Infinite hierarchies and prior distributions

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(December, 2000.)

Abstract. This paper introduces a way of constructing non-informative priors for Bayesian analysis, by taking a limit of priors arising from hierarchical constructions as the number of levels in the hierarchy converges to infinity. Results are proved showing that for location families, and other related cases, limits are often not dependent on the exact form of the increment distribution used.

KEYWORDS: hierarchial priors, non-informative priors.

1. Introduction.

Suppose that we had independent data from an $\mathbf{Exp}(\theta_0^{-1})$ distribution. In a Bayesian framework, we suppose that apriori $\theta_0 \sim \mathbf{Exp}(\theta_1^{-1})$, and that with uncertainty on the hyper-parameter θ_1, we might give it also a prior, $\mathbf{Exp}(\theta_2^{-1})$ say. In fact at each level of the hierarchy we can hedge our bets by imposing a further level of prior uncertainty. Suppose we impose N levels of the hierarchy by fixing the hyper-parameter θ_N and sequentially setting

$$\theta_i \sim \mathbf{Exp}(\theta_{i+1}^{-1})$$

* (Corresponding author) Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England. Email: g.o.roberts@lancaster.ac.uk.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Supported in part by NSERC of Canada.

$i = N-1, N-2, \ldots, 1, 0$. In terms of the data, the only thing that matters is the marginal prior of θ_0, obtained (if it were possible) by integrating out the hierarchical parameters.

In such a situation, it is natural to consider the prior distribution of θ_0 as $N \to \infty$. In this case and many others, no proper distributional limit exists, but the limit can sometimes still be described in terms of an improper prior distribution. This paper constructs such improper prior distributions, as limits of marginal priors produced from hierarchical structures of certain types. Of particular interest is the fact (Proposition 2 and the results of Section 4) that in some cases at least, the improper prior distribution produced by this limiting operation is invariant of the distributions imposed in the construction of the hierarchy, thus supporting the use of such prior distributions as canonical non-informative priors.

Let λ be an unknown parameter to be investigated as a result of a statistical experiment. In a Bayesian approach to inference about λ, a prior $G(\cdot)$ is imposed on λ to reflect a priori beliefs about the parameter. However where there is little or no prior information available about λ, the choice of G can be problematic. A common practical solution is to use a multi-level prior structure, the simplest of which sets $\lambda \sim G_1(\cdot|\lambda_1)$ with $\lambda_1 \sim G_2(\cdot)$. This has the effect of flattening the tails of the marginal prior distribution of λ, and thus of creating a less informative prior. Even more non-informative priors can be obtained by using higher order multi-level models.

In this paper, we shall look at infinite homogeneous hierarchies. We shall mostly consider the following hierarchical setup. Suppose $\lambda_i \sim G(\lambda_{i+1}, \cdot)$ for $i = 0, 1, 2, \ldots, N$. We shall investigate the effect of letting $N \to \infty$, on the distribution of $\lambda_0 (= \lambda)$, and the use of the corresponding marginal distribution as a prior distribution. Our approach is very different to that of general hierarchical modelling, since we do not attempt to model the prior structure by the hierarchy constructed. Instead, we are interested in the effect that hierarchical structure has on the marginal distribution of λ. Our approach could be considered as a way of dealing with priors where the hierarchical prior structure is not known, or latent, or simply as a device for producing natural non-informative priors for a particular distribution.

The limiting distributions obtained from letting $N \to \infty$ will often be improper; however, we can still study them in the sense of computing their limiting ratio of probabilities, and considering L^1 convergence in a suitable sense in the presence of observed data. In this context, the notion of non-informativity will be expressed by homogeneity of the hierarchical model, that is the distribution G has to remain invariant to the changes in the level of the hierarchy. Crucial to our argument will be that, at least in some cases, the form of the limiting prior distribution can be considered to be independent to the specific form of G.

Hierarchical structures with a large number (N say) of hierarchical levels describe a marginal prior for λ which is very heavy tailed, and moreover dependence between λ and λ_N is very small. In fact it is frequently observed in practice that posterior distributions of parameters of interest are often rather robust to the specification of high level hyperparameters. Some related issues were considered by DeGroot and Goel (1981), and Gustafson (1996). In particular, they proved that the influence λ_N on the distribution of λ_0 will always be a non-increasing function of N. However, they do not provide any specific information about the distribution of λ_0 in this context.

The construction of non-informative priors for use in Bayesian modelling is an important and difficult one (see for example the discussions in Bernardo and Smith, 1994). The approach adopted in this paper seems very different to existing techniques (for example reference analysis), though as we shall see, leads to the same natural prior choice as (for example) Jeffreys prior in a number of cases.

Multi-level priors are also use to construct priors appropriate for structured situations such as exchangeable parameters. Though such structured priors are not the primary focus of this paper, in Example 9, we shall give an example illustrating how our approach adapts naturally to the construction of priors in that context also.

The computation of such distributions involve interesting classical probability, including large deviations, stable laws, and Markov chain theory. Furthermore, such distributions are independent of the experiment to be performed, and may also provide a possible prior distribution for a given model. In addition, the procedure can be carried out totally independently for all parameters in the model, thus avoiding any problems of consistency between different hierarchical structures.

We shall study the question of the limiting behaviour of the marginal prior, for different choices of the parametric family of distributions $G(\lambda, \cdot)$. Formal definitions and motivation are given in Section 2. We then show (Section 3) using martingale theory that for a certain general class of distributions, the resulting distribution on λ_0 is *flat*, corresponding to Lebesgue measure on $\mathbf{R}$. For a scale family of distributions (Section 4), the resulting distribution has density proportional to $1/x$.

We also show (Section 5) that for a more general class of distributions, for the location family problem, the resulting distribution is related to the derivative of a *large deviations rate function*. For a different class of distributions, we show (Section 6) using the theory of stable laws that the resulting limit is again flat, regardless of the drift of the distribution. For another class of distributions, we show (Section 7) using ergodic Markov chain theory that the resulting distribution is related to the stationary distribution of a resulting Markov chain.

We also phrase (Section 8) our results in terms of the weak convergence of measures. This allows us to interpret our results in terms of standard Bayesian operations such as posterior expectations of functionals. This is followed (Section 9) by a discussion of how the prior distributions resulting from these infinite hierarchies are related to standard choices of prior distributions, for example Jeffreys non-informative prior (Jeffreys, 1946). Some concluding comments are offered in Section 10.

Throughout, our approach will be mathematically fairly general, though not to the extent of allowing mathematical complexity to obscure the statistical relevance. In particular, we will often assume the existence of densities with respect to Lebesgue measure where clearly more general results are possible. In addition, the work in this paper raises many interesting questions about what happens in more complicated and structured hierarchical situations.

