3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–13

Marginalised Spectral Mixture Kernels with Nested Sampling

Anonymous Authors
Anonymous Institution

Abstract

Gaussian Process (GPs) models are a rich distribution over functions with inductive biases controlled by a kernel function. Learning occurs through optimisation of the kernel hyperparameters using the marginal likelihood as the objective (ML-II). This work analyses the benefits of marginalising kernel hyperparameters using nested sampling (NS), a technique well-suited to sample from complex, multi-modal distributions. We benchmark against Hamiltonian Monte Carlo (HMC) on time-series regression tasks and find that a principled approach to quantifying hyperparameter uncertainty substantially improves the quality of prediction intervals.

1. Introduction

Gaussian processes (GPs) represent a powerful non-parametric and probabilistic framework for performing regression and classification. GPs are typically trained using ML-II, a procedure which denotes the maximisation of the GP marginal likelihood. ML-II yields point estimates of hyperparameters for the user chosen covariance / kernel function. The posterior predictive distribution is then evaluated at the ML-II point estimates. The form of the kernel function influences the geometry of the marginal likelihood surface. For instance, periodic kernels give rise to multiple local minima as functions with different periodicities can be compatible with the data. Expressive kernels which are derived by adding/multiplying together primitive kernels to encode different types of inductive biases typically have many hyperparameters, exacerbating the local minima problem. Concretely, the classical approach to training suffers from two issues we try and analyse in this work: 1) Point estimate hyperparameters give overconfident predictions by not accounting for hyperparameter uncertainty; 2) Non-convexity of the marginal likelihood surface can lead to poor estimates located at local minima. Further, the presence of multiple modes affects the interpretability of kernel hyperparameters.

The spectral mixture (SM) kernel proposed in Wilson and Adams (2013) is an expressive class of kernels derived from the spectral density reparameterisation of the kernel using *Bochner's Therorem* Bochner (1959). The SM kernel has prior support over all stationary kernels which means it can recover sophisticated structure provided sufficient spectral components are used. Several previous works Kom Samo and Roberts (2015); Remes et al. (2017, 2018) have attempted to further enhance the flexibility of spectral mixture kernels, such as the introduction of a time-dependent spectrum. However, we postulate that the key limitation in the SM kernel's performance lies not its stationarity or expressivity, but in the optimisation procedure. We show that the form of the SM kernel gives rise to multiple symmetric modes in the marginal likelihood making optimisation extremely unstable and prone to overfitting. Our two main contributions are:

- Highlight the failure modes of ML-II training in the context of spectral mixture kernels. We provide insights into the effectiveness of ML-II training in weak and strong data regimes.
- Propose the relevance of Nested Sampling (NS) as a way to effectively sample from the hyperparameter posterior in the presence of multiple modes.

Although this work focuses on the SM kernel class, many of the insights and general methodology are applicable to other stationary kernels in the context of Gaussian processes.

© A. Authors.

2. Background

2.1. Marginalised Gaussian Processes

Given some input-output pairs $(X, \boldsymbol{y}) = \{\boldsymbol{x_i}, y_i\}_{i=1}^N$ where y_i are noisy realizations of latent function values f_i with Gaussian noise, $y_i = f_i + \epsilon_i$, $\epsilon_i \in \mathcal{N}(0, \sigma_n^2)$, we seek to infer some as-yet unseen values y^*. Let $k_\theta(\boldsymbol{x_i}, \boldsymbol{x_j})$ denote a positive definite kernel function parameterized with hyperparameters $\boldsymbol{\theta}$. Following the prescription of Lalchand and Rasmussen (2020), the marginalised GP framework is given by

$$\textit{Hyperprior:} \quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}); \quad \textit{GP Prior:} \quad \boldsymbol{f}|X, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, K_\theta); \quad \textit{Likelihood:} \quad \boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma_n^2 \mathbb{I}) \tag{1}$$

where K_θ denotes the $N \times N$ covariance matrix, $(K_\theta)_{i,j} = k_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The predictive distribution for unknown test inputs $X^\star$ integrates over the joint posterior[1],

$$p(\boldsymbol{f}^\star|\boldsymbol{y}) = \iint p(\boldsymbol{f}^\star|\boldsymbol{f}, \boldsymbol{\theta}) p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{\theta} \simeq \frac{1}{M} \sum_{j=1}^{M} p(\boldsymbol{f}^\star|\boldsymbol{y}, \boldsymbol{\theta_j}) \tag{2}$$

where $\boldsymbol{f}$ has been integrated out analytically and $\boldsymbol{\theta}$ is dealt with numerically, yielding the final mixture of Gaussians form of the posterior predictive in marginalised framework. Throughout this work we shall adopt a Gaussian likelihood, hence the only intractable integrand we need to consider is the hyperparameter posterior $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$.

2.2. Spectral Mixture Kernels

Definition 1 (SM kernel) *A spectral mixture kernel is a positive definite stationary kernel given by,*

$$k(\boldsymbol{\tau}) = \sum_{q=1}^{Q} w_q \cos(2\pi \boldsymbol{\tau}^T \boldsymbol{\mu}_q) \prod_{d=1}^{D} \exp(-2\pi^2 \tau_d^2 \sigma_q^{2^{(d)}}) \tag{3}$$

where $\tau_d, \mu_q^{(d)}, \sigma_q^{(d)}$ are the dth components of the D dimensional vectors $\boldsymbol{\tau}, \boldsymbol{\mu}_q$ and $\boldsymbol{\sigma}_q$ respectively. The vector of kernel hyperparameters $\boldsymbol{\theta} = \{w_q, \boldsymbol{\mu}_q, \boldsymbol{\sigma}_q\}_{q=1}^Q$ is typically unknown, we account for this uncertainty by treating them as random variables with priors and sampling from their intractable posterior.

The SM kernel is derived by leveraging *Bochner's Therorem* (Bochner, 1959) which provides an explicit relationship between the spectral density $S(\nu)$ and the kernel function k as the integral $\int S(\nu) e^{2\pi i \nu^T \tau} d\nu$. If the $S(\nu)$ is modelled as a weighted mixture of Q Gaussian pairs,

$$S(\nu) = \sum_{q=1}^{Q} \frac{w_q}{2} \left[G(\nu, \mu_q, \sigma_q) + G(\nu, -\mu_q, \sigma_q) \right] . \tag{4}$$

and plugged into integral mentioned above, yields the SM kernel in eq. 3.[2]Here the weight w_q specifies the variance contributed by the qth component while $G(\nu, \mu, \sigma)$ denotes a Gaussian function with mean μ_q and standard deviation σ_q. To avoid confusion with other standard deviations, and convey its physical significance, we shall refer to σ_q as the *bandwidth*.

