Bayesian heavy-tailed models and conflict resolution: a review

Anthony O'Hagan[1] and Luis Pericchi[2]

[1]University of Sheffield,
[2]University of Puerto Rico, Rio Pedras

April 5, 2011

Abstract

We review a substantial literature, spanning 50 years, concerning the resolution of conflicts using Bayesian heavy-tailed models. Conflicts arise when different sources of information about the model parameters (e.g. prior information, or the information in individual observations) suggest quite different plausible regions for those parameters. Traditional Bayesian models based on normal distributions or other conjugate structures typically resolve conflicts by centring the posterior at some compromise position, but this is not a realistic resolution when it means that the posterior is then in conflict with the different information sources. Bayesian modelling with heavy-tailed distributions has been shown to produce more reasonable conflict resolution, typically by favouring one source of information over the other. The less favoured source is ultimately wholly or partially rejected as the conflict becomes increasingly extreme.

The literature reviewed here provides formal proofs of conflict resolution by asymptotic rejection of some information sources. Results are given for a variety of models, from the simplest case of a single observation relating to a single location parameter up to models with many location parameters, location and scale parameters, or other kinds of parameters. However, these results do not begin to address models of the kind of complexity that are routinely used in practical Bayesian modelling. In addition to reviewing the available theory, we also identify clearly the gaps in the literature that need to be filled in order for modellers to be able to develop applications with appropriate 'built-in robustness'.

Keywords: rejection of information, partial rejection of information, built-in robustness, heavy-tailed modelling, theory of conflict resolution, outliers

1 Introduction

The problem of how to handle conflicting information sources is quite a general one in statistics. It arises when two different pieces of information suggest, or are consistent with, very differing values of unknown parameters. One common situation is the presence of outliers in data. If, for instance, we have data $\mathbf{x} = (1.5, 2.6, 0.3, 0.9, 2.2, 25.5)$ from a distribution with unknown mean μ, then the observation 25.5 suggests a value for μ that is far from the values indicated by the other five observations. The last observation seems to be an outlier, and raises the question of how to reconcile the conflict between it and the remaining observations. A huge literature has developed around this question, largely involving techniques to decide whether to reject the outlier, and so make inference about μ on the basis of just the first five observations, or to include it, making inference from the whole sample treating the outlier as of equal weight to any of the other observations. The classic text on outliers is that of Barnett and Lewis (1994), and although much has been published since it remains an excellent source for the various approaches in this field.

In Bayesian statistics, we can have conflict when there is only one observation if it conflicts with the prior information. In more complex applied problems, we can be combining many different sources of information of diverse natures, and conflicts can arise in many subtle ways. Not only is it then difficult to determine how to handle conflicts but it can also be challenging even to spot them.

In this paper we review the literature on Bayesian methods to resolve conflicts in an automatic way, through the use of heavy-tailed distributions.

1.1 Bayesian outlier rejection

First notice that if the data $\mathbf{x}$ given above were to be modelled as a sample from the $N(\mu, 1)$ distribution, then the likelihood function is maximised at $\hat{\mu} = 5.5$, the sample mean. In a Bayesian analysis with a weak (uniform) prior distribution the posterior distribution would be centred at $E(\mu \,|\, \mathbf{x}) = 5.5$ with a standard deviation of about 0.4, leading to a 95% interval for μ of $(4.7, 6.3)$. If we were to make inference using only the first five observations we would have a mean of 1.5 and a 95% interval $(0.6, 2.4)$, whereas inference based on the single final observation would have mean 25.5 and 95% interval $(23.5, 27.5)$. The conflict between the first five observations and the last one is very evident in the widely differing 95% intervals that they imply, but another remarkable feature of this example is that the posterior inference from the whole sample is again quite different. It leads to a 95% interval that is not even close to overlapping with the intervals implied by the two subsamples separately. This is a feature of using normal distributions. The analysis of the whole sample gives the outlier the same weight as any other observation, effectively ignoring the conflict.

The first discussion of outlier rejection in Bayesian inference was by de Finetti (1961). He commented on the above analysis and noted that essentially the same features would hold if the observations had an unknown, but

common, error variance σ^2. However, he remarked that if the observations had different, independent, unknown variances, then the outlier would be estimated as having a large error variance, and so would be given less weight. The greater the conflict, in the sense of the outlier being further from the remaining observations, the less weight it would get. In the limit, as the outlier became infinitely separated from the other observations it would be given zero weight, which de Finetti described as Bayesian outlier rejection. De Finetti's argument was heuristic; he derived no formulae and did not consider specific cases. Lindley (1968), in response to a discussion contribution from E.M.L. Beale, gives an approximate analysis based on the leading term of an expansion for the posterior in the case of a t prior distribution. It was not until the contributions of Dawid (1973), Hill (1975) and O'Hagan (1979) that formal analysis was given and conditions presented under which this kind of outlier rejection behaviour would indeed arise.

1.2 Example with a t distribution

The connection between this kind of outlier rejection and heavy tails arises from the uncertainty about the error variance for each observation. Suppose that $x_i \sim N(\mu, \sigma_i^2)$, so that observation i has error variance σ_i^2, and now let σ_i^2 have distribution $F(\cdot)$. Assuming, as de Finetti did, that the σ_i^2s are independent, then the marginal density of x_i is of the form known as a scale mixture of normals. It has density

$$f(x_i \,|\, \mu) = \int \sigma^{-1} \phi((x_i - \mu)/\sigma) dF(\sigma^2) \ ,$$

where $\phi(\cdot)$ is the standard normal density function. It is clear that de Finetti's argument requires $F(\cdot)$ to give non-zero probability to arbitrarily large values of σ^2, so that the weight attached to an outlier can become arbitrarily small. The best known family of scale mixtures of normals is the t family, where $F(\cdot)$ is an inverse gamma distribution, and t distributions have heavier tails than normal distributions.

Returning to the example of data $\mathbf{x} = (1.5, 2.6, 0.3, 0.9, 2.2, 25.5)$ we now let the observations x_i have independent t distributions with 5 degrees of freedom, $x_i \sim t_5(\mu, 1)$. Specifically,

$$f(x_i \,|\, \mu) \propto \{5 + (x_i - \mu)^2\}^{-3}.$$

We assume again a uniform prior density for μ. Finally, we let the sixth observation be denoted by z and consider the posterior distribution of μ given data $\mathbf{x} = (1.5, 2.6, 0.3, 0.9, 2.2, z)$ as $z \to \infty$.

Figure 1 shows the resulting posterior densities for $z = 2$ (black curve), $z = 3$ (purple curve), $z = 5$ (blue curve), $z = 10$ (red curve) and $z = 25.5$ (green curve). We see that as z increases the posterior density initially ($z = 3, 5$) moves with it to be centred on higher values of μ, but then ($z = 10, 25.5$) it moves back, converging towards the posterior distribution that would apply based on only

the first five observations. Table 1 presents the posterior mean and variance values for these values of z and for the limiting posterior ($z = \infty$). We see not only the mean initially increasing and then converging back to its asymptotic value but we also see the variance initially increasing to a maximum around $z = 5$ before falling back towards the asymptotic value.

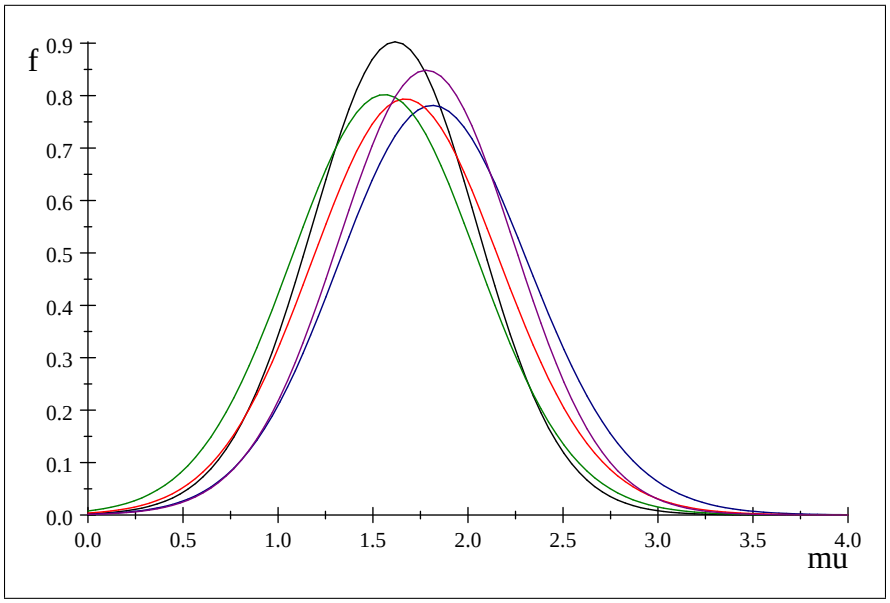

Figure 1. Posterior densities for five value of z.

z	$E(\mu \mid z)$	$var(\mu \mid z)$
2	1.61	0.201
3	1.78	0.226
5	1.83	0.271
10	1.68	0.262
25.5	1.57	0.256
∞	1.51	0.255

Table 1. Posterior means and variances for six values of z.

The influence of the outlying observation in this example is heavily discounted beyond $z = 5$, and at $z = 25.5$ the posterior distribution has effectively rejected the observation, being centred nowhere near the sample mean of 5.5, but instead very close to the posterior that would be obtained from just the first five observations.

1.3 Heavy-tailed modelling

The primary purpose of this article is to review the research showing how, when one or more of the distributions in a Bayesian analysis has heavy tails, and

when conflict becomes increasingly strong it is resolved by one or more of the information sources being rejected. The pioneering paper of Dawid (1973) and most subsequent research in this field does not assume particular distributional forms, but instead proves general results applicable for any distributions having specified heavy-tailed properties. Heavy-tailed distributions are by no means confined to the class of t distributions, nor even to scale mixtures of normals, but applications of this theory in Bayesian analyses of real problems has almost invariably involved t distributions. Their authors have usually simply replaced the normal distributions wherever they would typically be used by t distributions, in the belief that this would provide automatic robust posterior behaviour in response to outliers and other conflicts. For instance, Wakefield, Smith, Racine-Poon and Gelfand (1994) replace a bivariate normal distribution for parameters of a linear growth model (independently for each individual in the population) by a bivariate t distribution. They also suggest replacing normal distributions by t distributions for the observations, saying that this will provide "an analysis that is robust to both data outliers and outlying individuals in the population." Similarly, Meinhold and Singpurwalla (1989) employ t distributions for a time series of observations and for the underlying evolution in a Kalman filter model. Their approach allows for rejection of individual outlying observations, but through an approximation that does not take account of the joint influence of several outliers.

