Bayesian Methods for Joint Exoplanet Transit Detection and Systematic Noise Characterization

Jamila S. Taaki1, Farzad Kamalabadi1, and Athol J. Kemball2

1 Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 306 N. Wright Street, MC 702, Urbana, IL 61801-2918, USA
2 Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801-3074, USA

Received 2019 October 6; revised 2020 April 27; accepted 2020 April 27; published 2020 May 29

Abstract

The treatment of systematic noise is a significant aspect of transit exoplanet data processing due to the signal strength of systematic noise relative to a transit signal. Typically, the standard approach to transit detection is to estimate and remove systematic noise independently of and prior to a transit detection test. If a transit signal is present in a light curve, the process of systematic noise removal may distort the transit signal by overfitting and thereby reduce detection efficiency. We present a Bayesian framework for joint detection of transit signals and systematic noise characterization and describe the implementation of these detectors as optimal Neyman–Pearson likelihood ratio tests. The joint detectors reduce to closed form as matched filters under the assumption of a Gaussian Bayesian prior for the systematic noise. The performance of the exploratory detectors was evaluated in injection tests and show \( \sim 2\% \) improvement in overall detection efficiency relative to the standard approach. We find that joint detection efficiency is specifically improved for short-period, low-transit-depth exoplanet transits, providing evidence in support of the hypothesis that joint detection may indeed help to mitigate overfitting. In addition, an initial feasibility test to detect known exoplanets in Kepler data using the joint detectors produced encouraging preliminary results.

Unified Astronomy Thesaurus concepts: Exoplanets (498); Exoplanet detection methods (489); Transit photometry (1709); Bayesian statistics (1900); Astronomy data analysis (1858); Red noise (1956); Time series analysis (1916)

1. Introduction

Transiting exoplanet detection telescopes and missions such as CoRoT\(^3\) (Léger et al. 2009), Kepler\(^4\) (Borucki et al. 2010), K2 (Howell et al. 2014), and TESS\(^5\) (Ricker et al. 2014) have been crucial to expanding the catalog of known exoplanets and their population statistics. To date, Kepler and K2 have produced approximately \( \sim 3000 \) confirmed exoplanet detections from \( \sim 200,000 \) observed light curves (NASA Exoplanet Archive 2019). The detectability of a transiting exoplanet is limited by both the system performance of the telescope as well as the statistical efficiency of the detection and estimation methods used during data processing. Continuous improvements in transit detection methods may reveal more exoplanets in existing data sets and also push the limits of observable exoplanet populations in current and future observations. In this paper, we consider one such exploratory innovation in transit detection and estimation, as described in further detail below.

A transit signal is embedded in a variety of astrophysical (Bryson et al. 2013) and instrumental signals (Jenkins et al. 2010a). Typical astrophysical noise signals tend to have a known physical origin, such as photon-counting noise, intrinsic stellar emission variability, or variability from eclipsing systems; hence, such noise contributions are often well described by deterministic (Torres et al. 2010) or stochastic models (Scargle 1981; Borucki et al. 1985). Noise arising from systematic error (hereinafter systematic noise) is instrument noise for which there exists no a priori data model and which is not reduced by averaging; it is often modeled nonparametrically. Some possible sources of systematic noise include unmodeled residual pointing errors (Foreman-Mackey et al. 2015), instrumental detector offsets, and seasonal instrumental variations. Systematic noise may have trends over a range of timescales including short-duration or transient outlier events.

A large number of algorithms exist for the inference and removal of systematic noise from wide-field transit telescope light curves. A detailed overview of a number of techniques is provided by Roberts et al. (2013), who also provide their own algorithm for systematics removal. The prevalent approach is cotrending: a set of basis signals representative of a set of light curves is obtained and linear combinations of these basis signals are used to form net estimates of systematic noise; these are then removed from all light curves processed (Kinemuchi et al. 2012). In the Trend Filtering Algorithm (Kovács et al. 2005), the basis is chosen as a random subset of the light curves themselves. The method of principal component analysis (PCA) is an alternative method to create a set of basis signals from a large set of light curves; a basis obtained from PCA is orthogonal and maximizes the variance of the light curves projected onto it (Jolliffe 2011). The PCA method is used by the Sys-Rem detrending algorithm (Mazeh et al. 2007) and the Simultaneous Additive and Relative Systems Algorithm (Ofr et al. 2010). A similar related method, singular value decomposition, is used by the Kepler Pre-Search Data Conditioning (PDC) module (Twicken et al. 2010; Smith et al. 2012; Stumpe et al. 2012). The PDC-MAP (Maximum A-Posteriori) algorithm (Smith et al. 2012; Stumpe et al. 2012) was later developed from PDC-LS (least squares; Twicken et al. 2010); the latter forms least-squares systematic noise estimates from a set of basis signals. The PDC-MAP algorithm forms an empirical prior over the set of basis signals and forms
Bayesian maximum a posteriori (MAP) systematic noise estimates.

Systematic noise estimates are typically formed without any assumption of whether or not a transit signal is present in the underlying light curves. However, the signal space of transits and systematic noise may not be unambiguously separated in either the time or frequency domain. For example, Jenkins et al. (2010a) discuss the temporal properties of early Kepler data and identify several short-timescale sources of systematic error in the photometric light curves. High-frequency (≤10 days) systematic noise is similarly identified in Kepler photometry by Petigura & Marcy (2012). This presents a challenge to sequential estimation of systematic noise and transit detection. If a transit signal is present in a light curve, the estimated systematic noise may be biased and the subsequent transit detection process may therefore be also adversely affected (Christiansen et al. 2013; Foreman-Mackey et al. 2015). This point motivates joint modeling of systematics and transits.

In this paper, we consider an exploratory Bayesian approach of jointly estimating systematic noise and transit signals with the goal of improving detection rates while reducing the false-alarm rate. This idea has been explored in a non-Bayesian setting; Foreman-Mackey et al. (2015) provide a method that finds joint linear maximum likelihood estimates of transit signals and systematic noise. The methods that we propose here are Bayesian in nature. The advantage of PDC-MAP over PDC-LS, as described above, has set a precedent for Bayesian systematic noise treatment that leads us to believe it is also advantageous in this setting. Frequentist estimation of systematic noise considers every systematic basis signal as a priori equally likely of occurring in a light curve. Arguably, this model is not a realistic description of an obtained basis and may produce poor systematic estimates. For example, a basis obtained via PCA may be ordered by its singular values, providing a measure of the dominance of various basis signals within the set of light curves. Furthermore, there is expected variability in the presence or relative influence of individual basis signals across a set of separate light curves. If this information is discarded, a basis signal that occurred in a small number, or in a particular subset, of light curves may be unrealistically included in estimates of other light curves; in turn this biased systematics estimate may lead to poor transit detection performance. In a Bayesian treatment, such as PDC-MAP, any ancillary information of this nature can be utilized to more realistic systematic noise estimates.

The work presented here on joint transit detection with a Bayesian treatment of systematic noise is summarized as follows. Two methods are derived to compute detection tests on raw light curves: (i) the first method marginalizes over a Bayesian prior describing the systematic noise while (ii) the second method forms fixed Bayesian estimates of the systematic noise. The use of marginalization in (i) allows an averaged detection test to be computed over a continuous set of systematic noise models. Our second approach (ii) forms fixed Bayesian estimates of the systematic noise under two models in which the signal does or does not contain a transit signal; these are then used as input into a detection test. We describe a detection framework based on these two methods and derive analytic detectors. In the work by Luger et al. (2017), a Bayesian joint model of similar form is used to derive a computationally tractable likelihood function which the authors propose may be used for a variety of astrophysical purposes including transit light-curve modeling. Their derivation follows a marginalization approach whereby the properties of Gaussianity lead to an analytic form for a likelihood function. In deriving Bayesian detection strategies, we also examine a marginal Gaussian likelihood function which leads to a simple analytic detector. However, the framework we propose and the derived detectors are general; further, we formulate the detector as a binary hypothesis test.

