Enhanced detection limits in the SHINE F150 survey through the Regime Switching Model

Optimizing thresholds and investigating environmental noise

M. Sabalbal[1], O. Absil[1,⋆], C.-H. Dahlqvist[1], and P. Delorme[2]

[1] STAR Institute, Université de Liège, Allée du Six Août 19C, 4000, Liège, Belgium
e-mail: mariam.sabalbal@uliege.be
[2] IPAG, Univ Grenoble Alpes, CNRS, Grenoble, France

Received 18 April 2025 / Accepted 4 December 2025

ABSTRACT

Context. In high-contrast imaging, a novel detection algorithm for angular differential imaging (ADI) sequences has recently been introduced: the Regime Switching Model (RSM). This advanced statistical tool enhances the distinction between planetary signals and bright speckles by simultaneously combining multiple ADI-based post-processing techniques.
Aims. In this study, we apply the RSM algorithm to analyze the F150 sample from the SHINE high-contrast imaging survey carried out with VLT/SPHERE, aiming to enhance detection limits and identify new exoplanet candidates. Additionally, we investigate how environmental conditions influence post-processed noise distributions and detection thresholds.
Methods. We generate detection maps and contrast curves for 213 observations in the F150 SHINE sample using the RSM algorithm. A clustering approach based on environmental parameters is used to group observations with similar noise characteristics. We propose two methods for defining radial detection thresholds in the RSM maps: fitting a log-normal distribution to the post-processed noise and maximizing the F1 score. We also assess the performance of various combinations of post-processing techniques within the RSM framework to identify optimal configurations.
Results. This study demonstrates the utility of clustering based on observational parameters, effectively distinguishing features like wind-driven halos and low-wind effects. Detection thresholds vary significantly across clusters, differing by up to a factor of 10, highlighting the importance of considering observational environments. Log-normal thresholds provide conservative, noise-aware limits, while F1 score-based thresholds offer observation-specific results, both showing compatibility overall. RSM improves detection limits by an average factor of two at 1″ and five at inner working angles compared to standard PCA processing. This study reports more than 30 newly detected signals, including one promising candidate awaiting second-epoch confirmation.

Key words. Planets and satellites: detection – Atmospheric effects – Methods: data analysis – surveys – Techniques: image processing

1. Introduction

The field of high-contrast imaging (HCI) has undergone remarkable advancements in recent years, driven by innovations in coronagraphy, adaptive optics, sophisticated observing strategies, and advanced post-processing techniques. These developments have made direct imaging a powerful tool for detecting hot, massive Jupiter-like exoplanets at wide orbital separations, capable of capturing planets that are up to 10^6 times dimmer than their host stars in the infrared. This capability has led to the detection and characterization of dozens of young exoplanets, offering invaluable insights into planetary formation processes (see Currie et al. 2023, for a recent review). Pushing detection limits further to identify fainter, closer-in planets would bridge the gap between direct imaging and indirect methods, providing richer information about planetary formation pathways, enabling detailed atmospheric characterization of a larger sample of exoplanets, and expanding our understanding of exoplanet demographics.

The Spectro-Polarimetric High-contrast Exoplanet Research (SPHERE, Beuzit et al. 2019) instrument is a second-generation extreme adaptive optics system at the Very Large Telescope (VLT), equipped with advanced coronagraphs feeding imaging and spectroscopic cameras. SPHERE has been central to several exoplanet imaging campaigns, including the SPHERE High-contrast Imaging survey for Exoplanets (SHINE, Chauvin et al. 2017a). SHINE targets around 400 young, nearby, and relatively bright stars to detect and characterize giant exoplanets, contributing to our understanding of their architectures and formation (see Desidera et al. 2021, for further details). The survey utilizes both of SPHERE's near-infrared imaging cameras – the dual band imager IRDIS (Dohlen et al. 2008) and the integral field spectrograph IFS (Claudi et al. 2010) –, employing angular differential imaging (ADI, Marois et al. 2006) for high-contrast observations. The SHINE survey has enabled key discoveries, such as HIP 65426b (Chauvin et al. 2017b) and PDS 70b (Müller et al. 2018), has refined constraints on giant planet occurrence rates (Langlois et al. 2021), demographics of young exoplanets within 300 AU (Vigan et al. 2021), and has spatially resolved several circumstellar disks (Lagrange et al. 2016; Feldt et al. 2017; de Boer et al. 2016). In their analysis of a 150-star subset from the SHINE survey, referred to as the F150 sample, Langlois et al. (2021) evaluated the survey's detection capabilities.

