Supplementary Material for
"Hamiltonian Assisted Metropolis Sampling"

Zexi Song and Zhiqiang Tan

I Auxiliary variable derivation of proposal schemes

We show that the proposal scheme (7) can also be derived through an auxiliary variable argument related to Titsias and Papaspiliopoulos (2018), combined with an over-relaxation technique as in Adler (1981) and Neal (1998). Compared with Titsias and Papaspiliopoulos (2018), our derivation deals with the augmented density of (x, u), instead of x alone. More importantly, our derivation incorporates an over-relaxation technique to accommodate all possible proposal schemes (7). Finally, when applied without the momentum, our approach with preconditioning leads to a general class of proposal schemes, including the modified pMALA algorithm in Section 2 and the second-order scheme in Titsias and Papaspiliopoulos (2018). These proposal schemes may involve different configurations of auxiliary variables, after the approximate variance for the target is matched by preconditioning.

The starting point of our derivation is to introduce auxiliary variables (y, v) and further augment the target density as $\pi(x, u, y, v) = \pi(x, u)\pi(y, v|x, u)$. The conditional density $\pi(y, v|x, u)$ can be defined from a random walk update,

$$(y, v)|(x, u) \sim \mathcal{N}((x, u), S), \tag{S1}$$

where S is a $(2k) \times (2k)$ variance matrix independent from (x, u). Given (x_0, u_0), consider the following steps to sample from the new target:

- sample $(y, v)|(x_0, u_0) \sim \pi(y, v|x_0, u_0)$ directly according to (S1),
- sample $(x_1, u_1)|(y, v) \sim \pi(x_1, u_1|y, v)$ by drawing (x^*, u^*) from a conditional proposal density $q(x^*, u^*|y, v, x_0, u_0)$ and accepting $(x_1, u_1) = (x^*, u^*)$ with the usual Metropolis–Hastings probability or otherwise setting $(x_1, u_1) = (x_0, u_0)$.

The two steps can be identified as Gibbs sampling and Metropolis–Hastings within Gibbs sampling respectively. Next, the proposal density $q(x^*, u^*|y, v, x_0, u_0)$ can be defined as an approximation to $\pi(x^*, u^*|y, v)$, based on an approximation to $\pi(x)$ by a normal density with an identity variance anchored at x_0:

$$\begin{aligned}\tilde{\pi}(x; x_0) &\propto \exp\left\{-U(x_0) - (x - x_0)^{\mathrm{T}}\nabla U(x_0) - \frac{1}{2}(x - x_0)^{\mathrm{T}}(x - x_0)\right\} \\ &\propto \mathcal{N}(x|x_0 - \nabla U(x_0), I).\end{aligned} \tag{S2}$$

Specifically, $\tilde{\pi}(x; x_0)$ is determined such that the gradient of $-\log \tilde{\pi}(x; x_0)$ at x_0 coincides with $\nabla U(x_0)$, the gradient of $U(x) = -\log \pi(x)$ at x_0. We take $q(x^*, u^*|y, v, x_0, u_0) = \tilde{\pi}(x^*, u^*|y, v; x_0)$, the induced conditional density by (S3) in Lemma S1. This result can be shown by similar calculation as in Gelman et al. (2014, Section 3.5).

Lemma S1 *Define* $\tilde{\pi}(x, u; x_0) \propto \tilde{\pi}(x; x_0) \exp(-u^{\mathrm{T}}u/2)$. *Then the joint density defined by* $\tilde{\pi}(x, u; x_0)$ $\times \pi(y, v|x, u)$ *induces the conditional density*

$$\tilde{\pi}(x, u|y, v; x_0) = \mathcal{N}(x, u|\mu_{x_0}, A), \tag{S3}$$

where $\pi(y, v|x, u)$ *is as in (S1), and*

$$A = (I + S^{-1})^{-1}, \quad \mu_{x_0} = A \left[\begin{pmatrix} x_0 - \nabla U(x_0) \\ \mathbf{0} \end{pmatrix} + S^{-1} \begin{pmatrix} y \\ v \end{pmatrix} \right].$$

Similarly as in Titsias and Papaspiliopoulos (2018), the auxiliary variables (y, v) can be integrated out to obtain a marginal scheme from (x_0, u_0) to (x^*, u^*) as

$$\begin{aligned} q(x^*, u^*|x_0, u_0) &= \int \tilde{\pi}(x^*, u^*|y, v; x_0) \pi(y, v|x_0, u_0)\, d(y, v) \\ &= \mathcal{N}\left(\begin{pmatrix} x^* \\ u^* \end{pmatrix} \middle| \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} - A \begin{pmatrix} \nabla U(x_0) \\ u \end{pmatrix}, 2A - A^2 \right), \end{aligned} \tag{S4}$$

where $2A - A^2 = AS^{-1}A + A$ for $A = (I + S^{-1})^{-1}$. Hence the proposal scheme (S4) from the auxiliary variable argument takes the same form as (7). This discussion also confirms the previous observation that when the target density $\pi(x)$ is $\mathcal{N}(\mathbf{0}, I)$, the proposal (x^*, u^*) in (8) is always accepted, because the normal approximation $\tilde{\pi}(x; x_0)$ becomes exact and hence (x^*, u^*) is obtained from just two-block Gibbs sampling.

There is, however, a caveat in the link between (7) and (S4). Using the auxiliary variables leads to the proposal (S4), with the relation $A = (I + S^{-1})^{-1}$. Because S is positive semi-definite as a variance matrix, this relation imposes the constraint that $A \leq I$. For the proposal scheme (7), it is only required that $\mathbf{0} \leq A \leq 2I$. When $I < A \leq 2I$, the scheme (7) remains valid, but cannot be deduced from (S4). Hence (7) encapsulates a broader class of proposal distributions than directly derived via auxiliary variables.

Next we show that the over-relaxation technique (Adler, 1981; Neal, 1998) can be exploited to define an auxiliary proposal density $q(x^*, u^*|y, v, x_0, u_0)$ more flexible than above, so that the entire class of proposal distributions (7) can be recovered. By over-relaxation based on normal

distributions, consider the proposal density

$$q_\alpha(x^*, u^*|y, v, x_0, u_0) = \mathcal{N}\left(x^*, u^*|\mu_{x_0} + \alpha((x_0^{\mathrm{T}}, u_0^{\mathrm{T}})^{\mathrm{T}} - \mu_{x_0}), (1-\alpha^2)A\right),$$

where μ_{x_0} and A are defined as in Lemma S1, and $-1 \leq \alpha \leq 1$ controls the degree of over-relaxation. Setting $\alpha = 0$ gives the previous choice $q(x^*, u^*|y, v, x_0, u_0) = \tilde{\pi}(x^*, u^*|y, v; x_0)$ from (S3) and leads to the marginal proposal density (S4).

Lemma S2 *Let $A_\alpha = (1-\alpha)A$. The marginal proposal density obtained by integrating out (y, v) from $q_\alpha(x^*, u^*|y, v, x_0, u_0)$ is*

$$\begin{aligned} q_\alpha(x^*, u^*|x_0, u_0) &= \int q_\alpha(x^*, u^*|y, v, x_0, u_0)\pi(y, v|x_0, u_0)\, d(y, v) \\ &= \mathcal{N}\left(\begin{pmatrix} x^* \\ u^* \end{pmatrix} \middle| \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} - A_\alpha \begin{pmatrix} \nabla U(x_0) \\ u \end{pmatrix}, 2A_\alpha - A_\alpha^2\right). \end{aligned} \tag{S5}$$

By the preceding result, the marginal scheme (S5) is still of the form (7), with A replaced by A_α. The matrix A_α is determined from α and S as $A_\alpha = (1-\alpha)(I + S^{-1})^{-1}$. The constraints $-1 \leq \alpha \leq 1$ and $S \geq \mathbf{0}$ imply that $\mathbf{0} \leq A_\alpha \leq 2I$. Conversely, any matrix $\mathbf{0} \leq A \leq 2I$ can be obtained as A_α for some $-1 \leq \alpha \leq 1$ and $S \geq \mathbf{0}$. The choice $A = 2I$ corresponds to the limit case $\alpha = -1$ and $S \to \infty$. In this sense, the proposal scheme (7) with any choice $\mathbf{0} \leq A \leq 2I$ can be identified as a marginal scheme from the auxiliary variable argument while incorporating over-relaxation.

In the remainder of this section, we discuss consequences of the foregoing development (without over-relaxation) when the momentum variable u is dropped. In this case, the proposal density from x_0 to x^* in (S4) reduces to

$$q(x^*|x_0) = \mathcal{N}(x^*|x_0 - \tilde{A}\nabla U(x_0), 2\tilde{A} - \tilde{A}^2),$$

where $\tilde{A}$ is a $k \times k$ symmetric matrix satisfying $\mathbf{0} \leq \tilde{A} \leq I$ before over-relaxation. Taking $\tilde{A} = \frac{\epsilon^2}{1+\sqrt{1-\epsilon^2}} I = (1 - \sqrt{1-\epsilon^2})I$ leads to the proposal scheme

$$x^* = x_0 - \frac{\epsilon^2}{1+\sqrt{1-\epsilon^2}}\nabla U(x_0) + Z, \quad Z \sim \mathcal{N}(\mathbf{0}, \epsilon^2 I), \tag{S6}$$

which is precisely the proposal scheme in modified pMALA with $\Sigma = I$ (i.e., modified MALA). In general, our auxiliary variable argument can be applied, with a normal approximation to $\pi(x)$ using an arbitrary choice of variance matrix Σ:

$$\tilde{\pi}(x; x_0) \propto \exp\left\{-U(x_0) - (x - x_0)^{\mathrm{T}}\nabla U(x_0) - \frac{1}{2}(x - x_0)^{\mathrm{T}}\Sigma^{-1}(x - x_0)\right\}. \tag{S7}$$

With the momentum dropped from (S1), the auxiliary variable y is defined as $y|x \sim \mathcal{N}(x, \tilde{S})$, where $\tilde{S}$ is a $k \times k$ variance matrix. By similar reasoning which leads to (S4) but using (S7), we obtain a proposal scheme from x_0 to x^* in the form

$$x^* = x_0 - \tilde{A}\nabla U(x_0) + Z, \quad Z \sim \mathcal{N}(0, 2\tilde{A} - \tilde{A}\Sigma^{-1}\tilde{A}), \tag{S8}$$

where $\tilde{A} = (\Sigma^{-1} + \tilde{S}^{-1})^{-1}$, a $k \times k$ symmetric matrix satisfying $\mathbf{0} \le \tilde{A} \le \Sigma$ (without over-relaxation), and $2\tilde{A} - \tilde{A}\Sigma^{-1}\tilde{A} = \tilde{A}\tilde{S}^{-1}\tilde{A} + \tilde{A}$. When the target distribution is $\mathcal{N}(\mathbf{0}, \Sigma)$, the proposal x^* is always accepted under Metropolis–Hastings sampling. Taking $\tilde{A} = \frac{\epsilon^2}{1+\sqrt{1-\epsilon^2}}\Sigma$ or equivalently $\tilde{S} = \{(1-\sqrt{1-\epsilon^2})/\sqrt{1-\epsilon^2}\}\Sigma \propto \Sigma$ leads to the proposal scheme in modified pMALA. As a special case, taking $\Sigma = C$ and $\tilde{A} = \frac{\epsilon^2}{1+\sqrt{1-\epsilon^2}}C$ or equivalently $\tilde{S} = \{(1-\sqrt{1-\epsilon^2})/\sqrt{1-\epsilon^2}\}C \propto C$ yields the proposal scheme (2) in pCNL.

It is interesting to compare pCNL and pMALA* (i.e., modified pMALA) with Titsias and Papaspiliopoulos (2018) in the Bayesian setting with $\pi(x) \propto \exp\{-U(x)\} = \exp\{\ell(x)\}\mathcal{N}(x|\mathbf{0}, C)$, where $\ell(x)$ is the log-likelihood and C a prior variance. The marginal algorithm, denoted as mGrad, with the proposal (3) is derived in Titsias and Papaspiliopoulos (2018) using auxiliary variables and a normal approximation to $\pi(x)$ similarly as above. In fact, the proposal scheme (3) can be deduced from the general class (S8) by taking $\Sigma = C$ and $\tilde{A} = \tilde{C}$ or equivalently $\tilde{S} = \frac{\delta}{2}I$. Compared with pCNL, the mGrad algorithm also uses the prior variance C for preconditioning, but involves a different configuration of the auxiliary variable y given x, where the conditional variance is $\tilde{S} = \frac{\delta}{2}I$ instead of being proportional to C. As discussed in Titsias and Papaspiliopoulos (2018), Section 3.4, the mGrad algorithm based on this choice of auxiliary variables can achieve certain advantages over pCNL in the context of posterior sampling with a latent Gaussian field model.

As mentioned in Section 2, pMALA* in general differs from pCNL and mGrad in allowing a preconditioning matrix Σ to capture both the prior and the likelihood. In this direction, we compare pMALA* with the second-order algorithm, denoted as mGrad2, in the Supplement of Titsias and Papaspiliopoulos (2018). The proposal scheme for mGrad2, after correcting a typo to match the first-order scheme (3) when $G = \mathbf{0}$, can be written as

$$x^* = \frac{2}{\delta}C^\dagger x_0 + C^\dagger(\nabla\ell(x_0) - Gx_0) + Z = x_0 - C^\dagger\nabla U(x_0) + Z, \quad Z \sim \mathcal{N}(\mathbf{0}, \frac{2}{\delta}C^{\dagger 2} + C^\dagger), \tag{S9}$$

where $C^\dagger = (\frac{2}{\delta}I + C^{-1} - G)^{-1}$, and G is the Hessian $\nabla^2\ell(x_0)$. To facilitate comparison, assume that G is an Hessian approximation, independent of x_0. The proposal scheme (S9), by direct calculation, can be obtained from (S8) by taking $\Sigma = (C^{-1} - G)^{-1}$ as an approximation to the variance of the target $\pi(x)$ and $\tilde{A} = C^\dagger = (\Sigma^{-1} + \frac{2}{\delta}I)^{-1}$, corresponding to $\tilde{S} = \frac{\delta}{2}I$ as the conditional variance of the

auxiliary variable y given x. Therefore, the second-order algorithm mGrad2 and modified pMALA use proposal schemes both in the class (S8), but with different configurations of the auxiliary variable y (hence difference choices of $\tilde{A}$), after the approximate variance Σ is matched. Further comparison of these two algorithms with general $\Sigma \neq C$ can be investigated in future work.

As a final note, it is helpful to mention that we refer to as a preconditioning matrix specifically the variance matrix Σ used in the normal approximation (S7). The resulting proposal scheme (S8) is rejection-free when the target distribution is normal with variance matrix Σ. This approach differs from how preconditioning is constructed in Titsias and Papaspiliopoulos (2018). The preconditioned version of the proposal (3) in their Eq. (8) can be expressed as (S8) with $\Sigma = C$ and $\tilde{S} = \frac{\delta}{2}V$ for some preconditioning matrix V. The prior variance C is used as the approximate variance for the target distribution but V is used as the conditional variance of the auxiliary variable y given x. This proposal scheme is rejection-free for the target $\mathcal{N}(\mathbf{0}, C)$, not in general for $\mathcal{N}(\mathbf{0}, V)$.

I.1 Proof of Lemma S2

Given the current variables (x_0, u_0), the variables (y, v) are generated as

$$(y, v)|(x_0, u_0) \sim \mathcal{N}((x_0, u_0),\, S). \tag{S10}$$

The variables (x^*, u^*) are then generated from q_α as

$$(x^*, u^*)|(y, v, x_0, u_0) \sim \mathcal{N}\left((1-\alpha)\mu_{x_0} + \alpha \begin{pmatrix} x_0 \\ u_0 \end{pmatrix},\, (1-\alpha^2)A\right), \tag{S11}$$

where $-1 \leq \alpha \leq 1$, and

$$A = (I + S^{-1})^{-1},\ \mu_{x_0} = A\left(\begin{pmatrix} x_0 - \nabla U(x_0) \\ \mathbf{0} \end{pmatrix} + S^{-1}\begin{pmatrix} y \\ v \end{pmatrix}\right). \tag{S12}$$

Then (x^*, u^*) and (y, v) are jointly normal given (x_0, u_0) and hence $(x^*, u^*)|(x_0, u_0)$ is also normally distributed. It suffices to determine its mean and variance.

First, we compute $\mathrm{E}(x^*, u^*|x_0, u_0)$. By (S10) and (S12),

$$\begin{aligned} \mathrm{E}[\mu_{x_0}|x_0, u_0] &= A\left(\begin{pmatrix} x_0 - \nabla U(x_0) \\ \mathbf{0} \end{pmatrix} + S^{-1}\begin{pmatrix} x_0 \\ u_0 \end{pmatrix}\right) \\ &= A\left((I + S^{-1})\begin{pmatrix} x_0 \\ u_0 \end{pmatrix} - \begin{pmatrix} \nabla U(x_0) \\ u_0 \end{pmatrix}\right) = \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} - A\begin{pmatrix} \nabla U(x_0) \\ u_0 \end{pmatrix}. \end{aligned} \tag{S13}$$

Therefore, by (S11) and (S13),

$$\begin{aligned}
&\mathrm{E}(x^*, u^* | x_0, u_0) = \mathrm{E}[\,\mathrm{E}(x, u^* | y, v, x_0, u_0)\,|x_0, u_0] \\
&= \mathrm{E}\left[(1-\alpha)\mu_{x_0} + \alpha \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} \middle| x_0, u_0 \right] = \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} - A_\alpha \begin{pmatrix} \nabla U(x_0) \\ u_0 \end{pmatrix},
\end{aligned}$$

where $A_\alpha = (1-\alpha)A$.

Next, we compute $\mathrm{Var}(x^*, u^* | x_0, u_0)$. By (S11)–(S12),

$$\begin{aligned}
&\mathrm{Var}[\,\mathrm{E}(x^*, u^* | y, v, x_0, u_0)\,|x_0, u_0] = \mathrm{Var}\left[(1-\alpha)\mu_{x_0} + \alpha \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} \middle| x_0, u_0 \right] \\
&= (1-\alpha)^2 \mathrm{Var}[\mu_{x_0} | x_0, u_0] = (1-\alpha)^2 A S^{-1} A = A_\alpha S^{-1} A_\alpha, \qquad \text{(S14)} \\
&\mathrm{E}[\,\mathrm{Var}(x^*, u^* | y, v, x_0, u_0)\,|x_0, u_0] = \mathrm{E}[(1-\alpha^2)A | x_0, u_0] \\
&= (1-\alpha^2)A = (1+\alpha)A_\alpha. \qquad \text{(S15)}
\end{aligned}$$

Combining (S14) and (S15) yields

$$\begin{aligned}
&\mathrm{Var}(x^*, u^* | x_0, u_0) \\
&= \mathrm{E}[\,\mathrm{Var}(x^*, u^* | y, v, x_0, u_0 (\,|x_0, u_0] + \mathrm{Var}[\,\mathrm{E}(x^*, u^* | y, v, x_0, u_0)\,|x_0, u_0] \\
&= A_\alpha S^{-1} A_\alpha + (1+\alpha) A_\alpha.
\end{aligned}$$

Finally, we show that $A_\alpha S^{-1} A_\alpha + (1+\alpha)A_\alpha = 2A_\alpha - A_\alpha^2$. Because $A = (I + S^{-1})^{-1}$, we have $A(I + S^{-1}) = I$ and hence $A^2 + AS^{-1}A = A$. Then

$$\begin{aligned}
&(1-\alpha)^2 A S^{-1} A = (1-\alpha)^2 A - (1-\alpha)^2 A^2 \\
&\Rightarrow A_\alpha S^{-1} A_\alpha = (1-\alpha) A_\alpha - A_\alpha^2 \\
&\Rightarrow A_\alpha S^{-1} A_\alpha + (1+\alpha) A_\alpha = 2A_\alpha - A_\alpha^2.
\end{aligned}$$

This completes the proof of Lemma S2.

II Demonstration of validity of UDL

We demonstrate that UDL is valid in leaving the augmented target $\pi(x, u)$ invariant. Similarly as HAMS, by Proposition 3, it suffices to verify that the acceptance probability stated for UDL in Section 2 can be written in the form of generalized Metropolis–Hastings probability (21) for the associated (forward) proposal density Q.

First, we calculate the generalized Metropolis–Hastings probability (21) with the (forward) proposal density Q from UDL. The proposal scheme in UDL is defined as

$$
\begin{aligned}
&\text{Sample } Z_1, Z_2 \sim \mathcal{N}(\mathbf{0}, M) \text{ independently,} \\
&u^+ = \sqrt{c}u_0 + \sqrt{1-c}Z_1, \\
&\tilde{u} = u^+ - \frac{\epsilon}{2}\nabla U(x_0), \quad x^* = x_0 + \epsilon M^{-1}\tilde{u}, \quad u^- = \tilde{u} - \frac{\epsilon}{2}\nabla U(x^*), \\
&u^* = \sqrt{c}u^- + \sqrt{1-c}Z_2.
\end{aligned}
$$

The noises (Z_1, Z_2) can be expressed as

$$
Z_1 = \left(\frac{M}{\epsilon}(x^* - x_0) + \frac{\epsilon}{2}\nabla U(x_0) - \sqrt{c}u_0\right)(1-c)^{-1/2}, \tag{S16}
$$

$$
Z_2 = \left(\frac{\sqrt{c}M}{\epsilon}(x_0 - x^*) + \frac{\epsilon\sqrt{c}}{2}\nabla U(x^*) + \sqrt{c}u^*\right)(1-c)^{-1/2}. \tag{S17}
$$

Suppose that the mapping above from (x_0, u_0) to (x^*, u^*) is applied from $(x^*, -u^*)$ to $(x_0, -u_0)$, but using new noises (Z_3, Z_4). By exchanging (x_0, u_0) and $(x^*, -u^*)$, the new noises (Z_3, Z_4) can be calculated as

$$
Z_3 = \left(\frac{M}{\epsilon}(x_0 - x^*) + \frac{\epsilon}{2}\nabla U(x^*) + \sqrt{c}u^*\right)(1-c)^{-1/2}, \tag{S18}
$$

$$
Z_4 = \left(\frac{\sqrt{c}M}{\epsilon}(x^* - x_0) + \frac{\epsilon\sqrt{c}}{2}\nabla U(x_0) - \sqrt{c}u_0\right)(1-c)^{-1/2}. \tag{S19}
$$

Then the forward and backward transitions of the proposals for UDL can be illustrated in a similar manner to (20) as

$$
\begin{pmatrix} x_0 \\ u_0 \end{pmatrix} \overset{(Z_1, Z_2)}{\longrightarrow} \begin{pmatrix} x^* \\ u^* \end{pmatrix}, \qquad \begin{pmatrix} x^* \\ -u^* \end{pmatrix} \overset{(Z_3, Z_4)}{\longrightarrow} \begin{pmatrix} x_0 \\ -u_0 \end{pmatrix}, \tag{S20}
$$

where the arrows denote the *same* mapping, depending on (Z_1, Z_2) or (Z_3, Z_4).

Because (Z_1, Z_2) are the only sources of randomness, the (forward) proposal density from (x_0, u_0) to (x^*, u^*) is

$$
\begin{aligned}
Q(x^*, u^* | x_0, u_0) &= \mathcal{N}(Z_1 | \mathbf{0}, M)\mathcal{N}(Z_2 | \mathbf{0}, M) \\
&\propto \exp\left(-\frac{1}{2}Z_1^{\mathrm{T}} M^{-1} Z_1 - \frac{1}{2}Z_2^{\mathrm{T}} M^{-1} Z_2\right).
\end{aligned} \tag{S21}
$$

Evaluation of the *same* proposal density from $(x^*, -u^*)$ to $(x_0, -u_0)$ gives

$$
\begin{aligned}
Q(x_0, -u_0 | x^*, -u^*) &= \mathcal{N}(Z_3 | \mathbf{0}, M)\mathcal{N}(Z_4 | \mathbf{0}, M) \\
&\propto \exp\left(-\frac{1}{2}Z_3^{\mathrm{T}} M^{-1} Z_3 - \frac{1}{2}Z_4^{\mathrm{T}} M^{-1} Z_4\right).
\end{aligned} \tag{S22}
$$

Using (S16) to (S22), the log ratio of proposal densities is

$$\log\left(\frac{Q(x_0,-u_0|x^*,-u^*)}{Q(x^*,u^*|x_0,u_0)}\right)=\frac{1}{2}(x^*-x_0)^{\mathrm{T}}(\nabla U(x^*)+\nabla U(x_0))$$
$$-\frac{\epsilon^2}{8}\left([\nabla U(x^*)]^{\mathrm{T}}M^{-1}\nabla U(x^*)-[\nabla U(x_0)]^{\mathrm{T}}M^{-1}\nabla U(x_0)\right)-\frac{1}{2}\left(u_0^{\mathrm{T}}M^{-1}u_0-(u^*)^{\mathrm{T}}M^{-1}u^*\right). \quad \text{(S23)}$$

Furthermore, the log ratio of target densities at (x_0,u_0) and $(x^*,-u^*)$ is

$$\log\left(\frac{\pi(x^*,-u^*)}{\pi(x_0,u_0)}\right)=U(x_0)-U(x^*)+\frac{1}{2}\left(u_0^{\mathrm{T}}M^{-1}u_0-(u^*)^{\mathrm{T}}M^{-1}u^*\right). \quad \text{(S24)}$$

From (S23) and (S24), the generalized Metropolis–Hastings probability (21) is

$$\min\left(1,\exp\left\{U(x_0)-U(x^*)+\frac{(x^*-x_0)^{\mathrm{T}}}{2}(\nabla U(x_0)+\nabla U(x^*))\right.\right.$$
$$\left.\left.-\frac{\epsilon^2}{8}\left([\nabla U(x^*)]^{\mathrm{T}}M^{-1}\nabla U(x^*)-[\nabla U(x_0)]^{\mathrm{T}}M^{-1}\nabla U(x_0)\right)\right\}\right). \quad \text{(S25)}$$

Second, we show that generalized Metropolis–Hastings probability (S25) reduces to the acceptance probability stated in Section 2:

$$\min\left(1,\exp(H(x_0,u^+)-H(x^*,u^-))\right). \quad \text{(S26)}$$

In fact, direct calculation using $u^-=u^+-\frac{\epsilon}{2}\left(\nabla U(x^*)+\nabla U(x_0)\right)$ yields

$$(u^-)^{\mathrm{T}}M^{-1}u^-=(u^+)^{\mathrm{T}}M^{-1}u^++\frac{\epsilon^2}{4}(\nabla U(x_0)+\nabla U(x^*))^{\mathrm{T}}M^{-1}(\nabla U(x_0)+\nabla U(x^*))$$
$$-\epsilon(u^+)^{\mathrm{T}}M^{-1}(\nabla U(x_0)+\nabla U(x^*)),$$

and hence

$$\frac{1}{2}(u^+)^{\mathrm{T}}M^{-1}u^+-\frac{1}{2}(u^-)^{\mathrm{T}}M^{-1}u^-$$
$$=\frac{\epsilon}{2}(u^+)^{\mathrm{T}}M^{-1}(\nabla U(x_0)+\nabla U(x^*)-\frac{\epsilon^2}{8}(\nabla U(x_0)+\nabla U(x^*))^{\mathrm{T}}M^{-1}(\nabla U(x_0)+\nabla U(x^*))$$
$$=\frac{1}{2}(x^*-x_0)^{\mathrm{T}}(\nabla U(x_0)+\nabla U(x^*))-\frac{\epsilon^2}{8}\left([\nabla U(x^*)]^{\mathrm{T}}M^{-1}\nabla U(x^*)-[\nabla U(x_0)]^{\mathrm{T}}M^{-1}\nabla U(x_0)\right). \quad \text{(S27)}$$

By the definition of the Hamiltonian, we have

$$H(x_0,u^+)-H(x^*,u^-)=U(x_0)-U(x^*)+\frac{1}{2}(u^+)^{\mathrm{T}}M^{-1}u^+-\frac{1}{2}(u^-)^{\mathrm{T}}M^{-1}u^-.$$

Substituting (S27) into the above, we see that (S26) equals (S25).