PDC-MAP as included in version 8.0 of the Kepler science pipeline formed part of a continuous improvement in PDC; those residual systematic errors remaining were believed to arise from the fact that the systematic basis vectors were not a statistically independent set (Stumpe et al. 2014). This was remedied by introducing a refined multiscale PDC-MAP algorithm (msMAP; Stumpe et al. 2014). This method uses wavelet filtering to separate the different temporal timescales of the systematic effects; PDC-MAP is then used independently in each subband, and the results synthesized to correct the light curve. Our algorithmic approach includes the ability to model correlations between systematic noise trends. We assume statistical correlations exist between systematic basis vectors and have a model prior that may include such information. This represents a complementary approach to that adopted by msMAP.

We demonstrate the performance of our detection methods by performing single-transit injection tests on Kepler data which exclude prior known exoplanet detections. We simulate a set of limb-darkened transit signals enumerated in Table 1. In the injection tests, the new joint Bayesian detection methods outperformed a standard sequential countrending and detection approach with a relative average ~3% improvement in the overall detection efficiency subject to the same rate of incorrect detections. In particular, the injection tests demonstrated the ability of these techniques to separate short-period transits (<10 days) in the presence of high-frequency systematics. We also demonstrated the feasibility of these methods on a small subset of Kepler light curves when the sample included prior exoplanet detections.

The paper is organized as follows. Section 2 introduces the light-curve signal model, the joint detection framework, and particular detector implementations, and describes a set of numerical tests to evaluate detector performance. In Section 3, we report the results from our injection and feasibility tests with Kepler data. The results are discussed in Section 4 and conclusions presented in Section 5.
2. Methods

Transit detection infers the presence of an exoplanet in orbit around a star from time-series observations of an optically unresolved star–planet system; contemporary reviews are provided by Deeg & Alonso (2018) and Moutou & Pont (2006). As a planet obscures the face of a star, the observed brightness of the system drops. The effect a transiting planet has on the overall observed brightness of a star–planet system is well approximated by parametric models (Seager & Mallen-Ornelus 2003), Mandel–Agol equations (Mandel & Agol 2002), or periodic box functions (Kovacs et al. 2002), where the models and associated parameter sets depend largely on the properties of the star and planet. Expressed here in the language of detection theory, the transit method seeks to detect the presence of a transit signal of prespecified functional form, but with unknown parameter values, in the observed light curve of a star–planet system.

Detection of a transit signal can be posed as a binary hypothesis test upon a candidate light curve: does the light curve contain a transit signal? Commonly, this test is performed using a matched-filter-based test once cotrending has been used to remove systematic noise from a light curve (Tingley 2003). Because we wish to jointly detect transit signals and model systematic noise in the current work, we form the hypothesis test on raw light curves. A raw light curve $y_i$ for a particular star indexed by $i \in I$ (where $I$ is the set of integers) is expressed as a vector of $N$ photometric flux measurements in time, before any cotrending for systematic noise, and herein-after considered also to be normalized and median subtracted. An example subset of such raw light curves, here from the Kepler mission, is shown in Figure 1. The raw light-curve vector $y_i$ contains systematic noise $I_i$ and a stellar signal $s_i$. For conciseness, residual sources of nonsystematic statistical error are considered included in $s_i$. This includes instrumental shot noise, considered here in the Gaussian limit (Grinstead & Snell 2012).

The test is as follows: the null hypothesis ($H_0$) posits that $y_i$ contains no transit signal. The alternative hypothesis ($H_1$) posits that $y_i$ does contain a transit signal $t \in T$. The set $T$ describes all detectable transit signals. The hypothesis test is therefore posed as

$$H_0: y_i = s_i + I_i,$$

$$H_1: y_i = t + s_i + I_i,$$

To decide between hypotheses, a likelihood ratio test (LRT; Kay 1993; Wasserman 2013) compares the probability of the raw light curve $y_i$ under either hypothesis model. If the LRT $L_i(y_i)$ exceeds a fixed threshold $\tau$, then a transit signal has been detected. More generally, we consider the LRT as a test statistic $T(y_i)$ on the data $y_i$. Typically, a Neyman–Pearson criterion (Kay 1993; Wasserman 2013), in which the probability of detection is maximized subject to a fixed rate of false alarm $\alpha$, is used to decide the threshold $\tau$. These quantities are related as

$$T(y_i) = L_i(y_i) = \frac{p_1(y_i)}{p_0(y_i)} \frac{H_i}{H_0} \geq \tau,$$

where $p_h(y_i) \equiv p(y_i|H_h)$ is the probability of observing light curve $y_i$ under a particular hypothesis model $h \in \{0,1\}$.

2.1. Raw Light-curve Signal Model

This section describes typical models for the raw light-curve components: the systematic noise, the stellar signal, and the transit signal.

2.1.1. Systematic Noise

A standard model for systematic noise is a linear reduced basis model (Kovács et al. 2005) in which systematic noise vectors $I_i$ are modeled as a linear combination of a set of $K$ systematic noise basis vectors $\{v_k\}$ weighted by coefficients $c_i^k$. Because the set of systematic noise signals $\{I_i\}$ are reasonably expected to be correlated between light curves of sufficiently common instrumental origin (Stumpe et al. 2012), they may be expressed in terms of a common set of systematic noise basis vectors $\{v_k\}$.

$$I_i = \sum_{k=1}^{K} c_i^k v_k.$$

Systematic basis noise vectors $\{v_k\}$ may be estimated via dimensionality reduction techniques (Cunningham & Ghahramani 2015) applied on $\{y_i\}$. PCA is a commonly used technique for cotrending and produces a set of orthogonal basis vectors (Mazeh et al. 2007; Twicken et al. 2010; Petigura & Marcy 2012; Smith et al. 2012; Stumpe et al. 2012; Foreman-Mackey et al. 2015). A non-Bayesian least-squares estimation of the coefficients $c_i$ is equivalent to maximum likelihood estimation (Kay 1999a) with an assumed Gaussian stellar noise model $s_i$, where the samples are independent and identically distributed. However, least-squares estimation, by minimizing the net rms error, is prone to overfitting residual transit signatures (Smith et al. 2012; Stumpe et al. 2012). As shown by Smith et al. (2012), a Bayesian estimation of $c_i$ can mitigate this issue by allowing the incorporation of a prior $p(c_i)$ on the coefficients. The prior here is conditioned on the index $i \in I$; this index accommodates latent variables, such as position within the CCD and the stellar magnitude, that produce clustering in systematic noise properties as noted above (Smith et al. 2012; Stumpe et al. 2012).
By definition, the cotrending model for systematic noise in Equation (4) does not fully capture systematic noise that is temporally uncorrelated between light curves. Examples of such systematic noise effects include cosmic-ray events, sudden pixel sensitivity drop-off, and electronic image artifacts (Jenkins et al. 2010a; Stumpe et al. 2012). While our model cannot account for outlier effects, a number of residual unmodeled signatures may be treated as effectively Gaussian and absorbed into the term $s_i$ in Equation (2).

2.1.2. Stellar Signal

In the current work, we assume that the stellar noise $s_i$ (in each light curve $i$) may be reasonably modeled by a Gaussian noise model $s_i \sim \mathcal{N}(\theta, \text{Cov}_{ij})$. Gaussian processes are reviewed in the monograph by Gallager (2013, ch. 3). Gaussian models are powerful nonparametric models of processes with complex or unknown generating phenomena and may be justified by the central limit theorem (Grinstein & Snell 2012). Gaussian models have been shown to be effective for modeling stellar variability (Pereira et al. 2019) and have been used in the context of exoplanet detection (Carter & Winn 2009; Rajpaul et al. 2015).

Stellar flux density time series display variability on a wide range of timescales (Conroy et al. 2018), and time-correlated fluctuations significantly affect the detectability of transit signals (Borucki et al. 1985; Jenkins 2002; Pont et al. 2006). A Gaussian noise model can incorporate time-correlated structure via its covariance matrix. A review of common correlated noise estimators for exoplanet light curves is provided by Cubillos et al. (2017).

Our model further assumes that stellar noise is stationary within a fixed time window. In a stationary Gaussian colored noise model, correlations depend only on the separation between two points in time. Equivalently, the values on the stellar covariance matrix $\text{Cov}_{ij}$ are constant along the diagonals, and the matrix is Toeplitz in form. Such a model admits a power spectral representation and allows spectral analysis (Kay 1999a), which is considerably more efficient than time-domain analysis.

