Nonlocal parameter degeneracy in the intrinsic space of gravitational-wave signals from extreme-mass-ratio inspirals

Alvin J. K. Chua[1,2,3,*] and Curt J. Cutler[4,†]

[1]*Department of Physics, National University of Singapore, Singapore 117551*
[2]*Department of Mathematics, National University of Singapore, Singapore 119076*
[3]*Theoretical Astrophysics Group, California Institute of Technology, Pasadena, California 91125, USA*
[4]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*

Extreme-mass-ratio inspirals will be prized sources for the upcoming space-based gravitational-wave observatory Laser Interferometer Space Antenna (LISA). The hunt for these is beset by many open theoretical and computational problems in both source modeling and data analysis. We draw attention here to one of the most poorly understood: the phenomenon of nonlocal correlations in the space of extreme-mass-ratio-inspiral signals. Such correlations are ubiquitous in the continuum of possible signals (degeneracy) and severely hinder the search for actual signals in LISA data. However, they are unlikely to manifest in a realistic set of putative signals (confusion). We develop an inventory of new analysis tools in order to conduct an extensive qualitative study of degeneracy—its nature, causes, and implications. Previously proposed search strategies for extreme-mass-ratio inspirals are reviewed in the light of our results, and additional guidelines are suggested for the scientific analysis of such sources.

I. INTRODUCTION

A. Background

In gravitational-wave (GW) astronomy, source modeling and data analysis come together on the manifold of possible signals that is described by a given waveform model and analysis setting. Our intuition and knowledge of such *signal spaces* guides the development of techniques and strategies in GW scientific analysis: for example, local correlations in signal space might be used to perform approximate inference via the Fisher information, while the global structure of these correlations has utility in the fitting of numerical waveform data for the construction of faster approximate models. Many such modeling and analysis methods have been developed in the forge of contemporary ground-based observing [1]; most will be highly relevant, if not directly transferable, to sources for near-future millihertz observatories such as the planned European Space Agency-NASA mission Laser Interferometer Space Antenna (LISA) [2]. However, the nature of signal space —and thus the optimal approach to adopt in scientific analysis—remains poorly understood for one particular class of millihertz source.

Extreme-mass-ratio inspirals (EMRIs) are the late-stage orbits of astrophysical binaries with a small mass ratio $\epsilon \lesssim 10^{-4}$. They arise as the capture of stellar-mass compact objects (white dwarfs, neutron stars, or black holes) by massive black holes in galactic nuclei [3] and will be an important class of source for the LISA detector [4,5]. The GW signals from EMRIs that involve central black holes of $\sim 10^5 - 10^7 M_\odot$ can persist throughout the planned four-year lifetime of the LISA mission, typically with $\sim 10^5$ observable cycles. At the same time, most EMRIs will likely exhibit extreme periapsis precession and Lense-Thirring precession since the motion occurs deep in the strong field of the rotating central mass; those that are formed through traditional capture channels may also enter the LISA sensitivity band with high eccentricity [6]. These effects endow EMRI signals with rich harmonic content over their many cycles.

With their combination of signal longevity and strong-field complexity, EMRIs (and their intermediate-mass-ratio cousins with $10^{-4} \lesssim \epsilon \lesssim 10^{-2}$) have no analog in any other channel of GW astronomy. Our best EMRI waveform models will be constructed through black-hole-perturbation theory and multiscale approaches [7,8] to leverage the extreme mass ratio and slow evolution, but the long duration of expected signals places exacting constraints on the accuracy and efficiency of calculations. Overcoming the theoretical and computational challenges in EMRI forward modeling remains an open and active area of research, with several promising recent developments [9–12]. In the inverse direction, the extraction and characterization of EMRI signals in LISA data has to contend with the information volume of the signal space, which is

---

*alvincjk@nus.edu.sg
†curt.j.cutler@jpl.nasa.gov

20–30 orders of magnitude larger than in the case of comparable-mass binaries [13,14]. Further understanding of the signal space has been limited by the lack of suitable modeling and analysis tools; little is known about its global correlation structure, or the representativeness of the simple proxy models [15–17] used to date.

B. Definition of key concepts

The structure of *correlations* in the EMRI signal space is relevant to LISA scientific analysis in two distinct ways: *confusion* and *degeneracy*. All three of these terms are commonly and loosely used in the GW and LISA literature to refer to a variety of related concepts. Here we shall define them more concretely for the purposes of this work (and, it is hoped, for future use by the community). Given a waveform model $h : \Theta \to \mathcal{D}$ with some fixed sampling rate and duration, the canonical noise-weighted cross-correlation $\langle \cdot | \cdot \rangle$ [18] between two signals $h_{1,2} := h(\boldsymbol{\theta}_{1,2})$ in the signal space $\mathcal{S} := h[\Theta]$ defines an inner product on the ambient data space $\mathcal{D}$ of fixed-length time/frequency series (in which $\mathcal{S}$ can be treated as an embedded submanifold). Of frequent interest is the normalized cross-correlation, or *overlap* $\Omega(\cdot, \cdot)$; this can be interpreted as (the cosine of) the angle between $h_{1,2}$ in $\mathcal{D}$ and serves as a natural measure of signal similarity in that space. We will informally use the term "correlation" to mean the overlap rather than the cross-correlation and further qualify the correlation between $h_{1,2}$ as *local* if it is accompanied by a "small" distance between $h_{1,2}$ in $\mathcal{S}$ (equivalently, between $\boldsymbol{\theta}_{1,2}$ in Θ). See Sec. II B for the explicit definitions of $\langle \cdot | \cdot \rangle$, Ω, and locality.

Confusion refers to the presence of non-negligible correlations among a finite set of putative signals in the LISA data stream. It is a defining feature of the full LISA catalog, which comprises many resolvable and unresolvable signals across multiple source types, and it provides the primary motivation for the LISA *global fit*—a governing strategy to account for the multitude of correlated signals by searching for and characterizing them at the same time. The main contributor to global confusion is the Galactic population of compact binaries, as there are millions of such sources whose signals will form a foreground of astrophysical noise in the LISA band [19]. In the case of EMRIs, extreme estimates for the number of subthreshold signals will also reduce the effective sensitivity of LISA, potentially quite significantly [20,21]. Confusion between specific classes of source (e.g., a single EMRI and the Galactic-binary foreground) has been studied in some depth as well. In [22], analytic arguments from large-deviations theory are used to show that when analyzing resolvable signals of a given source type, a combination of unresolvable signals from a different source type can be well approximated as Gaussian noise.

Our focus in this work is more on *self-confusion*, where the signals under consideration are due to members from a single class of source, i.e., all sources (signals) are well described by points in the domain (image) of a single waveform model. The archetypal example of self-confusion is again provided by the Galactic binaries. Each source is described by a narrow band signal, so any signal-space correlations are effectively local; the problem is then one of resolving individual signals, with leading treatments employing the transdimensional sampling of a multisource likelihood [23]. Self-confusion is less of an immediate problem for EMRIs because of their highly uncertain event rates [4]. It has yet to be established whether self-confusion might occur for a realistic number ($\lesssim 10^4$) of detectable signals, but common intuition is that it will not, due to the sheer volume of the signal space. This is largely borne out by the results of an idealized calculation that we present in Sec. III.

We use the term "degeneracy" to mean the presence of non-negligible and *nonlocal* correlations in the continuous signal space of a given waveform model. The nonlocal condition excludes its more informal usage in GW data analysis to describe extensive but essentially local regions of high correlation. Degeneracy manifests as disjoint secondary maxima in the overlap surface $\Omega(h_{\mathrm{inj}}, \cdot)$ over the model parameter space, where the overlap is between a reference signal injection h_{inj} and the signal template at each point. We will also exclude from our definition the characteristic sinc-like "ringing" of the overlap surface near the injection, as this is generally low-amplitude and still relatively local. Degeneracy does not affect the narrow band Galactic binaries, or massive-black-hole mergers with their rapid evolution and merger-dominated signal-to-noise ratio (SNR). It is a unique issue for EMRIs and has significant implications even in the case of a lone EMRI signal in the data.

EMRI degeneracy occurs when different and nonlocal combinations of the *source-intrinsic* parameters result in similar values for a subset of the three initial fundamental frequencies, as well as their starting time derivatives up to some order (see Sec. IV B 1). This causes the phasing of the dominant harmonic mode in both the injection and the degenerate template to be aligned for much of the inspiral duration. Such a criterion is highly improbable for a realistic set of putative signals, but we show in Sec. IV that it is satisfied in a surprisingly large number of disjoint regions across the space of all possible signals. An alignment of dominant and subdominant modes can also cause degeneracy [24,25]—but to a lesser extent since the relative distribution of mode amplitudes must also be similar for two signals to have a high overlap. This latter case might become more relevant when *extrinsic* parameters are considered (again, see Sec. IV B 1).

Early hints of degeneracy in the EMRI signal space arose during the Mock LISA Data Challenges [26], where participants either reported the presence of a few secondary peaks in the posterior surface, or misidentified one of them as the

primary peak (even with fairly localized priors). At SNRs above the common estimate of 20 for the EMRI detection threshold [17], these posterior *secondaries* are exponentially suppressed relative to the primary peak. However, they can pose practical difficulties for "uninformed" search and inference algorithms if the overlap against the injection at their locations is high, and/or if they are unexpectedly distant from the primary peak, and/or if they are inordinately numerous. Such information about the nature of secondaries has remained unavailable thus far; the global (log-)likelihood surface is extremely challenging to *map out* even for a single injection, much less a globally representative set of injections. This is in turn due to the computational limitations of existing waveform models, as well as standard sampling and clustering methods being ill suited to the task of high-dimensional mapping and visualization.

C. Synopsis of results

In this work, we investigate both confusion and degeneracy for EMRI signals, within the setting of LISA data analysis. The confusion study in Sec. III is the more straightforward of the two, as it is a conventional calculation that makes use of existing tools, and its results are perhaps somewhat unsurprising. We examine the pairwise overlaps among a set of N detectable two-year signals, whose source parameters are distributed according to a simple astrophysical model (modified by selection based on some SNR threshold). For astrophysically realistic estimates of N, we conclude that self-confusion is unlikely to pose a problem for the extraction and characterization of EMRI signals. No signal in a set of $N \approx 200$ should resemble any of the others beyond a $\sim 1\%$ overlap level, while the root-mean-square correlation among signals is approximately constant at $\sim 0.1\%$ for $N \lesssim 200$.

The bulk of this work is devoted to an extensive (but by no means comprehensive) study of EMRI degeneracy. Such a study is hindered by several factors: (i) the computational cost of generating analysis-length waveforms, even with simple proxy models; (ii) the focus of modern sampling algorithms on optimization and density estimation rather than mapping; (iii) the general difficulty of clustering and visualizing high-dimensional data; and (iv) the intricate structure in the overlap surface that borders on noise in any region away from the injection parameters (which turns out to make the problem of localizing, counting, or even defining "secondaries" somewhat ill posed). To overcome or at least circumvent these difficulties, we introduce various new tools (see Sec. II) such as a stripped-down version of a fast semirelativistic model [17], an approximation to the inner product $\langle \cdot | \cdot \rangle$, an exploratory sampling density for the recovery of high-overlap points, and a bespoke algorithm for the clustering of such points into approximately disjoint secondaries.

With these tools, we are able to unambiguously demonstrate the existence of degeneracy in the EMRI signal

space for signals as long as two years and can also shed some light on the severity and prevalence of secondaries. Our study is detailed in Sec. IV; we map out the overlap surface $\Omega(h_{\text{inj}}, \cdot)$ for a single representative injection at varying magnification levels, from a starting region that is tightly centered on the posterior bulk to a final region whose Euclidean volume is $\gtrsim 10^{12}$ times larger. We find that secondaries exhibit varying degrees of connectivity to one another and take on an assortment of shapes and scales with no immediately discernible global patterns. However, both their number density and their overlap against the injection appear to fall off with distance from the injection. Furthermore, case studies of specific secondaries also indicate that the posterior surface in those regions is unlikely to resemble the posterior corresponding to an actual injection at the same location.

In Secs. V and VI, we discuss the implications of our qualitative findings for EMRI data analysis, broadly review previously proposed search strategies, and provide our own suggestions to guide future work. We identify pathological scenarios that might arise from the unfortunate interaction of degeneracy with detector noise or the presence of multiple actual signals, but argue that they are highly improbable. Thus the main issues we foresee are practical in nature: the computational difficulty of stochastic search and the verification of candidate signals. Past research on EMRI data analysis has mostly attempted to address the former, with some degree of success; here we propose several complementary strategies to tackle both issues, informed by the results of our confusion and degeneracy studies. These take the form of *post hoc* vetoes for candidate signals, as well as a modified "veto likelihood" that is designed for search.

II. INVENTORY OF TOOLS

A. Waveform and response models

Fully relativistic EMRI waveform models that are both efficient and extensive enough for data-analysis studies are still under active development. The current state of the art for efficiency-oriented waveforms is a fast model describing eccentric inspirals in Schwarzschild spacetime [10], whose frequency evolution is accurate up to leading *adiabatic* order in the mass ratio ϵ. In terms of extensiveness, recent work has introduced several models for various classes of inspirals in Kerr spacetime: eccentric and equatorial with adiabatic evolution [27], fully generic (i.e., eccentric and inclined) with a post-Newtonian (PN) approximation to adiabatic evolution [28], and fully generic with adiabatic evolution [29]. Such techniques will soon be combined to construct fast and fully generic adiabatic Kerr models—these in turn are precursors to the accurate, efficient, and extensive *postadiabatic* models [30] that will enable us to achieve the EMRI-related science goals of LISA.

None of the existing models from above are suitable for the present work, which requires large-scale Monte Carlo simulations along with a sufficiently representative depiction of the full Kerr signal space. We instead employ a semirelativistic model for fully generic Kerr inspirals, which combines adiabatic-fitted frequency evolution [31] with Newtonian instantaneous amplitudes (in the Peters-Mathews formalism [32]). This model is known as the augmented analytic "kludge" (AAK) [17]; it is based largely on an influential earlier construction for LISA data-analysis studies [15], but correctly accounts for relativistic frequencies [33]. We refer the reader to [17] for a full description of this model, along with a short review of its kludge predecessors [15,16].

As our study on EMRI confusion involves a straightforward analysis with an unambiguous answer, it is worth doing with slightly more "realistic" tools. There we use the AAK waveform integrated with a fast model for the time-delay-interferometry (TDI) [34] response $h_{A,E,T}$ of LISA to the signal, as implemented in [25,35]. In the degeneracy part of this work, we opt instead for the much simpler long-wavelength LISA response $h_{I,II}$ [36] (essentially, approximating LISA as a point detector relative to the gravitational wavelength). One reason is that the AAK–TDI model was only developed well after the start of our investigation—but in any case, the choice of response model should not significantly alter the structure of source-intrinsic degeneracy, which is our focus here. This is due to the weak coupling between intrinsic and extrinsic parameters, as discussed in Sec. IV A.

The full AAK model (i.e., including LISA response) is parametrized by a set of 14 source parameters, which can be further partitioned into the intrinsic parameters

$$\boldsymbol{\theta}_{\mathrm{int}} := (\mu, M, a/M, p_0/M, e_0, \iota, \Phi_0, \gamma_0, \alpha_0) \qquad (1)$$

and the extrinsic parameters

$$\boldsymbol{\theta}_{\mathrm{ext}} := (\theta_K, \phi_K, \theta_S, \phi_S, D), \qquad (2)$$

where all quantities are dimensionless. Explicitly:
  (i) (μ, M) are the detector-frame component masses in solar masses and with $\epsilon = \mu/M \ll 1$;
  (ii) a is the Kerr spin length scale for the central mass;
  (iii) (p_0, e_0, ι) are the quasi-Keplerian semilatus rectum, eccentricity, and inclination for the osculating geodesic to the inspiral at reference time t_0;
  (iv) $(\Phi_0, \gamma_0, \alpha_0)$ are phase angles describing the position of the small mass at reference time t_0;
  (v) (θ_K, ϕ_K) are polar and azimuthal angles describing the spin orientation in ecliptic coordinates;
  (vi) (θ_S, ϕ_S) are polar and azimuthal angles describing the sky location in ecliptic coordinates;
  (vii) D is the luminosity distance in Gpc.

Henceforth we will abuse the symbol $\boldsymbol{\theta}$ to denote any ordered combination of these parameters, as long as the specific parameters in question are clear from context.

There are several points to note about this choice of parametrization. First, the distinction between intrinsic and extrinsic EMRI parameters can be somewhat arbitrary and also depends on whether one is working in the modeling or analysis context. It is a largely semantic distinction, though, and here we take more of a modeling viewpoint (as opposed to [17])—but note that the sets $\{p_0, e_0, \iota\}$ and $\{\Phi_0, \gamma_0, \alpha_0\}$ each technically have one observer-dependent degree of freedom, corresponding to temporal translation and spatial rotation, respectively. Second, the inclination angle used is $\iota := \tan^{-1}(\sqrt{Q}/L_z)$ rather than $I := \pi/2 - \mathrm{sgn}(L_z)\theta_{\min}$ (where $(Q, L_z, \theta_{\min})$ are the Carter constant, the projection of angular momentum onto the spin axis and the turning point of polar motion). Usage of the latter is becoming standard in the modeling community, but $\iota \approx I$ across much of the Kerr geodesic space [37]. The rate of change for ι due to GW radiation is also generally small [38] and is thus approximated as zero in the AAK model. Finally, future postadiabatic models are likely to have a qualitatively similar parametrization, apart from: (i) the time evolution of (M, a) [8], which would simply change $(M, a) \to (M_0, a_0)$; and (ii) secular or resonant effects due to the spin of the small mass [39–44], although more detailed studies are required to determine how measurable this quantity will be with LISA.

The AAK model is efficiency oriented, with a computational wall time of ∼10 s for a two-year signal sampled at 0.1 Hz. While this is adequate for the confusion study, our investigation of EMRI degeneracy calls for multiple Monte Carlo simulations, each with up to billions of template evaluations (if this seems excessive, recall that our aim here is high-resolution mapping—not search or inference). For the degeneracy study, we thus rely on an amplitude-and-phase representation of the AAK model, further stripped down to just four strong harmonic modes. Although the semirelativistic and quadrupolar AAK waveform falls well short of realistic harmonic content, this still reduces the number of modes by about an order of magnitude. The amplitude and phase trajectories also vary smoothly over the radiation-reaction timescale M/ϵ and are downsampled by a factor of $\gtrsim 10^3$ for direct use in an approximate inner product (see Sec. II B). With these simplifications, the wall time for a single template evaluation (plus typical operations on said template) is slashed to $\lesssim 10$ ms.

An explicit description of harmonic modes in the AAK model will depend on the specific choice of harmonic basis. The model is constructed from Keplerian orbits with artificially induced precession; this is reflected in the phase parametrization (Φ, γ, α), where Φ is the quasi-Keplerian mean anomaly and (γ, α) are two precession-related angles. Data-analysis studies that involve the AAK (and its

predecessor [15]) typically use the frequencies $(\dot{\Phi}, \dot{\gamma}, \dot{\alpha})$ as the harmonic basis, where $\dot{\gamma} + \dot{\alpha}$ and $\dot{\alpha}$ are the rates of periapsis and Lense-Thirring precession, respectively. The phases (Φ, γ, α) have a simple relation to the *fundamental phases* $(\varphi_r, \varphi_\theta, \varphi_\phi)$, i.e., the generalized coordinates associated with the action-angle variables for Kerr geodesic motion [45]:

$$(\varphi_r, \varphi_\theta, \varphi_\phi) = (\Phi, \Phi + \gamma, \Phi + \gamma + \alpha). \tag{3}$$

Another harmonic basis is thus provided by the Kerr *fundamental frequencies* $(\omega_r, \omega_\theta, \omega_\phi)$—the derivatives of $(\varphi_r, \varphi_\theta, \varphi_\phi)$ with respect to coordinate time. We choose the latter basis for this work. The phase and angular frequency of a mode (m, k, n) are given, respectively, by

$$\varphi_{mkn} := m\varphi_\phi + k\varphi_\theta + n\varphi_r, \tag{4}$$

$$\omega_{mkn} := m\omega_\phi + k\omega_\theta + n\omega_r, \tag{5}$$

where ω is used here and henceforth to denote dimensionful frequencies with units of Hz.

For illustrative purposes, we introduce here the reference signal injection h_{inj} that is used in the degeneracy study of Sec. IV; its intrinsic parameters are

$$\boldsymbol{\theta}_{\text{inj}} = (10, 10^6, 0.5, 9.5, 0.2, \pi/6, 0, 0, 0), \tag{6}$$

with randomly chosen extrinsic parameters. The masses and spin are assigned "central" values in the LISA-relevant range, while the initial semilatus rectum is fixed by requiring that the small mass plunges (reaches the Kerr separatrix) at time $t_p = t_0 + 2$ y. We focus on low-to-moderate starting eccentricity, due to the trimming of modes, and examine the slightly inclined prograde case. Over its two-year duration, much of the power in the signal is contributed by the four modes

$$m = 2, \qquad k = 0, \qquad -1 \le n \le 2, \tag{7}$$

and, in particular, by the (2,0,0) mode. This can be visualized through the time-frequency plots in Figs. 1 and 2, for the h_I channel of the long-wavelength response $h_{\text{I,II}}$.

The exact AAK implementation used in this work is that from the now-discontinued EMRI Kludge Suite (v0.5.2) [46], which is largely faithful to the original presentation in [17]. During the course of this work, an updated AAK model with 5PN evolution [27] and graphics processing unit (GPU) support was included in the Fast EMRI Waveforms software package that will house the next-generation of EMRI models [47,48]. The AAK–5PN waveform features several improvements, such as exact fundamental frequencies (rather than their PN expansions) and an evolving inclination, and is also significantly accelerated over the
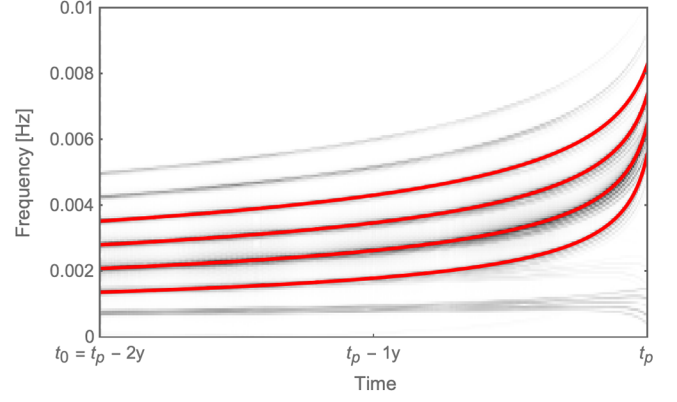

FIG. 1.   Spectrogram of reference signal (h_I channel only) with intrinsic parameters (6). Short-time Fourier amplitude (gray scale value) is in log scale. Overlaid in red are frequency trajectories for the four strong harmonic modes (7).
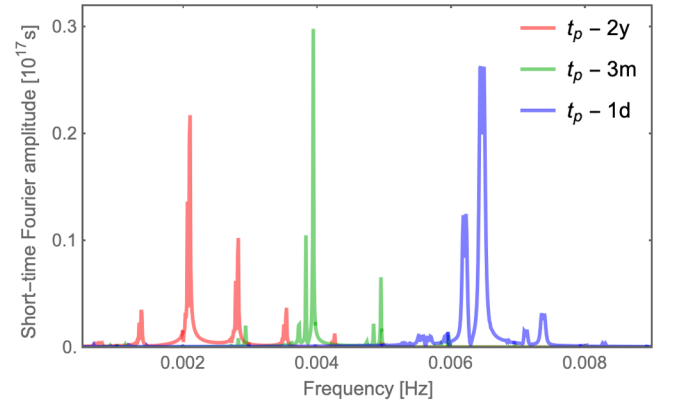

FIG. 2.   Linear-scale cross sections of the spectrogram in Fig. 1, at two years/ three months / one day before plunge. The four modes (7) are mostly representative of the signal, but strong sidebands become more resolvable in the final months.

CPU version in [46]. Future follow-up studies without the approximations made in this work will likely employ that model, or a relativistic 5PN model [28] that is currently being implemented within [48]. However, our present approach is still an order of magnitude faster than the GPU generation and manipulation of templates at full sampling resolution—and more importantly, we do not expect our main results to change qualitatively for other EMRI models (see extended discussion in Sec. IV B 3).

B. Similarity measures

In GW data analysis, the detection of a signal h in noisy time-series strain data $x = h + n$ relies on the linear filtering of x against a signal template that matches h. If the detector noise n is approximated as a zero mean and stationary process, then this matched-filtering procedure can be expressed as a noise-weighted cross-correlation

$$\langle x|h\rangle := 4\mathrm{Re}\sum_{\chi}\sum_{f>0}^{f_N}\delta f\frac{\tilde{x}_\chi(f)^*\tilde{h}_\chi(f)}{S_{n,\chi}(f)}, \qquad (8)$$

where the outer sum is over all independent data channels χ (for this work, $\chi = \mathrm{A,E,T}$ or $\chi = \mathrm{I,II}$); f_N is the Nyquist frequency; δf is the frequency resolution; overtildes denote discrete Fourier transforms (multiplied by the time resolution δt); and $S_{n,\chi}$ is the one-sided power spectral density of the channel noise n_χ (provided here by analytic models that correspond to the science requirements for the LISA mission [49–51]). The noise assumptions and the form of Eq. (8) give rise to the identities

$$\mathrm{E}[\langle n|a\rangle] = 0, \qquad (9)$$

$$\mathrm{E}[\langle n|a\rangle\langle n|b\rangle] = \langle a|b\rangle, \qquad (10)$$

valid for all time series a, b with the same length as n.

