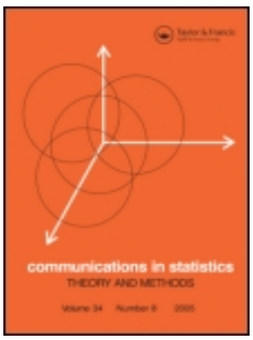

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process

P. H. Garthwaite, Y. Fan & S. A. Sisson

To cite this article: P. H. Garthwaite, Y. Fan & S. A. Sisson (2016) Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process, Communications in Statistics - Theory and Methods, 45:17, 5098-5111, DOI: 10.1080/03610926.2014.936562

To link to this article: https://doi.org/10.1080/03610926.2014.936562

© 2016 The Author(s). Published by Taylor & Francis

Published online: 05 Jul 2016.

Submit your article to this journal

Article views: 2304

View related articles

View Crossmark data

Citing articles: 18 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lsta20

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2016, VOL. 45, NO. 17, 5098–5111
http://dx.doi.org/10.1080/03610926.2014.936562

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process

P. H. Garthwaite[a], Y. Fan[b], and S. A. Sisson[b]

[a]Department of Mathematics and Statistics, Open University, Milton Keynes, UK; [b]School of Mathematics and Statistics, University of New South Wales, New South Wales, Sydney, Australia

ABSTRACT

We present an adaptive method for the automatic scaling of random-walk Metropolis–Hastings algorithms, which quickly and robustly identifies the scaling factor that yields a specified overall sampler acceptance probability. Our method relies on the use of the Robbins–Monro search process, whose performance is determined by an unknown steplength constant. Based on theoretical considerations we give a simple estimator of this constant for Gaussian proposal distributions. The effectiveness of our method is demonstrated with both simulated and real data examples.

ARTICLE HISTORY
Received 15 January 2014
Accepted 12 June 2014

KEYWORDS
Markov chain Monte Carlo; Optimal scaling; random-walk Metropolis–Hastings; Robbins–Monro.

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are routinely used in Bayesian statistical inference. In particular, the Metropolis–Hastings algorithm is highly popular due to its simplicity and general applicability (Brooks et al., 2011). The most frequently implemented variant, the random-walk Metropolis–Hastings (RWMH) sampler, uses a Gaussian proposal distribution centered on the current value of the Markov chain, with some specified scale parameter $\sigma^2 > 0$. The overall acceptance rate, and hence efficiency of RWMH, depends strongly on the value of σ^2, which typically produces a smaller acceptance probability for a proposed move when it is large, and a larger acceptance probability when it is small.

In this article, we propose the use of a stochastic search algorithm – the Robbins–Monro process (Robbins and Monro, 1951) – to automatically tune the scale parameter of a Gaussian RWMH algorithm for a prespecified value of the RWMH acceptance probability. The scale parameter is optimized to achieve a user-specified overall sampler acceptance probability. Theoretical arguments for target distributions of certain forms suggest optimal acceptance rates of 0.234 (Roberts et al., 1997) and 0.44 (Roberts and Rosenthal, 2001) for multivariate and univariate proposal distributions, respectively. In essence, the resulting adaptive sampler will increase the value of σ if the previous MCMC proposed move was likely to have been accepted and decrease σ if the proposal was likely to have been rejected. The amount by which σ is changed, termed the 'step size', decreases log-linearly with the number of iterations in the Markov chain.

CONTACT P. H. Garthwaite paul.garthwaite@open.ac.uk Department of Mathematics and Statistics, Open University, Milton Keynes MK7 6AA, UK.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

Published with license by Taylor & Francis Group, LLC © P. H. Garthwaite, Y. Fan, and S. A. Sisson
This is an Open Access article. Non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly attributed, cited, and is not altered, transformed, or built upon in any way, is permitted. The moral rights of the named author(s) have been asserted.

Within the context of the RWMH algorithm, the Robbins–Monro process guarantees the convergence of an estimate of σ for any specified acceptance probability, while other more ad-hoc search methods (e.g., Roberts and Rosenthal, 2009) do not have such convergence properties. The primary contribution of this paper is to develop a simple estimator of the optimum steplength constant of the Robbins–Monro process, based on theoretical arguments. This constant controls the magnitude of the step size and therefore the rate of convergence of the Robbins–Monro process. Andrieu and Thoms (2008) review a variety of adaptive MCMC methods, including several that use variants of the Robbins–Monro process. However, none of the algorithms they describe estimate the optimal value of the steplength constant. A previous attempt to estimate this constant (Andrieu and Robert, 2001) required combining information from three separate Markov chains, making the accuracy of the estimate questionable. This work is not mentioned in the review of Andrieu and Thoms (2008). Our proposed procedure offers a far simpler estimate for the steplength constant and requires minimal changes to the vanilla RWMH algorithm. It can therefore be easily included in existing MCMC software. For other recent work on adaptive MCMC methods, see, e.g., Roberts and Rosenthal (2009), Haario et al. (2005), and Craiu et al. (2009).

2. Robbins–Monro and estimation of the steplength constant

Consider a binary response with probability of success $p(\sigma)$, where $\sigma > 0$ is a parameter that can be controlled. It is assumed that $p(\sigma)$ is a monotonic function of σ and here it is appropriate to suppose that the function is monotonically *decreasing*. This assumption usually holds in RWMH, as a smaller scale parameter, σ, generally corresponds to a larger acceptance rate, $p(\sigma)$, and vice versa. The aim is to find the value of σ that gives a specified probability of success, p^*. Let σ^* denote this value, so that $p(\sigma^*) = p^*$. Following Vihola (2011), we conduct the search for σ^* on the logarithmic scale, so that $\theta = \ln(\sigma)$ and $\theta^* = \ln(\sigma^*)$, to constrain the scale parameter σ to positive values. The Robbins–Monro process conducts a stochastic search in which a sequence of Bernoulli trials is implemented. Let θ_i denote the estimate of θ^* at the ith trial, $i = 1, 2, \ldots$. The standard Robbins–Monro process (Robbins and Monro, 1951) updates $\theta_i \to \theta_{i+1}$ according to the rule

$$\theta_{i+1} = \begin{cases} \theta_i + c(1 - p^*)/i & \text{if the } i^{th} \text{ trial is a success} \\ \theta_i - cp^*/i & \text{if the } i^{th} \text{ trial is a failure,} \end{cases}$$

where the size of the change is controlled by a chosen *steplength* constant $c > 0$. If p_i is the probability that the ith trial is a success, then the expected size of the ith step is $c(p_i - p^*)/i$. The randomness introduced by the Bernoulli trial can then be avoided by writing

$$\theta_{i+1} = \theta_i + c(p_i - p^*)/i. \tag{1}$$

With RWMH, the Bernoulli trial consists of generating a value from the proposal distribution at the ith sampler iteration, and p_i is the corresponding acceptance probability.

