BAYESIAN TECHNIQUES IN THE PULSAR TIMING ARRAY

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Supervisors
Dr. Patricio Maturana-Russel
Dr. Willem van Straten
Prof. Renate Meyer
Prof. Sergei Gulyaev

June 2026

By
EL Mehdi Zahraoui
Department of Mathematical Sciences, School of Engineering, Computer and Mathematical Sciences

Abstract

Pulsar Timing Arrays (PTAs) utilize the extreme rotational stability of millisecond pulsars to detect the nanohertz stochastic gravitational-wave background (SGWB). Robust detection requires rigorous Bayesian modeling to disentangle the SGWB from complex, highly covariant sources like pulsar, solar wind, and systematic noise. However, calculating the Bayesian evidence within the high-dimensional parameter spaces of modern PTA datasets is computationally expensive, creating a bottleneck as array sizes and model complexities grow. This thesis addresses this limitation by developing highly scalable and accurate evidence-estimation techniques. Specifically, it introduces Generalized Steppingstone Sampling (GSS), which utilizes posterior-informed reference distributions to significantly reduce marginal likelihood estimation costs. Building upon this, the thesis presents the Morph approximation—a novel technique that decomposes high-dimensional posterior distributions into lower-dimensional, highly correlated blocks to serve as an optimal proposal distribution for bridge sampling. These methodologies were evaluated across diverse statistical benchmarks and major gravitational-wave datasets. Applied to recent NANOGrav and European PTA datasets, GSS and MorphZ accurately evaluated Bayes factors and reinforced the statistical evidence for an SGWB while minimizing computational overhead by several orders of magnitude. Furthermore, these tools enabled a comprehensive exploration of the noise-model hypothesis space of the Parkes Pulsar Timing Array (PPTA). This analysis revealed that most pulsars exhibit diffuse support for multiple noise models, demonstrated that data reprocessing fundamentally alters noise sensitivity, and established that Bayesian model averaging is necessary to prevent mis-modeling from obscuring true gravitational-wave signatures. Ultimately, this work provides a computationally tractable Bayesian framework that enables the rapid evaluation of broad hypothesis spaces and ensures the accurate characterization of noise-model uncertainty, which is essential for maximizing PTA sensitivity for future discoveries.

Contents

Abstract **2**

Attestation of Authorship **11**

Publications **12**

Acknowledgements **14**

Artificial Intelligence Declaration **16**

Co-Authorship Form **17**

1 Introduction **19**

2 Literature Review **23**
2.1 Pulsars . . . 23
2.1.1 Discovery . . . 23
2.1.2 Radio Observations . . . 24
2.1.3 Properties . . . 26
2.1.4 Millisecond Pulsars . . . 28
2.1.5 Pulsar Timing . . . 30
2.2 Gravitational Waves . . . 33
2.2.1 GW Sources . . . 34
2.2.1.1 Stellar-Mass Compact Binaries . . . 34
2.2.1.2 Supermassive Black-Hole Binaries . . . 35
2.2.1.3 Individually Resolvable Massive Black-Hole Binaries 35
2.2.1.4 Gravitational-Wave Memory Bursts . . . 36
2.2.1.5 Cosmological Phase Transitions . . . 36
2.2.2 Gravitational Wave Background . . . 37
2.2.3 Gravitational-wave Detectors . . . 38
2.2.4 Pulsar Timing Array . . . 40
2.3 Bayesian Inference . . . 42
2.3.1 Bayes Rule . . . 42
2.3.2 Bayesian Modeling . . . 44
2.3.3 Computational Bayesian Inference . . . 46

2.3.4 Bayesian Model Comparison 48
2.4 PTA Bayesian framework 49
2.4.1 PTA Likelihood 49
2.4.1.1 White noise 51
2.4.1.2 Achromatic red noise 51
2.4.1.3 Dispersion-measure variations 52
2.4.1.4 Solar-wind dispersion 53
2.4.1.5 Additional chromatic and phenomenological terms . 54
2.4.1.6 Common-spectrum processes 56
2.4.2 PTA Priors 58
2.4.2.1 Noise model priors 58
2.4.2.2 Common-spectrum priors. 61
2.4.3 PTA Samplers 63
2.4.4 PTA Evidence 65
2.5 Summary and Limitations 68

3 Manuscript 1 **70**
3.1 Introduction 72
3.2 Marginal likelihood estimation 75
3.2.1 Steppingstone Sampling 77
3.2.2 Generalized Steppingstone Sampling 79
3.3 Simulation study 81
3.4 Application of GSS to the PTA 83
3.4.1 Application to NANOGrav 85
3.4.1.1 Model selection for single pulsar analysis 86
3.4.1.2 Evidence estimation for the GWB analysis 89
3.4.2 Application to EPTA+InPTA 90
3.5 Discussion 93

4 Manuscript 2 **96**
4.1 Introduction 98
4.2 Marginal likelihood estimation 102
4.3 Method 103
4.3.1 Morph approximation 104
4.3.2 MorphZ 106
4.4 Application to statistical examples 109
4.5 Application to the Pulsar Timing Array 114
4.6 Application to LVK Compact Binary Coalescences 116
4.7 Discussion 120
4.7.1 Constructing the Morph approximation 120
4.7.2 Evidence estimation using MorphZ 122
4.8 Conclusion 124
4.9 KL divergence between P and $\mathcal{M}_{\mathcal{B}_L}$ in terms of total correlation . . . 126

5 Manuscript 3 **129**
5.1 Introduction . . . 130
5.2 Bayesian inference . . . 134
5.3 Data . . . 135
5.4 Analysis . . . 137
5.5 Results . . . 139
5.5.1 Precision of the marginal-likelihood estimates . . . 139
5.5.2 Summary of PPTA DR2 model space . . . 141
5.5.3 Classification of PPTA DR2 model spaces . . . 144
5.5.3.1 A decisive model space: PSR J1909−3744 . . . 144
5.5.3.2 A competitive model space: PSR J1022+1001 . . . 145
5.5.3.3 Posterior reorganization: PSR J1730−2304 . . . 146
5.5.4 Solar-wind emergence in PPTA DR2 . . . 148
5.6 Conclusion . . . 149
5.7 acknowledgments . . . 151
5.8 Posterior model-probability comparisons . . . 152
5.9 Solar-wind posterior distributions . . . 154

6 Discussion and Conclusion **157**
6.1 Discussion . . . 157
6.2 Future Work . . . 158
6.3 Implications for PTA Sensitivity . . . 159
6.4 Conclusions . . . 160

Glossary **183**

List of Tables

2.1 Typical observed parameter ranges for radio MSPs [Bhattacharyya et al., 2022]. 30

3.1 Mean and standard deviation of 100 independent log-marginal likelihood estimates for NANOGrav single pulsar noise models obtained via GSS with $K = 8$. 85
3.2 Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{IBF}})$ estimates for the parameter inclusion of DMGP model and RN model in NANOGrav single pulsar noise models. 87
3.3 Mean and standard deviation of 100 independent log-marginal likelihood estimates with $K = 16$ for different common red process models in the NANOGrav dataset. 90
3.4 Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{\mathrm{M/CURN}})$ estimates comparing Binned ORF and GWB model to CURN model across EPTA+InPTA datasets. 92
3.5 Mean and standard deviation of 100 log-marginal likelihood estimates for each common red process model applied to the EPTA+InPTA datasets with K = 16. 93

4.1 A comparison of the accuracy as $|\Delta| = |\overline{\log(\hat{z})}_{\mathrm{Method}} - \log(z)_{\mathrm{True}}|$, along the standard deviation of 100 $\log(\hat{z})$, and the efficiency as the number of likelihood calls for each estimate between Nested Sampling, MorphZ, and MSS for the 3 statistical examples. 108
4.2 Comparison of $\log(z)$ estimates using GSS, MorphZ for different PTA models with dimension d. The mean and standard deviation of 100 $\log(z)$ estimates per each model are displayed along the number of likelihood calls per estimate. 112
4.3 Ensemble sumary of $\Delta\log(\hat{z})$ per method relative to $\log(\hat{z})_{\mathrm{NS}}$ over 100 BBH simulations (see Figure 4.4). IQR denotes the 25–75% interval. 116
4.4 $\Delta\log(\hat{z})$ summary per method relative to $\log(\hat{z})_{\mathrm{NS}}$ for the GW150914 event. 119

5.1 Priors definition of the noise components used in the noise models for the PTA DR2 analysis. 135

5.2 Summary of the log marginal-likelihood estimates for the DR2 2020 and DR2 2023 datasets. The median and standard deviation of $\sigma_{\log(\widehat{z})}$ the standard deviation of 5×286 independent evidence estimates. The averaged coefficient of variation $\overline{\mathrm{CV}}_{\log(\widehat{z})}$ across these evidence estimates. 141
5.3 Solar-wind electron density at 1 au, n_{earth} (cm^{-3}), $n_{\mathrm{earth}}^{\mathrm{DR2}}$ from the highest-posterior SW-containing model under the 2023 DR2 analysis and $n_{\mathrm{earth}}^{\mathrm{DR3}}$ the PPTA DR3 analysis [Reardon et al., 2023a]. Posterior medians and central 68% credible intervals are shown; pulsars are sorted by absolute ecliptic latitude $|\beta_e|$ in degrees. 149

List of Figures

2.1 Pen chart recording of PSR B1919+21 pulses, first observed by Jocelyn Bell Burnell. Credit: National Radio Astronomy Observatory. 24
2.2 Pulsar catalog plotted from ATNF Pulsar Database showcasing regular pulsars in green and MSPs in cyan. Adapted from [Hobbs et al., 2004]. 27
2.3 A diagram of the pulsar timing pipeline from signal reception to the construction of timing residuals used in PTA analyses. Adapted from [Verbiest and Shaifullah, 2018]. 31
2.4 Representative sensitivity of PTA collaborations compared with other GW instruments such as advanced LIGO and future LISA. Figure generated with `https://jeremy-baier.github.io/GWplotter/` [Moore et al., 2014]. 39

3.1 Comparison of log-marginal likelihood estimated as a function of K number of β chains for the Gaussian model with TI, SS, and GSS methods. 83
3.2 Convergence of the marginal likelihood estimated using GSS for the Gaussian model with d = 2000 and v = 0.01. Top panel: Comparison of log-marginal likelihood estimated as a function of K number of β chains with n = 10. Bottom panel: Comparison of log-marginal likelihood estimated as a function of n number of independent samples per β chains with K = 4 84
3.3 Estimations as a function of K number of β chains for the NANOGrav dataset via GSS. Top panel: log-marginal likelihood estimated for the IRN model. Middle panel: log-marginal likelihood estimated for the CURN model. Bottom panel: $\log(\widehat{\mathrm{BF}}_{\mathrm{CURN/IRN}})$ comparing CURN to IRN. 88
3.4 Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{\mathrm{M_1/M_2}})$ estimates comparing different common red process models in the NANOGrav dataset where the arrow starts at model $\mathrm{M_2}$ and ends at model $\mathrm{M_1}$. . . 91
3.5 Evolution of $\log(\widehat{\mathrm{BF}}_{\mathrm{M_1/M_2}})$ comparing different common red process models as a function of the dataset size of EPTA+InPTA dataset with DR2full+ the full dataset and DR2new the shortest subset. 92

4.1 Convergence and accuracy of *Gaussian-shells* $\log(z)$ estimates versus $n_{\rm live}$ the number of live points. The posterior samples from one NS run with $n_{\rm live}$ is used for MorphZ estimates plotted at $n_{\rm live}$, with 3000 likelihood calls (for BS proposal samples) per each MorphZ estimate. *Top:* Estimated $\log(z)$, NS (black) and the MorphZ with Morph approximations of second order ($\mathcal{M}_2$, gray) and third order ($\mathcal{M}_3$, red). Symbols denote means over 100 run; error bars indicate 1σ uncertainties. The horizontal dotted line marks the true $\log(z)$. *Bottom:* Root-mean-square error (RMSE) of $\log(z)$ relative to the true value 110
4.2 Empirical vs. BS RE $\log(z)$ error across models 113
4.3 Comparison of $\log(z)$ estimates for the EPTA DR2FULL PSR+CURN model using GSS and MorphZ. The box plots show the distribution of $\log(z)$ for each Morph approximation order. The horizontal dashed line and shaded band indicate the mean and $\pm1\sigma$ range of the reference GSS estimates with 64 temperatures. Red diamonds (right axis) denote the forward Kullback-Leibler divergence $D_{\rm KL}(P|\mathcal{M}_{\mathcal{B}_L})$ between the posterior and each corresponding order of the Morph approximation. . 115
4.4 Comparison of dynesty, MorphZ $_{\text{(PT)}}$, and MorphZ $_{\text{(NS)}}$ across 100 independent BBH simulation (see Table 4.3). The mean and standard deviation of 100 $\log(z)$ estimate per simulation is used to compute $\Delta\log(\hat{z}) = \log(\hat{z})_{\rm method} - \log(\hat{z})_{\rm NS}$ for both independent MorphZ estimates. Note that $\log(z)_{\rm NS}$ is used as a reference estimate because the true $\log(z)$ is unknown. 118
4.5 Relative difference of the total sum of simulated weights and runtime comparison of SGM and ILP across different orders of the Morph approximation for 100 independent sum of weight maximization simulation. Each simulation assign a weight drawn from Beta$(\alpha, 2)$ with $\alpha \sim$ Uniform$(5, 9)$ for each possible group from $\binom{50}{L}$. Top: Relative difference of the total sum of chosen weights, $(((\mathrm{SGM}-\mathrm{ILP})/\mathrm{ILP})\times100\%)$, shown as box-and-whisker plots for $L = 2 - 5$ with the corresponding N_b number of blocks of length L ; the horizontal dashed line marks zero difference. Bottom: Wall-clock time in seconds for SGM (black) and ILP (red) on the same instances; markers denote mean and $(\pm1\sigma)$. The dashed line marks $1\mathrm{s}$. 123

5.1 Summary of the model space for 22 pulsars noise analysis. For each pulsar, the model with the highest posterior probability $p_{\rm top}$ is shown for both the 2020 DR2 (left) and 2023 DR2 (right) pipelines. Bars are colour-coded by class: decisive ($p_{\rm top} \geq 0.7$; cyan), competitive ($0.35 \leq p_{\rm top} < 0.7$; light orange), and diffuse ($p_{\rm top} < 0.35$; light grey). Abbreviated top models are annotated; EF and EC are present in every $p_{\rm top}$ model and omitted for clarity. Stars mark pulsars that change classes between datasets. 136

5.2 Posterior model probabilities for one representative pulsar from each class under the 2020 DR2 (grey) and 2023 DR2 (cyan) pipelines. *(a)* Decisive: one model dominates at $\sim 90\%$. *(b)* Competitive: two SW-containing models each hold $\sim 40\%$. *(c)* Diffuse: the posterior fragments in 2020 but reorganizes around DMv+SW in 2023. 140
5.3 Corner plots of the SP parameters of the highest-posterior 2023 noise model, evaluated using the 2020 DR2 (grey) and the 2023 DR2 (teal) pipelines: (a) PSR J1909−3744, representing a decisive model space; (b) PSR J1022+1001, representing a competitive space; and (c) PSR J1730−2304, which transitions from a diffuse to a competitive space. 143
5.4 Comparison of the inferred solar-wind electron density at Earth, $n_{\rm earth}^{\rm SW}$, for ten PPTA pulsars in the PPTA DR2 (2023) and PPTA DR3 analyses[Reardon et al., 2023a]. The violin profiles show the distributions of $n_{\rm earth}^{\rm SW}$ posterior summaries from Table 5.3, while the markers and vertical bars indicate the posterior medians and central 68% credible intervals. The horizontal dashed line marks the reference value $n_{\rm earth} = 4\,{\rm cm}^{-3}$. Pulsars are arranged by increasing absolute ecliptic latitude, $|\beta_e|$, which is reported below each pulsar. 147
5.5 Posterior model probabilities obtained with the 2020 DR2 pipeline (grey) and the 2023 DR2 pipeline (teal) for pulsars not shown in the main text. This Figure shows the first 12 pulsars. 153
5.6 Posterior model probabilities obtained with the 2020 DR2 pipeline (grey) and the 2023 DR2 pipeline (teal) for the remaining seven pulsars not shown in the main text. 155
5.7 Joint and marginalized posterior distributions for the parameters of the preferred noise models of the seven additional pulsars showing statistically significant solar-wind variations. Results obtained with the 2020 PPTA DR2 pipeline are shown in grey, while those obtained with the 2023 PPTA DR2 pipeline are shown in teal. 156

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature of candidate

Publications

These publications represents our total research contribution during the course of this thesis.

Chapter 3
Zahraoui, E. M., Maturana-Russel, P., van Straten, W., Meyer, R., & Gulyaev, S., 2025. Generalized stepping-stone sampling: efficient marginal likelihood estimation in gravitational wave analysis of pulsar timing array data. *Monthly Notices of the Royal Astronomical Society*, 540(4), pp.3818–3825.
Status: Published.

Chapter 4
Zahraoui, E. M., Maturana-Russel, P., Vajpeyi, A., van Straten, W., Meyer, R., & Gulyaev, S., 2026. Enhancing evidence estimation through informed probability density approximation. *Physical Review D*, 113(8), 083014.
Status: Published.

Chapter 5
Zahraoui, E. M., Maturana-Russel, P., van Straten, W., Meyer, R., & Gulyaev, S., 2026. Insights into the pulsar timing array hypothesis space. *Target journal: Physical Review D*.
Status: Published.

Additional Publications
During the course of my candidature, I also contributed to the following publications which are not included as chapters in this thesis:

Zahraoui, E. M., Rüdisser, H. T., Shaifullah, G. M., Tiburzi, C., Grießmeier, J.-M., *et al.*, 2026. Probing the magnetic field of a coronal mass ejection with PSR J1022+1001. *arXiv preprint*, arXiv:2603.18051.
Status: Accepted for Astronomy & Astrophysics.

Pan, Z., **Zahraoui, E. M.**, Maturana-Russel, P., & Cabrera-Guerrero, G., 2026. LIGO Core-Collapse Supernova Detection Using Convolutional Neural Networks. *Sensors*, 26(6), 1749.
Status: Published.

Acknowledgements

This thesis is dedicated to every brain and hand that helped shape my journey. Whether you stood near or from afar, your support has been the anchor keeping me steady through the shifting tides of life.

Among those who navigated these waters with me, my deepest gratitude begins with my mother, Nadia, whose unconditional love and support formed the bedrock beneath my every achievement. To my brother, my truest philosophical companion: you laid the early stones of my intellectual pathway and taught me not just to navigate the currents, but to accept no limits and reach for the highest peaks. To my little sister, whose bright spark and endless laughter were the sunlight breaking through the heavy clouds of this endeavor. And to my father, the quiet artist of my upbringing, whose open mind and creative spirit gave me the freedom to craft my own path.

If my family provided the foundation for this work, my friends provided the much-needed sanctuary. My brother and friend Cesar, your move to New Zealand brought light and laughter to the most exhausting days; thank you for every moment, trip and walk that helped clear my mind. I am equally grateful to Hamza, a kindred spirit who is nothing short of family. Even from afar, you have been an amazing companion, a relentless motivator, and my safest space to talk through life and share ideas. This deep appreciation extends to the rest of the core group I grew up with Simo, Hamza and Nabil. We have shared a lifetime of memories. Whether sharing laughs in person or connecting through screens across vast distances, your constant presence made this distance disappear and always felt like home. I also want to thank Lamia for the shared memories since our time in France, and for being such a sharp, willing mind to share my research with. Your scientific debates continuously pushed me to view my own work from entirely new perspectives. I am also deeply grateful to Salah, Anas, Reda, Arthur, and all the friends who filled my studies in France with such wonderful and joyous moments.

The daily reality of this PhD was shared with my friend, brother, and colleague, Avi. We practically lived through this entire experience side by side. From collaborating on science and making games, to traveling, dancing, and our endless brotherly nagging, you brought a perfect, chaotic balance to my life that kept me moving forward. Alongside Avi, I was incredibly lucky to share so much laughter and travel with a wider, wonderful group: Amine, for his quiet warmth; Rachel, for her brilliant Kiwi heart; Ajay, for always bringing excitement; and Nicholas, the Belgian of the group, for beautifully capturing it all through his camera lens. You all made this chapter truly unforgettable.

To my coffee mate, Hans: our timeline together might be short, but it feels like we have been friends for a lifetime. Thank you for the endless caffeine and the camaraderie that fueled these last few months. I must also thank Chiara, Nazeela, Sheryas, Jianan, and Mica for beautifully blurring the lines between colleagues and friends. The fun times we shared, both inside and outside work, brought a wonderful lightness to my routine.

My profound gratitude belongs to my supervisor, Patricio. Your support went far beyond what anyone could expect from a supervisor. Thank you for your endless availability, your generous advice, and the privilege of working so closely together; I have learned a tremendous amount from you. This immense gratitude extends to my co-supervisors, Willem, Renate, and Sergei. It has been an honor to be guided by such a distinguished and formidable team. Your collective wisdom and invaluable advice empowered me to confidently carve out my own independent research path. I owe the very existence of this endeavor to Caterina, who opened that first door for me. Thank you for the unwavering support and caring hand you extended from day one all the way to the finish line; your visits were always a bright spot, arriving like a breath of fresh hope. Within our research group, a special thank you goes to Matt and Kate for your generous advice and steadfast daily support. Finally, perhaps the greatest highlight of my academic journey was the unforgettable day I spent with the legendary John Skilling. Every conversation was a masterclass in history and inspiration, providing a deep intellectual thrill that reinvigorated my passion for this work.

To anyone I may have missed in these pages: please know that your kindness, encouragement, and shared moments are forever etched in my memory. This milestone is not just an ending, but the beginning of a bright future, and I am deeply grateful to step into it with all of you.

Artificial Intelligence Declaration

Specify the chapter number(s) in your research proposal/research component where AI was used.
Chapter number(s): **Chapter 1 to 6**
Briefly explain why the AI tool was used (e.g., idea generation, summarization, data analysis, image creation, etc.).
Purpose of AI Use: Code debugging, Search engine, writing suggestions, grammar correction, Latex debugging.
Clearly state the AI tools used.
AI Tool(s) Used: ChatGPT, Gemini, VS code.
Prompts or queries entered into the AI tool (only the instructive portion of the prompt is required). Include all relevant variations.
Prompts Used: -
A summary or description of the AI-generated output. (Optional: attach output in an appendix.)
Output Received: -
Describe how the AI output was used, edited, or transformed.
Post-AI Processing Methods: -

You may use this free text box to explain the intent for the use of the AI generated content. *For example, rather than stating what prompt(s) were used, you may articulate the purpose for which the prompt was used (supported by examples) here:*
Artificial intelligence is now integrated with any tool one considers using from a simple Google search, coding in an IDE or writing in Overleaf. For this thesis, AI was used generally as a tool to help with code debugging, as search engine, writing suggestions, grammar correction and Latex debugging. These tools are very helpful in many cases and can increase productivity. Nevertheless, most of its outputs or suggestions always needs significant modifications or corrections. The work presented in this thesis respects the guidelines stated in AUT's Postgraduate Handbook 2026.This research work is fully our intellectual property and a product of our Human work.

STUDENT DECLARATION			
By signing you are confirming that the AI use stated in the table(s) above are accurate and follow AUT's recommended guidelines detailed in the AI Hub and Postgraduate Handbook.			
Student Name:	**El Mehdi Zahraoui**	**Signature:**	**Date:** **06/06/2026**

Co-Authorship Form

Co-authorship Contributions within this Thesis

STUDENT AND SUPERVISOR APPROVALS				
By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.				
Student Name	EL Mehdi Zahraoui	**Signature**	**Date**	**6/6/2026**
Supervisor Name	Patricio Maturana-Russel	**Signature**	**Date**	**6/6/2026**

Chapter Number:	**Three**
Manuscript Title:	**Generalized Steppingstone Sampling: Efficient marginal likelihood estimation in gravitational wave analysis of Pulsar Timing Array data.**
Publication Status:	**Published**
Reference if published:	**https://doi.org/10.1093/mnras/staf953**

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
EL Mehdi Zahraoui	Conception, Application and Research output, Drafting
Patricio Maturana-Russel	Conception, Application and Research output, Supervision, Drafting, Funding
Willem van Straten	Contribution of knowledge, Supervision, Drafting, Funding
Renate Meyer	Contribution of knowledge, Supervision, Drafting, Funding
Sergei Gulyaev	Supervision, Drafting, Funding

Chapter Number:	**Four**
Manuscript Title:	**Enhancing evidence estimation through informed probability density approximation.**
Publication Status:	**Published**
Reference if published:	**https://doi.org/10.1103/1554-y6ns**

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
EL Mehdi Zahraoui	Conception, Application and Research output, Drafting
Patricio Maturana-Russel	Contribution of knowledge, Research output, Drafting
Avi Vajpeyi	Application to LVK, Drafting
Willem van Straten	Contribution of knowledge, Supervision, Drafting
Renate Meyer	Supervision, Drafting
Sergei Gulyaev	Supervision, Drafting

Chapter Number:	**Five**
Manuscript Title:	**Insights into the pulsar timing array hypothesis space.**
Publication Status:	**Submitted for Publication**
Reference if published:	

AUTHOR SURNAME: (order as per manuscript)	CONTRIBUTION (May copy from the guidelines above)
EL Mehdi Zahraoui	Conception, Application and Research output, Drafting
Patricio Maturana-Russel	Conception, Supervision, Research output, Drafting
Willem van Straten	Supervision, Drafting
Renate Meyer	Supervision, Drafting
Sergei Gulyaev	Supervision, Drafting

Chapter 1

Introduction

Since their first discovery in 1967 [Hewish et al., 1968], pulsars have been the subject of extensive research in astrophysics. These compact, rotating neutron stars emit electromagnetic radiation across the entire electromagnetic spectrum, with beams that can be observed as regular pulses as they sweep across the Earth [Gold, 1968]. Their extreme density, rapid rotation, strong magnetic fields, and long-term stability make them natural laboratories for testing fundamental physics. The subsequent development of precision pulsar timing turned these objects into invaluable tools for studying gravity, dense matter, binary evolution, interstellar plasma, and the large-scale structure of space-time.

The radio emission from Millisecond pulsars (MSPs) is of central importance to the pulsar timing array field. MSPs are exceptionally valuable because their stable radio pulses can be timed with extraordinary precision over many decades [Matsakis et al., 1997]. By comparing measured pulse Times of Arrival (TOAs) with a deterministic timing model, astronomers obtain timing residuals that encode a wealth of information [Lorimer, 2008]. These residuals may contain instrumental noise, propagation effects, intrinsic pulsar noise, Solar-System systematics, clock errors, and gravitational-wave signals. A Pulsar Timing Array (PTA) uses an ensemble of MSPs as a Galactic-scale

detector for nanohertz gravitational waves, particularly those produced by populations of SMBHBs [Taylor, 2021]. In the nanohertz band, a Stochastic Gravitational-Wave Background (SGWB) manifests not only as a common uncorrelated red process in pulsar residuals, but also through a distinctive spatial signature described in general relativity by the Hellings and Downs (HD) correlations [Hellings and Downs, 1983].

Over the last decade, PTA experiments have progressed from simply setting upper limits on nanohertz gravitational waves to finding statistical evidence for a common low-frequency process and spatial correlations consistent with a stochastic gravitational-wave background [Agazie et al., 2023b, collaboration et al., 2024, Reardon et al., 2023a]. Progress toward this milestone has been driven by longer observing baselines, improved radio instrumentation, wider bandwidths, better calibration, and the discovery of more stable MSPs. These efforts have been underpinned by Bayesian noise modeling, which provides a coherent probabilistic framework for combining timing data, physical signal models, instrumental effects, and multiple sources of noise uncertainty Lentati et al. [2014b], Van Haasteren et al. [2009].

However, as PTA datasets grow in length and complexity, the statistical problem becomes increasingly challenging. Both the accuracy and sensitivity of a PTA depend strongly on the carefully considered and optimal choice of noise models. Many processes such as unaccounted red noise, chromatic propagation effects, solar-wind variations, or instrumental systematics can easily absorb, mimic, or hide the correlations expected from a gravitational-wave background. The adopted statistical methods must be sophisticated enough to distinguish between true gravitational-wave signals and both pulsar-specific and systematic noise sources. The Bayesian inference employed in PTA provides the natural framework for this problem, allowing both parameter and model uncertainty to be quantified within a single analysis, where all noise model choices are currently informed by Bayesian model comparison. An optimal tuning of a PTA network might require robust Bayesian model comparison and more, which forms the

core focus of this thesis.

Once an array becomes sensitive to a common red noise signal, Bayesian evidence provides the standard metric to test whether a gravitational-wave model explains this signal better than other noise or systematic hypotheses. Therefore, evidence estimation is central to both stages of PTA analysis: building reliable noise models for individual pulsars and assessing whether the detected signal supports a gravitational-wave interpretation. In such high-dimensional PTA problems, computing the Bayesian evidence is exceptionally difficult. Likelihood calculations are computationally expensive, posteriors often contain strong correlations and degeneracies, and the model space grows rapidly as additional pulsars, noise processes, and spatial-correlation hypotheses are included. To address the growing computational challenge of PTA analysis, this thesis develops and applies efficient evidence-estimation methods that scale with the growing PTA experiments while maintaining high accuracy.

The thesis is structured around two main manuscripts that significantly reduce the amount of computation required to evaluate the Bayesian likelihood. Additionally, a third manuscript demonstrates how these fast and accurate methods enable new insights into the hypothesis landscape of pulsar noise modeling. The first manuscript introduces generalized steppingstone sampling as a new and practical method for PTA marginal-likelihood estimation. By utilizing posterior-informed reference distributions, this approach significantly reduces the computational cost of evidence estimation; it is successfully applied to both single-pulsar noise analysis and gravitational-wave-background analysis. Building upon this foundation, the second manuscript extends the methodology through proposing a new informed probability approximation and the MorphZ estimator. MorphZ constructs a tractable approximation of the posterior that preserves important parameter dependencies, which is then utilized in bridge sampling to derive efficient evidence estimates directly from posterior samples. Both methods reduce the evidence estimation cost by several orders of magnitude compared to the

methods currently used in PTA analyses.

The third manuscript applies these newly developed evidence-estimation tools to the single-pulsar noise-model hypothesis space. This work evaluates posterior support across multiple competing noise models using two versions of the Parkes Pulsar Timing Array (PPTA) second data release. The analysis demonstrates that several pulsars retain competitive support for multiple noise profiles, and that reprocessed data can reveal previously hidden signals. These findings strongly motivate the use of Bayesian model averaging to effectively propagate noise-model uncertainty into broader gravitational-wave searches. Together, these three manuscripts demonstrate that maximizing PTA sensitivity and future gravitational-wave discoveries depends just as heavily on reliable Bayesian computation as it does on acquiring better data.

The remainder of this thesis is organized as follows. Chapter 2 provides the necessary background on pulsars, gravitational waves, Bayesian inference, and PTA data analysis, with particular emphasis on Bayesian evidence estimation. Chapters 3–5 form the main body of the thesis and present three manuscripts. The first manuscript introduces Generalized Steppingstone Sampling (GSS) for PTA marginal-likelihood estimation; the second develops the MorphZ evidence estimator based on the morph approximation; and the third applies these methods to the noise-model hypothesis space of the PPTA second data release. Finally, Chapter 6 synthesizes the results of the thesis, discusses their implications for PTA Bayesian analysis, and outlines directions for future work.

Chapter 2

Literature Review

2.1 Pulsars

2.1.1 Discovery

The history of pulsars is deeply rooted in the earlier theoretical development of compact stellar remnants. In 1934, Baade and Zwicky argued that supernovae could leave behind neutron stars, objects of dense nuclear matter and small radius [Baade and Zwicky, 1934]. Oppenheimer and Volkoff later studied the equilibrium of cold degenerate neutron matter and showed that such stars would be relativistically compact [Oppenheimer and Volkoff, 1939]. For several decades these objects remained largely theoretical, but the combination of rapid rotation and strong magnetic fields was already being invoked to explain the energetics of the Crab Nebula [Pacini, 1967].

In 1967, Jocelyn Bell and Antony Hewish identified a recurring signal in an interplanetary scintillation survey at $81.5\,\mathrm{MHz}$ (see Figure 2.1), with successive transits arriving about four minutes earlier each solar day, as expected for a fixed celestial source. The signal had a period of $1.337\,\mathrm{s}$ and was later associated with PSR B1919+21 [Hewish et al., 1968], which marked the first observation of a pulsar. The regularity

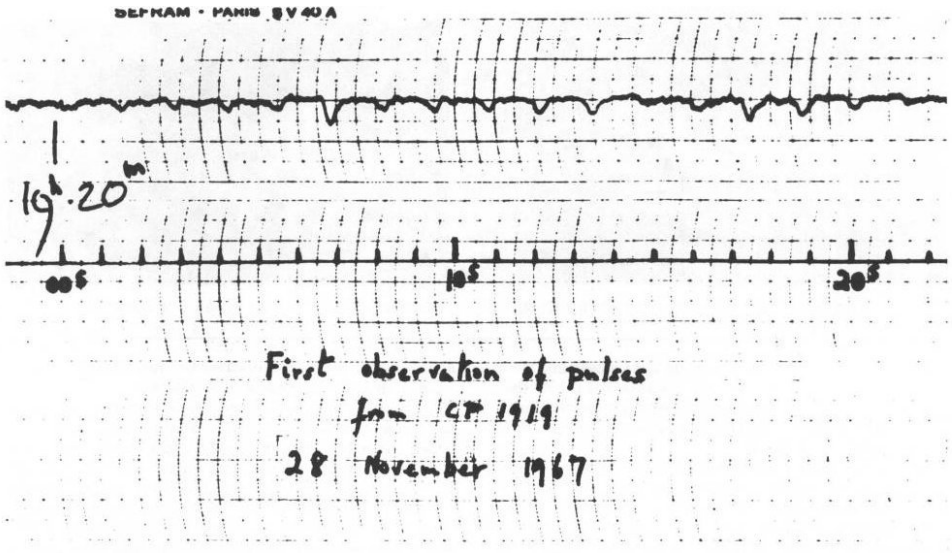


Figure 2.1: Pen chart recording of PSR B1919+21 pulses, first observed by Jocelyn Bell Burnell. Credit: National Radio Astronomy Observatory.

of the pulses made it natural to interpret the source as a rotating compact star, and the rotating neutron-star interpretation rapidly displaced earlier possibilities such as white-dwarf pulsation or terrestrial interference [Gold, 1968]. The radio beam is tied to the magnetic axis of the neutron star and sweeps across the observer once per rotation, producing the lighthouse-like sequence of pulses that defines the pulsar phenomenon.

This discovery transformed neutron stars from theoretical endpoints of stellar evolution into observable astrophysical laboratories. The following discoveries of the Crab and Vela pulsars strengthened this interpretation, since their association with young supernova remnants connected pulsars to massive-star death. These pulsars also provided direct evidence that young neutron stars lose rotational energy through electromagnetic and particle outflows.

2.1.2 Radio Observations

Radio pulsar observing systems must account for the fact that the signal propagates through ionized material before reaching the telescope. The Interstellar Medium (ISM),

the Solar Wind (SW), and ionospheric and instrumental effects all modify the observed pulse. The most important propagation effect for timing is cold-plasma dispersion. A pulse observed at radio frequency f is delayed to the speed of light in a vacuum by an amount proportional to f^{-2},

$$\Delta t_{\mathrm{DM}} = K_{\mathrm{DM}}\,\mathrm{DM}\,f^{-2}, \tag{2.1}$$

where K_{DM} is the dispersion constant and

$$\mathrm{DM} = \int_0^d n_e(l)\,dl \tag{2.2}$$

measures the integrated free-electron density along the line of sight [Lorimer and Kramer, 2005]. Consequently, lower-frequency signals arrive later than higher-frequency signals. This chromatic effect is corrected through incoherent or coherent de-dispersion, but the same f^{-2} scaling becomes important again in PTA analyses when time-variable DM and solar-wind delays are modeled as stochastic or deterministic noise components.

After de-dispersion, the data are integrated as a function of pulsar longitude, or rotational phase, to build an average pulse profile. This process, commonly known as *folding*, increases the signal-to-noise ratio by summing many single pulses, usually over observing times of minutes to an hour. The resulting profiles can be frequency-resolved, polarization-calibrated, and averaged over many hours into observing-system-dependent templates. Their quality depends on bandwidth, integration time, pulse width, receiver noise, calibration, radio-frequency interference, and pulse-to-pulse stochasticity such as jitter [Cordes and Shannon, 2010]. These instrumental and propagation effects set the measurement precision available for the timing procedure described in Section 2.1.5, where the pulse profiles are converted into TOAs and timing residuals.

Before a pulsar observation can be used for precision timing, the observatory data

must also be associated with accurate clock information, telescope metadata, observing frequency, backend configuration, and calibration information. These details are not merely observational bookkeeping: in long timing datasets they determine which offsets, noise parameters, or system-specific corrections must be later included in the timing and PTA likelihood models. Thus radio observations provide the calibrated pulse profiles and metadata from which pulsar timing begins, while the construction of residuals and the fitting of a timing model are treated in the next subsection.

2.1.3 Properties

Pulsars are rotating, magnetized neutron stars whose observed pulse period P is identified with the stellar rotation period. Their defining long-term property is spin-down: for most isolated rotation-powered pulsars $\dot{P} > 0$, indicating that rotational kinetic energy is being lost. Under the standard magnetic-dipole braking approximation, the characteristic age τ_c, inferred surface dipole field $B_{\rm s}$, and spin-down luminosity $\dot{E}$ are commonly written as

$$\tau_c = \frac{P}{2\dot{P}}, \tag{2.3}$$

$$B_{\rm s} \simeq 3.2 \times 10^{19} \sqrt{P\dot{P}}\,\mathrm{Gauss}, \tag{2.4}$$

$$\dot{E} = 4\pi^2 I \frac{\dot{P}}{P^3}, \tag{2.5}$$

where I is the neutron-star moment of inertia [Taylor and Manchester, 1977, Lorimer and Kramer, 2005]. These quantities are not direct measurements of age, magnetic field, or energy-loss mechanism in every system, because they assume a simplified braking law and neglect effects such as magnetic-field evolution, kinematic corrections, or accretion history. They are nevertheless useful parameters for broad classification of pulsars in terms of their locations in the P–$\dot{P}$ diagram (see Figure 2.2), such as

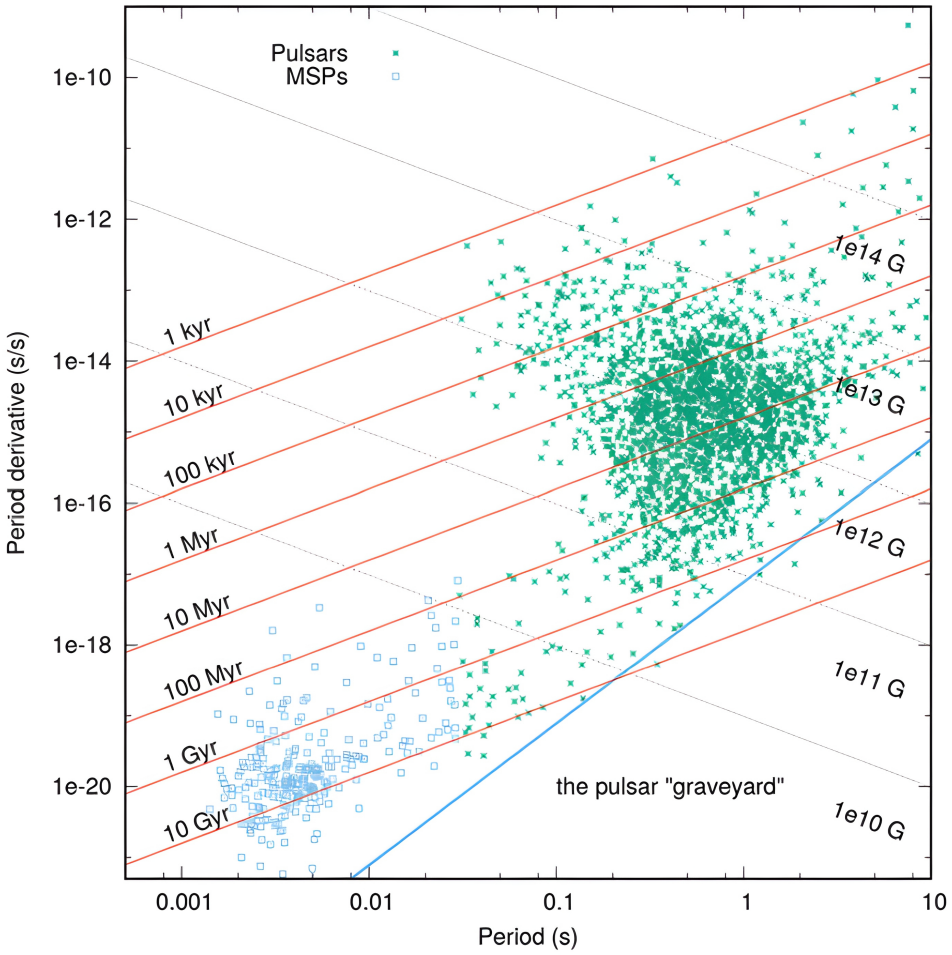


Figure 2.2: Pulsar catalog plotted from ATNF Pulsar Database showcasing regular pulsars in green and MSPs in cyan. Adapted from [Hobbs et al., 2004].

distinguishing young, energetic pulsars from old recycled millisecond pulsars. The radio emission mechanism is still not completely understood, but it is known to be coherent, highly beamed, and generated in a relativistic magnetospheric plasma rather than by ordinary incoherent synchrotron emission [Gold, 1969, Sturrock, 1971, Ruderman, 1976]. The axial tilt between the magnetic and rotation axes produces a pulse when the emission beam intersects the Earth Line of Sight (LoS), forming the lighthouse picture of pulsar emission [Pacini, 1968, Gold, 1969]. Individual pulses can vary strongly, but the average pulse profile obtained after integrating over many rotations is often stable enough to act as a high signal-to-noise template [Lorimer and Kramer, 2005]. This

stability is the practical reason that pulsars can be timed: the profile defines a fiducial phase, while the sequence of measured pulse times of arrival (TOAs) can be fitted using a deterministic timing model [Taylor, 1992b, Hobbs et al., 2006b].

The observed pulsar population spans several orders of magnitude in spin period, from the fastest known millisecond pulsars with periods of order $1\,\mathrm{ms}$ to slow pulsars with periods of up to ~ 76 seconds [Sob'yanin, 2023]. Ordinary non-recycled pulsars occupy the central region of the P–$\dot{P}$ diagram (see Figure 2.2) and typically have inferred magnetic field strengths around 10^{11}–$10^{13}\,\mathrm{G}$, whereas millisecond pulsars occupy the lower-left region with much smaller $\dot{P}$ and inferred field strengths of order 10^{7}–$10^{9}\,\mathrm{G}$.

2.1.4 Millisecond Pulsars

For more than a decade after the first pulsar discoveries, the known pulsar population was dominated by objects with spin periods ranging from hundreds of milliseconds to seconds. This changed with the discovery of PSR B1937+21, a pulsar with a period of approximately $1.558\,\mathrm{ms}$ [Backer et al., 1982]. The discovery introduced a new class of objects, Millisecond pulsars (MSPs), usually defined as rotation-powered pulsars with a period $P < 30\,\mathrm{ms}$, extremely small spin-period derivatives, and weak inferred surface magnetic fields compared with ordinary young pulsars [Lorimer, 2008, Manchester, 2017]. Their location in the P–$\dot{P}$ diagram shows that they are old systems, but their rapid rotation requires a mechanism to spin them back up after the initial supernova birth.

The recycling scenario still stand as the standard explanation, where an old neutron star in a binary system accretes mass from a companion and gains angular momentum. During an X-ray binary phase, this process reduces the magnetic field and spins up the star to millisecond periods. This interpretation was proposed shortly after the discovery

of the first MSP and developed into the modern evolutionary link between low-mass X-ray binaries and radio MSPs [Alpar et al., 1982, Radhakrishnan and Srinivasan, 1982]. Earlier work had already suggested that accretion from a companion could spin up a neutron star in a close binary [Bisnovatyi-Kogan and Komberg, 1976]. The later discovery of accreting millisecond X-ray pulsars provided a direct observational bridge between accretion-powered systems and rotation-powered radio MSPs [Wijnands and van der Klis, 1998].