2. Definitions and motivation.

We consider an infinite hierarchical model as follows. We let $G(\lambda, \cdot)$ be some fixed probability distribution family on $\mathbf{R}$, taking a parameter $\lambda \in \mathbf{R}$. We shall write the probability of a set A, under the distribution $G(\lambda, \cdot)$, as $G(\lambda, A)$. To avoid problems of periodicity, etc., we shall sometimes assume (without claiming any total generality in our approach) that $G(\lambda, \cdot)$ has a density with respect to Lebesgue measure on $\mathbf{R}$, i.e.

$$G(\lambda, dx) \;=\; g(\lambda, x)\, dx\,, \qquad x \in \mathbf{R}\,. \tag{1}$$

We define our model as follows. For $N \in \mathbf{N}$, we set $\lambda_N^N = a_0$, where a_0 is a fixed constant, and then iteratively let

$$[\lambda_i^N \,|\, \lambda_{i+1}^N, \lambda_{i+2}^N, \ldots, \lambda_N^N] \;\sim\; G(\lambda_{i+1}^N, \cdot)\,, \qquad i = N-1, N-2, \ldots, 1, 0\,. \tag{2}$$

We are interested in the limiting (possibly improper) distribution of λ_0^N, as $N \to \infty$. Specifically, given $G(\lambda, \cdot)$ and a_0, we are interested in the limit

$$R(A, B) \;=\; \lim_{N\to\infty} \frac{\mathbf{P}(\lambda_0^N \in A)}{\mathbf{P}(\lambda_0^N \in B)}\,, \qquad A, B \subseteq \mathbf{R}\,. \tag{3}$$

We shall also consider a density version of (3), by writing

$$\begin{aligned} r(x, y) \;&\equiv\; \lim_{\delta\searrow 0} R((x-\delta, x+\delta), (y-\delta, y+\delta)) \\ &= \lim_{\delta\searrow 0} \lim_{N\to\infty} \frac{\mathbf{P}(x-\delta < \lambda_0^N < x+\delta)}{\mathbf{P}(y-\delta < \lambda_0^N < y+\delta)}\,, \qquad x, y \in \mathbf{R}\,, \end{aligned} \tag{4}$$

whenever the limits exist. Thus $r(x, y)$ represents (essentially) the density ratio for the limiting prior distribution. Of course, there is some redundancy in the double index in both (3) and (4), since for example $r(x, y) = r(x, z)\, r(z, y)$ assuming all these limits exist and are finite.

We note that, if $G(\lambda, \cdot)$ defines a null-recurrent transition probability kernel, then it is well known that G has a unique sub-invariant measure, which is necessarily also invariant (see e.g. Meyn and Tweedie, 1993, Proposition 10.4.2). It follows easily that if $R(A, B)$ exists, then it is necessarily equal to $\pi(A)/\pi(B)$. (We note, however, that the limit $R(A, B)$

may still not exist in this case; see e.g. the counter-example in Chung, 1974. However, the limiting ratio of the *average* of the λ_i^N must still exist and equal $\pi(A)/\pi(B)$, see e.g. Meyn and Tweedie, 1993, Theorem 17.3.2.) This observation allows us to identify the possible limiting prior distribution without having to consider any detailed limiting arguments.

Under some circumstances, we can compute $r(x,y)$ directly as the limit of ratios of densities of the λ_0^N. For example, letting $\mathcal{L}(\cdot)$ denote the law of a random variable, suppose that $\mathcal{L}(\lambda_0^N)$ has a density $f_N(\cdot)$ with respect to some sigma-finite measure ν. Then we have the following.

Proposition 1. *Let $s_N(x,y) = f_N(x)/f_N(y)$, and suppose that $s(x,y) = \lim_{N\to\infty} s_N(x,y)$ exists for each x and y. Suppose further that*

$$\lim_{\delta\searrow 0}\lim_{N\to\infty}\sup_{\substack{x\in\mathbf{R}\\|x-y|<\delta}} s_N(x,y) \;=\; 1\,, \qquad y\in\mathbf{R}\,. \tag{5}$$

(which follows, for example, if the convergence of $s_N(x,y)$ to $s(x,y)$ is uniform over x and y in compact sets, and furthermore each f_N is continuous). Then $r(x,y)$ exists for all x and y, and in fact $r(x,y) = s(x,y)$.

Proof. We have that

$$\frac{\mathbf{P}(x-\delta<\lambda_0^N<x+\delta)}{\mathbf{P}(y-\delta<\lambda_0^N<y+\delta)} \;=\; \frac{f_N(x)}{f_N(y)}\;\frac{\int_{x-\delta}^{x+\delta} s_N(u,x)\nu(du)}{\int_{y-\delta}^{y+\delta} s_N(v,y)\nu(dv)}\,.$$

Now, by (5), as $N\to\infty$ and then $\delta\searrow 0$, each of the above integrals is asymptotic to 2δ. Hence, the ratio of integrals goes to 1. The result follows. ■

3. The martingale location family case.

Recall that $G = \{G(\lambda,\cdot)\}$ is a *location family* of probability measures on a vector space (e.g. $\mathbf{R}$) if the measures satisfy the relation

$$G(\lambda, A) = G_0(A-\lambda)\,, \qquad A\subseteq\mathbf{R}$$

for some probability measure G_0. (That is, $G(\lambda,A) = G_0(\{x-\lambda;\ x\in A\})$.) Where G_0 has a density g_0 with respect to Lebesgue measure, G gives rise to a family of densities

$g = \{g(\lambda; \cdot)\}$ such that $g(\lambda; x) = g_0(x - \lambda)$. We shall say that $G(\lambda, \cdot)$ is a *martingale location family* if G_0 satisfies $\int t\, G_0(dt) = 0$.

By (2), this implies that the sequence $\lambda_N^N, \lambda_{N-1}^N, \lambda_{N-2}^N, \ldots$ is in fact an additive martingale. In this case, if $\{\lambda_i^N\}$ are defined by (2), then $\lambda_i^N = \lambda_{i-1}^N + U_i$, where U_i are i.i.d. random variables with $\mathbf{E}(U_i) = 0$. We have

Proposition 2. *Suppose that $G(\lambda, \cdot)$ is additive martingale, with $G(\lambda, dx) = g_0(x-\lambda)dx$ where g_0 is bounded, and with $G(\lambda, \cdot)$ having finite positive variance v. Then for any $a_0 \in \mathbf{R}$, the limiting density of λ_0^N is flat. That is, $R(A, B) = leb(A)/leb(B)$ for whenever $leb(B) > 0$ (where leb is Lebesgue measure), and furthermore $r(x, y)$ exists and equals 1 for all $x, y \in \mathbf{R}$.*

