Selection Effects in Periodic X-Ray Data from Maximizing Detection Statistics

Reed Essick

Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, Ontario, N2L 2Y5, Canada; reessick@perimeterinstitute.ca

Received 2021 November 11; revised 2022 January 30; accepted 2022 February 1; published 2022 March 15

Abstract

The Neutron Star Interior Composition Explorer (NICER) records exceptional data on pulsars’ energy-dependent X-ray pulse profiles. However, in searching for evidence of pulsations, Guillot et al. (2019) introduce a procedure to select an ordered subset of data that maximizes a detection statistic (the H-test). I show that this can degrade subsequent analyses using an idealized model with stationary expected count rates from both noise and signal. Specifically, the data-selection procedure biases the inferred mean count rate to be too low and the inferred pulsation amplitude to be too high, and the size of these biases scales strongly with the amount of data that is rejected and the true signal amplitude. The procedure also alters the H-test’s null distribution, rendering nominal significance estimates overly optimistic. While the idealized model does not capture all the complexities of real NICER data, it suggests that these biases could be important for NICER’s observations of J0740+6620 and other faint pulsars (observations of J0030+0451 are likely less affected). I estimate that these effects may introduce a bias of $\mathcal{O}(10\%)$ on average in the inferred modulation depth of lightcurves like J0740+6620’s, and may be as large as $\mathcal{O}(50\%)$ for fainter pulsars. However, the change for a single data set like J0740+6620 is expected to be a shift between $-5\%$ and $+20\%$. This could imply that the lower limit on J0740+6620’s radius is slightly larger than it should be, although preliminary investigations suggest the radius constraints shift to larger radii by $\mathcal{O}(1\%)$ with the same overall statistical precision using real J0740+6620 data.

Unified Astronomy Thesaurus concepts: X-ray astronomy (1810); Neutron stars (1108); Compact objects (288); Nuclear astrophysics (1129); Pulsars (1306)

1. Introduction

Over the past few years, the Neutron Star Interior Composition Explorer (NICER; Gendreau et al. 2016; Prigozhin et al. 2016) released simultaneous measurements of the mass and radius of two X-ray pulsars: J0030+0451 (Miller et al. 2019; Riley et al. 2019) and J0740+6620 (Miller et al. 2021; Raaijmakers et al. 2021; Riley et al. 2021; Wolff et al. 2021). These observations inform the understanding of Neutron Star (NS) composition and structure, primarily through inferred constraints on the Equation of State (EoS) of extremely dense matter. Together with Gravitational Wave (GW) observations of the coalescence of compact binaries containing NSs, such as the observation of GW170817 (Abbott et al. 2017, 2018, 2019) with the advanced LIGO (Aasi et al. 2015) and Virgo (Acernese et al. 2014) interferometers, NICER X-ray observations set some of the most stringent constraints available to-date on the properties of matter at supranuclear densities ($\gtrsim 2.8 \times 10^{14}$ g cm$^{-3}$). See, e.g., Chatziioannou (2020) for a review.

Several authors have also remarked on the fact that the radii inferred from NICER observations tend to lie at the upper end of what is consistent with constraints from GW170817. For example, Figure 3 in Landry et al. (2020) and Figure 3 in Legred et al. (2021) both show the inferred masses and radii of individual objects from X-ray observations when considered separately and when considered jointly with GW data. For both pulsars, the X-ray data alone tend to prefer larger radii, as reported. This may just be a statistical fluctuation, as the posteriors are still broad, there are only a handful of events, and there is no real tension between GW and X-ray observations. However, as I will show, NICER’s current data-selection procedure may introduce an artificial preference for larger radii. To the best of my knowledge, this effect has not been previously described in the literature. As the number of observed systems continues to grow, it will be important to quantify all possible selection effects and associated biases if one is to obtain precise (and accurate!) constraints on the EoS.

In particular, recent constraints on J0740+6620 (the most massive NS with a confident mass measurement to-date) suggest $R \gtrsim 12$ km for $M \approx 2.08 M_\odot$ (Miller et al. 2021). This is a relatively large value, and it is of interest to know how robust the lower limit for the radius is. Miller et al. (2021) discuss this in some detail, concluding that their lower limit should not change much due to possible systematic errors. However, they do not consider the possible impact of selection effects within the data-selection procedure.

This article’s aim is to explain why there may be need for concern with current NICER procedures by examining an idealized model. I stress that this does not prove there are issues with any of the published constraints, but hope that it motivates further study to ensure the NICER observations can be utilized to their full potential.

