Draft version November 6, 2017
Typeset using LaTeX **twocolumn** style in AASTeX61

A NEW STANDARD FOR ASSESSING THE PERFORMANCE OF HIGH CONTRAST IMAGING SYSTEMS

Rebecca Jensen-Clem,[1, *] Dimitri Mawet,[2] Carlos A. Gomez Gonzalez,[3, 4] Olivier Absil,[3, †] Ruslan Belikov,[5] Thayne Currie,[6] Matthew A. Kenworthy,[7] Christian Marois,[8, 9] Johan Mazoyer,[10, 11] Garreth Ruane,[2] and Angelle Tanner[12]

[1] *Astronomy Department, University of California, Berkeley, Berkeley, CA 94720, USA*
[2] *Department of Astrophysics, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91101, USA*
[3] *Space sciences, Technologies and Astrophysics Research (STAR) Institute, Université de Liège, 19 Allée du Six Août, B-4000 Liège, Belgium*
[4] *Université Grenoble Alpes, IPAG, F-38000 Grenoble, France*
[5] *NASA Ames Research Center, Moffett Field, CA 94035, USA*
[6] *National Astronomical Observatory of Japan, Subaru Telescope, 650 A'ohoku Pl., Hilo, HI 96720, USA*
[7] *Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands*
[8] *National Research Council of Canada Herzberg, 5071 West Saanich Rd, Victoria, BC, Canada V9E 2E7*
[9] *University of Victoria, 3800 Finnerty Rd, Victoria, BC, Canada V8P 5C2*
[10] *Johns Hopkins University, Zanvyl Krieger School of Arts and Sciences, Department of Physics and Astronomy, Bloomberg Center for Physics and Astronomy, 3400 North Charles Street, Baltimore, MD 21218, USA*
[11] *Space Telescope Science Institute, 3700 San Martin Dr, Baltimore MD 21218, USA*
[12] *Mississippi State University, Department of Physics & Astronomy, Hilbun Hall, Starkville, MS, 39762, USA*

ABSTRACT

As planning for the next generation of high contrast imaging instruments (e.g. WFIRST, HabEx, and LUVOIR, TMT-PFI, EELT-EPICS) matures, and second-generation ground-based extreme adaptive optics facilities (e.g. VLT-SPHERE, Gemini-GPI) are halfway through their principal surveys, it is imperative that the performance of different designs, post-processing algorithms, observing strategies, and survey results be compared in a consistent, statistically robust framework. In this paper, we argue that the current industry standard for such comparisons – the contrast curve – falls short of this mandate. We propose a new figure of merit, the "performance map," that incorporates three fundamental concepts in signal detection theory: the true positive fraction (TPF), false positive fraction (FPF), and detection threshold. By supplying a theoretical basis and recipe for generating the performance map, we hope to encourage the widespread adoption of this new metric across subfields in exoplanet imaging.

[*] Miller Fellow
[†] F.R.S.-FNRS Research Associate

1. INTRODUCTION

The contrast curve describes an imaging system's sensitivity for a given detection significance in terms of the planet/star flux ratio and angular separation. A consistent methodology for computing the contrast curve, however, is lacking: a variety of approaches to throughput, small sample-size, and non-Gaussian noise corrections are represented in the literature (e.g. Marois et al. 2008a; Wahhaj et al. 2013; Mawet et al. 2014; Pueyo 2016; Otten et al. 2017). As inner working angles are pushed below $5\lambda/D$, these details dominate the calculation of the contrast curve. Secondly, the contrast curve's information content is limited: by fixing the detection significance for all separations, the contrast curve conceals important trade offs between the choice of detection threshold, false positive rates, and detection completeness statistics.

The purpose of this paper is to critically examine the contrast curve and present alternative figures of merit for the ground and space-based exoplanet imaging missions of the coming decades. In Section 2, we summarize the key points of signal detection theory, which provide the basis for our discussion of performance metrics. Section 3 describes the strengths and weaknesses of the contrast curve as a general purpose performance metric. Finally, Sections 4 and 5 give our proposal for a new figure of merit based on signal detection theory.

2. OVERVIEW OF SIGNAL DETECTION THEORY

Our task as planet hunters is to decide whether the data at each location in a "high contrast" image meets our threshold for a planet detection. Regardless of the details of the dataset (e.g. field rotation, spectral coverage, etc.), the presence of noise will interfere with the accuracy of our detection decisions. Signal detection theory provides a precise framework for describing the relationships between detections, non-detections, and detection thresholds.

If we assume that a planet is present at a location of interest in our data (the H_1, or "signal present" hypothesis), and we succeed in detecting that planet, our result is a true positive (TP). If we fail to detect the planet, our result is a false negative (FN). Clearly, we aim to maximize the number of true positives while minimizing the number of false negatives. Hence, we define a true positive fraction, or TPF:

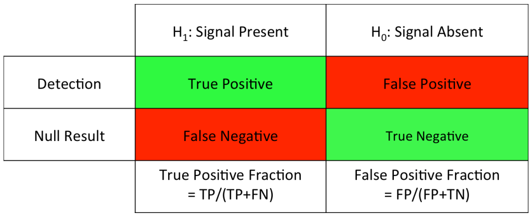
$$\mathrm{TPF} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \int_{\tau}^{+\infty} pr(x|H_1)dx \qquad (1)$$

where τ is the detection threshold and $pr(x|H_1)$ is the probability density function (PDF) of the data x under the hypothesis H_1. Our goal is to approach TPF= 1.

If we instead assume that no planet is present in the data (the H_0, or "signal absent" hypothesis), and we fail to make a detection, our result is a true negative (TN). If we incorrectly claim to detect a planet, however, our result is a false positive (FP). We are then interested in achieving a false positive fraction (FPF) close to zero:

$$\mathrm{FPF} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} = \int_{\tau}^{+\infty} pr(x|H_0)dx. \qquad (2)$$

These various hypotheses and outcomes are summarized in the "confusion matrix" (Figure 1). An early review of signal detection theory is given by Swets et al. (1961).



