Submitted to the Annals of Statistics

STABILITY OF THE GIBBS SAMPLER FOR BAYESIAN HIERARCHICAL MODELS

BY OMIROS PAPASPILIOPOULOS[*] AND GARETH ROBERTS

University of Warwick and Lancaster University

arXiv:0710.4234v1 [stat.ME] 23 Oct 2007

We characterise the convergence of the Gibbs sampler which samples from the joint posterior distribution of parameters and missing data in hierarchical linear models with arbitrary symmetric error distributions. We show that the convergence can be uniform, geometric or sub-geometric depending on the relative tail behaviour of the error distributions, and on the parametrisation chosen. Our theory is applied to characterise the convergence of the Gibbs sampler on latent Gaussian process models. We indicate how the theoretical framework we introduce will be useful in analyzing more complex models.

1. Introduction. Hierarchical modelling is a widely adopted approach to constructing complex statistical models. The appeal of the method lies in the simplicity in specifying a highly multivariate model by joining many simple and tractable models, the foundational justification based on the ideas of partial exchangeability, the flexibility to extend or simplify the model in the light of new information, and the ease of inference using powerful Markov chain Monte Carlo (MCMC) methods which have been developed to this end during the last two decades. Thus, hierarchical models have been used in many areas of applied statistics such as geostatistics [8], longitudinal analysis [9], disease mapping [3], and financial econometrics [23] to name just a few.

[*]Research funded by EPSRC grant GR/S61577/01

AMS 2000 subject classifications: Primary 65C05; secondary 60J27

Keywords and phrases: Geometric ergodicity, capacitance, collapsed Gibbs sampler, state-space models, parametrisation, Bayesian robustness

A rather general form of a two-level hierarchical model is

$$
\begin{aligned}
Y &\sim \mathcal{L}(Y|X) \\
X &\sim \mathcal{L}(X|\Theta)\,,
\end{aligned} \tag{1}
$$

where $\mathcal{L}(X)$ and $\mathcal{L}(Y \mid X)$ denote the distribution of X and the conditional distribution of Y given X respectively. We will refer to Y as the data, X as the missing data and Θ as the parameters. In a Bayesian context the model is completed by specifying a prior distribution for Θ. Typically the dimension of X is much larger than that of Θ and it can increase with the size of the data set. Most of the applications cited above fit into (1) by imposing the appropriate structure on $\mathcal{L}(Y \mid X)$ and $\mathcal{L}(X \mid \Theta)$. It is straightforward to construct models with more levels.

Bayesian inference for (1) involves the posterior distribution $\mathcal{L}(X, \Theta \mid Y = y)$. This is typically analytically intractable, but it can be sampled relatively easily using the Gibbs sampler [29], by simulating iteratively from the two conditional distributions $\mathcal{L}(X \mid \Theta, Y = y)$, and $\mathcal{L}(\Theta \mid X, Y = y)$. It has been demonstrated both theoretically and empirically that the convergence (to be formally defined in Section 3) of the Gibbs sampler relates to the structure of the hierarchical model and particularly to the dependence between the updated components, X and Θ. Nevertheless, the exact way in which the model structure interferes with the convergence remains largely unresolved. Concrete theoretical results exist only for Gaussian hierarchical models, but we will see that these results do not extend to more general cases. Although interesting characterizations of the convergence rate in terms of the dependence between X and Θ exist when the Gibbs sampler is geometrically ergodic [1], there exist no general results which establish geometric ergodicity for the Gibbs sampler. The difficulty in obtaining such general results lies in the intrinsic dependence of the convergence of the Gibbs sampler on the model structure.

In this paper we show explicitly how the relative tail behaviour of $\mathcal{L}(Y \mid X)$ and $\mathcal{L}(X \mid \Theta)$ determines the stability of the Gibbs sampler, i.e. whether the convergence is uniform, geometric or sub-geometric. Moreover, we show that the relative tail behaviour dictates the type of parametrisation that should be adopted. In order to retain tractability and formulate interpretable and easy to check conditions we restrict attention to the class of linear hierarchical models with general error distributions; the precise model structure is given in Section 2.1. Nevertheless, our main theoretical results, in particular Theorems 3.3, 3.4, 3.5 and 6.3, and the methodology for proving them are expected to be useful in a much more general context than the one considered here.

Consideration of the class of linear non-Gaussian hierarchical models is not merely motivated by mathematical convenience. These models are very useful in real applications, for example in longitudinal random effects modelling [9, 13], time series analysis [4, 12, 28] and spatial modelling [8]. They also are a fundamental tool in the robust Bayesian analysis [7, 20, 22, 30]. Furthermore, we will see that the stability of the Gibbs sampler for linear non-Gaussian models is very different compared to the Gaussian case, the local dependence between X and Θ being crucial in the non-Gaussian case. Notice that several other models can be approximately written as linear non-Gaussian models. Actually, this work has been motivated by the behaviour of MCMC for non-Gaussian Ornstein-Uhlnebeck stochastic volatility models [23].

The paper is organised as follows. Section 2.1 specifies the models we will be concerned with and it establishes some basic notation. Section 2.2 discusses Gibbs sampling under different parametrisations of the model and Section 2.3 motivates the theory and the methodology developed in this paper by a simple example. Section 3 is the theoretical core of this paper; the section commences with a short review of stability concepts for the Gibbs sampler; Section 3.1 recalls the existing results for Gaussian linear models; Section 3.2 develops stability theory for hierarchical models and states three main theorems for the stability of the Gibbs sampler; based on these theorems Section 3.3 provides the characterization of the stability of the Gibbs sampler under different parametrisations for a broad class of linear hierarchical models; Section 3.4 considers an alternative augmentation scheme when one of the error distributions is a scale mixture of normals and compares the convergence of a three-component Gibbs sampler with that of its collapsed two-component counterpart. Section 4 extends the theory to hierarchical models which involve latent Gaussian processes. Section 5 discusses extensions and contains some practical guidelines. Section 6 contains the proofs of all theorems and propositions. The proofs are based on establishing geometric drift conditions and minorization conditions and using capacitance arguments in conjunction with Cheeger's inequality.

2. Models, parametrisations and motivation.

2.1. *Linear hierarchical models.* The models we consider in this paper are of the following form, where $\mathbf{Y}_i$ is $m_i \times 1$, $\mathbf{C}_i$ is $m_i \times p$, $\mathbf{X}_i$ is $p \times 1$, $\mathbf{D}$ is $p \times 1$ and Θ is a scalar:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{C}_i \mathbf{X}_i + \mathbf{Z}_{1i}\,, \quad i = 1, \ldots, m \\ \mathbf{X}_i &= \mathbf{D}\Theta + \mathbf{Z}_{2i}\,. \end{aligned} \tag{2}$$

$\mathbf{Z}_{1i}, i = 1, \ldots, m$, are iid with distribution $\mathcal{L}(\mathbf{Z}_1)$, $\mathbf{Z}_{2i}, i = 1, \ldots, m$, are iid with distribution $\mathcal{L}(\mathbf{Z}_2)$, and $\mathcal{L}(\mathbf{Z}_1)$ and $\mathcal{L}(\mathbf{Z}_2)$ are symmetric distributions around $\mathbf{0}$ (a vector of 0s with the appropriate dimension). In the sequel, bold-face letters will correspond to vectors and matrices, capital letters to random variables and lower-case letters to their realisations. In this setting $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$. The first equation in (2) will be termed the *observation equation* and the second the *hidden equation*.

It is often conveniently assumed that both $\mathcal{L}(\mathbf{Z}_1)$ and $\mathcal{L}(\mathbf{Z}_2)$ are Gaussian. However there are several applications where this assumption is clearly inappropriate, especially if we wish to make the inference about $\mathbf{X}$ robust in the presence of prior-data conflict. It is known [see e.g. 20, 22, 30, and references therein] that if the tails of $\mathcal{L}(\mathbf{Z}_1)$ are heavier than the tails of $\mathcal{L}(\mathbf{Z}_2)$ then inference for $\mathbf{X}$ is robust to outlying observations, whereas if

$\mathcal{L}(\mathbf{Z}_2)$ has heavier tails than $\mathcal{L}(\mathbf{Z}_1)$ inference for $\mathbf{X}$ is less influenced by the prior in case of data-prior conflict; these robustness is absent from Gaussian models. This type of robust modelling has been undertaken in time-series analysis, see for example [12].

2.2. *Gibbs sampling and parametrisations.* As is common in this framework, we place an improper flat prior on Θ, which in this context leads to a proper posterior. Bayesian inference for (2) involves the joint posterior distribution $\mathcal{L}(\mathbf{X}, \Theta \mid \mathbf{Y} = \mathbf{y})$, which will abbreviate to $\mathcal{L}(\mathbf{X}, \Theta \mid \mathbf{Y})$. Although it is often analytically intractable, it can be sampled easily using the Gibbs sampler.