III Generalized Metropolis–Hastings sampling

We give a broader definition of generalized Metropolis–Hastings sampling in Section 4, to accommodate both continuous and discrete variables.

Let $\pi(y)$ be a pre-specified probability density function on $\mathcal{Y}$, with respect to possibly a product of Lebesgue and counting measures. Assume that $J : \mathcal{Y} \to \mathcal{Y}$ is an invertible mapping, such that for any set $C \subset \mathcal{Y}$ and integrable function h,

$$\int_{J(C)} \pi(y)\,\mathrm{d}y = \int_C \pi(y)\,\mathrm{d}y, \tag{S28}$$

$$\int_{J(C)} h(J^{-1}y)\pi(y)\,\mathrm{d}y = \int_C h(y)\pi(y)\,\mathrm{d}y, \tag{S29}$$

where J^{-1} denote the inverse mapping of J, and $J(C) = \{Jy : y \in C\}$. While (S28) is restated from (33), condition (S29) is in general stronger than (S28) and analogous to saying that the Jacobian determinant of mapping J is ± 1 and $\pi(J^{-1}y) = \pi(y)$ in the case where $\mathcal{Y}$ is an Euclidean space endowed with the Lebesgue measure. With this interpretation of Jy, generalized Metropolis–Hastings sampling is still defined as in Section 4. More importantly, Proposition 3 can be seen to remain valid, by substituting (S29) for all the change-of-variables calculation in the proof.

Next we show that the irreversible jump sampler (I-Jump) in Ma et al. (2018) can be obtained as a special case of generalized Metropolis-Hastings sampling, when a binary auxiliary variable $s \in \{1, -1\}$ is introduced for sampling from an original target density $\pi(x)$ on $\mathcal{X}$. Given current variables (x_0, s_0), an iteration of I-Jump can be described as follows, where $f(\cdot|x_0)$ and $g(\cdot|x_0)$ are two possibly different proposal densities.

Irreversible jump sampler (I-Jump).

- Sample $w \sim \text{Uniform}[0, 1]$.
- If $s_0 = 1$, sample $x^* \sim f(\cdot|x_0)$ and compute

$$\rho(x^*|x_0) = \min\left(1, \frac{\pi(x^*)g(x_0|x^*)}{\pi(x_0)f(x^*|x_0)}\right);$$

  else sample $x^* \sim g(\cdot|x_0)$ and compute

$$\rho(x^*|x_0) = \min\left(1, \frac{\pi(x^*)f(x_0|x^*)}{\pi(x_0)g(x^*|x_0)}\right).$$

- If $w < \rho(x^*|x_0)$, then set $(x_1, s_1) = (x^*, s_0)$; else set $(x_1, s_1) = (x_0, -s_0)$.

To recast I-Jump, consider the augmented target density $\pi(x, s) = \pi(x)/2$ on the product space $\mathcal{Y} = \mathcal{X} \times \{1, -1\}$, that is, x and s are independent and s takes value 1 or -1 with equal probabilities.

The mapping defined by $J(x, s) = (x, -s)$ satisfies conditions (S28)–(S29). Define the proposal density Q as

$$Q(x^*, s^*|x_0, s_0) = \begin{cases} f(x^*|x_0), & \text{if } s^* = s_0 = 1, \\ g(x^*|x_0), & \text{if } s^* = s_0 = -1, \\ 0, & \text{if } s^* \neq s_0. \end{cases}$$

Then the acceptance probability in I-Jump can be expressed as

$$\rho(x^*, s^*|x_0, s_0) = \min\left(1, \frac{\pi(x^*, -s^*)Q(x_0, -s_0|x^*, -s^*)}{\pi(x_0, s_0)Q(x^*, s^*|x_0, s_0)}\right).$$

by noticing that $s^* = s_0$ and $\pi(x^*, -s^*)/\pi(x_0, s_0) = \pi(x^*)/\pi(x_0)$. Therefore, the I-Jump algorithm can be seen as generalized Metropolis–Hastings sampling.

As a concrete example of I-Jump, Ma et al. (2018) proposed an irreversible MALA (I-MALA) algorithm. The proposal schemes $f(\cdot|x_0)$ and $g(\cdot|x_0)$ are defined as discretizations of irreversible continuous Markov processes. Each proposal scheme can be related to (36) in our G2MS algorithm with y_0 replaced by x_0:

$$x^* = x_0 - B\nabla U(x_0) + Z, \quad Z \sim \mathcal{N}(\mathbf{0}, B + B^{\mathrm{T}} - BB^{\mathrm{T}}).$$

For $B = \epsilon^2 B_0$ with $\epsilon \approx 0$, the preceding scheme is approximately

$$x^* = x_0 - \epsilon^2(D_0 + C_0)\nabla U(x_0) + Z, \quad Z \sim \mathcal{N}(\mathbf{0}, 2\epsilon^2 D_0), \tag{S30}$$

where $D_0 = (B_0 + B_0^{\mathrm{T}})/2$ is symmetric and $C_0 = (B_0 - B_0^{\mathrm{T}})/2$ is skew-symmetric. It is interesting that the form of (S30) matches the proposal schemes derived by discretizing general Markov procsses in Ma et al. (2018).

Although both HAMS and I-MALA can be subsumed by generalized Metropolis–Hastings sampling, there remain important differences. The HAMS algorithm uses momentum as an auxiliary variable and hence is able to exploit symmetry in the momentum distribution, whereas I-MALA relies on lifting with a binary variable (Gustafson, 1998; Vucelja, 2016) and needs to split the original variable x to specify symmetric and skew-symmetric matrices D_0 and C_0 when defining proposal schemes based on irreversible Markov processes in x. Further research is desired to compare and connect these algorithms.

IV Proofs

IV.1 Proof of Propositions 1 and 2

The results follow from Proposition 3, by the discussion at the end of Section 4.

IV.2 Proof of Proposition 3

First, the transition kernel $K(y_1|y_0)$ can be expressed as

$$K(y_1|y_0)\,\mathrm{d}y_1 = Q(y_1|y_0)\rho(y_1|y_0)\mathrm{d}y_1 + (1 - r(y_0))\delta_{Jy_0}(\mathrm{d}y_1), \tag{S31}$$

where $r(y_0) = \int Q(y_1|y_0)\rho(y_1|y_0)\mathrm{d}y_1$, δ_y denotes point mass at y, and as in (34),

$$\rho(y_1|y_0) = \min\left(1, \frac{\pi(J^{-1}y_1)Q(Jy_0|J^{-1}y_1)}{\pi(y_0)Q(y_1|y_0)}\right).$$

Then for $y_1 \neq Jy_0$,

$$\pi(y_0)K(y_1|y_0) = \pi(y_0)Q(y_1|y_0)\rho(y_1|y_0).$$

Replacing (y_0, y_1) with $(J^{-1}y_1, Jy_0)$ above shows that for $Jy_0 \neq y_1$,

$$\pi(J^{-1}y_1)K(Jy_0|J^{-1}y_1) = \pi(J^{-1}y_1)Q(Jy_0|J^{-1}y_1)\rho(Jy_0|J^{-1}y_1),$$

where

$$\rho(Jy_0|J^{-1}y_1) = \min\left(1, \frac{\pi(y_0)Q(y_1|y_0)}{\pi(J^{-1}y_1)Q(Jy_0|J^{-1}y_1)}\right).$$

By direct calculation, we see that for $Jy_0 \neq y_1$,

$$\begin{aligned}\pi(y_0)Q(y_1|y_0)\rho(y_1|y_0) &= \pi(J^{-1}y_1)Q(Jy_0|J^{-1}y_1)\rho(Jy_0|J^{-1}y_1)\\ &= \min\left(\pi(y_0)Q(y_1|y_0), \pi(J^{-1}y_1)Q(Jy_0|J^{-1}y_1)\right),\end{aligned} \tag{S32}$$

that is,

$$\pi(y_0)K(y_1|y_0) = \pi(J^{-1}y_1)K(Jy_0|J^{-1}y_1). \tag{S33}$$

Equation (S33) also holds trivially if $Jy_0 = y_1$. Hence (S33) holds whether $Jy_0 = y_1$ or not. By the invariance property $\pi(J^{-1}y_1) = \pi(y_1)$, equation (S33) reduces to (35).

The proof that $\pi(y)$ is a stationary distribution is a generalization of Tierney (1994). It suffices to show that for any set $C \subset \mathcal{Y}$,

$$\int_C \left(\int \pi(y_0)K(y_1|y_0)\,\mathrm{d}y_0\right)\,\mathrm{d}y_1 = \int_C \pi(y_1)\,\mathrm{d}y_1. \tag{S34}$$

By (S31), the left-hand side of (S34) can be calculated as

$$
\begin{aligned}
&\int_C \left(\int \pi(y_0) Q(y_1|y_0) \rho(y_1|y_0) \, \mathrm{d}y_0 \right) \mathrm{d}y_1 + \int_{J^{-1}(C)} (1 - r(y_0)) \pi(y_0) \, \mathrm{d}y_0 \\
&= \int_C \left(\int Q(Jy_0|J^{-1}y_1) \rho(Jy_0|J^{-1}y_1) \, \mathrm{d}y_0 \right) \pi(J^{-1}y_1) \, \mathrm{d}y_1 + \int_{J^{-1}(C)} (1 - r(y_0)) \pi(y_0) \, \mathrm{d}y_0 \\
&= \int_C r(J^{-1}y_1) \pi(J^{-1}y_1) |\det(J^{-1})| \, \mathrm{d}y_1 + \int_{J^{-1}(C)} (1 - r(y_0)) \pi(y_0) \, \mathrm{d}y_0 \\
&= \int_C r(J^{-1}y_1) \pi(J^{-1}y_1) |\det(J^{-1})| \, \mathrm{d}y_1 + \int_C (1 - r(J^{-1}y_0)) \pi(J^{-1}y_0) |\det(J^{-1})| \, \mathrm{d}y_0 \\
&= \int_C \pi(J^{-1}y_1) |\det(J^{-1})| \, \mathrm{d}y_1 = \int_{J(C)} \pi(y_1) \, \mathrm{d}y_1,
\end{aligned}
$$

which yields the right-hand side of (S34) by the invariance property (33). The first equality follows from (S32), the second from the definition of $r(\cdot)$ and the change of variables, and the third and fifth both from the change of variables.

IV.3 Proof of Corollary 1

The result follows from Corollary 3, by the discussion at the end of Section 4.

IV.4 Proof of Corollary 3

The backward proposal scheme (37) becomes $Jy_0 = (I - A)y^* + Z^*$. The new noise Z^* can be directly calculated using (39) as

$$
Z^* = Jy_0 - (I - A)y^* = (2A - A^2)Jy_0 - (I - A)Z.
$$

Suppose that $y_0 \sim \mathcal{N}(\mathbf{0}, I)$ and y^* is generated by (39), $y^* = (1-A)Jy_0 + Z$, with $Z \sim \mathcal{N}(\mathbf{0}, 2A - A^2)$ independently of y_0. Then the conditional density of y^* given y_0 is $p(y^*|y_0) = \mathcal{N}(Z|\mathbf{0}, 2A - A^2)$. Moreover, by direct calculation, Z^* is distributed as $\mathcal{N}(\mathbf{0}, 2A - A^2)$, independently of y^*. Hence the conditional density of Jy_0 given y^* is $p(Jy_0|y^*) = \mathcal{N}(Z^*|\mathbf{0}, 2A - A^2)$. By the change of variables, the conditional density of y_0 given y^* is also $p(y_0|y^*) = \mathcal{N}(Z^*|\mathbf{0}, 2A - A^2)$ because $|\det(J)| = 1$. Therefore, the acceptance probability (38) reduces to 1, because $\pi(y_0)p(y^*|y_0) = \pi(y^*)p(y_0|y^*)$: both $\pi(y_0)p(y^*|y_0)$ and $\pi(y^*)p(y_0|y^*)$ give the joint density of (y_0, y^*).

IV.5 Proof of Lemma 1

The HAMS-A proposal described in Section 3.3 is

$$\begin{aligned}
&\tilde{Z} = Z - a\nabla U(x_0) + \sqrt{ab}u_0, \quad Z \sim \mathcal{N}(\mathbf{0}, a(2-a-b)I), \\
&x^* = x_0 + \tilde{Z}, \quad u^* = -u_0 + \sqrt{\frac{b}{a}}\tilde{Z} + \phi(\tilde{Z} + \nabla U(x_0) - \nabla U(x^*)), \\
&Z^* = \tilde{Z} - a\nabla U(x^*) - \sqrt{ab}u^*.
\end{aligned}$$

We express x^*, u^* and Z^* in terms of x_0, u_0, Z and $\nabla U(x^*)$:

$$x^* = x_0 - \nabla U(x_0) + \sqrt{ab}u_0 + Z, \tag{S35}$$

$$u^* = [\phi - \phi a - \sqrt{ab}]\nabla U(x_0) - \phi\nabla U(x^*) + [\phi\sqrt{ab} + b - 1]u_0 + \left[\phi + \sqrt{\frac{b}{a}}\right] Z, \tag{S36}$$

$$\begin{aligned}
Z^* = {} & [ab + \phi a\sqrt{ab} - \phi\sqrt{ab} - a]\nabla U(x_0) + (\sqrt{ab}\phi - a)\nabla U(x^*) \\
& + [2\sqrt{ab} - \phi ab - b\sqrt{ab}]u_0 + \left[1 - \phi\sqrt{ab} - b\right] Z.
\end{aligned} \tag{S37}$$

Suppose that the target density $\pi(x)$ is $\mathcal{N}(\mathbf{0}, \gamma^{-1}I)$. Then x^*, u^* and Z^* from (S35)–(S37) can be expressed in terms of only x_0, u_0 and Z as

$$x^* = (-a\gamma + 1)x_0 + \sqrt{ab}u_0 + Z, \tag{S38}$$

$$u^* = [\underbrace{a\phi\gamma^2}_{(i)} - (a\phi + \sqrt{ab})\gamma]x_0 + [\underbrace{-\phi\sqrt{ab}\gamma}_{(iii)} + \phi\sqrt{ab} + b - 1]u_0 + \left[\underbrace{-\phi\gamma}_{(v)} + \phi + \sqrt{\frac{b}{a}}\right] Z, \tag{S39}$$

$$\begin{aligned}
Z^* = {} & [\underbrace{(a^2 - \phi a\sqrt{ab})\gamma^2}_{(ii)} + (ab - 2a + \phi a\sqrt{ab})\gamma]x_0 \\
& + [\underbrace{(\phi ab - a\sqrt{ab})\gamma}_{(iv)} + 2\sqrt{ab} - b\sqrt{ab} - \phi ab]u_0 \\
& + [\underbrace{(\phi\sqrt{ab} - a)\gamma}_{(vi)} + 1 - b - \phi\sqrt{ab}]Z.
\end{aligned} \tag{S40}$$

The quantity inside the exponential in (28) is

$$\begin{aligned}
& H(x_0, u_0) - H(x^*, u^*) + \frac{Z^{\mathrm{T}}Z - (Z^*)^{\mathrm{T}}Z^*}{2a(2-a-b)} \\
& = \frac{\gamma}{2}x_0^{\mathrm{T}}x_0 - \frac{\gamma}{2}(x^*)^{\mathrm{T}}x^* + \frac{1}{2}u_0^{\mathrm{T}}u_0 - \frac{1}{2}(u^*)^{\mathrm{T}}u^* + \frac{Z^{\mathrm{T}}Z}{2a(2-a-b)} - \frac{(Z^*)^{\mathrm{T}}Z^*}{2a(2-a-b)}.
\end{aligned} \tag{S41}$$

Substituting (S38)–(S40) into the above shows that (S41) can be expressed as a quadratic form in x_0, u_0 and Z:

$$(x_0^{\mathrm{T}}, u_0^{\mathrm{T}}, Z^{\mathrm{T}})G(\gamma)(x_0^{\mathrm{T}}, u_0^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}},$$

where $G(\gamma)$ is a 3×3 block matrix. For $i,j=1,2,3$, the (i,j)th block of $G(\gamma)$ is of the form $g_{ij}(\gamma)I$, where $g_{ij}(\gamma)$ is a scalar, polynomial of γ, with coefficients depending on (a,b,ϕ).

Now we compute the coefficients of the leading terms (terms corresponding to highest power of γ) of $g_{11}(\gamma)$, $g_{22}(\gamma)$ and $g_{33}(\gamma)$. Because we focus on only the leading terms, it is sufficient to examine (S38)–(S40) and account for the coefficients of x_0, u_0, Z, labeled as $(i),..,(v)$, which lead to the highest power of γ in $g_{11}(\gamma)$, $g_{22}(\gamma)$ and $g_{33}(\gamma)$. The coefficient of the leading term of $g_{11}(\gamma)$ associated with $x_0^{\mathrm{T}}x_0$ is

$$\begin{aligned}
&-\frac{(i)^2}{2}-\frac{(ii)^2}{2a(2-a-b)}=-\frac{1}{2}(a\phi)^2\gamma^4-\frac{(a^2-\phi a\sqrt{ab})^2\gamma^4}{2a(2-a-b)}\\
&=\frac{\gamma^4}{2(2-a-b)}(-a^2\phi^2(2-a-b)+2\phi a^2\sqrt{ab}-\phi^2a^2b-a^3)\\
&=\frac{\gamma^4a^2}{2(2-a-b)}(\phi^2(a-2)+\phi 2\sqrt{ab}-a). \qquad \text{(S42)}
\end{aligned}$$

The coefficient of the leading term of $g_{22}(\gamma)$ associated with $u_0^{\mathrm{T}}u_0$ is

$$\begin{aligned}
&-\frac{(iii)^2}{2}-\frac{(iv)^2}{2a(2-a-b)}=-\frac{1}{2}(\phi\sqrt{ab})^2\gamma^2-\frac{(\phi ab-a\sqrt{ab})^2\gamma^2}{2a(2-a-b)}\\
&=\frac{\gamma^2}{2(2-a-b)}(2\phi ab\sqrt{ab}-a^2b-\phi^2ab^2-2\phi^2ab+\phi^2a^2b+\phi^2ab^2)\\
&=\frac{\gamma^2ab}{2(2-a-b)}(\phi^2(a-2)+\phi 2\sqrt{ab}-a). \qquad \text{(S43)}
\end{aligned}$$

The coefficient of the leading term of $g_{33}(\gamma)$ associated with $Z^{\mathrm{T}}Z$ is

$$\begin{aligned}
&-\frac{(v)^2}{2}-\frac{(vi)^2}{2a(2-a-b)}=-\frac{1}{2}\phi^2\gamma^2-\frac{(\phi\sqrt{ab}-a)^2\gamma^2}{2a(2-a-b)}\\
&=\frac{\gamma^2}{2(2-a-b)}(2\phi\sqrt{ab}-a-\phi^2b-2\phi^2+a\phi^2+b\phi^2)\\
&=\frac{\gamma^2}{2(2-a-b)}(\phi^2(a-2)+\phi 2\sqrt{ab}-a). \qquad \text{(S44)}
\end{aligned}$$

Notice that (S42)–(S44) involve ϕ only through the same quadratic function of ϕ:

$$h(\phi)=\phi^2(a-2)+\phi 2\sqrt{ab}-a.$$

For $a>0, b\geq 0$ and $a+b\leq 2$, we have $h(\phi)\leq 0$, because $(2\sqrt{ab})^2+4a(a-2)=4a(a+b-2)\leq 0$. Hence $|h(\phi)|$ is minimized at $\phi=-2\sqrt{ab}/(2(a-2))=\sqrt{ab}/(2-a)$.

IV.6 Proof of Lemma 2

We use the following choice of A in (10): $a_1=2-\tilde{a}, a_2=\sqrt{\tilde{a}\tilde{b}}, a_3=2-\tilde{b}$ with the constraints on $\tilde{a},\tilde{b}$ that $\tilde{a}>0, \tilde{b}\geq 0$ and $\tilde{a}+\tilde{b}\leq 2$. The noise terms are proportional: $Z_2=-\sqrt{\tilde{b}/\tilde{a}}Z_1$. The new

noises Z_1^* and Z_2^*, defined by (15), (24), and (17), can be expressed in terms of $u_0, \nabla U(x_0), \nabla U(x^*)$ and Z_1 as

$$\begin{aligned}
Z_1^* &= \underbrace{\sqrt{\tilde{a}\tilde{b}}(\tilde{b} - \phi\sqrt{\tilde{a}\tilde{b}})}_{\theta_1} u_0 + \underbrace{(\tilde{a} + \tilde{a}\tilde{b} - 2 - \phi(\tilde{a}-1)\sqrt{\tilde{a}\tilde{b}})}_{\theta_2} \nabla U(x_0) \\
&\quad + \underbrace{(\tilde{a} - 2 + \phi\sqrt{\tilde{a}\tilde{b}})}_{\theta_3} \nabla U(x^*) + \underbrace{(\tilde{b} + 1 - \phi\sqrt{\tilde{a}\tilde{b}})}_{\theta_4} Z_1, \\
Z_2^* &= \underbrace{(2 - \tilde{b})(\tilde{b} - \phi\sqrt{\tilde{a}\tilde{b}})}_{\psi_1} u_0 + \underbrace{(\sqrt{\tilde{a}\tilde{b}}(1-\tilde{b}) - \phi(\tilde{a}-1)(2-\tilde{b}))}_{\psi_2} \nabla U(x_0) \\
&\quad + \underbrace{(-\sqrt{\tilde{a}\tilde{b}} + \phi(2-\tilde{b}))}_{\psi_3} \nabla U(x^*) + \underbrace{(\sqrt{\tilde{b}/\tilde{a}}(1-\tilde{b}) - \phi(2-\tilde{b}))}_{\psi_4} Z_1.
\end{aligned}$$

Suppose there exists $r \in \mathbb{R}$ such that $Z_2^* = rZ_1^*$ for arbitrary values of x_0, u_0 and Z_1. Then the coefficients, denoted above as $\theta_1, \ldots, \theta_4, \psi_1, \ldots, \psi_4$, satisfy

$$r\theta_1 = \psi_1, \quad r\theta_2 = \psi_2, \quad r\theta_3 = \psi_3, \quad r\theta_4 = \psi_4. \tag{S45}$$

We study the following possibilities.

First, suppose that $\theta_1 \neq 0$. Then $r = \frac{\psi_1}{\theta_1} = \frac{2-\tilde{b}}{\sqrt{\tilde{a}\tilde{b}}}$ by (S45). Substituting this into $r\theta_4 = \psi_4$ in (S45) yields

$$\begin{aligned}
&r\theta_4 = \psi_4 \Rightarrow \frac{2-\tilde{b}}{\sqrt{\tilde{a}\tilde{b}}}(\tilde{b} + 1 - \phi\sqrt{\tilde{a}\tilde{b}}) = \sqrt{\tilde{b}/\tilde{a}}(1-\tilde{b}) - \phi(2-\tilde{b}) \\
&\Rightarrow \frac{(2-\tilde{b})(\tilde{b}+1)}{\sqrt{\tilde{a}\tilde{b}}} = \sqrt{\tilde{b}/\tilde{a}}(1-\tilde{b}) \Rightarrow (2-\tilde{b})(\tilde{b}+1) = \tilde{b}(1-\tilde{b}) \\
&\Rightarrow \tilde{b} - \tilde{b}^2 + 2 = \tilde{b} - \tilde{b}^2 \Rightarrow 0 = 2,
\end{aligned}$$

which is a contradiction. Hence $\theta_1 = \psi_1 = 0$, which gives two possibilities: either $\tilde{b} = 0$ or $\phi = \sqrt{\tilde{b}/\tilde{a}}$.

Next suppose that $\tilde{b} = 0$. Then $\theta_4 = 1$ and $\psi_4 = -2\phi$, and hence $r = \psi_4/\theta_4 = -2\phi$ by (S45). Moreover, $\theta_2 = \tilde{a} - 2$ and $\psi_2 = -2\phi(\tilde{a}-1)$, and

$$r\theta_2 = \psi_2 \Rightarrow -2\phi(\tilde{a}-2) = -2\phi(\tilde{a}-1),$$

which implies that $\phi = 0$. Thus if $\tilde{b} = 0$, then $\phi = 0$ as well. This gives the trivial case that $r = 0$ and $Z_2^* \equiv 0$.

Finally suppose that $\phi = \sqrt{\tilde{b}/\tilde{a}}$. Then $Z_2^* = rZ_1^*$ is satisfied with $r = -\sqrt{\tilde{b}/\tilde{a}}$ by the following

calculation:

$$\theta_1 = \psi_1 = 0,$$
$$\theta_2 = \tilde{a} + \tilde{a}\tilde{b} - 2 - \tilde{b}(\tilde{a} - 1) = \tilde{a} + \tilde{b} - 2,$$
$$\psi_2 = \sqrt{\tilde{a}\tilde{b}}(1 - \tilde{b}) - \sqrt{\frac{\tilde{b}}{\tilde{a}}}(\tilde{a} - \tilde{b})(2 - \tilde{b}) = -\sqrt{\frac{\tilde{b}}{\tilde{a}}}(\tilde{a} + \tilde{b} - 2) = r\theta_2,$$
$$\theta_3 = \tilde{a} - 2 + \tilde{b}, \quad \psi_3 = -\sqrt{\tilde{a}\tilde{b}} + \sqrt{\frac{\tilde{b}}{\tilde{a}}}(2 - \tilde{b}) = -\sqrt{\frac{\tilde{b}}{\tilde{a}}}(\tilde{b} - 2 + \tilde{a}) = r\theta_3,$$
$$\theta_4 = \tilde{b} + 1 - \tilde{b} = 1, \quad \psi_4 = \sqrt{\frac{\tilde{b}}{\tilde{a}}}(1 - \tilde{b}) - \sqrt{\frac{\tilde{b}}{\tilde{a}}}(2 - \tilde{b}) = -\sqrt{\frac{\tilde{b}}{\tilde{a}}} = r\theta_4.$$

Therefore $Z_2^* = rZ_1^*$ if and only if $r = -\sqrt{\tilde{b}/\tilde{a}}$ and $\phi = \sqrt{\tilde{b}/\tilde{a}}$, which also includes the trivial case, $r = \phi = \tilde{b} = 0$.

IV.7 Proof of Lemma 3

By the Gaussian-calibrated rejection-free property, $(x_1, u_1) = (x^*, u^*)$ when the target density $\pi(x)$ is $\mathcal{N}(\mathbf{0}, I)$. We give a proof for HAMS-A and HAMS-B separately.

