Merkatas, Christos; Särkkä, Simo

A Gibbs Sampler for Bayesian Nonparametric State-Space Models

Published in:
2024 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Proceedings

DOI:
10.1109/ICASSP48485.2024.10446518

Published: 18/03/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Merkatas, C., & Särkkä, S. (2024). A Gibbs Sampler for Bayesian Nonparametric State-Space Models. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 - Proceedings* (pp. 13236-13240). (Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10446518

A Gibbs sampler for Bayesian nonparametric state-space models

Christos Merkatas*[1] and Simo Särkkä[1]

[1]Department of Electrical Engineering and Automation, Aalto University, Finland

Abstract

A common assumption in state space models is that the state and observation noise is Gaussian. However, there are cases where this assumption is violated and is chosen for computational convenience. In this article, we present a state space model whose noise processes are modeled via highly flexible density functions based on Bayesian nonparametric priors with decreasing weights. We are focusing on a system identification problem were the aim is to estimate the parameters and the states of the (possibly) nonlinear dynamical system along with its noise processes using Gibbs sampling. Experiments in simulated data show that the nonparametric model outperforms parametric models especially when the distributions of the noise processes depart from Gaussianity.

KEYWORDS: Bayesian nonparametrics, State-space models, Gibbs sampling.

1. Introduction. State-space models (SSMs) [16] are by now one of the most common approaches to model time-dependent data with applications ranging from the analysis of biological signals to system identification and financial application. In the most general form, these models are defined as

$$\begin{aligned} x_t &= g(\theta, x_{t-1}) + q_t, \quad t = 1, 2, \ldots, T \\ y_t &= h(\theta, x_t) + r_t, \quad t = 1, 2, \ldots, T, \end{aligned} \tag{1}$$

where g, h are functions parametrized by θ, the sequence of (x_t) is the hidden state, (y_t) are the observations and (q_t) and (r_t) are i.i.d samples from a distribution which is typically chosen to be Gaussian.

When g, h are linear functions with known parameters, and the noise processes $(q_t), (r_t)$ are Gaussian, the model in Equation (1) is amenable to Kalman filtering/smoothing equations. When the parameters are unknown, multiple models have been proposed for inference in SSMs of the form in Equation (1). In the Bayesian setup [10], the parameters are assigned prior distributions i.e. $\theta \sim \pi(\theta)$ and inference is achieved via Markov chain Monte Carlo (MCMC) methods [3]. In the SSM context, sequential Monte Carlo (SMC) and particle MCMC (PMCMC)

*Thanks to Research Council of Finland for funding.

methods [2; 5] are by now the state-of-the-art methods for inferring jointly the distribution of the parameters, and, the joint distribution of the states.

However, in many applications, the assumption of Gaussian errors can not be easily justified and it is typically used for computational convenience. In order to relax the assumption of Gaussian noise, recent advances in SSMs take a Bayesian nonparametric (BNP) [13] approach to model the noise processes. For example, focusing on linear dynamic systems, in [4] the Dirichlet process [8] is used as a mixing measure to model the densities of the noise processes as Dirichlet process mixtures (DPM) [15]. In [14], the DPM has been used to model the density of the noise process for the observations in nonlinear dynamic systems providing a SMC algorithm for density estimation of the noise process and the distribution of the states.

We propose a Bayesian nonparametric state space model whose noise processes are modeled as mixtures of nonparametric priors with decreasing weights [9; 7; 11]. We introduce a Gibbs sampler for density estimation of the noise processes and the distribution of the unknown parameters θ in the model (1). This forms the basis for a Gibbs kernel that leaves invariant the conditional distribution of the states given the data and the parameters which we sample via a conditional SMC (CSMC) step [2]. We focus on graphical models of state space models but the method can be extended to more general graphical models using SMC techniques for graphs [1]. While the DPM has been previously used to model state and measurement noise in linear and nonlinear models [4; 14], the use of nonparametric priors with decreasing weights has not been previously explored in the literature. Our contention is that decreasing weights are sufficient for offline state and parameter estimation.