The truth is more complex. First, conflicts can be resolved in more than one way. Where two information sources conflict, we can resolve this by rejecting one information source or the other. We can also reject neither source. The posterior distribution may, as the separation between the sources becomes wider, continue to encompass values of the parameter consistent with both of the information sources, so that the posterior variance increases to infinity. It may also be 'resolved' in the way that normal distributions often do, by the posterior distribution concentrating on a compromise value that is supported by neither source. Any particular choice of model will resolve conflicts in a particular way, and it is important to understand how different modelling choices lead to different resolutions.

Second, the theory has to date been developed in rather simple kinds of models. Generalisation to more complex models is not immediate. Indeed, applications almost invariably propose models for which the theory has not been developed — we do not know yet how conflicts are resolved in complex heavy-tailed models.

So in addition to reviewing the research in the field, this article has a second important purpose, namely to identify the limitations and gaps in the literature that need to be addressed in order to understand fully the effects of heavy-tailed modelling in complex applications.

2 Single observation, single location parameter

2.1 Tails and duality

We begin with the simplest case, in which we have a single observation x having density $f(x-\theta)$, so that θ is a location parameter. We let θ have prior density $g(\theta)$. We can think of x as composed of the location parameter θ plus 'observation error' $\phi = x - \theta$. Notice that $x = \theta + \phi$, and that θ and ϕ are independent random variables with densities g and f respectively.

Our interest is in the limiting behaviour of the posterior distribution for θ when x becomes large. That limiting behaviour will depend on to what extent the posterior distribution attributes the large value of $x = \theta + \phi$ to θ being large or to ϕ being large, which in turn depends on the forms of their densities, g and f. An important feature of this model is a duality between θ and ϕ that was pointed out by Dawid (1973). Whatever results we can prove about the posterior distribution of θ in this system will apply instead to ϕ if we reverse the roles of f and g. In particular, suppose that for given f and g we can show that as $x \to \infty$ the posterior density of θ tends to its prior density g, so that the information in the observation x is rejected in the limit. Now suppose that we switch f and g, so that now f is the prior density of θ, then the posterior density of ϕ will tend to g, and hence the posterior density of θ is asymptotically $g(x-\theta)$. In this case, it is the prior information about θ that is asymptotically rejected. This illustrates the fact that conflicts between sources of information can always be resolved in more than one way. In this simple case of a single observation and a single location parameter, we can reject either of the two sources of information in favour of a limiting posterior distribution based only on the other information source.

A second kind of duality arises when we consider what happens as $x \to -\infty$. For large positive x, the posterior depends only on the right-hand tails of f and g, since the implication is that either θ or ϕ (or perhaps both) are large and positive. Any result for $x \to \infty$ can be converted to one for $x \to -\infty$ by changing signs in both distributions and looking at their left-hand tails.

2.2 Illustrative examples

Figures 2, 3 and 4 illustrate the possibilities. First consider Figure 2, where θ and ϕ are both given standard normal distributions. In Figure 2 the large image shows the joint density of θ and ϕ, which has the spherical bivariate normal form. The line marked A corresponds to $\theta + \phi = 0.5$ and so the posterior distribution of θ given $x = 0.5$ is the conditional density along this line (projected onto the θ axis). The panel (a) in Figure 2 shows this posterior density. Similarly, line B represents $\theta + \phi = 6$ and panel (b) is the conditional density along line B, i.e. the posterior density of θ given $x = 6$. We see how when we use normal distributions the posterior distribution is normal for any x; it has constant posterior variance and a posterior mean that increases with x. For $x = 6$ we have conflict between the prior distribution which suggests $\theta \in [-2, 2]$ (with

approximately 95% probability) and the observation which suggests $\theta \in [4, 8]$. The posterior distribution is centred at $\theta = 3$ with 95% interval approximately $[2, 4]$. This is a less extreme example than the one of six observations introduced in Section 1, but still shows the same questionable resolution of this conflict.

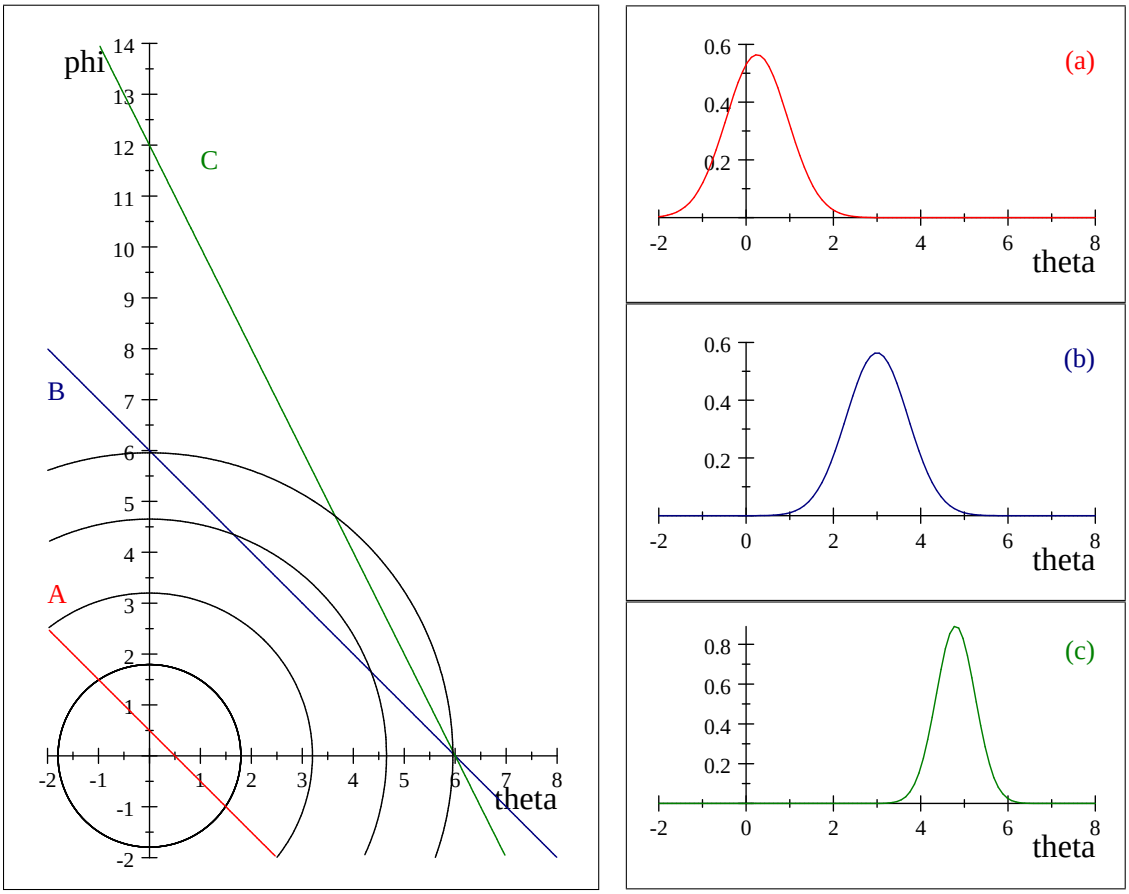

Figure 2. Joint and posterior densities in the case of normal f and g.

In the main panel of Figure 2 we have assumed equal variances for θ and ϕ, but variance is just a matter of scaling and we can readily use the same picture to look at a case of unequal variances. Line C and panel (c) correspond to $\theta + 0.5\phi = 6$. This represents the case where θ has the standard normal distribution but the error 0.5ϕ has a normal distribution with variance 0.25. The posterior density in panel (c) therefore applies to this case; it is again normal, with mean 4.8 and variance 0.2.

In Figure 3, the prior density g is still standard normal, but now f is the Cauchy density. The lines A, B and C are the same as in Figure 2, and the

densities (a), (b) and (c) are the corresponding posterior densities for θ obtained by conditioning along those lines. Thus, in (a) we see the posterior distribution after observing $x = 0.5$; there is no conflict here and the posterior distribution is similar to panel (a) of Figure 2. In (b), however, we see a very different resolution of the conflict created by observing $x = 6$. The posterior distribution is close to the Cauchy prior distribution, corresponding to rejection of x as uninformative about θ. We can see in the joint distribution that this result comes from the way the heavy-tailed Cauchy distribution stretches the contours of the joint distribution out of the circular form of Figure 2.