1. where we implicitly condition over inputs $X, X^\star$ for compactness.
2. The supplementary of Wilson and Adams (2013) has a full proof.

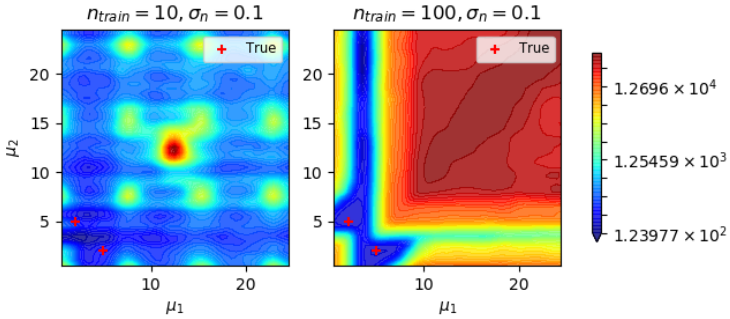

Figure 1: Visualising the negative log marginal likelihood surface as a function of mean frequencies in a 2 component spectral mixture kernel GP. The training data was generated from latent functions evaluated on the input domain [-1,1] and σ_n refers to the intrinsic data noise level. The two identical peaks correspond to different re-orderings of the 2 component mean frequency vector with a true value of [2,5].We can see that even with a 10 fold data increase, the flat ridge on the surface implies that gradient based optimisation might end up trapped in a bad local optima. Further, with small data sets the peaks are not pronounced enough and there is significant probability mass in regions of the hyperparameter space away from the true values.

Fig. 1 shows a two-dimensional slice of the 7-dimensional[3] marginal likelihood surface for a 2-component SM kernel, given two datasets of different size. One of the striking features of these surfaces lies in their symmetry due to the kernel's invariance to the ordering of its components. The marginal likelihood of a SM kernel with Q spectral components possesses $Q!$ identical regions of parameter space. A naive attempt to explore the full posterior distribution of a spectral mixture kernel would try to quantify probability mass across these degenerate regions, a much more computationally intensive task than is necessary. One solution is to only sample from one region, and ignore its symmetric counterparts. To achieve this, we can adopt an approach known as forced identifiability Buscicchio et al. (2019) to ensure that the components are defined in sequential order with respect to their frequency μ.

2.3. Nested Sampling

The nested sampling algorithm was developed by Skilling (2004) (see also Skilling et al. (2006)) as a means of estimating the model evidence $\mathcal{Z} = \int \psi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, where ψ denotes the likelihood [4], and $p(\boldsymbol{\theta})$ the prior. This may be recast as a one-dimensional integral over the unit interval, $\mathcal{Z} = \int_0^1 \psi(X) dX$, irrespective of the dimensionality of $\boldsymbol{\theta}$. Here X is the quantile function associated with the likelihood: it describes the volume of the prior lying below the likelihood value ψ. The extreme values of the integrand, $\psi(X{=}0)$ and $\psi(X{=}1)$, therefore correspond to the minimum and maximum likelihood values found under the support of the prior $\pi(\boldsymbol{\theta})$. Sampling proceeds in accordance with Algorithm 1 (appendix 1). While the sequence of samples provides an estimate of $\mathcal{Z}$, an invaluable quantity in the context of Bayesian model selection, they also represent importance weighted samples of the posterior.

We employ the PolyChord algorithm (Handley et al., 2015; Handley et al., 2015), which performs slice sampling (Neal, 2003) at each iteration. Unless otherwise stated, we use 100 live points, which are bounded in a set of ellipsoids (Feroz et al., 2009). These bounding surfaces allow the macroscopic structure of the likelihood contours to be traced, enabling a much more efficient sampling process. This approach has proven particularly adept at navigating multi-modal likelihood surfaces (Allison and Dunkley, 2014). These attractive properties have motivated numerous scientific applications, including the detection of gravitational waves (Veitch et al., 2015), the categorisation of cosmic rays (Cholis et al., 2015), and the imaging of a supermassive black hole (Akiyama et al., 2019). The

3. A two component SM kernel has two weights, bandwidths and frequencies in additional to the noise level for a 1d dataset.
4. for hyperparameter inference, the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ plays the role of the likelihood, so $\psi(\boldsymbol{\theta}) = e^{\mathcal{L}(\boldsymbol{\theta})}$

Kernel	Spectral	Spectral	Spectral	NKN
Inference	ML-II	HMC	NS	ML-II
airline	7.25	5.83	**5.22**	5.6
births	5.17	5.23	**4.96**	5.42
call centre	11	7.32	**7.29**	7.76
gas production	15.2	12.4	**11**	12.4
internet	**11.3**	11.4	13.2	12.6
mauna	**1.5**	3.32	1.8	3.4
radio	2.19	2.12	**2.07**	4.12
solar	1.4	0.82	**0.58**	2.38
sulphuric	5.13	5.15	5.11	6.32
temperature	2.8	**2.49**	2.52	4.2
unemployment	12.8	11.2	10.6	**8.5**
wages	159	8.71	6.59	**4.28**
wheat	8.47	6.44	6.58	**6.24**
Mean	18.7 ±1.2	6.34 ±0.03	**5.96** ±0.03	6.40

Table 1: NLPD values for various GP methods across a range of time series tasks.

algorithm described in 1 is implemented as part of the dynesty package (Speagle, 2020). We use the 'rslice' sampling option, along with the default number of five slices, and adopt the default criterion for convergence. This is defined as the point at which the estimated posterior mass contained within the set of live points falls below 1% of the total.

3. Experiments

3.1. Ground Truth Recovery

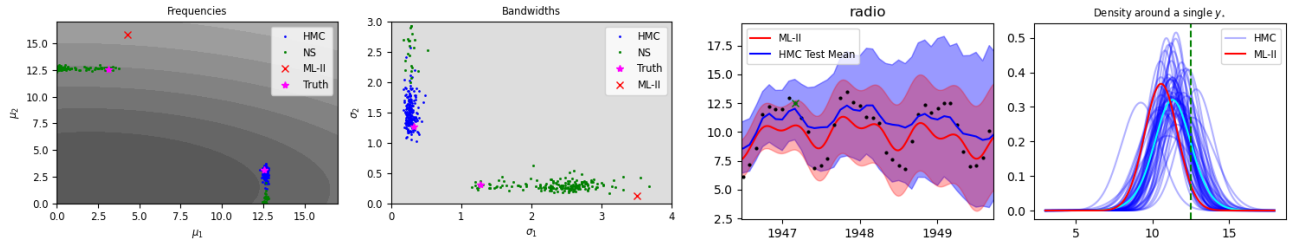

Figure 2: A comparison of hyperparameter estimation, where the ground truth is indicated by the magenta star. The grey shading indicates the prior. *Far left:* Recovering the mean frequency parameters of the two spectral components. Recovering the two bandwidth parameters of the two spectral components. The true hyperparameters are $[\mu_1, \mu_2] = [3.14, 12.56]$ and $[\sigma_1, \sigma_2] = [1.27, 0.32]$. For frequencies and bandwidths we note the symmetry i.e. the estimates can converge on $[\mu_1, \mu_2]$ or $[\mu_2, \mu_1]$, and that the nested sampling algorithm successfully identifies both modes whereas HMC can only sample from one. *Left:* Test means and 95% intervals, the green cross is the test point whose mixture density is enumerated on the right. *Far right:* An illustration of how the predictive density is comprised of a mixture of Gaussians, yielding broader intervals and a correspondingly better NLPD.