Stellar processes do, however, evolve in time and can exhibit nonstationarity as exemplified by our Sun (Borucki et al. 1985; Jenkins 2002). For this reason, we confine our stellar noise model to be stationary only within a single observational quarter of Kepler data (approximately 90 days). This assumption is motivated in part by the analysis of common stellar periodic variability in Kepler data on timescales of days to weeks (Basri et al. 2010).

The extension of our detectors to multiple Kepler observing quarters and associated implications for our statistical assumptions regarding signal covariance timescales are described in Appendix B.

2.1.3. Transit Signal

A transit signal is typically approximated as a periodic box function (Kovacs et al. 2002). The periodic box function is parameterized as follows: $\alpha$ describes the fractional drop in light relative to the stellar signal, $P$ is the orbital period of the planet, $d$ is the transit duration, $n$ is the time index, and $t_0$ is the phase (epoch) of the signal:

$$t_{\alpha,P,t_0,d}[n] = \begin{cases} \alpha, & \text{if } (n - t_0) \text{ mod } p \leq d. \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

A typical technique to estimate transit signal parameters is to quantify the space of possible parameter values to create a discrete set of candidate transit signals $T$ and to search through this set while performing detection tests (Jenkins et al. 2002). The set of possible transit signals $T$ is not described herein by a Bayesian prior as, in a prior, the population statistics of exoplanets are not fully known. As noted above, other parametric forms can be used in place of a periodic box function (Mandel & Agol 2002; Seager & Mallen-Orobas 2003).

2.1.4. Signal Matrix Form

The signal model may be expressed in matrix form. A similar form is presented in Smith et al. (2012), which we adapt to include the presence of a transit signal. As before, time is indexed by $n$ ranging from 1 to $N$ and $i$ is the light-curve index. $V$ is an $N \times K$ matrix with columns formed from $\{v_k\}$:

$$y = t + Vc_t + s_i \quad (6)$$

$$\begin{bmatrix} y_1[1] \\ y_1[2] \\ \vdots \\ y_1[N] \end{bmatrix} = \begin{bmatrix} t[1] \\ t[2] \\ \vdots \\ t[N] \end{bmatrix} + \begin{bmatrix} v_1 & v_2 & \cdots & v_K \end{bmatrix} \begin{bmatrix} c_1^1 \\ c_2^1 \\ \vdots \\ c_K^1 \end{bmatrix} + \begin{bmatrix} s_1[1] \\ s_1[2] \\ \vdots \\ s_1[N] \end{bmatrix} \quad (7)$$

2.2. Joint Detection Methods

In order to compute the LRT as described in Equation (3), it is necessary to compute $p_h(y): h \in \{0, 1\}$, the probability of a light curve conditioned on a hypothesis model. Under $H_1$, the light curve $y$ contains a transit signal $t \in T$, where we assume that each possible transit signal is a priori equally likely to occur. Hence, for the remainder of this section, we consider the LRT to be computed with respect to a fixed transit signal $t$.

With the transit signal fixed, the signal model describing $y$ for either hypothesis contains two unknown components: the systematic noise $I_t = Vc_t$ and the stellar signal $s_i$. Computing the conditional likelihood of a light curve $y$ for a particular $t$, $p_h(y|t): h \in \{0, 1\}$ requires addressing the dependence on these unknown terms. We use two standard approaches in Bayesian analysis for this purpose, namely marginalization over the systematics and estimation of the systematic noise. This results in two distinct detection methodologies.

With Bayesian marginalization (Loredo 1992), the joint probability distribution describing the observations and signal model is integrated with respect to the systematic noise to produce two marginal likelihoods for either hypothesis model. The ratio of these marginal likelihoods forms the LRT. In the second approach, systematics are estimated conditioned on the hypothesis model. This amounts to finding two distinct Bayesian systematics estimates both under the belief that there

---

A stationary Gaussian noise model is equivalent to a wide-sense-stationary Gaussian noise model, as Gaussian models are fully parameterized by their first and second moments (Papoulis & Pillai 2002).
is a transit signal in the light curve and that there is no transit signal. These fixed systematics estimates are used to compute the likelihoods \( p_h(y) \); \( h \in \{0, 1\} \), the ratio of which forms the LRT.

This section derives generic forms of the LRT obtained through Bayesian marginalization and fixed estimation. Specific implementation details are provided in following sections.

### 2.2.1. Matched Filter

Before proceeding further, a brief overview of the matched filter is provided as it plays an essential part in our detection framework. For more detail, we refer the reader to the monographs by Kay (1993) and Poor (2013).

The matched filter is the optimal Neyman–Pearson detector for a deterministic signal \( t \) in Gaussian noise \( n \sim N(0, \text{Cov}_n) \) for an observed signal \( y \). Using the form provided by Jenkins et al. (2002), the matched filter test statistic \( T(y) \) is given by

\[
T(y) = \frac{y^T \text{Cov}^{-1} t}{\sqrt{t^T \text{Cov}^{-1} t}} \geq \tau.
\]

(8)

This form of the matched filter may be derived from the LRT formulation in Equation (3) under the assumption of Gaussianity. The matched filter is desirable from an implementation standpoint as the distribution of the matched filter test statistic conditioned on the null hypothesis is normal: \( T(y)|H_0 \sim N(0, 1) \). Thus, a detection threshold \( \tau \) should achieve the same false-alarm rate for different noise covariance matrices \( \text{Cov}_n \), and signals \( t \). Second, if \( n \) is wide-sense-stationary, the matched filter may be efficiently computed in the Fourier domain (Kay 1999b).

#### 2.2.2. Detector A: Marginalization over Systematic Noise

This detector computes \( p_h(y) \); \( h \in \{0, 1\} \) by marginalizing over the systematic noise prior probability \( p(c_i) \). For a particular hypothesis, the conditional probability of the observed signal \( y \), conditioned on systematic noise coefficients \( c_i \) is given by

\[
p_0(y|c_i) = p(s_i = y_i - Vc_i|c_i),
\]

(9)

\[
p_1(y|c_i) = p(s_i = y_i - t - Vc_i|c_i).
\]

(10)

Because the \( t \) and \( y \) are deterministic, the only stochastic term in these distributions is \( s_i \), which is modeled as approximately Gaussian. See Sections 2 and 2.1.1 for a discussion of contributing non-Gaussian terms in \( s_i \). The LRT may be computed by marginalizing over the known prior \( p(c_i) \) for both of these conditional probability functions:

\[
L_i(y_i) = \frac{p_1(y_i)}{p_0(y_i)} = \frac{\int_{c_i \in C_i} p_1(y_i|c_i)p(c_i) \, dc_i}{\int_{c_i \in C_i} p_0(y_i|c_i)p(c_i) \, dc_i}.
\]

(11)

\[
= \frac{\int_{c_i \in C_i} p(y_i = t + Vc_i + s_i|c_i)p(c_i) \, dc_i}{\int_{c_i \in C_i} p(y_i = Vc_i + s_i|c_i)p(c_i) \, dc_i},
\]

(12)

where \( C_i \) is the domain of \( c_i \).