The article is organized as follows. In Section 2 we introduce the nonparametric model and the necessary augmentations to devise finite likelihoods. The steps of the proposed Gibbs algorithm are then presented in Section 3. In Section 4 we resort to simulation. We compare the performance of our model against a parametric model whose noise processes are assumed to be Gaussian. Finally, Section 5 concludes the article.

2. The model. In this section we consider SSMs of the type given in Equation (1). For simplicity we assume that the hidden states and the observations are one-dimensional but the formulation can trivially be extended to the multivariate case.

We model the densities of the noise processes as infinite mixtures of Normal kernels with decreasing weights that is

$$\begin{aligned} f_Q(q) &= \int_{\mathbb{R}^+} \mathrm{N}(q \,|\, 0, Q^{-1}) P_Q(\mathrm{d}Q), \\ f_R(r) &= \int_{\mathbb{R}^+} \mathrm{N}(r \,|\, 0, R^{-1}) P_R(\mathrm{d}R) \end{aligned} \tag{2}$$

where Q and R are the precisions i.e. inverse variances of the normal distribution. From now on, in order to avoid notation clutter, we will drop the subscripts from densities and limits of integration when there is no confusion.

The mixing measures P are discrete random probability measures with weights that are

almost surely decreasing and they admit the discrete representation

$$P(\mathrm{d}Q) = \sum_{k=1}^{\infty} w_k^q \delta_{Q_k}, \quad P(\mathrm{d}R) = \sum_{k=1}^{\infty} w_k^r \delta_{R_k}, \tag{3}$$

where the atoms $(Q_k), (R_k)$ are independent collections of i.i.d random variables from some base distributions P_0^Q, P_0^R which we assume admit densities w.r.t. the Lebesque measure p_0^Q, p_0^R, respectively. Recently, it has been shown [11] that a class of nonparametric priors with decreasing weights, that is sufficient for density estimation purposes, can be constructed by taking the weights to be $w_j^\bullet = \mathbb{P}(Z^\bullet \geq j)/\mu^\bullet$, $j = 1, 2, \ldots$ where Z is a discrete random variable and $\mu^\bullet = \mathbb{E}Z^\bullet$. Here, the bullet represents replication for the q and r variables.

Following [11], we introduce auxiliary variables z^q, z^r for a single component of the noise processes which lead to the augmented densities

$$\begin{aligned} f(q, z^q) &= p(z^q \,|\, \phi)/\mu^q \sum_{k=1}^{\infty} \mathbf{1}(k \geq z^q) \mathrm{N}(q \,|\, 0, Q_k^{-1}), \\ f(r, z^r) &= p(z^r \,|\, \phi)/\mu^r \sum_{k=1}^{\infty} \mathbf{1}(k \geq z^r) \mathrm{N}(r \,|\, 0, R_k^{-1}), \end{aligned} \tag{4}$$

where $p(z^\bullet \,|\, \phi)$ is the probability mass function corresponding to $\mathbb{P}(Z^\bullet > j)$ which depends on parameters ϕ. From the additivity of the errors, the transition density of the states and the observation become

$$\begin{aligned} f(x_t \,|\, x_{t-1}, \theta, Q_{1:\infty}) &= \sum_{k=1}^{\infty} w_k^q \, \mathrm{N}(x_t \,|\, g(\theta, x_{t-1}), Q_k^{-1}), \\ f(y_t \,|\, x_t, \theta, R_{1:\infty}) &= \sum_{k=1}^{\infty} w_k^r \, \mathrm{N}(y_t \,|\, h(\theta, x_t), R_k^{-1}). \end{aligned}$$