In the two left hand panels of Fig.2 we summarise the performance of ML-II, HMC and nested sampling inference in their ability to recover the true setting of the hyperparameters for a synthetic dataset with 100 datapoints, a noise amplitude of $\sigma_n = 0.1$ and a signal-to-noise ratio of ≈ 3.2 on a fixed domain [-1,1]. The magenta star $\star$ denotes the true value and the red cross $\times$ denotes the ML-II estimate. The HMC and NS sampling schemes are both better at recovering the ground truth hyperparameters than the point estimates. Further, the nested sampling scheme is able to obtain samples from both modes inherent in the marginal likelihood.

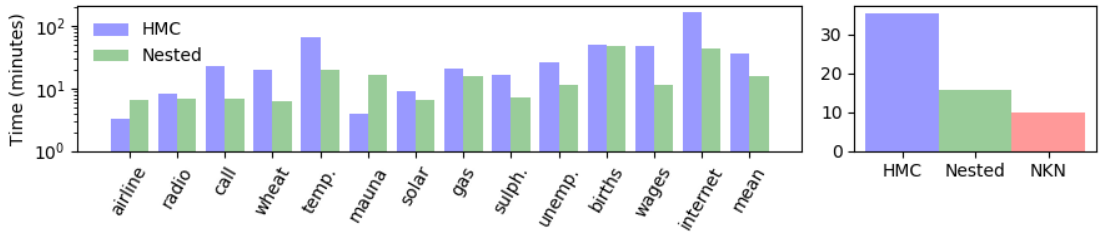

Figure 3: *Left:* Training time for each of the time series benchmarks. *Right:* Average training time (across data sets). It is interesting to note that nested sampling is only marginally more compute intensive than performing conventional ML-II inference with the neural kernel network.

3.2. Time series benchmarks

We evaluate the predictive performance of marginalised GPs with the SM kernel against thirteen time series, as used in Lloyd et al. (2014)[5]. The time series are of variable length, with up to $1,000$ data points in each. Our fiducial kernel has seven components ($Q = 7$), yielding a 22-dimensional hyperparameter space to explore. For reference, we also include results from the Neural Kernel Network (Sun et al., 2018) where a flexible kernel is learnt incrementally through a weighted composition of primitive base kernels. This is trained with the Adam optimiser for 100,000 iterations and a learning rate of 10^{-3}. In Table 3.1 we report the negative log of predictive density (NLPD) across test data. The evaluation was conducted with a 60/40 train/test split. The quoted uncertainties are estimated by repeating each set of experiment three times with a different random seed. We find that the spectral mixture kernel exhibits significant performance gains when using a sampling-based inference method compared to the conventional ML-II approach. HMC and NS offer similar performance levels, though NS carries an advantage of faster evaluation times and an estimate of the model evidence. More detailed results for various configurations of the nested sampler can be found in Table 3.

Fig. 3 depicts training times in wall clock seconds for our time-series experiments, all of which were run on a single Nvidia GTX1070 GPU. For HMC this corresponds to a single chain with 500 warm-up iterations and 500 samples. Nested sampling was performed with 100 live points, and the sampling continued until the convergence criterion was met. Across these benchmarks, HMC was found to be slower by approximately a factor of two compared to nested sampling. Both methods could enjoy significant gains in speed with schemes such as those proposed in Loper et al. (2020).

3.2.1. *Why do wider predictive intervals give better NLPD?*

The right hand panel of Fig. 2 is a zoomed in plot of the 'RADIO' predictions. It helps convey exactly why the NLPDs corresponding to the marginalised GP schemes are better. The plot shows test predictions and respective 95% intervals. The blue band corresponds to prediction intervals under the marginalised HMC scheme; the 95% quantiles are estimated empirically by sampling from the Gaussian mixture distribution at each test input X^*. The far right panel shows the mixture density for a single test point, the predictive mixture form places higher probability mass on the true test point than the narrower ML-II intervals.

3.3. 2d Pattern Extrapolation

To provide a demonstration of how the inference methods are readily adapted to higher dimensional kernels, we revisit the challenge presented in Sun et al [26]. The two-dimensional ground truth function is given by $y = (\cos 2x_1 \times \cos 2x_2)\sqrt{|x_1 x_2|}$. We train with just 50 points (as opposed to 100 in the original reference), which are chosen at random across $[-6, +6]$ in the $xy-$domain. The

5. The raw data is available at https://github.com/jamesrobertlloyd/gpss-research/tree/master/data/tsdlr-renamed

test points are defined on a 20×20 grid. In Figure 4 we visualise the mean values of the posterior predictive distribution from three different inference methods: ML-II, HMC and nested sampling (NS). Visually, the reconstruction under the marginalised GP schemes (HMC / NS) appears to be superior to ML-II. Further, the 95% confidence intervals (not visualised) differ markedly, as is evident from the NLPD values. These are given by 216, 2.56, and 2.62 for ML-II, HMC and NS respectively (*lower* is better). For reference we also trained the neural kernel network (Sun et al [26]), which achieved an NLPD of 3.8. This experiment marks a significant increase in the dimensionality of our parameters space, since in two dimensions each spectral component has five degrees of freedom. Yet the marginalised Gaussian processes comfortably outperform competing methods. ML-II was trained with Adam (learning rate=0.05) with 10 restarts and 2,000 iterations.