Proof. Since g_0 is a bounded density, it is in L^2, hence so is its characteristic function. Let $U_i = \lambda_i^N - \lambda_{i-1}^N$ as above. Then, by the density central limit theorem (see Feller, 1971, p. 516), the density of $\frac{1}{\sqrt{Nv}}(U_1 + \ldots + U_N)$ converges pointwise (and uniformly) to the standard normal density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, hence so does the density of $\frac{1}{\sqrt{Nv}} \lambda_0^N = \frac{1}{\sqrt{Nv}}(a_0 + U_1 + \ldots + U_N)$. By the change-of-variable theorem (see e.g. Kelly, 1994, p. 326; Billingsley, 1995, p. 217),

$$\frac{\sqrt{Nv} f_N(x)}{\phi(x/\sqrt{Nv})} \to 1$$

pointwise and uniformly as $N \to \infty$, where f_N is the density of $\mathcal{L}(\lambda_0^N)$. But as $N \to \infty$, clearly $\phi(x/\sqrt{Nv}) \;/\; \phi(y/\sqrt{Nv}) \to 1$ for any fixed x and y. It follows that for any two bounded sets A and B with $leb(B) > 0$, we have

$$\frac{\mathbf{P}(\lambda_0^N \in A)}{\mathbf{P}(\lambda_0^N \in B)} \to \frac{leb(A)}{leb(B)},$$

as $N \to \infty$. Hence, $R(A, B) = leb(A) \;/\; leb(B)$. The result $r(x, y) = 1$ now follows from the definition (4). ∎

Example 3. CENTERED NORMAL: Here $G(\lambda, \cdot) = N(\lambda, v)$ for fixed $v > 0$. This is clearly additive martingale, so by Proposition 2 we have $\mathcal{L}(\lambda_0^N)$ asymptotically flat, i.e. $r(x, y) = 1$ for all x and y.

Example 4. CENTERED UNIFORM: Here $G(\lambda, \cdot) = \mathbf{Unif}[\lambda - A,\ \lambda + A]$, where $A > 0$ is fixed. This is again additive martingale, so that Proposition 2 applies and we again conclude that $\mathcal{L}(\lambda_0^N)$ is asymptotically flat.

Remark. As will be discussed in Section 9 below, the asymptotically flat priors for the martingale location family case coincide with the corresponding Jeffreys prior (Jeffreys, 1946).

We note that similar reasoning applies in the multi-dimensional case, and we conclude the following.

Proposition 5. *Suppose that $G(\lambda, \cdot)$ is additive martingale in $\mathbf{R}^d$, with $G(\lambda, d\mathbf{x}) = g_0(\mathbf{x} - \lambda)d\mathbf{x}$ where g_0 is bounded and has finite positive variance v. Then for any $\mathbf{a}_0 \in \mathbf{R}^d$, the limiting density of λ_0^N is again flat.*

4. Transformations and the martingale scale family case.

Here we extend the results of Section 3 to a related scale family of distributions. To begin with, assume that G is a general transition kernel. Let $f : \mathbf{R} \to \mathbf{R}$ be a smooth monotone function, and suppose that $G_f(\lambda, A) \equiv G(f^{-1}(\lambda), f^{-1}(A))$. That is, G_f is the transition kernel obtained by applying the function f:

$$\mathbf{P}_{G_f}(\lambda_i^N \in A \,|\, \lambda_{i+1}^N = x) \ = \ \mathbf{P}_G(f^{-1}(\lambda_i^N) \in A \,|\, \lambda_{i+1}^N = f^{-1}(x))\,. \qquad (6)$$

Then we have the following.

Proposition 6. *Let G be a general transition kernel, and let G_f be as defined above. Then the limiting density of $\mathcal{L}(\lambda_0^N)$ is proportional to $\pi(f^{-1}(y))/|f'(f^{-1}(y))|$, where $\pi(\cdot)$ is the limiting density obtained by iterating the transition described by G.*

Proof. $1/|f'(f^{-1}(y))|$ just represents the Jacobian of the transformation defined by f. The result follows by applying the transformation to the sequence of limiting probabilities and again using the change-of-variable formula. ∎

In particular, say that G represents a *log-martingale scale family* if all the λ_i^N are positive, and if G_f represents a martingale location family as defined above when $f(x) = \log x$. Applying Proposition 6 with this f function, we obtain the following.

Corollary 7. Suppose that $G(\lambda, \cdot)$ is a log-martingale scale family. Then for any $a_0 > 0$, density of λ_0^N is proportional to $1/x$. That is, $r(x, y) = y/x$ for all $x, y \in \mathbf{R}$.

Proof. This follows from Proposition 6, since the Jacobian of the logarithm function is $1/x$. ∎

Remark. As will be discussed in Section 9 below, the $1/x$ priors for the log-martingale scale family case coincide with the corresponding Jeffreys prior (Jeffreys, 1946).

Example 8. EXPONENTIAL: Let $G(\lambda, \cdot) = \mathbf{Exp}(\lambda)$ be the exponential family (with density $\lambda e^{-\lambda x}$ for $x > 0$). Let $E_1, E_2, \ldots$ be i.i.d. $\sim \mathbf{Exp}(1)$. Then we may write $\lambda_i^N = E_{i+1}/\lambda_{i+1}^N$. Iterating this, we have that $\lambda_i^N = (E_{i+1}/E_{i+2})\lambda_{i+2}^N$. But (E_{i+1}/E_{i+2}) is a non-negative random variable with $\mathbf{E}\left(\log(E_{i+1}/E_{i+2})\right) = 0$. Hence, if we take N even and consider $\{\lambda_{2i}^N\}$ in place of $\{\lambda_i^N\}$, we see from Corollary 7 that the asymptotic density of λ_0^N will be proportional to $1/x$. It is straightforward to verify (by considering λ_{N-1}^N in place of a_0 and integrating with respect to $\mathcal{L}(\lambda_{N-1}^N)$) that this conclusion remains true if we allow N to take on both odd and even values.

Remark. Our approach produces priors which are independent of the chosen parameterisation, in the following sense. Suppose instead of putting a prior on λ, we attempted to put a prior on $f(\lambda)$. If we then used the induced kernel G_f as above, then the resulting prior on $f(\lambda)$ would coincide with the prior distribution of $f(\lambda)$ using the original prior on λ. For example, if λ is a scale parameter, and $f(x) = \log x$, then $\log \lambda$ is a location parameter, and G_f is a location kernel. Moreover, the prior on $\log \lambda$ is the same as the distribution of $\log \lambda$ using the original prior on λ.

Example 9. AN EXCHANGEABLE MODEL: The examples we have considered so far are based on simple linear hierarchical structures. Many of the ideas we have introduced can be translated to a more general setting. Here we just consider a simple example for a non-informative prior for exchangeable random variables. Although the theoretical results do not apply strictly to this context, this example serves to illustrate how the ideas here can be extended to more complicated hierarchical structures.

