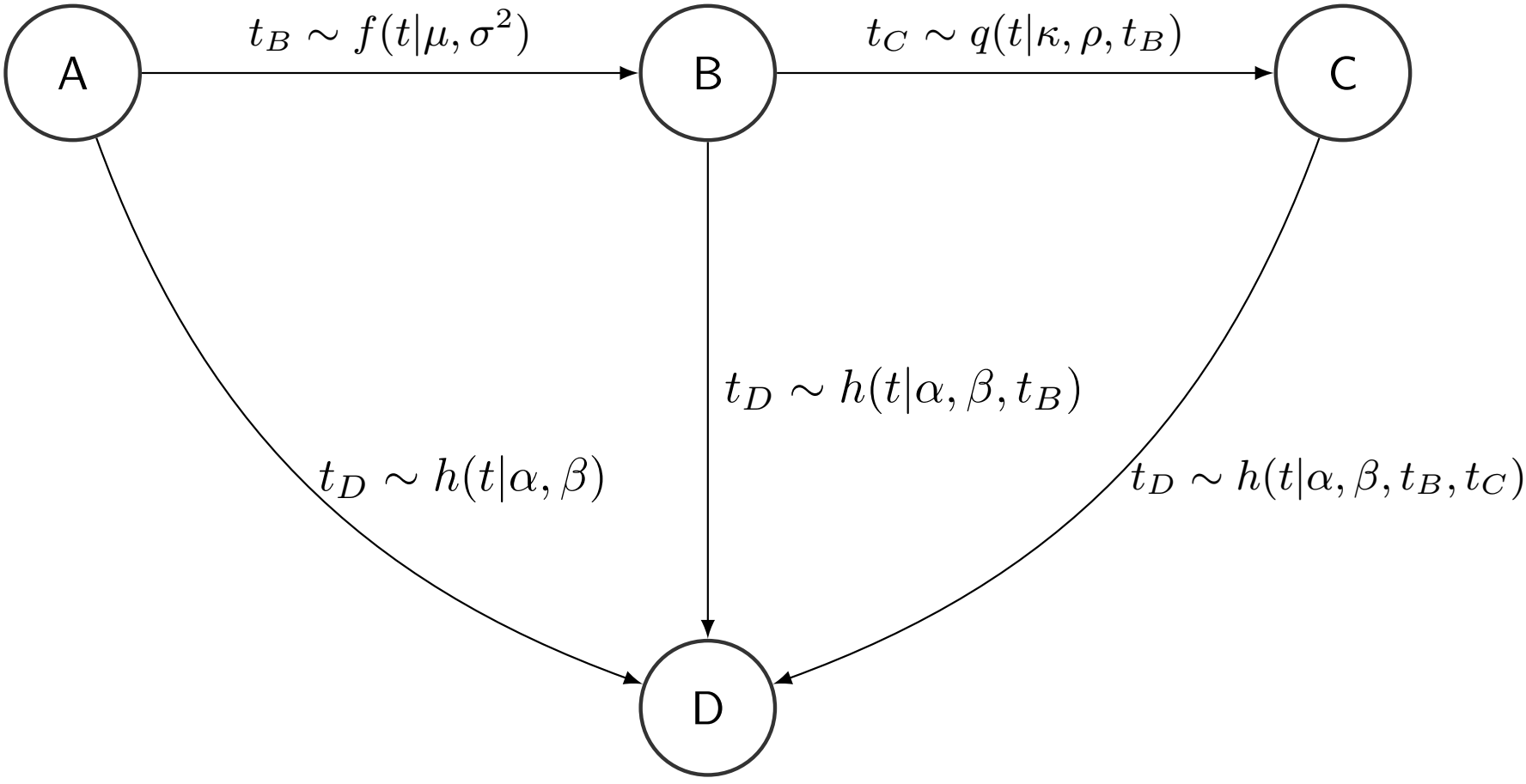



Figure: $A$, disease-free; $B$, the pre-clinical stage; $C$ clinical; absorbing state $D$ (death)