#### 2.2.3. Detector B: Joint Transit and Systematic Noise Estimation

This method obtains two fixed systematic noise estimates \( \hat{c}_0 \) and \( \hat{c}_0 \), with and without the presence of a transit signal, respectively. These Bayesian estimates are obtained from the posterior distributions \( p_h(c_i|y) \); \( h \in \{0, 1\} \) currently as MAP estimates. These can then be used to compute \( p_h(y) \); \( h \in \{0, 1\} \) and the LRT. Herein lies the main distinction from prior cotrending approaches; there, a single systematic noise estimate, roughly equivalent to \( \hat{c}_0 \), is used to compute a detection test. Cotrending is performed before detection, and the systematic noise estimate is formed without modeling a particular transit signal. Such an approach favors the null hypothesis in a detection test as \( \hat{c}_0 \) maximizes the null hypothesis posterior. Here instead we find the MAP estimates for systematic noise under both transit signal hypotheses and compare the likelihoods computed from these estimates in the form of the LRT. The conditional MAP estimates of the systematic noise under either hypothesis take the form

\[
\hat{c}_{MAP} = \arg \max_{c_i \in C_i} p_1(y_i|c_i)
\]

\[
= \arg \max_{c_i \in C_i} \ln p_1(y_i|c_i) + \ln p(c_i))
\]

\[
= \arg \max_{c_i \in C_i} \ln p_0(y_i|c_i) + \ln p(c_i))
\]

(13)

(14)

We insert these estimates into the LRT test statistics in Equation (2) as

\[
H_0: y_i = V\hat{c}_{MAP} + s_i,
\]

\[
H_1: y_i = t + V\hat{c}_{MAP} + s_i.
\]

(15)

Because the systematic noise signal for either hypothesis is fixed, the only stochastic variable is the Gaussian stellar noise \( s_i \). The detection test may be equivalently restated as

\[
H_0: \hat{y}_i = s_i
\]

\[
H_1: \hat{y}_i = k_i + s_i.
\]

(16)

(17)

This hypothesis test on \( \hat{y}_i = y_i - V\hat{c}_{MAP} \) describes the detection of a known signal \( k_i = t - V\hat{c}_{MAP} + V\hat{c}_{MAP} \) in Gaussian noise \( s_i \). Therefore, as described above in Section 2.2.1, the optimal detector is the matched filter with test statistic \( T_i(y_i) \):

\[
T_i(y_i) = \frac{\hat{y}_i^T \text{Cov}_{s_i}^{-1} k_i}{\sqrt{k_i^T \text{Cov}_{s_i}^{-1} k_i}} \geq \tau.
\]

(20)

### 2.3. Design and Implementation of the Joint Detectors

The detectors defined in Section 2.2 admit several choices in their concrete implementation, including assumptions regarding the statistical distribution of the underlying variables defining the systematic noise and stellar signal and the inference methods used to estimate their parameter values. In this regard, our design and implementation decisions have been wherever possible scientifically or empirically motivated and guided also by computational feasibility. The assumed statistical distribution of the systematic noise coefficient vector \( c_i \), as well as its dimension \( (K) \), largely determines the analytic and computational tractability of the detectors. A small number of distributions produce closed-form posterior distributions and thereby analytic forms of these detectors; specifically, the implementation for detectors A and B is described here under the assumption of Gaussian statistics. The inference methods
used to estimate the detector parameter values are described at the close of this section.

### 2.3.1. Systematic Noise Prior: Gaussianity

A systematic noise prior $p(e_i)$ in Gaussian form $e_i \sim \mathcal{N}(\mu_{c,i}, \text{Cov}_{c,i})$ produces detectors in closed analytic form. A secondary motivation for adopting a Gaussian prior is that it has desirable nonparametric modeling properties when the underlying distribution is not fully known. Specifically, it is inherently regularizing and also allows correlations to be easily modeled. We motivate the choice of a Gaussian prior for $e_i$ by examining the sample density of coefficient values $\{\hat{e}_i\}$ obtained via least-squares fits of $\{v_k\}$ to $\{y\}$. Notwithstanding the issues with frequentist estimation of systematic noise noted earlier, it is assumed here that the least-squares fits $\{\hat{e}_i\}$ are approximately unbiased samples of the coefficient distribution of the population such that they may be used in the inference of parameters describing $p(e_i)$. Figure 2 shows a subset of systematic basis vectors $\{v_k\}$ obtained here for CCD module 8 data during Kepler observing quarter 2 using PCA with model order $K = 20$. This model order was chosen empirically as sufficient to be representative of the systematic noise while low enough as compared to the size of the light-curve population $|I| = 5000$ so as to avoid overfitting. In Figure 3, a histogram of the set of sample least-squares coefficient values $\{\hat{e}_i\}$ is shown for the basis vectors in Figure 2; superscript 1 denotes the first element of each coefficient vector. A best-fit Gaussian coefficient prior is overlaid on the histogram. The histogram in Figure 3 has approximately Gaussian form but with a narrower central mode and a broader tail in the distribution; collectively, these broaden the overlaid Gaussian fit. The extremal outliers indicate excursions from our assumptions underlying the systematic noise model. We examine the impact of any non-Gaussianity in the systematic noise prior $p(e_i)$ on detector performance in Section 4.

### 2.3.2. Detector A: Marginalization over Systematic Noise (Gaussian)

Detector A computes an LRT by marginalizing over the systematic noise distribution, as described in Section 2.2.2. A closed analytic form of this detector is derived here under the assumption of a Gaussian prior for the coefficients of the systematic noise basis vectors. If the coefficient prior $p(e_i)$ is Gaussian $e_i \sim \mathcal{N}(\mu_{c,i}, \text{Cov}_{c,i})$, so too are the marginal probabilities of the observed signal $p_h(y)$: $h \in \{0, 1\}$ (see Equation (9) and Equation (10)):

$$p_0(y) \sim \mathcal{N}(V_{0,c,i}, \text{Cov}_{c,i} + V \text{Cov}_{c,i} V^T),$$

$$p_1(y) \sim \mathcal{N}(t + V_{1,c,i}, \text{Cov}_{c,i} + V \text{Cov}_{c,i} V^T),$$

where $\text{Cov}_{c,i}$ is the covariance of the zero-mean stellar signal $s$. The marginal detector formed from Equation (12) is therefore equivalent under these assumptions to the detection of a known signal $t$ in the presence of Gaussian noise $z_i \sim \mathcal{N}(\theta, \text{Cov}_{c,i} + V \text{Cov}_{c,i} V^T)$ within the data $\hat{y}_i = y_i - V_{h,c,i}$. The matched filter is therefore optimal (Section 2.2.1) with test statistic $T_i(\hat{y}_i)$:

$$T_i(\hat{y}_i) = \frac{\hat{y}_i^T C_{c,i}^{-1} t}{\sqrt{t^T C_{c,i}^{-1} t}} \leq H_0.$$

### 2.3.3. Detector B: Joint Transit and Systematic Noise Estimation (Gaussian)

Detector B produces MAP systematic estimates conditioned on the null and alternate hypothesis $z_{h,i}^{\text{MAP}}$: $h \in \{0, 1\}$ as defined in Equations (13) and (15). These distinct conditional estimates are used as input to the detection test in Equation (20). Closed-form expressions for MAP/MMSE$^7$

---

$^7$ Because both prior and likelihood are Gaussian, so too is the posterior as a Gaussian is a conjugate distribution (Wasserman 2013). For a Gaussian distribution, MAP and MMSE estimates are equivalent.
estimates \( \hat{\theta}_{h,i} \): \( h \in \{0, 1\} \) are provided based on a Gaussian systematic noise prior \( \epsilon_i \sim N(\mu_{c,i}, \text{Cov}_{c,i}) \) and a zero-mean stellar signal \( \eta_i \) with covariance \( \text{Cov}_{\eta,i} \). These estimates can be found either by computing the expectation with respect to the posterior distribution \( E_p(\epsilon|y) \) or by maximizing the log of the posterior as shown in Smith et al. (2012):

\[
\hat{\theta}_{h,i}^{\text{MAP/MMSE}} = (V^T \text{Cov}_{\epsilon,i}^{-1} + \text{Cov}^{-1}_{\epsilon,i})^{-1} \times (V^T \text{Cov}_{\epsilon,i}^{-1} \eta_i + \text{Cov}_{\epsilon,i}^{-1} \mu_{c,i}),
\]

\[
\hat{\theta}_{i,i}^{\text{MAP/MMSE}} = (V^T \text{Cov}_{\epsilon,i}^{-1} + \text{Cov}^{-1}_{\epsilon,i})^{-1} \times (V^T \text{Cov}_{\epsilon,i}^{-1} (\eta_i - t) + \text{Cov}^{-1}_{\epsilon,i} \mu_{c,i}).
\]

### 2.3.4. Detector Parameter Estimation

In their implementation, the joint detectors A and B described above require several parameters characterizing the systematic and stellar noise to be estimated or held fixed. This is not trivial as there are no clean, well-separated data. The parameters that need to be estimated include those of the Gaussian prior \( \epsilon_i \sim N(\mu_{c,i}, \text{Cov}_{c,i}) \) for the coefficients \( \epsilon_i \) of the systematic noise basis vectors \( \eta_k \), \( k \in K \), for fixed model order \( K \). In addition, the parameters of the statistical distribution of the stellar signal \( \eta_i \) need to be estimated.

