Adaptive sequential Monte Carlo for structured cross validation in Bayesian hierarchical models

Geonhee Han[*]

Graduate School of Arts and Sciences, Columbia University[†]

Graduate School of Public Policy, The University of Tokyo

Andrew Gelman

Department of Statistics and Department of Political Science, Columbia University

11 Aug 2025

Abstract

Importance sampling (IS) is commonly used for cross validation (CV) in Bayesian models, because it only involves reweighting existing posterior draws without needing to re-estimate the model by re-running Markov chain Monte Carlo (MCMC). For hierarchical models, standard IS can be unreliable; the out-of-sample generalization hypothesis may involve structured case-deletion schemes which significantly alter the posterior geometry. This can force costly MCMC re-runs and make CV impractical. As a principled alternative, we tailor adaptive sequential Monte Carlo to sample along a path of posteriors that leads to the case-deleted posterior. The sampler is designed to support various hypotheses by accommodating diverse CV designs, and to streamline the workflow by automating path construction and systematically minimizing MCMC intervention. We demonstrate its utility with three types of predictive model assessment: longitudinal leave-group-out CV, group K-fold CV, and sequential one-step-ahead validation.

Keywords: Cross validation, Bayesian hierarchical models, Sequential Monte Carlo, Predictive model evaluation, Bayesian workflow

---

[*]gh2610@columbia.edu.

[†]The majority of this research was carried out while GH was a graduate student at Columbia University GSAS QMSS.

1 Introduction

Evaluating the fit of a Bayesian model by identifying discrepancies between the model and the data is a crucial step of the Bayesian workflow (Gelman et al., 2020). In particular, *predictive model assessment* focuses on how well a model can predict new and unseen data, often assessed via cross validation (Stone, 1976; Geisser, 1975; Geisser and Eddy, 1979; Arlot and Celisse, 2010; Vehtari and Ojanen, 2012; Piironen and Vehtari, 2016).

With Bayesian models, cross validation (CV) is known to be computationally intensive due to the need for re-estimating the posterior distributions for datasets that omit subsets of observations. For example, naïve leave-one-out cross validation (LOO-CV) requires separate posterior estimations for each omitted observation, typically performed using computationally expensive methods such as Markov chain Monte Carlo (MCMC); this makes naïve LOO-CV computationally impractical for large datasets or complex models. A popular remedy is to use importance sampling (IS) and its variants (Gelfand and Dey, 1994; Peruggia, 1997; Epifani et al., 2008; Vehtari et al., 2017; Lobo et al., 2020), which approximate the case-deleted posterior by re-weighting posterior samples obtained from the full dataset, circumventing the need for repeated re-estimation and providing substantial computational savings.

There is considerable interest in efficient CV for structured Bayesian hierarchical models, such as those with longitudinal, spatial, or temporal structure. For example, with models for grouped data, identifying groups that are challenging to predict with leave-group-out CV (Merkle et al., 2019; Liu and Rue, 2023; Adin et al., 2024; Zhang et al., 2024) can highlight specific groups where the hierarchical model struggles to predict and motivate model expansions (Gelman et al., 2020, Chap. 6.2).

A potential challenge with IS for such Bayesian models is its instability in estimates, such as due to possibly infinite variance in importance weights (Vehtari et al., 2017; Millar, 2018; Silva and Zanella, 2023; Chang et al., 2024). CV in structural Bayesian models often requires non-standard design of blocking structures and out-of-sample prediction schemes that account for intricate dependencies (Gelman et al., 2014b; Roberts et al., 2017). Such

case-deletion schemes can result in distant posteriors that (a) a vanilla IS estimator would struggle to approximate accurately and reliably, and (b) would inevitably necessitate additional runs of MCMC to re-approximate the case-deleted posterior, which is extremely impractical. Some examples are spatial, temporal, and nested multilevel structures (e.g., phylo-genetic models) which involve dependent observations that are highly informative to the posterior geometry: see Bürkner et al. (2020), Bürkner et al. (2021), Lobo et al. (2020), and Martínez-Minaya and Rue (2024).

Research on computational methods to efficiently perform CV with structural blocking or case-deletion schemes remains limited. Recent work by Liu and Rue (2023) introduced methods for approximating a leave-group-out estimand in latent Gaussian models, leveraging the conditional independence of observations given linear Gaussian predictors. Their approach uses direct numerical integration by exploiting the inherent tractability of latent Gaussian models. In Bayesian hierarchical models, Zhang et al. (2024) focus on estimating cross validated means rather than the (log) predictive density. Mixture estimators have been introduced by Silva and Zanella (2023) for the computation of LOO-CV estimands, where the asymptotic variance of the weights is finite, although additional simulation from a proposal (often of a non-standard form) is required, essentially necessitating re-runs of, say, MCMC. Efforts to avoid MCMC re-runs through moment matching were explored by Paananen et al. (2021, 2024), while the authors also concede that affine transformation may be insufficient to produce suitable proposals and suggest that more complex methods may be needed, beyond LOO-CV. Other existing work explores case-deleted posterior approximations using a local sensitivity approach for sensitivity analysis (Ghosh et al., 2020; Broderick et al., 2023; Nguyen et al., 2024; Huang et al., 2024), rather than model evaluation.

Our goal is to develop a computational approach applicable to a wide range of structural Bayesian models and CV schemes, which can be executed as a byproduct of a single MCMC run on a full non-case-deleted dataset, complementing the prominent MCMC-based Bayesian workflow. The method adopts the adaptive sequential Monte Carlo (aSMC) sampler (Del Moral et al., 2006; Jasra et al., 2011), and bridges distant posteriors by au-

tomatically constructing a sequence of auxiliary intermediate distributions leading to the target case-deleted posterior(s). The sampler is applicable to a wide range of models and CV schemes, while allowing one to avoid the costly MCMC re-runs whenever possible, and further being equipped with design-efficient sample-generating capabilities, unlike existing IS methods, even when the target posteriors are detected to be distant.

The structure of the paper is as follows. Section 2 describes the setup and explores various structural CV schemes. In Section 3, we consider the aSMC approach. Section 4 demonstrates the application of the method in three examples involving grouped, time-series, and spatial data. Section 5 concludes with remarks and discussions.

2   Predictive evaluation of Bayesian hierarchical models

2.1   Bayesian hierarchical model

We index the groups by $g \in \{1, \ldots, G\}$ and the observations in each group by $i \in \{1, \ldots N_g\}$. Consider a Bayesian hierarchical model where $y_{g,i}$ represents the i-th observation within group g (á la Gelman et al., 2014a, Section 5.2), defined as

$$\phi \stackrel{ind}{\sim} p(\cdot), \qquad \theta_g \,|\, \phi \stackrel{ind}{\sim} p(\cdot \,|\, \phi), \qquad y_{g,i} \,|\, \theta_g, \phi \stackrel{ind}{\sim} p(\cdot \,|\, \theta_g, \phi),$$

where ϕ is a global parameter (hyperprior) and θ_g is a group-specific parameter. None are restricted to being univariate. The posterior distribution of the parameters $\Theta = (\phi, \theta_{1:G})$ is proportional to the joint distribution

$$p(\phi) \prod_{g=1}^{G} p(\theta_g \,|\, \phi) \prod_{i=1}^{N_g} p(y_{g,i} \,|\, \theta_g, \phi),$$

up to a normalizing constant. We assume that $y \mapsto p(y_{g,i} = y \,|\, \theta_g, \phi)$ may be evaluated.

Examples.   Some examples of models which may involve non-standard structural CV schemes are as follows.

- **Grouped models**: With grouped or panel data, group-specific parameters capture variation across units in a group or over time,

$$y_{g,i} = \boldsymbol{x}_{g,i}^{\mathsf{T}} \boldsymbol{\beta}_g + \varepsilon_{g,i},$$

where $\boldsymbol{\beta}_g \overset{ind}{\sim} p(\cdot)$ represents the coefficients specific to group g, and $\boldsymbol{x}_{g,i}$ denotes covariates. $\varepsilon_{g,i}$ are conditionally independent, and the conditional likelihood of $y_{g,i}$ is evaluable.

- **(Hierarchical) spatial regression**: Structured covariation within groups, such as spatial correlation, may involve

$$\boldsymbol{y}_{g,i} = \boldsymbol{X}_{g,i} \boldsymbol{\beta}_g + \boldsymbol{\omega}_{g,i} + \boldsymbol{\varepsilon}_{g,i},$$

Spatial dependence may involve, e.g.,

$$\boldsymbol{\omega}_{g,i} \overset{ind}{\sim} \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{V}_g), \qquad \boldsymbol{\varepsilon}_{g,i} \overset{ind}{\sim} \mathrm{MVN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

where $\mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean and covariance $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and covariance $\boldsymbol{V}_g$ expresses the g-specific intra-dependency. Here, we treat the (g, i)-th response as multivariate.

- **Dynamic normal linear models**: Temporal dynamics within groups may involve

$$\boldsymbol{y}_{g,t} = \boldsymbol{X}_{g,t} \boldsymbol{\beta}_{g,t} + \boldsymbol{\varepsilon}_{g,t}^{(y)}, \qquad \boldsymbol{\varepsilon}_{g,t}^{(y)} \overset{ind}{\sim} \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$
$$\boldsymbol{\beta}_{g,t} = \boldsymbol{\beta}_{g,t-1} + \boldsymbol{\varepsilon}_{g,t}^{(\beta)}, \qquad \boldsymbol{\varepsilon}_{g,t}^{(\beta)} \overset{ind}{\sim} \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{V}),$$

and $\boldsymbol{\beta}_{g,0} \overset{ind}{\sim} p(\cdot)$. Here, we have identified $\phi := (\boldsymbol{\Sigma}, \boldsymbol{V})$ and $\theta_g := \boldsymbol{\beta}_{g,0:T} := (\boldsymbol{\beta}_{g,0}, \ldots, \boldsymbol{\beta}_{g,T})^{\mathsf{T}}$, and $\boldsymbol{\beta}_{g,t}$ evolves smoothly over time with $p(\theta_g \,|\, \phi) = p(\boldsymbol{\beta}_{g,0} \,|\, \phi) \prod_{t=1}^{T} p(\boldsymbol{\beta}_{g,t} \,|\, \boldsymbol{\beta}_{g,t-1}, \boldsymbol{V})$.

2.2 Case-deletion schemes and computation

A widely used method to evaluate the fit of a Bayesian model is to assess its out-of-sample predictive performance (Roberts, 1965; Guttman, 1967; Geisser and Eddy, 1979; Vehtari and Ojanen, 2012). One of many approaches is within-sample CV (Stone, 1977), with

advancements in computationally efficient techniques such as approximate LOO-CV using IS (Gelfand and Dey, 1994; Peruggia, 1997; Epifani et al., 2008; Vehtari et al., 2017). For a more detailed overview of these methods, see Vehtari et al. (2016). We present examples of possible structural schemes in Bayesian hierarchical models below to provide an overview and highlight potential computational challenges associated with structural CV.