The current MSP population is strongly shaped by binary evolution. Many MSPs have white-dwarf companions, while a smaller but important sub-population belongs to compact interacting binaries. Black-widow systems have very low-mass companions, typically of order 0.01–$0.06\,M_\odot$, whereas redbacks have non-degenerate or partially stripped companions of order 0.2–$0.8\,M_\odot$ in short orbits of order hours to a day [Fruchter, 1989, Roberts, 2012]. These spider systems often show radio eclipses, excess dispersion near eclipse boundaries, and evidence of interaction between the pulsar wind and material from the companion. Transitional millisecond pulsars provide a more direct link between accretion and rotation power by switching between a radio pulsar state and a sub-luminous accretion-disc state [Archibald et al., 2010, Papitto et al., 2013]. In contrast to these low-mass evolution pathways, mildly recycled MSPs with neutron star companions originate from high-mass binary systems. These double neutron star systems typically feature longer spin periods (tens of milliseconds) and highly eccentric orbits, serving as critical testbeds for general relativity and the direct progenitors of compact binary mergers [Hulse and Taylor, 1975, Burgay, 2005].

For PTA experiments, MSPs are valuable because their short periods, narrow profiles, and small $\dot{P}$ allow TOAs to be measured with high precision over decade-long baselines. Population studies also indicate that the observed Galactic MSP population is incomplete, with sensitive surveys and targeted searches of gamma-ray sources continuing to uncover new systems [Lorimer, 2008, Manchester, 2017]. The typical observed

Parameters	Range of values (units)
Spin period (P)	1.4 – 30 ms
Spin period derivative ($\dot{P}$)	10^{-22} – 10^{-18}
Magnetic field strength (B)	10^{7} – 10^{9} G
Characteristic age (τ_c)	10^{7} – 10^{11} years
Distance (d)	0.11 – 49 kpc
Companion mass	0.009 – $1.39(M_{\odot})$
Orbital period (P_b)	0.065 – 669 days

Table 2.1: Typical observed parameter ranges for radio MSPs [Bhattacharyya et al., 2022].

parameter ranges for radio millisecond pulsars are summarized in Table 2.1.

2.1.5 Pulsar Timing

Pulsar timing starts with the calibrated pulse profiles that are produced as described in Section 2.1.2. Its purpose is to connect the rotational phase of a pulsar coherently over many observing epochs. The main steps of this process, from signal reception to the construction of timing residuals, are illustrated schematically in Figure 2.3. Since individual pulses are variable, timing relies on the stability of the average profile obtained after many rotations. A high signal-to-noise template profile is then constructed, and each observed profile is compared with this template, usually through template matching or Fourier-domain cross-correlation, to estimate a pulse time of arrival (TOA) and its uncertainty [Hankins and Rickett, 1975, Taylor, 1992a, Lorimer and Kramer, 2005].

The measured TOAs are initially recorded at the observatory and must be transformed to a common inertial reference frame before they can be interpreted as measurements of pulsar rotational phase. This transformation includes clock corrections from the observatory standard to terrestrial time, corrections for the motion of the observatory with respect to the Solar-System barycentre, and propagation delays associated with the Solar-System gravitational potential. The dominant deterministic corrections include the Roemer delay from light travel across the Solar System, the Einstein delay due to

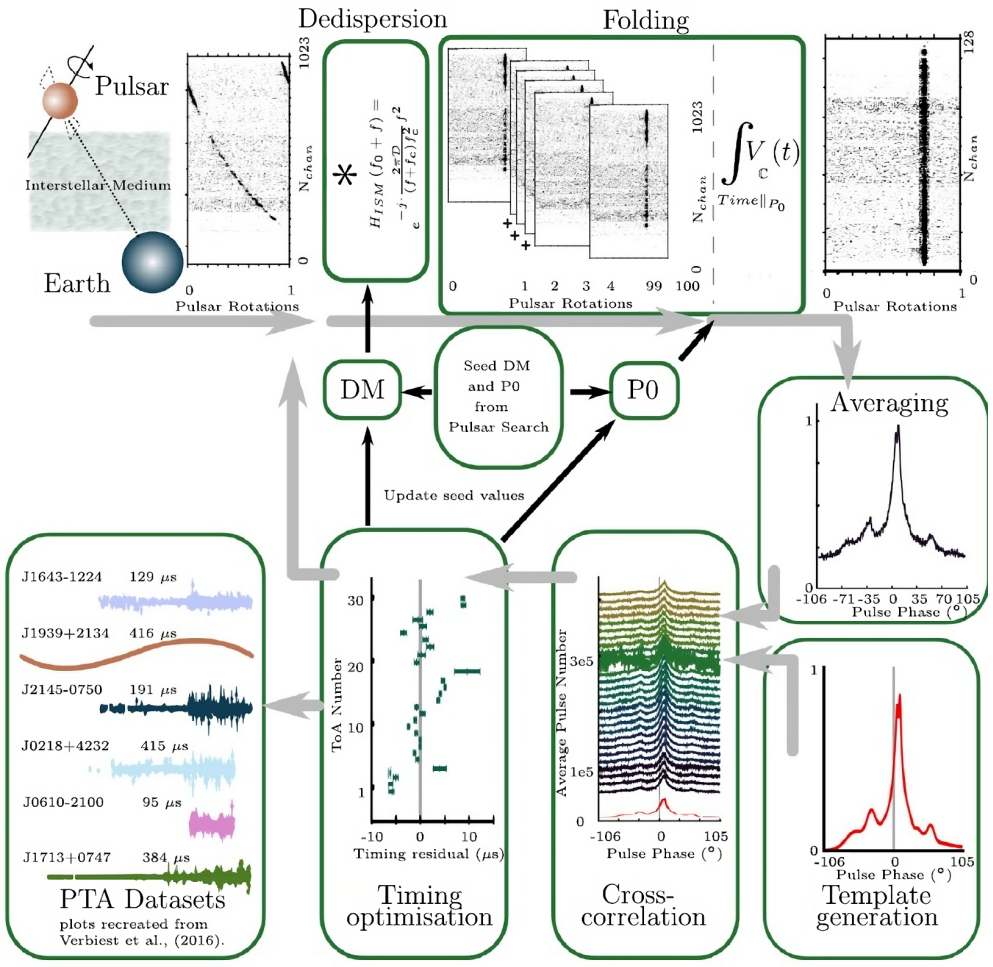


Figure 2.3: A diagram of the pulsar timing pipeline from signal reception to the construction of timing residuals used in PTA analyses. Adapted from [Verbiest and Shaifullah, 2018].

gravitational redshift and time dilation, and the Shapiro delay from propagation through gravitational potentials [Hobbs et al., 2006a]. Additional corrections are required for dispersive propagation through the interstellar medium and, for binary pulsars, for the motion of the pulsar around the binary barycentre.

After these transformations, the pulsar rotational phase is described by a timing model. For an isolated pulsar, the simplest part of this model is a Taylor expansion of the rotational phase $\phi(t)$ about a reference epoch t_0,

$$\phi(t) = \phi_0 + \nu(t - t_0) + \frac{1}{2}\dot{\nu}(t - t_0)^2 + \frac{1}{6}\ddot{\nu}(t - t_0)^3 + \cdots \tag{2.6}$$

where $\nu = 1/P$ is the spin frequency and $\dot{\nu}, \ddot{\nu}, \ldots$ are higher derivatives that describe the secular spin evolution. In a full timing solution, this spin model is fitted simultaneously with astrometric parameters such as sky position, proper motion, and parallax; propagation parameters such as the dispersion measure and its possible time variation; and, where relevant, binary orbital parameters. Instrument-dependent offsets between observing systems are often included as additional "jump" parameters, because long timing datasets are usually assembled from multiple telescopes, receivers, backends, and observing frequencies.

The fitting process compares the measured TOAs t_i^{obs} with the TOAs predicted by the timing model t_i^{model}. The differences,

$$\delta t_i = t_i^{\mathrm{obs}} - t_i^{\mathrm{model}}, \tag{2.7}$$

are called the timing residuals. The timing parameters are refined by minimizing any systematic structure in these residuals, usually through weighted least-squares fitting or related likelihood-based procedures. Standard timing packages such as TEMPO,

TEMPO2, and the more recent Python-based PINT package implement these barycentric corrections, timing-model calculations, and parameter fits [Hobbs et al., 2006a, 2009, Luo et al., 2021]. These tools take as input a set of TOAs, their measurement uncertainties, and an initial parameter file, and output an updated timing solution together with the corresponding pre-fit and post-fit residuals.

The residuals are not only a diagnostic of the fit, but rather the primary data product used in later high-precision pulsar analyses. Their structure can contain contributions from radiometer noise, pulse-jitter noise, calibration errors, imperfect profile modelling, intrinsic rotational irregularities, dispersion-measure variations, scattering variations, and unmodelled binary or astrometric effects [Verbiest et al., 2016, Cordes and Shannon, 2010, Lentati et al., 2014a]. For millisecond pulsars, the small spin-down rates and stable average profiles make the residuals especially precise, which is why they are central to pulsar timing array analyses. In that context, the final post-fit residuals from many pulsars form the time series to which PTA signal and noise models are applied.

2.2 Gravitational Waves

Gravitational waves (GWs) arise naturally in general relativity as propagating perturbations of the space-time metric generated by accelerating mass-energy distributions with a time-varying quadrupole moment [Einstein, 1916, 1918]. Their physical status was debated for decades, because coordinate choices can mimic wave-like behaviour in the metric. Later theoretical work clarified that gravitational waves carry energy and can produce measurable relative motion between freely falling test particles [Eddington, 1922, Pirani, 1956, Bondi, 1957]. The first strong empirical evidence came not from direct interferometric detection but from pulsar timing: the orbital decay of the Hulse–Taylor binary pulsar agreed with the energy loss due to gravitational radiation as predicted by the general theory of relativity [Hulse and Taylor, 1975, Taylor and

Weisberg, 1982]. The astronomers who discovered this binary system were awarded the Nobel prize in Physics for 1993. This historical connection is important for PTAs, because the same timing precision that first demonstrated GW emission from one binary now allows many pulsars to be used as a detector for nanohertz gravitational waves.

2.2.1 GW Sources

Gravitational-wave sources can be broadly classified according to the physical mechanism that generates the time-varying quadrupole moment, the characteristic duration of the signal, and the frequency band in which the radiation is observed. Compact binaries form the central source class across much of gravitational-wave astronomy, but the accessible population changes with frequency: ground-based interferometers observe stellar-mass compact binaries (Hz–KHz), space-based detectors target massive black-hole binaries and Galactic compact binaries (μHz–mHz), while pulsar timing arrays are sensitive to the radiation produced mainly by supermassive black-hole binaries (SMBHB) (nHz–μHz). In the following, we discuss the different GW sources and their properties.

2.2.1.1 Stellar-Mass Compact Binaries

The first direct detections of gravitational waves came from the inspiral, merger, and ringdown of stellar-mass compact binaries observed by ground-based laser interferometers [Abbott et al., 2016]. These systems include binary black holes, binary neutron stars, and neutron-star–black-hole binaries. Their signals sweep upward in frequency as the binary loses orbital energy and angular momentum through gravitational-wave emission, producing the characteristic chirp signal. The waveform can be divided into an inspiral regime, where post-Newtonian methods are accurate; a merger regime, which requires numerical relativity; and a ringdown regime, described by the quasi-normal modes of

the remnant compact object [Thorne, 1987, Blanchet, 2014]. These sources occupy the high-frequency gravitational-wave band, from tens of hertz to kilohertz, and are therefore outside the direct sensitivity of pulsar timing arrays.

2.2.1.2 Supermassive Black-Hole Binaries

Supermassive black holes, with masses of order 10^6–$10^{10} M_\odot$, are believed to reside in the centres of most massive galaxies. During hierarchical galaxy formation, galaxy mergers naturally lead to the formation of massive black-hole binaries, which subsequently evolve through dynamical friction, stellar scattering, gas interactions, and finally gravitational-wave emission [Begelman et al., 1980, Volonteri and Begelman, 2010]. In the pulsar-timing band, the most relevant systems are supermassive black-hole binaries with masses of approximately 10^8–$10^{10} M_\odot$, whose orbital periods are years to decades. These binaries are expected to be the dominant astrophysical source class for PTAs [Rajagopal and Romani, 1995, Jaffe and Backer, 2003, Wyithe and Loeb, 2003, Sesana, 2013].

If many such binaries contribute weak and unresolved signals, their incoherent superposition produces a stochastic gravitational-wave background. For a population of circular, gravitational-wave-driven binaries, this background has the canonical characteristic-strain scaling discussed in Section 2.2.2. Deviations from the canonical spectrum can arise from environmental coupling, eccentricity, or the transition from a background dominated by many binaries to a regime where a small number of loud systems contribute most of the power [Sesana et al., 2008, Sesana, 2013].

2.2.1.3 Individually Resolvable Massive Black-Hole Binaries

Although PTAs are generally expected to first detect a stochastic background, sufficiently nearby or massive supermassive black-hole binaries may stand above the unresolved background and become individually resolvable. In this case, the signal is approximately

continuous over the observing span, because the binary evolves slowly compared with the duration of a PTA experiment. The corresponding timing response depends on the binary sky location, orbital frequency, inclination, polarization angle, phase, chirp mass, and luminosity distance [Sesana et al., 2009]. A key feature of PTA observations is that the signal contains an Earth term, common to all pulsars, and a pulsar term, which depends on the distance to each pulsar and the time at which the wavefront passed it. This makes individual-source searches sensitive not only to the binary parameters but also to pulsar-distance uncertainties.

2.2.1.4 Gravitational-Wave Memory Bursts

A further source category associated with massive black-hole binaries is the gravitational-wave burst with memory. PTAs are generally not sensitive to the final high-frequency oscillatory merger waveform of the most massive binaries, but they can be sensitive to the non-oscillatory memory component produced near coalescence. The memory effect corresponds to a permanent change in the metric after the passage of the wave, which appears in pulsar timing data as a ramp-like feature in the timing residuals [Christodoulou, 1991, Favata, 2010, Van Haasteren and Levin, 2010]. Such signals are transient in strain but broadband in timing residuals, and therefore provide an indirect way for PTAs to probe the final merger of supermassive black-hole binaries.

2.2.1.5 Cosmological Phase Transitions

First-order phase transitions in the early Universe can generate gravitational waves through bubble nucleation, bubble collisions, magnetohydrodynamic turbulence, and sound waves in the primordial plasma [Witten, 1984, Caprini et al., 2016]. The characteristic frequency of the resulting background depends on the temperature and duration of the phase transition. While electroweak-scale transitions are usually associated with the millihertz band relevant to space-based detectors, lower-temperature or hidden-sector

transitions may produce signals in the nanohertz band accessible to PTAs [Schwaller, 2015, Arzoumanian et al., 2021].

2.2.2 Gravitational Wave Background

A Stochastic Gravitational-Wave Background (SGWB) arises when many individually weak or unresolved sources overlap incoherently, such that the signal is better described statistically than as a deterministic waveform. For an isotropic, Gaussian, stationary background, the strain power spectral density $S_h(f)$, fractional energy density $\Omega_{\rm GW}(f)$, and characteristic strain $h_c(f)$ provide equivalent descriptions of the background. The fractional energy density per logarithmic frequency interval is

$$\Omega_{\rm GW}(f) = \frac{1}{\rho_c}\frac{d\rho_{\rm GW}}{d\ln f}, \tag{2.8}$$

where $\rho_c = 3H_0^2/(8\pi G)$ is the closure density with H_0 the Hubble constant and G the gravitational constant Taylor [2021]. The characteristic strain is related to the strain power spectral density by

$$h_c^2(f) = fS_h(f). \tag{2.9}$$

Since, the energy density of an SGWB as a function the strain Power Spectral Density (PSD) $S_h(f)$ can be written as :

$$\Omega_{\rm SGWB}(f) = \frac{2\pi^2}{3H_0^2}f^3S_h(f). \tag{2.10}$$

Therefore, the energy density of an SGWB as a function $h_c(f)$ is given by

$$\Omega_{\rm SGWB}(f) = \frac{2\pi^2}{3H_0^2}f^2h_c^2(f), \tag{2.11}$$

following the standard SGWB convention [Romano and Cornish, 2017].

For many PTA analyses the characteristic strain is parameterized as a power law,

$$h_c(f) = A_\alpha \left(\frac{f}{f_{\rm ref}} \right)^\alpha , \tag{2.12}$$

where A_α is the characteristic strain amplitude at a reference frequency, usually $f_{\rm ref} = 1\,{\rm yr}^{-1}$, and α is the strain spectral index. For a population of circular, GW-driven SMBHBs, the canonical prediction is $\alpha = -2/3$, corresponding to a timing-residual power spectral index often written as $\gamma = 13/3$ [Phinney, 2001, Sesana et al., 2008]. Deviations from this simple spectrum can arise from environmental coupling, eccentricity, discreteness of the source population, anisotropy, or a non-SMBHB origin. Thus, the SGWB is not only a detection target but also a probe of massive black-hole demographics and binary evolution.

Stochastic backgrounds are not unique to the PTA band. Any population of weak, unresolved sources can generate a random gravitational-wave signal. In the ground-based band, the unresolved population of stellar-mass compact binaries is expected to form an astrophysical stochastic background [Abbott et al., 2018]. In the millihertz band, the laser Laser Interferometer Space Antenna (LISA) is expected to observe a confusion foreground from compact Galactic binaries, especially white-dwarf binaries, together with possible contributions from massive black-hole binaries and extreme mass-ratio inspirals [Nelemans et al., 2001, Amaro-Seoane et al., 2017]. These backgrounds illustrate the general principle that gravitational-wave source classes are best understood jointly with the detector band in which they appear.

2.2.3 Gravitational-wave Detectors

Gravitational-wave detectors cover complementary frequency bands because the characteristic GW frequency is set by the dynamical timescale of the source and by the size of the detector. Ground-based laser interferometers use kilometre-scale arms and are

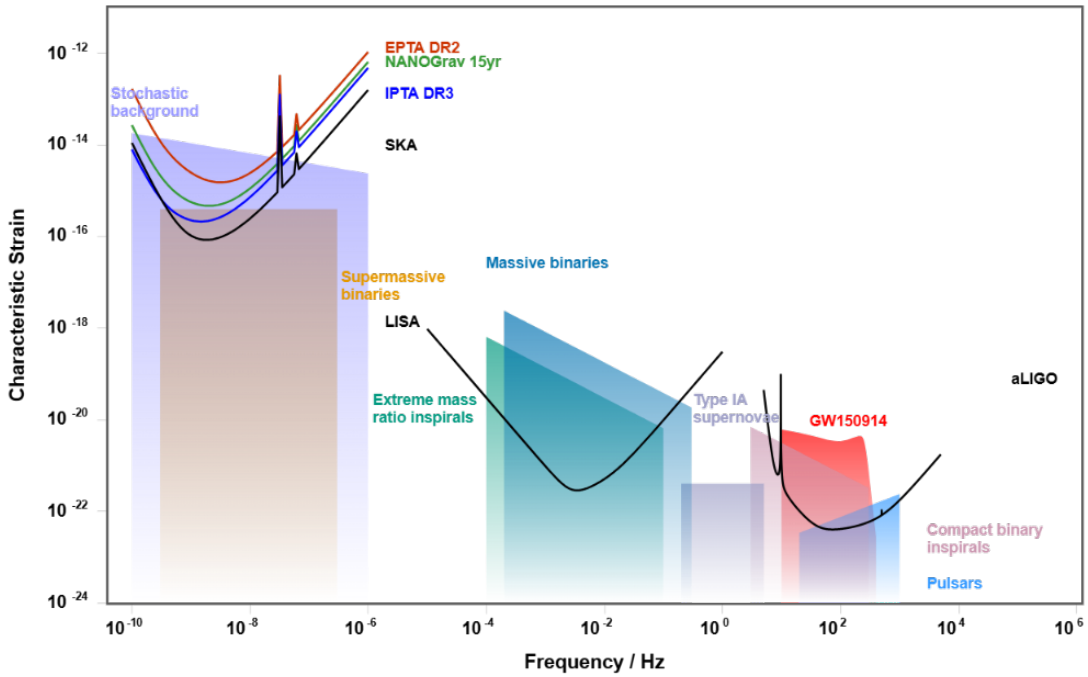


Figure 2.4: Representative sensitivity of PTA collaborations compared with other GW instruments such as advanced LIGO and future LISA. Figure generated with `https://jeremy-baier.github.io/GWplotter/` [Moore et al., 2014].

sensitive mainly from a few Hz to a few kHz; their low-frequency reach is limited by seismic and gravity-gradient noise while the high frequency is limited by instrumental noise. Space-borne interferometers avoid the seismic wall and use million-kilometre baselines to probe the millihertz band. Cosmic Microwave Background (CMB) polarization searches probe frequencies much lower than $\sim 10^{-15}$ Hz associated with the early Universe, while PTAs occupy the nanohertz band between these regimes [Moore et al., 2014, Romano and Cornish, 2017].

The PTA detector is qualitatively different from an interferometer because its arms are the lines of sight from the Earth to Galactic MSPs. The relevant observing timescale is the length and cadence of the timing campaign: the lowest accessible frequency is approximately the inverse of the data span, while the highest is set by the observing cadence. A decade-long campaign with roughly weekly observations is naturally sensitive to frequencies of order 10^{-9}–10^{-6} Hz [Lorimer, 2008, Manchester, 2017]. This

explains why PTAs target SMBHBs rather than stellar-mass binaries: the nanohertz band corresponds to orbital periods of years, appropriate for very massive binaries at large separations.

Sensitivity is not determined by a single pulsar alone. It improves with longer baselines, smaller TOA uncertainties, improved noise modelling, broader radio bandwidths, and an increasing number of pulsars $N_{\rm psr}$. The number of pulsar pairs grows as $N_{\rm psr}(N_{\rm psr}-1)/2$, so adding stable MSPs improves both raw sensitivity and the ability to reconstruct the angular correlation pattern.

2.2.4 Pulsar Timing Array

The idea of using pulsars as gravitational-wave detectors developed soon after the discovery of pulsars themselves. Sazhin considered the influence of ultra-long gravitational waves on electromagnetic pulse propagation, and Detweiler applied the idea to pulsar timing data, setting an early upper limit on the SGWB energy density [Sazhin, 1978, Detweiler, 1979]. Hellings and Downs then showed that an array of pulsars could be used to search for an isotropic background through the distinctive angular cross-correlation between pulsar residuals [Hellings and Downs, 1983]. The background induces timing residuals in different pulsars with a correlation that depends only on the angular separation ζ_{ij} between the pulsars. In general relativity, this correlation is described for two distinct pulsars by the Hellings–Downs function,

$$\Gamma(\zeta_{ij}) = \frac{1}{2} + \frac{3}{2}x_{ij}\ln x_{ij} - \frac{1}{4}x_{ij}, \qquad x_{ij} = \frac{1-\cos\zeta_{ij}}{2}, \tag{2.13}$$

with an additional pulsar self-noise term for autocorrelations [Hellings and Downs, 1983, Romano and Cornish, 2017]. This angular correlation is crucial because a common red spectrum alone is not unique to gravitational waves; clock errors, ephemeris errors, and other systematics can also produce common low-frequency structure. The quadrupolar

Hellings–Downs correlation is therefore the geometric signature that distinguishes a Gravitational-Wave Background (GWB) from other common processes. Foster and Backer later articulated the modern PTA concept of timing an ensemble of millisecond pulsars as a gravitational-wave detector [Foster and Backer, 1990].

PTA analysis begins with timing each pulsar independently, but its central data product is the joint covariance of all timing residuals. Signals are usually classified by their spatial correlation. A clock error produces a monopolar correlation because all TOAs are shifted in the same way. Solar-System ephemeris errors produce dipolar correlations because an incorrect Earth–SSB vector projects differently along each pulsar line of sight. A tensor-polarized SGWB in general relativity produces a quadrupolar Hellings–Downs correlation [Hobbs, 2013, Champion et al., 2010, Hellings and Downs, 1983]. Other correlated processes, including dark-matter models, non-standard GW polarizations, anisotropic backgrounds, or Solar-System systematics, can produce different angular structures [Khmelnitsky and Rubakov, 2014, Gair et al., 2014].

This classification also clarifies why PTA detection is difficult. The expected SGWB signal is a very low-frequency, low-amplitude process that must be separated from intrinsic pulsar red noise, chromatic ISM and solar-wind propagation delays, noise calibration errors, clock uncertainties, and ephemeris uncertainties. Mis-modelling one of these components can bias the inferred amplitude or mimic part of the spatial correlation pattern. In particular, Solar-System ephemeris errors can leak into correlations that may be confused with a gravitational-wave signal if the analysis focuses on only a restricted set of spatial modes [Tiburzi et al., 2016]. Modern PTA inference therefore fits pulsar-specific noise, common uncorrelated red processes, and spatially correlated signals in a single statistical framework.

The major PTA collaborations are the North American Nanohertz Observatory for Gravitational Waves (NANOGrav), the Parkes Pulsar Timing Array (PPTA), the European Pulsar Timing Array (EPTA), the MeerKAT Pulsar Timing Array (MPTA),

the Chinese Pulsar Timing Array (CPTA), the Indian Pulsar Timing Array (InPTA), the International Pulsar Timing Array (IPTA), and the Gamma-ray Pulsar Timing Array (GPTA) which combines data and expertise across regional arrays [McLaughlin, 2013b, Manchester et al., 2013, Desvignes et al., 2017, Verbiest et al., 2016, Miles et al., 2022, Chen et al., 2025, Tarafdar et al., 2022, Collaboration*†, 2022]. These collaborations use independent telescopes, backends, calibration pipelines, and noise models, making cross-checks between datasets essential. As timing baselines grow and the number of stable MSPs increases, PTA science is shifting from setting upper limits to characterizing the spectrum, spatial correlations, and astrophysical interpretation of the common nanohertz signal[Agazie et al., 2023b, collaboration et al., 2024, Reardon et al., 2023b].

2.3 Bayesian Inference

2.3.1 Bayes Rule

Bayesian inference provides a principled framework for relating observed data to the physical, instrumental, and stochastic processes that may have produced them. In the context of PTA analysis, the data $\boldsymbol{X}$ consist of the timing residuals which contain multiple sources of covariant noise sources and weak stochastic signals. Because these signals cannot be isolated through standard methods, the Bayesian framework is strictly necessary to model this dataset by robustly marginalizing over hundreds of nuisance parameters while consistently propagating measurement uncertainties. This framework allows prior knowledge, model assumptions, and measurement uncertainties to be combined consistently with the information contained in the data.

For a fixed model M, updating the prior information about a parameter vector $\boldsymbol{\theta}$

after observing data $\boldsymbol{X}$ is described by the Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\boldsymbol{X}|M)}, \tag{2.14}$$

where $p(\boldsymbol{\theta}|\boldsymbol{X}, M)$ is the posterior distribution of parameter vector $\boldsymbol{\theta}$ under the assumed model M, $p(\boldsymbol{X}|\boldsymbol{\theta}, M)$ is the likelihood function, and $p(\boldsymbol{\theta}|M)$ is the prior distribution of the parameters under model M. The denominator $p(\boldsymbol{X}|M)$ is the normalizing constant, known as the marginal likelihood or Bayesian evidence. In Manuscript 3 and 4, we use the notation $L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ and $\pi(\boldsymbol{\theta}|M)$ respectively for the likelihood function and the prior distribution. For basic parameter estimation, this normalizing constant is often not required explicitly, since many sampling algorithms use only the unnormalized product of likelihood and prior. However, its crucial role in model comparison is treated in detail in Section 2.3.4.

Bayes' theorem is also fundamentally at the level of the model space. Let

$$\mathcal{M} = \{M_1, M_2, \ldots, M_K\} \tag{2.15}$$

denote a set of competing models. In a PTA context, these models represent different physical assumptions about the data, such as pulsar noise, the presence of a common uncorrelated red process, spatially correlated processes, or the inclusion of instrumental noise terms. Each model M_i is characterized by a parameter vector $\boldsymbol{\theta}_i$, prior $p(\boldsymbol{\theta}_i|M_i)$, likelihood $p(\boldsymbol{X}|\boldsymbol{\theta}_i, M_i)$, and prior model probability $p(M_i)$. The posterior probability of the model is given by

$$p(M_i|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|M_i)p(M_i)}{\sum_{j=1}^{K} p(\boldsymbol{X}|M_j)p(M_j)}. \tag{2.16}$$

Here, $p(\boldsymbol{X}|M_i)$ is the marginal likelihood of model M_i, later denoted as $z(M_i)$. Therefore, Bayesian inference updates both parameter probabilities within a specific model

and model probabilities across the broader hypothesis space. This distinction is critical because uncertainty in PTA analysis arises not only from the unknown values of $\boldsymbol{\theta}$; but also from uncertainty regarding which phenomenological description adequately represents the data [Hinne et al., 2020].

2.3.2 Bayesian Modeling

Bayesian analysis begins with the specification of a model that mathematically describes the data-generation process. In this framework, a complete model consists of two inseparable components: the likelihood function, which connects theoretical predictions to the observed data, and the prior distribution, which defines the probability density across the parameter space before any data are observed [Andreon and Weaver, 2015]. In observational sciences, a rigorous likelihood must account not only for the underlying physical phenomena but also for the measurement process itself, including instrumental noise, nuisance parameters, selection biases, and calibration uncertainties.

The construction of a likelihood is guided primarily by the scientific question and the physics of the experiment. A robust model captures the dominant processes required to explain the dataset while remaining transparent enough that its underlying assumptions can be mathematically or physically verified. In practice, Bayesian modeling often follows an iterative approach: one begins with a simple foundational model and incrementally introduces complexity only when justified by the data or physical motivation. This procedure is not equivalent to selecting arbitrary combinations until a preferred result appears; rather, each added component should correspond to a plausible source of signal or noise in the data. This acknowledges the foundational statistical philosophy that all models are approximations of reality. The relevant scientific question is not whether an adopted model is perfectly true, but whether it is the most adequate representation among competing alternatives for the given data and scientific objective

[Gelman and Shalizi, 2013, Andreon and Weaver, 2015].

Equally critical to the model definition is the prior distribution. Priors are the explicit mathematical mechanism for incorporating existing domain expertise, physical boundary constraints, theoretical expectations, and previous independent measurements into the inference. Informative priors are appropriate when they are justified by external information and reflect the spirit of the Bayesian philosophy [Gelman and Shalizi, 2013, Eadie et al., 2023]. While "flat" or uniform priors are frequently chosen under the assumption that they are strictly uninformative, this interpretation is often misleading. A prior that is uniform in one parameterization is generally not uniform under a non-linear transformation. For example, a uniform prior on a logarithmic scale (a Jeffreys prior) assigns equal probability per decade, which is typically more physically appropriate for scale parameters—such as energies, variances, or amplitudes—than a uniform prior on a linear scale.

The specification of the prior is especially vital when the ultimate goal is model comparison. Because the marginal likelihood integrates the likelihood over the entire prior volume, the prior explicitly dictates the "size" of the model's predictive space. A prior that is unnecessarily broad will assign substantial probability mass to regions of parameter space that predict the data poorly, which severely penalizes the model's overall evidence. Conversely, an overly restrictive prior may artificially favor a model by improperly excluding plausible alternatives. Consequently, a comprehensive Bayesian model requires that priors be stated explicitly, justified physically, and subjected to rigorous sensitivity analyses to ensure the robustness of the final scientific conclusions [Gelman et al., 1996, Eadie et al., 2023].

2.3.3 Computational Bayesian Inference

In practice, the posterior distribution in Eq. 2.14 is available in closed form only for special conjugate cases. For the complex models required in PTA analysis, the posterior must be approximated using sampling techniques. Most algorithms require only the unnormalized posterior,

$$q(\boldsymbol{\theta}|\boldsymbol{X}, M) = p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M). \tag{2.17}$$

To explore this distribution, Markov chain Monte Carlo (MCMC) methods are widely adopted. The Metropolis–Hastings (MH) algorithm constructs a Markov chain by proposing a new point in parameter space and accepting or rejecting it based on the ratio of the posterior densities, allowing samples to eventually converge on the target distribution [Metropolis et al., 1953, Hastings, 1970]. However, as modern PTA datasets expand to include more pulsars and longer baselines, the dimensionality of $\boldsymbol{\theta}$ grows to several hundreds of parameters. In these high-dimensional spaces, the posterior geometry becomes highly irregular; for example, intrinsic pulsar red noise is often highly covariant with common stochastic signals, creating narrow, degenerate ridges. Standard MCMC algorithms frequently fail to mix across posterior modes, becoming trapped in local maxima.

To overcome this, parallel tempering extends MCMC by sampling a discrete ladder of tempered distributions that interpolate between the prior and the posterior. "Hot" chains flatten the posterior peaks, allowing the sampler to explore the parameter space freely, while swaps with "cold" chains allow new information about posterior location to permeate down to the target posterior chain [Geyer, 1991, Hansmann, 1997]. More generally, path sampling treats the interpolation between distributions as a continuous path, providing a framework for estimating ratios of normalizing constants through thermodynamic integration [Gelman and Meng, 1998]. This continuous formulation

also motivates extensions in which the inverse temperature is sampled jointly with the model parameters, rather than restricted to a fixed discrete ladder. While parallel tempering is standard in PTA literature for handling multimodal distributions, it remains computationally expensive.

Alternative exploration algorithms, such as Nested Sampling (NS) [Skilling, 2006] and Sequential Monte Carlo (SMC) [Chopin, 2002, Del Moral et al., 2006], are particularly attractive for model comparison because they simultaneously produce posterior samples and an estimate of the Bayesian evidence. Nested sampling achieves this by rewriting the evidence integral as an integral over enclosed prior mass, evolving a set of "live points" toward regions of increasing likelihood. Implementations like *MultiNest* and *dynesty* are widely used in astrophysics to navigate multimodal and degenerate posteriors [Feroz et al., 2009, 2013, Speagle, 2020], though their efficiency can degrade in high-dimensional problems. Similarly, SMC approximates the posterior by initializing a population of particles from the prior and moving them through a sequence of intermediate distributions. At each stage, these particles are reweighted, resampled, and mutated—commonly using MCMC transition kernels with adaptive temperature increments. Because SMC particle mutations can be carried out in parallel, it offers a computationally advantageous alternative on modern hardware while still estimating the marginal likelihood [Hinne, 2025].

The choice of sampler is therefore a crucial determinant for proper inference of parameters. Furthermore, regardless of the chosen method, the numerical approximation of the posterior must be rigorously checked. A biased sampler or an unconverged chain can lead to fundamentally incorrect scientific conclusions, even when the underlying mathematical model is perfectly specified.

2.3.4 Bayesian Model Comparison

Bayesian model comparison uses mainly the marginal likelihood to formally quantify how well a given model predicts the observed data. For a model M with parameter space Θ_M, the marginal likelihood (or Bayesian evidence) is the integral of the likelihood over the entire prior volume:

$$z(M) \equiv p(\boldsymbol{X}|M) = \int_{\Theta_M} p(\boldsymbol{X}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}. \tag{2.18}$$

Unlike maximum likelihood estimation, the evidence does not simply reward the best-fitting point in parameter space. Because it integrates over Θ_M, models with unnecessary parameterization are intrinsically penalized if much of their prior volume predicts the data poorly. This penalization is known as the Bayesian Occam's razor, making the evidence the most robust metric for comparing phenomenological models of differing dimensionality [MacKay, 2003].

For two competing models M_1 and M_2, the Bayes Factor (BF) is defined as the ratio of their marginal likelihoods:

$$BF_{12} = \frac{z(M_1)}{z(M_2)}. \tag{2.19}$$

The posterior odds are given by

$$\frac{p(M_1|\boldsymbol{X})}{p(M_2|\boldsymbol{X})} = BF_{12}\frac{p(M_1)}{p(M_2)}, \tag{2.20}$$

so the Bayes factor is the factor by which the data update the prior odds between two models [Jeffreys, 1998, Kass and Raftery, 1995]. If the prior model probabilities are equal, then the posterior odds reduce to the Bayes factor.

Evaluating Equation 2.18 constitutes the primary computational bottleneck in Bayesian inference. Because the integral spans the full prior support, its computational

cost scales exponentially with the dimensionality of Θ_M (the "curse of dimensionality"). While various numerical methods exist to estimate this quantity—such as path sampling , stepping-stone sampling, nested sampling, and SMC—each suffers from specific failure modes when applied to massive, highly covariant parameter spaces [Gelman and Meng, 1998, Xie et al., 2011, Skilling, 2006].

As highlighted in the following section, the physical extent of full-array PTA gravitational-wave searches pushes contemporary evidence estimators to their absolute limits. The failure of standard numerical integration techniques in these high-dimensional regimes necessitates the development of novel, highly scalable, and structurally aware evidence estimation methods.

2.4 PTA Bayesian framework

2.4.1 PTA Likelihood

The Bayesian PTA likelihood takes the post-fit timing residuals introduced in Eq. 2.7 as the data vector. In this stage, the timing models have been subtracted from the residuals, where the timing parameters are known with a finite precision. PTA analyses therefore linearize the deterministic timing model $\vec{\beta}$ around a reference ephemeris $\vec{\beta}_0$,

$$\vec{t}_{\rm det}(\vec{\beta}) = \vec{t}_{\rm det}(\vec{\beta}_0) + \mathbf{M}\vec{\epsilon}, \tag{2.21}$$

where $\mathbf{M}$ is the timing-model design matrix describing the partial derivatives of the timing model and $\vec{\epsilon}$ is the vector of small offsets from the reference timing parameters [Van Haasteren et al., 2009, Lentati et al., 2014b, Taylor, 2021]. This linearized term carries the remaining uncertainty in the deterministic ephemeris, while the stochastic part of the model accounts for measurement noise, pulsar-intrinsic processes, propagation effects, Solar-System effects, and possible gravitational waves.

For a single pulsar, the standard linearized residual model can be written as

$$\vec{\delta t} = \mathbf{M}\vec{\epsilon} + \sum_{\lambda} \mathbf{F}_{\lambda}\vec{a}_{\lambda} + \vec{n}, \tag{2.22}$$

where $\vec{n}$ is the white noise vector, and λ labels stochastic processes such as achromatic red noise, dispersion-measure variations, chromatic noise, a common uncorrelated red process, or a gravitational-wave background. The matrix $\mathbf{F}_{\lambda}$ contains the basis functions for process λ, usually a truncated Fourier basis, and $\vec{a}_{\lambda}$ contains the corresponding Gaussian-process weights. Defining

$$\vec{r} = \vec{\delta t} - \mathbf{M}\vec{\epsilon} - \sum_{\lambda} \mathbf{F}_{\lambda}\vec{a}_{\lambda}, \tag{2.23}$$

with $\vec{r}$ the residuals vector that converges to white noise $\vec{n}$ in ideal conditions. The PTA likelihood is usually assumed to be Gaussian,

$$p(\vec{\delta t}|\vec{\epsilon}, \{\vec{a}_{\lambda}\}, \vec{\eta}) = \frac{\exp\left[-\frac{1}{2}\vec{r}^{\top}\mathbf{N}^{-1}\vec{r}\right]}{\sqrt{\det(2\pi\mathbf{N})}}, \tag{2.24}$$

where $\mathbf{N}$ is the white-noise covariance matrix and $\vec{\eta}$ denotes the hyperparameters controlling the noise spectra and covariance amplitudes [Lentati et al., 2014b, van Haasteren and Vallisneri, 2014, Johnson et al., 2024]. In practice, the timing-model offsets and Fourier weights are often analytically marginalized, producing the marginalized likelihood used in modern PTA packages such as ENTERPRISE [Ellis et al., 2019, Taylor et al., 2021]. The following sub-sections describe the main components entering this likelihood, using the dispersion convention of Eq. 2.1 and the SGWB notation introduced in Section 2.2.2.

2.4.1.1 White noise

White noise describes TOA uncertainty that is uncorrelated between observing epochs and approximately flat in the timing-residual frequency domain. The dominant contribution is radiometer noise, which is typically estimated by the template-matching procedure using the noise in the off-pulse baseline, but additional short-timescale effects such as pulse jitter, radio-frequency interference, calibration errors, and pulse-shape mismatch can also contribute [Cordes and Shannon, 2010, Lam et al., 2018, Agazie et al., 2023a]. PTA analyses therefore inflate or supplement the formal TOA uncertainties using empirical parameters. For a backend or receiver system α, one common convention is

$$N_{ij,\alpha}^{\mathrm{WN}} = \left[E_\alpha^2 \sigma_i^2 + Q_\alpha^2\right] \delta_{ij} + J_\alpha^2 U_{ij}, \tag{2.25}$$

where E_α is EFAC, Q_α is EQUAD, J_α is ECORR, σ_i is the formal TOA uncertainty, and $U_{ij,\alpha} = 1$ only when TOAs i and j belong to the same observing epoch and system. EFAC rescales the quoted TOA errors, EQUAD adds an additional independent white-noise contribution in quadrature, and ECORR describes an epoch-correlated component that is fully correlated among frequency channels observed at the same epoch but uncorrelated between different epochs [Arzoumanian et al., 2015, Goncharov et al., 2021, Agazie et al., 2023a].

2.4.1.2 Achromatic red noise

Red noise denotes a time-correlated process whose power is concentrated at low Fourier frequencies. In an individual pulsar, achromatic red noise is often interpreted as spin noise or timing noise, since it does not depend on the observing radio frequency and may arise from rotational irregularities or other pulsar-intrinsic processes [Shannon and Cordes, 2010, Goncharov et al., 2021]. The distinction between intrinsic red noise, a common-spectrum uncorrelated process, and an isotropic tensor SGWB is therefore

made through the inter-pulsar covariance: no inter-pulsar correlation for intrinsic red noise, a common spectrum without spatial correlations for a CURN-like process, and the Hellings–Downs function of Eq. 2.13 for the SGWB.

In the Fourier-sum Gaussian-process representation, the red-noise covariance is

$$\mathbf{C}_{\mathrm{RN}} = \mathbf{F}_{\mathrm{RN}} \mathbf{\Phi}_{\mathrm{RN}} \mathbf{F}_{\mathrm{RN}}^{\mathsf{T}}, \tag{2.26}$$

with diagonal weight-space covariance

$$[\mathbf{\Phi}_{\mathrm{RN}}]_{\mu\nu} = S_{\mathrm{RN}}(f_\mu) \Delta f \, \delta_{\mu\nu}. \tag{2.27}$$

A standard choice is a power-law power spectral density,

$$S_{\mathrm{RN}}(f) = \frac{A_{\mathrm{RN}}^2}{12\pi^2} \left(\frac{f}{f_{\mathrm{yr}}} \right)^{-\gamma_{\mathrm{RN}}} \mathrm{yr}^3, \qquad f_{\mathrm{yr}} = 1\,\mathrm{yr}^{-1}, \tag{2.28}$$

where A_{RN} is the amplitude at $1\,\mathrm{yr}^{-1}$ and γ_{RN} is the spectral index [Lentati et al., 2014b, van Haasteren and Vallisneri, 2014, Goncharov et al., 2021, Johnson et al., 2024]. Modern analyses may also use a free-spectrum model, where each Fourier bin is assigned an independent amplitude. This reduces the assumptions imposed by a two-parameter power law, at the cost of introducing more parameters.

2.4.1.3 Dispersion-measure variations

Dispersion-measure (DM) variations are chromatic propagation delays caused by changes in the free-electron column density defined in Eq. 2.2. Since the leading cold-plasma delay follows the ν^{-2} scaling of Eq. 2.1, a stochastic DM process can be

represented using the same Fourier basis as red noise, multiplied by a radio-frequency-dependent factor. For TOA i observed at radio frequency ν_i,

$$F_{\mathrm{DM}}(f_k, t_i, \nu_i) = F(f_k, t_i)\left(\frac{\nu_i}{\nu_{\mathrm{ref}}}\right)^{-2}, \tag{2.29}$$

where ν_{ref} is commonly chosen to be $1.4\,\mathrm{GHz}$. The DM covariance is then

$$\mathbf{C}_{\mathrm{DM}} = \mathbf{F}_{\mathrm{DM}}\mathbf{\Phi}_{\mathrm{DM}}\mathbf{F}_{\mathrm{DM}}^{\mathsf{T}}, \qquad [\mathbf{\Phi}_{\mathrm{DM}}]_{\mu\nu} = S_{\mathrm{DM}}(f_\mu)\Delta f\,\delta_{\mu\nu}, \tag{2.30}$$

with $S_{\mathrm{DM}}(f)$ often modeled as a power law with amplitude A_{DM} and spectral index γ_{DM} [You et al., 2008, Lentati et al., 2014b, Taylor, 2021, Iraci et al., 2025].