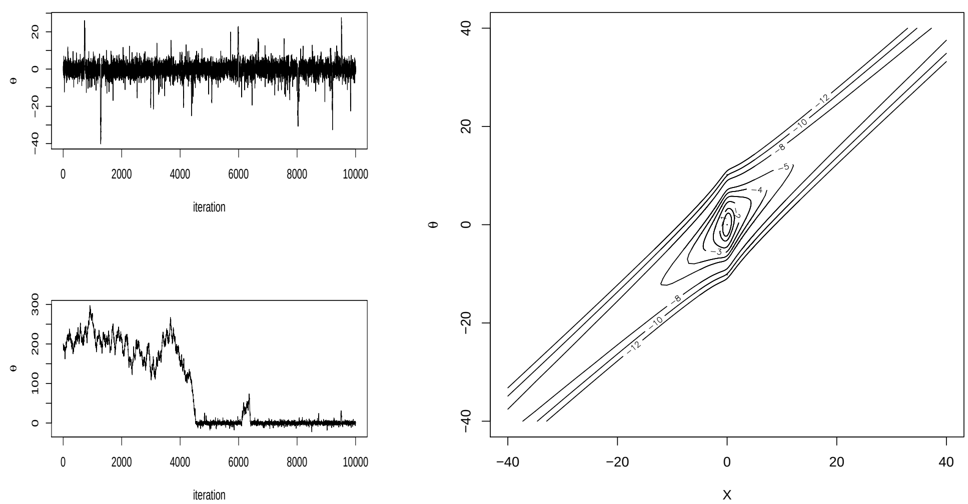
The parametrisation $\mathcal{P}_0 := (\mathbf{X}, \Theta)$ is termed the *centred parametrisation.* This terminology was first used in the linear Gaussian context by [10]. Following [21] we shall use the term more generally to refer to a parametrisation where the parameters and the data are conditionally independent given the missing data. We can use the Gibbs sampler to collect samples from $\mathcal{L}(\mathbf{U}, \Theta \mid \mathbf{Y})$ where $\mathbf{U} = h(\mathbf{X}, \Theta)$, for some invertible transformation h, and then transform the draws to obtain samples from $\mathcal{L}(\mathbf{X}, \Theta \mid \mathbf{Y})$. In the rest of the paper we will use $\mathcal{P}$ to refer to a general parametrisation $(\mathbf{U}, \Theta)$. It is known [16] that the convergence (to be formally introduced in Section 3) of the Gibbs sampler improves as the dependence between the updated components, $\mathbf{U}$ and Θ, decreases. Hence, the development of general re-parametrisation strategies has been actively researched, see [21] for a recent account. In that work, the authors introduce the *non-centred reparametrisation* $\mathcal{P}_1 := (\tilde{\mathbf{X}}, \Theta)$, which replaces $\mathbf{X}$ with $\tilde{\mathbf{X}} := h(\mathbf{X}, \Theta)$, where h is a transformation which makes Θ and $\tilde{\mathbf{X}}$ apriori independent. In the context of linear hierarchical models $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_m)$, where $\tilde{\mathbf{X}}_i = h(\mathbf{X}_i, \Theta)$, and $h(\mathbf{x}, \theta) := \mathbf{x} - \mathbf{D}\theta$. We will see that $\mathcal{P}_0$ and $\mathcal{P}_1$ present two natural choices.

The prolific expansion in the use of Gibbs sampling for inference in hierarchical models during the 1990s was fuelled by the apparent rapid convergence of the algorithm in many cases. However, to date, there has been little theoretical analysis linking the stability of the Gibbs sampler to the structure of hierarchical models. A notable exception are the explicit convergence results for Gaussian linear hierarchical models obtained in [24] and summarised in Section 3.1. The following example is revealing as to what might go wrong when considering non-Gaussian linear models, and motivates the methodology and theory developed in this article.

2.3. *A motivating example.* Consider a simplified version of (1) where $m = m_1 = C_1 = D = 1$,

$$\begin{aligned} Y &= X + Z_1 \\ X &= \Theta + Z_2. \end{aligned} \tag{3}$$

Assume that $\mathcal{L}(Z_1) = \mathrm{Ca}(0,1)$, a standard Cauchy distribution, $\mathcal{L}(Z_2) = \mathrm{N}(0,5)$, and $y = 0$ is observed. Figure 2.3a shows the sampled values of Θ after two independent runs of the Gibbs sampler, each of 10^4 iterations. The top one is started from the mode, $\Theta_0 = 0$, and superficially it appears to be mixing well: the autocorrelation in the series becomes negligible after 10 lags, and most convergence diagnostic tests would assess that the chain has converged. Nevertheless, the chain never exits the set $(-40, 40)$, although this is an event with stationary probability about 0.015. The second run, Figure 2.3a bottom, is started from $\Theta_0 = 200$, and the chain spends more than 4,000 iterations wondering around Θ_0. The contour plot of the joint posterior log-density of X and Θ in Figure 2.3b, provides an explanation: the contours look roughly spherical near the mode, but they become asymptotically concentrated around $x = \theta$ as $|\theta| \to \infty$. Thus, restricted to an area around the mode, X and Θ look roughly independent, but in the tails they are highly dependent. In fact, $\mathcal{L}(X - \theta \mid Y, \Theta = \theta) \to \mathrm{N}(0,5)$ as $|\theta| \to \infty$, and we show in Section 3.3 that the Gibbs sampler which updates X and Θ converges sub-geometrically. In contrast, $\mathcal{L}(\tilde{X} \mid Y, \Theta = \theta) \to \mathcal{L}(\tilde{X})$, as $|\theta| \to \infty$, and as we show in Section 3.3 the Gibbs sampler which updates $\tilde{X}$ and Θ is uniformly ergodic.



(a) (b)

FIG 1. *(a): two runs of the Gibbs sampler under $\mathcal{P}_0$ for the model (3) started at $\Theta_0 = 0$ (top) and $\Theta_0 = 200$ (bottom). (b): contours of the joint posterior log-density of X and Θ.*

3. Convergence of the Gibbs sampler for linear hierarchical models. Given the parametrisation $\mathcal{P} = (\mathbf{U}, \Theta)$, the two-component Gibbs sampler simulates iteratively from $\mathcal{L}(\mathbf{U}|\mathbf{Y}, \Theta = \Theta_{n-1})$, and $\mathcal{L}(\Theta|\mathbf{Y}, \mathbf{U} = \mathbf{U}_n)$, where Θ_0 is a starting value and $n \geq 1$ denotes the iteration number. This algorithm generates a Markov chain $\{(\mathbf{U}_n, \Theta_n)\}$ with stationary distribution $\mathcal{L}(\mathbf{U}, \Theta \mid \mathbf{Y})$. The marginal chain $\{\Theta_n\}$ is also Markov and reversible with respect to $\mathcal{L}(\Theta \mid \mathbf{Y})$ (Lemma 3.1. of [16]). Moreover, it can be shown [26] that the convergence rate of the joint chain coincides with the convergence rate of the marginal chain, $\{\Theta_n\}$. Notice that this result does not hold for Gibbs samplers which update more than two components. In the sequel, for any random variables W and V, and probability law μ, we will use the short-hand notation,

$$\mathcal{L}(V \mid W \sim \mu) := \int \mathcal{L}(V \mid W = w)\mu(dw).$$

We will consider the convergence of $\{\Theta_n\}$ through the total variation norm, defined as

$$\|\mathcal{L}_h(\Theta_n \mid \mathbf{Y}, \Theta_0) - \mathcal{L}(\Theta|\mathbf{Y})\| = \sup_{|g|\leq 1} |\mathbf{E}_h\{g(\Theta_n) \mid \mathbf{Y}, \Theta_0\} - \mathbf{E}\{g(\Theta)|\mathbf{Y}\}|.$$

$\mathcal{L}_h(\Theta_n \mid \mathbf{Y}, \Theta_0)$ is the distribution of the chain after n steps started from Θ_0, and $\mathbf{E}_h\{g(\Theta_n) \mid \mathbf{Y}, \Theta_0\}$ is the expected value of a real bounded function g with respect to this distribution. $\mathcal{L}_h(\Theta_n \mid \mathbf{Y}, \Theta_0)$ clearly depends on the parametrisation $\mathbf{U} = h(\mathbf{X}, \Theta)$, since,

$$\mathcal{L}_h(\Theta_1 \mid \mathbf{Y}, \Theta_0) = \mathcal{L}\Big\{ \; \Theta \; \Big| \; \mathbf{Y}, \mathbf{U} \sim \mathcal{L}(\mathbf{U} \mid \mathbf{Y}, \Theta = \Theta_0)\Big\}.$$

Under standard regularity conditions (Theorem 13.0.1 of [19]) the total variation norm converges to 0 as $n \to \infty$. We say that $\{\Theta_n\}$ is *geometrically ergodic* when there exist an $r < 1$ and some function $M(\cdot)$, such that

$$\|\mathcal{L}_h(\Theta_n \mid \mathbf{Y}, \Theta_0) - \mathcal{L}(\Theta|\mathbf{Y})\| \leq M(\Theta_0) r^n. \tag{4}$$

The smallest r for which (4) holds, say r_h, is known as the *rate of convergence* of $\{\Theta_n\}$. However, the actual distance from stationarity will in general depend on the starting point and this is represented by the term $M(\Theta_0)$ in (4). When $M(\cdot)$ is bounded above, $\{\Theta_n\}$ is called *uniformly ergodic*. Uniform ergodicity is a valuable property, since it ensures that the convergence of the chain does not depend critically on the initial value chosen. Whilst this does not guarantee rapid convergence, it ensures that the "burn-in" problem cannot become arbitrarily bad from certain starting points.

Geometric ergodicity is a qualitative stability property, and geometrically ergodic algorithms may still converge slowly and give Monte Carlo estimates with high variance (for example when $r_h \approx 1$). However, algorithms which fail to be geometrically ergodic can lead to various undesirable properties, including the break down of the central limit theorem for ergodic average estimates. In this case the simulation can be unreliable and the drawn samples might poorly represent the target distribution.