Then, the conditional likelihood for the model (1) based on a sample $y_{1:T}$ given a trajectory $x_{1:T}$ is given by

$$\begin{aligned} &f(y_{1:T} \,|\, x_{1:T}, \theta, x_0, w_{1:\infty}^q, w_{1:\infty}^r, Q_{1:\infty}, R_{1:\infty}) \\ &\quad = \prod_{t=1}^{T} \left\{ \sum_{k=1}^{\infty} w_k^q \, \mathrm{N}(x_t \,|\, g(\theta, x_{t-1}), Q_k^{-1}) \right\} \\ &\qquad \times \left\{ \sum_{k=1}^{\infty} w_k^r \, \mathrm{N}(y_t \,|\, h(\theta, x_t), R_k^{-1}) \right\}. \end{aligned} \tag{5}$$

For notational convenience we drop the conditioning in the notation and we denote the likehood in Equation (5) as $f(y_{1:T})$.

The infinite mixtures in the likelihood will also appear in the posterior making inference intractable. In order to have a Gibbs sampler with finite number of updates, we augment the posterior with clustering variables that indicate the component of the mixture that each state and observation comes from. In particular, for each state x_t we introduce d_t^q and for each

observation y_t we introduce d_t^r. Then, the likelihood admits a finite representation as

$$
\begin{aligned}
&f(y_{1:T}, d^r_{1:T}, z^r_{1:T}, d^q_{1:T}, z^q_{1:T}) \\
&\quad = \prod_{t=1}^{T} \frac{p(z_t^q)}{\mu^q}\, \mathbf{1}(d_t^q \le z_t^q)\, \mathrm{N}(x_t \,|\, g(\theta, x_{t-1}), Q^{-1}_{d_t^q}) \\
&\qquad \times \frac{p(z_t^r)}{\mu^r}\, \mathbf{1}(d_t^r \le z_t^y)\, \mathrm{N}(y_t \,|\, h(\theta, x_t), R^{-1}_{d_t^r}).
\end{aligned}
$$

3. Posterior inference. Having the augmented likelihood, it is now possible to derive Gibbs samplers with a finite number of steps for posterior inference. We are interested in the estimation of the densities of the noise components, the parameters that control the dynamics θ and finally, the state distributions $[x_t \,|\, y_{1:T}]$ for $1 \le t \le T$ in offline manner. Thus, the sampler has to swipe over the collection of variables

$$
\begin{aligned}
&\theta, (d_t^q, z_t^x),\ (d_t^r, z_t^y),\ x_t, \quad 1 \le t \le T \\
&w_k^q, Q_k, \quad 1 \le k \le z^{q*} = \max_t z_t^q, \\
&w_l^r, R_l, \quad 1 \le l \le z^{r*} = \max_t z_t^r,
\end{aligned}
$$

until the desired number of samples has been reached. Having initialized the variables we sample from the following distributions:

1. For the random measure of the state we construct the weights via $w_k^q = \mathbb{P}(z^q \ge k)/\mu^q$, $1 \le k \le z^{q*}$ while for the random measure of the observations the weights are constructed as $w_l^r = \mathbb{P}(z^r \ge k)/\mu^r$, $1 \le l \le z^{r*}$.

2. We then sample the atoms of the random measures. The full conditional distributions of the precisions are given by

$$
f(Q_k \,|\, \cdots) = p_0^Q(Q_k) \prod_{d_t^q = k} \mathrm{N}(x_t \,|\, g(\theta, x_{t-1}), Q_k^{-1}),
$$

for $1 \le k \le z^{q*}$ and for $1 \le l \le z^{r*}$

$$
f(R_l \,|\, \cdots) = p_0^R(R_l) \prod_{d_t^r = l} \mathrm{N}(x_t \,|\, h(\theta, x_{t-1}), R_l^{-1}).
$$

3. The clustering variables, conditionally on the auxiliary variables for the state are sampled from

$$
\mathrm{pr}(d_t^q = k \,|\, z_t^q, \cdots) \propto \mathrm{N}(x_t \,|\, g(\theta, x_{t-1}), Q_k^{-1}) \mathbf{1}(k \le z_t^q).
$$