Different PTA collaborations model the stochastic DM physics in different ways. NANOGrav commonly uses a piecewise-constant DMX model in which DM offsets are fitted in time windows as part of the timing model, while EPTA-style and PPTA-style analyses often include a stochastic DM Gaussian process in the likelihood [Arzoumanian et al., 2018, Taylor, 2021, Goncharov et al., 2021]. In either case, DM variations are important because they can covary with achromatic red noise and with the low-frequency gravitational-wave signal if the frequency coverage is insufficient. Low-frequency observations are therefore especially valuable, their strong ν^{-2} response helps constrain DM variations and disentangle them from achromatic red processes.

2.4.1.4 Solar-wind dispersion

The solar wind is another chromatic propagation contribution, produced by the ionized outflow from the Sun. Since it is a dispersive plasma effect, its timing delay uses the same cold-plasma scaling introduced in Eq. 2.1, but its amplitude depends strongly on the angular separation between the pulsar and the Sun and is most important for pulsars close to the ecliptic. A simple deterministic description writes the solar-wind dispersion

measure as a function of the pulsar–Sun–observer angle ρ,

$$\mathrm{DM_{SW}}(\rho) = n_\oplus \frac{\rho}{r_\oplus \sin\rho} [1\,\mathrm{AU}]^2, \tag{2.31}$$

where $n_\oplus$ is the electron density at $1\,\mathrm{AU}$ and $r_\oplus$ is the observatory–Sun distance [Hobbs et al., 2006a, Tiburzi et al., 2019, Hazboun et al., 2022, Iraci et al., 2025]. The corresponding timing delay is obtained by substituting $\mathrm{DM_{SW}}$ into Eq. 2.1.

The spherical, static solar-wind approximation is not sufficient for the highest-precision PTA data, because the solar wind varies over the solar cycle and differs between fast and slow streams. Recent low-frequency PTA analyses therefore include both a deterministic solar-wind component, controlled by $n_\oplus$, and a stochastic solar-wind Gaussian process, sometimes denoted $\mathrm{SW_{GP}}$, with a power-law spectrum

$$S_{\mathrm{SW}}(f) = A_{\mathrm{SW}}^2 \left(\frac{f}{f_{\mathrm{yr}}}\right)^{-\gamma_{\mathrm{SW}}} \mathrm{yr}^3. \tag{2.32}$$

This model is particularly relevant for low-frequency data from instruments such as LOFAR and NenuFAR, where the chromatic response to solar-wind dispersion is amplified [Hazboun et al., 2022, Susarla et al., 2024, Iraci et al., 2025].

2.4.1.5 Additional chromatic and phenomenological terms

Beyond achromatic red noise and ν^{-2} DM variations, PTA residuals can contain additional phenomenological terms that are motivated by instrumental effects, pulse-profile evolution, or interstellar propagation. A general chromatic Gaussian process can be written with

$$F_\chi(f_k, t_i, \nu_i) = F(f_k, t_i) \left(\frac{\nu_i}{\nu_{\mathrm{ref}}}\right)^{-\chi}, \tag{2.33}$$

where $\chi = 0$ corresponds to achromatic red noise, $\chi = 2$ to DM-like delays, and $\chi \simeq 4$ to a scattering-like chromatic process. The EPTA low-frequency analysis labels the

fixed-$\chi = 4$ process as CN4, while the PPTA analysis also considered cases where the chromatic index is inferred from the data [Goncharov et al., 2021, Iraci et al., 2025]. Such terms are useful because scattering, refraction, frequency-dependent DM, and imperfect wideband profile modeling can all produce residual chromatic structure that is not captured by a pure DM model [Shannon and Cordes, 2017, Lam et al., 2018, Agazie et al., 2023a].

Band noise and system noise are further phenomenological extensions. Instead of applying a red process to all TOAs of a pulsar, one applies it only to TOAs selected by observing band, receiver, backend, or another instrumental flag. These terms can absorb band-limited propagation effects, backend-dependent calibration errors, or other systematics that would otherwise bias the inferred achromatic red noise [Lentati et al., 2016, Goncharov et al., 2021].

Transient features are sometimes modeled as a jump with an amplitude that decays exponentially with time,

$$s_{\mathrm{E}}(t,\nu) = A_{\mathrm{E}} \exp\left[-\frac{t-t_0}{\tau_{\mathrm{E}}}\right] H(t-t_0) \left(\frac{\nu}{\nu_{\mathrm{ref}}}\right)^{-\chi_{\mathrm{E}}}, \tag{2.34}$$

where A_{E} is the initial amplitude of the jump, t_0 is the event epoch, τ_{E} is the decay timescale, H is a step function, and χ_{E} describes the chromaticity. Such exponentially decaying events have been associated with DM events, scattering changes, or magnetospheric/profile changes in individual pulsars such as PSR J1713+0747 and PSR J1643–1224 [Keith et al., 2013, Shannon and Cordes, 2017]. These additions illustrate a central feature of PTA likelihood construction: the model must be flexible enough to model pulsar-specific or instrument-specific noises.

2.4.1.6 Common-spectrum processes

The noise processes described above are modeled for each pulsar prior to the common spectrum searches. In the next step, PTA searches for an achromatic red process whose temporal spectrum is common to all pulsars. This search tests whether the array contains excess low-frequency power with a shared amplitude and spectral shape. In this sense, all extra shared low-frequency power are considered a common process while their characteristics differ depending on its physical source. Evidence for a shared red process is not by itself evidence for a gravitational-wave origin, since the same phenomenology could arise from unmodelled pulsar noise, clock errors, Solar-System ephemeris errors, or other low-frequency systematics [Arzoumanian et al., 2020, Taylor, 2021, Agazie et al., 2023b].

For a PTA with pulsars indexed by a, b and Fourier frequencies $f_i = i/T$, the covariance of achromatic Fourier coefficients can be written as

$$[\mathbf{\Phi}]_{(ai)(bj)} = \delta_{ij} \left[\delta_{ab}\, \phi_{ai} + \Gamma_{ab}\, \Phi_i\right], \tag{2.35}$$

where ϕ_{ai} describes the intrinsic red-noise spectrum of pulsar a, Φ_i is the spectrum of the common process, and Γ_{ab} is the Overlap Reduction Function (ORF) specifying the spatial covariance between pulsars. The same temporal spectrum can therefore be paired with different ORF hypotheses. The simplest spectrum is the monopole,

$$\Gamma_{ab}^{\mathrm{MON}} = 1, \tag{2.36}$$

which is associated with clock errors, while a dipole,

$$\Gamma_{ab}^{\mathrm{DIP}} = \cos \xi_{ab}, \tag{2.37}$$

is associated with Solar-System ephemeris errors [Tiburzi et al., 2016, Vallisneri et al., 2020, Taylor, 2021]. A common uncorrelated red-noise model, often denoted CURN, uses

$$\Gamma_{ab}^{\mathrm{CURN}} = \delta_{ab}, \tag{2.38}$$

so that each pulsar receives an independent realization of a red process with the same population-level spectrum. For SGWB, the Hellings–Downs correlation are described by:

$$\Gamma_{ab}^{\mathrm{HD}} = \frac{1}{2}\delta_{ab} + \frac{3}{2}x_{ab}\ln x_{ab} - \frac{1}{4}x_{ab} + \frac{1}{2}, \qquad x_{ab} = \frac{1-\cos\xi_{ab}}{2}, \tag{2.39}$$

where ξ_{ab} is the angular separation between pulsars a and b [Hellings and Downs, 1983]. Additional spatial model are used to model the SGWB such as the spline ORF, Binned ORF, and polynomial reconstructions through a Chebyshev or Legendre expansion [Antoniadis et al., 2023a, Agazie et al., 2023b, Taylor, 2021]

The standard spectral choice is a power law,

$$\Phi_i = S_{\mathrm{CP}}(f_i)\Delta f, \qquad \Delta f = \frac{1}{T}, \tag{2.40}$$

with

$$S_{\mathrm{CP}}(f) = \frac{A_{\mathrm{CP}}^2}{12\pi^2}\left(\frac{f}{f_{\mathrm{yr}}}\right)^{-\gamma_{\mathrm{CP}}} f_{\mathrm{yr}}^{-3}, \qquad f_{\mathrm{yr}} = 1\,\mathrm{yr}^{-1}. \tag{2.41}$$

For a stochastic background generated by a stationary ensemble of circular, gravitational-wave-driven supermassive black-hole binaries, the characteristic strain satisfies

$$h_c(f) = A_{\mathrm{GWB}}\left(\frac{f}{f_{\mathrm{yr}}}\right)^{\alpha}, \qquad \gamma_{\mathrm{GWB}} = 3 - 2\alpha, \tag{2.42}$$

so that $\alpha = -2/3$ corresponds to $\gamma_{\mathrm{GWB}} = 13/3$ [Phinney, 2001, Sesana, 2013, Taylor, 2021]. This fixed spectral index is often used as a fiducial SMBHB-motivated model, while allowing γ_{CP} to vary provides a less restrictive test of whether the data prefer a

steeper or shallower red process.

Power-law spectral models are compact and physically interpretable, but they cannot model intrinsic spectral structure. PTA analyses therefore also use free-spectrum models,

$$\Phi_i = \rho_i, \tag{2.43}$$

where each Fourier bin has an independent amplitude. Such models are useful for visualizing where the spectral power is concentrated and for diagnosing whether the power-law assumption is adequate [Arzoumanian et al., 2020, Agazie et al., 2023b]. The number of Fourier harmonics used to model a red noise process also must be chosen. In the NANOGrav 15-year analysis, intrinsic red noise is modelled with 30 frequency bins, whereas two-parameter common power-law models use only the lowest ~ 14 bins to avoid bias from frequencies where the common process falls below the white-noise level; free-spectrum models can include more bins because each frequency is assigned an independent amplitude [Agazie et al., 2023b, Johnson et al., 2024].

2.4.2 PTA Priors

2.4.2.1 Noise model priors

The prior structure in PTA inference follows the hierarchy of the likelihood. Timing-model offsets, white-noise parameters, Gaussian-process hyperparameters, chromatic indices, solar-wind parameters, and common-process parameters each require different prior choices. PTA analyses often compare noise models and gravitational-wave hypotheses through Bayes factors, the prior volume assigned to each component can directly affect model selection and the evidence for SGWB signal.

For the linearized timing-model offsets $\vec{\epsilon}$, the usual assumption is that the timing ephemeris is already tightly constrained by the timing analysis. The offsets are therefore given an effectively improper uniform prior, equivalently a zero-mean Gaussian prior

with very large or infinite variance,

$$p(\vec{\epsilon}) \propto \text{constant}, \qquad \vec{\epsilon} \sim \mathcal{N}(\vec{0}, \infty). \tag{2.44}$$

This prior is analytically convenient because the timing-model offsets can be marginalized together with the Gaussian-process weights [Van Haasteren et al., 2009, van Haasteren and Vallisneri, 2014, Taylor, 2021].

The Fourier coefficients of each stochastic process are assigned zero-mean Gaussian priors. For a process λ,

$$p(\vec{a}_\lambda|\vec{\eta}_\lambda) = \mathcal{N}\left(\vec{0}, \mathbf{\Phi}_\lambda(\vec{\eta}_\lambda)\right), \tag{2.45}$$

where $\vec{\eta}_\lambda$ contains the hyperparameters of the spectrum, for example $(A_\lambda, \gamma_\lambda)$. Thus the amplitude and spectral-index priors are priors on the covariance of the Gaussian process, not directly on the realization of the red-noise time series. For power-law red noise, NANOGrav prior-recovery tests use

$$\gamma \sim \mathcal{U}(0, 7), \qquad \log_{10} A \sim \mathcal{U}(-18, -11), \tag{2.46}$$

while the EPTA low-frequency analysis uses broad log-uniform amplitude ranges such as

$$\log_{10} A_{\rm RN},\, \log_{10} A_{\rm DM},\, \log_{10} A_{{\rm CN}\chi} \sim \mathcal{U}(-20, -11), \tag{2.47}$$

$$\gamma_{\rm RN},\, \gamma_{\rm DM},\, \gamma_{{\rm CN}\chi} \sim \mathcal{U}(0, 7). \tag{2.48}$$

The exact amplitude interval can differ between production analyses, intrinsic-noise studies, and gravitational-wave searches, but broad uniform priors in spectral index and log-uniform priors in amplitude are the standard choice for stochastic PTA processes [Johnson et al., 2024, Iraci et al., 2025].

White-noise priors are usually assigned separately for each receiver/backend system or observing band. A representative set of prior choices is

$$\mathrm{EFAC} \sim \mathcal{U}(0.01, 10), \tag{2.49}$$

$$\log_{10} \mathrm{EQUAD} \sim \mathcal{U}(-9, -5), \tag{2.50}$$

$$\log_{10} \mathrm{ECORR} \sim \mathcal{U}(-9, -2), \tag{2.51}$$

with EQUAD and ECORR expressed in seconds [Iraci et al., 2025] . In some workflows, white-noise parameters are first inferred in a single-pulsar noise analysis and then fixed to maximum-posterior or posterior-summary values in later, more expensive model-selection steps. This is a pragmatic computational choice, but it should be recognized as an approximation because white noise can covary with shallow red processes, band/system noise, and chromatic terms [Goncharov et al., 2021].

Chromatic indices may either be fixed or sampled. The fixed choices $\chi = 0$, 2, and 4 correspond respectively to achromatic red noise, DM-like variations, and a scattering-like CN4 process. When the chromatic index is inferred, a broad prior such as

$$\chi \sim \mathcal{U}(0, 10) \tag{2.52}$$

is used to allow the data to distinguish DM-like, scattering-like, and more general frequency scalings. Exponential events use priors on the event amplitude, timescale, epoch, and chromaticity, for example

$$\log_{10} A_{\mathrm{E}} \sim \mathcal{U}(-10, -2), \tag{2.53}$$

$$\log_{10} \tau_{\mathrm{E}} \sim \mathcal{U}(0, 2.5), \tag{2.54}$$

$$t_{0,\mathrm{E}} \sim \mathcal{U}(t_{\min}, t_{\max}), \tag{2.55}$$

$$\chi_{\mathrm{E}} \sim \mathcal{U}(0, 10), \tag{2.56}$$

where the allowed epoch range is chosen around the candidate event [Goncharov et al., 2021, Iraci et al., 2025].

Solar-wind priors depend on whether the analysis uses only a deterministic solar-wind density or also a stochastic solar-wind Gaussian process. In the EPTA low-frequency analysis, the deterministic density and stochastic SWGP parameters were assigned priors of the form

$$n_{\oplus} \sim \mathcal{U}(0, 30), \tag{2.57}$$

$$\log_{10} A_{\rm SW} \sim \mathcal{U}(-10, 1), \tag{2.58}$$

$$\gamma_{\rm SW} \sim \mathcal{U}(-2, -1). \tag{2.59}$$

These priors reflect the fact that solar-wind modelling is strongly informed by ecliptic latitude, observing cadence, radio-frequency coverage, and the availability of low-frequency data [Hazboun et al., 2022, Susarla et al., 2024, Iraci et al., 2025].

2.4.2.2 Common-spectrum priors.

The prior choices for common-spectrum processes are specified separately from the choice of spatial correlation. For a fair comparison between CURN, Hellings–Downs, monopole, and dipole hypotheses, the temporal-spectrum priors are usually kept identical while only the overlap-reduction function is changed. For a power-law common process with a variable spectral index, a standard choice is

$$\log_{10} A_{\rm CP} \sim \mathcal{U}(-18, -11), \qquad \gamma_{\rm CP} \sim \mathcal{U}(0, 7), \tag{2.60}$$

where $A_{\rm CP}$ denotes the common-process strain amplitude at $f_{\rm yr} = 1\,{\rm yr}^{-1}$. These standard priors on $\gamma_{\rm CP}$ leads to a divergence in the high frequency power for $\gamma_{\rm CP} < 1$. A more sensible prior , such as $\gamma_{\rm CP} \sim \mathcal{U}(1, 7)$, prevents placing weights on unphysical signals.

When the spectrum is fixed to the SMBHB-motivated value $\gamma_{\rm GWB} = 13/3$, the amplitude prior may be narrowed, for example

$$\log_{10} A_{\rm GWB} \sim \mathcal{U}(-18, -14), \qquad \gamma_{\rm GWB} = \frac{13}{3}. \tag{2.61}$$

For a free-spectrum common process, the independent power coefficients are assigned priors such as

$$\log_{10} \rho_i \sim \mathcal{U}(-18, -8), \qquad i = 1, \ldots, N_{\rm CP}, \tag{2.62}$$

with the number of common-spectrum frequency bins chosen according to the spectral model and the frequency range over which the common process is expected to dominate over white noise [Agazie et al., 2023b, Johnson et al., 2024].

Priors are also required when the spatial correlation itself is modeled flexibly. A binned or spline reconstruction assigns independent parameters to the correlation values at selected angular separations, while polynomial reconstructions use coefficients in a Chebyshev or Legendre expansion. Typical choices use broad uniform priors on the normalized correlation coefficients, for example

$$y_k \sim \mathcal{U}(-0.9, 0.9), \tag{2.63}$$

for spline knots, or

$$c_k \sim \mathcal{U}(-1, 1), \qquad \ell_k \sim \mathcal{U}(-1, 1), \tag{2.64}$$

for Chebyshev and Legendre coefficients [Gair et al., 2014, Antoniadis et al., 2023a]. Consequently, Bayes factors between HD and CURN models should be interpreted as evidence for the spatial correlation pattern only after accounting for the spectral model, the number of Fourier bins, and the prior volume assigned to competing common-process hypotheses.

2.4.3 PTA Samplers

The posterior distributions encountered in PTA analyses differs from low to high dimensional, strongly correlated, and often multimodal. Even after analytic marginalization over the linear timing-model offsets and Gaussian-process Fourier coefficients introduced in Section 2.4.1, a full PTA gravitational-wave search may still contain hundreds of hyperparameters describing intrinsic pulsar red noise, white noise, chromatic propagation effects, common uncorrelated red processes, and spatially correlated gravitational-wave signals. This section focuses on how these computational ideas are used in PTA analyses.

The most widely used posterior-exploration algorithm in current PTA analyses is Markov chain Monte Carlo, implemented through packages such as PTMCMC sampler and used together with the likelihood-building tools in ENTERPRISE and ENTERPRISE_EXTENSIONS [Ellis and van Haasteren, 2017, Ellis et al., 2019, Taylor et al., 2021]. In this setting the target density is the posterior of Eq. 2.14, with the PTA likelihood of Eq. 2.24 or its analytically marginalized form. Parallel tempering is commonly used for the reasons described in Section 2.3.3: hot chains explore the prior volume more freely, while the cold chain samples the posterior relevant for inference. This is particularly important in searches that compare spatially correlated and common uncorrelated red processes, or in single-pulsar noise analyses where chromatic and achromatic red processes may be partially degenerate [Geyer, 1991, Taylor, 2021, Johnson et al., 2024].

The efficiency of these samplers depends strongly on the proposal distributions. In PTMCMC sampler, the default proposal mixture includes Adaptive Metropolis (AM), Single-Component Adaptive Metropolis (SCAM), and Differential Evolution (DE) proposals. AM uses the empirical covariance accumulated from the chain history to propose jumps along correlated parameter directions, while SCAM moves along

one principal direction at a time, which is useful in high-dimensional spaces where full-dimensional jumps may be rejected too often [Haario et al., 2001, 2005]. DE uses differences between previous samples to build jumps adapted to the posterior geometry [Braak, 2006]. In PTA applications these default proposals are supplemented by prior draws, empirical proposals, and problem-specific jumps for common-process amplitudes, spectral indices, and pulsar-by-pulsar red-noise parameters [Taylor, 2021, Johnson et al., 2024].

A useful practical feature of PTA samplers is the grouping of strongly correlated parameters. For example, the amplitude and spectral index of a power-law red process are often proposed together, and separate proposal groups may be constructed for each pulsar's intrinsic red-noise parameters. Empirical distributions derived from single-pulsar noise analyses can also be used as proposal distributions in full-array gravitational-wave searches. This two-stage strategy reflects the structure of the PTA problem: single-pulsar analyses first identify and constrain the most important intrinsic noise components, while the full-array analysis then searches for a common process and possible inter-pulsar correlations.

Nested sampling has also been used in PTA analyses, especially for single-pulsar noise-model comparison and evidence cross-checks. In the PPTA DR2 noise analysis, for example, PTMCMC sampler was used for parameter estimation, while PolyChord-Lite was used to compute model evidences through nested sampling [Skilling, 2006, Handley et al., 2015]. Earlier PTA tools such as TempoNest also used nested sampling for Bayesian pulsar-timing inference [Lentati et al., 2014a]. The evidence-estimation role of nested sampling is discussed again in Section 2.4.4; here the important point is that its practical usefulness is limited by the dimensionality and geometry of the PTA likelihood.

Sampler validation is essential because PTA conclusions can depend on subtle

differences between noise models. Recent NANOGrav pipeline tests include prior-recovery tests, injection-recovery tests, convergence diagnostics, and comparisons between different Bayes-factor estimators [Johnson et al., 2024]. These checks are not merely technical, but they ensure that model preference is driven by the timing data and likelihood assumptions, rather than by poor mixing, insufficient exploration, or numerical differences between samplers.

2.4.4 PTA Evidence

Using the notation of Section 2.3.4, PTA evidence calculations evaluate $z(M)$ for competing timing-array models with $X = \vec{\delta t}$. These models may differ by the presence or absence of intrinsic pulsar red noise, chromatic noise, solar-wind terms, a common uncorrelated red process, or a Hellings–Downs-correlated gravitational-wave background. The Bayes factor BF_{ij} of Eq. 2.19 is therefore not only a numerical summary of fit quality; it is a test of which stochastic processes are required by the data after averaging over the prior volume assigned to each model.

The dominant model-comparison method in recent NANOGrav gravitational-wave analyses is product-space sampling, usually referred to in the PTA literature as the *HyperModel* method. The idea is to embed several competing models inside a single hypermodel $\mathcal{H}_*$ with a discrete model indicator, commonly denoted *nmodel*. The hypermodel parameter vector is the union of the unique parameters appearing in all candidate models,

$$\theta_* = \{\theta_1, \theta_2, \ldots, \theta_K\}, \qquad m \in \{1, \ldots, K\}, \tag{2.65}$$

but at any given likelihood evaluation only the parameters associated with the active model m contribute to the likelihood. The inactive parameters remain part of the chain but are not constrained by that likelihood evaluation [Carlin and Chib, 1995, Johnson

et al., 2024].

The marginal posterior probability of the model indicator is

$$p(n_i|\vec{\delta t}, \mathcal{H}_*) = \int p(\boldsymbol{\theta}_*, n_i|\vec{\delta t}, \mathcal{H}_*)\, \mathrm{d}\boldsymbol{\theta}_*. \tag{2.66}$$

For two models M_i and M_j, the posterior odds are obtained from the fraction of posterior samples assigned to each model,

$$\mathcal{O}_{ij} = \frac{p(n_i \mid \vec{\delta t}, \mathcal{H}_*)}{p(n_j \mid \vec{\delta t}, \mathcal{H}_*)} \tag{2.67}$$

$$= \frac{p(n_i \mid \mathcal{H}_*) z(M_i)}{p(n_j \mid \mathcal{H}_*) z(M_j)}. \tag{2.68}$$

If the prior probabilities assigned to the models are equal, then the posterior odds reduce to the Bayes factor,

$$BF_{ij} = \frac{p(n_i \mid \vec{\delta t}, \mathcal{H}_*)}{p(n_j \mid \vec{\delta t}, \mathcal{H}_*)}. \tag{2.69}$$

Uncertainties are commonly estimated by bootstrapping the thinned posterior samples of the model indicator, e.g., when one model is strongly favoured, the sampler may spend too little time in the disfavoured model to estimate the odds accurately. In this case, a constant log-weight can be added to one model's log-likelihood to improve mixing across the model index, and the weight is then removed in post-processing [Johnson et al., 2024] .

This hypermodel approach is attractive for PTA analyses because it directly compares physically distinct hypotheses within the same sampling run. For example, one may compare a common-spectrum uncorrelated red-noise process against a Hellings–Downs-correlated gravitational-wave background, or compare a pulsar noise model with and without an additional chromatic component. It also provides a natural framework for model averaging, since the posterior model probabilities are explicitly sampled.

However, the method is not cost-free: it requires good mixing not only within each model, but also across the model index. Poor transitions between models can bias the Bayes factor, especially when posterior support is concentrated in very different regions of parameter space.

Reweighting provides another way to estimate model probabilities when posterior samples from an approximate model are already available. Suppose samples are drawn from an approximate posterior $q(\boldsymbol{\theta})$ associated with a model M_q, but one wishes to estimate expectations or Bayes factors for a target model M_t. If the two distributions have overlapping support, each sample can be assigned a weight

$$w(\boldsymbol{\theta}) = \frac{p(\vec{\delta t} \mid \boldsymbol{\theta}, M_t) p(\boldsymbol{\theta} \mid M_t)}{p(\vec{\delta t} \mid \boldsymbol{\theta}, M_q) p(\boldsymbol{\theta} \mid M_q)}, \tag{2.70}$$

and the weighted samples approximate the target posterior. In PTA applications, this can be useful when the approximate model is much faster to evaluate than the target model, for example when reweighting common-spectrum uncorrelated red-process samples toward a spatially correlated Hellings–Downs model [Hourihane et al., 2023, Johnson et al., 2024]. The limitation is that the method fails when the approximate and target posteriors have poor overlap, since a small number of samples then dominate the weights and the effective sample size becomes too small.

For nested hypotheses, the Savage–Dickey density ratio gives a particularly simple Bayes-factor estimator. If model M_1 contains an additional signal controlled by an amplitude parameter A, and model M_0 is recovered at $A = A_0$, then

$$BF_{10} = \frac{p(A = A_0 | M_1)}{p(A = A_0 | \vec{\delta t}, M_1)}. \tag{2.71}$$

In PTA gravitational-wave searches, the switch-off point is often taken to be the lower sampling boundary of the common-process amplitude, for example $\log_{10} A_{\rm low} \simeq -18$,

chosen so that the corresponding signal is much smaller than the intrinsic pulsar noise. This method has been widely used to estimate the evidence for a fixed-spectral-index common uncorrelated red process or gravitational-wave background [Dickey, 1971, Taylor, 2021, Agazie et al., 2023b]. Its interpretation must be treated carefully, however, because evidence for a non-zero common amplitude is not automatically evidence for the Hellings–Downs spatial correlations that identify the signal as gravitational-wave in origin.

Nested sampling estimates the evidence more directly by transforming the multi-dimensional evidence integral into a one-dimensional integral over the enclosed prior volume,

$$z(M) = \int_0^1 \mathcal{L}(X_{\rm NS})\,\mathrm{d}X_{\rm NS}, \tag{2.72}$$

where $\mathcal{L}$ is the likelihood as a function of the prior volume, and $X_{\rm NS}$ denotes the prior volume enclosed by a likelihood contour and is distinct from the data symbol X used in Section 2.3.1. The algorithm evolves a set of live points from the prior toward regions of increasing likelihood and accumulates the evidence from the discarded points [Skilling, 2006]. In PTA analyses, nested sampling has been used in tools such as TEMPONEST and POLYCHORDLITE for pulsar-noise studies and evidence cross-checks [Lentati et al., 2014a, Goncharov et al., 2021]. For full-array gravitational-wave searches, however, the large number of pulsars and noise parameters makes nested sampling expensive, so it is often more practical for reduced models, single-pulsar analyses, or validation of hypermodel Bayes factors.

2.5 Summary and Limitations

The literature reviewed in this chapter demonstrates that the search for gravitational waves using Pulsar Timing Arrays is fundamentally an exercise in precision Bayesian

model comparison. Pulsar timing residuals are contaminated by a complex combination of stochastic processes—including white noise, intrinsic red noise, chromatic propagation effects, and solar-wind dispersion—which must be carefully disentangled from common uncorrelated red processes and potential Hellings–Downs correlated gravitational-wave signals. Consequently, the central challenge extends beyond mere signal detection; it requires identifying the optimal combination of stochastic processes supported by the data while robustly marginalizing over parameter uncertainty. The Bayesian evidence, thus emerges as an indispensable metric for both single-pulsar noise characterization and full-array gravitational-wave inference.

However, as highlighted in Section 2.4.4, contemporary evidence estimation strategies suffer from significant methodological limitations that restrict their utility at scale. HyperModel and product-space approaches are frequently hindered by poor mixing across model indices, while reweighting techniques rely heavily on a strong overlap between approximate and target posteriors. Furthermore, Savage–Dickey density ratio estimators are strictly constrained to nested hypotheses, and traditional nested sampling becomes computationally prohibitive for the high-dimensional likelihoods characteristic of modern PTAs. As datasets expand in both size and complexity, these algorithmic constraints create a severe computational bottleneck, ultimately impeding the systematic propagation of noise-model uncertainty into broader gravitational-wave searches.

This thesis directly addresses this bottleneck by introducing highly scalable, fast and accurate evidence estimation methods. Consequently, these methods enables the exploration of the noise model hypothesis landscape of PTA. These contributions establish a more scalable, model-aware, and statistically reliable framework for precision inference in future gravitational-wave-background analyses.

Chapter 3

Manuscript 1

Generalized Steppingstone Sampling

This manuscript presents generalized steppingstone sampling (GSS) as an efficient marginal likelihood estimator for Bayesian model selection in pulsar timing array analyses. Our work addresses the high computational cost and limited flexibility of existing evidence-estimation approaches in the PTA, particularly when comparing expensive and high dimensional models. By using posterior-informed reference distributions, GSS provides accurate and low-cost estimates of marginal likelihoods and Bayes factors, enabling robust comparisons of PTA noise and gravitational-wave models. On 11^{th} June 2025, the manuscript was published in the *Monthly Notices of the Royal Astronomical Society* under the title *Generalized Steppingstone Sampling: Efficient marginal likelihood estimation in gravitational wave analysis of Pulsar Timing Array data*.

Manuscript details

Title Generalized Steppingstone Sampling: Efficient marginal likelihood estimation in gravitational wave analysis of Pulsar Timing Array data.

Authors El Mehdi Zahraoui, Patricio Maturana-Russel, Willem van Straten, Renate Meyer, and Sergei Gulyaev.

Journal Monthly Notices of the Royal Astronomical Society.

Doi `https://doi.org/10.1093/mnras/staf953`.

Abstract

Globally, Pulsar Timing Array (PTA) experiments have revealed tentative evidence of a stochastic gravitational wave background (SGWB) signal in PTA data. Apart from acquiring more observations, the sensitivity of PTA experiments can be increased by improving the accuracy of noise modeling. In PTA data analysis, noise modeling is conducted primarily using Bayesian statistics, relying on the marginal likelihoods and Bayes factors to assess evidences. We introduce generalized steppingstone sampling (GSS) as an efficient and accurate marginal likelihood estimation method for the PTA-Bayesian framework. This method enables low-cost estimates with high accuracy, especially when comparing expensive models such as the Hellings-Downs (HD) model or the overlap reduction function model (ORF). We demonstrate the efficiency and the accuracy of GSS for model selection and evidence calculation by reevaluating the evidence of previous analyses from the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) 15 yr data set and the European PTA (EPTA) second data release. We find similar evidence for the GWB compared to the one reported by the NANOGrav 15-year data set. Compared to the evidence reported for the EPTA second data release, we find a substantial increase in evidence for a GWB signal across all data sets.

Keywords

pulsar – gravitational waves – methods: statistical – methods: data analysis

3.1 Introduction

The pulsar timing array (PTA) emerged from the concept of a single pulsar-Earth baseline to observe gravitational waves (GW) in the low-frequency nanohertz band [Sazhin, 1978]. The PTA experiments were designed as a network of pulsars that could unveil the influence of GWs on the arrival times of their pulses in the form of quadrupolar correlations [Hellings and Downs, 1983]. Millisecond pulsars (MSPs) became the key to achieving this objective due to their long-term stability, rivaling Earth's atomic clocks [Matsakis et al., 1997]. Since the initial MSP timing analysis [Stinebring et al., 1990], the red noise (RN) in the MSPs' dataset has pointed to a promising future as observational radio astronomy and computational capability constantly improve. The next step for the PTA community focused on gathering more data and refining the pulsar timing model to test and derive a realistic sensitivity threshold to observe GWs [Verbiest et al., 2009, Cordes and Shannon, 2010, Hazboun et al., 2019]. Recently, major PTA collaborations such as the North American Nanohertz Observatory for Gravitational Waves [NANOGrav; McLaughlin, 2013a], the Australian Parkes PTA [PPTA; Manchester et al., 2013], the European PTA [EPTA; Ferdman et al., 2010], the Indian PTA [InPTA; Joshi et al., 2018], the MeerKAT PTA [MeerKAT; Miles et al., 2022], and the Chinese PTA [CPTA; Xu et al., 2023] reported positive statistical support for gravitational wave background (GWB) in the latest data releases [Reardon et al., 2023b, Agazie et al., 2023b, Antoniadis et al., 2023a]. The PTA community has established a rigorous PTA-Bayesian framework [Van Haasteren et al., 2009], consistently expanding and acquiring new numerical and statistical techniques

[Lentati et al., 2014a, Johnson et al., 2024]. This framework has enabled a simultaneous fit for different noise models and GWB signals while having the ability to compare the evidence of each different model [van Haasteren and Vallisneri, 2014]. Before evaluating the GWB evidence, it was proven that a common red signal (CRS) is evident across different PTA datasets: NANOGrav [Arzoumanian et al., 2020], EPTA [Chen et al., 2021], and PPTA [Goncharov et al., 2022]. These findings resulted in a surge of interest in defining the nature of these CRSs and potentially revealing the buried GWB signal. Through this process, an extensive number of models were used to mitigate different sources of red noise that can affect the PTA's sensitivity either from pulsar contribution such as pulsar spin noise [Verbiest et al., 2009] and dispersion measure (DM) variations [Jones et al., 2017], or other noise sources such as spatial correlations due to the solar system ephemeris and the atomic clocks used in this experiment [Tiburzi et al., 2016]. To test and account for all noise models, the full PTA-Bayesian analysis results in a large model space. Down the line, incorporating new MSP datasets and updated noise models will further expand this space significantly.

In the Bayesian framework, hypotheses comparison is done by evaluating the *Bayes factor*, which is the ratio of *marginal likelihoods* of different models. This criterion is used to assess the evidence for each hypothesis [Newton and Raftery, 1994]. For the PTA, model selection and evidence estimation are currently carried through various Bayesian techniques [Agazie et al., 2023b, Antoniadis et al., 2023a]. Among these methods, Thermodynamic Integration (TI) [Lartillot and Philippe, 2006] stands out as the cornerstone method for marginal likelihood estimation. In parallel, nested sampling [Skilling, 2006] is implemented to provide not only estimates of marginal likelihood but also to generate posterior samples. Additionally, reweighting [Hourihane et al., 2023] was defined to estimate the marginal likelihood for an expensive model using an approximately matching model. Finally, the Hypermodel method, implemented using the Savage-Dickey density ratio approximation as a form of product space sampling for

nested models [Carlin and Chib, 1995, Lodewyckx et al., 2011], is routinely used across the whole PTA analyses. Johnson et al. [2024] discuss these methods in the context of the PTA, where nested sampling is used for models with small parameter spaces. In this case, nested sampling yields less accurate estimates with a similar cost to the Hypermodel method. In parallel, TI estimates become expensive for large PTA models and can lead to biased estimates if the required computational power is lacking. Consequently, TI is employed only where the Hypermodel fails to provide evidence estimates. Although the Hypermodel method was established as the predominant method, it is limited to nested models and requires assigning arbitrary weights to each model to avoid the Markov chain getting trapped in a particular model. Alternatively, we introduce generalized steppingstone sampling [GSS; Fan et al., 2011] as a new method to GW astronomy and in particular the PTA-Bayesian framework. GSS is intended to bypass most of the current issues with marginal likelihood estimation and provide an accurate marginal likelihood estimate at a lower computational price. This article is structured as follows. Section 3.2 outlines model selection through the Bayes factor and briefly defines the TI, SS, and GSS methods. Section 3.3 demonstrates a comparison between these methods' performance applied to a Gaussian model. Thereafter, section 3.4 first discusses the implementation of GSS within the PTA framework. Then, our application to different PTA analyses will revisit the GWB evidence and will demonstrate the benefits of marginal likelihood estimation in making a robust model selection and monitoring the evidence on different CRS models. Finally, section 3.5 will discuss the prospects of GSS in the PTA-Bayesian framework and the significance of different PTA application results.

3.2 Marginal likelihood estimation

The Bayesian framework is based on *Bayes'* theorem, which is given by

$$p(\boldsymbol{\theta}|X, M) = \frac{L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{z(X|M)} \tag{3.1}$$

for a model M and dataset X, where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $p(\boldsymbol{\theta}|X, M)$ the posterior probability density of $\boldsymbol{\theta}$, $L(X|\boldsymbol{\theta}, M)$ the likelihood function and $\pi(\boldsymbol{\theta}|M)$ the prior density. The constant $z(X|M)$, commonly denoted as z, is the *marginal likelihood*. The marginal likelihood is a multi-dimensional integral over the parameter space Θ defined by

$$z = \int_{\Theta} L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}, \tag{3.2}$$

where $\pi(\boldsymbol{\theta}|M)$ is a proper prior, i.e., $\int_{\Theta} \pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} = 1$. The integral z is a prior weighted average of the likelihood which quantifies the goodness of fit of model M to data X. For a set of two models M_1 and M_2 fitted to a dataset X, the marginal likelihoods $z(X|M_1)$ and $z(X|M_2)$ can be used to compare these models such that the model with the higher marginal likelihood is preferred. Hence, the *Bayes factor* (BF), defined by

$$\mathrm{BF}_{(M_1/M_2)} = \frac{z(X|M_1)}{z(X|M_2)}, \tag{3.3}$$

is used as a Bayesian criterion to compare M_1 to M_2 and perform model selection. Specifically, the *Bayes factor* is equivalent to the ratio of the posterior distributions of the models when the prior probabilities for the models are equal. In contrast, alternative performance tests are used to compare models such as the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). Although the BIC and the AIC penalize unnecessary complexity by applying a uniform penalty for each parameter, these approaches ignore the importance of priors in the Bayesian analysis. The Bayes

factor allows for tailored penalties based on the chosen priors which helps to ensure that overly complex models are penalized appropriately, reducing the risk of overfitting.

Apart from simple conjugate cases, the marginal likelihood has no analytical solution. In practice, considerable efforts have been made to numerically approximate z through Markov chain Monte Carlo (MCMC) methods such as the harmonic mean [Newton and Raftery, 1994], bridge sampling [Meng and Wong, 1996], and nested sampling [Skilling, 2006]. Path sampling [Gelman and Meng, 1998] or thermodynamic integration [Lartillot and Philippe, 2006] was the cornerstone technique of what would be later known as the method of power posterior [Friel and Pettitt, 2008]. This method defines a geometric path that connects the prior with the posterior through the power posterior densities

$$p_\beta(\boldsymbol{\theta}|X, M) = \frac{L(X|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta} \quad \text{with} \quad 0 \leq \beta \leq 1, \tag{3.4}$$

where β is the inverse temperature for the power posterior density p_β. When $\beta = 0$, the density p_β is equal to the proper prior distribution, which gives $z_0 = 1$. When $\beta = 1$, the density p_β is equal to the posterior distribution resulting in $z_1 = z$. Correspondingly, thermodynamic integration relies on the identity

$$\log(z) = \int_0^1 E_{p_\beta}[\log(L(X|\boldsymbol{\theta}, M))]\mathrm{d}\beta, \tag{3.5}$$

where E_{p_β} denotes the expected value with respect to p_β. For a sequence of β values, samples from each power posterior can be used to estimate the expected values by the sample averages. Thus, the integral in Equation (3.5) can be approximated using any technique for numerical integration, e.g., the trapezoidal rule. Thermodynamic integration can yield a good marginal likelihood estimate; nevertheless, it requires a high computational expense to produce a single estimate. This computational expense

obliged modifications to the method to reduce the estimation cost. Steppingstone sampling was proposed as an attempt to reduce TI's cost, which resulted in a far more efficient and accurate method , i.e., GSS, which is considered one of the most accurate marginal likelihood estimation methods.

3.2.1 Steppingstone Sampling

The approximation of a continuous integral induces a discretization bias in TI leading to an inaccurate estimate of the marginal likelihood. Hence, Steppingstone Sampling (SS) was first proposed in the phylogenetics field [Xie et al., 2011] and later introduced to LIGO GW analysis [Maturana-Russel et al., 2019b] as an alternative method to enable unbiased z estimates and improve model selection. The SS method considers an unnormalized version of the power posterior density, given in (3.4), defined by

$$q_\beta = L(X|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M). \tag{3.6}$$

This density is normalized by z_β yielding a normalized power posterior density

$$p_\beta = \frac{q_\beta}{z_\beta}, \text{ with } z_\beta = \int_\Theta q_\beta \mathrm{d}\boldsymbol{\theta}. \tag{3.7}$$

Note that z_0 represents the normalizing constant of a proper prior, which is equal to 1, whereas z_1 is the normalizing constant of the posterior, i.e., the marginal likelihood. Taking this into account, SS redefines the marginal likelihood as a product of $K-1$ intermediate ratios given by

$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}} \frac{z_{\beta_2}}{z_{\beta_1}} \ldots \frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}} = \prod_{k=1}^{K-1} \frac{z_{\beta_k}}{z_{\beta_{k-1}}}, \tag{3.8}$$

where $\beta_0 = 0 < \cdots < \beta_{k-1} < \beta_k < \cdots < \beta_{K-1} = 1$. Importance sampling can be used to estimate each of the ratios $r_k = z_{\beta_k}/z_{\beta_{k-1}}$, where $p_{\beta_{k-1}}$ acts as an excellent importance sampling distribution when $p_{\beta_{k-1}}$ is similar to p_{β_k}, which occurs for a sufficient large K. It can be shown that the ratios r_k can be expressed as

$$r_k = E_{p_{\beta_{k-1}}}\left[\left(L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\right)^{\beta_k-\beta_{k-1}}\right]. \tag{3.9}$$

In practice, these ratios are estimated using an unbiased Monte Carlo (MC) estimator for each ratio r_k, given by

$$\widehat{r}_k = \frac{1}{n}\sum_{i=1}^{n} L(X|\boldsymbol{\theta}_{k-1,i}, M)^{\beta_k-\beta_{k-1}}. \tag{3.10}$$

Therefore, the SS estimate of the marginal likelihood is calculated by

$$\widehat{z} = \prod_{k=1}^{K-1}\widehat{r}_k = \prod_{k=1}^{K-1}\frac{1}{n}\sum_{i=1}^{n} L(X|\boldsymbol{\theta}_{k-1,i}, M)^{\beta_k-\beta_{k-1}}, \tag{3.11}$$

with $\boldsymbol{\theta}_{k-1,i}$ samples drawn from $p_{\beta_{k-1}}$ and n the number of drawn samples. To avoid numerical errors, it is preferable to work with $\log\widehat{z}$. However, this transformation introduces a bias to the $\widehat{z}$ estimate which can be overcome by increasing K, the number of β chains that corresponds to n samples from $p_\beta(\boldsymbol{\theta}|X, M)$ [Xie et al., 2011]. The choice of the β values and their dispersion can enhance the efficiency of SS and even more for TI [Maturana-Russel et al., 2019b, Xie et al., 2011]. For β_k defined as

$$\beta_k = \frac{k}{K}100\% \text{ quantile of Beta}(\alpha, 1) \tag{3.12}$$

for $k = 1, \ldots, K-1$, it has been shown that the efficiency can be increased with $\alpha = 0.3$ for both methods [Xie et al., 2011]. In general, the prior is more diffuse than the posterior, leading to much greater differences between consecutive power posterior

densities near the prior, i.e., for β close to 0. Consequently, the SS method exhibits greater differences between consecutive ratios $\widehat{r}_k$ when β is close to 0. As a result, more power posterior densities are required near the prior, meaning that more β values close to 0 are needed, which is effectively achieved by the quantiles of the Beta(0.3, 1) distribution.