The optimum choice of the steplength constant is $c^* = -1/[dp(\sigma)/d\theta]_{\theta=\theta^*}$, where the derivative is evaluated at the target value $\theta = \theta^*$ (Hodges and Lehmann, 1955). The method has good asymptotic properties (Hodges and Lehmann, 1955; Wetherill, 1963; Schwabe and Walk, 1996). In particular, as $i \to \infty$, $\theta_i - \theta^*$ is asymptotically distributed as $N(0, p^*(1 - p^*)c^2c^*/i(2c - c^*))$, provided that $c > c^*/2$. If c is set equal to its optimal value, c^*, then the asymptotic variance of θ_i equals the Cramer–Rao lower bound to the variance of any non parametric unbiased estimator of θ^* (Wetherill, 1975). Moreover, the asymptotic variance is relatively insensitive to the precise value chosen for c (Wetherill, 1963), especially if c overestimates c^* so that steps are larger than their optimal size: the variance is one-third greater than

its lower bound when $c = 2c^*$ or $c = 2c^*/3$. In general, the optimal value c^* is not known and must be estimated.

In the context of the Metropolis–Hastings algorithm, suppose that the posterior target distribution is $f(\mathbf{x}) \propto f^{\#}(\mathbf{x})$, where $f^{\#}(\cdot)$ is known. Let $g(\cdot \mid \mathbf{x}, \sigma)$ be the proposal distribution when currently at $\mathbf{x}$, and σ denote the scale parameter. Then

$$p(\mathbf{x}, \sigma) = \int \min\left\{\frac{f^{\#}(\mathbf{y})\, g(\mathbf{x} \mid \mathbf{y}, \sigma)}{f^{\#}(\mathbf{x})\, g(\mathbf{y} \mid \mathbf{x}, \sigma)},\ 1\right\} g(\mathbf{y} \mid \mathbf{x}, \sigma)\, d\mathbf{y}$$

is the probability of accepting a proposed move from $\mathbf{x}$, using proposals from $g(\cdot \mid \mathbf{x}, \sigma)$. We assume that $g(\cdot \mid \mathbf{x}, \sigma)$ is a normal distribution and that, for any $\mathbf{x}$, $p(\mathbf{x}, \sigma)$ is a monotonic decreasing function of σ. The overall acceptance probability of the sampler is given as $p(\sigma) = \int p(\mathbf{x}, \sigma) f(\mathbf{x})\, d\mathbf{x}$, and under standard regularity conditions

$$\frac{dp(\sigma)}{d\theta} = \int\int \min\left\{\frac{f^{\#}(\mathbf{y})}{f^{\#}(\mathbf{x})},\ 1\right\} \left(\frac{dg(\mathbf{y} \mid \mathbf{x}, \sigma)}{d\sigma}\right) \frac{d\sigma}{d\theta}\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x}, \tag{2}$$

where $\theta = \ln(\sigma)$. The quantity (2) evaluated at θ^* determines the optimal value of the steplength constant.

However, even in the usual Robbins–Monro context it is difficult to estimate the steplength constant from variation in $p(\sigma)$ (Wetherill, 1963; Ruppert, 1991). In the present context, estimating c^* from variation in $p(\mathbf{x}, \sigma)$ is orders of magnitude harder, as $p(\mathbf{x}, \sigma)$ is as sensitive to change in $\mathbf{x}$ as to change in σ. In the following, we develop a procedure that avoids this problem.

Garthwaite and Buckland (1992) and Garthwaite (1996) provide an algorithm for finding confidence limits in Monte Carlo tests, in which the steplength constant is not estimated through variation in $p(\sigma)$. Instead, the estimate of c^* is based on the distance between the current estimate of one endpoint of the confidence interval and the point estimate of the quantity for which the interval is required. The ratio of this distance to the optimal steplength constant, c^*, is reasonably similar across a broad range of distributions – sufficiently similar to provide an adequate estimate of the steplength constant and an efficient search algorithm; see Lee and Young (2003). For the RWMH algorithm, our results suggest that the relationship between c^* and p^* may be sufficiently similar across distributions for c^* to be adequately estimated from p^*. Propositions 1–5 motivate the choice of the estimator for c^*. Propositions 2 and 4 are proved in Appendix A.

Proposition 1. *Suppose that $g(\mathbf{y}|\mathbf{x}, \sigma)$ is an m-dimensional multivariate Gaussian proposal distribution, $\mathbf{y} \sim MVN(\mathbf{x}, \sigma^2\mathbf{A})$, where $\mathbf{A}$ does not depend on σ. Then a lower bound on c^* is $(mp^*)^{-1}$.*

Proof. Differentiating $g(\mathbf{y}|\, \mathbf{x}, \sigma)$ gives $dg(\mathbf{y}|\, \mathbf{x}, \sigma)/d\sigma = \{\sigma^{-3}(\mathbf{y} - \mathbf{x})'\mathbf{A}^{-1}(\mathbf{y} - \mathbf{x}) - m\sigma^{-1}\} g(\mathbf{y}|\, \mathbf{x}, \sigma)$. Also, $d\sigma/d\theta = \sigma$. Substituting Eq. (2) gives

$$dp(\sigma)/d\theta = -mp(\sigma) + \phi, \tag{3}$$

where

$$\phi = \sigma^{-2} \int\int \min\left\{\frac{f^{\#}(\mathbf{y})}{f^{\#}(\mathbf{x})},\ 1\right\} (\mathbf{y} - \mathbf{x})'\mathbf{A}^{-1}(\mathbf{y} - \mathbf{x})\, g(\mathbf{y} \mid \mathbf{x}, \sigma)\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x}. \tag{4}$$

Since $\phi > 0$ is positive, $dp(\sigma)/d\theta > -mp(\sigma)$. It follows that $c^* > (mp^*)^{-1}$, as $c^* = -1/[dp(\sigma)/d\theta]_{\theta=\theta^*}$ and $p(\sigma^*) = p^*$. □

Proposition 2. *Let $m = 1$ and suppose the conditions of Proposition 1 hold and that $f(\cdot)$ has finite variance. Then $c^* \to 1/p^*$ as $\sigma^* \to \infty$, where $(\sigma^*)^2\mathbf{A}$ is the variance of the proposal distribution.*

Proposition 3. *Suppose that $g(\mathbf{y}|\,\mathbf{x}, \sigma)$ is the Gaussian distribution defined in Proposition 1. Also, suppose that $p(\sigma)$ has a continuous first derivative and a finite second derivative within an interval $(0, \delta)$ and that $p(\sigma) \to 1$ as $\sigma \to 0$. Then $c^* \approx 1/(1 - p^*)$ as $p^* \to 1$.*

Proof. For small σ, $p(\sigma) \approx p(\epsilon) + \sigma\, dp/d\sigma$ for arbitrarily small $\epsilon > 0$. The result follows as $p(\sigma^*) = p^*$, $p(\epsilon) \to 1$ as $\epsilon \to 0$ and $[\sigma\, dp(\sigma)/d\sigma]_{\sigma=\sigma^*} = [dp(\sigma)/d\theta]_{\theta=\theta^*} = -1/c^*$. □

Proposition 4. *Suppose that the target distribution is unimodal and the conditions of Proposition 2 hold. Then another lower bound on c^* is $(1 - p^*)^{-1}$.*

We now determine a simple relationship between c^* and p^* that will be taken as representative of the relationship for target distributions in general. Under mild regularity conditions on the target distribution, $p^* \to 0$ only as $\sigma^* \to \infty$. Hence, when $m = 1$, Proposition 2 implies that the relationship should satisfy $c^* \to 1/p^*$ as $p^* \to 0$. Also, from Propositions 1 and 3, c^* should exceed $1/(mp^*)$ for all p^* and $c^* \to 1/(1 - p^*)$ as $p^* \to 1$. When $m = 1$, the simplest function that meets these conditions is

$$\hat{c}^* = 1/\left[p^*\left(1 - p^*\right)\right]. \tag{5}$$

We examined the true relationship between c^* and p^* for a broad range of univariate target distributions, based on a univariate $N(x, \sigma^2)$ random-walk proposal. Specifically, we considered the standard normal, t (with 5 d.f.), Cauchy, uniform, logistic, and double exponential distributions, a gamma(5, 1) and beta(3, 7) distribution, and the bimodal and trimodal normal mixtures $\frac{1}{2}[N(0, 1) + N(5, 5)]$ and $\frac{1}{3}[N(5, 1) + N(10, 2) + N(15, 3)]$. For each target distribution, and for a range of values of $p^* \in [0.05, 0.95]$, Monte Carlo methods were used to determine c^* by first solving $p(\sigma) = \int p(\mathbf{x}, \sigma) f(\mathbf{x})\, d\mathbf{x}$ for σ and then evaluating $dp/d\theta$ at that value of σ, using (2). Large sample sizes were used to ensure that Monte Carlo variability was negligible.

