arXiv:2610.20860v2 [stat.ML] 16 Oct 2026

Time-partitioning Gaussian Process Mixture Model

Guilherme Piantino*
Insper Institute of Education and Research, São Paulo

Hedibert F. Lopes
Insper Institute of Education and Research, São Paulo

Abstract

Nonstationarity and heterogeneous dynamics are common in financial and macroeconomic time series. Volatility alternates between calm and turbulent regimes, interest rates follow monetary-policy cycles, and major events trigger persistent structural shifts. We propose the Time-partitioning Gaussian Process Mixture Model (TPGPMM), a Bayesian nonparametric framework for regime identification and nonstationary prediction that infers such regimes and their within-regime dependence from data, without Markovian assumptions. It combines (i) an adaptive Pitman–Yor process prior on ordered compositions, mapped to block lengths with contiguity enforced by construction, and (ii) regime-specific Gaussian processes for temporal dependence, yielding a nonstationary predictive distribution. By inferring its discount parameter, the prior adapts to homogeneous and highly heterogeneous regime lengths.

Unlike sequential or Markov-switching methods, it flags persistent structural shifts as change points rather than reacting to transient shocks. A split–merge Metropolis-within-Gibbs sampler with exact Gibbs updates of regime boundaries jointly learns change points and hyperparameters in batch, quantifying uncertainty over the partition and within-regime dynamics. With an RBF kernel and a fixed, data-informed noise level, TPGPMM produces a sparse, confident segmentation. On simulated data it recovers the true regimes, and on the CBOE Volatility Index and the 10-year US Treasury yield it identifies regimes aligned with documented market and monetary-policy events. For prediction, variants estimating the noise level fit better than change-point baselines and fit and forecast better than the main specification, at the cost of a less sparse or less event-aligned segmentation; with a regime mean, they fit better than a fitted stationary Gaussian process and have lower forecast RMSE except at the shortest horizons.

1 Introduction

Macro-financial time series rarely obey a single stable data-generating process. Volatility moves between calm and turbulent regimes, interest rates track monetary-policy cycles, and both are punctuated by structural breaks around major economic events. Identifying these regimes—and separating persistent shifts from transient shocks—is central to risk management, forecasting, and policy analysis, yet the number, timing, and duration of regimes are typically unknown a priori and must be inferred from the data. Mixture models offer a natural starting point, generating each observation from one of several latent components [McLachlan and Peel, 2000]; but classical finite mixtures require fixing the number of components in advance. Bayesian nonparametric methods relax this by placing a prior on an infinite mixture, so that the effective number of components is inferred from the data. A commonly used prior is the Dirichlet process (DP), introduced by Ferguson [1973], which induces the Dirichlet process mixture (DPM), with posterior computation and density estimation developed by Escobar and West [1995] and Neal [2000]. In the Gaussian case this yields the flexible, complexity-adaptive Dirichlet process Gaussian mixture model (DPGMM) [Görür and Rasmussen, 2010, Rasmussen, 1999].

In time-series applications, partition-based Bayesian nonparametric methods have often been applied in the context of outlier detection, as in Quintana and Iglesias [2003]. This is largely because standard DP-mixture formulations rely on exchangeability, an assumption that is inappropriate for ordered data where temporal dependence is intrinsic. To overcome this limitation, researchers have extended these frameworks to allow the random mixing distribution to evolve with covariates. Key developments include the dependent

*guilhermejlp@al.insper.edu.br

Dirichlet process [MacEachern, 2026], order-based dependent Dirichlet processes [Griffin and Steel, 2006], and kernel stick-breaking processes [Dunson and Park, 2008].

In contrast, much of the time-series regime literature is built on state-space and switching formulations—including Markov-switching autoregressions [Hamilton, 1989] and hidden Markov models [Rabiner, 1989]. Bayesian nonparametric extensions allow an unknown number of regimes while typically retaining a Markovian latent state structure [Fox et al., 2011, Ko et al., 2015]. Partition-based change-point models [Barry and Hartigan, 1993, Chib, 1998, Fearnhead, 2006] and related exchangeable partition probability function (EPPF)–based constructions also often adopt simple or Markovian within-segment models; a notable example is Martínez and Mena [2014], who combine a prior on ordered compositions with discretely observed Ornstein–Uhlenbeck diffusion processes.

Motivated by these limitations, we develop the Time-partitioning Gaussian Process Mixture Model (TPGPMM), a Bayesian nonparametric framework designed to capture multiple regime changes and nonstationary dependence in time-ordered data without imposing a rigid Markovian structure. Standard Gaussian processes (GPs) provide a robust probabilistic framework for nonparametric regression, but the stationarity of their covariance kernels limits their ability to represent regime shifts and abrupt changes. The TPGPMM addresses this limitation by inferring both the number and locations of regimes directly from the data, representing nonstationarity through a flexible piecewise GP structure.

TPGPMM is designed for two purposes. The first is *regime identification*: the posterior over contiguous partitions gives the number and timing of regime changes, with their uncertainty, and separates persistent shifts from transient shocks, which is what risk management and policy analysis require. The second is *nonstationary prediction*: given the partition, the regime-specific GPs form a nonstationary GP, and averaging over the posterior gives a mixture of such GPs whose predictive distribution adapts to the regime structure, both within the sample and beyond its end. The two purposes favour different choices of the observation-noise level. A conservative fixed level yields a sparse segmentation, whereas estimating the noise yields better fit and forecasts, so we evaluate the model for each purpose separately.

This connects our work to the broader literature on nonstationary GPs. Nonstationarity is often introduced via spatially varying kernels, such as process convolution constructions [Higdon et al., 1999] and nonstationary Matérn-type formulations [Paciorek and Schervish, 2003, Katzfuss, 2013], or by modelling input-dependent hyperparameters, including heteroskedastic GP frameworks [Heinonen et al., 2016, Binois et al., 2018]. Another line of work partitions the input space to fit local GPs with distinct hyperparameters, using approaches such as Voronoi tessellations [Kim et al., 2005], treed GPs [Gramacy and Lee, 2008], and local approximate GPs [Gramacy and Apley, 2015], which constructs a separate local GP approximation around each prediction location. A complementary strategy learns nonlinear transformations (warpings) of the input space to induce stationarity in a latent representation, originating with Sampson and Guttorp [1992] and Schmidt and O'Hagan [2003]. Building on this, Damianou and Lawrence [2013] introduced deep Gaussian processes (DGPs), which compose multiple GP layers to learn hierarchical warpings; recent work continues to develop full Bayesian inference for these models [Sauer et al., 2023].

Our work is most closely related to mixtures and segmentations of GPs. Rasmussen and Ghahramani [2001] proposed infinite mixtures of GP experts using DP-based mixing, while Tresp [2000] studied mixtures of GPs with gating functions. However, mixture-of-experts formulations generally allow non-contiguous assignments and are not tailored to contiguous regime segmentation in ordered data. In the change-point setting, Saatçi et al. [2010] and Garnett et al. [2010] introduced sequential online Bayesian methods in which each segment is modelled by a GP and change-points are governed by a hazard function. Compared to these alternatives, TPGPMM differs in two key respects: (i) the partition prior is exchangeable on compositions and position-neutral, so segmentation is driven entirely by the block-size distribution rather than by a Markovian run-length recursion; and (ii) inference is performed in batch, jointly across the entire series, so each candidate boundary is evaluated using global posterior evidence. As we show in Section 7, this combination identifies persistent structural shifts as change points rather than reacting to every transient volatility shock. We evaluate TPGPMM on simulated benchmarks and on two macro-financial series, the Chicago Board Options Exchange (CBOE) Volatility Index (VIX) and the 10-year United States (US) Treasury yield. For regime identification (Section 7), the main specification, with an RBF kernel and a fixed noise level, recovers a sparse segmentation aligned with major market and monetary-policy events. For nonstationary prediction (Section 8), the specifications that estimate the noise level fit better than the change-point baselines and fit and forecast better than the main one, at the cost of a less sparse or less event-aligned segmentation, and

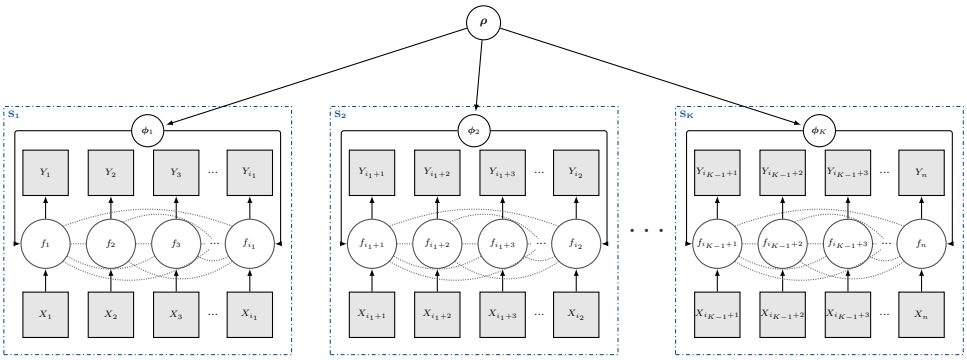


Figure 1: Graphical representation of the TPGPMM. Dashed blue boxes denote conditional independence across regimes, separating the global partition (ρ) from local parameters ($\boldsymbol{\phi}_k$). Within each regime, the gray lines connecting latent nodes (f) illustrate the non-Markovian GP covariance structure.

with a regime mean they fit better than a fitted stationary GP and have lower forecast RMSE except at the shortest horizons.

2 The model

We observe a time-ordered sequence $\mathbf{y} = (Y_1, \ldots, Y_n)$ at fixed, equally spaced inputs $X_t = t - 1$. Dependence among the observations is introduced through an ordered partition of the index set into K contiguous blocks, each of which is interpreted as a distinct regime. We represent the segmentation by its composition:

$$\rho = (n_1, \ldots, n_K), \qquad n_k \in \{m_{\min}, m_{\min} + 1, \ldots\}, \qquad \sum_{k=1}^{K} n_k = n, \tag{1}$$

where $m_{\min}$ is a minimum regime length ($m_{\min} = 3$ throughout). Given ρ, we define the cumulative endpoints

$$i_0 = 0, \qquad i_k = \sum_{j=1}^{k} n_j, \quad k = 1, \ldots, K,$$

so that $0 = i_0 < i_1 < \cdots < i_K = n$. The internal regime boundaries are $\{i_1, \ldots, i_{K-1}\}$, and the corresponding contiguous blocks are

$$S_k = \{i_{k-1} + 1, \ldots, i_k\}, \qquad n_k = |S_k|. \tag{2}$$

Each block S_k is associated with block-specific hyperparameters $\boldsymbol{\phi}_k \in \Phi$. We define the time-indexed hyperparameter process $\{\boldsymbol{\phi}_t\}_{t=1}^{n}$ by

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_k, \qquad t \in S_k,$$

so that $\boldsymbol{\phi}_t$ is piecewise constant over time and can change only at the internal boundaries. Conditional on ρ, we collect the block-specific hyperparameters as $\boldsymbol{\phi}_{1:K} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$ and assume

$$p(\boldsymbol{\phi}_{1:K} \mid \rho) = \prod_{k=1}^{K} p(\boldsymbol{\phi}_k), \tag{3}$$

so that the block hyperparameters are independent and identically distributed given the partition.

For each block k, let $f_k(\cdot)$ denote the latent function with GP prior governed by $\boldsymbol{\phi}_k$, and define the latent evaluations $f_t = f_k(X_t)$, $t \in S_k$, with global latent vector $\mathbf{f} = (f_1, \ldots, f_n)$. Observations are conditionally independent given $\mathbf{f}$, with Y_t depending only on f_t through a Gaussian noise model detailed in Section 4.

Figure 1 shows how dependence flows from the partition ρ, which determines the contiguous segments S_k, to the regime-specific hyperparameters $\boldsymbol{\phi}_k$, which characterise the non-Markovian GP dependence; combined with the inputs X_t, these govern the latent function evaluations f_t that generate the observed sequence $\mathbf{y}$.

3 Prior on the segmentation

The prior distribution on the segmentation ρ is specified via a Gibbs-type prior on random partitions, specifically the Pitman–Yor process (PYP) [Pitman and Yor, 1997], which includes the DP as a special case. Rather than fixing K, this formulation allows the number and location of regimes to be inferred directly from the data. The partition probability is a product of block terms that depend only on the block sizes, times a weight that depends only on the number of blocks; this is the defining form of a Gibbs-type prior [Gnedin and Pitman, 2006, De Blasi et al., 2015], which induces a coherent joint distribution over the number and lengths of regimes consistent with the underlying EPPF. (Because of the weight on K, the prior is not a product partition model in the sense of Barry and Hartigan [1993].)

While Gibbs-type priors typically describe unordered exchangeable partitions, temporal order must be preserved in this setting. We therefore consider compositions of n, which are in one-to-one correspondence with ordered partitions into contiguous blocks. To obtain a prior on ordered segmentations, we map the exchangeable partition induced by the PYP to an ordered composition using the exchangeable random order construction of Pitman [2006]. This prior closely relates to the one used in Martínez and Mena [2014]. By strictly reweighting the combinatorial order, this construction preserves the marginal distribution of K while ensuring position neutrality (i.e., conditional on block sizes, all orders are equally likely). This is advantageous because it avoids encoding time-asymmetry and the prior governs only the number and length of segments, while the data likelihood drives the specific placement of changes. As shown in Appendix A, this results in

$$\Pr(\rho) = \frac{n!}{K! \prod_{k=1}^{K} n_k!} \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1}. \tag{4}$$

Given ρ, the corresponding segmentation is obtained by assigning the blocks deterministically to consecutive intervals, thereby enforcing contiguity by construction.

We require each regime to contain at least $m_{\min}$ observations and therefore condition the prior on this minimum-length requirement:

$$\Pr_{m_{\min}}(\rho \mid \sigma, \theta) = \frac{\Pr(\rho \mid \sigma, \theta)\, \mathbf{1}\{n_k \geq m_{\min}\ \forall k\}}{Z_n(\sigma, \theta)}, \tag{5}$$

Here, $Z_n(\sigma, \theta)$ is the prior probability that every regime meets the requirement, computed using the recursion over K in Appendix A. Because it depends only on (σ, θ), this normalising constant cancels in partition updates when these parameters are held fixed, but must be included when they are updated. Omitting it would implicitly replace the specified hyperprior with one proportional to $p(\sigma)p(\theta)Z_n(\sigma, \theta)$.

We opt to work with the PYP instead of a simpler DP because the DP is the limiting case where $\sigma \to 0$; estimating σ allows the data to determine which kind of prior is used, providing a parsimonious yet adaptable prior for segmentation. While the DP implies logarithmic growth in the number of blocks and favours homogeneous segment lengths, the PYP allows for polynomial growth and induces a heavy-tailed distribution where many short segments can coexist with a few long ones.

4 Gaussian process likelihood within blocks

Having specified the prior on the regime boundaries, we now turn to the within-regime dependence structure. Conditional on the partition and the induced contiguous blocks, we model the observations within each regime using independent GPs. Each block k is associated with a latent function f_k with a GP prior

$$f_k \mid \boldsymbol{\phi}_k \sim \mathcal{GP}\big(0,\, C_{\boldsymbol{\phi}_k}(\cdot,\cdot)\big), \tag{6}$$

and observations within the block follow

$$Y_t \mid f_k(X_t), \sigma_{\mathrm{n}}^2 \sim \mathcal{N}\big(f_k(X_t), \sigma_{\mathrm{n}}^2\big), \qquad t \in S_k, \tag{7}$$

where the observation noise variance σ_{n}^2 is shared across all blocks and $C_{\boldsymbol{\phi}_k}(\cdot,\cdot)$ denotes a covariance (kernel) function. To prevent the model from explaining regime shifts as noise changes, we share the observation-noise variance across regimes. This way, regimes differ only through their GP hyperparameters.

We adopt a squared exponential, or radial basis function (RBF), kernel, which induces smooth latent functions without assuming parametric Markov transition dynamics:

$$C_{\boldsymbol{\phi}_k}(x, x') = s_k^2 \exp\left(-\frac{(x - x')^2}{2\ell_k^2}\right), \tag{8}$$

with block-specific length-scale ℓ_k and signal variance s_k^2, so that $\boldsymbol{\phi}_k = (\ell_k, s_k^2)$. Independent log-normal priors are placed on ℓ_k and s_k^2, with explicit hyperparameters given in Section 6. This flexible kernel allows the GP posterior to capture regime levels nonparametrically, bypassing explicit mean parameters.

For the main specification, we fix the observation-noise standard deviation to help distinguish observation noise from within-regime variation. The GP signal amplitude and the noise level can be difficult to separate from a single time series, so fixing the noise scale reduces ambiguity in the length-scale and amplitude parameters that drive the segmentation. A conservative noise level is intended to absorb short-horizon fluctuations and discourage spurious change-points. The chosen value and its empirical justification are given in Section 6.

To assess the role of the noise assumption and of the within-regime model, we also consider three alternative specifications in which the noise level is estimated rather than fixed. The first keeps the RBF kernel. The second replaces it with a Matérn-3/2 kernel [Stein, 1999], which induces continuous but rougher latent functions. The third adds to the Matérn-3/2 model a mean for each regime, which is integrated out analytically; we refer to it as the regime-mean alternative. These specifications allow us to examine how noise estimation, kernel smoothness and an explicit regime mean affect segmentation and prediction. Their kernels, block covariances and priors are given in Appendix B.

By the conditional-independence assumption (3), the collection $\{(\boldsymbol{\phi}_k, f_k)\}_{k=1}^K$ is independent given $(\rho, \sigma_{\mathrm{n}}^2)$. Integrating out f_k yields the marginal likelihood for each block,

$$L_k(\boldsymbol{\phi}_k) \;\equiv\; p(\mathbf{y}_{S_k} \mid \mathbf{x}_{S_k}, \boldsymbol{\phi}_k, \sigma_{\mathrm{n}}^2) = \mathcal{N}\Big(\mathbf{y}_{S_k}\,;\, \mathbf{0},\, K_k(\boldsymbol{\phi}_k) + \sigma_{\mathrm{n}}^2 I_{|S_k|}\Big), \tag{9}$$

where $K_k(\boldsymbol{\phi}_k) = \big[C_{\boldsymbol{\phi}_k}(X_i, X_j)\big]_{i,j \in S_k}$ is the $|S_k| \times |S_k|$ covariance matrix induced by the kernel on the inputs in S_k. For any positive semi-definite kernel and $\sigma_{\mathrm{n}}^2 > 0$ the matrix $K_k(\boldsymbol{\phi}_k) + \sigma_{\mathrm{n}}^2 I_{|S_k|}$ is positive definite, ensuring the likelihood is well-defined. Because the inputs are equally spaced and the kernel is stationary, it is also a symmetric Toeplitz matrix, which the sampler exploits (Section 5.1). Assuming conditional independence across blocks given the inputs $\mathbf{x}$, the partition ρ, the block-specific hyperparameters, and the shared noise variance, the full likelihood factorises as

$$p\Big(\mathbf{y} \mid \mathbf{x}, \rho, \boldsymbol{\phi}_{1:K}, \sigma_{\mathrm{n}}^2\Big) = \prod_{k=1}^{K} L_k(\boldsymbol{\phi}_k). \tag{10}$$

5 Posterior Computation

Combining the segmentation prior (5) with the block-wise GP likelihood (10), and with the hyperpriors $p(\boldsymbol{\phi}_k)$, $p(\sigma)$, $p(\theta)$ specified in Section 6, the joint posterior is

$$\pi(\rho, \boldsymbol{\phi}_{1:K}, \sigma, \theta \mid \mathbf{y}) \;\propto\; \mathrm{Pr}_{m_{\min}}(\rho \mid \sigma, \theta)\, p(\sigma)\, p(\theta) \prod_{k=1}^{K} \big[\, p(\boldsymbol{\phi}_k)\, L_k(\boldsymbol{\phi}_k)\big], \tag{11}$$

with L_k defined in (9). When σ_{n} is estimated, it joins the parameters on the left and its prior $p(\sigma_{\mathrm{n}})$ multiplies the right-hand side. Since the marginal posterior of ρ is analytically intractable, we perform inference jointly on $(\rho, \boldsymbol{\phi}_{1:K}, \sigma, \theta)$ via Markov chain Monte Carlo (MCMC) and obtain marginals for ρ by Monte Carlo.