- All individuals start in state $A$
- This is followed by a transition to either the clinical stage of the disease or death
- Following Parmigiani (1993) , we model the transitions as,

$$t_B \sim f(t|\mu, \sigma^2) = \text{LogNormal}(\mu, \sigma^2)$$

$$t_C \sim q(t|\kappa, \rho) = \text{LogLogistic}(\kappa, \rho)$$

$$t_D \sim h(t|c, d) = \text{Weibull}(\alpha, \beta)$$

- An individual is characterized by the triple $t = (t_B, t_C, t_D)$
- 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)
- We use 10,000 samples of $\theta_i \sim \pi_I(\theta)$ to represent the distributions

# Choosing an optimal screening design

- The task is to select an optimal screening schedule $a = (t_0, \delta)$
- The loss function is defined as follows (a function of the times $t = (t_B, t_C, t_D)$:

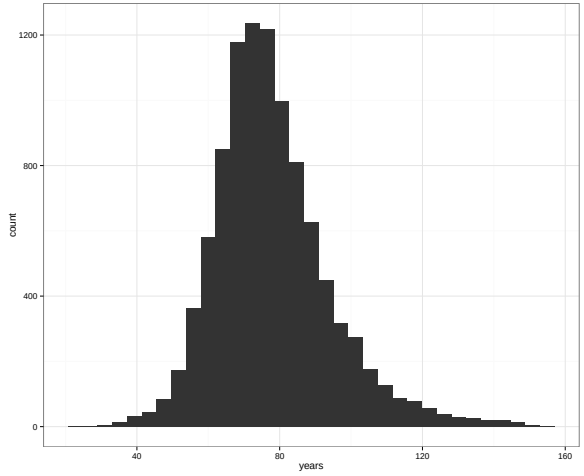
$$L_a(t_0, \delta) = r \cdot n_a(t) + 1_C$$

  where $n_a$ is the number of screening schedules an individual will receive during their lifetime, until they die or enter into the asymptomatic stage of the disease

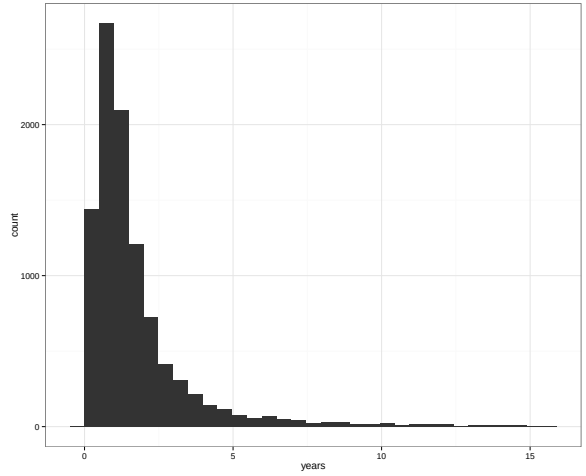
  $1_C$ is the indicator function, taking value 1 for the event that the pre-clinical tumor is not detected by screening or occurs before $t_0$, and zero otherwise

- $r$ trades off the cost of one screen against the cost incurred by the onset of the clinical disease, following others we take $r = 10^{-3}$