For HAMS-A, the lag-1 auto-covariance matrix is

$$C_A = \mathrm{Cov}((x^*, u^*), (x_0, u_0)) = \begin{pmatrix} (1-a)I & \sqrt{ab}I \\ -\sqrt{ab}I & (b-1)I \end{pmatrix}.$$

The eigenvalues of C_A are the eigenvalues of C_A with $I = 1$, each with multiplicities k. Henceforth we assume $I = 1$. The two eigenvalues of C_A are

$$\lambda_1 = \frac{1}{2}(b - a + \sqrt{\Delta}), \quad \lambda_2 = \frac{1}{2}(b - a - \sqrt{\Delta}),$$

where

$$\Delta = (a + b - 2)^2 - 4ab = \{2 - (\sqrt{a} - \sqrt{b})^2\}\{2 - (\sqrt{a} + \sqrt{b})^2\}.$$

Given $a \in (0, 2)$, we show that the choice of $b \in (0, 2 - a)$ which minimizes $\max(|\lambda_1|, |\lambda_2|)$ is $b^* = (\sqrt{2} - \sqrt{a})^2$, where $|\cdot|$ denotes the modulus. For this choice b^*, $\Delta = 0$ and the two eigenvalues are identical, $\lambda_1^* = \lambda_2^* = 1 - \sqrt{2a}$. We distinguish three cases.

(i) Suppose $(\sqrt{a} + \sqrt{b})^2 > 2$. Then λ_1 and λ_2 are complex, and

$$|\lambda_1|^2 = |\lambda_2|^2 = \lambda_1\lambda_2 = b + a - 1$$
$$> (\sqrt{2} - \sqrt{a})^2 + a - 1 = (\sqrt{2a} - 1)^2 = \lambda_1^{*2}.$$

(ii) Suppose $(\sqrt{a}+\sqrt{b})^2 < 2$ and $b \geq a$. Then $\lambda_1 (> 0)$ and λ_2 are real, and $\max(|\lambda_1|, |\lambda_2|) = \lambda_1$. For fixed a, the derivative of λ_1 with respect to b is

$$\frac{\mathrm{d}\lambda_1}{\mathrm{d}b} = \frac{1}{2}\left(1 + \frac{b-a-2}{\sqrt{\Delta}}\right) \leq \frac{1}{2}\frac{(2-a-b)+(b-a-2)}{\sqrt{\Delta}} = \frac{-a}{\sqrt{\Delta}} < 0,$$

where the first inequality uses $\sqrt{\Delta} \leq 2-a-b$. Then λ_1 is decreasing in b, which is upper-bounded by $b^* = (\sqrt{2}-\sqrt{a})^2$. Hence $\lambda_1 > \lambda_1^*$.

(iii) Suppose $(\sqrt{a}+\sqrt{b})^2 < 2$ and $b \leq a$. Then λ_1 and $\lambda_2 (< 0)$ are real, and $\max(|\lambda_1|, |\lambda_2|) = -\lambda_2$. For fixed a, the derivative of λ_2 with respect to b is

$$\frac{\mathrm{d}\lambda_2}{\mathrm{d}b} = \frac{1}{2}\left(1 - \frac{b-a-2}{\sqrt{\Delta}}\right) = \frac{1}{2}\left(1 + \frac{2+a-b}{\sqrt{\Delta}}\right) > 0.$$

Then λ_2 is increasing in b, which is upper-bounded by $\min(a, b^*)$. If $b^* \leq a$, then $|\lambda_2| = -\lambda_2 > -\lambda_2^* = |\lambda_2^*|$. If $a < b^*$, then $|\lambda_2| = -\lambda_2$ is greater than the value of $-\lambda_2$ corresponding $b = a$, which is identical to the value of λ_1 (due to $b = a$) and still greater than $|\lambda_1^*|$ by the conclusion from (ii).

Combining the three cases shows that $\max(|\lambda_1|, |\lambda_2|) \geq |\lambda_1^*| = |\lambda_2^*|$.

For HAMS-B, we work with equations (13)–(14) with $a_1 = 2-\tilde{a}, a_3 = 2-\tilde{b}$ and $a_2 = \sqrt{\tilde{a}\tilde{b}}$, that is, before the reparametrization from $(\tilde{a}, \tilde{b})$ to (a, b). Then the lag-1 auto-covariance matrix is

$$C_B = \mathrm{Cov}((x^*, u^*), (x_0, u_0)) = \begin{pmatrix} (-1+\tilde{a})I & \sqrt{\tilde{a}\tilde{b}}I \\ -\sqrt{\tilde{a}\tilde{b}}I & (1-\tilde{b})I \end{pmatrix}.$$

The two eigenvalues of C_B are

$$\lambda_1 = \frac{1}{2}(\tilde{a} - \tilde{b} + \sqrt{\Delta}), \quad \lambda_2 = \frac{1}{2}(\tilde{a} - \tilde{b} - \sqrt{\Delta}),$$

where $\Delta = (\tilde{a}+\tilde{b}-2)^2 - 4\tilde{a}\tilde{b}$. The two negative eigenvalues, $(-\lambda_1, -\lambda_2)$, depend on $(\tilde{a}, \tilde{b})$, in the same way as the two eigenvalues of C_A depend on (a, b) in the preceding proof for HAMS-A. Hence the maximum modulus of eigenvalues is also minimized by the choice $\tilde{b} = (\sqrt{2}-\sqrt{\tilde{a}})^2$ given $\tilde{a}$. By the reparametrization $\tilde{a} = 2-a$ and $\tilde{b} = ab/(2-a)$, the resulting choice of b given a is $b = \frac{a(2-a)}{(\sqrt{2}+\sqrt{2-a})^2}$.

IV.8 Simplification of preconditioning for Algorithm 3

As discussed in Section 3.6 for preconditioning, we apply the linear transformations $\tilde{x} = L^{\mathrm{T}}x$ and $\nabla U(\tilde{x}) = L^{-1}\nabla U(x)$ to HAMS-A/B in Algorithm 2. We show the the resulting algorithm, stated as Algorithm 4 here, can be rearranged in an equivalent but computationally more efficient form as Algorithm 3.

Algorithm 4: HAMS-A/HAMS-B (with preconditioning non-simplified)

Initialize x_0, u_0

for $t = 0, 1, 2, ..., N_{iter}$ **do**

Sample $w \sim \text{Uniform}[0,1]$ and $\zeta \sim \mathcal{N}(\mathbf{0}, I)$

Transform $\tilde{x}_t = L^{\mathrm{T}} x_t$

$\tilde{x}^* = \tilde{x}_t - aL^{-1}\nabla U(x_t) + \sqrt{ab}u_t + \sqrt{a(2-a-b)}\zeta$

Propose $x^* = (L^{\mathrm{T}})^{-1}\tilde{x}^*$

if *HAMS-A* **then**

Propose $u^* = \left(\frac{2b}{2-a} - 1\right) u_t - \frac{\sqrt{ab}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*)) + \frac{2\sqrt{b(2-a-b)}}{2-a}\zeta$

$\zeta^* = \left(1 - \frac{2b}{2-a}\right)\zeta - \frac{\sqrt{a(2-a-b)}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*)) + \frac{2\sqrt{b(2-a-b)}}{2-a} u_t$

if *HAMS-B* **then**

Propose $u^* = u_t - \frac{\sqrt{ab}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*))$

$\zeta^* = \zeta - \frac{\sqrt{a(2-a-b)}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*))$

$\rho = \exp\left\{H(x_t, u_t) - H(x^*, u^*) + \frac{1}{2}\zeta^{\mathrm{T}}\zeta - \frac{1}{2}(\zeta^*)^{\mathrm{T}}\zeta^*\right\}$

if $w < \min(1, \rho)$ **then**

$(x_{t+1}, u_{t+1}) = (x^*, u^*)$ # Accept

else

$(x_{t+1}, u_{t+1}) = (x_t, -u_t)$ # Reject

Suppose that the equivalence holds for (x_t, u_t). By the relation $\nabla U(\tilde{x}_t) = L^{-1}\nabla U(x_t)$ and the definition of ξ in Algorithm 3, we have

$$\begin{aligned}\tilde{x}^* &= \tilde{x}_t - a\nabla U(\tilde{x}_t) + \xi \\ &= \tilde{x}_t - aL^{-1}\nabla U(x_t) + \sqrt{ab}u_t + \sqrt{a(2-a-b)}\zeta.\end{aligned}$$

Hence, when the proposal is accepted, $x_{t+1} = x^* = (L^{\mathrm{T}})^{-1}\tilde{x}^*$ in both algorithms. By the relation $\tilde{\xi} = \nabla U(\tilde{x}) + L^{-1}\nabla U(x^*) = L^{-1}(\nabla U(x_t) + \nabla U(x^*))$, we see that when the proposal is accepted, the expressions of u_{t+1} are the same in both algorithms. When the proposal is rejected, $(x_{t+1}, u_{t+1}) = (x_t, -u_t)$ is also the same in the two algorithms.

To show the equivalence holds for (x_{t+1}, u_{t+1}), it remains to check that the acceptance probabilities are equal in the two algorithms. We need to show

$$U(x_t) - U(x^*) + \frac{1}{2-a}(\tilde{\xi})^{\mathrm{T}}(\xi - \frac{a}{2}\tilde{\xi}) = H(x_t, u_t) - H(x^*, u^*) + \frac{1}{2}\zeta^{\mathrm{T}}\zeta - \frac{1}{2}(\zeta^*)^{\mathrm{T}}\zeta^*,$$

which is equivalent to

$$\frac{2}{2-a}(\tilde{\xi})^{\mathrm{T}}(\xi - \frac{a}{2}\tilde{\xi}) = u_t^{\mathrm{T}}u_t - (u^*)^{\mathrm{T}}u^* + \zeta^{\mathrm{T}}\zeta - (\zeta^*)^{\mathrm{T}}\zeta^*,$$

because $H(x_t, u_t) - H(x^*, u^*) = U(x_t) - U(x^*) + \frac{1}{2}u_t^{\mathrm{T}}u_t - \frac{1}{2}(u^*)^{\mathrm{T}}u^*$.

Consider the algorithm HAMS-B. We use the following fact

$$u_t^{\mathrm{T}} u_t - (u^*)^{\mathrm{T}} u^* = (u_t - u^*)^{\mathrm{T}}(u_t + u^*), \quad \zeta^{\mathrm{T}}\zeta - (\zeta^*)^{\mathrm{T}}\zeta^* = (\zeta_t - \zeta^*)^{\mathrm{T}}(\zeta_t + \zeta^*). \tag{S46}$$

By direct calculation, we have

$$u_t - u^* = \frac{\sqrt{ab}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*)) = \frac{\sqrt{ab}}{2-a}\tilde{\xi}, \tag{S47}$$

$$(u_t - u^*)^{\mathrm{T}}(u_t + u^*) = \frac{\sqrt{ab}}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(2u_t - \frac{\sqrt{ab}}{2-a}\tilde{\xi}\right), \tag{S48}$$

and

$$\zeta - \zeta^* = \frac{\sqrt{a(2-a-b)}}{2-a} L^{-1}(\nabla U(x_t) + \nabla U(x^*)) = \frac{\sqrt{a(2-a-b)}}{2-a}\tilde{\xi}, \tag{S49}$$

$$(\zeta - \zeta^*)^{\mathrm{T}}(\zeta + \zeta^*) = \frac{\sqrt{a(2-a-b)}}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(2\zeta - \frac{\sqrt{a(2-a-b)}}{2-a}\tilde{\xi}\right). \tag{S50}$$

Combining (S46)–(S49) yields

$$\begin{aligned}
&u_t^{\mathrm{T}} u_t - (u^*)^{\mathrm{T}} u^* + \zeta^{\mathrm{T}}\zeta - (\zeta^*)^{\mathrm{T}}\zeta^* \\
&= (\tilde{\xi})^{\mathrm{T}}\left(\frac{2\sqrt{ab}}{2-a}u_t + \frac{2\sqrt{a(2-a-b)}}{2-a}\zeta - \left(\frac{ab}{(2-a)^2} + \frac{a(2-a-b)}{(2-a)^2}\right)\tilde{\xi}\right) \\
&= \frac{2}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(\sqrt{ab}u_t + \sqrt{a(2-a-b)}\zeta - \frac{a}{2}\tilde{\xi}\right) \\
&= \frac{2}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(\xi - \frac{a}{2}\tilde{\xi}\right).
\end{aligned} \tag{S51}$$

Hence the acceptance probabilities match for HAMS-B in Algorithms 3 and 4.

Finally consider the algorithm HAMS-A. Define intermediate variables

$$\begin{aligned}
u^\dagger &= \left(\frac{2b}{2-a} - 1\right)u_t + \frac{2\sqrt{b(2-a-b)}}{2-a}\zeta, \\
\zeta^\dagger &= \left(1 - \frac{2b}{2-a}\right)\zeta + \frac{2\sqrt{b(2-a-b)}}{2-a}u_t.
\end{aligned}$$

Then the following identities hold:

$$(u^\dagger)^{\mathrm{T}}u^\dagger + (\zeta^\dagger)^{\mathrm{T}}\zeta^\dagger = u_t^{\mathrm{T}}u_t + \zeta^{\mathrm{T}}\zeta, \tag{S52}$$

$$\sqrt{ab}u^\dagger + \sqrt{a(2-a-b)}\zeta^\dagger = \sqrt{ab}u_t + \sqrt{a(2-a-b)}\zeta\,(=\xi). \tag{S53}$$

Identity (S52) follows, because after expanding the inner products on the left hand side, the cross terms cancel out and the squared terms have coefficients

$$\left(\frac{2b}{2-a} - 1\right)^2 + \left(\frac{2\sqrt{b(2-a-b)}}{2-a}\right)^2 = 1.$$

Identity (S53) follows because by direct calculation

$$u^\dagger - u_t = \frac{2\sqrt{2-a-b}}{2-a}(\sqrt{2-a-b}\, u_t + \sqrt{b}\zeta),$$
$$\zeta^\dagger - \zeta = \frac{2\sqrt{b}}{2-a}(-\sqrt{b}\zeta + \sqrt{2-a-b}\, u_t).$$

Moreover, it can be verified by definition that

$$u^\dagger - u^* = \frac{\sqrt{ab}}{2-a}\tilde{\xi}, \quad \zeta^\dagger - \zeta^* = \frac{\sqrt{a(2-a-b)}}{2-a}\tilde{\xi}.$$

Then (S46)–(S51) remain valid with u_t and ζ replaced by $u^\dagger$ and $\zeta^\dagger$. From these equations together with the identities (S52)–(S53), we find

$$\begin{aligned}
&u_t^{\mathrm{T}} u_t - (u^*)^{\mathrm{T}} u^* + \zeta^{\mathrm{T}}\zeta - (\zeta^*)^{\mathrm{T}}\zeta^* \\
&= (u^\dagger)^{\mathrm{T}} u^\dagger - (u^*)^{\mathrm{T}} u^* + (\zeta^\dagger)^{\mathrm{T}}\zeta^\dagger - (\zeta^*)^{\mathrm{T}}\zeta^* \\
&= \frac{2}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(\sqrt{ab}u^\dagger + \sqrt{a(2-a-b)}\zeta^\dagger - \frac{a}{2}\tilde{\xi}\right) \\
&= \frac{2}{2-a}(\tilde{\xi})^{\mathrm{T}}\left(\xi - \frac{a}{2}\tilde{\xi}\right).
\end{aligned}$$

Hence the acceptance probabilities match for HAMS-A in Algorithms 3 and 4.

V Details for simulation studies

V.1 Expressions for multilevel logistic regression

Consider the multilevel logistic regression described in Section 5.1. The latent variables (or random effects) are $\mathbf{x} = (x_1, x_2, ..., x_{78})^{\mathrm{T}}$, and the parameters are $\theta = (\beta^{\mathrm{T}}, \sigma^{\mathrm{T}})^{\mathrm{T}}$ with the fixed effects $\beta = (\beta_1, \ldots, \beta_5)^{\mathrm{T}}$ and standard-deviation components $\sigma = (\sigma_1, \ldots, \sigma_5)^{\mathrm{T}}$. Let D_1 be the model matrix corresponding to the fixed effects β and D_2 that of the random effects $\mathbf{x}$. The model can be equivalently written as

$$P(y_i = 1) = \mathrm{expit}(\eta_i), \quad i = 1, ..., n\,(= 2015),$$
$$\text{the vector of linear predictors is } \eta = D_1\beta + D_2\mathbf{x}.$$

For convenience, denote $\eta^{(1)} = D_1\beta$ and $\eta^{(2)} = D_2\mathbf{x}$. The latent variables are marginally normal $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, C)$, with covariance matrix

$$C = \mathrm{diag}\left(\underbrace{\sigma_1^2, ..., \sigma_1^2}_{4}, \underbrace{\sigma_2^2, ..., \sigma_2^2}_{4}, \underbrace{\sigma_3^2, ..., \sigma_3^2}_{16}, \underbrace{\sigma_4^2, ..., \sigma_4^2}_{49}, \underbrace{\sigma_5^2, ..., \sigma_5^2}_{5}\right).$$

The priors on the parameters are $\beta \sim \mathcal{N}(\mathbf{0}, V)$, $V = \text{diag}(10^4, 10^4, 10^4, 10^4, 10^4)$, and $\sigma_j \propto 1$ for $j = 1, ..., 5$. Then the joint density is given by

$$\begin{aligned} p(\mathbf{x}, \theta|\mathbf{y}) &\propto \pi(\theta) \cdot \mathcal{N}(\mathbf{x}|\mathbf{0}, C) \cdot p(y|\mathbf{x}, \theta) \\ &\propto \exp\left(-\frac{1}{2}\beta^{\mathrm{T}} V^{-1}\beta\right) |\det(C)|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}} C^{-1}\mathbf{x}\right) \exp\left[\sum_{i=1}^{n}\{y_i\eta_i - \log(1+e^{\eta_i})\}\right]. \end{aligned}$$

For latent variable sampling from $p(\mathbf{x}|\mathbf{y}, \theta)$, the potential function is

$$U(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}} C^{-1}\mathbf{x} - \sum_{i=1}^{n}\{y_i\eta_i^{(2)} - \log(1+e^{\eta_i})\},$$

where the dependency on $(\mathbf{y}, \theta)$ is suppressed. The gradient and Hessian are

$$\begin{aligned} \nabla U(\mathbf{x}) &= C^{-1}\mathbf{x} - D_2^{\mathrm{T}}\{\mathbf{y} - \text{expit}(\eta)\}, \\ \nabla^2 U(\mathbf{x}) &= C^{-1} + D_2^{\mathrm{T}}\text{diag}\{\text{expit}(\eta)(1 - \text{expit}(\eta))\}D_2. \end{aligned}$$

Here expit(η) is evaluated element-wise. As mentioned in Section 5.1, the preconditioning matrix for all methods (except mGrad) is fixed at $\Sigma^{-1} = M = \nabla^2 U(\mathbf{0})$. For mGrad, the C matrix is the prior variance used in (3).

The conditional density of parameters given $(\mathbf{y}, \mathbf{x})$ is

$$p(\beta, \sigma|\mathbf{y}, \mathbf{x}) = \exp\left[-\frac{1}{2}\beta^{\mathrm{T}} V^{-1}\beta + \sum_{i=1}^{n}\{y_i\eta_i^{(1)} - \log(1+e^{\eta_i})\}\right] |\det(C)|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}} C^{-1}\mathbf{x}\right).$$

Notice that C is only a function of σ. Then β and σ are independent given $(\mathbf{y}, \mathbf{x})$. The conditional density of σ is

$$\begin{aligned} &p(\sigma|\mathbf{y}, \mathbf{x}) \propto \det(C)|^{-1/2} \exp(-x^T C^{-1} x/2) \\ &= \sigma_1^{-4}\sigma_2^{-4}\sigma_3^{-16}\sigma_4^{-49}\sigma_5^{-5} \exp\left(\frac{-\frac{1}{2}\sum_{k=1}^{4} x_k^2}{\sigma_1^2} + \frac{-\frac{1}{2}\sum_{k=5}^{8} x_k^2}{\sigma_2^2} + \frac{-\frac{1}{2}\sum_{k=9}^{2} 4x_k^2}{\sigma_3^2}\right. \\ &\quad \left. + \frac{-\frac{1}{2}\sum_{k=25}^{7} 3x_k^2}{\sigma_4^2} + \frac{-\frac{1}{2}\sum_{k=74}^{7} 8x_k^2}{\sigma_5^2}\right). \end{aligned}$$

Using a density transformation $\sigma \to \sigma^2$, we obtain

$$\begin{aligned} p(\sigma^2|\mathbf{y}, \mathbf{x}) \propto (\sigma_1^2)^{-5/2}(\sigma_2^2)^{-5/2}(\sigma_3^2)^{-17/2}(\sigma_4^2)^{-25}(\sigma_5^2)^{-3} \exp\left(\frac{-\frac{1}{2}\sum_{k=1}^{4} x_k^2}{\sigma_1^2} + \frac{-\frac{1}{2}\sum_{k=5}^{8} x_k^2}{\sigma_2^2}\right. \\ \left. + \frac{-\frac{1}{2}\sum_{k=9}^{2} 4x_k^2}{\sigma_3^2} + \frac{-\frac{1}{2}\sum_{k=25}^{7} 3x_k^2}{\sigma_4^2} + \frac{-\frac{1}{2}\sum_{k=74}^{7} 8x_k^2}{\sigma_5^2}\right). \end{aligned}$$

Therefore, given $(\mathbf{y}, \mathbf{x})$, each σ_j^2 follows inverse-Gamma distribution:

$$\sigma_1^2 \sim \text{InvGamma}(a = \frac{3}{2}, b = \frac{1}{2}\sum_{k=1}^{4} x_k^2), \quad \sigma_2^2 \sim \text{InvGamma}(a = \frac{3}{2}, b = \frac{1}{2}\sum_{k=5}^{8} x_k^2),$$
$$\sigma_3^2 \sim \text{InvGamma}(a = \frac{15}{2}, b = \frac{1}{2}\sum_{k=9}^{24} x_k^2), \quad \sigma_2^4 \sim \text{InvGamma}(a = 24, b = \frac{1}{2}\sum_{k=25}^{73} x_k^2),$$
$$\sigma_5^2 \sim \text{InvGamma}(a = 2, b = \frac{1}{2}\sum_{k=74}^{78} x_k^2).$$

Here we use the parameterization:

$$X \sim \text{InvGamma}(a, b) \iff p_X(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-b/x).$$

The conditional density of β is

$$p(\beta|\mathbf{y}, \mathbf{x}) \propto \exp\left[-\frac{1}{2}\beta^{\mathrm{T}} V^{-1}\beta + \sum_{i=1}^{n}\{y_i \eta_i^{(1)} - \log(1 + e^{\eta_i})\}\right].$$

The potential function, gradient, and Hessian are

$$U(\beta) = \frac{1}{2}\beta^{\mathrm{T}} V^{-1}\beta - \sum_{i=1}^{n}\{y_i \eta_i^{(1)} - \log(1 + e^{\eta_i})\},$$
$$\nabla U(\beta) = V^{-1}\beta - D_1^{\mathrm{T}}(\mathbf{y} - \text{expit}(\eta)),$$
$$\nabla^2 U(\beta) = V^{-1} + D_1^{\mathrm{T}}\text{diag}\{\text{expit}(\eta)(1 - \text{expit}(\eta))\}D_1,$$

where the dependency on $(\mathbf{y}, \mathbf{x})$ is suppressed. The preconditioning matrix we use for all methods (except mGrad) is

$$\Sigma^{-1} = M = V^{-1} + D_1^{\mathrm{T}}\text{diag}[\text{expit}(\eta^{(1)})(1 - \text{expit}(\eta^{(1)}))]D_1,$$

obtained by evaluating $\nabla^2 U(\beta)$ with $\mathbf{x} = \mathbf{0}$ and β at some fixed value. For mGrad, we use V in place of the prior variance C in (3).

With all the expressions above, we perform Gibbs sampling with three blocks, $p(\mathbf{x}|\mathbf{y}, \beta, \sigma)$, $p(\beta|\mathbf{y}, \mathbf{x})$, and $p(\sigma|\mathbf{y}, \mathbf{x})$. The density $p(\sigma|\mathbf{y}, \mathbf{x})$ is sampled using the R package `invgamma`, whereas $p(\mathbf{x}|\mathbf{y}, \beta, \sigma)$ and $p(\beta|\mathbf{y}, \mathbf{x})$ are sampled using MCMC.

V.2 Expressions for stochastic volatility model

As used in the experiments in Section VI.3, the stochastic volatility model is defined as

$$x_t = \phi x_{t-1} + \eta_t, \quad t = 2, ..., T, \quad x_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1 - \phi^2}\right),$$
$$y_t = z_t \beta \exp(x_t/2), \quad z_t \overset{iid}{\sim} \mathcal{N}(0, 1), \quad \eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad t = 1, ..., T.$$

Denote $\mathbf{x} = (x_1, ..., x_T)^{\mathrm{T}}, \mathbf{y} = (y_1, ..., y_T)^{\mathrm{T}}, \mathbf{z} = (z_1, ..., z_T)^{\mathrm{T}}$ and $\theta = (\beta, \sigma, \phi)^{\mathrm{T}}$. The joint density of $(\mathbf{x}, \mathbf{y}, \theta)$ is

$$p(\mathbf{x}, \mathbf{y}, \theta) = \pi(\theta) \cdot \underbrace{p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1}, \phi, \sigma)}_{\mathcal{N}(\mathbf{x}|\mathbf{0}, C)} \cdot \overbrace{\prod_{t=1}^{T} p(y_t | x_t, \beta)}^{\mathcal{N}(\mathbf{y}|\mathbf{0}, \beta^2 \exp(\mathbf{x}))}$$

$$\propto \pi(\theta) |\det(C)|^{-1/2} \exp\left\{ -\frac{1}{2} \mathbf{x}^{\mathrm{T}} C^{-1} \mathbf{x} \right\} \beta^{-T} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (x_t + \beta^{-2} y_t^2 \exp(-x_t)) \right\}.$$

The matrix C and its inverse are given by

$$C = \frac{\sigma^2}{1 - \phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-2} & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-3} & \phi^{T-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{T-4} & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi^{T-2} & \phi^{T-3} & \phi^{T-4} & \cdots & 1 & \phi \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & \phi & 1 \end{pmatrix}$$

$$\iff C^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 & 0 \\ -\phi & 1 + \phi^2 & -\phi & \cdots & 0 & 0 \\ 0 & -\phi & 1 + \phi^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \phi^2 & -\phi \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{pmatrix}.$$

The conditional posterior of the latent variables is

$$p(\mathbf{x} | \mathbf{y}, \theta) \propto \exp\left\{ -\frac{1}{2} \mathbf{x}^{\mathrm{T}} C^{-1} \mathbf{x} \right\} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (x_t + \beta^{-2} y_t^2 \exp(-x_t)) \right\}.$$

Then the negative log-density (or potential function) is

$$U(\mathbf{x}) = \frac{1}{2} \mathbf{x}^{\mathrm{T}} C^{-1} \mathbf{x} + \frac{1}{2} \sum_{t=1}^{T} (x_t + \beta^{-2} y_t^2 \exp(-x_t)),$$

where dependency on $(\mathbf{y}, \theta)$ is suppressed in the notation. The gradient is

$$\nabla U(\mathbf{x}) = C^{-1} \mathbf{x} - \frac{1}{2} \beta^{-2} \mathbf{y} \exp(-\mathbf{x}) + \frac{1}{2} \mathbf{1},$$

where $\mathbf{1}$ is a vector of all 1's. The Hessian is

$$\nabla^2 U(\mathbf{x}) = C^{-1} + \frac{1}{2} \mathrm{diag}[\beta^{-2} \mathbf{y}^2 \exp(-\mathbf{x})].$$

The square $\mathbf{y}^2$ is taken component-wise. Using the relation between $\mathbf{y}$ and $\mathbf{x}$, the diagonal elements in the second term can be expressed as

$$\beta^{-2}\mathbf{y}^2\exp(-\mathbf{x}) = \beta^{-2}\exp(-\mathbf{x})\mathbf{z}^2\beta^2\exp(\mathbf{x}) = \mathbf{z}^2.$$

Hence

$$\mathrm{E}[\nabla^2 U(\mathbf{x})] = C^{-1} + \frac{1}{2}I,$$

which leads to the preconditioning in Section VI.3. The expectation above is taken over the marginal distribution of $\mathbf{z}$.