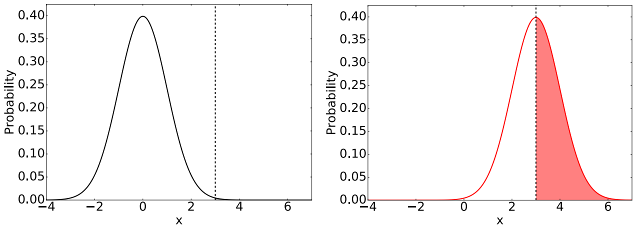
	H_1: Signal Present	H_0: Signal Absent
Detection	True Positive	False Positive
Null Result	False Negative	True Negative
	True Positive Fraction = TP/(TP+FN)	False Positive Fraction = FP/(FP+TN)

Figure 1. The confusion matrix

To make these relationships concrete, consider a post-processed image in which the intensities, x, in a series of photometric apertures located at a certain distance from the central star are drawn from a normal distribution ($\mu = 0$ and $\sigma = 1$, where the choice of an annular region is justified by the symmetry of the star's point spread function). The PDF of the noise is shown in Figure 2a. Now let us assume that our goal is to detect a planet with a mean intensity of $x = 3$ inside the annulus of interest. Because the intensity in the photometric aperture at the planet's location is also affected by the noise, it is described by a PDF identical to that of the noise, but with a mean of $x = 3$ (here, we ignore the contribution of the planet's shot noise). The PDF of the signal is shown in Figure 2b.

Given our knowledge of the PDFs of the noise and the signal, we now wish to choose a detection threshold. Let us assume that because our detection follow-up resources (e.g. telescope time) are limited, we wish to achieve a false positive fraction of 0.001. We therefore choose a detection threshold of 3σ because a fraction 0.001 of the area of the noise PDF falls above this value (2a, dotted line). A second consequence of this choice of detection threshold is that we will only detect half of all planets with a mean intensity of $x = 3$ (TPF= 0.5; 2b, dotted line). If we wish to increase the TPF, we must lower the detection threshold and hence unfavorably raise the FPF.

Our choice of detection threshold therefore allows us to trade between the FPF and TPF, within the con-



(a) Noise distribution (b) Signal distribution

Figure 2. (a) The normally distributed PDF of a noise source with a mean of zero and standard deviation of one. Here, the detection threshold is arbitrarily set to 3σ (dashed line), which corresponds to $x = 3$ for this distribution. Because the noise PDF falls above the detection threshold a fraction 0.001 of the time, the false positive fraction in this example is 0.001. (b) The Gaussian PDF of a signal source with a mean of $x = 3$ and a standard deviation of one. Because half of the signal distribution's area falls above the detection threshold, the true positive fraction is 0.5.

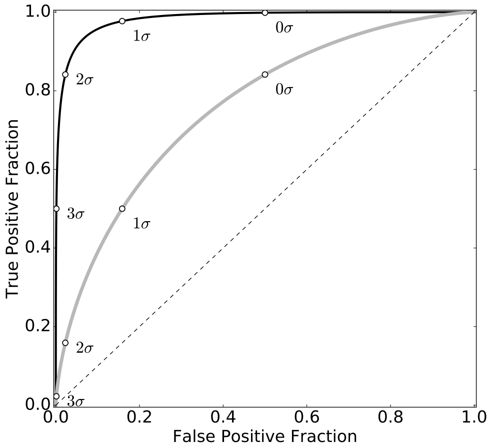


Figure 3. Black line: an ROC curve corresponding to a range of detection thresholds applied to the normal noise and signal distributions in Figure 2. The (TPF, FPF) locations corresponding to thresholds of $0\sigma - 3\sigma$ are labeled to demonstrate the trade-offs between these key parameters. Grey line: the equivalent ROC curve for a signal distribution centered at $x = 1$.

straints imposed by the noise PDF and the signal mean. The receiver operator characteristic (ROC) curve allows us to visualize this trade by plotting the TPF as a function of the FPF, with each parameter varying between 0 and 1 as we move the detection threshold from large to small values (Tanner & Swets (1954) gives an early example of an ROC curve; Krzanowski & Hand (2009) provide an updated discussion of the topic). The black line in Figure 3 shows the ROC curve associated with our example. The (TPF, FPF) pair corresponding to our example threshold of 3σ is labeled, along with a broader range of possible detection threshold choices. We note that the detection threshold must be less than the mean of the noise distribution to produce FPF values greater than 0.5. Because the mean is zero in this example in this example, such thresholds are negative. While mathematically consistent, negative thresholds have no observational relevance.

The shape of the ROC curve is determined by the shape of the noise distribution as well as the signal mean. For example, if we change the mean of the signal distribution in Figure 2b from $x = 3$ to $x = 1$, we obtain the gray ROC curve shown in Figure 3. Because the noise distribution was unchanged, the black and gray curves' (TPF,FPF) pairs corresponding to detection thresholds of $0\sigma - 3\sigma$ share identical FPF values. Alternatively, if we had chosen a positively skewed rather than a normal noise distribution, the nearly vertical part of the black ROC curve at small FPFs would tilt to the right.

We may now describe our goal of characterizing the detection statistics of an exoplanet imager in the vocabulary of signal detection theory: we wish to determine the maximum FPF and minimum TPF that satisfy our resource limitations and science goals – in other words, we must choose a target location in (TPF, FPF) space. Our goal in designing an instrument, observing strategy, or post-processing routine is to produce a noise distribution whose ROC curve will reach that location for a signal of interest.