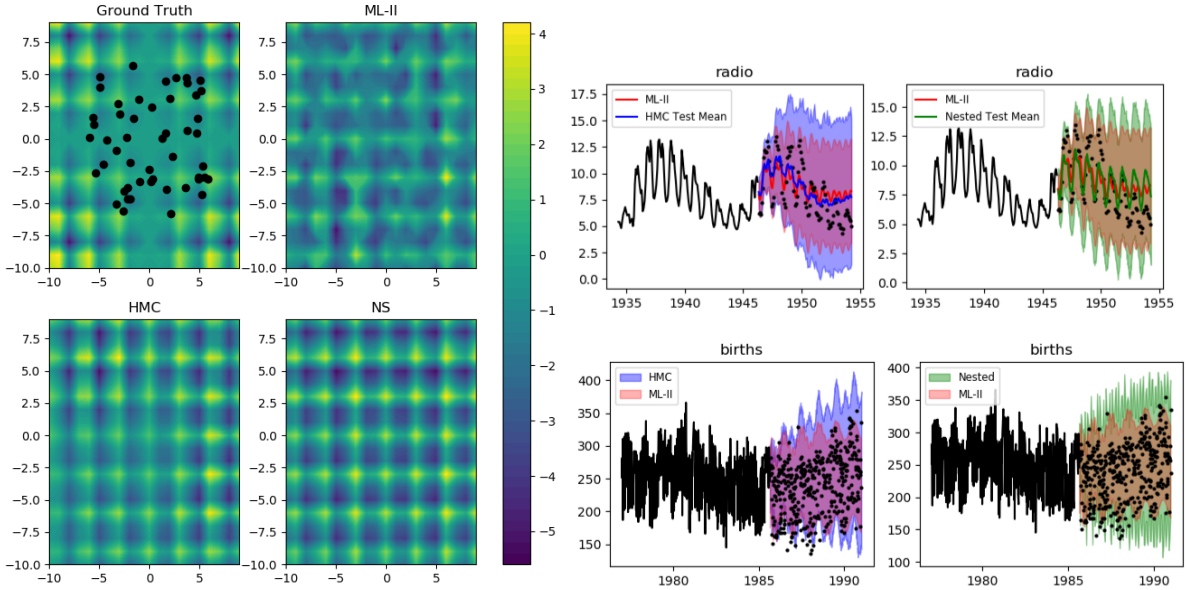

Figure 4: *Left:* Learning a two-dimensional pattern with a 10 component spectral mixture kernel. The black dots denote the locations of the 50 training points. *Top Left:* Posterior predictive mean under ML-II. *Bottom Left:* Posterior predictive mean under HMC and nested sampling (NS). *Right:* An illustration of the systematic underestimation of predictive uncertainty when adopting point estimates of kernel hyperparameters. Here we show 95% test confidence intervals derived from the spectral mixture kernel. Black lines denote training data and the black dots denote test data. 'BIRTHS' data set - the ML-II intervals capture 84% of the test data where HMC and NS capture 93% and 99% of the test points respectively. 'RADIO' data set (also showing test mean prediction) - the three methods show similar performance however, the average predictive density under test data is higher under the sampling methods.

4. Discussion

While a pathological marginal likelihood geometry can pose problems for both gradient based optimisation and sampling; sampling schemes are able to quantify them better if the practical difficulties of deploying them (like chains stuck in bad optima) are overcome. The nested sampling scheme, which does not require gradient information, provides remarkably good approximations to gold-standard HMC samples at a fraction of the cost of running HMC. Further, it is crucial to ask if marginalisation increases expressivity and representational power by accounting for a more diverse range of models than point estimates? A marginalised GP with a standard SM kernel not only improved upon its conventional counterpart, but also surpassed the performance of the Neural Kernel Network model (Sun et al., 2018) in the time series tasks, despite possessing far fewer hyperparameters. This indicates there is merit in the question.

References

Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.

Rupert Allison and Joanna Dunkley. Comparison of sampling techniques for bayesian parameter estimation. *Monthly Notices of the Royal Astronomical Society*, 437(4):3918–3928, 2014.

Salomon Bochner. *Lectures on Fourier integrals*, volume 42. Princeton University Press, 1959.

Riccardo Buscicchio, Elinore Roebber, Janna M. Goldstein, and Christopher J. Moore. Label switching problem in bayesian analysis for gravitational wave astronomy. *Physical Review D*, 100 (8), Oct 2019. ISSN 2470-0029. doi: 10.1103/physrevd.100.084041. URL http://dx.doi.org/10.1103/PhysRevD.100.084041.

Ilias Cholis, Carmelo Evoli, Francesca Calore, Tim Linden, Christoph Weniger, and Dan Hooper. The galactic center gev excess from a series of leptonic cosmic-ray outbursts. *Journal of Cosmology and Astroparticle Physics*, 2015(12):005, 2015.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

F. Feroz, M. P. Hobson, and M. Bridges. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *MNRAS*, 398(4):1601–1614, October 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. polychord: nested sampling for cosmology. *MNRAS*, 450:L61–L65, June 2015. doi: 10.1093/mnrasl/slv047.

WJ Handley, MP Hobson, and AN Lasenby. Polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398, 2015.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Yves-Laurent Kom Samo and Stephen Roberts. Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*, 2015.

Vidhi Lalchand and Carl Edward Rasmussen. Approximate inference for fully bayesian gaussian process regression. volume 118 of *Proceedings of Machine Learning Research*. PMLR, 08 Dec 2020.

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

Jackson Loper, David Blei, John P. Cunningham, and Liam Paninski. General linear-time inference for gaussian processes on one dimension, 2020.

Radford M. Neal. Slice sampling. *Ann. Statist.*, 31(3):705–767, 06 2003. doi: 10.1214/aos/1056562461. URL https://doi.org/10.1214/aos/1056562461.

Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2, 2011.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pages 4642–4651, 2017.

Sami Remes, Markus Heinonen, and Samuel Kaski. Neural non-stationary spectral kernel. *arXiv preprint arXiv:1811.10978*, 2018.

John Skilling. Nested Sampling. In Rainer Fischer, Roland Preuss, and Udo Von Toussaint, editors, *American Institute of Physics Conference Series*, volume 735 of *American Institute of Physics Conference Series*, pages 395–405, November 2004. doi: 10.1063/1.1835238.

John Skilling et al. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4): 833–859, 2006.

Joshua S. Speagle. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *MNRAS*, 493(3):3132–3158, April 2020. doi: 10.1093/mnras/staa278.

Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, and Roger Grosse. Differentiable compositional kernel learning for gaussian processes. *arXiv preprint arXiv:1806.04326*, 2018.