5.1 Inference Algorithm

Standard MCMC moves that adjust individual change points at random often exhibit poor mixing and high computational cost. We therefore adopt a split–merge strategy that proposes global modifications to the

partition structure in a single step. This approach parallels split–merge samplers for mixture models [Jain and Neal, 2004] and their extensions, enabling the chain to escape local modes more effectively. Because the number of blocks K is itself a random variable, structural moves are implemented within a reversible-jump MCMC (RJMCMC) framework [Green, 1995, Richardson and Green, 1997]: a split draws one auxiliary variable per hyperparameter and the merge is its exact inverse, so the move is dimension matched.

The Toeplitz structure makes two further moves cheap. The Cholesky factor of a leading sub-matrix of a covariance is the leading block of its Cholesky factor, and a symmetric Toeplitz matrix is unchanged by reversing the order of its rows and columns. One factorisation of a block's covariance therefore gives the log-likelihood of every prefix and every suffix of the block. We use this to draw each boundary exactly from its full conditional, and to propose split points in proportion to the target given the parent's hyperparameters, an informed proposal in the sense of Zanella [2020].

Our sampler targets (11) using a split–merge Metropolis-within-Gibbs scheme. Each iteration updates the partition by a split or merge, draws two regime boundaries from their full conditionals, updates the kernel hyperparameters of three random blocks by random-walk Metropolis, and updates the Pitman–Yor parameters σ and θ (and the noise level, when it is estimated). Because latent functions are integrated out, likelihood evaluations are block-wise and structural proposals recompute only the affected blocks. Proposal details and acceptance ratios are given in Appendix C, their derivation in Appendix D, and the procedure is summarised in Algorithm 1.

Algorithm 1 TPGPMM split–merge Metropolis-within-Gibbs sampler

Require: Data $(\mathbf{X}, \mathbf{y})$, hyperpriors, total iterations M, burn-in B, thinning T

Initialise $\rho^{(0)}$, $\boldsymbol{\phi}^{(0)}_{1:K}$, $\sigma^{(0)}$, $\theta^{(0)}$
for $m = 1, \dots, M$ **do**
  (A) Propose a split (informed split point) or a merge and accept/reject by the RJMCMC ratio in Appendix C, recomputing L_k only for the affected blocks
  (A′) Draw two randomly chosen boundaries from their full conditionals
  (B) For three blocks k chosen at random, update $\log \ell_k$ and $\log s_k^2$ by random-walk Metropolis
  (C) Update σ on the logit scale and θ on the log scale (including Z_n); if σ_{n} is estimated, update $\log \sigma_{\mathrm{n}}$
  (D) If $m \bmod 200 = 0$, refresh all L_k from scratch
  if $m > B$ **and** $(m - B) \bmod T = 0$ **then**
    Store $(\rho^{(m)}, \boldsymbol{\phi}^{(m)}_{1:K}, \sigma^{(m)}, \theta^{(m)})$
  end if
end for

5.2 Posterior Summaries

Let $\{(\rho^{(m)}, \boldsymbol{\phi}^{(m)}_{1:K^{(m)}}, \sigma^{(m)}, \theta^{(m)})\}_{m=1}^{M^\star}$ denote the retained post-burn-in, post-thinning samples. Several non-predictive summaries follow directly from these draws.

We estimate the posterior probability of a regime boundary at index t by

$$\widehat{\tau}_t = \frac{1}{M^\star} \sum_{m=1}^{M^\star} \mathbf{1}\{t \text{ is an internal boundary of } \rho^{(m)}\}. \tag{12}$$

Because posterior mass for one boundary is often spread over neighbouring days, we also report the probability of at least one boundary within ± 10 trading days of t.

To summarise which observations are consistently assigned to the same regime, we construct the posterior similarity matrix $\widehat{S} \in [0,1]^{n \times n}$ with entries $\widehat{S}_{ij} = \frac{1}{M^\star} \sum_m \mathbf{1}\{i \sim_{\rho^{(m)}} j\}$, where $i \sim_\rho j$ denotes that observations i and j belong to the same block under ρ.

We also summarise the marginal posteriors of the Pitman–Yor parameters σ and θ using their means, medians and equal-tailed 95% credible intervals computed from the retained draws. The posterior of K is obtained as the empirical distribution of $\{|\rho^{(m)}|\}_{m=1}^{M^\star}$.

The construction of TPGPMM's in-sample posterior predictive density $p(Y_t \mid \mathbf{y})$ and of its forecasts is given in Section 8.1.

6 Experimental setup

We evaluate TPGPMM for its two purposes. Section 7 assesses regime identification, on simulated data with known change-points and on two macro-financial series whose regimes we compare with documented events. Section 8 assesses nonstationary prediction, within the sample and beyond its end, on the same two series. The data, baselines, model specifications and computational settings are described below. Appendices E and F contain implementation and metric details.

The first empirical series is the log-transformed VIX, $y_t = \log(\mathrm{VIX}_t)$, which we use as a proxy for financial stress. The daily observations span January 2015 to December 2023 (n = 2,264). This timeframe spans low-volatility phases, sudden spikes, and prolonged crises, providing a strong testbed for structural-change detection. We also analyse the daily 10-year US Treasury constant-maturity yield (Federal Reserve Economic Data, FRED, series DGS10) over the same period (n = 2,251), modelling the standardised yield level (centred and divided by its sample standard deviation). The Treasury yield provides a complementary setting: a trending interest-rate level shaped by monetary-policy cycles, alongside the mean-reverting, stress-sensitive VIX. Both are noisy financial time series. On this scale the fixed noise level $\sigma_{\mathrm{n}} = 0.1$ corresponds to about 9 basis points of yield. For forecasting, both series are extended with the trading days of 2024, which are held out from fitting.

For regime identification, we compare TPGPMM with Bayesian online change-point detection (BOCPD) with a Normal–Inverse-Gamma run-length model [Adams and MacKay, 2007], the sticky hierarchical Dirichlet process switching linear dynamical system (HDP-SLDS) [Fox et al., 2011], and the Martinez–Mena method [Martínez and Mena, 2014]. We fit Martinez–Mena (MM) under both specifications and with the settings of the original paper (Appendix E.3): Model B (MM-B), with the autocorrelation parameter held fixed at zero, treats observations as independent and identically distributed given each block's parameters; Model A (MM-A) allows Markovian autoregressive dependence by assigning the autocorrelation parameter a $\mathrm{U}(0, 1)$ prior. For nonstationary prediction, we compare with single stationary GPs with RBF and Matérn-3/2 kernels, fitting their hyperparameters by maximising the marginal likelihood. The fit comparisons additionally include the local approximate GP (laGP) of Gramacy and Apley [2015] and the posterior predictive distributions of HDP-SLDS and Martinez–Mena.

Throughout the experiments, TPGPMM denotes the main specification introduced in Section 4, with an RBF kernel and fixed $\sigma_{\mathrm{n}} = 0.1$. The alternatives estimate σ_{n}: TPGPMM (σ_{n} estimated) keeps the RBF kernel, TPGPMM (Matérn-3/2, σ_{n} estimated) changes the kernel, and TPGPMM (Matérn-3/2, regime mean, σ_{n} estimated), the regime-mean alternative, adds an integrated regime mean to the Matérn-3/2 model.

The fixed noise level is informed by first- and second-difference estimates of the short-horizon noise level [Rice, 1984, Gasser et al., 1986]. These estimators are robust to smooth within-regime trends and give standard deviations of 0.056 and 0.048 for log-VIX, and 0.041 and 0.034 for the standardised Treasury yield. The chosen value of 0.1 is of the same order of magnitude as these estimates, but deliberately conservative: about twice the estimates for log-VIX and two to three times those for the yield. We use the main specification for regime identification (Section 7) and compare all four specifications for nonstationary prediction (Section 8).

For the kernel hyperparameters, we assign independent log-normal priors: $\log \ell_k \sim \mathcal{N}(\log 10,\, 0.6^2)$ and $\log s_k^2 \sim \mathcal{N}(\log 1,\, 0.8^2)$; for the regime-mean alternative, $\log s_k^2 \sim \mathcal{N}(\log \widehat{\mathrm{var}}(\mathbf{y}), 1)$, and the regime mean has prior mean $\bar{y}$ and prior variance $\widehat{\mathrm{var}}(\mathbf{y})$. The partition follows the Pitman–Yor process in (5) with $\sigma \sim \mathrm{Beta}(1, 1)$ and $\theta \sim \mathrm{Gamma}(5,\, 0.5)$ (prior mean 10). When the observation noise is estimated, $\log \sigma_{\mathrm{n}} \sim \mathcal{N}(\log 0.1, 1)$.

Posterior inference uses four chains for each empirical series, each with its own random stream. The chains start from $K_0 \in \{1, 5, 15, 30\}$ equal blocks and run for 60,000 iterations on VIX and 20,000 on the Treasury yield. We discard the first 50% as burn-in and retain every fifth draw. Each simulated dataset is fitted with one chain of 15,000 iterations. Posterior summaries are computed using $M^\star = 800$ retained draws and forecasts using 500. The convergence analysis of the sampler (acceptance rates, effective sample sizes (ESS), and rank-normalised $\widehat{R}$ for K, σ, θ) is given in Appendix H.

The computational cost of inference benefits from conditional independence across blocks, which reduces the cost of GP likelihood evaluation from $\mathcal{O}(n^3)$ to $\mathcal{O}(\sum n_k^3)$. We use Cholesky factorization with 10^{-8} diagonal jitter. Each proposal recomputes the likelihoods of the affected blocks, and a full refresh every 200 iterations limits numerical drift. Boundary updates reuse the Cholesky factorisations to evaluate the prefix and suffix likelihoods for all candidate positions. On one core of an Apple Silicon machine, a 60,000-iteration VIX chain (n = 2,264) takes about 8 minutes for TPGPMM and up to 50 minutes for the alternatives

(Table 11); the four chains run in parallel.

7 Regime identification

This section evaluates the segmentation of TPGPMM, the specification designed for it, and compares it with the change-point baselines. The estimated-noise alternatives are discussed where they segment differently.

7.1 Simulation Study

We evaluate change-point detection and the adaptation of the Pitman–Yor discount parameter σ on two synthetic scenarios at $n = 900$ with true number of regimes $K = 10$. The homogeneous scenario splits the series into ten segments of equal length 90; the power-law scenario consists of one dominant block of length 810 plus nine shorter blocks of length 10. The scenarios share the same specification and differ only in the segment-size distribution; the exact data-generating configuration is given in Appendix G. Each scenario is run over 50 replications. For every replication and every method, we record segmentation-accuracy metrics $\hat{K}$, precision, recall, F_1, and change-point mean absolute error (CP_MAE) at a 5-step tolerance (all defined in Appendix F).

Table 1: Simulation-study change-point recovery summary across 50 replications per scenario, for TPGPMM, its three alternatives with estimated noise and the baselines. SD denotes standard deviation.

Scenario	Method	$\hat{K}$ (mean ± SD)	Precision	Recall	F1	CP_MAE
homogeneous	**TPGPMM**	10.3 ± 0.7	0.968	0.998	0.981	0.21 ± 1.41
homogeneous	TPGPMM (σ_n estimated)	10.3 ± 0.6	0.973	0.998	0.984	0.20 ± 1.41
homogeneous	TPGPMM (Matérn-3/2, σ_n estimated)	10.0 ± 0.2	0.996	0.996	**0.996**	0.40 ± 1.98
homogeneous	TPGPMM (Matérn-3/2, regime mean, σ_n estimated)	10.0 ± 0.2	0.996	0.996	**0.996**	0.40 ± 1.98
homogeneous	BOCPD	11.1 ± 1.3	0.844	0.938	0.885	4.98 ± 5.56
homogeneous	HDP-SLDS	15.5 ± 4.7	0.563	0.833	0.658	8.56 ± 8.33
homogeneous	MM-A	9.8 ± 0.5	0.998	0.973	0.985	3.38 ± 4.73
homogeneous	MM-B	18.9 ± 2.6	0.493	0.964	0.650	1.84 ± 2.01
power-law	**TPGPMM**	10.0 ± 0.3	0.996	0.996	0.996	0.05 ± 0.22
power-law	TPGPMM (σ_n estimated)	10.0 ± 0.2	0.996	0.998	**0.997**	0.03 ± 0.16
power-law	TPGPMM (Matérn-3/2, σ_n estimated)	9.9 ± 0.3	0.998	0.989	0.993	0.11 ± 0.34
power-law	TPGPMM (Matérn-3/2, regime mean, σ_n estimated)	9.9 ± 0.3	0.998	0.989	0.993	0.11 ± 0.34
power-law	BOCPD	9.4 ± 1.1	0.988	0.918	0.947	2.48 ± 1.18
power-law	HDP-SLDS	9.6 ± 4.2	0.614	0.567	0.554	18.92 ± 64.19
power-law	MM-A	9.8 ± 0.4	1.000	0.976	0.987	1.22 ± 0.42
power-law	MM-B	14.0 ± 1.8	0.675	0.958	0.788	1.39 ± 0.67

Table 1 reports TPGPMM, its three alternatives with estimated noise and the baselines. In the power-law scenario TPGPMM achieves an F_1 of 0.996 and a CP_MAE of 0.05, recovering the true $K = 10$ in 92% of the replications ($\hat{K} = 10.0 \pm 0.3$). In the homogeneous scenario it over-segments slightly ($\hat{K} = 10.3 \pm 0.7$, exact K in 72% of the replications), so its precision is lower and its F_1 (0.981) is slightly below that of MM-A (0.985), although its CP_MAE is lower than that of every baseline.

The alternatives show what estimating the noise changes when the model is well specified. All three recover the noise level (posterior mean σ_n between 0.095 and 0.100, truth 0.1). With the RBF kernel, which is the data-generating kernel, estimating the noise gives results close to TPGPMM's (F_1 of 0.984 and 0.997), including the slight over-segmentation of the homogeneous scenario ($\hat{K} = 10.3$). The two Matérn-3/2 alternatives are the most accurate in that scenario ($F_1 = 0.996$, exact K in 96% of the replications), even though their kernel differs from the data-generating one. They find the same modal partition in 98 of the 100 datasets, so the regime mean makes no difference to segmentation here. On simulated data the noise is therefore identifiable; on the real series it is not, and estimating it with the RBF kernel fragments the segmentation (Section 7.2).

Among the baselines, MM-A is the closest, with $\hat{K}$ tightly concentrated around the truth (9.8 ± 0.5 and 9.8 ± 0.4) and F_1 of 0.985 and 0.987. It rarely adds a spurious boundary but misses some and localises them less precisely (CP_MAE 3.38 and 1.22). BOCPD performs reasonably on power-law ($F_1 = 0.947$)

but over-segments on homogeneous ($F_1 = 0.885$, $\hat{K} = 11.1 \pm 1.3$). MM-B and HDP-SLDS over-segment substantially in both scenarios, with lower precision and inflated CP_MAE.

The posterior of the Pitman–Yor discount σ adapts to the underlying segment-size distribution (Figure 2). In the homogeneous scenario, where the true partition is uniform, the posterior mean is small ($\bar{\sigma} = 0.121$ on average), close to the DP limit. In the power-law scenario, where one dominant block coexists with many short ones, it moves well away from zero ($\bar{\sigma} = 0.526$), consistent with a heavier-tailed Pitman–Yor regime.

Figure 2 compares the posterior mean of σ across the 50 replicate seeds of the simulation study, separately for the two scenarios and for all four specifications. For every specification the two distributions are completely disjoint: across specifications, every homogeneous-scenario seed produces $\overline{\sigma} \in [0.104, 0.148]$ while every power-law seed produces $\overline{\sigma} \in [0.516, 0.540]$, and the four specifications give almost the same values. This per-seed separation provides direct evidence that the posterior on σ adapts to the segment-size distribution of the data—staying near the Dirichlet-process limit when blocks are uniform and moving distinctly above zero when block sizes are heavy-tailed—rather than being pinned by the prior. Because it does not depend on the kernel, the noise model or the regime mean, the adaptation is a property of the partition prior. Including the truncation constant Z_n in (5) matters here: without it the effective prior on σ has mean 0.04 at $n = 900$, which pulls both scenarios towards the Dirichlet-process limit.

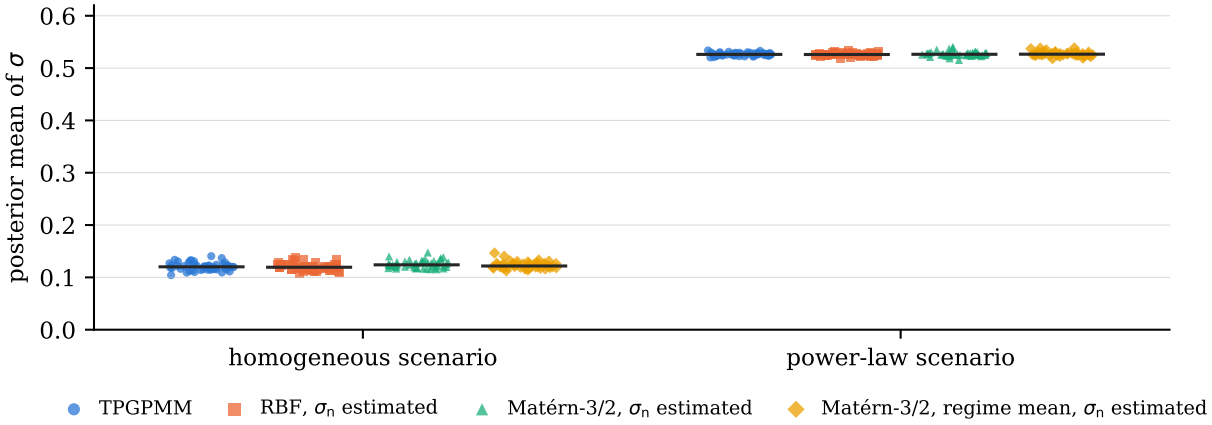


Figure 2: Distribution of the posterior mean of σ across 50 replicate seeds in the homogeneous and power-law scenarios, for TPGPMM and its three alternatives with estimated noise (one colour and marker per specification). Each point is the posterior mean for one seed, jittered within each specification; horizontal lines mark the medians.