In Fig. 1, the dotted lines plot c^* against p^* for each of these distributions. The closeness of the 10 lines indicates that the relationship is broadly similar across these distributions, although the dashed line, corresponding to the Cauchy distribution, is slightly higher for low p^*. The thick solid line illustrates relationship (5) indicating its suitability, although many other choices would also be satisfactory and could be made without greatly affecting the performance of the method. A useful feature of (5) is that it generally yields an estimate of c^* that is a little too large rather than too small: in the Robbins–Monro context, it is better to overestimate c^* than to underestimate it. The lowest lines in Fig. 1 are the lower bounds given by Propositions 1 and 4. Together they provide a bound that can be quite tight for much of the range of p^*. They show that our choice of c^* will never be excessively large for any unimodal distribution.

In terms of multivariate distributions, we follow Roberts et al. (1997) who considered m-dimensional target distributions of the form

$$f(\mathbf{x}) = h(x_1)h(x_2)\ldots h(x_m) \tag{6}$$

for some univariate smooth density h, where $\mathbf{x} = (x_1, \ldots, x_m)'$. They showed that with an m-dimensional Gaussian proposal distribution, $\mathbf{y} \sim MVN(\mathbf{x}, \sigma^2\mathbf{I}_m)$, then $p(\sigma) = 2\Phi(-\sigma Bm^{1/2}/2)$ as $m \to \infty$, where $B > 0$ is a constant that depends on h. Roberts and

Rosenthal (2001) derive similar results for the case where the target distribution is a multivariate Gaussian with an m-dimensional multivariate Gaussian proposal distribution, $\mathbf{y} \sim MVN(\mathbf{x},\ \sigma^2\mathbf{A})$, or if $f(\mathbf{x}) = \prod_{i=1}^{m} C_i h(C_i x_i)$, where the $\{C_i\}$ are i.i.d draws from some fixed distribution. They show that as $m \to \infty$, then $p(\sigma) \to 2\Phi(-\beta\sigma)$ for some $\beta > 0$. The following proposition gives c^* whenever $p(\sigma)$ has this form. An attraction of the result is that c^* does not depend on β.

Proposition 5. *Suppose that $p(\sigma) = 2\Phi(-\beta\sigma)$, where $\beta > 0$ is a constant and Φ is the cdf of the standard normal distribution. Denoting $\alpha = -\Phi^{-1}(p^*/2)$, then*

$$c^* = (2\pi)^{1/2} \exp\left(\alpha^2/2\right)/(2\alpha).$$

Proof. Differentiating $p(\sigma) = 2\int_{-\infty}^{-\beta\sigma} (2\pi)^{-1/2} \exp(-z^2/2)dz$ gives $dp(\sigma)/d\theta = \sigma\, dp(\sigma)/d\sigma = -2\beta\sigma(2\pi)^{-1/2}\exp\{-(\beta\sigma)^2/2\}$. Write $\alpha = \beta\sigma^*$, so that $\alpha = -\Phi^{-1}(p^*/2)$. The proposition follows as $c^* = -1/[dp(\sigma)/d\theta]_{\theta=\theta^*}$. □

To determine how c^* varies with the dimension of multivariate target distributions, m, we examined target distributions of the form (6), and m-dimensional multivariate-$t(\nu)$ distributions. For fixed p^*, and with an m-dimensional Gaussian proposal distribution, $\mathbf{y} \sim N(\mathbf{x}, \sigma^2\mathbf{I}_m)$, experimentation indicated that c^* was close to a linear function of $1/m$. Assuming that Proposition 5 holds as $m \to \infty$, and that (5) holds when $m = 1$, then this suggests the linear function

$$\hat{c}^* = \left(1 - \frac{1}{m}\right)\frac{(2\pi)^{1/2}e^{\alpha^2/2}}{2\alpha} + \frac{1}{mp^*(1-p^*)}, \tag{7}$$

where $\alpha = -\Phi^{-1}(p^*/2)$.

To examine the utility of (7) for target distributions of the form $f(\mathbf{x}) = \prod_{i=1}^{m} h(x_i)$, each univariate distribution considered in Fig. 1 was taken in turn as the component distribution $h(\cdot)$. As before, Monte Carlo methods were used to determine c^* for a range of values of $p^* \in [0.05, 0.95]$, and the resulting true relationships are shown in Fig. 2 (panels (a)–(d))

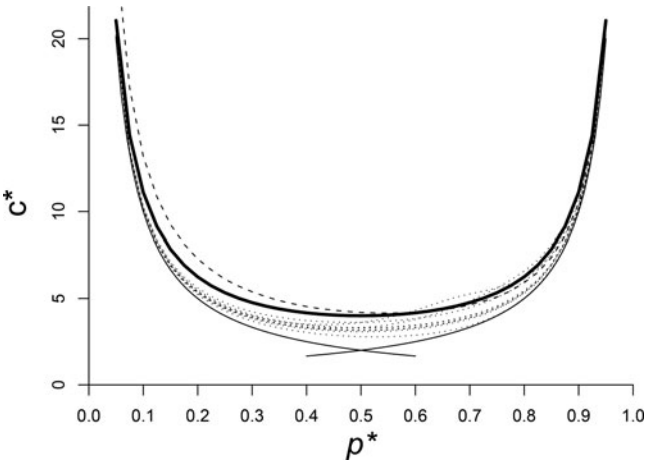

Figure 1. Plots of c^* against p^* for 10 univariate distributions: (dotted lines) standard normal, t (with 5 d.f.), uniform, logistic, and double exponential distributions, a gamma (5, 1) and beta (3, 7) distribution, bimodal and tri-modal normal mixtures, and (dashed line) a standard Cauchy. The thick solid line denotes the relationship $\hat{c}^* = 1/[p^*(1-p^*)]$. The lowest two lines that intersect at $p^* = 0.5$ are the lower bounds $1/p^*$ and $1/(1-p^*)$.