John Veitch, Vivien Raymond, Benjamin Farr, W Farr, Philip Graff, Salvatore Vitale, Ben Aylott, Kent Blackburn, Nelson Christensen, Michael Coughlin, et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library. *Physical Review D*, 91(4):042003, 2015.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

Appendix A. Further Experimental Results

A.1. Synthetic data

As a first demonstration of how the different methods perform, we draw four samples from a two-component SM kernel with known weights, frequencies and bandwidths. For each latent function sample we construct noisy training data across three noise levels $[0.01, 0.1, 0.5]$ and three training sizes $[10, 30, 100]$, on a fixed input domain $[-1, 1]$. In this way we seek to quantify the quality of predictions and prediction uncertainty under weakly identified regimes characterised by very few data points in a fixed domain to strongly identified regimes characterized by a dense distribution of data in a fixed domain. Further, the intrinsic noise level σ_n^2 of the data can also impact inference in weakly and strongly identified data regimes. In order to analyse the impact of σ_n^2 and training size we calculate the average performance across each of three different noise levels for each training set size. We train under each of the candidate inference methods (ML-II, HMC, Nested) for each of the 9×4 data sets created and report prediction performance in terms of the average negative log predictive density (NLPD) for test data in Figure 5. ML-II uses five random restarts with a initialisation protocol tied to the training data. Following protocols from Wilson and Adams (2013), the SM weights (w_i) were initialised to the standard deviation of the targets $\boldsymbol{y}$ scaled by the number of components ($Q = 2$). The SM bandwidths (σ_i) were initialised to points randomly drawn from a truncated Gaussian $|\mathcal{N}(0, \max d(x, x')^2)|$ where $\max d(x, x')$ is the maximum distance between two training points and mean frequencies (μ_i) were drawn from $\text{Unif}(0, \nu_N)$ to bound against degenerate frequencies. HMC used $\text{LogNormal}(0, 2)$ priors for all the hyperparameters. The ML-II experiments used `gpytorch` while the HMC experiments used the NUTS sampler in `pymc3`.

In Figure 5 we notice that ML-II struggles with small training data sets and catastrophically underestimates the noise level. The two sampling methods, HMC and NS, perform comparably well, though it is notable that a single NS run correctly identifies both modes in the likelihood surface, while a single HMC chain can only discover one.

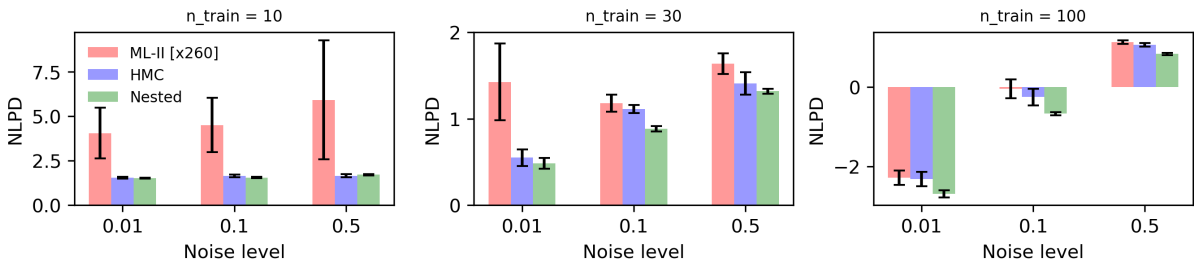

Figure 5: The impact of marginalising hyperparameters on the predictive performance (NLPD) of a spectral mixture kernel. The error bars at each noise level correspond to the mean NLPD and standard error of the mean across a diversity of latent functions. Nested sampling significantly outperforms ML-II in all nine experiments. The subplots contain the performance summaries under the different configurations. *Left:* Sparse distribution of training data ($n_train = 10$). *Middle:* Moderate sized distribution of training data ($n_train = 30$). *Right:* Dense distribution of training data ($n_train = 100$) - all on a fixed domain [-1,+1].

Fig.6 shows the test mean squared error across the three inference schemes. The sampling schemes largely dominate the ML-II method when the hyperparameters are well identified ($n_train = 100$). The data generating configurations are the same as the ones described in the main paper.

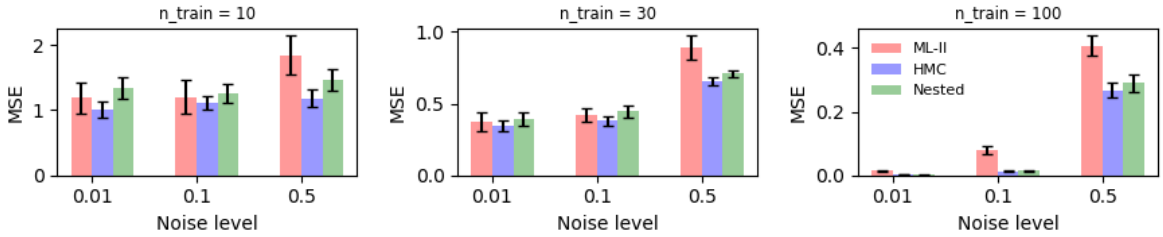

Figure 6: Mean-squared error for synthetic data sets under different noise levels and training set sizes.

A.2. Ground Truth Recovery

In this subsection we summarise the performance of ML-II, HMC and Nested sampling inference in recovering the true setting of the hyperparameters under two different noise settings, for a fixed training size $n_train = 100$ on a fixed domain [-1,1]. The top row of panels in Fig.7 indicate a low noise setting $\sigma_n = 0.01$ and the bottom row indicates a higher noise setting of $\sigma_n = 0.1$. The magenta star $\star$ denotes the true value and the red cross $\times$ denotes the ML-II estimate. The sampling schemes HMC and Nested are both better at recovering the target than the point estimate. While ML-II estimates the noise-level to a high precision when the noise is low (top-row, $\sigma_n = 0.01$), it does not fare so well when noise level is raised by an order of magnitude (bottom row, $\sigma_n = 0.1$). In this case, the estimate of the intrinsic noise level is off by several orders of magnitude. The sampling schemes prove to be far more robust in recovering the frequencies, bandwidths and noise level, especially when operating in the low signal-to-noise regime.