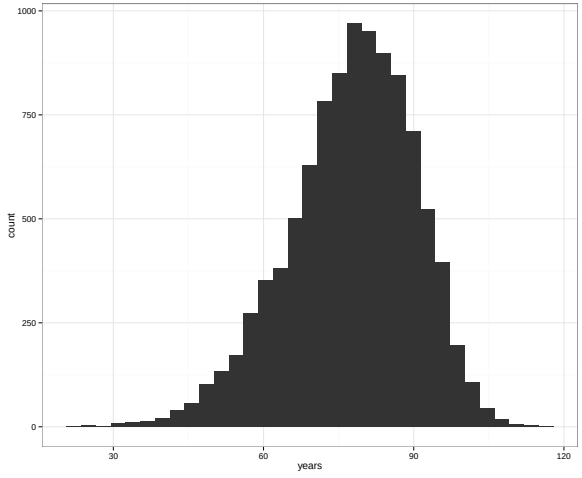
# Model – $\pi_I(\theta)$



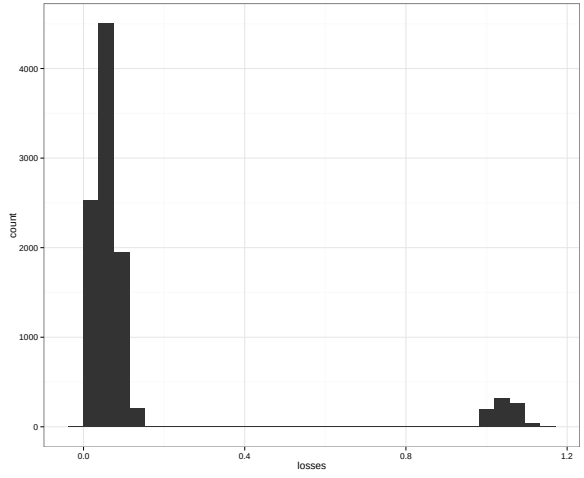

(a) Transition time: $A$ to $B$




(b) Sojourn times: $B$ to $C$.




(c) Lifetimes generated from $\text{Weibull}(7.2, 82.6)$.




(d) Histogram of losses with schedule $(40, .75)$.

# Loss distributions for various schedules

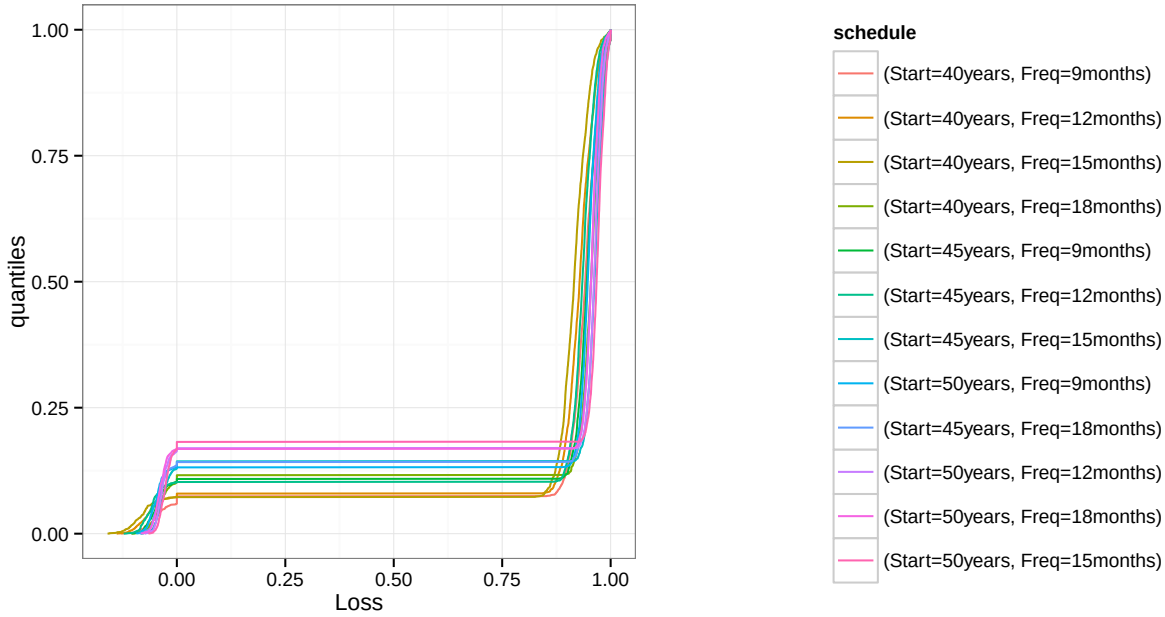



Figure: Loss distribution for 12 potential schedules following Ruggeri et al. (2005)