These results highlight the role of two key ingredients in the TPGPMM architecture: a global Pitman–Yor-induced prior on contiguous compositions, which concentrates posterior mass on well-separated boundaries, and a within-block likelihood flexible enough to absorb regime-internal variation without creating spurious change points. In the synthetic setting, MM-A's Ornstein–Uhlenbeck process is expressive enough to generate good performance, whereas BOCPD and HDP-SLDS rely on local Markovian or run-length dynamics, and MM-B retains the partition prior but removes within-block dependence.

7.2 VIX

We benchmark the posterior segmentation of log-VIX against six major market events, listed in Table 2, and use them as economic references to interpret the inferred regimes rather than as model constraints.

Figure 3 separates the five methods along two axes: the confidence of the change-point detections and the sparsity of the inferred segmentation. TPGPMM produces few, confident detections: its change-point probability exceeds 0.5 at 7 locations and 0.2 at 13. MM-A is as sparse (7 peaks above 0.5 and 9.8 regimes on average), but most of its boundaries fall between November 2016 and February 2018: it isolates the short spikes of August 2017 and February 2018 and places no boundary at the China sell-off (0.00) or the COVID-19 panic (0.00). MM-B has only 3 peaks above 0.5 but spreads moderate mass over many locations (14 above

Table 2: Six benchmark market events used to interpret the posterior VIX segmentation.

Event (Date)	Market Impact
China "Black Monday" (2015-08-24)	China slowdown triggered U.S. liquidity strains.
Brexit Vote (2016-06-24)	Surprise "Leave" vote sparked global equity sell-offs.
"Volmageddon" (2018-02-05)	VIX spike driven by short-volatility deleveraging.
COVID Panic (2020-03-16)	Pandemic lockdowns caused acute funding strains.
Russia–Ukraine (2022-02-24)	Invasion led to risk sell-offs and commodity surges.
US Bank Turmoil (2023-03-10)	Silicon Valley Bank failure sparked bank solvency and rate volatility.

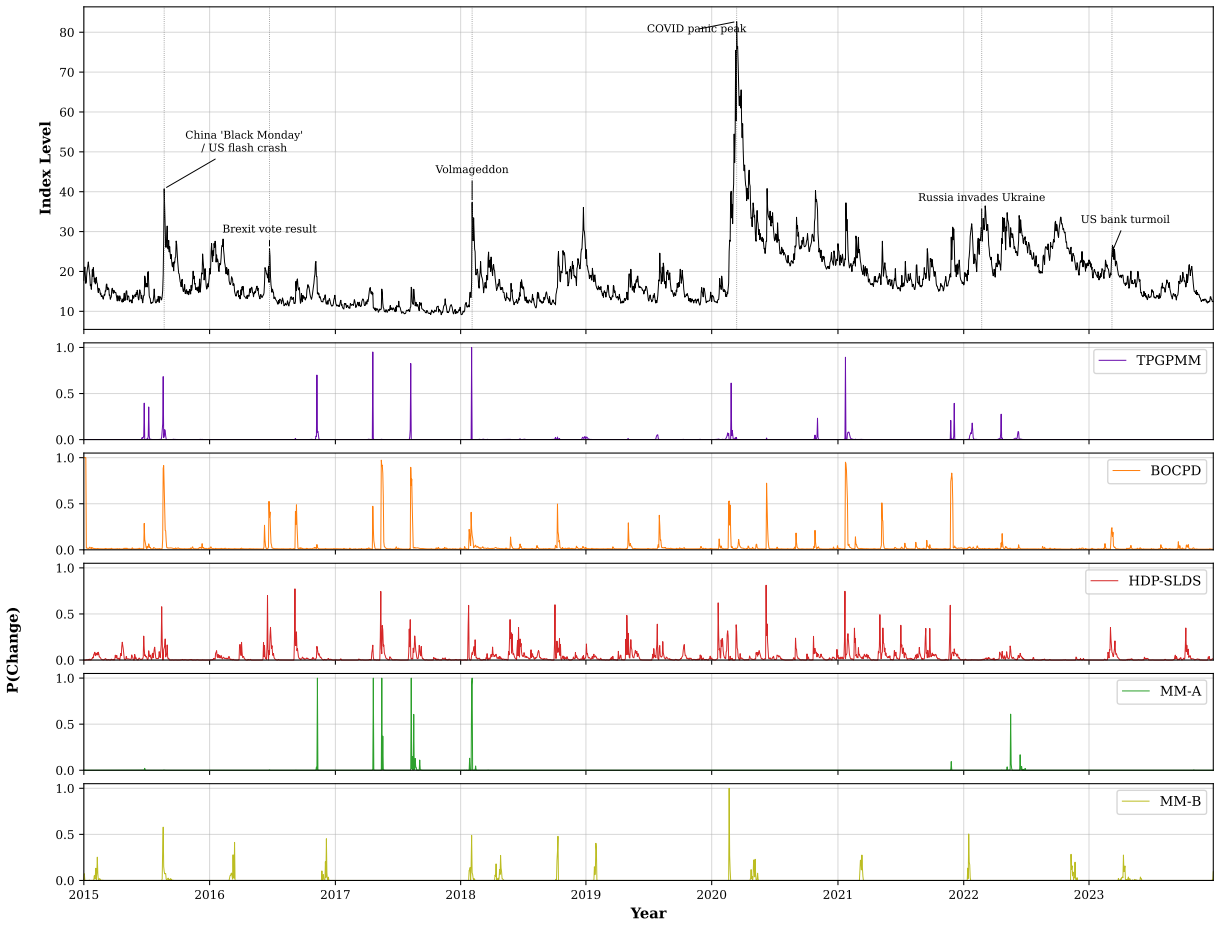


Figure 3: Analysis of the VIX dataset. Top panel: volatility index with the six benchmark market events. Lower five panels: posterior change-point probabilities $\widehat{\tau}_t$ for TPGPMM and the benchmark models.

0.2). BOCPD and HDP-SLDS produce denser and more diffuse boundary probabilities (20 and 38 peaks above 0.2), with HDP-SLDS in particular yielding a near-continuous stream of low-confidence transition candidates.

The posterior of TPGPMM places a mean of 13.4 regimes on the VIX (95% interval $[11, 17]$), and its modal partition has $K = 11$ regimes, with change-points on 2015-06-29, 2015-08-21, 2016-11-09, 2017-04-24, 2017-08-10, 2018-02-05, 2020-03-02, 2021-01-27, 2021-12-07 and 2022-04-22. The probability of a boundary within ± 10 trading days of these dates is 0.96, 1.00, 1.00, 0.99, 0.98, 1.00, 0.99, 0.99, 0.68 and 0.42. Three of the six benchmark events are matched:

- **China, August 2015** (1.00): a new regime opens on 21 August, the first day of the sell-off that

culminated on 24 August.

- **Volmageddon, February 2018** (1.00): a new regime opens on 5 February 2018.
- **COVID-19, March 2020** (0.89): a new regime opens on 2 March 2020, at the start of the sell-off, two weeks before the 16 March peak.

The 2016 Brexit vote (0.00), the Russia–Ukraine invasion (0.01) and the 2023 bank turmoil (0.00) receive no boundary mass: these short-lived spikes are absorbed as within-regime variation of regimes that began earlier. TPGPMM thus identifies persistent structural shifts rather than isolated excursions, at the cost of not marking every major event.

The alternatives with estimated noise behave differently. With the RBF kernel and estimated noise ($\sigma_{\mathrm{n}} = 0.063$), the posterior fragments the series into 29.0 regimes on average. With the rougher Matérn-3/2 kernel ($\sigma_{\mathrm{n}} = 0.049$), the posterior is again parsimonious (8.7 regimes) and matches China and Volmageddon (1.00 and 1.00), but it absorbs the COVID-19 panic (0.00) into a long, elevated-volatility regime that runs from February 2018 to June 2022.

The BOCPD posterior at index t is a distribution over the run-length (time since the last change point) and tends to spread mass over windows of plausible boundary locations. Similarly, the Sticky HDP-SLDS uses a Markovian regime-switching prior on latent states but lacks global partition coherence, which can produce local and noisy transition probabilities.

Martinez–Mena also employs a Pitman–Yor-induced prior on ordered compositions; consequently, the differences between MM and TPGPMM are driven by the within-block likelihood rather than the prior. In the log-VIX application, the autocorrelation parameter of MM-A saturates near the unit root (posterior mean $\phi = 0.997$), so each regime behaves like a random walk whose accumulated drift absorbs persistent level shifts: MM-A merges the China sell-off and COVID-19 into long regimes while still isolating a few short spikes. MM-B, whose observations are independent within a block, reacts instead to transient level changes and uses 16.7 regimes on average. TPGPMM, by contrast, models within-block dependence as a GP with non-Markovian covariance. Transient spikes can be absorbed into the within-block GP posterior, and the marginal likelihood penalizes adding new hyperparameters (ℓ_k, s_k^2) unless persistence and amplitude genuinely differ across a candidate boundary. This mechanism explains why TPGPMM's segmentation in Figure 3 is both sparse and aligned with persistent episodes.

To assess partition stability, we compute the Posterior Similarity Matrix (PSM) for each method, where entry (i, j) is the posterior probability that observations i and j belong to the same regime (Figure 4). TPGPMM yields a clean block-diagonal structure. Eight of its ten modal boundaries are sharp (probability 0.96 or more). The main uncertainty concerns a possible split of the 2018–2020 regime during 2019 (boundary probabilities 0.21–0.28) and the boundaries of late 2021 and 2022, which receive probabilities between 0.36 and 0.68. MM-A and MM-B also yield sharply block-diagonal PSMs. MM-A's is dominated by two long blocks, from 2015 to November 2016 and from February 2018 to May 2022, which merge the China sell-off and the Brexit vote, and the COVID-19 panic and the Russia–Ukraine invasion; MM-B's has many smaller blocks. BOCPD exhibits cross-block bleeding consistent with its diffuse run-length posterior, while HDP-SLDS collapses to a near-diagonal pattern indicative of many short, weakly distinguished regimes.

Across the five methods, only TPGPMM produces a segmentation that is at once sparse, confident and aligned with persistent market episodes. The results suggest that parsimonious change-point detection requires both a global prior on contiguous partitions and a within-block likelihood flexible enough to absorb transient variation without introducing unnecessary regimes.

7.2.1 Interpretation of the inferred regimes

To obtain an interpretable summary of the latent structure, we extract the *modal partition* $\widehat{\rho}$, defined as the most frequently sampled composition $\rho^{(m)}$ in the MCMC output. It accounts for only 0.19% of the draws, because boundary positions vary by a few days between draws; the boundary probabilities reported above show that the boundaries themselves are well identified. Conditional on this fixed partition, each block's kernel hyperparameters $\boldsymbol{\phi}_k = (\ell_k, s_k^2)$ are set to the posterior mean of $\{\boldsymbol{\phi}_k^{(m)}\}$ over the subset of MCMC draws with $\rho^{(m)} = \widehat{\rho}$, and the per-block posterior mean function is then computed by standard GP conditioning on the within-block data with the noise variance σ_{n}^2. We refer to the resulting fit as the *maximum a posteriori*

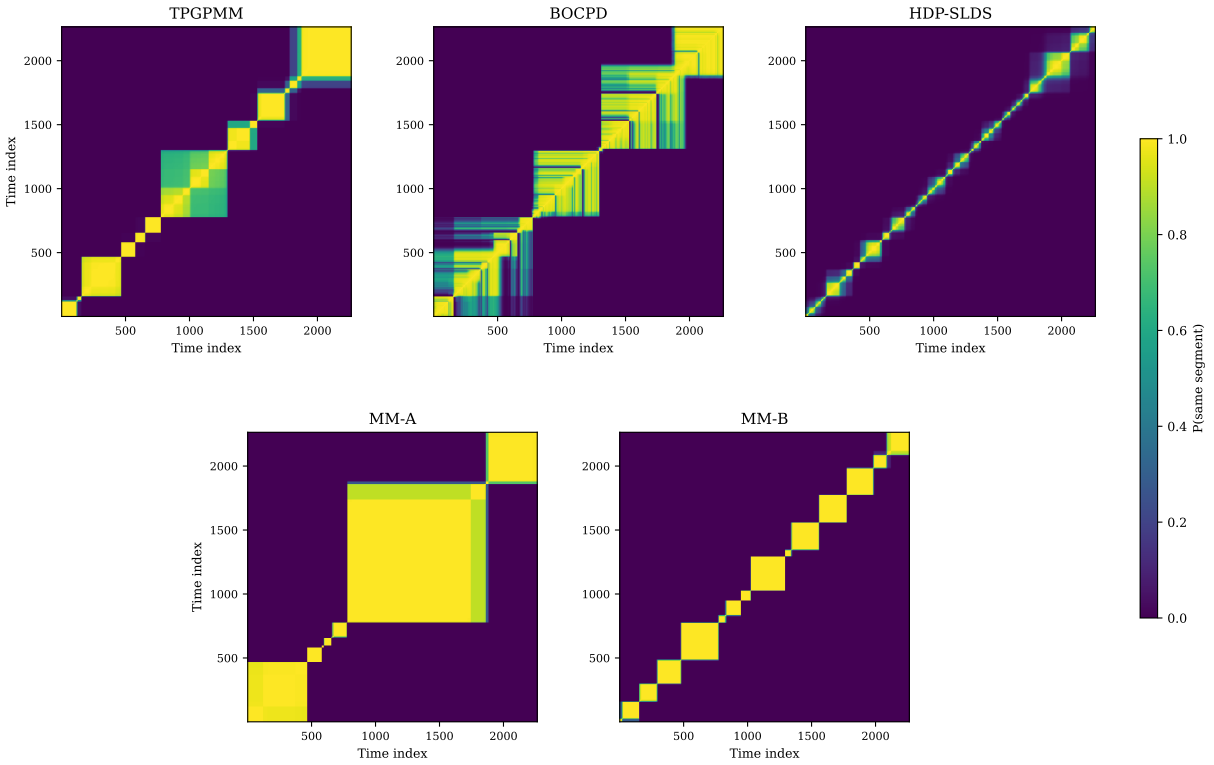


Figure 4: Posterior similarity matrices for the VIX time series under each method.

(MAP) segmentation for brevity, although it is technically a partition-MAP with within-block posterior-mean kernels.

Figure 5 shows the MAP segmentation for the VIX series. The eleven inferred regimes recover an economically coherent narrative; the dates of the boundaries coincide with the events named below, which we use only as interpretation:

- **Block 1** is the calm first half of 2015 (VIX around 15), with a long length-scale.
- **Block 2** opens on 29 June 2015, when Greece closed its banks, and runs to the eve of the August sell-off.
- **Block 3** runs from the China sell-off of August 2015 to the 2016 US election. It is a volatile phase with a short length-scale ($\ell = 11.7$) that absorbs the Brexit vote.
- **Blocks 4–6** capture the exceptionally calm period from November 2016 to January 2018, with the lowest VIX levels in the sample (around 10–12) and long length-scales. Their boundaries fall in late April 2017, after the first round of the French presidential election, and on 10 August 2017, a short-lived spike.
- **Block 7** opens with Volmageddon on 5 February 2018 and runs to the end of February 2020, absorbing the late-2018 sell-off.
- **Block 8** is the COVID-19 regime, from 2 March 2020 to January 2021, with the highest level in the sample (VIX around 30).
- **Block 9** opens with the retail-trading volatility spike of late January 2021 and has the largest signal variance ($s^2 = 5.2$).

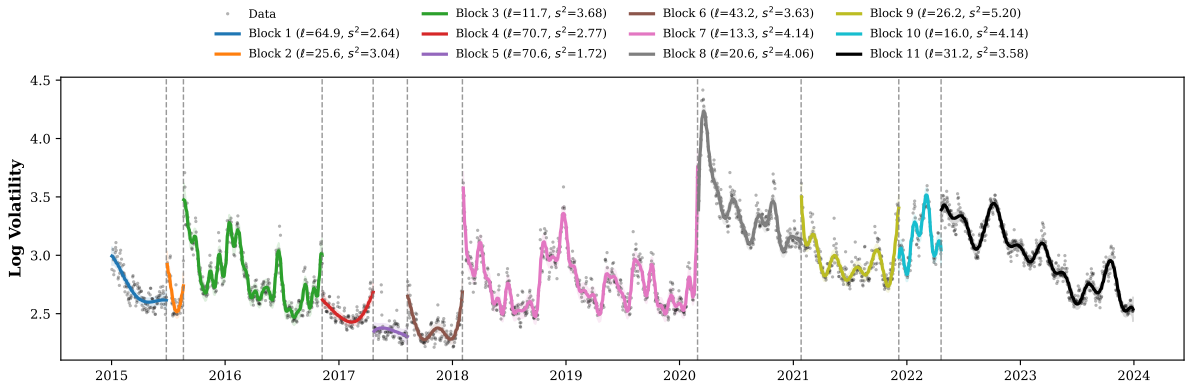


Figure 5: MAP segmentation of the VIX series. Coloured lines show the per-block GP posterior means under the modal partition $\widehat{\rho}$, with kernel hyperparameters fixed to the posterior mean across MCMC draws matching that partition. Dashed vertical lines mark the inferred internal change-points. The legend gives the averaged kernel hyperparameters (ℓ_k, s_k^2) for each of the eleven blocks.

- **Blocks 10 and 11** cover December 2021 to April 2022, which contains the Russia–Ukraine invasion, and the 2022 bear market and its normalisation through 2023, which contains the bank turmoil of March 2023. Their boundaries are the least certain (0.68 and 0.42).

Throughout this discussion, "high-volatility" and "low-volatility" regimes refer to the level of $\log(\text{VIX}_t)$ implied by the within-regime posterior mean of f_k, not to a model parameter: TPGPMM characterises each regime by its kernel hyperparameters (ℓ_k, s_k^2)—persistence and amplitude—and the level itself emerges from conditioning on the within-block data.

7.3 10-year Treasury yield

The posterior of TPGPMM places a mean of 7.0 regimes on the standardised yield (95% interval $[5, 10]$), with a 5-regime MAP partition. We interpret it against three monetary-policy events, listed in Table 3.

Table 3: Benchmark monetary-policy events used to interpret the posterior 10-year Treasury segmentation.

Event (Date)	**Policy / market impact**
Fed mid-cycle easing (2019-07-31)	First of three 2019 rate cuts as growth and trade-war risks mounted.
COVID emergency cuts (2020-03-15)	Policy rate cut to the zero lower bound; large-scale asset purchases.
Hawkish pivot (2021-12-15)	The Federal Open Market Committee (FOMC) accelerated tapering, signalling the 2022 tightening cycle.