An ROC curve, however, only represents a single noise distribution (i.e. image location) and signal level. In the sections that follow, we will discuss methods for representing the performance of a full image.

3. CONTRAST CURVES AS PERFORMANCE METRICS

3.1. *The Definition of the Contrast Curve*

The contrast curve is a means of representing the true and false positive fractions associated with a range of signals and positions in a final image. Schematically, we can define the contrast as:

$$\text{contrast} = \left(\frac{\text{factor} \times \text{noise}}{\text{stellar aperture photometry}}\right)\left(\frac{1}{\text{throughput}}\right) \tag{3}$$

where the numerator is the detection threshold, expressed as a multiple of the noise distribution's width. Often, the width of the noise distribution (here, the "noise") is chosen to be the standard deviation of the resolution element intensities at a given separation from the star (e.g. Figure 4), while the multiplicative "factor" is chosen to be three or five to produce a 3σ or 5σ contrast curve. In Figure 2, factor = 3 and noise

$= \sigma = 1$. The detection threshold is then converted to a fraction of the parent star's brightness via the "stellar aperture photometry" term. Finally, the "throughput" term corrects this brightness ratio for any attenuation of the off-axis signal relative to the star's (e.g. due to field-dependent flux losses imposed by the coronagraphic system and post-processing algorithms). The final contrast is therefore the planet-to-star flux ratio of a planet whose brightness is equal to the detection threshold. Figure 2 illustrates that the TPF associated with such a signal is 0.5. Hence, the contrast curve can be interpreted as the signal for which we achieve 50% completeness given our choice of detection threshold in the numerator. The numerator also fixes the false positive fraction – for example, choosing factor= 3 for a white noise distribution gives FPF = 0.001. Finally, it is important to note that the contrast curve's statistics refer to planet detectability, and not to the photometric accuracy associated with any given planetary signal.

3.2. *Where the Contrast Curve Falls Short*

Both practical and fundamental shortcomings, however, undermine the utility of the contrast curve as a general purpose performance metric. First, the contrast is inflexible: by fixing the true positive fraction to 0.5 and the false positive fraction to a value set by the numerator, we cannot explore the (TPF, FPF, detection threshold) trade space. Even if we were to plot multiple contrast curves on the same figure to show different detection thresholds, we could not escape the arbitrary choice of TPF= 0.5. Similarly, if we were to plot a 90% detection completeness curve as a function of separation, we could not access a range of false positives fractions. Finally, fixing the TPF, FPF, and detection threshold for all separations may not be desirable for all applications – because the number of resolution elements, the PDF of the noise, and the predicted population of planets all vary as a function of separation, a particular imaging program's science goals may be better served by a detection threshold that also varies with separation.

More problematic, however, is the calculation of the terms in Equation 3. As mentioned above, the "noise" term is typically chosen to be the standard deviation of resolution elements in a region of the image, whose shape and size widely varies in the literature. This approach is valid if two conditions are met: 1) if the region includes enough statistically independent realizations of the noise to allow for an accurate measure of the distribution's standard deviation, and 2) if the underlying noise distribution is Gaussian. While there is no hard and fast rule for deciding whether the first condition is met, statisticians generally consider 30 independent samples to be the boundary between large and small sample statistics (Wilcox 2009). For the case of 1 λ/D-wide annular regions, 30 samples corresponds to a separation of $\sim$ 5λ/D. Below this threshold, the sample standard deviation is an increasingly uncertain estimate of the width of the underlying noise distribution (Student 1908; Mawet et al. 2014). The mitigating strategy proposed by Mawet et al. (2014), however, also requires that condition #2 (Gaussian noise) is met. Aime & Soummer (2004) and many others have shown that uncorrected low-order wavefront aberrations cause the noise at small separations to follow a positively skewed modified rician distribution rather than a normal distribution (Perrin et al. 2003; Bloemhof 2004; Fitzgerald & Graham 2006; Soummer et al. 2007; Hinkley et al. 2007; Marois et al. 2008a). While numerous observing and post-processing strategies have been employed to whiten this skewed distribution (e.g. Liu 2004; Marois et al. 2006; Lafrenière et al. 2007; Amara & Quanz 2012; Soummer et al. 2012), their success at small separations is limited by the temporal and spectral variability of the noise (Appendix A discusses the difficulty of testing for normality using methods such as the Shapiro-Wilk test). The result is that the noise distribution at small angles retains an unknown skewness at small separations that increases the false positive fraction compared to a Gaussian distribution. Hence, neither condition for the use of the standard deviation as a proxy for the FPF is met at small separations[1]. In Section 5.2 we will address alternative methods for probing the distribution of the noise without the assumption of normality.

3.3. *Inconsistencies in Contrast Curve Computations*

We further note that the contrast and its constituent terms are inconsistently computed in the literature, in particular the noise and throughput terms. While many authors (e.g. Wahhaj et al. 2013) account for spatially correlated speckle statistics by defining the noise to be the standard deviation of resolution elements in an annulus, others do not. For example, Otten et al. (2017) define the noise in relation to the standard deviation of pixel values inside of a single 1 λ/D aperture of interest. The region within a few λ/D of the inner working angle, however, is fundamentally sensitive to azimuthally

[1] It is worth noting that some authors interpret the numerator of Equation 3 as an empirical signal to noise threshold without reference to the distribution of the noise or a false positive fraction. This interpretation, however, robs the contrast curve of much of its practical use – the knowledge that we can achieve TPF= 0.5 for a given planet:star flux ratio does not guide our observing or science if the associated false positive fraction can fall anywhere from zero to one.

correlated speckle noise: effects such as pointing jitter, thermal variations, and non-common path aberrations induce low order wavefront aberrations, and hence close-in, variable speckles, on the timescale of an observation (Shi et al. 2016). Secondly, the definition of the term "throughput" is context dependent. Authors computing contrast curves for angular differential imaging (ADI) datasets typically define the throughput in terms of the flux losses imposed by signal self-subtraction (e.g. Wahhaj et al. 2013). However, in discussions of coronagraph design trades, throughput refers to the often field-dependent flux losses caused by the coronagraphic system itself (e.g. Guyon et al. 2006; Krist et al. 2015). Finally, the small sample correction presented by Mawet et al. (2014) has been adopted by some authors (e.g. Wertz et al. 2017), but not others (e.g. Uyama et al. 2016). Such a variety of methodologies inhibit meaningful comparisons of instrument performance.