For the parameters, the priors are

$$\pi(\beta) \propto \beta^{-1}, \quad \sigma^2 \sim \text{Inv-}\chi^2(10, 0.05), \quad \frac{\phi+1}{2} \sim \text{Beta}(20, 1.5).$$

Then σ and ϕ are also transformed by $\sigma = \exp(\gamma)$ and $\phi = \tanh(\alpha)$. The resulting potential for the transformed parameters is

$$U(\beta,\alpha,\gamma) = (T+1)\log\beta - 20.5\log(1+\tanh\alpha) - 2\log(1-\tanh\alpha)\frac{1}{2}\mathbf{x}^{\mathrm{T}}C^{-1}\mathbf{x} + \frac{1}{2}\sum_{t=1}^{T}\beta^{-2}y_t^2\exp(-x_t),$$

where dependency on $(\mathbf{y}, \mathbf{x})$ is suppressed in the notation. The gradient is

$$\begin{aligned}
\frac{\partial U(\beta,\alpha,\gamma)}{\partial\beta} &= \frac{T+1}{\beta} - \frac{\sum_{t=1}^{T} y_t^2\exp(-x_t)}{\beta^3}, \\
\frac{\partial U(\beta,\alpha,\gamma)}{\partial\alpha} &= 22.5\tanh\alpha - 18.5 - \exp(-2\gamma)x_1^2\tanh\alpha(1-\tanh^2\alpha), \\
&\quad - \exp(-2\gamma)\sum_{t=2}^{T}(x_t - \tanh\alpha x_{t-1})x_{t-1}(1-\tanh^2\alpha), \\
\frac{\partial U(\beta,\alpha,\gamma)}{\partial\gamma} &= -\mathbf{x}^{\mathrm{T}}C^{-1}\mathbf{x} - \frac{1}{2}\exp(-2\gamma) + 10 + T.
\end{aligned}$$

Finally the expected Hessian computed with respect to the marginals of $\mathbf{x}$ and $\mathbf{z}$ is

$$\mathrm{E}[\nabla^2 U(\beta,\alpha,\gamma)] = \begin{pmatrix} (2T-1)/\beta & 0 & 0 \\ 0 & \exp(-2\gamma)+2T & 2\tanh\alpha \\ 0 & 2\tanh\alpha & 21.5 - 19.5\tanh^2\alpha + (T-1)(1-\tanh^2\alpha) \end{pmatrix}.$$

When sampling the parameters, we use $M = \Sigma^{-1} = \mathrm{E}[\nabla^2 U(\beta,\alpha,\gamma)]$ for preconditioning.

V.3 Expressions for log-Gaussian Cox model

Denote $\mathbf{x} = (x_{ij}), \mathbf{y} = (y_{ij}), i, j = 1, ..., m$ and let C be the matrix corresponding to the covariance function as described in Section 5.2. The joint posterior density is

$$p(\mathbf{x}, \sigma^2, \beta|\mathbf{y}) \propto$$
$$\pi(\sigma^2)\pi(\beta)(det|C|)^{-1/2} \exp\left\{-\frac{1}{2}x^{\mathrm{T}}C^{-1}x\right\} \exp\left\{\sum_{i,j}(y_{ij}(x_{ij}+\mu) - n^{-1}\exp(x_{ij}+\mu))\right\}.$$

The potential function from the conditional posterior of the latent variables given $(\mathbf{y}, \sigma^2, \beta)$ is

$$U(\mathbf{x}) = \frac{1}{2}x^{\mathrm{T}}C^{-1}x - \sum_{i,j}(y_{ij}x_{ij} - n^{-1}\exp(x_{ij}+\mu)),$$

where dependency on $(\mathbf{y}, \sigma^2, \beta)$ is suppressed in the notation. The gradient is

$$\nabla U(\mathbf{x}) = C^{-1}\mathbf{x} - \mathbf{y} + n^{-1}\exp(\mathbf{x}+\mu).$$

The Hessian is

$$\nabla^2 U(\mathbf{x}) = C^{-1} + n^{-1}\mathrm{diag}[\mathbf{x}+\mu].$$

Because marginally $\mathbf{x} \sim \mathcal{N}(0, C)$, we take the expectation

$$\mathrm{E}[\nabla^2 U(\mathbf{x})] = C^{-1} + n^{-1}\mathrm{diag}[\sigma^2/2 + \mu],$$

which is used for preconditioning in Section 5.2.

For the parameters, we use the priors $\sigma^2 \sim \mathrm{Gamma}(2, 0.5)$ and $\beta \sim \mathrm{Gamma}(2, 0.5)$ and the transformations $\sigma^2 = \exp(\varphi_1), \beta = \exp(\varphi_2)$. Then the potential function from the conditional posterior of transformed parameters given $(\mathbf{y}, \mathbf{x})$ is

$$U(\varphi_1, \varphi_2) = \frac{1}{2}(\exp(\varphi_1) + \exp(\varphi_2)) - 2(\varphi_1 + \varphi_2) + \frac{1}{2}\mathbf{x}^{\mathrm{T}}C^{-1}\mathbf{x} + \frac{1}{2}\log\det(C),$$

where dependency on $(\mathbf{y}, \mathbf{x})$ is suppressed in the notation. The gradient is

$$\frac{\partial U(\varphi_1, \varphi_2)}{\partial \varphi_1} = \frac{\exp(\varphi_1)}{2} - 2 + \frac{n}{2} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}C^{-1}\mathbf{x},$$

$$\frac{\partial U(\varphi_1, \varphi_2)}{\partial \varphi_2} = \frac{\exp(\varphi_2)}{2} - 2 + \frac{1}{2}\mathrm{tr}\left(\frac{\partial C}{\partial \varphi_2}\right) - \frac{1}{2}\mathbf{x}^{\mathrm{T}}C^{-1}\frac{\partial C}{\partial \varphi_2}C^{-1}\mathbf{x},$$

where

$$\frac{\partial C}{\partial \varphi_2}[(i,j),(i',j')] =$$
$$m^{-1}\exp(\varphi_1)\exp(-\varphi_2)\sqrt{(i-i')^2 + (j-j')^2}\exp(-\sqrt{(i-i')^2 + (j-j')^2}/(m\exp(\varphi_2))).$$

The marginal expected Hessian is

$$\mathrm{E}[\nabla^2 U(\varphi_1, \varphi_2)] = \begin{pmatrix} \frac{1}{2}(\exp(\varphi_1) + n) & \frac{1}{2}\mathrm{tr}(C^{-1}\frac{\partial C}{\partial \varphi_2}) \\ \frac{1}{2}\mathrm{tr}(C^{-1}\frac{\partial C}{\partial \varphi_2}) & \frac{1}{2}(\exp(\varphi_1) + \mathrm{tr}(C^{-1}\frac{\partial C}{\partial \varphi_2}C^{-1}\frac{\partial C}{\partial \varphi_2})) \end{pmatrix}.$$

When sampling the parameters, we use $M = \Sigma^{-1} = \mathrm{E}[\nabla^2 U(\varphi_1, \varphi_2)]$ for preconditioning.

V.4 Step size tuning

As mentioned in Section 5, we periodically adjust step size ϵ based on the acceptance rate during the burn-in period. When acceptance is too low (smaller than a lower threshold), we decrease ϵ by the mapping $\epsilon \leftarrow \max(1 - \sqrt{1-\epsilon}, \frac{\epsilon}{1+\delta})$; when acceptance is too high (larger than a upper threshold), we increase ϵ by the mapping $\epsilon \leftarrow \epsilon + \epsilon \cdot \min(1-\epsilon, \delta)$, where δ is an adjustment value taken to be $\delta = 0.2$ in all our simulations. The increase and decrease mappings are, by design, inverse of each other, as illustrated in Figure S1. The two mappings are mostly linear, but are curved when ϵ is close to 1 to ensure that ϵ is always between 0 and 1 after the update.

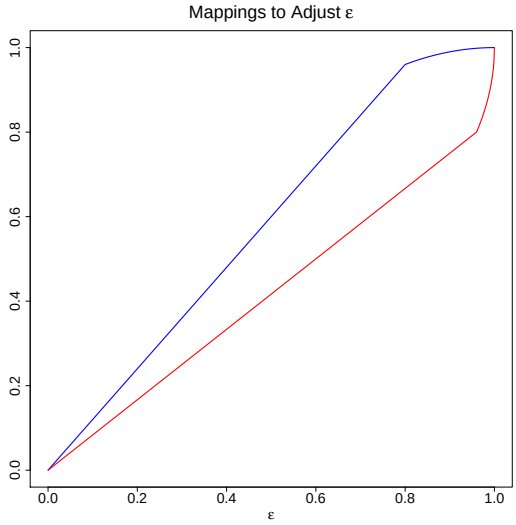


Figure S1: Tuning of step size ϵ with $\delta = 0.2$. Blue curve is mapping used to increase ϵ. Red curve is mapping used to decrease ϵ

VI Additional simulation results

We present experiments with a multivariate normal distribution and a stochastic volatility model, and additional simulation results including pMALA*, GMC and mGrad from the experiments with the multilevel logistic regression and log-Gaussian Cox model.

VI.1 Multivariate normal distribution

Consider the problem of sampling from a 100 dimensional normal distribution with high correlations: $\pi(x) = \mathcal{N}(\mathbf{0}, C)$ where the entries of C are

$$C[i,j] = 0.9^{|i-j|}, \quad i,j = 1, ..., 100.$$

We do not employ any preconditioning here, although we still refer to pMALA and pMALA* as such. This experiment is used to compare different algorithms when the variance of the target distribution may not be readily approximated. Hence potential advantages associated with the rejection-free property are removed from HAMS-A/B.

In terms of tuning, we set $\epsilon = 0.19$ for HAMS-A, HAMS-B, UDL, GMC, pMALA and pMALA* to maintain acceptance rates around 70%. Through empirical trials we find that HAMS-A, UDL and GMC have good performance using a large carryover (c value), while HAMS-B favors a relatively small carryover. Hence we set $c = 0.95$ for HAMS-A, UDL and GMC, $c = 0.25$ for HAMS-B. For HMC, we set $nleap = 50$ and $\epsilon = 0.17$ which also yields a 70% acceptance rate. For RWM, we set $\epsilon = 0.06$ and the resulting acceptance is around 40%. To account for the additional computation cost due to leapfrog steps, HMC is run for 200 iterations and all other methods are run for $200 \times 50 = 10000$ iterations. The simulation process is repeated for 100 times with a fixed starting value of $\mathbf{0}$.

Figure S2 shows boxplots of sample means and variances of 100 coordinates and sample covariances of 100 coordinates with the first coordinate after centered about the true values. Hence deviations from 0 (marked by red lines) show divergence from the truth. From the boxplots, we see that HAMS-A, UDL and GMC are comparable to each other. They are mostly accurate in the means and covariances while slightly underestimate the variances. Sample means of HAMS-B are correctly centered but exhibit more variation. HAMS-B underestimates the variances more than HAMS-A, UDL, and GMC, and also the covariances associated with the first several coordinates. Compared to HAMS-B, pMALA shows similar underestimation of variances and covariances, but has an even wider spread in sample means. For pMALA*, because $\epsilon = 0.18$ is small, its performance is similar to that of the unmodified pMALA. While HMC is good in terms of sample means, it underestimates variances and is inaccurate in covariances with a considerable number of outliers. RWM performs poorly to capture neither variance nor covariance.

Figure S3 shows trace plots of first 2000 iterations (first 40 iterations for HMC) from an individual run. The first two coordinates are plotted and red ellipses mark regions containing 95% probability of the marginal target density. HAMS-A best fills up the area. UDL and GMC are also reasonable but leave a small part in the upper right blank. HAMS-B, pMALA and pMALA* all cover smaller

areas with parts of the corners missing. The HMC trace misses the top right quadrant and its movement is only aligned to the long axis of the ellipse. RWM performs poorly and covers the least amount of the area.

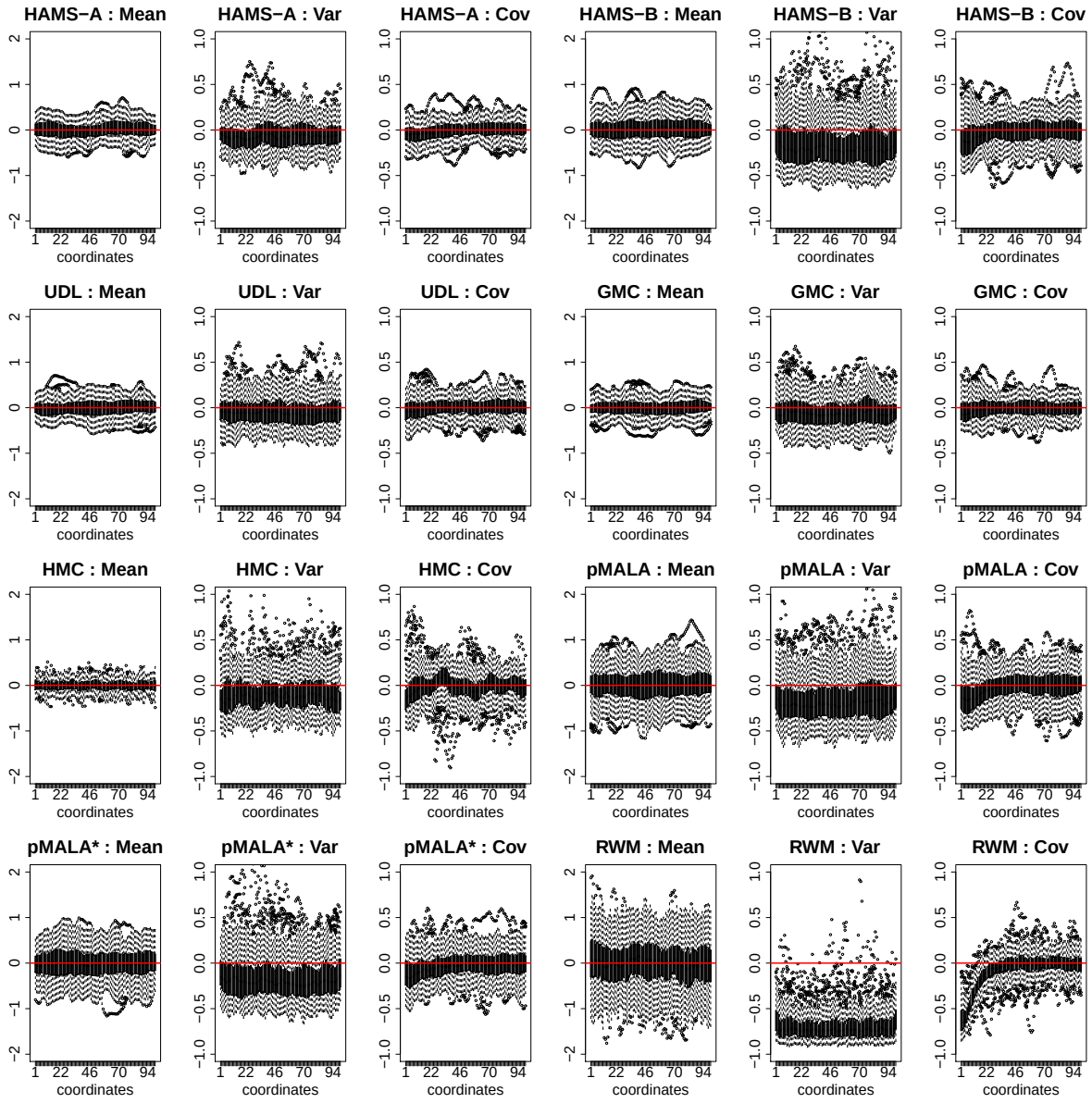


Figure S2: Time-adjusted and centered boxplots of sample means, variances, and covariances of 100 coordinates over 100 repetitions for sampling from the multivariate normal distribution. Red lines indicate zero.

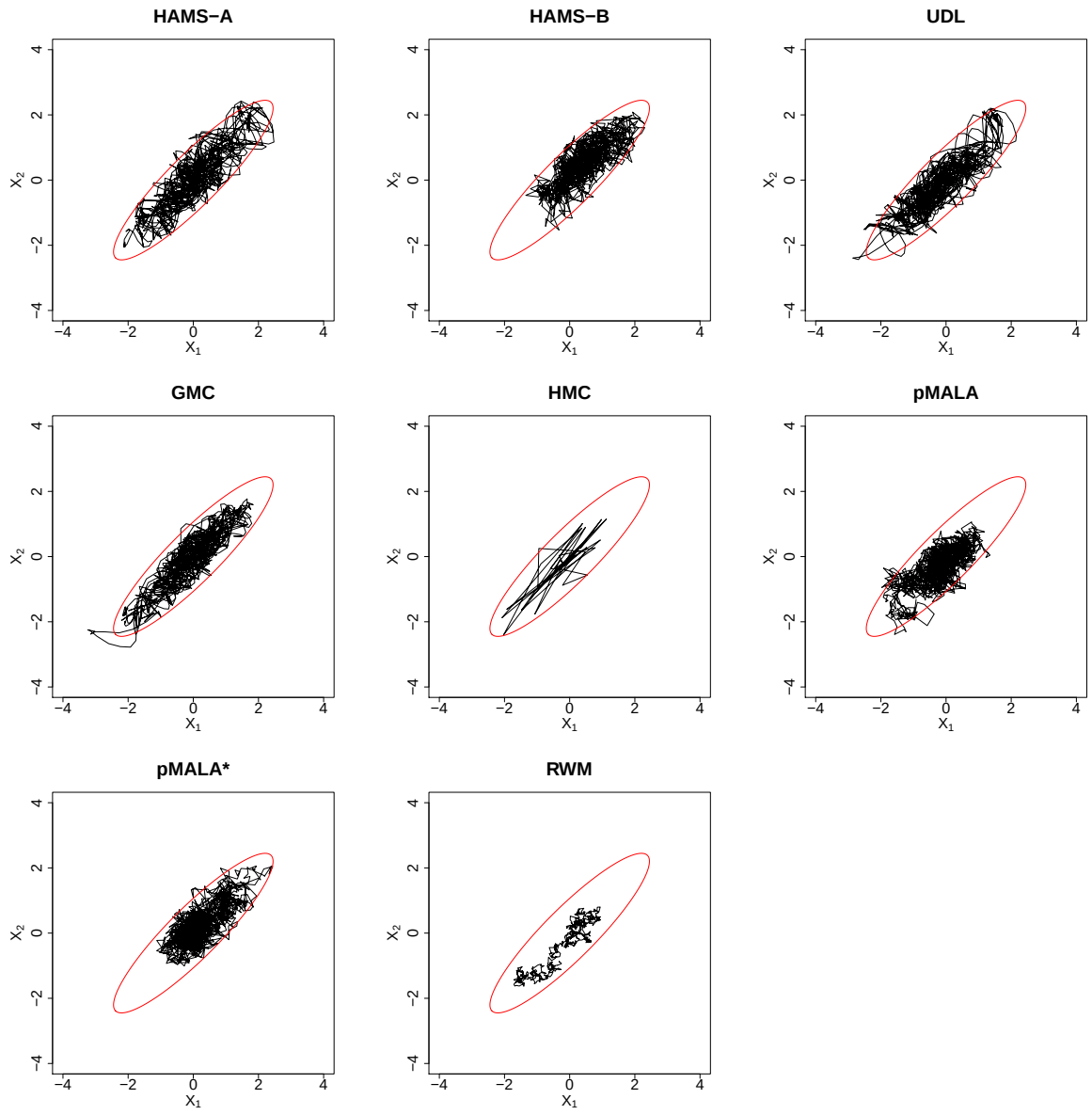


Figure S3: Time-adjusted trace plots of the first two coordinates from first 2000 iterations (first 40 iterations for HMC) for sampling from the multivariate normal distribution. Red ellipses indicate 95% probability regions.

VI.2 Multilevel logistic regression

Consider the setting of multilevel logistic regression in Section 5.1. For sampling latent variables only, Figure S4 shows the average acceptance rates (red curves) and step sizes ϵ (black curves) during the burn-in period, using the tuning procedure described in Section V.4. The upper and lower thresholds of acceptance rates for such adjustments are marked by the dashed lines. For mGrad as described by (3), the step-size parameter δ is not bounded between 0 and 1, unlike ϵ in all other methods. The black curve plotted for mGrad is $\delta/0.03$. We follow Titsias and Papaspiliopoulos (2018) and tune mGrad to achieve acceptance rate between 50% and 60%. (We explored using the 60–80% thresholds for mGrad, but the performance is worse; hence all results of mGrad shown here and subsequently are tuned to achieve 50–60% acceptance rates.) From Figure S4, we see that our tuning achieves target acceptance rates except for HAMS-A/B and pMALA*, where the step sizes ϵ increase to almost 1 but high acceptance rates are maintained. A general explanation for this phenomenon is that these three methods use coefficient $\frac{\epsilon^2}{1+\sqrt{1-\epsilon^2}}$ instead of $\frac{\epsilon^2}{2}$ for gradient updates and satisfy the Gaussian-calibrated rejection-free property. Compared with the log-Gaussian Cox and stochastic volatility examples, our preconditioning in this example may be more accurate, even though the preconditioning matrix is derived in a simple manner, differen from Girolami and Calderhead (2011).

Table S1 shows the runtime and ESS comparison, expanded from Table 1: HAMS-B is the best; pMALA* is comparable to HAMS-A and improves upon pMALA considerably, due to its ability to achieve high acceptance rates with large step sizes; GMC is better than UDL; and mGrad only outperforms HMC and RWM. Figure S6 shows that average sample means are similar among all methods except for RWM. However, according to Figure S5 and S7, the spreads of sample means are very different: HAMS-B is the most consistent across repeated simulations, followed by HAMS-A and pMALA*, and other methods have much more variability. Additional trace plots are provided in Figures S8-S10, we see that HAMS-B, HAMS-A and pMALA* have the best mixing and HAMS-B shows the most noticeable negative autocorrelations.

Next we present additional results of posterior sampling including GMC, pMALA* and mGrad. Notice that mGrad can be used in this setting, because both latent variables $\mathbf{x}$ and parameters β can be associated with latent Gaussian field models. For posterior sampling in the log-Gaussian Cox and stochastic volatility models, mGrad is not applicable. To make meaningful comparison, we set $C = I$ for mGrad when sampling $p(\mathbf{x}|\mathbf{y}, \beta, \sigma)$ and $p(\beta|\mathbf{y}, \sigma, \mathbf{x})$ in the first stage. Then in the second stage, we evaluate and fix the matrix C using the sample mean of σ from the first stage for $p(\mathbf{x}|\mathbf{y}, \beta, \sigma)$, and use V for $p(\beta|\mathbf{y}, \sigma, \mathbf{x})$. A complete summary of all parameters is provided in Table S2, expanded

from Table 2. Boxplots of posterior means are shown in Figure S11, and posterior density plots are shown in Figure S12. According to Figure S11, HAMS-A is the best across all parameters; HMC is also good but overestimates σ_1, σ_4 and σ_5 compared HAMS-A. GMC and UDL have larger spread in σ; mGrad has decent performance in σ but suffers in β, which might be attributed to the suboptimal preconditioning only using the prior variance V.

Figure S13 shows trace plots of $\beta_1, \beta_5, \sigma_4$ and σ_5 from an individual run which demonstrates our two-stage scheme for posterior sampling, where each stage consists of two sub-stages (hence four sub-stages). In these plots, four sub-stages are divided by blue vertical lines. In the first sub-stage, we apply no preconditioning and adjust step size ϵ. In the second sub-stage we fix ϵ and collect samples for crude parameter estimates; we then evaluate preconditioning matrices using the sample means of parameters from the second sub-stage and fix them. In the third sub-stage we apply preconditioning and adjust ϵ. In the fourth sub-stage, we fix ϵ and continue applying preconditioning to collect working samples.

Method	Time (s)	ESS_1 (min, median, max)	$\frac{\text{minESS}_1}{\text{Time}}$	ESS_2 (min, median, max)	$\frac{\text{minESS}_2}{\text{Time}}$
HAMS-A	21.2	(13034, 17315, 22773)	614.8	(2750, 5242, 9950)	129.7
HAMS-B	20.3	(43659, 57046, 71749)	2149.1	(11639, 17191, 29344)	573.1
UDL	20.1	(1815, 2557, 3336)	90.3	(541, 796, 1416)	26.9
GMC	20.1	(2553, 3378, 4223)	127.2	(671, 985, 1945)	33.4
HMC	257.8	(10085, 15846, 32386)	39.1	(28, 255, 1068)	0.1
pMALA	19.4	(1337, 1838, 2254)	69.0	(342, 523, 970)	17.6
pMALA*	18.9	(11457, 14795, 19089)	605.5	(2870, 4735, 7852)	151.7
mGrad	17.5	(419, 712, 5170)	23.9	(103, 223, 1777)	5.9
RWM	6.9	(11, 22, 36)	1.5	(0.3, 1.1, 1.9)	0.04

Table S1: Runtime and ESS comparison for sampling latent variables in the multilevel logistic regression (including mGrad). Results are averaged over 50 repetitions.