2.2.1 Leave one-in-group out (LOO)

The LOO-CV scheme, as outlined in Vehtari et al. (2017), can be applied to the above Bayesian hierarchical model as follows. The leave-(h, j)-out posterior, corresponding to excluding the observation $y_{h,j}$ with $h \in \{1, \ldots, G\}$ and $j \in \{1, \ldots, N_h\}$, is defined as:

$$p_{-(h,j)}(y^*, \boldsymbol{\Theta}) \propto p(\phi) \prod_{g=1}^{G} p(\theta_g \,|\, \phi) \prod_{i=1}^{N_g} p(y_{g,i} \,|\, \theta_g, \phi)^{\mathbb{I}\{(g,i) \neq (h,j)\}} p(y^* \,|\, \theta_g, \phi)^{\mathbb{I}\{(g,i)=(h,j)\}},$$

where y^* is the posterior predictive for observation (h, j), treated as unobserved, along with the model parameters $\boldsymbol{\Theta}$.

Under the logarithmic scoring rule (Gneiting and Raftery, 2007), out-of-sample predictive accuracy for the excluded unit is evaluated via its log posterior predictive distribution. Under the hierarchical model, the posterior predictive distribution for a new replication within group g under the leave-(g, i)-out posterior is obtained by integrating out the parameters $\boldsymbol{\Theta}$,

$$p_{-(g,i)}(y) = \int p_{-(g,i)}(y_{g,i}^* = y, \boldsymbol{\Theta}) \, d\boldsymbol{\Theta}.$$

Then, given the collection $(y_{g,i})_{g,i}$, the log pointwise predictive density for *new* within-group observations (Vehtari et al., 2017; Gelman et al., 2014a, Chap. 7) would be

$$\ell^{(\text{LOO})} := \sum_{g=1}^{G} \sum_{i=1}^{N_g} \log p_{-(g,i)}(y_{g,i}).$$

A naïve approach to compute the estimand requires (re-)executing posterior inference $N_1 + \ldots + N_G$ times, once for each leave-(g, i)-out posterior. This strategy is clearly computationally expensive and impractical. A more efficient strategy uses importance

weighting, taking advantage of posterior samples from the *baseline* non-case-deleted posterior. One in practice approximates the leave-(g, i)-out posterior by giving the draws $\Theta^{(r)} = (\phi^{(r)}, \theta_{1:G}^{(r)}) \sim p(\cdot \mid \boldsymbol{y}_{1:G})$ over $r = 1, \ldots, R$ (where R is the number of posterior MCMC draws) an importance ratio

$$w_{g,i}^{(r)} := p(y_{g,i} \mid \theta_g^{(r)}, \phi^{(r)})^{-1}, \qquad W_{g,i}^{(r)} := \frac{w_{g,i}^{(r)}}{\sum_{r=1}^{R} w_{g,i}^{(r)}},$$

where $w_{g,i}^{(r)}$ is the (g, i)-importance ratio and $W_{g,i}^{(r)}$ its self-normalized weight for the r-th draw. The LOO estimand is then approximated by the right-hand side approximation,

$$\ell^{(\mathrm{LOO})} \approx \sum_{g=1}^{G} \sum_{i=1}^{N_g} \log \sum_{r=1}^{R} W_{g,i}^{(r)} p(y_{g,i}^* = y_{g,i} \mid \theta_g^{(r)}, \phi^{(r)}) =: \hat{\ell}^{(\mathrm{LOO})}, \tag{1}$$

which is computable if the predictive distribution in the summand over (g, i) can be evaluated, which is usually true.

2.2.2 Leave group out (LGO)

Case deletion in hierarchical models can extend beyond individual within-group observations to entire groups (e.g., Section 4.1). The leave-g-out posterior, where g denotes the excluded group, is defined as

$$p_{-(g,:)}(y_{1:N_g}^*, \Theta) \propto p(\phi) \prod_{h \neq g} p(\theta_h \mid \phi) \prod_{i=1}^{N_g} p(y_{h,i} \mid \theta_h, \phi)^{\mathbb{I}\{h \neq g\}} p(y_i^* \mid \theta_h, \phi)^{\mathbb{I}\{h = g\}}.$$

In other words, we condition on all but $\boldsymbol{y}_g = y_{g,1:N_g} = (y_{g,1}, \ldots y_{g,N_g})^{\mathsf{T}}$, which we simply write as $\boldsymbol{y}_{-g}$. Using this distribution, the posterior predictive density $p_{-(g,:)}(\cdot)$, evaluated at $\boldsymbol{y}_g$, defines a new estimand,

$$\ell^{(\mathrm{LGO})} := \sum_{g=1}^{G} \log p_{-(g,:)}(\boldsymbol{y}_g), \tag{2}$$

where $p_{-(g,:)}(\cdot)$ is the joint posterior predictive distribution marginalized over the leave-g-out posterior,

$$p_{-(g,:)}(\boldsymbol{y}_g) = \int p(\boldsymbol{y}_g^* = \boldsymbol{y}_g \mid \theta_g, \phi) p(\Theta \mid \boldsymbol{y}_{-g}) \, d\Theta.$$

The procedure is also described by Merkle et al. (2019) as the *approximate leave-one-cluster-out CV*. Unlike the LOO scheme for hierarchical models, which evaluates individual conditionally independent observations within a group (Section 2.2.1), LGO log pointwise predictive density assesses the joint predictive accuracy for a hypothetical replication of the entire group.

Under this setup, the unnormalized importance ratio is

$$w_g := p(\boldsymbol{y}_g \,|\, \theta_g, \phi)^{-1} = \prod_{i=1}^{N_g} p(y_{g,i} \,|\, \theta_g, \phi)^{-1},$$

where the second equality follows from the conditional independence of within-group observations $\boldsymbol{y}_g$. An estimate of the LGO log pointwise predictive density is computed by weighting then aggregating the respective joint predictive density,

$$\ell^{(\text{LGO})} \approx \hat{\ell}^{(\text{LGO})} := \sum_{g=1}^{G} \hat{\ell}_g^{(\text{LGO})} := \sum_{g=1}^{G} \log \sum_{r=1}^{R} W_g^{(r)} p(\boldsymbol{y}_g^* = \boldsymbol{y}_g \,|\, \theta_g^{(r)}, \phi^{(r)}). \tag{3}$$

The self-normalized weights $W^{(r)}$ are computed as before.

2.2.3 Backward-sequential leave end out (LEO)

For settings with inherent ordering (e.g., time-series data under dynamic models, in Section 4.2), it often makes sense to align blocking schemes with that ordering to better assess predictive performance on future sequentially arriving data. Below are non-exhaustive examples of backward-sequential exclusion of the most recent observations.

Within-group. Fix a group index g, and assume that i indexes time. We define the sequential LEO posterior for group g, which conditions upon all of $\boldsymbol{y}_{-g}$ and the subset $y_{g,1:t} = (y_{g,1}, \ldots, y_{g,t})^\mathsf{T}$ up to time t,

$$p_{-(g,\ t+1:T_g)}(y_{t+1:T_g}^*, \boldsymbol{\Theta}) \propto p(\phi) \prod_{h=1}^{G} p(\theta_h \,|\, \phi) \prod_{i=1}^{T_h} \begin{cases} p(y_i^* \,|\, \theta_h, \phi) & \text{if } (h = g) \text{ and } (i \in (t+1 : T_h))) \\ p(y_{h,i} \,|\, \theta_h, \phi) & \text{otherwise} \end{cases}.$$

$T_g = N_g$ now represents the horizon. We then reverse the time index as $t = T_g - 1, \ldots, 0$ for backward-sequential exclusion.

The unnormalized importance weight associated with this LEO scheme given t is $w_g = \prod_{i=t+1}^{T_g} p(y_{g,i} \,|\, \theta_g, \phi)^{-1}$. A natural estimand is the multi-step ahead log pointwise predictive density,

$$
\ell_g^{\text{(LEO)}} := \log \underset{\boldsymbol{\Theta}, y^*}{\mathbb{E}} \left(p_{-(g,\, t+1:T_g)}(y_{t+1}^*, \,\ldots,\, y_{t+h-1}^*, \, \underbrace{y_{t+h}^* = y_{g,t+h}}_{\text{evaluate } y_{g,t+h}}, \, \boldsymbol{\Theta}) \,\Big|\, \boldsymbol{y}_{-g}, y_{g,1:t} \right)
$$

$$
\approx \log \sum_{r=1}^{R} W_g^{(r)} p_{-(g,\, t+1:t+h)}([y_{(t+1):(t+h-1)}^*]^{(r)}, \, y_{t+h}^* = y_{g,t+h}, \, \boldsymbol{\Theta}^{(r)}) =: \hat{\ell}_g^{\text{(LEO)}},
$$

after marginalizing out the posterior predictive after y_{t+h}^*. The operator $\underset{\boldsymbol{\Theta}}{\mathbb{E}}$ denotes expectation with respect to $\boldsymbol{\Theta}$. In most modeling situations, it would likely be the case that $G = 1$, and a joint model (e.g., dynamic generalized linear model: West et al., 1985; West and Harrison, 1997, Chap. 16) specifies the inter-temporal and inter-variable dependence of the observed multivariate sequence.

Across-group. For $G > 1$, and assuming that all groups have equal trajectory lengths $T_g = T$ for simplicity, the LEO posterior can be generalized across all groups by similarly indexing backward as $t = T - 1, \ldots, 0$ and defining the posterior by leaving out $(y_{1,t+1:T}, \ldots, y_{G,t+1:T})$, where the importance weight for this scheme is the product $\prod_{g=1}^{G} w_g = \prod_{g=1}^{G} \prod_{i=t}^{T} p(y_{g,i} \,|\, \theta_g, \phi)^{-1}$ which leads to the across-group evaluation $\ell^{\text{(LEO)}} := \sum_{g=1}^{G} \ell_g^{\text{(LEO)}}$.

The above are illustrative cases; more general schemes may be relevant when group-level series vary in length and/or are sampled at mixed temporal resolutions/frequencies, and these should be explicitly accommodated in the case-deletion scheme beyond the standard scheme.

2.2.4 Leave subset out (LSO)

The preceding CV schemes can be generalized by defining a set of indices $\mathcal{I}_k \subseteq \mathcal{I} := \bigcup_{g=1}^{G} \{g\} \times \{1, \ldots, N_g\}$ at which the corresponding observations are deleted from the baseline posterior. The unnormalized importance weight for this general case is $w_k := \prod_{(g,i) \in \mathcal{I}_k} p(y_{g,i} \,|\, \theta_g, \phi)^{-1}$.