In the current work, we estimate the systematic noise basis vectors \( \{\eta_k\} \) using PCA as implemented in the Python module sklearn.\(^8\) To suppress the inclusion of stellar noise or dominating outlier light curves, the basis is constructed from 90% of the total light curves—those which have the lowest variance in absolute value. The parameter values of the Gaussian coefficient prior \( \epsilon_i \sim N(\mu_{c,i}, \text{Cov}_{c,i}) \) are estimated directly from the set of coefficient estimates \( \{\epsilon_i\} \) obtained from least-squares fits of the systematic basis vectors \( \{\eta_k\} \) to the raw light curves \( \{\eta_i\} \). Here we assume that for a population \( I \) the same coefficient covariance may be used, denoted by \( \text{Cov}_{\epsilon,I} \). An example sample covariance is shown in Figure 4. It can be seen that in this example the coefficient values are correlated (by the presence of nonzero off-diagonal elements). The PCA method finds an orthogonal set of basis vectors \( \{\eta_k\} \), but there is no reason to expect independent systematic noise signals themselves to be orthogonal. Hence, the orthogonalization procedure may distribute a systematic noise signal across multiple basis vectors. This will likely produce correlations between coefficients.

While the coefficient covariance \( \text{Cov}_{\epsilon,I} \) is estimated from the global population of least-squares fits, the coefficient mean \( \mu_{c,I} \) is taken to be the least-square coefficient vector \( \hat{\epsilon} \) obtained for the particular light curve under consideration.

As noted above, we assume that the stellar noise \( \eta_i \) is wide-sense-stationary (WSS) and zero-mean. This implies the stellar covariance matrix is Toeplitz and fully parameterizes the stellar noise. We therefore only need to estimate the stellar spectrum \( (N \text{ free parameters}) \) as opposed to a full covariance matrix \( (N^2 \text{ free parameters}) \). The spectral estimation technique is not prescribed by the form of the detectors; here we use a smoothed periodogram (Kay 1999b) on the least-squares cotrended light curves.

\(^8\) https://scikit-learn.org/stable/

![Figure 4](image.png) Normalized sample covariance of coefficient value estimates \( \{\hat{\epsilon}_i ; i \in I\} \) obtained from least-squares fits of \( \{\eta_k ; k \in K\} \) to light curves \( \{\eta_i ; i \in I\} \). These data are from the Kepler CCD module 8 during quarter 2. The \( \epsilon_0 \) variance indicates that \( \{\eta_1\} \) is a dominant systematic trend across the light-curve population. Correlations are also observed between basis signals. For ease of presentation, the covariance is only shown for 8 of 20 coefficient values and has been normalized by dividing the matrix by the maximum value.

### 2.4. Numerical Simulations and Detector Performance Evaluation

The joint detectors considered in this paper are defined in statistical and algorithmic terms in Section 2.2. Their specific design and implementation in the current work is described in Section 2.3. To evaluate the performance of the detectors, we have conducted numerical single-transit injection tests using a subset of raw Kepler light curves from which confirmed exoplanet detections have been excluded. Confirmed exoplanets are those for which the initial detection has been validated by the project by a secondary analysis or follow-up observations. In addition, an initial feasibility test of the detectors against Kepler data containing known exoplanet detections was also performed. This section describes the nature of these numerical studies augmented by a brief discussion of computational optimization and complexity issues relevant to these tests.

#### 2.4.1. Injection Tests

We evaluate our detection performance using single-transit injection tests (Giarranti et al. 2000; Wengrave et al. 2005; Burke et al. 2006; Burke & Catanzarite 2017) on raw Kepler SAP light curves (Jenkins et al. 2010b) selected to exclude known exoplanet detections. We use the long-cadence Kepler data in this analysis for which there is a 29.4 minute integration time (Jenkins et al. 2010a). A subset of 20,000 of such Kepler light curves were selected, comprising 5000 light curves from each of the following Kepler CCD module and observing quarter pairs [M6:Q10, M8:Q2, M14:Q9, M18:Q4].\(^9\) These module and quarter pairs were selected randomly over time and CCD module position. Within each pair, the light curves were

\(^9\) SAP flux light curves were downloaded using the Lightkurve library (Lightkurve Collaboration et al. 2018).
sorts by angular separation from the module reference point used by the MAST Kepler data archive, and the first 5000 were selected from the sorted list. We exclude any light curves associated with exoplanets defined as confirmed by the NASA Exoplanet Archive.\(^{10}\) A single synthetic transit is injected once per light curve; transit signals were simulated using the Python transit\(^{11}\) library developed by D. Foreman-Mackey. These synthetic transit signals include limb darkening (Mandel & Agol 2002; Kipping 2013a) and a complete description of the transiting Keplerian orbital elements. The injected signals are drawn from a distribution of exoplanet population parameters given in Table 1. This population parameter distribution is informed by that used by Foreman-Mackey et al. (2015) and Kipping (2013a, 2013b), but it is not identical.

We adopted zero orbital eccentricity in the current work, among other changes.

2.4.2. Standard Model Processing

As a reference detector, we adopt the standard heuristic of sequential cotrending followed by detection (Stumpe et al. 2012). We term this the standard model in what follows and provide our own implementation of this detector in the current work. In the standard model, the cotrending is performed assuming the light curve contains no transit signal: \(\hat{y}_i = s_i + V_{c,i}\), analogous to hypothesis \(H_0\) (Equation (16)). An MAP/MMSE estimator for the systematic noise is constructed equivalent to \(\hat{\epsilon}_0^{\text{MAP/MMSE}}\) (Equation (15)) and applying the same Bayesian priors for systematic noise and stellar noise as used by detectors A and B. Detection is then performed on the cotrended light curves \(\hat{y}_i = y_i - \hat{V}_{c,i}^{\text{MAP/MMSE}}\) using the matched filter in Equation (8) with \(\text{Cov}_n = \text{Cov}_{s,i}\).

2.4.3. Transit Search Space Optimization

As discussed in Section 2.1.3, transit detection requires testing every candidate transit signal \(t \in T\) to find that which maximizes the test statistic \(T(y)\). The transit space \(T\) is typically populated by periodic box functions over a range of candidate orbital periods \(P\), transit durations \(d\), and epoch times \(t_0\) in the functional form described by Equation (5). Transit depth \(\alpha\) is omitted here as our detector forms generally do not make use of this parameter. The dimensionality of the transit parameter search space is therefore intrinsically large and the transit detection problem computationally expensive. This computational cost can be reduced sharply by constraining the range of epochs \(\{t_0\}\) for a candidate transit with a certain period and duration using the method of phase correlation (Averbuch & Keller 2002). We adopt this method in our numerical studies due to the significant reduction in computational cost. The phase-correlation method and its application to epoch estimation is described in Appendix A.

The phase-correlation method, however, requires the use of cotrended light curves for sufficient accuracy in the estimated epochs \(\{t_0\}\). This raises the concern that the transit signal may not be detected optimally due to the use of the cotrended, as opposed to raw, light curves. To verify that this approach does not decrease detection efficiency, we compared detection results with and without phase correlation using single-transit injection tests over 5000 light curves from the broader injection test data described in Section 2.4.1 selected here from [M8: Q2]. We define detection efficiency in this context as the rate of correct detection of the known injected signals as described in Section 3.1. In each case, the standard reference detector defined in Section 2.4.2 was used; as described above, this detector comprises sequential cotrending and detection steps. In the test without phase correlation, a three-dimensional transit signal parameter search space was used as defined in Table 2; the standard detector operated on the raw light curves over this gridded search space. In the phase-correlation test, a transit epoch \(t_0\) was estimated from each least-squares cotrended light curve using the phase-correlation method. The standard detector was then applied to the raw light curves holding \(\{t_0\}\) fixed to the phase-correlation estimate but searching over a residual two-dimensional search space in period and duration as tabulated in Table 2.