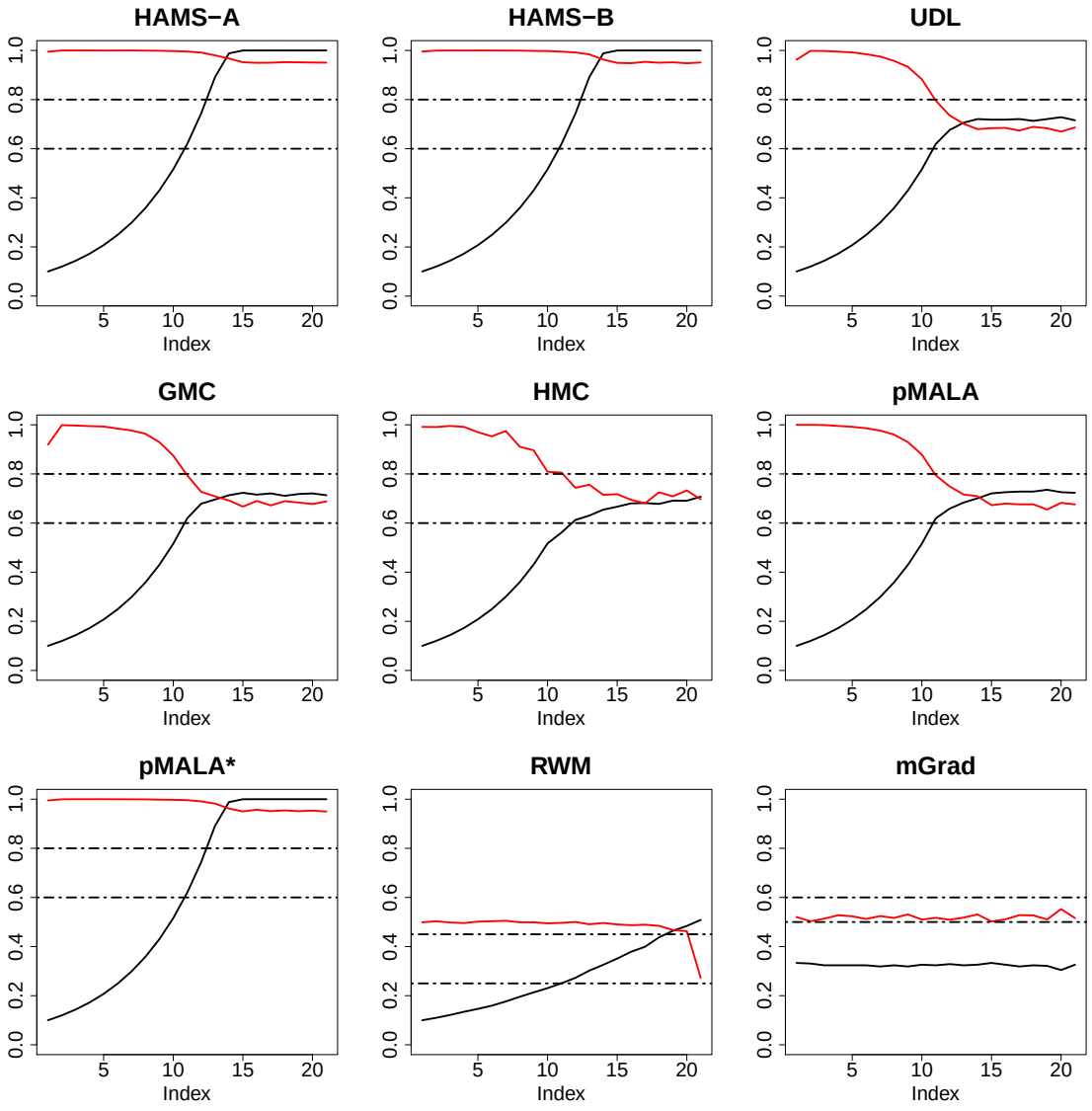


Figure S4: Average step sizes (black) and acceptance rates (red) for sampling latent variables in the multilevel logistic regression. For every 250 iterations, acceptance rates are calculated and step sizes adjusted. Results are averaged over 50 repetitions.

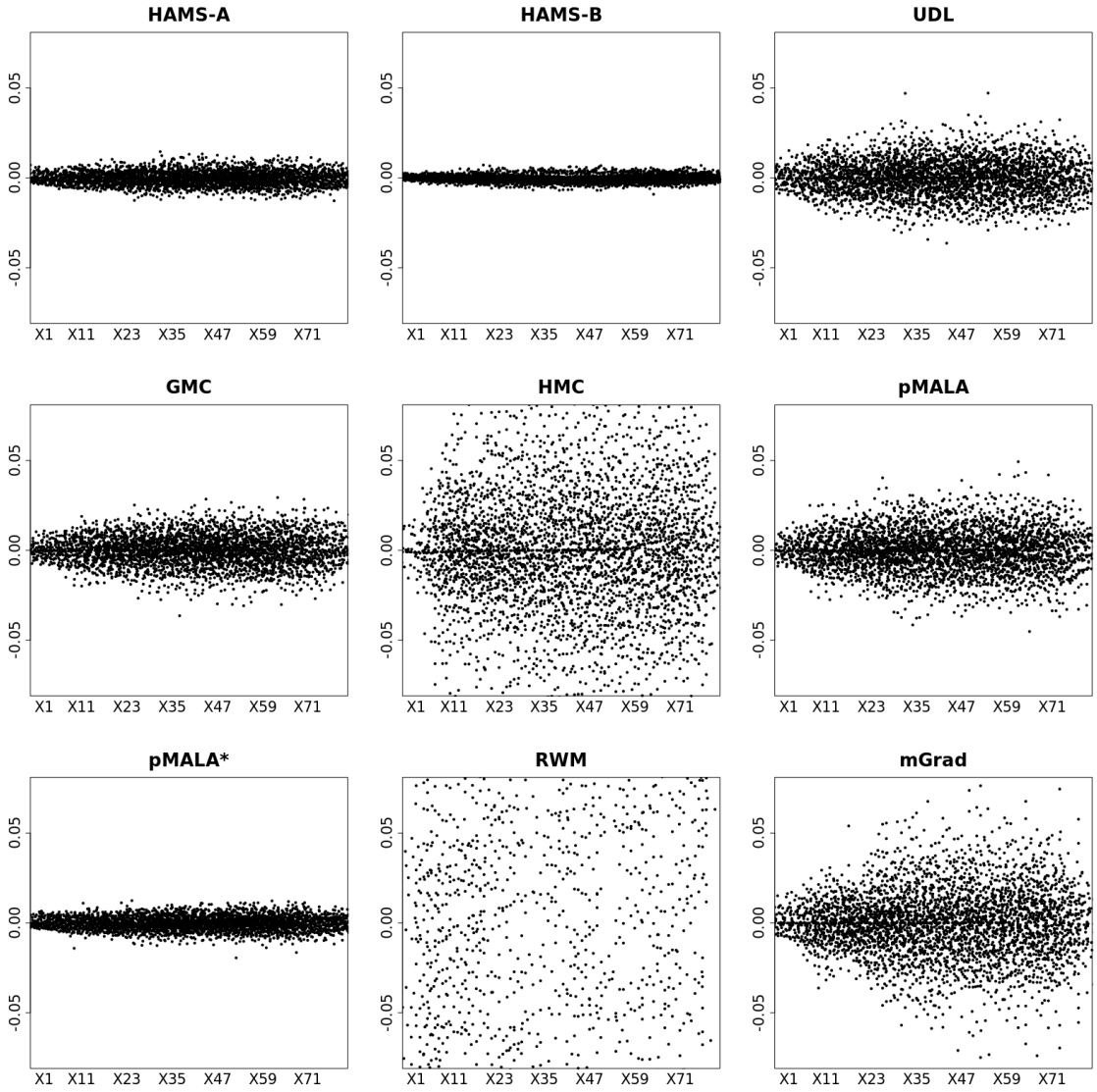


Figure S5: Time-adjusted and centered plots of sample means of all latent variables over 50 repetitions for sampling latent variables in the multilevel logistic regression.

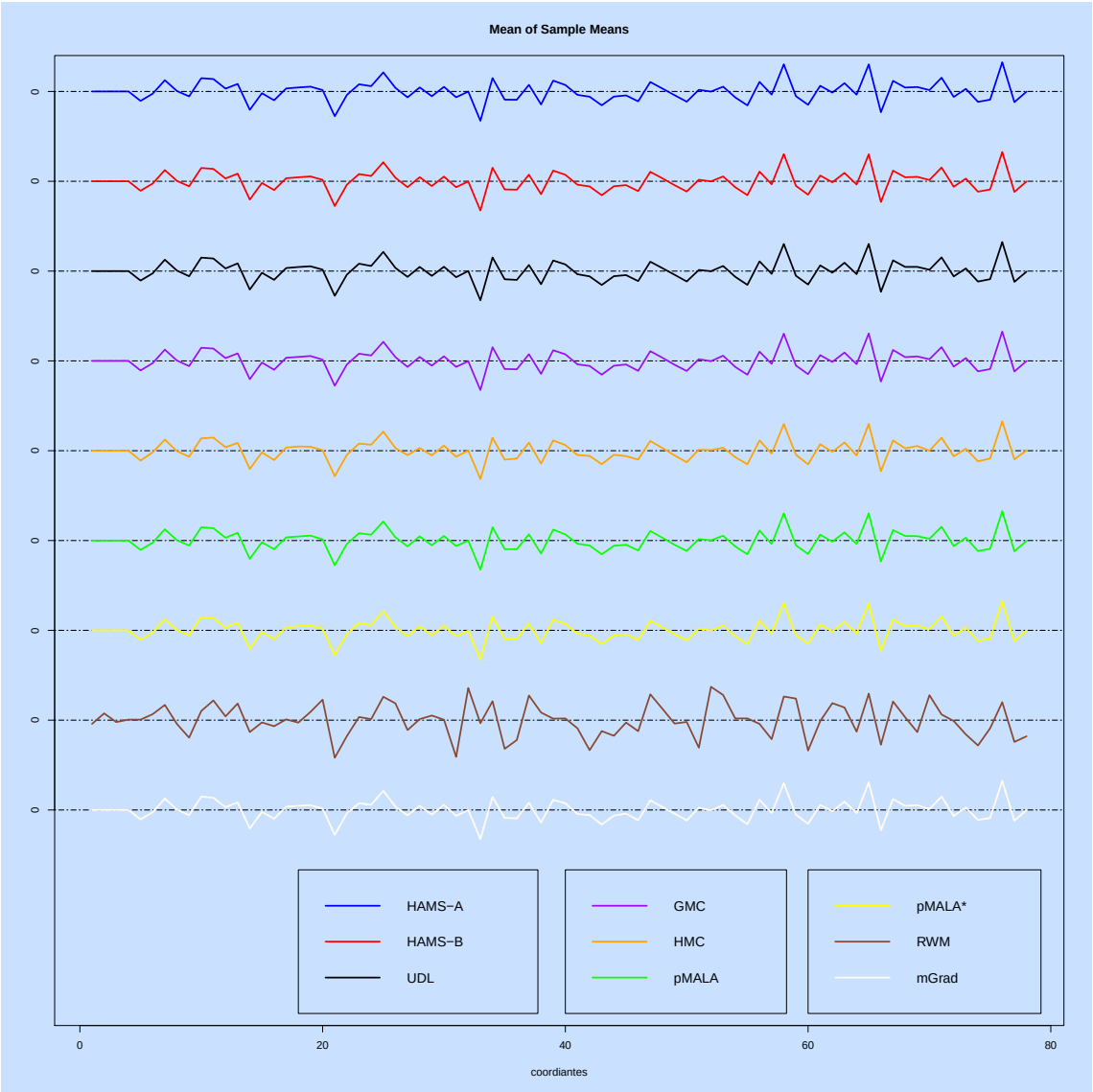


Figure S6: Time-adjusted averages of sample means of all latent variables over 50 repetitions for sampling latent variables in the multilevel logistic regression.

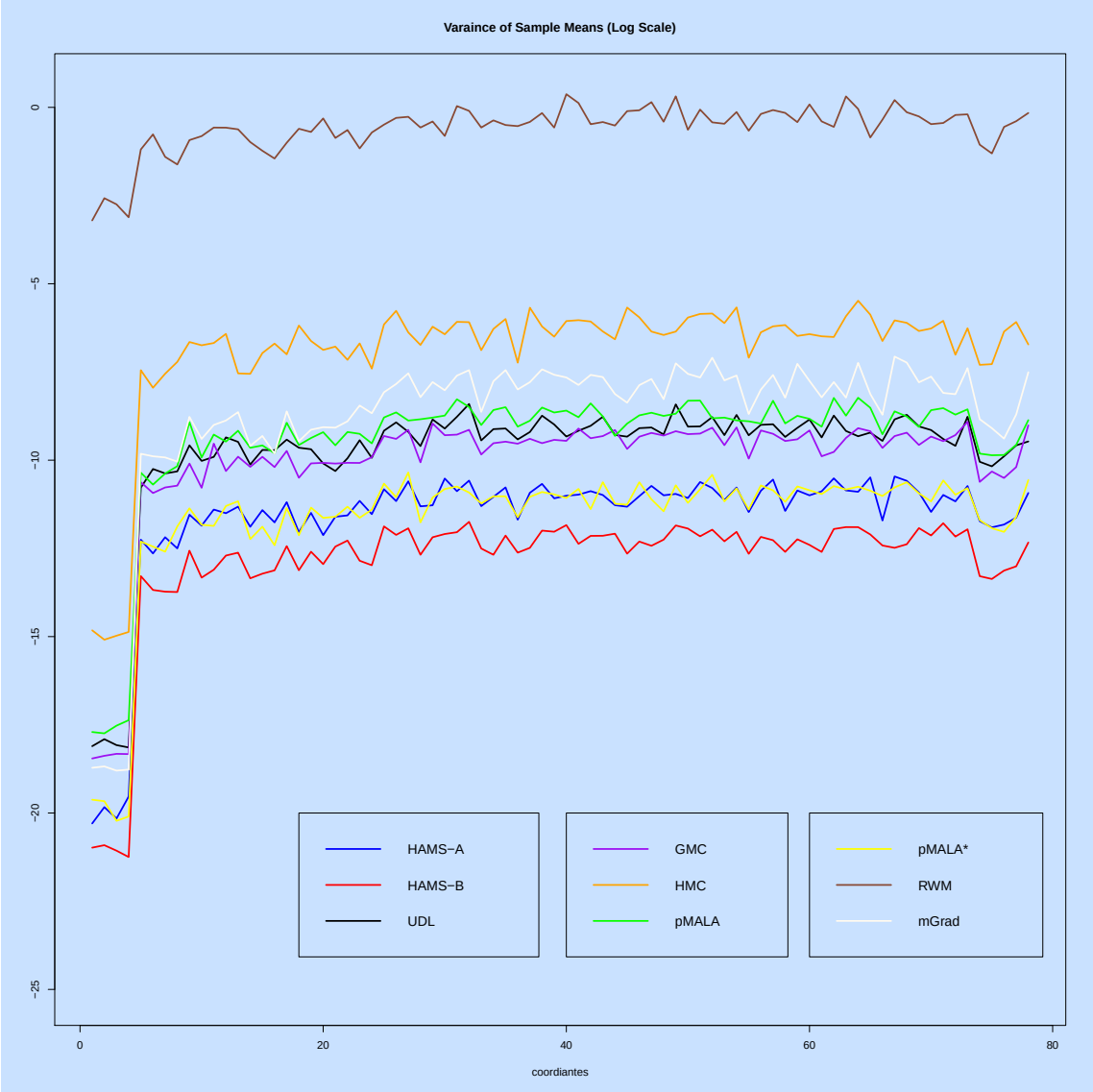


Figure S7: Time-adjusted variances of sample means (log-scale) of all latent variables over 50 repetitions for sampling latent variables in the multilevel logistic regression.

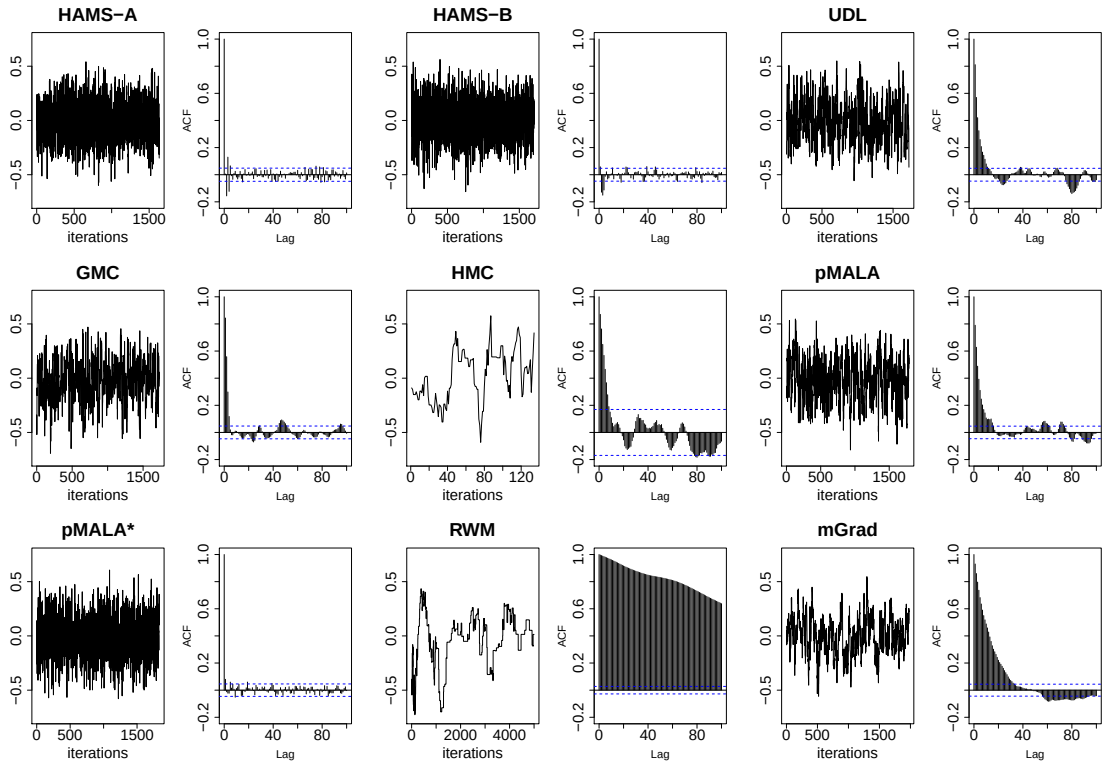


Figure S8: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the multilevel logistic regression.

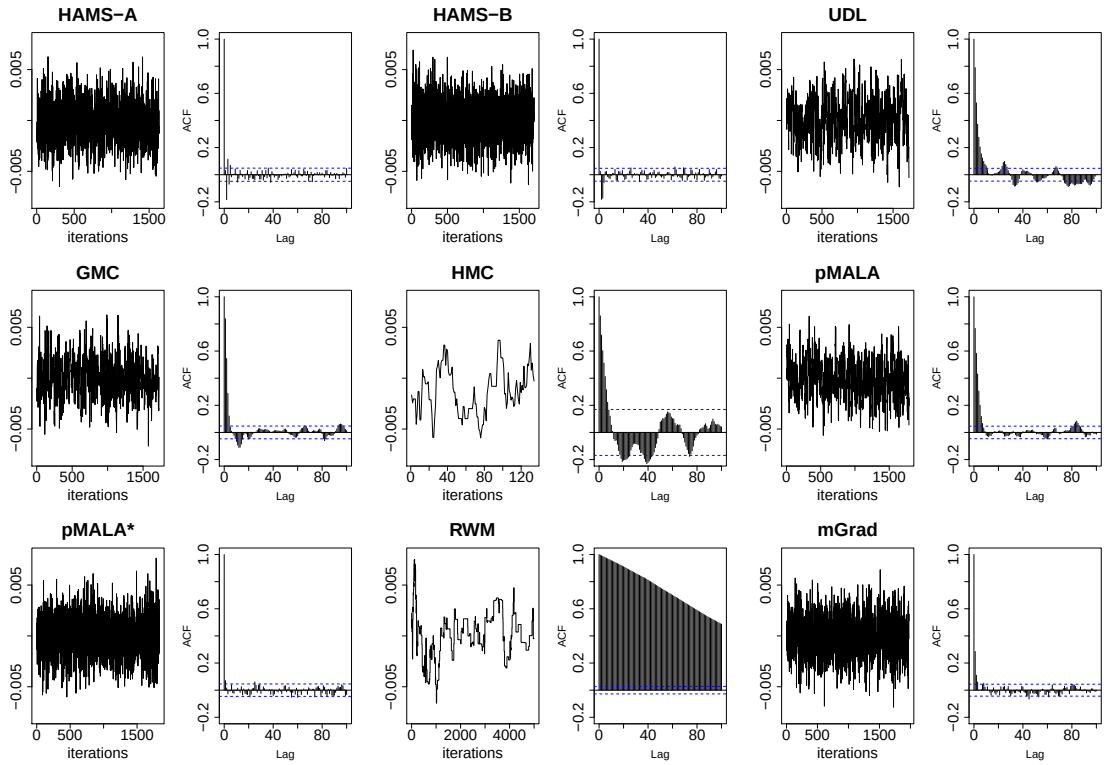


Figure S9: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the multilevel logistic regression.

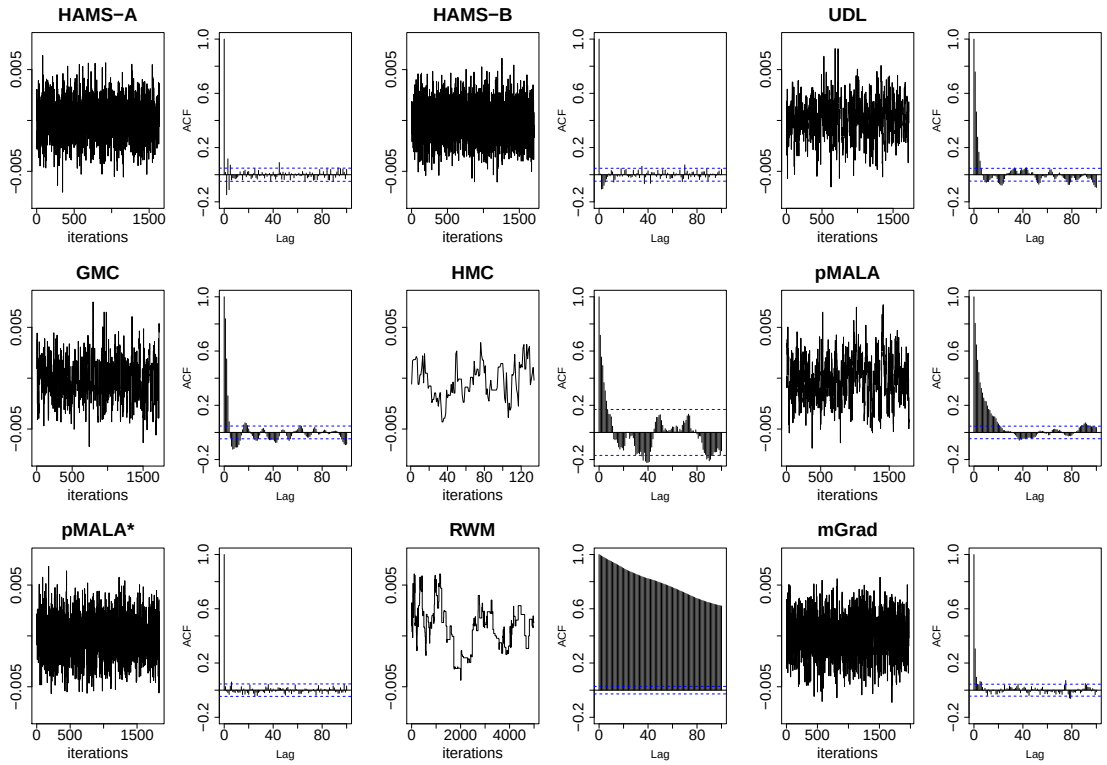


Figure S10: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the multilevel logistic regression.

Method	Time (s)	Sample Mean					
		β_1 (sd)	β_2 (sd)	β_3 (sd)	β_4 (sd)	β_5 (sd)	σ_1 (sd)
HAMS-A	43.9	-3.38 (0.315)	-1.67 (0.018)	-0.09 (0.004)	-0.18 (0.012)	6.77 (0.487)	0.15 (0.046)
HAMS-B	43.8	-3.37 (0.522)	-1.69 (0.0022)	-0.09 (0.004)	-0.17 (0.012)	6.81 (0.382)	0.20 (0.120)
UDL	43.9	-3.47 (0.437)	-1.68 (0.022)	-0.09 (0.004)	-0.17 (0.015)	6.83 (0.630)	0.17 (0.065)
GMC	43.8	-3.31 (0.591)	-1.67 (0.019)	-0.09 (0.003)	-0.18 (0.012)	6.88 (0.838)	0.15 (0.061)
HMC	404.8	-3.42 (0.118)	-1.68 (0.005)	-0.09 (0.001)	-0.17 (0.010)	6.86 (0.092)	0.22 (0.0025)
pMALA	44.4	-3.21 (1.151)	-1.67 (0.036)	-0.09 (0.005)	-0.18 (0.007)	6.55 (0.805)	0.19 (0.072)
pMALA *	44.3	-3.46 (0.377)	-1.69 (0.022)	-0.09 (0.005)	-0.17 (0.022)	6.96 (0.621)	0.19 (0.073)
mGrad	44.7	-3.51 (0.615)	-1.70 (0.071)	-0.09 (0.007)	-0.14 (0.104)	7.00 (1.218)	0.19 (0.068)
RWM	22.7	-3.33 (0.489)	-1.68 (0.036)	-0.09 (0.006)	-0.18 (0.024)	6.76 (0.743)	0.25 (0.227)

Method	Sample Mean				$\frac{\text{minESS}_1}{\text{Time}}$ (β, σ)	$\frac{\text{minESS}_2}{\text{Time}}$ (β, σ)
	σ_2 (sd)	σ_3 (sd)	σ_4 (sd)	σ_5 (sd)		
HAMS-A	0.27 (0.090)	0.14 (0.043)	0.22 (0.023)	0.39 (0.079)	(1.885, 0.461)	(0.222, 0.093)
HAMS-B	0.32 (0.209)	0.13 (0.046)	0.23 (0.023)	0.47 (0.345)	(1.711, 0.422)	(0.092, 0.038)
UDL	0.24 (0.167)	0.14 (0.070)	0.21 (0.055)	0.42 (0.278)	(1.649, 0.713)	(0.113. 0.029)
GMC	0.28 (0.116)	0.13 (0.055)	0.21 (0.063)	0.44 (0.387)	(1.773. 0.462)	(0.070, 0.027)
HMC	0.33 (0.042)	0.14 (0.015)	0.22 (0.014)	0.44 (0.037)	(1.844, 0.297)	(0.233, 0.087)
pMALA	0.45 (0.558)	0.14 (0.050)	0.22 (0.041)	0.65 (0.629)	(1.412, 0.366)	(0.017, 0.026)
pMALA *	0.28 (0.071)	0.13 (0.045)	0.23 (0.042)	0.42 (0.167)	(0.749, 0.240)	(0.173, 0.070)
mGrad	0.26 (0.092)	0.15 (0.029)	0.22 (0.035)	0.40 (0.226)	(0.199, 0.426)	(0.033, 0.058)
RWM	0.29 (0.064)	0.16 (0.075)	0.22 (0.055)	0.48 (0.292)	(1.632, 0.537)	(0.168, 0.047)

Table S2: Comparison of posterior sampling in the multilevel logistic regression. Standard deviations of sample means are in parentheses. Results are averaged over 20 repetitions.