Some further specific examples are *K-fold* or *group K-fold CV*, where the observations $(y_{g,i})_{g,i}$ are divided into K mutually exclusive partitions such that $\mathcal{I} = \bigsqcup_{k=1}^{K} \mathcal{I}_k$. The group K-fold is implemented by considering partitions $\mathcal{I}_{1:K}$ which appropriately accounts for strata or grouping structure (e.g., group K-fold in Section 4.3).

Multiple groups $\varnothing \subsetneq \mathcal{G} \subsetneq \{1, \ldots, G\}$ may likewise be excluded in what we can term *leave-groups-out*, which induces posterior predictive distributions and corresponding replications over multiple groups.

Agenda We illustrated that non-standard case-deletion schemes can yield importance weights formed by reciprocals of products of conditional likelihoods, and that this arises naturally in Bayesian hierarchical models and their (C)V schemes, such as LGO, LEO, and LSO (including K-fold and group K-fold). Instability is expected in these settings, especially when, as is likely, the baseline posterior exhibits thinner tails than its case-deleted counterparts(e.g., leading to high- or infinite-variance weights: Epifani et al., 2008; Vehtari et al., 2017; Silva and Zanella, 2023; Vehtari et al., 2024). In the next section, we design an aSMC sampler that automatically constructs a sequence of intermediate distributions bridging the baseline and the target distribution, to facilitate a stable approximation of geometries that are otherwise difficult to traverse directly.

3 Sequential Monte Carlo approach

3.1 Bridging distant posteriors via Markov kernels

Let the baseline unnormalized posterior be

$$\gamma_0(\boldsymbol{\Theta}) := p(\phi) \prod_{g=1}^{G} p(\theta_g \,|\, \phi) \prod_{i=1}^{N_g} p(y_{g,i} \,|\, \theta_g, \phi),$$

and the target unnormalized posterior be

$$\gamma_k(\boldsymbol{\Theta}) := p(\phi) \prod_{g=1}^{G} p(\theta_g \,|\, \phi) \prod_{i=1}^{N_g} p(y_{g,i} \,|\, \theta_g, \phi)^{\mathbb{I}\{(g,i) \notin \mathcal{I}_k\}}.$$

The index $k \in \{1, \ldots, K\}$ references the set of observation indices $\mathcal{I}_k$ that are to be deleted from the baseline posterior (e.g., LGO with $\mathcal{I}_g = \{g\} \times \{1, \ldots, N_g\}$). Then, induce K targets $p_1, \ldots, p_K$ from the respective unnormalized posteriors,

$$p_k(\boldsymbol{\Theta}) = \frac{\gamma_k(\boldsymbol{\Theta})}{Z_k},$$

where $Z_k = \int \gamma_k(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}$ is an unknown normalizing constant.

For each of these K targets, we now prepare a sequence of (intermediate) distributions $\gamma_{k,\ell}$ for $\ell \in \{0, 1, \ldots, L_k\}$ such that $\gamma_{k,0} = \gamma_0$ and $\gamma_{k,L_k} = \gamma_k$. Fixing k henceforth and following Del Moral et al. (2006), in a common product space, we introduce backward Markov kernels $(\mathcal{L}_{k,\ell-1})_{\ell=1}^{L_k}$ as

$$\tilde{p}_k(\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_{L_k}) := p_{k,L_k}(\boldsymbol{\Theta}_{L_k}) \prod_{\ell=1}^{L_k} \mathcal{L}_{k,\ell-1}(\boldsymbol{\Theta}_{\ell-1} \leftarrow \boldsymbol{\Theta}_\ell),$$

which admits $p_{k,L_k} = p_k$ as its marginal with respect to $\boldsymbol{\Theta}_{L_k}$. We then introduce forward Markov kernels $(\mathcal{K}_{k,\ell})_{\ell=1}^{L_k}$ such that

$$\tilde{q}_k(\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_{L_k}) := p_{k,0}(\boldsymbol{\Theta}_0) \prod_{\ell=1}^{L_k} \mathcal{K}_{k,\ell}(\boldsymbol{\Theta}_\ell \leftarrow \boldsymbol{\Theta}_{\ell-1}).$$

It follows from the Radon–Nikodym theorem that

$$\mathbb{E}_{\tilde{p}_k}(f(\boldsymbol{\Theta}_\ell)) = \mathbb{E}_{\tilde{q}_k}\left(f(\boldsymbol{\Theta}_\ell)\frac{\tilde{p}_k}{\tilde{q}_k}(\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_\ell)\right), \tag{4}$$

where $\tilde{p}_k/\tilde{q}_k$ is as follows. First, define the forward kernels $(\mathcal{K}_{k,\ell})_{\ell=1}^{L_k}$ as invariant kernels (e.g., MCMC) targeting the respective intermediate distributions $(p_{k,\ell})_{\ell=1}^{L_k}$. Further define the backward kernels as the time reversal of forward kernels, that is $\mathcal{L}_{k,\ell-1}(\boldsymbol{\Theta}_{\ell-1} \leftarrow \boldsymbol{\Theta}_\ell) := \mathcal{K}_{k,\ell}(\boldsymbol{\Theta}_\ell \leftarrow \boldsymbol{\Theta}_{\ell-1})\gamma_{k,\ell}(\boldsymbol{\Theta}_{\ell-1})/\gamma_{k,\ell}(\boldsymbol{\Theta}_\ell)$, we obtain incremental weights (Dai et al., 2022) within the expectation of the form

$$\frac{\tilde{p}_k}{\tilde{q}_k}(\boldsymbol{\Theta}_0, \ldots, \boldsymbol{\Theta}_\ell) \propto \prod_{l=1}^{\ell} \frac{\gamma_{k,l}(\boldsymbol{\Theta}_l)}{\gamma_{k,l-1}(\boldsymbol{\Theta}_{l-1})} \frac{\mathcal{L}_{k,l-1}(\boldsymbol{\Theta}_{l-1} \leftarrow \boldsymbol{\Theta}_l)}{\mathcal{K}_{k,l}(\boldsymbol{\Theta}_l \leftarrow \boldsymbol{\Theta}_{l-1})} =: \prod_{l=1}^{\ell} w_{k,l}(\boldsymbol{\Theta}_{l-1}),$$

where we have defined

$$w_{k,\ell}(\boldsymbol{\Theta}_{\ell-1}) := \frac{\gamma_{k,\ell}(\boldsymbol{\Theta}_{\ell-1})}{\gamma_{k,\ell-1}(\boldsymbol{\Theta}_{\ell-1})}. \tag{5}$$

(a) Leave-group-out (LGO) (b) Leave-subset-out (LSO) (c) Leave-end-out (LEO)
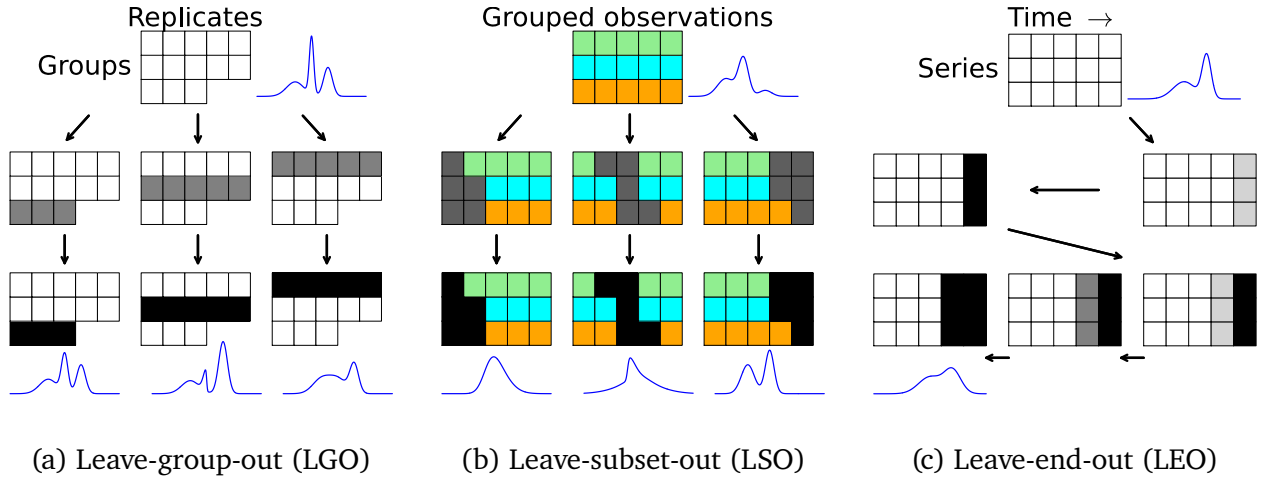
Figure 1: Different forms of case-deletion for Bayesian hierarchical models and corresponding aSMC-based approximation strategies, based on a single set of baseline (non-case-deleted) samples. Each block denotes a conditionally independent unit. Black blocks do not contribute to the model likelihood; gray blocks contribute *partially*. **Leave-subset-out** removes a carefully chosen subset of conditionally independent (possibly intra-dependent multivariate) observations. **Leave-group-out** is a special case. **Leave-end-out** backward-sequentially deletes temporally ordered observations (where each block may be multivariate and non-factorizable). **Across all schemes**, the sampler adaptively constructs auxiliary distributions as necessary to support reliable approximation.

We therefore arrive at an approximately executable sampler over the augmented space $\Theta_{0:L_k}$ provided access to draws from the baseline distribution $p_{k,0}$. The access holds in practice, and aligns with the usual computational setup, in the sense that a single *baseline* draw from the complete non-case-deleted posterior is available from established tools such as `Stan` (Carpenter et al., 2017). Note that we are abbreviating the dependence of the parameters on the index k for simplicity because this is fixed.

3.2   Parameterizing case deletions

Fixing k, a design choice lies in the sequence $\gamma_{k,1:L_k} = (\gamma_{k,1}, \ldots, \gamma_{k,L_K})$. See Figure 1; with LOO- or LGO-CV (Sections 2.2.1 and 2.2.2), the goal is to approximate the leave-g-out posteriors as the final distribution p_{g,L_g}; with backward-sequential LEO (Section 2.2.3), the intermediate distributions should contain those induced by leaving the data points backward, and they are also of interest, and not solely the final distribution and its marginal draw. We briefly explore how to parameterize such structural case-deletions.

3.2.1 Tempering for joint group deletion

Tempering is a natural choice to bridge the baseline and target (Swendsen and Wang, 1986; Marinari and Parisi, 1992; Hukushima and Nemoto, 1996; Neal, 2001). Taking the LGO posterior, for instance, we use an augmented likelihood contribution for group g (Agostinelli and Greco, 2013; Kallioinen et al., 2023) of the form

$$\rho_g(n) := p(\boldsymbol{y}_g | \theta_g, \phi)^{\varphi_g(n)},$$

where $\varphi_g : [0, N_g] \to [0, 1]$ is a continuous decreasing path such that $\varphi_g(0) = 1$ and $\varphi_g(N_g) = 0$; we parameterize a *geometric path* of distributions (Neal, 1993; Gelman and Meng, 1998) from the non-case deleted to leave-g-out posterior. (With LSO, set $N_k = 1$.)