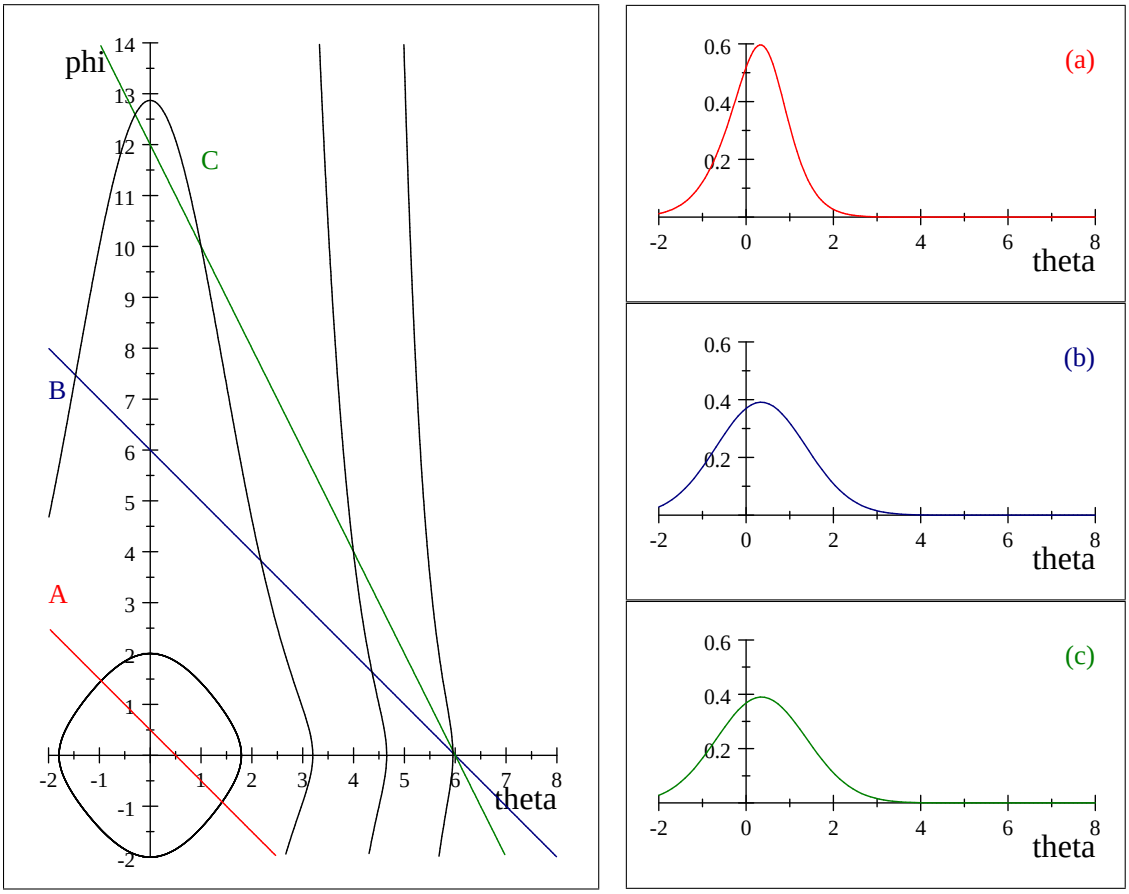

Figure 3. Joint and posterior densities in the case of Cauchy f and normal g.

The distribution (c) in Figure 3 is also interesting. Remember that line C mimics the case where the observation error standard deviation is halved relative to line B, but in both cases for an observation $x = 6$. Therefore the observation becomes more informative, and in Figure 2 this resulted in a posterior distribution for θ that is centred closer to the observation. In Figure 3, however,

the posterior density (c) is similar to (b), and so also represents rejection of the observation. The message is that variance is important in determining the posterior distribution in the absence of conflict (or when conflict is ignored), but that tail behaviour drives the result when we have conflict.

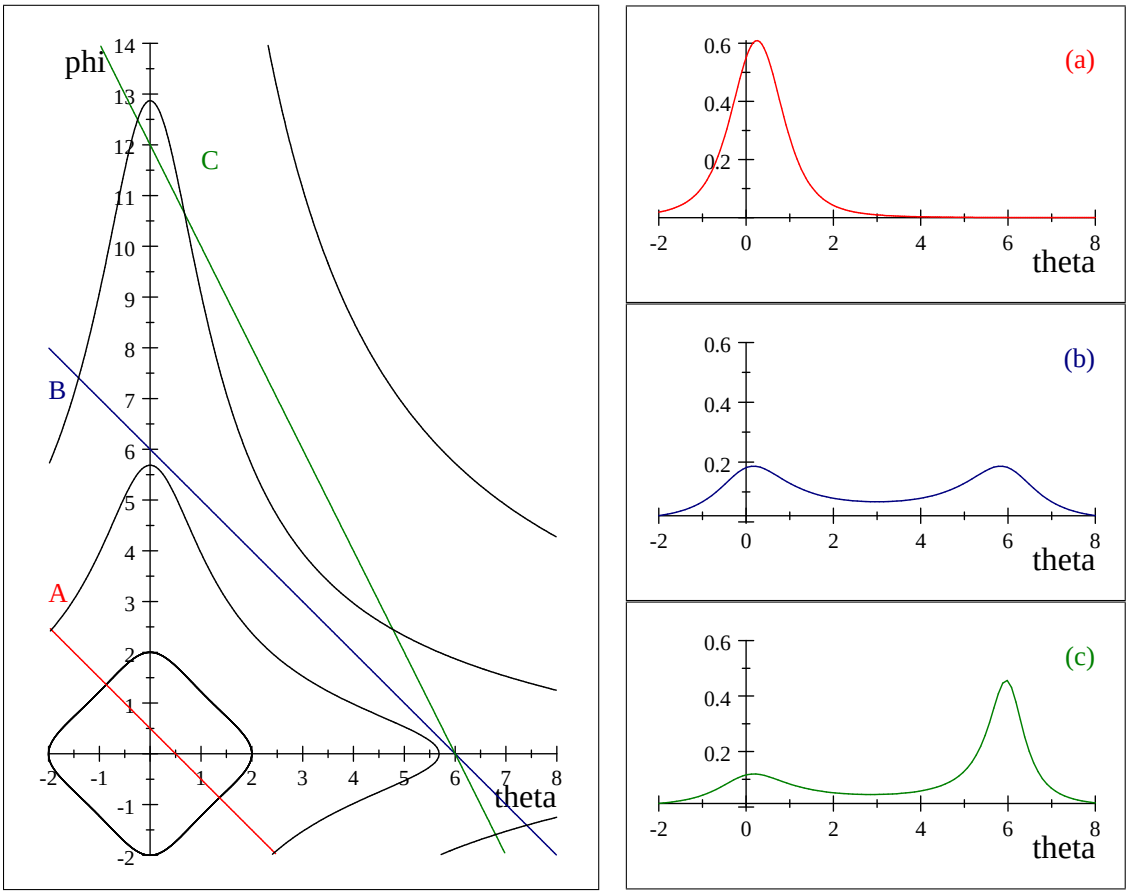

Figure 4. Joint and posterior densities in the case of Cauchy f and g.

These findings are reinforced in Figure 4, where both f and g are now Cauchy distributions. For the non-conflicting observation $x = 0.5$, the posterior density (a) is again like those of Figures 2 and 3. But panel (b) is again different. Now the posterior is bimodal. One mode corresponds to rejection of the observation as in Figure 2(b), but the other mode is centred on $\theta = 6$ and represents rejection of the prior information. Because the tail behaviours of f and g are identical, we have no basis to say which source of information should be rejected. That was also the case in Figure 2, but the resolution now is different and more credible — the posterior supports values of θ that are consistent with the prior or the data, but intermediate values of θ which are not consistent with either source

of information are not supported. The posterior distribution (c) is basically the same as (b), the only difference being that now the observation is supposed to have smaller error, and so the mode around $\theta = 6$ is narrower.

2.3 Asymptotic density results

Several results are available in the literature characterising combinations of f and g for which the limiting posterior distribution of θ as x tends to infinity is the prior density g. From duality, these results also characterise when the limiting posterior density is $f(x-\theta)$. Indeed, several of the results are couched in terms of rejection of the prior distribution, because where there is concern that the prior information may not be reliable the appropriate resolution of a conflict may be to reject the prior: for consistency of discourse we have converted these results herein to the dual conditions for rejection of the observation.

The theorems concern the right-hand tails of f and g, but analogous conditions on the left-hand tails will characterise the limiting posterior as x tends to minus infinity.

2.3.1 Dawid's conditions

First, Dawid (1973) gave the following sufficient set of conditions.

(A1) Given $\varepsilon > 0$, $h > 0$, there exists A such that if $y > A$ then

$$|f(y') - f(y)| < \varepsilon f(y)$$

whenever $|y' - y| < h$.

(A2) For some constants B, M,

$$0 < f(y') < Mf(y)$$

whenever $y' > y > B$.

(A3) $\int^{\infty} k(\theta)g(\theta)d\theta < \infty$, where $k(\theta) = \sup_x \{f(x-\theta)/f(x)\}$.

O'Hagan (1979) sought to simplify Dawid's condition (A3), which may be difficult to verify in practice because of the need to derive $k(\theta)$. He showed that if (A2) and (A3) were replaced by the following (slightly stronger) conditions then together with (A1) they would still be sufficient for the posterior density of θ to tend to $g(\theta)$ as $x \to \infty$.

(B2) (a) $f(y)$ is continuous and positive for all y.

(b) There exists a B such that for all $y \geq B$

i. $f(y)$ is decreasing in y,

ii. $d \log f(y)/dy$ exists and is increasing in y.

(c) There exists a C such that, for all $y \leq C$, $f(y)$ is increasing in y.

(B3) $\int^{\infty} \{f(\theta)\}^{-1} g(\theta) d\theta < \infty$.

In simple terms, the role of condition (A1) is to ensure that f is heavy-tailed by requiring that for large enough y the density $f(y)$ becomes arbitrarily flat, while conditions (A2) and (B2) are regularity conditions on f. The purpose of conditions (A3) and (B3) is to ensure that, even though g may also be heavy-tailed, its right-hand tail is thinner than that of f.

The most widely used examples of heavy-tailed distributions are the t distributions. In general, we will say that f is a t density with d degrees of freedom if it has the form

$$f(y) \propto \left\{d + s^{-2}(y-m)^2\right\}^{-(d+1)/2} ,$$

where m and s^{-2} are fixed parameters. (A non-zero value for m would represent a bias in the error distribution, while s^{-2} determines its precision.) This distribution is easily found to satisfy (A1), (A2) and (B2). Dawid's function $k(\theta)$ is asymptotically proportional to θ^{d+1}, which is of course also the asymptotic form of $\{f(\theta)\}^{-1}$, and so (A3) and (B3) are satisfied provided the prior distribution possesses moments of order $d+1$. This will in particular be true if g is normal, but also if g is another t distribution with degrees of freedom $d' > d+1$. Remembering that a Cauchy distribution is t with one degree of freedom, these results confirm what is found in Figures 2 and 3. Notice also that the asymptotic rejection of the observation occurs regardless of any bias and precision parameters, although they will influence the rate of approach to the limit.