To keep nomenclature simple we will identify a parametrisation $\mathcal{P} = (\mathbf{U}, \Theta)$ with the Gibbs sampler which updates $\mathbf{U}$ and Θ. Thus, we say that a parametrisation $\mathcal{P}$ is geometrically (respectively uniformly) ergodic, if the Gibbs sampler implemented using this parametrisation is geometrically (respectively uniformly) ergodic.

3.1. *Gaussian models.* The Gibbs sampler for the Gaussian linear model is geometrically ergodic with rate given in [24]. In the simplified model (3) assume that $\mathcal{L}(Z_i) = N(0, \sigma_i^2), i = 1, 2$, and define $\kappa = \sigma_2^2/(\sigma_2^2 + \sigma_1^2)$. Then, [21] building on the results of [24] showed that, when $U = h(X, \Theta) = X - \rho\Theta$,

$$r_h := r_\rho = \frac{(\rho - (1-\kappa))^2}{\rho^2\kappa + (1-\rho)^2(1-\kappa)} = \{\mathrm{corr}(U, \Theta \mid Y)\}^2 \tag{5}$$

which gives rise to the two special cases of interest, $r_0 = 1 - \kappa$, $r_1 = \kappa$. In this setting, the dependence between U and Θ is appropriately quantified by the correlation coefficient, and (5) shows that the larger the correlation the worse the convergence. Many refinements and generalizations of these results can be found in [24], [21] and [17]. Notice that both $\mathcal{P}_0$ and $\mathcal{P}_1$ are geometrically ergodic. $\mathcal{P}_0$ converges rapidly when the observation equation is "more precise" than the hidden equation, that is $\sigma_1 << \sigma_2$, and it converges slowly when the hidden equation is relatively precise. $\mathcal{P}_1$ converges rapidly when the hidden equation is relatively more precise.

3.2. *General theory for linear hierarchical models.* This section gives general results which can be used to characterise the stability of the Gibbs sampler on linear hierarchical models of the form (2) where the X_is are univariate and $D = 1$. Our results are valid when $m > 1$ and $m_i > 1$ (see Remark 1 in page 13), however in order to keep the notation simple we will work with the simplified model (3), where all Y, X and Θ are scalars. $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ are arbitrary symmetric distributions with continuous bounded everywhere positive densities, f_1 and f_2 respectively; common examples include the Gaussian, the Cauchy and the double exponential. This section gives the general results, while Section 3.3 applies them to characterise the convergence of the Gibbs sampler for (a broad class of) linear non-Gaussian hierarchical models. Section 4 deals with extensions where the X_is are vectors of dependent variables, therefore covering state-space and spatial models. Nevertheless, the results even for the more structured models follow relatively easily from the results of this section. All proofs are deferred to Section 6.

We begin by introducing a collection of posterior robustness concepts, which are related with the behaviour of the conditional posterior distribution $\mathcal{L}(U \mid Y, \Theta = \theta)$ as $|\theta| \to \infty$. All these concepts have statistical interpretations but they turn out to provide the required mathematical conditions for characterising the stability of the Gibbs sampler, as we show in Theorems 3.3, 3.4 and 3.5 below.

DEFINITION 3.1. *The parametrisation $\mathcal{P} = (U, \Theta)$ is called:*

1. *partially tight in parameter (PTIP), if for all y, there is some $k > 0$ such that,*

$$\limsup_{|\theta|\to\infty} \mathbf{P}(|U| > k | Y = y, \Theta = \theta) < 1, \tag{6}$$

2. *geometrically tight in parameter (GTIP), if there exist positive constants, a, b (independent of θ) such that for all θ,*

$$\mathbf{P}(|U| > x | Y = y, \Theta = \theta) \leq a e^{-bx}.$$

GTIP not only implies that $\mathcal{L}(U \mid Y, \Theta = \theta)$ is a tight family of distributions, but also that the tail probabilities are bounded exponentially. (We recall that a family of distributions on the real line, say F_θ, indexed by a scalar θ, is called *tight* when $\lim_{k\to\infty} \sup_\theta F_\theta([-k,k]^c) = 0$.) Clearly, GTIP is much stronger condition than PTIP. We consider also the following model robustness concepts.

DEFINITION 3.2. *We say that the linear hierarchical model (3) is*

1. *robust in parameter (RIP), if*

$$\lim_{|\theta|\to\infty} \mathcal{L}(X|Y=y,\Theta=\theta) = \mathcal{L}(Z_1+y),$$

2. *robust in data (RID), if*

$$\lim_{|\theta|\to\infty} \mathcal{L}(\tilde{X}|Y=y,\Theta=\theta) = \mathcal{L}(\tilde{X}),$$

3. *data uniformly relevant (DUR), if there exist positive constants d, k such that for all $|\theta|>k$,*

$$|\mathbf{E}\{X|Y=y,\Theta=\theta\}| \le |\theta| - d,$$

4. *parameter uniformly relevant (PUR), if there exist positive constants d, k such that for all $|\theta|>k$,*

$$sgn(\theta)\mathbf{E}\{X-y|Y=y,\Theta=\theta\} \ge d.$$

These definitions characterise the hierarchical model according to how inference for X (conditionally on $\Theta=\theta$) is affected by a large discrepancy between the data y and the prior guess θ. When the model is RIP inference for X ignores θ, and it is symmetric around y. Conversely, when the model is RID inference for X ignores the data and becomes symmetric around θ. When the model is DUR (PUR) the data (the parameter) always influences the conditional expectation of X. Notice that when the model is RIP $\mathcal{P}_0$ is PTIP (although not necessarily GTIP), and when it is RID $\mathcal{P}_1$ is PTIP. The example in Section 2.3 describes a RID model. A model can be both DUR and PUR (for example the Gaussian linear model).

THEOREM 3.3. *Consider the linear hierarchical model (3) where the error densities f_1 and f_2 are continuous, bounded and everywhere positive. If $\mathcal{P}_0$ ($\mathcal{P}_1$) is PTIP, then it is uniformly ergodic.*

THEOREM 3.4. *Consider the linear hierarchical model (3) where the error densities f_1 and f_2 are continuous, bounded and everywhere positive. If the model is RID then $\mathcal{P}_0$ is not geometrically ergodic, and if the model is RIP then $\mathcal{P}_1$ is not geometrically ergodic.*

Distribution	Code	Density $g(x)$ up to proportionality
Cauchy	C	$\sigma^2/(1+x^2)$
Double exponential	E	$\exp\{-\lvert x\rvert/\sigma\}$
Gaussian	G	$\exp\left\{-(x/\sigma)^2/2\right\}$
Exponential power distribution	L	$\exp\left\{-\lvert x/\sigma\rvert^\beta\right\}, \quad \beta > 2$

TABLE 1
Distributions for the error terms and their densities. In the paper they are coded according to the letter in the middle column.

The proof Theorem 3.4 is based on the general Theorem 6.3 about Markov chains on the real line, which is stated and proved in Section 6.

THEOREM 3.5. *1. If the model is DUR, $\mathcal{P}_1$ is GTIP, and $\mathcal{L}(Z_2)$ has finite moment generating function in a neighbourhood of* 0*, then $\mathcal{P}_0$ is geometrically ergodic. 2. If the model is PUR, $\mathcal{P}_0$ is GTIP, and $\mathcal{L}(Z_1)$ has finite moment generating function in a neighbourhood of* 0*, then $\mathcal{P}_1$ is geometrically ergodic.*

The theorems are proved by establishing a geometric drift condition. The requirements of GTIP for $\mathcal{P}_1$ ($\mathcal{P}_0$) and finite moment generating function for $\mathcal{L}(Z_2)$ ($\mathcal{L}(Z_1)$) are in order to tilt exponentially the linear drift condition provided by DUR (PUR).

3.3. *Characterising the stability of the Gibbs sampler according to the distribution tails of the error terms.* In this section, building upon the general theory of Section 3.2, we characterise the stability of the Gibbs sampler on the linear hierarchical model (3) for different specifications of $\mathcal{L}(Z_1), \mathcal{L}(Z_2)$. Although we consider the error distributions in Table 1, our proofs remain valid for much broader families of distributions (see Remark 2 on page 13). Notice that the exponential power distribution contains both the Gaussian ($\beta = 2$) and the double exponential ($\beta = 1$) as special cases. Here we consider densities with tails lighter than Gaussian ($\beta > 2$). For the use of this distribution in Bayesian robustness see [5].

We shall specify linear models giving first $\mathcal{L}(Z_1)$ and then $\mathcal{L}(Z_2)$, for instance the (C, E) model corresponds to (3) with Cauchy distribution for

Stability of $\mathcal{P}_0$					
		$\mathcal{L}(Z_1)$			
		C	E	G	L
$\mathcal{L}(Z_2)$	C	U	U	U	U
	E	N	G/U	U	U
	G	N	G	G	G
	L	N	G	G	G

Stability of $\mathcal{P}_1$					
		$\mathcal{L}(Z_1)$			
		C	E	G	L
$\mathcal{L}(Z_2)$	C	U	N	N	N
	E	U	U/G	G	G
	G	U	U	G	G
	L	U	U	G	G

TABLE 2
Stability $\mathcal{P}_0$ (left) and $\mathcal{P}_1$ (right) for the linear hierarchical model (3) for specifications of the distribution of the error terms as in Table 1.

Z_1, and double exponential distribution for Z_2. For each model we have two parametrisations, thus two algorithms, $\mathcal{P}_0$ and $\mathcal{P}_1$. When we refer to the stability of an algorithm we shall write U, G, and N to refer to uniform, geometric and non-geometric (i.e. sub-geometric) ergodicity, respectively.