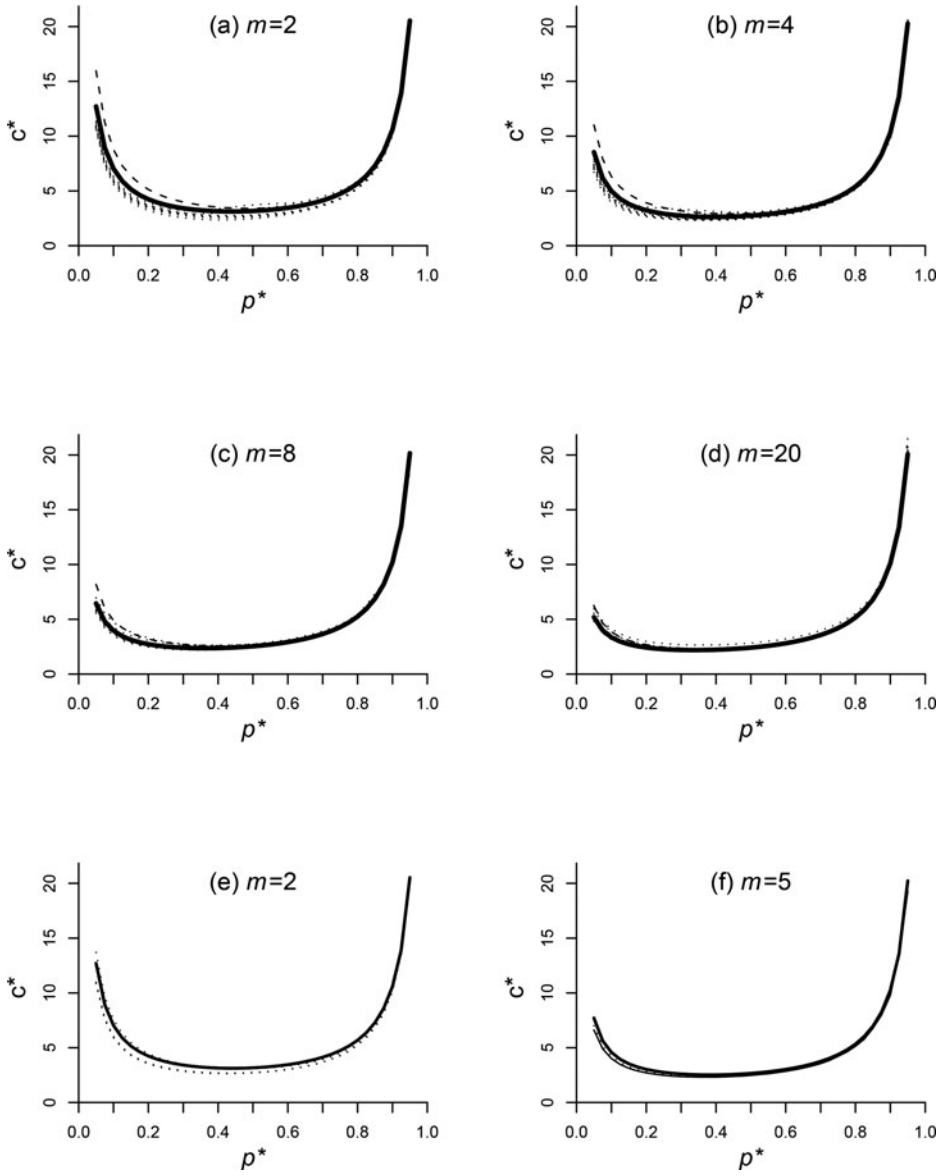

Figure 2. Plots of c^* against p^* for m-dimensional multivariate distributions. Panels (a)–(d) represent distributions of the form $f(\mathbf{x}) = \prod_{i=1}^{m} h(x_i)$, where $h(\cdot)$ is given by the 10 univariate distributions in Figure 1 (dotted lines). The dashed line denotes the standard Cauchy distribution. The solid lines denote the relationship given by (7). Panels (a)–(d) represent $m = 2$, 4, 8, and 20 dimensions. Panels (e) and (f) represent m-dimensional $t(\nu)$ distributions with ν degrees of freedom. Panel (e) shows $m = 2$ with $\nu = 2$, 10, and 25 and panel (f) shows $m = 5$ with $\nu = 5$, 10, and 25.

for $m = 2$, 4, 8, and 20 dimensions. Similarly, Fig. 2 (panels (e) and (f)) examines the relationship between c^* and p^* for multivariate $t(\nu)$ distributions with $\nu = 2$, 10, 25 (for $m = 2$) and $\nu = 5$, 10, 25 (for $m = 5$). For each model dimension, the relationship between c^* and p^* is similar for all distributions. Function (7) (solid line) exhibits a strong similarity with the other curves, implying it models the relationship across model dimensions well for these models. As with (5), the form of (7) generally represents a practically convenient overestimate of c^*.

In summary, Figs. 1 and 2 show that c^* is largely determined by the values of p^* and m for a range of distributions. Propositions 1–5 suggest that this should hold more generally. Thus, (7) gives a good estimate of c^* to be used in an implementation of the Robbins–Monro process within a Gaussian RWMH sampler. If $\widehat{\sigma}_c$ is the estimate of σ^* after i steps of a Robbins–Monro search with a steplength constant of c, then $\text{Var}(\widehat{\sigma}_c) = p^*(1-p^*)c^2/\{i(2c/c^* - 1)\}$ (Hodges and Lehmann, 1955). This is minimized when $c = c^*$ so the efficiency of a search is defined to be

$$100\% \times \text{Var}(\widehat{\sigma}_{c^*})/\text{Var}(\widehat{\sigma}_c) = 100\% \times (2c - c^*)c^*/c^2.$$

In Fig. 1, the solid line shows the value of $c = \hat{c}^*$ given by (5) for any value of p^* and the dotted lines show the values of c^* for each of 10 distributions. For these distributions, the efficiency of the search is at least 96% when $p^* = 0.234$ and at least 91% when $p^* = 0.440$.

3. Search algorithms for optimal scaling

For the univariate target distribution $f(x) \propto f^{\#}(x)$, under a Gaussian RWMH sampler with proposal distribution $y \sim N(x, \sigma^2)$, we suppose that $p^* = 0.44$ is appropriate (Roberts and Rosenthal, 2001). Similarly, for an m-dimensional multivariate target distribution, $f(\mathbf{x}) \propto f^{\#}(\mathbf{x})$, where all components of $\mathbf{x}$ are updated simultaneously using a Gaussian random-walk proposal $MVN(\mathbf{x}, \sigma^2\mathbf{A})$, for some positive-definite matrix $\mathbf{A}$, we suppose that $p^* = 0.234$ (Roberts et al., 1997). In order to find a value of σ that gives an overall sampler acceptance probability of p^*, we use the Robbins–Monro search process to improve the estimate of $\theta^* = \ln(\sigma^*)$ after each iteration of the Markov chain. If θ_i denotes the estimate of θ^* after the ith step of the search, we determine p_i, the probability that the proposed move was accepted. The value of θ_i is then updated by (1), where c is given by estimate (7).

Starting values for a search can be arbitrary, such as $\theta_1 = 0$, or more considered, such as one based on an estimated standard deviation of $f(\cdot)$. Either way, θ_1 need not be well chosen, as the Robbins–Monro process can be monitored and a search restarted if the starting value seems poor (e.g., Garthwaite, 1996; Matsui and Ohashi, 1999). On a restart, the most recent estimate of θ^* is taken as the starting value and the value of i is reset. Note that we start/restart a search with $i = n_0$, where n_0 is a moderate size so as to avoid too rapid steplength changes in the early stages of a search. We choose n_0 to be the integer closest to $5/\{p^*(1-p^*)\}$, which works well in practice. We also restart the search if the estimate of θ^* changes by $\ln(3)$ from its value from when the search started, or last restarted, so the search restarts if the estimate of σ changes by more than a factor of 3. Many other criteria for restarts would also be suitable, as the only requirement is to restart if a poor starting value was used.