In Figure 4 the two t distributions have equal degrees of freedom, and so no rejection occurs, either of the observation or of the prior information. However, the results of Dawid (1973) and O'Hagan (1979) suggest that the same situation will arise when the degrees of freedom differ by 1 or less, yet empirically if we compute posterior distributions as in Figure 4 for a case where f and g are t densities with degrees of freedom $d < d' \leq d+1$ we still find the posterior distribution tending to g as $x \to \infty$. This highlights the fact that these conditions are sufficient for rejection of the observation but not necessary.

Meinhold and Singpurwalla (1989) prove that rejection of the observation occurs in the limit for two t distributions whose degrees of freedom differ by an arbitrarily small amount. We now consider some rather more general results in which f and g are not limited to being t densities.

2.3.2 Credence and regular variation

O'Hagan (1990) introduces the notion of *credence* according the following definition.

Definition 1 *A density $f(y)$ has credence c if there exist $K \geq k > 0$ such that for all $y \in \mathbb{R}$,*

$$k \leq (1+y^2)^{c/2} f(y) \leq K .$$

Thus f has credence c if it is bounded above and below by multiples of a t density with $c-1$ degrees of freedom (and in particular the t density itself has credence $d+1$). He then proves that if f has credence c and g has credence $c' > c$ then

(a) for any given $d > 0$, there exists an A such that for all $|d| > A$ the posterior density is bounded for all $|\theta| \leq d$ above and below by multiples of g, and

(b) for all $r : \mathbb{R}^+ \to \mathbb{R}^+$ such that $r(x) \to \infty$ as $x \to \infty$, the posterior probability that $P(|\theta| > r(|x|)$ tends to 0 as $|x| \to \infty$.

This result does not quite show that the posterior density tends to the prior density g. Because the tails of f are only constrained to be flat within upper and lower bounds result (a) says only that the posterior is bounded by multiples of g. However, (b) shows that posterior probability outside $\pm ax$ (or even $\pm a \ln x$) for any positive a (no matter how small) tends to zero, and so the observation is indeed asymptotically rejected. Furthermore, a theorem of Hill (1975) implies that if the posterior converges uniformly to any density it must be g.

The t distributions are examples of a broader class of distributions $f(y)$ whose extreme right-hand tails behave like a power of y, and are often said to have polynomial tails. This wider class provides useful additional flexibility for modelling compared with just the class of t distributions. The class of distributions with finite credence is broader than the class of t distributions but there are other ways to define distributions with polynomial tails. Andrade and O'Hagan (2006) consider distributions with *regularly varying* tails.

Definition 2 *The right-hand tail of a density $f(y)$ is regularly varying with index ρ if*

$$\frac{f(\lambda y)}{f(y)} \to \lambda^{\rho}$$

as $y \to \infty$ for all $\lambda > 0$.

The regular variation index for a proper density must be negative, so Andrade and O'Hagan say that f has *RV-credence* c if its right-hand tail is regularly varying with index $-c$. A t distribution with d degrees of freedom has credence $d+1$ and also RV-credence $d+1$, but the two definitions are not equivalent. Andrade and O'Hagan (2006) introduce the new conditions

(C1) f has RV-credence c,

(C3) g has RV-credence $c' > c+1$,

and prove that (C1), (B2) and (C3) together imply the conditions (A1), (A2) and (A3) of Dawid (1973), and hence that the posterior density will tend to the prior density g.

Another similar result is the Generalised Polynomial Tails Theorem of Fúquene, Cook and Pericchi (2009). Their conditions are

(D1&2) There exist constants A_1, c, C_1, C_2 and C_3 such that for all $y > A_1$

$$C_1 y^{-c} \leq f(y) \leq C_2 y^{-c} \ ,$$
$$\frac{d}{dy} f(y) \leq C_3 y^{-c-1} \ .$$

(D3) There exist constants A_2, $c' > c$ and C_4 such that for all $m > A_2$

$$\int_{\theta > m} g(\theta) d\theta \leq C_4 m^{-c'} \ .$$

Conditions (D1&2) say that the right-hand tail of f is polynomial, being bounded above and below by multiples of y^{-c}, and impose regularity so that the tail does not wiggle too much. Condition (D3) says in effect that the tail of g is thinner than a polynomial of form $\theta^{-(c+1)}$. Under conditions (D1&2) and (D3), Fúquene, Cook and Pericchi (2009) prove that the observation will be rejected as $x \to \infty$. In the case of t distributions, (D3) is like (C3) and Dawid's conditions in the sense of only guaranteeing rejection when the degrees of freedom of f and g differ by more than 1.

2.3.3 Regular log-convex tails

Recently, Desgagné and Angers (2007) have introduced another class of distributions. We will say that a density f has a *regular log-convex* right-hand tail if it is positive, bounded above and satisfies both the condition (A1) and a new condition:

(E2) There exist constants $A_2 > 0$ and $M > 1$ and proper density functions f^* and f^+ such that for all $y > A_2$

$$\frac{f^2(y/2)}{f(y) f^+(y)} \leq M \ ,$$
$$\frac{d^2}{dy^2} \log f^*(y) \geq \frac{d^2}{dy^2} \log f^+(y) \geq 0 \ ,$$

where f^* must be such that there exist constants $B > 0$ and $0 < K_1 < K_2 < \infty$ for which $K_1 \leq f(y)/f^*(y) \leq K_2$ whenever $y > B$.

Condition (E2) is deliberately weakened through the introduction of densities f^* and f^+. It is satisfied for a t density simply by setting both f^* and f^+ equal to f, but the additional flexibility widens the class of distributions appreciably. For instance the introduction of f^* allows the tail of f to be not strictly log-convex but simply to be bounded by one that is. Desgagné and Angers (2007) then prove that the observation is asymptotically rejected as $x \to \infty$ if f has a regular log-convex right-hand tail and

(E3) $\lim_{y \to \infty} g(y)/f(y) = 0$.

Although the conditions for f and g to have regular log-convex tails are complex, the final condition (E3) is particularly straightforward. In particular, it shows immediately that rejection occurs for t densities whose degrees of freedom differ by an arbitrarily small amount.

We have considered four different formulations: Dawid's (with O'Hagan's variant of his conditions), credence, RV-credence and regular log-convexity. Each approach proves rejection of the information in the density f when it conflicts with the information in g, under different combinations of f and g. The case when both f and g are t densities is covered in every one of the formulations, but each encompasses combinations that are not covered by the others.

- The theory of credence requires both f and g to have finite credence, whereas other approaches allow g to be any kind of distribution subject only to it being lighter-tailed than f in the sense of condition (A3), (B3), (D3) or (E3). Andrade and O'Hagan allow g to be rapidly varying as well as regularly varying, which for instance includes the case of g being normal.

- Distributions with finite credence can have tails that 'wiggle' in ways that are not allowed by Dawid's conditions or regular variation.

- A density $f(y)$ whose tails are proportional to a t density with d degrees of freedom multiplied by a slowly-varying function like $\log y$ does not have any credence value but is covered by the regular variation approach with RV-credence $d + 1$. Desgagné and Angers prove that such a f is heavier-tailed than the t distribution with d degrees of freedom and the observation will be rejected in the limit if g has that t distribution.

- There are other kinds of distribution, for instance those whose tails decay like $\exp(-y^b)$ for $0 < b < 1$, that are covered by Dawid's conditions and by Desgagné and Angers' log-convexity conditions, but these are lighter-tailed than the t distributions.

For practical Bayesian modelling interest usually focuses on more heavy-tailed distributions, for which there may be a more rapid transition to the appropriate form of resolution as conflicts arise. In practice it is the distributions with polynomial tails that are of greatest importance, particularly the t distributions, and particularly those with low degrees of freedom.

2.4 Asymptotic results for moments

Dawid showed that subject to the additional condition

(A4) $\int^{\infty} m(\theta)k(\theta)g(\theta)d\theta < \infty$

the posterior expectation of $m(\theta)$ tends to its prior expectation. In the case of f having a t distribution with d degrees of freedom, this means that the posterior

moments of order up to n converge to the corresponding prior moments if g is normal, or if g is a t density with degrees of freedom $d' > d + p + 1$. The same will be true under the alternative conditions of O'Hagan (1979) and Andrade and O'Hagan (2006). In fact, again we find empirically that this convergence holds if $d > d + p$, but O'Hagan (1990) only obtains the same requirement, $d' > d + p + 1$, using the credence approach. The tighter condition $d' > d + p$ is proved specifically for the case of two t distributions by Fan and Berger (1992), at least for the mean and variance ($p = 1$ or 2), but this issue is now fully resolved by Desgagné and Angers (2007). Their condition for convergence of the posterior expectation of $m(\theta)$ to its prior expectation is simply

(E4) $\lim_{y\to\infty} m(y)g(y)/f(y) = 0$.

2.5 Some related research

Some authors have considered conditions under which a particular source of information will *not* be rejected. For instance, O'Hagan (1979) showed that if f is a normal density then the observation will not be rejected, no matter what form g might take. He remarks that for the double exponential distribution $f(y) \propto \exp(-a\,|y|)$ the observation x is not rejected for any g but its influence on the posterior distribution is bounded. This is explored further by Pericchi and Smith (1992) and Mitchell (1994), and generalised to other distributions with bounded influence by Pericchi and Sansó (1995). (Note that much of the work on a single observation is focused on the dual case of rejecting a heavy-tailed prior distribution rather than a heavy-tailed observation, but we have recast those findings in the framework of heavy-tailed f for ease of comparison with the principal results above.)

Others have provided results on the posterior distribution for finite x under some kinds of heavy-tailed distributions for f or g. O'Hagan (1981) showed that when g is normal and f is heavy-tailed, so that as $x \to \infty$ the observation will ultimately be rejected, then for some finite x the posterior variance will reach a local maximum before decreasing to its asymptotic value. He characterised this as resulting from 'indecision', the posterior accommodating both conflicting information sources before resolving the conflict by rejecting the observation. Fan and Berger (1992) present a number of results for the case when both f and g are t distributions, such as conditions for the posterior density to be unimodal and an expression for how large the local maximum of the posterior variance can be. Meeden and Isaacson (1977) obtain some results on the rate of convergence of the posterior mean to the prior mean. Goldstein (1983) shows that the convergence is monotone in distribution when g is strongly unimodal (log-concave).