I begin in Section 2 by reviewing the detection statistic employed by NICER to identify periodic signals: the H-test (de Jager et al. 1989; de Jager & Büsching 2010). I then review the data-selection procedure adopted by the NICER collaboration for J0740+6620 as described in Guillot et al. (2019) and Wolff et al. (2021). In particular, I examine the possible impact of the choice to retain only the subset of data that maximizes the H-test. For example, de Jager et al. (1989) note that the H-test is more sensitive to signals with larger relative oscillation amplitudes and narrower beaming patterns. By selecting the data that maximize the significance of the inferred periodicity
in the (unknown) lightcurve, the inferred size of oscillations in the selected data may be artificially inflated. In turn, this could lead to a smaller inferred compactness and larger inferred radius at a fixed mass. Section 4 investigates this quantitatively within the context of an idealized model. I find that Guillot et al. (2019)’s data-selection procedure biases the inferred signal parameters to smaller phase-averaged rates and larger modulation depths. These effects are larger for weaker signals. The data-selection procedure also alters the null distribution of the H-test, implying that standard significance estimates cannot be used. I summarize my findings in Section 5, quantitatively estimating the implications for current measurements of J0030+0451 and J0740+6620 in turn.

2. The H-test for Weak, Periodic Signals

I first review the H-test and its use in detecting periodic signals with unknown lightcurves in sparse X-ray observations (fundamental period of the lightcurve is much shorter than the average time between events). Readers should refer to de Jager et al. (1989) and de Jager & Büsching (2010) for more details.

The H-test provides a robust null test for uniformity on a circle, and as such, is particularly useful when searching for the presence of periodic features in phase-folded X-ray data. de Jager et al. (1989) showed that the H-test is typically as sensitive as, or more sensitive than, other tests when the signal is small and the lightcurve is unknown. The H-test works by considering deviations in the lightcurve \( f(\phi) \) from a uniform distribution, defining

\[
\psi(f) = \int_0^{2\pi} d\phi \left( f(\phi) - \frac{1}{2\pi} \right)^2
\]

and modeling the observed lightcurve with a Fourier Series Estimator with \( n \) harmonics

\[
\hat{f}_m = \frac{1}{2} \left( 1 + \sum_{k=1}^m (\hat{\alpha}_k \cos(k\phi) + \hat{\beta}_k \sin(k\phi)) \right),
\]

with the Fourier coefficients estimated via Monte Carlo sums over the set of \( n \) observed phase data \( \{\phi_i\} \)

\[
\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n \cos(k\phi_i)
\]

\[
\hat{\beta}_k = \frac{1}{n} \sum_{i=1}^n \sin(k\phi_i).
\]

For a given harmonic \( m \), define the deviation statistic for a set of phase observations

\[
Z_m^2(\{\phi_i\}) = 2\pi n \psi(\hat{f}_m) = 2n \sum_{k=1}^m [\hat{\alpha}_k^2 + \hat{\beta}_k^2].
\]

The H-test is then defined as a maximization over different harmonics

\[
H(\{\phi_i\}) = \max_{1 \leq m \leq 20} [Z_m^2(\{\phi_i\}) - 4m + 4].
\]

If \( H \) is sufficiently large, then the null hypothesis that the data are uniformly distributed (no periodicity) is rejected, and the presence of a periodic signal is inferred. de Jager & Büsching (2010) find that the survival function for \( H \) is exponential:

\[
P(H > H_{obs}) = \exp(-0.4H_{obs}).
\]

Additionally, de Jager et al. (1989) discuss possible selection effects associated with the H-test. They note that the H-test will tend to favor sources that have deeper relative oscillation amplitudes and narrower radiation beams (lower pulse duty cycles). In the context of NICER’s observations, both effects will tend to prefer less compact stars (less self-lensing implies deeper oscillations and narrower beams). Selecting subsets of data based on what maximizes the H-test may introduce a preference for data that minimizes the inferred stellar compactness (maximizes the inferred radius at a fixed mass).

3. Review of NICER’s Data-selection Procedure

NICER is sensitive to soft X-rays (0.2–12 keV) and provides both exceptional energy and timing resolution (Gendreau et al. 2016; Prigozhin et al. 2016). Typical timing uncertainty is <100 ns, which is much less than the rotation period of any observed pulsar. This provides a measurement of the energy-dependent pulsation pattern of X-ray sources by directly measuring the energy and the time of arrival of individual photons. The instrument is mounted on the International Space Station (ISS) and, given its environment (location of the ISS, space weather, etc.) and current target (pointing with respect to the Sun, Earth, and Moon, etc.), is subject to a variety of possible noise sources. The noise sources can vary over time, and as such, not all data recorded by NICER are equivalent. For this reason, the NICER team divides the total observation into small segments (each spanning 10–100 s; Guillot et al. 2019), calling each segment a good time interval (GTI). As described in Guillot et al. (2019) and elsewhere, the final set of GTIs included in downstream analyses for some pulsars is chosen by maximizing the H-test significance (e.g., this procedure was used for J0740+6620 but not J0030+0451).