It seems conceivable that a search might oscillate between σ tripling in value and reducing in value by two-thirds, so that the search is continually restarting. Our implementation includes simple safeguards to limit the number of restarts, though they have never been needed. Together with a finite number of restarts, the search algorithm satisfies the *diminishing adaptations* criterion (Roberts and Rosenthal, 2009; Rosenthal, 2011), in that changes to the scale parameter vanish as the length of the Markov chain goes to infinity. Ergodicity of the Markov chain is preserved if an adaptive scheme also satisfies the *containment* (or bounded convergence) condition (Bai et al., 2008; Roberts and Rosenthal, 2009). Containment typically holds under very general conditions for all but pathological counterexamples (Bai et al., 2008; Rosenthal, 2011). Readers uncomfortable with these conditions could simply stop adapting after a fixed, finite amount of time. See Vihola (2011) for further discussion on convergence of related samplers.

Table 1. Sampler performance based on 200 chain replicates of length 2,000 iterations, under a target acceptance probability of $p^* = 0.44$, for each of the 10 univariate target distributions in Fig. 1. Optimal values of σ^*, and 0.05, 0.5, and 0.95 quantiles of the empirical distribution of the estimates of σ^* are given, together with quantiles of the sampler acceptance rates, based on the last 1,000 iterations.

		Estimated σ^* quantile			Acceptance rate quantile		
Target distribution	Optimal σ^*	0.05	median	0.95	0.05	median	0.95
$N(0, 1)$	2.42	2.32	2.43	2.56	0.413	0.436	0.465
$t(5)$	2.71	2.58	2.72	2.84	0.411	0.437	0.465
Cauchy	4.39	3.82	4.25	5.00	0.391	0.443	0.492
Logistic	4.05	3.90	4.06	4.22	0.416	0.440	0.464
Double exponential	2.70	2.59	2.72	2.88	0.409	0.437	0.465
Gamma(5,1)	4.98	4.76	4.98	5.22	0.415	0.441	0.463
Beta(3,7)	0.335	0.321	0.338	0.355	0.412	0.437	0.461
Uniform	0.806	0.756	0.813	0.854	0.412	0.435	0.461
Bimodal normal mixture	6.07	5.674	6.070	6.412	0.413	0.440	0.467
Tri-modal normal mixture	7.86	8.157	8.671	9.157	0.416	0.443	0.470

3.1. Example: Univariate RWMH updates

The above search algorithm was applied in turn to each of the ten univariate target distributions considered earlier. Two hundred samplers of length 2,000 iterations were run for each target distribution, and a search for the value of σ^* that gave an overall acceptance probability of $p^* = 0.44$ was conducted within each chain. Each search was initialized by randomly setting $\sigma_1 \sim \text{Exp}(1)$ and $\theta_1 = \ln(\sigma_1)$. The final estimate of σ^* and the acceptance rate of the sampler over the last 1,000 iterations was recorded for each chain. The results are summarized in Table 1.

For each target distribution, the second column in Table 1 provides the theoretical value of σ^* and the next three columns give the 0.05, 0.50 and 0.95 quantiles of the final estimate of σ^* from each search. The last three columns present the same quantiles of the acceptance rates in the final 1,000 steps of each search. The results indicate that the Robbins–Monro search has low bias and good accuracy: the median estimate of σ^* is close to σ^* for each of the ten distributions, the median values of the sampler acceptance rate are close to their target of 0.44, and the 0.05 and 0.95 quantiles for both σ^* and the acceptance rate are close together. This performance exceeds practical requirements, as the efficiency of the Metropolis–Hastings algorithm is not sensitive to the precise value of p^*.

To demonstrate that the search procedure is relatively insensitive to a poor starting value for θ_1, Fig. 3 illustrates the search with a gamma(5,1) target distribution when the starting value is either much too large or much too small. The position of each restart is marked with a cross, and it can be seen that the restarts yield fast convergence. Each path is close to its optimal value of θ within 500 iterations.

3.2. Example: Multivariate RWMH updates

We follow the example of Roberts and Rosenthal (2009) in which the target distribution is $f(\mathbf{x}) = MVN(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{M}\mathbf{M}'$ and $\mathbf{M}$ is an $m \times m$ matrix whose elements are generated randomly from a $N(0, 1)$ distribution. As typically $\boldsymbol{\Sigma}$ will be close to singular, we increase each diagonal element by 1%. In many circumstances the convergence rate of the Markov chain can only be optimized if $\mathbf{A}$ is proportional to $\text{Cov}(x)$. Although $\boldsymbol{\Sigma}$ is typically unknown, its value may be estimated as the Markov chain runs (Craiu et al., 2009).

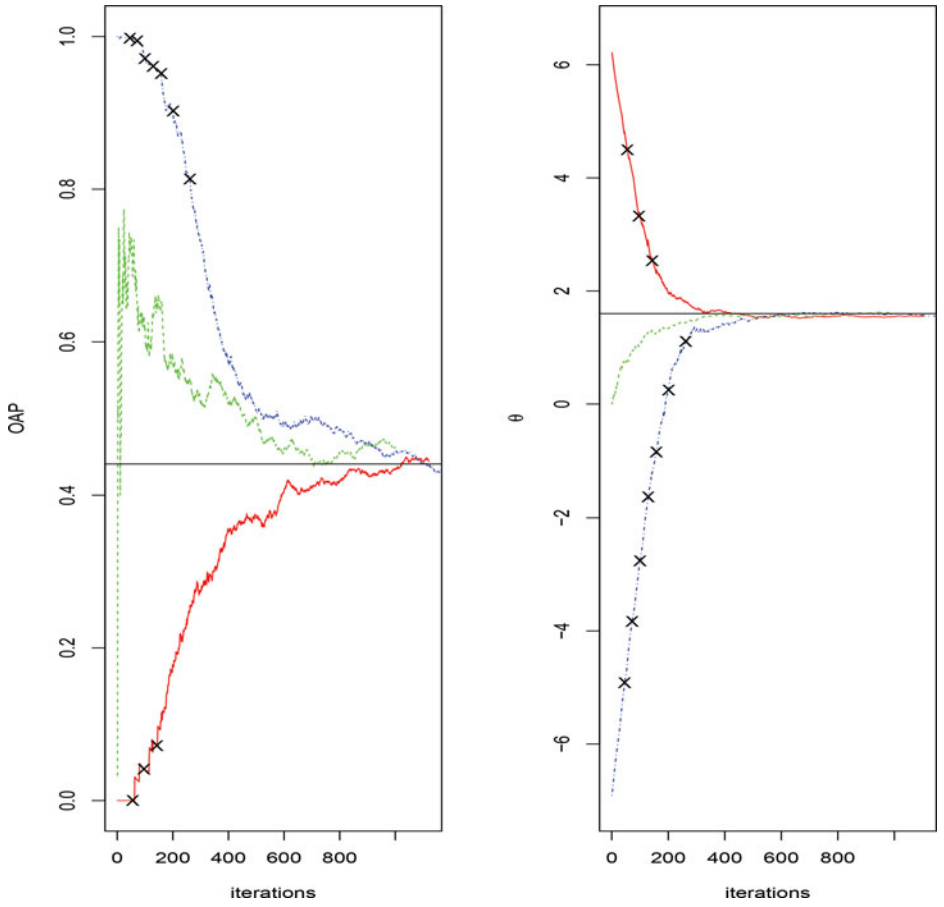

Figure 3. Trace plots of acceptance rate, in the left panel, and $\theta = \ln(\sigma)$, in the right panel, for starting points of $\theta = \ln(0.001), \ln(1), \ln(500)$, for a gamma(5,1) target distribution. Target values are indicated by horizontal line. Crosses indicate search restarts.