A potential drawback is that existing MCMC algorithms, which were effective in targeting the baseline posterior ($n = 0$), may become unsuitable as the invariant kernel within aSMC. For example, a well-mixing Gibbs sampler that efficiently targets the baseline by exploiting conjugacy might become inapplicable when the power-scaled coefficients in the interior $(0, 1)$ induce likelihoods that do not admit conveniently simulatable conjugate priors. Some exceptions are presented in Kallioinen et al. (2023) and in Section 4.2.

3.2.2 Ordered continuous within-group deletions

A *continuous case deletion* over discrete $i \in \{1, \ldots, N_g\}$ with an augmented likelihood

$$\rho_g(n) := \prod_{i=1}^{N_g} p(y_{g,i} | \theta_g, \phi)^{\varphi_{g,i}(n)},$$

and the likelihood power-scaling factors $\varphi_{g,i} : [0, N_g] \to [0, 1]$ continuously parameterized by n, such as $\varphi_{g,i}(n) = \min\{\max\{0, i - n\}, 1\}$, defines a path such that $\rho_g(0) = p(\boldsymbol{y}_g | \theta_g, \phi)$ and $\rho_g(N_g) = 1$, and in particular $\rho_g(n) = \prod_{i=n+1}^{N_g} p(y_{g,i} | \theta_g, \phi)$ for integer n.

Unlike tempering, this is convenient when an efficient tailored Gibbs sampler exploiting conjugacy is available; the distribution induced from $n \in \{1, \ldots, N_g\}$ would simply be the posterior with observations $y_{g,i}$ such that $i \leq n$ are left out, which makes it suitable for cases with ordering on the observations (e.g., Figure 1c: by enforcing that such discrete

checkpoints are present along the trajectory). However, the approach is not applicable if conditional independence is violated.

3.3 Adaptive mechanisms

Choosing the backward kernels as time reversals in Section 3.1 yields the incremental weight function $w_{k,\ell}(\Theta_{\ell-1}) = \gamma_{k,\ell}(\Theta_{\ell-1})/\gamma_{k,\ell-1}(\Theta_{\ell-1})$. Note that this depends only on the previous step $\Theta_{\ell-1}$ and not on the ℓ-th step Θ_ℓ. This should be leveraged to design user-friendly adaptive mechanisms to simplify workflows.

3.3.1 Automating bridging

Eliminating the need to explicitly specify the intermediate distributions $(p_{k,1}, \ldots, p_{k,L_k})_{k=1}^K$ is advantageous to achieve what is described in Figure 1, automatically.

(a) In LGO, only marginal draws at the final step from the respective leave-g-out posteriors are of interest.

(b) In LEO, draws from the sub-intermediate distributions between the backward-sequentially case deleted distributions are only auxiliary.

The former was suggested in the related work by Bornn et al. (2010), though not implemented, and the latter has been touched upon by Bürkner et al. (2020) tangentially.

We implement these as follows. Given the previous-step particles $\Theta_{k,\ell-1}^{(r)} \sim p_{k,\ell-1}(\cdot)$ (now with index k) given the case deletion parameter $n_{k,\ell-1}$, we measure (the lack of) weight diversity by the effective sample size (ESS: Kong et al., 1994),

$$\mathrm{ESS}_{k,\ell} := \frac{1}{\sum_{r=1}^R (W_{k,\ell}^{(r)})^2}, \qquad W_{k,\ell}^{(r)} = \frac{w_{k,\ell}^{(r)}}{\sum_{r=1}^R w_{k,\ell}^{(r)}}, \qquad w_{k,\ell}^{(r)} = \frac{\gamma_{k,\ell}(\Theta_{k,\ell-1}^{(r)})}{\gamma_{k,\ell-1}(\Theta_{k,\ell-1}^{(r)})}.$$

Asymptotic connections between ESS and χ^2-divergence between the target and proposal distributions have been discussed in Agapiou et al. (2017). Under the case deletion parameterizations discussed in Section 3.2, the unnormalized weights can then be explicitly expressed as a function of the power coefficient in $[0, 1]$. We can solve for $n_{k,\ell} \in (0, N_k]$ (or

up to the pre-determined sub-intermediate *checkpoint* $n_{k,\ell+1} \leq N_k$ for LEO) to determine the next target distribution such that ESS meets a specified threshold, as ESS decreases in $n_{k,\ell}$ between the sub-intermediates; safe root-finding algorithms such as the bisection or Brent's method can be used (Beskos et al., 2016). See also Cornebise et al. (2008), Jasra et al. (2011), and Del Moral et al. (2012).

3.3.2 Diagnostics

It is unclear *a priori* whether subset deletion necessitates particle rejuvenation via an invariant kernel. For instance, when $N_g = 1$, the excluded observation may induce minimal posterior shift, and the sampler may proceed without intermediate steps (i.e., $L_k = 1$). In such cases, given that the invariant kernel constitutes the most computationally intensive component, IS may suffice; an option is to invoke the invariant kernel selectively contingent on diagnostics. Since the baseline draws from step $L_k - 1 = 0$ can be used to compute the next step ($L_k = 1$) importance weights, the reliability of the IS estimate can immediately be assessed prior to applying the invariant kernel, using any suitable metric. We use the generalized Pareto $\hat{k}$ diagnostic (e.g., $\hat{k} > 0.7$ as recommended by Vehtari et al., 2024 and Millar, 2018), along with the ESS criterion. If both indicate stability, we proceed with the fast(er) Pareto-smoothed IS (PSIS). Otherwise, the invariant kernel is triggered to rejuvenate the particles; the approach embeds the conventional heuristic of *re-running MCMC* only when necessary (e.g., Bürkner et al., 2020) into the aSMC sampler.

3.4 Choice of estimands

The idea behind the sequential approximation is motivated by the identity in (4). Taking (3) for instance, estimating $\ell^{\text{(LGO)}}$ could be considered as a special case in which the target functions are defined as

$$f_g(\boldsymbol{\Theta}) := p(\boldsymbol{y}_g^* = \boldsymbol{y}_g \,|\, \theta_g, \phi),$$

and the final approximating quantity $\hat{\ell}^{\text{(LGO)}}$ is obtained as a sum of the approximate logarithmic scores $\hat{\ell}_g^{\text{(LGO)}}$. We emphasize, from an algorithmic point of view (the focus of this

paper) that the case-deletion scheme and the selection of an estimand can be treated as distinct and independent operations under the user's control. In light of this, we advocate choosing the estimand to reflect hypothetical data replications and the specific aspects of the out-of-sample generalization that are most relevant (Gelman et al., 2014b).

It is also often preferable to select scoring rules familiar to subject-matter experts, over relying solely on the predictive densities as default measures whose differences or variabilities may be difficult to interpret. For example, in Bayesian econometric time-series applications, brute-force LEO is often used to evaluate Bayesian point forecasts relative to frequentist counterparts using standard error metrics (e.g., Faust and Wright, 2013), or density forecasts via log predictive likelihood (e.g., Koop et al., 2019).

That predictive density is not necessarily the default highlights that out-of-sample model checking need not center on predicting new observations. Beyond predictive densities, other targets of interest may be the shared parameter $f_g(\Theta) := \phi$ and the group-specific parameter $f_g(\Theta) := \theta_g$. Extrapolation to a new group $G+1$ may be considered via leave-g-out-integrable $f_g(\Theta, \boldsymbol{y}^*_{G+1}, \theta_{G+1})$ by operating on the augmented posterior with a new group $G + 1$ which admits the original leave-g-out posterior as its marginal.

Finally, we note the asymptotic equivalence of Bayesian CV and the widely available information criterion (WAIC: Watanabe, 2010), as well as the correspondence of different forms of WAIC to different forms of LOO-CV and LGO-CV estimands (Gelman et al., 2014b; Merkle et al., 2019). With regard to the choice of target functions, we further acknowledge discussions that are generally in favor of the use of marginal likelihoods over conditional likelihoods, due to improved numerical stability and accurate approximation (of WAIC: Li et al., 2015), and the fact that marginal measures align better with the regularity condition for the asymptotic equivalence to hold (Millar, 2018). The target function may then be appropriately selected to target these marginal estimands, provided that marginal likelihoods can be evaluated; a further step may need to be implemented to approximate the integral, such as by quadrature (Merkle et al., 2019, Appendix C).

3.5 Summary and relation to existing approach

Algorithm 1 details the adaptive approach. For clarity, a diagrammatic illustration of the sampler with three CV schemes is provided in Figure 1.

Algorithm 1: `Adaptive SMC for structural Bayesian cross validation`

Input: MCMC draws $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(R)} \sim p(\boldsymbol{\Theta} \,|\, \boldsymbol{y}_{1:G})$
Result: $\hat{\ell}_{1:K}, \hat{\ell}$
for $k = 1, \ldots, K$ *in parallel* **do**
    Initialize index $\ell \leftarrow 0$;
    Initialize case deletion parameter $n_{\ell=0} \leftarrow 0$;
    Initialize particles $(\boldsymbol{\Theta}_0^{(r)}, W_0^{(r)}) \leftarrow (\boldsymbol{\Theta}^{(r)}, 1/R)$;   `// Index k omitted`
    while $n_{\ell-1} < N_k$ **do**
        $\ell \leftarrow \ell + 1$;
        Solve $n_\ell \in (n_{\ell-1}, N_k]$;   `// Section 3.3.1`
        Compute $(W_\ell^{(1)}, \ldots, W_\ell^{(R)})$ from n_ℓ ;   `// Equation 5`
        if $n_\ell < N_k$ **then**
            Deduce $\gamma_{k,\ell}$ from n_ℓ ;   `// Section 3.2`
1   Compute $W_\ell^{(r)} \propto w_\ell^{(r)} = (\gamma_{k,\ell}/\gamma_{k,\ell-1})(\boldsymbol{\Theta}_{\ell-1}^{(r)})$;
2            $A_\ell^{(r)} \sim$ `Resample`$(W_\ell^{(1)}, \ldots, W_\ell^{(R)})$;
3            $\boldsymbol{\Theta}_\ell^{(r)} \sim \mathcal{K}_{k,\ell}(\,\cdot\, \leftarrow \boldsymbol{\Theta}_{\ell-1}^{(A_\ell^{(r)})})$ in parallel ;   `// Invariant kernel (3.1)`
4            $(\boldsymbol{\Theta}_\ell^{(r)}, W_\ell^{(r)}) \leftarrow (\boldsymbol{\Theta}_\ell^{(r)}, 1/R)$;
        else
            $\hat{k}, (\widehat{W}_\ell^{(1)}, \ldots, \widehat{W}_\ell^{(R)}) \leftarrow$ `ParetoSmooth`$(W_\ell^{(1)}, \ldots, W_\ell^{(R)})$;
            if $\hat{k} < 0.7$ **then**
                $(\boldsymbol{\Theta}_\ell^{(r)}, W_\ell^{(r)}) \leftarrow (\boldsymbol{\Theta}_{\ell-1}^{(r)}, \widehat{W}_\ell^{(r)})$;   `// Optional (3.3.2)`
            else
                Rejuvenate $\boldsymbol{\Theta}_\ell^{(r)}$ in parallel ;   `// as in lines 1 to 4`
            end
            $L_k \leftarrow \ell$;   `// `$n_\ell = N_k$`; end loop`
        end
        $\hat{\ell}_{k,\ell} \leftarrow \log \sum_{r=1}^R W_\ell^{(r)} f_k(\boldsymbol{y}_{\mathcal{I}_k}, \boldsymbol{\Theta}_\ell^{(r)})$;   `// Sections (2.2; 3.4)`
    end
    $\hat{\ell}_k \leftarrow \hat{\ell}_{k,L_k}$;
    `// Further continue for LEO`
end
$\hat{\ell} \leftarrow$ `Aggregate`$(\hat{\ell}_{1:K}) \overset{\text{e.g.}}{=} \sum_{k=1}^K \hat{\ell}_k$;
return $\hat{\ell}_{1:K}, \hat{\ell}$