Guillot et al. (2019) define a procedure for ordering GTIs by rough estimates of their backgrounds (GTIs with smaller estimated backgrounds first) and including only those GTIs that increase the cumulative H-test significance (see their Figure 1). Specifically, they assume that the total count rate is dominated by the background (signals are weak), and they order GTIs by the ratio of their total number of counts and their durations. GTIs are then included cumulatively until the corresponding value of \( H \) is maximized. Guillot et al. (2019) advocate discarding all remaining data and only retaining the ordered subset that maximizes \( H \). Figure 1 shows an example trajectory of \( H \) in the idealized model as more data are included, with a maximum reached with only \( \approx 80% \) of the data.

This ordering is intended to include GTIs with smaller intrinsic background rates first. However, I show in Section 4 that this procedure causes several issues even in idealized situations. It is reasonable to expect these issues to persist in more complicated situations (see Appendix), although further study is needed.

Guillot et al. (2019) additionally perform a grid search over which energies are considered (between 0.2 and 2.0 keV) in order to further maximize the detection significance (see their Table 3). I do not consider the implications of such an optimization in this work, as I de facto consider a single energy channel. However, it is likely that optimization over the energy channels could introduce additional biases.
4. Data Selection for Sparse X-Ray Observations

I now consider an idealized Poisson model for sparse X-ray observations. That is, I consider differential Poisson rates for the signal and background as a function of rotation phase,

\[ \lambda_i = \text{constant} \]

\[ = a_0 + \sum_{m=1}^{M} a_m \cos(m\phi + \delta_m), \]

where I assume the noise is uncorrelated with the rotation phase and express the signal as a periodic function with \( M \) harmonics. Furthermore, the expected rate of counts of type \( i \) in the interval \([\phi_1, \phi_2]\) is

\[ \Lambda_i(\phi_1, \phi_2) = \int_{\phi_1}^{\phi_2} d\phi \lambda_i(\phi), \]

and the corresponding expected number of counts within the same phase interval throughout an observation with duration \( T \) is \( \Lambda_i(\phi_1, \phi_2)T \). I then model the likelihood for \( N_{\text{phs}} \) binned phase measurements via

\[ \log p(c_i | \lambda_n, \lambda_s) = \sum_{m=1}^{N_{\text{phs}}} \log(\Lambda(\phi_i, \phi_{i+1})T) - \log(c_i!) \]

\[ - \Lambda(0, 2\pi)T, \]

where \( c_i \) is the number of observed counts within \([\phi_i, \phi_{i+1})\), \( \phi_0 = 0, \phi_{N_{\text{phs}}+1} = 2\pi \), and

\[ \Lambda(\phi_1, \phi_2) = \Lambda_n(\phi_1, \phi_2) + \Lambda_s(\phi_1, \phi_2) \]

is the total expected rate of events between \( \phi_1 \) and \( \phi_2 \) from both signal and noise.

Furthermore, I consider a sequence of \( N_{\text{seg}} \) independent, identically distributed (i.i.d.) segments, each with duration \( \tau \) such that \( T = \tau N_{\text{seg}} \). I proceed by generating random realizations of observed data from this model separately for each segment, and then consider the impact on the global inference when I employ various data-selection algorithms.

Throughout the following, I assume \( \tau = 100 \text{ s} \) for all segments and \( \tau \lambda_0 = 50 \). I investigate various signal strengths ranging between \( \tau a_0 = 0.1 \) and 10. On average, then, each segment contains \( O(2\pi \cdot 50) \) counts. For simplicity, I consider a single injected lightcurve with three nonzero harmonics:

\[ a_1/a_0 = 0.3, \delta_1 = 0.0 \]

\[ a_2/a_0 = 0.5, \delta_2 = 0.0 \]

\[ a_4/a_0 = 0.2, \delta_4 = 0.75, \]

and scale \( a_0 \) to control the overall signal strength. Figure 1 shows an example realization with \( N_{\text{seg}} = 1000 \) and \( \tau a_0 = 1 \).

The main text focuses on experiments with a constant noise rate. The Appendix discusses possible implications of variable noise rates in more detail.

4.1. Biases and Correlations

To begin, I consider the impact of the data-selection process on the inferred signal parameters by computing maximum-likelihood estimates (MLEs) of the signal parameters for many realizations of the experiment. For computational expediency, I assume \( a_m \ll \lambda_n + a_0 \forall m \geq 1 \) and expand Equation (11) to second order in \( a_m \), which allows me to analytically determine MLEs for \( \lambda_0 = \lambda_n + a_0, a_m \cos(\delta_m)/\lambda_0, \) and \( a_m \sin(\delta_m)/\lambda_0 \). I report \( \hat{\delta}_m \geq 0 \) derived from the MLEs in this alternate parameterization.