The MAP partition (Table 4) splits the sample into five economically coherent regimes:

- **Normalisation and hiking cycle,** from 2015 to mid-2019. It covers the December 2015 lift-off, the 2016–2018 rate hikes and the balance-sheet runoff.
- **Mid-cycle easing,** from mid-2019. The timing of this boundary is uncertain: the modal partition places it on 30 July, the day before the first 2019 cut (0.22 within ±10 days), but more posterior mass (0.69) falls in early September 2019.
- **COVID and quantitative easing,** from 20 March 2020, five days after the emergency cuts of 15 March, with probability 1.00.
- **Tightening onset,** from late February 2022, three weeks before the first hike (probability 0.47).

- **Restrictive policy,** from mid-June 2022, when the Fed moved to 75-basis-point hikes. The mass for this boundary is spread over June and July 2022.

The COVID cuts are identified with certainty, and the 2019 easing and the 2022 tightening are detected, but with uncertain timing. The hawkish pivot of December 2021 receives little mass of its own (0.11). The Matérn-3/2 alternative with estimated noise dates the tightening more sharply (probability 0.93 in February 2022) but gives no boundary for the 2019 easing (0.01). As with VIX, transient moves that do not persist are absorbed within regimes.

Table 4: MAP segmentation of the standardised 10-year Treasury yield under TPGPMM. Mean yield (in percent) and daily-change standard deviation (in basis points) are computed within each inferred block; the boundary posterior probability is the posterior probability of a change-point within ±10 trading days of the block's left boundary (regime 1 has no left boundary).

Regime	Interpretation	Dates	N	Mean yield	Daily-change SD	Boundary post. prob.
1	Normalisation and hiking cycle	2015-01-02 to 2019-07-29	1,144	2.32%	4.22 bp	—
2	Mid-cycle easing	2019-07-30 to 2020-03-19	160	1.65%	6.50 bp	0.22
3	COVID; quantitative easing	2020-03-20 to 2022-02-25	486	1.19%	4.26 bp	1.00
4	Tightening onset	2022-02-28 to 2022-06-10	73	2.61%	7.15 bp	0.47
5	Restrictive policy	2022-06-13 to 2023-12-29	388	3.77%	7.74 bp	0.38

8 Nonstationary prediction

This section evaluates TPGPMM as a nonstationary predictive model and asks whether it predicts better than a stationary GP. We compare the posterior predictive distributions of all four specifications with stationary GPs whose hyperparameters are fitted to the data: their in-sample fit on 2015–2023, and their forecasts over the held-out year 2024. With a regime mean and estimated noise, the nonstationary model fits the training series more closely than the stationary GPs and has lower forecast RMSE except at the shortest horizons. Here the estimated-noise alternatives are the relevant models, and TPGPMM shows what its conservative noise level costs.

The results reveal a trade-off between sparse segmentation and predictive accuracy. TPGPMM produces the sparse, confident segmentation described in Section 7, with 13.4 regimes on VIX and 7.0 on the Treasury yield on average. Estimating the noise improves prediction. Every alternative fits more closely than TPGPMM in sample on both series; on VIX, their CRPS/sd is 0.093, 0.068 and 0.046, against 0.122. The alternatives also forecast more accurately in most comparisons across horizons and metrics, and the regime-mean alternative in all of them. These predictive gains are accompanied by changes in the inferred segmentation. With the RBF kernel, estimating the noise fragments the series into 29.0 regimes on VIX and 40.2 on the Treasury yield, and the regime-mean alternative uses 51 and 13. The Matérn-3/2 alternative remains parsimonious (8.7 and 8.2 regimes) but misses the COVID-19 regime on VIX and the 2019 easing on the Treasury yield.

8.1 Posterior predictive distribution

To characterise the latent function through a mixture of GPs and quantify in-sample uncertainty, we approximate the in-sample posterior predictive distribution via Monte Carlo integration, averaging over the MCMC draws of the partition and GP hyperparameters. The summaries reported here evaluate $p(Y_t \mid \mathbf{y})$ at the same inputs X_t used to fit the model; standard GP conditioning extends the construction to any $x^\star$ at no additional cost.

For a given MCMC sample m, observation Y_t falls within some block k under $\rho^{(m)}$, and the conditional predictive distribution is Gaussian by standard GP conditioning on the within-block data:

$$p\big(Y_t \mid \rho^{(m)},\, \boldsymbol{\phi}_k^{(m)},\, \sigma_{\mathrm{n}}^{2(m)},\, \mathbf{y}\big) = \mathcal{N}\big(Y_t \mid \mu_t^{(m)}, \nu_t^{(m)}\big), \tag{13}$$

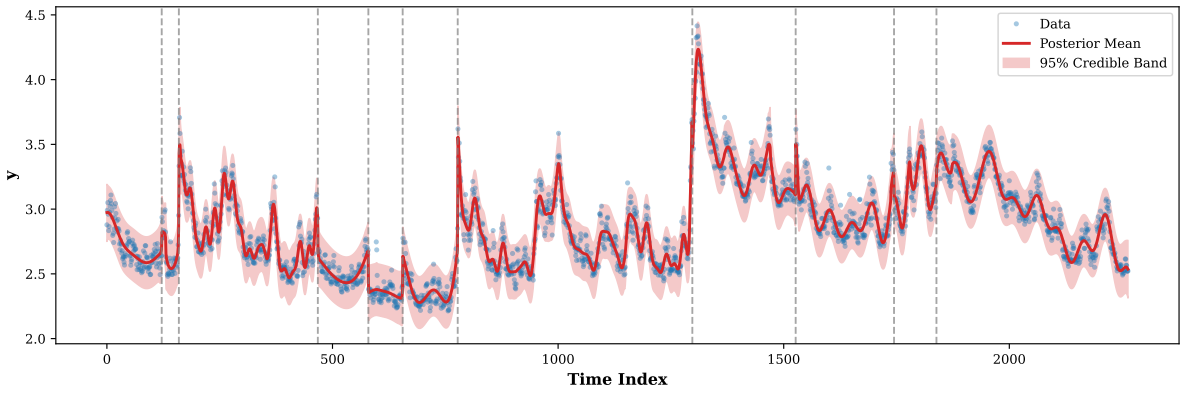


Figure 6: Posterior predictive mixture under TPGPMM on log-VIX. The red line is the posterior mean of Y_t, the shaded band is the 95% credible band, and the dashed vertical lines mark the modal-partition change-points. The construction follows Equation (16).

with

$$\mu_t^{(m)} = (\mathbf{k}_*^{(m)})^\top \big(K_k^{(m)} + \sigma_{\mathrm{n}}^{2(m)}\, I\big)^{-1} \mathbf{y}_{S_k}, \tag{14}$$

$$\nu_t^{(m)} = k_{**}^{(m)} - (\mathbf{k}_*^{(m)})^\top \big(K_k^{(m)} + \sigma_{\mathrm{n}}^{2(m)}\, I\big)^{-1} \mathbf{k}_*^{(m)} + \sigma_{\mathrm{n}}^{2(m)}, \tag{15}$$

where $K_k^{(m)}$ is the kernel matrix for block k at the training inputs, $\mathbf{k}_*^{(m)}$ is the cross-covariance vector between X_t and the within-block training inputs, and $k_{**}^{(m)}$ is the prior variance at X_t, all computed using $\boldsymbol{\phi}_k^{(m)}$. Marginalising over the joint posterior of $(\rho, \boldsymbol{\phi}_{1:K}, \sigma, \theta, \sigma_{\mathrm{n}})$ yields the mixture

$$p(Y_t \mid \mathbf{y}) \;\approx\; \frac{1}{M^\star} \sum_{m=1}^{M^\star} \mathcal{N}\big(Y_t \mid \mu_t^{(m)}, \nu_t^{(m)}\big). \tag{16}$$

The predictive mean is the average of the $\mu_t^{(m)}$ and credible intervals are the exact quantiles of the mixture (16). This naturally propagates uncertainty in the partition, the within-regime GP hyperparameters and, when it is estimated, the noise level, and widens the bands in regions of structural ambiguity. Forecasts are built the same way: each draw forecasts from its own final regime, and the draws are averaged.

Figure 6 shows the posterior predictive mixture on the log-VIX series. The red curve is the posterior mean of Y_t, the shaded band is the pointwise 95% credible band, and the dashed vertical lines mark the modal-partition change-points discussed in Section 7.2. The fitted predictive mean tracks the series closely within each regime, including during the COVID-19 peak.

8.2 In-sample fit

Table 5 reports the *in-sample* posterior-predictive fit on the log-VIX series. On the scale-normalised metrics, which are the comparable ones because the baselines operate on different targets, TPGPMM (RMSE/sd 0.208, CRPS/sd 0.122) fits better than laGP, HDP-SLDS and MM-B, about as well as MM-A (0.224, 0.121), and less closely than the fitted stationary GPs (RBF: 0.146, 0.081; Matérn-3/2: 0.115, 0.065), because its conservative noise level deliberately attributes daily fluctuations to noise. MM-A has no observation noise, so its predictive of y_t conditions on y_{t-1} within the regime rather than on y_t (Appendix E.3). The alternatives with estimated noise fit more closely than all the change-point baselines, and the regime-mean alternative fits best of all (0.076, 0.046). The fixed noise level that stabilises TPGPMM's segmentation therefore costs predictive accuracy, and we use TPGPMM for segmentation and the regime-mean alternative for prediction. Because in-sample fit rewards flexibility, the out-of-sample comparison is the forecasting exercise of Section 8.3.

Table 5: *In-sample* posterior-predictive fit on the log-VIX series ($n \approx 2264$ trading days). RMSE, NLPD and CRPS are the in-sample posterior-predictive metrics; RMSE/sd, ΔNLPD and CRPS/sd are their scale-normalised variants. Because the baselines operate on different targets (second column), the normalised columns are the comparable ones. Lower is better throughout. The standard GPs have hyperparameters fitted by maximum marginal likelihood; the laGP and HDP-SLDS rows are from the original study, and Martínez–Mena A and B are fitted with the settings of Martínez and Mena [2014]. Martínez–Mena A has no observation noise, so its predictive of y_t conditions on y_{t-1} within the regime (one step ahead) rather than on y_t itself.

Method	Target	RMSE	NLPD	CRPS	RMSE/sd	ΔNLPD	CRPS/sd	n
TPGPMM	log VIX	0.0714	−1.1042	0.0417	0.2082	−1.4537	0.1215	2264
TPGPMM (σ_n estimated)	log VIX	0.0564	−1.4339	0.0318	0.1642	−1.7834	0.0928	2264
TPGPMM (Matérn-3/2, σ_n estimated)	log VIX	0.0409	−1.7048	0.0233	0.1193	−2.0543	0.0678	2264
TPGPMM (Matérn-3/2, regime mean, σ_n estimated)	log VIX	0.0261	−2.0406	0.0159	0.0760	−2.3901	0.0463	2264
Standard GP (RBF)	log VIX	0.0502	−1.5255	0.0277	0.1464	−1.8750	0.0808	2264
Standard GP (Matérn-3/2)	log VIX	0.0394	−1.7117	0.0223	0.1149	−2.0612	0.0651	2264
laGP	log VIX	0.1328	−0.7501	0.0691	0.3869	−1.0996	0.2013	2264
HDP-SLDS	$\log(r_t^2+\epsilon)$	0.7444	1.2443	0.3903	0.2900	−1.1174	0.1520	2263
Martínez–Mena A	log VIX	0.0769	−1.2057	0.0414	0.2241	−1.5552	0.1206	2264
Martínez–Mena B	log VIX	0.1441	−0.4771	0.0820	0.4199	−0.8266	0.2388	2264

On the Treasury yield, Table 6 reports the in-sample posterior-predictive fit for the four specifications and the baselines. On the comparable scale-normalised metrics, TPGPMM (RMSE/sd 0.061, CRPS/sd 0.037) fits better than laGP (0.092, 0.046), HDP-SLDS and MM-B but less closely than MM-A (0.059, 0.032), whose regimes, with $\phi = 0.998$, track the yield from one day to the next, and than the fitted stationary GPs (CRPS/sd 0.025 for RBF and 0.018 for Matérn-3/2). The alternatives with estimated noise fit more closely than all the change-point baselines, and the regime-mean alternative fits best (0.021, 0.013).

Table 6: *In-sample* posterior-predictive fit on the standardised 10-year Treasury yield ($n \approx 2{,}251$). Columns as in Table 5.

Method	Target	RMSE	NLPD	CRPS	RMSE/sd	ΔNLPD	CRPS/sd	n
TPGPMM	std. yield	0.0605	−1.1785	0.0371	0.0605	−2.5972	0.0371	2251
TPGPMM (σ_n estimated)	std. yield	0.0357	−1.8700	0.0203	0.0357	−3.2887	0.0203	2251
TPGPMM (Matérn-3/2, σ_n estimated)	std. yield	0.0244	−2.1798	0.0143	0.0244	−3.5985	0.0143	2251
TPGPMM (Matérn-3/2, regime mean, σ_n estimated)	std. yield	0.0215	−2.2681	0.0129	0.0215	−3.6868	0.0129	2251
Standard GP (RBF)	std. yield	0.0452	−1.6580	0.0250	0.0452	−3.0767	0.0250	2251
Standard GP (Matérn-3/2)	std. yield	0.0320	−1.9500	0.0181	0.0320	−3.3688	0.0181	2251
laGP	std. yield	0.0919	−1.2189	0.0460	0.0919	−2.6377	0.0460	2251
HDP-SLDS	$\log(\Delta y_t^2+\epsilon)$	1.9097	1.4495	0.7079	0.3094	−1.7894	0.1147	2250
Martínez–Mena A	std. yield	0.0588	−1.5288	0.0318	0.0588	−2.9476	0.0318	2251
Martínez–Mena B	std. yield	0.1428	−0.4421	0.0843	0.1429	−1.8608	0.0843	2251

8.3 Out-of-sample forecasting

Out-of-sample forecasting is the more demanding test. Table 7 evaluates 1–5 trading-day-ahead forecasts of log-VIX on a held-out 2024 window: all models are fit *once* on the 2015–2023 training series and forecast forward with a daily rolling origin, conditioning on the observed data up to each origin but without re-estimation. TPGPMM forecasts from each draw's final inferred regime and averages over draws; the standard GPs are single stationary GPs with hyperparameters fitted on the training window.

- **The regime-mean alternative** is the most accurate model in CRPS at every horizon and in RMSE from two days ahead. At one day the fitted Matérn GP has a slightly lower RMSE (0.470 against 0.476). The margins are small: at five days the RMSE/sd is 0.940 against 0.946, and the CRPS/sd 0.482 against 0.496.
- **Without a regime mean,** every TPGPMM specification is less accurate than the fitted standard GP

with the same kernel, at every horizon. This includes TPGPMM itself, whose RMSE/sd is 0.706 at one day and 1.400 at five, against 0.503 and 1.003 for the fitted RBF GP.

- **Among the TPGPMM specifications,** all three alternatives with estimated noise forecast more accurately than TPGPMM in CRPS at every horizon. In RMSE they do so at every horizon, except the RBF alternative at four and five days.

On VIX, the regime structure therefore matches the best stationary GP in forecasting only when each regime has its own mean and the noise is estimated. The value of TPGPMM itself lies in the segmentation.

Table 7: Out-of-sample forecasting on log-VIX. All models are fit once on the 2015–2023 training window and forecast 1–5 trading days ahead across 2024 (252 days) with a daily rolling origin, conditioning on observed data up to each origin but not re-estimated. TPGPMM forecasts are posterior mixtures over 500 draws, each forecasting from its final regime; the standard GPs are single stationary GPs with hyperparameters fitted by maximum marginal likelihood on the training window. Entries are horizon-wise RMSE and CRPS normalised by the test standard deviation; lower is better, best in bold.

	RMSE/sd					CRPS/sd				
Model	$h=1$	2	3	4	5	$h=1$	2	3	4	5
TPGPMM	0.706	0.893	1.069	1.237	1.400	0.369	0.473	0.574	0.671	0.761
TPGPMM (σ_n estimated)	0.636	0.846	1.046	1.243	1.435	0.322	0.436	0.544	0.649	0.752
TPGPMM (Matérn-3/2, σ_n estimated)	0.548	0.780	0.978	1.146	1.288	0.272	0.397	0.501	0.590	0.674
TPGPMM (Matérn-3/2, regime mean, σ_n estimated)	0.476	**0.664**	**0.796**	**0.886**	**0.940**	**0.227**	**0.325**	**0.392**	**0.441**	**0.482**
Standard GP (RBF)	0.503	0.726	0.886	0.978	1.003	0.249	0.360	0.440	0.496	0.533
Standard GP (Matérn-3/2)	**0.470**	0.670	0.808	0.898	0.946	0.230	0.332	0.402	0.455	0.496

On the Treasury yield we follow the same protocol (Table 8).

- **In RMSE,** the regime-mean alternative and the fitted Matérn GP are close. The GP is more accurate at one and two days (0.248 against 0.250 at one day), and the regime-mean alternative from three days ahead (0.595 against 0.614 at five days).

- **In CRPS,** the fitted Matérn GP is more accurate at every horizon (0.347 against 0.377 at five days). The regime-mean alternative's predictive bands are too wide, which the CRPS penalises: all 2024 observations fall inside its 95% five-day interval.

- **Within the RBF family,** TPGPMM is less accurate than the fitted RBF GP at one day but more accurate from three days ahead (0.729 against 0.909 at five days); with estimated noise, the RBF version is more accurate at every horizon. The Matérn-3/2 version without a regime mean is less accurate than the fitted Matérn GP at every horizon.

- **Among the TPGPMM specifications,** the RBF and regime-mean alternatives forecast more accurately than TPGPMM at every horizon in both metrics. The Matérn-3/2 alternative does so in RMSE at every horizon except four days, but in CRPS only at one and two days.

As on VIX, the regime structure gives forecasts comparable to a well-tuned stationary GP rather than clearly better ones. Its advantages are the interpretable segmentation and, for the regime-mean alternative, the fit.

9 Conclusions

We introduced the Time-partitioning Gaussian Process Mixture Model, a Bayesian nonparametric framework that infers a contiguous segmentation of a time-ordered series jointly with regime-specific Gaussian processes, learning the kernel and partition hyperparameters, including the Pitman–Yor discount σ, directly from the data, with the observation noise fixed at a conservative, data-informed level in the main specification. The model serves two purposes. For regime identification, TPGPMM localises the simulated change points more accurately than every change-point baseline, with an F_1 at the level of the best one, Martinez–Mena A, and the alternatives with estimated noise perform comparably. On the log-VIX series it produces a confident

Table 8: Out-of-sample forecasting on the standardised 10-year Treasury yield over 2024 (250 days), with the protocol of Table 7.