THEOREM 3.6. *The stability $\mathcal{P}_0$ and $\mathcal{P}_1$ is given in Table 2.*

Remark 1. The determining factor in classifying the stability of a parametrisation is the tail behaviour of $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$. Thus, Theorem 3.6 generalises to the case of multiple random effects and observations:

$$\begin{aligned} Y_{ij} &= X_i + Z_{1ij},\ j = 1, \ldots, m_i \\ X_i &= \Theta + Z_{2i},\ i = 1, \ldots, m \end{aligned}$$

where $Z_{1\cdot\cdot}$ and $Z_{2\cdot}$ are independently distributed identically to $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ respectively. This extension is immediate where obvious sufficient statistics exist (the C and N cases). However, since proving formally the full generalisation would be extremely tedious (although in the same lines as in Section 6), we do not attempt it here.

Remark 2. The same results can be obtained when any of the distributions considered in Table 2 is replaced by another symmetric distribution with the same tail behaviour, which possess a bounded continuous everywhere positive density.

Remark 3. Different results hold when a proper prior for Θ is imposed. In this case the convergence improves.

Remark 4. The results of Theorem 3.6 are independent of the actual value of y. This does not necessarily hold in other contexts.

Remark 5. In the (E, E) model, the stability depends on the ratio of the scale parameters in $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$. Depending on this ratio convergence can be either geometric or uniform (see Section 6 for details).

Remark 6. The following heuristic can be derived from Table 2: convergence of $\mathcal{P}_0$ is best when $\mathcal{L}(Z_1)$ has lighter tails than $\mathcal{L}(Z_2)$, and worst when it has heavier tails. The situation for $\mathcal{P}_1$ is the reverse. Both algorithms become more stable the lighter the tails of $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ become.

3.4. *Convergence of the grouped Gibbs sampler.* An alternative augmentation scheme and sampling algorithm can be adopted when one of the error distributions, say $\mathcal{L}(Z_2)$ for convenience, is Gaussian and the other, say $\mathcal{L}(Z_1)$, is a scale mixture of Gaussian distributions. Several symmetric distributions belong in this class, for instance the Student-t (thus the Cauchy) and the double exponential [2]. In this case, Z_1 can be represented as $Z_1 = V/Q$, where V has a standard Gaussian distribution and Q is positive and independent of V. We can treat Q as missing data and construct a *three-component* Gibbs sampler which updates iteratively X, Q and Θ from their conditional distributions. (When $\mathbf{X} = (X_1, \ldots, X_m)$ then $\mathbf{Q} = (Q_1, \ldots, Q_m)$ where Q_i is independent from Q_j for every $i \neq j$). A major computational advantage of this approach is that $\mathcal{L}(X \mid Y, \Theta, Q)$ is Gaussian and it can be easily sampled. Notice that Q and Θ are independent given X, thus we can implement the Gibbs sampler using a *grouped* scheme [15] where Θ and Q are updated in one block. It is of interest to know whether the convergence of this grouped Gibbs sampler is better than the convergence of the *collapsed* Gibbs sampler (as defined in [15]), where Q has been integrated out. The "Three-schemes Theorem" of [15] states that the norm of the transition operator of the grouped Gibbs sampler is larger than the one which corresponds to the collapsed Gibbs sampler. This result, however, is not enough to guarantee that the collapsed sampler will have better convergence rate.

In order to give a concrete answer, we consider the important special case, where $\mathcal{L}(Z_1)$ is the Cauchy distribution, therefore $Q \sim \mathrm{Ga}(1/2, 1/2)$. We have the following proposition, whose proof is based on Theorem 6.3.

PROPOSITION 3.7. *The grouped Gibbs sampler is not geometrically ergodic.*

This result remains true for a number of random effects $m > 1$, and it will hold for more general Student-t distributions. This result has important practical implications especially in algorithms for latent Gaussian models, considered in Section 4. It is also significant that it contrasts the result obtained by [27], who establishes geometric ergodicity for variance component models (of which the model considered here is a special case). However, the result in [27] is true when the number of data Y_{ij}, m_i, per random effect X_i is larger than some number bigger than one, whereas in Lemma 3.7 we take $m_i = 1$.

4. Latent Gaussian process models. In this section we consider a rather specific though useful model and demonstrate that the results of Section 3.2 can be extended quite readily to this context giving some clear-cut conclusions and advice for practical implementation. The results below are certainly not the most general possible, but it is hoped that the method of proof will indicate how analogous models might be addressed.

THEOREM 4.1. *Consider the latent Gaussian process model:*

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} + \mathbf{Z}_1 \\ \mathbf{X} &= \mathbf{1}\Theta + \mathbf{\Sigma}^{1/2}\mathbf{Z}_2 \end{aligned}$$

where $\mathbf{Z}_1 = \{Z_{11}, \ldots Z_{1p}\}$ *is a vector of independent and identically distributed standard Cauchy random variables,* $\mathbf{Z}_2 = \{Z_{21}, \ldots Z_{2p}\}$ *is a vector of independent and identically distributed standard Gaussian random variables, and* $\mathbf{1}$ *is a vector of* 1*'s.* $\mathbf{\Sigma}$ *is assumed known and a flat is prior is assigned to* Θ*. Then 1.* $\mathcal{P}_0$ *fails to be geometrically ergodic; 2.* $\mathcal{P}_1$ *is uniformly ergodic.*

As we remarked on page 13, the result holds when the Cauchy is generalised to a Student-t with any degrees of freedom. The MCMC for latent Gaussian process models is often implemented using a different augmentation scheme. As in Section 3.4, we can augment the model with $\mathbf{Q} = (Q_1, \ldots, Q_p)$, where $\mathcal{L}(Q_i) = \mathrm{Ga}(1/2, 1/2)$. However, a similar argument as in the proof of Proposition 3.7 shows that the Gibbs sampler which updates $\mathbf{X}, \mathbf{Q}$ and Θ is not geometrically ergodic.

As a numerical illustration we consider a linear non-Gaussian state-space model: $X_1, \ldots, X_p$ are consecutive draws from an AR(1) model, which are observed with Cauchy error. We have simulated $p = 100$ data from this model using $\Theta = 0$. The update of Θ given $\mathbf{X}$ is from a Gaussian distribution, however the update of $\mathbf{X}$ given Θ and $\mathbf{Y}$ is non-trivial. We update all the states together using a highly efficient Langevin algorithm, see [6] for details. Moreover, we perform several updates of $\mathbf{X}$ for every update of Θ so that our results are not critically affected by not being able to simulate directly from $\mathcal{L}(\mathbf{X} \mid \mathbf{Y}, \Theta)$. Figure 4 depicts our theoretical findings. $\mathcal{P}_0$ has a random walk-like behaviour in the tails, whereas $\mathcal{P}_1$ returns rapidly to the modal area. On the other hand, $\mathcal{P}_0$ mixes better than $\mathcal{P}_1$ around the mode. Note that the instability of $\mathcal{P}_0$ in the tails is not due to lack of information about Θ but due to the robustness properties of the model.

In this context it is definitely advisable to mix between $\mathcal{P}_0$ and $\mathcal{P}_1$, i.e to use a hybrid sampler which at every iteration with some probability updates $(\Theta, \mathbf{X})$ and with the remaining probability it updates $(\Theta, \tilde{\mathbf{X}})$. This hybrid sampler will inherit the uniform ergodicity from $\mathcal{P}_1$ but it will also mix well

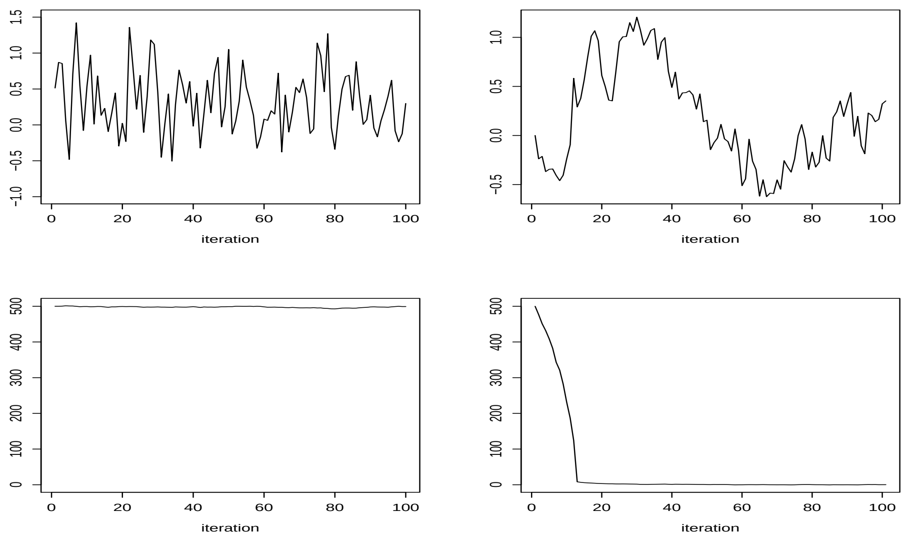


FIG 2. *Two runs of $\mathcal{P}_0$ (left) and $\mathcal{P}_1$ (right) with two different starting values: $\Theta_0 = 0$ (top) and $\Theta_0 = 500$ (bottom).*

around the modal area.

5. Discussion. We have obtained rigorous theoretical results for the stability of the Gibbs sampler which explores the posterior distribution arising from a broad class of linear hierarchical models. We have also proved results regarding more complicated hierarchical models with latent Gaussian processes, and we have compared different sampling schemes. We have shown how the model structure dictates which parametrisation should be adopted for improving the convergence of the Gibbs sampler.