There is a strong correlation between the number of segments retained after the data-selection procedure \( (N_{\text{seg}}^\text{ret}) \) and \( \lambda_0 \). That is, I infer smaller \( \hat{\lambda}_0 \) if I retain fewer segments. This makes sense, as Guillot et al. (2019)’s data-selection procedure first orders segments by increasing average count rate. Therefore, unless all the data are included, the procedure preferentially selects the data with fewer counts. This is strongly correlated with \( \lambda_0 \), as

\[ \hat{\lambda}_0 = \frac{1}{2\pi} \sum_{i} c_i \]
is directly proportional to the total number of counts. Figure 1 demonstrates this with binned phase data from a single experiment. Figure 2 shows the correlations between $N_{\text{seg}}^{(\text{ret})}$ and the inferred lightcurve model for many experiments.

Figure 2 demonstrates additional correlations between $N_{\text{seg}}^{(\text{ret})}$ and $\hat{a}_m$. As fewer segments are retained, one tends to systematically overestimate $a_m$. This can be interpreted as the H-test’s preference for subsets of data that maximize the signal’s modulation. That is, when $N_{\text{seg}}^{(\text{ret})}$ is small, both the numerator and the denominator in the ratio $\hat{a}_m / \lambda_0$ are biased, making the ratio larger. This ratio defines the relative size of the oscillations within the lightcurve. Furthermore, the H-test’s estimate for the optimal number of harmonics to include is often larger than the true number of harmonics when $N_{\text{seg}}^{(\text{ret})}$ is low. Maximizing the H-test, then, selects data that appear to contain signals that are both larger and more complex than the true signal.

In line with this picture, I do not observe a large bias in $\hat{\delta}_m$. Even though particular combinations of phases may make the lightcurve narrower and therefore easier to detect, this does not appear to be easily achieved with random noise realizations. As such, $\delta_m$ does not affect the detectability of the signal as much as $a_m$ does.

Because I find a strong correlation between the inferred mean count rate and the amount of data retained when segments are first ordered by the number of counts they contain, I also explore a possible alternative in Figure 2. I consider first randomly ordering segments (as opposed to ordering by increasing count) and then cumulatively including segments until $H$ is maximized. While the bias in $\hat{\delta}_0$ is lessened (except in a few extreme cases), I still observe a bias in $\hat{a}_m$, that scales with $N_{\text{seg}}^{(\text{ret})}$ (compare ordered and unordered subsets in Figure 2).

In order to better understand the persistent bias in $\hat{a}_m$ introduced by H-test maximization, I consider the distribution of $\hat{a}_m$ conditioned on $N_{\text{seg}}^{(\text{ret})}$ in Figure 3. I expect any bias to grow monotonically, as I retain smaller fractions of the data based on the joint distributions in Figure 2. By estimating the distributions $p(\hat{a}_m | N_{\text{seg}}^{(\text{ret})} = 10^3)$, I quantify the smallest the bias could be. Figure 3 shows this when $a_0 / \lambda_0 = 2\%$. I see that both the ordered and unordered subsets demonstrate nearly identical distributions, and that these distributions are biased toward larger $\hat{a}_m$ than when I always use all the data. That is, the mean of the distribution is larger than it would be if I always included all the data. I define the relative bias as the ratio of expectation values

$$b(\hat{a}_m | N_{\text{seg}}^{(\text{ret})}) = \frac{\langle \hat{a}_m | N_{\text{seg}}^{(\text{ret})} \rangle}{\langle \hat{a}_m | \text{All Data} \rangle} - 1$$

and estimate the size of the bias as a function of $N_{\text{seg}}^{(\text{ret})}$ in Figure 3.

Put another way, even if the H-test maximization procedure instructs one to retain all the data, the resulting estimator is still biased relative to what one would obtain if they always included all the data. As such, I conclude that H-test maximization is guaranteed to introduce a bias in $\hat{a}_m$ regardless of whether any data are actually discarded.

I find that the mean of my estimators are $\approx 7\%$ larger than when I always include all the data with $N_{\text{seg}}^{(\text{ret})} = 10^3$ and $a_0 / \lambda_0 = 2\%$. The bias grows as $N_{\text{seg}}^{(\text{ret})}$ decreases, and I find biases of $\gtrsim 10\%$ when I retain 90\% of the data and $O(30\%)$ when I retain only half the data.

4.2. Scaling of Bias with Signal Strength

The size of the bias scales strongly with the fraction of data that is retained ($N_{\text{seg}}^{(\text{ret})} / N_{\text{seg}}$). Typically, I find that maximizing $H$ for data containing larger injected signals tends to retain larger fractions of the data. This makes sense, as larger signals are more easily detected in individual segments, and therefore it

---

1. My estimator for $\hat{a}_m$ is positive semidefinite and its distribution is skew right. It therefore predicts larger values on average than the true injection, even when I always include all the data. This effect is larger for weaker signals. For this reason, I define the size of the biases introduced by H-test maximization by normalizing them against the mean obtained when I always use all the data. This avoids overstating the effect, as might have been the case if they were normalized by the injected value.
is more likely that each additional segment will increase $H$. Conversely, maximizing $H$ for data with weak signals tends to discard larger fractions of data and introduce larger biases, on average.