	RMSE/sd					CRPS/sd				
Model	$h = 1$	2	3	4	5	$h = 1$	2	3	4	5
TPGPMM	0.331	0.433	0.530	0.632	0.729	0.194	0.247	0.301	0.359	0.414
TPGPMM (σ_{n} estimated)	0.282	0.386	0.484	0.595	0.699	0.158	0.216	0.272	0.339	0.403
TPGPMM (Matérn-3/2, σ_{n} estimated)	0.265	0.428	0.525	0.639	0.725	0.149	0.246	0.320	0.406	0.483
TPGPMM (Matérn-3/2, regime mean, σ_{n} estimated)	0.250	0.383	**0.448**	**0.531**	**0.595**	0.141	0.222	0.274	0.330	0.377
Standard GP (RBF)	0.291	0.447	0.597	0.761	0.909	0.161	0.246	0.331	0.429	0.517
Standard GP (Matérn-3/2)	**0.248**	**0.365**	0.452	0.545	0.614	**0.137**	**0.201**	**0.248**	**0.304**	**0.347**

segmentation whose boundaries match three of the six benchmark events (the China sell-off, Volmageddon and COVID-19), while shorter-lived spikes are absorbed within regimes; on the Treasury yield it identifies the COVID-19 cuts with certainty and the 2019 easing and the 2022 tightening with uncertain timing. For nonstationary prediction, the specifications that estimate the noise fit better than the change-point baselines, and the regime-mean alternative fits better than a fitted stationary GP and has lower forecast RMSE except at the shortest horizons (Section 8). The two purposes trade sparsity against predictive accuracy: fixing the noise at a conservative level keeps TPGPMM's segmentation sparse and aligned with the events, while estimating it improves fit and forecasts but either fragments the segmentation or misses events.

For segmentation, the results point to three ingredients. The Pitman–Yor-induced prior on contiguous compositions adapts between the Dirichlet-process limit and heavier-tailed segment-size structures; the non-Markovian GP likelihood absorbs short-range fluctuations as within-regime dynamics and penalises unnecessary boundaries through the marginal likelihood; and the conservative noise level keeps daily fluctuations from being modelled as new regimes. Without any one of them the segmentation changes: estimating the noise with the RBF kernel fragments the series; Martinez–Mena A, with the same prior but Markovian regimes, is as sparse but places its boundaries at short spikes rather than at the China sell-off and COVID-19; Martinez–Mena B, with independent observations, reacts to transient level changes; and BOCPD and HDP-SLDS produce diffuse, frequent boundaries.

TPGPMM is most useful for series that decompose into persistent regimes with non-Markovian within-regime dynamics, where transient shocks should be absorbed rather than flagged and each regime warrants a full functional description rather than only a mean or variance shift, as in regime analysis for risk management and policy evaluation; it is correspondingly less suited to rapidly switching or recurring states. With respect to limitations, we highlight three. First, inference is computationally demanding: each iteration factorises the covariances of the regimes it updates, at a cost cubic in their lengths, so specifications with fewer and longer regimes are slower. On one core, a 60,000-iteration chain on the VIX takes about 8 minutes for TPGPMM, 5 and 7 minutes for the RBF and regime-mean alternatives, and 50 minutes for the Matérn-3/2 alternative, which has the fewest and longest regimes. This is practical for series of several thousand observations, which suits macro-financial regime analysis on daily, monthly or quarterly data, whereas multivariate extensions and more complex kernels would raise the cost; we deliberately prioritised an accurate approximation of the posterior (Appendix I) over scaling to very long series. Second, although forecasts average over posterior draws, each draw forecasts from its own final regime, so the predictive mixture reverts towards the recent regime level and cannot anticipate a new regime; the regime-mean alternative improves on a well-tuned stationary GP only by small margins and beyond the shortest horizons, and on the Treasury yield the GP remains more accurate in CRPS, so a persistent component shared across regimes, such as a random-walk or long-range Ornstein–Uhlenbeck kernel, is the natural extension. Third, the composition prior is defined for a fixed series length, so online use would require a sequential construction.

References

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Richard Arratia, A. D. Barbour, and Simon Tavaré. *Logarithmic Combinatorial Structures: A Probabilistic Approach*. EMS Monographs in Mathematics. European Mathematical Society, 2003. doi: 10.4171/000.

Daniel Barry and J. A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993. doi: 10.1080/01621459.1993.10594323.

Mickaël Binois, Robert B. Gramacy, and Mike Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018. doi: 10.1080/10618600.2018.1458625.

Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86 (2):221–241, 1998. doi: 10.1016/S0304-4076(97)00115-2.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, 2013.

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H. Mena, Igor Prünster, and Matteo Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2015. doi: 10.1109/TPAMI.2013.217.

David B. Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008. doi: 10.1093/biomet/asn012.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. doi: 10.1080/01621459.1995.10476550.

Paul Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006. doi: 10.1007/s11222-006-8450-8.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973. doi: 10.1214/aos/1176342360.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, 2011. doi: 10.1109/TSP.2010.2102756.

Roman Garnett, Michael A. Osborne, Steven Reece, Alex Rogers, and Stephen J. Roberts. Sequential Bayesian prediction in the presence of changepoints and faults. *The Computer Journal*, 53(9):1430–1446, 2010. doi: 10.1093/comjnl/bxq003.

Theo Gasser, Lothar Sroka, and Christine Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633, 1986. doi: 10.1093/biomet/73.3.625.

Alexander Gnedin and Jim Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006. doi: 10.1007/s10958-006-0335-z.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Dilan Görür and Carl Edward Rasmussen. Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010. doi: 10.1007/s11390-010-9355-8.

Robert B. Gramacy and Daniel W. Apley. Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015. doi: 10.1080/10618600.2014.914442.

Robert B. Gramacy and Herbert K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008. doi: 10.1198/016214508000000689.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. doi: 10.1093/biomet/82.4.711.

Jim E. Griffin and Mark F. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006. doi: 10.1198/016214505000000727.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989. doi: 10.2307/1912559.

Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 732–740, 2016.

David Higdon, Jenise Swall, and John Kern. Non-stationary spatial modeling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 761–768. Oxford University Press, 1999. doi: 10.1093/oso/9780198504856.003.0036.

Sonia Jain and Radford M. Neal. A split–merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004. doi: 10.1198/1061860043001.

Matthias Katzfuss. Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24(3): 189–200, 2013. doi: 10.1002/env.2200.

Hyoung-Moon Kim, Bani K. Mallick, and Chris C. Holmes. Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668, 2005. doi: 10.1198/016214504000002014.

Stanley I. M. Ko, Terence T. L. Chong, and Pulak Ghosh. Dirichlet process hidden Markov multiple change-point model. *Bayesian Analysis*, 10(2):275–296, 2015. doi: 10.1214/14-BA910.

Steven N. MacEachern. Dependent Dirichlet processes. *Bayesian Analysis*, pages 1–46, 2026. doi: 10.1214/26-BA1610. Advance publication.

Asael Fabian Martínez and Ramsés H. Mena. On a nonparametric change point detection model in Markovian regimes. *Bayesian Analysis*, 9(4):823–858, 2014. doi: 10.1214/14-BA878.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley, New York, 2000. doi: 10.1002/0471721182.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. doi: 10.1080/10618600.2000.10474879.

Christopher J. Paciorek and Mark J. Schervish. Nonstationary covariance functions for Gaussian process regression. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

Jim Pitman. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer, Berlin, 2006. doi: 10.1007/b11601500.

Jim Pitman and Marc Yor. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997. doi: 10.1214/aop/1024404422.

Fernando A. Quintana and Pilar L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):557–574, 2003. doi: 10.1111/1467-9868.00402.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Carl Edward Rasmussen. The infinite Gaussian mixture model. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, 1999.

Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

John Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230, 1984. doi: 10.1214/aos/1176346788.

Sylvia Richardson and Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59 (4):731–792, 1997. doi: 10.1111/1467-9868.00095.

Yunus Saatçi, Ryan D. Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.

Paul D. Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992. doi: 10.1080/01621459.1992.10475181.

Annie Sauer, Robert B. Gramacy, and David Higdon. Active learning for deep Gaussian process surrogates. *Technometrics*, 65(1):4–18, 2023. doi: 10.1080/00401706.2021.2008505.

Alexandra M. Schmidt and Anthony O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(3):743–758, 2003. doi: 10.1111/1467-9868.00413.

Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York, 1999. doi: 10.1007/978-1-4612-1494-6.

Volker Tresp. Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2000.

Ryan Darby Turner. *Gaussian Processes for State Space Models and Change Point Detection*. PhD thesis, Department of Engineering, University of Cambridge, 2011.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021.

Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020. doi: 10.1080/01621459.2019.1585255.

A Derivation of the Pitman–Yor prior on ordered segmentations

A Gibbs-type prior is characterised by an EPPF of the form

$$p_K^{(n)}(n_1, \ldots, n_K) = V_{n,K} \prod_{k=1}^{K} (1-\sigma)_{n_k-1},$$

where $\sigma < 1$ is a discount parameter, $(1-\sigma)_{n_k-1}$ denotes the rising factorial $(1-\sigma)(2-\sigma)\cdots(n_k-1-\sigma)$, and the nonnegative weights $V_{n,K}$ satisfy the backward recursion

$$V_{n,K} = (n - K\sigma)\, V_{n+1,K} + V_{n+1,K+1}, \qquad V_{1,1} = 1.$$

Different choices of $(\sigma, V_{n,K})$ recover familiar Bayesian nonparametric priors. In particular:

Dirichlet process. For $\sigma = 0$,

$$V_{n,K} = \frac{\theta^{K-1}}{(\theta+1)_{n-1}},$$

which implies logarithmic growth of K.

Pitman–Yor process. For $\sigma \in (0,1)$,

$$V_{n,K} = \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{(\theta+1)_{n-1}},$$

which implies polynomial growth of K.

For $\sigma \geq 0$, Gibbs-type priors are nonparametric and provide explicit control over the growth of the number of blocks through the discount parameter σ.

We derive the prior in (4) from the EPPF of the Pitman–Yor process. Let Π_n denote an exchangeable random partition of $[n] = \{1, \ldots, n\}$ with K blocks, and suppose that the block sizes are $(n_1, \ldots, n_K)$, where $n_k \geq 1$ and $\sum_{k=1}^{K} n_k = n$. For an exchangeable partition with EPPF $p_K^{(n)}$, the probability of any particular unordered partition of $[n]$ with these block sizes is

$$p_K^{(n)}(n_1, \ldots, n_K),$$

where the EPPF is symmetric in its arguments.

To induce a distribution on ordered compositions, we use the exchangeable random order construction of Pitman [2006]. Conditional on the unordered partition and on K, the K blocks are placed in a uniformly random order. Hence each ordered arrangement of the K blocks has conditional probability $1/K!$.

Now fix an ordered composition

$$\rho = (n_1, \ldots, n_K).$$

The number of ordered assignments of the labels $\{1, \ldots, n\}$ to blocks of sizes $n_1, \ldots, n_K$ is the multinomial coefficient

$$\binom{n}{n_1, \ldots, n_K} = \frac{n!}{\prod_{k=1}^{K} n_k!}.$$

Each ordered assignment corresponds to an unordered exchangeable partition together with a particular ordering of its K blocks. Therefore, summing over all ordered label assignments with block-size sequence ρ gives

$$\Pr(\rho) = \binom{n}{n_1, \ldots, n_K} \frac{1}{K!} p_K^{(n)}(n_1, \ldots, n_K).$$

This is precisely the induced probability mass function on compositions under the exchangeable random order construction.

For the Pitman–Yor process with discount parameter $\sigma \in [0, 1)$ and strength parameter $\theta > -\sigma$, the EPPF is

$$p_K^{(n)}(n_1, \ldots, n_K) = \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1},$$

with the convention that the empty product equals one when $K = 1$. Substituting this EPPF into the composition probability above yields

$$\Pr(\rho) = \frac{n!}{K! \prod_{k=1}^{K} n_k!} \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1}.$$

This is the expression in (4). Summing it over the compositions with K blocks recovers the Pitman–Yor law of K; we checked this against the Chinese-restaurant recursion to machine precision.

Finally, in the segmentation model, an ordered composition $\rho = (n_1, \dots, n_K)$ determines a unique contiguous segmentation of $[n]$ by assigning the first block to observations $1, \dots, n_1$, the second block to observations $n_1 + 1, \dots, n_1 + n_2$, and so on. Thus the probability mass assigned to the ordered composition is inherited directly by the corresponding contiguous segmentation. The resulting prior preserves the Pitman–Yor distribution on the number and sizes of blocks while restricting the support to temporally ordered contiguous regimes.

Two properties. Writing $w_\sigma(m) = (1-\sigma)_{m-1}/m!$, the prior factorises as

$$\Pr(\rho \mid \sigma, \theta) = \frac{n!}{(\theta+1)_{n-1}} \cdot \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{K!} \cdot \prod_{k=1}^{K} w_\sigma(n_k).$$

The middle factor depends on K only, so the prior is of Gibbs type but not a product partition model. The family is also not projective in n: removing the last observation from a draw at size $n+1$ does not give the prior at size n, because in a uniformly random order the last block is a uniformly chosen block, whereas the block containing a fixed element of an exchangeable partition is size-biased. For $n = 5, 8, 11$ the total-variation distance between the two laws is 0.06–0.10. The model is therefore defined for a fixed series length, which is why an online version needs a different construction (Section 9).

Truncation constant. The normalising constant of the truncated prior (5) is

$$Z_n(\sigma, \theta) = \sum_{K \geq 1} \frac{n!}{(\theta+1)_{n-1}} \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{K!} A_K(\sigma), \qquad A_K(\sigma) = \sum_{\substack{n_1 + \cdots + n_K = n \\ n_k \geq m_{\min}}} \prod_{k=1}^{K} w_\sigma(n_k).$$

The sums $A_K(\sigma)$, $K = 1, \dots, \lfloor n/m_{\min} \rfloor$, are computed by repeated convolution of the sequence $w_\sigma(m)\mathbf{1}\{m \geq m_{\min}\}$ with rescaling, on a grid of σ values, and interpolated linearly in σ; the product is evaluated as $\prod_{i<K}(\theta + i\sigma) = \sigma^{K-1}\Gamma(\theta/\sigma + K)/\Gamma(\theta/\sigma + 1)$. The table takes about ten seconds for $n = 2{,}264$ and matches brute-force enumeration exactly for $n = 12$ and 18. For $\sigma = 0$ and $n \to \infty$, $Z_n \to e^{-1.5\theta}$ when $m_{\min} = 3$, so omitting it has a large effect on θ.

B Specifications of the within-regime model

The four specifications share the partition prior (5), the minimum regime length, the priors on (σ, θ) and the prior $\log \ell_k \sim \mathcal{N}(\log 10, 0.6^2)$. They differ in the kernel, in whether each regime has its own mean, and in whether the observation noise is fixed or estimated (Table 9).

Table 9: Within-regime specifications. The amplitude prior is $\log s_k^2 \sim \mathcal{N}(0, 0.8^2)$ except for the regime-mean alternative, for which it is $\mathcal{N}(\log \widehat{\mathrm{var}}(\mathbf{y}), 1)$.

Specification	Kernel	Regime mean	Noise level σ_{n}
TPGPMM	RBF	no	fixed at 0.1
TPGPMM (σ_{n} estimated)	RBF	no	estimated
TPGPMM (Matérn-3/2, σ_{n} estimated)	Matérn-3/2	no	estimated
TPGPMM (Matérn-3/2, regime mean, σ_{n} estimated)	Matérn-3/2	yes	estimated

Estimated observation noise. The noise level is shared across regimes, as in (7), with prior

$$\log \sigma_{\mathrm{n}} \sim \mathcal{N}(\log 0.1,\ 1),$$

centred on the value fixed in the main specification. Because σ_{n}^2 enters every block covariance, each update of σ_{n} recomputes all block likelihoods (Appendices C.3 and D).

Matérn-3/2 kernel. The RBF kernel (8) is replaced by

$$C_{\boldsymbol{\phi}_k}(x, x') = s_k^2 \left(1 + \frac{\sqrt{3}\,|x - x'|}{\ell_k}\right) \exp\!\left(-\frac{\sqrt{3}\,|x - x'|}{\ell_k}\right), \tag{17}$$

with the same block hyperparameters $\boldsymbol{\phi}_k = (\ell_k, s_k^2)$ and priors [Stein, 1999]. Its sample paths are continuous and once mean-square differentiable, whereas those of the RBF kernel are infinitely differentiable, so it can follow rougher within-regime variation. The kernel is stationary, so the block covariances remain symmetric Toeplitz and the sampler of Section 5.1 applies unchanged.

Regime mean. The regime-mean alternative writes the latent function of block k as a constant level plus a zero-mean GP with the Matérn-3/2 kernel (17),

$$f_k(x) = \mu_k + g_k(x), \qquad g_k \mid \boldsymbol{\phi}_k \sim \mathcal{GP}\big(0, C_{\boldsymbol{\phi}_k}\big), \qquad \mu_k \sim \mathcal{N}(m_0, v_0),$$

with μ_k independent across regimes and of g_k. Integrating out μ_k together with f_k replaces the block marginal likelihood (9) by

$$L_k(\boldsymbol{\phi}_k) = \mathcal{N}\Big(\mathbf{y}_{S_k}\,;\, m_0\mathbf{1},\; K_k(\boldsymbol{\phi}_k) + v_0\mathbf{1}\mathbf{1}^\top + \sigma_{\mathrm{n}}^2 I_{|S_k|}\Big).$$

Adding the constant v_0 to every entry keeps the covariance symmetric Toeplitz, so the sampler is again unchanged. We set $m_0 = \bar{y}$ and $v_0 = \widehat{\mathrm{var}}(\mathbf{y})$, the sample mean and variance of the series, and centre the amplitude prior on the sample variance, $\log s_k^2 \sim \mathcal{N}(\log \widehat{\mathrm{var}}(\mathbf{y}), 1)$. The predictive distributions of Section 8.1 apply with $\mathbf{y}_{S_k}$ replaced by $\mathbf{y}_{S_k} - m_0\mathbf{1}$, every kernel value increased by v_0, and m_0 added to the predictive mean.

C MCMC implementation details

This appendix gives the full specification of the four update blocks summarised in Section 5.1. We write $a_k = \log \ell_k$ and $b_k = \log s_k^2$, and φ_ℓ, φ_s for the normal prior densities of a_k, b_k, including their normalising constants. The normalising constants matter because a split or a merge changes the number of hyperparameters. A derivation of every ratio is given in the accompanying note on the sampler equations.

C.1 Block A: structural updates of ρ

We enforce a minimum block length $m_{\min} = 3$. The structural move at each iteration is a *split*, which inserts a new internal boundary inside an existing splittable block, or a *merge*, which removes an internal boundary, fusing two adjacent blocks. If $K = 1$ a split is proposed; otherwise a split is proposed with probability $q = 1/2$ and a merge otherwise. We write q_K for the probability of proposing a split when there are K blocks and $s(\rho)$ for the number of splittable blocks (those with $n_j \geq 2m_{\min}$). Each structural move is followed by exact Gibbs draws of two regime boundaries.