The test data here comprise actual raw light curves from which confirmed exoplanets have been excluded; however, the data cannot be shown provably to exclude any hitherto undetected transit signals. As such, we define a quasi-false-alarm rate as the rate of incorrect detection with respect to the injected signal set. The detection tests using the phase-correlation method show an improvement in detection efficiency of 14% and a reduction in the quasi-false-alarm rate of 16% (at a detection threshold \(\tau = 8.4\)) over the detection tests for which a direct search was performed. A comparison of detection efficiency broken down by transit parameter values is

\(^{10}\) List of confirmed exoplanets from NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu).
\(^{11}\) http://dfm.io/transit
shown in Figure 5 and demonstrates no marked reduction in detection for weak signals. The phase-correlation method described in Appendix A by definition has a maximum accuracy $\Delta t_0$ in epoch $t_0$ of one long-cadence sample $\Delta t_{LC}$ (a single “pixel”). Nonadditive noise will reduce the accuracy of the method. By considering the minimum required correlation between a measured transit and a parameterized transit model $(t_0, d, P)$, Jenkins et al. (2010c) provide an analysis motivating a default search spacing in the epoch of $\Delta t_0 = d/10$ for the Transiting Planet Search (TPS) module in the Kepler science pipeline. Our direct search here used a step size $\Delta t_0 = d/2$ (see Table 2), which is suboptimal relative to the maximum epoch accuracy of the phase-correlation method, thereby possibly explaining the improved detection efficiency of the latter method here. We stress here that these tests only demonstrate that the phase-correlation optimization is suitable for the transit search space considered here; generalization to broader applicability is left to future work.

The ratio of computational cost between the test without phase correlation and that using phase correlation was $\sim 10^2$ : 1. As discussed further in Section 4, the freed computational resources allow more refined searches in the remaining transit parameters and can be argued to improve overall accuracy in that sense. As a result of the positive outcome of this verification test and the significant associated reduction in computational cost, we used the phase-correlation method described in Appendix A to estimate transit epochs $(t_0)$ in our full injection tests described in Section 2.4.1. For the full injection tests, the transit signal parameter search space is informed broadly by Jenkins et al. (2002). However, given our use of the phase-correlation method and the associated freed computational resources, we search (in units of long-cadence samples $\Delta t_{LC} = 29.4$ minutes) over the period range $[20, 2125]$ with a 0.25 step size and within the following set of transit durations (in long-cadence samples): $[2, 3, 4, 5, 6, 7, 9, 10, 12]$. Detector B additionally requires a search over transit-depth parameter $\alpha$. This arises from the estimation of transit-dependent systematics as in Equation (25) for $\sigma_{\alpha}^2$. In contrast, the matched filter depends purely on the shape of the transit signal (defined by $t_0, d, P$) and not the signal strength $\alpha$. One can see this property by considering a scaled signal $\alpha$ in the matched filter function, Equation (8); the scaling $\alpha$ immediately cancels. For the rest of the detectors, the only step that is dependent on a transit signal is a matched filter step, ergo they do not depend on a transit-depth parameter. For detector B, we search over four equally spaced transit depths $\alpha \in \{0.2, 0.5, 0.8, 1.1\}$, scaled by the maximum range of the least-squares cotrended light curve under consideration. Our choice of transit-depth sampling is exploratory but proved practical. We note, however, that it sets a limit on the detectability of signals with transit depths below 20% of the cotrended light curve. Future work will consider optimized sampling schemes for transit depth including estimated noise levels.

### 2.4.4. Computational Complexity

The computational complexity of the detectors and their key constituent operations acting on a single light curve are summarized in Table 3. The computational complexities are expressed in terms of the length of the light curve $N$ and the size of the transit signal search space $|T|$ as defined in the introduction of Section 2. These complexities are specific to the case of a Gaussian systematics prior described in Section 2.3.1.

A time-domain matched filter defined in the form of Equation (8) has a computational complexity dominated by the product of an $[N \times N]$ matrix and a length $N$ vector; therefore, its complexity is $O(N^2)$. A matched filter implemented in the Fourier domain involves computing fast Fourier transforms (FFT) and inner products of length $N$ vectors (Kay 1999b). Between these operations, the FFT is more computationally intensive, hence the Fourier-domain matched filter is $O(N \log N)$ (Bracewell & Bracewell 2000).

The computational complexity of MAP/MMSE systematics estimation is defined by the form of Equation (25). For a particular light curve, we can reuse many of the computed terms during multiple transit tests. Considering transit-dependent terms, the dominant computational term is the product of the $[N \times N]$ matrix $\text{Cov}_{\alpha}^{-1}$ and the length $N$ vector $t$; consequently this computation is $O(N^2)$. Once per light curve, a matrix inversion of the $N \times N$ matrix $\text{Cov}_{\alpha}$ is performed, and this operation is $O(N^3)$. However, it is not leading order because generally $|T|N^2 \gg N^3$. As such, MAP/MMSE systematics estimation has computational complexity $O(N^2)$ per light curve per transit.

The detector complexities are determined by the form of the matched filter used and scaled by the size of the transit search space. The standard detector (Section 2.4.2) is the most computationally efficient detection strategy as per transit. The only computation performed is a Fourier-domain matched filter. The standard detector searches a transit space of size $|T|$ and therefore has a computational complexity $O(|T|N \log N)$. Detector A (Section 2.3.2) searches for a transit signal contained in non-WSS Gaussian noise; therefore, a time-domain matched filter must be used for each transit test. The search space is of size $|T|$ and the net computational complexity is $O(|T|N^2)$.

Detector B (Section 2.3.3) uses a larger search space than the other detection strategies as it includes transit depth $\alpha$; this search space was described in Section 2.4.3 and is of size $|T^\alpha|$. In addition to this increased search space, this detector must compute a MAP/MMSE systematics estimate once per transit, which is then used as input into a Fourier-domain matched filter. The net computational complexity is therefore $O(|T^\alpha|N^2)$.

Approximate elapsed wall-clock run times are summarized in Table 4. Transit detection tests were parallelized with one light curve assigned to each CPU core. All runs were performed on the Blue Waters petascale system at UIUC.

| Operation | Complexity |
|-----------|------------|
| Time-domain matched filter | $O(N^2)$ |
| Fourier-domain matched filter | $O(N \log N)$ |
| MAP/MMSE systematics estimation | $O(N^2)$ |
| Standard detector (Gaussian prior) | $O(|T|N \log N)$ |
| Detector A (Gaussian prior) | $O(|T|N^2)$ |
| Detector B (Gaussian prior) | $O(|T^\alpha|N^2)$ |

*Note. Where $|T|$ is the search space size. For detector B, the search space size $|T^\alpha|$ is generally larger as one must additionally search over candidate transit depths $\alpha$. |
Table 4

| Detector            | Run time per light curve (hr) | Run time per light curve per transit (s) |
|---------------------|-------------------------------|-----------------------------------------|
| Standard detector   | 1.5                           | 0.06                                    |
| Detector A          | 5                             | 0.18                                    |
| Detector B          | 30                            | 0.28                                    |

Note. The transit search space sizes for these runs are $|R| \approx 10^5$ and $|T| \approx 4 \times 10^5$. All detectors used a Gaussian prior. All run times are approximate elapsed wall-clock run times.

NCSA (Bode et al. 2013). This is a Cray XE/XK system with a peak performance of 13.34 PF.\footnote{https://bluewaters.ncsa.illinois.edu/hardware-summary}

\subsection*{2.4.5. Feasibility Test: Kepler Data Containing Exoplanets}

We conducted an initial feasibility test using detectors A and B and the standard detector on a subset of Kepler light curves that did not exclude known exoplanets. These tests were designed to demonstrate initial detection feasibility only on real exoplanet transit signatures. Detection tests were performed over the transit search space identical to that used in injection tests. The transit signal parameter search space is described in Section 2.4.3. A total of 2000 light curves were analyzed in this test, 1000 light curves were selected from CCD module 2 over observing quarters [Q2, Q10, Q14], and an additional 1000 light curves were selected from CCD module 12 over observing quarters [Q3, Q7, Q15]. These CCD modules and observing quarters were chosen randomly over time and across the CCD module. As was performed for the injection tests, the light curves for each module were first sorted by angular separation from the module reference point used by the MAST Kepler data archive before the first 1000 were selected. No explicit selection for CCD module output was applied: module 2 data included outputs 3 and 4, while module 12 data included only output 4. For our detection tests, we did not use stitched quarters but instead performed separate detection tests on each of the quarters and computed an averaged test statistic (per transit over time).