Figure 4 shows the distributions for the retained fraction of segments for different signal strengths. Data sets with louder signals suffer less from the biases introduced by the data-selection process, simply because one is less likely to discard large fractions of the data. Nonetheless, biases are always present, as the procedure still discards at least some of the data some of the time, regardless of the signal strength.

Furthermore, the type of persistent bias shown in Figure 3 occurs for all signal strengths, but it is larger for weaker signals. My simplified model predicts a bias of $\approx 7\%$ when $N_{\text{seg}}^{\text{ret}} = 10^3$ with $a_0/\lambda_n = 2\%$. This bias is exacerbated when $N_{\text{seg}}^{\text{ret}}$ decreases. The minimum bias is also larger when the signal is weaker, closer to $O(50\%)$ when $a_0/\lambda_n = 0.2\%$.

### 4.3. Maximization Alters the Null Distribution

Finally, the maximization process changes the null distribution of the $H$ statistic. de Jager & Büsching (2010) show that the $H$ statistic is exponentially distributed when all data are included and there is no signal present (Equation (7)). Figure 5 recovers this prediction within my idealized model. However, the maximized $H$-test statistic produced by Guillot et al. (2019)’s data-selection procedure follows a different distribution, even in the absence of a signal. This means that...
significance estimates corresponding to values of \( H \) obtained via Guillot et al. (2019)’s procedure cannot use the nominal distribution proposed by de Jager & Büsching (2010). Indeed, in the absence of a signal, I find an average nominal \( p \)-value of 5.4% using the distribution of \( H \) produced by Guillot et al. (2019)’s procedure. Ten percent of the time, the nominal \( p \)-value is less than 0.3%. Out of \( 10^7 \) experiments, the smallest nominal \( p \)-value observed by maximizing \( H \) was \( 3.2 \times 10^{-5} \), nearly two orders of magnitude smaller than it should have been.

Therefore, one must take care when assessing the significance of possible signals using the maximized H-test. While the value of \( H \) is increased by the data-selection procedure, this does not necessarily mean that it is more likely there is a signal present in the data. Instead, the apparent increase in \( H \) may be due only to the reordering and downselection of GTIs.

In general, there is still a shift toward larger \( H \) when signals are present. However, the relative importance is smaller for stronger signals.

5. Discussion

While the results of Section 4 suggest current data-selection procedures require reexamination, I reiterate that real NICER data are more complex than my idealized model. The precision made available by NICER observations will continue to be invaluable for a broad community, and the goal of this Letter is only to help maximize the scientific value of those observations.

Nonetheless, the idealized model presented in Section 4 shows that, even in the best-case scenario, the current data-selection procedure can lead to biases in the overall inferred count rate and the morphology of the inferred signal. One might expect these effects to scale approximately as \( N_{\text{count}}^{-1/2} \) and therefore be less than 1% for current NICER observations with \( \gtrsim 10^5 \) counts. However, the total signal can be of the same order of magnitude for dim pulsars like J0740+6620 (signal amplitude is only a few percent). The bias introduced in the total inferred count rate may be a significant fraction of the true signal. That being said, a large bias in the total inferred count rate does not necessarily imply the inferred signal count rate is significantly biased.

Nonetheless, I find that selecting a subset of data that maximizes the H-test detection statistic can bias estimators toward larger apparent modulations with phase, discard larger fractions of data for weaker signals, and change the H-test’s null distribution, rendering standard significance estimates overly optimistic. None of these effects are present if all available data are always used, and they are all directly attributable to the H-test maximization procedure.

I discuss these effects in the context of current NICER data in Section 5.1 before speculating on possible remedies in Section 5.2. I again note that, while I focus only on the selection of GTIs, Guillot et al. (2019) also optimize the H-test over energy bands. Further consideration should be given to the possible impact of optimization over both the energy band and GTIs.

5.1. Impact on Published NICER Observations

I again stress that real NICER data are more complicated than my idealized model. While it would be foolhardy to attempt anything but the roughest error propagation from my model to the actual NICER results, it is still worth pointing out a few things.

5.1.1. J0740+6620

First, it is worth noting that Wolff et al. (2021) state, “[t]he GTI selection is particularly effective for faint pulsars with count rates \( \sim 0.01 \text{ c/s} \) as PSR J0740+6620.” This matches my observation that the bias and the increase in the H-test significance are more important for small signals. Wolff et al. (2021) also states that the total set of data for J0740+6620 originally had 574,405 events between 0.31 and 1.22 keV and the downselected set had only 521,004 events between 0.31 and 1.18 keV. This corresponds to a retained fraction of \( \approx 90.7\% \) of the counts. Assuming the signal is only a few percent of the noise (like in Figure 2), then this might corresponds to a bias on \( \lambda_n \) that is roughly 50% of the size of the true signal’s mean rate \( a_0 \). However, some (or all) of this bias could be absorbed by the noise model (smaller \( \lambda_n \)).