Our results are certainly not the most general possible, though the method of proof we have used indicates clearly how analogous problems might be addressed. As an example of this, it is easy to extend the conclusions of Table 2 to the case where the light-tailed distributions are replaced by (say) uniform distributions on finite ranges. The robustness concepts of PTIP, GTIP, RIP and RID are already stated in a general form, while the concepts of DUR and PUR can be translated in a natural way using Lyapunov drift conditions. Families of models to which we are currently investigating extensions of our methods, include stochastic volatility models prevalent in finance. This is the subject of on-going research by the authors.

The general heuristic is clear - the stability of the centred and non-centred algorithms, $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively, depends on the relative tail behaviour of $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$, with the centred method being more stable when $\mathcal{L}(Z_1)$ is relatively light tailed, and the non-centered being more stable when $\mathcal{L}(Z_2)$ is relatively light tailed. An additional conclusion of Table 2 is that, as expected, both algorithms possess comparatively more stable convergence properties the lighter the tails of $\mathcal{L}(Z_1)$ and $\mathcal{L}(Z_2)$ become.

The main message of the paper for the MCMC practitioner is a positive one: the competition between $\mathcal{P}_0$ and $\mathcal{P}_1$ works to the user's benefit. Our results suggest that a combination of $\mathcal{P}_0$ and $\mathcal{P}_1$ is often desirable. When the tails of the error distributions are very different we have found that one of the algorithms might be very good for visiting the tails of the target distribution whereas the other for exploring the modal area (as for example we demonstrate in Figure 4). Therefore, it is advisable to use a hybrid Gibbs sampler which at every iteration with some probability updates (Θ, X) and with the remaining probability it updates $(\Theta, \tilde{X})$. Moreover, by linking the stability of the Gibbs sampler to the robustness properties of the hierarchical model we provide intuition which can be found useful for models outside the scope of this paper.

Another interesting product of this work is that linear re-parametrisations, which can substantially improve the convergence rate in (approximately) Gaussian models, might be of little relevance when the tail behaviour of $\mathcal{L}(Z_1)$ is very different from $\mathcal{L}(Z_2)$. For example, in (C,G) model, where the observation error is Cauchy and the prior for X is Gaussian, we can prove that the Gibbs sampler which updates $U = X - \rho\Theta$ and Θ is sub-geometrically ergodic for all $\rho < 1$, whereas it is uniformly ergodic for $\rho = 1$ as we already know from Theorem 3.6. This emphasizes the special role of $\mathcal{P}_1$, which differs because of the prior independence it induces on $\tilde{X}$ and Θ.

This result suggests that conditional augmentation (as in [18]) algorithms might fail to be geometrically ergodic when $\mathcal{P}_0$ does.

All the results presented here are specific to the Gibbs sampler, however our findings are clearly relevant to contexts where certain direct simulation steps have to be replaced by appropriate Metropolis-Hastings steps (as for example in the simulation illustration in Section 4).

It is worth mentioning that once we have established geometric ergodicity for an algorithm, it is important to obtain computable bounds on the rate of convergence. We have not attempted to do so, since it is outside the focus of this paper. For advances in this direction see for example [11, 27].

One interesting feature resulting from this paper is that the marginal chain $\{\Theta_n\}$ of the Gibbs sampler on linear non-Gaussian models often behaves asymptotically (i.e in the tails) like a random auto-regression of the form:

$$\Theta_n \;=\; \rho_n \Theta_{n-1} + \epsilon_n$$

where ρ_n is a random variable taking values in $[0,1]$, and ϵ_n is an error term. For instance in the (G, G) case of Theorem 3.6 for $\mathcal{P}_0$ ($\mathcal{P}_1$) ρ_n is deterministically equal to r_0 (r_1) defined in Section 3.1. The cases where we demonstrate that the algorithm is random-walk like correspond to taking $\rho_n = 1$ (*almost surely*). Furthermore in a number of cases, ρ_n is genuinely random. For instance, in the (E, E) case with identical rates, $\rho_n \sim U[0,1]$. In the (C,C) case, we find that ρ_n takes the value 0 or 1 with probabilities determined by the scale parameters of the Cauchy distributions involved.

An extension of our ideas is possible for hierarchical models with more levels. For instance consider the linear structure given by

$$\begin{aligned} Y &= \Theta_1 + Z_1 \\ \Theta_i &= \Theta_{i+1} + Z_{i+1}, \quad i = 1, \dots d-1 \,, \end{aligned} \tag{7}$$

with a flat prior on Θ_d. Since Y is the only information available, the posterior tails of $\Theta_1, \Theta_2 \dots$ become progressively heavier. If at any stage, Z_i has lighter tails than Z_{i-1}, then whenever Θ_{i-1} and Θ_{i+1} strongly disagree, the conditional distribution of Θ_i given $Y, \mathbf{\Theta}_{-i}$ will virtually ignore Θ_{i-1} and hence the data. This will lead to potential instabilities in the chain in components $\Theta_i, \Theta_{i+1}, \dots, \Theta_d$. We call this phenomenon the *quicksand* principle, and this is the subject of ongoing investigation by the authors.

6. Proofs of main results. In the sequel we will use π to denote the density of any stationary measure, in particular $\pi(\theta \mid y)$ and $\pi(x \mid y, \theta)$ will be the Lebesgue densities of $\mathcal{L}(\Theta \mid Y = y)$ and $\mathcal{L}(X \mid Y = y, \Theta = \theta)$ respectively. With $p(\cdot,\cdot)$ we denote the transition density of a Markov chain, and with Θ_0 and Θ_1 the consecutive values of the marginal chain $\{\Theta_n\}$.

PROOF OF THEOREM 3.3. We show the result for $\mathcal{P}_0$, since the corresponding result for $\mathcal{P}_1$ can be proved in an analogous way. In particular, we show that when $\mathcal{P}_0$ is PTIP, the transition density of the the marginal chain $\{\Theta_n\}$, is such that $\inf_{\theta_0} p(\theta_0, \theta_1) > 0$, and p is also continuous in θ_1. This guarantees uniform ergodicity by Theorem 16.0.2 of [19].

$$\begin{aligned} p(\theta_0, \theta_1) &= \int f_2(|x - \theta_1|)\pi(x \mid y, \theta_0)dx \geq \int_{-k}^{k} f_2(|x - \theta_1|)\pi(x \mid y, \theta_0)dx \\ &\geq \inf_{|x| \leq k} f_2(|x - \theta_1|)\ \mathbf{P}(|X| \leq k | Y = y, \Theta = \theta_0), \end{aligned}$$

for k such that (6) holds. Since f_1 and f_2 are everywhere positive, bounded and continuous, $\mathbf{P}(|X| \leq k | Y = y, \Theta = \theta_0)$ is also positive and continuous in θ_0, therefore by the PTIP property it follows that $\inf_{\theta_0} \mathbf{P}(|X| \leq k | Y = y, \Theta = \theta_0) > 0$. Moreover, $\inf_{|x| \leq k} f_2(|x - \theta_1|)$, is positive and continuous in θ_1, thus the result follows. □

The proof of Theorem 3.4 requires Theorem 6.3, hence it is proved on page 24. The proof of Theorem 3.5 requires the following lemmas.

LEMMA 6.1. *1. If (3) is DUR and the parametrisation $(\tilde{X}, \Theta)$ is GTIP, then for all sufficiently small $\alpha > 0$,*

$$\begin{aligned} \mathbf{E}\left\{e^{\alpha X} | Y, \Theta = \theta\right\} &\leq e^{\alpha\theta}(1 - \alpha d/2), \quad \text{for } \theta > k \\ \mathbf{E}\left\{e^{-\alpha X} | Y, \Theta = \theta\right\} &\leq e^{-\alpha\theta}(1 - \alpha d/2), \quad \text{for } \theta < -k, \end{aligned}$$

where k, d are defined in Definition 3.2.

2. *If (3) is PUR and the parametrisation (X, Θ) is GTIP, then for all sufficiently small $\alpha > 0$,*

$$\begin{aligned} \mathbf{E}\left\{e^{\alpha(y-\tilde{X})}|Y=y,\Theta=\theta\right\} &\leq e^{\alpha\theta}(1-\alpha d/2), \quad \textit{for } \theta > k \\ \mathbf{E}\left\{e^{-\alpha(y-\tilde{X})}|Y=y,\Theta=\theta\right\} &\leq e^{-\alpha\theta}(1-\alpha d/2), \quad \textit{for } \theta < -k, \end{aligned}$$

PROOF. 1. We will prove only the first inequality, for $\theta > k$, since the other is proved in a similar fashion. We define $G_\theta(t) = \mathbf{E}\left\{e^{t\ (X-\theta)} \mid Y, \Theta = \theta\right\}$, which is finite for all sufficiently small $t > 0$, say $0 < t < t_0$ for some t_0, and for all θ, since by the GTIP assumption $\mathcal{L}(|X-\theta| \mid Y, \Theta = \theta)$ has exponential or lighter tails. By a second order Taylor series expansion of $G_\theta(t)$ around $t = 0$, we obtain for some $0 < t_1 < t_0$, and for $\theta > k$,

$$\begin{aligned} G_\theta(t) &= 1 + t\,\mathbf{E}\{X-\theta \mid Y,\Theta=\theta\} + \frac{t^2}{2}\,\mathbf{E}\left\{(X-\theta)^2 e^{t_1\ (X-\theta)}|Y,\Theta=\theta\right\} \\ &\leq 1 - td + \frac{t^2}{2}\mathbf{E}\left\{(X-\theta)^2 e^{t_1(X-\theta)}|Y,\Theta=\theta\right\}. \end{aligned}$$

Now pick $\alpha < t_1$ small enough so that for all $\theta > k$ $\alpha\mathbf{E}\left\{(X-\theta)^2 e^{t_1(X-\theta)}|Y,\Theta=\theta\right\} < d$. Such α exists due to the GTIP assumption. Then, $G_\theta(\alpha) \leq 1-\alpha d/2$, and the result follows. 2. It is proved as 1, recognising that $\tilde{X} = X - \theta$. $\square$

LEMMA 6.2. 1. *If (3) is DUR and the parametrisation $(\tilde{X}, \Theta)$ is GTIP, then for all sufficiently small $\alpha > 0$,*

$$\mathbf{E}\left\{e^{\alpha|X|}|Y,\Theta=\theta\right\} \leq e^{\alpha|\theta|}(1-\alpha d/2) + K, \quad \textit{for } |\theta| > k,$$

where k, d are defined in Definition 3.2, and $0 < K < \infty$.