[⋆] F.R.S.-FNRS Research Director

Utilizing SPHERE's IRDIS H-band mode, the median detection performance achieved 5σ contrasts of 11.8 mag at 200 mas and 13.1 mag at 800 mas. These results were obtained by applying standard post-processing techniques, including principal component analysis (PCA, Soummer et al. 2012; Amara & Quanz 2012) and the template-based locally optimized combination of images (TLOCI, Marois et al. 2014).

In recent years, more advanced post-processing algorithms such as ANDROMEDA (Cantalloube et al. 2015), PACO (Flasseur et al. 2018, 2020, 2024), SODINN / NA-SODINN (Gomez Gonzalez et al. 2018; Cantero et al. 2023), RSM (Dahlqvist et al. 2020, 2021b,a), or 4S (Bonse et al. 2025) have been developed, with the goal of optimizing noise modeling and enhancing planetary signal detection – particularly for faint companions at close separations. Notably, Chomez et al. (2023) applied PACO to the SHINE F150 sample, improving the median contrast limits by a factor five at inner working angles and a factor two overall, and demonstrating the impact of more refined statistical modeling.

Building on this progress, we revisit the SHINE F150 sample using the RSM algorithm, which has shown excellent performance in the Exoplanet Imaging Data Challenge (EIDC, Cantalloube et al. 2020) by achieving high F1 scores and a low false positive rate. Unlike spatial-only modeling approaches, RSM captures both the spatial and temporal evolution of pixel intensities across residual, derotated frames using a Markov regime-switching framework applied to outputs from multiple PSF subtraction techniques. This enables it to fully exploit the temporal structure of high-contrast imaging sequences, offering improved sensitivity and robustness in exoplanet detection. In this study, we define detection thresholds tailored to RSM maps, accounting for noise variations across observing conditions. We compare two thresholding strategies, one assuming a log-normal noise distribution, and another based on maximizing the F1 score, and assess their relative performance. We further evaluate how different post-processing algorithm combinations affect RSM sensitivity and present updated detection limits for the SHINE survey based on this reanalysis.

The structure of this paper is as follows: Sect. 2 outlines the F150 sample selection criteria. Section 3 describes the observational conditions in the dataset. Section 4 explains the RSM algorithm and its application in this study. Section 5 examines two methods for setting detection thresholds and their atmospheric context. Section 6 presents contrast curves for different detection thresholds and post-processing combinations, highlighting the optimal approach within the RSM framework, and discusses the improvements in the SHINE detection limits. Section 7 presents the point sources detected by RSM in the F150 sample and compares them with those identified by other algorithms, such as PACO, including a discussion of newly detected candidates.

2. The data sample

In this study, we use the pre-reduced F150 sample from the SHINE survey, observed in H23 bands with ADI mode, covering an approximately 9″ × 9″ field of view. The High Contrast Data Center (HC-DC, Delorme et al. 2017; Galicher et al. 2018; Beuzit et al. 2019; Maire et al. 2016) pre-processed these observations, resulting in a dataset of 343 observations. Atmospheric conditions during these observations varied significantly, with seeing values ranging from 0.″4 to 3″ and Strehl ratios from 0.1 to 0.95. To ensure data quality and consistency, a pre-selection of datasets was necessary based on these atmospheric parameters.

The preliminary selection of data was based on several parameters, including seeing, Strehl ratio, raw contrast, number of frames, FWHM, and an assessment of the features in the PSF and the science cubes. Information like seeing, wind speed, coherence time were primarily derived from differential image motion monitor (DIMM, Sarazin & Roddier 1990) measurements, as recommended by Milli et al. (2017), while the Strehl ratio was empirically measured from the PSF. Additional parameters, such as raw contrasts for science cubes and frame quality assessments, were provided by the HC-DC reduction pipeline. In cases where DIMM seeing measurements were unavailable, they were inferred from the available seeing values provided by the SPHERE real-time computer SPARTA (Fedrigo et al. 2006) using the relationship between DIMM and SPARTA seeing established in Milli et al. (2017).