A.3. Time Series

In table 2 we present further results from the benchmark time series experiments. Instead of making full use of the data, we consider only the first 100 points as training data, followed by testing with

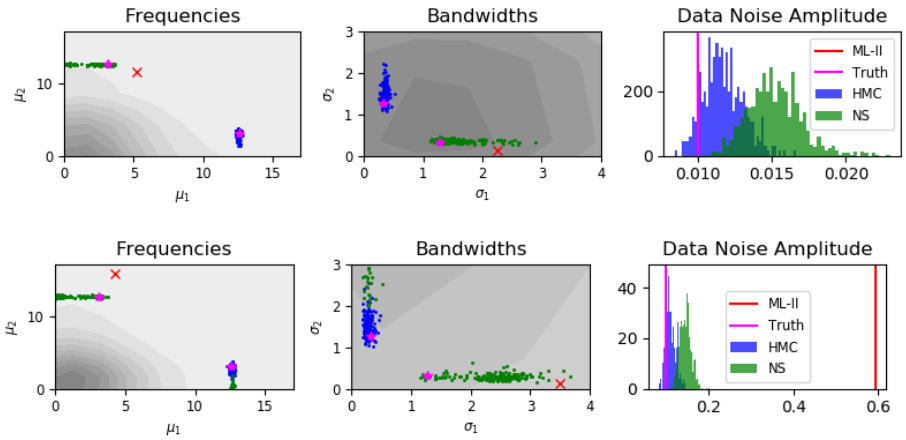

Figure 7: A comparison of hyperparameter estimation, where the ground truth is indicated by the magenta star. The grey shading indicates the prior. *Left:* Recovering the mean frequency parameters of the two spectral components. *Middle:* Recovering the two bandwidth parameters of the two spectral components. *Right:* Recovering the data noise level (σ_n). The true hyperparameters are $[\mu_1, \mu_2] = [3.14, 12.56]$ and $[\sigma_1, \sigma_2] = [1.27, 0.32]$. For frequencies and bandwidths we note the symmetry i.e. the estimates can converge on $[\mu_1, \mu_2]$ or $[\mu_2, \mu_1]$, and that the nested sampling algorithm successfully identifies both.

the subsequent 30 points. As with the results from the full training set, significant performance gains are found when marginalising the hyperparameters of the spectral mixture kernels. However in this case, the nested sampling algorithm doesn't offer a performance advantage over HMC. We speculate this may be due to the simpler likelihood surface associated with the smaller set of training data. Fewer modes in the surface would facilitate exploration via HMC.

Table 2: NLPD values for various GP methods across a range of truncated time series tasks, where only the first 100 points are used for training.

Kernel	Spectral	Spectral	Spectral	NKN
Inference	ML-II	HMC	Nested Sampling	ML-II
airline	**4.90**	5.07	5.36	4.90
births	5.34	**4.66**	4.67	5.29
call centre	8.64	**5.70**	6.54	8.62
gas production	6.15	6.23	**5.97**	6.15
internet	12.5	**11.2**	11.3	12.5
mauna	2.14	1.06	**0.96**	2.87
radio	10.0	2.02	**1.95**	2.27
solar	0.03	**-0.19**	-0.17	0.22
sulphuric	4.36	5.43	**5.42**	**4.57**
temperature	3.39	2.80	**2.65**	2.99
unemployment	8.60	**6.13**	6.31	8.56
wages	**1.67**	1.74	2.19	3.44
wheat	5.86	**4.59**	4.97	12.6
Mean	5.66 ±0.3	**4.34** ±0.08	4.47 ±0.06	5.77

Table 3 shows how the configuration of the nested sampler impacts its performance. The fiducial configuration is shown in the first column, and the uncertainty in the derived mean NLPD is obtained by repeating this experiment for several different random seeds. The three subsequent columns correspond to changes in the number of live points, combining multiple runs, and using a linear prior to describe the frequency parameter μ. The only change which makes a statistically significant alteration to the performance is found to be the choice of the prior.

Table 3: NLPD values for various configurations of the nested sampling algorithm.

	Fiducial	Dense	Multi-run	Prior
Live points	100	500	100	100
Runs	1	1	5	1
Prior	Log	Log	Log	Uniform
airline	5.36	4.66	5.36	8.88
births	4.67	4.68	4.65	4.73
Call centre	6.54	6.17	6.45	6.46
Gas production	5.97	6.38	5.92	8.10
internet	11.3	11.3	11.3	11.4
mauna	0.96	1.15	1.09	3.73
radio	1.95	2.03	1.84	2.19
solar	-0.17	-0.20	-0.19	-0.28
sulphuric	5.42	5.58	5.15	6.07
temperature	2.65	2.64	2.63	2.94
unemployment	6.31	6.21	6.31	6.26
wages	2.19	2.24	2.20	1.97
wheat	4.97	5.20	5.39	5.42
Mean	4.47 ±0.03	4.46 ±0.02	4.47 ±0.03	5.22 ±0.05

Appendix B. Candidate Inference Methods

B.1. ML-II

A Gaussian noise setting facilitates an analytically tractable marginal likelihood,

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{\theta})d\boldsymbol{f} = \int \mathcal{N}(\mathbf{0}, K_\theta)\mathcal{N}(\boldsymbol{f}, \sigma_n^2\mathbb{I})d\boldsymbol{f} = \mathcal{N}(\mathbf{0}, K_{\boldsymbol{\theta}} + \sigma_n^2\mathbb{I}) \tag{5}$$

using the definitions from (1). ML-II estimates are obtained by maximizing $\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$ over the kernel hyperparameters, $\boldsymbol{\theta}_\star = \text{argmax}_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta})$.

B.2. Hamiltonian Monte Carlo (HMC)

Hamiltonian Monte Carlo (Duane et al., 1987) is a fundamental tool for inference in intractable Bayesian models. HMC relies on using gradient information to suppress random walk behaviour inherent to samplers like Metropolis-Hastings and its variants. The sampler operates in an augmented space $(\boldsymbol{\theta}, \boldsymbol{p})$ of position variables and momentum variables where the position variables are the variable of interest. The momentum variables are sampled from an independent $\mathcal{N}(0, 1)$ and can safely be marginalised out from the joint samples to obtain just the samples of interest. New proposals are generated by simulating Hamiltonian dynamics in the joint 2D phase space $(\boldsymbol{\theta}_i, \boldsymbol{p}_i) \to (\boldsymbol{\theta}_{i+1}, \boldsymbol{p}_{i+1})$ (where D is the dimensionality of $\boldsymbol{\theta}$) for preset number of steps called the path length (L). Hamiltonian dynamics are simulated using the leap-frog symplectic integrator which relies on a step size (ϵ). Further, each iteration requires the gradients of the log marginal likelihood w.r.t the hyperparameters.

In the case of a GP, this means each iteration of the leap-frog integrator requires the inversion of K_θ, L times to simulate one proposal. Refer to Neal et al. (2011) for a detailed tutorial. In the experiments we use a self-tuning variant of HMC called the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) where the path length is adapted for every iteration. NUTS is frequently shown to work as well as a hand-tuned HMC; hence in this way we avoid the compute overhead in tuning for good values of the step-size (ϵ) and path length (L). We use the version of NUTS available in the python package `pymc3`.