2. *If (3) is PUR and the parametrisation (X, Θ) is GTIP, then for all sufficiently small $\alpha > 0$,*

$$\mathbf{E}\left\{e^{\alpha|y-\tilde{X}|}|Y=y,\Theta=\theta\right\} \leq e^{\alpha|\theta|}(1-\alpha d/2) + K, \quad \textit{for } |\theta| > k,$$

where k, d are defined in Definition 3.2, and $0 < K < \infty$.

PROOF. 1. We prove the result for $\theta > 0$ exploiting the first inequality given in Lemma 6.1. The case $\theta < 0$ is proved analogously but exploiting the second inequality of Lemma 6.1. Notice that

$$\mathbf{E}\left\{e^{\alpha|X|}|Y,\Theta=\theta\right\} \leq \mathbf{E}\left\{e^{\alpha X}|Y,\Theta=\theta\right\} + \int_{-\infty}^{0} e^{-\alpha x}\pi(x \mid y,\theta)dx,$$

thus, due to Lemma 6.1 we only need to show that the second term of the sum above can be bounded above for all θ. Recall a, b from the GTIP Definition 3.2. Choose $\alpha < b$. Using integration by parts, we find that the second summand is bounded above by, $e^{-b\theta}[a+\alpha/(b-\alpha)]$, which can easily be bounded above for all $\theta > k$. 2. It is proved as 1, recognising that $\tilde{X} = X - \Theta$. □

PROOF OF THEOREM 3.5 1. We prove the result establishing a geometric drift condition for the marginal chain $\{\Theta_n\}$, using the function $V(\theta) = e^{\alpha|\theta|}$, for appropriately chosen $\alpha > 0$. Notice first that $\mathcal{L}(\Theta \mid Y, X = x) \equiv \mathcal{L}(\Theta \mid X = x)$ is symmetric around x and has a finite moment generating function in a neighbourhood of the origin. Thus, working as in Lemma 6.1 and Lemma 6.2, we can show that for all sufficiently small $\alpha > 0$, there exists $K_1 > 0$ and $\epsilon > 0$, such that,

$$\mathbf{E}\{e^{\alpha|\Theta|} \mid X = x\} \leq \left(1 + \alpha^2\epsilon\right) e^{\alpha|x|} + K_1.$$

Then, for $|\theta_0| > k$, and appropriate $K_1 > 0, K > 0$,

$$\begin{aligned}
\mathbf{E}\{e^{\alpha|\Theta_1|} \mid Y, \Theta_0 = \theta_0\} &= \mathbf{E}\{\mathbf{E}\{e^{\alpha|\Theta_1|} \mid X_1\} \mid Y, \Theta_0 = \theta_0\} \\
&\leq \mathbf{E}\{(1+\alpha^2\epsilon)e^{\alpha|X_1|} + K_1 \mid Y, \Theta_0 = \theta_0\} \\
&\leq (1+\alpha^2\epsilon)(1-\alpha d/2)e^{\alpha|\theta_0|} + K \\
&\leq (1-\alpha\delta)e^{\alpha|\theta_0|} + K.
\end{aligned}$$

Now since standard arguments (see for example [25]) show that compact sets are small for this problem, the Gibbs sampler is shown to be geometrically ergodic by Theorem 15.0.1 of [19].

2. The second result is proved almost identically. Notice that $\mathcal{L}(\Theta \mid Y = y, \tilde{X} = x)$ is symmetric around $y - x$ and possesses finite moment generating

function in a neighbourhood of 0, thus as we showed above, for all sufficiently small $\alpha > 0$, there exists a $K_1 > 0$ such that,

$$\mathbf{E}\{e^{\alpha|\Theta|} \mid Y = y, \tilde{X} = x\} \leq \left(1 + \alpha^2 \epsilon\right) e^{\alpha|y-x|} + K_1.$$

Using Lemma 6.2 and arguing as in 1 proves the theorem. □

Before proving Theorems 3.4 and 3.6 we need the following general result about Markov chains on the real line.

THEOREM 6.3. *Let $\{W_n\}$ be an ergodic and reversible with respect to a density π, Markov chain on $\mathbf{R}$ with transition density $p(x, y)$ which is random walk-like in the tails, in the sense that there is a continuous positive symmetric density q such that*

$$\lim_{|x|\to\infty} p(x, x + z) = q(z), \;\; z \in \mathbf{R}. \tag{8}$$

Then

1. *π has heavy tails, in the sense that*

$$\lim_{x\to\infty} \frac{\log \int_x^\infty \pi(u)du}{x} = \lim_{x\to\infty} \frac{\log \int_{-\infty}^{-x} \pi(u)du}{-x} = 0 \; ; \tag{9}$$

2. *$\{W_n\}$ is not geometrically ergodic.*

PROOF 1. We will prove the result for $x \to \infty$, since the case $x \to -\infty$, is proved in the same way. Fix $z, \delta \in \mathbf{R}^+$, and let W denote a random variable which has density π. By (8), there exists $k > 0$ such that for $x > k$

$$\frac{p(x + z, x)}{p(x, x + z)} \leq (1 + \delta) \; .$$

This uses the fact that $q(z) > 0$. Thus by reversibility, and for $x > k$,

$$\frac{\pi(x)}{\pi(x + z)} = \frac{p(x + z, x)}{p(x, x + z)} \leq (1 + \delta) \; ,$$

so that

$$\pi(x + z) \geq (1 + \delta)^{-1} \pi(x) \; . \tag{10}$$

Integrating (10) over $x > k$, gives that

$$\mathbf{P}(W > k + z) \geq (1 + \delta)^{-1}\mathbf{P}(W > k)\ . \tag{11}$$

Iterating this expression, and after some algebra, we get that

$$\lim_{n\to\infty} \frac{\log \mathbf{P}(W > k + nz)}{n} \geq -\delta,$$

which, since δ can be chosen arbitrarily small, proves the statement.

2. The second follows from the following standard capacitance argument; see [25] for similar arguments for MCMC algorithms and [14] for an introduction to Cheeger's inequality using capacitance. Cheeger's inequality for reversible Markov chains implies that geometric ergodicity must fail if we can find $k > 0$, such that the probability

$$\mathbf{P}\left(|W_1| \leq k \mid W_0 \sim \pi_{(-k,k)^c}\right)$$

is arbitrarily small, where we use $\pi_{(-k,k)^c}$ to denote the density π restricted and re-normalised to the set $\{|x| > k\}$. Notice that (11) implies that for sufficiently large k, for $|x| > k$, and any $l > 0$, there

$$\mathbf{P}(|W_1| > x + l | W_0 > k) \geq (1 + \delta)^{-1} \geq 1 - \delta\ .$$

Now choose l sufficiently large that $\int_l^\infty q(u)du < \delta$ then for all $|x| > k$,

$$\mathbf{P}\left(|W_1| < k\right) \leq \mathbf{P}(|W_1| < k \mid W_0 \sim \pi_{(-k,k)^c}) + \mathbf{P}(|W_1 - W_0| > l)$$

which converges as $|x| \to \infty$ to a limit bounded by 3δ. Since δ is arbitrary, the result is proved. □

PROOF OF THEOREM 3.4 we prove the theorem for the case where the model is RID, since the proof when the model is RIP is identical. We will show that under the assumptions the marginal chain $\{\Theta_n\}$ generated by the centred Gibbs sampler is random walk-like, thus by Theorem 6.3 $\mathcal{P}_0$ is not geometrically ergodic. By assumption, $\lim_{|\theta|\to\infty} \mathcal{L}(\tilde{X}|Y, \Theta = \theta) = \mathcal{L}(\tilde{X})$, which is symmetric around 0, and let F denote its corresponding distribution function. Therefore $\mathbf{P}(X \leq \theta + z \mid Y, \Theta = \theta) \to F(z)$, as $|\theta| \to \infty$. Notice that,

$$p(\theta_0, \theta_0+z) = \int f_2(|x-\theta_0-z|)dF(x \mid Y, \Theta = \theta_0) = \int f_2(|u-z|)dF(u+\theta_0 \mid Y, \Theta = \theta_0),$$

therefore, since f_2 is bounded, $p(\theta_0, \theta_0 + z) \to \int f_2(|u - z|)dF(u) = q(z)$, as $|\theta_0| \to \infty$, where q is a symmetric density around 0. □

PROOF OF THEOREM 3.6 Throughout the proof we shall use the following notation: f_1 and f_2 denote the density of Z_1 and Z_2 respectively (at least up to proportionality), and we define

$$f_\theta(x) \quad = \quad f_1(|y-x|)f_2(|x-\theta|),$$

thus, $\pi(x \mid y, \theta) = f_\theta(x)/c_\theta$, where c_θ is the normalisation constant. Any scale parameter involved in f_i will be denoted by $\sigma_i,\ i = 1, 2$.