By de Finetti's theorem, exchangeable random variables $\lambda_1, \ldots \lambda_k$ can be written as conditionally independent and identically distributed random variables, conditional on θ say. Then suppose we assume that

$$\lambda_i \sim N(\lambda_0, s_0)\,, \qquad i = 1, 2, \ldots, k\,,$$

where λ_0 has the non-informative centered martingale location family flat prior, and s_0 has the martingale scale family prior. We can therefore write

$$p(\lambda_0, s_0) \;\propto\; \frac{1}{s_0}, \quad \lambda_0 \in \mathbf{R},\; s_0 \in \mathbf{R}^+\,,$$

and

$$\begin{aligned} p(\lambda_0, s_0, \lambda_1, \ldots, \lambda_k) \;&\propto\; \frac{1}{s_0} \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi s_0}} e^{-(\lambda_i - \lambda_0)^2/2s_0} \\ &= \frac{1}{s_0}(2\pi s_0)^{-k/2} \exp\left(-(2s_0)^{-1} \sum_{i=1}^{k} (\lambda_i - \lambda_0)^2\right) \\ &= \frac{1}{s_0}(2\pi s_0)^{-k/2} \exp\left(-(2s_0)^{-1}\left[D + k(\bar{\lambda} - \lambda_0)^2\right]\right) \end{aligned}$$

for $\lambda_i \in \mathbf{R}$ and $s_0 \in \mathbf{R}^+$, where $D = \sum_{i=1}^{k}(\lambda_i - \bar{\lambda})^2$ and $\bar{\lambda} = \frac{1}{k}\sum_{i=1}^{k} \lambda_i$.

Integrating out λ_0, and neglecting multiplicative constants, we obtain that

$$p(s_0, \lambda_1, \ldots, \lambda_k) \;\propto\; (s_0)^{-(k/2)-1} e^{-D/2s_0} (s_0/k) \;\propto\; (s_0)^{-k/2} e^{-D/2s_0}\,.$$

Finally, we obtain the marginal non-informative prior for $\lambda_i,\ 1 \leq i \leq k$ by integrating with respect to s_0:

$$
\begin{aligned}
p(\lambda_1, \ldots, \lambda_k) \ &\propto \ \int_0^\infty (s_0)^{-k/2} e^{-D/2s_0}\, ds_0 \ = \ \int_0^\infty (D/2u)^{-k/2} e^{-u}\, (D\, du/2u^2) \\
&= \ (D/2)^{-(k/2)+1} \int_0^\infty u^{(k/2)-1} e^{-u} du \ = \ (D/2)^{-(k/2)+1} \Gamma(k/2) \\
&\propto \ \frac{1}{\left(\sum_{i=1}^k (\lambda_i - \bar{\lambda})^2 \right)^{\frac{(k-2)}{2}}} \, .
\end{aligned}
$$

Clearly this technique can be used to construct non-informative priors in other situations where some structure can be assumed, for instance for pairwise interaction priors.

5. The non-centered location family case.

If the structure of $\{\lambda_i^N\}$ is not martingale, but still has the additivity property $G(\lambda, dx) = G_0(dx - \lambda)$, then the situation is more complicated than that described in Proposition 2. The following example illustrates the phenomena involved.

Example 10. UNCENTERED NORMAL: Here $G(\lambda, \cdot) = N(\lambda + m, v)$ for fixed $m \in \mathbf{R}$ and $v > 0$. Then setting f_0^N to be the density for λ_n, we have that $f_0^N(x) = \frac{1}{\sqrt{2\pi n v}} e^{-(x-nm)^2/2nv}$, whence $f_0^N(x)/f_0^N(y) = e^{(y^2 - x^2 + 2nm(x-y))/2nv}$. Therefore, by Proposition 1, we have $r(x,y) = \lim_{n\to\infty} f_0^N(x)/f_0^N(y) = e^{m(x-y)/v}$.

It turns out that this example is a special case of a rather more general result which is can be analysed by using a slight modification of the classic large deviations result, Cramér's theorem (see e.g. Deuschel and Stroock, 1989; Dembo and Zeitouni, 1993; Varadhan, 1984).

We proceed initially with a heuristic argument. Suppose that $\{X_i\}_{i=1,2,\ldots}$ is a sequence of i.i.d. random variables. It will be necessary to assume the existence of certain moments; in particular, we assume a finite mean m.

Let $Y_i = X_i - m$ (so that $\mathbf{E}(Y_i) = 0$). Let $I(x)$ be the large deviations rate function for the Y_i, that is

$$
\mathbf{P}[\bar{Y}^{(n)} < -y] \sim \exp\{-I(-y)n\} \ ,
$$

where $\bar{Y}^{(n)} = \sum_{i=1}^{n} Y_i/n$. Then we might expect

$$\mathbf{P}[\sum_{i=1}^{n} Y_i < -mn + x] = \mathbf{P}[\bar{Y} < -m + (x/n)] \sim \exp\{-nI(-m + x/n)\} \sim \exp\{-xI'(-m)\} \; .$$

Of course, this relies on the sub-exponential terms in the approximations above not interfering at all, but it turns out that this argument can be made rigorous rather generally as we shall see. Note that in the uncentered normal case, $I(q) = q^2/2v$, so that $I'(-m) = -m/v$; hence the result holds in that case.

We now proceed to make this argument more precise. Let $Y_i = X_i - m$, let μ be the distribution of Y_i, and let μ_n be the distribution of $\frac{1}{n}(Y_1 + \ldots + Y_n)$. For definiteness, take $\mu(dy) = e^{-\phi(y)}dy$ for some function ϕ. Finally define $L(\lambda) = \log \mathbf{E}(e^{\lambda Y_i})$, and let $L^*(y) = \sup_\lambda (\lambda y - L(\lambda))$.

Let $\lambda(y) = \operatorname{argsup}_\lambda (\lambda y - L(\lambda))$, so that $L(\lambda(y)) = L^*(y)$. Note that $\lambda(\cdot)$ is a continuous function on the interval $(\inf \operatorname{supp} Y_i,\ \sup \operatorname{supp} Y_i) = (\inf \operatorname{supp} X_i - m,\ \sup \operatorname{supp} X_i - m)$; we shall assume that $\inf \operatorname{supp} X_i < 0 < \sup \operatorname{supp} X_i$ so that

$$\lambda(q) < \infty \quad \text{and} \quad L^*(q) < \infty \quad \text{for } q \text{ in a neighbourhood of } -m\,. \tag{7}$$

Classical inequalities (see e.g. Deuschel and Stroock, 1989, pp. 5–6) then say that for $q \in \mathbf{R}$,

$$\mu_n[q, \infty) \;\leq\; e^{-nL^*(q)} \mu_n(\infty, q] \;\leq\; e^{-nL^*(q)} \tag{8}$$

and for any $\delta > 0$,

$$\mu_n(q - \delta, q + \delta) \;\geq\; e^{-n(L^*(q) - \lambda(q)\delta)} \tilde{\mu}_n^q(q - \delta, q + \delta)\,. \tag{9}$$

Here $\tilde{\mu}_n^q$ is the distribution of $\frac{1}{n}(Z_1 + \ldots + Z_n)$ where $\{Z_i\}$ are i.i.d. $\sim \tilde{\mu}^q$, where

$$\tilde{\mu}^q(dy) \;=\; e^{\lambda(q)y - L(\lambda(q))} \mu(dy)\,.$$

Hence (cf. Deuschel and Stroock, 1989, p. 6),

$$\int x \tilde{\mu}^q(dx) \;=\; \frac{d}{dt} L(t)\Big|_{t=\lambda(q)} \;=\; q$$

(since $\frac{d}{dt}(tq - L(t))\Big|_{t=\lambda(q)} = 0$). It then follows by the weak law of large numbers that

$$\tilde{\mu}_n^q(q-\delta, q+\delta) \;\to\; 1\,, \qquad n \to \infty\,.$$

From this and equations (8) and (9), the classical Cramèr's Theorem follows easily:

$$-\inf_{q\in S^\circ} L^*(q) \;\leq\; \liminf_{n\to\infty} \frac{1}{n}\log \mu_n(S) \;\leq\; \limsup_{n\to\infty} \frac{1}{n}\log \mu_n(S) \;\leq\; -\inf_{q\in \overline{S}} L^*(q)$$

In particular, the large deviation rate function for the Y_i is given by $I(q) = L^*(q)$.