For the observations, the discrete distribution is given by

$$
\mathrm{pr}(d_t^r = l \,|\, z_t^r, \cdots) \propto \mathrm{N}(y_t \,|\, h(\theta, x_t), R_l^{-1}) \mathbf{1}(l \le z_t^r).
$$

4. Conditionally on the clustering variables, the auxiliary variables $z^\bullet$ are sampled from

$$
p(Z_t^\bullet = z_t^\bullet \,|\, d_t^\bullet = j, \cdots) \propto p(z_t^\bullet \,|\, \phi) \mathbf{1}(j \le z_t^\bullet), \quad 1 \le t \le T.
$$

5. The parameters θ of the state equation have conditionals given by

$$f(\theta \mid \cdots) \propto f(\theta) \prod_{t=1}^{T} \mathrm{N}(x_t \mid g(\theta, x_{t-1}), Q_{d_t^q}^{-1}),$$

and for the observations

$$f(\theta \mid \cdots) \propto f(\theta) \prod_{t=1}^{T} \mathrm{N}(y_t \mid h(\theta, x_t), R_{d_t^r}^{-1}).$$

When g, h are affine functions, these full conditionals are normal distributions. In this article we consider nonlinear functions g, h thus, the full conditional is of the form

$$f(\theta \mid \cdots) \propto \kappa(\theta) \prod_{t=1}^{T} \zeta_t(\theta),$$

with κ being a density and ζ_t being non-negative invertible functions of θ. We can sample from this full conditional using auxiliary variables as proposed in [6].

6. For state estimation, we are interested in the distribution of $p(x_{1:T} \mid y_{1:T})$. If we denote with

$$\Theta = \{\theta, (d_t^q, z_t^x), (d_t^r, z_t^y), w_k^q, Q_k, w_l^r, R_l\},$$

the collection of variables excluding the states, we can define a Gibbs kernel that leaves invariant the conditional distribution $p(x_{1:T} \mid \Theta, y_{1:T})$. We can sample this conditional distribution using CSMC scheme with multinomial resampling [2; 5, Chapter 16]. The important part to note here is, that for the algorithm to be valid, the particle filter ran in the CSMC must use as proposal density at each time step $t, 1 \leq t \leq T$, the component of the mixture at each time step indicated by the clustering variable d_t^q, that is, $\mathrm{N}(x_t \mid g(\theta, x_{t-1}), Q_{d_t^q}^{-1})$ for all the particles N. A similar argument holds for the distribution of the observations at each time step.

7. Finally, for the estimation of the noise processes, the sampler must sweep over the noise predictive densities $f(q_{T+1} \mid q_1, \ldots, q_T)$ and $f(r_{T+1} \mid r_1, \ldots, r_T)$. These are sampled by first taking random variables $u^q, u^r \sim U(0, 1)$ and the Q_k^*, R_l^* for which

$$\sum_{\ell=1}^{k-1} w_\ell^q < u^q \leq \sum_{\ell=1}^{k} w_\ell^q, \quad \sum_{\ell=1}^{l-1} w_\ell^r < u^r \leq \sum_{\ell=1}^{l} w_r^r. \tag{6}$$

Then a sample from $q_{T+1} \sim \mathrm{N}(\cdot \mid 0, (Q_k^*)^{-1})$ and $r_{T+1} \sim \mathrm{N}(\cdot \mid 0, (R_l^*)^{-1})$ [11].

4. Illustrations. In this section we apply the proposed Gibbs sampler, from now on called NP-PG, in simulated data. The base distributions are taken to be gamma distributions with $p_0^Q = \mathrm{Ga}(Q \mid a_q, b_q)$ and $p_0^R(R) = \mathrm{Ga}(R \mid a_r, b_r)$. For the auxiliary variables z^q, z^r we choose the geometric distribution with probabilities ϕ^q, ϕ^r respectively. We complete the model by assigning independent beta distributions on ϕ's.