The proposed approach can be viewed as a direct extension of previous works (Gelfand

and Dey, 1994; Peruggia, 1997; Epifani et al., 2008; Bornn et al., 2010; Vehtari et al., 2017; Bürkner et al., 2020) using (PS)IS for approximate LOO-CV. The algorithm complements these works in that we operate on a continuum of distributions which are easier to approximate. The auxiliary intermediate distributions are determined fully automatically to streamline the workflow. The selection of PSIS and MCMC re-runs (in the sense that the MCMC kernel is applied) is guided via the ESS criterion (Kong et al., 1994; Agapiou et al., 2017) and partly the generalized Pareto shape diagnostic (Vehtari et al., 2024) to minimize unnecessary MCMC re-runs where appropriate. When the MCMC kernel is invoked, it serves as a design-efficient alternative to fully re-running MCMC, under reasonable parallel resources (as we demonstrate in Section 4), since it enables particle-wise parallelization without extensive burn-in or full-chain regeneration. We thereby extend the non-adaptive SMC approach of Bornn et al. (2010) for LOO-CV specifically on Bayesian LASSO (Park and Casella, 2008), and then Vehtari et al. (2017) and Bürkner et al. (2020) to subsume the workflow of MCMC re-runs as an efficient systematic component of the sampler. The sampler also supports various CV designs, including LGO (Merkle et al., 2019; Liu and Rue, 2023), LEO (Bürkner et al., 2020), and LSO (e.g., K-fold).

4 Applications

We illustrate the proposed approach using real data examples. Throughout the examples, unless otherwise noted, we obtain 1000 samples from the non-case-deleted posterior via 4000 iterations of the dynamic (Hoffman and Gelman, 2014) Hamiltonian Monte Carlo (HMC), discarding the initial 1000 and applying a thinning factor of 3 upon inspecting empirical autocorrelation. We then use HMC as the aSMC invariant kernel with 1–3 iterations (5 for Gibbs) chosen in consideration of differences in autocorrelation, which is available from the baseline draw. The aSMC sampler uses the 1000 resulting samples as its initial high-quality marginal draw and operates with an ESS ratio threshold of 0.5 (which is typical). We note that, since the MCMC kernel is applied per particle, the run time increases proportionally with the number of kernel applications; it is therefore prudent to set the

number of iterations at a minimal sufficient level for effective low-autocorrelation particle rejuvenation. That said, because rejuvenations are independent particle-wise, distributed computing can substantially reduce run time, and in principle down to that required for a single-particle rejuvenation provided the parallel resources. We employ 8-thread multi-threading unless otherwise noted. Moreover, to further promote efficiency, we advocate repurposing information already available from baseline dynamic HMC posterior draws at no additional cost (e.g., autocorrelations, as above, and covariances for the HMC mass matrix). `Julia` code is available at https://github.com/geonhee619/aSMC-BayesCV.

4.1 Hierarchical example

4.1.1 Radon exposure multilevel regression

This section considers a hierarchical example in which group sizes vary substantially with within-group observation counts N_g ranging from 1 to 116. Following Vehtari et al. (2017), consider the Bayesian multilevel model that describes the measurement of radon in households in Minnesota,

$$y_i \overset{ind}{\sim} \text{normal}(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_{g[i]}, \sigma), \qquad \boldsymbol{\beta}_g \overset{ind}{\sim} \text{MVN}(\boldsymbol{\Gamma} \boldsymbol{u}_g, \boldsymbol{\Sigma}),$$

where y_i represents the measurements of the radon concentration on a logarithmic scale. normal(μ, σ) is the univariate normal distribution with location and scale (μ, σ). The measurement-level predictor $\boldsymbol{x}_i = (1, x_i)^{\mathsf{T}}$ includes an intercept and a binary indicator x_i set to one if the measurement was taken on the first floor and zero if taken in the basement. The group- or county-level predictor $\boldsymbol{u}_g = (1, u_g)^{\mathsf{T}}$ is observed, where u_g is the soil uranium level in county g also on a logarithmic scale. The indices run over $i \in \{1, \ldots, N\}$ and $g \in \{1, \ldots, G\}$, where $N = 919$ denotes the number of observations and $G = 85$ represents the number of counties. For a more complete description of the data and setup, we refer to Price et al. (1996) and Gelman and Hill (2006).

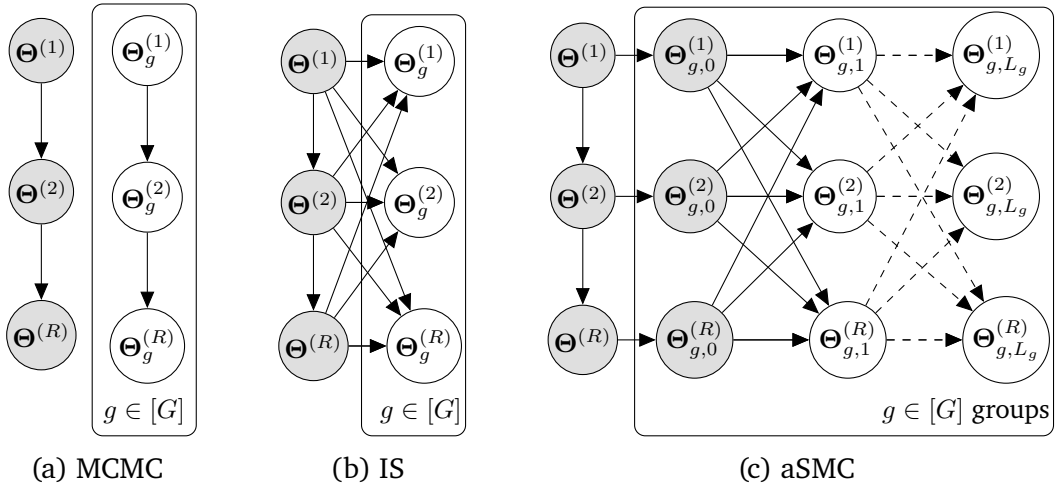Figure 2: Diagrammatic comparison of three strategies for approximate *leave-group-out cross-validation*. The solid arrow ($\rightarrow$) represents the algorithmic input-output relationship. The draws $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(R)} \sim p(\cdot \mid \boldsymbol{y}_{1:G})$ are from the baseline non-case-deleted posterior distribution, MCMC, say. At the post-MCMC stage, we seek draws from the $\boldsymbol{y}_g$-deleted posterior using what is available (gray), to approximate latent quantities not directly accessible (white), for $g \in [G] := \{1, \ldots, G\}$. (a) MCMC constructs Markov chains that leave the G leave-g-out posteriors invariant (with appropriate burn-in and thinning). (b) IS exploits baseline samples as input to produce one-step weighted approximations $p(\boldsymbol{\Theta}_g \mid \boldsymbol{y}_{-g}) \approx \sum_{r=1}^{R} W^{(r)} \delta_{\boldsymbol{\Theta}^{(r)}}(\boldsymbol{\Theta}_g)$. (c) aSMC yields multi-step weighted approximations $\sum_{r=1}^{R} W_{L_g}^{(r)} \delta_{\boldsymbol{\Theta}_{L_g-1}^{(r)}}(\cdot)$. The dashed arrow ($\dashrightarrow$) indicates that $L_g \geq 1$ is determined adaptively and is thus unknown *a priori*.

4.1.2 Leave-group-out cross validation

We consider LGO-CV; see Figures 1a and 2. The parameters in the non-case deleted posterior would be $\boldsymbol{\Theta} = (\boldsymbol{\beta}_{1:G}, \sigma, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ partitioned with $\phi = (\boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \sigma)$ and $\theta_g = \boldsymbol{\beta}_g$ according to the notation in Section 2. The unnormalized importance weights are

$$w_h = \prod_{i:g[i]=h} \text{normal}(y_i \mid \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}_h, \sigma)^{-1}.$$

Although the dataset, with 919 observations and 85 counties, is not excessively large, a single run of MCMC to obtain draws from the non-case-deleted posterior takes approximately 15 minutes. Naïvely extending this to compute the LGO estimands could result in a total run time of up to 21 hours. Reducing inefficiencies lets applied modelers focus their workflow on more meaningful diagnostic checks and model expansion.