- We considered 12 potential screening schedules
- The optimal schedule, assuming everything is true, minimising expected loss under $\pi_I$, is

$$\hat{a} = (t_0 = 40, \delta = 0.75)$$

# Local-minimax loss as KL radii changes

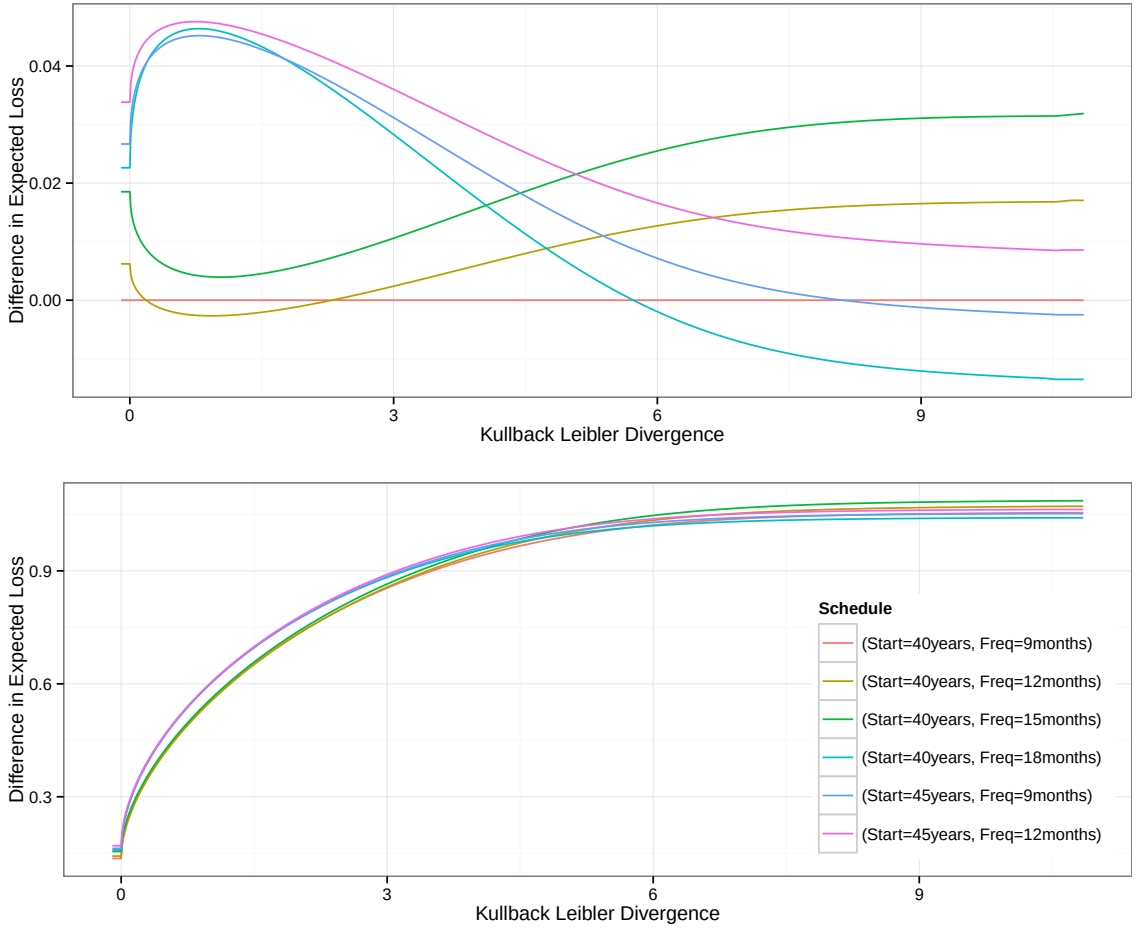



Figure: Change in optimality of selected schedules as a function of KL radius with respect to most optimal under $\pi_I$(1: 40 years, 9 months). We plot the difference in expected loss between each action (schedule) and the optimal action. Note for a KL radius greater than $\approx 5.5$ schedule 1 is suboptimal.