3.2.2 Generalized Steppingstone Sampling

The performance of the steppingstone sampling reinforced the interest in decreasing the cost of marginal likelihood estimation while maintaining high accuracy. The generalized SS was proposed by Fan et al. [2011] to achieve this objective. The GSS method introduces an additional distribution, referred to as the reference distribution π_0, which should be an approximation of the posterior. This constrains the parameter space explored by the MCMC method, making GSS more efficient than SS. In practice, the reference distribution can be defined as the product of convenient distributions tuned by posterior samples. GSS introduces the reference distribution π_0 in the unnormalized posterior given in (3.6), redefining it as follows

$$q_\beta = [L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)]^\beta [\pi_0(\boldsymbol{\theta}|M)]^{1-\beta}. \tag{3.13}$$

Thus, the power posterior in (3.7) defines a path between the reference distribution and the posterior. Note that for $\beta = 0$, p_β is equal to the reference distribution, while for $\beta = 1$, p_β is equal to the posterior. The product of ratios that defines the SS estimator in (3.8) remains the same in the GSS estimator; however, it can be shown that the ratios are now defined as follows

$$r_k = E_{p_{\beta_{k-1}}}\left[\left(\frac{L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{\pi_0(\boldsymbol{\theta}|M)}\right)^{\beta_k - \beta_{k-1}}\right]. \tag{3.14}$$

The MC estimator of the ratio r_k is given by

$$\widehat{r}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{L(X|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\pi_0(\boldsymbol{\theta}_{k-1,i}|M)}\right)^{\beta_k-\beta_{k-1}}, \tag{3.15}$$

with n samples $\boldsymbol{\theta}_{k-1,i}$ drawn from $p_{\beta_{k-1}}$.

Finally, the marginal likelihood is estimated by multiplying over the estimated ratios as follows

$$\widehat{z} = \prod_{k=1}^{K-1}\widehat{r}_k = \prod_{k=1}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{L(X|\boldsymbol{\theta}_{k-1,i}, M)\pi(\boldsymbol{\theta}_{k-1,i}|M)}{\pi_0(\boldsymbol{\theta}_{k-1,i}|M)}\right)^{\beta_k-\beta_{k-1}}. \tag{3.16}$$

The MC estimator for each ratio r_k provides an unbiased estimate of the ratio. However, this does not imply that their product is unbiased unless the estimated ratios are independent. In practice, the interaction between Markov chains in the parallel tempering algorithm in the GSS case is unnecessary due to prior knowledge of the posterior distribution. As a result, independent Markov chains can be generated, leading to independent estimated ratios and thereby eliminating bias in the GSS estimate.

The uncertainty of the GSS estimate can be approximated using the delta method–as proposed in the SS case [Xie et al., 2011]–which relies on the assumption that the samples are independent. In practice, this assumption might not be met, resulting in a biased uncertainty estimate. Given that independent GSS estimates can be obtained at a relatively low computational cost, empirical uncertainty can be directly estimated. This method is adopted in the following applications.

GSS inherits the benefits of the SS method and is equal to SS if $\pi_0(\boldsymbol{\theta}|M) = \pi(\boldsymbol{\theta}|M)$. For GSS to be efficient, one can tailor a reference distribution π_0 that approximates the posterior as the product of independent distributions, each reflecting the marginal distribution of each parameter θ_j, which is an element of the parameter vector $\boldsymbol{\theta}$. If posterior samples are available, these samples can be used to calibrate each marginal

probability density of the reference distribution for each θ_j, where each distribution can be set, for example, to $\mathcal{N}(\bar{\theta}_j, S_j^2)$, with $\bar{\theta}_j$ and S_j being the mean and standard deviation of these samples, respectively.

The GSS method bears similarity to the recently proposed weighting method [Hourihane et al., 2023]. It can be shown that for a target model T and an approximate model A with equal parameter space, the reweighting method is a particular case of the GSS method. In this case, the reference distribution π_0 is equal to *likelihood* × *prior* of the approximate model A with $K = 2$. The reweighting method is particularly useful for the PTA models when the sampling from model T is computationally expensive compared to the approximate model A.

3.3 Simulation study

In this simulation study, we will implement the GSS method and compare its performance to that of SS and TI. Since the marginal likelihood z of real cases generally does not have an analytical solution, we resort to a simple Gaussian model where the marginal likelihood is known [Lartillot and Philippe, 2006]. This Gaussian model is parameterized by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$ with a null data vector X, where the prior on $\boldsymbol{\theta}$ is a product of standard normal distributions $\theta_j \sim \mathcal{N}(0, 1)$, and the likelihood is defined by

$$L(X|\boldsymbol{\theta}) = \prod_{j=1}^{d} \exp\left(-\frac{\theta_j^2}{2v}\right), \tag{3.17}$$

with d the number of dimensions, and a known variance v. This setup enables access to the analytical form for each component of the Bayesian model. The power posterior density can be expressed as a product of d normal distribution $\mathcal{N}(0, v/(v + \beta))$. The density p_β is equal to the posterior distribution when $\beta = 1$ yielding a marginal likelihood

equal to

$$z = \left(\frac{v}{1+v}\right)^{d/2}. \tag{3.18}$$

Since the analytical power posteriors are available, we sample directly from these densities to avoid MCMC sampling. For SS and TI, n independent samples are generated for each β_k chain, and then K the number of chains is increased until the numerical estimate reaches the true value. In contrast, the GSS estimator is executed in two steps. First, independent samples from the posterior distribution (calibration samples N_{cal}) are used to tune the reference distribution, then n samples are generated for each β_k from the power density $\mathcal{N}_{\beta_k,j}(\mu_{\beta_k,j},\, \sigma^2_{\beta_k,j})$ with

$$\begin{aligned} \mu_{\beta_k,j} &= \frac{\bar{\theta}_j(1-\beta_k)}{S_j^2\,\beta_k\left(\frac{1+v}{v}\right)+(1-\beta_k)}, \text{and} \\ \sigma^2_{\beta_k,j} &= \left(\beta_k\left(\frac{1+v}{v}\right)+\frac{(1-\beta_k)}{S_j^2}\right)^{-1}, \end{aligned} \tag{3.19}$$

where $\bar{\theta}_j$ and S_j^2 are, respectively, the mean and the variance of the calibration samples from the j^{th} dimension.

For this comparison, we set the number of dimensions to $d = 50$ and the variance to $v = 0.01$, resulting in a $\log(z) = -115.38$. With β_k uniformly dispersed along the quantiles of Beta$(0.3, 1)$, we fix $n = 1000$ independent sample, $K \in [4, 8, 16, 32, 64]$ number of β chains for each method. For the GSS estimator, the same setup is used, however, we set $N_{\text{cal}} = 1000$ and $n = 10$. To evaluate the uncertainty of each method's estimate, 1000 independent estimates are generated for each K. Figure 3.1 shows that TI requires more than 32 chains and 16 chains for SS to converge to the true $\log(z)$, while the GSS estimator yields an accurate estimate of $\log(\widehat{z}) = -115.37$ with $K = 4$. This comparison demonstrates how the $\log(z)$ estimation using SS and TI can be biased even when error constraints are available. Thus, only increasing K will ensure convergence and an accurate $\log(z)$ estimate.

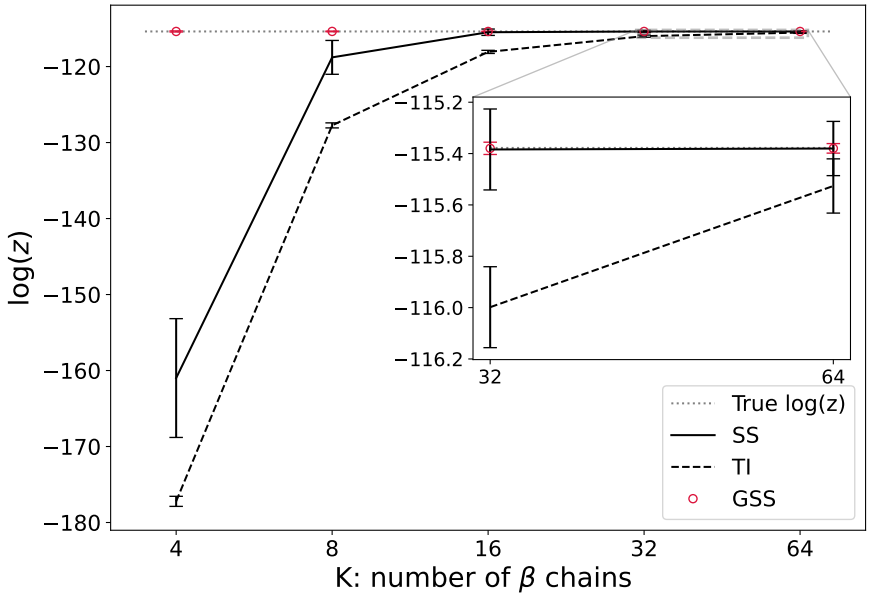


Figure 3.1: Comparison of log-marginal likelihood estimated as a function of K number of β chains for the Gaussian model with TI, SS, and GSS methods.

The GSS estimator's performance remains consistent with a higher number of dimensions as shown in Figure 3.2 where the number of dimensions $d = 2000$ and $N_{\text{cal}} = 500$. Figure 3.2 shows that one can either use fewer samples and increase K or use few β chains and increase n the number of samples per chain. In practice, numerical experiments suggest that the computational cost should be allocated to the number of β chains since a higher number of K increases the similarity between the consecutive importance distributions [Maturana-Russel, 2017].

3.4 Application of GSS to the PTA

In this section, we will illustrate how the GSS integrates into the PTA-Bayesian framework, and demonstrate the different benefits of estimating the *marginal likelihood* in a PTA analysis. For this application, a Metropolis-Hastings algorithm [Chib and Greenberg, 1995] was implemented to perform the sampling for the GSS method, where MPI for Python [Dalcín et al., 2005] was employed to parallelize the sampling of β chains.

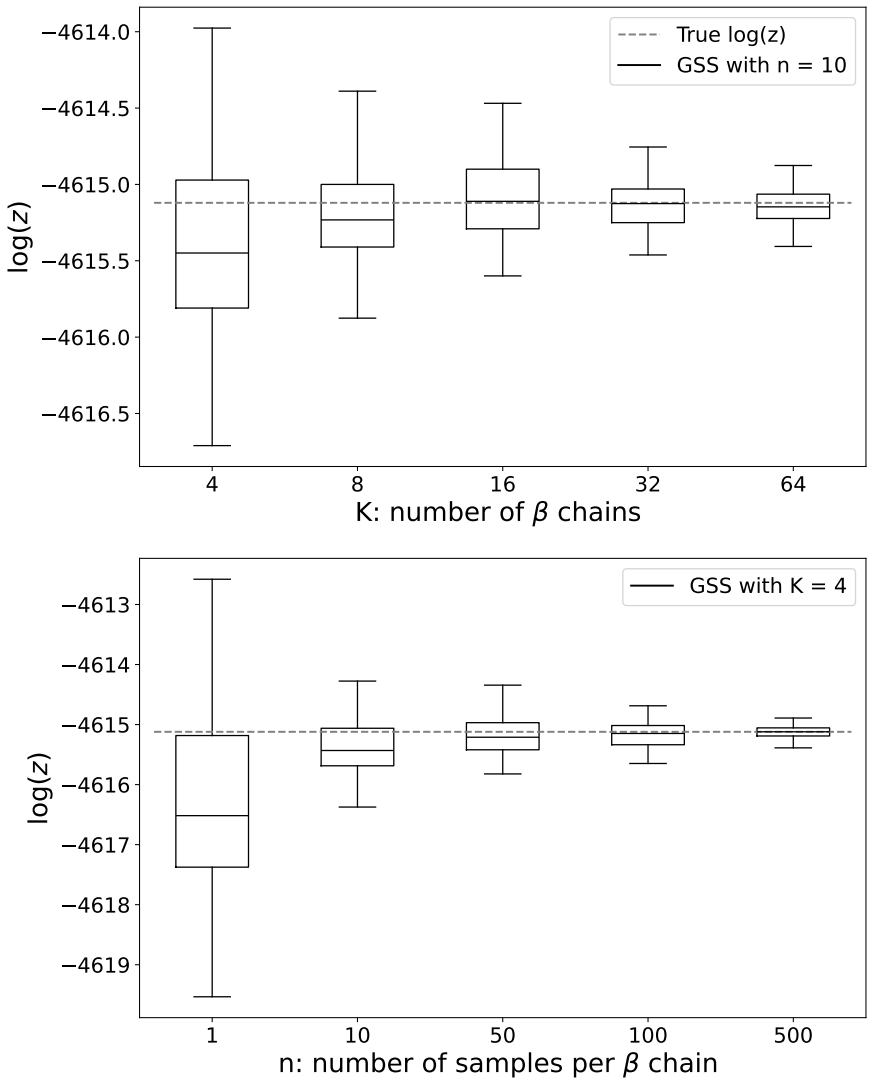


Figure 3.2: Convergence of the marginal likelihood estimated using GSS for the Gaussian model with $d = 2000$ and $v = 0.01$. Top panel: Comparison of log-marginal likelihood estimated as a function of K number of β chains with $n = 10$. Bottom panel: Comparison of log-marginal likelihood estimated as a function of n number of independent samples per β chains with $K = 4$.

Table 3.1: Mean and standard deviation of 100 independent log-marginal likelihood estimates for NANOGrav single pulsar noise models obtained via GSS with $K = 8$.

	WN		WN+DMGP		WN+RN		WN+RN+DMGP	
PSR	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std
J1012+5307	275708.30	0.20	275708.35	0.25	275744.84	0.48	275745.76	0.61
J1600-3053	261778.81	0.57	261778.87	0.65	261784.37	0.81	261784.82	0.87
J1705-1903	111504.39	0.29	111600.51	0.32	111562.63	0.44	111633.47	0.41
J1713+0747	750942.80	3.24	750946.68	1.18	751285.52	0.78	751284.93	2.13

This implementation was designed to adequately fit into the Bayesian framework used in PTA analyses [ENTERPRISE; Ellis et al., 2019, Taylor et al., 2021]. In the following, the performance of GSS is assessed based on K, the number of β chains, and n_{ESS} the Effective Sample Size (ESS) per β chain. We will use the GSS method to produce multiple independent estimates for each analysis. These multiple estimates allow the direct comparison of all models by constructing a distribution of the Bayes factor of the compared models. In practice, single random draws of $\log(\widehat{z})$ from the marginal likelihood distributions for model M_1 and M_2 are used to compute each $\log(\text{BF}_{M_1/M_2})$, the standard deviation of the resulting $\log(\text{BF}_{M_1/M_2})$ distribution is used to measure uncertainty.

3.4.1 Application to NANOGrav

This section demonstrates the application of the GSS to estimate the marginal likelihood for different models used in the PTA single pulsar noise analysis and GWB analysis. We reproduce the analysis and the public NANOGrav 15yr Data Release and the NANOGrav tutorials [Agazie et al., 2023b] using ENTERPRISE. The consistency of the reproduced Bayesian analysis is compared with the results discussed in the NANOGrav 15 yr Data Set.

3.4.1.1 Model selection for single pulsar analysis

The following four pulsars were selected from NANOGrav [Agazie et al., 2023b] as candidates for single-pulsar noise analysis: PSRs J1012+5307, J1600-3053, J1705-1903, and J1713+0747. Around $100,000$ posterior samples are generated for each combination of the following noise models: White Noise (WN), Dispersion Measure Gaussian Process (DMGP), and Red Noise (RN). A 20,000 samples of the burn-in period is discarded from the posterior samples used to calibrate the reference distribution π_0. In this case, the posterior samples are reasonably unimodal; therefore, for each model a normal distribution is used to approximate each parameter's marginal posterior distribution. The product of these distributions defines the reference distribution π_0, as it usually provides a sufficiently accurate approximation to the posterior for the GSS method. The mean and the variance of each parameter's set of samples are used to calibrate π_0, then a mean effective sample size $n_{\text{ESS}} \approx 10$ is generated for each β chain with $K = 8$ for each single estimate. Table 3.1 reports the mean and standard deviation of 100 independent $\log(\widehat{z})$ for the combination of the three noise models. These values of $\log(\widehat{z})$ are used to compute the $\log(\widehat{\text{BF}})$ for comparing models' relative performance.

For each pulsar, the WN model is used as a baseline for model comparison, the best-performing model compared to WN is reported with their $\log(\widehat{\text{BF}}_{\text{M/WN}}) \pm 1\sigma$. Across all four pulsars, strong statistical evidence is observed for the Red Noise model. For PSR J1012+5307, WN+RN is favored with $\log(\widehat{\text{BF}}) = 36.47 \pm 0.56$. When combined with the other two models, the DMGP model has weak statistical evidence since $\log(\widehat{\text{BF}}) < 1$. For PSR J1705-1903 and PSR J1713+0747, the model comparison shows high statistical significance between each fitted model, where the most significant model is WN+RN+DMGP with $\log(\widehat{\text{BF}}) = 129.07 \pm 0.49$ for PSR J1705-1903, and WN+RN with $\log(\widehat{\text{BF}}) = 342.58 \pm 3.29$ for PSR J1713+0747. PSR J1600-3053 demonstrates the least RN significance compared to the other 3 pulsars with $\log(\widehat{\text{BF}}) = 5.56 \pm 0.99$. The

	DMGP		RN	
PSR	$\log(\widehat{\text{IBF}})$	std	$\log(\widehat{\text{IBF}})$	std
J1012+5307	0.99	0.87	73.94	0.82
J1600-3053	0.52	1.41	11.42	1.43
J1705-1903	166.92	0.75	91.20	0.74
J1713+0747	3.46	4.14	680.88	4.17

Table 3.2: Mean and standard deviation of 1000 independent $\log(\widehat{\text{IBF}})$ estimates for the parameter inclusion of DMGP model and RN model in NANOGrav single pulsar noise models.

evidence of the excess noise becomes more significant for the model WN+RN+DMGP with $\log(\widehat{\text{BF}}) = 6.00 \pm 1.05$.

One can further examine the performance of each model by inspecting the influence of its parameters $\boldsymbol{\delta}$ when included. This influence can be demonstrated for the PTA models using the previous combinations of red noise models. This evaluation can be conducted using the *Inclusion Bayes factor* [IBF; Hinne et al., 2020] which examines the proportional change of the marginal likelihood when a set of parameters $\boldsymbol{\delta}$ are included in a set of models. Let $\mathcal{M}_{\boldsymbol{\delta}^1}$ and $\mathcal{M}_{\boldsymbol{\delta}^0}$ be respectively the sets of models that include and exclude $\boldsymbol{\delta}$ from their respective parameter vector. The is given by

$$\text{IBF} = \frac{\sum_{M \in \mathcal{M}_{\boldsymbol{\delta}^1}} z(X|M)}{\sum_{M \in \mathcal{M}_{\boldsymbol{\delta}^0}} z(X|M)}, \tag{3.20}$$

which denotes the sum of marginal likelihoods of models including $\boldsymbol{\delta}$ divided by the sum of marginal likelihoods of models excluding it. Table 3.2 shows the IBF to compare the overall performance of RN and DMGP parameters. IBF reinforces the strong statistical evidence of RN across all pulsars while including DMGP parameters delivers a better performance only for PSR J1705-1903 and PSR J1713+0747. Since pulsar customized noise models are used for the full GWB analysis, the IBF can provide additional statistical evidence and help with a robust model selection where decisions are critical.

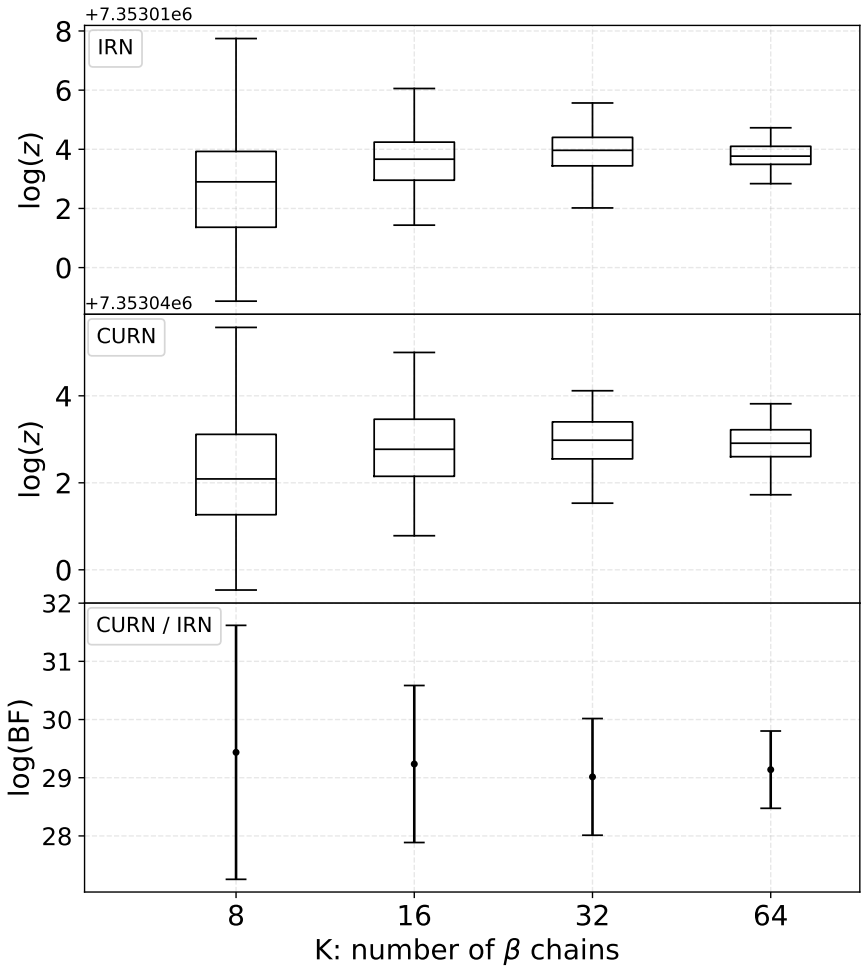


Figure 3.3: Estimations as a function of K number of β chains for the NANOGrav dataset via GSS. Top panel: log-marginal likelihood estimated for the IRN model. Middle panel: log-marginal likelihood estimated for the CURN model. Bottom panel: $\log(\widehat{\mathrm{BF}}_{\mathrm{CURN/IRN}})$ comparing CURN to IRN.

3.4.1.2 Evidence estimation for the GWB analysis

In parallel with our previous analysis, the GSS estimator can also be used to investigate the evidence of GWB in the PTA. We reproduce the posterior samples used in the NANOGrav analysis where all CRS models share the log-spectral amplitude $\log(A_M)$ and the spectral index γ_M of the power law as free parameters for the four following models: intrinsic Intrinsic Red Noise (IRN), Common Uncorrelated Red Noise (CURN), spline overlap reduction function (Spl ORF), and the Hellings and Downs (HD) serving as the probe for GWB. The default setup and codes provided by the NANOGrav public data release are used in this demonstration. We refer to Agazie et al. [2023b] for extended details on the setup and the models. In this case, around $120,000$ samples are generated for each of the 4 models. Before estimating $\log(z)$, the posterior samples are compared with the previous analyses for consistency. The means and the standard deviations of the generated samples are in agreement with the reported results in Agazie et al. [2023b].

To configure the GSS, a 20,000 samples of the burn-in period is discarded from the posterior samples used to tune the reference distribution π_0. By tuning the reference distribution and using it as a sampling proposal distribution, the GSS estimator enables us to skip the new burn-in period in the GSS sampling. In Figure 3.3, each $\log(z)$ estimate is computed using $n_{\text{ESS}} \approx 50$ per each β chain, while the number of parallel chains grows with $K \in [8, 16, 32, 64]$. A higher number of n_{ESS} is used to cross-check the consistency of the $\log(\widehat{z})$ estimates. As demonstrated in Figure 3.3, the GSS estimator produces a consistent median over 100 $\log(\widehat{z})$ starting from $K = 16$ while constraining the uncertainty as the number of chains increases, a higher number of n_{ESS} can improve the estimate of the median $\log(\widehat{z})$ for $K = 8$ as shown in Figure 3.2. The $\log(\widehat{\text{BF}})$ is consistent through all K and the slight variation would not affect the inference.

	$\log(\widehat{z})$	std
IRN	7353013.52	1.07
CURN	7353042.72	0.84
Spl ORF	7353037.75	1.18
HD	7353047.80	0.91

Table 3.3: Mean and standard deviation of 100 independent log-marginal likelihood estimates with K = 16 for different common red process models in the NANOGrav dataset.

To evaluate the GWB evidence, 100 independent estimates were computed for each of the four models, Table 3.3 reports the mean and the standard deviation of 100 estimates of $\log(\widehat{z})$. Figure 3.4 reports the mean and standard deviation where 1000 estimates of $\log(\widehat{\text{BF}})$ were derived to compare the models. Compared to IRN, the other three models are highly supported reinforcing the strong evidence of an extra red process in the PTA dataset. Furthermore with a $\log(\widehat{\text{BF}}_{\text{CURN/IRN}})$ of 29.18 ± 1.36, a common uncorrelated red noise process is evident in the dataset reconfirming the previous PTA analyses [Arzoumanian et al., 2020, Chen et al., 2021, Goncharov et al., 2022]. Currently, CURN is favored over the Spl ORF model with $\log(\widehat{\text{BF}}_{\text{CURN/Spl ORF}}) = 5.04 \pm 1.41$. For the GWB evidence, $\log(\widehat{\text{BF}}_{\text{HD/CURN}}) = 5.11 \pm 1.24$ is consistent with the NANOGrav reported estimation of 5.42 ± 4.25, whereas the uncertainty is more constrained by the GSS method.

3.4.2 Application to EPTA+InPTA

The EPTA+InPTA collaboration has publicly released its second dataset with the analysis for GW evidence [Antoniadis et al., 2023a]. This analysis combines the dataset of 25 pulsars into four subsets: DR2full covers 24.7 years of EPTA data, DR2new is limited to the last 10.3 years of observations with the upgraded radio telescopes, DR2full+ and DR2new+ add 0.7 years of InPTA data to the previous subsets. The EPTA+InPTA study uses customized pulsar noise models reported in Antoniadis et al. [2023b]. We will reproduce the analysis using the 4 subsets for the following models

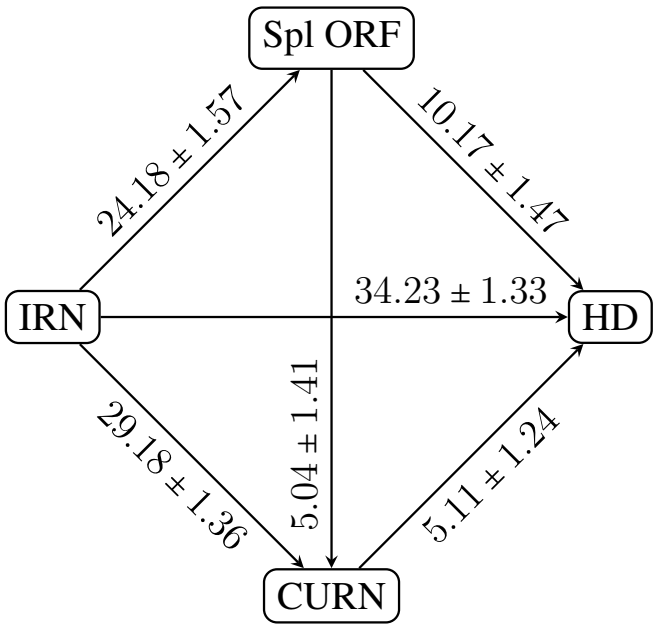


Figure 3.4: Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{\mathrm{M_1/M_2}})$ estimates comparing different common red process models in the NANOGrav dataset where the arrow starts at model $\mathrm{M_2}$ and ends at model $\mathrm{M_1}$.

CURN, GWB (referred to as HD in section 3.4.1.2), and Binned ORF by using the default setup provided in the second data release. Around $120,000$ posterior samples are generated for each model, and then their means and standard deviations are examined for consistency with the reported ones in Antoniadis et al. [2023a]. A burn-in period of the first 20,000 samples was discarded from the posterior samples used to calibrate the reference distribution π_0. A 100 $\log(\widehat{z})$ per model are estimated with K = 16 and $n_{\mathrm{ESS}} \approx 20$ samples per β chain. To assess the stability of the estimation, n_{ESS} and K are increased and $\log(\widehat{z})$ estimates are cross-checked for consistency. Table 3.5 reports the mean and the standard deviation of 100 $\log(\widehat{z})$ estimates for the 3 models across all EPTA+InPTA datasets. To compare each model's performance, Table 3.4 reports the mean and the standard deviation of 1000 $\log(\mathrm{BF_{M/CURN}})$ independent estimates of the other two models compared to the CURN model. For DR2full and DR2full+, the $\log(\mathrm{BF_{GWB/CURN}})$ supports the GWB model over the CURN model. This support is further reinforced when fitting only to the new generation dataset DR2new. However,

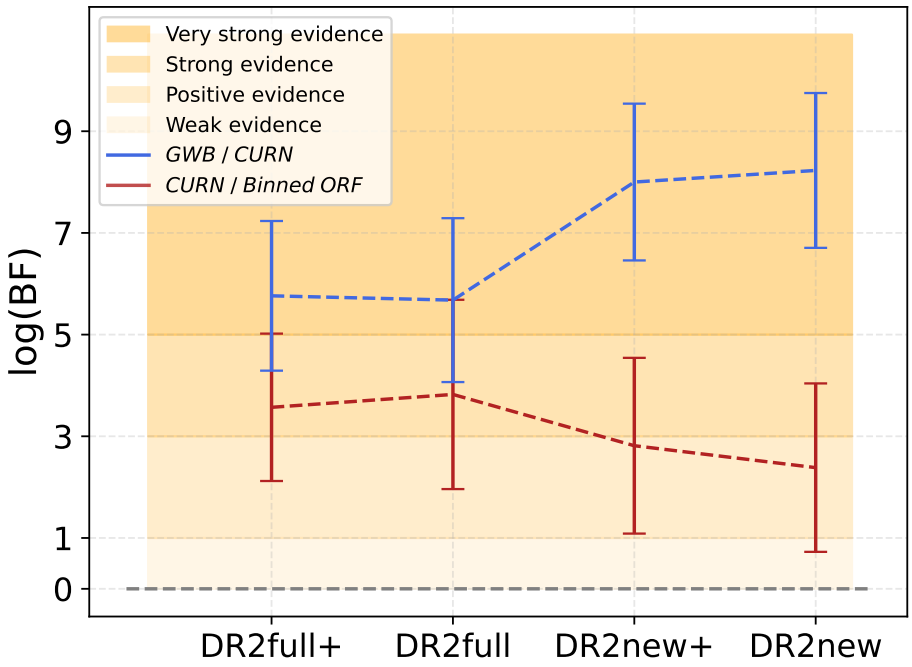


Figure 3.5: Evolution of $\log(\widehat{\mathrm{BF}}_{\mathrm{M_1/M_2}})$ comparing different common red process models as a function of the dataset size of EPTA+InPTA dataset with DR2full+ the full dataset and DR2new the shortest subset.

Table 3.4: Mean and standard deviation of 1000 independent $\log(\widehat{\mathrm{BF}}_{\mathrm{M/CURN}})$ estimates comparing Binned ORF and GWB model to CURN model across EPTA+InPTA datasets.

Model	DR2new		DR2new+		DR2full		DR2full+	
	$\log(\widehat{\mathrm{BF}})$	std	$\log(\widehat{\mathrm{BF}})$	std	$\log(\widehat{\mathrm{BF}})$	std	$\log(\widehat{\mathrm{BF}})$	std
PSRN+Binned ORF	-2.35	1.65	-2.81	1.76	-3.62	1.42	-3.80	1.81
PSRN+GWB	8.24	1.47	7.89	1.48	5.79	1.42	5.70	1.57

CURN model is favored over the Binned ORF model, and this evidence decreases for DR2new. The addition of the InPTA dataset in DR2full+ and DR2new+ slightly decreases the evidence on the GWB model and slightly increases the evidence favoring CURN over Binned ORF. Figure 3.5 summarizes this evolution of evidence in the EPTA+InPTA dataset. Although the InPTA dataset is recent, the slight decrease in evidence of GWB might arise from adding 1 year of dataset asymmetrically to only 12 pulsars out of 25 candidates, which will directly affect the correlations between pulsars.

Table 3.5: Mean and standard deviation of 100 log-marginal likelihood estimates for each common red process model applied to the EPTA+InPTA datasets with K = 16.

	DR2new		DR2new+		DR2full		DR2full+	
	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std	$\log(\widehat{z})$	std
PSRN+CURN	493684.31	1.02	525786.67	1.08	607134.56	1.20	639232.25	1.03
PSRN+Binned ORF	493681.86	1.30	525783.84	1.36	607130.75	1.31	639228.67	1.05
PSRN+GWB	493692.55	1.14	525794.58	0.98	607140.21	1.12	639238.08	1.01

3.5 Discussion

In the simulation study, we highlighted the differences between methods of power posterior and showed their relative performance. The GSS method has outperformed both TI and SS and proved efficient for higher dimensional cases. GSS can benefit from the previous efforts usually done in the parameter estimation phase. The success of a cheaper GSS estimation relies on having a good posterior approximation, which shortens the path to an accurate estimate. Without such an approximation, the GSS estimator will require more effort (but still less than SS and TI) to achieve an accurate estimate. These $\log(z)$ estimates result in an accurate evaluation of the $\log(\text{BF})$ and a robust interpretation of evidence supporting different hypotheses. This evidence can further be tested using the IBF criteria, which examine the influence of a particular set of parameters on the evidence.

Due to its lower computational cost, the GSS method enabled multiple $\log(z)$ estimates with high accuracy for each PTA-Bayesian analysis in this study. For single noise analysis, GSS has shown its ability to offer model selection without constraints, compared to the Hypermodel, which can fail to provide evidence evaluation in cases such as estimation for RN cases [Smarra et al., 2023]. In the NANOGrav dataset, the four pulsars have shown different levels of significance for RN and DMGP signals, where PSR J1713+0747 showed extreme evidence for RN. Note that PSR J1713+0747 contributed the most to the common red noise signal in the EPTA+InPTA dataset [Antoniadis et al., 2023a]. For NANOGrav GWB evidence, the GSS method has provided

a consistent and more accurate estimate than the Hypermodel method's estimate of evidence supporting a GWB signal compared to a common red noise signal. Furthermore, GSS enabled a direct comparison of expensive models, efficiently eliminating the arbitrary weights required by the Hypermodel method. The EPTA GWB analysis consistently reconfirms the evidence supporting a GWB signal in the PTA dataset. This evidence becomes more significant for the new-generation dataset as shown in Figure 3.5.

The PTA experiments rely on model selection to tune each part of the Bayesian noise model which increases the number of Bayesian analyses required to achieve better sensitivity. The current GSS setup substantially minimizes the cost of a single $\log(z)$ estimate for both light and expensive models, and would further benefit in the future from better samplers. The efficiency of the GSS method will help reduce the cost of evidence estimation for the continuously growing model portfolio for GWs and new physics in the PTA [Afzal et al., 2023, Smarra et al., 2023].

Acknowledgments

EMZ, PMR, WS, and RM gratefully acknowledge support by the Marsden Fund Council grant MFP-UOA2131 from New Zealand Government funding, managed by the Royal Society Te Apārangi. This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and from the Victorian Higher Education State Investment Fund (VHESIF) provided by the Victorian Government.

Software: `ENTERPRISE` [Ellis et al., 2019, Taylor et al., 2021], `enterprise_extension` [Taylor et al., 2021], `PTMCMC` [Ellis and van Haasteren, 2017] , `MPI4PY` [Dalcín et al., 2005], `Jupyter` [Kluyver et al., 2016], `matplotlib`

[Hunter, 2007], `numpy` [Harris et al., 2020], `scipy` [Virtanen et al., 2020a], `arviz` [Kumar et al., 2019].

Data Availability

Data used in this analysis are available on GSS-estimator GitHub [1].

[1] https://github.com/nz-gravity/GSS-estimator

Chapter 4

Manuscript 2

MorphZ

This manuscript introduces MorphZ, a post-processing marginal likelihood estimator based on the *Morph* approximation. Our work enables accurate Bayesian evidence estimates at substantially reduced computational cost by using a posterior-informed approximation as proposal distributions in optimal bridge sampling. It is designed to work directly from posterior samples, making it sampler-agnostic and easy to integrate into existing inference workflows. The manuscript demonstrates the method across statistical benchmarks, pulsar timing array analyses, compact-binary-coalescence gravitational-wave simulations, and the GW150914 event. On 10^{th} April 2026, this manuscript was published in *Physical Review D* under the title *Enhancing evidence estimation through informed probability density approximation*.

Manuscript details

Title Enhancing evidence estimation through informed probability density approximation.

Authors El Mehdi Zahraoui, Patricio Maturana-Russel, Avi Vajpeyi, Willem van Straten, Renate Meyer, and Sergei Gulyaev.

Journal Physical Review D.

DOI `https://doi.org/10.1103/1554-y6ns`.

Abstract

We introduce the *Morph* approximation, a class of product approximations of probability densities that selects low–order disjoint parameter blocks by maximizing the sum of their total correlations. We use the posterior approximation via Morph as the importance distribution in optimal bridge sampling. We denote this procedure by MorphZ, which serves as a post–processing estimator of the marginal likelihood. The MorphZ estimator requires only posterior samples, and is fully agnostic regarding the choice of sampler. We evaluate MorphZ's performance across statistical benchmarks, pulsar timing array (PTA) models, compact binary coalescence (CBC) gravitational-wave (GW) simulations and the GW150914 event. Across these applications, spanning low to high dimensionalities, MorphZ yields accurate evidence estimates at substantially reduced computational cost relative to standard approaches. We have found that when these approaches fail to provide accurate estimates, MorphZ has proven to either resolve the estimation failure or significantly improve the results. Its bridge sampling relative error diagnostic provides conservative uncertainty estimates. Because MorphZ operates directly on posterior draws, it complements exploration-oriented samplers by enabling fast and reliable evidence estimation, while it can be seamlessly integrated into existing inference workflows.

4.1 Introduction

Across modern astrophysics and cosmology, Bayesian inference has become foundational for both parameter estimation and hypothesis testing. In this framework, prior beliefs about model parameters are updated to posterior beliefs as new data are acquired, providing a coherent probabilistic framework for quantifying uncertainty in both the parameters and the resulting model predictions. A key quantity is the Bayesian evidence (marginal likelihood), which underpins probabilistic model comparison and hypothesis assessment. Thanks to its ability to combine heterogeneous datasets and incorporate prior information consistently, the Bayesian framework is well suited to experiments across a wide range of scales, and is particularly advantageous for decade-spanning observational programs.

Among these experiments, the pulsar timing array (PTA) stands out as a well-suited experiment that employs Bayesian analysis techniques to update the status of gravitational-wave background detection. Accordingly, several of the latest PTA data releases provide Bayesian evidence favoring the presence of a gravitational-wave background (GWB) in their datasets Reardon et al. [2023b], Agazie et al. [2023b], Antoniadis et al. [2023a]. These data releases are a product of Herculean efforts involving observations of hundreds of millisecond pulsars (MSPs), detailed per-pulsar noise characterization Agazie et al. [2023b], Antoniadis et al. [2023a], Zic et al. [2023], and joint analysis aimed at GWB detection Reardon et al. [2023b], Agazie et al. [2023b], Antoniadis et al. [2023a], Miles et al. [2023]. Currently, the PTA model selection and evidence estimation are performed using different techniques [Agazie et al., 2023b, Antoniadis et al., 2023a]: thermodynamic integration [Lartillot and Philippe, 2006], nested sampling [Skilling, 2006], model reweighting [Hourihane et al., 2023], Savage-Dickey density ratio approximation [Carlin and Chib, 1995, Lodewyckx et al., 2011]. The efficiency and accuracy of evidence estimation methods employed by the PTA

are key for robust detection claims and lower computational costs. The future Square Kilometre Array (SKA) Smits et al. [2009], Stappers et al. [2018] dataset will significantly increase the complexity of hierarchical noise and signal models, making efficient evidence estimation a practical necessity to keep the computational cost tractable.

In parallel with pulsar timing arrays, ground-based gravitational-wave astronomy provides a compelling setting in which Bayesian model selection plays a key role in astrophysical inference. Within the LIGO–Virgo–KAGRA (LVK) collaboration, the Bayesian evidence plays a central role in adjudicating between competing waveform models, ranking candidate events, identifying incoherent noise transients, and performing tests of general relativity Puecher et al. [2024], Roy et al. [2025], Vajpeyi et al. [2022]. Recent analyses have shown that subtle differences in waveform physics, such as the inclusion of spin precession or higher-order modes, can lead to biased scientific conclusions unless model choices are explicitly validated through evidence-based comparison Puecher et al. [2024], Roy et al. [2025]. In addition, the evidence can also be used for waveform model averaging, weighting each waveform family in proportion to its support to marginalize over waveform systematics rather than relying on a single preferred model [Ashton and Khan, 2020]. As the complexity of Compact Binary Coalescence (CBC) signal hypotheses continues to grow, LVK inference pipelines increasingly rely on accurate and scalable evidence estimation. This need will intensify for next-generation instruments, including the Einstein Telescope Punturo et al. [2010], Cosmic Explorer Reitze et al. [2019], and LISA Amaro-Seoane et al. [2017], which will produce significantly larger datasets and more intricate likelihoods. Efficient and reliable computation of evidence is therefore essential for model selection and hypothesis testing across current and future gravitational-wave observatories.

The Bayesian evidence, commonly known as the *marginal likelihood*, is a multidimensional integral over model parameters that is generally estimated numerically. The marginal likelihood literature provides a wide variety of approaches, spanning analytic

approximations, sampling-based estimators Chib and Greenberg [1995], Meng and Wong [1996], Skilling [2006], and path-based techniques Friel and Pettitt [2008], Xie et al. [2011]. Each newly proposed method aims to lower the computational expense of marginal likelihood estimation while improving estimator accuracy in challenging Bayesian models. Among these methods, several accurate evidence estimators rely on importance sampling distributions tuned to the posterior samples, such as bridge sampling (BS; Meng and Wong [1996], Gronau et al. [2017]), nested importance sampling (NIS; Chopin and Robert [2010], Feroz et al. [2013]), Generalized Annealing importance sampling (GAIS; Maturana-Russel [2017]), and Generalized Steppingstone Sampling (GSS; Fan et al. [2011]). Chan and Eisenstat [2015] show that an optimal importance sampling distribution is the distribution that minimizes the cross-entropy (equivalently, the *Kullback–Leibler* divergence) relative to the posterior. In practice, selecting an importance distribution on a case-by-case basis limits its practicality for routine estimation. Common heuristics include modeling the posterior as a product of marginal distributions fitted to posterior samples or a multivariate Gaussian with a covariance matrix estimated from those samples Fan et al. [2011], Chan and Eisenstat [2015]. These heuristics limit the importance sampling distribution to either preserve the marginal shapes or to assume Gaussian marginals while capturing linear correlations. Alternatively, better importance distributions employ different types of copulas tuned to posterior samples Silva and Lopes [2008], Dellaportas and Tsionas [2019], Kauermann and Meyer [2014], Craiu [2011], Schmidl et al. [2013], South et al. [2019]. These copula-based importance densities capture skewness and tail dependence better than a multivariate normal and are usually cheap to fit to posterior samples. However, they were only tested on moderate-dimensional regression and random-effects models Silva and Lopes [2008], and need to be further tested on high-dimensional and multimodal problems. Additionally, training a normalizing flow on the posterior samples has proven

to be effective and more flexible as an importance distribution for up to 21 dimensions Polanska et al. [2024]; however, scalability to higher dimensionalities and more complex posteriors remains to be systematically assessed.