Equation (8) satisfies the conditions for an inner product on the data space $\mathcal{D}$ of fixed-length time/frequency series, which unlocks a useful geometric picture of other common concepts in GW data analysis. For one, the overlap between two signals $h_{1,2}$ is simply their normalized inner product, or the cosine of the angle between them in $\mathcal{D}$:

$$\Omega(h_1, h_2) := \frac{\langle h_1|h_2\rangle}{\sqrt{\langle h_1|h_1\rangle\langle h_2|h_2\rangle}}. \qquad (11)$$

The optimal SNR of a signal template h is its norm:

$$\rho_{\mathrm{opt}}(h) := \frac{\mathrm{E}[\langle x|h\rangle]}{\sqrt{\mathrm{E}[\langle n|h\rangle\langle n|h\rangle]}} = \sqrt{\langle h|h\rangle}, \qquad (12)$$

and its detection SNR is the scalar projection of x on h:

$$\rho_{\mathrm{det}}(h) := \frac{\langle x|h\rangle}{\sqrt{\langle h|h\rangle}}. \qquad (13)$$

Maximum-likelihood estimation with the standard GW likelihood $L(\boldsymbol{\theta}|x)$ (see Sec. II C) boils down to distance minimization between x and the signal space $\mathcal{S}$ in $\mathcal{D}$:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}}\langle x - h(\boldsymbol{\theta})|x - h(\boldsymbol{\theta})\rangle. \qquad (14)$$

Finally, the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}_{\mathrm{inj}})$ [52] for $L(\boldsymbol{\theta}|x) \propto p(x|\boldsymbol{\theta}_{\mathrm{inj}})$ coincides precisely (component-wise) with the pullback by $h(\boldsymbol{\theta})$ of the flat metric on $\mathcal{D}$:

$$[\mathcal{I}(\boldsymbol{\theta}_{\mathrm{inj}})]_{ij} = \langle\partial_i h(\boldsymbol{\theta})|\partial_j h(\boldsymbol{\theta})\rangle|_{\boldsymbol{\theta}_{\mathrm{inj}}}. \qquad (15)$$

This geometric picture extends to the concept of local correlations in signal space and thus to that of degeneracy—the presence of non-negligible and nonlocal

correlations. From the informal definition in Sec. I B, the correlation between two signals is local if there is a "small" distance between their corresponding points in parameter space Θ (with respect to the pullback metric at one of the points). To make this slightly more concrete, we will say that two parameter points $\boldsymbol{\theta}_{1,2}$ are local with respect to each other if their metric distance in Θ approximately equals the Euclidean distance in $\mathcal{D}$ between their associated signals $h_{1,2}$, i.e.,

$$\sqrt{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T\mathcal{I}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} \approx \sqrt{\langle h_1 - h_2|h_1 - h_2\rangle}, \qquad (16)$$

where $\mathcal{I}$ is also approximately invariant: $\mathcal{I}(\boldsymbol{\theta}_1) \approx \mathcal{I}(\boldsymbol{\theta}_2)$. In other words, $h_{1,2}$ are local if they coexist in a region where the signal manifold is nearly flat. From a data-analysis perspective, the definition (16) holds in a region around the maximum-likelihood parameters, where the likelihood is near-Gaussian, and to a lesser extent in the immediate vicinity of this region, with decaying likelihood oscillations in parameter directions that affect frequency. Degeneracy on the other hand occurs when the specific embedding $h(\boldsymbol{\theta})$ admits (regions of) signals that have high overlap against some $h_{\mathrm{inj}} \in \mathcal{S}$, but whose parameters are nonlocal to the neighborhood of $\boldsymbol{\theta}_{\mathrm{inj}} \in \Theta$.

The inner product $\langle\cdot|\cdot\rangle$ is defined for time series at full sampling resolution and can be computationally unwieldy even without accounting for the cost of template generation. For the degeneracy study in Sec. IV, we introduce an approximation to $\langle\cdot|\cdot\rangle$ that acts directly on amplitude and phase trajectories from the AAK model. Beyond its original partial decomposition into harmonic modes of only radial motion, the AAK waveform with LISA response $h_{\mathrm{I,II}}$ is more fully decomposed as

$$h_{\mathrm{I}}(t) + ih_{\mathrm{II}}(t) = \sum_j A_j(t)\exp(i\varphi_j(t)), \qquad (17)$$

where $1 \leq (j := n + 2) \leq 4$ is a reindexing of the four strong modes (7), and A_j is a complex amplitude for mode j [46]. The mode phasing φ_j is obtained from the time integration of frequency trajectories $\omega_j(t)$, whose availability also allows us to fold the noise weighting of $\langle\cdot|\cdot\rangle$ directly into the time-domain signal [15]. Using overbars to denote noise-weighted quantities, $\bar{h}_{\mathrm{I,II}}$ is given by the analog of Eq. (17) with $A_j \to \bar{A}_j$, where

$$\bar{A}_j(t) := \frac{A_j(t)}{\sqrt{S_n(f_j(t))}}; \qquad (18)$$

i.e., the mode amplitudes at time t are essentially reduced by the noise estimate $S_n := S_{n,\mathrm{I}} = S_{n,\mathrm{II}}$ at the corresponding instantaneous mode frequencies $f_j := \omega_j/(2\pi)$.

In the above representation, the mode amplitude and phase trajectories are smooth enough to be downsampled

significantly. For our degeneracy study, we consider a signal injection specified by the source parameters (6), with an analysis duration of $T = t_p - t_0 = 2$ y and a fixed trajectory time step of $\Delta t = T/10^3$. All injection-template comparisons are performed on the common analysis interval $\{t : t_0 \leq t \leq t_0 + T\}$, using the same trajectory timestamps. Note that φ_j is technically ill defined at a set of times in this interval for templates that plunge before $t_0 + T$, but we also have $A_j = 0$ at those times. An approximation to the noise-weighted inner product $\langle \cdot | \cdot \rangle$ between two signals h, h' is then defined as

$$(h|h') := \mathrm{Re} \sum_{jj'} \sum_{t \in \tau_{jj'}} \Delta t \bar{A}_j(t)^* \bar{A}'_{j'}(t), \qquad (19)$$

$$\tau_{jj'} := \{t : |\varphi_j(t) - \varphi'_{j'}(t)| < 1\}, \qquad (20)$$

with $\tau_{jj'}$ being the set of times at which mode j is "in phase" with mode j'. The phase tolerance value of 1 in Eq. (20) is chosen such that $(\cdot | \cdot)$ gives an upper envelope for $\langle \cdot | \cdot \rangle$ between two sinusoidal signals, as it varies with the difference in their frequencies (see Fig. 3). More generally, we have

$$(h|h') \gtrsim 2 \sum_{\chi = \mathrm{I,II}} \sum_{t=t_0}^{t_0+T} \delta t \bar{h}_\chi(t) \bar{h}'_\chi(t) \approx \langle h|h' \rangle. \qquad (21)$$

Since Eq. (19) is undefined for general time series in the data space $\mathcal{D}$, it is not strictly an inner product, but rather approximates the restriction of $\langle \cdot | \cdot \rangle$ to the signal space $\mathcal{S}$. For convenience, however, we will call $(\cdot | \cdot)$ the approximate "inner product" and $\langle \cdot | \cdot \rangle$ the full inner product. The approximate inner product is only used in the degeneracy study of Sec. IV and even there only in bulk calculations, for which the full inner product would be unfeasible. Qualitative results obtained with $(\cdot | \cdot)$ are verified with $\langle \cdot | \cdot \rangle$ if practical; e.g., after a set of high-overlap points from
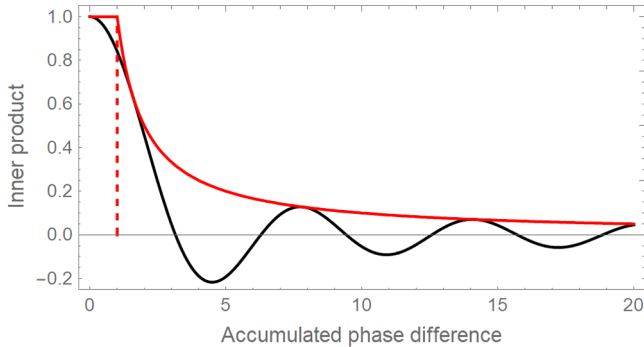

FIG. 3. Full (black) and approximate (red) inner product between $\cos(\omega t)$ and $\cos(\omega' t)$ over duration $T = 1$, as a function of accumulated phase difference $(\omega - \omega')T$. Vertical red dashed line indicates specified phase tolerance of 1.

sampling has been distilled to a smaller set of representative points by clustering, we report their overlaps with respect to $\langle \cdot | \cdot \rangle$. Thus we will, in general, refer to both $(\cdot | \cdot)$ and $\langle \cdot | \cdot \rangle$ simply as the inner product and will only distinguish between them (and in derived quantities such as the overlap) if important qualitative differences arise due to the approximation.

C. Sampling densities

Of central importance in GW inference is the standard Whittle likelihood function $L(\boldsymbol{\theta}|x) \propto p(x|\boldsymbol{\theta}_{\mathrm{inj}})$ (with the distinction between the random argument $\boldsymbol{\theta}$ and the fixed quantity $\boldsymbol{\theta}_{\mathrm{inj}}$ being explicitly highlighted). As in Eq. (8), we take the data as $x = h_{\mathrm{inj}} + n$ with zero mean and stationary noise, but here n is additionally assumed to be a Gaussian process [53]. The natural logarithm of L (informally, the "log-likelihood") is given by

$$\ln L_{\mathrm{full}}(\boldsymbol{\theta}|x) := -\frac{1}{2} \langle x - h(\boldsymbol{\theta})|x - h(\boldsymbol{\theta}) \rangle. \qquad (22)$$

Functional form notwithstanding, Eq. (22) describes a non-Gaussian density function on the parameter space Θ; however, from the definition of locality (16), a Gaussian approximation to L is reasonably valid in the local neighborhood of $\boldsymbol{\theta}_{\mathrm{ML}}$ from Eq. (14). Since Eq. (22) is defined in terms of the full inner product $\langle \cdot | \cdot \rangle$, we will refer to it in the context of this work as the full (log-)likelihood.

The full likelihood is not used in our studies beyond the initial validation of an approximate likelihood, which is simply given by Eq. (22) with $\langle \cdot | \cdot \rangle \to (\cdot | \cdot)$:

$$\ln L_{\mathrm{app}}(\boldsymbol{\theta}|x) := -\frac{1}{2} (x - h(\boldsymbol{\theta})|x - h(\boldsymbol{\theta})). \qquad (23)$$

With the reduced cost of both the template-generation and inner-product operations, L_{app} is $\gtrsim 10^3$ times faster to evaluate than L_{full}. Efficiency considerations aside, Eq. (23) also facilitates the degeneracy study in Sec. IV by smoothing out local oscillations in the posterior surface, such that the maxima identified by our algorithms are generally due to nonlocal correlations. By design, the profile of L_{app} over Θ corresponds roughly to the upper envelope of L_{full}; this relation is not exact as depicted for the inner products in Fig. 3 since the signals here have evolving frequency, and the log-likelihood contains normalization terms in addition to the inner product.

For conceptual ease, let us consider the specific noise realization $n = 0$ without much loss of generality, such that $x = h_{\mathrm{inj}}$ lies in the signal space $\mathcal{S}$ with SNR $\rho_{\mathrm{inj}} := \rho_{\mathrm{opt}}(h_{\mathrm{inj}}) = \rho_{\mathrm{det}}(h_{\mathrm{inj}})$. Equation (23) may be written as

$$\ln L_{\mathrm{app}}(\boldsymbol{\theta}|x) = \rho_{\mathrm{inj}} \rho_{\boldsymbol{\theta}} \Omega(x, h(\boldsymbol{\theta})) - \frac{1}{2}(\rho_{\mathrm{inj}}^2 + \rho_{\boldsymbol{\theta}}^2), \qquad (24)$$

where $\rho_{\boldsymbol{\theta}} := \rho_{\text{opt}}(h(\boldsymbol{\theta}))$. In the tail regions of the likelihood, where $\mathrm{E}[\Omega(x,h)] \approx 0$, we see that $L_{\text{app}} \overset{\propto}{\sim} \exp\left(-1/2\rho_{\boldsymbol{\theta}}^2\right)$ is dominated by large-scale gradients arising from the slow variation of $\rho_{\boldsymbol{\theta}}$ over Θ [relative to $(x|h(\boldsymbol{\theta})]$. As these gradients can hinder a global analysis of intrinsic parameter degeneracy, it is useful to introduce a *matched-SNR* version of L_{app}, with templates that have been renormalized to match the injection SNR:

$$\ln L_{\text{match}}(\boldsymbol{\theta}|x) := \rho_{\text{inj}}^2(\Omega(x,h(\boldsymbol{\theta})) - 1). \qquad (25)$$

This matched-SNR likelihood is proportional to the overlap surface $\Omega(h_{\text{inj}},\cdot)$ over Θ; in the vicinity of $\boldsymbol{\theta}_{\text{inj}}$, we also have $L_{\text{match}} \approx L_{\text{app}}$. Our degeneracy study leverages the absence of large-scale SNR gradients in Eq. (25) to search more effectively for secondary peaks in L_{match} (and the overlap surface), which then have the same locations as secondaries in L_{app} (and the posterior surface). This is because the intrasecondary variation of $\rho_{\boldsymbol{\theta}}$ is generally small, as opposed to its intersecondary variation.

As it turns out, straightforward sampling of the matched-SNR likelihood is unsuitable for finding, resolving, and visualizing secondaries in L_{match}; even if these secondaries were cleanly localized and separated (they are not), they would be exponentially suppressed at realistic values of $\rho_{\text{inj}} \gtrsim 10$ and contiguously congealed at artificially lowered values—or high temperatures, in the language of annealing Markov-chain Monte Carlo (MCMC) methods [54,55]. Our solution is to introduce an *exploratory* "likelihood" that is used only for global exploration and to obtain a large set of high-overlap points for follow-up clustering analysis. This is given by L_{match} with both injection and template normalized to unit SNR, plus artificially suppressed tails:

$$\ln L_{\text{expl}}(\boldsymbol{\theta}|x) := \frac{1}{\rho_{\text{inj}}^2}\ln L_{\text{match}}(\boldsymbol{\theta}|x) + \text{ramp}(\boldsymbol{\theta}|x), \qquad (26)$$

$$\text{ramp}(\boldsymbol{\theta}|x) := 100 \ \min\{\Omega(x,h(\boldsymbol{\theta})) - 0.5, 0\}. \qquad (27)$$

Equation (26) is designed to reduce the volume in the tail regions of the matched-(unit-)SNR likelihood, so as to better locate and flesh out individual secondaries. The exact specifications of the ramp function (27) are empirically chosen, and not necessarily optimal. While we will continue referring to Eq. (26) as a (Bayesian) likelihood, it can no longer be linked to a natural probabilistic statement based on noise and signal assumptions. This last point also holds true for a *veto* likelihood that we propose as a mitigation of EMRI degeneracy, but whose presentation we delay until Sec. VI (as it is a consequence, rather than component, of the degeneracy study).

D. Clustering algorithm

With the exploration aspect of signal-space mapping covered by the stochastic sampling of the exploratory likelihood in Sec. II C, the remaining analysis boils down to visualizing and making sense of the resultant high-dimensional data. Central to this is the task of clustering. Some of the commonly used sampling algorithms in GW data analysis [56–58] employ in-built clustering to aid convergence on multimodal distributions; this form of clustering is more of an intermediate means to an end, with the byproduct set of identified "clusters" being of limited utility for realistic examples. A specialized clustering algorithm is required here, but the main off-the-shelf options are unsuitable as well. *Centroid-based* methods such as k-means clustering [59] and its many variants use geometric proximity to define cluster membership. They typically require the number of clusters to be prespecified and are thus immediately inadequate for our purposes. *Density-based* methods [60] perform clustering based on the density of points in the data set. In this context, they would rely on the sampler being able to populate secondaries in the correct proportion—which is, in general, nontrivial to guarantee even for "clean" multimodal distributions. Such methods also tend to rely on arbitrarily tuned criteria for cluster membership, which can significantly affect the cluster count itself.

We introduce here a bespoke clustering algorithm that can be classified as a *connectivity-based* method [59]. For a general signal space, the overlap between two signals provides a measure of connectivity between their associated parameter points. It can thus be used directly in clustering applications; for example, to group posterior samples in multisource transdimensional-MCMC searches for Galactic-binary signals in LISA data [23], by assigning sample points to the same cluster if there is a high overlap between their associated templates. This works well for resolving sources in the absence of parameter degeneracy because it agrees with the natural measure of connectivity defined by the metric distance (16) in Θ (which is required for sensible results). When nonlocal correlations are present, this sort of clustering might instead be used as a veto for spurious source candidates that arise due to secondaries (see Sec. VI A).

In the context of our degeneracy study, however, an alternative notion of connectivity is required. As discussed in Sec. II C, the key quantity in the understanding of posterior secondaries is the overlap function $\Omega(h_{\text{inj}},\cdot)$. We are not restricted to the overlap information in the raw data set that is obtained by sampling L_{expl} since our access to a generative model for the overlap can be used to provide additional information during clustering. Let us first define the *connection* ℓ between two points $\boldsymbol{\theta}_{1,2} \in \Theta$: a vector of l equally spaced evaluations of $\Omega(h_{\text{inj}},\cdot)$ along the connecting line between $\boldsymbol{\theta}_{1,2}$ (inclusive). We set $l = 7$ in our algorithm, which is empirically determined to be the

minimum value required for satisfactory convergence of results. The connection ℓ is then used to construct a symmetric *premetric* d on parameter space, satisfying only $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq 0$, $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) = 0$ and $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = d(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$ [61]. This premetric is defined as

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \coloneqq 1 - \frac{\min_i \ell_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\min\{\ell_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \ell_l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}}, \quad (28)$$

such that two points with "predistance" zero have no intermediate point with a lower overlap value than either.

With the existence of a premetric, we may define various degrees of "connectedness" for use in the clustering algorithm. We will say that two points are *strictly connected* if they have a predistance $d = 0$, that they are *connected* if $d < 0.5$, and that they are *not connected* if $d \geq 0.5$. The additional category of being connected, rather than only strictly connected or not strictly connected, is required in our algorithm to handle intrasecondary structure. In other words, we are defining secondaries as disjoint clusters of points that are uniquely connected to some representative point (generally a cluster maximum), instead of simply calling all local maxima secondaries (because these can and often do coexist in contiguous regions of high overlap). Since we will only be clustering high-overlap points with $\Omega(h_{\text{inj}}, \cdot) \gtrsim 0.5$, any two connected points will have no intermediate point with $\Omega(h_{\text{inj}}, \cdot) \lesssim 0.25$, which admits an interpretation of connectedness in terms of absolute overlap. The predistance between two points also depends more strongly on their overlap values when $d \geq 0.5$, such that disconnected points with higher overlaps tend to be "further apart." This allows connectedness to be used in locating cluster maxima, in addition to the more obvious function of determining cluster membership.

Using the above definitions of connectedness as a notion of connectivity leads to a clustering problem with a significant degree of noise. In graph-theory terms, this is because the graph for a set of parameter points (vertices) and their pairwise connectedness (edges) has high vertex connectivity [62] and cannot be partitioned into disjoint subgraphs without removing a large number of vertices. Another issue is that standard connectivity-based algorithms are generally computationally expensive, with their complexity scaling as $\mathcal{O}(N^2)$ or worse for a data set of size N—for example, if one were to compute the full similarity matrix of pairwise predistances for use in spectral-clustering methods [63]. To combat these difficulties, we take a greedy approach [64] to the identification of candidate *nodes* (representative points for clusters), with global and local steps to ensure that these nodes pick out any clear maxima without the need for full-blown optimization. Another design objective is for the algorithm to run in $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$ time. A single iteration of our clustering algorithm is described below in pseudocode:

(1) At the ith iteration, there exists an $(i-1) \times N$ matrix containing the predistance of all points in the data set to each of the existing $i - 1$ nodes. Each point has a minimal predistance to the set of existing nodes. (For the set of nodes themselves and the points that are strictly connected to them, this minimal predistance will be zero.) Choose the point with the largest minimal predistance as the preliminary ith node. This is the point that is "least connected" to the existing nodes, and the process of finding it is akin to a global-search step.

(2) Compute the predistance of all points to the preliminary ith node.

(3) Define the strict ith cluster as the set of points that are strictly connected to the preliminary node and not connected to any of the existing $i - 1$ nodes. Choose the point in the strict ith cluster with the highest overlap value as the actual ith node. This process is akin to a local-maximization step.

(4) If the preliminary ith node is not the same point as the actual ith node, then "recenter" the cluster by computing the predistance of all points to the actual ith node. In principle, the actual node might still not be the highest-overlap point in its cluster after recentering; however, additional recentering steps are found to offer marginal gain for their cost.

(5) Append the list of predistances for the ith node to the predistance matrix as a new row.

(6) Compute the cluster coverage, which is the fraction of points that are connected to at least one of the existing i nodes. If the coverage equals unity, then end the algorithm. Defining the stopping criterion in terms of being connected rather than strictly connected causes the cluster count to depend on the arbitrary connectedness threshold $d = 0.5$, but this just corresponds to setting a limit on what is defined as a clear cluster—which is never fully avoidable in any nontextbook clustering task.

The only special iteration is the initial one, where both the preliminary and actual nodes in steps 1 and 3 are simply chosen to be the point in the data set with the highest overlap value $\Omega(h_{\text{inj}}, \cdot)$. Computing new rows of the predistance matrix in steps 2 and 4 dominates the algorithm cost, which thus scales as $\mathcal{O}(N)$ times the number of identified nodes. With the final predistance matrix in hand, it is straightforward to construct a set of clusters *post hoc*. The final (strict) cluster for each node is defined as the set of points that are (strictly) connected to the node and not (strictly) connected to any of the other nodes. Note that the strict cluster here differs slightly from that defined in step 3, where the short list of candidates for the actual node is more stringent by design. For illustrative purposes, we provide in Fig. 4 a heuristic example that depicts the application of the algorithm to an artificial one-dimensional overlap surface.
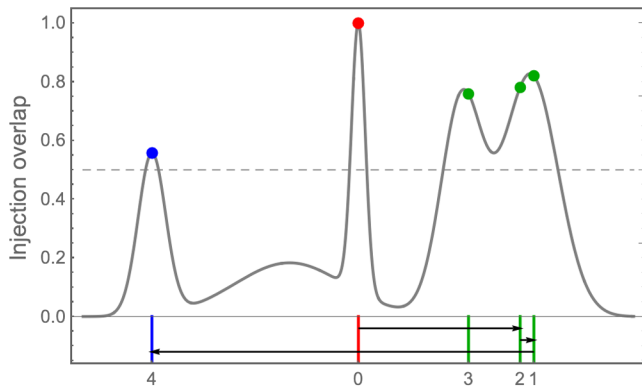
FIG. 4. Application of the clustering algorithm to an artificial one-dimensional overlap surface and a data set of $N = 5$ high-overlap points (with zero-based index). At the first iteration, point 0 (the injection) is the identified node, and the cluster coverage is 0.2. At the second iteration, point 2 ($\Omega = 0.78$) is the preliminary node with largest minimal predistance $d = 0.95$; note that this is neither the point with the next highest overlap, nor the point with the largest metric distance. Point 1 ($\Omega = 0.82$) is the actual node after recentering, and the cluster coverage is 0.8 (points 2 and 3 are both connected to point 1). At the third iteration, point 4 is the identified node with largest minimal predistance $d = 0.88$ (to point 0), and the algorithm terminates with a cluster coverage of 1. Color grouping of points indicates the final clusters for each node.

Our clustering algorithm uses information about $\Omega(h_{\text{inj}}, \cdot)$ beyond the original data set, and thus its performance does not rely strongly on the distribution of the data. The number and composition of clusters are both robust to the choices of connection length $l \gtrsim 7$ and initial node (provided it is close to a local maximum), with variations in the results that are far smaller than errors due to imperfect sampling. Nontrivial structure in the high-dimensional overlap surface can be unearthed even from relatively small data sets with $N \lesssim 10^4$, as demonstrated particularly clearly in Sec. IV A 5. Although the algorithm was developed for our specific purposes, it essentially only requires the surface under examination to be evaluable separately from the data and thus might find utility in more general clustering applications. In the degeneracy study of Sec. IV, we employ it both for large-scale mapping to identify disjoint secondary regions of high overlap and for localized mapping in these regions to resolve and visualize the finer structure.

III. CONFUSION STUDY

Before turning our attention to correlations in the continuous space of possible signals, we seek to address a simpler (two-part) question: For an astrophysically realistic set of EMRI signals in LISA data, (i) how correlated are the most correlated pair of signals; and (ii) how does the overall degree of correlation among the set vary with the size of the set? The answer to this bears on both EMRI search and

inference; large overlaps between actual signals will hinder their resolution during search and will necessitate a combined inference of their source parameters. This in turn affects the design of data-analysis strategies, as well as the prospects of EMRI science. The conclusion of our study is unequivocal, at least for a middle-ground estimate of 200 detectable signals—self-confusion is unlikely to be an issue for EMRIs.

A. Astrophysical model

Models for the astrophysical population of EMRIs that LISA will observe are weakly constrained by existing knowledge and thus span a broad range of predictions. Some of the main uncertainties lie in the massive-black-hole mass function, in the fraction of such black holes hosted in dense stellar cusps, as well as in the intrinsic rate of EMRI formation per massive black hole [4]. In this work, we consider a population of detectable sources from a single representative model since the distribution of these sources is dominated by detector-specific selection effects anyway. Furthermore, our intuition *a priori* is that the results of the confusion study will be virtually independent of model choice, as they are simply determined by the intrinsic nature of signal space—specifically, the large information volume expected to be retained by any astrophysically restricted subspace.

We now define a population model that corresponds approximately to M1 from [4], which is often used as a representative model in LISA Science Group work-package studies (e.g., [65]). Our model specifies independent distributions for each of the AAK parameters (1) and (2), but with the source-frame masses (μ', M') and redshift z in place of the detector-frame masses (μ, M) and luminosity distance D. The map $(\mu', M', z) \rightarrow (\mu, M, D)$ is given by the usual relations

$$\mu(\mu', z) = (1 + z)\mu', \qquad M(M', z) = (1 + z)M', \qquad (29)$$

as well as [66]

$$D(z) = (1 + z) \int_0^z dz' \frac{c}{H(z')}, \qquad (30)$$

$$H(z) = H_0\sqrt{(1 - \Omega_\Lambda)(1 + z)^3 + \Omega_\Lambda}, \qquad (31)$$

with $c/H_0 = 4280$ Mpc and $\Omega_\Lambda = 0.7$.

The probability density of the joint distribution for $\boldsymbol{\theta}$ is the product of the individual parameter densities $p(\boldsymbol{\theta}_i)$, which are supported on the corresponding sets $\Delta\boldsymbol{\theta}_i \subset \mathbb{R}$ (to consolidate notation, $\Delta\boldsymbol{\theta}_i$ can have measure zero). These sets are listed in Table I for each source parameter; note the reparametrization of all polar angles to their cosine values. With the exception of M' and z, all parameters are either fixed to, or distributed uniformly, on $\Delta\boldsymbol{\theta}_i$. Although the

TABLE I. Source-parameter extents and distributions for the simple astrophysical-population model in Sec. III A.