Indeed, within the context of J0740+6620, XMM-Newton (Jansen et al. 2001) observations were used to estimate
a_0 separately from the NICER data. That is, assuming completely separate noise models for XMM and NICER, the analysis used historical XMM blank-sky observations to estimate the energy-dependent noise rate for XMM (\(\lambda^{(\text{XMM})}_0\)) and observations of J0740+6620 to estimate the total count rate (signal and noise) from the pulsar (\(\lambda^{(\text{XMM})}_0\)). Assuming there are no biases in XMM data’s selection, this should provide an unbiased estimate for the mean signal rate \(\hat{a}_m^{(\text{XMM})} = \lambda^{(\text{XMM})}_0 - \lambda^{(\text{XMM})}_n\).

While Guillot et al. (2019)’s data-selection procedure may still bias NICER’s estimate of the total mean count rate \(\hat{\lambda}^{(\text{NICER})}_0\), that bias could be completely accommodated by the flexible noise model assumed within the NICER analysis. That is, the bias in the mean count rate observed in Figure 2 can be absorbed in \(\hat{\lambda}^{(\text{NICER})}_0\) without affecting the constraint on \(a_0\) from XMM. As such, one may suspect that the main impact of the Guillot et al. (2019) procedure could be the bias toward larger \(\hat{a}_m\) evident in Figures 2 and 3, rather than the bias toward lower \(\hat{\lambda}_0\).

Assuming my simulation with \(a_0/\hat{\lambda}_0 = 2\%\) is characteristic of the observed signal strength for J0740+6620, I estimate possible biases in \(\hat{a}_m\) of \(O(10\%)\) when 90\% of segments are retained. If XMM data provides an unbiased estimate of \(\hat{a}_m\), one then can expect an \(O(10\%)\) bias in the relative modulation depth: \(a_m/a_0\). In the context of my simplified model, the bias for \(\hat{a}_m\) is larger than \(\hat{a}_0\), suggesting that the primary signal harmonic may be more biased than other harmonics.

However, the mapping from \(a_m/a_0\) to the compactness is not linear, and small increases in the compactness (smaller radii) can quickly flatten the lightcurve (much smaller \(a_m/a_0\)), whereas low compactness (large radii) may not be as easily constrained (only slightly larger \(a_m/a_0\)). As such, my estimate of a \(O(10\%)\) effect on the relative modulation depth may not correspond to a shift in the lower bound on J0740+6620’s radius of similar size. Nonetheless, it would be of interest to determine exactly how large this shift could be, as the difference between a 2\(M_\odot\) star with a radius of, e.g., 12.5 km versus 13.0 km (\(\sim 4\%\) shift) could have nontrivial implications for the dense matter EoS, such as the existence of strong phase transitions with large onset densities or nearly causal sound speeds at high densities (Drischler et al. 2021a, 2021b; Legred et al. 2021).

An early version of this Letter was shared with the NICER lightcurve working group, and they reanalyzed J0740+6620 using all available GTIs and an essentially unchanged energy band (0.30–1.23 keV versus 0.31–1.22 keV), finding an \(O(1\%)\) increase in both the upper and the lower bound on the radius compared to the published results (M. C. Miller et al. 2021, private communication); i.e., the statistical uncertainty was not degraded by including all the data. This is in line with the predictions of my idealized model, which suggests changes in \(\hat{a}_m\) anywhere between \(-5\%\) and \(+20\%\) are possible when one compares the ordered subset of GTIs to the result using all the data for a single experiment. Nonetheless, just because the constraints obtained from a single observation are not strongly affected by Guillot et al. (2019)’s procedure, this does not mean that no bias exists. Observations of other pulsars, particularly dim stars, may still be significantly affected.

As I have noted several times, my simplified model assumes a stationary noise process (noise counts are i.i.d in each segment). While this is not the case for real NICER data, the actual distribution of rates and durations of GTIs is not publicly available. One may expect the bias to be different in the presence of nonstationary noise if the noise level changes dramatically between GTIs, but the exact behavior will almost certainly depend on the exact distribution of rates and GTI durations. However, the stationary-noise model shows that maximizing the H-test always introduces a bias, regardless of the true noise level. The only question is whether the is large enough to be of practical importance. The Appendix speculates further about the possible impact of nonstationary noise, but I leave a more precise quantification of the impact of realistic nonstationary noise to future work.