Here we use an adaptive MCMC strategy (Roberts and Rosenthal, 2009; Rosenthal, 2011) whereby after each iteration, $\mathbf{A}$ is set equal to the current sample estimate of $\boldsymbol{\Sigma}$, given by

$$\widehat{\boldsymbol{\Sigma}}_i = \begin{cases} \mathbf{I}_m, & i \leq 100 \\ \frac{1}{i-1}\sum_{j=1}^{i}(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)' & i > 100. \end{cases} \tag{8}$$

We follow Haario et al. (2001) and use $\widehat{\boldsymbol{\Sigma}}_i + \epsilon \mathbf{I}_m$ (where $\epsilon > 0$) as a positive-definite estimate of $\boldsymbol{\Sigma}$. Specifically, we use $\epsilon = \sigma_i^2/i$ so that $\mathbf{A}_i = \widehat{\boldsymbol{\Sigma}}_i + \sigma_i^2\mathbf{I}_m/i$ is the estimate of $\boldsymbol{\Sigma}$. Thus, we use the proposal distribution $MVN(\mathbf{x}_{i-1},\ \sigma_i^2\mathbf{A}_i)$ after i steps of the Robbins–Monro search.

When the dimension of $\mathbf{x}$ is large, a substantial number of sampler iterations may be needed before the estimate of $\boldsymbol{\Sigma}$ stabilizes (Roberts and Rosenthal, 2009). For the Robbins–Monro process to achieve the target sampler overall acceptance probability, p^*, it must not converge before the estimate of $\boldsymbol{\Sigma}$ stabilizes. To achieve this, we deliberately slow down the convergence of the Robbins–Monro search by replacing i with $\max\{200, i/m\}$ in the denominator in (1). Alternative heuristics would also be satisfactory. Trace plots of $\widehat{\boldsymbol{\Sigma}}_i$ can be monitored to assess their stability and speed of convergence.

We consider three versions of the RWMH sampler. The first (the *RM method*) follows the above, where σ_i is estimated using the Robbins–Monro procedure. The second sampler (the

Table 2. RWMH sampler performance summaries using a multivariate normal proposal. Values correspond to means and standard errors of each quantity over the last 50,000 chain iterations, based on 10 sampler replicates. Columns denote σ^2 (the mean value of σ_i^2; optimal value is 0.1133); the overall acceptance probability (OAP); posterior mean and standard deviation; the integrated autocorrelation time (ACT); and the average squared jumping distances (ASD) for the parameter x_1.

			Statistics for x_1			
	σ^2	OAP	Mean	sd	ACT	ASD
RM method	0.114 (0.01)	0.233 (0.0002)	0.08 (0.24)	7.42 (0.20)	78.24 (1.00)	1.07 (0.08)
Optimal $\boldsymbol{\Sigma}$	0.113 (–)	0.239 (0.001)	− 0.15 (0.38)	7.62 (0.15)	75.37 (0.71)	1.47 (0.02)
Fixed-scaling	0.113 (–)	0.235 (0.003)	0.08 (0.43)	7.37 (0.20)	78.56 (1.13)	1.07 (0.07)

Optimal method) implements a RWMH sampler with the theoretically optimal (fixed) proposal distribution $MVN(\mathbf{x}_{i-1}, 2.38^2\boldsymbol{\Sigma}/m)$ (Roberts et al., 1997). This sampler is unavailable in practice, as the true value of $\boldsymbol{\Sigma}$ is typically unknown, but its performance provides a benchmark for the other methods. The final sampler (the *fixed-scaling* method) is the same as the RM method, but with $\sigma_i^2 = 2.38^2/m$ fixed at its optimal value.

Setting $m = 50$ and $p^* = 0.234$, we implement ten replicate samplers, each of length 100,000 iterations, for each of the sampler variants. Results are based on discarding the first half of a chain as burn-in. As a measure of algorithm efficiency, we follow Roberts and Rosenthal (2009) in monitoring the integrated auto-correlation time (ACT) (e.g., Roberts and Rosenthal, 2001). We also monitor the average squared jumping distance (ASD) between the iterates of the chain. A smaller value of the ACT indicates less auto-correlation, and hence greater efficiency. Similarly, the larger the jumping distance, the faster the mixing of the chain. All ACT and ASD values are calculated using full length of the chain (including the burn-in period).

A summary of the results is provided in Table 2, which includes specific results for the first coordinate, x_1. The RM method consistently estimates σ^2 with good accuracy and the OAPs for all methods are close to the target value of 0.234. The optimal method benefits from the unrealistic advantage of knowing Σ, and it has a noticeably better ASD than the other methods. However, all three methods give good estimates of the mean and standard deviation of x_1 (whose true values are 0 and 7.48, respectively).

3.3. Metropolis–Hastings within Gibbs: respiratory infection in children

To illustrate a multivariate application, we model respiratory infection in Indonesian children (Diggle et al., 1995; Lin and Carroll, 2001). The data contain longitudinal measurements on 275 children, with a binary indicator for respiratory infection. Covariates include age, height, indicators for vitamin A deficiency, gender, stunting, and visit numbers. Following Zhao et al. (2006) and Fan et al. (2008), we use a Bayesian logistic additive mixed model of the form

$$\text{logit}\left\{P\left(\text{respiratory infection}_{ij} = 1\right)\right\} = \beta_0 + U_i + \boldsymbol{\beta}'\mathbf{X}_{ij} + f\left(\text{age}_{ij}\right)$$

for $1 \le i \le 275$ children and $1 \le j \le n_i$ repeated measures within a child. The random child effect is $U_i \overset{\text{ind.}}{\sim} N(0, \sigma_U^2)$, $\mathbf{X}_{ij}$ is the measurement vector of the 11 covariates,

$$f(\text{age}) = \beta_0 + \beta_1 \text{age} + \mathbf{Z}_{\text{age}}\mathbf{u} \quad \text{with} \quad \mathbf{Z}_{\text{age}} = \left[\left|\underset{1\le\kappa\le K}{\text{age} - \kappa_k}\right|^3\right]\left[\left|\underset{1\le k,k'\le K}{\kappa_k - \kappa_{k'}}\right|^3\right]^{-1/2},$$

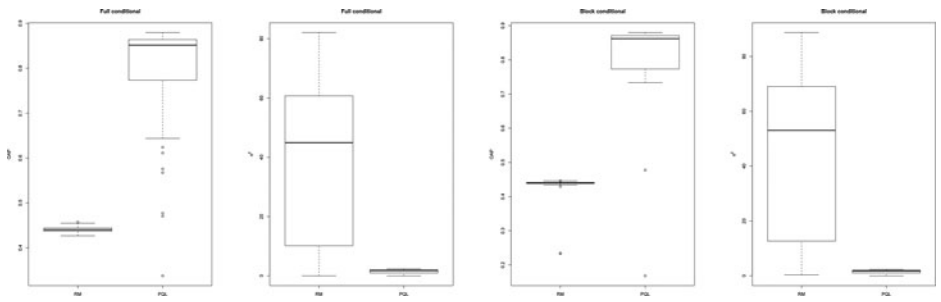

Figure 4. Boxplots of overall acceptance rates and mean σ^2 values for all univariate Robbins–Monro (RM) and penalized quasi-likelihood (PQL)-based searches, based on the second half of the sampler output. Left and right panels, respectively, indicate full-conditional and block-conditional samplers.

where $\mathbf{u} \sim N(0, \sigma_u^2\mathbf{I})$, κ_k is the $(k+1)/(K+2)$th quantile of the unique predictor values, and $K = 20$. All remaining model specifications follow Zhao et al. (2006) and Fan et al. (2008).