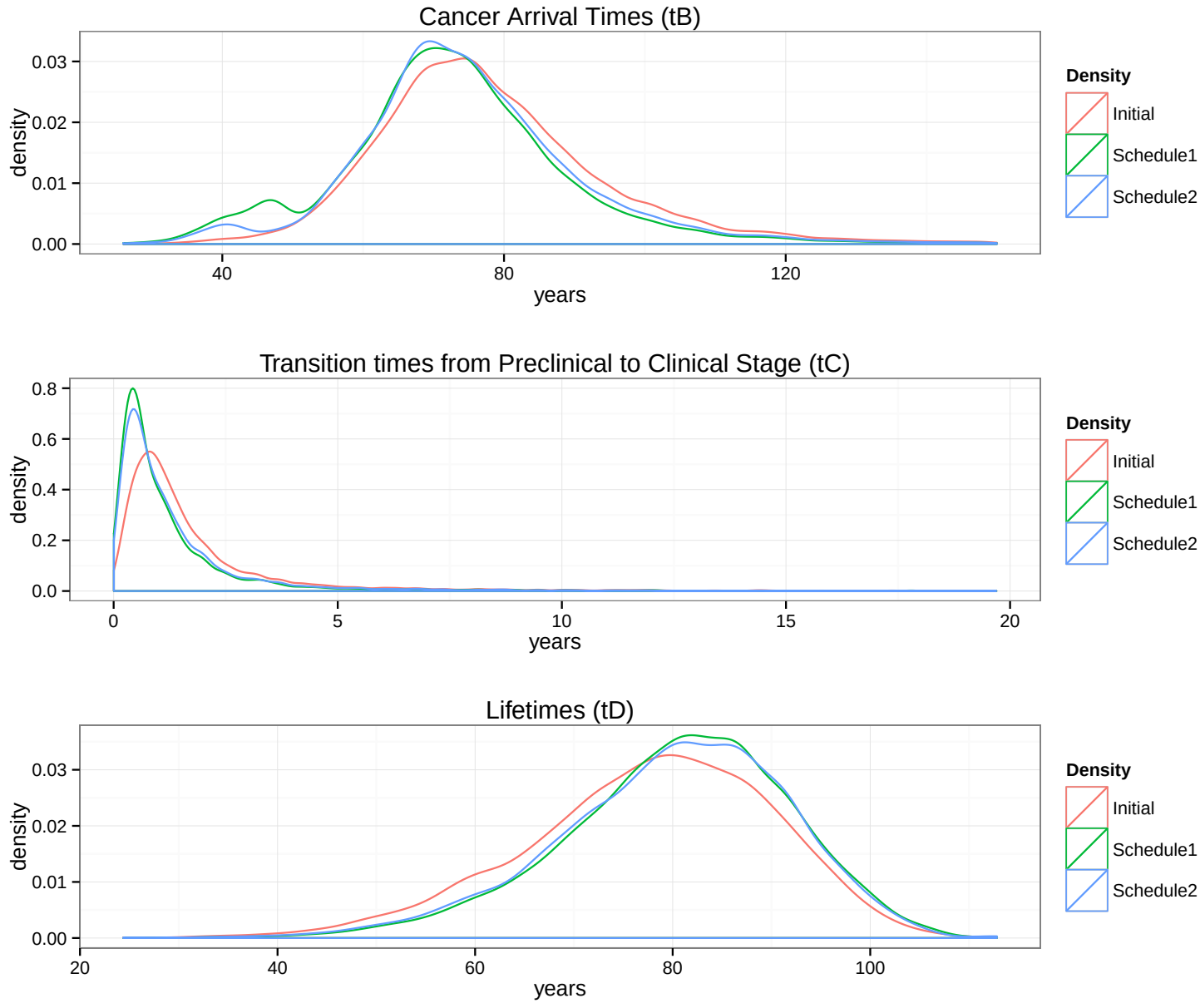



Figure: Interpreting Kullback-Leibler divergence: marginal densities over $\Theta$. KL radius $C = 1$