θ_i	$\Delta\theta_i$	Distribution
μ'	$\{10\}$	Fixed
M'	$[3, 30] \times 10^5$	Eq. (32)
a/M	$\{0.99\}$	Fixed
p_0/M	$\{11\}$	Fixed
e_0	$(0, 0.2]$	Uniform
$\cos\iota$	$(-1, 0) \cup (0, 1)$	Uniform
Φ_0	$[-\pi, \pi]$	Uniform
γ_0	$[-\pi, \pi]$	Uniform
α_0	$[-\pi, \pi]$	Uniform
$\cos\theta_K$	$(-1, 1)$	Uniform
ϕ_K	$[-\pi, \pi]$	Uniform
$\cos\theta_S$	$(-1, 1)$	Uniform
ϕ_S	$[-\pi, \pi]$	Uniform
z	$(0, 4.5]$	Eq. (33)

initial semilatus rectum is chosen as $p_0 = 11M$ for all sources (to accommodate retrograde inspirals), each is assumed to be observed for $T = 2$ y, i.e., their evolution is extended backward in time from the reference time $t_0 = 0$ if they plunge before $t_0 + T$. Near-spherical inspirals with $e_0 \approx 0$ are excluded from the analysis for technical reasons, as well as near-equatorial ($\cos\iota \approx \pm 1$) or near-polar ($\cos\iota \approx 0$) ones.

Following M1 from [4], the density for M' is almost independent of redshift and thus taken to be [67]

$$p(M') \propto (M')^{-1.3} \mathbf{1}_{\Delta M'}(M'), \qquad (32)$$

where $\mathbf{1}_{\Delta\theta_i}$ denotes the indicator function of $\Delta\theta_i$. Finally, for the redshift itself, we have [68]

$$p(z) \propto \frac{D(z)^2 \dot{\sigma}(z)}{(1+z)^3 H(z)} \mathbf{1}_{\Delta z}(z), \qquad (33)$$

where $\dot{\sigma}(z)$ is the EMRI event rate per unit proper time, per unit comoving volume. The variation of $\dot{\sigma}$ is neglected in our model, since it is highly uncertain to begin with, and is also expected to be nearly constant at low redshifts of $z \lesssim 2$ (where most detectable events occur).

B. Monte Carlo analysis

Let us consider a set of N detectable EMRIs, obtained in practice by applying a minimum-SNR cutoff to a sufficiently large set of $\geq N$ sources that is distributed according to the population model in Sec. III A. Each simulated source is described by a random draw from the joint distribution in Table I; its associated optimal SNR is computed using the AAK–TDI waveform model [35] from Sec. II A and the full inner product (8) with a noise model corresponding to the LISA science requirements. We adopt the standard SNR-threshold value of 20 [17], which results
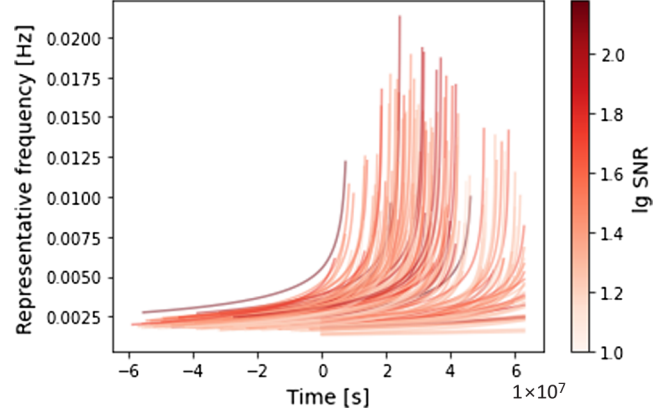

FIG. 5. Time evolution of representative frequency ω_ϕ/π over two years, for 200 detectable sources drawn from the population model in Sec. III A with a SNR threshold of 20.

in a detection efficiency of 0.06. A set of $N = 200$ detectable sources is simulated for this study, as a middle ground between estimates of ~ 1 and $\sim 10^4$ such sources from a range of astrophysical models [4,17]. Their signals are visualized as a population in Fig. 5, where only the time evolution of the representative (dominant) frequency ω_ϕ/π is plotted for each signal.

For this (and any) set of N independent and identically distributed EMRIs, we may define a vector-valued statistic v_N of pairwise overlaps, i.e., v_N is a vector comprising $N_p := N(N-1)/2$ components, each taking a value within the interval $[-1, 1]$. Note the important conceptual distinction between the components of v_N and the sequence of overlaps for N_p pairs of independent identically distributed (IID) sources; the latter is itself a set of IID random variables, but does not relate to the question at hand (except in the trivial case $N = 2$). The sampling distribution of the statistic v_N is intractable—analytically and even numerically for modest N. Nevertheless, a single realization of v_N still provides some insight into the matter of EMRI self-confusion. Figure 6 shows a histogram of the
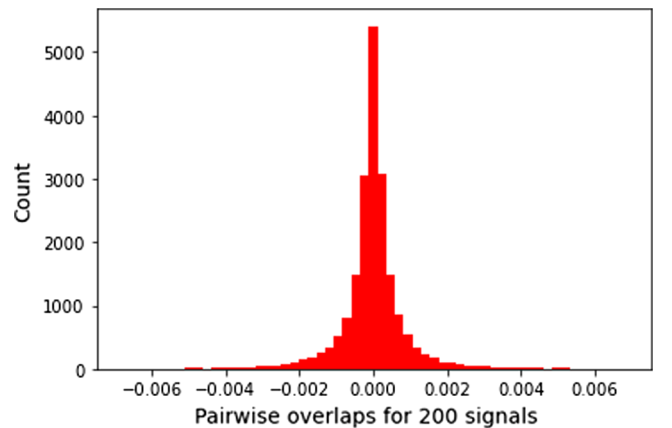

FIG. 6. Histogram of the 19900 pairwise-overlap values for the 200 detectable signals depicted in Fig. 5.

$N_p = 19900$ pairwise-overlap values for our set of 200 sources. (We reiterate that this histogram cannot be construed as a sampling distribution since the components of v_N are not IID.) As might be expected, the set of overlaps is tightly and symmetrically centered about zero, with a "standard deviation" of 0.001 and a maximum absolute value of 0.012. This addresses the first part of our question: in a set of ≈ 200 detectable sources, we expect no two of their signals to have a $\gtrsim 1\%$ resemblance.

The above root-mean-square overlap value of 0.001 may be explained by the following order-of-magnitude estimate. First, idealize each of the EMRI signals visualized in Fig. 5 as comprising only its depicted dominant harmonic mode (2,0,0), with time-evolving frequency $f := \omega_\phi/\pi$. Pairs of signals whose frequency trajectories do not cross have an approximately vanishing overlap, while the overlap value for those that do cross is largely determined by the length of time that both signals are concurrently oscillating at similar frequencies, i.e., some interval centered on the instant of crossing. This duration is given by $\tau \approx |\Delta\dot{f}|^{-1/2}$, where $\Delta\dot{f}$ is the difference in the frequency derivatives of the two signals at the crossing time. At leading order in a PN expansion, $\dot{f}$ for a binary-inspiral signal is $(3/8)(f/T_m)$, where T_m ($\lesssim T = 2$ y) is the time to merger from frequency f. For a typical (crossing) pair of signals in Fig. 5, $\Delta\dot{f}$ then falls within an order of magnitude of $(3/8)(f_c/T)$ (where f_c is the crossing frequency), so let us simply say that $|\Delta\dot{f}| \approx f_c/T$. Finally, the overlap is approximately $\tau/T \approx (f_c T)^{-1/2}$. Since f_c falls between 10^{-3} Hz and 10^{-2} Hz for the vast majority of signal pairs in Fig. 5, we recover an estimated overlap of $\sim 10^{-3}$ as expected.

For the second part of our question, it is useful to form the matrix M_N of pairwise overlaps (including self-overlaps) for a set of N sources. This matrix has unit diagonal elements and off-diagonal elements corresponding to the components of v_N. Thus it is effectively a correlation matrix for the set of sources (viewed as random variables), and various standard matrix norms can provide different summary notions of the overall correlations among the set; e.g., the Frobenius norm $||\cdot||_F$ (the vector norm of the vectorized matrix), or the spectral norm $||\cdot||_S$ (the largest eigenvalue of the matrix). We use such norms to examine the behavior of M_N as sources are incrementally added to a starting population, which we take to be a subset of 30 detectable sources from our full set of 200. More precisely, we consider nested subsets of our full set such that M_N is a submatrix of M_{N+1} for N ranging from 30 to 199.

An adjusted Frobenius norm $(||M_N||_F^2 - N)^{1/2}$ (essentially just $|v_N|$) is plotted in the top panel of Fig. 7, as a function of N from 30 to 200. We bound the observed sequence of norms by two other sequences for hypothetical M_N, where the root mean squares of the off-diagonal terms are held constant at $\{0.9, 1.1\} \times 10^{-3}$. A linear trend is evident, which indicates that the root mean square of

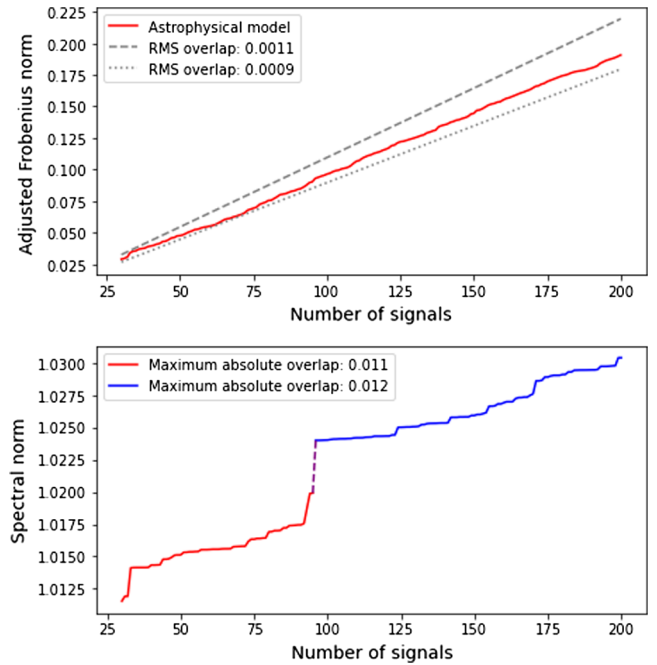

FIG. 7. Matrix norms for the pairwise overlaps among N of the 200 detectable signals in Fig. 5, for $30 \leq N \leq 200$. Top: Adjusted Frobenius norm, indicating that the root-mean-square overlap is approximately constant at ≈ 0.001. Bottom: Spectral norm, illustrating how the overall correlation of the set is driven by individual signals. The largest jump coincides with the addition of a signal that raises the maximum absolute overlap among the set (dashed purple line).

pairwise overlaps remains approximately constant at ≈ 0.001 as the number of sources is raised from 30 to 200. As $N \to \infty$, linear scaling will not hold as the set of signals saturates the (astrophysically restricted) signal space. This can be made more intuitive by examining the spectral norm of M_N (bottom panel of Fig. 7), which would be unity for a completely uncorrelated set of sources. A large jump in $||M_N||_S$ occurs when a signal with high overlap against any of the previous ones is added to the set. This occurrence naturally becomes more frequent as $N \to \infty$ and is what drives the eventual superlinear scaling of overall correlation against N. The value of N at which superlinear scaling becomes an issue (say, leading to a root-mean-square overlap of $\gtrsim 0.1$) remains undetermined due to computational constraints, although we conjecture that such a scenario will not arise for astrophysically relevant values of $N \lesssim 10^4$.

C. Discussion

As context, we now briefly describe how the present work relates to some previous studies on EMRI confusion noise. The spectral density of the effective noise from unresolvable EMRI signals in LISA data was first estimated in [20], in order to quantify its impact on LISA's ability to detect sources of other types—principally, Galactic binaries

and massive-black-hole mergers. Since EMRI event rates as functions of the source parameters were (and remain) poorly understood, the analysis relied on assumptions based on the theoretical models available at the time; the resultant noise curve was then multiplied by a range of overall rate factors (for different classes of compact objects from white dwarfs to black holes) to give approximate bounds on the potential impact of an EMRI confusion background. The main conclusion of [20] is that unresolvable EMRIs will likely be a significant contributor to LISA's noise budget and, for the highest event rates, might actually dominate the total noise at frequencies where detector noise is minimal.

More recently, an updated analysis of EMRI confusion noise for LISA was performed with several key differences [21]: (i) the current LISA baseline noise curve is used, with a trough that is located at higher frequencies; (ii) the assumed scaling of event rates with the central mass is based on newer numerical simulations; and (iii) the EMRI confusion noise curve is computed for a range of astrophysical population models in the literature. While the updated noise curves differ in shape from that found in [20], the study arrives at the same basic conclusion: EMRI confusion noise will be significant for LISA and might be dominant at its most sensitive frequencies if EMRI event rates are on the high end of estimates. In both [20,21], the stochastic process n_{EMRI} that describes the sum of unresolvable signals from an astrophysical EMRI background is treated as approximately Gaussian. This assumption of near-Gaussianity is largely justified by the study in [22]; there, the Edgeworth expansion and large-deviations theory are used to show that the distribution of the cross-correlation between n_{EMRI} and a Galactic-binary or massive-black-hole-merger signal template is indeed very close to Gaussian.

The above papers all deal with sums of N unresolvable EMRI signals and their impact on searches for resolvable signals. In our confusion study, we are instead concerned with the distinguishability of a set of N resolvable signals, which can be described summarily by the root mean square of their pairwise overlaps. These two concepts are of course related, but quite distinct. To draw a connection between them, consider the mean square of the unnormalized cross-correlations (rather than overlaps) between some normalized template and N unresolvable signals with some astrophysical distribution of subthreshold SNRs. This quantity times N is then approximately the variance in the detection SNR of that template due to the associated EMRI confusion noise and thus will be significant if it is $\gtrsim 1$. We do not include such an analysis in this work, although sums of N resolvable signals are explicitly studied in Sec. VA2. There, however, the focus is on how often such sums have a higher cross-correlation with some completely separate template than with any individual template in the sum.

IV. DEGENERACY STUDY

In this study, we examine a single representative signal injection with intrinsic source parameters given by Eq. (6) and delve deeply into the tail structure of its associated posterior distribution. The bulk properties of secondaries in the posterior will no doubt vary for different injections across the parameter space. Instead of attempting to characterize the entire statistical manifold of posterior distributions (which will provide conclusions that are model specific to a greater extent), we focus here on using our sole posterior to build up general strategies for EMRI posterior mapping. This is largely uncharted territory; for example, even visualizing the local posterior around secondaries turns out to be quite counterintuitive (see Secs. IV A 4 and IV A 5). Also, our results indicate that our original aim of localizing and counting secondaries (in order to estimate their coverage of parameter space) is not particularly well posed. Nevertheless, we expect that the qualitative statements we are able to make from this study will, in general, be model independent and thus representative of the EMRI signal space. Various arguments to that end are put forth in Sec. IV B.

A. Mapping analysis

We restrict our analysis to the six intrinsic parameters,

$$\boldsymbol{\theta} = (\lg \mu, \lg M, a/M, p_0/M, e_0, \cos \iota), \qquad (34)$$

and the associated six-dimensional subspaces of Θ and $\mathcal{S}$, for fixed values of the remaining parameters. (Note the reparametrization $(\mu, M, \iota) \to (\lg \mu, \lg M, \cos \iota)$.) This is motivated partly by computational constraints and partly by our initial focus on the degeneracy in frequency evolution—which is most strongly determined by these six parameters. We have verified from a sampling of the full-dimensional (approximate) posterior that local covariances between the sets of intrinsic and extrinsic EMRI parameters are low, at least in the case of the simple long-wavelength response used here. This is also observed in a more recent but unrelated EMRI study [69]. In other words, the conditional posterior for the intrinsic parameters (given fixed extrinsic parameters) is locally not too dissimilar from the marginal posterior. Nonlocally, additional degrees of degeneracy might arise due to the interplay of intrinsic and extrinsic parameters; this possibility is discussed briefly in Sec. IV B 1.

The three phase angles $(\Phi_0, \gamma_0, \alpha_0)$ are also excluded from consideration in our mapping analysis. With only the four strong harmonic modes from Eq. (7), the stripped-down AAK model loses all sidebands due to Lense-Thirring precession and thus its ability to constrain α_0. For simplicity, we set the other two phase angles to zero in the injection and all templates, such that the phasing in all signals lines up at time $t_0 = 0$ (the initial time in the

degeneracy study). This alignment essentially forces any crossing of mode phasing between two signals to occur at the earliest time, which generally maximizes their overlap over (Φ_0, γ_0)—even if the signals are nondegenerate since the frequencies for the four modes (7) take on their minimal values at that time. An inclusion of (Φ_0, γ_0) in the analysis would likely reveal a small degree of additional degeneracy in those parameter directions, but this would not significantly alter the intrinsic degeneracy observed here (as initial phasing does not affect frequency evolution in this simple model).

We explore the signal space over a variety of hyper-rectangular regions in (the six-dimensional subspace of) Θ, centered on $\boldsymbol{\theta}_{\mathrm{inj}}$ from Eq. (6). To be explicit here, the injection parameters in the form of Eq. (34) are

$$\boldsymbol{\theta}_{\mathrm{inj}} = (1, 6, 0.5, 9.5, 0.2, 0.866), \qquad (35)$$

with the signal renormalized to an optimal SNR of 20. Results are presented for three regions in particular: a starting region around the primary posterior peak, another with 10 times its extent in each parameter direction, and a third with ~ 100 times its extent. In the latter two regions, we obtain a set of representative locations for secondary peaks in the overlap surface; from the line of reasoning given in Sec. II C, these correspond almost exactly to secondaries in the actual posterior for the full likelihood (with a flat or diffuse prior). Two high-overlap secondaries are further singled out as case studies, where we directly map out the (approximate) posterior in their vicinity, instead of the overlap surface. A summary plot that depicts the relative scales and locations for all of these analyses is provided at the end of Sec. IV A.

Throughout the degeneracy study, we employ a combination of different implementations for the two main classes of stochastic sampling algorithms used in GW data analysis: nested sampling (as implemented in `PolyChord` [57] and `Dynesty` [58]) and parallel-tempering MCMC (as implemented in `PTMCMCSampler` [70]). The primary motivation for this redundancy is to provide a cross-algorithm and cross-implementation validation of our sampling results, where feasible. For the posterior sampling (using L_{app}) in Secs. IV A 1, IV A 4, and IV A 5, we run the samplers to the point of convergence—typically determined by built-in stopping criteria, but also easily verified through cross-sampler checks. Convergence to the target distribution is less relevant for the exploratory searches in Secs. IV A 2 and IV A 3. There we simply seek to obtain $\sim 10^5 - 10^6$ independent samples from L_{expl}, a large fraction of which correspond to signals with $\Omega(h_{\mathrm{inj}}, \cdot) > 0.5$ [since that is the threshold used in Eq. (27)]. For the most extensive searches in Sec. IV A 3, this can require $\gtrsim 10^9$ likelihood evaluations.

1. Starting region: The posterior bulk

Our starting region $\mathcal{R}_0$ is chosen to encompass the Gaussian-analogous 2-σ contours of the primary posterior peak, with some room. More precisely, we consider a "prior" with probability density proportional to the indicator function on the Cartesian product of intervals,

$$\mathcal{R}_0 := \prod_i [\boldsymbol{\theta}_{\mathrm{inj},i} - \delta\boldsymbol{\theta}_i/2, \boldsymbol{\theta}_{\mathrm{inj},i} + \delta\boldsymbol{\theta}_i/2], \qquad (36)$$

where the half-extents $\delta\boldsymbol{\theta}/2$ are 2–3 times larger than the sample standard deviations. (This is then only a prior in practice but not in principle, as it is defined *post hoc*.) At an SNR of 20, suitable values for the prior extents are

$$\delta\boldsymbol{\theta} = (0.5, 1, 3, 15, 0.3, 5) \times 10^{-3}. \qquad (37)$$

The Euclidean volume enclosed within $\mathcal{R}_0$ is a minuscule fraction of the six-dimensional parameter subspace: $\lesssim 10^{-18}$ since we have $\delta\boldsymbol{\theta}_i/\Delta\boldsymbol{\theta}_i \lesssim 10^{-3}$ for global extents $\Delta\boldsymbol{\theta}_i$ corresponding to the range of LISA-relevant signals.

The sampling of L_{app} over the region $\mathcal{R}_0$ is extremely straightforward, as the locality condition in Eq. (16) ensures that L_{app} is near-Gaussian. A direct visual examination of the conditional densities $L_{\mathrm{app}}(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{j \neq i} = \boldsymbol{\theta}_{\mathrm{inj},j})$ reveals that the non-Gaussianity takes the form of a flatter peak and heavier tails, which is largely due to the usage of $(\cdot | \cdot)$ rather than $\langle \cdot | \cdot \rangle$; we refer the reader again to Fig. 3 for intuition about why this is the case. The traditional way of visualizing a sampled $(d > 2)$-dimensional posterior distribution is simply to examine plots of its $d(d-1)/2$ bivariate marginal distributions (and the d univariate marginals). In Fig. 8, we instead characterize the sampled posterior using four visual indicators: (i) the maximum *a posteriori* (MAP) estimate; (ii) level sets of the marginal posterior densities p_{ij}, defined analogously to the 1-σ level set for a bivariate Gaussian density:

$$\left\{ (\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) : p_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = e^{-1/2} \max_{(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)} p_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \right\}; \qquad (38)$$

(iii) the sample mean and covariance; and (iv) a level set of the joint posterior density p, defined analogously to the 1-σ level set for a six-dimensional Gaussian density. This last indicator is itself five-dimensional, so it is more conveniently represented by the projections of posterior samples within the corresponding superlevel set

$$\left\{ \boldsymbol{\theta} : p(\boldsymbol{\theta}) > e^{-1/2} \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) \right\}. \qquad (39)$$

This whole exercise may seem somewhat trivial, as all four visual indicators will coincide exactly when p describes a multivariate Gaussian—the MAP estimate with the sample mean, the marginal level sets with the 1-σ
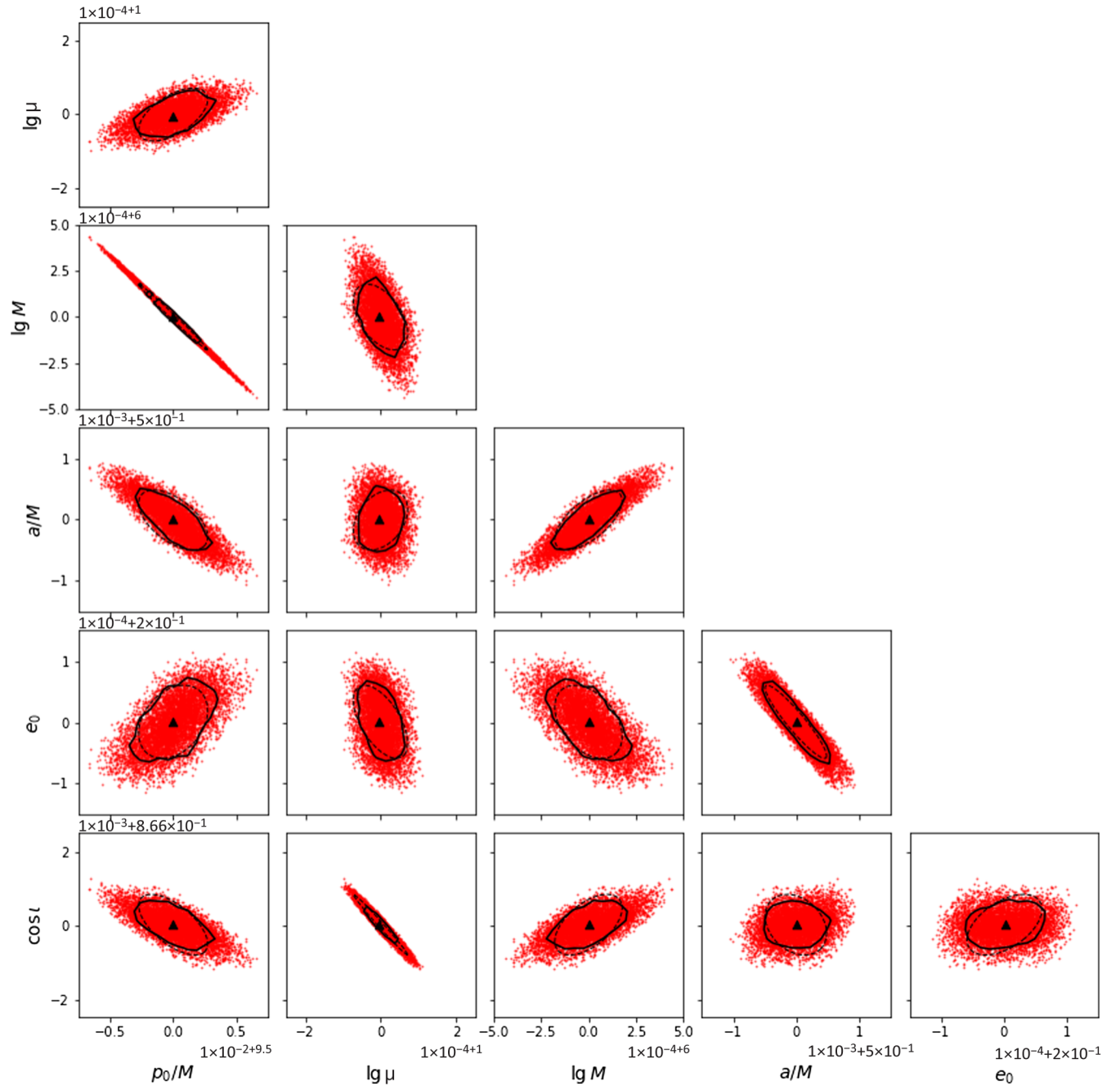
FIG. 8. Various visual indicators of posterior distribution over $\mathcal{R}_0$ (around the posterior bulk), computed using the approximate likelihood and a flat hyperrectangular prior. Black triangle: MAP estimate (this agrees with the injection parameters at the prior centroid). Solid black curves: level sets (38) of the marginal posterior densities p_{ij} (1-σ level for a bivariate Gaussian distribution). Dashed black ellipses: 1-σ ellipses corresponding to projections of the sample mean and covariance. Red points: projections of posterior samples in the superlevel set (39) of the joint posterior density p (1-σ level for a six-dimensional Gaussian distribution). All four visual indicators will coincide exactly for a multivariate Gaussian distribution.

sample covariance ellipses, and the convex hulls of the projected superlevel set with said ellipses. The indicators do, however, provide a useful way of visualizing high-dimensional non-Gaussian posteriors, especially when the deviation from Gaussianity is severe (see Secs. IV A 4 and IV A 5). For the posterior over $\mathcal{R}_0$, the incongruence of the final indicator clearly highlights the non-Gaussianity that

would not be immediately apparent from examining traditional contour plots for the marginal distributions (essentially, the second indicator). Nevertheless, the posterior bulk appears to be relatively well behaved, i.e., unimodal as expected. We verify this by applying our clustering algorithm to the posterior samples in the superlevel set (the red points in Fig. 8). All samples are not just connected

to the MAP point, but strictly connected with a maximal predistance of zero.