In this section, we have described three shortcomings of the contrast curve: 1) its inability to illustrate the (TPF, FPF, detection threshold) trade space, 2) its potential inconsistency with the shape of the underlying noise distribution, and 3) its inconsistent treatment in the literature. In the sections that follow, we will discuss strategies for computing the FPFs and TPFs associated with an unknown noise distribution and present a new figure of merit for the performance of high dynamic range imaging systems.

3.4. *The Raw Contrast*

The above discussions concern what we might call an "observer's" definition of the contrast. Users of exoplanet imaging testbeds, however, refer to the "raw contrast," which is typically defined as

$$\text{raw contrast} = \frac{\text{mean}[R(x,y)]}{\max[\text{PSF}_{\text{star}}(x,y)]} \tag{4}$$

where mean$[R(x,y)]$ is the mean number of photons per pixel over a region of interest (for example a dark hole) and max$[\text{PSF}_{\text{star}}(x,y)]$ is the number of photons in the pixel corresponding to the peak of a stellar PSF offset to a representative location inside of the region of interest. The key difference between the raw contrast and the observer's contrast is that the raw contrast does not refer to an astrophysical flux ratio – a raw contrast of 10^{-10} does not indicate that a planet with an astrophysical flux ratio of 10^{-10} is in any sense detectable. Rather, it simply indicates that the mean intensity of the background in a certain region is 10^{10} smaller than the peak of the offset stellar PSF. Hence, while the raw contrast is a useful shorthand for describing an instrument's starlight suppression, it should not be interpreted as a detection limit. Obtaining a detection limit by estimating the noise inside of the region of interest carries with it the attendant dangers of small sample statistics and non-Gaussian noise described in Section 3.2 as well as exposure time dependencies and signal throughput effects.

4. REPRESENTING THE (FPF, TPF, SEPARATION) TRADESPACE WITH THE PERFORMANCE MAP

In Section 3.2, we argued that the contrast curve's limited information content – the astrophysical flux ratios of those planets that give TPF=0.5 for a single detection threshold as a function of separation – obscures the much richer (FPF, TPF, separation) trade-space. Here, we propose two modifications to the contrast curve: 1) a detection threshold (and hence FPF) that varies with separation, and 2) the inclusion of all possible TPFs as a heatmap.

When the detection threshold is held constant with separation, the radial distribution of false positives is not uniform because the number of resolution elements varies with separation. If the expected number of false positives N_{FP} is given by $N_{\text{FP}}(r) = \text{FPF} \times 2\pi r$ for separation r, then a constant detection threshold (and hence a constant FPF) allows more total false positives at wide separations than at small separations. If we instead keep the radial distribution of false positives constant, we allow the detection threshold to adapt to the changing number of resolution elements with separation (see Ruane et al. (2017) for a similar approach).

Next, we plot the astrophysical flux ratios of those planets that give any desired TPF as a function of separation. Rather than choosing a single TPF contour, we propose to show the full $0 \leq \text{TFP} \leq 1$ space as a heatmap. A representative TPF contour can be overplotted for clarity.

We call this modified figure the performance map (e.g. Figure 6). We argue that the performance map highlights the most scientifically and programmatically relevant quantities, namely the TPFs of the signals of interest for a given number of false positives. The contrast curve, on the other hand, highlights the detection threshold, which has no intrinsic meaning beyond pointing to a false positive fraction.

5. GENERATING THE PERFORMANCE MAP

Constructing the performance map requires knowledge of the false positive fraction which in turn requires knowledge of the underlying noise distribution. As discussed in Section 3.2, the distribution of the noise at small separations is often unknown. In the following subsections, we consider two limiting cases: 1) the PDF of the noise is Gaussian (Section 5.1), and 2) the PDF of the noise is completely unknown (Section 5.2).

Following Mawet et al. (2014), we define a resolution element to be a circular aperture with a diameter of λ/D. The number of resolution elements, N_r at a distance r from the central star is $2\pi r$, where r is also expressed in terms of λ/D (Figure 4). We consider only whole numbers of resolution elements (e.g. six resolution elements at 1λ/D.).

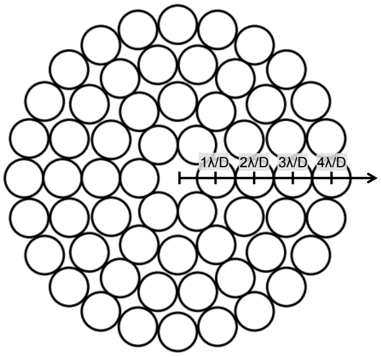


Figure 4. The number of resolution elements of width λ/D at a distance r from the central star is $2\pi r$, where here we consider only whole numbers of resolution elements.

To illustrate the construction of a performance map in detail, we consider a set of HR8799 observations taken by the Spectro-Polarimetric High-contrast Exoplanet REsearch (SPHERE, Beuzit et al. 2008) at the Very Large Telescope (VLT). The data were acquired in December of 2014 during science verification of the Infra-Red Dual-band Imager and Spectrograph (IRDIS, Dohlen et al. 2008) instrument, and have been extensively described in the literature (Zurlo et al. 2016; Apai et al. 2016; Wertz et al. 2017). We adopt a 200 frame broadband H filter ($1.48 - 1.77\,\mu$m) sequence from December 4th 2014, where the detector integration time was 8 s and the total amount of parallactic angle rotation was $8^\circ.7$. We choose to include only the data taken on the left-hand side of the IRDIS detector.