B.3. Nested Sampling Algorithm

Algorithm 1: Nested Sampling for hyperparameter inference

Initialisation: Draw n_L 'live' points $\{\boldsymbol{\theta}\}_{i=1}^{n_L}$ from the prior $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta})$, set model evidence $\mathcal{Z} = 0$.

while *stopping criterion is unmet* **do**

- Compute $\psi_i = \min(\psi(\boldsymbol{\theta}_1), \ldots \psi(\boldsymbol{\theta}_N))$, the lowest marginal likelihood from the current set of live points.
- Sample a new live point $\boldsymbol{\theta}'$ subject to $\psi(\boldsymbol{\theta}') > \psi_i$
- Remove the point $\boldsymbol{\theta}_i$ corresponding to the lowest marginal likelihood ψ_i, moving it to a set of 'saved' points
- Assign estimated prior mass at this step $\hat{X}_i = e^{-\frac{i}{N}}$
- Assign a weight for the saved point, $V_i = \hat{X}_{i-1} - \hat{X}_i$
- Accumulate evidence, $\mathcal{Z} = \mathcal{Z} + \psi_i V_i$
- Evaluate stopping criterion, if triggered then break;

end

Add final n_L points to the 'saved' list of K samples:

- Weight of each of these final points is assigned to $p_i = \hat{X}_K / n_L \forall i = K, \ldots, n_L + K$ `// final slab of enclosed prior mass`
- Final evidence is given by, $\mathcal{Z} = \sum_{i=1}^{n_L+K} \psi_i V_i$
- Importance weights for each sample are given by, $p_i = \psi_i V_i / \mathcal{Z}$

return *set of samples* $\{\boldsymbol{\theta}_i\}_{i=1}^{n_L+K}$*, along with importance weights* $\{p_i\}_{i=1}^{n_L+K}$ *and evidence estimate* $\mathcal{Z}$*.*

Appendix C. Hyperpriors

In this section we outline a suitable set of priors for the three fundamental parameters of a single spectral component. As defined in (4), these are the mean frequency μ, bandwidth σ, and weight w. For most of these hyperparameters, we find it is sufficient to impose a weakly informative prior of the form $\{\sigma, w, \sigma_n^2\} \sim \text{LogNormal}(0, 2)$. However the spectral component's characteristic frequency μ deserves closer attention.

Two properties of the data strongly influence our perspective on which frequencies we can expect to observe: the fundamental frequency and the highest observable frequency. The fundamental frequency ν_F is the lowest frequency observable within the data, and is given by the inverse of the interval spanned by the observed x locations. Meanwhile the maximum frequency ν_N represents

the highest observable frequency. For gridded data, this is naturally determined by the Nyquist frequency, which is half the sampling frequency.

It is crucial to bear in mind that the spectral density we wish to model is that of the underlying process, and not the spectral density of the data. These two quantities are often very different, due to the limited scope of the observations. For example, the change in stock prices over the period of a single day cannot exhibit frequencies above 10^{-6}Hz. Yet the process will have received contributions from long term fluctuations, such as those due to macroeconomic factors, whose periods can span many years. If we make no assumption regarding the relationship between the process we wish to model and the finite range over which it is observed, then a priori, some as-yet undiscovered frequency within the process ought to be considered equally likely to lie above or below the fundamental frequency. Furthermore, given the large initial uncertainty in frequency μ, it is appropriate to adopt a prior which spans many orders of magnitude.

Towards very low frequencies, $\mu \ll \nu_F$, a sinusoid contributes very little variance to the observations - an annual modulation makes a tiny contribution to a minute's worth of data. As we consider frequencies much lower than the fundamental frequency, it therefore becomes less likely that we will detect their contributions. We model this suppressed probability of observation with a broad Gaussian in log frequency for the regime $\mu < \nu_F$. Meanwhile, at frequencies above the Nyquist frequency, $\mu > \nu_F$, we encounter a degenerate behaviour: these sinusoids are indistinguishable from their counterparts at lower frequencies: they are said to be aliases of each other. As a result of this aliasing behaviour, the likelihood surface is littered with degenerate modes with identical likelihood values. From a computational perspective, it is advantageous to restrict our frequency parameter to a much narrower range than is permitted by our prior, while maintaining the same probability mass. As illustrated in the supplementary, mapping these higher frequencies down to their corresponding alias at $\mu < \nu_N$ yields a uniform prior on μ.

$$\mu/\nu_F \sim \begin{cases} \text{Lognormal}(0, 7), & \text{for } \mu < \nu_F \,, \\ \text{Uniform}(1, \nu_N/\nu_F), & \text{for } \nu_F < \mu < \nu_N \,. \end{cases} \tag{6}$$

C.1. Spectral Priors

Figure 8 shows the parameter space for the frequency and bandwidth of a single spectral component. The likelihood surfaces adjacent to any of the dashed lines are mirror images of each other. It is therefore preferable to avoid exploring multiple copies of these regions when performing nested sampling, as it will attempt to locate the duplicate modes, dispersing the live points.

To give a clearer picture of how the parameters of the spectral mixture kernel are inferred via nested sampling, we take as a example the radio experiment. The posterior distribution in this case can be seen in Figure 9. The corresponding joint posterior distribution of the 22 hyperparameters are displayed in Fig.10. We make use of the dynesty package in order to perform nested sampling. The full set of configuration parameters used can be found in Table 4.

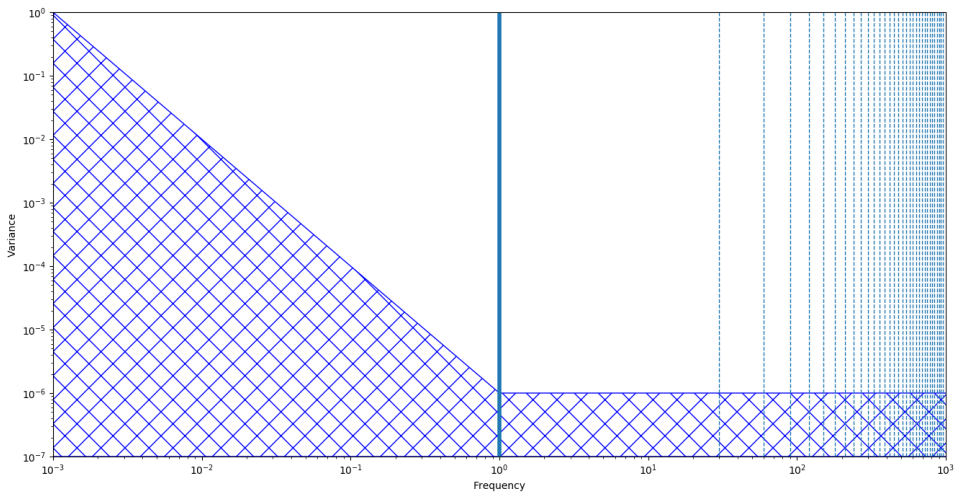

Figure 8: Schematic of the observability of a spectral component as a function of frequency and variance. The fundamental frequency is denoted by the solid vertical line, while dashed vertical lines indicate multiples of the Nyquist frequency. The hatched region denotes the regime where the variance is deemed too low to be observed.