Figure 3 first shows the empirical distribution of log predictive likelihood from LGO posteriors, comparing the top and bottom three groups ranked by within-group sample
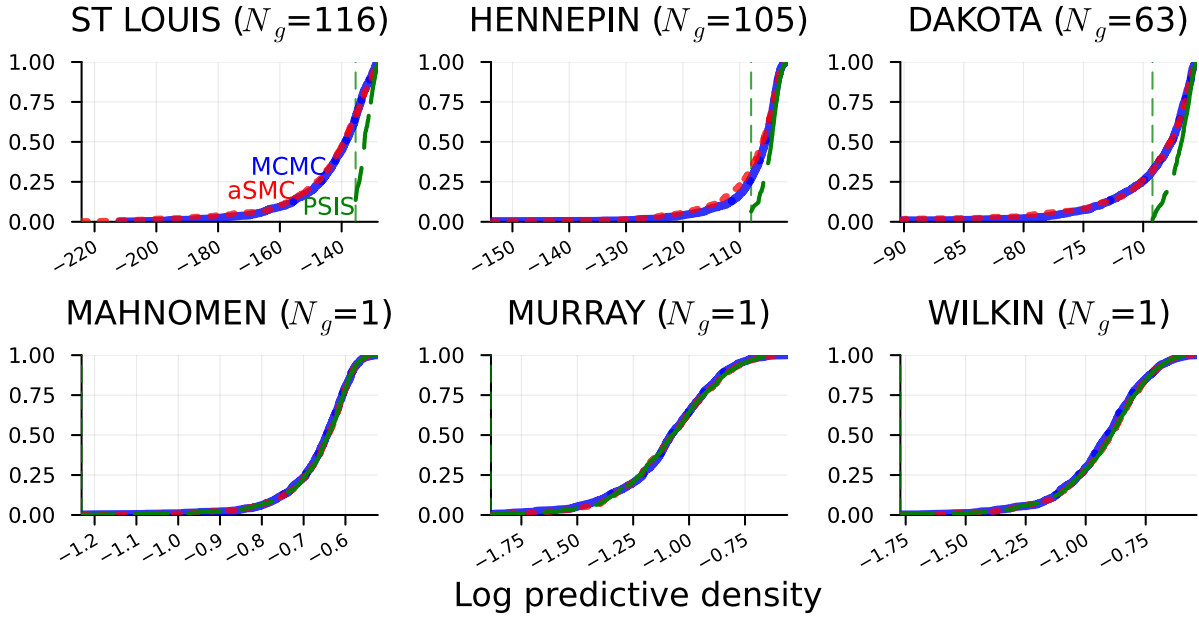
Figure 3: *Hierarchical example*. Comparison of posterior distribution function approximations of LGO log predictive likelihood. **Note**: Displayed in the top row are the three most populous groups in the data; the bottom row shows the three least. The solid line corresponds to the MCMC approximation. The dotted vertical line indicates the point at which the PSIS approximation ends.

size. Note that we are taking the logarithm; this is to facilitate visible comparison. Treating MCMC as reference, aSMC produces approximations highly close to those of MCMC, especially when the within-group observations N_g is high, where IS estimators fail to approximate the tails of the group-deleted posteriors. Results are nearly identical for groups with a single observation, which is expected.

Figure 4 plots the absolute error against within-group observations N_g, using the MCMC-based estimates as reference. On the low end, the two estimators produce practically identical results. As N_g increases, the quality of PSIS estimates degrades, while aSMC maintains a more reliable performance.

Figure 4 also visualizes the paths. More intermediate distributions are typically configured for groups with large within-group sample sizes, and often none for smaller ones. Adaptive bridging therefore streamlines the workflow by automating both the design of the sequential path and the decision of when to invoke the MCMC kernel.

Figure 5 then compares the run time of aSMC versus re-running MCMC. As opera-

Figure 4: **Left**: Error comparison for LGO log predictive likelihood estimates, benchmarked against the brute-force strategy of re-running MCMC for each group (county). **Right**: Realized trajectories of distributions automatically determined by aSMC. **Note**: Annotated numbers denote the determined likelihood-contribution power coefficients. The bisection method is used to perform the root-finding step (see Section 3.3.1). Counties are ordered by the number of within-group observations N_g (in parentheses).

Figure 5: Comparison of run time. Counties are ordered by MCMC run time. The numbers annotated on the aSMC run times are the total number of intermediate distributions excluding the baseline and final LGO distribution. MCMC run time variation arises from HMC adaptation tailored to each LGO posterior geometry, performed during the burn-in phase.

tions can be parallelized across counties for both approaches, we compare the run time per county. aSMC is faster than MCMC for all counties. The run time is negligible for counties without intermediate distributions, in which case the procedure reduces to PSIS, which is fast. In cases with at least one intermediate distribution, run time increases with the number of distributions but remains faster than MCMC re-runs, yielding substantially faster total run time with comparable estimates.

4.2   Time-series example

This section illustrates model validation of a Bayesian state-space model using the backward-sequential LEO scheme (Section 2.2.3).

4.2.1   Yield curve forecasting

Forecasting the term structure of interest rates plays a central role in macroeconomic analysis, as the yield spread has consistently demonstrated predictive ability for future macroeconomic conditions. Estrella and Hardouvelis (1991) documented the yield spread, the difference between the rates of the 10-year Treasury bond and the three-month Treasury bill, as an effective predictor of future macroeconomic variables. Hamilton and Kim (2002) highlighted the predictive capacity of the yield spread for real GDP growth.

Diebold and Li (2006) introduced a time-varying factor representation of the term structure of interest rates as the dynamic Nelson–Siegel (DNS) model. The DNS models the yield for a specific maturity τ as

$$\mu_t(\tau) = \beta_t^{(l)} + \beta_t^{(s)} \left(\frac{1 - \exp(-\lambda_t \tau)}{\lambda_t \tau} \right) + \beta_t^{(c)} \left(\frac{1 - \exp(-\lambda_t \tau)}{\lambda_t \tau} - \exp(-\lambda_t \tau) \right),$$

where factors $\boldsymbol{\beta}_t = (\beta_t^{(l)}, \beta_t^{(s)}, \beta_t^{(c)})^\mathsf{T}$ represent the time-varying level (long-term yields), slope (short- and long-term spread), and curvature (midterm hump). λ_t controls the exponential decay rate of the curve. The factors evolve smoothly as

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t^{(\beta)}, \qquad \boldsymbol{\varepsilon}_t^{(\beta)} \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}^{(\beta)}).$$

Bayesian extensions to the DNS model have since been proposed (e.g., Laurini and Hotta, 2010; Abanto-Valle et al., 2012). We focus on assessing the Bayesian rendition given data of the monthly yield from Japanese government bonds. The dataset spans from September 1999 to January 2024 and includes maturities $\tau \in \mathcal{T} = \{2, 5, 10, 20, 30\}$, which was the longest available time frame with complete data for these maturities at the time of analysis.

We complete the Bayesian model specification first by the measurement equation,

$$\boldsymbol{y}_t = \begin{bmatrix} \mu_t(\tau_1) \\ \vdots \\ \mu_t(\tau_K) \end{bmatrix} + \boldsymbol{\varepsilon}_t^{(y)}, \qquad \boldsymbol{\varepsilon}_t^{(y)} \overset{iid}{\sim} \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{(y)}),$$

where $\boldsymbol{y}_t$ denotes the yields observed across maturities $\mathcal{T}$, with K indicating the number of maturities. $\boldsymbol{\varepsilon}_t^{(y)}$ captures measurement noise. The priors we impose are: the initial state $\boldsymbol{\beta}_0 = (\beta_0^{(l)}, \beta_0^{(s)}, \beta_0^{(c)})^\mathsf{T} \sim \mathrm{MVN}(\boldsymbol{m} = \boldsymbol{0}, \boldsymbol{P}^{-1} = 10\boldsymbol{I}_3)$, noise covariances $\boldsymbol{\Sigma}^{(y)} \sim \mathrm{IW}(\nu_0^{(y)} = 2K, \boldsymbol{S}_0^{(y)} = \boldsymbol{I}_K)$, $\boldsymbol{\Sigma}^{(\beta)} \sim \mathrm{IW}(\nu_0^{(\beta)} = 2(3), \boldsymbol{S}_0^{(\beta)} = \boldsymbol{I}_3)$ (where $\mathrm{IW}(\nu_0, \boldsymbol{S}_0)$ is the inverse Wishart distribution with degrees of freedom ν_0 and scale matrix $\boldsymbol{S}_0$). We set $\lambda_t = 0.0609$ (Diebold and Li, 2006) to simplify estimation, as the main focus of this section is model validation. Under the notation in Section 2, we have $\phi = (\boldsymbol{\Sigma}^{(\beta)}, \boldsymbol{\beta}_{0:T}, \boldsymbol{\Sigma}^{(y)})$, with $G = 1$ (e.g., single country) and index $i = t$; each conditionally independent $y_{(g=1),t} := \boldsymbol{y}_t$ is treated as K-variate.

4.2.2 Backward-sequential LEO

The LEO scheme sequentially deletes the final dependent observation, as illustrated in Figure 1c. The corresponding case-deleted posteriors are deterministically injected as intermediate distributions, and between these, sub-intermediate distributions are further introduced adaptively by the sampler. Continuous case deletions are applied (see Section 3.2.2), as the sub-intermediate models admit fast full conditional Gibbs sampling, executed in parallel across particles via 12-thread multi-threading; see Appendix for further details. To obtain 1000 baseline particles, we ran 12000 iterations of the Gibbs sampler with 2000 burn-in samples and a thinning factor of 10 (in consideration of autocorrelation).

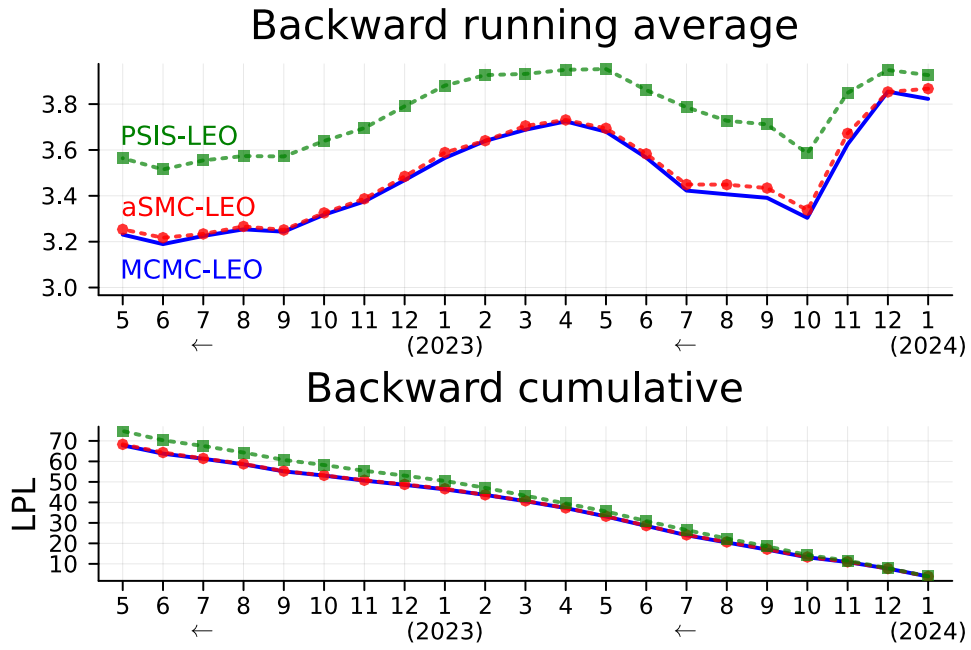## Backward running average

Backward cumulative

Figure 6: *Time-series example*. Comparison of log predictive likelihood (LPL) approximation. Both are computed *backwards* as observations are sequentially deleted backwards. The leftmost value corresponds to the estimate (a) using the least amount of in-sample data points from the model's perspective (as subsequent data points are deleted) and (b) the most amount of data points from perspective of one-step-ahead log predictive likelihood estimation (as the deleted cases are treated as out-of-sample data).