Following Wertz et al. (2017), we use an off-axis broadband H image of β Pictoris (January 30th 2015, PI: A.-M. Lagrange) as our PSF template due to the absence of an off-axis exposure in the original observing sequence. We fit a 2D Gaussian function to the β Pictoris template PSF to obtain FWHM $= 4.0$ pixels $= 0''.049$ for a plate scale of 12.251 mas (Wertz et al. 2017). Because this measured FWHM is slightly larger than the diffraction limit would suggest ($0.98 \times \lambda$/D$= 0''.040$), we conservatively adopt the FWHM as the resolution element diameter rather than $1\,\lambda$/D.

For the purposes of this demonstration, we are interested in estimating the FPFs and planet-injected TPFs. Hence, we begin our reduction by subtracting HR8799 bcde from the dataset. This is accomplished via the Vortex Image Processing (VIP, Gomez Gonzalez et al. 2017) package's functions for injecting negative fake companions into the data and optimizing their flux and positions using a Nelder-Mead based minimization.

Next, we use VIP's implementation of the PCA-ADI algorithm to subtract a reconstructed datacube from our set of 200 images. The reconstructed cube was generated using three principal components (chosen to maximize the SNR of HR8799 c in a full reduction of the dataset prior to planet subtraction). We median-combine the residual datacube to obtain our final reduced image. We compute the algorithmic throughput (signal self-subtraction) as a function of separation by injecting fake planets at separation intervals of 1 FWHM and azimuthal intervals of 120°. For each separation interval, the data is PCA-ADI reduced, and the signals' flux attenuation in the three azimuthally separated apertures are averaged.

Here, we consider only the first ten separation intervals after the inner working angle (in this case FWHM$=$ $2 - 11$). Figure 5 shows a 3σ contrast curve generated with VIP (where algorithmic throughput and small sample statistics are properly accounted for).

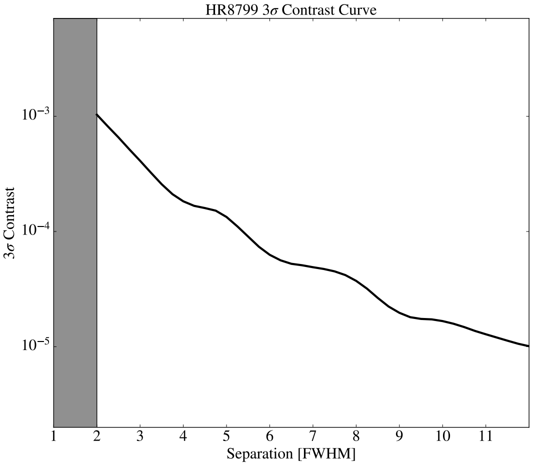


Figure 5. A contrast curve representing the observation of HR8799 with SPHERE described in Section 5.

5.1. *The Gaussian Assumption*

We first consider the most straightforward path to constructing the performance map: assuming that the

PDF of the noise follows a Gaussian distribution with a width corresponding to the measured standard deviation of the resolution elements as a function of separation (acknowledging the uncertainty in the standard deviation due to small sample statistics). Because any calculation of the FPF requires the hypothesis H_0 (signal absent), we are assuming that any detections are false, despite the reality that there may be true planets in the data.

To choose the FPF (and hence the detection threshold) for each separation, we must first choose the total number of false positives that we are willing to accept in the FWHM$= 2-11$ region of interest. For example, perhaps we have sufficient telescope time to follow up one false positive for every ten observations. Hence, we can accept 0.1 false positives per image, or FPF$= \mathbf{0.01}/N_r$. For each separation we then derive the corresponding detection threshold that will connect the FPFs to the TPFs of the injected signals. Here, the threshold is given by the quantile (inverse CDF) function of the Student T distribution with N_r degrees of freedom and a width scaled by the measured standard deviation of the resolution elements at r. We can then inject planet signals to determine the TPF of a given signal at a given separation. For the purposes of this simplified demonstration, the TPF is computed using the CDF of the scaled Student T distribution representing the noise, but shifted by the throughput-corrected test signal.

The resulting performance map is shown in Figure 6 with the TPF$= 0.5$ contour overplotted.

We emphasize that the "depth" of the TPF$= 0.5$ contour in Figure 6 is different from that of the contrast curve in Figure 5 because the performance map is illustrating a lower false positive rate in this example. Furthermore, the performance map allows the detection threshold to vary with separation, while the contrast curve holds the detection threshold fixed.

5.2. *The Empirical Performance Map*

In the preceding section, we considered an ideal scenario in which the PDF of the noise was known and the false positive fractions could be computed analytically. In Section 3.2, however, we argued that the PDF of the noise at small separations is difficult to determine given the effects of imperfect speckle subtraction.

To address this effect, we now consider an extreme case where the PDF of the noise is completely unknown, and the FPFs must be determined empirically: for each separation, we will simply count the number of resolution elements that exceed a test detection threshold. For a single 1 λ/D-thick annulus in the final, post-processed image, the possible values of the empirical false positive fraction are therefore constrained to i/N_r, where i is an integer between zero and N_r (inclusive). Accessing desirable FPFs between zero and $1/N_r$ requires additional realizations of the noise – for example, data from the same instrument can provide additional resolution elements if the distribution of the noise is assumed to be constant with time. Ruffio et al. (2017) describe the application of this technique to the Gemini Planet Imager Extra Solar Survey (GPIES) campaign. Another possibility for the case of ADI data is obtaining an additional image "for free" by reversing the order of the parallactic angle assignments (Wahhaj et al. 2013). This produces an image with similar azimuthal noise characteristics to the science image, doubling the number of noise realizations. Further angle randomization, however, will artificially whiten the speckle noise and fail to capture the temporal speckle evolution that de-rotation translates into azimuthal variation.