5.1.2. J0030+0451

Observations of J0030+0451 may be less affected by such issues. At first glance, J0030+0451 has much larger modulations (\(\gtrsim 30\%\) peak-to-peak; see Bogdanov et al. 2019) than J0740+6620 (\(\lesssim 8\%\) peak-to-peak; see Wolff et al. 2021). Therefore, one expects the impact of maximizing the H-test significance to be less for J0030+0451. Furthermore, while selection cuts were placed on J0030+0451 data based on the count rate in 16 s segments, the threshold was not chosen to maximize the H-test significance (see discussion in Bogdanov et al. 2019; Miller et al. 2019). Indeed, Miller et al. (2019) discard any segments with count rates above 3 Hz, noting that the average count rate is 0.7 Hz. For 16 s segments, this threshold corresponds to approximately 11\(\sigma\) fluctuations in the number of counts per segment. Such a selection may still introduce a bias, but it is likely vanishingly small and unimportant for all practical considerations.

5.2. Possible Paths Forward

Fundamentally, these biases occur because the likelihood does not reflect the true data generation process. Selecting a subset of data (that is not chosen completely randomly) will inevitably bias the inference, as the unaltered Poisson likelihood does not account for the additional data-selection process.

One possible solution is to simply use all data that were recorded. However, as I have said several times, real NICER data are more complicated than my idealized model, and the actual background can vary over time. Using all the data may simply exchange one source of bias (selecting a subset of data) for another (misspecification of the noise model) or increased statistical uncertainty from segments with very large count rates (see Appendix for more discussion). As such, one may still wish to implement some sort of data-selection procedure based on the counts in short segments.

More work will be needed in order to determine the extent of the selection effects described herein for more complicated data in the presence of optimization over both GTIs and energy channels. A thorough understanding of such effects may allow them to be modeled and accounted for within future analyses.

I am very thankful for the many detailed discussions with Cole Miller during the preparation of this manuscript. I also thank the broader NICER Collaboration for entertaining the implications of this idealized model for the much more complicated analyses they perform that enable exceptional science with X-ray observations.

Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation,
Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities. I also thank the Canadian Institute for Advanced Research (CIFAR) for support. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation grant PHY-1607611.

This work was made possible by the numpy (Harris et al. 2020), scipy (Virtanen et al. 2020), and matplotlib (Hunter 2007) software packages. All code used to generate synthetic data within this study is publicly available at (https://github.com/reedessick/h-test-maximization).

Appendix

Simplified Model of Nonstationary Noise

In the main text, I investigate the implications of selecting subsets of data by maximizing the H-test statistic via simulated data that were generated assuming a constant Poisson noise rate within each segment. This was done for several reasons, the foremost of which is simplicity. The stationary noise model allows for relatively clean interpretations of the effects. However, of equal importance is the fact that the NICER collaboration has not published the distribution of noise rates for real GTIs. The precise impact of the H-test maximization may depend strongly on this distribution.

For example, if the noise rate varies only over a small range \( \tau \Delta \lambda_n \lesssim \sqrt{\lambda_n} \), then the change in the expected counts per GTI from the variable rate may be smaller than the statistical variation from Poisson uncertainty. In this case, while the noise rate is not truly stationary, stationarity may be a reasonable approximation. Conversely, if the noise rate can take one of two values in each segment, either \( \lambda_n^\text{low} \) or \( \lambda_n^\text{high} \) with \( \lambda_n^\text{low} \ll \lambda_n^\text{high} \), then H-test maximization may always discard all segments with \( \lambda_n = \lambda_n^\text{high} \). In this case, the inference would effectively only span GTIs with \( \lambda_n = \lambda_n^\text{low} \). This, again, reduces to the stationary noise model.

In reality, the distribution of \( \lambda_n \) is likely (at least somewhat) smooth. Although the details of this distribution are not publicly available, I nevertheless introduce a simple model in which the noise rate varies between segments but is always drawn from

\[
p(\lambda_n) = \begin{cases} 
1/\lambda_{\text{ref}} & \text{if } \left( \frac{1}{7} \leq \lambda_n / \lambda_{\text{ref}} \leq \frac{3}{7} \right), \\
0 & \text{otherwise}
\end{cases}
\]

so that the expected value (averaged over segments) is \( \lambda_{\text{ref}} \). I stress that this model may not be any more justified than assuming identical noise rates in each segment, but its consequences are informative. Figure 6 demonstrates the behavior of my MLEs under this model.

Figure 6 immediately shows that biases are still introduced by H-test maximization in the presence of nonstationary noise. Indeed, the general trends identified within the context of the stationary noise model all remain. While the distribution of \( \hat{\lambda}_0 \) is widened by the nonstationary noise for all data sets, the most significant differences seem to be that the ordered subset’s estimator \( \hat{\lambda}_0 \) depends much more strongly on \( N_{\text{seg}}^{\text{ret}} \) and the ordered subset also tends to retain a smaller fraction of the data more often. Even if the bias in \( \hat{\lambda}_0 \) may be entirely absorbed by the noise model, the correlations between \( \hat{\lambda}_m \) and \( N_{\text{seg}}^{\text{ret}} \) are comparable to the stationary noise simulations. This implies the biases introduced by Guillot et al. (2019)’s data-selection procedure might have a more significant impact in the presence of nonstationary noise because smaller fractions of the data are retained, on average.