2. Starting region ×10⁶

We now consider the Cartesian product of intervals

$$\mathcal{R}_1 := \prod_i [\boldsymbol{\theta}_{\text{inj},i} - 5\delta\theta_i, \boldsymbol{\theta}_{\text{inj},i} + 5\delta\theta_i], \qquad (40)$$

which is a six-dimensional hyper-rectangle with 10^6 times the Euclidean volume of $\mathcal{R}_0$. It is useful to examine the sampling of posteriors over this region for the various likelihood functions in Sec. II C. Even in this intermediate region, large-scale SNR gradients are already manifest in L_{app}; this shifts the posterior bulk toward certain edges of the prior (and away from the injection parameters, even if the sampler manages to "find" that point along the way). When sampling from L_{match} over $\mathcal{R}_1$, the posterior bulk does remain tightly centered on the injection parameters, and the samplers we use are able to locate it without any fine-tuning. However, no information about any secondaries in the region can be gleaned from the sampled posterior, even for L_{match} at higher annealing temperatures (as discussed in Sec. II C).

By sampling from the exploratory likelihood L_{expl}, we obtain a large set of points in $\mathcal{R}_1$ whose associated signals have a high overlap against the injection. We consider only the subset of points with approximate overlaps of >0.5 for clustering. The approximate overlap typically overestimates full overlaps that are >0.1 by a factor of $\lesssim 3$, such that these points actually have full overlaps of $\gtrsim 0.25$ (which is still reasonably considered "non-negligible"). Since we know the cluster content in $\mathcal{R}_0$, any points that fall in the previously analyzed region are also removed; this is done in Sec. IV A 3 as well, where $\mathcal{R}_1$ is excised instead. We find it convenient to perform such an excision after sampling, rather than through a modification of the prior. The resultant set of $N \gtrsim 10^5$ points is generally still too large for our clustering algorithm to handle in an acceptable time frame—although the computational cost scales as $\mathcal{O}(N)$, each iteration involves template generation and thus is more expensive than a typical step in a standard algorithm. We further distill the set to exactly $N = 10^4$ points, by sorting it in order of overlap and then downsampling the sequence (such that the distribution of the data is loosely preserved).

In the region $\mathcal{R}_1$ (minus $\mathcal{R}_0$), the clustering algorithm identifies 30 secondary nodes, with full overlap values ranging from 0.45 to 0.72. The locations of these nodes, projected onto each of the 15 parameter-pair planes, are plotted in Fig. 9. It is not particularly useful to present the actual clusters associated with the nodes, due to visual congestion. To give a flavor of the overlap surface in this region (and some assurance that the nodes are indeed

adequate representations of local overlap maxima), we instead consider the six pairwise connections ℓ among the injection parameters and the three nodes in $\mathcal{R}_1$ with the highest overlap values. These connections are shown in the inset of Fig. 9, with an increased resolution of $l = 50$ (see Sec. II D) for visualization purposes.

To verify the robustness of our results against the stochastic error in sampling, we repeat the analysis on the output of a second sampling run and compare the first set (I) of 30 nodes to the new set (II). Set II turns out to have the exact same count; surprisingly, however, the locations of its nodes do not match up well with those in set I. We seek to pair up nodes in the two sets and to determine which pairs are more "distinct." This is done by computing the distance matrix of pairwise predistances between two sets of points—not to be confused with the predistance matrix from the clustering algorithm, which is essentially a (partial) distance matrix between a single set of points and itself. With the distance matrix between sets I and II in hand (Fig. 10; left panel), it is straightforward to identify pairs of nodes that are in one-to-one correspondence with each other. More precisely, if node a from set I has node b as its closest node in set II, then we demand that node a is also the closest node in set I to node b. There are 19 such pairs (which we index in decreasing order of set-I overlap, such that the set-I point in pair 0 has the highest overlap value in set I). The distance matrix may then be trimmed accordingly and sorted by pair index, to visually indicate the interconnectivity among the reduced set of node pairs (Fig. 10; right panel).

The locations of the 19 node pairs and intrapair connections are plotted in the top panel of Fig. 11, after projection onto the mass-mass plane. Another surprise at this stage is the variety of orientations and length scales on display; all of the pairs are strictly or nearly strictly connected (see diagonal elements of reduced distance matrix in Fig. 10), indicating that the local maxima they represent also take on an assortment of shapes and sizes. While these pairs clearly correspond to secondaries in accordance with our search criteria, we do not find any that can be cleanly localized—i.e., where the posterior peak is contained exclusively in a topologically connected subset of parameter space. We will revisit the problem of trying to localize secondaries in Secs. IV A 4 and IV A 5. For now, we examine the connections for a couple of example node pairs, extended outward until the (approximate) overlap $\Omega(h_{\text{inj}}, \cdot)$ is ≈ 0. These extended connections are shown in the middle and bottom panels of Fig. 11 for both the approximate and full overlaps, and they provide some intuition (albeit cross sectional) as to the structure of the overlap surface at and around secondaries.

3. Starting region ×10¹² (and beyond)

A similar sampling and clustering analysis is performed for the Cartesian product of intervals,
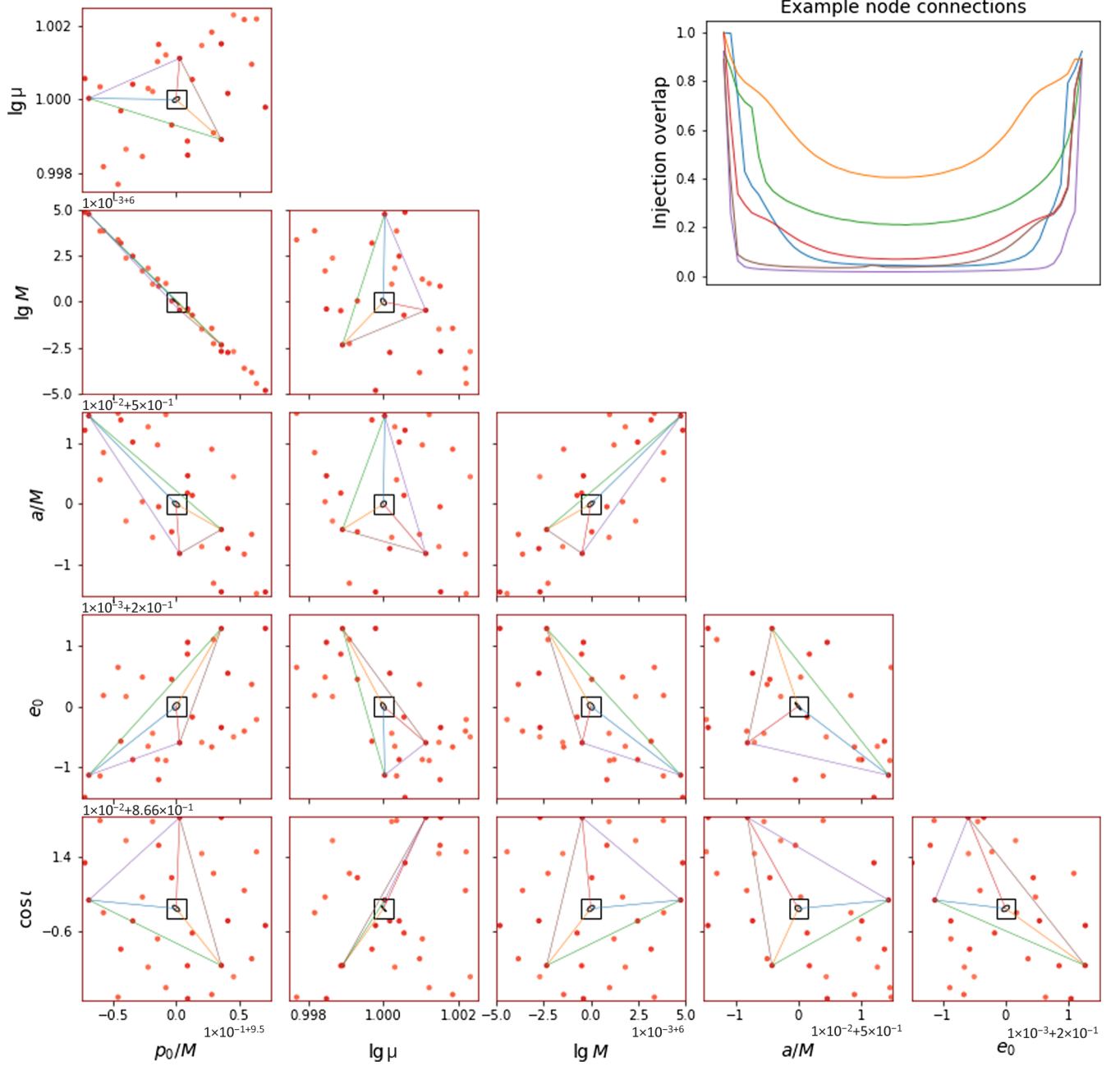
FIG. 9. Locations of 30 secondary nodes (red points) identified in the intermediate region $\mathcal{R}_1 \backslash \mathcal{R}_0$. Relative overlap values at these nodes are indicated by color saturation. The black square and ellipse in each panel corresponds to the plot range and dashed ellipse in each panel of Fig. 8. Inset: the six pairwise connections among the injection parameters and the three highest-overlap nodes in $\mathcal{R}_1$, i.e., the overlap value against the injection along the corresponding colored lines in each panel.

$$\mathcal{R}_2 := \prod_i [\boldsymbol{\theta}_{\mathrm{inj},i} - 50\,\delta\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{inj},i} + 50\,\delta\boldsymbol{\theta}_i]. \qquad (41)$$

Note that $\mathcal{R}_2$ is truncated by the upper bound $\cos\iota = 1$, and thus the enclosed volume is $\lesssim 10^{12}$ times that of $\mathcal{R}_0$. In this large region, the direct sampling of L_{match} now presents difficulties for the samplers; they are unable to locate the posterior bulk either at termination, or after $\sim 10^9$ likelihood evaluations. This result likely indicates the rough scale at

which posterior sampling can become challenging without prior localization of the source parameters (even if large-scale SNR gradients are accounted for)—although it may be possible to improve the performance of specific samplers through further tuning.

As in Sec. IV A 2, the previous analysis region (now $\mathcal{R}_1$) is excised after sampling to avoid double-counting, and the data set for clustering is trimmed to 10^4 points with approximate overlaps of >0.5. A total of 675 secondary
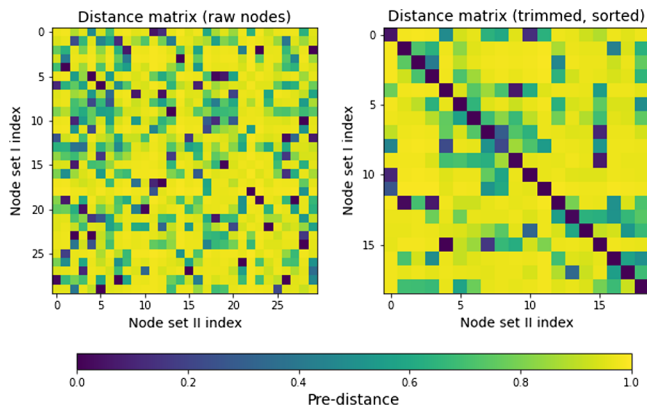
FIG. 10. Distance matrix between the two sets of secondary nodes identified in $\mathcal{R}_1$. Left: raw sets with 30 nodes each. Right: trimmed and sorted sets with 19 node pairs.

nodes are identified in the region $\mathcal{R}_2 \backslash \mathcal{R}_1$, with full overlap values ranging from 0.23 to 0.76. Their projected locations are shown in Fig. 12. The broad correlation structure of secondaries in the large region—as visually indicated by the plane convex hull for each projection of the set—generally does not resemble that observed in Fig. 9, or the local structure in Fig. 8 (one exception being the strong correlation between $\lg M$ and p_0/M). Another notable feature in Fig. 12 is that the distribution of nodes appears somewhat "patchy," although this may be attributable to imperfect sampling rather than the actual absence of secondaries in some of the empty regions.

With a large set of secondary nodes from the combined analysis in $\mathcal{R}_{1,2}$, it is natural to examine the relationship between overlap (against the injection) and distance (from the injection parameters) for any discernible trends. Instead of using the Euclidean distance in parameter space Θ or the (computationally intractable) geodesic distance in signal space $\mathcal{S}$, we consider the distance with respect to the pullback metric at the injection parameters. This is given componentwise by Eq. (15), where the Fisher information $\mathcal{I}(\boldsymbol{\theta}_{\mathrm{inj}})$ is in turn approximated by the sample covariance of the posterior bulk in Sec. IV A 1. A plot of overlap against the injection-metric distance is shown in the inset of Fig. 12, for the 30 set-I nodes in $\mathcal{R}_1 \backslash \mathcal{R}_0$ and the 675 nodes in $\mathcal{R}_2 \backslash \mathcal{R}_1$. There is a slight negative correlation between overlap and distance, as well as between number density and distance; however, these trends are not particularly pronounced (and even less so if the Euclidean distance is used).

The results from the analysis of $\mathcal{R}_2$ are admittedly less reliable than those for $\mathcal{R}_1$ since we do not (and cannot) increase sampling resolution by anywhere close to the factor of 10^6 that is required to compensate for the larger volume. Even if we were able to, the size of the data set for
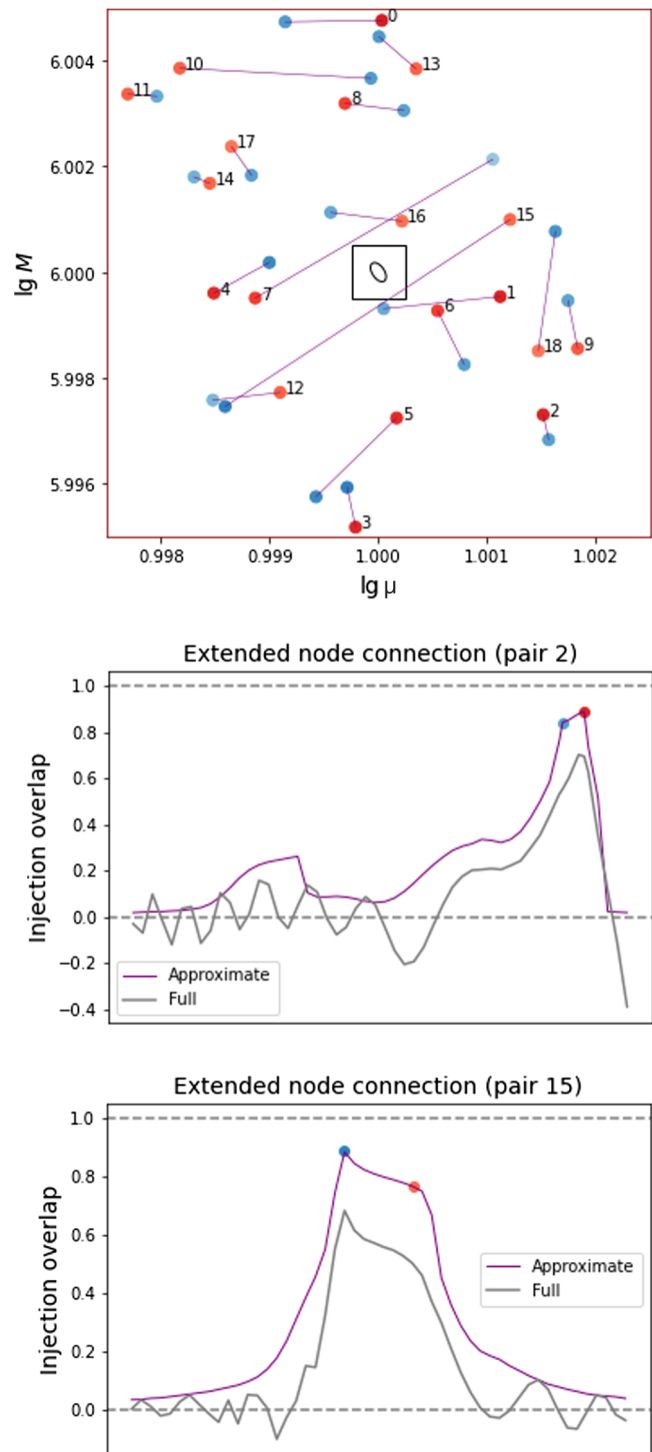

FIG. 11. Trimmed and sorted pairs from sets I (red) and II (blue) of secondary nodes identified in $\mathcal{R}_1$. Relative overlap values at these nodes are indicated by color saturation. Top: projection of node-pair locations onto the $(\lg \mu, \lg M)$ plane (corresponding to Fig. 9, second row, second panel). Middle: extended connection for pair 2, which has the third-highest set-I overlap value. Bottom: extended connection for pair 15, which has the largest extent in the mass-mass plane.
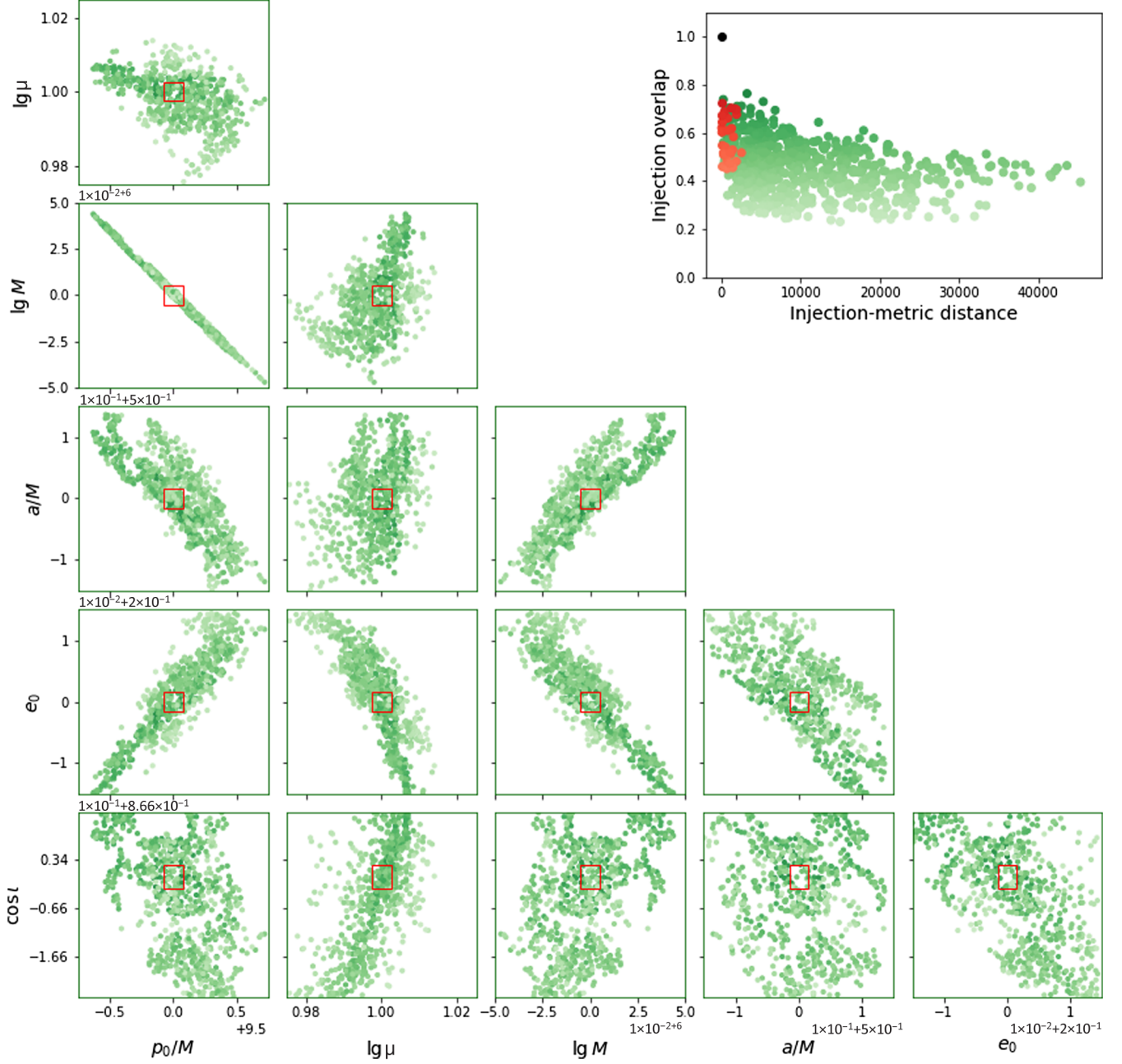
FIG. 12.   Locations of 675 secondary nodes (green points) identified in the large region $\mathcal{R}_2 \backslash \mathcal{R}_1$. Relative overlap values at these nodes are indicated by color saturation. The red rectangle in each panel corresponds to the plot range in each panel of Fig. 9. Inset: full overlap values for these nodes, plotted against their distance from the injection parameters (black point) with respect to the pullback (Fisher) metric there. The set-I nodes shown in Fig. 9 (red points) are also included in this plot.

clustering would also have to be raised accordingly for high-resolution sampling to be meaningful. As a heuristic assessment of sampling error and its increased impact in extended regions, we briefly discuss our analysis of a final region: the Cartesian product of intervals,

$$\mathcal{R}_3 := \prod_i [\boldsymbol{\theta}_{\mathrm{inj},i} - 100\, \delta\boldsymbol{\theta}_i,\, \boldsymbol{\theta}_{\mathrm{inj},i} + 100\, \delta\boldsymbol{\theta}_i], \qquad (42)$$

whose results we do not present explicitly here (although we do use one of the identified secondary nodes as a case study in Sec. IV A 4). Two identically initialized sampling runs are conducted for L_{expl} over this region, each with $\gtrsim 10^6$ independent samples drawn after $\gtrsim 10^9$ likelihood evaluations. It is immediately clear from their density plots that the two sets of posterior samples are concentrated in very different subregions and thus that clustering will give inconsistent results as well. This also points to a systematic

undercounting of clusters, which worsens as the analysis region is expanded. Nevertheless, the convex hulls of the raw posterior samples from the two runs are similar in both shape and volume ($\sim 10^{-6}$ relative to $\mathcal{R}_3$), which gives a more reliable indication of where secondaries are absent, as well as a very conservative upper bound on their coverage of parameter space.

4. Case study A

We now examine in detail the actual posterior surface over the vicinity of an example secondary node in $\mathcal{R}_3$—specifically, the node with the highest overlap value found outside $\mathcal{R}_2$. Relative to $\boldsymbol{\theta}_{\text{inj}}$ from Eq. (35), this node (labeled A) is given by

$$\boldsymbol{\theta}_A - \boldsymbol{\theta}_{\text{inj}} = (-30, 13, 228, -53, 13, -482) \times 10^{-3}. \quad (43)$$

Its associated signal has an approximate overlap of 0.81 against the injection, but a full overlap value of only 0.39. This is nevertheless a significant example of degeneracy, especially considering its large injection-metric distance from the injection parameters ($\gtrsim 2 \times 10^5$), and it is not inconceivable for an uninformed search algorithm to flag the secondary represented by node A as a separate candidate source (with a detection SNR of $\approx 0.4\rho_{\text{inj}}$).

The approximate likelihood L_{app} is sampled over a flat-prior region $\mathcal{R}_A$ centered on $\boldsymbol{\theta}_A$, with the same extents as $\mathcal{R}_0$. A traditional inspection of the marginal posterior densities, either of their Gaussian-analogous 1-σ contours (the solid black curves in Fig. 13) or their density plots (not shown), presents nothing particularly out of the ordinary; the posterior might be construed at this stage as a moderately deformed Gaussian distribution, centered on a point that is close to node A. We find, however, that any attempt to localize this (or any other) secondary by simply enlarging and recentering the prior region—"chasing" the secondary, as we not-so-affectionately term it—turns out to be futile. Hints as to why this is the case are provided in Fig. 13 by the other visual indicators introduced in Sec. IV A 1. While the sample mean and covariance are at least somewhat consistent with the marginal 1-σ contours, the projections of the superlevel set (points within the Gaussian-analogous 1-σ contour for the joint posterior) are not. The MAP estimate also falls in a location that defies reasonable prediction.

Our findings seem to indicate that secondary A cannot be localized; i.e., there does not exist a topologically connected subset of parameter space $U \ni \boldsymbol{\theta}_A$ with some neighborhood $V \supset U$ such that the posterior probability is approximately zero over $V \backslash U$. We do not make this argument rigorous, but offer an anecdotally supported conjecture that it holds generally for all secondaries, as well as a partial explanation for the phenomenon (see Sec. IV B 2). Another important qualitative distinction between the primary posterior peak and a typical secondary

might be the presence of fine structure. The posterior surface over the region $\mathcal{R}_A$ is not quite unimodal and appears to comprise multiple local maxima that are strongly, but not strictly, connected to one another. This is inferred from the fact that the maximal predistance between node A and all posterior samples in the superlevel set is ≈ 0.1. Node A itself is at least a stationary point of the posterior, from visual examination of the conditional densities $L_{\text{app}}(\theta_i | \boldsymbol{\theta}_{j \neq i} = \theta_{A,j})$, but we are unable to rule out the saddle point case due to severe instability in the numerical derivatives.

As we enlarge our field of view of the secondary A by a factor of eight (i.e., sample from L_{app} over a larger flat-prior region $\mathcal{R}'_A$ with extents $8\,\delta\boldsymbol{\theta}$), we encompass additional local maxima that are now disconnected to one another. The existence and proximity of these other secondaries is likely the main impediment to the localization of secondary A; indeed, their presence cannot even be discerned from the traditional approach of considering only the marginal posteriors (see solid black curves in Fig. 14). To resolve and visualize the multiple secondaries in this region, we again apply our clustering algorithm to all samples in the superlevel set of the posterior over $\mathcal{R}'_A$. Five nodes (including node A) are identified, and their associated clusters are shown grouped by color in Fig. 14. These clusters do not exhibit a high degree of order in their spatial structure and are also underpopulated due to increased interconnectivity, i.e., many points in the superlevel set are connected to more than one node, and thus do not appear in any cluster by definition. Nevertheless, the clustering analysis provides clear evidence of additional maxima in the region (see inset of Fig. 14) and demonstrates how secondaries can congeal into larger ones especially when viewed after projection (marginalization).