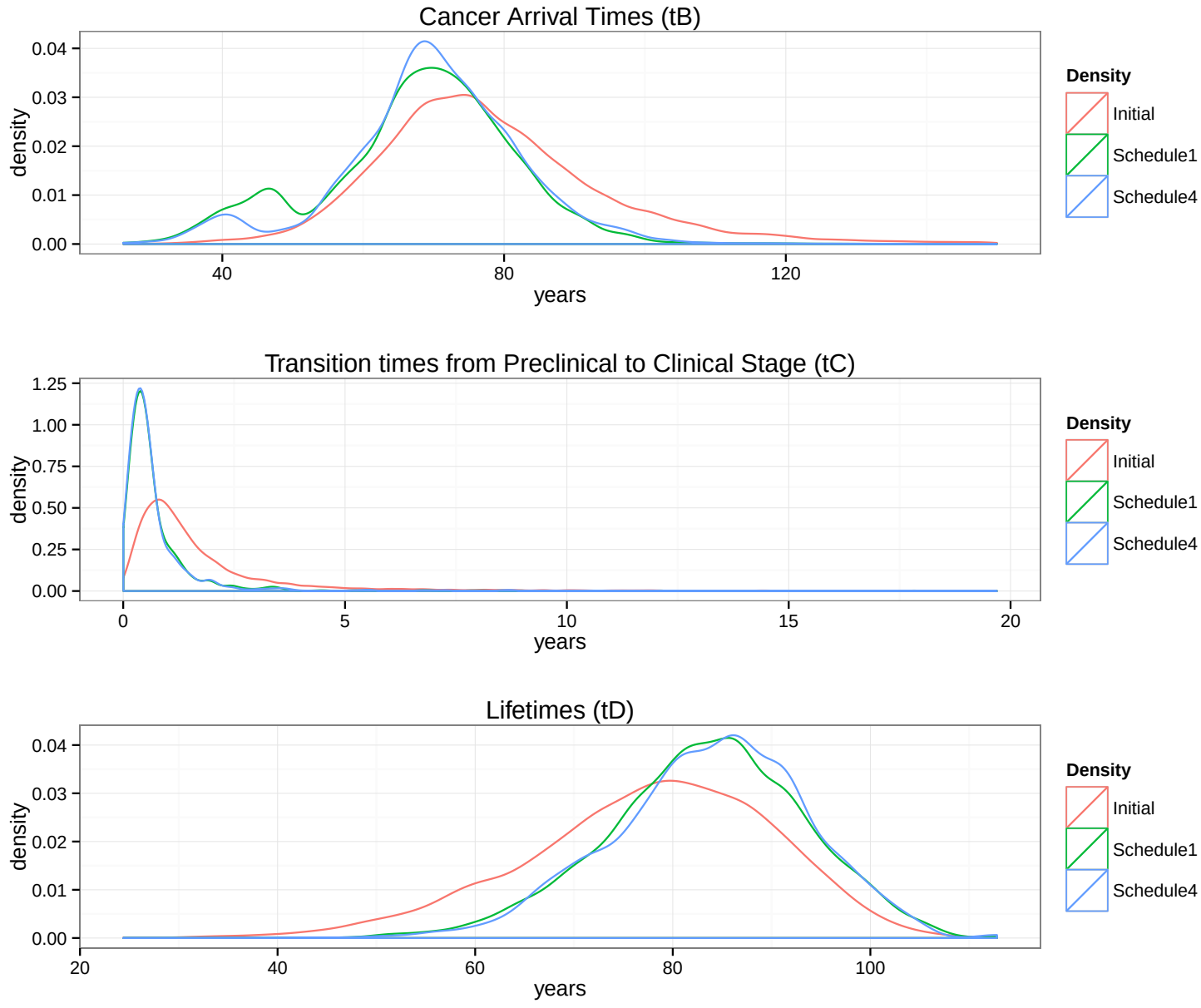



Figure: Interpreting Kullback-Leibler divergence: marginal densities over $\Theta$. KL radius $C = 5$