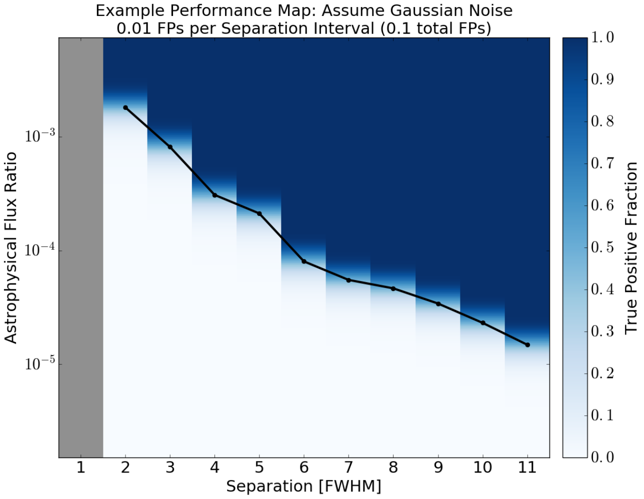


Figure 6. An example performance map where the FPFs have been calculated under the assumption that the noise is Gaussian at all separations. Here, the detection thresholds that connect the FPFs to the TPFs of the injected signals were chosen to give 0.01 false positives per separation interval, or 0.1 total false positives in the FWHM$= 2 - 11$ region.

To generate a performance map from a single image using this empirical FPF approach, we first make a list of FPFs for a range of detection thresholds and separations by the following recipe:

1. Draw rings of FWHM-diameter apertures around the central star (see Figures 4) and sum the fluxes inside of the apertures. The result is a list of $2\pi r$ aperture sums for each separation r.

2. Choose a detection threshold.

3. For each separation, find the fraction of resolution elements whose sum exceeds the detection threshold. These are the FPFs.

4. Vary the detection threshold and repeat Step 3 to produce all possible FPF values for all separations.

Using the same set of detection thresholds and separations as the preceding recipe, we can find the associated TPFs for a range of planet signals of interest. This is accomplished by the following steps:

1. Sum the flux inside of a FWHM-diameter aperture around the unocculted stellar PSF[2].

2. Choose an astrophysical flux ratio and multiply by the star's aperture sum (previous step) to obtain the planet's signal.

3. For each separation, multiply the planet's signal by the algorithmic throughout (previous paragraph), and add the result to each resolution element.

4. Choose a detection threshold from the same list of threshold used to generate the FPFs above.

5. For each separation, find the fraction of resolution elements whose sum exceeds the threshold. These are the TPFs.

6. For the same range of detection thresholds used to calculate the FPFs, repeat Step 5.

7. Repeat Steps 2-6 for different astrophysical flux ratios.

To plot the performance map, we elect to consider the smallest detection thresholds associated with the least non-zero FPFs ($1/N_r$), giving 1.0 false positive per $1\,\lambda$/D separation interval. For each injected signal at each separation, we then plot the TPFs corresponding to these detection thresholds. Figure 7 shows the resulting performance map. For each separation, we also plot the signal with the TPF nearest to TPF= 0.5 for these detection thresholds.

We may now compare the total number of false positives in the empirical performance map above with that of the contrast curve. The choice the 3σ detection threshold used to compute the contrast curve implies a false positive fraction of 0.0013 under the assumption that the noise is Gaussian. To obtain the total number of false positives in the image under this assumption, the false positive fraction is multiplied by the total number of whole resolution elements, N_T. For the $2-11\,\lambda$/D region of the image considered here, $N_T = 403$ and the total number of false positives is $N_{FP} = FPF \times N_T \approx 0.54$. In comparison, the empirical performance map approach makes no assumptions about the PDF of the noise, and gives one false positive per separation, or $N_{FP} = 10$ in this example. Given additional realizations of the noise (e.g. observations of other targets in a homogeneous survey), however, the least non-zero FPF is $1/(N_r N_f)$, where N_f is the total number of frames. In this example, reducing N_{FP} from 10 to the 3σ white noise case of $N_{FP} = 0.54$ would require $10/0.54 \approx 19$ additional images to increase the total number of resolution elements available at each separation.

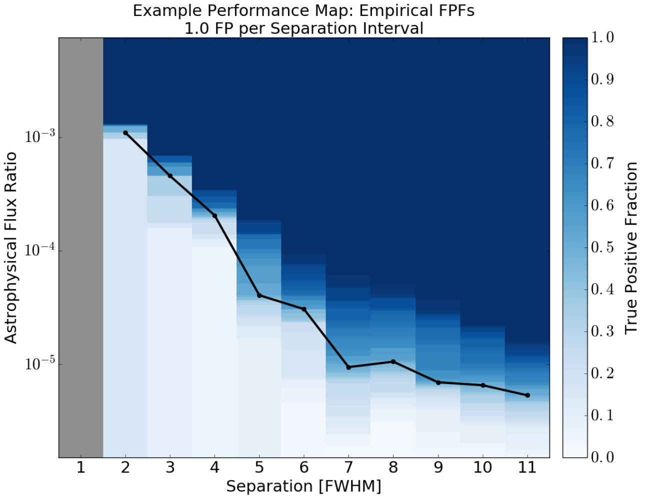


Figure 7. The performance map shows the astrophysical flux ratio versus the separation, color-coded by the true positive fraction. The solid black line represents the approximate TPF = 0.5 contour.

While the empirical approach described here does require many observations to reach the small FPFs promised by the Gaussian noise assumption, it will eventually yield the correct connections between the FPFs, detection thresholds, and TPFs as the number of images increases, regardless of the underlying PDF of the noise. Hence, such an approach is particularly appealing for large surveys, and less appealing for a single observation.