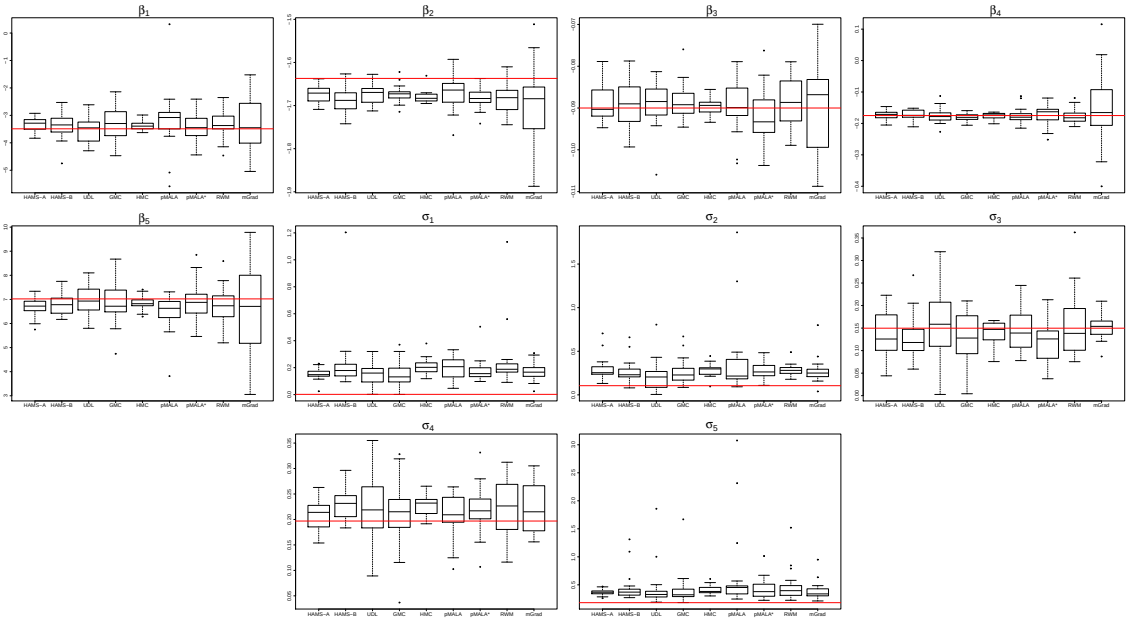
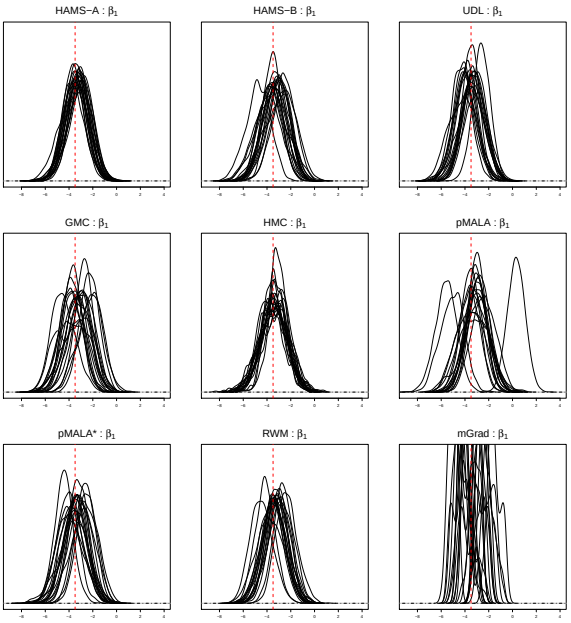
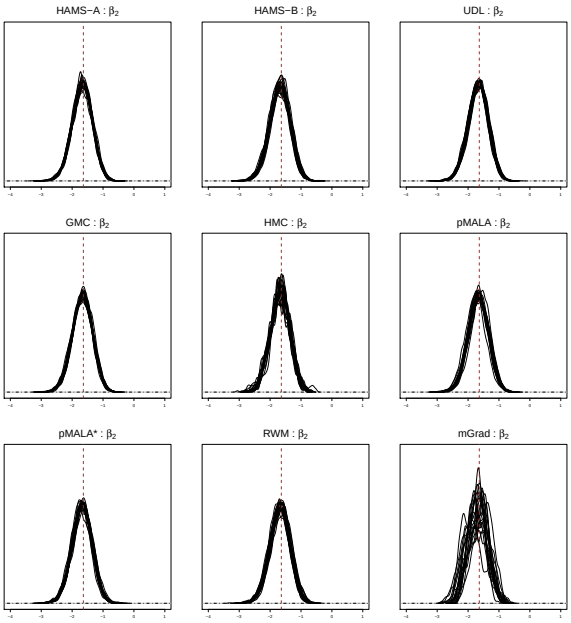


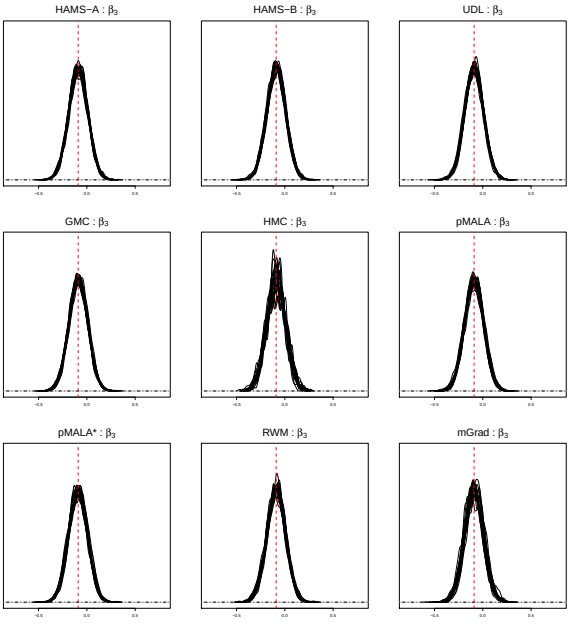
Figure S11: Time-adjusted boxplots of posterior means of parameters over 20 repetitions for posterior sampling in the multilevel logistic regression. The estimates obtained using `lme4` are marked by red lines.



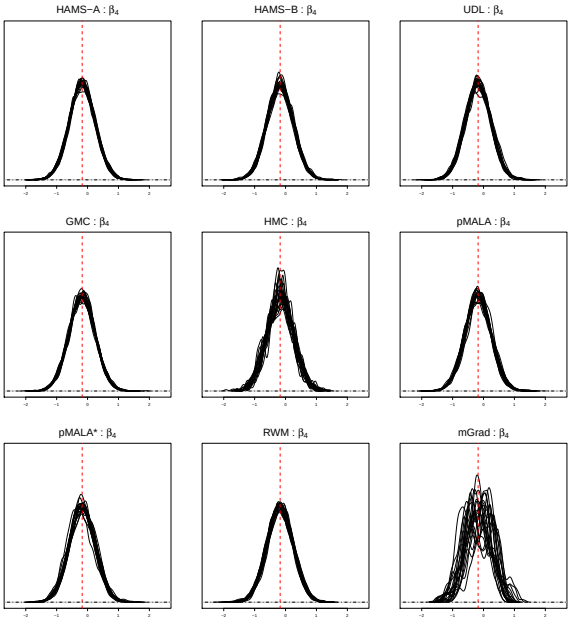
(a) Densities of β_1



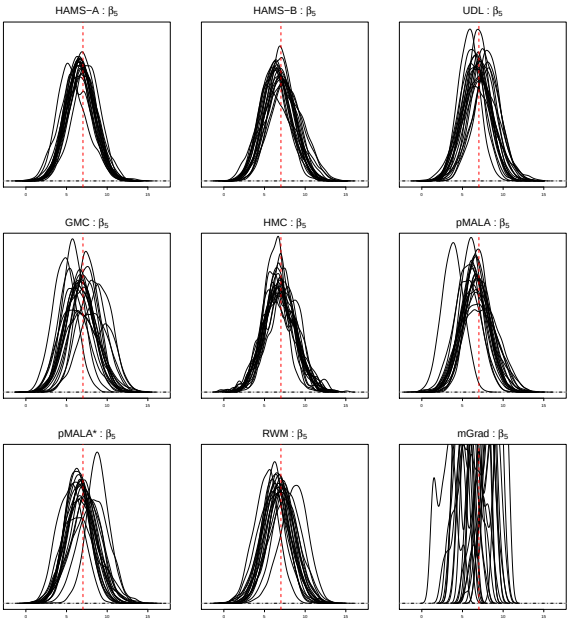
(b) Densities plots of β_2



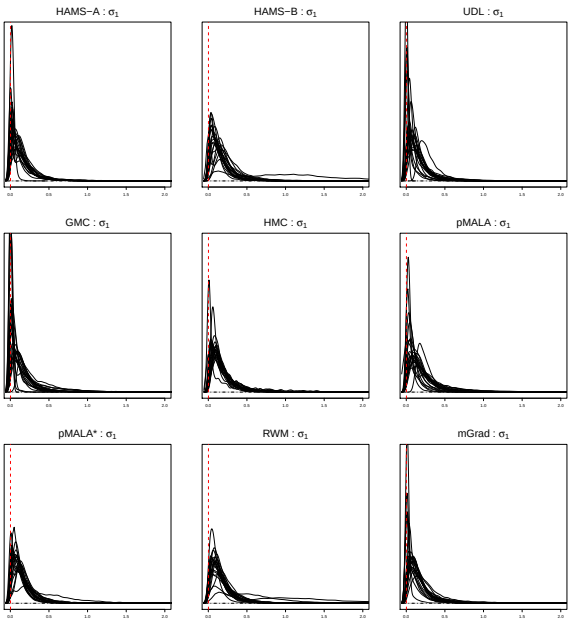
(c) Densities of β_3



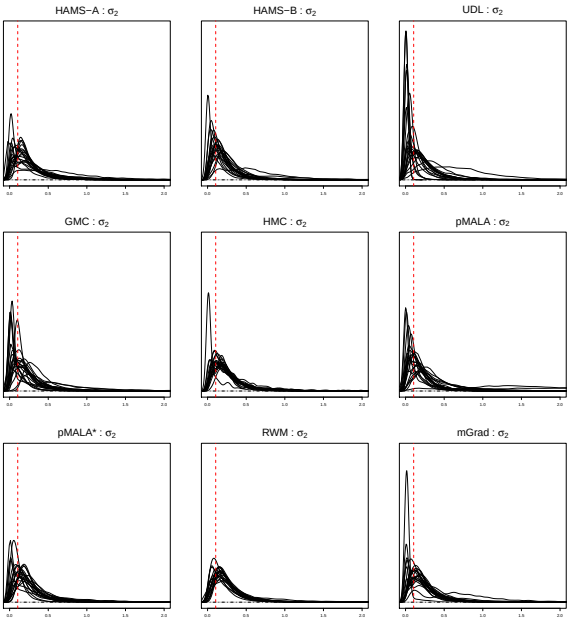
(d) Densities of β_4



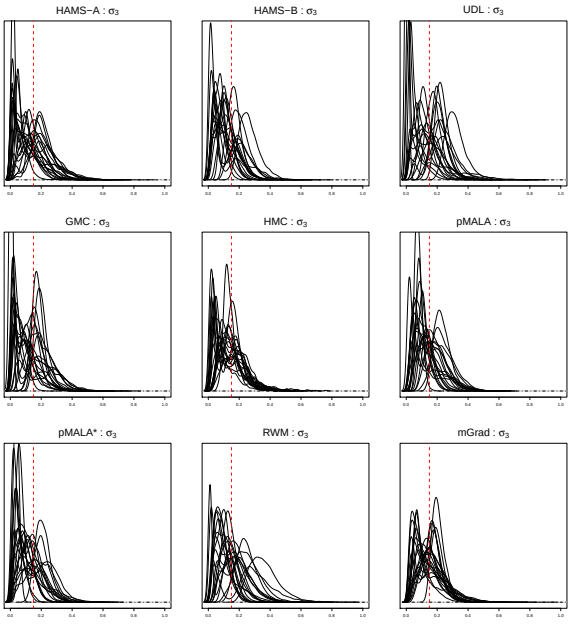
(e) Densities of β_5



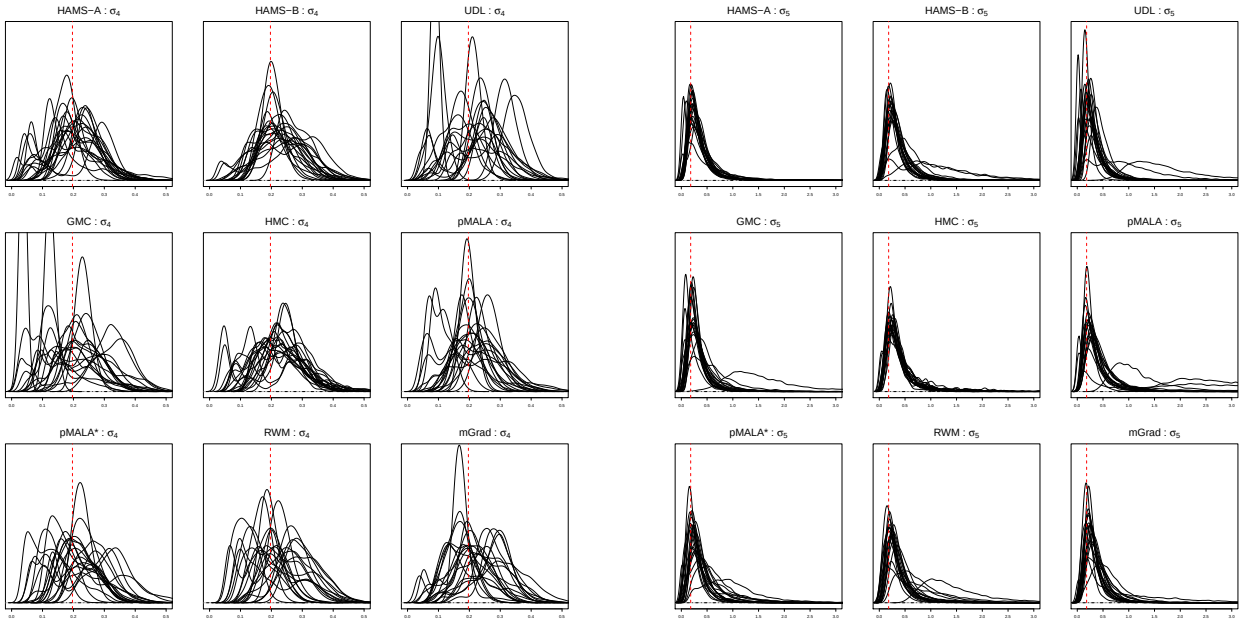
(f) Densities of σ_1



(g) Densities of σ_2



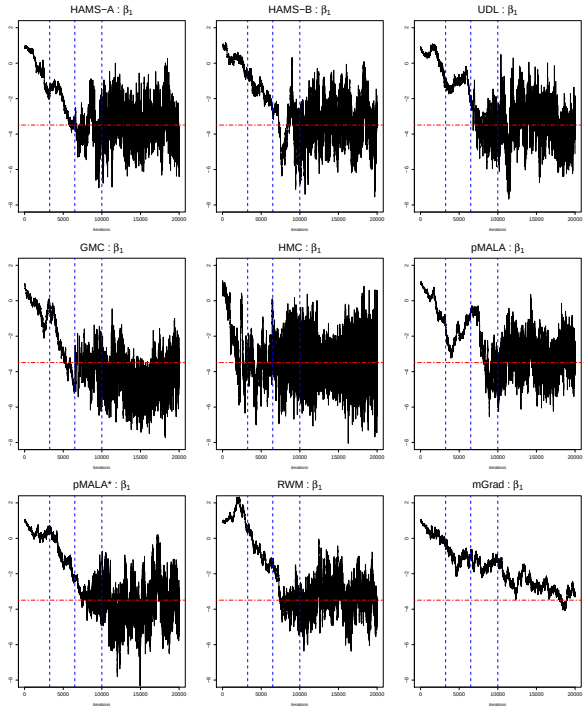
(h) Densities of σ_3



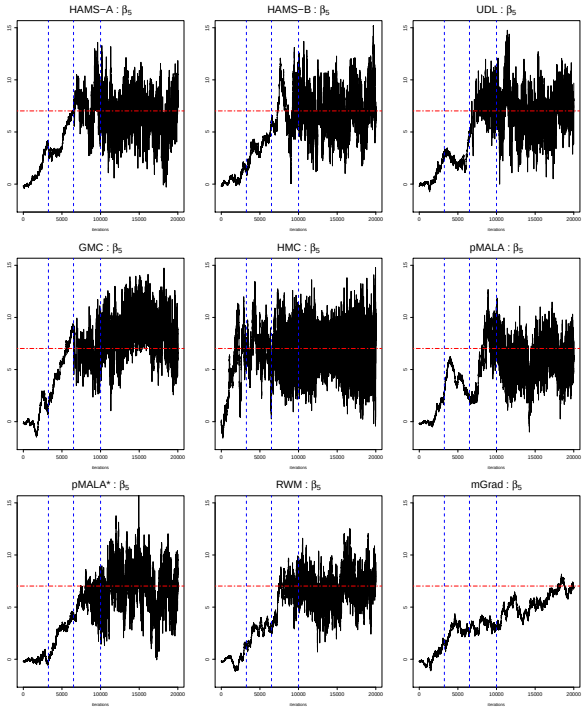
(j) Densities of σ_4

(k) Densities of σ_5

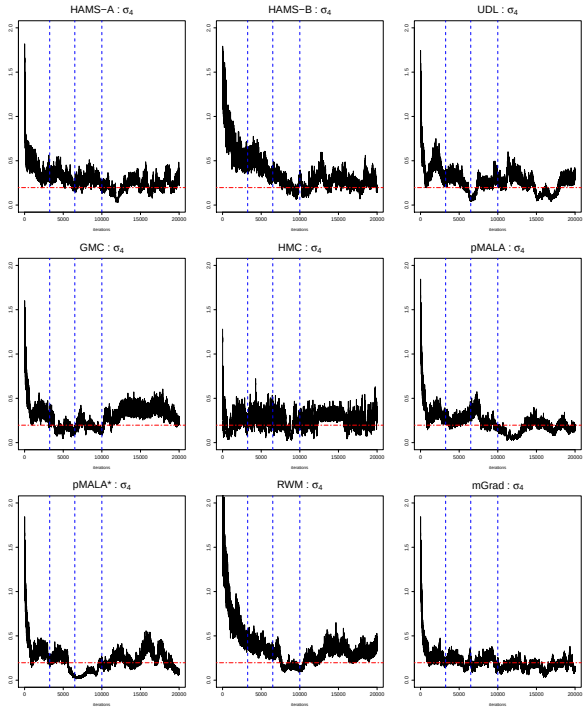
Figure S12: Time-adjusted posterior density plots of parameters (20 repetitions overlaid) in the multilevel logistic regression. The estimates obtained using `lme4` are marked by vertical lines.



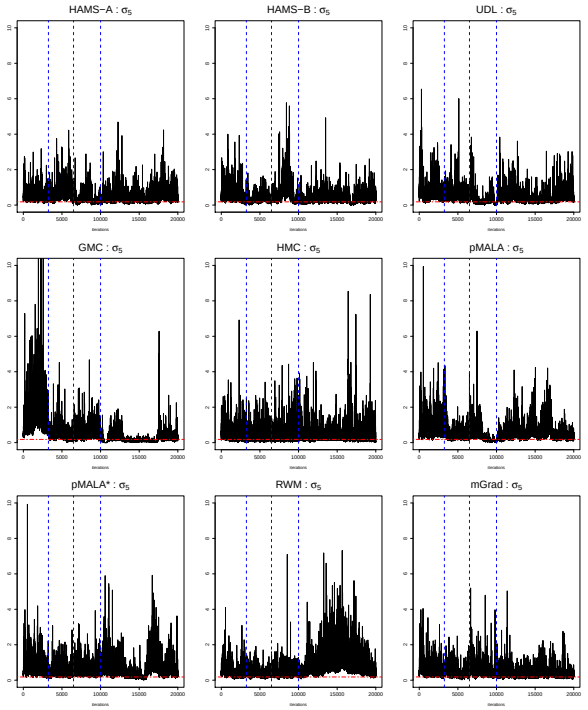
(a) Trace plots of β_1



(b) Trace plots of β_5



(c) Trace plots of σ_4



(d) Trace plots of σ_5

Figure S13: Trace plots of $\beta_1, \beta_5, \sigma_4$ and σ_5, from an individual run for posterior sampling in the multilevel. The estimates obtained using `lme4` are marked by red horizontal lines. There are four sub-stages divided by blue vertical lines. The first two are without preconditioning, with 3250 iterations each. The last two are with preconditioning, with 3500 and 10000 iterations respectively. The first three sub-stages are counted as burn-in.

VI.3 Stochastic volatility model

Consider a stochastic volatility model, where latent volatilities are generated as

$$x_t = \phi x_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2), \quad t = 2, 3, ..., T, \tag{S54}$$

with $x_1 \sim \mathcal{N}(0, \sigma^2/(1-\phi^2))$, and the observations are generated as

$$y_t = z_t \beta \exp(x_t/2), \quad z_t \sim \mathcal{N}(0, 1), \qquad t = 1, ..., T. \tag{S55}$$

The parameters of interest are $\theta = (\beta, \sigma, \phi)^{\mathrm{T}}$. We simulate $T = 1000$ observations from (S54)–(S55) using parameter values $\beta = 0.65, \sigma = 0.15$ and $\phi = 0.98$. Let $\mathbf{x} = (x_1, ..., x_T)^{\mathrm{T}}$ and $\mathbf{y} = (y_1, ..., y_T)^{\mathrm{T}}$. Two sets of experiments are conducted. First, we fix parameter values and sample latent variables from $p(\mathbf{x}|\mathbf{y}, \theta)$. Then we perform Bayesian analysis and sample both the parameters and latent variables from $p(\mathbf{x}, \theta|\mathbf{y})$.

For the first experiment, we fix parameters at their true values and perform sampling for latent variables only. The joint distribution of $(x_1, \ldots, x_T)$ is $\mathcal{N}(\mathbf{0}, C)$, where the entries of the variance matrix are $C[i, j] = \phi^{|i-j|}\sigma^2/(1-\phi^2)$. The inverse C^{-1} retains a simple tri-diagonal form. Following Girolami and Calderhead (2011), the inverse variance $\mathrm{Var}^{-1}(\mathbf{x})$ can be approximated by $-\mathrm{E}[\nabla^2 \log p(\mathbf{x}|\mathbf{y}, \theta)] = C^{-1} + \frac{1}{2}I$. Hence for preconditioning, we set $M = \Sigma^{-1} = C^{-1} + \frac{1}{2}I$ for all methods except mGrad, which uses the prior variance C as the preconditioning matrix. As mentioned in Section 5, we use $nleap = 50$ for HMC as in Girolami and Calderhead (2011) and choose c given ϵ by (30)–(31) for HAMS-A/B, UDL, and GMC. All algorithms are run for 5000 burn-in iterations, and then samples are collected from 5000 iterations. The simulation process is repeated for 50 times.

Average acceptance rates and step sizes are shown in Figure S14, where black curves are $\delta/4$ for mGrad and ϵ for the others. Our tuning achieves target acceptance rates for all methods, with HAMS-A/B and pMALA* using larger ϵ values, which can be explained by the Gaussian-calibrated rejection-free property. Table S3 shows that HAMS-A is the best in ESS_1, HAMS-B is the best in ESS_2, and both lead the remaining methods. Among the rest, pMALA* is the best followed by mGrad, then UDL and GMC which are comparable to each other, and finally pMALA, HMC and RWM. According to Figure S15, HAMS-A and HAMS-B have the least amount of variation in sample means across repetitions. In Figure S16, all methods except RWM have comparable average sample means, while RWM overestimates. The variance comparison in Figure S17 agrees with the spreads shown in Figure S15. Trace plots of three difference latent dimensions from an individual are shown in Figures S18–S20. In terms of mixing behaviors, HAMS-A/B are the best, followed

by pMALA* and mGrad, then UDL and GMC. After adjusting for time, trace plots of HMC show much fewer draws, which reflects the high computational cost of HMC.

In the second experiment, we perform Bayesian analysis and sample both latent variables and parameters from the posterior $p(\mathbf{x}, \theta|\mathbf{y})$. The priors are, independently, $\pi(\beta) \propto \beta^{-1}$, $\sigma^2 \sim \text{Inv-}\chi^2(10, 0.05)$ and $(\phi + 1)/2 \sim \text{Beta}(20, 1.5)$. Moreover, we use the transformations $\sigma = \exp(\gamma)$ and $\phi = \tanh(\alpha)$ to ensure that $\sigma > 0$ and $|\phi| < 1$. We employ Gibbs sampling, alternating between the two blocks $p(\mathbf{x}|\mathbf{y}, \theta)$ and $p(\theta|\mathbf{y}, \mathbf{x})$. We do not include mGrad because the parameters are not from a latent Gaussian model in this case. As previously mentioned, we first run each algorithm without any preconditioning to obtain a crude estimate of the parameters, and then fix the preconditioning matrix evaluated at this estimate (see Section V.2 for associated expressions). For HMC, the numbers of leapfrog steps are 50 for latent variables and 6 for parameters as in Girolami and Calderhead (2011). The initial values of parameters are dispersed over the following intervals $\beta \in [0.5, 2]$, $\sigma \in [0.1, 1]$, and $\phi \in [0, 0.3]$. For all methods, 10000 draws are collected after 10000 iterations, which include two sub-stages without preconditioning and one sub-stage of tuning with preconditioning. The simulation process is repeated for 20 times.

As shown in Table S4, the posterior means of the parameters are all very close (except for RWM). But HAMS-A has the smallest standard deviations in β and ϕ followed by HAMS-B. Hence HAMS-A/B give more consistent results, as corroborated by Figure S21. While for σ, pMALA has a smaller standard deviation, it is inferior to HAMS-A/B in both ESS_1 and ESS_2. Figure S22 shows time-adjusted density plots for the parameters. Each plot shows densities from 20 repeated runs overlaid together. Clearly, HAMS-A yields the most consistent density curves for all three parameters, followed by HAMS-B, UDL, and GMC which sometimes produce outlying curves, especially in β and σ.

Method	Time (s)	ESS_1 (min, median, max)	$\frac{\text{minESS}_1}{\text{Time}}$	ESS_2 (min, median, max)	$\frac{\text{minESS}_2}{\text{Time}}$
HAMS-A	239.4	(2420, 3669, 6668)	10.11	(433, 1073, 2250)	1.85
HAMS-B	238.6	(1915, 57046, 71749)	8.03	(597, 999, 2645)	2.50
UDL	239.3	(657, 2557, 3336)	2.74	(181,304, 648)	0.76
GMC	240.3	(752, 3378, 4223)	3.13	(216, 377, 710)	0.90
HMC	5258.5	(1125, 12481, 19699)	0.21	(24, 162, 1493)	0.004
pMALA	282.7	(374, 1838, 2254)	1.32	(91, 175, 395)	0.32
pMALA*	281.6	(1740, 14795, 19089)	6.18	(415, 889, 1914)	1.47
mGrad	237.1	(948,1588, 2611)	4.00	(266, 474, 912)	1.12
RWM	116.8	(7, 12, 20)	0.06	(0.3, 0.6, 1.5)	0.002

Table S3: Runtime and ESS comparison for sampling latent variables in the stochastic volatility model. Results are averaged over 50 repetitions.