In the example below, our goal is to estimate the distribution of the states conditionally on all parameters as well as to perform density estimation for the noise processes and estimate

the parameters of the maps that define the dynamics of the system. We compare our results with those obtained from a particle Gibbs sampler which assumes Gaussian state and observation noise processes with unknown precisions (Gaussian-PG). All algorithms run for $10,000$ iterations with the first $1,000$ samples discared as a burn-in period. For both algorithms we have used $N = 50$ particles, with a backward sampling step to generate the fixed trajectory at each iteration.

4.1. A noisy logistic map. We consider a time series of length $T = 250$ generated from the state space model

$$\begin{aligned} x_t &= g(\theta, x_{t-1}) + q_t, \quad t = 1, \dots, T \\ y_t &= h(\xi, x_t) + r_t, \quad t = 1, \dots, T \end{aligned} \tag{7}$$

where $x_0 = 0.5$, $g(\theta, x) = \sum_{p=0}^{2} \theta_p x^p$ and $h(\xi, x) = \xi\, x$. Here, we choose $\theta := (\theta_0, \theta_1, \theta_2) = (1, 0, -1.38)$ and $\xi = 1$. For this values of θ the dynamics are bounded and converge to an 8-cycle [12]. The noise processes for the states and the observations are given by the mixtures

$$f(q) = \frac{2}{3}\mathrm{N}(q \,|\, 0, 0.005) + \frac{1}{3}\mathrm{N}(q \,|\, 0, 0.01), \tag{8}$$

$$f(r) = \frac{2}{3}\mathrm{N}(r \,|\, 0, 0.05^2) + \frac{1}{3}\mathrm{N}(r \,|\, 0, 0.03^2). \tag{9}$$

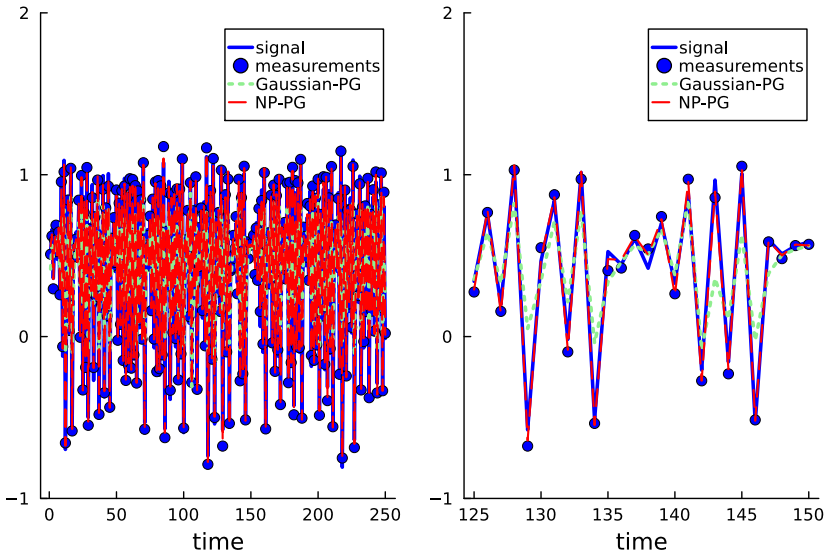

Figure 1: Scatter plot of the observations with true signal superimposed (blue solid line) is shown in left panel. State estimation using the particle Gibbs algorithm for the nonparametric model is superimposed in red-dash line. The estimation of the states when a Gaussian distribution is assumed for the noise is shown in green. In the right panel we plot the same results focusing on time steps $t = 125, \dots, 150$.

In Figure 1 (left panel) we present the time series generated from the random logistic map along with the true state (blue color). The state estimation using NP-PG is superimposed in

red dashed line. The state estimation from Gaussian-PG is shown in green. In the right panel of Figure 1 we zoom in the time steps $t = 125, \ldots, 150$.