6. HYPOTHESIS TESTING

In the discussion above, our calculation of the true and false positive fractions required a choice of hypothesis:

[2] As mentioned above, our example dataset lacks an unocculted image, but we fit the positions and fluxes of the HR8799 planets using a later off-axis observation of β Pictoris. For the purposes of this example, we estimate HR8799's unocculted aperture sum based on the fitted flux of HR8799 b and the H-band planet-to-star flux ratio given in Marois et al. (2008b).

either H_1 (signal present; planet-injected data), or H_0 (signal absent; planet-free data). This approach allowed us to characterize the performance of the instrument by probing the range of possible TPFs and FPFs for various positions and signals.

We may also consider a different objective: deciding whether a particular bright spot in our final science image is a planet. In this scenario, we must decide which hypothesis, H_1 or H_0, applies to our location of interest. While hypothesis testing is beyond the scope of this paper, we refer to the detailed discussions in Kasdin & Braems (2006), Section 5, and Young et al. (2013).

7. CONCLUSION

As the cost and complexity of ground and space-based exoplanet imaging missions increase, so too must the fidelity and relevance of our diagnostic tools improve. We argue that the drawbacks of the contrast curve – its lack of transparency, flexibility, and connection to the data – motivate a re-evaluation of its use as a general purpose performance metric. Our proposed "performance map" is one among many possible methods for visualizing the true and false positive fractions associated with a high dynamic range image. The performance map is an opportunity for displaying the results of planet search programs in a consistent and statistically correct way as well as comparing the performance of various post-processing algorithms within a well-defined statistical framework. By encouraging the scrutiny of this new metric, we hope to improve the prediction and evaluation of the performance of the next generation of high contrast imaging instruments.

This work has been supported by the Keck Institute for Space Studies and the NASA Exoplanet Exploration Program Study Analysis Group #19: Exoplanet Imaging Signal Detection Theory and Rigorous Contrast Metrics. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant No. DGE-1144469. G. Ruane is supported by an NSF Astronomy and Astrophysics Postdoctoral Fellowship under award AST-1602444. C.G.G. and O.A. ackowledge funding from the European Research Council Under the European Union's Seventh Framework Program (ERC Grant Agreement n. 337569) and from the French Community of Belgium through an ARC grant for Concerted Research Action.

Software: Vortex Image Processing (Gomez Gonzalez et al. 2017)

REFERENCES

Aime, C., & Soummer, R. 2004, The Astrophysical Journal Letters, 612, L85

Amara, A., & Quanz, S. P. 2012, Monthly Notices of the Royal Astronomical Society, 427, 948

Apai, D., Kasper, M., Skemer, A., et al. 2016, The Astrophysical Journal, 820, 40

Beuzit, J.-L., Feldt, M., Dohlen, K., et al. 2008, in , 701418–701418–12

Bloemhof, E. E. 2004, Optics Letters, 29, 159

Dohlen, K., Langlois, M., Saisse, M., et al. 2008, in , 70143L–70143L–10

Fitzgerald, M. P., & Graham, J. R. 2006, The Astrophysical Journal, 637, 541

Gomez Gonzalez, C. A., Wertz, O., Absil, O., et al. 2017, AJ, 154, 7

Guyon, O., Pluzhnik, E. A., Kuchner, M. J., Collins, B., & Ridgway, S. T. 2006, The Astrophysical Journal Supplement Series, 167, 81

Hinkley, S., Oppenheimer, B. R., Soummer, R., et al. 2007, The Astrophysical Journal, 654, 633

Kasdin, N. J., & Braems, I. 2006, The Astrophysical Journal, 646, 1260

Krist, J., Nemati, B., & Mennesson, B. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 2, 011003

Krzanowski, W. J., & Hand, D. J. 2009, ROC Curves for Continuous Data (CRC Press), google-Books-ID: UZHwdiwOs4QC

Lafrenière, D., Marois, C., Doyon, R., Nadeau, D., & Artigau, E. 2007, The Astrophysical Journal, 660, 770

Liu, M. C. 2004, Science, 305, 1442

Marois, C., Doyon, R., Racine, R., & Nadeau, D. 2000, Publications of the Astronomical Society of the Pacific, 112, 91

Marois, C., Lafrenière, D., Doyon, R., Macintosh, B., & Nadeau, D. 2006, The Astrophysical Journal, 641, 556

Marois, C., Lafrenire, D., Macintosh, B., & Doyon, R. 2008a, The Astrophysical Journal, 673, 647

Marois, C., Macintosh, B., Barman, T., et al. 2008b, Science, 322, 1348

Mawet, D., Milli, J., Wahhaj, Z., et al. 2014, The Astrophysical Journal, 792, 97

Otten, G. P. P. L., Snik, F., Kenworthy, M. A., et al. 2017, The Astrophysical Journal, 834, 175

Perrin, M. D., Sivaramakrishnan, A., Makidon, R. B., Oppenheimer, B. R., & Graham, J. R. 2003, The Astrophysical Journal, 596, 702

Pueyo, L. 2016, The Astrophysical Journal, 824, 117

Ruane, G., Mawet, D., Kastner, J., et al. 2017, arXiv:1706.07489

Ruffio, J.-B., Macintosh, B., Wang, J. J., et al. 2017, ApJ, 842, 14

Shapiro, S. S., & Wilk, M. B. 1965, Biometrika, 52, 591

Shi, F., Balasubramanian, K., Hein, R., et al. 2016, Journal of Astronomical Telescopes, Instruments, and Systems, 2, 011021

Soummer, R., Ferrari, A., Aime, C., & Jolissaint, L. 2007, The Astrophysical Journal, 669, 642

Soummer, R., Pueyo, L., & Larkin, J. 2012, The Astrophysical Journal Letters, 755, L28