For each model, we first prove the result for $\mathcal{P}_0$ and subsequently for $\mathcal{P}_1$. We will prove the statements corresponding to the upper triangular elements of the $\mathcal{P}_0$ and $\mathcal{P}_1$ tables. This is without loss of generality, since we can write (3) as

$$\begin{aligned} \tilde{X} &= Y - \Theta - Z_1 \\ \tilde{X} &= Z_2\ . \end{aligned}$$

Since the actual value of Y does not affect convergence (as can be verified by our proofs below), we may as well set it to be 0, and since $\mathcal{L}(Z_1), \mathcal{L}(Z_2)$ are symmetric around 0, the model written above under a non-centred parametrisation coincides with (3) under a centred parametrisation but with the error distributions interchanged. We first prove the results concerning the diagonal elements.

The (C, C) model

We prove the result by verifying the PTIP property. The result then follows by Theorem 3.3. Notice that in this model, $c_\theta = \int_{-\infty}^{\infty} f_\theta(x)dx =$

$2\int_{-\infty}^{(y+\theta)/2} f_\theta(x)dx$. We show that $\mathcal{P}_0$ is PTIP by demonstrating that for arbitrary $k > 0$,

$$\liminf_{|\theta|\to\infty} \int_{y-k}^{y+k} f_\theta(x)/c_\theta \; dx > 0 \; .$$

By symmetry, it is enough to prove this statement for large positive θ values, so from now on we shall assume that $\theta > y$.

For $x < (y+\theta)/2$, $1+(y-\theta)^2 \le 1+4(x-\theta)^2 \le 4(1+(x-\theta)^2)$, so that $c_\theta \le 4/\pi(1+(y-\theta)^2)$. Moreover, notice that when $x \in (y-k, y+k)$, then there exist a $d > 0$ (depending on k, y), such that for all $\theta > d$,

$$\frac{1+(y-\theta)^2}{1+(x-\theta)^2} \ge \frac{1+(y-\theta)^2}{1+(y+k-\theta)^2} \ge 1/2.$$

Therefore, for $\theta > d$,

$$\begin{aligned}\int_{y-k}^{y+k} f_\theta(x)/c_\theta \; dx \;\; &\ge \;\; \int_{y-k}^{y+k} \frac{1+(y-\theta)^2}{4\pi(1+(y-x)^2)(1+(x-\theta)^2)} dx \\ &\ge \;\; \frac{1}{8}\int_{y-k}^{y+k} \frac{1}{\pi(1+(y-x)^2)} > 0,\end{aligned}$$

which proves the result. The result for $\mathcal{P}_1$ is proved identically.

The (E,E) model

Without loss of generality we assume that $f_1(x) \propto \exp\{-|x|\}$, and $f_2(x) \propto \exp\{-|x|/\sigma\}, \sigma > 0$. The stability of the Gibbs sampler depends on whether $\sigma < 1$, $\sigma = 1$ or $\sigma > 1$, thus we consider these cases separately. Again by symmetry it is enough to consider $y < \theta$.

1. $\sigma = 1$: here we can write

$$f_\theta(x) = \begin{cases} \frac{1}{4}e^{2x-y-\theta}, & x < y \\ \frac{1}{4}e^{-(\theta-y)}, & y \le x \le \theta \\ \frac{1}{4}e^{y+\theta-2x}, & x > \theta \, . \end{cases}$$

From this it is easy to demonstrate that $E(\Theta_1|\Theta_0 = \theta_0) = (y+\theta_0)/2$. Since all compact sets are small for the Markov chain $\{\Theta_n\}$ this is enough to demonstrate geometric ergodicity by Theorem 15.0.1 of [19].

2. $\sigma > 1$: here we can write:

$$f_\theta(x) = \begin{cases} \frac{1}{4} e^{(1+\sigma)x - y - \sigma\theta}, & x < y \\ \frac{1}{4} e^{y - \sigma\theta + (\sigma-1)x}, & y \le x \le \theta \\ \frac{1}{4} e^{y + \sigma\theta - (1+\sigma)x}, & x > \theta\,. \end{cases}$$

Direct algebra shows that

$$\mathbf{E}\{X - \theta \mid Y, \Theta = \theta\} = p_1(\theta)(Y-1) + [p_2(\theta) + p_3(\theta) - 1]\theta + p_2(\theta) r(\theta) + \frac{p_3(\theta)}{\sigma+1} - \frac{p_2(\theta)}{\sigma-1},$$

where $p_1(\theta) + p_2(\theta) + p_3(\theta) = 1$, and as $\theta \to \infty$, $p_2(\theta) \to (\sigma + 1)/(2\sigma), p_1(\theta) \to 0, r(\theta) \to 0$. Therefore,

$$\lim_{\theta\to\infty} \mathbf{E}\{X - \theta | Y, \Theta = \theta\} \le \frac{-2}{\sigma^2 - 1}\ ,$$

and the model is DUR. Since $\mathcal{P}_1$ is easily seen to be GTIP, by part 1 of Theorem 3.5, $\mathcal{P}_0$ is geometrically ergodic.

3. $\sigma < 1$: Here, in an analogous way to the above, we can demonstrate that $\mathcal{P}_0$ is RIP therefore by Theorem 3.3, $\mathcal{P}_0$ is uniformly ergodic.

Due to symmetry, the results for $\mathcal{P}_1$ are proved in a similar fashion, notice however, that $\mathcal{P}_1$ is uniformly ergodic when $\sigma > 1$.

The (G, G) model

This is covered in [21, 24] and reviewed in Section 3.1.

The (L, L) model

We assume that $f_1(x) \propto \exp\{-|x/\sigma_1|^\beta\}$, $f_2(x) \propto \exp\{-|x/\sigma_2|^\beta\}$, and we let $a = \beta/(\beta-1)$. Again by symmetry we just consider the case $y < \theta$. For large θ, $\mathcal{L}(X|Y, \Theta = \theta)$ converges weakly and in L^1 to a point mass at $\rho\theta + (1-\rho)y$ where

$$\rho = \frac{\sigma_1^{-a}}{\sigma_2^{-a} + \sigma_1^{-a}}.$$

As a result, neither $\mathcal{P}_0$ nor $\mathcal{P}_1$ are GTIP, so it is not possible to establish geometric ergodicity using the DUR and PUR properties (which hold for this model) in conjunction with Theorem 3.5. Instead, we have to construct directly a geometric drift condition. However, this is rather easy. Notice that since $\mathcal{L}(\Theta \mid X = x)$ is symmetric around x, we can find a $b > 0$ such that $\mathbf{E}\{|\Theta| \mid X = x\} \leq |x| + b$. Moreover, for any $\epsilon > 0$, there is some $k > 0$, such that for all $\theta| > k$, $\mathbf{E}\{|X - y| \mid Y = y, \Theta = \theta\} \leq (1 + \epsilon)\rho|\theta - y|$, thus

$$\mathbf{E}\{|\Theta_1 - y| \; |\Theta_0 = \theta_0\} \leq b + \rho(1 + \epsilon)|\theta_0 - y|$$

which implies geometric ergodicity for $\mathcal{P}_0$ since compact sets can easily be seen to be small. The result for $\mathcal{P}_1$ is proved identically.

The $(C, G), (E, C)$ and (L, C) models
We show that the model is RIP, therefore since $\mathcal{P}_0$ is PTIP, by Theorem 3.3 $\mathcal{P}_0$ is uniformly ergodic, and by Theorem 3.4 $\mathcal{P}_1$ is not geometrically ergodic. Notice, however, that for any x, using dominated convergence we can show that $c_\theta / f_2(|x - \theta|) \to 1$, as $|\theta| \to \infty$. The argument is that, for any u, $f_2(|u - \theta|)/f_2(|x - \theta|) \to 1$, and the ratio is bounded above (as a function of θ) by a function of u which is integrable with respect to f_1, as long as f_1 has exponential tails or lighter, which is the case in the models considered here. However, since $f_\theta / c_\theta \to f_1(|y - x|)$, and this limit is a proper density, it follows that the corresponding distribution functions converge and $\mathcal{L}(X \mid Y = y, \Theta = \theta) \to \mathcal{L}(|Z_1 - y|)$ as $|\theta| \to \infty$.

The (G, E) model
Calculations show that

$$\lim_{\theta \to \infty} \mathcal{L}(X|Y, \Theta = \theta) = N(y + \sigma_1^2/\sigma_2, \sigma_1^2), \text{ and } \lim_{\theta \to -\infty} \mathcal{L}(X|Y, \Theta = \theta) = N(y - \sigma_1^2/\sigma_2, \sigma_1^2),$$

therefore $\mathcal{P}_0$ is PTIP (but not RIP) and by Theorem 3.3 uniformly ergodic. The above result, however, shows that the model is PUR, and since all conditions of Theorem 3.5 are satisfied, $\mathcal{P}_1$ is geometrically ergodic.

The (L, E) model
The result is proved as above.

The (L, G) model

Here (perhaps surprisingly) $\mathcal{P}_0$ is not PTIP but the model is DUR and PUR, and both $\mathcal{P}_0$ and $\mathcal{P}_1$ are GTIP so that Theorem 3.5 can be applied.