Choy and Smith (1997) considered scale mixtures of normal distributions for f, as originally envisaged by de Finetti (1961), in the special case when g is normal. They showed that the family of stable distributions are heavy-tailed and result in rejection of the observation, while the exponential power family with densities $f(y) \propto \exp(-a\,|y|^b)$ lead to bounded influence rather than

rejection for values of b greater than 1 (the double-exponential distribution) but less than 2 (the normal distribution). Exponential power distributions are also covered as a special case in Angers (2000), who extends O'Hagan's definition of credence. Angers defines $f(y)$ to have p-credence (b, a, c, d) if it can be bounded by multiples of

$$\exp(-a(y^*)^b)(y^*)^{-c}\log^{-d}(y^*)\ , \tag{1}$$

where (to avoid singularities at the origin) $y^* = \max(|y|, y_0)$ for some positive y_0. By convention, $a = 0$ when $b = 0$. Angers refers to the distributions with densities proportional to (1) as the *generalised exponential power* (GEP) family. His results generalise those of O'Hagan (1990), whose credence c corresponds to p-credence $(0, 0, c, 0)$, and in particular show that the exponential power family with p-credence $(b, a, 0, 0)$ is heavy-tailed for $b < 1$ (a finding that is also covered by O'Hagan, 1979, and proved in a different way by Choy and Walker, 2003). Desgagné and Angers (2007) redefine p-credence: $f(y)$ is now said to have p-credence (b, a, c, d) in its right-hand tail if

$$\lim_{y\to\infty} \frac{f(y)}{\exp(-a(y^*)^b)(y^*)^{-c}\log^{-d}(y^*)} = K$$

for some K. With this modification they show that the distributions with p-credence (b, a, c, d) for $b < 1$ have regular log-convex tails, so that their basic results apply for any f and g in this wide class of densities.

2.6 Gaps in the theory

For the important case where both f and g are t densities, or where one is a t and the other is normal, we have quite comprehensive results on conflict resolution. Bayesian modellers may wish to use other kinds of heavy-tailed distribution, but it is likely that all cases of practical importance will be covered by the theory of Desgagné and Angers (2007). Their results cover in particular the wide class of distributions with tails in the GEP family, and do not require densities to be symmetric or to have the same tail behaviour in both tails.

3 General location parameter models

3.1 Multiple observations, single location parameter

The next simplest situation is where we have n observations, $x_1, x_2, \ldots, x_n$, rather than just one. We suppose that they have densities $f_i(x_i - \theta)$, $i = 1, 2, \ldots, n$, so that θ is a location parameter for every observation but they can have different distributions. The case when $f_i = f$ for all i is the important special case of a sample of iid observations. The prior distribution for θ is g as before. Interest now focuses on rejection of outlying observations, with the posterior distribution asymptotically tending to the posterior that would arise from the non-outlying observations alone. In this context, it is useful to think

of there being $n+1$ sources of information which might conflict with each other, and to write the prior distribution as

$$f_0(x_0 - \theta) = g(\theta) \ .$$

In this formulation, x_0 is a mean or location value for the prior distribution, and the prior conflicts with the data if x_0 is far from the other x_is. Notice that the right-hand tail of g becomes the left-hand tail of f_0.

First suppose that x_n is a single outlier, so that we imagine $x_n \to \infty$ while all the other $x_0, x_1, \ldots, x_{n-1}$ remain fixed. This can be reduced to the problem of Section 2 if we redefine g to be the posterior distribution of θ after observing the $n-1$ fixed observations, and f is the density f_n of the outlying observation. The conditions discussed in Section 2.3 now determine when the outlying observation is rejected, but in order to apply these we need to be able to identify the properties of the right-hand tail of the redefined g. If we wish to study rejection of several outliers then we also need to consider the tail behaviour of groups of observations.

Both O'Hagan (1979) and Desgagné and Angers (2007) address outlier rejection in these general terms. Consider m groups of observations, identified by subsets S_j, $j = 1, 2, \ldots, m$, of the indices. Thus, $\cup_{j=1}^{m} S_j = \{0, 1, 2, \ldots, n\}$ and $S_j \cap S_{j'} = \varnothing$ when $j \neq j'$. We suppose that the observations in group 1 remain fixed while the other groups move increasing far apart from the first group and from each other. Formally, for $i \in S_j$ we write $x_i = \bar{x}_j + z_i$, so that $\bar{x}_j$ is a reference point for group j and the z_is denote deviations of the observations from their respective reference points. Then we let the reference points $\bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m$ tend to ∞ or to $-\infty$ such that the separations $|\bar{x}_j - \bar{x}_{j'}|$ all tend to infinity, while the z_is remain fixed. For instance, we could have $\bar{x}_j = u_j w$, where $u_1 = 0, u_2, \ldots, u_m$ are fixed and distinct real values and let $w \to \infty$. We now seek conditions under which the posterior distribution of θ tends to the posterior

$$g^*(\theta) \propto \prod_{i \in S_1} f_i(x_i - \theta)$$

that would arise given only the information sources in group 1.

O'Hagan (1979) proves that if f_i has credence c_i, $i = 0, 1, 2, \ldots, n$ then the asymptotic rejection of all the other groups occurs provided $C_1 = \max_j C_j$, where $C_j = \sum_{i \in S_j} c_i$. That is, the group with the largest total credence dominates and ultimately rejects all the others. In the special case of iid observations, the largest group will dominate. (The prior information could count as one of the observations if f_0 has the same credence as the other f_is. Alternatively, it would be ignored if we assume a uniform prior, since that corresponds to $c_0 = 0$.)

Desgagné and Angers (2007) consider the case of $m = 2$ or 3. First, if the members of group 2 all tend to $+\infty$, and if the densities f_i have regular log-convex right tails for $i \in S_2$ then group 1 dominates group 2 provided

$$\lim_{y \to \infty} \frac{\prod_{i \in S_1} f(x_i - y)}{\prod_{i \in S_2} f(y)} = 0 \ .$$

Next, if members of group 3 all tend to $-\infty$, and if the densities f_i have regular log-convex left tails for $i \in S_3$ then group 1 dominates group 3 provided

$$\lim_{y \to -\infty} \frac{\prod_{i \in S_1} f(x_i - y)}{\prod_{i \in S_3} f(y)} = 0 \ .$$

Group 1 dominates over both groups 1 and 2, with the posterior distribution of θ tending to that arising from just observing group 1, if both sets of conditions hold. Desgagné and Angers (2007) give corresponding results on limits of posterior expectations. They also provide simpler conditions based on p-credence values for the case where all the f_is have GEP tails.

Although Desgagné and Angers consider only 3 groups and O'Hagan considers only symmetric densities, it seems probable that both results could be generalised to deal with many groups and the possibility of different right and left tails.

It is important to recognise in this model, as in all practical Bayesian modelling, that different tail thicknesses can lead to different resolutions of conflict. For example, if g is a t distribution with credence 9, while each of the f_is is a t distribution with credence 2 (a Cauchy distribution), then $n = 4$ observations that are similar to each other but very far from the prior mean will be rejected, but if $n = 5$ it is the prior that is rejected. Such a resolution of conflict would be reasonable in many situations where we would want to give precedence to the prior over a small number of conflicting observations but when the weight of sample evidence is enough we should abandon the prior. Simply using always a Cauchy distribution for every source of information (for instance, because it is a t distribution with just about the heaviest tail and so leads to the swiftest resolution) is not good modelling practice.

3.2 Multivariate location

For problems with more than one parameter, the theory is far less complete. The first generalisation of the single location parameter that we can consider is the multivariate location parameter, where one or more observations $\mathbf{x}_i$ are distributed with densities $f_i(\mathbf{x}_i - \theta)$, $i = 1, 2, \ldots, n$, where now $\mathbf{x}_i$ and θ are vectors of p elements, and θ has prior density $g(\theta)$. Hill (1975) addresses the case of a single observation $\mathbf{x}$ in this framework. He presents conditions under which the posterior distribution tends to the prior distribution as $||\mathbf{x}|| \to \infty$, under a given norm $||.||$. His conditions are rather complex, but it should be possible to generalise the various results presented above for $p = 1$ to this case under natural generalisations of their conditions. However, additional complications arise.

Hill's $||\mathbf{x}|| \to \infty$ supposes the observation tending to infinity in any direction and his conditions entail $f(\mathbf{y})$ having uniformly heavy tails as $||\mathbf{y}|| \to \infty$, but it is easy to see that different asymptotics might apply in different directions. O'Hagan and Le (1994) point out that tails of a multivariate distribution can be of different degrees of thickness in different directions. They contrast two

bivariate distributions, the bivariate t distribution with density

$$f_1(y_1, y_2) \propto (1 + y_1^2 + y_2^2)^{-3}$$

and the product of independent univariate t densities

$$f_2(y_1, y_2) \propto (1 + y_1^2)^{-2}(1 + y_2^2)^{-2} \ .$$

If we hold y_1 fixed and let $y_2 \to \infty$, then f_1 decays as y_2^{-6} while f_2 decays like y_2^{-4}, so f_2 has a heavier tail in this direction (and similarly if we hold y_2 fixed and let $y_1 \to \infty$). Yet if we set $y_1 = y_2 = y$ and let $y \to \infty$, then f_1 again decays like y^{-6} while f_2 decays as y^{-8}. In this direction f_1 is heavier tailed than f_2. The bivariate t distribution f_1 has uniform tail thickness in every direction; its contours are circular like those in Figure 2. The product of independent t distributions, in contrast, has contours which become increasingly star-shaped as we move into the tails, like those in Figure 4; the tail thickness is different in the y_1 and y_2 axis directions from in any other direction.

O'Hagan and Le (1994) introduce a family of bivariate heavy-tailed distributions, the bivariate T family, generalising the two forms above. The $T(c, c_1, c_2)$ distribution is defined to have density function

$$f(y_1, y_2) \propto (1 + y_1^2 + y_2^2)^{-c/2}(1 + y_1^2)^{-c_1/2}(1 + y_2^2)^{-c_2/2} \ . \tag{2}$$

O'Hagan and Le provide several numerical examples to illustrate different asymptotics for a single bivariate observation when both f and g have bivariate T distributions. Theoretical results supporting the numerical examples are given by Le and O'Hagan (1998), who parameterise the bivariate T distributions using $s_1 = c + c_1$, $s_2 = c + c_2$ and $s_3 = c + c_1 + c_2$. For instance, they prove that (2) defines a proper distribution if $s_1 > 1$, $s_2 > 1$ and $s_3 > 2$.