To ensure a sample of stars with observing conditions ranging from fair to excellent, observations that met any of the following conditions were also discarded: seeing larger than 2″, Strehl ratio less than 0.5, FWHM greater than 5.5 pixels, number of frames in the ADI sequence less than 40, and nonphysical raw contrasts at 500 mas (negative or greater than one). In case of multiple observations of the same star within five consecutive nights, only the observation with the best conditions was retained. Finally, we assessed the quality of both coronagraphic and non-coronagraphic cubes using a combination of tools: the HC-DC frame quality assessment, correlation functions from the VIP package (Gomez Gonzalez et al. 2017; Christiaens et al. 2023), and visual inspection. Cubes containing a majority of corrupted frames were discarded. In addition, science cubes in which the star was located outside the coronagraph mask, or where the off-axis PSF showed strong secondary lobes caused by the low-wind effect (see Milli et al. 2018), were also excluded.

These filtering steps resulted in a total of 213 observations for 150 stars. These constitute the data sample used in this paper. Additional information about the list of targets and their observing conditions is provided in Section 8.

3. Clustering based on environmental conditions

The quality of observations and specific features within coronagraphic images can be characterized by several environmental and instrumental parameters. Factors such as seeing, wind speed, coherence time, Strehl ratio, and raw contrast collectively indicate observation quality and provide insights into SPHERE-specific features. For instance, the wind-driven halo (Cantalloube et al. 2020) arises under high wind velocity and short coherence times, while the low-wind effect (LWE, Milli et al. 2018) appears at low wind speeds. These features can severely degrade SPHERE image quality, reducing the achievable contrast levels (see Cantalloube et al. 2019, for more details).

We adopted the clustering method outlined in Dahlqvist et al. (2022), grouping observations with similar characteristics using the k-means clustering algorithm from the `scikit-learn` package. This approach not only helps identify distinct classes of SPHERE observations but also reduces computational load when optimizing PSF subtraction techniques and RSM parameters (as described in Section 4). The parameters selected for clustering include the number of frames in the ADI sequence, seeing, wind speed, parallactic angle range, Strehl ratio, raw contrast at 500 mas, and coherence time. Observations were grouped into six clusters. The number of clusters was determined by maximizing the silhouette score, a metric used to evaluate the quality of clustering. For each cluster, k-means automatically selected a

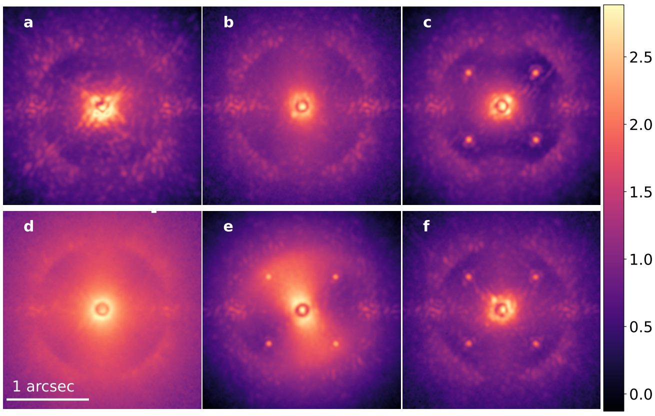


Fig. 1: Image representation of single frames from various observations across different clusters, labeled (a) to (f) corresponding to clusters 1 to 6 presented in log scale.

representative observation, i.e., the dataset closest to the cluster centroid. In the following, the terms cluster center and representative observation are used interchangeably. The parameters for PSF subtraction and the RSM algorithm can then be optimized for these representative data sets, and generalized to other observations within the same cluster. To ensure unbiased parameter selection, we made sure that none of the cluster centers contained datasets with bright astrophysical signals. A pixel correlation test on science cubes validates the effectiveness of clustering in capturing noise features (see Appendix A), and supports the use of the k-means elected centers for parameter optimization. Figure 1 displays a 200 × 200 pixel crop of representative individual frames from different clusters. These images illustrate the varying noise behavior across clusters, with some frames showing secondary lobes at close separations from the star indicative of the low-wind effect (Fig. 1a), while others exhibit strong elongated (aka butterfly) patterns characteristic of the wind driven halo (Fig. 1e).