We consider two Metropolis–Hastings within Gibbs samplers. The first implements separate univariate Robbins–Monro searches for each of the $t = 1, \ldots 306$ parameters based on independent RWMH updates with proposal distributions $N(x_{t,i-1}, \sigma_t^2)$, and $p^* = 0.44$. The second scheme additionally block-updates the 11 $\boldsymbol{\beta}$ parameters and the 20 knot locations $\{\kappa_k\}$ with $MVN(\mathbf{x}_{i-1}^{(\beta)}, \sigma_{(\beta)}^2\mathbf{A}^{\beta})$ and $MVN(\mathbf{x}_{i-1}^{(\kappa)}, \sigma_{(\kappa)}^2\mathbf{A}^{\kappa})$ proposals, respectively, where $p^* = 0.234$ in each case. In both schemes, σ_U^s and σ_u^2 are updated using Gibbs updates. Fan et al. (2008) used penalized quasi-likelihood (PQL) (Breslow and Lin, 1995) to obtain an approximate maximum-likelihood estimate of the covariance matrix of all 306 parameters for the above model. To provide comparison with the Robbins–Monro method, we re-implement the previous two samplers by fixing the σ_t^2 as the diagonal elements of the PQL matrix, replacing $\mathbf{A}^{\beta}$ and $\mathbf{A}^{\kappa}$ by the appropriate blocks of the PQL matrix, and optimally setting $\sigma_{(\beta)}^2 = 0.51$ and $\sigma_{(\kappa)}^2 = 0.28$. Chains of length 10,000 and 50,000 were used for the full- and block-conditional samplers, respectively.

Results for all univariate proposal distributions are summarized in Fig. 4, which illustrates overall acceptance probabilities and the final mean estimates of $(\sigma^*)^2$ based on the second half of each chain, for each of the 306 or 275 parameters. Results for the full-conditional sampler and block-conditional sampler are given in the left and right panels, respectively. Within each panel, the left boxplot displays results for the Robbins–Monro method, and the right boxplot the PQL matrix approach.

Both Robbins–Monro samplers led to acceptance rates that were very close to the target of 0.44, ranging from 0.427 to 0.457 and 0.429 to 0.446 for the full- and block-conditional samplers, respectively. To achieve these rates, the search varied σ_t^2 substantially: the mean values of σ_t^2 ranged from 0.002 to 82.08 in the full-conditional sampler. In contrast, the PQL-based samplers produced more variable acceptance rates. Ranging from 0.337 to 0.880, these rates were frequently and substantially different from the ideal target. The Robbins–Monro searches were also effective with the multivariate proposals. With a target of $p^* = 0.234$, the block updates achieved an acceptance rate of 0.233 for the 11 $\boldsymbol{\beta}$ coefficients and 0.234 for the block of 20 knots $\{\kappa_k\}$.

4. Discussion

This paper provides new results on the optimal steplength constant, c^*, of the Robbins–Monro process, when used in the context of automatically determining the scaling factor of the

RWMH algorithm. Figures 1 and 2 demonstrate that the true values of c^* are remarkably similar for a wide range of distributions. Hence, function (7), which is within the span of these optimum values, will work well in practice. While (7) will not be optimal for all target distributions, Propositions 1–5 and our examples suggest that this choice of c^* should work well in many applications.

Our adopted rules regarding restarts (Section 3) are obviously arbitrary to a degree, and alternative choices may also work well. However, there is no clear optimal restart strategy, as any such approach would vary with the (unknown) accuracy of the starting value for σ: a very inaccurate starting value would benefit from a procedure that restarts quickly, whereas an accurate starting value would benefit from a procedure that rarely restarts unnecessarily. Our choices aim to balance these ideas, while excluding the possibility of infinite restarts.

From a practical perspective, in our experience, the performance of the RWMH algorithm using a sub-optimal steplength constant in the Robbins–Monro process will typically provide acceptable results, provided that the steplength constant is in the "ball park" of the optimal value, c^*. From this perspective, the precision obtained through estimates (5) and (7) is probably greater than that required in terms of efficiency gain in a typical RWMH sampler, particularly for low-dimensional problems. However, problems in higher dimensions require larger numbers of random-walk block-update steps, and hence a larger number of scale parameters to specify – see our respiratory infection data analysis in Section 3.3. In these settings, the cumulative impact of multiple sub-optimal steplength constants will impact on the performance of the sampler. Good estimates of the optimal steplength constants are accordingly invaluable for general algorithm implementation.

Code in the *R* programming language for the examples in this paper is available at `http://www.maths.unsw.edu.au/~yanan/RM.html`.

Appendix A. Proof of propositions 2 and 4

Proof of Proposition 2. $\int \min(f^{\#}(\mathbf{y})/f^{\#}(\mathbf{x}),\ 1)\, g(\mathbf{y}\,|\,\mathbf{x},\sigma)\, d\mathbf{y} \le 1$, so

$$\int\int \min\left\{\frac{f^{\#}(\mathbf{y})}{f^{\#}(\mathbf{x})},\ 1\right\} \mathbf{x}'\mathbf{x}\, g(\mathbf{y}\,|\,\mathbf{x},\sigma)\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x} \le \int \mathbf{x}'\mathbf{x}\, f(\mathbf{x})\, d\mathbf{x}. \tag{A.1}$$

As $f(.)$ has finite variance, both sides of (A.1) are finite. Similarly, as $g(\mathbf{y}\,|\,\mathbf{x},\sigma) = g(\mathbf{x}\,|\,\mathbf{y},\sigma)$, we have that $\int\int \min(f^{\#}(\mathbf{y})/f^{\#}(\mathbf{x}),\ 1)\, \mathbf{y}'\mathbf{y}\, g(\mathbf{y}\,|\,\mathbf{x},\sigma)\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x}$ is also finite. It follows that ϕ in equation (4) is $O(\sigma^{-2})$.