We will say that f has T-credence $\mathbf{s} = (s_1, s_2, s_3)$ if it can be bounded above and below by multiples of the $T(s_1 + s_2 - s_3, s_3 - s_2, s_3 - s_1)$ density. Suppose that f has T-credence $\mathbf{s}$ and g has T-credence $\mathbf{s}'$. Le and O'Hagan (1998) prove the following results.

- If $s_1 < s_1'$, $s_2 < s_2'$ and $s_3 < s_3'$ then f has heavier tails in every direction than g. As $\mathbf{x} \to \infty$ in any direction the probability that θ is in any neighbourhood of $(0, 0)$ tends to 1. In this case the observation is rejected.

- If $s_1 \geq s_1'$, $s_2 \leq s_2'$ and $\min(s_3, s_3') > s_1' + s_2$ then as $\mathbf{x} \to \infty$ in any direction the probability that θ is in any neighbourhood of $(0, x_2)$ tends to 1. In this case the observation is rejected in respect of learning about θ_1 but it is the prior information that is rejected in respect of learning about θ_2.

Le and O'Hagan prove these and other results to illustrate the variety of conflict resolutions that are possible with f and g in this family of bivariate heavy-tailed distributions. Their findings, however, barely scratch the surface of the behaviour of heavy-tailed multivariate distributions. First, even for this

bivariate family, their theorems do not exhaust all the possible combinations of $\mathbf{s}$ and $\mathbf{s}'$ triples; second, this family does not by any means exhaust the possible bivariate forms of heavy tailed density; third, it is surely true that even more complexity will be possible in three or more dimensions.

It seems likely that quite general results could be proved to the effect that if f has heavier tails than g in every direction, then as $||\mathbf{x}||$ tends to infinity for any norm the observation will be rejected and the posterior density will tend to g in the limit. However, not only are there many more ways to resolve the conflict but these ways are also practically important as the next section demonstrates.

3.3 Exchangeable locations

A closely related model is the one-way analysis of variance model, which we write here in a general hierarchical form.

1. Observations x_i have densities $f_i(x_i - \theta_i)$ for $i = 1, 2, \ldots, p$. Often we have replication and would suppose that observations x_{ij} have densities $f_i(x_{ij} - \theta_i)$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n_i$.

2. Parameters θ_i have independent densities $g_i(\theta_i - \xi)$ given ξ, $i = 1, 2, \ldots, p$.

3. The hyperparameter ξ has density $h(\xi)$.

Typically, the θ_is all have the same density given ξ. This is the case of exchangeable parameters. Similarly, the observations typically have a common density, but the extra generality of allowing different densities f_i and g_i may sometimes be useful in practice. Note that we can reduce this model to the multivariate location model of Section 3.2 by integrating out the hyperparameter ξ. Thus

$$g(\theta) = \int h(\xi) \prod_{i=1}^{p} g_i(\theta_i - \xi) \, d\xi$$

and $f(\mathbf{x} - \theta) = \prod_{i=1}^{p} f_i(x_i - \theta_i)$. However, the resulting distribution g may be complex to deal with, and we may also be interested in inference about ξ. It is therefore usual not to collapse the hierarchical model.

The analysis when all of the distributions are normal is straightforward and widely used in Bayesian applications, but more reasonable posterior behaviour in conflict situations will be achieved by using heavy-tailed distributions. Based on the results in previous sections, we can speculate on how conflict will be resolved with the aid of some diagrams.

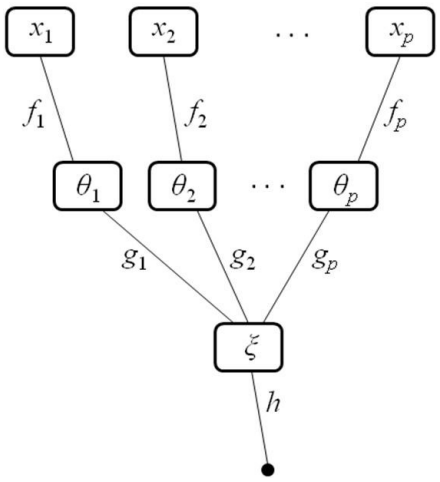

Figure 5. Exchangeable means, no conflict.

Figure 5 shows a situation when there is no conflict. In this diagram, the horizontal locations of the boxes represent the observations and the posterior parameter estimates. The vertical locations are simply arranged to show the different levels of the hierarchy. The lines between boxes indicate the distributions, linking each box to its location parameter in the next level of the hierarchy, and we can think of each line as acting like a spring to pull the corresponding pair of boxes together. The observations x_i have fixed values. They are grouped together, with no apparent outliers. The corresponding estimates of the θ_is are shrunk together by the hierarchical model, because in addition to being linked to the corresponding x_i they are linked to ξ. The relative strengths (precisions) of the distributions determine how much shrinkage occurs. Finally, ξ is linked to its prior mean, another fixed point in the diagram, and we can see that this has pulled ξ a little way from the centre of the θ_is. This is the kind of situation that would apply with normal distributions, and would be qualitatively the same with heavy-tailed distributions when there is no conflict. Conflict can arise by moving any of the fixed points sufficiently far apart.

Figure 6 shows such a situation, where x_p is an outlier, lying far from the other x_is (and from the prior mean of ξ). The resolution of this conflict depends on the weights of the different tails. In Figure 6, it is supposed that f_p has thinner right-hand tail than g_p, and so the prior has been rejected. This is indicated in Figure 6 by the absence of the link represented by g_p. (We can suppose that the spring has broken under the strain, and think of tail thickness as determining the ability of a spring to stretch without deforming and ultimately breaking.) As a result, θ_p is not shrunk towards ξ, and its posterior distribution is given solely by the observation term $f_p(x_p - \theta_p)$.

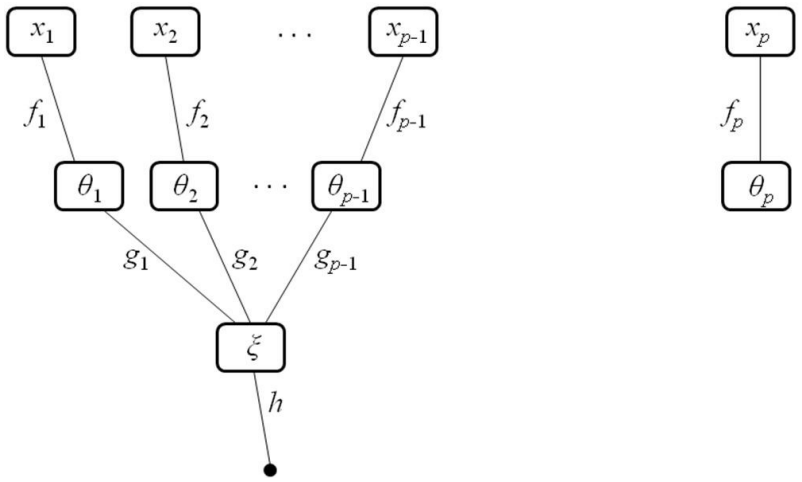

Figure 6. Exchangeable means with an outlier.

The conflict would be resolved differently if g_p had thinner tails than f_p. Then it would be the f_p link that broke, with the result that the posterior distribution of θ_p would be based only on that of ξ. If there is replication, then outlying observations within each group might be rejected, and then the balance between rejecting the data and rejecting the prior might depend on how many (non-outlying) observations are in a group. This kind of heuristic reasoning was presented in O'Hagan (1988) and illustrated with a numerical example. But such behaviour cannot be proved just from the theory regarding a single location parameter or results in the preceding Section 3.2. However intuitively reasonable the suggestions may be, they are unproven and could be wrong. For instance, the tail thickness of h may matter.

Angers and Berger (1991) prove that the behaviour described in Figure 6 arises in the specific case where the $f_i s$ are normal and the g_is are Cauchy. Choy and Smith (1997) give numerical examples of the same behaviour when the g_is have other heavy-tailed distributions.

3.4 Gaps in the theory

We have quite complete theory for the case of many observations and a single location parameter, as set out in Section 3.1. Unfortunately, in models with two or more parameters we have only a few sparse results. We need more general theory of multivariate heavy-tailed distributions, addressing more of the potential complexity that is opened up by allowing different tail thicknesses in different directions. We need some theory dealing with the interaction of heavy-tailed distributions at different levels of a hierarchical model. Even the simplest of all hierarchical models has not been addressed fully, and yet the

norm in practical Bayesian statistics is to build much more complex hierarchical models. The combination of hierarchical models and more general heavy-tailed distributions is illustrated by the proposal in O'Hagan (1988) of the following exchangeable prior distribution for $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$:

$$f(\theta) = \prod_{i<i'} u(\theta_i - \theta_{i'}) \ ,$$

where u is a t distribution or other heavy-tailed distribution. A diagram like Figure 5 for this case would have a line (spring) connecting every pair of θ_is. O'Hagan argues that then a group of outlying x_is could lead to the θ_i estimates being shrunk within each group but with no shrinkage between groups. This would be a very natural resolution of conflict between groups, but does not happen with the hierarchical forms discussed in Section 3.3. O'Hagan illustrates this behaviour with a numerical example, but no proof is offered.

In the usual linear regression model $y_i = \mathbf{x}_i'\beta + e_i$, the parameters β are often thought of as analogous to location parameters, generalising the case of $\mathbf{x}_i'\beta = \theta$. West (1984) considers this model, allowing both the e_is and β to have heavy-tailed distributions. He argues, on the basis of the score or influence functions, that individual observations or individual components of the prior distribution may be rejected in cases of conflict. Again, though, no formal proof or theory has been developed for linear models.