To visualize the clustering effectiveness based on selected parameters, we applied Principal Component Analysis (PCA) for dimensionality reduction (see Appendix B), enabling clear separation of clusters in the latent space. For a more physically grounded interpretation of the nature of the various clusters, Fig. 2 presents their visualization based on key environmental parameters, revealing distinct traits for each cluster:

- cluster 1 (Fig. 1a, 11% of observations) is associated with low wind speed (Fig. 2b) and high coherence time (Fig. 2a), suggesting the presence of the low-wind effect;
- cluster 2 (Fig. 1b, 48% of observations) represents observations under generally fair to good conditions;
- cluster 3 (Fig. 1c, 8% of observations) is characterized by a large number of frames (Fig. 2d);
- cluster 4 (Fig. 1d, 3% of observations) is defined by low raw contrast, low coherence time, and high seeing (Figs. 2a, 2c);
- cluster 5 (Fig. 1e, 15% of observations) shows high wind speed, high seeing, and low coherence time (Figs. 2a, 2b), which makes it susceptible to strong wind driven halo;
- cluster 6 (Fig. 1f, 15% of observations) is associated with large parallactic angle ranges (Fig. 2d).

This successful clustering of data into meaningful categories will allow us to later analyze noise distributions and detection limits based on well-defined observational conditions, enabling a more targeted approach to interpreting results.