Figure 6 illustrates the cumulative and running-average log predictive likelihoods for one-step-ahead forecasts. The target function is the log predictive likelihood, to facilitate visual comparison. The cumulative likelihoods are computed in reverse, consistent with the backward-sequential LEO scheme; the running average yields a backward estimate of the one-step-ahead log predictive likelihood.

Focusing on the running average, the aSMC sampler closely approximates the estimates obtained via the brute-force MCMC approach. IS estimates degrade with longer horizons and therefore more intensive deletions; the diagnostic measure $\hat{k}$ is almost always above 0.7. aSMC sampler avoids this degradation by rejuvenating the particles where appropriate.

Figure 7 shows where the sampler applied the invariant MCMC kernel. Notably, at some points, no sub-intermediate MCMC kernel interventions were required (e.g., from month 4 to 3 of the year 2023), while for others, multiple sub-intermediate interventions
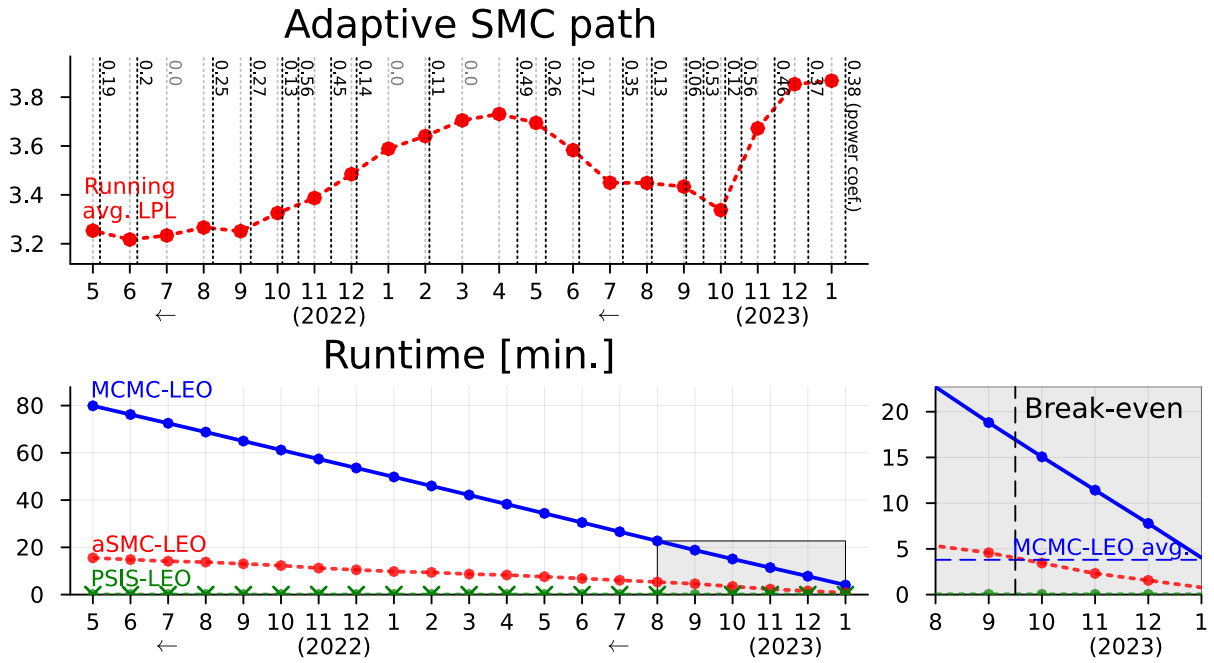
Figure 7: **Top**: Intermediate distributions adaptively determined by the aSMC sampler; the power coefficient on the corresponding pseudo-latest likelihood contribution is indicated. The **(red) dashed line** represents the backward running average LPL for aSMC as in Figure 6. **Bottom**: Cumulative run time comparison, interpreted *backwards*, as the latest observations are sequentially deleted backwards. For PSIS, we denote the point at which it was diagnosed as $\hat{k} > 0.7$ using a cross ($\times$). **Bottom right**: Zoomed-in. The break-even point incates the *furthest* time point from the right at which cumulative aSMC run time remains below the average run time of MCMC.

were applied (e.g., from month 11 to 10 of the year 2022); the latest data point provides different levels of information over time and the sampler adjusts in response to maintain a sound per-period approximation.

Figure 7 also presents the cumulative run time for each method. IS is the fastest, as it involves only re-weighting the samples. This speed comes at the cost of poor approximation quality, particularly for longer case-deletion horizons, as shown in Figure 6. To compare MCMC and aSMC, note that per time-point computations in MCMC are parallelizable, whereas aSMC chains the computation due to backward deletion. The break-even point indicates that aSMC is faster up to four backward-sequential deletions. Since this exceeds one, the discrepancy in cumulative run time continues to widen. This further suggests that applying comparable parallelism to aSMC (in complement with MCMC) has the

26

potential to yield faster run times than MCMC alone, while maintaining approximation quality comparable to MCMC.

4.3 Spatial example

4.3.1 Panel data of retail goods sales

We conclude with a spatial modeling example. The spatial dataset from the M5 competition (Makridakis et al., 2022) contains unit sales at the item level from ten stores located in California (CA), Texas (TX), or Wisconsin (WI). Each item is classified within a unique department, which is further classified under a unique product category. For instance, the item HOBBIES_2_001 belongs to the department HOBBIES_2 and falls under the category HOBBIES. An exhaustive list of department identifiers is: HOUSEHOLD_1, HOUSEHOLD_2, HOBBIES_1, HOBBIES_2, FOOD_1, FOOD_2, and FOOD_3. For a complete description of the data, we refer to https://www.kaggle.com/competitions/m5-forecasting-accuracy.

We work with item-level sales trajectories summarized by their average daily change in ten store locations (CA_1 to CA_4, TX_1 to TX_3, and WI_1 to WI_3). For simplicity, the analysis focuses on items numbered 001 to 030 per department, producing a balanced panel of $K = 209$ items across $S = 10$ store locations. To capture variation in item-level unit sales and their spatial co-movement patterns, we estimate a Bayesian hierarchical model that allows for spatial dependence across observations,

$$\boldsymbol{y}_k \overset{ind}{\sim} \text{MVN}(\boldsymbol{\mu} + \alpha_{g[k]}\mathbf{1}_S, \boldsymbol{\Sigma}), \qquad \boldsymbol{\mu} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}_S), \qquad \alpha_g \overset{iid}{\sim} \text{normal}(0, 1),$$

where $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,S})^\mathsf{T}$ captures spatial variation in unit sales across S stores for item k, $\boldsymbol{\mu}$ denotes store-specific means, $g[k]$ identifies the department associated with item k (e.g., $k = \text{FOOD\_1\_001}$ maps to $g[k] = \text{FOOD\_1}$), $\alpha_{g[k]}$ gives the group (department) mean for item k, and $G = 30$ is the total number of departments. Since the exact store locations are not disclosed, the conditional covariance is estimated by $\boldsymbol{\Sigma} \sim \text{IW}(\nu_0 = 2S, \boldsymbol{S}_0 = \boldsymbol{I}_S)$. Under the notation in Section 2, we may write $\phi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\theta_g = \alpha_g$, with each index k corresponding to $(g[k], i[k])$; each conditionally independent S-variate observation $\boldsymbol{y}_k$ is identified as $y_{g[k],i[k]}$.

4.3.2 Group multi-fold cross validation over spatially dependent units

Several CV designs are possible, and the proposed aSMC framework accommodates each.

(a) **Multifold**: Given conditional independence over k, one may randomly partition item indices and evaluate out-of-sample replication at the unit level.

(b) **LGO**: Alternatively, one may target department-level generalization via LGO, dropping entire groups (departments), such as the example in Section 4.1.

(c) **Group multifold**: For illustration, we consider a different setting. Each item maps to a product department, so we apply a grouped 10-fold CV scheme that ensures coverage of each department across all folds; we reframe the generalization exercise from predicting new generic items or items in new departments to predicting new items from *existing* departments across ten spatially dependent store locations.

We compute the predictive likelihood by leaving out the subsets $\mathcal{I}_j$, where $j \in \{1, \ldots, 10\}$. Each subset $\mathcal{I}_j$ is constructed to contain (approximately) $1/J$ of the item identifiers sampled from each department to ensure a balanced representation. This leads to the unnormalized weights

$$w_j = \prod_{k \in \mathcal{I}_j} \text{MVN}(\boldsymbol{y}_k \,|\, \boldsymbol{\mu} + \alpha_{g[k]} \mathbf{1}_S, \boldsymbol{\Sigma})^{-1}.$$

Given that each subset involves approximately 20 ($S = 10$)-variate deletions, a single-step IS approach is unlikely to yield reliable estimates; the aSMC sampler is therefore applied via the LSO design described in Section 2.2.4 and visualized in Figure 1b. With the baseline MCMC run taking roughly 12.2 minutes to yield 1000 (thinned and post burn-in) samples, total run time for completing MCMC approximations across all folds can extend up to about 2 hours. Although not prohibitively expensive, it seems burdensome for evaluating a single model, considering the iterative nature of applied modeling, where diagnostics often inform model extensions or refinements.