# DP sampler and quantile of loss within KL ball

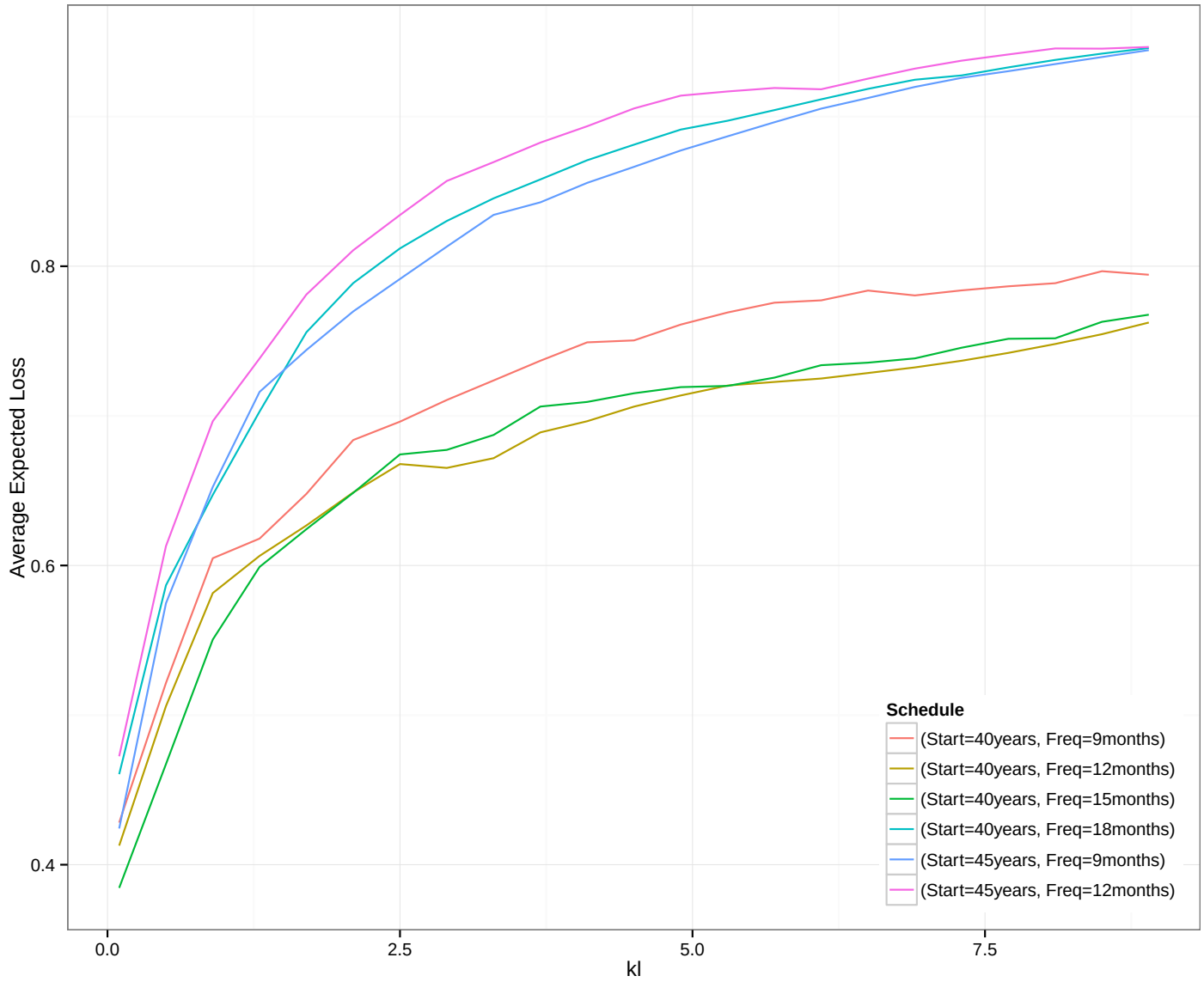



Figure: DP sampler: 10% quantile of expected loss as a function of KL divergence.

# Summary

- Checking robustness and sensitivity of answers to modelling assumptions should be a key component of any data analysis
- In the Bayesian approach robustness can be explored through a class of distributions $\Gamma$ around the prior or posterior model
- We showed how using re-weighted MC samples generated from the model, $\theta_i \sim p(\theta|x)$, we can explore robustness within a KL ball around the approximating model
  - Deterministic re-weighting of MC samples estimates the local-minimax outcome within a ball of fixed radii from $p(\theta|x)$
    - Predictive tempering arises as robust local-minimax solution
  - Stochastic re-weighting of MC samples using a Coupled Dirichlet Process, provides samples from the space of all distributions within some KL ball around $p(\theta|x)$
- Moreover the KL is the only coherent divergence measure, and the local-minimax leads to predictive tempering