Table 4: A summary of the configuration settings used to perform nested sampling. Most of these adhere to the default set-up in dynesty, with the most significant changes being in the sampling method, and a reduction in the number of live points.

	Fiducial	Default
method	rslice	auto
live points	100	500
Bound	multi	multi
slices	5	5
dlogz	0.01	0.01
max iter	None	None
max call	None	None
min eff	3	10
vol dec	0.5	0.5

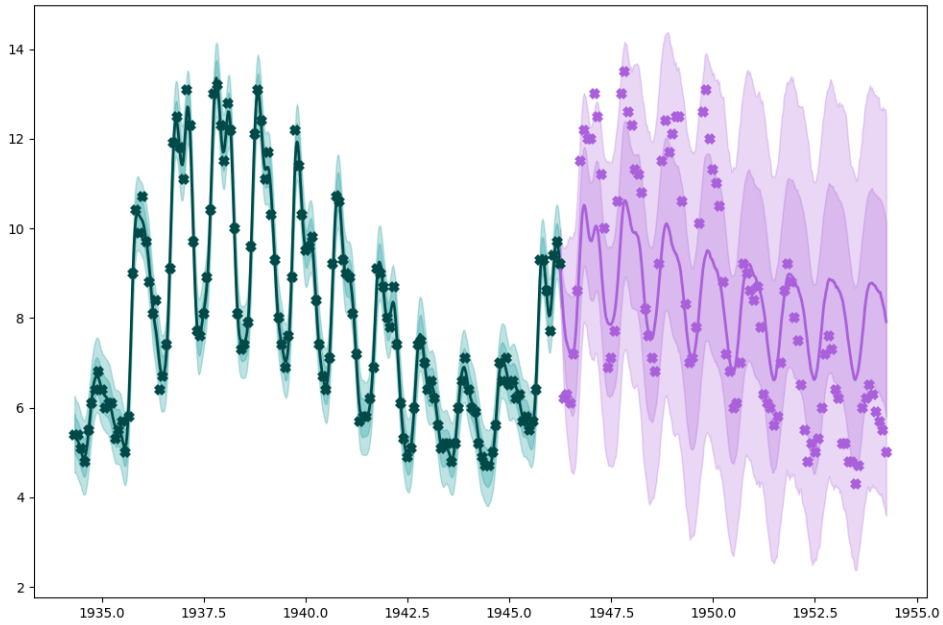

Figure 9: 68 and 95 per cent confidence intervals for the radio dataset, derived from nested sampling of a seven-component spectral mixture kernel.

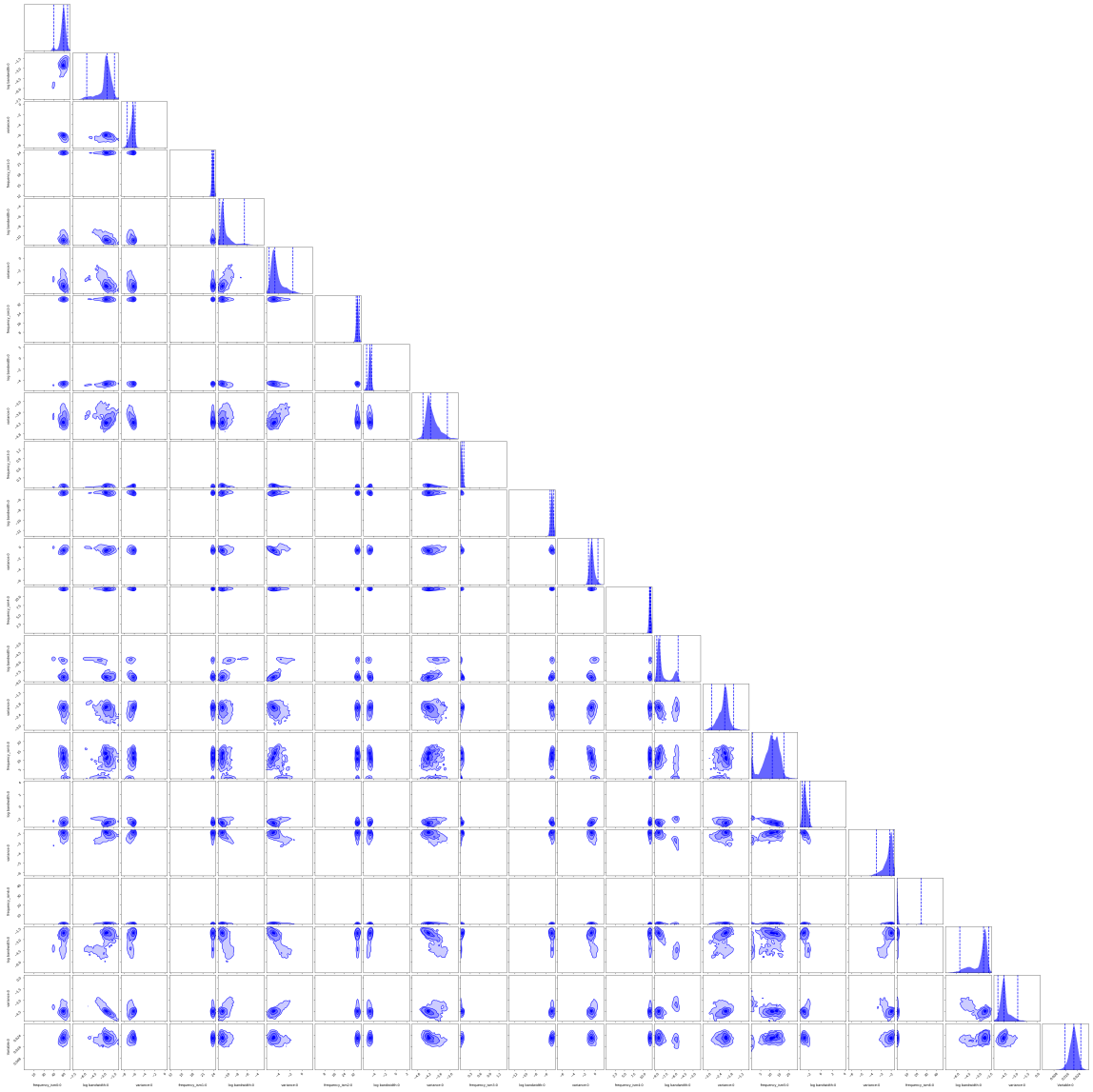

Figure 10: Joint posterior distributions for the 22 hyperparameters associated with Figure 9.