Figure 8 summarizes the results. We measured the discrepancy between the reference log predictive likelihood, obtained by re-running MCMC for each fold, and the approximations produced by aSMC and PSIS, using relative error. The aSMC approach yields a
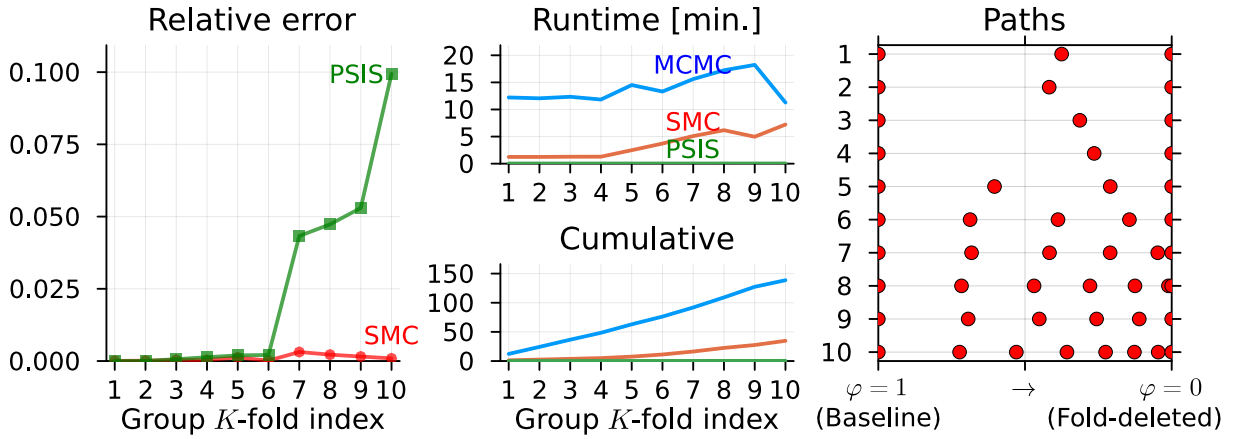
Figure 8: *Spatial example*. Summary of group K-fold results. **Left**: Relative error of the log predictive likelihood approximations. Lower relative error indicates that the estimates were closer to the second-best strategy of re-running MCMC to re-approximate the posterior. **Center**: Fold-wise and cumulative comparison of run times. **Right**: The adaptively determined paths of distribution when running aSMC for each of the folds.

consistently lower relative error. Since all methods allow parallel computation over folds, aSMC offers substantial time savings relative to MCMC while offering near-identical approximation quality. Some folds in the grouped CV setting appear to have contributed to larger posterior shifts, with relative errors occasionally reaching 5–10% under PSIS. For these folds, the automatically constructed intermediate distributions in aSMC helped maintain low error.

5   Summary and discussion

We have introduced an aSMC sampler to (cross) validate Bayesian hierarchical models. The method was motivated by a computational challenge in Bayesian hierarchical setups, particularly when case-deletion schemes are applied to one or more groups (or subsets) involving multiple and/or non-factorizable observations. In such scenarios, conventional IS-based approximations can be unreliable, as only an inadequate fraction of the finite posterior draws from the non-case-deleted distribution lie in the higher-mass regions of the case-deleted posterior. Re-running MCMC becomes the second-best fallback option, which itself is often costly and impracticable with modern complex Bayesian hierarchical

models.

The algorithm was designed to be automatic in the sense that (a) the user need not specify the path of distributions as input, (b) the selection of (PS)IS and MCMC re-runs is automatically determined, and (c) MCMC re-runs targeting the adaptively chosen path bridging the baseline and case-deleted posterior are implemented as an efficient, parallelizable component of the algorithm. Using three real data examples involving LGO CV, group K-fold CV, and LEO validation, we have shown that the sampler can efficiently and automatically approximate diverse CV schemes and facilitate the Bayesian workflow.

Although the sampler was designed to be user-friendly, several parameters must still be specified in advance. For example, the number of times the MCMC kernel is applied currently needs to be set manually by the user. This presents a trade-off; more iterations may yield higher-quality samples due to the asymptotic exactness of invariant kernels, but they also increase run time. We have advocated leveraging parallel resources where available and repurposing information from baseline posterior draws for efficient tuning at no additional cost (e.g., using autocorrelations to determine the minimal sufficient number of MCMC iterations, and covariances to initialize the HMC mass matrix). Whether this strategy is optimal remains unclear, and alternative approaches are worth exploring. Margossian et al. (2024) propose diagnostics for parallel MCMC convergence in a high-chain, low-iteration regime, which is structurally analogous to the SMC setup with many particles and parallelized updates. Incorporating such diagnostics to develop a tuning-free, user-friendly and more trustworthy algorithm is a promising direction for future refinement.

Acknowledgments

Disclosure Statement

There are no competing interests to declare.

SUPPLEMENTARY MATERIAL

`Appendix.pdf`: Brief expositions of (a) Gibbs sampler ("Gibbs sampler for the dynamic Nelson–Siegel (DNS) model") and (b) sensitivity analysis ("Additional sensitivity analysis"). (PDF file)

`DynamicImage.gif`: A dynamically rendered image accompanying the sensitivity analysis in the appendix (subsection "Additional sensitivity analysis"). (GIF file)

`aSMC-BayesCV`: The folder includes files (data, Julia code, and outputs) required to run the methodology and reproduce results presented in the article. Please refer to `README.md` (https://github.com/geonhee619/ASMC-BayesCV) for complete descriptions of setup instruction and execution flow. (folder)

References

Abanto-Valle, C. A., Lachos, V. H., and Ghosh, P. (2012). A Bayesian approach to term structure modeling using heavy-tailed distributions. *Applied Stochastic Models in Business and Industry*, 28(5):430–447. 24

Adin, A., Krainski, E. T., Lenzi, A., Liu, Z., Martínez-Minaya, J., and Rue, H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, 62:100843. 2

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431. 14, 18

Agostinelli, C. and Greco, L. (2013). A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Computational Statistics*, 28(1):319–339. 13

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79. 2

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). On the convergence of adaptive sequential Monte Carlo methods. *Annals of Applied Probability*, 26(2):1111–1146. 15

Bornn, L., Doucet, A., and Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64. 14, 18

Broderick, T., Giordano, R., and Meager, R. (2023). An automatic finite-sample robustness metric: When can dropping a little data make a big difference? *https://arxiv.org/abs/2011.14999*. 3

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for bayesian time series models. *Journal of Statistical Computation and Simulation*, 90(14):2499–2523. 3, 14, 15, 18

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2021). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, 36(2):1243–1261. 3

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32. 12

Chang, J. C., Li, X., Xu, S., Yao, H.-R., Porcino, J., and Chow, C. (2024). Gradient-flow adaptive importance sampling for Bayesian leave one out cross-validation with application to sigmoidal classification models. *https://arxiv.org/abs/2402.08151*. 2

Cornebise, J., Moulines, E., and Olsson, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing*, 18:461–480. 15

Dai, C., Heng, J., Jacob, P. E., and Whiteley, N. (2022). An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association*, 117(539):1587–1600. 11

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436. 3, 11

Del Moral, P., Doucet, A., and Jasra, A. (2012). On adaptive resampling strategies for sequential monte carlo methods. *Bernoulli*, 18(1). 15

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364. 23, 24

Epifani, I., MacEachern, S. N., and Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2. 2, 6, 10, 18

Estrella, A. and Hardouvelis, G. (1991). The term structure as a predictor of real economic activity. *Journal of Finance*, 46(2):555–76. 23

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting,* volume 2 of *Handbook of Economic Forecasting,* pages 2–56. Elsevier. 16

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association,* 70(350):320–328. 2

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association,* 74(365):153–160. 2, 5

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact ealculations. *Journal of the Royal Statistical Society. Series B,* 56(3):501–514. 2, 6, 17

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014a). *Bayesian Data Analysis.* Chapman and Hall/CRC, 3 edition. 4, 6

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. 19

Gelman, A., Hwang, J., and Vehtari, A. (2014b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing,* 24(6):997–1016. 2, 16

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science,* 13(2):163–185. 13

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *https://arxiv.org/abs/2011.01808.* 2

Ghosh, S., Stephenson, W., Nguyen, T. D., Deshpande, S., and Broderick, T. (2020). Approximate cross-validation for structured models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems,* volume 33, pages 8741–8752. 3

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and Eestimation. *Journal of the American Statistical Association*, 102(477):359–378. 6

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B*, 29(1):83–100. 5

Hamilton, J. and Kim, D. H. (2002). A Reexamination of the Predictability of Economic Activity Using the Yield Spread. *Journal of Money, Credit and Banking*, 34(2):340–60. 23

Hoffman, M. D. and Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623. 18

Huang, J. Y., Burt, D. R., Nguyen, T. D., Shen, Y., and Broderick, T. (2024). Approximations to worst-case data dropping: Unmasking failure modes. *https://arxiv.org/abs/2408.09008*. 3

Hukushima, K. and Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608. 13

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22. 3, 15

Kallioinen, N., Paananen, T., Bürkner, P.-C., and Vehtari, A. (2023). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34(1). 13

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential Imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288. 14, 18

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154. 16

Laurini, M. P. and Hotta, L. K. (2010). Bayesian extensions to Diebold-Li term structure model. *International Review of Financial Analysis*, 19(5):342–350. 24

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2015). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, 26(4):881–897. 16

Liu, Z. and Rue, H. (2023). Leave-group-out cross-validation for latent Gaussian models. *https://arxiv.org/abs/2210.04482*. 2, 3, 18

Lobo, V. G., Fonseca, T. C., and Moura, F. A. (2020). Bayesian cross-validation of geostatistical models. *Spatial Statistics*, 35:100394. 2, 3

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364. 27

Margossian, C. C., Hoffman, M. D., Sountsov, P., Riou-Durand, L., Vehtari, A., and Gelman, A. (2024). Nested $\hat{R}$: Assessing the convergence of Markov chain Monte Carlo When running many short chains. *Bayesian Analysis*. 30

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19(6):451–458. 13

Martínez-Minaya, J. and Rue, H. (2024). A flexible Bayesian tool for CoDa mixed models: Logistic-normal distribution with Dirichlet covariance. *Statistics and Computing*, 34(3). 3

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3):802–829. 2, 8, 16, 18

Millar, R. B. (2018). Conditional vs. marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, 28(2):375–385. 2, 15, 16

Neal, R. M. (1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto. 13

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139. 13

Nguyen, T. D., Giordano, R., Meager, R., and Broderick, T. (2024). Sensitivity of MCMC-based analyses to small-data removal. *https://arxiv.org/abs/2408.07240*. 3

Paananen, T., Bürkner, P., Vehtari, A., and Gabry, J. (2024). Leave-one-out cross-validation for non-factorizable models. *https://mc-stan.org/loo/articles/loo2-non-factorizable.html*. 3

Paananen, T., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2). 3

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686. 18

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, 92(437):199–207. 2, 6, 18

Piironen, J. and Vehtari, A. (2016). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735. 2

Price, P. N., Nero, A., and Gelman, A. (1996). Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Health Physics*, 71(6):922–936. 19

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929. 2

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, 60(309):50–62. 5

Silva, L. A. and Zanella, G. (2023). Robust leave-one-out cross-validation for high-dimensional Bayesian models. *Journal of the American Statistical Association*, 119(547):2369–2381. 2, 3, 10

Stone, M. (1976). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–133. 2

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B*, 39(1):44–47. 5

Swendsen, R. and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57:2607–2609. 13

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27:1413–1432. 2, 6, 10, 18, 19

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research*, 17(103):1–38. 6

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228. 2, 5

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58. 10, 15, 18

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594. 16

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer, 2nd edition. 9

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalised linear models and Bayesian forecasting (with discussion). *Journal of the American Statistical Association*, 80(389):73–97. 9

Zhang, A., Daniels, M. J., Li, C., and Bao, L. (2024). Approximate cross-validated mean estimates for Bayesian hierarchical regression models. *Journal of Computational and Graphical Statistics*. 2, 3