The marginal likelihood computational cost decreases as the proposal more closely matches the posterior Chan and Eisenstat [2015]. Accordingly, we present a novel approach based on a class of product approximations to probability densities that more fully exploits available posterior information. This approximation enhances marginal likelihood estimators making evidence available at a fraction of the cost relative to popular approaches in astrophysics. This new family of product approximations is built on the notion of *total correlation* Watanabe [1960], an information-theoretic measure that captures both linear and non-linear dependencies; we refer to it as *Morph* approximation. By reducing a high-dimensional posterior to a product of low-dimensional factors, the *Morph* approximation achieves scalable inference and strong performance from low to high dimensionalities. To construct the *Morph* approximation, we use standard Kernel Density Estimation (KDE) algorithms Virtanen et al. [2020b,a], yielding an approximation that preserves empirical marginal shapes and captures the dominant dependencies at low cost. In combination with bridge sampling, the MorphZ estimator delivers efficient and accurate marginal likelihood estimates, outperforming commonly used methods on challenging problems and remaining effective for high-dimensional and expensive-likelihood settings. This approach prioritizes the posterior coverage; once coverage is sufficient, the evidence can be computed at minimal additional cost.

The remainder of this paper is organized as follows. In Section 4.2, we introduce the marginal likelihood and the corresponding estimation methods. Section 4.3 defines the Morph approximation, the MorphZ estimator, and describes the procedure used to construct this estimator. Section 4.4 evaluates the estimator's performance on a set of challenging statistical benchmarks. Sections 4.5 and 4.6 demonstrate the efficiency and accuracy of the MorphZ estimator through applications to pulsar timing array

and LIGO-VIRGO-KAGRA (LVK) gravitational wave analyses. In Section 4.7, we address practical aspects of the Morph approximation and bridge sampling. Finally, Section 4.8 concludes with an assessment of the effectiveness of MorphZ and outlines future directions.

4.2 Marginal likelihood estimation

The *marginal likelihood* or evidence is defined by *Bayes'* rule as

$$z(X|M) = \int_{\Theta} L(X|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}, \tag{4.1}$$

where $L(X|\boldsymbol{\theta}, M)$ is the likelihood function of the model M over the dataset X, and $\pi(\boldsymbol{\theta}|M)$ is a proper prior density of the parameter vector $\boldsymbol{\theta}$. The marginal likelihood $z(X|M)$, or simply z from here onward, is a quantity routinely used to measure the quality of the Bayesian fit. If two models, M_0 and M_1, fitted to a dataset X, are assumed to be equally probable, then the ratio of their marginal likelihoods is defined as the *Bayes factor*,

$$\mathrm{BF}_{(M_0/M_1)} = \frac{z(X|M_0)}{z(X|M_1)}. \tag{4.2}$$

This Bayesian criterion selects, between M_0 and M_1, the model best supported by the observed data X. The Bayes factor embodies *Occam's razor*, discouraging excessive complexity and mitigating overfitting. The marginal likelihood can be evaluated analytically only for special choices of priors and likelihoods (conjugate prior-likelihood pairs) whose product integrates in closed form. In general, the multi-dimensional integral is intractable and must be approximated numerically. To this end, multiple Markov chain Monte Carlo (MCMC) methods have been proposed to estimate the evidence, such as the harmonic mean [Newton and Raftery, 1994] , importance sampling [Kloek and Van Dijk, 1978], nested sampling [Skilling, 2006], path sampling [Gelman and Meng,

1998], and steppingstone sampling [Xie et al., 2011]. In the subsequent literature, an importance density fitted to posterior draws has been introduced as an extension of these methods to increase the accuracy of the marginal-likelihood estimates (bridge sampling Gelman and Meng [1998], nested importance sampling Chopin and Robert [2010], generalized steppingstone sampling Fan et al. [2011]). These extension methods have proven to reduce the computational cost while being more accurate than the original MCMC methods.

In this work, we introduce the *Morph approximation*: an information-based product approximation of the posterior density that can serve as an optimal importance proposal. We combine the Morph approximation with optimal bridge sampling Gelman and Meng [1998], Gronau et al. [2017], which yields an automated and highly efficient estimator for the marginal likelihood (MorphZ) that performs well in typical applications and demonstrates improved performance in statistically challenging scenarios (see Section 4.4).

4.3 Method

We begin by introducing the formal mathematical definitions underlying the Morph approximation. In this framework, we consider a probability density over a set of random variables (which, in Bayesian applications, corresponds to the posterior distribution). The Morph approximation constructs a tractable, normalized probability density from low-order factors while retaining its most important dependence structure. This is achieved by decomposing the full probability density into a product of lower-dimensional disjoint factors, each defined on a block of variables of equal length. These blocks are formed from groups of variables that are strongly correlated, and are chosen so as to maximize the total correlation captured within the blocks. The resulting low-dimensional factors can then be estimated using standard statistical techniques

from available samples and combined to form an efficient proposal distribution for subsequent bridge-sampling evidence estimation.

4.3.1 Morph approximation

Let $P(\boldsymbol{\theta})$ be a d-dimensional joint probability distribution with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\intercal \in \mathbb{R}^d$ the parameter vector. A product approximation of $P(\boldsymbol{\theta})$ is defined as a product of lower-order distributions, such that the product is a probability extension of these lower-order distributions that satisfies the unity sum property Lewis II [1959].

Definition 1 *Let $\Gamma = \{1, \ldots, d\}$ represent the set of integers from 1 to d, fix $L \in \Gamma$ by choosing one element from Γ , then $m = \lfloor d/L \rfloor \in \mathbb{N}^*$ is the integer number of times that L divides d, with $\mathbb{N}^*$ the natural number set excluding zero. Let*

$$\begin{aligned} \mathcal{F}_L = \{\mathcal{B}_L = \{b_1, \ldots, b_m\} : b_i \subseteq \Gamma, |b_i| = L, \\ b_i \cap b_j = \varnothing \ \ \forall i \neq j\}, \end{aligned} \tag{4.3}$$

with $\mathcal{B}_L$ the partial partition that includes all non-intersecting blocks b_i of Γ containing L elements. For $\mathcal{B}_L \in \mathcal{F}_L$, let the singletons residual index set $\mathcal{S}$ be defined by

$$\mathcal{S}(\mathcal{B}_L) = \Gamma \setminus \bigcup_{b \in \mathcal{B}_L} b, \tag{4.4}$$

with $\setminus$ being the set minus operator.
For $b = (k_1, \ldots, k_L) \subseteq \Gamma$, $k_i \in \Gamma$, $i = 1, \ldots, L$, let $\boldsymbol{\theta}_b = (\theta_{k_1}, \ldots, \theta_{k_L})^\intercal$. A Morph approximation of order L is a product approximation of $P(\boldsymbol{\theta})$ such that

$$\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta}) = \prod_{b \in \mathcal{B}_L} P_b(\boldsymbol{\theta}_b) \prod_{s \in \mathcal{S}(\mathcal{B}_L)} P_s(\theta_s), \tag{4.5}$$

with P_b the joint probability distribution of the sub-vector $\boldsymbol{\theta}_b$ of length L.

Although all $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ are valid product approximations of $P(\boldsymbol{\theta})$, these approximations differ in their closeness to $P(\boldsymbol{\theta})$. Therefore, one can choose an optimal Morph approximation where $\mathcal{M}_{\mathcal{B}_L}$ covers the important dependencies between variables. Accordingly, the total correlation, also known as multi-information Watanabe [1960], Studený and Vejnarová [1998], is a suitable measure to weight the importance of the dependencies between variables in each $\boldsymbol{\theta}_b$, $\forall b \subseteq \Gamma$. The total correlation $\mathcal{C}$ of $\boldsymbol{\theta}_b$ as a function of entropy H is given by

$$\mathcal{C}(\boldsymbol{\theta}_b) = D_{KL}\left(P_b || \prod_{k\in b} P_k\right) = \sum_{k\in b} H(\theta_k) - H(\boldsymbol{\theta}_b), \tag{4.6}$$

where D_{KL} is the *Kullback-Leibler divergence* between P_b, the joint probability density, and $\prod_{k\in b} P_k$, the element-wise product of the marginal densities of the block b. The *Kullback-Leibler divergence* between $P(\boldsymbol{\theta})$ and $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ as a function of $\mathcal{C}$ can be represented by

$$D_{KL}\left(P||\mathcal{M}_{\mathcal{B}_L}\right) = \mathcal{C}(\boldsymbol{\theta}) - \sum_{b\in B} \mathcal{C}(\boldsymbol{\theta}_b), \tag{4.7}$$

where $\mathcal{C}(\boldsymbol{\theta})$ is the total correlation of the parameter vector $\boldsymbol{\theta}$ and $\sum_{b\in B} \mathcal{C}(\boldsymbol{\theta}_b)$ the sum of the total correlations of the disjoint-blocks (see appendix 4.9 for the derivation). Therefore, one can use $\sum_{b\in B} \mathcal{C}(\boldsymbol{\theta}_b)$ as a measure of the total covered dependencies in $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ where its maximization increases the closeness of $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$.

Definition 2 *$\mathcal{M}_{\mathcal{B}^*_L}$ is an optimal Morph approximation of order L if*

$$\mathcal{B}^*_L \in argmax_{\mathcal{B}_L\in\mathcal{F}_L} \sum_{b\in\mathcal{B}_L} \mathcal{C}(\boldsymbol{\theta}_b). \tag{4.8}$$

An optimal Morph distribution of order L is, by definition, normalized and incorporates the maximum dependencies between variables in blocks of length L, while maintaining a simple approach.

4.3.2 MorphZ

We introduce MorphZ, an estimator for marginal likelihood, which uses the Morph approximation as a proposal distribution for the bridge sampling (BS) estimator Meng and Wong [1996]. The bridge sampling estimator is given by

$$z = \frac{\mathbb{E}_{g(\boldsymbol{\theta})}[L(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta})h(\boldsymbol{\theta})]}{\mathbb{E}_{post}[h(\boldsymbol{\theta})g(\boldsymbol{\theta})]}, \tag{4.9}$$

where $h(\boldsymbol{\theta})$ is the bridge function, $g(\boldsymbol{\theta})$ is the proposal distribution (or the importance distribution when $h(\boldsymbol{\theta}) = 1$) and $\mathbb{E}_{g(\boldsymbol{\theta})}$ and $\mathbb{E}_{post}$ are the expected values with respect to the proposal and posterior distribution, respectively. Specifically, the bridge function is set to the optimal bridge function Meng and Wong [1996] defined by

$$h(\boldsymbol{\theta}) = C \cdot \frac{1}{f_1\, L(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) + f_2\, z\, g(\boldsymbol{\theta})}, \tag{4.10}$$

with $f_1 = N_1/(N_1 + N_2)$ and $\mathrm{f}_2 = N_2/(N_1 + N_2)$ the fraction of N_1 posterior and N_2 proposal samples, respectively, and C a constant that cancels out in the ratio in Equation 4.9. An iterative scheme is used to estimate the marginal likelihood from an initial guess of z Gronau et al. [2017]. The accuracy of BS estimates can be quantified through an approximation of the relative mean-squared error Frühwirth-Schnatter [2004], where both posterior and proposal samples can be used at no-extra sampling cost Gronau et al. [2017]. A Python-based implementation [1] of the BS estimator has been developed for this study. Our implementation follows the detailed description of the bridge sampling estimator and its Relative Error (RE) in Gronau et al. Gronau et al. [2017]. We use the Morph approximation $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ as a proposal distribution $g(\boldsymbol{\theta})$. For simplicity, we will refer to $\mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})$ as $\mathcal{M}_L(\boldsymbol{\theta})$. We adopt the sample-splitting technique Overstall and Forster [2010], which consists of dividing the

[1] MorphZ: `https://github.com/EL-MZ/MorphZ`

Algorithm 1 SEEDED GREEDY MAXIMIZATION

Select m pairwise-disjoint blocks of length L maximizing $\sum_{b\in\mathcal{B}_L} C(\boldsymbol{\theta}_b)$.

Inputs: Dimension n; block length L; number of seeds $K \geq 1$; total correlation $C(\boldsymbol{\theta}_b) \geq 0$ for each block b. Let $\mathcal{W}$ be the full collection of b candidates, i.e., *all* $\binom{n}{L}$ *length L combinations of the n dimensions*.

Output: A set $B_{\text{best}} \subseteq \mathcal{B}_L$ of at most m pairwise-disjoint blocks with the largest $\sum_{b\in\mathcal{B}_L} C(\boldsymbol{\theta}_b)$ between the top K seeds.

1. **Precompute and sort.** Set $m \leftarrow \lfloor n/L \rfloor$. Let $\mathcal{W} \leftarrow$ the list of all possible blocks b sorted in decreasing $C(\boldsymbol{\theta}_b)$. Define the predicate: $\mathsf{disjoint}(b, \mathit{used}) \iff b \cap \mathit{used} = \varnothing$.
2. **Loop over the top K seeds.** For $j = 1, \ldots, \min\{K, |\mathcal{W}|\}$:
   (i) *Start a fresh construction.*
   $B \leftarrow \varnothing$; $\mathit{used} \leftarrow \varnothing$; $\mathit{total} \leftarrow 0$.
   (ii) *Place the seed.*
   $\mathit{seed} \leftarrow \mathcal{W}[j]$; $B \leftarrow B \cup \{\mathit{seed}\}$; $\mathit{used} \leftarrow \mathit{used} \cup \mathit{seed}$; $\mathit{total} \leftarrow \mathit{total} + C(\boldsymbol{\theta}_{\mathit{seed}})$.
   (iii) *Greedy fill.*
   Scan $\mathcal{W} \setminus \{\mathit{seed}\}$ in order while $|B| < m$.
   For each b: if $\mathsf{disjoint}(b, \mathit{used})$ then $B \leftarrow B \cup \{b\}$; $\mathit{used} \leftarrow \mathit{used} \cup b$; $\mathit{total} \leftarrow \mathit{total} + C(\boldsymbol{\theta}_b)$.
   (iv) *Keep the best construction.*
   If *total* > *Best total* then *Best total* $\leftarrow$ *total* and $B_{\text{best}} \leftarrow B$.
3. **Return** B_{best}.

available posterior samples into one batch for constructing the Morph approximation and another for the bridge sampling phase.

The kernel density estimator (KDE) is a fast standard method for density estimation of lower-dimensional problems Sheather and Jones [1991], Botev et al. [2010b]. The Morph approximation factorizes a higher-dimensional probability density into a product of low-dimensional factors, each of which can be estimated efficiently with KDEs. Throughout this work, the multivariate KDE with a Gaussian kernel Virtanen et al. [2020b] is used to estimate the densities $P_b(\boldsymbol{\theta}_b)$ and $P_s(\boldsymbol{\theta}_s)$ which are used both in $\mathcal{M}_L(\boldsymbol{\theta})$ and $\mathcal{C}(\boldsymbol{\theta}_b)$.

To construct $\mathcal{M}_L(\boldsymbol{\theta})$, we compute $\mathcal{C}(\boldsymbol{\theta}_b)$ for every candidate block $b \subset \Gamma$. Then, a

			Nested Sampling			MorphZ			MSS		
Model	d	$\log(z)_{\text{True}}$	calls	$\lvert\Delta\rvert$	std	calls	$\lvert\Delta\rvert$	std	calls	$\lvert\Delta\rvert$	std
Egg-box	2	235.856	6.9×10^4	0.025	0.106	5×10^3	0.035	0.01	4×10^3	0.062	0.72
Peak-plateau	20	0.693	1.8×10^6	0.332	0.277	5×10^3	0.051	0.06	4×10^3	1.555	1.6
Gaussian-shells	30	-60.13	6.5×10^6	0.921	0.317	5×10^3	0.074	0.01	4×10^3	0.141	0.35

Table 4.1: A comparison of the accuracy as $|\Delta| = |\overline{\log(\hat{z})}_{\text{Method}} - \log(z)_{\text{True}}|$, along the standard deviation of 100 $\log(\hat{z})$, and the efficiency as the number of likelihood calls for each estimate between Nested Sampling, MorphZ, and MSS for the 3 statistical examples.

Seeded Greedy Maximization (SGM) (see Algorithm 1): sort all blocks by *decreasing* $\mathcal{C}$; iterate over the top K blocks as seeds; for each seed, greedily add the highest-scoring non-overlapping blocks until m blocks are selected; return the run with the largest sum of total correlations. For $L = 2$, where the optimum is computationally tractable, SGM achieves or closely approaches optimality on our benchmarks; while for $L > 2$, it maintains similarly high performance relative to the best computed sum of total correlations (see Figure 4.5). The product structure of the Morph approximation allows each factor to propose samples independently. Taking one draw from each factor gives joint samples that one can evaluate under the posterior. In practice, fast KDEs enable us to generate large arrays of samples almost instantly. We leverage this feature to vectorize KDE draws and assemble batched joint proposals in a single pass.

To compare the efficiency of the Morph approximation under different sampling methods, we introduce a particular case of the GSS method Fan et al. [2011] that incorporates the Morph approximation as the proposal distribution. We refer to this method as *Morphed Steppingstone Sampling (MSS)*. The Python implementation of GSS presented in Ref. Zahraoui et al. [2025], combined with the Morph approximation, is utilized in Section 4.4.

4.4 Application to statistical examples

In this section, we compare both MorphZ and MSS to nested sampling through challenging statistical problems that reflect real inference scenarios. These statistical problems are considered benchmarks for assessing the accuracy of marginal likelihood estimation. Two of these benchmarks consist of the egg-box likelihood and the Gaussian shells likelihood Feroz et al. [2013]. As an additional benchmark, a different version of the peak likelihood originally discussed by Skilling [Skilling, 2006], represents one of the most challenging problems for evidence estimation methods. In this case, only Nested sampling can efficiently draw samples from the posterior distribution, yet requires a higher cost for an accurate marginal likelihood estimate Maturana-Russel et al. [2019a].

We set up each model, where the data is a null vector, as follows:

- ***Egg-box***: The two-dimensional likelihood is given by

$$L(\boldsymbol{\theta}) = \exp\left[\left(2 + \cos\left(\frac{\theta_1}{2}\right) + \cos\left(\frac{\theta_2}{2}\right)\right)^5\right], \tag{4.11}$$

  with uniform priors $\mathcal{U}(0, 10\pi)$ for both parameters.

- ***Gaussian-shells***: The d-dimensional likelihood is defined as

$$L(\boldsymbol{\theta}) = \text{circ}(\boldsymbol{\theta}, \boldsymbol{c_1}, r_1, w_1) + \text{circ}(\boldsymbol{\theta}, \boldsymbol{c_2}, r_2, w_2), \tag{4.12}$$

  where the circ function is given by

$$\text{circ}(\boldsymbol{\theta}, \boldsymbol{c}, r, w) = \frac{1}{\sqrt{2\pi w^2}} \exp\left[-\frac{(|\boldsymbol{\theta} - \boldsymbol{c}| - r)^2}{2w^2}\right], \tag{4.13}$$

  with $\boldsymbol{\theta}$ the parameter vector, $\boldsymbol{c}$ the center point, r the radius, and w the Gaussian radial profile of width. Following Feroz et al. [2009], we set the model to $d = 30$, $r_1 = r_2 = 2$, and $w_1 = w_2 = 0.1$ with a uniform prior $\mathcal{U}(-6, 6)^d$.

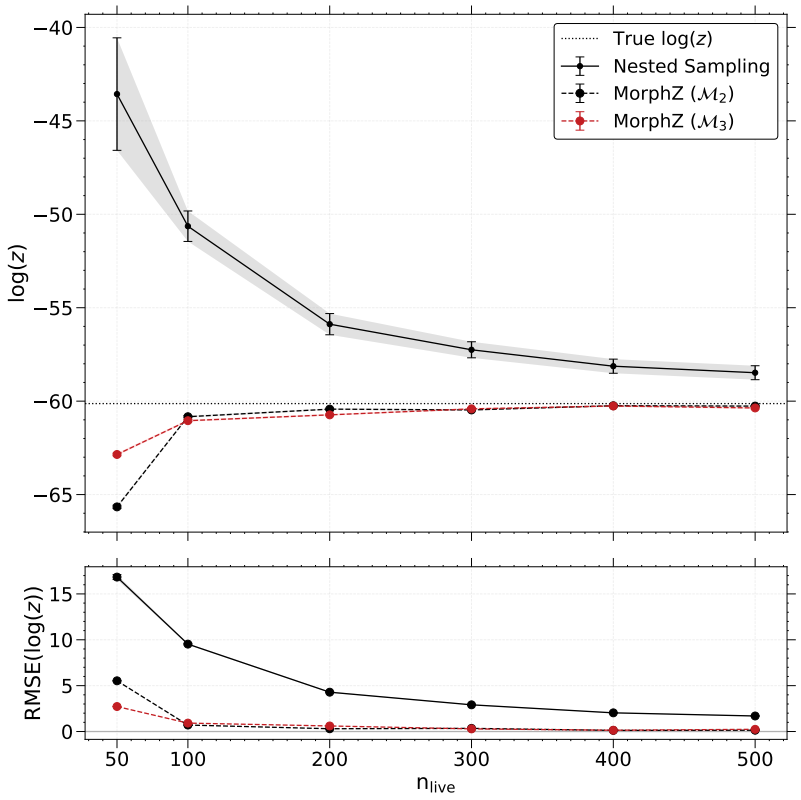


Figure 4.1: Convergence and accuracy of *Gaussian-shells* $\log(z)$ estimates versus n_{live} the number of live points. The posterior samples from one NS run with n_{live} is used for MorphZ estimates plotted at n_{live}, with 3000 likelihood calls (for BS proposal samples) per each MorphZ estimate. *Top:* Estimated $\log(z)$, NS (black) and the MorphZ with Morph approximations of second order ($\mathcal{M}_2$, gray) and third order ($\mathcal{M}_3$, red). Symbols denote means over 100 run; error bars indicate 1σ uncertainties. The horizontal dotted line marks the true $\log(z)$. *Bottom:* Root-mean-square error (RMSE) of $\log(z)$ relative to the true value .

- ***Peak-plateau***: The d-dimensional likelihood is expressed as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{d} \frac{1}{v\sqrt{2\pi}} \exp\left[-\frac{\theta_i}{2v^2}\right] + A \prod_{i=1}^{d} \frac{1}{u\sqrt{2\pi}} \exp\left[-\frac{\theta_i}{2u^2}\right], \tag{4.14}$$

  where $\boldsymbol{\theta}$ is the parameter vector, $v = 0.1$ the variance for the Gaussian plateau, $u = 0.01$ the variance for the Gaussian peak, and $A = 1$ the scale of the peak with uniform priors $\mathcal{U}(-0.05, 0.05)^d$ Maturana-Russel [2017] .

For all models, the nested sampling (NS) estimates are obtained using dynesty

Skilling [2006], Speagle [2020], Koposov et al. [2023]. For each model, we run NS with the number of live points set to $n_{\text{live}} = 500$ and an adequate stopping criterion. Since the Morph approximation requires posterior samples, we reuse the posterior samples produced from one of the NS runs for each model. For the MorphZ, we construct a second-order Morph approximation $\mathcal{M}_2$, with the standard *Silverman*'s rule for the KDE bandwidth. A total of 4000 NS posterior samples are split evenly: one half to construct $\mathcal{M}_2$, and the other half for the bridge sampling. For the MSS, we reuse the same Morph approximation and set the number of temperatures to 4.

Table 4.1 reports the true evidence $\log(z)_{\text{True}}$, the absolute distance between the mean $\overline{\log \hat{z}}$ of 100 estimate and $\log(z)_{\text{True}}$, the standard deviation of the estimates, and the number of likelihood calls (evaluations) used for each single estimate. For the *Egg-box* problem, both Morph-based methods produce $\log(z)$ estimates consistent with the NS estimate, demonstrating robustness to the high multi-modality of this likelihood. For the *Gaussian-shells* problem, MorphZ and MSS outperform NS. Their evidence estimates are more accurate and obtained at roughly two orders of magnitude lower computational cost. For the *Peak-plateau*, only the MorphZ estimator succeeded in producing an estimate closer to the true $\log(z)$, while NS requires a higher number of live points n_{live} to achieve similar accuracy.

Since MSS is based on path sampling, the *Peak-plateau* is sharply peaked over a narrow region but nearly flat over a wide plateau, so accurate path sampling requires a substantially larger number of likelihood evaluations to explore both regions adequately Skilling [2006]. Maturana-Russel 2017 Maturana-Russel [2017] shows that GSS can accurately estimate the marginal likelihood of the *Peak-plateau* problem with $A = 100$ using 48 temperatures. However, for $A = 1$ GSS fails to estimate $\log(z)$ even with 100 temperatures. In this case, the tails contain more probability than the spike region compared to the case of $A = 100$. In the current setup, MSS underestimates $\log(z)$ with high variance. Across all models, we have shown that MorphZ estimated

			GSS			MorphZ		
PTA	Model Name	d	calls	$\overline{\log(\hat{z})}$	std	calls	$\overline{\log(\hat{z})}$	std
NANOGrav	PSR J1705-1903 WN+RN	8	8×10^3	111 562.63	0.44	1×10^1	111 562.60	0.06
NANOGrav	PSR J1012+5307 WN+RN+DMGP	16	8×10^3	275 745.76	0.61	2×10^1	275 745.53	0.06
NANOGrav	PSR J1713+0747 WN+RN	32	8×10^3	751 285.52	0.78	3×10^2	751 284.62	0.05
EPTA	DR2 NEW PSRN+GWB	70	3.2×10^4	493 692.55	1.14	5×10^2	493 693.40	0.39
EPTA	DR2 FULL PSRN+CURN	74	3.2×10^4	607 134.56	1.20	5×10^2	607 135.48	0.15
NANOGrav	15 yr CURN	136	8×10^4	7 353 042.72	0.84	3×10^3	7 353 041.33	0.32
NANOGrav	15 yr HD	136	8×10^4	7 353 047.80	0.91	4×10^3	7 353 046.79	0.31

Table 4.2: Comparison of $\log(z)$ estimates using GSS, MorphZ for different PTA models with dimension d. The mean and standard deviation of 100 $\log(z)$ estimates per each model are displayed along the number of likelihood calls per estimate.

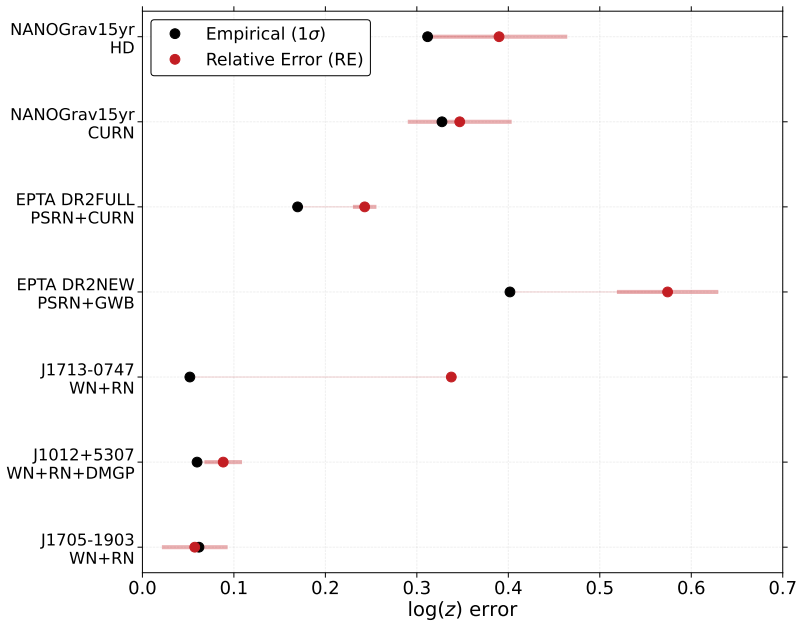


Figure 4.2: Comparison of bridge sampling RE and empirical errors for $\log(z)$ across models in Table 4.2. For each model, the red dot marks and the faint red bar shows respectively the mean and $\pm 1\sigma$ of 100 RE, and the black square marks the empirical error as the standard deviation of 100 estimates of $\log(z)$. The horizontal red segment connects the approximate and empirical values; indicating the degree of under/overestimation.

the marginal likelihood accurately and at a lower cost compared to nested sampling and MSS, while MSS has a slightly higher variance than MorphZ. Moreover, MSS requires the likelihood, as a function of the prior mass, to be concave to avoid failure. In practice, the likelihood as a function of the prior mass is usually approximately concave, in which case path sampling is effective.

In Figure 4.1, we compare nested sampling and MorphZ estimates for the *Gaussian-shells* example. First, NS is used to obtain posterior samples with increasing n_{live} the number of live points. Then at each n_{live}, MorphZ uses the same posterior samples drawn from one of the NS runs to construct second and third order Morph approximation (respectively $\mathcal{M}_2$ and $\mathcal{M}_3$ to show MorphZ consistency), along with 3000 likelihood calls (for BS proposal samples) per $\log(z)$ estimate. As n_{live} increases, the NS estimates converge gradually to the true $\log(z)$ reaching a $|\Delta| < 1.7$ at $\mathrm{n}_{live} = 500$. In contrast, both MorphZ approximations $\mathcal{M}_2$ and $\mathcal{M}_3$ are already consistent with the true value

with $|\Delta| < 0.9$ at $n_{live} = 100$ and remain more accurate thereafter. The RMSE shows that NS errors decrease with n_{live} yet remain significantly higher than those of MorphZ. Note that even at $n_{live} = 50$, where the posterior coverage is incomplete, MorphZ complements NS by leveraging the available posterior information provided by NS and closing the gap to an accurate estimate.

One can use NS with a reduced number of live points to explore parameter space and generate posterior samples, and then apply MorphZ in post-processing to obtain an estimate closer to the true $\log(z)$, substantially cutting computational costs.

4.5 Application to the Pulsar Timing Array

Recently, GSS was applied to different PTA analyses, demonstrating substantial reduction in computational cost for accurate evidence estimation Zahraoui et al. [2025]. In this application, we compare the efficiency of MorphZ for model evidence estimation across problems whose dimensionality ranges from 8 to 136. A second-order Morph approximation $\mathcal{M}_2$ is constructed using SGM, based on 2000 posterior samples for $8 - 32$ dimensions and 5000 samples for $70 - 136$ dimensions, and adopt *Silverman*'s rule for the KDE bandwidth. In these settings, SGM iterates over the complete set of pairwise seeds. For MorphZ, the bridge-sampling settings are those discussed in Section 4.3.2.

Table 4.2 presents the mean and standard deviation of 100 $\log(z)$ estimates computed with GSS and MorphZ, together with the number of likelihood evaluations per estimate (calls). For low-dimensional models $d = 8 - 16$, MorphZ produced an accurate $\log(z)$ estimate using only 10-20 likelihood evaluations. For mid-dimensional models $d = 32 - 74$, MorphZ consistently estimated $\log(z)$ with $300 - 500$ samples, even for costly models like the `EPTA DR2 NEW PSRN+GWB` model. For high-dimensional

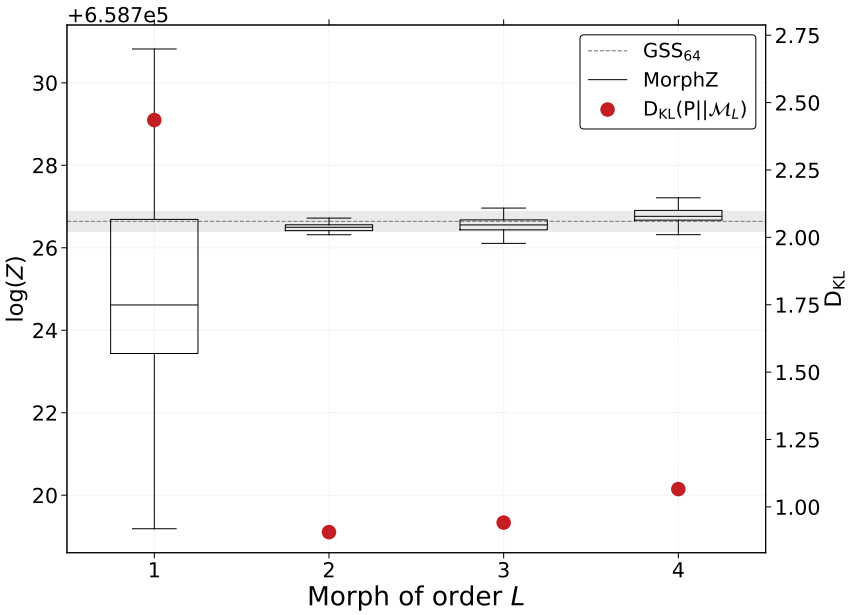


Figure 4.3: Comparison of $\log(z)$ estimates for the EPTA DR2FULL PSR+CURN model using GSS and MorphZ. The box plots show the distribution of $\log(z)$ for each Morph approximation order. The horizontal dashed line and shaded band indicate the mean and $\pm 1\sigma$ range of the reference GSS estimates with 64 temperatures. Red diamonds (right axis) denote the forward Kullback-Leibler divergence $D_{\mathrm{KL}}(P|\mathcal{M}_{\mathcal{B}_L})$ between the posterior and each corresponding order of the Morph approximation.

problems $d = 136$, MorphZ reduced the computational cost by a factor of 20 compared to GSS, making it particularly attractive for very expensive models such as the `NANOGrav 15yr HD` model.

In Figure 4.2, we compare the empirical error of $\log(z)$, defined as 1σ of 100 estimates, with the BS relative-error (RE) of each of the 100 estimates. Across all the models listed in Table 4.2, the relative error is either consistent, or is slightly overestimated compared to the empirical one. This mild overestimation can act as a safeguard, making RE conservative when empirical error is expensive to obtain, as in the case of `NANOGrav 15yr HD`. Overall, MorphZ achieves a two to three orders of magnitude reduction in likelihood calls relative to GSS, while providing more accurate $\log(z)$ and conservative relative error estimates, demonstrating efficiency over problems with varying difficulties.

In Figure 4.3, we compare different orders of the Morph approximation used in

Table 4.3: Ensemble sumary of $\Delta \log(\hat{z})$ per method relative to $\log(\hat{z})_{\mathrm{NS}}$ over 100 BBH simulations (see Figure 4.4). IQR denotes the 25–75% interval.

Method	Median	$\mathrm{IQR}_{25-75\%}$	Mean	std	$\lvert\Delta \log(\hat{z})\rvert < 0.5$	$\lvert\Delta \log(\hat{z})\rvert < 1.0$
SS	1.049	0.316 - 1.629	-0.373	4.788	12.1%	30.3%
$\mathrm{MorphZ}_{(\mathrm{NS})}$	0.034	-0.130 - 0.153	0.010	0.226	96.0%	100.0%
$\mathrm{MorphZ}_{(\mathrm{PT})}$	0.096	-0.005 - 0.223	0.106	0.219	94.9%	100.0%

MorphZ to estimate the marginal likelihood of `EPTA DR2 NEW PSRN+CURN` model. A reference $\log(z)$ is estimated using GSS with 64 temperatures and $30,000$ samples per temperature. For second and higher orders of the Morph approximation, the $\log(z)$ estimates are consistent with those from GSS while exhibiting similar variance at a cost of only 3000 likelihood evaluations per estimate. The Kullback-Leibler divergence shows that incorporating dependencies between posterior samples lowers the distance between the posterior and $\mathcal{M}_{L\geq 2}$. This additional information leads to more accurate estimates of $\log(z)$ and a reduction in its variance. The comparison of $\mathrm{D_{KL}}(P(\boldsymbol{\theta})||\mathcal{M}_{L\geq 2}(\boldsymbol{\theta}))$ does not, however, accurately reflect the theoretical prediction, because an exactly optimal Morph approximation can only be obtained for $L = 2$. We discuss this in more detail in Section 4.7.

4.6 Application to LVK Compact Binary Coalescences

In this section, we assess the accuracy of MorphZ for compact binary coalescence (CBC) and the GW150914 event for gravitational-wave inference. We compare MorphZ evidence estimates against nested sampling (NS; Skilling [2006], Speagle [2020], Koposov et al. [2023]) and Steppingstone Sampling (SS; Maturana-Russel et al. [2019b]) results while maintaining identical data, priors, likelihoods, and waveform models. All CBC inferences are carried out using bilby Ashton et al. [2019], Romero-Shaw et al. [2020]. To ensure full reproducibility, all scripts, prior definitions, and detailed sampler configuration used in this study are provided in the accompanying code repository (see

Section 4.8).

We consider two datasets. (i) A simulation study of 100 independent binary black hole (BBH) signal injections, each corresponding to a distinct synthetic GW source with different intrinsic parameters. The extrinsic parameters (ψ, RA, Dec, coalescence time, phase, luminosity distance, and inclination) are fixed to provide a controlled comparison with reduced extrinsic uncertainty. (ii) GW150914, analyzed with distance and time marginalization. The Power Spectral Density (PSD) for GW150914 are obtained from the public LVK parameter-estimation release Collaboration and Collaboration [2022, 2024], while the simulation study uses the bilby default design-sensitivity PSD. The 100 simulated CBC analyses uses 4 s of data ($injection$(CBC GW) + $noise$(Design-sensitivity PSD)), and the standard frequency-domain CBC Whittle likelihood with distance and time marginalization enabled Thrane and Talbot [2019]. We employ the `IMRPhenomPv2` waveform model Husa et al. [2016], Khan et al. [2016] and analyse data from the LIGO Hanford (H1) and Livingston (L1) detectors Abbott et al. [2016]. Nested sampling is performed with dynesty using $n_{\rm live}$ = 2000, $n_{\rm act}$ = 20, and the bilby implementation of the `rwalk` proposal. Additionally, parallel-tempered (PT) bilby-MCMC Ashton and Talbot [2021] is used independently with $n_{\rm temps}$ = 8, to retain posterior samples with $n_{\rm samples}$ = 2000 after adaptive tuning, thinning by γ = 0.2, proposal cycle GWA as implemented in bilby-MCMC Ashton and Talbot [2021], and $(L1, L2)$ = $(100, 5)$ sub-steps. The temperature ladder is set using the `Tmax_from_SNR` prescription with $T_{\rm max}$ = 20. SS is employed to estimate the evidence using the bilby-MCMC runs, a setting known to retain residual bias unless a finer temperature resolution is used.

To ensure a fair comparison across methods, we compare MorphZ estimates using each sampler's posterior (identical likelihoods and priors). This results in two independent estimates: MorphZ$_{\rm (NS)}$ and MorphZ$_{\rm (PT)}$. For both estimates, a second order Morph approximation $\mathcal{M}_2$ is constructed using 5000 NS and 2000 PT posterior samples and

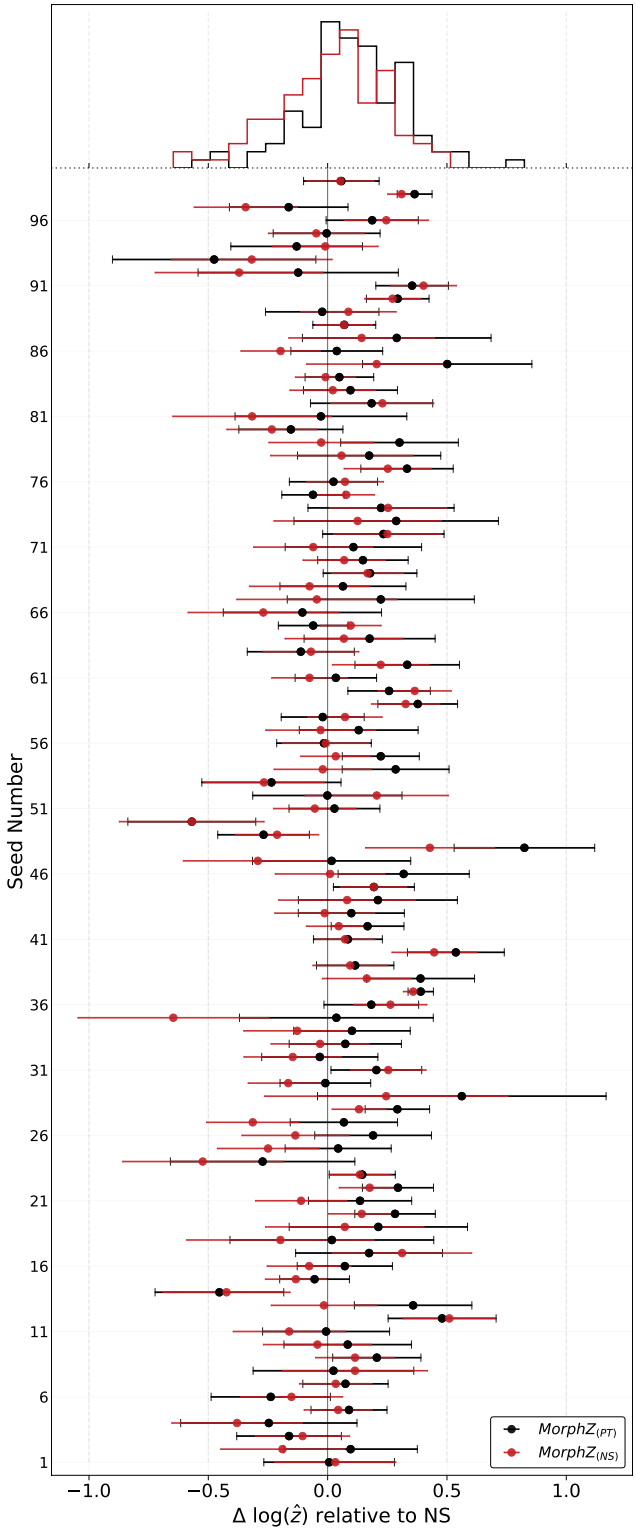


Figure 4.4: Comparison of dynesty, MorphZ (PT), and MorphZ (NS) across 100 independent BBH simulation (see Table 4.3). The mean and standard deviation of 100 $\log(z)$ estimate per simulation is used to compute $\Delta \log(\hat{z}) = \log(\hat{z})_{\text{method}} - \log(\hat{z})_{\text{NS}}$ for both independent MorphZ estimates. Note that $\log(z)_{\text{NS}}$ is used as a reference estimate because the true $\log(z)$ is unknown.

Table 4.4: $\Delta \log(\hat{z})$ summary per method relative to $\log(\hat{z})_{\mathrm{NS}}$ for the GW150914 event.

	NS		SS		MorphZ$_{(\mathrm{NS})}$		MorphZ$_{(\mathrm{PT})}$	
Dataset	$\log(\hat{z})$	σ_{NS}	$\Delta \log(\hat{z})$	σ_{SS}	$\Delta \log(\hat{z})$	std	$\Delta \log(\hat{z})$	std
GW150914	-7268.391	0.108	-4.007	0.122	-0.178	0.366	-0.246	0.415

Silverman's rule for KDE, and 5000 likelihood calls for the BS phase. 100 MorphZ estimates are obtained for each case (each individual estimate takes a few seconds to compute). For each dataset, we report the evidences from NS, SS, and both MorphZ independent estimates. We compare methods primarily via $\Delta \log(z)$ relative to NS.

For the 100 BBH simulations we compute the evidence difference $\Delta \log(\hat{z})$ = $\log(\hat{z})_{\mathrm{method}}$ – $\log(\hat{z})_{\mathrm{NS}}$. The $\log(z)_{\mathrm{NS}}$ is used as a reference estimate of the evidence, although it does not necessarily coincide with the true $\log(z)$ and may itself be subject to error. For NS and SS, $\sigma_{\mathrm{NS}} \approx \sqrt{H/n_{\mathrm{live}}}$ is nested sampling standard error, where H is the information gained about the parameters from the data, measured by the posterior relative to the prior Skilling [2006], and σ_{SS} is the bootstrap estimated error Ashton and Talbot [2021]. All $|\Delta \log(\hat{z})|$ for MorphZ$_{(\mathrm{NS})}$ and MorphZ$_{(\mathrm{PT})}$ are shown in Figure 4.4. Table 4.3 summarizes these results. Both MorphZ variants show small offsets, typically with $|\Delta \log(\hat{z})| \lesssim 0.1$. The population spread is also small with standard deviations around 0.22. More than 95% of simulations satisfy $|\Delta \log(\hat{z})| < 0.5$, and the whole population reaches $|\Delta \log(\hat{z})| < 1$ enabling unbiased model comparison between signal and noise models. By contrast, steppingstone sampling evidences runs show a much larger scatter and a clear bias, which is consistent with expectations for SS estimates computed with a coarse temperature ladder. These results show that the number of PT temperatures used to obtain the posterior samples remains sufficient for MorphZ to obtain accurate evidence estimates; however, they are not enough to obtain reliable SS estimates.