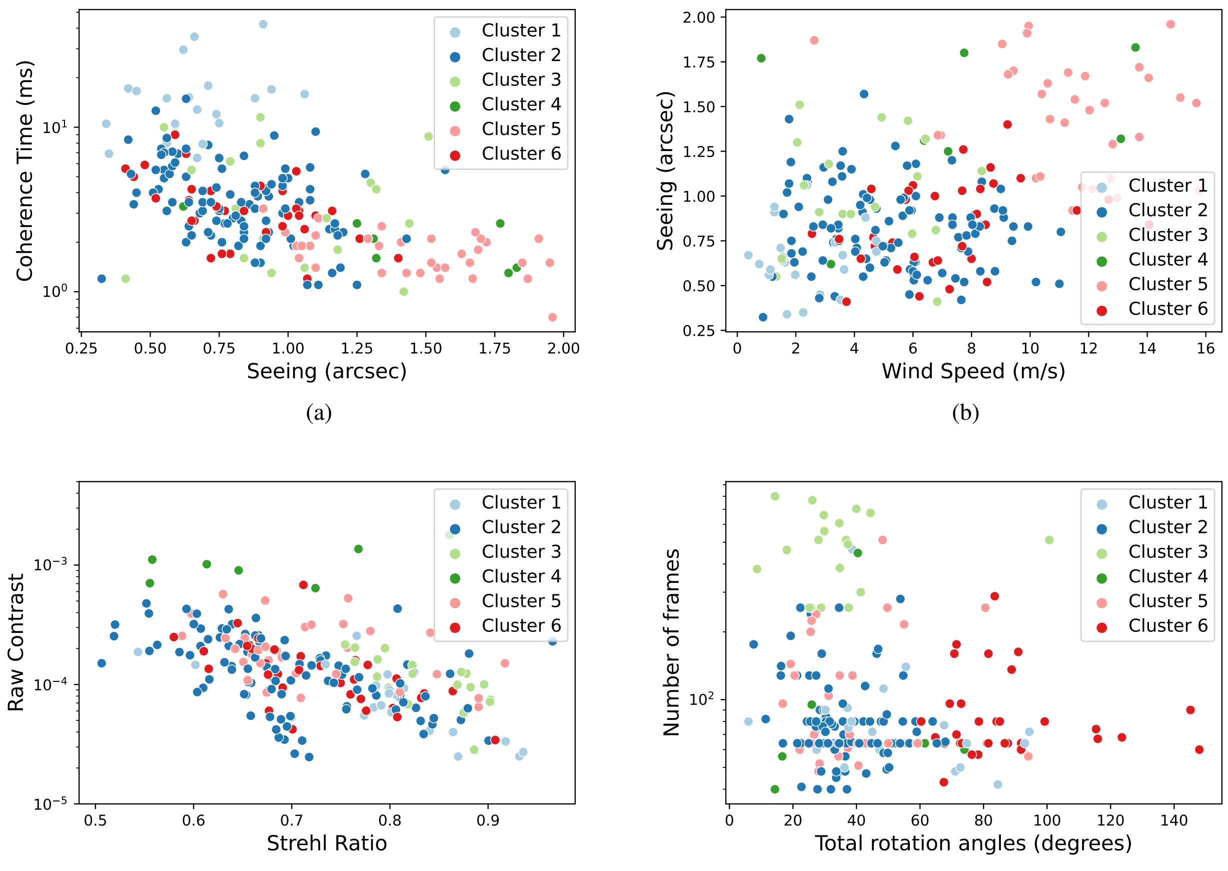
4. The RSM algorithm

The Regime Switching Model (RSM) is a well-established econometric algorithm that was successfully adapted to HCI by Dahlqvist et al. (2020) to enhance the detection of faint companions at small angular separations from their host stars. In the RSM algorithm, patches of pixels from derotated, residual frames after PSF subtraction are used to construct time series that capture spatial and temporal noise behavior at a given separation. When a planetary signal is present, these time series exhibit significant deviations from the noise patterns across the pixel patch and instead follow a planetary model that is both spatially and temporally structured. The algorithm models the data with two regimes: a noise regime, describing the mean and variance of noise at a given separation, and a planetary regime, incorporating the noise profile to which is added a model of the PSF whose intensity, β, is tuned. Each central pixel of the patch in a given annulus is assigned a likelihood of belonging to either regime. The RSM algorithm employs a Markov Chain model, where the probability of a pixel belonging to the planetary regime depends on the likelihood of the pixel-centered patch alignment with a planetary model, the probability of adjacent pixels, and a transition probability that connects these points. This integration of spatial and temporal constraints, combined with the transition probabilities, enables the algorithm to reliably distinguish speckles from true astrophysical signals. Additionally, the RSM algorithm utilizes outputs from various PSF subtraction techniques to construct a unified time series for each annulus. By drawing on the temporal noise diversity across different post-processing methods, this approach enhances the sensitivity to faint companions.

RSM has demonstrated significant advancements over current post-processing techniques, as highlighted in multiple studies (Dahlqvist et al. 2020, 2021b, 2022). Furthermore, RSM shows excellent performance compared to other post-processing algorithms, achieving the lowest false positive rate in the Exoplanet Imaging Data Challenge (EIDC, Cantalloube et al. 2020), underscoring its reliability in distinguishing planetary signals from noise.

One of the most sensitive aspects of the RSM algorithm involves selecting optimal parameters for both PSF subtraction techniques and RSM itself, including the planetary flux multiplicative factor δ (where planetary flux is expressed as a multiple of the standard deviation of pixel intensity at a given separation, $\beta = \delta\sigma$, see equation 1 in Dahlqvist et al. 2020), PSF model crop size, and the region used for noise estimation within an annulus. Dahlqvist et al. (2021a) introduced an automated optimization process within RSM to identify these optimal settings by maximizing contrast for PSF subtraction parameters and minimizing false positives for RSM parameters using the reversed parallactic angles, thereby enhancing the algorithm's reliability in exoplanet detection. In this study, we employ RSM with a simplified parameter set, as recommended by Dahlqvist et al. (2022). In the following, we refer to the value of each pixel in the final RSM map – previously defined as the RSM probability – as the RSM score, to avoid any potential ambiguity.