5. Case study B

For our second case study, we select one of the highest-overlap nodes found in the region $\mathcal{R}_2 \backslash \mathcal{R}_1$. Relative to $\boldsymbol{\theta}_{\text{inj}}$ from Eq. (35), this node (labeled B) is given by

$$\boldsymbol{\theta}_B - \boldsymbol{\theta}_{\text{inj}} = (3, -12, -54, 170, 1, 64) \times 10^{-3}. \quad (44)$$

It has a full (approximate) overlap value of 0.72 (0.89), and its injection-metric distance from the injection parameters is ≈ 8000. The analysis in Sec. IV A 4 is repeated around $\boldsymbol{\theta}_B$—first in a small region $\mathcal{R}_B$ with the same size as $\mathcal{R}_0$ and $\mathcal{R}_A$ (Fig. 15), then in a larger region $\mathcal{R}'_B$ with the same size as $\mathcal{R}'_A$ (Fig. 16).

As its overlap value would indicate, the signal template at node B strongly resembles the signal injection. The two signals can no longer be distinguished in a plot analogous to the inset of Fig. 13, and their dominant (2,0,0) modes remain "in phase" over the full analysis duration (see inset of Fig. 15). This increased resemblance over node A might partly explain why the visual indicators in Fig. 15 are
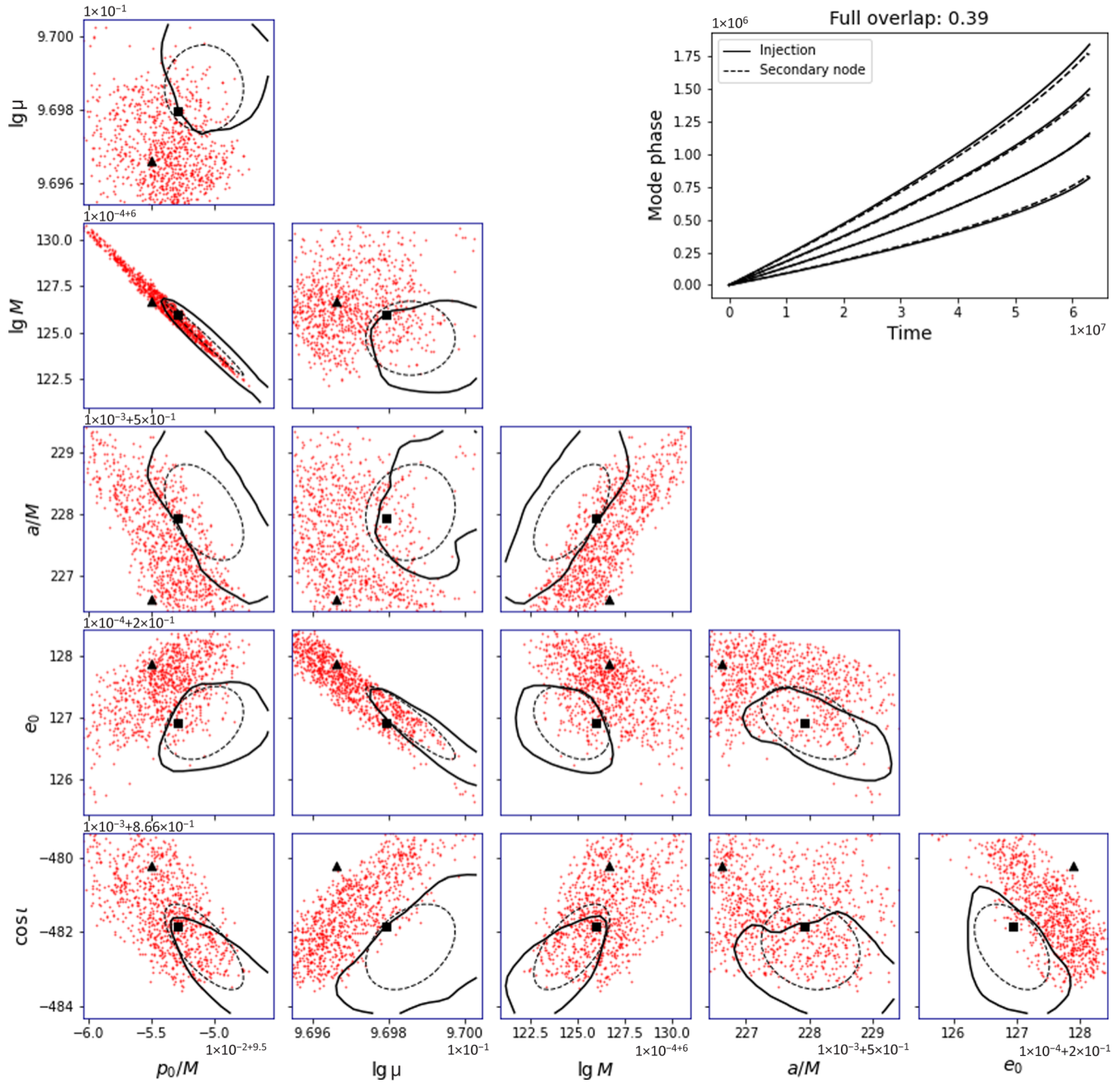
FIG. 13. Various visual indicators of posterior distribution over $\mathcal{R}_A$ (a small region around secondary node A), computed using the approximate likelihood and a flat hyper-rectangular prior. The prior extents are the same as in Fig. 8. Black square: node A (the prior centroid). Black triangle: MAP estimate. Solid black curves: level sets (38) of the marginal posterior densities p_{ij} (1-σ level for a bivariate Gaussian distribution). Dashed black ellipses: 1-σ ellipses corresponding to projections of the sample mean and covariance. Red points: projections of posterior samples in the superlevel set (39) of the joint posterior density p (1-σ level for a six-dimensional Gaussian distribution). All four visual indicators will coincide exactly for a multivariate Gaussian distribution. Inset: comparison of phase trajectories $\varphi_j(t)$ between the signal injection and the signal template at node A. Only the dominant (2,0,0) modes (third from top) are in phase for a substantial fraction of the analysis duration.

slightly less inconsistent with one another than in Fig. 13. However, secondary B is still distinctly non-Gaussian, and like secondary A it resists all attempts at localization. Fine structure is present here as well—for all samples in the superlevel set of the posterior over $\mathcal{R}_B$, the maximal

predistance from node B is ≈ 0.1. For the posterior over $\mathcal{R}'_B$, we see once again from Fig. 16 that the marginal densities show only a single contiguous secondary peak without any hints of internal structure. Clustering reveals seven nodes (including node B) in this region; more
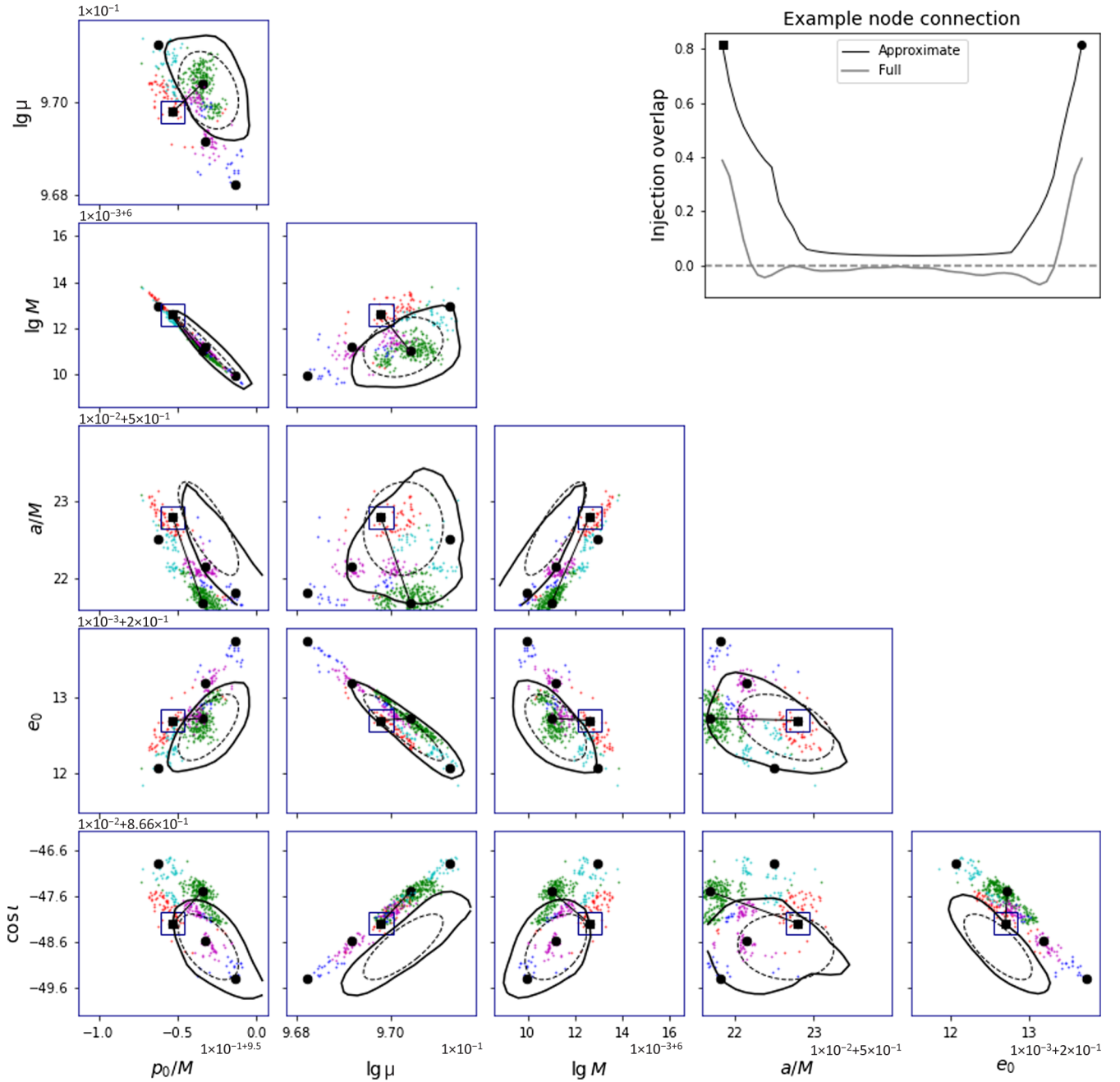
FIG. 14. Various visual indicators of posterior distribution over $\mathcal{R}'_A$ (a larger region around secondary node A), computed using the approximate likelihood and a flat hyper-rectangular prior. The blue square in each panel corresponds to the plot range in each panel of Fig. 13. Black square: node A (the prior centroid). Black points: representative nodes for four other secondaries identified in $\mathcal{R}'_A$. Solid black curves: level sets (38) of the marginal posterior densities p_{ij}. Dashed black ellipses: 1-σ ellipses corresponding to projections of the sample mean and covariance. Colored points: clusters of posterior samples in the superlevel set (39) of the joint posterior density p. These are the subsets of samples that are connected to each node and not connected to any of the other nodes. Inset: connection between node A and the highest-overlap node in $\mathcal{R}'_A$, i.e., the overlap value against the injection along the black line in each panel.

interestingly, it is now possible to identify clear patterns in the shapes, sizes, and relative locations of the associated clusters. The general existence of ordered structure in the posterior surface is expected *a priori*, and its verification in this case study also serves to demonstrate the efficacy of our clustering algorithm.

6. Summary plot

The relative scales of the various subanalyses in Sec. IVA are depicted by the summary plot of Fig. 17, where the starting region $\mathcal{R}_0$ (along with $\mathcal{R}_{A,B}$) is unresolvable and represented as a point. We draw particular attention to the secondary associated with pair 15 from
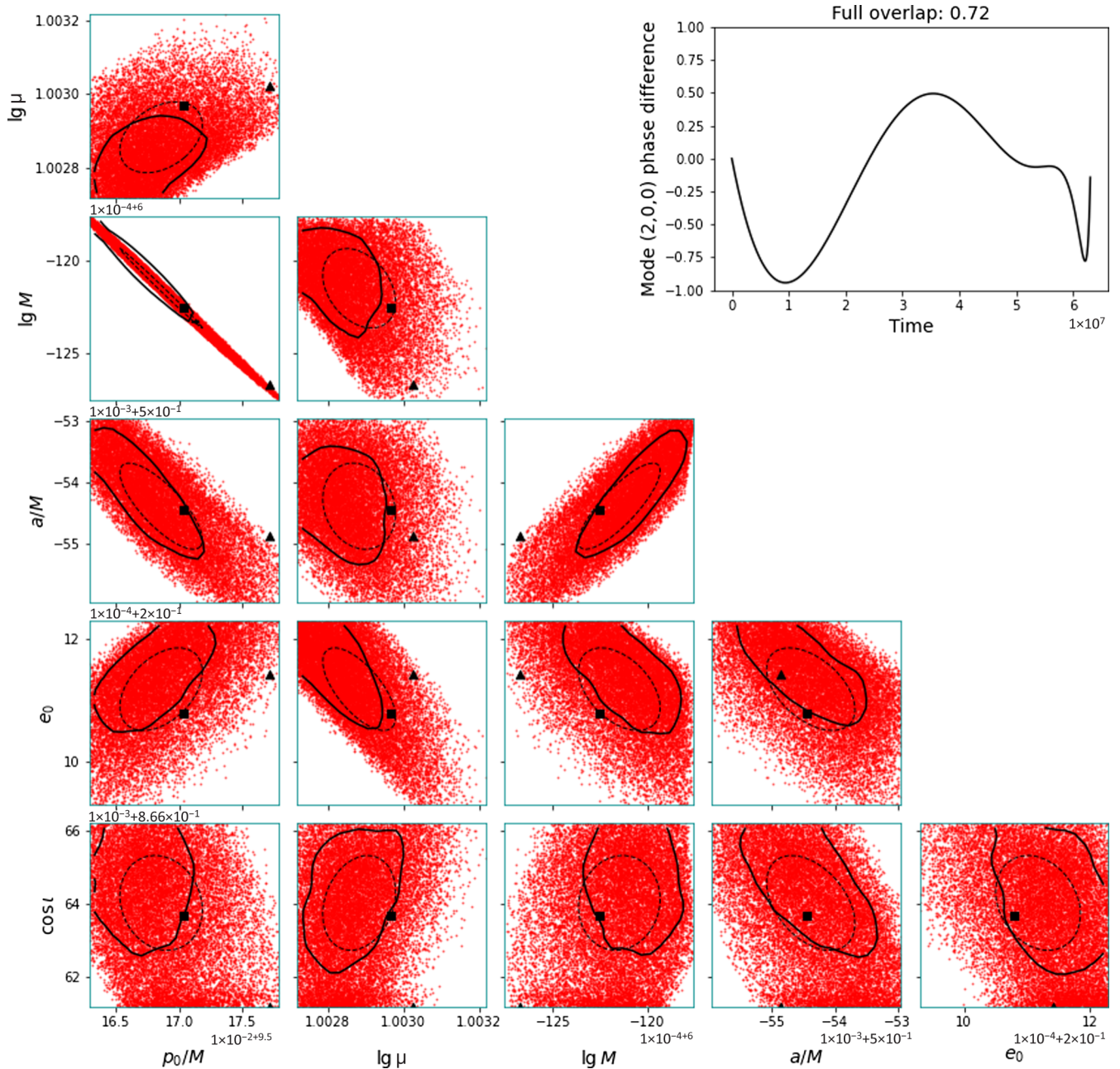
FIG. 15.   Various visual indicators of posterior distribution over $\mathcal{R}_B$ (a small region around secondary node B), computed using the approximate likelihood and a flat hyper-rectangular prior. The prior extents are the same as in Figs. 8 and 13. Black square: node B (the prior centroid). Black triangle: MAP estimate. Solid black curves: level sets (38) of the marginal posterior densities p_{ij} (1-σ level for a bivariate Gaussian distribution). Dashed black ellipses: 1-σ ellipses corresponding to projections of the sample mean and covariance. Red points: projections of posterior samples in the superlevel set (39) of the joint posterior density p (1-σ level for a six-dimensional Gaussian distribution). All four visual indicators will coincide exactly for a multivariate Gaussian distribution. Inset: phase difference between the dominant (2,0,0) modes of the signal injection and the signal template at node B, showing that the two signals are approximately in phase over the full analysis duration.

Sec. IV A 2 (see bottom panel of Fig. 11). This is the pair of nodes in $\mathcal{R}_1$ with the largest separation—overtly in the mass parameters, as shown in the top panel of Fig. 11, but in many other parameter pairs as well. Its extended connection indicates at least a ridgelike region of high posterior probability relative to the immediate neighborhood and

possibly one with non-negligible width (which cannot be determined from the connection alone). As seen in Fig. 17, this secondary actually extends well beyond $\mathcal{R}_1$ and up to a significant fraction of $\mathcal{R}_2$.

Our findings in Sec. IV A are clear evidence of highly nontrivial degeneracy in the EMRI signal space—this
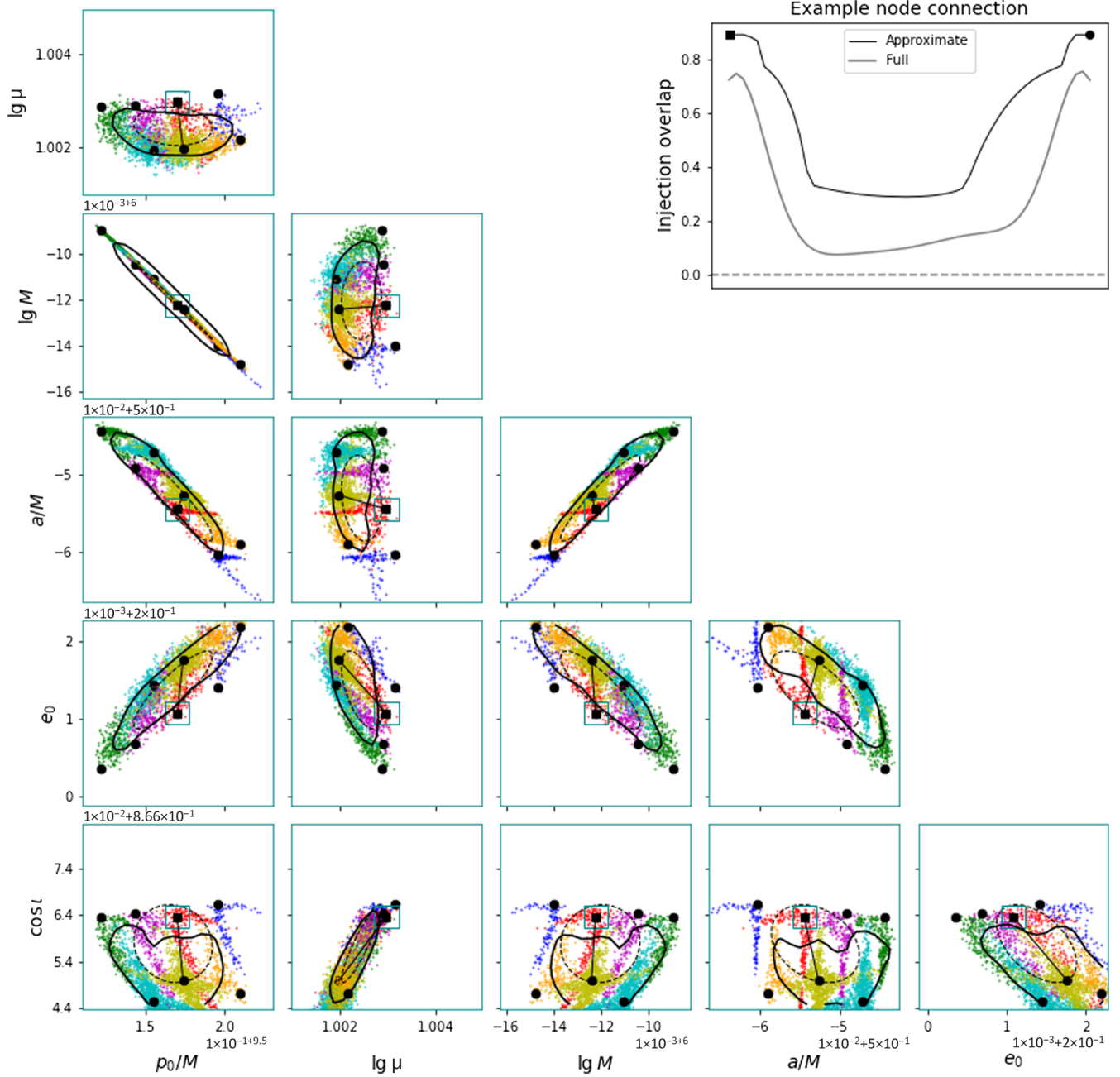
FIG. 16. Various visual indicators of posterior distribution over $\mathcal{R}'_B$ (a larger region around secondary node B), computed using the approximate likelihood and a flat hyper-rectangular prior. The green square in each panel corresponds to the plot range in each panel of Fig. 15. Black square: node B (the prior centroid). Black points: representative nodes for six other secondaries identified in $\mathcal{R}'_B$. Solid black curves: level sets (38) of the marginal posterior densities p_{ij}. Dashed black ellipses: 1-σ ellipses corresponding to projections of the sample mean and covariance. Colored points: clusters of posterior samples in the superlevel set (39) of the joint posterior density p. These are the subsets of samples that are connected to each node and not connected to any of the other nodes. Inset: connection between node B and the highest-overlap node in $\mathcal{R}'_B$, i.e., the overlap value against the injection along the black line in each panel.

manifests as the presence of many posterior secondaries with different shapes, sizes, and degrees of connectivity to one another. While the tools we have brought to bear in the study can provide insight into specific occurrences of degeneracy, the cost and reliability of high-resolution sampling (and the follow-up clustering of data) remains

the impediment to a fully global analysis. There is, nevertheless, ample room for improvement, and the present work provides the foundation for extended studies in the future. Such studies will be especially relevant after science-adequate waveform models become available (see additional comments in Secs. IV B 3 and VII).
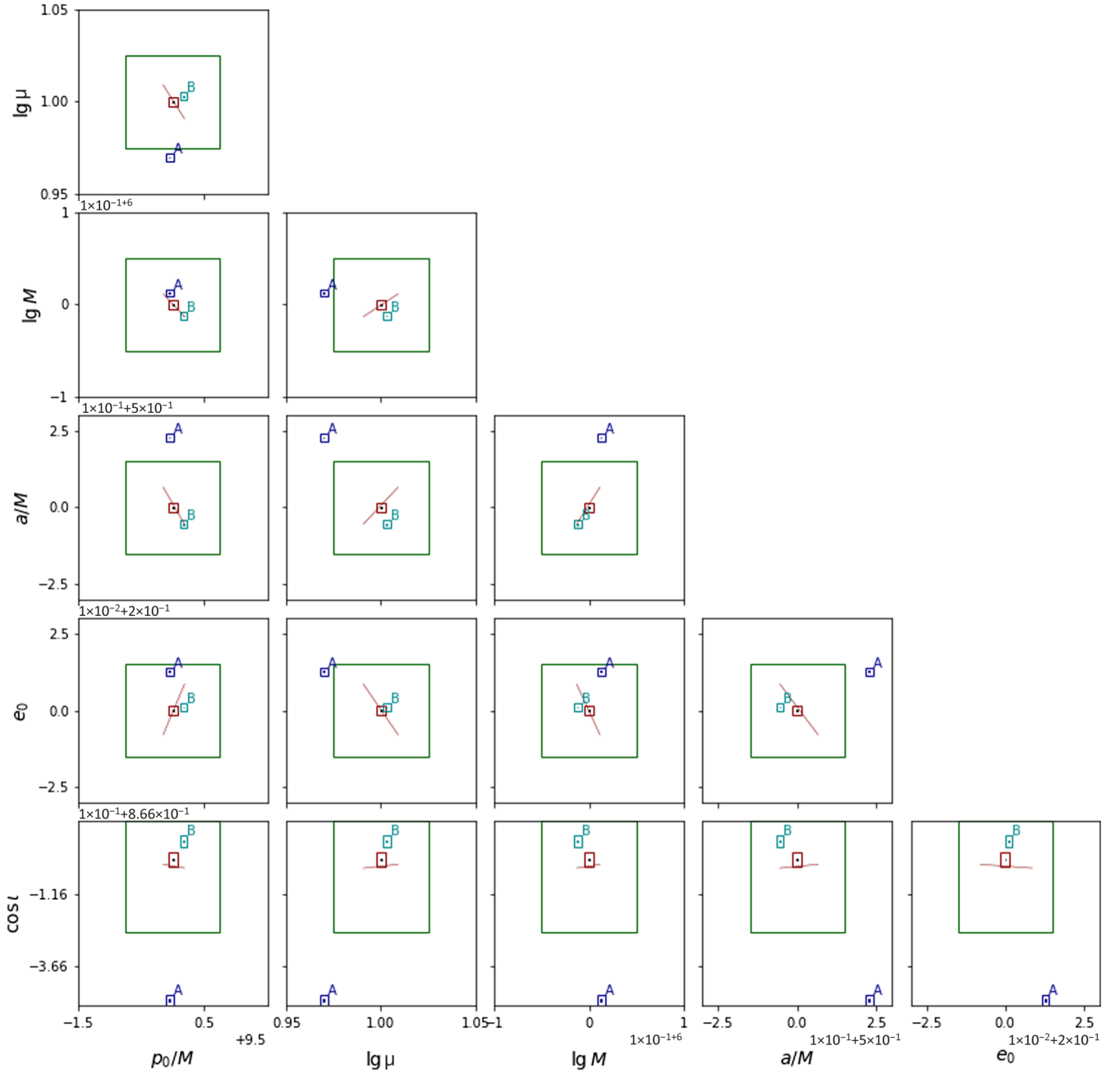
FIG. 17. Summary plot showing the various analysis subregions in Sec. IV A: $\mathcal{R}_0$ (black points, corresponding to the plot range in each panel of Fig. 8); $\mathcal{R}_1$ (red rectangles, Fig. 9); $\mathcal{R}_2$ (green rectangles, Fig. 12); $\mathcal{R}_3$ (plot range in each panel here); $\mathcal{R}_A$ (blue points, Fig. 13); $\mathcal{R}'_A$ (blue rectangles, Fig. 14); $\mathcal{R}_B$ (cyan points, Fig. 15); $\mathcal{R}'_B$ (cyan rectangles, Fig. 16). The red line in each panel indicates the projected extent of the extended connection for node-pair 15 in Fig. 11.