Split ($K \to K + 1$). Choose j uniformly from the splittable blocks and draw the length c of the left child, $m_{\min} \leq c \leq n_j - m_{\min}$, from

$$\psi_j(c) = \frac{0.1}{n_j - 2m_{\min} + 1} + 0.9\,\pi_j(c),$$

where

$$\pi_j(c) \propto w_\sigma(c)\, w_\sigma(n_j - c)\, \mathcal{N}(\mathbf{y}_{1:c}^{(j)}; \boldsymbol{\phi}_j)\, \mathcal{N}(\mathbf{y}_{c+1:n_j}^{(j)}; \boldsymbol{\phi}_j)$$

is the conditional distribution of the split point given the parent's hyperparameters [Zanella, 2020]; the uniform component keeps every split point reachable. All the prefix and suffix likelihoods come from one Cholesky factorisation of the parent's covariance (Section 5.1). The children's hyperparameters are

$$a_L = a_j + w_R u_\ell, \quad a_R = a_j - w_L u_\ell, \quad b_L = b_j + w_R u_s, \quad b_R = b_j - w_L u_s, \qquad u_\ell, u_s \sim \mathcal{N}(0, \tau^2),$$

with $w_L = c/n_j$, $w_R = 1 - w_L$ and $\tau = 0.5$. This map is invertible, with inverse $a_j = w_L a_L + w_R a_R$ and $u_\ell = a_L - a_R$ (and likewise for b), and its Jacobian on the log scale is one. The split is accepted with probability $\min\{1, A\}$, where

$$A = \frac{\theta + K\sigma}{K+1} \frac{w_\sigma(c)\, w_\sigma(n_j - c)}{w_\sigma(n_j)} \frac{\varphi_\ell(a_L)\varphi_\ell(a_R)\varphi_s(b_L)\varphi_s(b_R)}{\varphi_\ell(a_j)\varphi_s(b_j)} \frac{\mathcal{N}(\mathbf{y}_L; \boldsymbol{\phi}_L)\,\mathcal{N}(\mathbf{y}_R; \boldsymbol{\phi}_R)}{\mathcal{N}(\mathbf{y}_{S_j}; \boldsymbol{\phi}_j)} \times \frac{(1 - q_{K+1})/K}{q_K\, \psi_j(c)/s(\rho)} \frac{1}{\mathcal{N}(u_\ell; 0, \tau^2)\,\mathcal{N}(u_s; 0, \tau^2)}.$$

The first two factors are the ratio of the partition priors (5), in which Z_n cancels. Equivalently, in terms of densities on (ℓ, s^2) the ratio carries the Jacobian $\ell_L \ell_R / \ell_j \cdot s_L^2 s_R^2 / s_j^2$.

Merge ($K \to K - 1$). Choose an internal boundary uniformly at random and merge the two adjacent blocks. The merged block receives the size-weighted geometric mean of the two children's hyperparameters, $a_j = w_L a_L + w_R a_R$ and $b_j = w_L b_L + w_R b_R$, and $u = (a_L - a_R,\, b_L - b_R)$. The merge is accepted with probability $\min\{1, A^{-1}\}$, with A evaluated at the split that would recreate the current state from the merged one, including ψ_j and $s(\cdot)$ of the merged composition. Because ψ_j depends only on the pre-split state, which the merge recovers exactly, the ratio is valid for any choice of ψ_j.

Boundary Gibbs draws. For two internal boundaries chosen at random, the boundary between blocks i and $i+1$, with combined length $T = n_i + n_{i+1}$ starting at position a, is drawn from its full conditional

$$p(c \mid \cdot) \;\propto\; w_\sigma(c)\, w_\sigma(T - c)\, \mathcal{N}(\mathbf{y}_{a+1:a+c}; \boldsymbol{\phi}_i)\, \mathcal{N}(\mathbf{y}_{a+c+1:a+T}; \boldsymbol{\phi}_{i+1}), \qquad c = m_{\min}, \ldots, T - m_{\min},$$

with the hyperparameters of both blocks held fixed. Unlike a uniform boundary shuffle, this move never proposes an implausible position, and it is cheap: the prefix likelihoods of block i and the suffix likelihoods of block $i+1$ each need a single Cholesky factorisation, so every candidate position is evaluated at the cost of one Metropolis proposal.

C.2 Block B: kernel hyperparameter update

Conditional on the current composition, we select three blocks at random and update $\log \ell_k$ and then $\log s_k^2$ of each by a random-walk Metropolis step,

$$a_k' = a_k + \lambda_k \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1), \qquad \lambda_k = \lambda \sqrt{100/n_k},$$

accepted with probability

$$\alpha = \min\Big\{1,\; \frac{\varphi_\ell(a_k')\, L_k(e^{a_k'}, s_k^2)}{\varphi_\ell(a_k)\, L_k(e^{a_k}, s_k^2)}\Big\}, \tag{18}$$

and likewise for b_k. Written with densities on ℓ_k the ratio carries the Jacobian factor ℓ_k'/ℓ_k. The scale λ is adapted towards a 35% acceptance rate during burn-in and fixed afterwards, so the retained chain is a valid Markov chain. The dominant cost is the $|S_k| \times |S_k|$ Cholesky factorisation required by L_k.

C.3 Block C: Pitman–Yor and noise updates

The discount σ is updated by a random-walk Metropolis step on the logit scale (proposal SD 0.35) and the strength θ on the log scale (proposal SD 0.4), ten times each per iteration. The acceptance ratios are

$$\frac{p(\sigma')\, \sigma'(1-\sigma')\, \Pr(\rho \mid \sigma', \theta)\, Z_n(\sigma, \theta)}{p(\sigma)\, \sigma(1-\sigma)\, \Pr(\rho \mid \sigma, \theta)\, Z_n(\sigma', \theta)}, \qquad \frac{p(\theta')\, \theta'\, \Pr(\rho \mid \sigma, \theta')\, Z_n(\sigma, \theta)}{p(\theta)\, \theta\, \Pr(\rho \mid \sigma, \theta)\, Z_n(\sigma, \theta')},$$

where $\sigma'(1-\sigma')/[\sigma(1-\sigma)]$ and θ'/θ are the Jacobians of the logit and log transformations. They involve only the partition prior and the hyperpriors, so no Cholesky factorisations are needed, but they do involve the truncation constant Z_n of Appendix A.

When the noise is estimated, one random walk on $\log \sigma_{\mathrm{n}}$ per iteration (step adapted during burn-in) is accepted with the ratio of $\prod_k L_k$ times the normal prior density of $\log \sigma_{\mathrm{n}}$; it recomputes every block likelihood.

C.4 Block D: bookkeeping

To prevent numerical drift in the per-block log-marginals, the full set of likelihoods is recomputed from scratch every 200 iterations. We discard the first half of each chain as burn-in and thin the retained chain every five iterations.

D Derivation of the sampler

This appendix derives each update of Appendix C from the detailed-balance condition and identifies, for each, the term whose omission changes the distribution the chain targets. For every such term we state the target that results without it; Table 10 summarises them, and Appendix H.1 reports the tests that confirm the sampler targets (11).

D.1 Target in log coordinates

We work with $a_k = \log \ell_k$ and $b_k = \log s_k^2$, so the hyperparameter priors are normal densities $\varphi_\ell(a) = \mathcal{N}(a; \mu_\ell, s_\ell^2)$ and $\varphi_s(b) = \mathcal{N}(b; \mu_s, s_s^2)$, *including* their normalising constants $(2\pi s_\ell^2)^{-1/2}$ and $(2\pi s_s^2)^{-1/2}$. In these coordinates the posterior (11), with σ_{n} fixed, is

$$\pi(\rho, a_{1:K}, b_{1:K}, \sigma, \theta \mid \mathbf{y}) \;\propto\; p(\sigma)\, p(\theta)\, \frac{\Pr(\rho \mid \sigma, \theta)}{Z_n(\sigma, \theta)} \prod_{k=1}^{K} \varphi_\ell(a_k)\, \varphi_s(b_k)\, L_k(e^{a_k}, e^{b_k}), \tag{19}$$

on the union over K of spaces of dimension $2K + 2$, with ρ restricted to compositions with blocks of at least $m_{\min}$ observations. Each update below leaves (19) invariant. The splits and merges connect every composition to the single-block one, so the chain is irreducible on the partitions, and the rejections make it aperiodic.

D.2 Split and merge

The reversible-jump condition. A split maps the state (ρ, x) with $x = (a_{1:K}, b_{1:K})$ to (ρ', x') of dimension $2(K + 1)$ by drawing auxiliary variables $u \sim g$ and applying a deterministic map $x' = T(x, u)$; the merge applies T^{-1} and draws nothing. Detailed balance between the two moves holds when T is a bijection, the dimensions match ($\dim x + \dim u = \dim x'$), and the split is accepted with probability $\min\{1, A\}$ [Green, 1995],

$$A = \frac{\pi(\rho', x')}{\pi(\rho, x)}\, \frac{q_{\mathrm{merge}}(\rho \mid \rho')}{q_{\mathrm{split}}(\rho' \mid \rho)}\, \frac{1}{g(u)}\, \left| \frac{\partial x'}{\partial (x, u)} \right|, \tag{20}$$

and the merge with probability $\min\{1, A^{-1}\}$, with A evaluated at the split that would undo it.

Dimension matching. A split replaces (a_j, b_j) by (a_L, b_L, a_R, b_R), so it needs exactly two auxiliary variables. With $w_L = c/n_j$ and $w_R = 1 - w_L$,

$$a_L = a_j + w_R u_\ell, \qquad a_R = a_j - w_L u_\ell, \qquad b_L = b_j + w_R u_s, \qquad b_R = b_j - w_L u_s.$$

Proposition 1. *The map $(a_j, u_\ell) \mapsto (a_L, a_R)$ is a bijection of $\mathbb{R}^2$ with inverse $a_j = w_L a_L + w_R a_R$, $u_\ell = a_L - a_R$, and its Jacobian determinant has absolute value one. The same holds for (b_j, u_s).*

Proof. $w_L a_L + w_R a_R = a_j + (w_L w_R - w_R w_L) u_\ell = a_j$ and $a_L - a_R = (w_R + w_L) u_\ell = u_\ell$, so the stated map is the inverse. The Jacobian matrix is $\begin{pmatrix} 1 & w_R \\ 1 & -w_L \end{pmatrix}$, whose determinant is $-(w_L + w_R) = -1$. □

The merged block therefore receives the size-weighted geometric means

$$\ell_j = \ell_L^{w_L} \ell_R^{w_R}, \qquad s_j^2 = (s_L^2)^{w_L} (s_R^2)^{w_R},$$

and the merge recovers u exactly. A construction that perturbs each child independently around the parent draws four variables for a move that adds two. Its split and merge are then not inverse maps, (20) has no valid Jacobian, and the chain does not target the posterior.

Natural scale. Written in (ℓ, s^2), the hyperpriors are log-normal, $p(\ell) = \varphi_\ell(\log \ell)/\ell$, and the Jacobian of the split is $\ell_L \ell_R/\ell_j \cdot s_L^2 s_R^2/s_j^2$. The two forms agree:

$$\frac{p(\ell_L)\, p(\ell_R)}{p(\ell_j)} \cdot \frac{\ell_L \ell_R}{\ell_j} = \frac{\varphi_\ell(a_L)\, \varphi_\ell(a_R)}{\varphi_\ell(a_j)}.$$

Prior ratio. From (4), a split increases K by one, multiplies $\prod_{i<K}(\theta + i\sigma)$ by $\theta + K\sigma$, divides by $K+1$ through $K!$, and replaces $w_\sigma(n_j)$ by $w_\sigma(c)\, w_\sigma(n_j - c)$. The truncation constant $Z_n(\sigma, \theta)$ is the same on both sides and cancels. The hyperparameter prior ratio is $\varphi_\ell(a_L)\varphi_\ell(a_R)\varphi_s(b_L)\varphi_s(b_R)/\{\varphi_\ell(a_j)\varphi_s(b_j)\}$, which has one more density factor in a and one in b in the numerator. The normalising constants therefore do *not* cancel.

Proposition 2. *If φ_ℓ, φ_s in the split–merge ratio are replaced by unnormalised Gaussian kernels, the chain targets* (19) *multiplied by* $(2\pi s_\ell s_s)^K$.

Proof. Each split multiplies the ratio by $\sqrt{2\pi}\, s_\ell \cdot \sqrt{2\pi}\, s_s$ and each merge divides it by the same factor, so the modified ratio is the correct ratio for a target that carries this factor once per block. □

With $s_\ell = 0.6$ and $s_s = 0.8$ the factor is 3.02^K, a strong bias towards more regimes that does not show up in fixed-dimension moves, where the constants cancel.

Proposal ratio. The split chooses the move with probability q_K, a splittable block with probability $1/s(\rho)$ and the split point with probability $\psi_j(c)$; the merge chooses the move with probability $1 - q_{K+1}$ and one of the K boundaries of ρ' uniformly. With $g(u) = \mathcal{N}(u_\ell; 0, \tau^2)\mathcal{N}(u_s; 0, \tau^2)$ and unit Jacobian, (20) gives the ratio A of Appendix C.1.

The informed split proposal is valid. ψ_j depends on the parent block's data, its hyperparameters ϕ_j and σ, all of which belong to the pre-split state. The merge reconstructs that state exactly (Proposition 1), so the probability of the reverse split can be evaluated at the merge. Detailed balance therefore holds for any ψ_j that gives positive probability to every $c \in \{m_{\min}, \ldots, n_j - m_{\min}\}$, which the uniform component guarantees. The choice of ψ_j and of τ affects only efficiency: $\pi_j(c)$ proposes split points where the data support them, and $\tau = 0.5$ lets the children's hyperparameters differ substantially, whereas very small τ makes most splits indistinguishable from the parent.

D.3 Boundary moves

Holding K, every other boundary and all hyperparameters fixed, the boundary between blocks i and $i+1$ (combined length T, starting after position a) can take the values $c \in \{m_{\min}, \ldots, T - m_{\min}\}$. In (19), $n!$, $(\theta+1)_{n-1}$, $\prod_{i<K}(\theta + i\sigma)/K!$, Z_n, the hyperpriors and the likelihoods of all other blocks do not depend on c. The full conditional is therefore

$$\pi(c \mid \text{rest}) \;\propto\; w_\sigma(c)\, w_\sigma(T-c)\, \mathcal{N}(\mathbf{y}_{a+1:a+c}; \boldsymbol{\phi}_i)\, \mathcal{N}(\mathbf{y}_{a+c+1:a+T}; \boldsymbol{\phi}_{i+1}),$$

a discrete distribution with at most T atoms that can be sampled exactly, so the move is a Gibbs update with no rejection. A uniform shuffle targets the same conditional but proposes most positions where it is negligible, so for long blocks nearly all its proposals are rejected. The Gibbs draw costs about as much as one shuffle proposal because of the following property.

Proposition 3. *Let $\Sigma = LL^\top$ be the $m \times m$ covariance of a block with equally spaced inputs and a stationary kernel, $\mathbf{v} = L^{-1}\mathbf{y}$ and $\tilde{\mathbf{v}} = L^{-1}J\mathbf{y}$, with J the reversal permutation. Then for every $c \le m$,*

$$\log \mathcal{N}(\mathbf{y}_{1:c}; \mathbf{0}, \Sigma_{1:c,1:c}) = \sum_{i \le c} \Big(-\tfrac{1}{2} v_i^2 - \log L_{ii} - \tfrac{1}{2} \log 2\pi \Big),$$

and the same sum with $\tilde{\mathbf{v}}$ gives the log-likelihood of the suffix $\mathbf{y}_{m-c+1:m}$.

Proof. The Cholesky factor of a leading principal sub-matrix is the leading block of L, and the leading block of L^{-1} is the inverse of that block, so the first c entries of $\mathbf{v}$ are the whitened prefix. Σ is symmetric Toeplitz, hence persymmetric ($J\Sigma J = \Sigma$), so the reversed suffix has the covariance of a prefix of the same length. □

One factorisation for block i and one for block $i+1$ thus give the likelihood of every candidate position. The same property gives $\pi_j(c)$ for the informed split from a single factorisation of the parent.

D.4 Kernel hyperparameters

The random walk proposes $a'_k = a_k + \lambda_k \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$, which is symmetric in a. Metropolis–Hastings with the target density in a gives the ratio $\varphi_\ell(a'_k) L_k(e^{a'_k}, s_k^2) / \{\varphi_\ell(a_k) L_k(e^{a_k}, s_k^2)\}$. In terms of the log-normal density on ℓ, the same ratio is $p(\ell'_k) L_k(\ell'_k)\, \ell'_k / \{p(\ell_k) L_k(\ell_k)\, \ell_k\}$.

Proposition 4. *If the random walk on $a = \log \ell$ is accepted with the natural-scale ratio*

$$\frac{p(\ell'_k)\, L_k(\ell'_k)}{p(\ell_k)\, L_k(\ell_k)},$$

without the factor ℓ'_k/ℓ_k, the chain targets a prior on $\log \ell_k$ shifted from $\mathcal{N}(\mu_\ell, s_\ell^2)$ to $\mathcal{N}(\mu_\ell - s_\ell^2, s_\ell^2)$, and likewise for $\log s_k^2$.

Proof. That ratio is the correct one for the density $p(e^a)L(e^a) = \varphi_\ell(a) e^{-a} L(e^a)$ in a, and $\varphi_\ell(a) e^{-a} \propto \exp\{-(a - \mu_\ell + s_\ell^2)^2 / (2 s_\ell^2)\}$. □

With our priors the shifts are -0.36 for $\log \ell_k$ and -0.64 for $\log s_k^2$: shorter length-scales and smaller amplitudes than the prior states.

The step $\lambda_k = \lambda \sqrt{100/n_k}$ depends on the block length, which this update does not change, so the proposal remains symmetric given ρ. The scale λ is adapted towards a 35% acceptance rate only during burn-in. After burn-in the transition kernel is fixed, so the retained draws come from a time-homogeneous Markov chain with stationary distribution (19), and no conditions on vanishing adaptation are needed.

D.5 Pitman–Yor parameters and the truncation constant

Given ρ, the conditional of (σ, θ) is proportional to $p(\sigma)\, p(\theta) \Pr(\rho \mid \sigma, \theta) / Z_n(\sigma, \theta)$. The random walks act on $z = \operatorname{logit} \sigma$ and $\log \theta$, whose densities carry the Jacobians $\sigma(1-\sigma)$ and θ. This gives the ratios of Appendix C.3. Unlike in the split and merge, Z_n depends on the parameters being updated and does not cancel.