For our purposes, these bounds aren't sufficiently sharp. Instead, we compute directly. We wish to compute

$$\mathbf{P}\Big(X_1 + \ldots + X_n \in (x-\delta, x+\delta)\Big) \;\Big/\; \mathbf{P}\Big(X_1 + \ldots + X_n \in (y-\delta, y+\delta)\Big)$$

for fixed $x, y \in \mathbf{R}$, as $n \to \infty$ and $\delta \searrow 0$.

Lemma 11. *Let $\{X_i\}$ be i.i.d. with density $e^{-\phi(\cdot)}$ and finite mean m, and with $\inf\operatorname{supp} X_i < 0 < \sup\operatorname{supp} X_i$. Let Y_i, $L(y)$, and $\lambda(y)$ be as above, and assume that $L(y)$ is finite in a neighbourhood of 0. Then for any $x \in \mathbf{R}$, we have as $\delta \searrow 0$ that*

$$\lim_{n\to\infty} \frac{\mathbf{P}\Big(X_1 + \ldots + X_n \in (x-\delta, x+\delta)\Big)}{2\delta \;/\; \sqrt{2\pi n v_{-m}}} \;=\; e^{mL(-m)) - \lambda(-m)(x + O(\delta))}\,,$$

where v_q is the variance of $\tilde{\mu}^q$.

Proof. We compute that:

$$\mathbf{P}\Big(X_1 + \ldots + X_n \in (x-\delta, x+\delta)\Big)$$

$$= \mathbf{P}\left(Y_1 + \ldots + Y_n \in (x - \delta - nm,\; x + \delta - nm)\right)$$

$$= \mathbf{P}\left(|Y_1 + \ldots + Y_n - (x - nm)| \leq \delta\right)$$

$$= \int_{|z_1+\ldots+z_n-(x-nm)|\leq\delta} \exp\Big[-\sum_{i=1}^n \phi(z_i)\Big] dz_1 \ldots dz_n$$

$$= \int_{|z_1+\ldots+z_n-(x-nm)|\leq\delta} \exp\Big[-\sum_{i=1}^n \left(\phi(z_i) + \lambda(-m)z_i\right)\Big] \times \exp\Big[-\lambda(-m)\sum_{i=1}^n z_i\Big] dz_1 \ldots dz_n$$

$$= e^{-\lambda(-m)(x-nm+O(\delta))} \times \int_{|z_1+\ldots+z_n-(x-nm)|\leq\delta} \exp\Big[-\sum_{i=1}^{n}\phi_{\lambda(-m)}(z_i)\Big]dz_1\ldots dz_n\,.$$

Here we have factored out a term which is almost surely within $O(\delta)$ of $e^{-\lambda(-m)(x-nm)}$. To finish the argument, we apply the central limit theorem to the remaining integral. This integral is with respect to the distribution $\tilde{\mu}^{-m}$. Now, the moment generating function of $\tilde{\mu}^{-m}$ satisfies that $\log\int e^{sx}\tilde{\mu}^{-m}(dx) = \text{const} + L(s+\lambda(-m))$. function exists in a neighbourhood of 0. By (7), $L(s+\lambda(-m)) < \infty$ for s in a neighbourhood of 0. Hence, $\tilde{\mu}^{-m}$ has moment generating function which is finite in a neighbourhood of 0, and therefore has finite variance: $v_{-m} < \infty$. We conclude that as $n \to \infty$,

$$\mathbf{P}\Big(X_1+\ldots+X_n \in (x-\delta, x+\delta)\Big) = \Big(1+o(1)\Big)\frac{2\delta}{n}\sqrt{\frac{n}{2\pi v_{-m}}} \times e^{mL(-m)} \times e^{-\lambda(-m)(x+O(\delta))}\,.$$

This gives the result. ∎

Using this lemma, we are able to give a result with generalises Example 10 to virtually arbitrary i.i.d. sequences with large deviation rate functions and finite moment generating functions.

Theorem 12. *Let $\{X_i\}$ be i.i.d. with density $e^{-\phi(\cdot)}$, finite mean m, and with* $\inf\operatorname{supp} X_i < 0 < \sup\operatorname{supp} X_i$. *Let $Y_i = X_i - m$, and assume that Y_i has moment generating function which exists at least in some neighbourhood of* 0. *Let $I(\cdot) = L^*(\cdot)$ be the large deviations rate function for the Y_i. Then for $x, y \in \mathbf{R}$,*

$$\lim_{\delta\searrow 0}\lim_{n\to\infty}\mathbf{P}\Big(X_1+\ldots+X_n \in (x-\delta,x+\delta)\Big) \Big/ \mathbf{P}\Big(X_1+\ldots+X_n \in (y-\delta,y+\delta)\Big)$$
$$= \exp\{-I'(-m)(x-y)\}\,,$$

assuming the derivative exists.

Proof. From the definition of $L^*(y)$, we have that $L^*(y) = \lambda(y)\,y - L(\lambda(y))$, so that $L^{*\prime}(-m) = \lambda(-m) + (-m)\lambda'(-m) - L'(\lambda(-m))\lambda'(-m) = \lambda(-m) + (-m)\lambda'(-m) - -m\lambda'(-m) = \lambda(-m)$. Hence, $L^{*\prime}(-m) = \lambda(-m)$. Thus, the conclusion of Lemma 11 can be written as

$$\mathbf{P}\Big(X_1+\ldots+X_n \in (x-\delta,x+\delta)\Big) \propto (1+O(\delta))\exp\Big[-xL^{*\prime}(-m) + O(1/n)\Big]$$

The result follows since $L^*(\cdot) = I(\cdot)$. ∎

We can state this result in terms of our λ_0^N variables as follows.