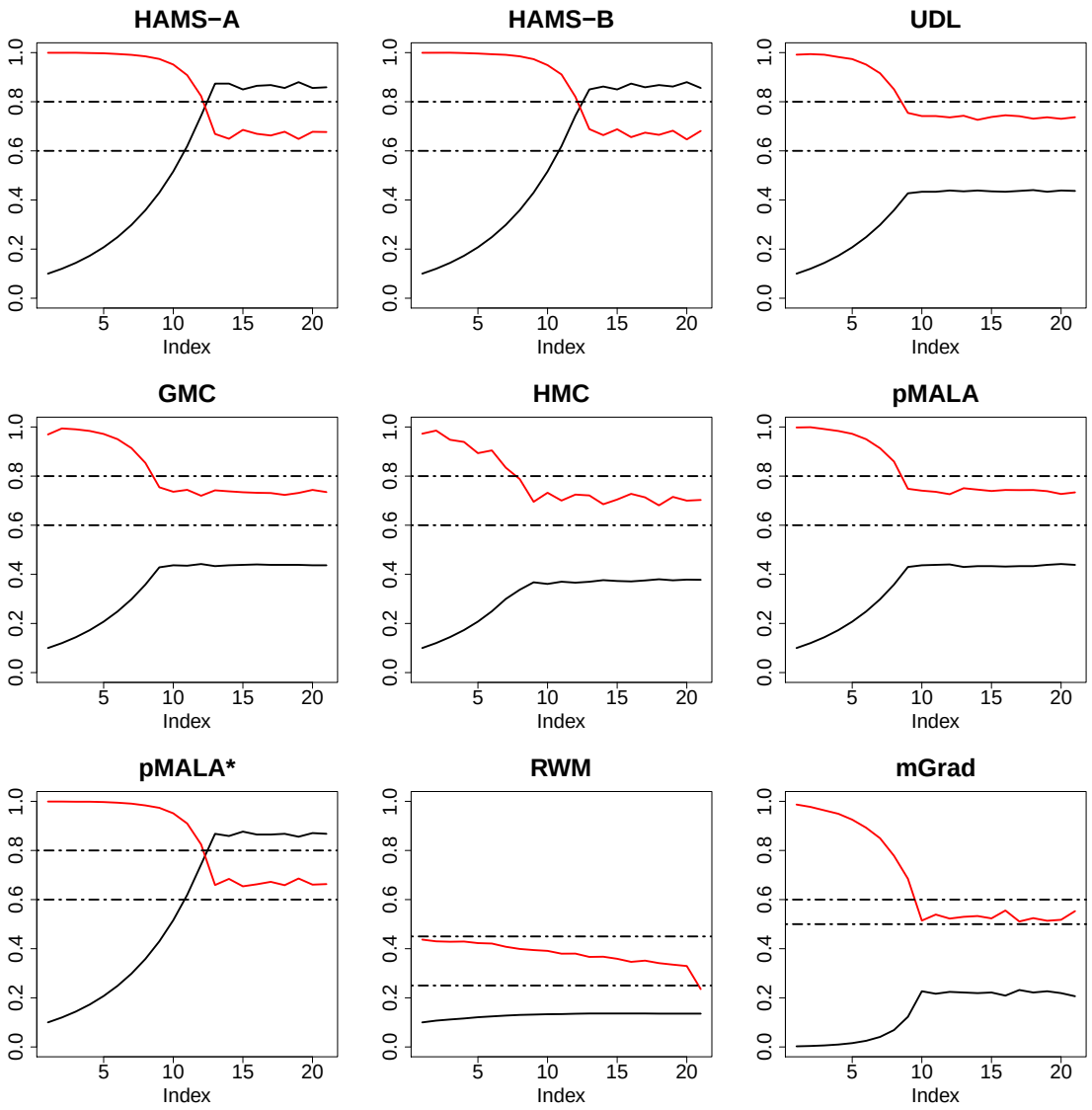


Figure S14: Average step sizes (black) and acceptance rates (red) for sampling latent variables in the stochastic volatility model. For every 250 iterations, acceptance rates are calculated and step sizes adjusted. Results are averaged over 50 repetitions.

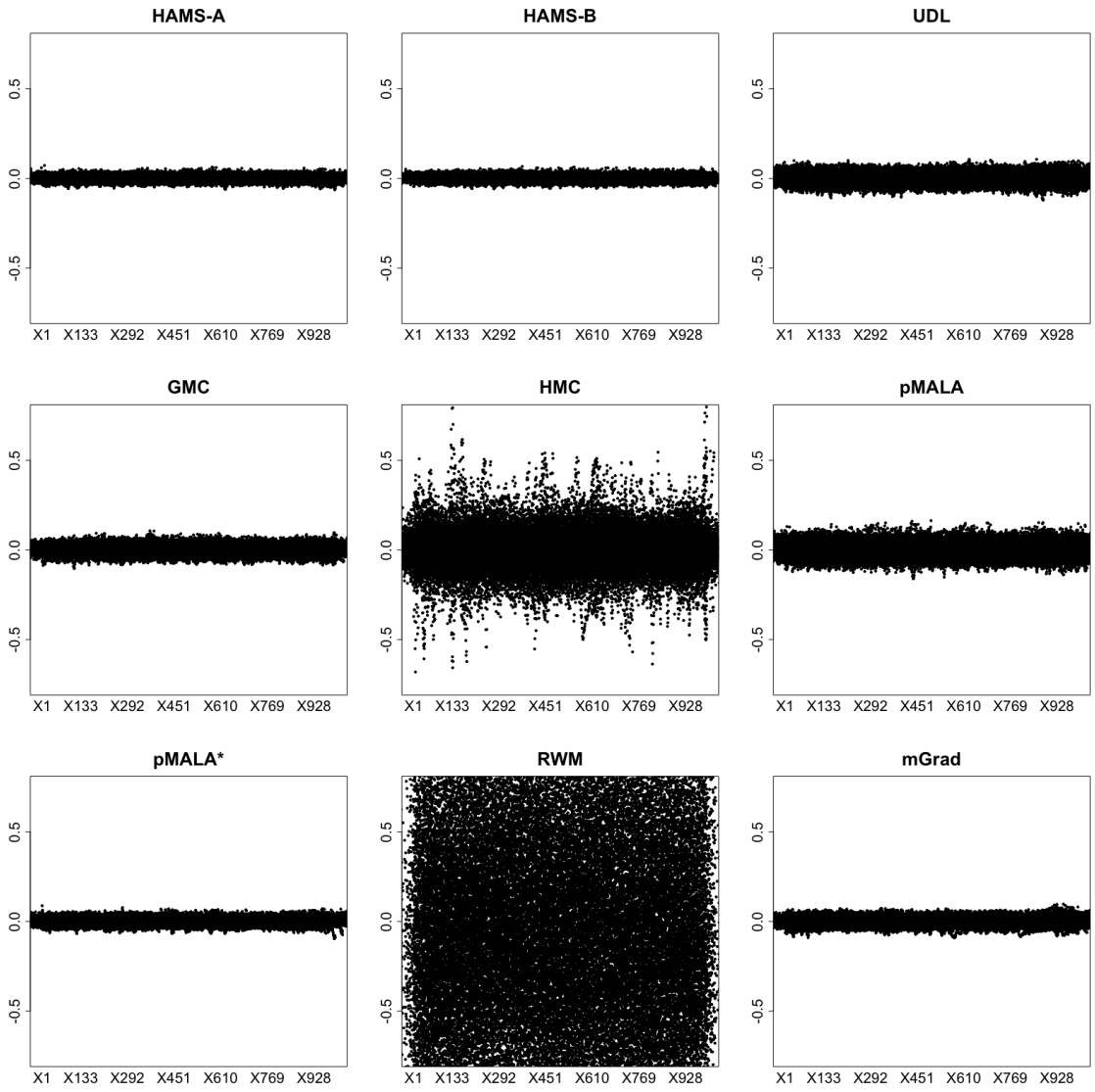


Figure S15: Time-adjusted and centered plots of sample means of all latent variables over 50 repetitions for sampling latent variables in the stochastic volatility model.

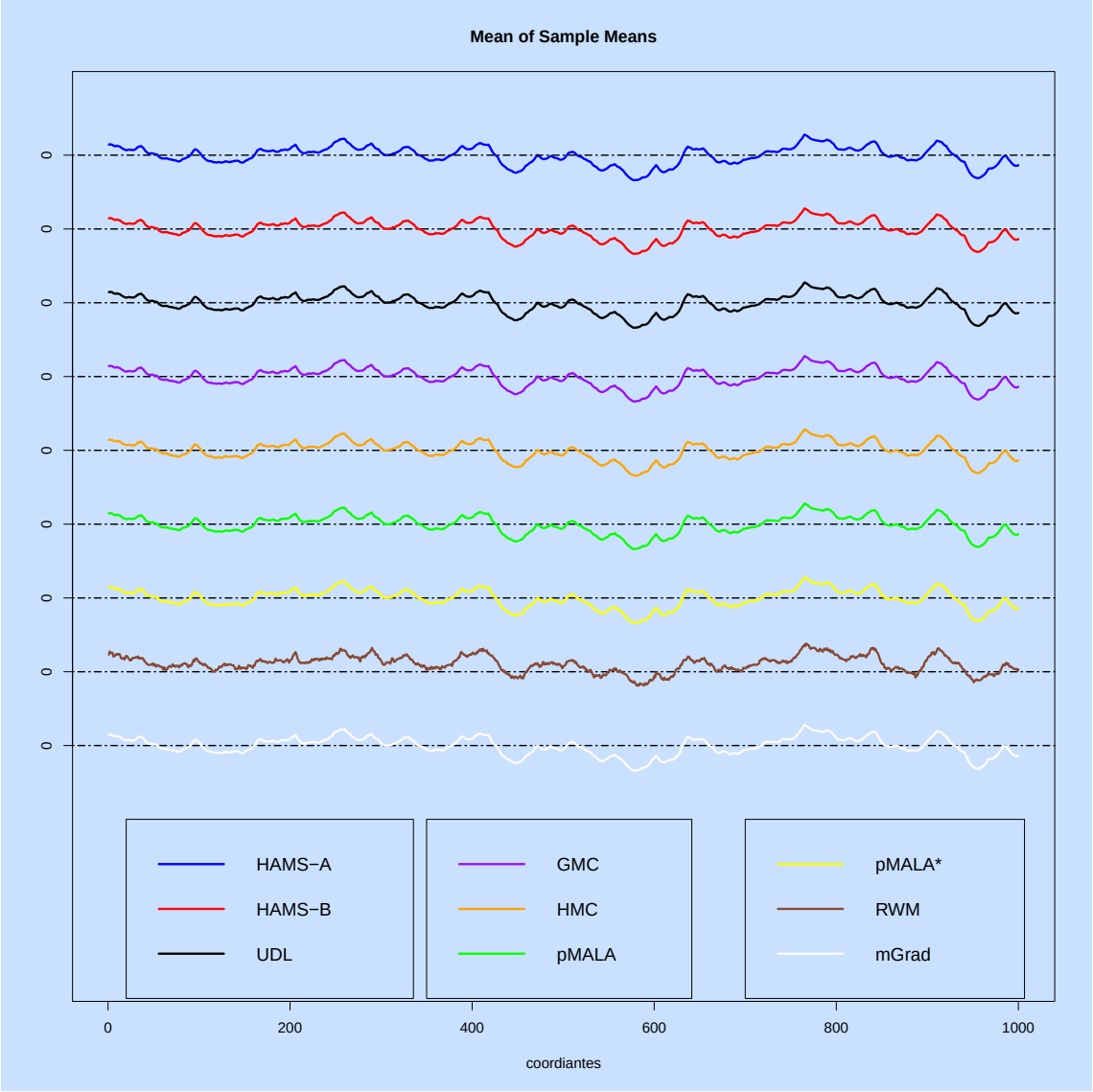


Figure S16: Time-adjusted averages of sample means of all latent variables over 50 repetitions for sampling latent variables in the stochastic volatility model.

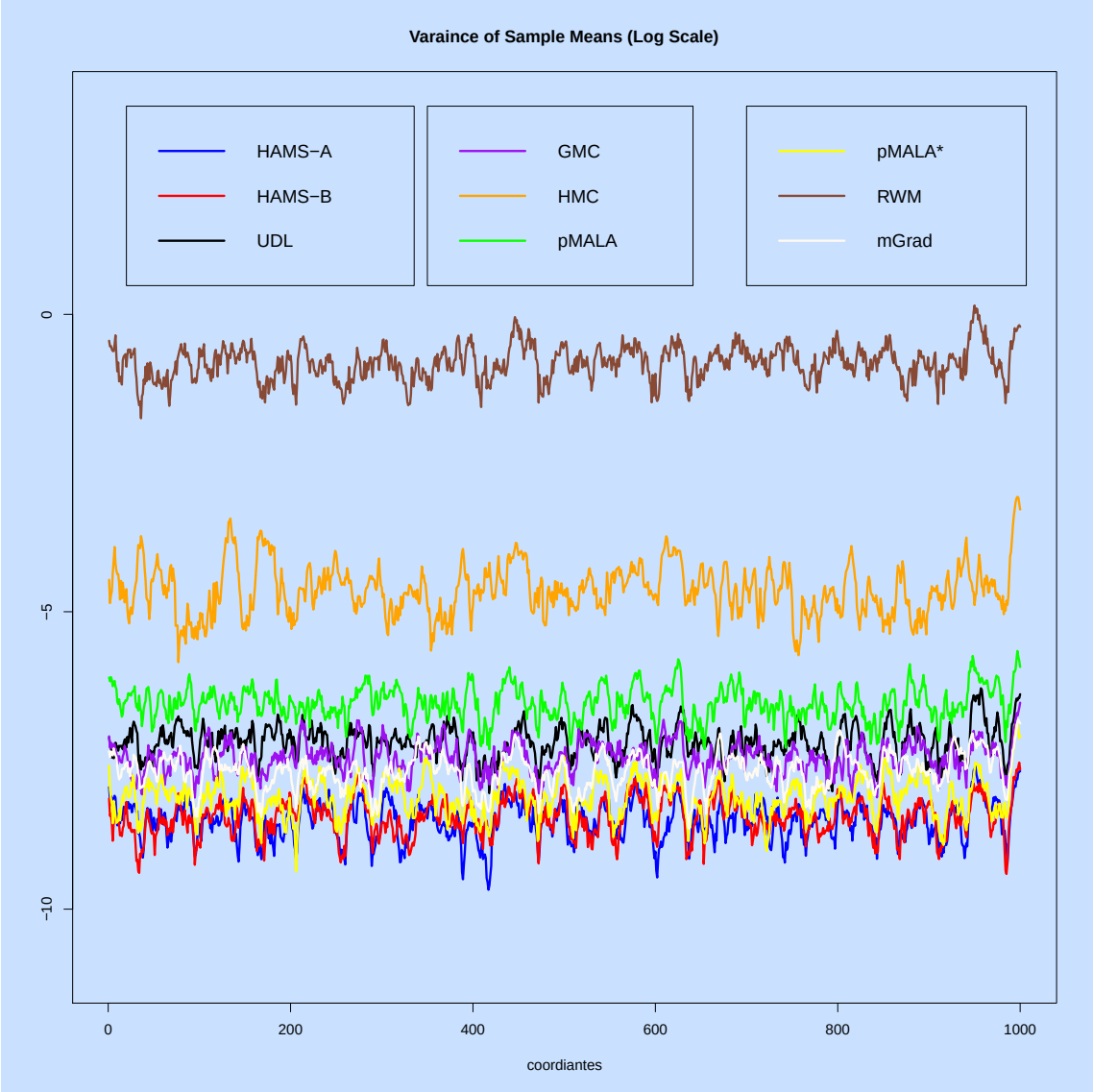


Figure S17: Time-adjusted variances of sample means (log-scale) of all latent variables over 50 repetitions for sampling latent variables in the stochastic volatility model.

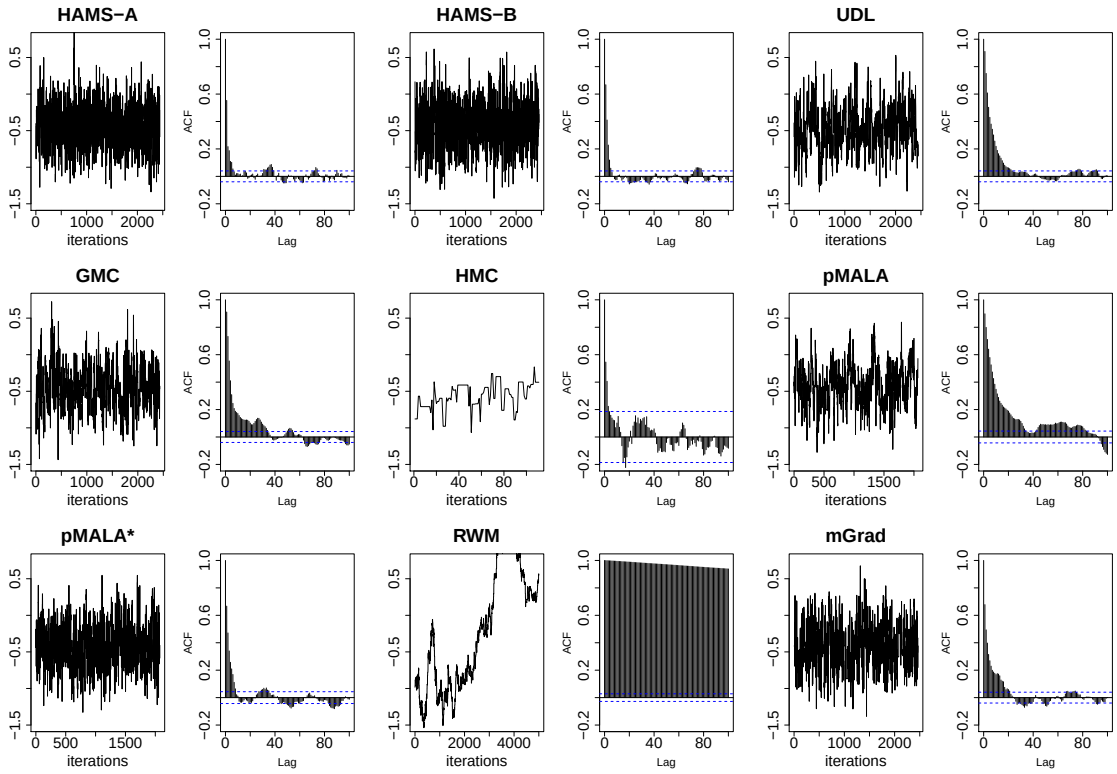


Figure S18: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the stochastic volatility model.

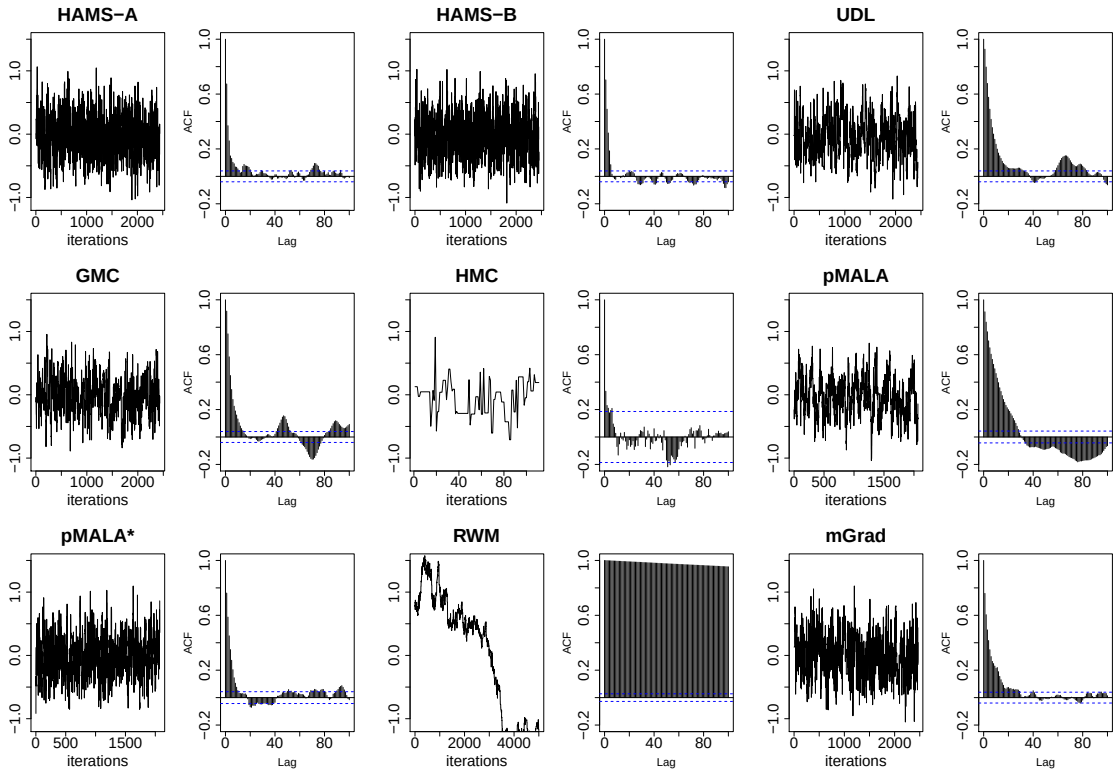


Figure S19: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the stochastic volatility model.

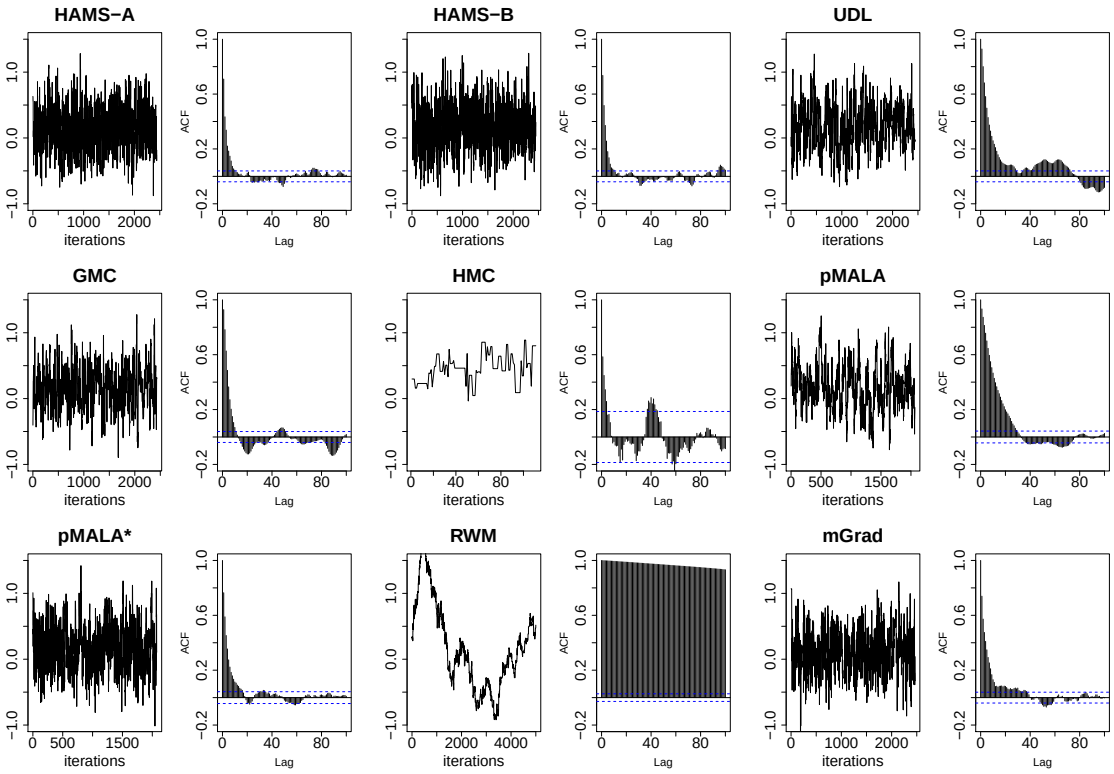


Figure S20: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the stochastic volatility model.

Method	Time (s)	Sample Mean			ESS_1	$\frac{\text{minESS}_1}{\text{Time}}$	ESS_2	$\frac{\text{minESS}_2}{\text{Time}}$
		β (sd)	σ (sd)	ϕ (sd)	(β, σ, ϕ)		(β, σ, ϕ)	
HAMS-A	1951.3	0.68 (0.034)	0.19 (0.006)	0.98 (0.001)	(30, 73, 220)	0.015	(7, 9, 58)	0.004
HAMS-B	1942.3	0.68 (0.037)	0.19 (0.007)	0.98 (0.001)	(25, 58, 188)	0.013	(6, 8, 38)	0.003
UDL	1945.8	0.68 (0.039)	0.20 (0.008)	0.98 (0.002)	(29, 37, 87)	0.015	(6, 9, 20)	0.003
GMC	1968.2	0.67 (0.059)	0.20 (0.007)	0.98 (0.003)	(35, 68, 169)	0.018	(3, 11, 10)	0.001
HMC	20920.2	0.69 (0.050)	0.19 (0.014)	0.98 (0.003)	(19, 12, 78)	0.001	(5, 1, 7)	0.00006
pMALA	2013.0	0.68 (0.039)	0.20 (0.005)	0.98 (0.001)	(15, 30, 76)	0.008	(5, 22, 34)	0.002
pMALA*	2015.2	0.70 (0.054)	0.19 (0.006)	0.98 (0.001)	(23, 54, 149)	0.012	(3, 10, 34)	0.002
RWM	1311.1	0.76 (0.050)	0.47 (0.229)	0.51 (0.149)	(89, 12, 7)	0.005	(0.16, 0.01, 0.32)	0.00001

Table S4: Comparison of posterior sampling in the stochastic volatility model. Standard deviations of sample means are in parentheses. Results are averaged over 20 repetitions.