Proposition 5. *A sampler that omits Z_n from the updates of (σ, θ) targets the posterior under the hyperprior $p(\sigma)\, p(\theta)\, Z_n(\sigma, \theta)$, normalised.*

Proof. Omitting Z_n multiplies the conditional density of (σ, θ) by $Z_n(\sigma, \theta)$. This is the same as replacing $p(\sigma)p(\theta)$ by $p(\sigma)p(\theta)Z_n(\sigma, \theta)$ in (19), and no other update depends on it. □

$Z_n(\sigma, \theta)$ is the prior probability that every block has at least $m_{\min}$ observations. It falls quickly as θ grows, because larger θ produces more and smaller blocks. For $\sigma = 0$ the block counts of each size are asymptotically independent Poisson variables with means θ/j [Arratia et al., 2003], so $Z_n \to \exp\{-\theta \sum_{j < m_{\min}} 1/j\} = e^{-1.5\theta}$ for $m_{\min} = 3$. Under our Beta$(1,1) \times$ Gamma$(5, 0.5)$ hyperprior, the tilted prior has mean $\sigma \approx 0.04$ and mean $\theta \approx 2.2$ for both $n = 900$ and $n = 2{,}264$, instead of 0.5 and 10. It pulls both parameters towards the Dirichlet-process limit regardless of the data. In the simulation study, omitting Z_n lowers the average posterior mean of σ from 0.121 and 0.526 (Section 7.1) to about 0.03 and 0.06, so the adaptation of σ to the segment sizes is largely masked. Appendix A gives the exact computation of Z_n.

D.6 Observation noise

When σ_{n} is estimated, the prior is normal on $\log \sigma_{\mathrm{n}}$ and the random walk is symmetric on that scale. The acceptance ratio is therefore the ratio of $\prod_k L_k$ times the ratio of the normal prior densities, with no Jacobian. Since σ_{n} enters every block covariance, every block likelihood is recomputed.

D.7 Summary

Table 10: Terms each update of the sampler must include, the distribution targeted without them, and how the sampler was checked. The last two rows concern efficiency, not correctness.

Term	Without it	Check
Dimension-matched split with exact inverse merge (Prop. 1)	Split and merge are not inverse maps; the reversible-jump ratio is undefined	Prior-only runs
Normalising constants of the hyperpriors (Prop. 2)	Target multiplied by $(2\pi s_\ell s_s)^K \approx 3^K$	Prior-only law of K
Log-scale Jacobian of the random walk (Prop. 4)	Prior of $\log \ell_k$ and $\log s_k^2$ shifted by $-s^2$ (-0.36, -0.64)	Prior-only marginals
Truncation constant Z_n in the (σ, θ) updates (Prop. 5)	Hyperprior tilted to mean $\sigma \approx 0.04$, $\theta \approx 2.2$	Prior-only KS tests; exact sampler
Adaptation during burn-in only	Retained chain is not a fixed Markov kernel	—
Gibbs draw of the boundaries (Prop. 3)	Uniform shuffles, mostly rejected	ESS (Appendix H)
Informed split proposal and $\tau = 0.5$	Uninformed or near-identical splits, mostly rejected	ESS (Appendix H)

The prior-only runs of Appendix H.1 switch the likelihood off, so the chain must reproduce the prior exactly. A missing term in the first four rows shows up there as a distorted law of K or of the hyperparameters; with all terms included, the total-variation distance from the exact law of K is 0.014. With the likelihood on, the exact grid sampler of Appendix I, which uses no MCMC moves at all, agrees with the chains on the posterior of K and on the boundary probabilities.

E Implementation details of benchmark methods

This appendix specifies the implementation used for each benchmark method in Sections 7 and 8. BOCPD, HDP-SLDS and laGP's in-sample fit are those of the original study of this model, and their results are reproduced from it. Martinez–Mena is refitted for this paper with the settings of Martínez and Mena [2014]. The standard-GP fits and forecasts were computed for this paper. Each method is applied to the same observation series $\mathbf{y}$ as TPGPMM unless otherwise stated.

E.1 BOCPD: Bayesian online change-point detection with a Normal-Inverse-Gamma conjugate underlying predictive model

The underlying predictive model of this baseline is a Normal–Inverse-Gamma conjugate model rather than a Gaussian process, so we refer to it as Bayesian online change-point detection [BOCPD; Adams and MacKay, 2007]. The pipeline follows the "Industry Portfolios" specification of Saatçi et al. [2010], with implementation details as in Turner [2011]:

Preprocessing (VIX only). Daily close-to-close returns $r_t = \mathrm{VIX}_t/\mathrm{VIX}_{t-1} - 1$ are first standardised using the sample mean $\widehat{\mu}$ and standard deviation $\widehat{\sigma}$ of the first $T_{\text{train}} = 250$ observations, $z_t = (r_t - \widehat{\mu})/\widehat{\sigma}$. Each standardised return is then Gaussianised by applying the probability-integral transform under a Student-$t(\nu)$ distribution with $\nu = 4$ degrees of freedom, followed by the inverse standard-normal cumulative distribution function (CDF): $\widetilde{z}_t = \Phi^{-1}\big(F_{t_4}(z_t)\big)$. This nonlinear warping maps the heavy-tailed marginal of financial returns into approximate standard normality, ensuring that the Gaussian assumptions of the conjugate model are satisfied. For the simulation study, the synthetic series is already Gaussian by construction and is fed to the model unchanged; no preprocessing is applied.

Within-segment model. Within each run, the (Gaussianised) observations are modelled as $\widetilde{z}_t \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$, with a Normal-Inverse-Gamma conjugate prior $\mu \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$ and $\sigma^{-2} \sim \text{Gamma}(\alpha_0, \beta_0)$. We fix $\mu_0 = 0$, $\kappa_0 = 1$, $\alpha_0 = 1$, $\beta_0 = 1$ throughout. Marginalising both μ and σ^2, the one-step-ahead predictive distribution is Student-t with degrees of freedom $\text{df}_n = 2\alpha_n$, location μ_n, and scale $\beta_n(\kappa_n + 1)/(\alpha_n \kappa_n)$, where the posterior parameters update analytically from the within-run sufficient statistics.

Run-length recursion. We use the standard forward recursion of Adams and MacKay [2007] with a constant hazard $H = 1/\lambda$ and $\lambda = 250$, corresponding to a geometric prior on inter-changepoint intervals with an expected gap of one trading year. The run-length posterior is truncated at $R_{\max} = 1500$, and hypotheses with probability below 10^{-4} are pruned at each step.

Change-point detection. We declare a change-point at time t when the recent-changepoint score $P(r_t \leq 5 \mid x_{1:t})$ exceeds a threshold of 0.5, with a cooldown of 5 trading days suppressing duplicate detections within the same event. A short warmup of $k + 1 = 6$ steps is applied so the cumulative score $P(r_t \leq k)$ is well-defined.

E.2 Sticky HDP-SLDS

We implement the Markov-switching stochastic volatility (MSSV) variant of the sticky HDP-SLDS of Fox et al. [2011]. The latent log-volatility process follows

$$x_t = \phi\, x_{t-1} + \mu^{(z_t)} + e_t, \qquad e_t \sim \mathcal{N}(0, \sigma_\eta^2),$$

with mode-specific process-noise mean $\mu^{(z_t)}$, shared dynamic coefficient ϕ, and a discrete mode sequence $z_t \in \{1, \ldots, L\}$. The observation model is $y_t = x_t + v_t$, where v_t is approximated by a 10-component mixture-of-Gaussians fit to the $\log \chi^2(1)$ density. The mode sequence follows a sticky hierarchical Dirichlet process hidden Markov model (HDP-HMM) prior with truncation level $L = 15$ and concentration hyperpriors $\alpha + \kappa \sim \text{Gamma}(1, 0.01)$, $\gamma \sim \text{Gamma}(1, 0.01)$, $\rho = \kappa/(\alpha + \kappa) \sim \text{Beta}(10, 1)$. The dynamic-parameter prior is $\phi \sim \mathcal{N}(0, 1)$, $\mu_k \sim \mathcal{N}(0, 1)$, $\sigma_\eta^2 \sim \text{IG}(3, 0.2)$.

Inference. A blocked Gibbs sampler alternates between (i) forward-filtering / backward-sampling of $x_{1:T}$ given $z_{1:T}$ and the parameters, (ii) sampling the mixture-component indicators of the measurement noise, (iii) sampling $(\phi, \boldsymbol{\mu}, \sigma_\eta^2)$ given $x_{1:T}$ and $z_{1:T}$, (iv) block sampling $z_{1:T}$ via forward-backward over the truncated HDP transition distributions, and (v) resampling the HDP transition rows and the global stick-breaking weights $\boldsymbol{\beta}$. The simulation study uses $n_{\text{iter}} = 1{,}500$ Gibbs iterations with a burn-in of 500; the VIX application uses $n_{\text{iter}} = 3{,}000$ with a burn-in of 1,000. After burn-in, we accumulate the per-time change-point score $\widehat{p}_t = \frac{1}{M} \sum_m \mathbf{1}\{z_t^{(m)} \neq z_{t-1}^{(m)}\}$ and, when needed, the posterior-predictive moments $E[y_t \mid \mathbf{y}]$ and $\text{Var}[y_t \mid \mathbf{y}]$.

E.3 Martinez–Mena: Pitman–Yor–OU change-point detection

We replicate the change-point procedure of Martínez and Mena [2014]. The method shares the Pitman–Yor-induced prior on ordered contiguous compositions used by TPGPMM (Section 3), but combines it with a markedly different within-block model and is fit through a dedicated split–merge–shuffle MCMC sampler. We evaluate two specifications from the original paper.

Composition prior. The prior on the segmentation $\rho = (n_1, \ldots, n_K)$ is the same Pitman–Yor-induced ordered-composition prior derived in Appendix A, with discount σ and strength θ.

Within-block likelihood. Conditional on the partition, observations within each block follow a discretely-observed Ornstein–Uhlenbeck process,

$$y_t = \mu_k + \phi\,(y_{t-1} - \mu_k) + \xi_t, \qquad \xi_t \sim \mathcal{N}\big(0,\, \lambda_k^{-1}(1 - \phi^2)\big), \qquad t \in S_k,$$

with regime-specific mean μ_k and precision λ_k, and a shared autocorrelation parameter ϕ. The within-block parameters are assigned the conjugate Normal–Gamma prior $\mu_k \mid \lambda_k \sim \mathcal{N}\big(0,\, (c\,\lambda_k)^{-1}\big)$, $\lambda_k \sim \text{Gamma}(a, b)$,

and integrated out analytically, yielding a closed-form per-block marginal likelihood $L_k(\phi) = p(\mathbf{y}_{S_k} \mid \phi, a, b, c)$ given by Eq. 9 of Martínez and Mena [2014]. We use $a = b = 1$ and $c = 0.1$ throughout, as in the paper.

Inference. A split–merge–shuffle MCMC sampler iterates over (i) a structural move on ρ (split, merge, or boundary shuffle, operating on the integrated within-block likelihood directly, with no within-block hyperparameters to perturb), (ii) random-walk Metropolis updates of the autocorrelation parameter ϕ on the logit scale when it is free, and (iii) random-walk updates of the Pitman–Yor strength θ on the log scale. The discount σ is held fixed across both specifications, following Martínez and Mena [2014].

Specifications. We evaluate the two specifications of the original paper's application [Martínez and Mena, 2014, Section 6], with its settings. **Model A** treats observations as Markovian within each block, with $\phi \sim \mathrm{U}(0, 1)$. **Model B** treats them as independent and identically distributed within each block ($\phi = 0$ held fixed). In both models the discount is fixed at $\sigma = 0.9$ and the strength has the shifted Gamma prior $\theta \sim \mathrm{Ga}(1, 1, -0.9)$, that is, $\theta + 0.9 \sim \mathrm{Gamma}(1, 1)$; $a = b = 1$ and $c = 0.1$; and we run 20,000 iterations after a burn-in of 10,000, for the VIX and Treasury series and for each simulated dataset.

Preprocessing. For VIX, $y_t = \log(\mathrm{VIX}_t)$ and the series is centred on its sample mean before fitting, so the Normal–Gamma prior with prior mean $\mu_0 = 0$ is well-calibrated. For the simulation study, the synthetic series y_t is also centred on its sample mean; no further transform is applied.

Change-point detection. We declare a change-point at index t when the marginal posterior probability that t is an internal boundary, $\widehat{\tau}_t^{\mathrm{MM}} = \frac{1}{M^\star} \sum_m \mathbf{1}\{t \text{ is an internal boundary of } \rho^{(m)}\}$, exceeds the threshold 0.5, with the same 5-step cooldown used by BOCPD and HDP-SLDS.

In-sample predictive. For each retained draw, (μ_k, λ_k) of every regime has the Normal–Gamma posterior given the regime's data and the draw's ϕ, and the predictive of y_t is Student-t; the draws are mixed as for TPGPMM. For Model B ($\phi = 0$) the predictive of every y_t in a regime is that of a new observation from the regime. Model A has no observation noise, so a predictive that conditions on y_t would be degenerate; instead y_t is predicted one step ahead from y_{t-1}, with mean $(1 - \phi)\,\mathrm{E}[\mu_k] + \phi\, y_{t-1}$, and the first observation of a regime from the regime's stationary distribution. With $\phi = 0$ the two coincide.

E.4 Gaussian-process benchmarks

Standard GP. A single stationary GP on the whole training series, with RBF or Matérn-3/2 kernel and hyperparameters (length-scale, signal variance, noise) set by maximising the marginal likelihood (ML-II). It is used for the in-sample scores and for the forecasts. The in-sample predictive conditions on the whole training series; for forecasting it conditions on the last 150 observations and is centred on their mean.

laGP. The local approximate GP of Gramacy and Apley [2015] (R package `laGP`), with a local design for each prediction point that starts from its 10 nearest neighbours and is extended to 60 observations by the ALC-ray criterion, length-scale and nugget by maximum likelihood, and predictions at the training inputs.

TPGPMM forecasts. For each of 500 posterior draws, the forecast is the GP predictive of that draw's final regime, conditioning on at most the last 150 observations of the regime; specifications without a regime mean are centred on the mean of that window. The draws are averaged, and scores use the mean and variance of the resulting mixture.

F Definitions of evaluation metrics

This appendix defines the metrics used in Sections 7 and 8. Throughout, $\mathbf{y} = (y_1, \ldots, y_n)$ is the observed series, and a method's posterior predictive at index t is summarised by its mean μ_t and variance ν_t.

F.1 Posterior-predictive metrics

Root mean squared error (RMSE).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \mu_t)^2}. \tag{21}$$

Same units as y. Sensitive only to the predictive mean, not the spread.

Negative log predictive density (NLPD). Under a Gaussian predictive $\mathcal{N}(\mu_t, \nu_t)$,

$$\mathrm{NLPD} = \frac{1}{n}\sum_{t=1}^{n}\left[\frac{1}{2}\log(2\pi\nu_t) + \frac{(y_t - \mu_t)^2}{2\nu_t}\right]. \tag{22}$$

Lower is better. NLPD penalises both location error and miscalibration of the predictive variance.

Continuous ranked probability score (CRPS). We use the closed-form Gaussian-approximation CRPS [Gneiting and Raftery, 2007],

$$\mathrm{CRPS}_t = \sigma_t\left[z_t\big(2\Phi(z_t) - 1\big) + 2\varphi(z_t) - \frac{1}{\sqrt{\pi}}\right], \tag{23}$$

with $\sigma_t = \sqrt{\nu_t}$, $z_t = (y_t - \mu_t)/\sigma_t$, and Φ, φ the standard-normal CDF and probability density function (PDF). The reported CRPS is the average $\frac{1}{n}\sum_t \mathrm{CRPS}_t$. Lower is better. Like NLPD, CRPS is a strictly proper scoring rule that rewards both accurate point predictions and well-calibrated uncertainty, but unlike NLPD it remains finite under heavy-tailed errors. For TPGPMM, μ_t and ν_t are the mean and variance of the mixture (16).

Scale-normalised variants. Because TPGPMM, laGP, and HDP-SLDS operate on different targets, we report scale-normalised versions to make magnitudes comparable across rows of Tables 5 and 6. Let $s_y = \mathrm{sd}(y)$ denote the empirical standard deviation of the target. Then

$$\mathrm{RMSE/sd} = \frac{\mathrm{RMSE}}{s_y}, \qquad \mathrm{CRPS/sd} = \frac{\mathrm{CRPS}}{s_y}, \tag{24}$$

$$\Delta\mathrm{NLPD} = \mathrm{NLPD} - \left[\tfrac{1}{2}\log(2\pi s_y^2) + \tfrac{1}{2}\right]. \tag{25}$$

The bracketed term is the NLPD that a marginal-Gaussian baseline $\mathcal{N}(\overline{y}, s_y^2)$ would obtain on its own data. Negative values of ΔNLPD therefore indicate that the model improves on the trivial constant-Gaussian fit; RMSE/sd and CRPS/sd less than 1 indicate improvements over a constant-mean predictor. For the out-of-sample forecasts, s_y is the standard deviation of the 2024 test window.

F.2 Change-point detection metrics

Let $\mathcal{T}^\star = \{t_1^\star, \ldots, t_{K-1}^\star\}$ be the set of true internal change-points and $\widehat{\mathcal{T}} = \{\widehat{t}_1, \ldots, \widehat{t}_{\widehat{K}-1}\}$ the set of inferred change-points. We say that a true change-point $t_j^\star$ is *matched* if there exists $\widehat{t}_i \in \widehat{\mathcal{T}}$ with $|\widehat{t}_i - t_j^\star| \le \tau$, where the matching tolerance τ is set to 5 trading days throughout. Each true change-point is matched to at most one inferred change-point and vice versa, in a greedy nearest-neighbour fashion, yielding n_{matched} matched pairs.

Inferred regime count. $\widehat{K} = |\widehat{\mathcal{T}}| + 1$, the pruned number of inferred regimes after applying any method-specific merge radius. For TPGPMM, $\widehat{\mathcal{T}}$ is the set of change-points of the modal composition.

Precision, recall, F_1.

$$\text{Precision} = \frac{n_{\text{matched}}}{|\widehat{\mathcal{T}}|}, \qquad \text{Recall} = \frac{n_{\text{matched}}}{|\mathcal{T}^\star|}, \qquad F_1 = \frac{2\,\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{26}$$

Precision penalises spurious detections; recall penalises missed change-points. $F_1 \in [0,1]$ is their harmonic mean and is the standard one-number summary because precision and recall are in tension—a method can score perfect recall by predicting change-points everywhere, at the cost of precision.

Change-point mean absolute error.

$$\text{CP_MAE} = \frac{1}{|\mathcal{T}^\star|}\sum_{j=1}^{|\mathcal{T}^\star|}\min_{\widehat{t}\in\widehat{\mathcal{T}}}\left|\widehat{t}-t_j^\star\right|. \tag{27}$$

The localisation error of the closest inferred change-point to each true change-point, averaged across true change-points. Reported in trading-day units. By convention CP_MAE $= 0$ when both sets are empty, and undefined (NaN) when $\widehat{\mathcal{T}} = \emptyset$ but $\mathcal{T}^\star \neq \emptyset$.

F.3 TPGPMM-specific summaries

For the simulation study we additionally summarise the marginal posteriors of the Pitman–Yor parameters from the TPGPMM chains. Given the retained samples $\{(\sigma^{(m)}, \theta^{(m)})\}_{m=1}^{M^\star}$,

$$\overline{\sigma} = \frac{1}{M^\star}\sum_{m=1}^{M^\star}\sigma^{(m)}, \qquad \text{CI}_{95}(\sigma) = \left[\sigma^{(0.025)},\, \sigma^{(0.975)}\right], \tag{28}$$

with $\sigma^{(q)}$ the q-th empirical quantile of the chain. The reported σ posterior mean in the body text is seed-averaged: for a batch with R replicate seeds, we report $\frac{1}{R}\sum_r \overline{\sigma}_r$. The same convention applies to θ and σ_{n} when reported.