$\square$

PROOF OF LEMMA 3.7 Consider the Gibbs sampler with initial value X_0 which updates (Θ, Q) first and then X. Direct calculation gives that $\mathcal{L}(Q \mid Y = y, X = x, \Theta = \theta) = \mathrm{Ga}(1, (y-x)^2/2)$, $\mathcal{L}(X \mid Y = y, \Theta = \theta, Q = q) = N(\theta/(q+1) + qy/(q+1), 1/(q+1))$, therefore $\mathcal{L}(X_1 - X_0 \mid Y = y, Q_1 = q) = N(q(y - X_0)/(q+1), 1 + 1/(q+1))$. However, since $q \to 0$ in probability, when $X_0 \to \infty$, the algorithm is random walk-like in the tails and by Theorem 6.3 fails to be geometrically ergodic. $\square$

PROOF OF THEOREM 4.1 It is easy to demonstrate that the model is RID,

$$\lim_{|\theta| \to \infty} \mathcal{L}(\tilde{\mathbf{X}} | \mathbf{Y}, \Theta = \theta) = N_p\left(\mathbf{0}, \Sigma\right).$$

Therefore, $\mathcal{P}_1$ is PTIP and by Theorem 3.3 is uniformly ergodic. Since

$$\Theta | \mathbf{X} \sim \left(\frac{\mathbf{1}\Sigma^{-1}\mathbf{X}\mathbf{1}}{\mathbf{1}\Sigma^{-1}\mathbf{1}}, \frac{1}{\mathbf{1}\Sigma^{-1}\mathbf{1}} \right)$$

this implies that for the Gibbs sampler using $\mathcal{P}_0$,

$$\lim_{|\theta_n| \to \infty} \mathcal{L}(\Theta_{n+1} - \theta_n | \Theta_n = \theta_n) = N\left(0, \frac{2}{\mathbf{1}\Sigma^{-1}\mathbf{1}}\right) \ ,$$

Therefore by Theorem 6.3, geometric ergodicity fails.

References.

[1] Yali Amit. On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.*, 38(1):82–99, 1991.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B*, 36:99–102, 1974.

[3] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43(1):1–59, 1991. With discussion and a reply by Besag.

[4] C. K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994.

[5] S. T. Boris Choy and Stephen G. Walker. The extended exponential power distribution and Bayesian robustness. *Statist. Probab. Lett.*, 65(3):227–232, 2003.

[6] O.F. Christensen, G.O. Roberts, and M. Sköld. Robust mcmc methods for spatial GLMM's. *J. Comput. Graph. Statist.*, 15:1–17, 2006.

[7] A. P. Dawid. Posterior expectations for large observations. *Biometrika*, 60:664–667, 1973.

[8] P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *J. Roy. Statist. Soc. Ser. C*, 47(3):299–350, 1998. With discussion and a reply by the authors.

[9] Peter Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

[10] Alan E. Gelfand, Sujit K. Sahu, and Bradley P. Carlin. Efficient parameterisations for normal linear mixed models. *Biometrika*, 82(3):479–488, 1995.

[11] Galin L. Jones and James P. Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.*, 16(4):312–334, 2001.

[12] Genshiro Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.*, 82(400):1032–1063, 1987. With comments and a reply by the author.

[13] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[14] G. Lawler and A. Sokal. Bounds on the l^2 spectrum for markov chains and markov processes. *Transations of the AMS*, 309:557–580, 1988.

[15] Jun S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.*, 89(427):958–966, 1994.

[16] Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[17] Jun S. Liu and Ying Nian Wu. Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.*, 94(448):1264–1274, 1999.

[18] Xiao-Li Meng and David van Dyk. The EM algorithm—an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, 59(3):511–567, 1997. With discussion and a reply by the authors.

[19] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[20] A. O'Hagan. On outlier rejection phenomena in bayes inference. *J. Roy. Statist. Soc. Ser. B*, 41:358–367, 1979.

[21] Omiros Papaspiliopoulos, Gareth O. Roberts, and Martin Sköld. Non-centered pa-

rameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 307–326. Oxford Univ. Press, New York, 2003. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

[22] L.R. Pericchi and A.F.M. Smith. Exact and approximate posterior moments for a normal location parameter. *J. Roy. Statist. Soc. Ser. B*, 54:793–804, 1992.

[23] G. O. Roberts, O. Papaspiliopoulos, and P. Dellaportas. Bayesian inference for Non-Gaussian Ornstein-Uhlenbeck Stochastic Volatility processes. *J. Roy. Statist. Soc. Ser. B*, 66:369–394, 2003.

[24] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B*, 59(2):291–317, 1997.

[25] G. O. Roberts and R. L. Tweedie. *Understanding MCMC*. Springer-Verlag, London, 2005. in preparation.

[26] Gareth O. Roberts and Jeffrey S. Rosenthal. Markov chains and de-initializing processes. *Scand. J. Statist.*, 28(3):489–504, 2001.

[27] Jeffrey S. Rosenthal. Rates of convergence for Gibbs sampling for variance component models. *Ann. Statist.*, 23(3):740–761, 1995.

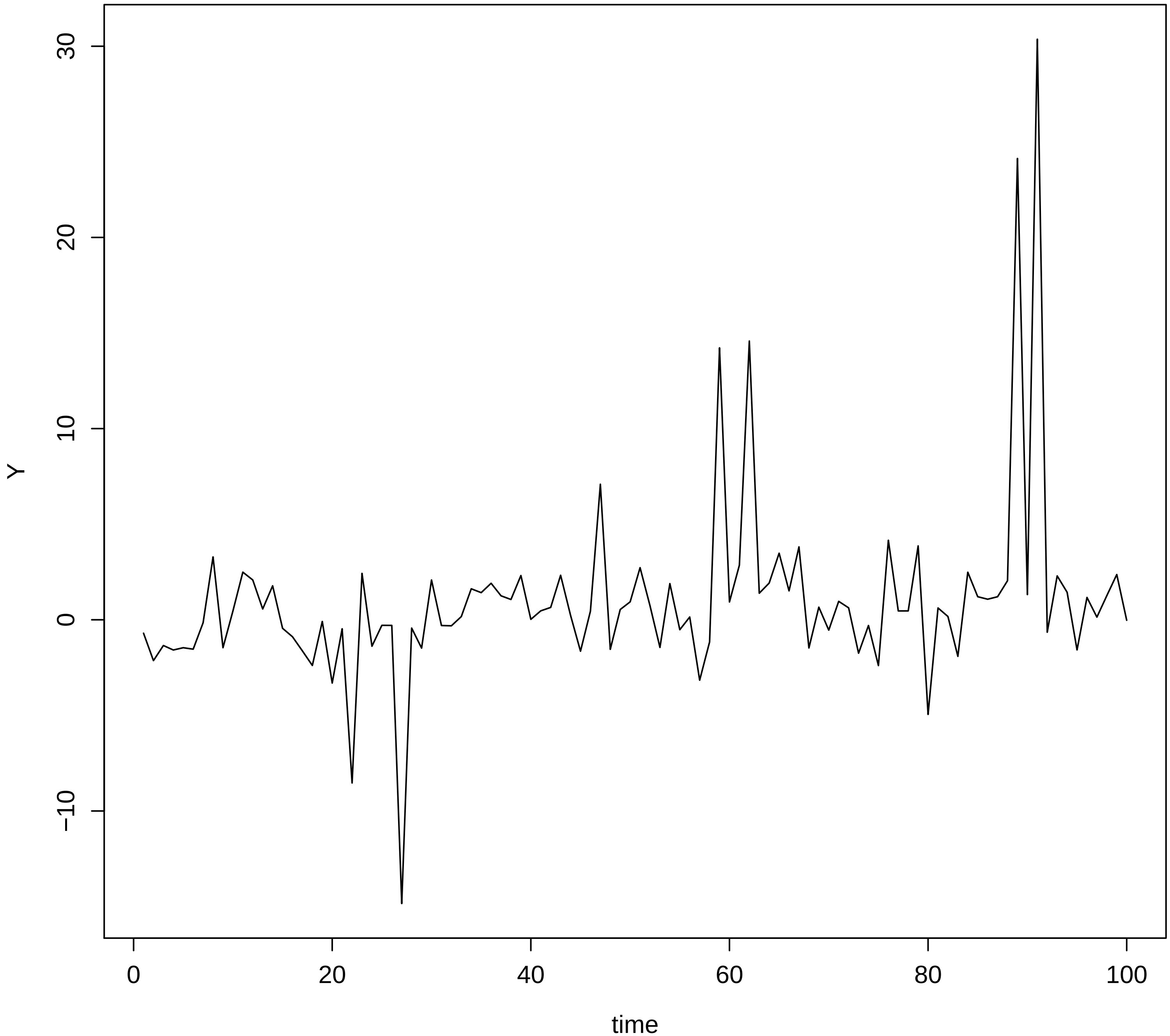
[28] Neil Shephard. Partial non-Gaussian state space. *Biometrika*, 81(1):115–131, 1994.

[29] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, 55(1):3–23, 1993.

[30] J. C. Wakefield, A. F. M. Smith, A. Racine-Poon, and A. E. Gelfand. Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *J. Roy. Statist. Soc. Ser. C*, 43:201–221, 1994.

Mathematics Institute
University of Warwick
Coventry, CV4 7AL
E-mail: o.papaspiliopoulos@warwick.ac.uk

Department of Mathematics & Statistics
Lancaster University
Lancaster, LA1 4YF
E-mail: g.o.roberts@lancaster.ac.uk


Y
time
30
20
10
0
−10
0
20
40
60
80
100