B. Interpretations

1. What actually causes degeneracy?

Broadly speaking, degeneracy arises when strong modes in the signal injection and template have similar initial frequencies and time derivatives of these frequencies, such that their phasing is aligned for much of the analysis duration. Treating the effect of initial phasing on degeneracy as negligible, we will find it useful to interpret any

given secondary (node) $\boldsymbol{\theta}_{\text{sec}}$ as an approximate root for a nonlinear system of equations in $\boldsymbol{\theta}$:

$$\left(\frac{d}{dt}\right)^k \omega_j(t;\boldsymbol{\theta})|_{t_0} \approx \left(\frac{d}{dt}\right)^k \omega_{j'}(t;\boldsymbol{\theta}_{\text{inj}})|_{t_0}, \quad (45)$$

where (j, j') is a single fixed pair of general mode indices, and the order k of the time derivative ranges from zero to some small positive integer. (To be more precise, the

notation $x \approx y$ in Eq. (45) represents $|x - y| < \epsilon$ for some given $\epsilon(j, j', k) \ll 1$.) In principle, a secondary might satisfy a system specified by a range of pairs (j, j'), i.e., two or more modes are matched well between injection and template. This is far less common in the signal space, however, and the injection overlap is likely to be more strongly determined by a single pair in any case.

From the results of our mapping analysis, we may draw several anecdotal conclusions about which modes are typically aligned between injection and template and to what degree. For all of the 705 secondary nodes identified in the large analysis region $\mathcal{R}_2$, it is the dominant mode in both injection and template that is matched, i.e., $j = j' = 2$ in Eq. (45), with j as defined just after Eq. (17). In other words, only the azimuthal fundamental frequency ω_ϕ and its time derivatives (up to at least second order) are similar in the injection and these secondaries. As an example, the fractional differences in $(d/dt)^k \omega_{r,\theta,\phi}$ (henceforth denoted by overdots) with respect to the injection values are shown in Fig. 18 for the case-study secondaries A and B. A matched azimuthal mode of motion is a sufficient condition for the existence of a strong secondary (at least in the regime of low eccentricity and modest initial separation, where $j = 2$ is the dominant mode), and our results seem to indicate that it is a necessary one as well. This is not too surprising since our considered injection has around 70–90% of its overall power in the dominant mode and 10–20% in the second strongest (see Fig. 2); thus the expected overlap from a perfect alignment of the dominant modes is $\gtrsim 0.7$, versus $\lesssim 0.2$ for the second-strongest modes.

As highlighted in [24,25], it is also possible for secondaries to arise from the matching of different modes between injection and template, i.e., $j \neq j'$ in Eq. (45). These are not encountered in our analysis—partly because we do not include the sideband modes from Lense-Thirring precession and partly because templates with a $j \neq j'$ matching of modes tend to occur at greater separations from the injection in parameter space (as the initial mode frequencies are now different). It is also unclear whether such secondaries will be more common than the type
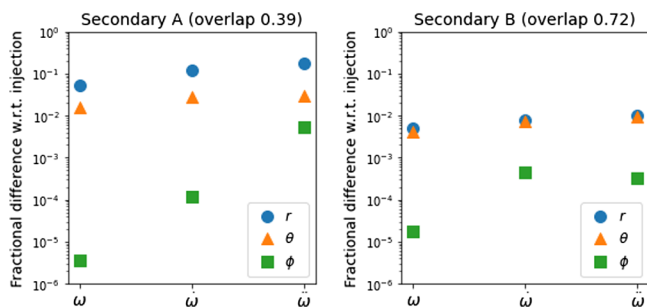

FIG. 18. Fractional differences in the fundamental frequencies and their low-order time derivatives at time t_0, for secondaries A [Eq. (43)] and B [Eq. (44)] with respect to the injection.

observed here; while they admit additional possible combinations of matched modes, their prevalence really depends on the measure of the set of approximate solutions to Eq. (45). Nevertheless, the $j \neq j'$ case (excluding strong sidebands) is generally less of a factor for matched-SNR searches because it does not cause secondaries that are very pronounced. For our considered injection, the expected overlap from a perfect alignment of the dominant and second-strongest modes is $\lesssim 0.4$.

When large-scale SNR gradients in the posterior surface are taken into account, the relationship between the strength of a secondary (relative to the posterior tails) and its overlap against the injection is no longer straightforward. For example, consider two secondaries that correspond to nodes $\theta_{\text{sec}}, \theta'_{\text{sec}}$ with the same injection-overlap value, but different optimal SNRs $(\rho_{\text{sec}}, \rho'_{\text{sec}}) = (\rho_{\text{inj}}, 2\rho_{\text{inj}})$. The difference in log-likelihood between θ_{sec} and a nearby tail region (where $\rho_{\text{opt}} \approx \rho_{\text{inj}}$ and $\Omega(h_{\text{inj}}, \cdot) \approx 0$) is half that of the difference between θ'_{sec} and its nearby tails (where $\rho_{\text{opt}} \approx 2\rho_{\text{inj}}$). In other words, low-overlap secondaries (be they from the $j = j'$ or $j \neq j'$ case) are more pronounced relative to the large-scale gradients if they occur in regions of high SNR. This fact becomes particularly relevant when extrinsic source parameters are added to the mix since they have a larger effect on SNR relative to their measurement precision and thus might lead to the (uneven) boosting of low-overlap secondaries within a typical search region.

2. Why are secondaries non-Gaussian?

Speaking from our own experience, it is natural to hold two related notions about the nature of EMRI degeneracy: (i) posterior secondaries are unimodal peaks that can be cleanly localized; and (ii) even congealed or deformed secondaries can still be characterized as "Gaussian" around a local maximum, with covariances given by the pullback metric at that point (i.e., the Fisher information matrix for a different posterior with that signal as the injection). This intuition is largely inherited from the behavior of the posterior bulk around the injection parameters—not just for EMRIs, but for GW sources in general. Our mapping analysis now provides empirical evidence that both notions are invalid, and here we give a simple theoretical argument to support this conclusion.

Recall from Sec. II that the standard GW log-likelihood (22) with $n = 0$ is proportional to the squared Euclidean distance between signal injection and template in the data space $\mathcal{D}$ (with equipped inner product $\langle \cdot | \cdot \rangle$). For θ near the global likelihood maximum θ_{inj} such that $|\delta\theta_{\text{inj}}| \ll |\theta_{\text{inj}}|$ with $\delta\theta_{\text{inj}} := \theta_{\text{inj}} - \theta$, we have

$$\ln L(\theta) \propto \langle h_{\text{inj}} - h(\theta) | h_{\text{inj}} - h(\theta) \rangle$$
$$= \delta\theta_{\text{inj}}^T \mathcal{I}_{\text{inj}} \delta\theta_{\text{inj}} + \mathcal{O}(|\delta\theta_{\text{inj}}|^3), \quad (46)$$

where the Fisher information $\mathcal{I}_{\text{inj}} := \mathcal{I}(\boldsymbol{\theta}_{\text{inj}})$ from Eq. (15) is symmetric and positive definite. With an uninformative prior, the posterior can thus be approximated as Gaussian over a local neighborhood of $\boldsymbol{\theta}_{\text{inj}} \in \Theta$.

An oft-overlooked subtlety is that Eq. (46) alone does not explain why GW posteriors can strongly resemble Gaussian distributions up to 1- or 2-σ. Over a sufficiently small region around a maximum point, the likelihood surface can be well described by only the leading-order piece of Eq. (46); this does not guarantee that the associated length scales are smaller than the region itself, which is a requirement for the surface to "look" Gaussian rather than flat. The other requirement is that the higher-order central moments must be similar to the multivariate-normal values for the leading-order likelihood. In particular, the third-order moments that determine skewness must nearly vanish. This is indeed satisfied in typical GW posteriors for the intrinsic source parameters since those moments scale approximately in size with the components of the third-order tensor $\langle \partial_{\boldsymbol{\theta}} h | \partial_{\boldsymbol{\theta}}^2 h \rangle$, which are ≈ 0 for parameters that affect signal phasing.

Now consider a secondary likelihood maximum $\boldsymbol{\theta}_{\text{sec}}$, with corresponding signal h_{sec}. For $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_{\text{sec}}$ such that $|\delta\boldsymbol{\theta}_{\text{sec}}| \ll |\boldsymbol{\theta}_{\text{sec}}|$ with $\delta\boldsymbol{\theta}_{\text{sec}} := \boldsymbol{\theta}_{\text{sec}} - \boldsymbol{\theta}$, we may write

$$
\begin{aligned}
\ln L(\boldsymbol{\theta}) &\propto \langle h_{\text{inj}} - h(\boldsymbol{\theta}) | h_{\text{inj}} - h(\boldsymbol{\theta}) \rangle \\
&= \langle h_{\text{inj}} - h_{\text{sec}} | h_{\text{inj}} - h_{\text{sec}} \rangle + 2\langle h_{\text{inj}} - h_{\text{sec}} | h_{\text{sec}} - h(\boldsymbol{\theta}) \rangle \\
&\quad + \langle h_{\text{sec}} - h(\boldsymbol{\theta}) | h_{\text{sec}} - h(\boldsymbol{\theta}) \rangle \tag{47}
\end{aligned}
$$

$$
\begin{aligned}
&= \delta\boldsymbol{\theta}_{\text{sec}}^T \mathcal{I}_{\text{sec}} \delta\boldsymbol{\theta}_{\text{sec}} + \delta\boldsymbol{\theta}_{\text{sec}}^T \langle h_{\text{inj}} - h_{\text{sec}} | H_{\text{sec}} \rangle \delta\boldsymbol{\theta}_{\text{sec}} \\
&\quad + \mathcal{O}(|\delta\boldsymbol{\theta}_{\text{sec}}|^3) + \text{const.}, \tag{48}
\end{aligned}
$$

where $\mathcal{I}_{\text{sec}} := \mathcal{I}(\boldsymbol{\theta}_{\text{sec}})$ and $H_{\text{sec}} := \partial_{\boldsymbol{\theta}}^2 h(\boldsymbol{\theta})|_{\boldsymbol{\theta}_{\text{sec}}}$. Note that H here denotes the Hessian tensor of the waveform model, not to be confused with the Hessian matrix of the leading-order likelihood (which is $-\mathcal{I}$). Also, the linear-in-$\delta\boldsymbol{\theta}_{\text{sec}}$ term in Eq. (47) vanishes because

$$
\langle h_{\text{inj}} - h_{\text{sec}} | \partial_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \rangle|_{\boldsymbol{\theta}_{\text{sec}}} \propto \partial_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}_{\text{sec}}} = 0. \tag{49}
$$

The third-order term in Eq. (47) includes the usual contribution from $\langle \partial_{\boldsymbol{\theta}} h | \partial_{\boldsymbol{\theta}}^2 h \rangle$, which is again ≈ 0 at $\boldsymbol{\theta}_{\text{sec}}$, but picks up an additional piece that scales in size with the components of $\langle h_{\text{inj}} - h_{\text{sec}} | \partial_{\boldsymbol{\theta}}^3 h \rangle \approx \langle h_{\text{inj}} | \partial_{\boldsymbol{\theta}}^3 h \rangle$. This quantity generally does not vanish since h_{inj} is not perfectly proportional to h_{sec}; although its impact on the local deviation from Gaussianity is specific to the behavior of h at $\boldsymbol{\theta}_{\text{sec}}$, it does indicate that secondaries with higher injection overlaps will tend to look more Gaussian (as might be expected). Higher-order likelihood terms aside, the local covariance structure for a secondary is altered even at leading order, as seen from the second term in Eq. (48). This quadratic form is not necessarily positive definite, but

let us say that the sum of the first two terms in Eq. (48) still is and that all higher-order terms are negligible. The inverse covariance matrix is then not $\mathcal{I}_{\text{sec}}$ as expected, but $\mathcal{I}_{\text{sec}} + \langle h_{\text{inj}} - h_{\text{sec}} | H_{\text{sec}} \rangle$, which encodes the local embedding curvature of the signal manifold $\mathcal{S}$ through its dependence on H_{sec}.

3. How representative are these results?

As discussed at the start of Sec. IV, a more extensive study of degeneracy is not really warranted at this stage since any results obtained with existing tools will change quantitatively for the next generation of science-adequate EMRI waveform models—or even qualitatively, if the nature of signal space is severely altered by distinctive higher-order effects such as transient self-force resonances [71–74]. Thus it is important to question the robustness of our results to the expected differences in waveform models and whether our conclusions are generally representative of the EMRI signal space. In particular, might systematic errors in PN (theoretical) or adiabatic-fitted (computational) evolution schemes introduce artificial degrees of degeneracy that would otherwise be nonexistent in more accurate models? Here we give a heuristic argument for why this is unlikely.

Regardless of the underlying evolution, a sufficient condition for a strong secondary to arise (at low eccentricity and modest initial separation) is simply Eq. (45) with $j = j' = 2$ and $k \leq 2$. We write this more concisely as

$$
(\omega_\phi, \dot{\omega}_\phi, \ddot{\omega}_\phi)(t_0; \boldsymbol{\theta}) \approx (\omega_\phi, \dot{\omega}_\phi, \ddot{\omega}_\phi)(t_0; \boldsymbol{\theta}_{\text{inj}}). \tag{50}
$$

The time evolution of the azimuthal fundamental frequency ω_ϕ (along with its radial and polar counterparts) over the inspiral is fully described by a trajectory of osculating geodesics $G(t) := (p(t)/M, e(t), \iota(t))$, whose governing equations are specific to the EMRI model in question. (Recall that the evolution of ι is only neglected in the AAK model.) In other words, we may decouple its explicit dependence on time: $\omega_\phi(t) = \omega_\phi(G(t))$. Both the instantaneous frequency $\omega_\phi(G)$ and the trajectory $G(t)$ depend explicitly on the intrinsic source parameters $\boldsymbol{\theta}$, but the former does so in a model-independent way (it is simply a characteristic of Kerr geodesic motion [45]). With such a description, we have

$$
\dot{\omega}_\phi(t) = \partial\omega_\phi(G) \cdot \dot{G}(t), \tag{51}
$$

$$
\ddot{\omega}_\phi(t) = \dot{G}(t)^T \cdot \partial^2\omega_\phi(G) \cdot \dot{G}(t) + \partial\omega_\phi(G) \cdot \ddot{G}(t), \tag{52}
$$

where all dependence on $\boldsymbol{\theta}$ is hidden for compactness.

We may examine the "approximate solution set" for the underdetermined nonlinear system (50) by casting it as an optimization problem. Some shorthand notation is useful here: let $\omega_0(\boldsymbol{\theta}) := \omega_\phi(t_0, \boldsymbol{\theta})$ and $\delta\omega_0(\boldsymbol{\theta}) := |\omega_0(\boldsymbol{\theta}) - \omega_0(\boldsymbol{\theta}_{\text{inj}})|/\omega_0(\boldsymbol{\theta}_{\text{inj}})$, with analogous notation for

the time derivatives of ω_ϕ. (The set $\{\delta\omega_0, \delta\dot\omega_0, \delta\ddot\omega_0\}$ corresponds to the ordinate values of the green squares in Fig. 18.) In essence, a secondary node is then simply a point that minimizes the objective function,

$$f(\boldsymbol{\theta}) := |(\delta\omega_0, \delta\dot\omega_0, \delta\ddot\omega_0)(\boldsymbol{\theta})|, \qquad (53)$$

for f below some unspecified threshold value.

The objective surface over parameter space will, of course, depend on the variation of $\dot G(t_0; \boldsymbol{\theta})$ and $\ddot G(t_0; \boldsymbol{\theta})$ through Eqs. (51) and (52), but the trajectories themselves are qualitatively similar across different EMRI models—modulo the inclusion of transient resonances, which cannot be treated with Eq. (50) anyway. To demonstrate that degeneracy does not arise solely from using any specific trajectory model, let us consider the case where $\dot G(t_0; \boldsymbol{\theta})$ and $\ddot G(t_0; \boldsymbol{\theta})$ are artificially assigned fixed values over parameter space (say, their values at $\boldsymbol{\theta}_{\rm inj}$). The objective surface is then driven purely by the variation of ω_ϕ and its partial derivatives with respect to G, but it is straightforward to verify that there still exist multiple distant minima in f (see Fig. 19). This strongly indicates that degeneracy occurs under generic conditions and is not merely an artifact of the adiabatic-fitted AAK trajectories used in our study.

While our conclusions on the nature of EMRI degeneracy should generalize to future waveform models, the single signal injection that we consider is not fully representative of injections in other regions of parameter space. Degeneracy is expected to be most severe at low eccentricity since the prevalence of strong posterior secondaries should be reduced when higher harmonic modes have a larger contribution to the overall signal power, i.e., the injection becomes more distinctive with added degrees of freedom to fit. Another consideration is the effect of the analysis duration T. In regions of parameter space where the inspiral lifetime is significantly longer than T, an extension of the latter will also mitigate degeneracy (this is not applicable here since all sources plunge after around $T = 2$ y). Finally, transient resonant "jumps" in the trajectories $G(t)$ will be a generically occurring feature of models with postadiabatic evolution [71] and thus of the true EMRI signal space; the presence of one such jump in the signal injection effectively shifts the evolution of its postjump part onto a different set of phase trajectories. This again increases the complexity of the injection and should reduce degeneracy as well.

V. IMPLICATIONS FOR DATA ANALYSIS

A. Fundamental implications

1. Interaction with detector noise

With the qualitative nature of degeneracy established in Sec. IV A, we now turn to various implications that our
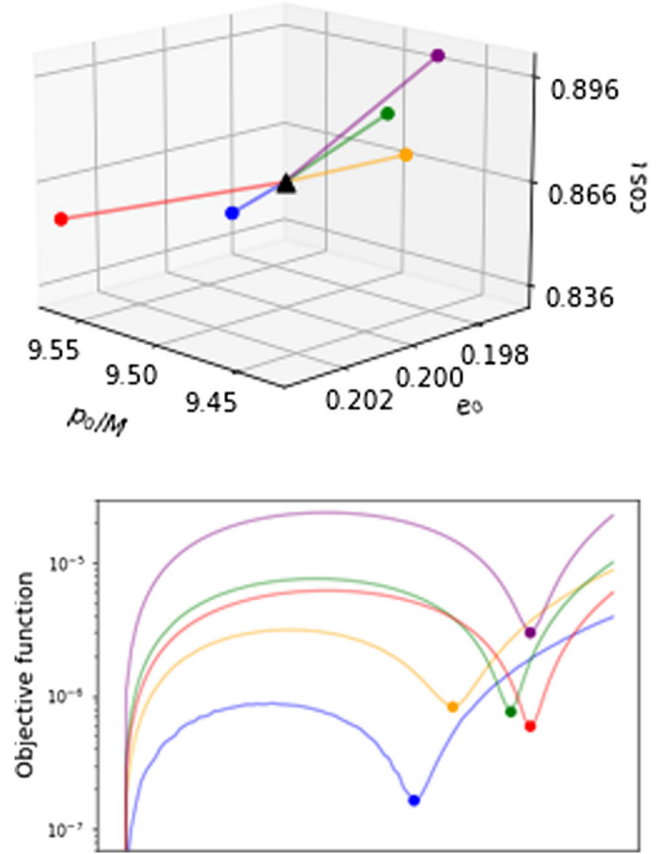

FIG. 19. Minimization of objective function (53) with constant $\dot G(\boldsymbol{\theta})$ and $\ddot G(\boldsymbol{\theta})$, over a five-dimensional cross section of the intermediate region $\mathcal{R}_1$ in Sec. IV A 2 (setting lg $\mu = 1$). Around 50 randomly drawn starting points are used, with each leading to a different local minimum; only five of these are presented here for clarity. Top: projection of minimum locations (colored points) onto the $(p_0/M, e_0, \cos\iota)$ space, together with the lines connecting them to the injection parameters (black triangle). Bottom: value of objective function along the (extended) connecting lines in the top panel.

results pose for EMRI data analysis (both in a general sense and in the context of the standing wisdom on search, inference, and modeling approaches). The first question we address is: How likely is it that detector noise will combine with the signal template at a secondary node to give a higher detection SNR than the template corresponding to the injection? Under the standard noise assumptions laid out in Sec. II B, the answer is straightforward—a false determination of the best-fit template due to noise is extremely improbable.

As in Eq. (8), let the data be $x = h_{\rm inj} + n$ and further assume that n is a zero mean and stationary Gaussian process. From Eqs. (9), (10), and (13), the detection SNR ρ_1 of the injection template $h_1 := h_{\rm inj}$ is normally distributed with mean $\rho_{\rm inj}$ and unit variance. Consider some secondary template h_2 with an injection overlap of Ω; its detection SNR ρ_2 is also normally distributed with mean

$\Omega\rho_{\text{inj}}$ and unit variance, while the correlation coefficient (normalized covariance) of ρ_1 and ρ_2 is Ω. The probability distribution of $\rho_2 - \rho_1$ is then $\mathcal{N}(\rho_{\text{inj}}(\Omega - 1), 2(1 - \Omega))$, and the probability that $\rho_2 \geq \rho_1$ is given by

$$P(\rho_2 \geq \rho_1) = \frac{1}{2}\left(1 - \text{erf}\left(\frac{1}{2}\rho_{\text{inj}}\sqrt{1 - \Omega}\right)\right). \quad (54)$$

Thus for threshold injections with $\rho_{\text{inj}} = 20$, noise will lead to false determination of the injection parameters $\gtrsim 1\%$ of the time only if the offending secondary has an injection overlap of $\gtrsim 0.97$. For the highest-overlap secondary identified in the analysis regions of Sec. IV A ($\Omega = 0.76$), the probability of this occurring is $\sim 10^{-12}$.

2. Interaction with multiple signals

The degeneracy study of Sec. IV focuses on nonlocal signal templates that strongly resemble a single signal injection. We now use analytic arguments to extend our results to the scenario where templates have non-negligible and nonlocal correlations with the sum of two injections (which themselves are nonlocal to each other). Let us restrict to the case where the injections are uncorrelated since (i) two injections with non-negligible overlap have a sum that lies approximately in the span of either, so they reduce effectively to the case of a single injection; and (ii) injections with near-zero overlap are far more typical anyway, as shown by the confusion study of Sec. III. Our analysis remains well motivated in the uncorrelated-injection case since it is quite plausible for some template to match the dominant mode in one injection with its own strongest mode, while matching the dominant mode in the other with its second strongest mode.

To pose a more precise question, let us consider two uncorrelated injections $h_{1,2}$ with optimal SNRs $\rho_{1,2}$, where $\rho_1 \geq \rho_2$ without loss of generality. Then we ask: What is the probability that some nonlocal template $h(\boldsymbol{\theta})$ (i.e., a template that is not local to either h_1 or h_2) has a high enough overlap with $h_1 + h_2$ such that its (expected) detection SNR is $\rho_{\boldsymbol{\theta}} \geq \rho_1$? If this scenario turned out to be even moderately likely, then clearly one would have to design search strategies with the possibility in mind. However, the order-of-magnitude argument we sketch below strongly suggests that the probability of such an occurrence is negligible.

Denoting the individual injection-template overlaps by $\Omega_1(\boldsymbol{\theta}) := \Omega(h_1, h(\boldsymbol{\theta}))$ and $\Omega_2(\boldsymbol{\theta}) := \Omega(h_2, h(\boldsymbol{\theta}))$, the detection SNR of $h(\boldsymbol{\theta})$ given the sum of injections $h_1 + h_2$ may be written as $\rho_{\boldsymbol{\theta}} = \Omega_1(\boldsymbol{\theta})\rho_1 + \Omega_2(\boldsymbol{\theta})\rho_2$. Our calculation relies on some assumptions about the functions $\Omega_{1,2}(\boldsymbol{\theta})$ over parameter space, which for random $\boldsymbol{\theta}$ may be approximated as two independent and identically distributed random variables $\Omega_{1,2}$. Based on the overall population of identified secondaries in Sec. IV A, we will assume that $\Omega_{1,2} \leq \Omega_{\text{max}} \approx 0.8$ (beyond the local neighborhoods of

$\boldsymbol{\theta}_{1,2}$). For a lower bound on $\Omega_{1,2}$, note first that our formal definition of detection SNR in Eq. (13) admits negative values; implicitly, we are treating any template as being fully specified by its parameters with no freedom for maximization over phase (e.g., the translation of a template in time is viewed as a distinct template corresponding to a different point in parameter space). We then consider only $\Omega_{1,2} \geq \Omega_{\text{min}} \approx 0$ in this calculation, which is conservative as $\rho_{\boldsymbol{\theta}}$ is constrained to be positive, and thus the probability $P(\rho_{\boldsymbol{\theta}} \geq \rho_1)$ is overestimated. This streamlines our argument and connects it more naturally to results obtained with the approximate inner product (which is also positive by definition).

Another important assumption is to specify the probability distribution of the random variables Ω. Our mapping analysis does not directly provide an empirical determination of the probability density function $p(\Omega)$, but it enables us to posit a rough functional form for p. For $\Omega_{\text{min}} \leq \Omega \leq \Omega_{\text{max}}$, we write

$$p(\Omega) := -\frac{1}{V_0}\frac{dV}{d\Omega}, \quad (55)$$

where $V(\Omega)$ is the volume of the set of points in Θ with injection-overlap values that are $\geq \Omega$, and $V_0 := V(\Omega_{\text{min}})$. Since $\Omega_{\text{min}} \approx 0$, V_0 is of a similar order of magnitude to the volume of the full (six-dimensional) parameter space Θ. The function V is defined with respect to some natural measure on Θ; we will choose a measure such that the volume associated with a typical posterior secondary is ~ 1, give or take a few orders of magnitude (to account for the diversity in the extents of secondaries with different injection overlaps). This allows us to make the simplification that $V(\Omega)$ is given by the number of secondaries with injection-overlap values $\geq \Omega$, so that $dV/d\Omega$ can be estimated from our mapping analysis. With such a measure, the volume of the posterior-bulk region examined in Sec. IV A 1 is also ~ 1 and thus $V_0 \sim 10^{18}$ from the discussion after Eq. (37).