Sparks, W. B., & Ford, H. C. 2002, The Astrophysical Journal, 578, 543

Student. 1908, Biometrika, 6, 1

Swets, J., Tanner, W. P., & Birdsall, T. G. 1961, Psychological Review, 68, 301

Tanner, W. P., & Swets, J. A. 1954, Psychological Review, 61, 401

Uyama, T., Hashimoto, J., Kuzuhara, M., et al. 2016, arXiv:1604.04697 [astro-ph], arXiv: 1604.04697

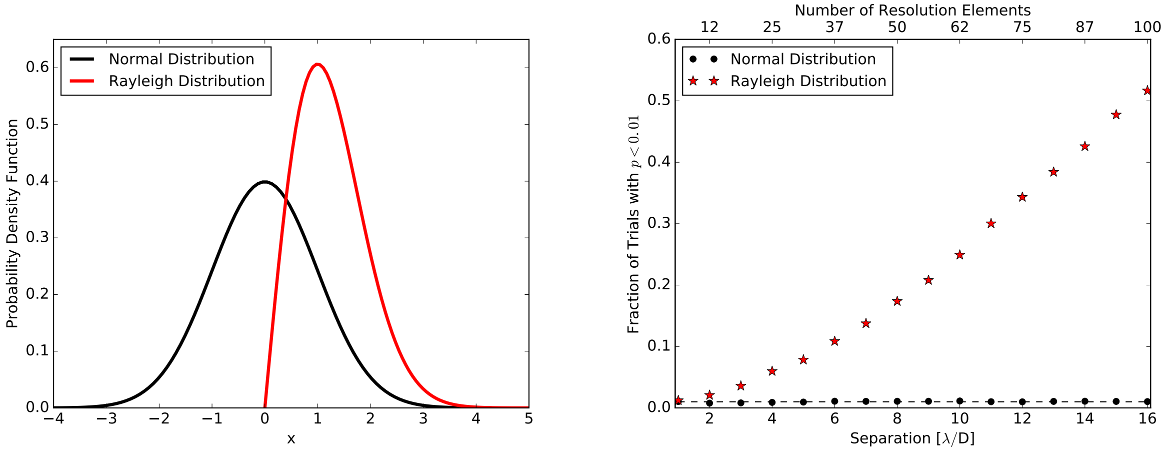
Wahhaj, Z., Liu, M. C., Biller, B. A., et al. 2013, The Astrophysical Journal, 779, 80

Wertz, O., Absil, O., Gomez Gonzalez, C. A., et al. 2017, Astronomy & Astrophysics, 598, A83

Wilcox, R. R. 2009, Basic Statistics: Understanding Conventional Methods and Modern Insights (Oxford University Press), google-Books-ID: gbr8JU41pncC

Young, E. J., Kasdin, N. J., & Carlotti, A. 2013, in , 88640S–88640S–13

Zurlo, A., Vigan, A., Galicher, R., et al. 2016, Astronomy & Astrophysics, 587, A57



(a) The PDF of a normal distribution (black line) and Rayleigh distribution (red line) with a scale parameter of two.

(b) The fraction of all trials that reject the null hypothesis ($p \leq 0.01$) for the normally distributed data (black circles) and Rayleigh distributed data (red stars). The black dotted line indicates the expected fraction of trails for which the normally distributed data is expected to reject the null hypothesis ($p = 0.01$).

Figure 8. Even though the Rayleigh distribution (scale parameter = 2) is highly skewed compared with the normal distribution, the Shapiro-Wilk test cannot reliably distinguish it from a normal distribution for the sample sizes shown here. For separations less than 15 λ/D, the Shapiro-Wilk test gives the wrong outcome (fails to reject the null hypothesis) for more than half of all trials.

APPENDIX

A. THE SHAPIRO-WILK TEST

For a given post-processed dataset, we may be interested in testing whether our data has been successfully whitened at small separations. The Shapiro-Wilk test (Shapiro & Wilk 1965) tests the null hypothesis that a dataset was drawn from a normal distribution; it returns a $p-$value that specifies that probability of obtaining the dataset given the null hypothesis. In order to test the utility of the Shapiro-Wilk test at small separations, we consider data drawn from two different distributions: a normal distribution (Figure 8a, black line), and a positively skewed Rayleigh distribution (Figure 8a, red line). At face value, we expect to easily reject the Shapiro-Wilk test's null hypothesis when testing data drawn from the dramatically non-white Rayleigh distribution.

We first compute the Shapiro-Wilk test $p-$value using a normally distributed dataset with $2\pi r$ elements. We then draw new sets of $2\pi r$ elements to repeat the test 10^4 times, giving 10^4 $p-$values per separation. We arbitrarily choose $p \leq 0.01$ as our threshold for rejecting the null hypothesis. As expected, we find that for all separations, the normally distributed test data gives $p \leq 0.01$ a fraction 0.01 of the time (Figure 8b, black points).

Next, we repeat this procedure for the Rayleigh distributed data. We find that these data reject the null hypothesis for a much larger fraction of trials than the normally distributed data (Figure 8b, red points). However, we quickly see a problem: at 15λ/D, the Rayleigh distributed data only rejects the null hypothesis about half of the time. This means that in any one science image, the probability of erroneously accepting the null hypothesis that the data is normally distributed is also 50%. At smaller separations, we draw the wrong conclusion most of the time – hence, the Shapiro-Wilk test cannot be used to test for normality at small separations.

Some tests (e.g. the Kolmogorov-Smirnov test) perform better in these respects than the Shapiro-Wilk test, while others (e.g. the Anderson-Darling test) are similarly problematic. The purpose of the example given here is to demonstrate that the outcomes of normality tests cannot be taken at face value, and must be rigorously validated in order to be applied to observational datasets.