Based on this, I again conclude that biases will almost surely be introduced by Guillot et al. (2019)’s data-selection procedure, regardless of the precise details of the noise model.

As a final note, I remind the reader that the motivation for H-test maximization is the observation that the total rate of counts varies between GTIs. This is, quite reasonably, taken as evidence that the noise process must vary over time. However, the likelihood employed by subsequent analyses (independent Poisson distributions for each energy+phase bin for counts summed over all GTIs) implicitly assumes the noise rate is constant. If one considers the joint distribution of the counts in
a single bin over $N$ segments, one obtains

$$p(c_1, \ldots, c_N) = \prod_i^N p(c_i) = \prod_i^N \left[ \frac{(\lambda_i \tau_i)^{c_i} e^{-\lambda_i \tau_i}}{c_i!} \right]. \quad \text{(A2)}$$

where $\tau_i$ and $\lambda_i$ may vary from segment to segment. The NICER analyses, and the MLEs I derive in this work, instead use

$$p(c_1, \ldots, c_N) \propto \left( \frac{\sum_i c_i}{\lambda_i} \right)^{\lambda_i} \sum \frac{e^{-\lambda_i \tau_i}}{i^i}. \quad \text{(A3)}$$

$\sum c_i$ is a sufficient statistics for the full set $\{c_i\}$ if and only if $\lambda_i = \lambda \forall i$. That is, the likelihoods assumed in, e.g., Miller et al. (2019, 2021) and Riley et al. (2019, 2021), implicitly assume the total expected rate of counts is constant in all GTIs. Because the signal rate is expected to be constant, this implies the noise rate is also assumed to be constant.

It is for this reason that the NICER analyses may be degraded by the inclusion of additional GTIs with high noise rates. One would expect that, for a likelihood that is consistent with the data-generation process, additional data can never hurt the inference. For example, if the noise model was fit separately for each GTI, then additional data from high-count-rate GTIs would not degrade the information from low-count-rate GTIs. Instead, the high-count-rate GTIs would simply be uninformative. I leave a more complete exploration of the implications of such assumptions to future work.

References

Aasi, J., Abbott, B. P., Abbott, R., et al. 2015, CQGra, 32, 074001
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017, PhRvL, 119, 161101
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2018, PhRvL, 121, 161101
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019, PhRvX, 9, 011001
Acernese, F., Agathos, M., Agatsuma, K., et al. 2014, CQGra, 32, 024001
Bogdanov, S., Gillott, S., Ray, P. S., et al. 2019, ApJL, 887, L25
Chatziioannou, K. 2020, GReGr, 52, 109
de Jager, O. C., & Büsching, I. 2010, A&A, 517, L9
de Jager, O. C., Raubenheimer, B. C., & Swanepoel, J. W. H. 1989, A&A, 221, 180
Drischler, C., Han, S., Lattimer, J. M., et al. 2021a, PhRvC, 103, 045808
Drischler, C., Han, S., & Reddy, S. 2021b, arXiv:2110.14896
Gendreau, K. C., Arzoumanian, Z., Adkins, P. W., et al. 2016, Proc. SPIE, 9905, 99051H
Guillot, S., Kerr, M., Ray, P. S., et al. 2019, ApJL, 887, L27
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hunter, J. D. 2007, CSE, 9, 90
Jansen, F., Lumb, D., Altieri, B., et al. 2001, A&A, 365, L1
Landry, P., Essick, R., & Chatziioannou, K. 2020, PhRvD, 101, 123007
Legred, I., Chatziioannou, K., Essick, R., Han, S., & Landry, P., 2021, PhRvD, 104, 063003
Miller, M. C., Lamb, F. K., Dittmann, A. J., et al. 2019, ApJL, 887, L24
Miller, M. C., Lamb, F. K., Dittmann, A. J., et al. 2021, ApJL, 918, L28
Prigozhin, G., Gendreau, K., Doty, J. P., et al. 2016, Proc. SPIE, 9905, 436
Raaijmakers, G., Greif, S. K., Hebeler, K., et al. 2021, ApJL, 918, L29
Riley, T. E., Watts, A. L., Bogdanov, S., et al. 2019, ApJL, 887, L21
Riley, T. E., Watts, A. L., Ray, P. S., et al. 2021, ApJL, 918, L27
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Wolff, M. T., Gillott, S., Bogdanov, S., et al. 2021, ApJL, 918, L26

ORCID iDs

Reed Essick @ https://orcid.org/0000-0001-8196-9267

The Astrophysical Journal, 927:195 (9pp), 2022 March 10

Essick