G Simulation-study configuration

This appendix gives the exact specification used to generate the synthetic scenarios in Section 7.1, sufficient to reproduce all reported TPGPMM numbers from the code release.

Data-generating process. For each block k, the latent function is sampled from a zero-mean Gaussian process with squared-exponential kernel of length-scale ℓ_k and signal variance s_k^2. The block-specific mean m_k is added, and observations are drawn as $y_t = f(X_t) + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \sigma_n^2)$ and $\sigma_n = 0.1$. The inputs are integer indices $X_t = t - 1$. Continuity at boundaries is not enforced.

The data-generating process includes an explicit per-block mean m_k in addition to the within-block kernel hyperparameters (ℓ_k, s_k^2), while the fitted TPGPMM has only the kernel hyperparameters as block-specific parameters. This is deliberate: the simulation tests whether TPGPMM recovers regime structure when regimes differ in mean as well as in covariance, a setting strictly more general than the model's prior. The chain has no parametric mean to fit, so the mean shift is absorbed by the within-block GP posterior; recovery of the true partitions in Section 7.1 indicates this absorption is effective for the level scales and segment lengths considered. The data are generated with a squared-exponential kernel and $\sigma_n = 0.1$, so TPGPMM uses the data-generating kernel family and noise level and differs from the data-generating process only in the block means. The Matérn-3/2 alternative with estimated noise is misspecified in the kernel and has to learn the noise, which makes it the more demanding test.

Per-block parameters. The two scenarios share identical per-block kernels and means; only the block lengths differ:

Block k	ℓ_k	s_k^2	m_k
1	25.0	0.85	0.0
2	8.0	0.70	1.8
3	6.0	0.75	−1.7
4	10.0	0.70	1.6
5	7.0	0.70	−1.8
6	12.0	0.65	1.7
7	9.0	0.75	−1.5
8	11.0	0.70	1.9
9	8.0	0.70	−1.6
10	7.0	0.70	1.7

Block lengths.

- *Homogeneous*: $(n_1, \ldots, n_{10}) = (90, \ldots, 90)$, total $n = 900$.
- *Power-law*: $(n_1, \ldots, n_{10}) = (810, 10, 10, 10, 10, 10, 10, 10, 10, 10)$, total $n = 900$.

Replication and seeds. Each scenario is replicated $R = 50$ times. For replication $r = 1, \ldots, R$, the data-generating seed is $s_{\text{data}}^{(r)} = 1000 + r$ and the MCMC seed is $s_{\text{model}}^{(r)} = 8800 + r$, for both scenarios. Each replication is fitted with one chain of 15,000 iterations, the first half discarded. Martinez–Mena A and B are fitted on the same datasets, with the same seeds. The BOCPD and HDP-SLDS results in Table 1 are taken from the original study, which used the same design with data seeds 1000–1049; for those two baselines the comparison is therefore between draws from the same data-generating process, not on identical datasets.

Evaluation. Change-point matches are declared within $\tau = 5$ trading days. Adjacent inferred change-points within a merge radius of 4 days are collapsed before computing precision, recall, F_1, and CP_MAE.

H Sampler diagnostics

To assess sampler mixing over the partition space, we report acceptance rates for the split and merge moves, the fraction of boundary Gibbs draws that move a boundary, effective sample sizes for the global quantities K, σ, and θ, and the rank-normalised split-$\widehat{R}$ across four chains started from very different partitions. Detailed numbers are collected in Table 11; the traces of the global quantities for the VIX chains are shown in Figure 7.

Table 11: Sampler diagnostics. Four chains per series started at $K_0 \in \{1, 5, 15, 30\}$; acceptance rates are averaged over chains; ESS is the rank-normalised bulk effective sample size pooled over chains and $\widehat{R}$ the rank-normalised split-$\widehat{R}$ [Vehtari et al., 2021], reported as the maximum over K, σ, θ; statistics on the second half of each chain.

Series	Specification	$\text{acc}_{\text{split}}$	$\text{acc}_{\text{merge}}$	$\text{moved}_{\text{Gibbs}}$	ESS_K	ESS_σ	ESS_θ	$\widehat{R}$ (max)	min/chain
VIX	TPGPMM	0.055	0.055	0.439	143	7336	3013	1.031	8.3
VIX	TPGPMM (σ_{n} estimated)	0.065	0.064	0.381	12	14083	51	1.266	5.3
VIX	TPGPMM (Matérn-3/2, σ_{n} estimated)	0.061	0.061	0.354	544	3452	10183	1.006	50.0
VIX	TPGPMM (Matérn-3/2, regime mean, σ_{n} estimated)	0.147	0.145	0.452	88	8470	280	1.047	6.5
Treasury	TPGPMM	0.126	0.127	0.780	387	1762	4441	1.011	13.8
Treasury	TPGPMM (σ_{n} estimated)	0.095	0.091	0.539	26	4604	158	1.139	2.0
Treasury	TPGPMM (Matérn-3/2, σ_{n} estimated)	0.115	0.115	0.667	147	1655	4148	1.023	18.1
Treasury	TPGPMM (Matérn-3/2, regime mean, σ_{n} estimated)	0.062	0.062	0.597	96	2952	1417	1.031	6.7

For TPGPMM the four chains agree: the largest $\widehat{R}$ over K, σ, θ is 1.031 on VIX and 1.011 on the Treasury yield, and the effective sample size of K is 143 and 387. The same holds for the Matérn-3/2 and regime-mean alternatives. The RBF model with estimated noise is the exception ($\widehat{R} = 1.27$ on VIX and 1.14 on the Treasury yield, with $\text{ESS}_K = 12$ on VIX): its chains settle at different combinations of the noise level and the

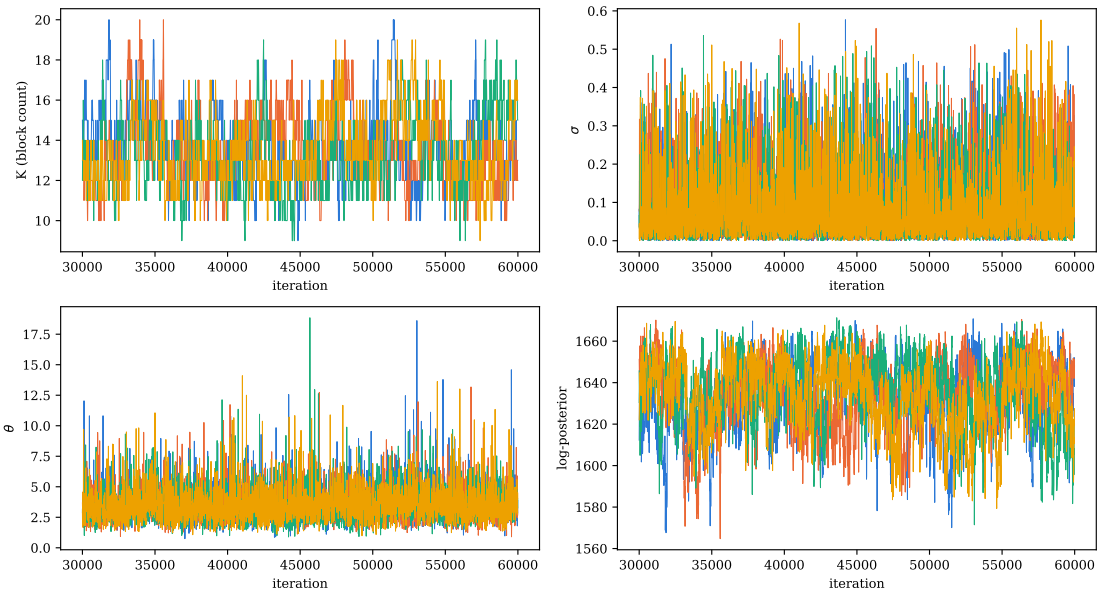


Figure 7: Post-burn-in traces of the global quantities (K, σ, θ, and the log-posterior) for the four VIX chains of TPGPMM, started from $K_0 = 1, 5, 15, 30$ equal blocks (first half of the 60,000 iterations discarded as burn-in). The four chains overlap in every panel.

number of regimes. This is the weak separability of within-regime amplitude and noise that motivates fixing σ_n (Section 4).

For comparison, a basic split–merge sampler for TPGPMM, with uniform split points, a uniform boundary shuffle and hyperparameter perturbations of scale 0.05 at a split, gives $\mathrm{ESS}_K = 5$ and split-$\widehat{R}_K = 1.51$ for a 20,000-iteration VIX chain. Three features account for the difference: the exact Gibbs draws of the boundaries, the informed split proposal, and the dimension-matched split with auxiliary variables of scale $\tau = 0.5$.

H.1 Verification of the sampler

Prior-only runs. With the likelihood switched off ($n = 20$, noise estimated), four chains of 60,000 iterations reproduce the prior: the total-variation distance between the sampled and the exact law of K is 0.014, and the Kolmogorov–Smirnov distances of σ and θ from their priors are 0.019 and 0.002. A sampler that omits Z_n gives mean $\theta \approx 3.0$ and mean $\sigma \approx 0.15$ in the same setting, against the prior means 10 and 0.5. Block likelihoods from the Toeplitz prefix computation match a direct multivariate normal density to relative error 10^{-13} for all kernels, with and without the regime mean.

Exact sampler. With the likelihood switched on, Appendix I compares the chains with an exact sampler that needs no MCMC moves. The two agree on the posterior of the number of regimes, the boundary probabilities and the Pitman–Yor parameters.

I Validation against an exact sampler

For TPGPMM's main specification, with an RBF kernel and fixed noise, the posterior of the partition can be computed without MCMC once the kernel hyperparameters are restricted to a grid. Its draws are independent, so it has no convergence issue, and its only approximation is the grid. It therefore provides a reference against which to check that the chains of Appendix C converge to the right distribution.

Discretised hyperprior. We replace the log-normal priors on (ℓ_k, s_k^2) by a discrete prior on a regular grid $\{(a_g, b_g)\}_{g=1}^{G}$ of $(\log \ell, \log s^2)$, covering $\mu_\ell \pm 3.3 s_\ell$ and $[\mu_s - 3s_s, \mu_s + 4s_s]$, with weights $\pi_g \propto \varphi_\ell(a_g)\varphi_s(b_g)$. Because the block hyperparameters are independent across blocks, they can be summed out block by block. Each block $y_{s+1:t}$ then contributes the collapsed likelihood

$$S(s,t) = \sum_{g=1}^{G} \pi_g \, \mathcal{N}\big(\mathbf{y}_{s+1:t}; \mathbf{0}, \Sigma_g[1{:}t-s,\, 1{:}t-s]\big),$$

and the posterior of (ρ, σ, θ) is

$$\pi(\rho, \sigma, \theta \mid \mathbf{y}) \;\propto\; p(\sigma)\, p(\theta)\, \frac{c_K(\sigma,\theta)}{Z_n(\sigma,\theta)} \prod_{k=1}^{K} w_\sigma(n_k)\, S(i_{k-1}, i_k), \qquad c_K(\sigma,\theta) = \frac{n!}{(\theta+1)_{n-1}} \, \frac{\prod_{i=1}^{K-1}(\theta + i\sigma)}{K!}.$$

As the grid is refined, this converges to the posterior of the continuous model (19).

All block likelihoods at once. Every block covariance at grid point g is a leading sub-matrix of the same $n \times n$ Toeplitz matrix Σ_g. Let L_g be its Cholesky factor and H the Hankel matrix whose column s holds $(y_{s+1}, \dots, y_n, 0, \dots, 0)$. By Proposition 3, the cumulative sums down each column of the matrix with entries

$$-\tfrac{1}{2}\big[(L_g^{-1}H)_{rs}\big]^2 - \log(L_g)_{rr} - \tfrac{1}{2}\log 2\pi$$

are the log-likelihoods of all blocks (s,t). One factorisation and one triangular solve per grid point therefore give the whole matrix S.

Forward recursion and backward sampling. The prior is a product of block terms times a function of K (Appendix A), so the recursion carries the number of blocks:

$$F_1(t) = w_\sigma(t)\, S(0,t), \qquad F_K(t) = \sum_{s:\, t-s \geq m_{\min}} F_{K-1}(s)\, w_\sigma(t-s)\, S(s,t),$$

and

$$p(\mathbf{y} \mid \sigma, \theta) = \frac{1}{Z_n(\sigma,\theta)} \sum_{K} c_K(\sigma,\theta)\, F_K(n).$$

We place (σ, θ) on a grid of 24 values of $\sigma \in [0, 0.9]$ and 120 log-spaced values of $\theta \in [0.05, 60]$, weighted by their priors. Each independent draw takes three steps [Fearnhead, 2006]:

1. draw (σ, θ) from its posterior on the grid;
2. draw K with probability proportional to $c_K(\sigma,\theta)F_K(n)$;
3. draw the boundaries backwards. The last block starts at s with probability

$$\Pr(s) \;\propto\; F_{K-1}(s)\, w_\sigma(n-s)\, S(s,n),$$

   and the earlier boundaries follow in the same way.

We use 4,000 draws and three grids, with $G = 210$, 783 and 1,760 points. On 12 cores the block likelihoods take 7–51 seconds and a complete run two to three minutes.

Results. Table 12 and Figure 8 compare the exact posterior with TPGPMM's four chains.

- **Number of regimes.** On VIX, the exact posterior mean of K is 13.48 on the 783-point grid and 13.69 on the finest grid, against 13.44 for the chains, whose Monte Carlo standard error is 0.15. The total-variation distance between the posteriors of K from the finest grid and from the chains is 0.049. On the Treasury yield the two means are 6.95 and 7.00 (standard error 0.07), with a total-variation distance of 0.024.

- **Boundary locations.** The probabilities of a boundary within ±10 trading days of each date agree closely along both series: the mean absolute difference is 0.007 on VIX and 0.002 on the Treasury yield, and the largest is 0.14 and 0.04.
- **Pitman–Yor parameters.** The posterior means of σ and θ are 0.082 and 3.58 (exact) against 0.090 and 3.56 (chains) on VIX, and 0.367 and 5.16 against 0.375 and 5.26 on the Treasury yield.
- **Effect of the grid.** The coarsest grid, with 210 points, overstates the number of regimes on VIX (14.43). The two finer grids differ from each other by about as much as they differ from the chains, so the remaining gap is of the order of the discretisation error. On the Treasury yield all three grids agree with the chains.

The chains therefore converge to the exact posterior, up to Monte Carlo error and the discretisation of the hyperprior. The check is limited to the main specification. With estimated noise, σ_{n} would add a third grid dimension, and the cost of the block-likelihood matrix grows linearly with the grid size.

Table 12: TPGPMM's chains against the exact grid sampler, which includes the truncation constant. TV_K is the total-variation distance between the two posteriors of K, and $\max|\Delta\tau|$ the largest difference between the probabilities of a boundary within ±10 trading days of each date. Exact results use 4,000 independent draws.

Series	Method	$E[K]$ (95% interval)	$E[\sigma]$	$E[\theta]$	TV_K	$\max\lvert\Delta\tau\rvert$
VIX	Exact, 210-point grid	14.43 [11, 18]	0.072	3.67	0.227	0.24
VIX	Exact, 783-point grid	13.48 [10, 18]	0.083	3.51	0.029	0.15
VIX	Exact, 1,760-point grid	13.69 [11, 18]	0.082	3.58	0.049	0.14
VIX	TPGPMM chains (MCMC)	13.44 [11, 17]	0.090	3.56	—	—
Treasury	Exact, 210-point grid	7.04 [5, 10]	0.359	5.01	0.045	0.09
Treasury	Exact, 783-point grid	6.97 [5, 10]	0.367	5.17	0.030	0.03
Treasury	Exact, 1,760-point grid	6.95 [5, 10]	0.367	5.16	0.024	0.04
Treasury	TPGPMM chains (MCMC)	7.00 [5, 10]	0.375	5.26	—	—

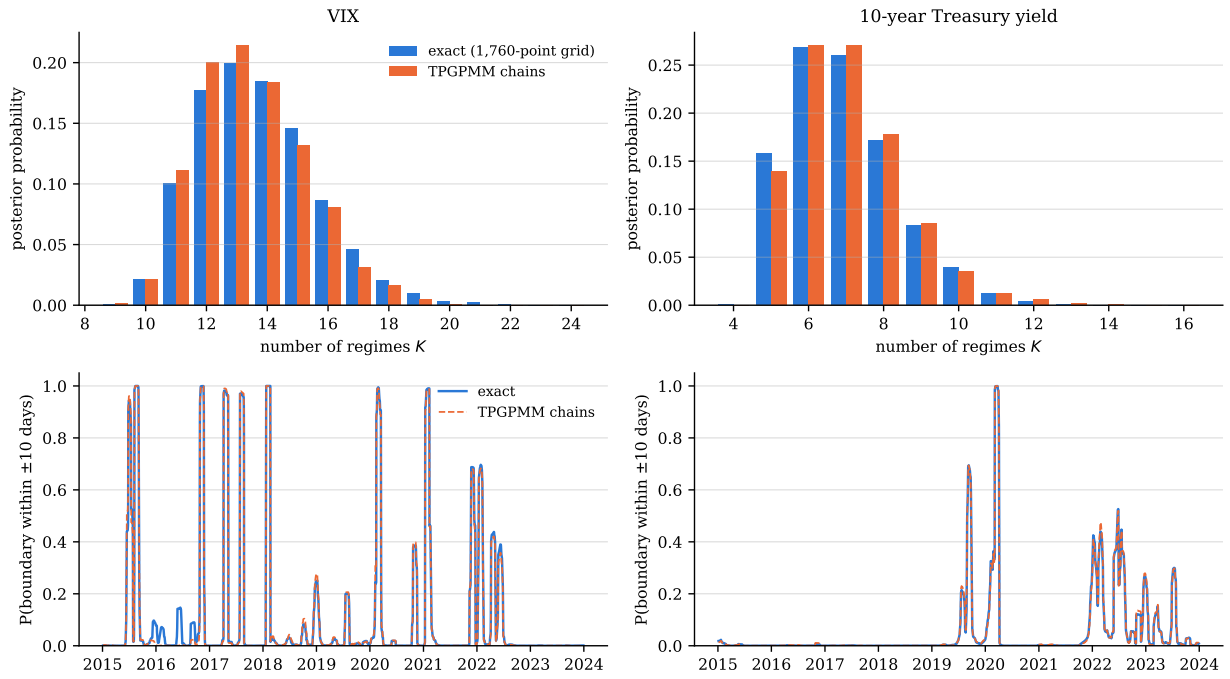


Figure 8: Exact grid sampler against TPGPMM's chains, for VIX (left) and the 10-year Treasury yield (right). *Top*: posterior of the number of regimes K. *Bottom*: posterior probability of a boundary within ± 10 trading days of each date. The exact results use the finest grid and 4,000 independent draws.