3.5 Scale mixtures of normals

One way to generate heavy-tailed distributions, and t distributions in particular, is as scale mixtures of normal distributions. Thus, if f_N is the density of the $N(0, \omega)$ distribution and we let ω have a density $p(\omega)$, then integrating out ω gives the density

$$f(y) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \omega^{-1/2} \exp(-\omega^{-1} y^2/2) p(\omega) \ d\omega \ .$$

If p is an inverse-gamma density, then f is a t density. Other choices of $p(\omega)$ can give exponential power distributions or stable distributions. If the tail of p is sufficiently heavy then f will be a heavy-tailed distribution. The representation as a scale mixture of normals can facilitate computation of the posterior distribution by MCMC; see for instance West (1981, 1984), Carlin and Polson (1991), Choy and Smith (1997). Essentially, rejection of an information source represented by a scale mixture arises through the posterior distribution of ω concentrating on larger values of ω. As the posterior distribution of ω 'goes to infinity', the information is deemed increasingly uninformative and it is ultimately rejected.

With multiple observations, heavy-tailed distributions f_i can be modelled as scale mixtures, each with its own ω_i. Therefore individual observations can be rejected when they conflict with the remaining observations, and the posterior estimates of the ω_is provide an indication of which observations are being discounted as outliers. However, it is useful to clarify here the distinction between

this use of scale mixtures and a similar construct. In all of the models we have consider thus far, we have effectively assumed known variances.

Suppose we have a sample from a normal distribution with unknown mean and variance. Thus x_i has the distribution $N(\theta, \sigma^2)$ with both θ and σ^2 unknown. If σ^2 has an inverse-gamma distribution, which is the standard conjugate prior distribution, then after integrating out σ^2 each x_i has a t distribution. But these are not independent t distributions. Instead, the x_is jointly have a multivariate t distribution. The difference is substantial when conflict arises. Instead of being able to reject individual observations as outliers (corresponding to large individual ω_i values), we can only reject the entire sample or none of it. This is because there is just a single mixing scale parameter σ^2. Large observations (conflicting with the prior distribution) will, given a suitably lighter-tailed prior, lead to a large posterior estimate of σ^2, and thereby the whole sample is discounted and eventually rejected. In practice, it is very important to give each information source its own ω_i (which usually means having independent heavy-tailed distributions), so that each may be rejected in the appropriate situation.

4 Adding a scale parameter

4.1 Single scale parameter

Almost all applied Bayesian models will have one or more unknown scale parameters. The presence of scale parameters raises two questions about heavy-tailed models and the resolution of conflicts. First, if we focus on posterior inference about the scale parameter(s), do heavy-tailed models result in rejection of sample information or prior information regarding the scale parameter(s)? Second, if we focus on location parameter(s), does the presence of unknown scale parameters change the theory in Sections 2 and 3?

Consider the situation where we have a single observation with density $\theta^{-1} f(x/\theta)$, so that θ is a scale parameter, and let the prior density of θ be $g(\theta)$. When x becomes large, a conflict arises between the observation, which suggests θ should have a value in the neighbourhood of x, and the prior which says θ should have values in the neighbourhood of the finite and fixed prior mean.

Before proceeding to study this further, it is useful to ask whether it can be dealt with by transformation to the location parameter case. For, if we write $x^* = \log x$ and $\theta^* = \log \theta$, then the density of x^* is $f^*(x^* - \theta^*)$ where $f^*(y) = e^y f(e^y)$. Section 2 then deals with the asymptotic behaviour of the posterior distribution of θ^* as $x^* \to \infty$. It will say when the observation is ultimately rejected and the posterior distribution of θ^* tends to its prior distribution, and will thereby say when, as $x \to \infty$ the posterior distribution of θ tends to its prior distribution g. And using duality, we can also say when the prior distribution is rejected and the posterior distribution of θ becomes centred around the observation x. However, this theory is of limited use in practice. One

reason is that in order for this kind of rejection to occur, we need either f^* or the corresponding prior distribution g^* to be heavy-tailed, but heavy-tailed distributions on the logarithm convert to extremely heavy-tailed distributions in the original formulation. For instance, if θ^* has a t prior distribution, then the prior distribution of θ is a log-t distribution. This distribution is so heavy-tailed that it does not have any moments, no matter how many degrees of the freedom the original t distribution for θ^* had. So one problem is that the theory of Section 2 requires f or g to have the form of distributions that are rarely used, and seem unrealistic, in practical Bayesian modelling. The other problem is that in order for x^* to become large enough for rejection to occur, x needs to be enormous, so rejection only arises when the conflict is very extreme.

We therefore consider the original problem, noting that there is a duality also in this case. Writing $\varphi = x/\theta$, we have $x = \theta\varphi$, where θ and φ have independent distributions. Therefore when x becomes large the conflict requires that either θ or φ (or both) must take a large value relative to their distributions, g and f respectively. We consider the case where the resolution will be to reject the observation, and so ask under what conditions the posterior distribution of θ will remain finite with probability 1 as $x \to \infty$.

This question is answered by Andrade and O'Hagan (2006) when f has regularly varying right tail. If the RV-credence of f is c they have a single condition for g to be lighter-tailed than f.

(F) For some $\delta > 0$, $\int^\infty \theta^{c+\delta-1} g(\theta)\, d\theta < \infty$.

In particular, if g also has regularly varying right tail with RV-credence c', then condition (F) simply requires $c' > c$. Andrade and O'Hagan (2006) prove that then the limiting posterior density of θ is

$$\frac{\theta^{c-1} g(\theta)}{\int_0^\infty \theta^{c-1} g(\theta)\, d\theta} \ . \tag{3}$$

Notice that this is not the prior distribution. Andrade and O'Hagan refer to (3) as a partial rejection of the observation. One way to see why this happens is to look again at the log transformation. Given that $f(y)$ has a tail like y^{-c} then the tail of $f^*(y)$ is asymptotically $\exp(-(c-1)y)$, so in the log scale we have a location parameter model with a tail like that of an exponential (or double exponential) distribution. This distribution is not sufficiently heavy-tailed to lead to rejection of the observation, but instead has bounded influence. That bounded influence is seen in the additional term θ^{c-1} in (3).

Another way to look at this result is to generalise the idea of rejection. This is most readily appreciated in the dual framework where it is the prior information that is rejected. In the location parameter models, rejection of the prior means a limiting posterior that is the same as would have been obtained from an improper uniform prior. A uniform prior is, however, not usually the preferred representation of weak prior information about a scale parameter; a density proportional to θ^{-1} is more often chosen for its invariance properties.

In the dual form of (3), the limiting posterior corresponds to an improper prior with density proportional to θ^{c-1}.

Andrade and O'Hagan (2006) also consider the case of multiple observations, and prove that when there are two or more outlying groups the group with largest total RV-credence dominates and all the other information sources are 'partially rejected'.

4.2 Location and scale parameters

We now consider the case of a model with both location and scale parameters. Suppose that x has density $\theta^{-1}f((x-\mu)/\theta)$, so that the location parameter is μ and the scale parameter is θ. What might be appropriate resolutions of the conflict between observation and prior distribution as $x \to \infty$? If the prior information on θ is strong and light-tailed, then we may expect the posterior for μ to behave as in the discussion of location parameters. Conversely, if the prior information on μ is strong and light-tailed we should expect the posterior distribution of θ to behave as in Section 4.1. It is nevertheless not obvious how the posterior distribution of the other parameter would behave (for instance if the observation is rejected in the posterior for μ will it also be rejected for the posterior of θ?), and when both parameters have heavy-tailed prior distributions, it is less clear what resolutions are available.

O'Hagan and Andrade (2011) deal with this model, but only for a specific form of prior distribution. They assume that the posterior distribution of μ given θ has the form $\theta^{-1}g(\mu/\theta)$, while the prior density for θ is $h(\theta)$. This prior structure imposes some association between μ and θ, effectively making θ a scale parameter for the prior distribution of μ as well as for the observation. It reflects the conjugate prior structure for normal observations with unknown location and scale, but it is nevertheless a restrictive form that will not always be reasonable in practice.

With this model, they assume that all three densities have regularly varying right tails, such that the RV-credences of f, g and h are c, c' and c'', respectively. They prove that as $x \to \infty$ the following three forms of resolution are possible.

1. If $c < \min(c', c'')$, then subject to additional regularity conditions on f the observation is 'partially rejected' and the posterior joint distribution of μ and θ is in the limit proportional to $\theta^{c-1}g(\mu/\theta)h(\theta)$.

2. If $c' < \min(c, c'')$, then subject to additional regularity conditions on g the prior information on μ is 'partially rejected' and the posterior joint distribution of μ and θ is in the limit proportional to $\theta^{c'-1}f((x-\mu)/\theta)h(\theta)$.

3. If $c'' < \min(c, c')$, then the prior information on θ is 'partially rejected' and the posterior joint distribution of μ and θ is in the limit proportional to $\theta^{-c''}f((x-\mu)/\theta)g(\mu/\theta)$.

The case of multiple observations is more complex. O'Hagan and Andrade (2011) consider only a single outlying observation, for which they can apply the

above results by absorbing the remaining observations into the prior information. With more than one outlying observation there is information about θ arising from differences between the outliers. It may be possible for the outliers to be rejected in terms of the information they provide about μ but for them still to provide information about θ. It may even be possible for the outliers to be rejected in terms of information about μ but for the prior information about θ to be rejected. However, these cases have not been explored and no theoretical results are available. Nor is there any literature on other prior structures.

One more article is worthy of mention here. Haro-López and Smith (1999) consider a p-dimensional vector observation $\mathbf{x} = (x_1, \ldots, x_p)$ with a joint density in the class of v-spherical distributions introduced by Fernández, Osiewalski and Steel (1995),

$$\theta^{-p} f(v(\mathbf{x} - \mu)/\theta) \ ,$$

where v is a scalar function with the property that $v(k\mathbf{a}) = kv(\mathbf{a})$ for any $k \geq 0$. If we integrate the scale parameter θ out of this model, we obtain a location-parameter model that generalises the work of Hill (1975), because the function v does not have to be a metric. Such distributions still have the same tail thickness in all directions, and so will also involve rejection of the whole observation $\mathbf{x}$ or none. However, Haro-López and Smith (1999) also provide results for both location and scale, particularly concerning when the observation $\mathbf{x}$ has bounded influence on the posterior expectation of a general function $m(\mu, \theta)$.