We next analyze the GW150914 event. Table 4.4 shows that both MorphZ independent estimates differ from the NS baseline by less than $|\Delta \log(\hat{z})| \lesssim 0.25$. Differences at the level of $|\Delta \log(\hat{z})| \lesssim 0.2$ are comparable to the natural run-to-run variation expected from nested sampling with $n_{\rm live} \sim 10^3$, since the evidence uncertainty scales as $\sigma_{\rm NS}$ for typical information values $H \sim 10{-}50$ Skilling [2006], Ashton et al. [2019]. The SS estimate from the GW150914 bilby-MCMC run is clearly biased relative to the nested-sampling baseline. [2]

These results show that MorphZ provides accurate and computationally efficient evidence estimates for CBC gravitational-wave analyses. Across both the simulation study and GW150914, MorphZ reproduces NS evidences within $|\Delta \log(\hat{z})|$ of order 0.1 to 0.25. Most importantly, MorphZ remains accurate when applied to posterior samples generated by low-temperature-resolution PT runs even though the corresponding SS evidences are biased. Because MorphZ requires only posterior samples and access to the likelihood and prior, it enables reliable evidence estimation without specialized evidence-targeted runs. This makes it well suited for rapid model selection, large simulation campaigns, and retrospective analyses of existing LVK posteriors.

4.7 Discussion

4.7.1 Constructing the Morph approximation

In this work, the Morph approximation showed its consistency and efficiency across different examples using kernel density estimates and the seeded greedy maximization Algorithm 1. The total correlation of all possible blocks is required to build such

[2]We assess posterior consistency for GW150914 using the Jensen–Shannon divergence (JSD) computed from 50-bin one-dimensional marginals. The median JSDs were 0.126 for dynesty compared with the GWTC-2.1 posterior, 0.121 for bilby-MCMC compared with GWTC-2.1, and 0.073 between our two samplers. The largest values (around 0.15 to 0.18) occurred for luminosity distance and $\chi_{\rm p}$. Crucially, these differences do not impact the goal of this work, which is accurate evidence $\log(z)$ estimation rather than precise agreement of posterior marginals.

an approximation. For block sizes $L \geq 3$, the evaluation of the total correlation over all possible blocks becomes increasingly expensive in dimensions exceeding ≈ 136 (see Section 4.5). Empirically, we find that thinning the sample size to the order of a few hundred keeps the cost significantly low while yielding similar total correlation estimates to those obtained with larger sample sizes. For $L \geq 3$, the total correlation can also be used to probe higher-order (linear and non-linear) dependencies among parameters, highlighting when they are physically linked by common mechanisms and potentially revealing new, otherwise hidden couplings in the model.

While this implementation of the Morph approximation is efficient, both the SGM algorithm and KDEs have intrinsic limitations. Although an optimal Morph approximation exists in principle for $2 \leq L \leq d$, it can be obtained in practice only for $L = 2$, where the problem reduces to maximum-weight graph matching. This problem is solved exactly via Blossom-based algorithms Edmonds [1965], Galil [1986]. While for $\mathcal{M}_{L>2}$, an approximately optimal solution can be acquired by solving an integer-linear-programming (ILP) set-packing formulation using, for instance, PuLP Mitchell et al. [2011], Forrest and Lougee-Heimer [2005], since matching-based methods no longer apply. The optimal solution problem becomes increasingly difficult. For example in Figure 4.5, the 50-dimensional case with $L = 5$, an algorithm has to select 10 blocks out of $\approx 1.22 \times 10^6$ possible choices that maximizes the total sum of the 10 blocks' weights. This performance comparison, presented in Figure 4.5, shows that SGM is significantly faster and outperforms ILP for $L \geq 3$. However, the SGM algorithm can be sensitive to the choice of initialization seed, may potentially converge to suboptimal local maxima, and its approximation error may increase for larger block sizes L. Since an optimal Morph approximation cannot be computed for orders three and higher, a direct quantitative comparison between different orders of the Morph approximation might not be faithful to the theoretical expectations for high dimensional problems (see example in Figure 4.3).

The construction of the Morph approximation might possibly generate singletons that are correlated and therefore not well represented, since they are assumed independent by definition in Equation 4.5. By default, the algorithm defines singletons as the parameters with the lowest multi-information, which may partially mitigate this under-representation. However, the quality of the marginal likelihood estimate is not significantly affected by this issue (see Figure 4.3). Nonetheless, several potential approaches could be considered to account for correlations among singletons. For instance, one could model the singletons as one block of length $\leq L - 1$ rather than a product of independent marginals.

In parallel, the performance of KDEs degrades rapidly with higher dimensions, where naïve KDEs become unreliable beyond $d \geq 10$ dimensions Tsybakov [2009], Virtanen et al. [2020b]. Although *Silverman*'s rule of thumb often suffices, KDE bandwidth choice is critical and largely determines the bias–variance balance; alternative methods can be used such as the Improved Sheather–Jones bandwidth Botev et al. [2010a] or bandwidth selection through isotropic cross-validation Silverman [2018], Duong and Hazelton [2005].

4.7.2 Evidence estimation using MorphZ

For marginal likelihood estimation, the Morph approximation serves as an importance sampling distribution for evidence estimators, and is especially effective for bridge sampling. We tested the Morph approximation on importance sampling (IS) estimator Chan and Eisenstat [2015] for the models presented in Table 4.2 and found that the estimator may occasionally be unstable for high-dimensional models (results not shown). IS requires an importance density with heavier tails than the posterior to avoid instability. In contrast, bridge sampling keeps the ratios bounded under both thinner and heavier tails mismatches provided proposal and posterior overlap. Beyond tail behavior, bridge

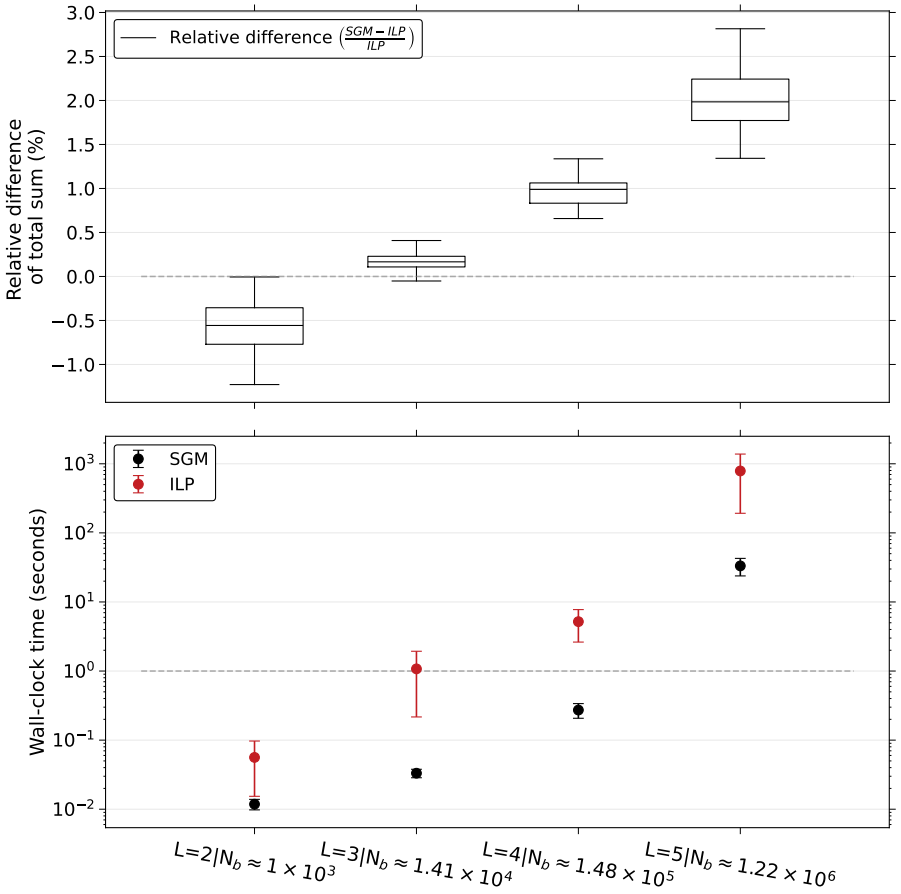


Figure 4.5: Relative difference of the total sum of simulated weights and runtime comparison of SGM and ILP across different orders of the Morph approximation for 100 independent sum of weight maximization simulation. Each simulation assign a weight drawn from Beta$(\alpha, 2)$ with $\alpha \sim$ Uniform$(5, 9)$ for each possible group from $\binom{50}{L}$. Top: Relative difference of the total sum of chosen weights, $(((\text{SGM} - \text{ILP})/\text{ILP}) \times 100\%)$, shown as box-and-whisker plots for $L = 2 - 5$ with the corresponding N_b number of blocks of length L ; the horizontal dashed line marks zero difference. Bottom: Wall-clock time in seconds for SGM (black) and ILP (red) on the same instances; markers denote mean and $(\pm 1\sigma)$. The dashed line marks 1s.

sampling places fewer constraints on the proposal distribution making it easier to construct suitable proposals. Using the Morph approximation together with the optimal bridge function and the iterative scheme, bridge sampling becomes nearly an automatic flow to implement and obtain accurate, efficient marginal-likelihood estimates, even for hierarchical models. This implementation reaches its full potential when posterior coverage is adequate, allowing the evidence to be obtained with only a few additional likelihood evaluations. For reliable evidence estimation, MorphZ requires at least minimal information about the posterior beyond the burn-in phase. The factored structure of the Morph approximation enables immediate generation of independent proposal draws in the current implementation. When posterior coverage is sparse, the newly accepted samples can be merged with previously collected samples to complete the posterior coverage.

4.8 Conclusion

In this work, we showed that MorphZ can deliver accurate evidence at substantially lower cost across statistical problems, real PTA noise and GWB inference, and LVK-CBC applications, spanning 2–136 parameters.

On statistical benchmarks, it matches or exceeds nested sampling accuracy while cutting likelihood calls by ~ 2 orders of magnitude; uniquely, it remains accurate on the challenging peak-plateau problem where path-based estimators struggle and standard NS requires more sampling efforts. MorphZ also improves NS estimates even when posterior coverage is still incomplete. In PTA models, it attains precise evidence with ∼10–20 calls for d=8–16, ∼300–500 for d=32–74, and $\sim 20\times$ fewer calls than GSS at d=136 even for expensive likelihoods, with bridge sampling relative errors that conservatively track empirical ones. In CBC analyses, MorphZ reproduces NS evidence values with $|\Delta \log(\hat{z})| \leq 0.1$ on simulations and $|\Delta \log(\hat{z})| \leq 0.25$ for GW150914, while

SS requires a greater number of temperatures for a more accurate estimate.

Currently, MorphZ can complement existing samplers by acting as a lightweight, post-processing evidence engine that needs only posterior draws. It can be paired with an NS run at reduced live points to explore and sample, then used to refine evidence accurately at far lower cost; likewise, it remains reliable when fed posterior samples from coarse-temperature PT-MCMC runs that are insufficient for unbiased SS estimates, enabling evidence estimation without dedicated evidence-targeted runs and facilitating retrospective analyses. Moreover, it can leverage posterior samples from fast, gradient-based samplers (e.g., HMC Duane et al. [1987], NUTS Hoffman and Gelman [2014], and Riemannian variants Girolami and Calderhead [2011]), turning their high-quality draws into accurate evidence estimates with minimal additional computation.

Looking ahead, the standard KDE used in Morph approximation construction, although efficient, could be replaced by richer models, such as copulas or normalizing flows, that better capture dependence and tail behavior, reducing bias and variance at a given sample size. Future work will also parallelize likelihood evaluation (vectorized and batched across bridges and sample blocks) to cut wall-clock time without increasing the number of calls, pushing MorphZ 's efficiency to new frontiers. Finally, bootstrap methods (parametric and nonparametric) will be explored to refine the relative-error estimates of bridge sampling and make the reported uncertainties track empirical variability more closely.

Acknowledgments

EMZ, PMR, AV, WS, and RM gratefully acknowledge support by the Marsden Fund Council grant MFP-UOA2131 from New Zealand Government funding, managed by the Royal Society Te Apārangi. This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part

from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and from the Victorian Higher Education State Investment Fund (VHESIF) provided by the Victorian Government.

Software: `ENTERPRISE` Ellis et al. [2019], Taylor et al. [2021], `enterprise_extension` Taylor et al. [2021], Ellis and van Haasteren [2017] ,Dalcín et al. [2005], `Jupyter` Kluyver et al. [2016], `matplotlib` Hunter [2007], `numpy` Harris et al. [2020], `scipy` Virtanen et al. [2020a], `arviz` Kumar et al. [2019].

Data and Software Availability

The MorphZ estimator is available as the open-source MorphZ Python package on GitHub (`https://github.com/EL-MZ/MorphZ`) and on PyPI (`https://pypi.org/project/morphZ/`) under the BSD 3-Clause License. The code used for the analysis, along with links to the public datasets, can be found on GitHub (`https://github.com/EL-MZ/MorphZ`)

4.9 KL divergence between P and $\mathcal{M}_{\mathcal{B}_L}$ in terms of total correlation

We recall the entropies

$$H(\boldsymbol{\theta}) = -\mathbb{E}_P[\log P(\boldsymbol{\theta})], \tag{4.15}$$

$$H(\boldsymbol{\theta}_b) = -\mathbb{E}_{P_b}[\log P_b(\boldsymbol{\theta}_b)], \tag{4.16}$$

$$H(\theta_s) = -\mathbb{E}_{P_s}[\log P_s(\theta_s)], \tag{4.17}$$

and define the block total correlation by

$$\mathcal{C}(\boldsymbol{\theta}_b) = \sum_{i \in b} H(\theta_i) - H(\boldsymbol{\theta}_b). \tag{4.18}$$

We now derive an expression for $D_{\mathrm{KL}}\big(P \| \mathcal{M}_{\mathcal{B}_L}\big)$ in terms of the total correlation. The *Kullback–Leibler* between P, the probability density, and $\mathcal{M}_{\mathcal{B}_L}$, the Morph approximation of order L, is given by

$$D_{\mathrm{KL}}\big(P \| \mathcal{M}_{\mathcal{B}_L}\big) = \mathbb{E}_P[\log P(\boldsymbol{\theta}) - \log \mathcal{M}_{\mathcal{B}_L}(\boldsymbol{\theta})] \tag{4.19}$$

$$= \mathbb{E}_P\left[\log P(\boldsymbol{\theta}) - \sum_{b \in \mathcal{B}} \log P_b(\boldsymbol{\theta}_b) - \sum_{s \in \mathcal{S}} \log P_s(\theta_s)\right] \tag{4.20}$$

$$= -H(\theta) + \sum_{b \in \mathcal{B}} \big(- \mathbb{E}_P[\log P_b(\boldsymbol{\theta}_b)]\big) \tag{4.21}$$

$$+ \sum_{s \in \mathcal{S}} \big(- \mathbb{E}_P[\log P_s(\theta_s)]\big). \tag{4.22}$$

Assuming that P_b and P_s denote the corresponding marginals of P (that is, P_b is the joint marginal distribution of the block b and P_s is the marginal distribution of the singleton s), we have

$$\mathbb{E}_P[\log P_b(\boldsymbol{\theta}_b)] = \mathbb{E}_{P_b}[\log P_b(\boldsymbol{\theta}_b)], \tag{4.23}$$

and similarly for P_s we have

$$\mathbb{E}_P[\log P_s(\theta_s)] = \mathbb{E}_{P_s}[\log P_s(\theta_s)]. \tag{4.24}$$

Hence,

$$D_{\mathrm{KL}}\big(P \| \mathcal{M}_{\mathcal{B}_L}\big) = -H(\boldsymbol{\theta}) + \sum_{b \in \mathcal{B}} H(\boldsymbol{\theta}_b) + \sum_{s \in \mathcal{S}} H(\theta_s) \tag{4.25}$$

$$\left(\text{since } \sum_{i \in \Gamma} H(\theta_i) = \sum_{s \in \mathcal{S}} H(\theta_s) + \sum_{b \in \mathcal{B}} \sum_{i \in b} H(\theta_i)\right) \tag{4.26}$$

$$= \Big(\sum_{i \in \Gamma} H(\theta_i) - H(\boldsymbol{\theta}) \Big) - \sum_{b \in \mathcal{B}} \Big(\sum_{i \in b} H(\theta_i) - H(\boldsymbol{\theta}_b) \Big) \tag{4.27}$$

$$= \Big(\sum_{i \in \Gamma} H(\theta_i) - H(\boldsymbol{\theta}) \Big) - \sum_{b \in \mathcal{B}} \mathcal{C}(\boldsymbol{\theta}_b). \tag{4.28}$$

Equivalently,

$$D_{\mathrm{KL}}(P \| \mathcal{M}_{\mathcal{B}_L}) = \mathcal{C}(\boldsymbol{\theta}) - \sum_{b \in \mathcal{B}} \mathcal{C}(\boldsymbol{\theta}_b), \tag{4.29}$$

where $\mathcal{C}(\theta) = \sum_{i \in \Gamma} H(\theta_i) - H(\boldsymbol{\theta})$ is the total correlation of the full vector, and Γ is the index set of all variables. All quantities are assumed finite.

Chapter 5

Manuscript 3

PTA hypothesis space

This manuscript investigates the hypothesis space of single-pulsar noise models in pulsar timing array analyses. It addresses the problem of model uncertainty in PTA noise characterization by comparing the posterior probabilities of competing noise models, rather than selecting a single preferred model. Using two versions of the Parkes Pulsar Timing Array second data release, the study shows that improved data processing increases sensitivity to weak signals such as solar-wind variations and reveals substantial model competition and degeneracies. The results indicate that many pulsars require Bayesian model averaging to propagate noise-model uncertainty into gravitational-wave-background searches. The manuscript is titled *Insights into the pulsar timing array hypothesis space*, has been published in *Physical Review D*.

Manuscript details

Title Insights into the pulsar timing array hypothesis space.

Authors El Mehdi Zahraoui, Patricio Maturana-Russel, Willem van Straten, Renate

Meyer, and Sergei Gulyaev.

Journal Physical Review D.

DOI https://journals.aps.org/prd/abstract/10.1103/lk8b-wq5d.

Abstract

We present novel insights into the pulsar-noise model space in pulsar timing array (PTA) experiments. Through a comparative analysis of the same Parkes PTA second data release observations processed with the 2020 and 2023 pipelines, we show that data processing materially changes the distribution of posterior support across competing noise hypotheses and increases sensitivity to weak contributions such as the solar wind. A solar-wind component appears in the highest-posterior hypothesis for ten pulsars under the 2023 pipeline, against four under the 2020 pipeline, while the corresponding mean electron density estimates are consistent with PPTA DR3 results. Furthermore, hypothesis-space analysis exposes model competition and degeneracies that are hidden by single-model summaries. Approximately 75% of the pulsars retain substantial posterior support for alternative noise descriptions. Therefore, understanding the nature of the pulsar noise hypothesis space is crucial for robust inference and nanohertz gravitational-wave background searches.

5.1 Introduction

Pulsar timing arrays (PTAs) were conceived as Galactic-scale detectors of gravitational waves (GWs) in the nanohertz frequency band Sazhin [1978], Hellings and Downs [1983]. By monitoring the pulse times of arrival from an ensemble of exceptionally stable millisecond pulsars (MSPs), PTAs search for the characteristic quadrupolar

angular correlations induced by an isotropic stochastic gravitational-wave background (GWB) in the pulsar timing residuals Foster and Backer [1990]. Over the past several decades, major collaborations including the North American Nanohertz Observatory for Gravitational Waves [NANOGrav; McLaughlin, 2013a], the Parkes Pulsar Timing Array [PPTA; Manchester et al., 2013], the European Pulsar Timing Array [EPTA; Ferdman et al., 2010], the Indian Pulsar Timing Array [InPTA; Joshi et al., 2018], the MeerKAT Pulsar Timing Array [MeerKAT PTA; Miles et al., 2022], and the Chinese Pulsar Timing Array [CPTA; Xu et al., 2023], have assembled increasingly longer and more precise pulsar-timing data sets.

Through Bayesian analyses, recent data releases have provided strong evidence for a common red noise spectrum process, with several analyses also reporting inter-pulsar spatial correlations consistent with those expected from a GWB Reardon et al. [2023a], Agazie et al. [2023b], Antoniadis et al. [2023a], Zic et al. [2023], Miles et al. [2023]. However, the timing residuals entering these analyses are themselves products of calibration, radio-frequency-interference excision, frequency averaging, pulse-profile template construction, and time-of-arrival estimation [Zic et al., 2023]. These processing choices are not statistically neutral: improved instrumental calibration and template matching can reduce apparent white noise and alter the inferred properties of red noise Rogers et al. [2024], while inadequately modelled frequency-dependent profile evolution and shape variability can bias measured arrival times Pennucci [2019], Lentati et al. [2017]. Because the adopted single-pulsar noise models can affect the inferred properties of a common red-spectrum process Reardon et al. [2023a], processing-induced changes in the timing residuals may propagate into PTA-wide gravitational-wave conclusions.

Bayesian inference has become a central component of PTA data analysis Van Haasteren et al. [2009], Lentati et al. [2013], Ellis [2013]. It is widely used both to characterize stochastic processes affecting individual MSPs Goncharov et al. [2021], Agazie et al. [2023b], Antoniadis et al. [2023a], Zic et al. [2023], Chen et al. [2025] and to

perform joint searches for a common GWB signal Reardon et al. [2023a], collaboration et al. [2024], Agazie et al. [2023b]. A robust assessment of the evidence for a GWB requires careful treatment of the stochastic processes that can affect PTA sensitivity Lentati et al. [2016]. These include intrinsic pulsar spin noise, commonly referred to as red noise (RN), variations in the interstellar dispersion measure (DM), and dispersive contributions from the solar wind (SW) Lentati et al. [2016], Goncharov et al. [2021], Hazboun et al. [2022]. Different combinations and parameterizations of these components give rise to a large discrete space of candidate noise models. Consequently, several studies have used Bayesian model comparison to identify preferred noise descriptions for individual pulsars and PTA data sets Jones et al. [2017], Goncharov et al. [2021], Chalumeau et al. [2022], Antoniadis et al. [2023b].

To navigate this large model space, PTA analyses have increasingly adopted model-index and trans-dimensional sampling methods. Product-space sampling, often referred to in PTA analyses as the hypermodel approach, embeds a collection of candidate models within a common augmented parameter space and introduces a discrete model-index parameter that determines which model is active Carlin and Chib [1995], Ellis et al. [2019], Ellis and van Haasteren [2017]. More recently, inference frameworks such as `tPTAbilby` have introduced binary indicator variables for candidate stochastic processes, allowing the number and combination of active noise components to be inferred directly from the data Marco et al. [2026]. These approaches permit parameter estimation and Bayesian model selection to be performed simultaneously within a single sampling run, while naturally incorporating the Bayesian penalty for unnecessary model complexity. Provided that the sampler mixes efficiently between models, posterior model probabilities can be estimated from their relative occupancy fractions, and Bayes factors can be obtained from ratios of these probabilities if the models are assumed apriori equally probable Ellis and van Haasteren [2017], Marco et al. [2026].

Recent advances in PTA marginal-likelihood estimation, including generalized

steppingstone sampling (GSS) [Zahraoui et al., 2025] and MorphZ [Zahraoui et al., 2026], have substantially improved the feasibility of estimating marginal likelihoods and, in turn, posterior model probabilities. Unlike product-space and trans-dimensional methods, which estimate model probabilities from the posterior occupancy of competing configurations [Carlin and Chib, 1995, Marco et al., 2026], these approaches evaluate the evidence for each model independently allowing a direct estimation of posterior model probabilities and Bayes factors [Kass and Raftery, 1995]. This enables parallel computation and model-specific convergence checks, and allows evidences to be reused when the candidate set or model priors change. Moreover, their efficiency extends to high-dimensional PTA models, including searches for common processes [Zahraoui et al., 2025, 2026].

Here, we use MorphZ to compare a fixed set of pulsar-noise models across the original PPTA DR2 Reardon et al. [2021] and its 2023 DR3 reprocessing Zic et al. [2023]. By examining the posterior model probabilities, we identify competing hypotheses, parameter degeneracies, and cases where the data cannot clearly distinguish between different combination of noise components. We then assess how changes in calibration, radio-frequency-interference mitigation, time-of-arrival estimation, and observing-system treatment affect inferred model preferences. This comparison reveals which noise-model conclusions are robust to reprocessing and which depend sensitively on the construction of the timing data set.

This paper is organized as follows. Section 5.2 introduces the Bayesian model comparison framework and defines the posterior model probabilities used throughout the analysis. Section 5.3 describes the original and reprocessed PPTA DR2 data sets. Section 5.4 presents the pulsar-noise models, prior choices, and marginal-likelihood calculations. Section 5.5 compares the resulting model spaces and examines the impact of data reprocessing on noise-model preferences. Finally, Section 5.6 summarizes the main findings and their implications for PTA gravitational-wave searches.

5.2 Bayesian inference

For a dataset D and a finite set of K competing hypotheses $\mathcal{H} = \{H_1, \ldots, H_K\}$, the posterior probability of hypothesis H_i is given by

$$p(H_i \mid D, \mathcal{H}) = \frac{p(D \mid H_i)\, p(H_i \mid \mathcal{H})}{\sum_{j=1}^{K} p(D \mid H_j)\, p(H_j \mid \mathcal{H})}, \tag{5.1}$$

where $p(H_i \mid \mathcal{H})$ is the prior probability assigned to H_i, and

$$Z_i \equiv p(D \mid H_i) = \int_{\Theta} p(D \mid \boldsymbol{\theta}_i, H_i)\, p(\boldsymbol{\theta}_i \mid H_i)\, \mathrm{d}\boldsymbol{\theta}_i, \tag{5.2}$$

where $\boldsymbol{\theta}_i \in \Theta_i$ is the parameter vector for model H_i, is its marginal likelihood, or evidence. The denominator of Eq. (5.1),

$$p(D \mid \mathcal{H}) = \sum_{j=1}^{K} Z_j\, p(H_j \mid \mathcal{H}), \tag{5.3}$$

is the model space normalizing constant. Because each fixed-model evidence Z_j is estimated separately, this normalizing constant is obtained without additional computation, enabling the estimation of posterior model probabilities.

The model priors $p(H_i \mid \mathcal{H})$ may represent different expectations about the occurrence of intrinsic red noise, dispersion-measure variations, solar-wind variations, or additional white-noise components. Uniform model priors assign equal prior support to every candidate model, whereas informed priors can favor simpler models or increase the prior probability of physically motivated noise components. Posterior model probabilities may therefore be recomputed under different model-prior choices without extra sampling cost, provided that $p(D \mid \boldsymbol{\theta}_i, H_i)$ and $p(\boldsymbol{\theta}_i \mid H_i)$ remain unchanged.

Table 5.1: Priors definition of the noise components used in the noise models for the PTA DR2 analysis.

Noise component	Parameters	Priors
WN	EFAC	$\mathcal{U}(0.01, 10)$
	EQUAD	$\log_{10}\mathcal{U}(10^{-9}, 10^{-5})$
	ECORR	$\log_{10}\mathcal{U}(10^{-9}, 10^{-2})$
RN	A_{RN}	$\log_{10}\mathcal{U}(10^{-20}, 10^{-11})$
	γ_{RN}	$\mathcal{U}(0, 7)$
DMv	A_{DMv}	$\log_{10}\mathcal{U}(10^{-20}, 10^{-11})$
	γ_{DMv}	$\mathcal{U}(0, 7)$
SW	n_{earth}	$\mathcal{U}(0, 30)$
	A_{SW}	$\log_{10}\mathcal{U}(10^{-11}, 10^{1})$
	γ_{SW}	$\mathcal{U}(-2, -1)$

5.3 Data

In this study, we analyze two instances of the Parkes Pulsar Timing Array (PPTA) second data release to assess the sensitivity of single-pulsar noise models to data processing choices prerequisite to the Bayesian analysis phase. The first is the original PPTA-DR2 Kerr et al. [2020], as used in the timing analysis of Reardon et al. [2021]. The second is the reprocessed version of the same observations, produced as part of the PPTA third data release Zic et al. [2023]. Both datasets share a common set of 25 millisecond pulsars observed with the 64-m Parkes radio telescope.

The original DR2 provides sub-banded arrival times derived from calibrated pulse profiles with 32 frequency channels per observation, dynamically sub-banded to maintain a minimum signal-to-noise ratio Kerr et al. [2020]. Times of arrival (ToAs) were obtained by cross-correlating profiles against band-specific analytic templates. The timing models and noise characterization are described in Reardon et al. [2021] and Goncharov et al. [2021], respectively. The data are referred to the TT(BIPM2018) timescale and the JPL DE436 solar system ephemeris.

The reprocessed DR2 was produced to enable consistent combination with new

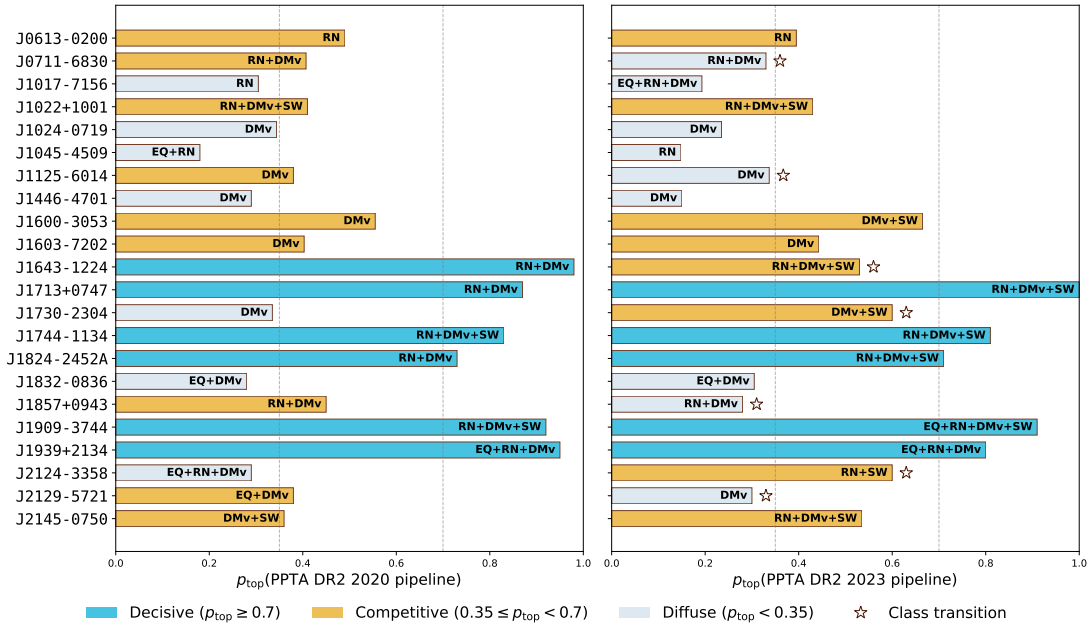


Figure 5.1: Summary of the model space for 22 pulsars noise analysis. For each pulsar, the model with the highest posterior probability $p_{\rm top}$ is shown for both the 2020 DR2 (left) and 2023 DR2 (right) pipelines. Bars are colour-coded by class: decisive ($p_{\rm top} \geq 0.7$; cyan), competitive ($0.35 \leq p_{\rm top} < 0.7$; light orange), and diffuse ($p_{\rm top} < 0.35$; light grey). Abbreviated top models are annotated; EF and EC are present in every $p_{\rm top}$ model and omitted for clarity. Stars mark pulsars that change classes between datasets.

ultra-wide-bandwidth (UWL) receiver observations Zic et al. [2023]. The reprocessing involved several changes, including improved radio-frequency interference excision using MeerGuard Lazarus et al. [2016], reduction to four frequency channels per band, new ToA templates derived from UWL wide-band pulse portraits Curyło et al. [2024], Fourier-domain Monte Carlo ToA estimation via pat Hotan et al. [2004], updated dispersion measure values and timing ephemerides, and adoption of the TT(BIPM2020) reference timescale. Additional frequency-dependent (FD) parameters were also required to absorb residual profile-evolution systematics introduced by the wide-band portrait templates. The full details of these changes are described in the DR3 release paper Zic et al. [2023]. In this work, we restrict our analysis to the 22 pulsars common to both datasets, enabling a direct comparison of the inferred noise models.

5.4 Analysis

In the following, we refer to the different hypotheses as noise models. We construct the noise models using ENTERPRISE Ellis et al. [2019], Taylor et al. [2021] via ENTERPRISE WARP [1]. Posteriors are sampled with EMCEE Foreman-Mackey et al. [2013] and cross-checked with POCOMC Karamanis et al. [2022a,b]; all marginal likelihoods are estimated with MORPHZ Zahraoui et al. [2026]. The posterior model probabilities are computed directly using these marginal likelihood estimates and a custom prior model probabilities.

We adopt the following notation: EF for EFAC (error scale factor), EQ for EQUAD (error added in quadrature), and EC for ECORR (extra correlated noise) as the white-noise parameters; and RN, DM_{v}, and SW for achromatic red noise, dispersion-measure variations, and solar-wind variations, respectivelyGoncharov et al. [2021]. Each stochastic process (SP) is modeled as a power law with two free parameters: the

[1] ENTERPRISE WARP: `https://enterprise-warp.readthedocs.io/`

log-amplitude $\log_{10} A_{\rm SP}$ and a spectral index $\gamma_{\rm SP}$. We use a SW model which includes an additional parameter: $n_{\rm earth}$ the mean free electron density at 1 AU Reardon et al. [2023a], Hazboun et al. [2022], Iraci et al. [2025]. Table 5.1 shows the prior distribution used for each noise component parameters.

For each pulsar and dataset, we evaluate K=13 candidate noise models after excluding white-noise only models. These excluded models have a zero posterior model probabilities for every pulsar in both datasets, due to a presence of excess noise in the dataset. These models span different combinations of EQ, EC, RN, and DMv. For the solar wind, we present only the four following models: EC+RN+SW, EC+DMv+SW, EC+RN+DMv+SW, and EQ+EC+RN+DMv+SW. These models are added to determine if the solar wind contribution is present in the dataset. For each pulsar, the thirteen models are shown in Figure 5.2 and in the appendix 5.8 (Figure 5.5 and 5.6), where EF is included in all models and is omitted from abbreviated labels. Deterministic signals and chromatic noise are included as free components for each relevant pulsar, following the description in Goncharov et al.Goncharov et al. [2021]. For each stochastic component, the number of Fourier harmonics is set to 30.

Our approach differs from the sequential procedure of Ref. Goncharov et al. [2021], where white-noise parameters are first fixed at their maximum-posterior values before red-noise components are selected iteratively. Such fixing influences the inference when white-noise and red-noise parameters are covariant [Goncharov et al., 2021]. In contrast, we evaluate 13 noise-component combinations simultaneously while keeping the white-noise parameters free, allowing the data rather than the procedure to determine the preferred model.

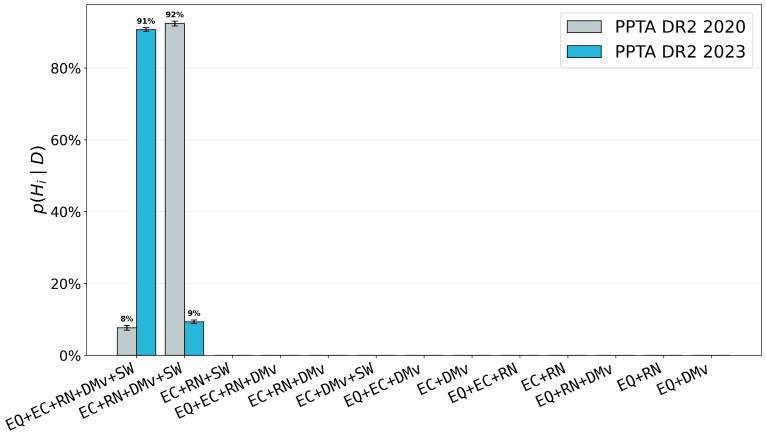
5.5 Results

We compare the candidate models through their posterior model probabilities defined in Eq. (5.1), assuming uniform prior probabilities over the $K = 13$ models considered for each pulsar. Throughout this section, the posterior probability of the highest-probability model is denoted by p_{top}. This quantity provides a compact description of the concentration of the model posterior, but it does not contain all of the information in the model space. In particular, two pulsars with the same value of p_{top} may differ substantially in how the remaining posterior probability is distributed among the other models.

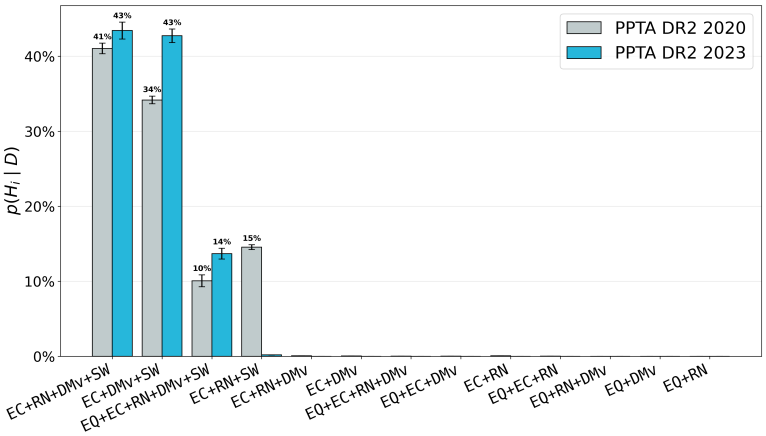
In the following, a pulsar is labeled *decisive* when $p_{\mathrm{top}} \geq 0.7$, *competitive* when $0.35 \leq p_{\mathrm{top}} < 0.7$, and *diffuse* when $p_{\mathrm{top}} < 0.35$. We use this classification, as shown in Figure 5.1, only as a descriptive summary of the model spaces of these datasets. These arbitrary thresholds are not used to accept the highest-probability model or to discard the remaining models. Instead, they distinguish model spaces dominated by one model from those in which the data retain substantial support for multiple noise descriptions.

5.5.1 Precision of the marginal-likelihood estimates

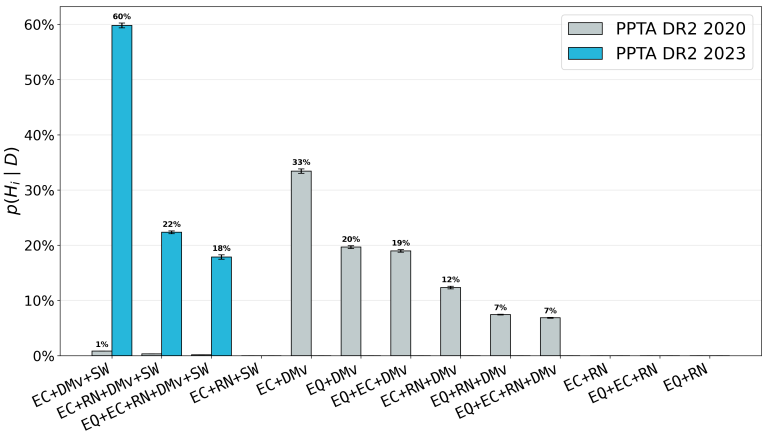
First, we assess the precision of the marginal-likelihood estimates used to construct the posterior model probabilities. Table 5.2 summarizes the run-to-run variation across five independent calculations for all model–pulsar combinations. For each dataset, the summary contains 1430 $\log(z)$ estimates. The median standard deviation across repeated calculations is 0.016 for the 2020 DR2 analysis and 0.013 for the 2023 DR2 analysis. The corresponding coefficient of variation values are $9.23 \cdot 10^{-5}$ and $1.78 \cdot 10^{-4}$, respectively, indicating a low variance between the independent runs. The close agreement between independent runs demonstrates that the marginal-likelihood estimates are numerically stable for both datasets. This stability is essential for the next stage of the



(a) PSR J1909−3744



(b) PSR J1022+1001



(c) PSR J1730−2304

Figure 5.2: Posterior model probabilities for one representative pulsar from each class under the 2020 DR2 (grey) and 2023 DR2 (cyan) pipelines. *(a)* Decisive: one model dominates at $\sim 90\%$. *(b)* Competitive: two SW-containing models each hold $\sim 40\%$. *(c)* Diffuse: the posterior fragments in 2020 but reorganizes around DMv+SW in 2023.

Table 5.2: Summary of the log marginal-likelihood estimates for the DR2 2020 and DR2 2023 datasets. The median and standard deviation of $\sigma_{\log(\widehat{z})}$ the standard deviation of 5×286 independent evidence estimates. The averaged coefficient of variation $\overline{\mathrm{CV}}_{\log(\widehat{z})}$ across these evidence estimates.

Dataset	$N_{\log(\widehat{z})}$	$\mathrm{med}_{\sigma_{\log(\widehat{z})}}$	$\mathrm{std}_{\sigma_{\log(\widehat{z})}}$	$\overline{\mathrm{CV}}_{\log(\widehat{z})}$
DR2 2020	1430	0.016	0.091	$9.23 \cdot 10^{-5}$
DR2 2023	1430	0.013	0.058	$1.78 \cdot 10^{-4}$

analysis, because it allows the uncertainty in the marginal-likelihood estimates to be propagated reliably into the posterior model probabilities shown in the figures.

The model probabilities uncertainties (see Figure 5.2) were obtained by a first-order propagation of the log-evidence uncertainties through the normalized exponential transformation. The log-evidence estimates are assumed to be independent, and model-prior probabilities were treated as fixed. Consequently, the uncertainty in each model probability includes contributions from the log-evidence uncertainties of all models entering the normalization, as described in Eq. 5.1. Consequently, the differences observed between competing models and between processing pipelines are not dominated by numerical fluctuations in the evidence calculations, but instead reflect genuine changes in the support provided by the data.

5.5.2 Summary of PPTA DR2 model space

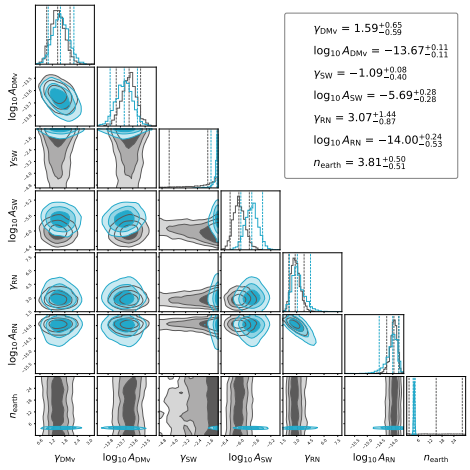
Figure 5.1 summarizes the model spaces of the 22 pulsars through p_{top} and the composition of the highest-probability model. Under the 2020 pipeline, six pulsars are classified as decisive, nine as competitive, and seven as diffuse. Under the 2023 pipeline, the corresponding numbers are five, eight, and nine. Thus, neither dataset generally identifies a unique noise model. The competitive and diffuse classes contain the majority of the pulsars under both processing pipelines.

Seven pulsars change classification between datasets. These transitions occur in

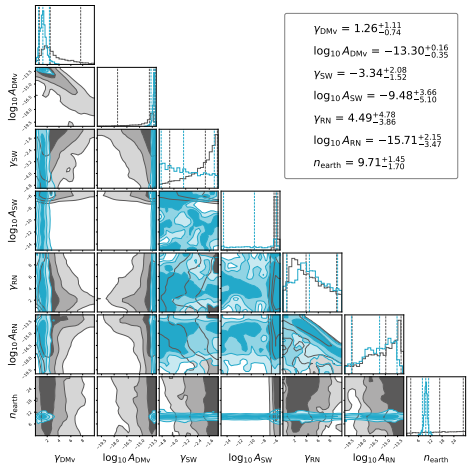
both directions. PSRs J0711−6830, J1125−6014, J1857+0943, and J2129−5721 move from the competitive class in 2020 to the diffuse class in 2023. For these pulsars, the reprocessed data distribute the posterior probability more broadly among the candidate models. PSR J1643−1224 moves from decisive to competitive, because its previously dominant model loses posterior probability to the SW variant models after reprocessing. Conversely, PSRs J1730−2304 and J2124−3358 move from diffuse to competitive, showing a concentration of posterior support for multiple SW models in the 2023 analysis.