Corollary 13. *Suppose that the $\{\lambda_i^N\}$ are defined with $G(\lambda, \cdot) = G_0(\cdot - \lambda)$, where $G_0(\cdot)$ has positive density and finite mean m, and has $\inf \operatorname{supp} G_0 < 0 < \sup \operatorname{supp} G_0$. Let $I(\cdot) = L^*(\cdot)$ be the large deviation rate function for Y_i, where $Y_i = X_i - m$ with $\{X_i\}$ i.i.d. $\sim G_0$. Assume Y_i has moment generating function which exists at least in some neighbourhood of 0. Then if $I'(-m)$ exists, then $r(x, y)$ exists for all $x, y \in \mathbf{R}$, with*

$$r(x, y) = \exp\{(y - x)I'(-m)\} .$$

Example 14. GAMMA: if $g(x) = \text{Gamma}(a, b; x) \propto x^{a-1}e^{-bx}$ for $x > 0$ (where $a, b > 0$), then the corresponding U_i are all non-negative, so we might expect λ_i to go to infinity linearly. Indeed, we compute that in this case, $\mathcal{L}(\lambda_n) \propto x^{na-1}e^{-bx}dx$, so that $f_0^n(x)/f_0^n(y) = (x/y)^{na-1}e^{-bx}$. As $n \to \infty$, this ratio converges to ∞ whenever $x > y$. In fact, here $R(A, B) = \infty$ whenever $\operatorname{esssup}(A) > \operatorname{esssup}(B)$.

Example 15. SHIFTED GAMMA: if $g(x) \propto (x + c)^{a-1}e^{-b(x+c)}$ for $x > -c$ (where $a, b, c > 0$), i.e. a gamma distribution shifted c units to the left, then the corresponding U_i are no longer non-negative, so the result is more interesting. Indeed, in this case, for $x, y > 0$, we have by Proposition 1 that

$$r(x, y) = \lim_{n\to\infty} \frac{(x + cn)^{na-1}e^{-b(x+cn)}}{(y + cn)^{na-1}e^{-b(y+cn)}}$$

$$= \lim_{n\to\infty} \left(1 + \frac{x - y}{y + cn}\right)^{na-1} e^{-b(x-y)}$$

$$= e^{(x-y)((a/c)-b)} .$$

Now, if $c = a/b$ (the mean of the gamma distribution), then we get a flat limit, corresponding to Proposition 2 again. On the other hand, as $c \searrow 0$ the ratio goes to infinity, corresponding to the previous example.

Finally in this section, we say that $G(\lambda, \cdot)$ is a *non-centered scale family* if $G(\log \lambda, \log \cdot)$ is a non-centered location family as above, i.e. if (6) is satisfied with $f(x) = \log x$ where G_f is a non-centered location family. We then have

Corollary 16. *Suppose that $G(\lambda, \cdot)$ is a non-centered scale family. Then $r(x, y) = (y/x)^{1+I'(m)}$ for all $x, y \in \mathbf{R}$, where m and $I(\cdot)$ are the mean and large-deviations rate function, respectively, corresponding to the non-centered location family $G(\log \lambda, \log \cdot)$.*

Proof. This follows immediately from Proposition 6 and Theorem 12. ■

6. Results using the theory of stable laws.

The previous section relies heavily on the fact that certain variances are finite. If these variances are not finite, then the resulting limits may involve non-Gaussian stable laws. We begin with an example.

Example 17. CAUCHY: Suppose that $G(\lambda, dx) = \frac{1}{\pi(1+(x-a)^2)} dx$ is a Cauchy distribution with drift a. Then the distribution of $\frac{1}{n}(\lambda_n - \lambda_0)$ is again this same Cauchy distribution. From this fact, is it straightforward to conclude that we will have $r(x, y) = 1$ for all x and y, i.e. that the the limiting distribution will again be *flat*.

The above example is part of a more general phenomenon, as we now show. Call a density *log-Lipshitz* if its logarithm is a Lipshitz function, i.e. if there is $\ell < \infty$ with $|\log f(x) - \log f(y)| \leq \ell |x - y|$ for all x and y. (Note that the Gaussian density is *not* log-Lipshitz, but the Cauchy density is.) Then we have

Proposition 18. *Let $G(\lambda, \cdot)$ be additive, where the increment distribution is a stable law with log-Lipshitz stable density (with respect to Lebesgue measure). Then $r(x, y) = 1$ for all x and y, i.e. the distribution of λ_0^N is asymptotically flat.*

Proof. Let α be the parameter of the stable law, say ν_α, having density f_α, and let ℓ be the log-Lipshitz constant. Let $U_i = \lambda_i - \lambda_{i-1}$ be the i^{th} increment, so that $\lambda_N - \lambda_0 = \sum_{i=1}^N U_i$. Then as $N \to \infty$, we have for some $b \in \mathbf{R}$ that

$$\frac{\sum_{i=1}^N U_i - Nb}{N^{1/\alpha}} \ \Rightarrow \ \nu_\alpha \, .$$

Hence, as $N \to \infty$, we have

$$\mathbf{P}\left(\lambda_N \in (x-\delta,\, x+\delta)\right)$$

$$= \mathbf{P}\Big(\frac{\sum_{i=1}^N U_i - Nb}{N^{1/\alpha}} \in (N^{-1/\alpha}(x-\delta+\lambda_0-Nm),\, N^{-1/\alpha}(x+\delta+\lambda_0-Nm)\Big)$$

$$\to \nu_\alpha\left(N^{-1/\alpha}(x-\delta+\lambda_0-Nm),\; N^{-1/\alpha}(x+\delta+\lambda_0-Nm)\right).$$

Hence, using the log-Lipshitz property, this probability is

$$\leq\; 2N^{-1/\alpha}\delta\; f_\alpha\left(N^{-1/\alpha}(x+\lambda_0-Nm)\right)\; e^{\ell 2\delta N^{-1/\alpha}},$$

and is also

$$\geq\; 2N^{-1/\alpha}\delta\; f_\alpha\left(N^{-1/\alpha}(x+\lambda_0-Nm)\right)\; e^{-\ell 2\delta N^{-1/\alpha}},$$

where ℓ is the log-Lipshitz constant. Now, these two expressions are independent of x, and have ratio approaching 1 as $\delta \searrow 0$. The result follows. ∎

Remark. Of course, if the increment distribution is instead in the *domain of attraction* of a stable law with log-Lipshitz density, then the distribution of λ_0^N will still be asymptotically flat, provided that it converges to the corresponding stable law in such a way that its probability ratios also converge.

Proposition 18 leads to the question of which stable laws have log-Lipshitz densities. We make the following conjecture.

Conjecture. All non-Gaussian stable laws have log-Lipshitz densities.

Of course, it is well known that all stable laws have densities of some sort (with respect to Lebesgue measure). The question here is whether or not these densities are necessarily log-Lipshitz. We believe that the conjecture is true, however we are unable to prove it or to locate an appropriate result in the literature on large deviations. For background on stable laws, see e.g. Feller (1971), Zolotarev (1986), and Bingham et al. (1987).