3. Results

\subsection*{3.1. Detector Performance: Injection Tests}

Detection efficiency and quasi-false-alarm rate are defined in the context of the recovery of injected signals in Section 2.4.3. A detection occurs whenever there is a test statistic above the detection threshold $\tau$.

We consider a correct detection of an injection signal to occur if the maximum test statistic above the threshold satisfies both of the following requirements: (i) the detected orbital period is within 3 hr of the true injected orbital period, and (ii) for an injected transit signal $t_+$ and detected transit signal $t_d$, the cosine similarity satisfies the condition $|\cos(t_+/|t_d|/|t_+|)| > 1/2$. This threshold ensures that for an injected and detected transit of identical duration, the error in estimated epoch does not exceed half the transit duration. We note that the limb-darkened injected transit has a different functional form from the detected periodic box transit function; the correlation value will therefore be slightly lower than expected for a correct match.

Because the purpose of these tests is a comparison of detection strategies, we do not seek to stringently reduce the false-alarm rate and thus require only two transit events (passes of an exoplanet) for a detection as opposed to the standard three transits (Burke & Catanzarite 2017).

When comparing Neyman–Pearson detectors, a detector is considered optimal if its detection rate is maximized for a fixed rate of false alarm (Kay 1993; Wasserman 2013). We adopt a detection threshold $\tau = 8.4$ in comparing the detection efficiency across the detectors considered here as it achieved a consistent quasi-false-alarm rate for these detectors of 13% ± 1%. The detection efficiency broken down by orbital period $P$ and ratio of planet-to-star radius $R_p/R_*$. of the injected transit are show in Figure 6 for the standard model. For the remaining detectors, we display the difference in detection efficiency relative to the standard model. This is depicted in Figure 7 for detector A and Figure 8 for detector B.
The detection rate as a function of quasi-false-alarm rate for the detectors is plotted in Figure 9.

### 3.2. Kepler Data Detections

As described in Section 2.4.5, an initial feasibility test was conducted using these detectors on a subset of 2000 Kepler light curves from which prior exoplanet detections were not excluded; this subset contained 17 confirmed exoplanets. Performance was measured at a detection threshold of $\tau = 7.5$, a lower detection threshold was used relative to the injection test value $\tau = 8.4$ given the expected suboptimal performance on real transit data. Detector A produced 252 detections, 9/17 of which are confirmed exoplanets and 76 are threshold crossing events (TCEs; Jenkins et al. 2010b). Detector B produced 219 detections, 9/17 of which are confirmed exoplanets and 67 are TCEs. The standard detector produced 187 detections, 8/17 of which are confirmed exoplanets and 59 are TCEs.

On visual inspection, we find no new convincing exoplanet candidates in the complete set of detections. A histogram of the detected orbital periods is shown in Figure 10, in which detections that are also TCEs are marked.

We emphasize that this initial feasibility test on Kepler data containing exoplanets is neither intended nor designed as a comparison of the statistical performance of these exploratory detectors against the Kepler science pipeline. The Kepler results are from a full multiquarter analysis and are used here only as a test of the initial feasibility of our detectors in recovering known exoplanets and demonstrating consistent results.

### 4. Discussion

As shown in Figures 6–8, in injection tests with a fixed threshold, the joint Bayesian detectors A and B (Section 2.2) achieve an overall detection efficiency improvement of $\sim 2\%$ over the reference standard processing model (Section 2.4.2). As noted above, the relative detection efficiencies were assessed at a comparable quasi-false-alarm rate for a fixed threshold.

As shown in Figure 9 for a fixed quasi-false-alarm rate, injection tests show that the $\sim 2\%$ improvement in detection efficiency for detectors A and B on average produced a $\sim 2\%$ increase in detection efficiency for the same quasi-false-alarm rate. The inset shows a magnified view of a region of the enclosing outer plot.

### Figure 9

A plot of detection efficiency against quasi-false-alarm rate in injection tests for decreasing threshold. Joint Bayesian detectors A and B on average produced a $\sim 2\%$ increase in detection efficiency for the same quasi-false-alarm rate. The inset shows a magnified view of a region of the enclosing outer plot.

### Figure 10

Histogram of detected Kepler orbital periods from a sample of 2000 light curves selected from CCD modules 2 and 12. Results for detectors A and B are shown as indicated. Those detections that are also TCEs are marked.

As shown in Figure 9 for a fixed quasi-false-alarm rate, injection tests show that the $\sim 2\%$ improvement in detection efficiency for detectors A and B relative to the standard model remains consistent for every quasi-false-alarm rate above 4%. Furthermore, below this rate, detectors A and B continue to outperform the standard model; this suggests that the joint detection strategies are Neyman–Pearson optimal (Wasserman 2013). As defined in Section 2.4.3, the quasi-false-alarm rate may overestimate the true false-alarm rate due to the presence of hitherto unknown actual detections in the Kepler data used for the injection tests despite the exclusion of confirmed exoplanet detections from these data. The quasi-false-alarm rate can, however, be argued as a reasonable proxy for the true false-alarm rate. Specifically, a false injection test detection does not require that the true injected transit signal be below the detection threshold, only that another transit signal produce a stronger test statistic. This condition can be expected to have low probability, however, given that prior undetected transit signals are likely to be weaker than the injected signals in general. In addition, although the quasi-false-alarm rate may overestimate the true false-alarm rate, we expect it to do so.
monotonically, lending validity to its use as a proxy in comparative detector studies.

Figure 9 shows that detector B marginally outperforms detector A. We speculate that this may be partially explained by deviations from Gaussianity between the sample and fitted systematic noise coefficient prior $p(c)$ discussed in Section 2.3.1. In a direct systems estimate (detector B), a broader prior $p(c)$ simply allows more variability about the mean to obtain a maximizing estimate under each hypothesis model. In a marginalization scheme (detector A), this may lead to weakened test statistics by including likelihoods for a number of improbable systematics estimates. In future work, we will explore methods to constrain the systematics prior to more closely approximate the central mode (see Figure 3).

Much of the improvement in detection efficiency for detectors A and B is concentrated in exoplanets with shorter orbital periods ($P < 10$ days) or with low ratios of planet-to-host-star radii $R_p/R_*$ ($<0.05\%$). Broadly, therefore, the improvements occur in short-period, low transit-depth populations within our sample. In an analysis of the detection efficiency of the Kepler pipeline (Christiansen et al. 2013, 2015), a drop-off in detectability was shown for exoplanets with orbital periods $P < 3$ days. The authors demonstrate that the process of removing harmonics (residual high-frequency stellar noise left over after PDC; Jenkins 2002; Jenkins et al. 2010b) prior to transit detection may distort short-period transit signals. The distribution of detection efficiency improvement as a function of orbital period for the detectors in the current work suggests that joint modeling of the systematic noise and transit signal mitigates this effect as it is better able to jointly differentiate between high-frequency noise and short-period transit signals. Specifically, we propose that a joint modeling approach, though computationally expensive, may be particularly effective when probing the aforementioned exoplanet populations. In general, it may be fruitful to use adaptive detection strategies in different parts of the transit parameter search space and in different signal-to-noise ratio (S/N) regimes. Further investigation of this effect over larger data samples is required. As noted earlier, Foreman-Mackey et al. (2015) have implemented a non-Bayesian joint estimation of systematic noise and the transit signal as a mitigating strategy for analogous overfitting, in their case primarily to address systematic errors due to pointing errors in the K2 mission. We echo their conclusion that these approaches have clear advantages in transiting exoplanet detection.

The Bayesian joint detectors described in the current work are computationally expensive (Section 2.4.4); however, we have demonstrated that such detectors can be used effectively in conjunction with the phase-correlation method (Averbuch & Keller 2002) applied to cotrended light curves. Phase correlation reduces the size of the transit signal parameter search space significantly and allows freed-up computational resources to be allocated to finer searches over other transit parameters such as orbital period. We note that it is possible that the use of phase correlation on cotrended light curves may have introduced a slight bias in favor of the standard method detector as the phase-correlation method finds the optimal phase estimate for a cotrended light curve. However, we do not believe this affects our conclusions from the current work. We note also that in the future, we propose exploring the use of ranked cross-correlation between the light curves and candidate transit signals in a generalized approach to optimize the identification of transit epochs.