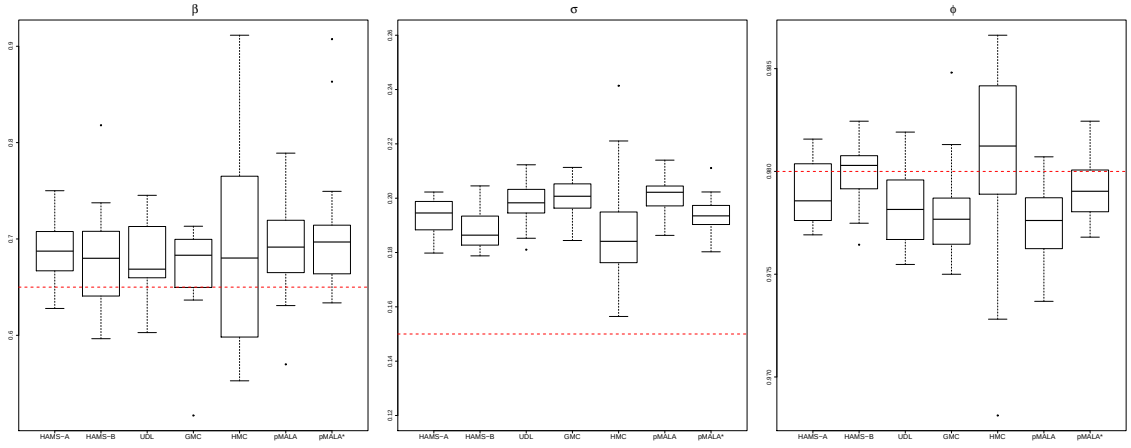
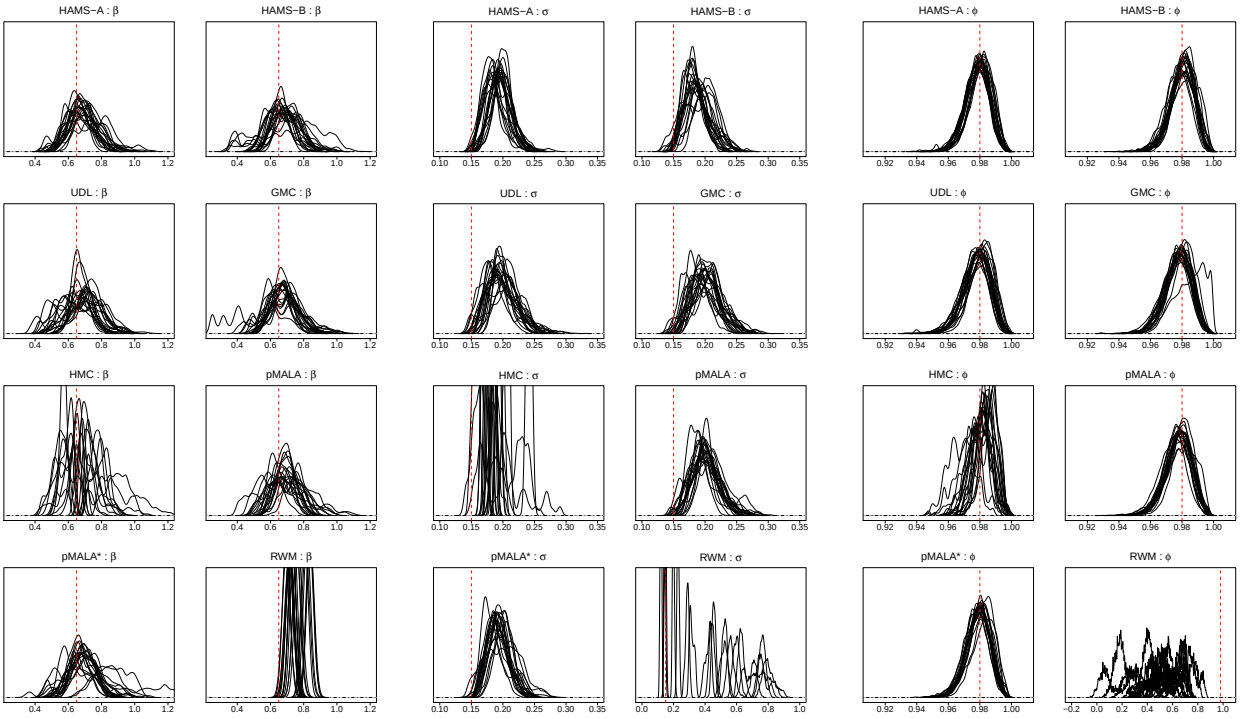
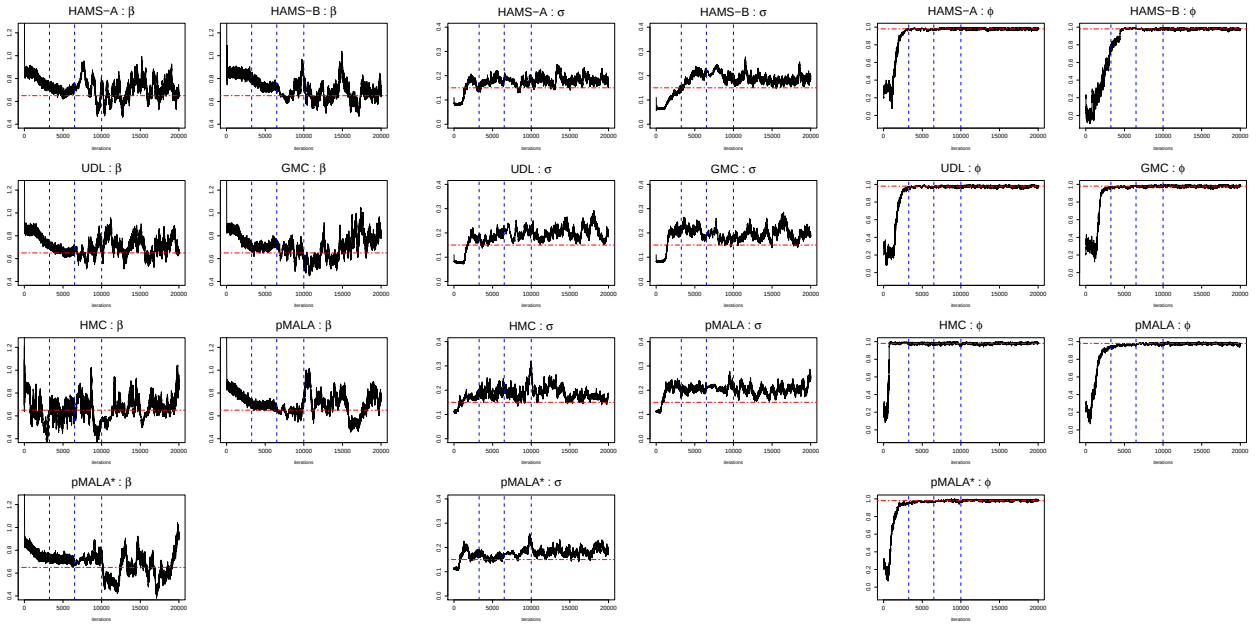


Figure S21: Time-adjusted boxplots of posterior means of parameters over 20 repetitions for posterior sampling in the stochastic volatility model. The data generating parameter values are marked by red lines.



(a) Densities of β (b) Densities of σ (c) Densities of ϕ

Figure S22: Time-adjusted posterior density plots of parameters (20 repetitions overlaid) in the stochastic volatility model. The true parameter values are marked by vertical lines.



(a) Trace plots of β (b) Trace plots of σ (c) Trace plots of ϕ

Figure S23: Trace plots of β, σ and ϕ, from an individual run for posterior sampling in the multilevel. The data generating parameter values are marked by red horizontal lines. There are four sub-stages divided by blue vertical lines. The first two are without preconditioning, with 3250 iterations each. The last two are with preconditioning, with 3500 and 10000 iterations respectively. The first three sub-stages are counted as burn-in.

VI.4 Log-Gaussian Cox model

We report additional simulation results for the log-Gaussian Cox model discussed in Section 5.2. When only sampling latent variables, mGrad is comparable to pMALA, pMALA* improves upon the original pMALA, and GMC shows comparable performance to UDL. From Figures S25–S27, while all methods have similar average sample means, HAMS-A and HAMS-B have the smallest variances in sample means.

Additional results from posterior sampling are provided in Figures S31–S33. GMC has similar performance to UDL for both parameters (σ^2, β). The average posterior mean of pMALA* is closer to the true value of σ^2 than HAMS-A, but the standard deviation from pMALA* is also larger. For β, pMALA* shows even larger spread. Compared to the stochastic volatility model, the effect of preconditioning can be seen more clearly from the trace plots in Figure S33.

Method	Time (s)	ESS_1 (min, median, max)	$\frac{\text{minESS}_1}{\text{Time}}$	ESS_2 (min, median, max)	$\frac{\text{minESS}_2}{\text{Time}}$
HAMS-A	2013.5	(1015, 1530, 3950)	0.50	(207, 464, 1084)	0.10
HAMS-B	1998.5	(629, 953, 1931)	0.31	(143, 290, 1005)	0.07
UDL	1997.8	(361, 576, 1187)	0.18	(87, 172, 563)	0.04
GMC	1999.1	(397, 625, 1465)	0.20	(98, 185, 532)	0.05
HMC	44425.1	(1011, 7381, 12824)	0.02	(29, 330, 3567)	0.001
pMALA	2862.4	(246, 382, 797)	0.09	(55, 113, 263)	0.02
pMALA*	2873.0	(611, 903, 1955)	0.21	(145, 272, 696)	0.05
mGrad	3064.6	(245, 410, 1666)	0.08	(55, 120, 550)	0.02
RWM	1217.6	(7, 11, 22)	0.01	(0.1, 0.3, 0.7)	0.0001

Table S5: Runtime and ESS comparison (including mGrad) for sampling latent variables in the log-Gaussian Cox model ($n = 4096$). Results are averaged over 50 repetitions.

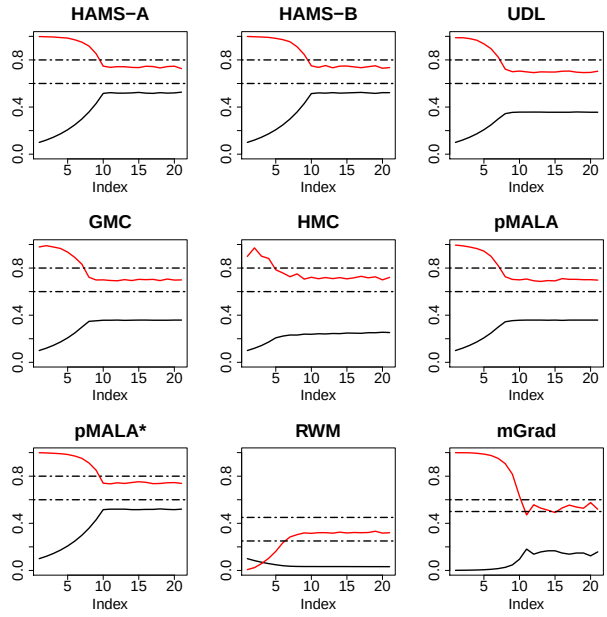


Figure S24: Average step sizes (black) and acceptance rates (red) for sampling latent variables in the log-Gaussian Cox model ($n = 4096$). For every 250 iterations, acceptance rates are calculated and step sizes adjusted. Results are averaged over 50 repetitions.

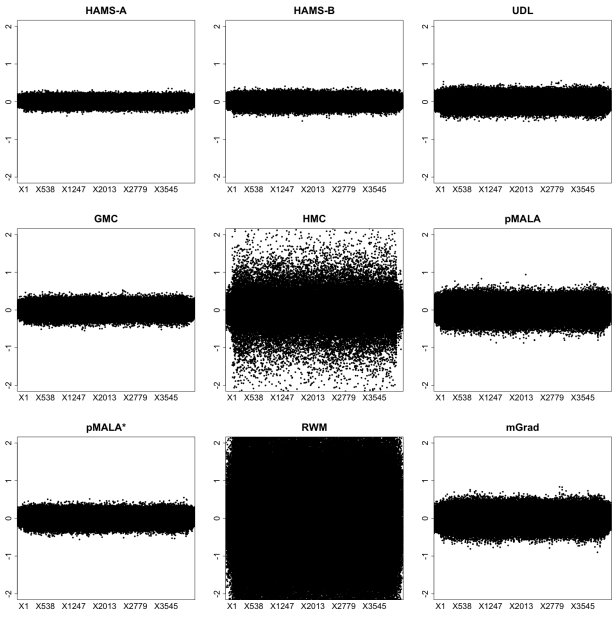


Figure S25: Time-adjusted and centered plots of sample means of all latent variables over 50 repetitions for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

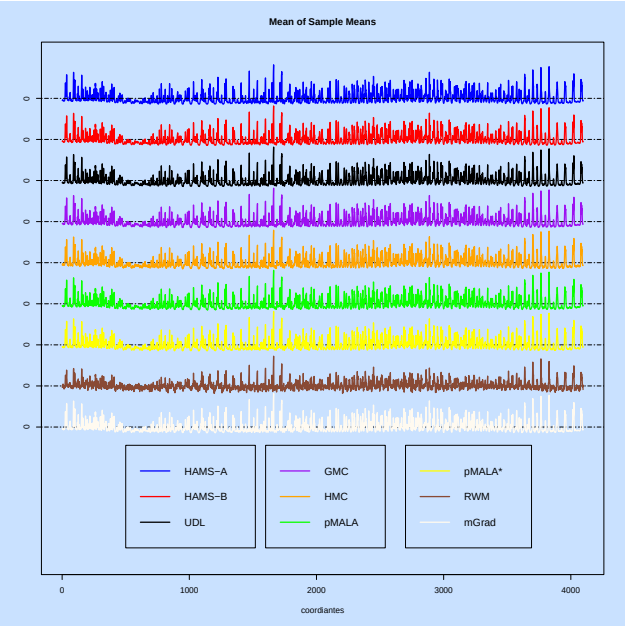


Figure S26: Time-adjusted averages of sample means (shifted) of all latent variables over 50 repetitions for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

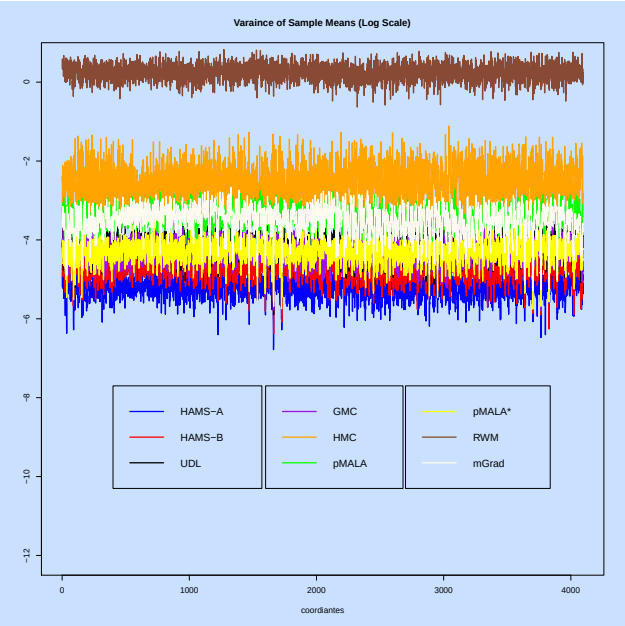


Figure S27: Time-adjusted variances of sample means (log-scale) of all latent variables over 50 repetitions for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

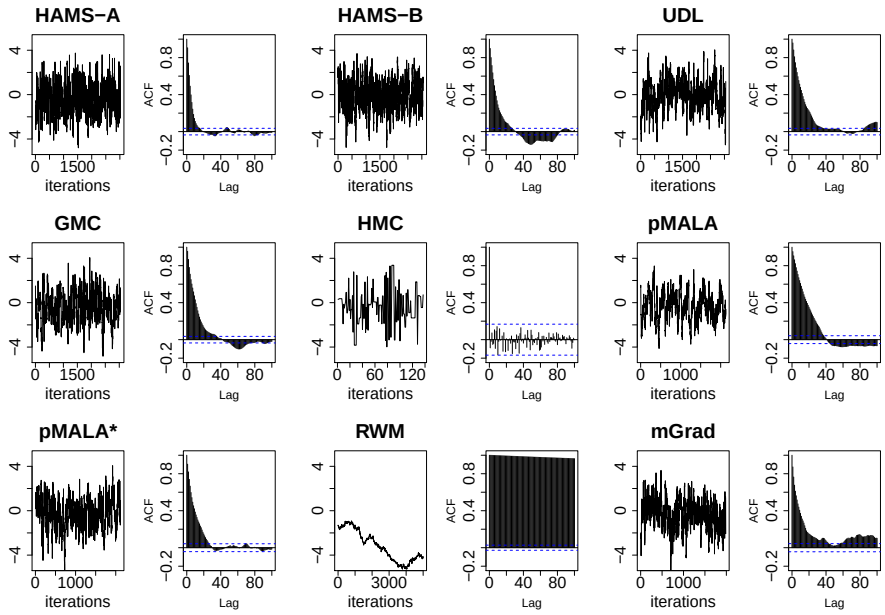


Figure S28: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

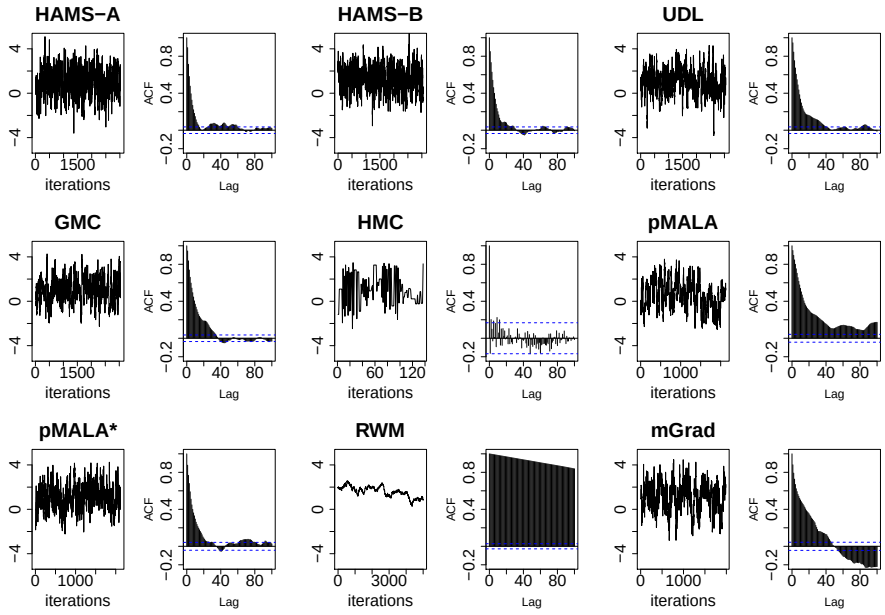


Figure S29: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

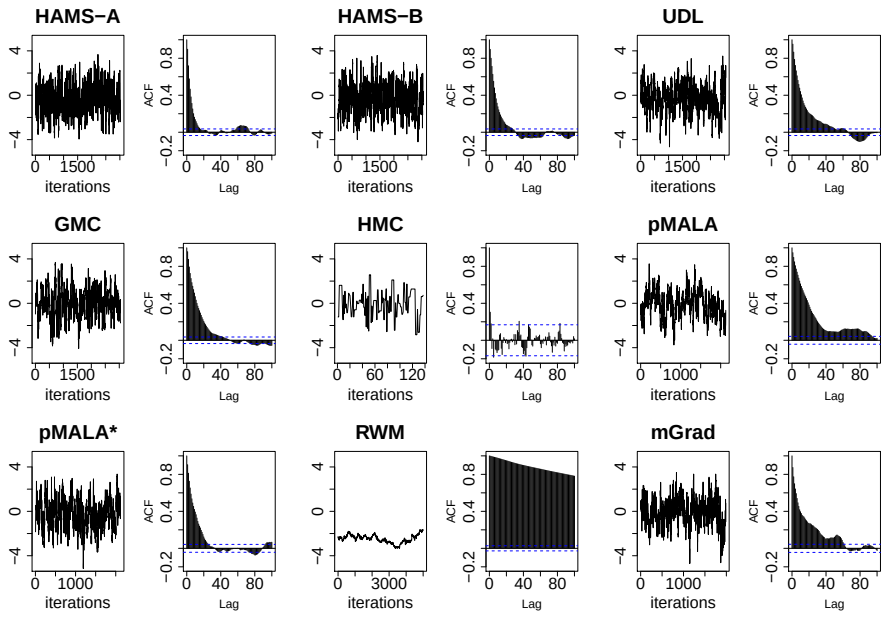


Figure S30: Time-adjusted trace and ACF plots of one latent variable from an individual run for sampling latent variables in the log-Gaussian Cox model ($n = 4096$).

Method	Time (1000 s)	Sample Mean		ESS_1	$\frac{\text{ESS}_1}{\text{Time}}$	ESS_2	$\frac{\text{ESS}_2}{\text{Time}}$
		σ^2 (sd)	β (sd)	(σ^2, β)	(σ^2, β)	(σ^2, β)	(σ^2, β)
HAMS-A	161.1	2.08 (0.086)	0.04 (0.010)	(178, 24)	(1.105, 0.147)	(10.9, 1.0)	(0.068, 0.006)
HAMS-B	161.0	3.26 (1.309)	0.57 (0.560)	(554, 468)	(3.441, 2.908)	(2.4, 0.7)	(0.015, 0.005)
UDL	161.0	2.11 (0.109)	0.04 (0.006)	(109, 31)	(0.677, 0.190)	(7.7, 2.0)	(0.048, 0.012)
GMC	160.8	2.18 (0.112)	0.03 (0.007)	(91, 31)	(0.566, 0.194)	(8.0, 1.2)	(0.050, 0.007)
HMC	1366.9	2.45 (0.850)	0.21 (0.436)	(342, 375)	(0.250, 0.274)	(1.9, 0.4)	(0.001, 0.0003)
pMALA	162.5	2.08 (0.207)	0.04 (0.019)	(75, 17)	(0.462, 0.103)	(2.8, 0.4)	(0.017, 0.003)
pMALA*	162.4	1.97 (0.092)	0.05 (0.029)	(116, 32)	(0.714, 0.195)	(19.0, 0.7)	(0.117, 0.004)
RWM	82.5	2.42 (1.074)	0.16 (0.158)	(304, 279)	(3.685, 3.377)	(1.1, 0.6)	(0.013, 0.007)

Table S6: Comparison of posterior sampling in the log-Gaussian Cox model ($n = 4096$). Standard deviations of sample means are in parentheses. Results are averaged over 15 repetitions.

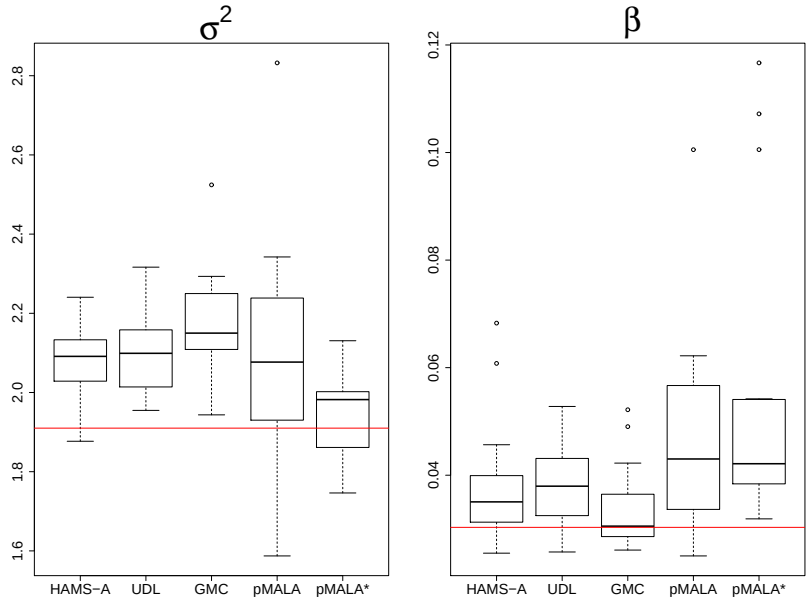
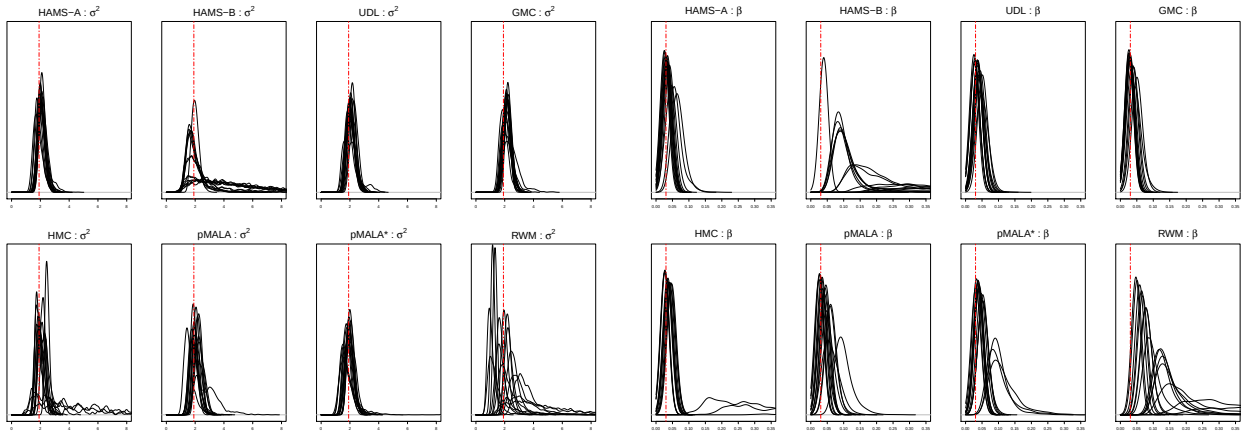
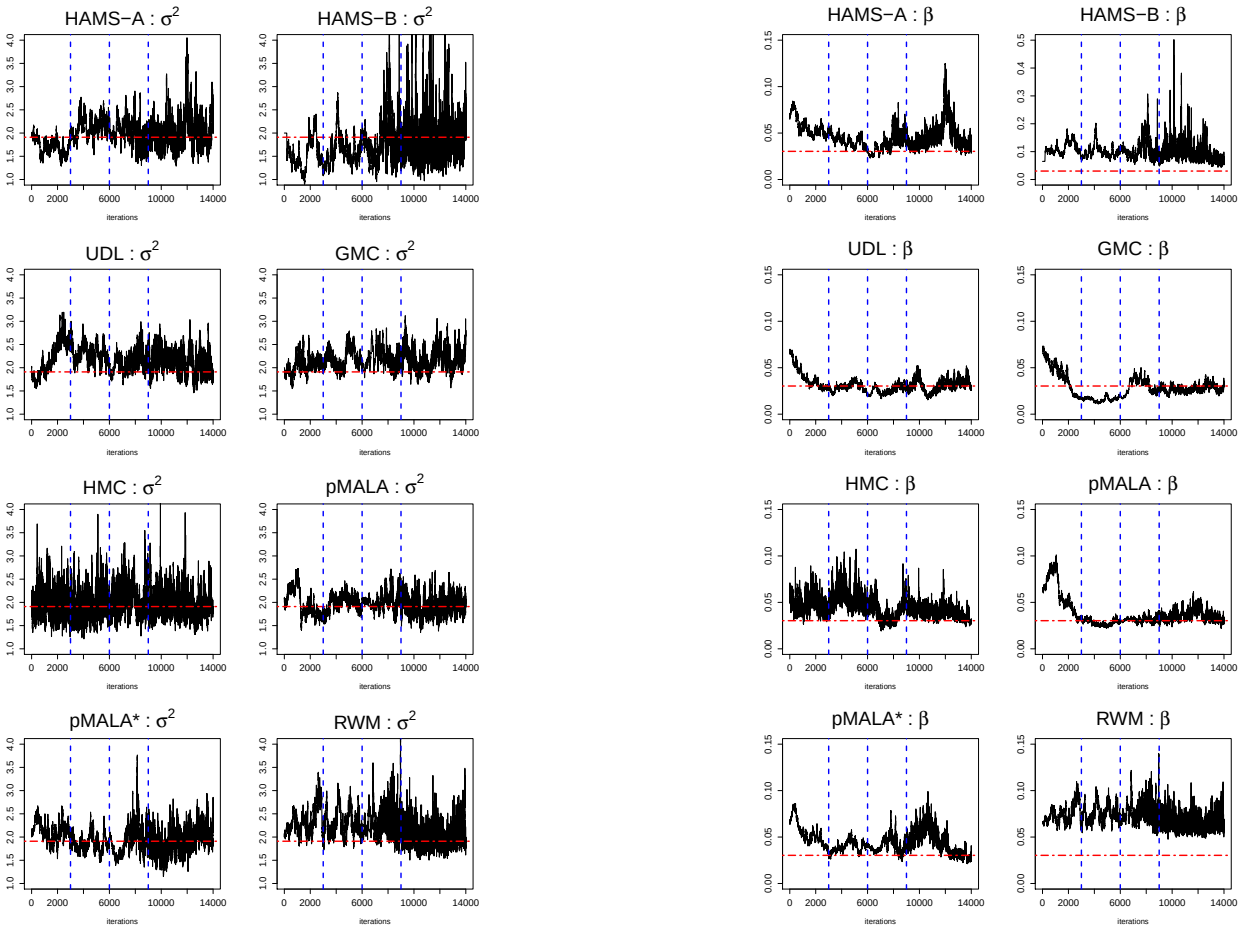


Figure S31: Time-adjusted boxplots of posterior means of parameters over 15 repetitions for posterior sampling in the log-Gaussian Cox model ($n = 4096$). HAMS-B, HMC and RWM are not included due to large outliers. The data generating parameter values are marked by red lines.



(a) Densities of σ^2 (b) Densities of β

Figure S32: Time-adjusted posterior density plots (15 repetitions overlaid) in log-Gaussian Cox model ($n = 4096$). The true parameter values are marked by vertical lines.



(a) Trace plots of σ^2

(b) Trace plots of β

Figure S33: Trace plots from an individual run for posterior sampling in the log-Gaussian Cox model ($n = 4096$). Data generating parameter values are marked by red horizontal lines. There are four sub-stages divided by blue vertical lines. The first two are without preconditioning, with 3000 iterations each. The last two are with preconditioning, with 3000 and 5000 iterations respectively. The first three sub-stages are counted as burn-in.