The presence of transitions in both directions shows that improved reprocessing does not make every model posterior more concentrated. Rather, the changes in calibration, interference mitigation, template construction, ToA estimation, and timing-model treatment alter the relative ability of the candidate stochastic processes to describe each pulsar. For some pulsars this resolves part of the model ambiguity, whereas for others it exposes additional competition that was less apparent in the original dataset.

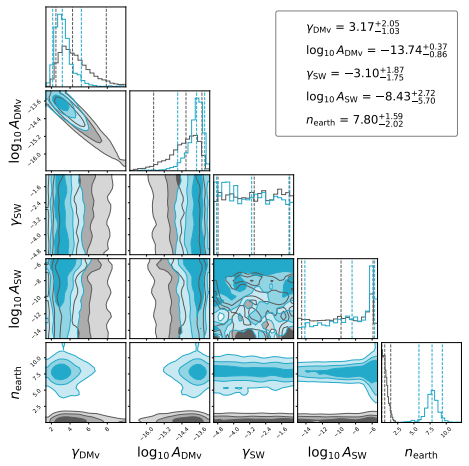
The most prominent change with the reprocessed data concerns the solar wind contribution. When models differing only in their white-noise components are grouped into the same stochastic process family, the most common highest probability family changes from RN+DMv in the 2020 analysis to RN+DMv+SW in the 2023 analysis. A solar wind component occurs in the highest-probability model for four pulsars under the original pipeline and for ten pulsars under the reprocessed pipeline. The ten pulsars are PSRs J1022+1001, J1600−3053, J1643−1224, J1713+0747, J1730−2304, J1744−1134, J1824−2452A, J1909−3744, J2124−3358, and J2145−0750. For these ten pulsars, Table 5.3 reports solar-wind-density estimates from the highest-posterior SW-containing model under the 2023 pipeline.



(a) PSR J1909−3744



(b) PSR J1022+1001



(c) PSR J1730−2304

Figure 5.3: Corner plots of the SP parameters of the highest-posterior 2023 noise model, evaluated using the 2020 DR2 (grey) and the 2023 DR2 (teal) pipelines: (a) PSR J1909−3744, representing a decisive model space; (b) PSR J1022+1001, representing a competitive space; and (c) PSR J1730−2304, which transitions from a diffuse to a competitive space.

5.5.3 Classification of PPTA DR2 model spaces

The complete model-posterior distributions in Figures 5.2, 5.5, and 5.6 show that the changes between pipelines take several distinct forms. For some pulsars, the same broad stochastic-process family remains preferred and the posterior concentration changes only moderately. In other cases, posterior support is redistributed among models that contain the same stochastic processes but differ in their white-noise components. A third group exhibits a qualitative change in the stochastic-process composition, most commonly through increased support for models containing SW.

Several model spaces remain broad under both pipelines. For example, the model posteriors of PSRs J1024−0719, J1045−4509, and J1446−4701 remain distributed among numerous models rather than becoming dominated by a single configuration. The absence of a decisive model in such cases is itself an inference from the data: several combinations of white noise, RN, and DMv remain plausible within the candidate set.

At the opposite extreme, pulsars such as J1713+0747, J1744−1134, J1909−3744, and J1939+2134 retain strongly concentrated model posteriors. Nevertheless, concentration of the model posterior does not necessarily imply complete stability under reprocessing. The identity of the preferred white-noise variant or the inclusion of a solar-wind component can change even when p_{top} remains in the decisive range. The three pulsars examined below illustrate the different forms of stability, competition, and posterior reorganization visible in the full sample.

5.5.3.1 A decisive model space: PSR J1909−3744

PSR J1909−3744 provides an example in which the inferred stochastic-process composition is stable while the preferred white-noise description changes. As shown in Figure 5.2(a), an RN+DMv+SW model family dominates the model space under both pipelines. The EC+RN+DMv+SW model carries approximately 92% of the posterior

probability in the 2020 analysis, whereas the corresponding model including EQ carries approximately 91% in the 2023 analysis.

The change in the highest-probability model is therefore associated primarily with the additional white-noise component rather than with a change in the inferred set of stochastic processes. This example demonstrates why it is useful to distinguish between the complete noise-model label and the underlying stochastic-process family. At the level of the latter, the model selection is highly stable.

The parameter posteriors in Figure 5.3(a) provide additional information that is not visible from the model probabilities alone. The RN and DMv parameters are constrained under both pipelines and remain mutually consistent. In contrast, the solar-wind parameters becomes more informative. In the 2020 analysis, the solar-wind posterior is comparatively broad and largely prior dominated. After reprocessing, the posterior becomes localized, yielding $n_{\rm earth} = 3.81^{+0.50}_{-0.51}\ {\rm cm}^{-3}$. Thus, even though the same stochastic-process family dominates both model spaces, the reprocessed observations provide substantially more information about the solar-wind contribution.

5.5.3.2 A competitive model space: PSR J1022+1001

PSR J1022+1001 illustrates a different form of model uncertainty. The model posterior is concentrated on four SW models in both datasets Figure 5.2(b). In the 2020 analysis, EC+RN+DMv+SW and EC+DMv+SW have posterior probabilities of approximately 0.41 and 0.34, respectively. In the 2023 analysis, the two models become nearly degenerate, each carrying approximately 0.43 of the posterior probability. The remaining support is divided between the EQ+EC+RN+DMv+SW and EC+RN+SW models. Within the candidate model set and under the adopted uniform model priors, the solar-wind contribution is robustly supported even though the complete noise model remains uncertain. The unresolved question is whether an additional achromatic RN process is required. Model competition therefore occurs mainly between the RN-containing

and RN-free models, rather than between models with and without a DM and SW contributions.

This interpretation is consistent with Figure 5.3(b). The DMv parameters and $n_{\rm earth}$ are well constrained by the reprocessed data, whereas the RN amplitude remains weakly constrained. Consequently, adding RN does not lead to a sufficiently distinct description for the data to select decisively between the two leading models. Conditional on the EC+RN+DMv+SW model, the inferred solar-wind density is $n_{\rm earth} = 9.71^{+1.45}_{-1.70}\ {\rm cm}^{-3}$. This density is consistent with the value obtained from the longer PPTA DR3 dataset Reardon et al. [2023a]. PSR J1022+1001 therefore demonstrates that strong support for an individual physical component can coexist with substantial uncertainty over the complete noise model.

5.5.3.3 Posterior reorganization: PSR J1730−2304

PSR J1730−2304 illustrates how reprocessing can reorganize a previously diffuse model posterior. Under the 2020 pipeline, the posterior probability is divided among several DMv-containing models, with no model reaching the competitive threshold. The hypothesis space therefore does not provide a clear preference for a particular combination of DMv, RN, and white-noise components.

In the 2023 analysis, the support becomes concentrated on models containing both DMv and SW. The EC+DMv+SW model carries approximately 60% of the posterior probability, causing the pulsar to move from the diffuse to the competitive class. The posterior does not become fully decisive, because non-negligible probability remains assigned to alternative SW-containing configurations. Nevertheless, the reprocessed data identify a much more restricted region of the model space than the original pipeline.

The parameter posterior in Figure 5.3(c) mirrors this model-level change. When both data sets are evaluated under the highest-posterior 2023 model, the solar-wind-density posterior is broad under the 2020 pipeline but becomes substantially more localized

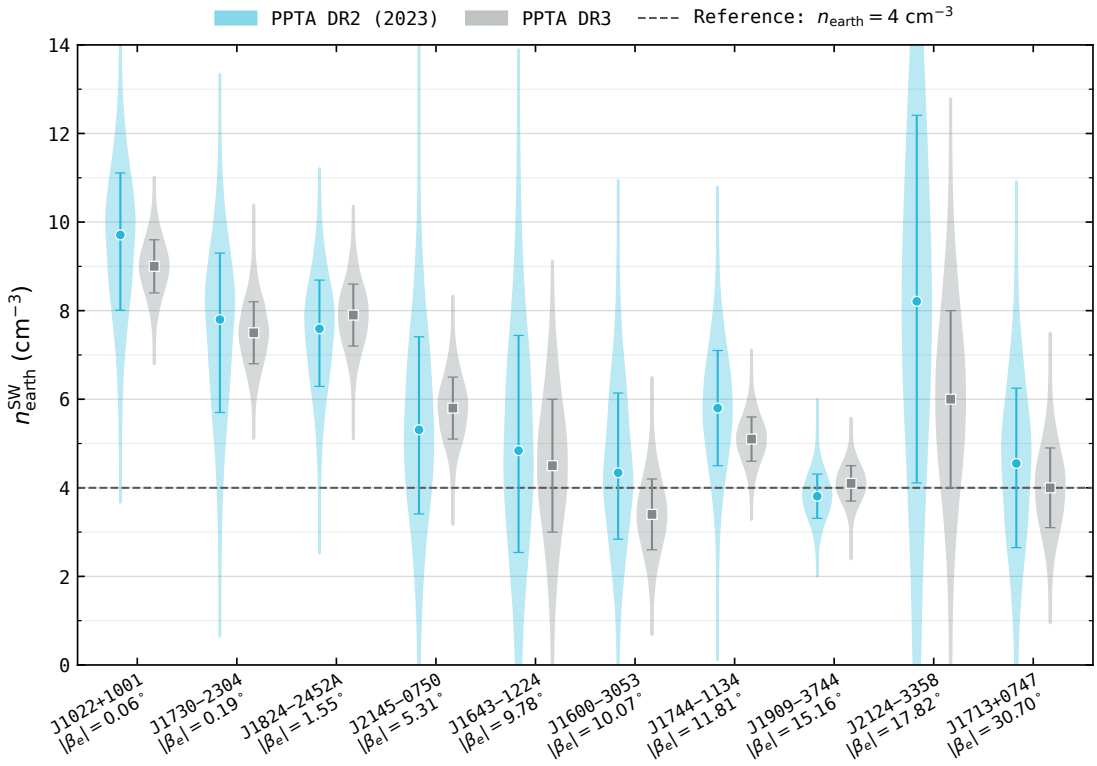


Figure 5.4: Comparison of the inferred solar-wind electron density at Earth, $n^{\rm SW}_{\rm earth}$, for ten PPTA pulsars in the PPTA DR2 (2023) and PPTA DR3 analyses[Reardon et al., 2023a]. The violin profiles show the distributions of $n^{\rm SW}_{\rm earth}$ posterior summaries from Table 5.3, while the markers and vertical bars indicate the posterior medians and central 68% credible intervals. The horizontal dashed line marks the reference value $n_{\rm earth} = 4\,{\rm cm}^{-3}$. Pulsars are arranged by increasing absolute ecliptic latitude, $|\beta_e|$, which is reported below each pulsar.

under the 2023 pipeline, with $n_{\rm earth}$ = $7.80^{+1.59}_{-2.02}$ $\rm cm^{-3}$. This is consistent with the PPTA DR3 Reardon et al. [2023a] estimate of 7.5 ± 0.7 $\rm cm^{-3}$. The agreement indicates that the solar-wind information recovered in the longer DR3 analysis is already present in the reprocessed DR2 observations.

5.5.4 Solar-wind emergence in PPTA DR2

Table 5.3 compares the 2023 DR2 estimates of $n_{\rm earth}$ with those obtained from the full PPTA DR3 dataset [Reardon et al., 2023a]. The table contains ten pulsars spanning absolute ecliptic latitudes from 0.06° to 30.70°. For every pulsar, the $n_{\rm earth}$ posterior median inferred from the reprocessed DR2 data is consistent with the corresponding DR3 estimate within the reported credible intervals.

Figure 5.4 displays the solar wind agreement between datasets, despite the shorter observing baseline of the DR2 subset and the absence of the UWL observations included in DR3 [Reardon et al., 2023a]. The DR2 credible intervals are generally wider, as expected from the reduced data volume, but they are centred on values compatible with the ones from the longer dataset. The comparison therefore shows that the increased support for SW models is not produced solely by the additional years of DR3 observations. Instead, improved processing of the observations already contained in DR2 is sufficient to reveal the solar-wind contribution.

The posterior distributions for the seven additional pulsars in Figure 5.7 support the same interpretation. Together with the three representative pulsars in Figure 5.3, they show that the 2023 analysis generally yields more localized solar-wind-density posteriors than the original pipeline. The improvement is not identical for every pulsar, and the widths of the distributions continue to depend on the available information for each line of sight. Nevertheless, the agreement with the independently obtained DR3 values across all ten pulsars establishes the consistency of the recovered signal.

Table 5.3: Solar-wind electron density at 1 au, $n_{\rm earth}$ ($\rm cm^{-3}$), $n_{\rm earth}^{\rm DR2}$ from the highest-posterior SW-containing model under the 2023 DR2 analysis and $n_{\rm earth}^{\rm DR3}$ the PPTA DR3 analysis [Reardon et al., 2023a]. Posterior medians and central 68% credible intervals are shown; pulsars are sorted by absolute ecliptic latitude $|\beta_e|$ in degrees.

Pulsar	$\lvert\beta_e\rvert$	$n_{\rm earth}^{\rm DR2}$	$n_{\rm earth}^{\rm DR3}$
J1022+1001	0.06	$9.71^{+1.4}_{-1.7}$	$9.0^{+0.6}_{-0.6}$
J1730−2304	0.19	$7.80^{+1.5}_{-2.1}$	$7.5^{+0.7}_{-0.7}$
J1824−2452A	1.55	$7.59^{+1.1}_{-1.3}$	$7.9^{+0.7}_{-0.7}$
J2145−0750	5.31	$5.31^{+2.1}_{-1.9}$	$5.8^{+0.7}_{-0.7}$
J1643−1224	9.78	$4.84^{+2.6}_{-2.3}$	$4.5^{+1.5}_{-1.5}$
J1600−3053	10.07	$4.34^{+1.8}_{-1.6}$	$3.4^{+0.8}_{-0.8}$
J1744−1134	11.81	$5.80^{+1.3}_{-1.3}$	$5.1^{+0.5}_{-0.5}$
J1909−3744	15.16	$3.81^{+0.5}_{-0.5}$	$4.1^{+0.4}_{-0.4}$
J2124−3358	17.82	$8.21^{+4.2}_{-4.7}$	$6.0^{+2.0}_{-2.0}$
J1713+0747	30.70	$4.55^{+1.7}_{-1.9}$	$4.0^{+0.9}_{-0.9}$

5.6 Conclusion

In this work, we have examined how the processing of a pulsar-timing dataset affects Bayesian noise-model inference by comparing the model spaces obtained from two processing versions of the PPTA second data release. For each of 22 pulsars, we evaluated the same set of 13 candidate noise models and used independently estimated marginal likelihoods to construct the complete model posterior. The comparison demonstrates that preprocessing choices can materially alter both the concentration and the composition of the inferred noise-model posterior. Seven pulsars change between the decisive, competitive, and diffuse classes. These transitions occur in both directions: for some pulsars, reprocessing concentrates support on a smaller set of models, whereas for others it distributes the support more broadly. Improved processing therefore does not merely increase the apparent decisiveness of model selection. Instead, it changes

which distinctions between the candidate stochastic processes are supported by the observations.

The clearest systematic change is the increased posterior support of SW-containing models. A solar-wind component appears in the highest-posterior model for ten pulsars in the 2023 analysis, compared with four under the original pipeline. The inferred values of the mean electron density at 1 AU are consistent with the values reported from the longer PPTA DR3 dataset for all ten pulsars Reardon et al. [2023a]. Although the credible intervals from the shorter DR2 baseline are generally wider, their agreement with the DR3 measurements shows that the solar-wind signal is already present in the DR2 observations.

The results also illustrate why support for a physical noise component should not be equated with decisive selection of a complete model. For J1022+1001, for example, nearly degenerate models provide strong and stable support for SW while differing over the inclusion of achromatic RN. Conversely, for J1909−3744, the stochastic-process family remains stable even though the preferred white-noise extension changes. The full model posterior separates these different forms of uncertainty, whereas a comparison based only on the highest-probability model would obscure them.

Most pulsars in both analyses have competitive or diffuse hypothesis spaces. Averaging over the two pipelines, approximately 75% of the pulsars are not adequately summarized by a single overwhelmingly preferred noise model. Selecting only the highest-probability model for these pulsars would overstate confidence in a particular noise description and would fail to propagate the uncertainty associated with alternative, data-supported models.

This issue is directly relevant to PTA searches for a common red-spectrum process and a gravitational-wave background. Noise components inferred for individual pulsars can be covariant with a common process, and a noise model selected in a preliminary analysis is usually treated as fixed in the subsequent PTA-wide analysis. Uncertainty

that is removed at the single-pulsar stage is therefore not automatically recovered later. The single-pulsar noise analysis should instead produce a posterior distribution over plausible noise models that can be marginalized over, or otherwise propagated, in the common-process analysis.

The numerical model probabilities reported here are conditional on the candidate model space and on the uniform model priors adopted in this work. The decisive, competitive, and diffuse labels are similarly intended as descriptive categories rather than universal model-selection thresholds. Because the marginal likelihood of each model has been estimated independently, the model posterior can be recomputed under alternative model-prior choices without repeating the parameter-inference calculations. This provides a practical route for examining the sensitivity of downstream conclusions to different physical expectations about the occurrence of RN, DMv, SW, and additional white-noise components.

More broadly, this study shows that data processing is a component of the systematic uncertainty budget of Bayesian PTA inference. Hypothesis-space analysis provides a direct means of exposing that uncertainty by retaining the relative support for every candidate model, identifying degeneracies between noise components, and distinguishing robust component-level conclusions from uncertain complete-model assignments. As described in van Haasteren van Haasteren [2025], propagating this information through Bayesian model averaging offers a more principled foundation for gravitational-wave inference than conditioning on a single selected noise model.

5.7 acknowledgments

EMZ, PMR, WS, and RM gratefully acknowledge support by the Marsden Fund Council grant MFP-UOA2131 from New Zealand Government funding, managed by the Royal Society Te Apārangi. We thank the Parkes Pulsar Timing Array collaboration for

making their data publicly accessible. This work uses the third data release of the Parkes Pulsar Timing Array [Zic et al., 2023], available on the CSIRO Data Access Portal [Zic et al., 2023]. We also thank Boris Goncharov for making the PPTA DR2 noise analysis publicly reproducible (`https://github.com/bvgoncharov/ppta_dr2_noise_analysis`). This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and from the Victorian Higher Education State Investment Fund (VHESIF) provided by the Victorian Government.

Software: `ENTERPRISE` Ellis et al. [2019], Taylor et al. [2021], `enterprise_extension` Taylor et al. [2021], Ellis and van Haasteren [2017], Dalcín et al. [2005], `Enterprise Warp`, `EMCEE` Foreman-Mackey et al. [2013], `PocoMC` Karamanis et al. [2022b], `MorphZ` Zahraoui et al. [2026], `Jupyter` Kluyver et al. [2016], `matplotlib` Hunter [2007], `numpy` Harris et al. [2020], `scipy` Virtanen et al. [2020a], and `arviz` Kumar et al. [2019].

5.8 Posterior model-probability comparisons

Figures 5.5 and 5.6 show the posterior model-probability comparisons under the 2020 DR2 and 2023 DR2 pipelines for the 19 pulsars not displayed in the main text.

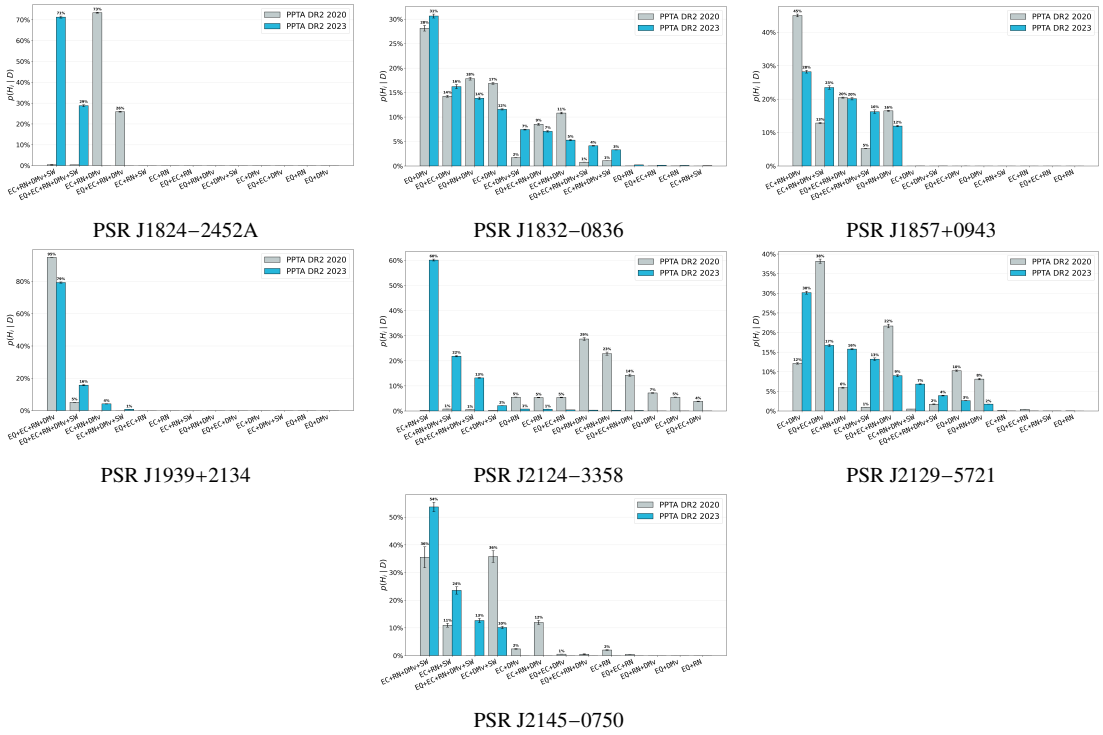
The notation follows that of the main text. EF is present in every model and is therefore omitted from the abbreviated model labels. EQ, EC, RN, DMv, and SW denote EQUAD, ECORR, achromatic red noise, dispersion-measure variations, and solar-wind variations, respectively.



Figure 5.5: Posterior model probabilities obtained with the 2020 DR2 pipeline (grey) and the 2023 DR2 pipeline (teal) for pulsars not shown in the main text. This Figure shows the first 12 pulsars.

5.9 Solar-wind posterior distributions

Figure 5.7 shows the joint and marginalized posterior distributions for the preferred noise models of seven additional pulsars whose highest-posterior model under the 2023 pipeline contains a solar-wind component.

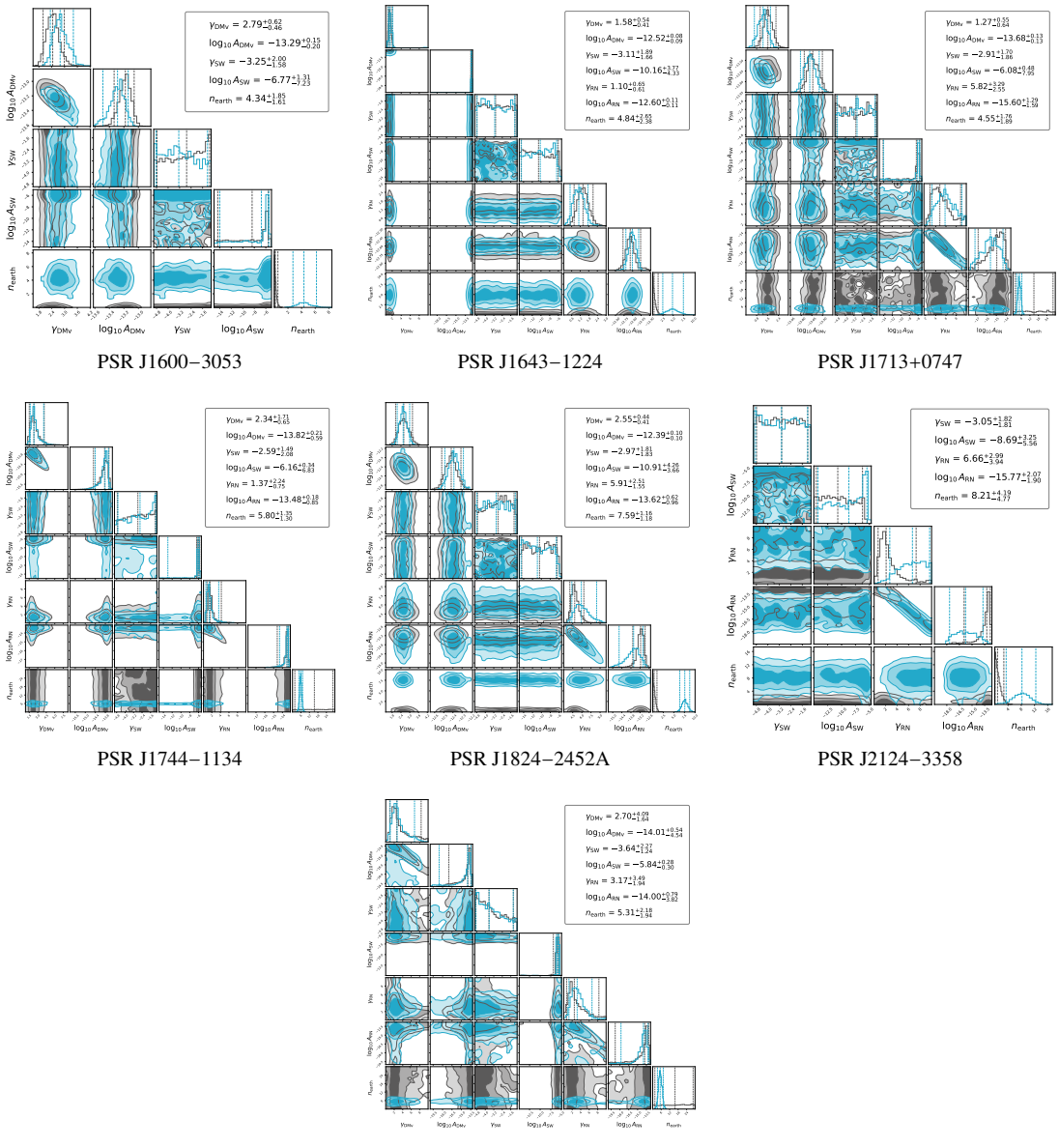


PSR J1824−2452A PSR J1832−0836 PSR J1857+0943

PSR J1939+2134 PSR J2124−3358 PSR J2129−5721

PSR J2145−0750

Figure 5.6: Posterior model probabilities obtained with the 2020 DR2 pipeline (grey) and the 2023 DR2 pipeline (teal) for the remaining seven pulsars not shown in the main text.



PSR J1600−3053 PSR J1643−1224 PSR J1713+0747

PSR J1744−1134 PSR J1824−2452A PSR J2124−3358

PSR J2145−0750

Figure 5.7: Joint and marginalized posterior distributions for the parameters of the preferred noise models of the seven additional pulsars showing statistically significant solar-wind variations. Results obtained with the 2020 PPTA DR2 pipeline are shown in grey, while those obtained with the 2023 PPTA DR2 pipeline are shown in teal.

Chapter 6

Discussion and Conclusion

6.1 Discussion

The central aim of this thesis has been to develop Bayesian techniques that make pulsar timing array inference more reliable, and computationally efficient. This aim was motivated by recent statistical evidence for a common low-frequency process consistent with a gravitational-wave background. Currently, the limiting difficulty is no longer just acquiring more and higher-quality timing data, but the ability to properly separate gravitational-wave signals from intrinsic pulsar noise, chromatic propagation effects, solar-wind variations, and instrumental systematics.

The work presented here addresses the Bayesian and computational challenges. Bayesian evidence estimation is one of these challenges, required both for selecting or weighting pulsar noise models and for comparing competing common-process hypotheses in full-array GWB searches. However, the marginal likelihood is notoriously difficult to compute accurately for the high-dimensional and expensive likelihoods that arise in PTA analyses. Chapters 3 and 4 addressed this by developing efficient posterior-informed evidence estimators, while Chapters 5 demonstrated how these methods can be used to understand the uncertainty across the PTA noise-model landscape.

The first major contribution of this work was introducing efficient marginal-likelihood estimation into the PTA Bayesian framework. Generalized stepping-stone sampling mitigated the inefficiency of traditional power-posterior methods by replacing the prior endpoint with a posterior-informed reference distribution. This shortened the path between distributions, significantly reducing the computational cost of accurate evidence estimates. This foundation led to the development of the Morph approximation and the MorphZ estimator, designed specifically to leverage available posterior samples. The Morph approximation preserves the strongest parameter dependencies while keeping the proposal density tractable for high-dimensional problems. Combined with bridge sampling, MorphZ provides a post-processing evidence estimator applicable not only to PTA analyses but across a vast range of Bayesian applications.

Consequently, we applied these tools to the single-pulsar noise-model hypothesis space in the Parkes Pulsar Timing Array second data release. Comparing the 2020 and 2023 processing pipelines revealed that noise-model support can change substantially upon data reprocessing, and that this ambiguity affects a large portion of the array. In both pipelines, roughly three-quarters of the pulsars did not overwhelmingly support a single model. The reprocessed data also uncovered stronger evidence for solar-wind contribution previously hidden in the 2020 dataset. The inferred solar-wind electron densities aligned with independent PPTA DR3 measurements, demonstrating that improved calibration directly enhances sensitivity to the chromatic low-frequency signals relevant for GWB searches.

6.2 Future Work

While the methodological advancements presented in this thesis establish a robust foundation for evaluating broad hypothesis spaces, maximizing PTA sensitivity requires continued exploration of model uncertainties. Building upon the framework established

in Section 5, the practical implementation of Bayesian model averaging (BMA) across the full array represents a primary direction for future research.

BMA provides a principled alternative to binary model selection by marginalizing over the single-pulsar noise-model space. Future work must focus on operationalizing this framework for full-array gravitational-wave searches without introducing prohibitive computational bottlenecks.

Preliminary testing on PSR J1022+1001, utilizing the PPTA DR2 pipeline, demonstrated the viability of this approach as a post-processing step. These initial tests showed that BMA can act as a natural extension of the hypothesis-space analysis shown in Chapter 5. Specifically, by utilizing the Morph approximation to construct a continuous representation of the model-averaged distribution, post-processed BMA results successfully matched the accuracy of computationally expensive direct-sampling methodologies.

Extending this specific technique from isolated test cases to full-array gravitational-wave searches represents a critical next step. Fully integrating Morph-assisted BMA into the global PTA pipeline will ensure that complex, ambiguous array noise is rigorously marginalized out, thereby maximizing the reliability of future gravitational-wave inferences.

6.3 Implications for PTA Sensitivity

The sensitivity of a pulsar timing array to nanohertz gravitational waves depends heavily on the ability to preserve a weak common signal while correctly assigning other sources of low-frequency power. Intrinsic spin noise, dispersion-measure variations, solar-wind delays, calibration errors, and instrumental systematics can all contribute power in the exact frequency range where a stochastic GWB is expected. If these processes are mis-modeled, their power can leak into the common-spectrum component and bias the

inferred gravitational-wave signal. Therefore, this work supports a model-aware view of PTA sensitivity. Where, the uncertainty associated with the modeling choices are also propagated into the full array SGWB analysis.

A single-pulsar Bayes factor may provide limited support for a low-frequency process in isolation, but that process remains relevant in a full-array analysis because it is covariant with the common red process. Stripping such a component out before the global analysis forces the common signal to absorb variance that belongs in the pulsar-specific noise model. Alternatively, the model-averaging framework provides a principled alternative by marginalizing over the single-pulsar noise-model space, retaining crucial uncertainty rather than converting it into a binary decision. Fast and accurate evidence estimators enable the exploration of these larger model spaces, ensuring that ambiguity is quantified and propagated directly into the final gravitational-wave inference.

6.4 Conclusions

This thesis has demonstrated that scalable Bayesian computation is central to the future of pulsar timing array science. Moving from the initial evidence of a common process toward the precision characterization of the nanohertz gravitational-wave background requires statistical methods capable of comparing complex models and exploring high-dimensional parameter spaces. By introducing efficient marginal-likelihood estimators such as GSS and the MorphZ this work has made rigorous Bayesian model comparison highly practical for expensive PTA likelihoods. Applying these tools to real array data revealed that single-pulsar noise models often harbor significant ambiguity, which can be resolved using novel Bayesian technique such as Bayesian model averaging.

Maximizing PTA sensitivity requires a triad of improved observations, careful physical modeling, and robust Bayesian computation. As timing data sets continue to

grow in volume and precision, the ability to rapidly evaluate broad hypothesis spaces will be the deciding factor in separating true astrophysical signals from complex array noise. The methodological advancements developed here provide a crucial foundation for that goal, ensuring that future gravitational-wave inferences are both computationally accessible and statistically sound.

References

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, et al. Search for tensor, vector, and scalar polarizations in the stochastic gravitational-wave background. *Physical review letters*, 120(20):201102, 2018.

B. P. Abbott et al. GW150914: The Advanced LIGO Detectors in the Era of First Discoveries. *Phys. Rev. Lett.*, 116(13):131103, April 2016. doi: 10.1103/PhysRevLett.116.131103.

A. Afzal, G. Agazie, A. Anumarlapudi, A. M. Archibald, Z. Arzoumanian, P. T. Baker, B. Bécsy, J. J. Blanco-Pillado, L. Blecha, K. K. Boddy, et al. The nanograv 15 yr data set: Search for signals from new physics. *The Astrophysical Journal Letters*, 951 (1):L11, 2023. doi: 10.3847/2041-8213/acdc91.

G. Agazie, A. Anumarlapudi, A. M. Archibald, Z. Arzoumanian, P. T. Baker, B. Bécsy, L. Blecha, A. Brazier, P. R. Brook, S. Burke-Spolaor, et al. The nanograv 15 yr data set: Detector characterization and noise budget. *The Astrophysical Journal Letters*, 951(1):L10, 2023a.

G. Agazie, A. Anumarlapudi, A. M. Archibald, P. T. Baker, B. Bécsy, L. Blecha, A. Bonilla, A. Brazier, P. R. Brook, S. Burke-Spolaor, et al. The nanograv 15 yr data set: Constraints on supermassive black hole binaries from the gravitational-wave background. *The Astrophysical Journal Letters*, 952(2):L37, 2023b.

M. A. Alpar, A. F. Cheng, M. A. Ruderman, and J. Shaham. A new class of radio pulsars. *Nature*, 300:728–730, 1982.

P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse, P. Bender, E. Berti, P. Binetruy, M. Born, D. Bortoluzzi, et al. Laser interferometer space antenna. *arXiv preprint arXiv:1702.00786*, 2017.

P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse, et al. Laser Interferometer Space Antenna. *arXiv e-prints*, art. arXiv:1702.00786, Feb. 2017. doi: 10.48550/arXiv.1702.00786.

S. Andreon and B. Weaver. Bayesian methods for the physical sciences. *Switzerland: Springer Inetrnational Publishing*, 52, 2015.

J. Antoniadis, P. Arumugam, S. Arumugam, S. Babak, M. Bagchi, A.-S. B. Nielsen, C. Bassa, A. Bathula, A. Berthereau, M. Bonetti, et al. The second data release from the european pulsar timing array-iii. search for gravitational wave signals. *Astronomy & Astrophysics*, 678:A50, 2023a.

J. Antoniadis, P. Arumugam, S. Arumugam, S. Babak, M. Bagchi, A.-S. B. Nielsen, C. Bassa, A. Bathula, A. Berthereau, M. Bonetti, et al. The second data release from the european pulsar timing array-ii. customised pulsar noise models for spatially correlated gravitational waves. *Astronomy & Astrophysics*, 678:A49, 2023b. doi: 10.1051/0004-6361/202346842.

A. M. Archibald, V. M. Kaspi, S. Bogdanov, J. W. T. Hessels, I. H. Stairs, S. M. Ransom, and M. A. McLaughlin. X-ray Variability and Evidence for Pulsations from the Unique Radio Pulsar/X-ray Binary Transition Object FIRST J102347.6+003841. *ApJ*, 722:88–95, Oct. 2010. doi: 10.1088/0004-637X/722/1/88.

Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. Chamberlin, S. Chatterjee, B. Christy, J. M. Cordes, N. Cornish, K. Crowter, P. B. Demorest, et al. The nanograv nine-year data set: observations, arrival time measurements, and analysis of 37 millisecond pulsars. *The Astrophysical Journal*, 813(1):65, 2015.

Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. Chamberlin, S. Chatterjee, B. Christy, J. M. Cordes, N. J. Cornish, F. Crawford, H. T. Cromartie, et al. The nanograv 11-year data set: High-precision timing of 45 millisecond pulsars. *The Astrophysical Journal Supplement Series*, 235(2):37, 2018.

Z. Arzoumanian, P. T. Baker, H. Blumer, B. Bécsy, A. Brazier, P. R. Brook, S. Burke-Spolaor, S. Chatterjee, S. Chen, J. M. Cordes, et al. The nanograv 12.5 yr data set: search for an isotropic stochastic gravitational-wave background. *The Astrophysical Journal Letters*, 905(2):L34, 2020.

Z. Arzoumanian, P. T. Baker, H. Blumer, B. Becsy, A. Brazier, P. R. Brook, S. Burke-Spolaor, M. Charisi, S. Chatterjee, S. Chen, et al. Searching for gravitational waves from cosmological phase transitions with the nanograv 12.5-year dataset. *Physical review letters*, 127(25):251302, 2021.

G. Ashton and S. Khan. Multiwaveform inference of gravitational waves. Phys. Rev. D, 101(6):064037, Mar. 2020. doi: 10.1103/PhysRevD.101.064037.

G. Ashton and C. Talbot. BILBY-MCMC: an MCMC sampler for gravitational-wave inference. MNRAS, 507(2):2037–2051, Oct. 2021. doi: 10.1093/mnras/stab2236.

G. Ashton, M. Hübner, P. D. Lasky, C. Talbot, K. Ackley, et al. BILBY: A User-friendly Bayesian Inference Library for Gravitational-wave Astronomy. ApJS, 241(2):27, Apr. 2019. doi: 10.3847/1538-4365/ab06fc.

W. Baade and F. Zwicky. On super-novae. *Proc. Nat. Acad. Sci.*, 20:254–259, 1934.

D. C. Backer, S. R. Kulkarni, C. Heiles, M. Davis, and W. Goss. A millisecond pulsar. *Nature*, 300(5893):615–618, 1982.

M. C. Begelman, R. D. Blandford, and M. J. Rees. Massive black hole binaries in active galactic nuclei. *Nature*, 287(5780):307–309, 1980.

S. Bhattacharyya, A. Papitto, and D. Bhattacharya. *Millisecond Pulsars*. Springer, 2022.

G. Bisnovatyi-Kogan and B. Komberg. Possible evolution of a binary-system radio pulsar as an old object with a weak magnetic field. *Soviet Astronomy Letters, vol. 2, July-Aug. 1976, p. 130-132. Translation. Pisma v Astronomicheskii Zhurnal, vol. 2, July 1976, p. 338-342.*, 2:130–132, 1976.

L. Blanchet. Gravitational radiation from post-newtonian sources and inspiralling compact binaries. *Living reviews in relativity*, 17(1):2, 2014.

H. Bondi. Plane gravitational waves in general relativity. *Nature*, 179(4569):1072–1073, 1957.

Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel Density Estimation via Diffusion. *Annals of Statistics*, 38(5):2916–2957, 2010a. doi: 10.1214/10-AOS799.

Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957, 2010b. doi: 10.1214/10-AOS799. URL `https://doi.org/10.1214/10-AOS799`.

C. J. T. Braak. A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249, 2006.

M. e. a. Burgay. Discovery of a relativistic double-neutron-star system. In F. Rasio and I. H. Stairs, editors, *Binary Radio Pulsars*, pages 53–58, San Francisco, 2005. Astronomical Society of the Pacific.

C. Caprini, M. Hindmarsh, S. Huber, T. Konstandin, J. Kozaczuk, G. Nardini, J. M. No, A. Petiteau, P. Schwaller, G. Servant, et al. Science with the space-based interferometer elisa. ii: Gravitational waves from cosmological phase transitions. *Journal of cosmology and astroparticle physics*, 2016(04):001–001, 2016.

B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal Of The Royal Statistical Society Series B: Statistical Methodology*, 57(3): 473–484, 1995.

A. Chalumeau, S. Babak, A. Petiteau, S. Chen, A. Samajdar, R. N. Caballero, G. Theureau, L. Guillemot, G. Desvignes, A. Parthasarathy, and et al. Noise analysis in the European Pulsar Timing Array data release 2 and its implications on the gravitational-wave background search. MNRAS, 509(4):5538–5558, Feb. 2022. doi: 10.1093/mnras/stab3283.

D. Champion, G. Hobbs, R. Manchester, R. Edwards, D. Backer, M. Bailes, N. Bhat, S. Burke-Spolaor, W. Coles, P. Demorest, et al. Measuring the mass of solar system planets using pulsar timing. *The Astrophysical Journal Letters*, 720(2):L201, 2010.

J. C. C. Chan and E. Eisenstat. Marginal likelihood estimation with the cross-entropy method. *Econometric Reviews*, 34(3):256–285, 2015.

S. Chen, R. Caballero, Y. Guo, A. Chalumeau, K. Liu, G. Shaifullah, K. Lee, S. Babak, G. Desvignes, A. Parthasarathy, et al. Common-red-signal analysis with 24-yr high-precision timing of the european pulsar timing array: inferences in the stochastic gravitational-wave background search. *Monthly Notices of the Royal Astronomical Society*, 508(4):4970–4993, 2021. doi: 10.1093/mnras/stab2833.

S. Chen, H. Xu, Y. Guo, B. Wang, R. N. Caballero, J. Jiang, J. Xu, Z. Xue, K. Lee, J. Yuan, et al. The chinese pulsar timing array data release i-single pulsar noise analysis. *Astronomy & Astrophysics*, 699:A165, 2025.

S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995. doi: 10.1080/00031305.1995.10476177.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3): 539–552, 2002.

N. Chopin and C. P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, 2010.

D. Christodoulou. Nonlinear nature of gravitation and gravitational-wave experiments. *Physical review letters*, 67(12):1486, 1991.

E. collaboration, I. Collaboration, et al. The second data release from the european pulsar timing array: V. search for continuous gravitational wave signals. *Astronomy and Astrophysics*, 690:A118, 2024.

L. S. Collaboration and V. Collaboration. GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run - Parameter Estimation Data Release, May 2022. URL `https://doi.org/10.5281/zenodo.6513631`.

L. S. Collaboration and V. Collaboration. GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run. Phys. Rev. D, 109(2):022001, Jan. 2024. doi: 10.1103/PhysRevD.109.022001.

F.-L. Collaboration*†. A gamma-ray pulsar timing array constrains the nanohertz gravitational wave background. *Science*, 376(6592):521–523, 2022.

J. Cordes and R. Shannon. A measurement model for precision pulsar timing. *arXiv preprint arXiv:1010.3785*, 2010.

R. V. Craiu. Tuning of Markov chain Monte Carlo algorithms using copulas. *UPB Scientific Bulletin, Series A: Applied Mathematics and Physics*, 73(1):5–12, 2011. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-80051894706&partnerID=40&md5=57101790b1cbb7dbbd8d0dae7572e913. Cited by: 4.

M. Curyło, T. T. Pennucci, M. Bailes, N. D. R. Bhat, A. D. Cameron, W. A. Coles, P. B. Demorest, G. Hobbs, A. Kapur, M. J. Keith, et al. Pulse portrait modelling and the wide-band timing of 26 PPTA millisecond pulsars. *Monthly Notices of the Royal Astronomical Society*, 528:3304–3327, 2024. doi: 10.1093/mnras/stae036.

L. Dalcín, R. Paz, and M. Storti. Mpi for python. *Journal of Parallel and Distributed Computing*, 65(9):1108–1115, 2005.

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

P. Dellaportas and M. G. Tsionas. Importance sampling from posterior distributions using copula-like approximations. *Journal of Econometrics*, 210(1):45–57, 2019.

G. Desvignes, R. Caballero, L. Lentati, J. Verbiest, D. Champion, B. Stappers, G. Janssen, P. Lazarus, S. Oslowski, S. Babak, et al. Vizier online data catalog: 42 millisecond pulsars high-precision timing (desvignes+, 2016). 2017.

S. Detweiler. Pulsar timing measurements and the search for gravitational waves. *Astrophysical Journal, Part 1, vol. 234, Dec. 15, 1979, p. 1100-1104.*, 234:1100–1104, 1979.