The above simplification is clearly invalid for $\Omega \gtrsim 0.75$ (where there are no secondaries), while selection effects in both sampling and clustering also lead to vastly lower counts for secondaries with overlaps $\lesssim 0.5$ (by design). Nevertheless, our data for $0.5 \lesssim \Omega \lesssim 0.75$ shows a clear trend for $dV/d\Omega$, which we will take to hold for $\Omega_{\text{min}} \leq \Omega \leq \Omega_{\text{max}}$. Let us assume that $dV/d\Omega$ is described by a power law in the latter range and write

$$\frac{dV}{d\Omega} := -C\Omega^{-\alpha}. \quad (56)$$

The combined set of 705 secondaries from the analyses in Secs. IV A 2 and IV A 3 (the colored points in the inset of Fig. 12) is distilled to a set of 251 secondaries with overlaps > 0.5; the distribution of this smaller set is fit well by Eq. (56) with $C \approx 26$ and $\alpha \approx 7$ (see Fig. 20). As it turns
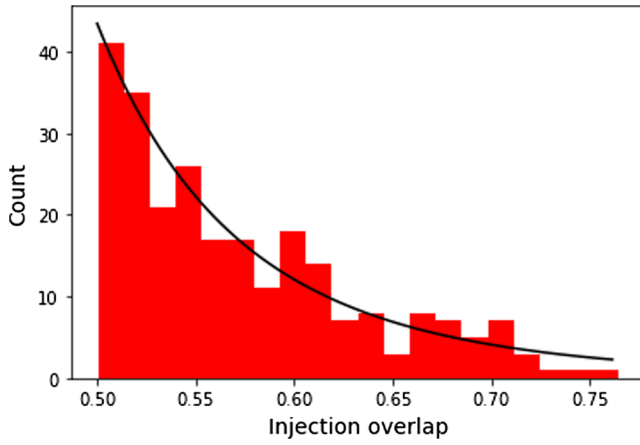
FIG. 20. Histogram of injection-overlap values for 251 signals (with overlaps >0.5) from inset of Fig. 12. The probability density for the histogram distribution is well fit by the power law (56) with $\alpha \approx 7$ and is robust to the choice of bin size.

out, the results from the confusion study suggest that this power law is a reasonable approximation over the range $\Omega_{min} \leq \Omega \leq \Omega_{max}$. Assuming the root-mean-square pairwise overlap of $\sim 10^{-3}$ from that study is indicative of the injection overlap for the majority of points in the full space, we might expect that $V(10^{-3})$ should be of a similar order of magnitude to $V_0 \sim 10^{18}$. This is indeed the case: for Eq. (56) with $C = 26$ and $\alpha = 7$, we find $V(10^{-3}) \approx 4 \times 10^{18}$. Conversely, we also require $\Omega_{min} \sim 10^{-3}$ such that the probability density (55) is properly normalized on the interval $[\Omega_{min}, \Omega_{max}]$ (its integral depends negligibly on Ω_{max}).

Now we define $\beta := \Omega_1 + r\Omega_2$, where $r := \rho_2/\rho_1 \leq 1$; this random variable takes values in the interval $[0, 2]$, and $\beta \geq 1$ corresponds to the case of interest $\rho_\theta \geq \rho_1$. Since Ω_1 and Ω_2 are IID random variables, we have

$$p(\beta) = \int_{\Omega_{min}}^{\Omega_{max}} \int_{\Omega_{min}}^{\Omega_{max}} d\Omega_1 d\Omega_2 \frac{C^2}{V_0^2} \Omega_1^{-\alpha} \Omega_2^{-\alpha} \delta(\beta - \Omega_1 - r\Omega_2)$$

$$= \frac{C^2}{V_0^2} r^{\alpha-1} \int_{\Omega_{min}}^{\Omega_{max}} d\Omega (\beta - \Omega)^{-\alpha} \Omega^{-\alpha}$$

$$\times \vartheta(\beta - \Omega - r\Omega_{min}) \vartheta(r\Omega_{max} + \Omega - \beta), \quad (57)$$

where δ and ϑ denote the Dirac-delta and Heaviside-theta functions, respectively. The most conservative (largest) estimate of the probability $P(\beta \geq 1)$ occurs for the equal-SNR case $r = 1$; with $V_0 = 10^{18}$, $C = 26$, $\alpha = 7$, $\Omega_{min} = 10^{-3}$, and $\Omega_{max} = 0.8$, we find $P(\beta \geq 1) \sim 10^{-30}$.

To what extent does the above estimate for the probability $P(\beta \geq 1)$ depend on both the maximal injection overlap Ω_{max} and the SNR ratio r? In Fig. 21, $P(\beta \geq 1)$ is plotted as a function of the (minimal) mismatch $1 - \Omega_{max}$ for $0.8 \leq \Omega_{max} < 1$ and $r \in \{1, 0.7, 0.4\}$. When $\Omega_{max} < 1 - 10^{-3}$, the probability scales approximately as r^6

[in agreement with Eq. (57)] and $(1 - \Omega_{max})^{-4.6}$. As $\Omega_{max} \rightarrow 1$, it asymptotes instead to a maximal value of $\approx 4 \times 10^{-20}$ for $r = 1$. In other words, even if we assume that the power law (56) is valid up to $\Omega_{max} = 1$, the probability that $\beta \geq 1$ is still completely negligible. From Eqs. (55) and (56), the maximal probability value corresponds approximately to $P(\Omega \geq 0.998)$ for a single injection-template overlap; essentially, if there exists a template closely resembling h_1 with a mismatch of $\lesssim 10^{-3}$, then virtually any h_2 can push β over unity.

We started this analysis by considering in isolation the sum of two signal injections. For a set of N injections, there are $N(N-1)/2$ such sums. The random variables β_i associated with these sums are, of course, correlated, but it is conservative to treat them as IID in this context. A plausible middle-ground estimate for the number of resolvable EMRI signals that might be present in the LISA data stream is $N = 200$, as considered in Sec. III. For such a set of signals, the probability that any $\beta_i \geq 1$ is increased only by a factor of $\sim 10^4$ (to $\sim 10^{-26}$, if we take $r = 1$ and $\Omega_{max} = 0.8$). Thus it is highly improbable for the sum of any two signals in the data to mimic a nonexistent third signal (in the continuum of possible signals) so well that the best-fit template actually corresponds to the latter, rather than either of the two actual signals.

What about the possibility of three actual signals coincidentally summing to mimic a nonexistent fourth signal? Intuition indicates that such an occurrence is even more improbable than the two-signal case, although an extension of the above analysis to three signals is somewhat nontrivial. Nevertheless, we may construct a back-of-the-envelope argument by restricting to specific subcases of the two scenarios. For a small set of M injections h_j with the same optimal SNR ρ, consider the probability that the
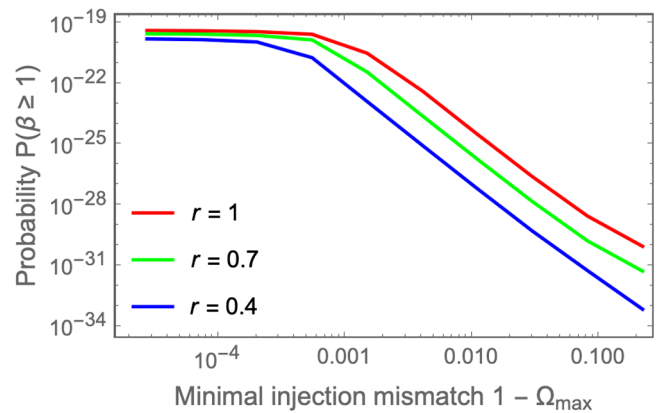

FIG. 21. Probability that the sum of two injections yields a nonlocal template with a higher detection SNR than the optimal SNR of either injection (i.e., the integral of Eq. (57) over the interval $[1, 2]$). This is plotted as a function of the minimal injection-template mismatch $1 - \Omega_{max}$ across the signal space, for several values of the injection-SNR ratio r.

overlap Ω_j of each injection with a given template $h(\boldsymbol{\theta})$ is $\geq 1/M$, which is a sufficient condition for $\rho_{\boldsymbol{\theta}} = \rho \sum_j \Omega_j \geq \rho$. From Eqs. (55) and (56), we have

$$P\left(\Omega \geq \frac{1}{M}\right)^M = \left(\frac{C(M^{\alpha-1} - 1)}{V_0(\alpha - 1)}\right)^M. \quad (58)$$

With $V_0 = 10^{18}$, $C = 26$, and $\alpha = 7$ as before, we find a probability of $\sim 10^{-31}$ for $M = 2$; this is just below the estimate of $P(\beta \geq 1) \sim 10^{-30}$ for the more general treatment with $r = 1$ (as expected).

Now, for a set of $N = 200$ actual signals in the data, we may estimate how likely it is that any $M \leq N$ signals fulfill the above sufficient condition and then examine the ratio of probabilities between the two representative cases $M = 3$ and $M = 2$. This ratio evaluates to

$$\frac{\binom{N}{3}P(\Omega \geq \frac{1}{3})^3}{\binom{N}{2}P(\Omega \geq \frac{1}{2})^2} \approx (3 \times 10^7)V_0^{-1} = 3 \times 10^{-11}, \quad (59)$$

indicating that $P(\rho_{\boldsymbol{\theta}} > \rho)$ in the case of three signals is infinitesimal relative to its value in the case of two signals (which we have already established as practically negligible). More generally, the ratio of probabilities between the case of any $M \leq N$ and that of $M = 2$ scales as γV_0^{2-M}, where $\lg \gamma$ is only slightly superlinear with M and has a much smaller absolute gradient than -18 throughout (i.e., the coefficient of M in the common logarithm of the other factor). Thus the logarithm of the ratio decreases monotonically with M, which convincingly rules out the likelihood of any number of actual signals coincidentally summing to mimic a nonexistent signal. In summary, we conclude that the interaction of degeneracy with multiple signals is unlikely to pose any fundamental difficulties for the extraction and characterization of resolvable EMRI signals in LISA data.

B. Search strategies

The search for signals in GW data really involves two distinct tasks: detection, which establishes the presence of a signal; and what we shall refer to as "identification," which maps the detected signal to astrophysically relevant source parameters (approximately or otherwise). Detection technically includes the assessment of statistical significance for candidate signals but may be treated, in principle, as simply the attainment of some threshold value for the detection SNR. Identification is the step that connects to inference; it poses less of a separate issue when the parameters have an easily invertible map to the waveform observables (as is the case for Galactic binaries [23]), or when standard inference techniques can already cover the entire space of possible signals (as is the case for massive-black-hole mergers [75,76]). For these sources and those of

ground-based observing, the task of identification is either unnecessary or trivial, leading to the familiar dichotomy of detection and inference (the latter is traditionally known as "parameter estimation").

In the case of EMRIs, detection is relatively straightforward. The presence of a signal can be uncovered with crude waveform models, with completely phenomenological ones, or even without models at all (see Sec. V B 2 for a short overview of such approaches). Parameter degeneracy may give rise to spurious candidate signals, but does not affect detection of any actual signal itself. On the other hand, identification becomes nontrivial since degeneracy endows the inverse problem with a highly disjoint set of near solutions. It remains unclear whether identification will eventually be necessary as an independent step since a sophisticated sampling algorithm that is well tuned to the EMRI problem might conceivably allow inference with a global prior; however, this has not been demonstrated before, much less in a reliable fashion. If it proves impractical to move directly from detection to inference, identification will be required to provide a good initial guess or to narrow down the prior region. Approximate waveform models might be used in rough searches to supply such information, but as shown in Sec. IV B 3, even a weakly physical model can still admit a multitude of near solutions to its inversion. If the model is overly simplistic, then it might succeed in smoothing out the search surface, but then may not provide sufficiently precise and/or accurate prior localization for inference.

The search for EMRIs is further complicated by the potential presence of more than one such signal in the data. Transdimensional methods [77], as used in Galactic-binary searches to find sets of signals with unspecified cardinality, are less viable for EMRIs due to degeneracy and the greater expense of waveform models. They are also not as crucial in this context since an astrophysically realistic set of EMRI signals is far less likely to "self-confuse" (as shown in Sec. III). A more direct approach for EMRIs would be to conduct a single-source search for the global best-fit template, corresponding to the strongest signal in the data. After performing precise inference with accurate models, this signal is subtracted from the data, which may then be searched for the second-strongest signal. (The same approach might be taken for massive-black-hole-merger signals since these are expected to be more resolvable from one another in the time domain.) However, the search for the strongest signal is still hindered by the presence of a second strong signal, which would manifest in the search surface as another set of apparent secondaries (but see Sec. VI).

1. General sampling algorithms

We now touch broadly on the implications of our results for the two main classes of sampling algorithms that will be important in LISA signal searches. The sampling of source posterior distributions with these algorithms is, of course,

a key component of inference as well, but we may view any nonlocal aspect of inference as being part of the search (identification) stage. Most forms of modeled search rely on the global exploration of some search surface—not necessarily the posterior—that is defined in terms of cross-correlations [Eq. (13)] or residuals [Eq. (14)] between data and template. As is well known, stochastic sampling is required to explore such a surface for EMRIs, even without accounting for the presence of degeneracy. This is due to the information volume of the signal space, which rules out grid sampling with precomputed template banks [13]. Let us focus on the case of a single-source search for the global probability maximum since any difficulties encountered there are only going to be amplified for strategies that look for multiple candidates at the same time (e.g., transdimensional sampling, or the online clustering of posterior samples).

Parallel-tempering MCMC methods [55] utilize multiple Markov chains of "walkers" to explore variants of the likelihood at different temperatures T_i (essentially L^{1/T_i}), doing so in parallel and with the exchange of information. Such algorithms allow walkers to move more efficiently across the search region than in traditional MCMC and thus aid sampling convergence for complex distributions. They are well suited to coping with the "noisy" tails of GW likelihoods [78,79] and should also be useful for posterior-based EMRI search. In practice, the strength and prevalence of EMRI degeneracy, coupled with the minuscule extent of the posterior bulk relative to the global prior, is likely to necessitate more walkers and/or temperatures than usual. As a rough indicator of the sort of numbers required for the EMRI problem, we report that for the prior region $\mathcal{R}_2$ in Sec. IV A 3, the posterior bulk cannot be found within 10^6 iterations using a standard geometric temperature ladder [80] (with $T_{\max} = T_{10} = \rho_{\mathrm{inj}}^2 = 400$), 10 walkers per temperature, and an assortment of jump proposals [70]. A fine-tuning of the ladder might be informed by exploratory studies to determine the pertinent temperature scales in the EMRI likelihood and to do so for a representative range of injections across the full parameter space.

Nested sampling [81] is a more recent paradigm for Bayesian posterior sampling and evidence calculations. It involves the iterative replacement of a set of "live" points with new points of higher likelihood, which provides an exponential contraction of the prior to the posterior. In the context of EMRI search, a very large number of live points is needed for sufficient resolution to chance upon the region containing the posterior bulk—which can easily be $<10^{-12}$ times the size of the prior region, even for a modest degree of prior localization. This might be computationally prohibitive, and furthermore, it is not clear how to incorporate problem-specific tuning in the proposal of new live points (the real engine under the hood of nested sampling, which only assumes that such points can be generated). Again, to indicate the required sampling resolution for the prior

region $\mathcal{R}_2$ in Sec. IV A 3, a slice-nested-sampling implementation [57] with 10^4 live points fails to converge on the posterior bulk for the (considerably smoother) likelihood L_{match} in Eq. (25).

2. EMRI-specific methods

A small number of modeled-search methods for EMRIs have been proposed in the literature. All of these were developed using the earliest kludge model [15], which supplies both the injection and the templates in each study. Nevertheless, the basic ideas that underpin such methods will apply to the search for realistic signals as well, provided that models with sufficiently accurate frequency evolution are used in the analysis. Both tasks of detection and identification are addressed simultaneously in a modeled search since each candidate signal is associated with some high-detection-SNR template from the model, which in turn has "known" parameters of interest (that must be specified in order to generate the template).

Semicoherent search is a broadly applicable strategy that is similar in effect to annealing-type sampling methods. As sketched out for EMRIs in [13], it is a modeled search with "flexible" templates that are phase-matched against N segments of the data with duration T/N, rather than the full data stream over its duration T. Note, however, that there are several other (mis)conceptions of semicoherent EMRI search—none of which have been described properly in the literature, if at all. The general idea of semicoherent filtering originates from continuous-wave searches in ground-based observing [82,83]. In the context of a precomputed template bank, it increases the "reach" of each template so that the density of the bank can be reduced. When used in stochastic searches (as required for EMRIs), it works essentially by smoothing out the search surface, making high-probability regions more extensive and easier to find. In terms of its interaction with degeneracy, the semicoherent approach will likely congeal secondaries and thus be useful for prior localization; the question is whether this localization is sufficient to transition directly into inference, and if not, then how many iterations of semicoherent search (with gradually increased segment durations) might be required.

Another class of method that is more specific to EMRIs exploits the fact that any point $\boldsymbol{\theta}_{\mathrm{sec}}$ with high injection overlap generally satisfies Eq. (45) for some $\{j, j', k\}$, which can thus be used to inform MCMC jump proposals. This is a central idea in [24], where the inversion of Eq. (45) with $\boldsymbol{\theta}_{\mathrm{inj}} \to \boldsymbol{\theta}_{\mathrm{sec}}$ [more precisely, the minimization of the associated objective function, as in Eq. (53)] is proposed for combinations of the azimuthal frequency ω_ϕ and the Lense-Thirring frequency $\omega_\phi - \omega_\theta$, along with their time derivatives up to second order. Only $j \neq j'$ is considered in the original proposal, which effectively puts the focus on jumping between templates where the dominant mode in one is aligned with a sideband of the dominant mode in the

other; it is straightforward to extend this to include the $j = j'$ case as well, in light of our results that highlight its importance. Finding all minima in the objective function is no less intractable than finding all posterior maxima, of course, and so only a single "root" is identified per jump—however, this may be of limited benefit if secondaries are extremely numerous. It is shown in [24] that such jumps help with localization to a broad region in parameter space, but it remains unclear whether they are sufficient on their own to cope with the denser set of secondaries, due to intrinsic degeneracy, when homing in on the posterior bulk.

In [25], similarly constrained MCMC jump proposals are integrated with ideas from the semicoherent approach in a search pipeline. Shorter half-year segments of the data are analyzed with either simplified or fully phenomenological templates in a number of separate analysis stages, but a key feature of the method is the inclusion of jumps that are informed by fixing the three fundamental frequencies (and not their derivatives), i.e., the inversion of Eq. (45) with $j \in \{r, \theta, \phi\}$ and $j = j'$. Segment information is combined by tracing the evolution of candidate signals in different segments to the same reference time t_0, and the overall list of candidates is then used to seed a localized search with full-duration templates. The entire method is stated to involve some degree of manual intervention and tuning. As in the case of [24], it is demonstrated on data sets with high-SNR injections as part of the Mock LISA Data Challenges [26]; the final parameters reported in [25] do not seem to agree with most of the injection parameters within their expected uncertainties (this is only an estimation on our part since no inference results are presented). Nevertheless, the method appears to provide solid localization for follow-up analysis.

Apart from approaches that use astrophysically parametrized EMRI models, there is also the option of performing minimally modeled or unmodeled searches. One such proposal [84] features a phenomenological model that is parametrized directly by the observable quantities in the signal: the dominant mode amplitudes (treated as constant), the three initial fundamental frequencies, and their initial time derivatives to third order. This approach could be useful for search since the model is extremely efficient and can overfit physical signals such that detection sensitivity is not impaired. Its main difficulty lies in mapping the phenomenological quantities to the source parameters for identification, which ultimately still requires the inversion of a physical model. Another proposed strategy is to directly search for EMRI-like frequency tracks in spectrograms of the data, using edge-detection algorithms [85]. Such excess-power searches are more limited for EMRIs, due to their low instantaneous SNR and their place in the SNR hierarchy of the global LISA catalog. They could be feasible for the detection of high-SNR sources, but again would have to rely on physical models for identification and thus remain susceptible to the issue of degeneracy as well.

C. Inference (and modeling) strategies

By partitioning the EMRI data-analysis problem into nonlocal search and local inference, all that is required in the latter stage for each candidate signal is to map out the vicinity of its source parameters in a localized prior region. Standard posterior sampling, as well as the multitude of techniques proposed for its acceleration, should be adequate for this purpose (but may still be less efficient if the prior region is too large). A high detection SNR for the candidate is insufficient to provide confidence that the examined parameter point corresponds to the parameters of some physical signal, but the nature of EMRI degeneracy actually works to our benefit here. One of the key conclusions from Sec. IV is that secondary posterior peaks are expected to feature pathologies that are absent in the posterior bulk, which could be a simple way of vetoing such solutions (see Sec. VI A).

However, we have up till now assumed a fitting factor [86] of unity between the injection and the template manifold, i.e., $h_{\text{inj}} \in \mathcal{S}$ such that $F := \max_{\boldsymbol{\theta}} \Omega(h_{\text{inj}}, h(\boldsymbol{\theta})) = 1$. This ensures that the posterior is well behaved near the best-fit point $\text{argmax}_{\boldsymbol{\theta}} \Omega(h_{\text{inj}}, h(\boldsymbol{\theta}))$, but it may not be achievable in practice even with future postadiabatic waveforms. A low fitting factor can impact on search through the reduction of expected detection SNRs and might pose a problem for inference even in the case where $F \gtrsim 0.5$. The latter is evident from the posterior around secondary nodes with injection overlaps as high as ≈ 0.8 (indicative of putative models where those nodes are the best-fit points)—and also from the results in [47], where the parameters of an injected signal with fully relativistic mode amplitudes are inferred from a template model with identical phasing but semi-relativistic amplitudes (leading to a best-fit overlap of ≈ 0.5 and a multimodal posterior). To complicate matters, F alone contains no information about the template manifold away from the best-fit point and so is insufficient as a determinant of whether and how much the posterior is deformed. Assuming the source parameters are still correctly and sufficiently localized by the search stage, $F < 1$ does not significantly challenge posterior estimation during inference, but may, of course, lead to an unacceptable degree of bias for precision applications of EMRI observations.

The well-understood need for low or manageable bias in high-precision inference [87], as well as the strong dependence of EMRI-specific search strategies on models with the correct frequency evolution, will impose constraints on modeling accuracy that have yet to be determined. Degeneracy does not impact directly on accuracy requirements, but is still relevant through its influence on data-analysis approaches. Studies on modeling accuracy are required at a level beyond simple dephasing arguments or fitting-factor calculations and are presently being undertaken as part of the LISA Science Group's work-package activities. In terms of modeling strategies, degeneracy would appear to severely handicap approaches such as

the construction of reduced-order-modeling surrogates [88–90] in the time or frequency domain (at least in the global sense). These waveform-level fits are already unlikely to be practical for EMRIs due to the length and complexity of signals, and the added severity of nonlocal correlations seems to rule out their viability beyond highly localized regions in parameter space. The data compression provided by reduced-order modeling itself is still useful in EMRI modeling, however; as demonstrated in [10], it only needs to be applied to the set of instantaneous mode amplitudes over the space of Kerr geodesics.

VI. SUGGESTIONS FOR DATA ANALYSIS

A. Simple *post hoc* vetoes

Notwithstanding the efficacy of search strategies, the main concern for EMRI search is that secondaries of actual signals might be falsely identified as candidate signals. There are two specific subscenarios to consider: (i) the actual signals themselves are found, and (ii) they are not. The former implies the existence of a set of ≥2 candidate signals; a simple check would then be to compute the pairwise overlaps among all of them (effectively, to cluster the candidates using the direct overlap as a measure of connectivity, as mentioned in Sec. II D). From the results of our confusion study in Sec. III, we do not expect ≳1% overlap between any two candidates. If this occurs, then the pair should be flagged as a possible manifestation of parameter degeneracy for follow-up analysis.

In the latter subscenario (which could arise after the above veto, or for a single incorrectly identified candidate), there exists a set of ≥1 candidate signals that are virtually uncorrelated with one another. If any of these candidates is a secondary for some actual signal that was not found due to search error, this should immediately be evident when attempting to estimate the posterior in the vicinity of its associated parameters. The obtained posterior is expected to be highly non-Gaussian and difficult to localize; further checks could involve comparison to either a Fisher-matrix analysis, or to a simulated posterior with the spurious candidate as an injection.

B. Veto "likelihood" function

Annealing-type sampling methods such as parallel-tempering MCMC involve exploring the likelihood at different temperatures (L^{1/T_i}) to efficiently converge on the posterior bulk. However, the relative log-likelihood gradients near the injection parameters and near some secondary node are unaltered by annealing, i.e., the ratio between $\ln L(\boldsymbol{\theta}_{\text{inj}}) - \tau(\boldsymbol{\theta}_{\text{inj}})$ and $\ln L(\boldsymbol{\theta}_{\text{sec}}) - \tau(\boldsymbol{\theta}_{\text{sec}})$ for local tail values $\tau(\boldsymbol{\theta}) := -\rho_{\boldsymbol{\theta}}^2$ (see Eq. (24) and related discussion) is preserved for all T_i. This means that secondary likelihood peaks remain present at all temperatures, and a large number of high-overlap ones can still cause problems in practice. Here we suggest a modified likelihood that suppresses secondary

peaks by accounting for the expected spread of individual-mode detection SNRs, in order to provide improved identification in the search stage. Specifically, we define a sampling density $L'(\boldsymbol{\theta})$ such that for a typical secondary node $\boldsymbol{\theta}_{\text{sec}}$, the difference $\ln L'(\boldsymbol{\theta}_{\text{sec}}) - \tau(\boldsymbol{\theta}_{\text{sec}})$ is much reduced from its original counterpart $\ln L(\boldsymbol{\theta}_{\text{sec}}) - \tau(\boldsymbol{\theta}_{\text{sec}})$. We also demand that $L'(\boldsymbol{\theta}_{\text{inj}}) \approx L(\boldsymbol{\theta}_{\text{inj}})$ in the vicinity of $\boldsymbol{\theta}_{\text{inj}}$, so as to facilitate a smooth transition into inference. More precisely, the volume of the corresponding "posterior bulk" should neither be much larger (to give good prior localization) nor much smaller (such that there is no need to compensate with increased sampling resolution).

We begin with a generic angular and frequency-based decomposition of any EMRI waveform into a small number $M(\boldsymbol{\theta})$ of strong harmonic modes:

$$h(\boldsymbol{\theta}) \approx \sum_{m=1}^{M(\boldsymbol{\theta})} h_m(\boldsymbol{\theta}), \qquad (60)$$

where $M(\boldsymbol{\theta})$ varies slowly over parameter space and really only depends strongly on the orbital eccentricity. For any given analysis, we may fix $M(\boldsymbol{\theta})$ to some maximal value M in practice. Throughout most of this work, we have used the AAK model with $m \equiv j$ and $M = 4$ [see Eq. (17)], but the decomposition (60) is compatible with all other standard approaches in EMRI modeling (see Sec. II A). Note, however, that a decomposition into angular harmonics alone (as done for comparable-mass-binary waveforms) is insufficiently discriminative for our proposal, which relies on there being a distinctive spread in individual-mode optimal SNRs for each source.