As described in Section 2.3, the joint detection framework presented here admits many implementation choices and optimizations. As an exploratory evaluation of the statistical performance of these Bayesian joint detectors, we took care to maintain consistency between the implementations to allow meaningful relative comparisons but did not fine-tune the detectors to produce optimal detection rates. For example, the same priors and epoch estimates were used across all models. Also, the detection efficiency is computed over a single observing quarter of Kepler data; in practice, a folded test statistic across multiple quarters would likely be less vulnerable to poor data quality in a single quarter. Similarly, we have not yet evaluated adjunct techniques to enhance detection efficiency, including methods such as outlier detection or harmonic filtering. These alternative implementation choices will be explored in future work.

The initial joint detector feasibility test with Kepler data containing known exoplanet detections is described in Section 3.2 and the results depicted in Figure 10. These preliminary tests show that detectors A and B were able to recover known exoplanets at a rate comparable or marginally superior to the standard model. All detectors show a large number of spurious detections, particularly at larger orbital periods within the search window. The spurious detections are primarily due to the short data segment used and the small sample of light curves from which the systematic noise prior is built. However, we also believe that the spurious long-period detections may be reduced with more careful optimization of the detectors for maximal detection efficiency. This was not within the scope of the current work. Specifically, we believe that the lack of outlier rejection may be a contributing factor to the spurious long-period detections. These detections may also be reduced simply by utilizing more data over longer observational periods. We stress, however, that the joint detection tests with Kepler data containing exoplanets are preliminary in nature and primarily, though successful, an initial feasibility test.

5. Conclusions

We have developed a Bayesian framework for the joint detection of systematic noise and exoplanet transit signals. We formulated our detection framework as an LRT and used a Neyman–Pearson optimality criterion. Two general Bayesian approaches were used, namely marginalization over the systematic noise (detector A) and conditional estimation of the systematic noise (detector B). Under the assumption of a Gaussian prior for the systematic noise, we show that these detectors can be expressed in closed form as matched filters. The performance of the joint detectors was evaluated in numerical recovery tests of injected transit signals added to raw Kepler light curves. The Kepler data in the injection tests excluded known exoplanet detections. Further, an initial feasibility test was performed by applying the detectors to a subset of Kepler data from which confirmed exoplanet detections had not been excluded. An additional standard detector which performed sequential cotrending and detection was defined as a comparator during the numerical tests.

The principal conclusions of the paper are as follows:
1. In the injection tests, the joint Bayesian detectors A and B show an improvement of ~2% in overall detection efficiency relative to the standard detector. As an initial exploratory assessment, without significant detector optimization, the joint detectors therefore show sufficient promise to warrant further detailed investigation. We have identified several proposed detector efficiency optimizations.

2. The joint detectors A and B show specific improvement in detection efficiency for exoplanets with both short orbital periods ($P < 10$ days) and low ratios of planet-to-host-star radius $R_p/R_*$ ($<0.05\%$). We conclude that joint estimation offers improved separation of residual high-frequency systematic noise and overlapping transit signals and mitigates overfitting. We believe this approach has future potential in this regime specifically as well as in low-$S/N$ environments.

3. The Bayesian joint detectors are computationally expensive but we have shown that they are tractable with contemporary high-performance computing resources. The Bayesian approach offers the advantage of full statistical generality regarding the form of the probability distribution adopted for the statistical noise and stellar signal and the parameter estimators used. We have demonstrated that phase correlation can be used in conjunction with this method to reduce significantly the transit parameter search space and thereby the net computational complexity.

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This paper includes data collected by the Kepler mission. Funding for the Kepler mission is provided by the NASA Science Mission directorate.

Appendix A

Phase-correlation Method

The phase-correlation method (Averbuch & Keller 2002) is a classic image-registration technique. In the current work, this method is used to estimate the epoch $t_0$ of a candidate transit signal from a coarsely cotrended light curve. This method estimates the offset $t_0$ between a signal $x(t)$ and a shifted version of this same signal $x(t-t_0)$. Denote the Fourier transform of $x(t)$ as $X(\omega) = \mathcal{F}\{x(t)\}$. The Fourier shift theorem (Bracewell & Bracewell 2000) yields

$$\mathcal{F}\{x(t-t_0)\} = X(\omega) e^{-jt_0}. \quad \text{(A1)}$$

The pixel-level phase-correlation method (Averbuch & Keller 2002) uses this property to form the normalized cross-power spectrum as an estimate of this phase shift:

$$\frac{X(\omega)X^*(\omega)e^{jt_0}}{|X(\omega)||X(\omega)e^{-jt_0}|} = e^{jt_0}. \quad \text{(A2)}$$

The inverse Fourier transform of this expression yields a delta function centered on the position of the shift $t_0$, thus providing a local maximum. This method is also effective if the observed signal contains additive noise or is improperly scaled (Averbuch & Keller 2002).

Appendix B

Extension to Multiple Quarters

In this appendix, we discuss the extension of our current detectors to the case of multiple observing quarters, although we stress that these methods were not applied in the current work.

The complexity of the detectors grows polynomially with $N$ as described in Section 2.4.4. As such, application to long time-series data becomes prohibitively expensive. We will briefly outline how the detectors may be applied to multiquarter searches with linear growth in complexity.

The form of the generalized matched filter described in Section 2.2.1 depends strongly on the form of the signal covariance matrix. If we assume data between quarters to be uncorrelated, a covariance matrix describing all observations will be of block-diagonal form. For $Q$ quarters, where each quarter has the covariance matrix $\text{Cov}_q$: $q \in Q$, we obtain a multiquarter covariance matrix $\text{Cov}$:

$$\text{Cov} = \begin{bmatrix} \text{Cov}_1 & \text{Cov}_{12} & \cdots & \text{Cov}_{1Q} \\ \text{Cov}_{21} & \text{Cov}_2 & \cdots & \text{Cov}_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}_{Q1} & \text{Cov}_{Q2} & \cdots & \text{Cov}_Q \end{bmatrix}. \quad \text{(B1)}$$

This allows a linear decomposition of the matched filter:

$$\frac{y^T \text{Cov}_n^{-1} t}{\sqrt{t^T \text{Cov}_n^{-1} t}} = \sum_{q \in Q} q^T \text{Cov}_q^{-1} t_q. \quad \text{(B2)}$$

In this form, the total computational cost is the complexity of a single quarter multiplied by the total number of quarters.

Given the definition of detector A in Section 2.2.2 under the systematic noise model of Section 2.1.1, and the stellar models in Section 2.1.2, the multiquarter covariance matrix $\text{Cov}$ will in fact be in block-diagonal form. For detector B, further optimization is possible due to the Toeplitz structure of the stellar covariance within a given quarter. In addition, for detector B, the Toeplitz nature of $\text{Cov}$ allows each term within the summand to be computed efficiently in the Fourier domain.

ORCID iDs

Jamila S. Taaki @ https://orcid.org/0000-0001-5475-1975
Athol J. Kemball @ https://orcid.org/0000-0001-6233-8347

References

Averbuch, A., & Keller, Y. 2002, ITASS, 4, 3608
Basri, G., Walkowicz, L. M., Batalha, N., et al. 2010, AJ, 141, 20
Bode, B., Butler, M., Dunning, T., et al. 2013, in Contemporary High Performance Computing, ed. J. W. Vetter (Boca Raton, FL: CRC Press), 339
Borucki, W. J., Koch, D., Basri, G., et al. 2010, Sci, 327, 977
Borucki, W. J., Scargle, J. D., & Hudson, H. S. 1985, ApJ, 291, 852
Bracewell, R., & Bracewell, R. 2000, The Fourier Transform and Its Applications (New York: McGraw Hill)
Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al. 2013, PASP, 125, 889
Burke, C. J., & Catanzarite, J. 2013, Planet Detection Metrics: Per-Target Flux-Level Transit Injection Tests of TPS for Data Release 25, Kepler Science Document, KSCI-19109-002
Burke, C. J., Gaudi, B. S., DePoy, D., & Pegge, R. W. 2006, AJ, 132, 210
Carter, J. A., & Winn, J. N. 2009, ApJ, 704, 51
Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, ApJS, 207, 35
Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2015, ApJ, 810, 95
Courty, C., Strader, I., van Dokkum, P., et al. 2018, ApJ, 864, 111