5 Other kinds of parameter

We have so far considered only models with location and/or scale parameters, but many important statistical models do not fit these restrictive forms. Perhaps the simplest such models are where the observations have distributions from the general exponential family. With the exception of the normal and the log-gamma distributions, exponential family distributions are not location-scale. Suppose that the observation x follows an exponential family distribution with canonical parameter θ with an arbitrary prior distribution for θ, and consider the posterior expectation of a function $m(\theta)$. Extending a result of Meeden and Isaacson (1977), Pericchi, Sansó and Smith (1993) show that if $m(\theta)$ is bounded for large θ by a power of θ then, subject to some regularity conditions, the posterior expectation of $m(\theta)$ tends to $m(\tilde{\theta})$ as $x \to \infty$, where $\tilde{\theta}$ is the posterior mode. They also show that $\tilde{\theta}$ may be found by solving a simple equation. It is now possible to explore cases under which the observation (or the prior distribution) is asymptotically rejected by studying the behaviour of $\tilde{\theta}$.

As an example, Pericchi et al (1993) consider a Poisson likelihood, $f(x|\theta) \propto \exp(\theta x - e^{\theta})$. They establish that: (i) if the prior is normal, the posterior mean of the mean parameter e^{θ} diverges from the observation, and thus the prior has unbounded influence; (ii) if the prior is a t distribution the posterior mean approaches x, and so the prior is discarded; (iii) if the prior is logistic with σ^2 then $E[e^{\theta}|x]$ behaves like $x - \frac{\pi}{\sigma\sqrt{3}}$, reflecting a situation of bounded influence.

6 Discussion

The substantial literature reviewed here is at the heart of a "Theory of Conflict Resolution," and one of our objectives in this work has been to stimulate the continuing development of such a theory. There are many different results showing how, when heavy-tailed distributions are used to represent different sources of information in a Bayesian model, then these sources of information can be wholly or partially rejected in the limit as conflicts between information sources become larger. Such results allow the modeller to achieve a kind of 'built-in robustness'. But we have also seen, as O'Hagan (1988) asserted; that it is not enough to simply employ arbitrary heavy-tailed distributions, such as the Cauchy distribution or a t distribution with 2 degrees of freedom. Instead the modeller needs to think about what conflicts may arise and how they should be resolved, and then should use the theory to apply heavy-tailed distributions to the various information sources with tail weights that achieve the required behaviour.

This theory is one essential component in the specification of the prior and the likelihood. When eliciting such distributions, the practitioner's substantive knowledge and past experience, together with careful introspection, will typically allow the centre and spread of the distribution to be determined rather well, but it is much more difficult to elicit meaningful beliefs about tails. In this context, the theory of conflict resolution is a powerful tool for determining the relative weights of tails. It is quite natural to ask the practitioner, for example, "What if the next observation is in conflict with prior expectations, would you believe the data (the prior is wrong) or the prior (the data is an outlier) or both (I would not decide yet, but will wait until more information is gathered)?" The three different answers for the "What if" question immediately settle the question of tail characteristics of likelihood and prior. Heavy-tailed distributions can be used precisely to achieve whatever resolutions of conflicts are judged to be appropriate. For instance, if it is felt that when observations conflict with the prior information the prior should be rejected, then this can be achieved by a suitably heavy-tailed prior distribution. Equally, if it is felt that extreme data should be discounted as outliers, then this can be achieved with appropriate heavy-tailed distributions for the data. And if the judgement is that a small number of observations conflicting with the prior might be discounted but that a larger number should lead to rejection of the prior then this, too, can be 'built-in' through careful choices of tail weights.

The existing theory gives us clear guidance on how to model prior and likelihood tails in order to obtain the desired behaviour for a "what if" question as simple as the one above. For more complex and realistic models, the Theory of Conflict Resolution in Bayesian Statistics is still lagging behind. The theoretical results in the literature are almost exclusively confined to models far simpler than those that are routinely used in practice, because in the last two decades Bayesian Statistics has undergone a spectacular development in computational capabilities. Nowadays, highly complex models are the standard, since complexity is not overly expensive. Comparatively, a much smaller effort has been

devoted to understanding conflict resolution.

The progression in increasingly complex models reviewed here already makes it clear just how little we know about how those more complex models will behave. The results in simple models do not carry over automatically. There is a pressing need for more research, both to fill the gaps in what has been done so far and to extend to ever more complex models. Specifically, we believe that the following gaps in the existing literature need to be filled in order to build understanding and confidence in the use of heavy-tailed models in Bayesian analysis.

- *Hierarchical models.* With three or more layers of hierarchy, and even with known variances throughout, we do not know how the posterior behaves when all layers have heavy-tailed models.

- *Unknown variances.* In models with a single location parameter and a single scale parameter, we do not know how the posterior behaves when there are multiple outlying observations. We do not even know what happens with a single observation when the joint prior distribution does not fit the structure assumed by O'Hagan and Andrade (2011), for instance when the parameters have independent priors. There is no theory at all for models with two or more scale parameters, such as routinely arise in hierarchical modelling.

- *More general models.* An important class of models in which heavy-tailed models have been used but for which there is almost no theory is the linear and generalised linear models. Research is needed here and in time-series models. Although there has been some work on observations with exponential family distributions, the whole area of models that do not fit the structure of location and/or scale parameters is more or less unexplored.

- *Other questions.* The more abstract question of what constitutes a source of information demands careful study. Observations with independent t distributions behave like separate information sources but multivariate t distributions behave like a single source. It is not just a matter of independence, because in the bivariate priors of (2) the parameters are generally not independent and yet it seems that each of the three components represents a separate source. which can be separately rejected.

One motivation for the authors in preparing this review was to stimulate research to address some of these questions. Computational power may come to our aid in this endeavour. On the one hand, it means that we are modelling complex systems to make powerful and subtle predictions without a fundamental insight into how the different components of modelling interact, at least in extreme circumstances. On the other hand, however, computation may be used to educate the intuition in extreme circumstances of conflict, to at least enable us to conjecture likely behaviour. Indeed, several such conjectures have already been identified in preceding sections. Research into the Theory of Conflict

Resolution may proceed on two fronts, perhaps being led by insights revealed through computation, but ultimately it is vital that intuition is followed up by formal proof.

References

Andrade, J.A.A. and O'Hagan, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis* **1**, 169–188. Available electronically at: http://ba.stat.cmu.edu/vol01is01.php.

Angers, J-F. (2000). P-credence and outliers. *Metron* **58**, 81–108.

Angers, J-F. and Berger, J.O. (1991). Robust hierarchical Bayes estimation of exchangeable means. *Canadian Journal of Statistics* **19**, 39–56.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd ed. Wiley: Chichester.

Carlin, B.P. and Polson, N.G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* **19**, 399–405.

Choy, S.T.B. and Smith, A.F.M. (1997). On robust analysis of a normal location parameter. *Journal of the Royal Statistical Society B* **59**, 463–474.

Choy, S.T.B. and Walker, S.G. (2003). The extended exponential power distribution and Bayesian robustness. *Statistics and Probability Letters* **65**, 227–232.

Dawid, A.P. (193). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

Desgagné, A. and Angers, J-F. (2007) Conflicting information and location parameter inference, *Metron* **65**, 67–97.

Fan, T.H. and Berger, J.O. (1992). Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. *Statistics & Decisions* **10**, 99–120.

Fernández, C., Osiewalski, J. and Steel, M.F.J. (1995). Modelling and inference with v-spherical distributions. *Journal of The American Statistical Association* **90**, 1331–1340.

de Finetti, B. (1961). The Bayesian approach to the rejection of outliers. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I pp. 199-210. Univ. California Press: Berkeley, California.

Fúquene, J.A., Cook. J.D. and Pericchi, L.R. (2009). A case for robust Bayesian priors with applications to clinical trials. *Bayesian Analysis* **4**, 817–846. Available electronically at: http://ba.stat.cmu.edu/vol04is04.php.

Goldstein, M. (1983). Outlier resistant distributions: Where does the probability go? *Journal of the Royal Statistical Society B* **45**, 355–357.

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

Hill, B.M. (1975). On coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage* (S.E. Fienberg and A. Zellner, eds.), 555–584. North-Holland.

Le, H. and O'Hagan, A. (1998). A class of bivariate heavy-tailed distributions. *Sankhya* **B 60**, 82–100.

Lindley, D.V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society B* **30**, 31–66.

Meeden, G. and Isaacson, D. (1977). Approximate behavior of the posterior distribution for a large observation. *Annals of Statistics* **5**, 899–908.

Meinhold, R.J. and Singpurwalla, N.D. (1989). Robustification of Kalman filter models. *Journal of The American Statistical Association* **84**, 479–486.

Mitchell, A.F.S. (1994). A note on posterior moments for a normal mean with double-exponential prior. *Journal of the Royal Statistical Society B* **56**, 605–610.

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

O'Hagan, A. (1981). A moment of indecision. *Biometrika* **68**, 329–330.

O'Hagan, A., (1988). Modelling with heavy tails. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.), 345–359. Oxford: Clarendon Press.

O'Hagan, A. (1990). Outliers and credence for location parameter inference. *Journal of the American Statistical Association* **85**, 172–176.

O'Hagan, A. and Andrade, J.A.A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics* (in press).

O'Hagan, A. and Le, H. (1994). Conflicting information and a class of bivariate heavy-tailed distributions. In *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P.R. Freeman and A.F.M. Smith, eds.), 311–327. New York: Wiley.

Pericchi, L.R. and Sansó, B. (1995). A note on Bounded Influence in Bayesian Analysis. *Biometrika* **82**, 223–225

Pericchi, L.R. and Smith, A.F.M. (1992). Exact and Approximate Posterior Moments for a Normal Location Parameter. *Journal of the Royal Statistical Society B* **54**, 793–804.

Pericchi, L.R., Sansó, B. and Smith, A.F.M. (1993). Posterior cumulant relationships in Bayesian inference involving the exponential family. *Journal of the American Statistical Association* **88**, 1419–1426.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. and Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics* **43**, 201–221.

West, M. (1981). Robust sequential approximate Bayesian estimation. *Journal of the Royal Statistical Society B* **43**, 157–166.

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