# Summary

- Of course, through time constraints there have been many areas we were forced to omit or gloss over
  - Did not have time to discuss diagnostics techniques using graphical methods and summary statistics
  - Merely skimmed the important areas of PAC-Bayesian and Gibbs posteriors which you loss functions (rather than log-likelihoods) to construct models
- See Watson & Holmes (2014) for further details and references

# Conclusions

- Optimal actions and decisions are conditional on models
- If the models are approximations then so are the answers and decisions
- Their is a rise in the development and use of approximate probabilistic models to address modern big-data applications
  - merits a reappraisal of Bayesian robustness
- Statisticians should be sensitive to decision stability – "shake your model"

*Acknowledgements: Medical Research Council, Wellcome Trust, EPSRC (*`www.i-like.org.uk`*)*

# Additional reasons for using Kullback-Leibler

- invariant to re-parameterisation
- interpretable as half the expected deviance
- information theoretic interpretation as number of bits of information to recover $\pi_I$ from $\pi^*$
- equals the expected loss in using $\pi_I$ to approximate $\pi^*$ when preferences are described by proper local scoring rules
- from the "robustness" properties of MLE we know that $\pi_I$ will converge to the closest distribution in KL divergence to Nature's true $\pi$
- the solution turns out to be analytic and computable
- KL bounds $L_1$ loss, so $KL(\pi^*, \pi_I) \geq \frac{1}{2}||\pi^* - \pi_I||_1^2$
- KL is the only coherent divergence measure

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.

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

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

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 $\varepsilon$-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.

Bernardo, José M, & Smith, Adrian FM. 2009. *Bayesian theory*. Vol. 405. John Wiley & Sons.

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

Bissiri, P.G., Holmes, C.C., & Walker, S.G. 2013. *A General Framework for Updating Belief Distributions*. http://arxiv.org/abs/1306.6430.

Box, George EP. 1953. Non-normality and tests on variances. *Biometrika*, **40**(3-4), 318–335.

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*.

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.

de Finetti, B. 1961. The Bayesian approach to the rejection of outliers. *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*, **1**, 199–210.

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

Fisher, Ronald Aylmer. 1920. A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. *Monthly Notices of the Royal Astronomical Society*, **80**, 758–770.

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

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.

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.

Haro-Lopez, R. A., & Smith, A. F. M. 1999. On robust Bayesian analysis for location and scale parameters. *Journal of Multivariate Analysis*, **70**, 30–56.

Huber, Peter J. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**(1), 73–101.

Huber, Peter J. 1972. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 1041–1067.

Huber, Peter J., & Ronchetti, E. M. 2009. *Robust statistics*. Wiley.

Kadane, J.B. (ed). 1984a. *Robustness of Bayesian analyses*. Studies in Bayesian econometrics. North-Holland.

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

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

Langford, John, & Schapire, Robert. 2005. Tutorial on Practical Prediction Theory for Classification. *Journal of machine learning research*, **6**(3).

Newcomb, Simon. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 343–366.

Neyman, Jerzy, & Scott, Elizabeth L. 1971. Outlier proneness of phenomena and of related distributions. *Optimizing Methods in Statistics. Academic Press, New-York*, 413–430.

O'Hagan, A. 1979. On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society, Series B*, **41**(3), 358–367.

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

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*.

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.

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.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. *In:* Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., & Mann, H. B. (eds), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford studies in mathematics and statistics. Stanford University Press.

Vidakovic, B. 2000. $\Gamma$-minimax: a paradigm for conservative robust Bayesians. *Pages 241–259 of:* Rios Insua, D., & Ruggeri, F. (eds), *Robust bayesian analysis*. Springer.

Wasserman, L. 1992. Recent methodological advances in robust Bayesian inference. **4**, 483–502.

Watson, James, & Holmes, Chris. 2014. Approximate Models and Robust Decisions. *arXiv preprint arXiv:1402.6118*.

West, M. 1984. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B*, **46**(3), 431–439.

Whittle, P. 1990. *Risk-sensitive Optimal Control*. Wiley.

Zhang, T. 2006a. From $\epsilon$-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.