In Figure 2 we present the true densities of the state noise process $f(q)$ given in Equation (9) (left panel), and the noise process of the observations $f(r)$ give in Equation (8) (right panel), with the KDEs based on the samples taken from the posterior predictive superimposed in blue color (NP-PG). The curves Gaussian-PG represent the Gaussian distributions with the precisions fixed to the mean of their posterior samples when the state and observations noise processes are assumed to be Gaussian. We note how the introduction of the nonparametric components in the Gibbs algorithm leads to density estimations of the noise processes that are able to capture the high peak of the true densities.

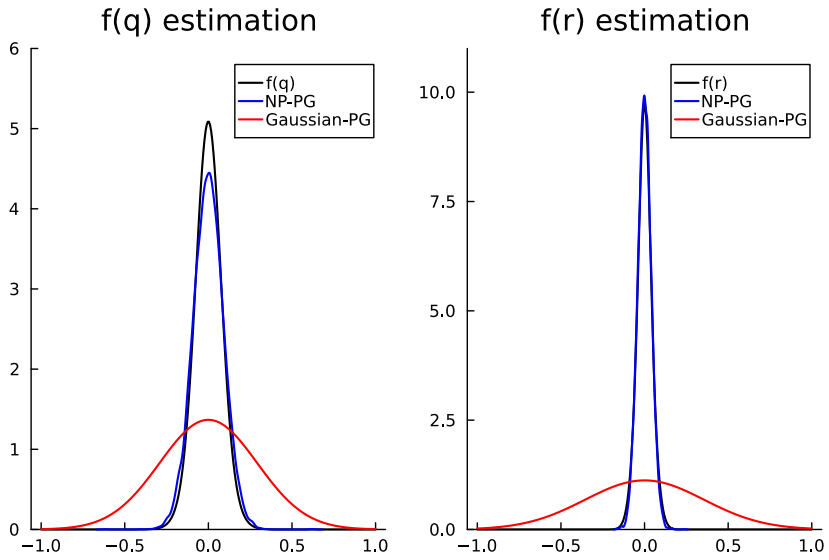

Figure 2: Density estimation for the noise processes.

In Figure 3 we present the ergodic means for the samples of the parameter θ. The overestimation of the variance using Gaussian-PG leads to erroneous estimations for θ. This is a more or less anticipated result as the full conditional for θ is directly influenced by the precision of the Gaussian distribution as is shown in Step 5 of the Gibbs sampler.

5. Conclusion. We have presented a state space model whose noise processes are modeled using Bayesian nonparametric mixtures and we have devised a Gibbs sampler for posterior inference. Simulations in simulated data arising from nonlinear dynamical systems show that decreasing weights are sufficient for inference in SSMs. The proposed method outperforms its parametric counterparts when the real noise distributions depart from normality.