7. Results using ergodic Markov chain theory.

Clearly, we may think of $\lambda_N, \lambda_{N-1}, \ldots$ as a Markov chain, with transition probabilities governed by $G(\lambda, \cdot)$. Recall (cf. Meyn and Tweedie, 1993, p. 312) that the Markov chain is ergodic with stationary distribution $\pi(\cdot)$ if

$$\int \pi(\lambda) G(\lambda, dy) d\lambda \;=\; \pi(dy)\,, \qquad y \in \mathbf{R}$$

and

$$\lim_{N \to \infty} \|\mathcal{L}(\lambda_0^N) - \pi(\cdot)\| \;\to\; 0\,.$$

In this case, we obtain

Proposition 19. *Suppose $G(\lambda, \cdot)$ gives an ergodic Markov chain with stationary distribution $\pi(\cdot)$. Then, for any $a_0 \in \mathbf{R}$, the limiting distribution of λ_0 is proportional to $\pi(\cdot)$. That is, $R(A, B) = \pi(A)/\pi(B)$ whenever $\pi(B) > 0$.*

Proof. By ergodicity, we have as $N \to \infty$ that $\mathbf{P}(\lambda_0^N \in A) \to \pi(A)$ and $\mathbf{P}(\lambda_0^N \in B) \to \pi(B)$. The result follows. ■

This situation does not often arise in statistical inference models. However, there are various examples of this phenomenon.

Example 20. RECIPROCAL POISSON: Here $\log(\lambda_{i+1}) \sim \mathbf{Pois}(1/\lambda_i)$. Here it is easily verified that $G(\lambda, 1) \geq e^{-1}$ for all λ. This is sufficient for positive recurrence (see for example Meyn and Tweedie, 1993). Hence, Proposition 19 applies, and we conclude that the limiting distribution of λ_0^N is proportional to the stationary distribution of this Markov chain.

8. Weak convergence results.

It is worth considering how the value of $r(x,y)$ relates to convergence of expectations of functionals $z(x)$ according to the posterior distribution. Let $L(x)$ denote the likelihood for given parameter value x (which also depends on the data, although we suppress this in the notation). Then we have the following.

Proposition 21. *Let $\lambda_i^n \sim G(\lambda_{i+1}^n, \cdot)$ as usual, with $\mathbf{P}(\lambda_0^n \in dx) = f_0^n(x)\,dx$. Suppose that (5) holds, and further that for some $c \in \mathbf{R}$ (e.g. $c = 0$), and some L^1 function Y, we have $L(x) f_0^n(x)/f_0^n(c) \leq Y(x)$ for all $x \in \mathbf{R}$, for all sufficiently large n. (For example, perhaps $L \in L^1$ and*

$$f_0^n(x) \leq K f_0^n(c) \tag{10}$$

for all x; we can then take $Y(x) = KL(x)$.) Let $z(x)$ be any bounded functional. Write $\mathbf{E}\left(z(\lambda_n)\right)$ for the expectation with respect to the posterior distribution, i.e.

$$\mathbf{E}\left(z(\lambda_n)\right) \equiv \frac{\int z(x)L(x)f_0^n(x)dx}{\int L(x)f_0^n(x)dx}\,.$$

Then as $n \to \infty$, we have

$$\mathbf{E}\left(z(\lambda_n)\right) \to \frac{\int z(x)L(x)r(x,c)dx}{\int L(x)r(x,c)dx}\,.$$

That is, the posterior expectation converges to the value suggested by the form $r(x,y)$ of the limiting prior density ratios.

Proof. We have that

$$\frac{\int z(x)L(x)f_0^n(x)dx}{\int L(x)f_0^n(x)dx} = \frac{\int z(x)L(x)(f_0^n(x)/f_0^n(c))dx}{\int L(x)(f_0^n(x)/f_0^n(c))dx}\,.$$

The result now follows from letting $n \to \infty$, applying the dominated convergence theorem to the numerator and denominator separately, and using Proposition 1. ■

This proposition leads to the question of when the dominating condition will hold. We have the following.

Proposition 22. *Suppose that $G(\lambda, dx) = g(x - \lambda)dx$, where g is continuous at 0 and is maximised at 0. Then (10) holds with $c = 0$. Hence, the conclusions of Proposition 21 hold whenever $L \in L^1$.*

Proof. Denote by $g^{(n)}$ the n-fold convolution of the density g. Since g is symmetric, so is $g^{(n)}$ for each n. Furthermore, the convolution $g^{(2n)} = g^{(n)} * g^{(n)}$ will automatically satisfy $(g^{(n)} * g^{(n)})(x) \leq (g^{(n)} * g^{(n)})(0)$ for all $x \in \mathbf{R}$. Indeed, by Cauchy-Schwarz we have

$$(g^{(n)} * g^{(n)})(x) \;=\; \int g^{(n)}(t) g^{(n)}(x-t)dt \;\leq\; \sqrt{\left(\int (g^{(n)}(t))^2 dt\right)\left(\int g^{(n)}(x-t)^2 dt\right)}$$

$$=\; \int g^{(n)}(t)^2 dt \;=\; g^{(2n)}(0) \;=\; (g^{(n)} * g^{(n)})(0)\,.$$

It follows that $g^{(n)}(x) \leq g^{(n)}(0)$ for all even n. Hence, equation (10) holds for even n, with $K = 1$.

To handle odd values of n, write

$$\frac{g^{(2n+1)}(x)}{g^{(2n+1)}(0)} \;=\; \frac{g^{(2n+1)}(x)}{g^{(2n)}(0)}\;\frac{g^{(2n)}(0)}{g^{(2n+1)}(0)}\,.$$

We shall bound each of these two factors separately.

For the first factor, we note that since $g^{(2n)}(z)/g^{(2n)}(0) \leq 1$ for all z, we have

$$\frac{g^{(2n+1)}(x)}{g^{(2n)}(0)} = \int_z \frac{g^{(2n)})(z)}{g^{(2n)})(0)} g(x-z)dz$$

$$\leq \int g(x-z)dz \leq 1\ .$$

For the second factor, note that by continuity, there exists a positive constant ϵ such that

$$\frac{g(z)}{g(0)} \geq \epsilon$$

for $|z| \leq \epsilon$. Furthermore, by Proposition 2, the measure defined by the density $g^{(2n)}(\cdot)/g^{(2n)}(0)$ has $r(x, y) = 1$ for all x and y. Hence, by Proposition 21 applied to even n only, $\lim_{n\to\infty} \int g(-z)\frac{g^{(2n)}(z)}{g^{(2n)}(0)}dz = \int g(-z)dz$. We obtain that

$$\liminf_{n\to\infty} \frac{g^{(2n+1)}(0)}{g^{(2n)}(0)} \;=\; \liminf_{n\to\infty} \int g(-z)\frac{g^{(2n)}(z)}{g^{(2n)}(0)}dz \;=\; \int g(-z)dz \;\geq\; \int_{-\epsilon}^{\epsilon} g(-z)dz \;\geq\; 2\epsilon^2 g(0)\,.$$

Taking reciprocals, we have that

$$\limsup_{n\to\infty} \frac{g^{(2n+1)}(0)}{g^{(2n)}(0)} \;\leq\; \big(2\epsilon^2 g(0)\big)^{-1} \,.$$

Combining these bounds, we see that (10) holds for all sufficiently large n, with $K = \max\Big(1,\; \big(2\epsilon^2 g(0)\big)^{-1}\Big)$. (Note that it is easily checked that we must have $2\epsilon^2 g(0) \leq 1$, so that in fact $K = \big(2\epsilon^2 g(0)\big)^{-1}$.) ∎

9. Prior distributions and non-informativity.

This paper has largely concentrated on the simple case of prior choice for location parameters (and related families). The results in these cases are already fairly involved. However, the general idea of infinite hierarchies is considerably more flexible. It is interesting to ask to what extent general statements can be made about this approach, without assuming any structural properties of the parameter in question. The partition of Markov chain kernels into transient and recurrent provides some insight.