Our desired likelihood L' must naturally make use of the discrepancy between the mode structure of a putative signal at each $\boldsymbol{\theta}$, i.e., the vector of mode optimal SNRs,

$$[v_{\text{opt}}(\boldsymbol{\theta})]_m := \rho_{\text{opt}}(h_m(\boldsymbol{\theta})), \qquad (61)$$

with ρ_{opt} as defined in Eq. (12), and the mode information that is actually recovered from the data x at that same point, i.e., the vector of mode detection SNRs,

$$[v_{\text{det}}(\boldsymbol{\theta})]_m := \rho_{\text{det}}(h_m(\boldsymbol{\theta})), \qquad (62)$$

with ρ_{det} as defined in Eq. (13). This allows the point $\boldsymbol{\theta}_{\text{inj}}$ (where $\text{E}[v_{\text{det}}] = v_{\text{opt}}$) to be differentiated from any typical $\boldsymbol{\theta}_{\text{sec}}$ (where only a single component of v is matched). In fact, the discrepancy between v_{opt} and v_{det} is already partly accounted for in the standard log-likelihood (22), which may be written more evocatively as

$$\ln L = \langle x|h \rangle - \frac{1}{2}\langle h|h \rangle - \frac{1}{2}\langle x|x \rangle$$
$$\approx v_{\text{opt}} \cdot v_{\text{det}} - \frac{1}{2}\langle h|h \rangle - \frac{1}{2}\langle x|x \rangle, \qquad (63)$$

where all dependence on $\boldsymbol{\theta}$ is implicit. The approximation in Eq. (63) is due only to the truncation of modes in Eq. (60), whereas $\langle h|h\rangle \approx |v_{\text{opt}}|^2$ if we further assume $\langle h_m|h_{m'}\rangle \approx 0$ for all $m \neq m'$ (which is generally valid).

Computation of the constant final term in Eq. (63) is not needed in practice except for evidence calculations, and the penultimate term varies much more slowly than the first with respect to most parameters that affect phasing. Thus the first term is the main determinant of the likelihood profile over parameter space—its value is closer to ρ_{inj}^2 at $\boldsymbol{\theta}_{\text{sec}}$ and closer to zero at $\boldsymbol{\theta}_{\text{tail}}$. A direct strategy is then to suppress the first term to zero at $\boldsymbol{\theta}_{\text{sec}}$, using a similaritylike quantity $Q\colon\Theta \to [0, 1]$ that depends on some measure of discrepancy between v_{opt} and v_{det}:

$$\ln L' := Q\langle x|h\rangle - \frac{1}{2}\langle h|h\rangle - \frac{1}{2}\langle x|x\rangle$$
$$= \ln L - (1 - Q)\langle x|h\rangle. \qquad (64)$$

The form of Eq. (64) seems to preclude statistically motivated definitions where L' is derived from natural probabilistic statements; nevertheless, such a modification should facilitate EMRI search, at the cost of invalidating the interpretation of L' as a proper Bayesian likelihood.

Our first suggestion for Q is

$$Q_1 := 1 - F_{\chi^2}(|v_{\text{opt}} - v_{\text{det}}|^2), \qquad (65)$$

where F_{χ^2} denotes the cumulative distribution function of a chi-squared random variable with M degrees of freedom. The veto likelihood L' with Q_1 works well to "flatten out" secondaries (see Fig. 22), while the cross-sectional profile of the density bulk around $\boldsymbol{\theta}_{\text{inj}}$ is only marginally narrower at the Gaussian-analogous 3-σ value $\ln L = -9/2$ (and well beyond). Another option for Q is
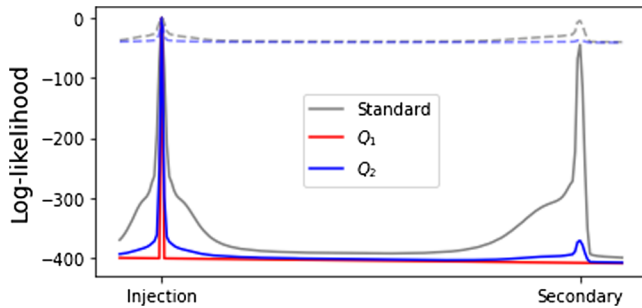

FIG. 22. Standard and veto log-likelihood values along an extended connecting line through the injection parameters and a strong secondary node from Sec. IV A 2—specifically, the set-I node of pair 2 (red point in middle panel of Fig. 11). The approximate inner product (19) is used for clarity, although the general concept holds for the full inner product (8). Dashed curves are annealed versions of the respectively colored standard and veto (Q_2) log-likelihoods, with a temperature of 10.

$$Q_2 := (\hat{v}_{\text{opt}} \cdot \hat{v}_{\text{det}})^{2q}, \qquad (66)$$

where an overhat denotes normalization with respect to the Euclidean inner product on $\mathbb{R}^M$ ($\hat{v} \cdot \hat{v} = 1$). The quantity q is empirically determined; it might be defined to depend on $\boldsymbol{\theta}$ via the properties of v_{opt}, but here we treat it as a tunable hyperparameter. This version of L' is also shown in Fig. 22, for $q = 20$. There are, of course, many other viable choices for the functional form of Q, which only needs to be ≈ 1 near $\boldsymbol{\theta}_{\text{inj}}$ and ≈ 0 everywhere else (or at least near $\boldsymbol{\theta}_{\text{sec}}$). Such veto likelihoods might be combined with other likelihood-smoothing approaches such as annealing or semicoherent filtering (which essentially broadens all likelihood peaks), in order to further aid EMRI search. We leave optimization of the basic concept in Eqs. (64)–(66) for future follow-up studies.

As decompositions of the form (60) are central to EMRI waveform modeling, the set of modes $\{h_m\}$ for the veto likelihood can first be predetermined (in terms of membership), then obtained as a byproduct of template generation during data analysis. Similarly, the inner-product operations $\langle h_m|h_m\rangle$ in v_{opt} (and v_{det}) may be bypassed with a precomputed fit. The veto likelihood is, however, penalized with added online cost through its reliance on v_{det}, which entails M evaluations of $\langle x|h_m\rangle$. All other operations have negligible cost relative to the inner product at full sampling resolution, and thus the final cost of the veto likelihood is approximately M times that of the standard likelihood. This may seem like a significant penalty even for modest M, but the hope is that it will be more than offset by the increased efficiency of search. Sampling algorithms will no longer have to explore a large number of probability peaks en route to the global peak; this in turn will increase the reliability of candidate detections and accelerate the transition to performing inference on individual signals.

VII. CONCLUSION

Self-confusion and degeneracy for EMRIs, as we have defined them in Sec. I B, are both manifestations of the fact that sources with very different parameters can have signals that strongly resemble one another. The former term is used in the context of a finite set of putative EMRI signals that might plausibly be present in LISA data, while the latter refers instead to the continuum of possible EMRI signals described by a physical waveform model. We have shown in Sec. III that self-confusion is unlikely to arise in reality—for a typical set of ≈ 200 detectable sources distributed according to a representative astrophysical model, the root mean square of pairwise overlaps among their two-year signals is $\sim 10^{-3}$. On the other hand, degeneracy is one of the main hindrances to the search for EMRI signals in LISA data, and we have devoted the bulk of this work to its characterization.

Several new analysis tools have been introduced in Sec. II for our study of degeneracy, where existing methods are inadequate. These include an approximate noise-weighted inner product that acts on highly downsampled waveform amplitude/phase trajectories, as well as a bespoke clustering algorithm that uses overlaps along connecting lines in parameter space to define the degree of connectivity between pairs of signal templates. In Sec. IV A, we have conducted an extensive mapping analysis of the posterior surface over parameter space for a representative signal injection, which has yielded a variety of qualitative results. The most notable of these are: (i) secondary posterior peaks are both strong (injection-template overlaps of $\lesssim 0.8$) and numerous ($\sim 10^3$ of them) in an encompassing region that is $\gtrsim 10^{12}$ times larger than the posterior bulk; (ii) they come in different shapes and sizes; and (iii) they are interconnected and thus cannot be localized. Furthermore, the conditions that give rise to secondaries are very natural, as discussed in Sec. IV B, and thus they are expected to be a generic feature of the EMRI signal space.

Our results hold several implications for EMRI data analysis, as we have discussed at length in Sec. V. The fundamental interaction of degeneracy with detector noise or with multiple actual signals is unlikely to cause false positives or negatives in the search for candidate signals, but this may still arise in practice due to the technical limitations of stochastic search methods. Regardless, previously proposed strategies for EMRI search are a promising first step in addressing degeneracy; we have provided a few complementary suggestions of our own in Sec. VI, to rule out false signals and to better identify the source parameters of actual ones. We have, however, mostly limited the scope of the current study on degeneracy to illuminate the problem itself, rather than to develop possible practical solutions. The new tools and perspectives we have introduced here should help to inform and direct follow-up work on the latter.

There is also utility in an improved characterization of the degeneracy problem, although this is arguably unneeded before the advent of efficient and extensive next-generation waveform models. For example, the global distribution of secondaries over the full-dimensional space of intrinsic and extrinsic parameters is still unknown for a single injection—much less a representative set of injections with different source parameters and analysis durations. Scaling up the mapping analysis in this way will require computational enhancements such as fast LISA-response models, alternative algorithms for obtaining a large set of high-overlap points, and GPU acceleration of the sampling and clustering steps. Once fast and fully generic Kerr models with adiabatic evolution schemes and mode content become available, it might be worthwhile to perform comprehensive surveys of secondaries, with the focus shifting to the quantitative characterization of degeneracy for these specific models. Such studies could inform the tuning of sampling algorithms and strategies for EMRI search. Proxy postadiabatic models that incorporate approximate or phenomenological resonant jumps might also be explored using the tools we have introduced, in order to investigate how degeneracy changes in the presence of transient resonances.

ACKNOWLEDGMENTS

[1] R. Abbott *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, Phys. Rev. X **11**, 021053 (2021).

[2] K. Danzmann *et al.*, Laser interferometer space antenna, 2017.

[3] Pau Amaro-Seoane, Relativistic dynamics and extreme mass ratio inspirals, Living Rev. Relativity **21**, 4 (2018).

[4] Stanislav Babak, Jonathan Gair, Alberto Sesana, Enrico Barausse, Carlos F. Sopuerta, Christopher P. L. Berry, Emanuele Berti, Pau Amaro-Seoane, Antoine Petiteau, and Antoine Klein, Science with the space-based interferometer LISA. V: Extreme mass-ratio inspirals, Phys. Rev. D **95**, 103012 (2017).

[5] Christopher P. L. Berry, Scott A. Hughes, Carlos F. Sopuerta, Alvin J. K. Chua, Anna Heffernan, Kelly Holley-Bockelmann, Deyan P. Mihaylov, M. Coleman Miller, and Alberto Sesana, The unique potential of extreme mass-ratio inspirals for gravitational-wave astronomy, 2019.

[6] Pau Amaro-Seoane, Carlos F. Sopuerta, and Marc Dewi Freitag, The role of the supermassive black hole spin in the estimation of the EMRI event rate, Mon. Not. R. Astron. Soc. **429**, 3155 (2013).

[7] Leor Barack and Adam Pound, Self-force and radiation reaction in general relativity, Rep. Prog. Phys. **82,** 016904 (2019).

[8] Adam Pound and Barry Wardell, Black hole perturbation theory and gravitational self-force, 2021.

[9] Adam Pound, Barry Wardell, Niels Warburton, and Jeremy Miller, Second-Order Self-Force Calculation of Gravitational Binding Energy in Compact Binaries, Phys. Rev. Lett. **124,** 021101 (2020).

[10] Alvin J. K. Chua, Michael L. Katz, Niels Warburton, and Scott A. Hughes, Rapid Generation of Fully Relativistic Extreme-Mass-Ratio-Inspiral Waveform Templates for LISA Data Analysis, Phys. Rev. Lett. **126,** 051102 (2021).

[11] Niels Warburton, Adam Pound, Barry Wardell, Jeremy Miller, and Leanne Durkan, Gravitational-Wave Energy Flux for Compact Binaries through Second Order in the Mass Ratio, Phys. Rev. Lett. **127,** 151102 (2021).

[12] Barry Wardell, Adam Pound, Niels Warburton, Jeremy Miller, Leanne Durkan, and Alexandre Le Tiec, Gravitational waveforms for compact binaries from second-order self-force theory, 2021.

[13] Jonathan R. Gair, Leor Barack, Teviet Creighton, Curt Cutler, Shane L. Larson, E. Sterl Phinney, and Michele Vallisneri, Event rate estimates for LISA extreme mass ratio capture sources, Classical Quantum Gravity **21,** S1595 (2004).

[14] Christopher J. Moore, Davide Gerosa, and Antoine Klein, Are stellar-mass black-hole binaries too quiet for LISA?, Mon. Not. R. Astron. Soc. **488,** L94 (2019).

[15] Leor Barack and Curt Cutler, LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy, Phys. Rev. D **69,** 082005 (2004).

[16] Stanislav Babak, Hua Fang, Jonathan R. Gair, Kostas Glampedakis, and Scott A. Hughes, 'Kludge' gravitational waveforms for a test-body orbiting a Kerr black hole, Phys. Rev. D **75,** 024005 (2007); **77,** 049902(E) (2008).

[17] Alvin J. K. Chua, Christopher J. Moore, and Jonathan R. Gair, Augmented kludge waveforms for detecting extreme-mass-ratio inspirals, Phys. Rev. D **96,** 044005 (2017).

[18] Curt Cutler and Eanna E. Flanagan, Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?, Phys. Rev. D **49,** 2658 (1994).

[19] G. Nelemans, L. R. Yungelson, and Simon F. Portegies Zwart, The gravitational wave signal from the galactic disk population of binaries containing two compact objects, Astron. Astrophys. **375,** 890 (2001).

[20] Leor Barack and Curt Cutler, Confusion noise from LISA capture sources, Phys. Rev. D **70,** 122002 (2004).

[21] Matteo Bonetti and Alberto Sesana, Gravitational wave background from extreme mass ratio inspirals, Phys. Rev. D **102,** 103023 (2020).

[22] Etienne Racine and Curt Cutler, Gaussianity of LISA's confusion backgrounds, Phys. Rev. D **76,** 124033 (2007).

[23] Tyson Littenberg, Neil Cornish, Kristen Lackeos, and Travis Robson, Global analysis of the gravitational wave signal from galactic binaries, Phys. Rev. D **101,** 123021 (2020).

[24] Neil J. Cornish, Detection strategies for extreme mass ratio inspirals, Classical Quantum Gravity **28,** 094016 (2011).

[25] Stanislav Babak, Jonathan R. Gair, and Edward K. Porter, An Algorithm for detection of extreme mass ratio inspirals in LISA data, Classical Quantum Gravity **26,** 135004 (2009).

[26] Stanislav Babak *et al.*, The mock LISA data challenges: From challenge 3 to challenge 4, Classical Quantum Gravity **27,** 084009 (2010).

[27] Ryuichi Fujita and Masaru Shibata, Extreme mass ratio inspirals on the equatorial plane in the adiabatic order, Phys. Rev. D **102,** 064005 (2020).

[28] Soichiro Isoyama, Ryuichi Fujita, Alvin J. K. Chua, Hiroyuki Nakano, Adam Pound, and Norichika Sago, Adiabatic waveforms from extreme-mass-ratio inspirals: an analytical approach, 2021.

[29] Scott A. Hughes, Niels Warburton, Gaurav Khanna, Alvin J. K. Chua, and Michael L. Katz, Adiabatic waveforms for extreme mass-ratio inspirals via multivoice decomposition in time and frequency, Phys. Rev. D **103,** 104014 (2021).

[30] Jeremy Miller and Adam Pound, Two-timescale evolution of extreme-mass-ratio inspirals: Waveform generation scheme for quasicircular orbits in Schwarzschild spacetime, Phys. Rev. D **103,** 064048 (2021).

[31] Jonathan R Gair and Kostas Glampedakis, Improved approximate inspirals of test-bodies into Kerr black holes, Phys. Rev. D **73,** 064037 (2006).

[32] P. C. Peters and J. Mathews, Gravitational radiation from point masses in a keplerian orbit, Phys. Rev. **131,** 435 (1963).

[33] Alvin J. K. Chua and Jonathan R. Gair, Improved analytic extreme-mass-ratio inspiral model for scoping out eLISA data analysis, Classical Quantum Gravity **32,** 232002 (2015).

[34] Massimo Tinto and Sanjeev V. Dhurandhar, Time-delay interferometry, Living Rev. Relativity **8,** 4 (2005).

[35] Alvin J. K. Chua, Natalia Korsakova, Christopher J. Moore, Jonathan R. Gair, and Stanislav Babak, Gaussian processes for the interpolation and marginalization of waveform error in extreme-mass-ratio-inspiral parameter estimation, Phys. Rev. D **101,** 044027 (2020).

[36] Curt Cutler, Angular resolution of the LISA gravitational wave detector, Phys. Rev. D **57,** 7089 (1998).

[37] Steve Drasco and Scott A. Hughes, Gravitational wave snapshots of generic extreme mass ratio inspirals, Phys. Rev. D **73,** 024027 (2006); **88,** 109905(E) (2013); **90,** 109905(E) (2014).

[38] Scott A. Hughes, The evolution of circular, nonequatorial orbits of Kerr black holes due to gravitational wave emission, Phys. Rev. D **61,** 084004 (2000); **63,** 049902 (E) (2001); **65,** 069902(E) (2002); **67,** 089901(E) (2003); **78,** 109902(E) (2008); **90,** 109904(E) (2014).

[39] Josh Mathews, Adam Pound, and Barry Wardell, Self-Force Calculations with a Spinning Secondary, 2021.

[40] Lisa V. Drummond and Scott A. Hughes, Precisely computing bound orbits of spinning bodies around black holes I: General framework and results for nearly equatorial orbits, 2022.

[41] Lisa V. Drummond and Scott A. Hughes, Precisely computing bound orbits of spinning bodies around black holes II: Generic orbits, 2022.

[42] Gabriel Andres Piovano, Andrea Maselli, and Paolo Pani, Model independent tests of the Kerr bound with extreme mass ratio inspirals, Phys. Lett. B **811,** 135860 (2020).

[43] Gabriel Andres Piovano, Andrea Maselli, and Paolo Pani, Extreme mass ratio inspirals with spinning secondary: A detailed study of equatorial circular motion, Phys. Rev. D **102,** 024041 (2020).

[44] Ondřej Zelenka, Georgios Lukes-Gerakopoulos, Vojtěch Witzany, and Ondřej Kopáček, Growth of resonances and chaos for a spinning test particle in the Schwarzschild background, Phys. Rev. D **101,** 024037 (2020).

[45] Wolfram Schmidt, Celestial mechanics in Kerr space-time, Classical Quantum Gravity **19,** 2743 (2002).

[46] EMRI Kludge Suite, v0.5.2, https://github.com/alvincjk/EMRI_Kludge_Suite.

[47] Michael L. Katz, Alvin J. K. Chua, Lorenzo Speri, Niels Warburton, and Scott A. Hughes, FastEMRIWaveforms: New tools for millihertz gravitational-wave data analysis, 2021.

[48] Fast EMRI Waveforms, v1.3.6, https://bhptoolkit.org/FastEMRIWaveforms. Code releases with individual DOI are available at 10.5281/zenodo.4005001.

[49] LISA Science Requirements, https://www.cosmos.esa.int/documents/678316/1700384/SciRD.pdf.

[50] Antoine Petiteau, Gerard Auger, Hubert Halloin, Olivier Jeannin, Eric Plagnol, Sophie Pireaux, Tania Regimbau, and Jean-Yves Vinet, LISACode: A scientific simulator of LISA, Phys. Rev. D **77,** 023002 (2008).

[51] Travis Robson, Neil J. Cornish, and Chang Liu, The construction and use of LISA sensitivity curves, Classical Quantum Gravity **36,** 105011 (2019).

[52] Michele Vallisneri, Use and abuse of the fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects, Phys. Rev. D **77,** 042001 (2008).

[53] Lee S. Finn, Detection, measurement and gravitational radiation, Phys. Rev. D **46,** 5236 (1992).

[54] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Optimization by simulated annealing, Science **220,** 671 (1983).

[55] J. Wang and R. H. Swendsen, Replica Monte Carlo simulation (revisited), Prog. Theor. Phys. Suppl. **157,** 317 (2005).

[56] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman, emcee: The MCMC Hammer, Publ. ASP **125,** 306 (2013).

[57] W. J. Handley, M. P. Hobson, and A. N. Lasenby, POLY-CHORD: Next-generation nested sampling, Mon. Not. R. Astron. Soc. **453,** 4385 (2015).

[58] Joshua S. Speagle, DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493,** 3132 (2020).

[59] Leonard Kaufman and Peter J Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons, New York, 2009), Vol. 344.

[60] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek, Density-based clustering, Wiley Interdiscip. Rev. **1,** 231 (2011).

[61] AV Arkhangel'Skii and VV Fedorchuk, *General Topology I: Basic Concepts and Constructions Dimension Theory* (Springer Science & Business Media, New York, 2012), Vol. 17.

[62] R. Diestel, *Graph Theory*, Graduate Texts in Mathematics (Springer, New York, 2000).

[63] Ulrike Von Luxburg, A tutorial on spectral clustering, Stat. Comput. **17,** 395 (2007).

[64] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction To Algorithms*, Introduction to Algorithms (MIT Press, Cambridge, MA, 2001).

[65] Pau Amaro Seoane *et al.*, The Effect of Mission Duration on LISA Science Objectives, 2021.

[66] P. J. E. Peebles, *Principles of Physical Cosmology*, Princeton Series in Physics (Princeton University Press, Princeton, NJ, 1993).

[67] Enrico Barausse, The evolution of massive black holes and their spins in their galactic hosts, Mon. Not. R. Astron. Soc. **423,** 2533 (2012).

[68] C. Cutler and J. Harms, BBO and the neutron-star-binary subtraction problem, Phys. Rev. D **73,** 042001 (2006).

[69] Lorenzo Speri, Andrea Antonelli, Laura Sberna, Stanislav Babak, Enrico Barausse, Jonathan R. Gair, and Michael L. Katz, Measuring accretion-disk effects with gravitational waves from extreme mass ratio inspirals, 2022.

[70] Justin Ellis and Rutger van Haasteren, jellis18/ptmcmcsampler: Official release, 2017.

[71] Eanna E. Flanagan and Tanja Hinderer, Transient Resonances in the Inspirals of Point Particles into Black Holes, Phys. Rev. Lett. **109,** 071102 (2012).

[72] Uchupol Ruangsri and Scott A. Hughes, Census of transient orbital resonances encountered during binary inspiral, Phys. Rev. D **89,** 084036 (2014).

[73] Christopher P. L. Berry, Robert H. Cole, Priscilla Cañizares, and Jonathan R. Gair, Importance of transient resonances in extreme-mass-ratio inspirals, Phys. Rev. D **94,** 124042 (2016).

[74] Lorenzo Speri and Jonathan R. Gair, Assessing the impact of transient orbital resonances, Phys. Rev. D **103,** 124032 (2021).

[75] Michael L. Katz, Sylvain Marsat, Alvin J. K. Chua, Stanislav Babak, and Shane L. Larson, GPU-accelerated massive black hole binary parameter estimation with LISA, Phys. Rev. D **102,** 023033 (2020).

[76] Sylvain Marsat, John G. Baker, and Tito Dal Canton, Exploring the Bayesian parameter estimation of binary black holes with LISA, Phys. Rev. D **103,** 083011 (2021).

[77] Peter J. Green, Reversible jump Markov Chain Monte Carlo computation and bayesian model determination, Biometrika **82,** 711 (1995).

[78] Tyson B. Littenberg and Neil J. Cornish, A Bayesian approach to the detection problem in gravitational wave astronomy, Phys. Rev. D **80,** 063007 (2009).

[79] W. D. Vousden, W. M. Farr, and I. Mandel, Dynamic temperature selection for parallel tempering in Markov Chain Monte Carlo simulations, Mon. Not. R. Astron. Soc. **455,** 1919 (2016).

[80] David J. Earl and Michael W. Deem, Parallel tempering: Theory, applications, and new perspectives, Phys. Chem. Chem. Phys. (Incorporating Faraday Transactions) **7,** 3910 (2005).

[81] John Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1,** 833 (2006).

[82] Patrick R. Brady and Teviet Creighton, Searching for periodic sources with LIGO. 2. Hierarchical searches, Phys. Rev. D **61**, 082001 (2000).

[83] Curt Cutler, Iraj Gholami, and Badri Krishnan, Improved stack-slide searches for gravitational-wave pulsars, Phys. Rev. D **72**, 042004 (2005).

[84] Yan Wang, Yu Shang, Stanislav Babak, Yu Shang, and Stanislav Babak, EMRI data analysis with a phenomenological waveform, Phys. Rev. D **86**, 104050 (2012).

[85] Jonathan R. Gair, Ilya Mandel, and Linqing Wen, Improved time-frequency analysis of extreme-mass-ratio inspiral signals in mock LISA data, Classical Quantum Gravity **25**, 184031 (2008).

[86] Theocharis A. Apostolatos, Search templates for gravitational waves from precessing, inspiraling binaries, Phys. Rev. D **52**, 605 (1995).

[87] Curt Cutler and Michele Vallisneri, LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms, Phys. Rev. D **76**, 104018 (2007).

[88] Scott E. Field, Chad R. Galley, Frank Herrmann, Jan S. Hesthaven, Evan Ochsner, and Manuel Tiglio, Reduced Basis Catalogs for Gravitational Wave Templates, Phys. Rev. Lett. **106**, 221102 (2011).

[89] Scott E. Field, Chad R. Galley, Jan S. Hesthaven, Jason Kaye, and Manuel Tiglio, Fast Prediction and Evaluation of Gravitational Waveforms Using Surrogate Models, Phys. Rev. X **4**, 031006 (2014).

[90] Nur E. M. Rifat, Scott E. Field, Gaurav Khanna, and Vijay Varma, Surrogate model for gravitational wave signals from comparable and large-mass-ratio black hole binaries, Phys. Rev. D **101**, 081502 (2020).