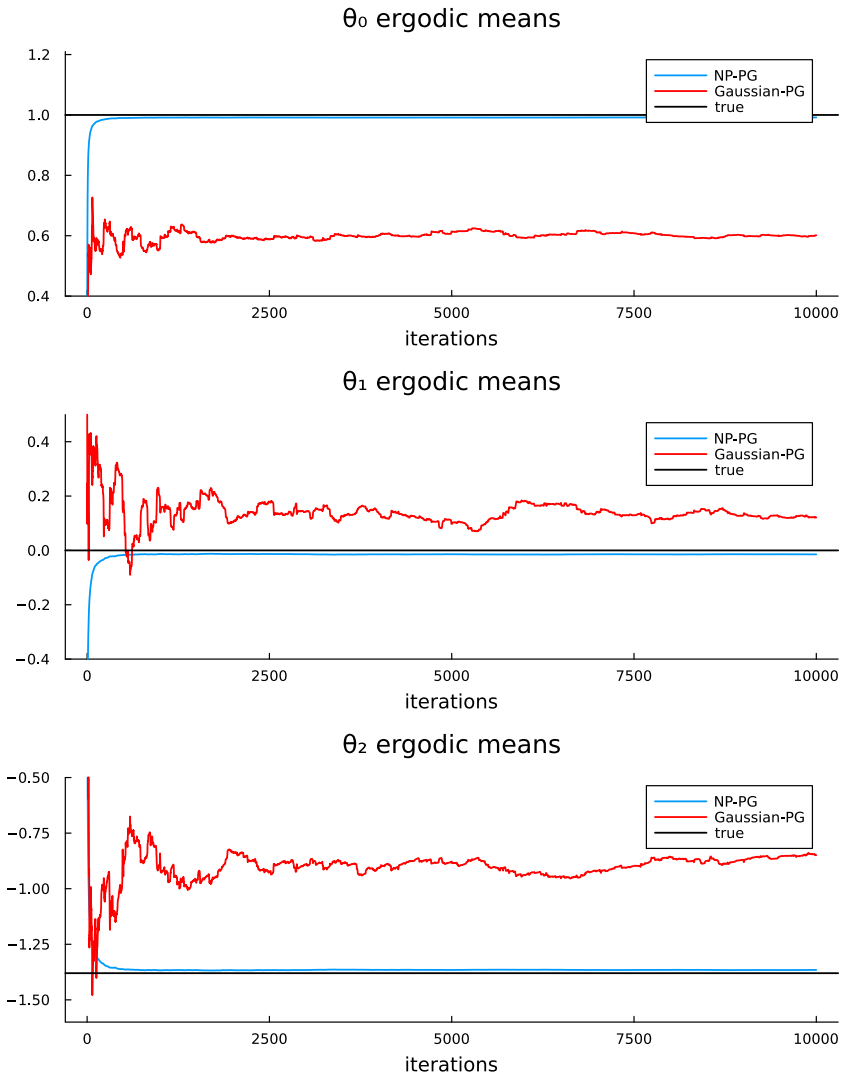

Figure 3: Ergodic means for the $\theta = (\theta_0, \theta_1, \theta_2)$ samples. Black lines represent the true value of θ_p.

References.

[1] Christian Andersson Naesseth, Fredrik Lindsten, and Thomas B Schön. Sequential Monte Carlo for graphical models. *Advances in neural information processing systems*, 27, 2014.

[2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, 2010.

[3] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[4] Francois Caron, Manuel Davy, Arnaud Doucet, Emmanuel Duflos, and Philippe Vanheeghe. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 56(1):71–84, 2007.

[5] Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

[6] Paul Damlen, John Wakefield, and Stephen Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

[7] Pierpaolo De Blasi, Asael Fabian Martínez, Ramsés H Mena, and Igor Prünster. On the inferential implications of decreasing weight structures in mixture models. *Computational Statistics & Data Analysis*, 147:106940, 2020.

[8] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[9] Ruth Fuentes-Garcia, Ramses H Mena, and Stephen G Walker. A new Bayesian nonparametric mixture model. *Communications in Statistics—Simulation and Computation®*, 39(4):669–682, 2010.

[10] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[11] Spyridon J Hatjispyros, Christos Merkatas, and Stephen G Walker. Mixture models with decreasing weights. *Computational Statistics & Data Analysis*, 179:107651, 2023.

[12] Spyridon J Hatjispyros, Theodoros Nicoleris, and Stephen G Walker. A Bayesian nonparametric study of a dynamic nonlinear model. *Computational statistics & data analysis*, 53(12):3948–3956, 2009.

[13] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[14] Nouha Jaoua, Emmanuel Duflos, Philippe Vanheeghe, and François Septier. Bayesian nonparametric state and impulsive measurement noise density estimation in nonlinear dynamic systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5755–5759, 2013.

[15] Albert Y Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, pages 351–357, 1984.

[16] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*, volume 17. Cambridge university press, 2023.