Let $S(r)$ be an m-dimensional sphere of radius r, centered at the mean of $f(\,.\,)$. Let $S^c(r) = \Omega - S(r)$ denote its complement. Given any ϵ, choose r such that $\int_{S^c(r)} f(\mathbf{y})\, d\mathbf{y} < \epsilon$. Then $mp(\sigma) \approx \int_\Omega \int_{S(r)} \min(f^{\#}(\mathbf{y})/f^{\#}(\mathbf{x}),\ 1)\, m\, g(\mathbf{y}\,|\,\mathbf{x},\sigma)\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x}$ and

$$\lim_{\sigma\to\infty} mp(\sigma) \approx \int_\Omega \int_{S(r)} \min\left\{\frac{f^{\#}(\mathbf{y})}{f^{\#}(\mathbf{x})},\ 1\right\} m\,(2\pi)^{-m/2}\sigma^{-m}\,|\mathbf{A}|^{-1/2}\, f(\mathbf{x})\, d\mathbf{y}\, d\mathbf{x},$$

since $\lim_{\sigma\to\infty} g(\mathbf{y}\,|\,\mathbf{x},\sigma) \to (2\pi)^{-m/2}\sigma^{-m}\,|\mathbf{A}|^{-1/2}$ for $\mathbf{y}\in S(r)$. As $mp(\sigma)$ is non zero for finite σ, $mp(\sigma)$ is $O(\sigma^{-m})$. Hence, $\lim_{\sigma\to\infty}\{\phi - mp(\sigma)\} = \lim_{\sigma\to\infty}\{-mp(\sigma)\}$ if $m = 1$, since ϕ is $O(\sigma^{-2})$. Then from (3), $\lim_{\sigma\to\infty} dp(\sigma)/d\theta = -mp(\sigma)$ and the proposition follows from $c^* = -1/[dp(\sigma)/d\theta]_{\theta=\theta^*}$. □

Proof of Proposition 4. Making the transformation $y = x + \sigma z$, $p(\sigma)$ may be written as

$$p(\sigma) = \int\int \min\left\{\frac{f^{\#}(x+\sigma z)}{f^{\#}(x)},\ 1\right\} (2\pi)^{-1/2} \exp\left(-z^2/2\right) f(x)\, dx\, dz.$$

Given σ, put $h(z) = 1 - \int_{x=-\infty}^{\infty} \min\{\frac{f^{\#}(x+\sigma z)}{f^{\#}(x)}, 1\}\, f(x)\, dx$. For any $z > 0$, there is a point $a(z)$ such that $f(x + \sigma z) < f(x)$ if and only if $x < a(z)$. The mode of $f(x)$ is between $a(z)$ and $a(z) + \sigma z$. We have

$$h(z) = 1 - \int_{-\infty}^{a(z)} f(x)\, dx - \int_{a(z)+\sigma z}^{\infty} f(x)\, dx = \int_{a(z)}^{a(z)+\sigma z} f(x)\, dx \tag{A.2}$$

and $h(z) = h(-z)$. From (A.2), $\{h(z)/\,|z|\}$ is a monotonic non increasing function of $|z|$ as $f(x)$ is unimodal. Consequently,

$$\frac{\int_{-\infty}^{\infty} z^2\, h(z)\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz}{\int_{-\infty}^{\infty} h(z)\ (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz} \le \frac{\int_{-\infty}^{\infty} z^2\, |z|\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz}{\int_{-\infty}^{\infty} |z|\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz}.$$

As $\int_{-\infty}^{\infty} z^2\, |z|\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz = 4$ and $\int_{-\infty}^{\infty} |z|\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz = 2$, we have

$$\int_{-\infty}^{\infty} z^2\, h(z)\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz \le 2 \int_{-\infty}^{\infty} h(z)\, (2\pi)^{-1/2} \exp\left(-z^2/2\right) dz. \tag{A.3}$$

The left-hand side of (A.3) is $1 - \phi$ and the right-hand side is $2(1 - p(\sigma))$. Thus, $\phi \ge (2p(\sigma) - 1)/\sigma$ and, from (3), $dp(\sigma)/d\theta \ge -(1 - p(\sigma))$. □

Funding

This work was supported by a project grant from the Medical Research Council, UK. YF and SAS are supported by the Australian Research Council under the Discovery Project scheme (DP0877432).

References

Andrieu, C., Robert, C.P. (2001). Controlled MCMC for optimal sampling. Technical report, University of Paris-Dauphine.

Andrieu, C., Thoms, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* 18:343–373.

Bai, Y., Roberts, G.O., Rosenthal, J.S. (2008). On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances in Applied Statistics* 21:1–54.

Breslow, N.E., Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 82:81–91.

Brooks, S.P., Gelman, A., Jones, G.L., Meng, X.-L. eds. (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC Press.

Craiu, R.V., Rosenthal, J.S., Yang, C. (2009). Learn from thy neighbor: parallel-chain adaptive MCMC. *J. Am. Stat. Assoc.* 104:1454–1466.

Diggle, P., Heagerty, P., Liang, K.-Y., Zeger, S. (1995). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Fan, Y., Leslie, D.S., Wand, M.P. (2008). Generalised linear mixed model analysis via sequential Monte Carlo sampling. *Electr. J. Stat.* 2:916–938.

Garthwaite, P.H. (1996). Confidence intervals from randomization tests. *Biometrics* 52:1387–1393.

Garthwaite, P.H., Buckland, S.T. (1992). Generating Monte Carlo confidence intervals by the Robbins-Monro process. *J. R. Stat. Soc. Ser. C* 41:159–171.

Haario, H., Saksman, E., Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 7:223–242.

Haario, H., Saksman, E., Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC output. *Comput. Stat.* 20:265–274.

Hodges, J.L., Lehmann, E.L. (1955). Two approximations to the Robbins-Monro process. In: *Proceedings of the Third Berkeley Symposium* (Vol. I, pp. 95–104).
Lee, S.M.S., Young, G.A. (2003). Prepivoting by weighted bootstrap iteration. *Biometrika* 90:393–410.
Lin, X., Carroll, R.J. (2001). Semiparametric regression for clustered data. *Biometrika* 88:1179–1185.
Matsui, S., Ohashi, Y. (1999). Analysis of recurrent events: application of a clinical trial of colony stimulating factor with the endpoint of febrile neutropenia. *Stat. Med.* 18:2409–2420.
Robbins, H., Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* 22:400–407.
Roberts, G.O., Gelman, A., Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* 7:110–120.
Roberts, G.O., Rosenthal, J.S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* 16:351–367.
Roberts, G.O., Rosenthal, J.S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat.* 18:349–367.
Rosenthal, J.S. (2011). Optimal proposal distributions and adaptive MCMC. In: Brooks, S.P., Gelman, A., Jones, G., Meng, X.-L., eds. *Handbook of Markov Chain Monte Carlo* (pp. 93–112). Boca Raton, FL: Chapman and Hall/CRC Press.
Ruppert, D. (1991). *Handbook of Sequential Analysis*. Chapter Stochastic Approximation (pp. 503–529). New York, NY: Martin Decker.
Schwabe, R., Walk, H. (1996). On a stochastic approximation procedure based on averaging. *Metrika* 44:165–180.
Vihola, M. (2011). On the stability and ergodicity of adaptive scaling Metropolis algorithms. *Stochastic Process. Appl.* 121(12):2839–2860.
Wetherill, G.B. (1963). Sequential estimation of quantal response curve. *Stat. J. R. Soc. B* 25:1–48.
Wetherill, G.B. (1975). *Sequential Methods in Statistics*(2nd ed.). Boca Raton, FL: Chapman and Hall.
Zhao, Y., Staudenmayer, J., Coull, B.A., Wand, M.P. (2006). General design Bayesian generalised linear mixed models. *Stat. Sci.* 21:35–51.