Again taking the location family case as a tractable example, it seems eminently reasonable to assume the martingale form of G leading to the limiting flat prior. Any other choice would indicate a prior bias towards either $+\infty$ or $-\infty$, reflecting the transience of the Markov chain to one of those limits. Further, extending to the case of general choice of G, one might argue that the notion of recurrence is consistent with non-informativity of the hierarchical construction, whereas transience expresses some kind of qualitative bias for the prior. Given this, a sensible restriction on G would be to assume recurrence.

Under the assumptions of aperiodicity and Harris recurrence, dependence between λ_0^N and λ_N^N necessarily diminishes to 0 as $N \to \infty$ as described by Orey's Theorem (see e.g. Meyn and Tweedie, 1993, Theorem 18.1.2). This is useful since it shows that recurrence concurs with another reasonable criterion for non-informativity, further supporting its use as a criterion for the construction of non-informativity.

We note that in certain cases (e.g. the centered location family case, and the exponential location family case, but not the non-centered location family case) our construction gives the standard Jeffreys prior, given by $\sqrt{\mathbf{E}(I)}$, where $I = -\frac{\partial^2}{(\partial\theta)^2}\log(\text{likelihood})$. It is unclear to what extent this agreement holds in general, and it is unlikely that our approach will turn out to be closely connected to either Jeffreys priors, reference analysis or other approaches for the choice of non-informative priors.

We have not considered the case where the latent hierarchical structure has different levels living on different spaces. This is clearly of some interest in mimicking the properties of real hierarchical models which might for instance have a variance and a mean determining the distribution of a mean, perhaps leading to the ith leave of the hierarchy containing 2^i parameters. It turns out that much of this can be covered by the existing theory described in this paper. For instance, consider the case where $\lambda_i \sim N(\lambda_{i+1}, \sigma^2_{i+1})$, where σ^2_{i+1} has its own infinite hierarchical structure. Then the conditional prior for λ_0 conditional on all the σ hierarchies is uniform by the centered martingale family case. Therefore the marginal prior for λ_0 is therefore also uniform on $\mathbf{R}$.

10. Concluding comments.

This paper discusses what happens when we consider an infinite hierarchical model, with $\lambda_i \sim G(\lambda_{i+1}, \cdot)$ for each i, where $G(\lambda, \cdot)$ is some parametric family of probability distributions. Some limited extensions to different hierarchical structures are also briefly discussed.

We have seen that the limiting distribution in such cases may be flat, or proportional to $1/x$, or other possibilities, depending on the properties of $G(\lambda, \cdot)$, but that in certain cases (for instance the location and scale family cases) the family of possible limits is limited. These results may provide some justification for the choice of certain prior distributions in Bayesian modeling.

There are two potential contributions of the ideas in this paper. The first describes a method of constructing non-informative priors in a non-informative setting in a way quite different from those currently available in the literature.

Secondly, in practice, identification of the limiting prior distribution could provide considerable computational advantages in avoiding the need for complicated hierarchical structures for prior distributions, for example in MCMC algorithms.

On the other hand, many important questions are also raised by our approach. We have made an extensive analysis of the location family case, where it is easier to produce explicit expressions for the results of our construction. A particularly appealing result of this analysis is that results obtained in a number of cases are essentially independent of the chosen hierarchical linking distribution (G). Indeed if we include the recurrence restriction discussed in Section 9, limiting flat priors are obtained in essentially all cases. This property of independence from the exact form of G will extend to other classes of problems, and the work in this paper suggests the need for a detailed study of these kind of invariance properties.

One area not covered by our paper in any detail is the construction of dependent priors (for example as used in variance component models) and their properties in infinite hierarchical structures. Example 9 illustrates at least how our approach can be easily extended to such situations. This example also illustrates the ease with which our approach can be combined with partial knowledge of hierarchical structure.

In this paper we have seen how our method of producing prior distributions is in agreement with Jeffreys prior in at leat two natural situations. However it will not always be the case that the two classes of priors coincide. Jeffreys priors have the property of second order agreement of highest posterior density regions with frequentist confidence intervals. Our hierarchical methodology will certainly fail to possess this property when the priors disagree. On the other hand, a number of advantages of our hierarchical approach have been described in the text. The real test of how effective our methodogy is, will come in more complicated examples. It will be particularly interesting to see how the approach introduced here behaves in more complex stochastic models.

Acknowledgements. We are very grateful to Jeremy Quastel for his assistance with the theory of large deviations. We thank Susie Bayarri, Ron Doney, Michael Escobar, Radford Neal, Tom Salisbury, and Richard Tweedie for helpful comments and suggestions.

REFERENCES

Bernardo, J.M. and Smith, A.F.M. (1994), Bayesian Theory. John Wiley & Sons, Chichester, England.

Billingsley, P. (1995), Probability and Measure, 3$^{\text{rd}}$ ed. John Wiley & Sons, New York.

Bingham, N.H., Goldie, C.M., and Teugels, J.L. (1987), Regular variation. Cambridge University Press.

Chung, K.L. (1974), A course in probability theory (2$^{\text{nd}}$ ed.). Academic Press, New York.

DeGroot, M.H. and Goel, P.K. (1981), Information about hyperparameters in hierarchical models. J. Amer. Stat. Assoc. **76**, 140–147.

Dembo, A. and Zeitouni, O. (1993), Large deviations techniques and applications. Jones and Bartlett, Boston, Massachusetts.

Deuschel, J.-D. and Stroock, D.W. (1989), Large deviations. Academic Press, Boston.

Feller W. (1971), An introduction to Probability Theory and its applications, Vol. *II*, 2$^{\text{nd}}$ ed. Wiley & Sons, New York.

Gustafson, P. (1996), Local sensitivity of inferences to prior marginals. JASA, 91, 774-781.

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. Series A **186**, 453–461.

Kelly, D.G. (1994), Introduction to probability. Macmillan, New York.

Meyn, S. and Tweedie, R.L. (1993), Markov chains and stochastic stability. Springer, New York.

Varadhan, S.R.S. (1984), Large deviations and applications. SIAM, Philadelphia, Pennsylvania.

Zolotarav, V.M. (1986), One dimensional stable distributions. Translations of Math. Monographs, Vol. **65**. Amer. Math. Soc., Providence, Rhode Island.