J. M. Dickey. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, pages 204–223, 1971.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

G. M. Eadie, J. S. Speagle, J. Cisewski-Kehe, D. Foreman-Mackey, D. Huppenkothen, D. E. Jones, A. Springford, and H. Tak. Practical guidance for bayesian inference in astronomy. *arXiv preprint arXiv:2302.04703*, 2023.

A. S. Eddington. The propagation of gravitational waves. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 102(716):268–282, 1922.

J. Edmonds. Paths, Trees, and Flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

A. Einstein. Näherungsweise integration der feldgleichungen der gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pages 688–696, 1916.

A. Einstein. Über gravitationswellen. *Sitzungsber. Preuss. Akad. Wiss.*, 154:1–14, 1918.

J. Ellis and R. van Haasteren. jellis18/ptmcmcsampler: Official release, Oct. 2017. URL `https://doi.org/10.5281/zenodo.1037579`.

J. A. Ellis. A Bayesian analysis pipeline for continuous GW sources in the PTA band. *Classical and Quantum Gravity*, 30(22):224004, Nov. 2013. doi: 10.1088/0264-9381/30/22/224004.

J. A. Ellis, M. Vallisneri, S. R. Taylor, and P. T. Baker. Enterprise: Enhanced numerical toolbox enabling a robust pulsar inference suite. *Astrophysics Source Code Library*, pages ascl–1912, 2019.

Y. Fan, R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. Choosing among partition models in bayesian phylogenetics. *Molecular biology and evolution*, 28(1):523–532, 2011.

M. Favata. The gravitational-wave memory effect. *Classical and Quantum Gravity*, 27 (8):084036, 2010.

R. D. Ferdman, R. van Haasteren, C. G. Bassa, M. Burgay, I. Cognard, A. Corongiu, N. D'Amico, G. Desvignes, J. W. Hessels, G. H. Janssen, et al. The european pulsar timing array: current efforts and a leap toward the future. *Classical and Quantum Gravity*, 27(8):084014, 2010. doi: 10.1088/0264-9381/27/8/084014.

F. Feroz, M. P. Hobson, and M. Bridges. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, 2009.

F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt. Importance nested sampling and the multinest algorithm. *arXiv preprint arXiv:1306.2144*, 2013.

D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The mcmc hammer. *PASP*, 125:306–312, 2013. doi: 10.1086/670067.

J. Forrest and R. Lougee-Heimer. CBC User Guide. In *Emerging Theory, Methods, and Applications*, pages 257–277. INFORMS, 2005. doi: 10.1287/educ.1053.0020.

R. S. Foster and D. C. Backer. Constructing a Pulsar Timing Array. ApJ, 361:300, Sept. 1990. doi: 10.1086/169195.

R. S. Foster and D. C. Backer. Constructing a pulsar timing array. *ApJ*, 361:300, 1990.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(3):589–607, 2008. doi: 10.1111/j.1467-9868.2007.00650.x.

A. S. Fruchter. *Pulsars Lost and Found: The Second Princeton–Arecibo Millisecond Pulsar Search*. PhD thesis, Princeton University, 1989.

S. Frühwirth-Schnatter. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7 (1):143–167, 2004.

J. Gair, J. D. Romano, S. Taylor, and C. M. Mingarelli. Mapping gravitational-wave backgrounds using methods from cmb analysis: Application to pulsar timing arrays. *Physical Review D*, 90(8):082001, 2014.

Z. Galil. Efficient Algorithms for Finding Maximum Matching in Graphs. *ACM Computing Surveys*, 18(1):23–38, 1986.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

A. Gelman and C. R. Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

C. J. Geyer. Markov chain monte carlo maximum likelihood. 1991.

M. Girolami and B. Calderhead. Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. doi: 10.1111/j.1467-9868.2010.00765.x.

T. Gold. Rotating neutron stars as the origin of the pulsating radio sources. *Nature*, 218 (5143):731–732, 1968.

T. Gold. Rotating neutron stars and the nature of pulsars. *Nature*, 221:25–27, 1969.

B. Goncharov, D. Reardon, R. Shannon, X.-J. Zhu, E. Thrane, M. Bailes, N. Bhat, S. Dai, G. Hobbs, M. Kerr, et al. Identifying and mitigating noise sources in precision pulsar timing data sets. *Monthly Notices of the Royal Astronomical Society*, 502(1): 478–493, 2021.

B. Goncharov, E. Thrane, R. M. Shannon, J. Harms, N. R. Bhat, G. Hobbs, M. Kerr, R. N. Manchester, D. J. Reardon, C. J. Russell, et al. Consistency of the parkes pulsar timing array signal with a nanohertz gravitational-wave background. *The Astrophysical Journal Letters*, 932(2):L22, 2022. doi: 10.3847/2041-8213/ac76bb.

Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, et al. A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97, 2017.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

H. Haario, E. Saksman, and J. Tamminen. Componentwise adaptation for high dimensional mcmc. *Computational Statistics*, 20(2):265–273, 2005.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. Polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398, 2015.

T. H. Hankins and B. J. Rickett. Pulsar signal processing. In *Methods in Computational Physics Volume 14 — Radio Astronomy*, pages 55–129, New York, 1975. Academic Press.

U. H. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

J. S. Hazboun, J. D. Romano, and T. L. Smith. Realistic sensitivity curves for pulsar timing arrays. *Physical Review D*, 100(10):104028, 2019. doi: 10.1103/PhysRevD.100.104028.

J. S. Hazboun, J. Simon, D. R. Madison, Z. Arzoumanian, H. T. Cromartie, K. Crowter, M. E. DeCesar, P. B. Demorest, T. Dolch, J. A. Ellis, R. D. Ferdman, E. C. Ferrara, E. Fonseca, P. A. Gentile, G. Jones, M. L. Jones, M. T. Lam, L. Levin, D. R. Lorimer, R. S. Lynch, M. A. McLaughlin, C. Ng, D. J. Nice, T. T. Pennucci, S. M. Ransom, P. S. Ray, R. Spiewak, I. H. Stairs, K. Stovall, J. K. Swiggum, W. Zhu, and T. N. Collaboration. Bayesian Solar Wind Modeling with Pulsar Timing Arrays. *The Astrophysical Journal*, 929(1):39, Apr. 2022. doi: 10.3847/1538-4357/ac5829. URL `https://doi.org/10.3847/1538-4357/ac5829`.

R. Hellings and G. Downs. Upper limits on the isotropic gravitational radiation background from pulsar timing analysis. *Astrophysical Journal, Part 2-Letters to the Editor, vol. 265, Feb. 15, 1983, p. L39-L42.*, 265:L39–L42, 1983.

A. Hewish, S. J. Bell, J. D. Pilkington, P. F. Scott, and R. A. Collins. Observation of a rapidly pulsating radio source. *Nature*, 217(5130):709–713, 1968.

M. Hinne. An introduction to sequential monte carlo for bayesian inference and model comparison—with examples for psychology and behavioral science. *Behavior Research Methods*, 57(5):125, 2025.

M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers. A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215, 2020. doi: 10.1177/2515245919898657.

G. Hobbs. The parkes pulsar timing array. *Classical and Quantum Gravity*, 30(22): 224007, 2013.

G. Hobbs, R. Manchester, A. Teoh, and M. Hobbs. The atnf pulsar catalog. In F. Camilo and B. M. Gaensler, editors, *Young Neutron Stars and Their Environments, IAU Symposium 218*, pages 139–140, San Francisco, 2004. Astronomical Society of the Pacific.

G. Hobbs, R. Edwards, and R. Manchester. TEMPO2: a New Pulsar Timing Package. *Chin. J. Atron. Astrophys., Suppl. 2*, 6:189–192, Dec. 2006a.

G. Hobbs, F. Jenet, K. J. Lee, J. P. W. Verbiest, D. Yardley, R. Manchester, A. Lommen, W. Coles, R. Edwards, and C. Shettigara. TEMPO2: a new pulsar timing package - III. Gravitational wave simulation. *MNRAS*, 394:1945–1955, Apr. 2009. doi: 10.1111/j.1365-2966.2009.14391.x.

G. B. Hobbs, R. T. Edwards, and R. N. Manchester. Tempo2, a new pulsar-timing package - i. an overview. *MNRAS*, 369:655–672, June 2006b. doi: 10.1111/j.1365-2966.2006.10302.x.

M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

A. W. Hotan, W. van Straten, and R. N. Manchester. PSRCHIVE and PSRFITS: An open approach to radio pulsar data storage and analysis. *Publications of the Astronomical Society of Australia*, 21:302–309, 2004. doi: 10.1071/AS04022.

S. Hourihane, P. Meyers, A. Johnson, K. Chatziioannou, and M. Vallisneri. Accurate characterization of the stochastic gravitational-wave background with pulsar timing arrays by likelihood reweighting. *Physical Review D*, 107(8):084045, 2023. doi: 10.1103/PhysRevD.107.084045.

R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. *ApJ*, 195: L51–L53, 1975.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(03):90–95, 2007. doi: 10.1109/MCSE.2007.55.

S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, et al. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. Phys. Rev. D, 93(4):044006, Feb. 2016. doi: 10.1103/PhysRevD.93.044006.

F. Iraci, A. Chalumeau, C. Tiburzi, J. Verbiest, A. Possenti, S. Susarla, M. Krishnakumar, G. Shaifullah, J. Antoniadis, M. Bagchi, et al. Combining the second data release of the european pulsar timing array with low-frequency pulsar data. *Astronomy & Astrophysics*, 704:A109, 2025.

A. H. Jaffe and D. C. Backer. Gravitational waves probe the coalescence rate of massive black hole binaries. *ApJ*, 583:616–631, 2003.

H. Jeffreys. *The theory of probability*. OuP Oxford, 1998.

A. D. Johnson, P. M. Meyers, P. T. Baker, N. J. Cornish, J. S. Hazboun, T. B. Littenberg, J. D. Romano, S. R. Taylor, M. Vallisneri, S. J. Vigeland, K. D. Olum, X. Siemens, J. A. Ellis, R. van Haasteren, S. Hourihane, G. Agazie, A. Anumarlapudi, A. M. Archibald, Z. Arzoumanian, L. Blecha, A. Brazier, P. R. Brook, S. Burke-Spolaor, B. Bécsy, J. A. Casey-Clyde, M. Charisi, S. Chatterjee, K. Chatziioannou, T. Cohen, J. M. Cordes, F. Crawford, H. T. Cromartie, K. Crowter, M. E. Decesar, P. B. Demorest, T. Dolch, B. Drachler, E. C. Ferrara, W. Fiore, E. Fonseca, G. E. Freedman, N. Garver-Daniels, P. A. Gentile, J. Glaser, D. C. Good, K. Gültekin, R. J. Jennings, M. L. Jones, A. R. Kaiser, D. L. Kaplan, L. Z. Kelley, M. Kerr, J. S. Key, N. Laal, M. T. Lam, W. G. Lamb, T. J. W. Lazio, N. Lewandowska, T. Liu, D. R. Lorimer, J. Luo, R. S. Lynch, C.-P. Ma, D. R. Madison, A. McEwen, J. W. McKee, M. A. McLaughlin, N. McMann, B. W. Meyers, C. M. F. Mingarelli, A. Mitridate, C. Ng, D. J. Nice, S. K. Ocker, T. T. Pennucci, B. B. P. Perera, N. S. Pol, H. A. Radovan, S. M. Ransom, P. S. Ray, S. C. Sardesai, C. Schmiedekamp, A. Schmiedekamp, K. Schmitz, B. J. Shapiro-Albert, J. Simon, M. S. Siwek, I. H. Stairs, D. R. Stinebring, K. Stovall, A. Susobhanan, J. K. Swiggum, J. E. Turner, C. Unal, H. M. Wahl, C. A. Witt, O. Young, and Nanograv Collaboration. NANOGrav 15-year gravitational-wave background methods. Phys. Rev. D, 109(10):103012, May 2024. doi: 10.1103/PhysRevD.109.103012.

A. D. Johnson, P. M. Meyers, P. T. Baker, N. J. Cornish, J. S. Hazboun, T. B. Littenberg, J. D. Romano, S. R. Taylor, M. Vallisneri, S. J. Vigeland, et al. Nanograv 15-year gravitational-wave background methods. *Physical Review D*, 109(10):103012, 2024.

M. L. Jones, M. A. McLaughlin, M. T. Lam, J. M. Cordes, L. Levin, S. Chatterjee, Z. Arzoumanian, K. Crowter, P. B. Demorest, T. Dolch, et al. The nanograv nine-year data set: measurement and analysis of variations in dispersion measures. *The Astrophysical Journal*, 841(2):125, 2017. doi: 10.3847/1538-4357/aa73df.

B. C. Joshi, P. Arumugasamy, M. Bagchi, D. Bandyopadhyay, A. Basu, N. Dhanda Batra, S. Bethapudi, A. Choudhary, K. De, L. Dey, et al. Precision pulsar timing with the ort and the gmrt and its applications in pulsar astrophysics. *Journal of Astrophysics and Astronomy*, 39(4):51, 2018. doi: 10.1007/s12036-018-9549-y.

M. Karamanis, F. Beutler, J. A. Peacock, D. Nabergoj, and U. Seljak. Accelerating astronomical and cosmological inference with preconditioned monte carlo. *Monthly Notices of the Royal Astronomical Society*, 516(2):1644–1653, 2022a.

M. Karamanis, D. Nabergoj, F. Beutler, J. A. Peacock, and U. Seljak. pocomc: A python package for accelerated bayesian inference in astronomy and cosmology. *arXiv preprint arXiv:2207.05660*, 2022b.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

G. Kauermann and R. Meyer. Penalized marginal likelihood estimation of finite mixtures of Archimedean copulas. *Computational Statistics*, 29(1):283–306, 2014.

M. Keith, W. Coles, R. Shannon, G. Hobbs, R. Manchester, M. Bailes, N. Bhat, S. Burke-Spolaor, D. Champion, A. Chaudhary, et al. Measurement and correction of variations in interstellar dispersion in high-precision pulsar timing. *Monthly Notices of the Royal Astronomical Society*, 429(3):2161–2174, 2013.

M. Kerr, D. J. Reardon, G. Hobbs, R. M. Shannon, R. N. Manchester, S. Dai, C. J. Russell, L. Toomey, B. Goncharov, H. Middleton, et al. The parkes pulsar timing array project: second data release. *Publications of the Astronomical Society of Australia*, 37:e020, 2020. doi: 10.1017/pasa.2020.11.

S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, et al. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. Phys. Rev. D, 93(4):044007, Feb. 2016. doi: 10.1103/PhysRevD.93.044007.

A. Khmelnitsky and V. Rubakov. Pulsar timing signal from ultralight scalar dark matter. *Journal of Cosmology and Astroparticle Physics*, 2014(02):019, 2014.

T. Kloek and H. K. Van Dijk. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks–a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas*, pages 87–90. IOS press, 2016. doi: 10.3233/978-1-61499-649-1-87.

S. Koposov, J. Speagle, K. Barbary, G. Ashton, E. Bennett, et al. joshspeagle/dynesty: v2.1.4. *Zenodo*, 2023.

R. Kumar, C. Carroll, A. Hartikainen, and O. Martin. Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, 4(33):1143, 2019. doi: 10.21105/joss.01143. URL `https://doi.org/10.21105/joss.01143`.

M. Lam, J. Ellis, G. Grillo, M. Jones, J. Hazboun, P. Brook, J. Turner, S. Chatterjee, J. Cordes, T. Lazio, et al. A second chromatic timing event of interstellar origin toward psr j1713+ 0747. *The Astrophysical Journal*, 861(2):132, 2018.

N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207, 2006.

P. Lazarus, R. Karuppusamy, E. Graikou, R. N. Caballero, D. J. Champion, K. J. Lee, J. P. W. Verbiest, and M. Kramer. Prospects for high-precision pulsar timing with the new effelsberg PSRIX backend. *Monthly Notices of the Royal Astronomical Society*, 458:868–880, 2016. doi: 10.1093/mnras/stw189.

L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren. Hyper-efficient model-independent Bayesian method for the analysis of pulsar timing data. Phys. Rev. D, 87(10):104021, May 2013. doi: 10.1103/PhysRevD.87.104021.

L. Lentati, P. Alexander, M. P. Hobson, F. Feroz, R. van Haasteren, K. Lee, and R. M. Shannon. Temponest: a bayesian approach to pulsar timing analysis. *Monthly Notices of the Royal Astronomical Society*, 437(3):3004–3023, 2014a.

L. Lentati, M. P. Hobson, and P. Alexander. Bayesian estimation of non-gaussianity in pulsar timing analysis. *Monthly Notices of the Royal Astronomical Society*, 444(4):3863–3878, 2014b.

L. Lentati, R. M. Shannon, W. A. Coles, J. P. Verbiest, R. van Haasteren, J. Ellis, R. Caballero, R. N. Manchester, Z. Arzoumanian, S. Babak, et al. From spin noise to systematics: stochastic processes in the first international pulsar timing array data release. *Monthly Notices of the Royal Astronomical Society*, 458(2):2161–2187, 2016.

L. Lentati, R. M. Shannon, W. A. Coles, J. P. W. Verbiest, R. van Haasteren, J. A. Ellis, R. N. Caballero, R. N. Manchester, Z. Arzoumanian, S. Babak, and et al. From spin noise to systematics: stochastic processes in the first International Pulsar Timing Array data release. MNRAS, 458(2):2161–2187, May 2016. doi: 10.1093/mnras/stw395.

L. Lentati, M. Kerr, S. Dai, M. Hobson, R. M. Shannon, G. Hobbs, M. Bailes, N. R. Bhat, S. Burke-Spolaor, W. Coles, et al. Wide-band profile domain pulsar timing analysis. *Monthly Notices of the Royal Astronomical Society*, 466(3):3706–3727, 2017.

P. M. Lewis II. Approximating probability distributions to reduce storage requirements. *Information and Control*, 2(3):214–225, 1959.

T. Lodewyckx, W. Kim, M. D. Lee, F. Tuerlinckx, P. Kuppens, and E.-J. Wagenmakers. A tutorial on bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, 55(5):331–347, 2011. doi: 10.1016/j.jmp.2011.06.001.

D. R. Lorimer. Binary and millisecond pulsars. *Living reviews in relativity*, 11(1):8, 2008.

D. R. Lorimer and M. Kramer. *Handbook of Pulsar Astronomy*. Cambridge University Press, 2005.

J. Luo, S. Ransom, P. Demorest, P. S. Ray, A. Archibald, M. Kerr, R. J. Jennings, M. Bachetti, R. van Haasteren, C. A. Champagne, et al. Pint: a modern software package for pulsar timing. *The Astrophysical Journal*, 911(1):45, 2021.

D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

R. Manchester. Millisecond pulsars, their evolution and applications. *Journal of Astrophysics and Astronomy*, 38(3):42, 2017.

R. Manchester, G. Hobbs, M. Bailes, W. Coles, W. Van Straten, M. Keith, R. Shannon, N. Bhat, A. Brown, S. Burke-Spolaor, et al. The parkes pulsar timing array project. *Publications of the Astronomical Society of Australia*, 30:e017, 2013.

V. D. Marco, N. Guttman, M. T. Miles, A. Zic, R. M. Shannon, and E. Thrane. A transdimensional sampling framework for pulsar timing noise modelling, 2026. URL https://arxiv.org/abs/2603.23817.

D. N. Matsakis, J. H. Taylor, and T. M. Eubanks. A statistic for describing pulsar and clock stabilities. *Astronomy and Astrophysics, v. 326, p. 924-928*, 326:924–928, 1997.

P. Maturana-Russel, B. J. Brewer, S. Klaere, and R. R. Bouckaert. Model selection and parameter inference in phylogenetics using nested sampling. *Systematic Biology*, 68 (2):219–233, 2019a.

P. Maturana-Russel, R. Meyer, J. Veitch, and N. Christensen. Stepping-stone sampling algorithm for calculating the evidence of gravitational wave models. *Physical Review D*, 99(8):084006, 2019b.

P. A. Maturana-Russel. *Bayesian inference in phylogenetics using Nested Sampling*. PhD thesis, University of Auckland, 2017. URL https://researchspace.auckland.ac.nz/docs/uoa-docs/rights.htm.

M. A. McLaughlin. The north american nanohertz observatory for gravitational waves. *Classical and Quantum Gravity*, 30(22):224008, Nov. 2013a. ISSN 1361-6382. doi: 10.1088/0264-9381/30/22/224008. URL http://dx.doi.org/10.1088/0264-9381/30/22/224008.

M. A. McLaughlin. The north american nanohertz observatory for gravitational waves. *Classical and Quantum Gravity*, 30(22):224008, 2013b.

X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

M. T. Miles, R. M. Shannon, M. Bailes, D. J. Reardon, M. J. Keith, A. D. Cameron, A. Parthasarathy, M. Shamohammadi, R. Spiewak, W. van Straten, S. Buchner, F. Camilo, M. Geyer, A. Karastergiou, M. Kramer, M. Serylak, G. Theureau, and V. Venkatraman Krishnan. The meerkat pulsar timing array: first data release. *Monthly Notices of the Royal Astronomical Society*, 519(3):3976–3991, 12 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac3644. URL `https://doi.org/10.1093/mnras/stac3644`.

M. T. Miles, R. M. Shannon, M. Bailes, D. J. Reardon, M. J. Keith, et al. The MeerKAT pulsar timing array: First data release. *Monthly Notices of the Royal Astronomical Society*, 519(3):3976–3991, 2023.

S. Mitchell, M. OSullivan, and I. Dunning. PuLP: a linear programming toolkit for Python. *The University of Auckland, Auckland, New Zealand*, 65:25, 2011.

C. J. Moore, R. H. Cole, and C. P. Berry. Gravitational-wave sensitivity curves. *Classical and Quantum Gravity*, 32(1):015014, 2014.

G. Nelemans, L. Yungelson, and S. P. Zwart. The gravitational wave signal from the galactic disk population of binaries containing two compact objects. *Astronomy & Astrophysics*, 375(3):890–898, 2001.

M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1):3–26, 1994. doi: 10.1111/j.2517-6161.1994.tb01956.x.

J. R. Oppenheimer and G. M. Volkoff. On massive neutron cores. *Phys. Rev.*, 55: 374–381, 1939. doi: 10.1103/PhysRev.55.374.

A. M. Overstall and J. J. Forster. Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54(12): 3269–3288, 2010.

F. Pacini. Energy emission from a neutron star. *Nature*, 216(5115):567–568, 1967.

F. Pacini. Rotating neutron stars, pulsars, and supernova remnants. *Nature*, 219: 145–146, 1968.

A. Papitto, C. Ferrigno, E. Bozzo, N. Rea, L. Pavan, L. Burderi, M. Burgay, S. Campana, T. Di Salvo, M. Falanga, et al. Swings between rotation and accretion power in a binary millisecond pulsar. *Nature*, 501(7468):517–520, 2013.

T. T. Pennucci. Frequency-dependent template profiles for high-precision pulsar timing. *The Astrophysical Journal*, 871(1):34, 2019.

E. S. Phinney. A Practical Theorem on Gravitational Wave Backgrounds. *ArXiv Astrophysics e-prints*, Aug. 2001.

F. A. Pirani. On the physical significance of the riemann tensor. *Acta Physica Polonica*, 15:389–405, 1956.

A. Polanska, M. A. Price, D. Piras, A. S. Mancini, and J. D. McEwen. Learned harmonic mean estimation of the Bayesian evidence with normalizing flows. *arXiv preprint arXiv:2405.05969*, 2024.

A. Puecher, A. Samajdar, G. Ashton, C. Van Den Broeck, and T. Dietrich. Comparing gravitational waveform models for binary black hole mergers through a hypermodels approach. Phys. Rev. D, 109(2):023019, Jan. 2024. doi: 10.1103/PhysRevD.109.023019.

M. Punturo, M. Abernathy, F. Acernese, B. Allen, N. Andersson, et al. The Einstein Telescope: a third-generation gravitational wave observatory. *Classical and Quantum Gravity*, 27(19):194002, Oct. 2010. doi: 10.1088/0264-9381/27/19/194002.

V. Radhakrishnan and G. Srinivasan. *Curr. Sci.*, 51:1096, 1982.

M. Rajagopal and R. W. Romani. Ultra–Low-Frequency Gravitational Radiation from Massive Black Hole Binaries. *ApJ*, 446:543–549, June 1995.

D. J. Reardon, R. M. Shannon, A. D. Cameron, B. Goncharov, G. B. Hobbs, H. Middleton, M. Shamohammadi, N. Thyagarajan, M. Bailes, N. D. R. Bhat, and et al. The Parkes pulsar timing array second data release: timing analysis. *Monthly Notices of the Royal Astronomical Society*, 507(2):2137–2153, Oct. 2021. doi: 10.1093/mnras/stab1990.

D. J. Reardon, A. Zic, R. M. Shannon, V. Di Marco, G. B. Hobbs, A. Kapur, M. E. Lower, R. Mandow, H. Middleton, M. T. Miles, et al. The gravitational-wave background null hypothesis: characterizing noise in millisecond pulsar arrival times with the parkes pulsar timing array. *The Astrophysical Journal Letters*, 951(1):L7, 2023a.

D. J. Reardon, A. Zic, R. M. Shannon, G. B. Hobbs, M. Bailes, V. Di Marco, A. Kapur, A. F. Rogers, E. Thrane, J. Askew, et al. Search for an isotropic gravitational-wave background with the parkes pulsar timing array. *The Astrophysical Journal Letters*, 951(1):L6, 2023b. doi: 10.3847/2041-8213/acdd02.

D. Reitze, R. X. Adhikari, S. Ballmer, B. Barish, L. Barsotti, et al. Cosmic Explorer: The U.S. Contribution to Gravitational-Wave Astronomy beyond LIGO. In *Bulletin of the American Astronomical Society*, volume 51, page 35, Sept. 2019. doi: 10.48550/arXiv.1907.04833.

M. S. Roberts. Surrounded by spiders! new black widows and redbacks in the galactic field. *Proceedings of the International Astronomical Union*, 8(S291):127–132, 2012.

A. F. Rogers, W. van Straten, S. Gulyaev, A. Parthasarathy, G. Hobbs, Z.-C. Chen, Y. Feng, B. Goncharov, A. Kapur, X. Liu, et al. Reducing instrumental errors in parkes pulsar timing array data. *The Astrophysical Journal*, 973(2):94, 2024.

J. D. Romano and N. J. Cornish. Detection methods for stochastic gravitational-wave backgrounds: a unified treatment. *Living reviews in relativity*, 20:1–223, 2017.

I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D'Emilio, G. Ashton, et al. Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO-Virgo gravitational-wave transient catalogue. MNRAS, 499(3): 3295–3319, Dec. 2020. doi: 10.1093/mnras/staa2850.

S. Roy, M. Haney, G. Pratten, P. T. H. Pang, and C. Van Den Broeck. An improved parametrized test of general relativity using the IMRPhenomX waveform family: Including higher harmonics and precession. *arXiv e-prints*, art. arXiv:2504.21147, Apr. 2025. doi: 10.48550/arXiv.2504.21147.

M. Ruderman. Crust-breaking by neutron superfluids and the Vela pulsar glitches. *ApJ*, 203:213–222, 1976.

M. Sazhin. Opportunities for detecting ultralong gravitational waves. *Sov. Astron*, 22: 36–38, 1978.

D. Schmidl, C. Czado, S. Hug, and F. J. Theis. A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Bayesian Analysis*, 8(1):1–22, 2013. doi: 10.1214/13-BA801. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-84875110570&doi=10.1214%2f13-BA801&partnerID=40&md5=c7e2bd67a4c23f5e1c8452b02a6a27aa`. Cited by: 17; All Open Access, Gold Open Access, Green Open Access.

P. Schwaller. Gravitational waves from a dark phase transition. *Physical review letters*, 115(18):181101, 2015.

A. Sesana. Systematic investigation of the expected gravitational wave signal from supermassive black hole binaries in the pulsar timing band. *MNRAS*, 433:L1–L5, June 2013. doi: 10.1093/mnrasl/slt034.

A. Sesana, A. Vecchio, and C. N. Colacino. The stochastic gravitational-wave background from massive black hole binary systems: implications for observations with Pulsar Timing Arrays. *MNRAS*, 390:192–209, Oct. 2008. doi: 10.1111/j.1365-2966.2008.13682.x.

A. Sesana, A. Vecchio, and M. Volonteri. Gravitational waves from resolvable massive black hole binary systems and observations with Pulsar Timing Arrays. *MNRAS*, 394: 2255–2265, Apr. 2009. doi: 10.1111/j.1365-2966.2009.14499.x.

R. M. Shannon and J. M. Cordes. Assessing the Role of Spin Noise in the Precision Timing of Millisecond Pulsars. *ApJ*, 725:1607–1619, Dec. 2010. doi: 10.1088/0004-637X/725/2/1607.

R. M. Shannon and J. M. Cordes. Modelling and mitigating refractive propagation effects in precision pulsar timing observations. *MNRAS*, 464:2075–2089, Jan. 2017. doi: 10.1093/mnras/stw2449.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

R. d. S. Silva and H. F. Lopes. Copula, marginal distributions and model selection: a Bayesian note. *Statistics and Computing*, 18(3):313–320, 2008.

B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

J. Skilling. Nested sampling for general bayesian computation. 2006.

C. Smarra, B. Goncharov, E. Barausse, J. Antoniadis, S. Babak, A.-S. B. Nielsen, C. G. Bassa, A. Berthereau, M. Bonetti, E. Bortolas, P. R. Brook, M. Burgay, R. N. Caballero, A. Chalumeau, D. J. Champion, S. Chanlaridis, S. Chen, I. Cognard, G. Desvignes, M. Falxa, R. D. Ferdman, A. Franchini, J. R. Gair, E. Graikou, J.-M. Grießmeier, L. Guillemot, Y. J. Guo, H. Hu, F. Iraci, D. Izquierdo-Villalba, J. Jang, J. Jawor, G. H. Janssen, A. Jessner, R. Karuppusamy, E. F. Keane, M. J. Keith, M. Kramer, M. A. Krishnakumar, K. Lackeos, K. J. Lee, K. Liu, Y. Liu, A. G. Lyne, J. W. McKee, R. A. Main, M. B. Mickaliger, I. C. Niţu, A. Parthasarathy, B. B. P. Perera, D. Perrodin, A. Petiteau, N. K. Porayko, A. Possenti, H. Q. Leclere, A. Samajdar, S. A. Sanidas, A. Sesana, G. Shaifullah, L. Speri, R. Spiewak, B. W. Stappers, S. C. Susarla, G. Theureau, C. Tiburzi, E. van der Wateren, A. Vecchio, V. V. Krishnan, J. Wang, L. Wang, and Z. Wu. Second data release from the european pulsar timing array: Challenging the ultralight dark matter paradigm. *Physical Review Letters*, 131:171001, Oct 2023. doi: 10.1103/PhysRevLett.131.171001. URL `https://link.aps.org/doi/10.1103/PhysRevLett.131.171001`.

R. Smits, M. Kramer, B. Stappers, D. R. Lorimer, J. Cordes, et al. Pulsar searches and timing with the Square Kilometre Array. *Astronomy & Astrophysics*, 493(3): 1161–1170, 2009.

D. Sob'yanin. Ultraslow psr j0901-4046 with an ultrahigh magnetic field of 3.2× 10 16 g. *Physical Review D*, 107(8):L081301, 2023.

L. F. South, A. N. Pettitt, and C. C. Drovandi. Sequential Monte Carlo samplers with independent Markov chain Monte Carlo proposals. *Bayesian Analysis*, 14(3):753–776, 2019. doi: 10.1214/18-BA1129. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070022076&doi=10.1214%2f18-BA1129&partnerID=40&md5=34dccdc5a44872ba1c26dab97d5fdb7d. Cited by: 27; All Open Access, Gold Open Access.

J. S. Speagle. dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3): 3132–3158, 2020.

B. W. Stappers, E. F. Keane, M. Kramer, A. Possenti, and I. H. Stairs. The prospects of pulsar timing with new-generation radio telescopes and the Square Kilometre Array. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2120):20170293, 2018.

D. Stinebring, M. Ryba, J. H. Taylor, and d. R. Romani. Cosmic gravitational-wave background: limits from millisecond pulsar timing. *Physical Review Letters*, 65(3): 285, 1990. doi: 10.1103/PhysRevLett.65.285.

M. Studený and J. Vejnarová. The Multiinformation Function as a Tool for Measuring Stochastic Dependence. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_10. URL https://doi.org/10.1007/978-94-011-5014-9_10.

P. A. Sturrock. A model of pulsars. *ApJ*, 164:529–556, 1971.

S. Susarla, A. Chalumeau, C. Tiburzi, E. Keane, J. Verbiest, J. Hazboun, M. Krishnakumar, F. Iraci, G. Shaifullah, A. Golden, et al. Exploring the time variability of the solar wind using lofar pulsar data. *Astronomy & Astrophysics*, 692:A18, 2024.

P. Tarafdar, K. Nobleson, P. Rana, J. Singha, M. Krishnakumar, B. C. Joshi, A. K. Paladi, N. Kolhe, N. D. Batra, N. Agarwal, et al. The indian pulsar timing array: First data release. *Publications of the Astronomical Society of Australia*, 39:e053, 2022.

J. Taylor and R. Manchester. Galactic distribution and evolution of pulsars. *Astrophysical Journal, Part 1, vol. 215, Aug. 1, 1977, p. 885-896.*, 215:885–896, 1977.

J. H. Taylor. Pulsar timing and relativistic gravity. *Philos. Trans. Roy. Soc. London A*, 341:117–134, 1992a.

J. H. Taylor. Pulsar timing and relativistic gravity. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 341(1660): 117–134, 1992b.

J. H. Taylor and J. M. Weisberg. A new test of general relativity: Gravitational radiation and the binary pulsar PSR 1913+16. *ApJ*, 253:908–920, 1982.

S. R. Taylor. The nanohertz gravitational wave astronomer. *arXiv preprint arXiv:2105.13270*, 2021.

S. R. Taylor, P. T. Baker, J. S. Hazboun, J. Simon, and S. J. Vigeland. enterprise_extensions, 2021. URL `https://github.com/nanograv/enterprise_extensions`. v2.4.3.

K. S. Thorne. Gravitational radiation. In S. Hawking and W. Israel, editors, *300 Years of Gravitation*, pages 330–458, Cambridge, 1987. Cambridge University Press.

E. Thrane and C. Talbot. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. PASA, 36:e010, Mar. 2019. doi: 10.1017/pasa.2019.2.

C. Tiburzi, G. Hobbs, M. Kerr, W. Coles, S. Dai, R. Manchester, A. Possenti, R. Shannon, and X. You. A study of spatial correlations in pulsar timing array data. *Monthly Notices of The Royal Astronomical Society*, 455(4):4339–4350, 2016.

C. Tiburzi, J. Verbiest, G. Shaifullah, G. Janssen, J. Anderson, A. Horneffer, J. Künsemöller, S. Osłowski, J. Donner, M. Kramer, et al. On the usefulness of existing solar wind models for pulsar timing corrections. *Monthly Notices of the Royal Astronomical Society*, 487(1):394–408, 2019.

A. B. Tsybakov. *Nonparametric estimators*, pages 1–76. Springer New York, New York, NY, 2009. ISBN 978-0-387-79052-7. doi: 10.1007/978-0-387-79052-7_1. URL `https://doi.org/10.1007/978-0-387-79052-7_1`.

A. Vajpeyi, R. Smith, E. Thrane, G. Ashton, T. Alford, et al. A follow-up on intermediate-mass black hole candidates in the second LIGO-Virgo observing run with the Bayes Coherence Ratio. MNRAS, 516(4):5309–5317, Nov. 2022. doi: 10.1093/mnras/stac2332.

M. Vallisneri, S. R. Taylor, J. Simon, W. M. Folkner, R. S. Park, C. Cutler, J. A. Ellis, T. J. W. Lazio, S. J. Vigeland, K. Aggarwal, et al. Modeling the uncertainties of solar system ephemerides for robust gravitational-wave searches with pulsar-timing arrays. *The Astrophysical Journal*, 893(2):112, 2020.

R. Van Haasteren and Y. Levin. Gravitational-wave memory and pulsar timing arrays. *Monthly Notices of the Royal Astronomical Society*, 401(4):2372–2378, 2010.

R. van Haasteren and M. Vallisneri. New advances in the gaussian-process approach to pulsar-timing data analysis. *Physical Review D*, 90(10):104012, 2014.

R. Van Haasteren, Y. Levin, P. McDonald, and T. Lu. On measuring the gravitational-wave background using pulsar timing arrays. *Monthly Notices of the Royal Astronomical Society*, 395(2):1005–1014, 2009.

R. van Haasteren. Use model averaging instead of model selection in pulsar timing. *Monthly Notices of the Royal Astronomical Society: Letters*, 537(1):L1–L6, 02 2025. ISSN 1745-3925. doi: 10.1093/mnrasl/slae108. URL https://doi.org/10.1093/mnrasl/slae108.

J. Verbiest, M. Bailes, W. Coles, G. Hobbs, W. Van Straten, D. Champion, F. Jenet, R. Manchester, N. Bhat, J. Sarkissian, et al. Timing stability of millisecond pulsars and prospects for gravitational-wave detection. *Monthly Notices of the Royal Astronomical Society*, 400(2):951–968, 2009. doi: 10.1111/j.1365-2966.2009.15508.x.

J. Verbiest, L. Lentati, G. Hobbs, R. van Haasteren, P. B. Demorest, G. Janssen, J.-B. Wang, G. Desvignes, R. Caballero, M. Keith, et al. The international pulsar timing array: first data release. *Monthly Notices of the Royal Astronomical Society*, 458(2): 1267–1288, 2016.

J. P. Verbiest and G. Shaifullah. Measurement uncertainty in pulsar timing array experiments. *Classical and Quantum Gravity*, 35(13):133001, 2018.

P. Virtanen, R. Gommers, E. Burovski, T. E. Oliphant, W. Weckesser, D. Cournapeau, alexbrc, P. Peterson, T. Reddy, M. Haberland, J. Wilson, A. Nelson, endolith, N. Mayorov, S. van der Walt, D. Laxalde, I. Polat, M. Brett, E. Larson, J. Millman, Lars, P. van Mulbregt, eric jones, C. Carey, E. Moore, R. Kern, T. Leslie, J. Perktold, K. Striega, and Y. Feng. scipy/scipy: Scipy 1.6.0, Dec. 2020a.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3): 261–272, 2020b.

M. Volonteri and M. C. Begelman. Quasi-stars and the cosmic evolution of massive black holes. *Monthly Notices of the Royal Astronomical Society*, 409(3):1022–1032, 2010.

S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

R. Wijnands and M. van der Klis. A millisecond pulsar in an X-ray binary system. 394: 344–346, July 1998. doi: 10.1038/28557.

E. Witten. *Phys. Rev. D*, 30:272–285, 1984.

J. S. B. Wyithe and A. Loeb. Low-Frequency Gravitational Waves from Massive Black Hole Binaries: Predictions for LISA and Pulsar Timing Arrays. *ApJ*, 590:691–706, June 2003.

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2): 150–160, 2011.

H. Xu, S. Chen, Y. Guo, J. Jiang, B. Wang, J. Xu, Z. Xue, R. N. Caballero, J. Yuan, Y. Xu, et al. Searching for the nano-hertz stochastic gravitational wave background with the chinese pulsar timing array data release i. *Research in Astronomy and Astrophysics*, 23(7):075024, 2023. doi: 10.1088/1674-4527/acdfa5.

X.-P. You, G. Hobbs, W. A. Coles, R. N. Manchester, and J. Han. The Effect of Interstellar and Interplanetary Medium on Precision Pulsar Timing. In Y.-F. Yuan, X.-D. Li, and D. Lai, editors, *AIP Conf. Series*, volume 968 of *AIP Conf. Series*, pages 181–184, Jan. 2008. doi: 10.1063/1.2840391.

E. M. Zahraoui, P. Maturana-Russel, W. van Straten, R. Meyer, and S. Gulyaev. Generalized stepping-stone sampling: efficient marginal likelihood estimation in gravitational wave analysis of pulsar timing array data. *Monthly Notices of the Royal Astronomical Society*, 540(4):3818–3825, 2025.

E. M. Zahraoui, P. Maturana-Russel, A. Vajpeyi, W. van Straten, R. Meyer, and S. Gulyaev. Enhancing evidence estimation through informed probability density approximation. *Phys. Rev. D*, 113:083014, Apr 2026. doi: 10.1103/1554-y6ns. URL `https://link.aps.org/doi/10.1103/1554-y6ns`.

A. Zic, D. J. Reardon, A. Kapur, G. Hobbs, R. Mandow, M. Curyło, R. M. Shannon, J. Askew, M. Bailes, N. D. R. Bhat, A. Cameron, Z.-C. Chen, S. Dai, V. Di Marco, Y. Feng, M. Kerr, A. Kulkarni, M. Lower, R. Luo, R. N. Manchester, M. Miles, R. Nathan, S. Osłowski, A. Rogers, C. Russell, M. Shamohammadi, R. Spiewak, N. Thyagarajan, L. Toomey, S. Wang, L. Zhang, S. Zhang, and X. Zhu. Parkes pulsar timing array third data release (part 2 of 2), 2023. URL `https://doi.org/10.25919/axvw-qa43`. Dataset.

A. Zic, D. J. Reardon, A. Kapur, G. Hobbs, R. Mandow, M. Curyło, R. M. Shannon, J. Askew, M. Bailes, N. D. R. Bhat, and et al. The Parkes Pulsar Timing Array third data release. *Publications of the Astronomical Society of Australia*, 40:e049, Dec. 2023. doi: 10.1017/pasa.2023.36.

A. Zic, D. J. Reardon, A. Kapur, G. Hobbs, R. Mandow, et al. The Parkes pulsar timing array third data release. *Publications of the Astronomical Society of Australia*, 40: e049, 2023.

Glossary

BF Bayes Factor 48

BMA Bayesian Model Averaging 159

CBC Compact Binary Coalescence 97, 99, 116, 117, 120, 124

CMB Cosmic Microwave Background 39

CPTA Chinese Pulsar Timing Array 42

CURN Common Uncorrelated Red Noise 89

DMGP Dispersion Measure Gaussian Process 86

EPTA European Pulsar Timing Array 41

ESS Effective Sample Size 85

GPTA Gamma-ray Pulsar Timing Array 42

GSS Generalized Steppingstone Sampling 2, 9, 22, 108, 111, 114–116, 124, 160

GWB Gravitational-Wave Background 41, 89

HD Hellings and Downs 20, 89

IBF Inclusion Bayes Factor 87, 93

InPTA Indian Pulsar Timing Array 42

IPTA International Pulsar Timing Array 42

IRN Intrinsic Red Noise 89

ISM Interstellar Medium 24

KDE Kernel Density Estimation 101

LISA Laser Interferometer Space Antenna 38

LoS Line of Sight 27

LVK LIGO-Virgo-KAGRA 4, 99, 116, 117, 120

MCMC Markov chain Monte Carlo 46

MPTA MeerKAT Pulsar Timing Array 41

MSP Millisecond pulsar 29

MSPs Millisecond pulsars 6, 8, 19, 20, 27–30, 39, 40, 42

MSS Morphed Steppingstone Sampling 108

NANOGrav North American Nanohertz Observatory for Gravitational Waves 41

NS Nested Sampling 47

ORF Overlap Reduction Function 56, 89

PPTA Parkes Pulsar Timing Array 22, 41

PSD Power Spectral Density 37

PTA Pulsar Timing Array 2, 19–22, 158

RE Relative Error 106

RN Red Noise 86

SGM Seeded Greedy Maximization 9, 108, 114, 121, 123

SGWB Stochastic Gravitational-Wave Background 2, 20, 37, 38, 40, 41, 50, 51, 57, 58

SMBHB Supermassive Black-Hole Binary 20, 34, 38, 57, 62

SMBHBs Supermassive Black-Hole Binaries 38, 40

SMC Sequential Monte Carlo 47, 49

SS Steppingstone Sampling 77

SW Solar Wind 25

TI Thermodynamic Integration 73

TOA Time of Arrival 30, 40, 51, 53

TOAs Times of Arrival 19, 25, 28–30, 32, 33, 41, 51, 55

WN White Noise 86