Our study incorporates several PSF subtraction techniques: annular principal component analysis (APCA, Absil et al. 2013), locally optimized combination of images (LOCI, Lafrenière et al. 2007), non-negative matrix factorization (NMF, Ren et al. 2018), as well a forward-model version of KLIP (FM-KLIP, Pueyo 2016) and LOCI (FM-LOCI, see more in Dahlqvist et al. 2021b). RSM scores are computed over a radial range of $0\farcs11$ (9 pixels) to $1\farcs1$ (90 pixels), except for forward-model tech-



(a)

(b)

(c)

(d)

Fig. 2: Projection of cluster distributions across parameters: seeing, coherence time, wind speed, raw contrast, Strehl ratio, total parallactic angles, and number of frames.

niques, which are restricted to a region around 0.′′35 (30 pixels) due to their computational cost and their tendency to produce results similar to APCA at larger separations. RSM scores are estimated using the forward-backward approach (as defined in Dahlqvist et al. 2021b), which integrates information from both past and future frames in the ADI time series. Parameters are optimized via Bayesian optimization, as described in Dahlqvist et al. (2021a).

To reduce computational load and mitigate potential issues with the RSM parameter optimization on certain pathological datasets, we propose to determine the optimal parameters for the cluster centers identified in Sect. 3 and to apply these settings consistently across all datasets within each cluster. Table C.1 in Appendix **??** details the specific parameters for each cluster. To assess the reliability of this approach, Appendix D compares the cluster center parameters with those computed individually for each dataset, providing insights into their alignment and the robustness of the clustering methodology. We exclude from the optimization process the selection of the optimal combination of PSF subtraction techniques within the auto-RSM framework, as delivered by the opti-combination method described in Dahlqvist et al. (2021a). Instead, we keep this parameter adjustable and explore how varying combinations of PSF subtraction techniques impact noise distributions and contrast curves across separations, as discussed in Sect. 6.

5. Noise behavior in RSM

In HCI, detection limits are commonly set at a 5σ level, corresponding to a false alarm probability (FAP) of 3×10^{-7} based on a Gaussian noise assumption in processed frames (Mawet et al. 2014). However, this approach may not accurately reflect the true noise characteristics in HCI datasets. Alternative approaches, based e.g. on Laplacian distributions (Pairet et al. 2019; Dahlqvist et al. 2020), have been proposed to address this limitation, especially at small separations. Moreover, methods like those in Jensen-Clem et al. (2018), Bonse et al. (2023), and Daglayan et al. (2024) avoid assuming any specific noise model, instead using adaptive techniques to derive thresholds directly from data, reducing the reliance on potentially mismatched distributions.

In the context of RSM, fitting a theoretical distribution to the noise in the final RSM maps (hereafter referred to as RSM noise) is challenging due to the nonlinear dependency of the final RSM score on the likelihood functions and transition probabilities, leading to the distortion of any pre-assumed distribution. Dahlqvist et al. (2021b) defines the detection threshold for the contrast curve as the brightest false positive (FP) in the final detection map. As noted in Appendix E, this threshold is broadly consistent with those defined in this section. However, it lacks robustness in the presence of astrophysical signals and is not designed to detect them. An alternative approach explored by

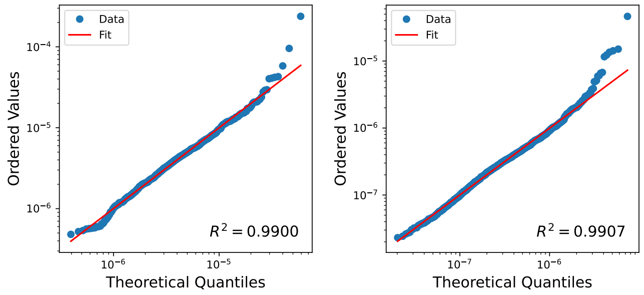


Fig. 3: Quantile-Quantile Plot of the RSM noise in Cluster 1 at 0″.25 (left) and 1″ (right) using APCA. The plot compares the RSM noise distribution to the Lognormal Distribution, excluding values above the 3×10^{-7} False Alarm Probability of the fit. The Coefficient of Determination (R^2) is also displayed, indicating a good fit.

Dahlqvist et al. (2022) , based on reversed parallactic angles, was also considered as a potential threshold applicable to both detection maps and contrast curves. However, it posed challenges by revealing a high number of false positives, ultimately limiting its utility as a reliable detection criterion.

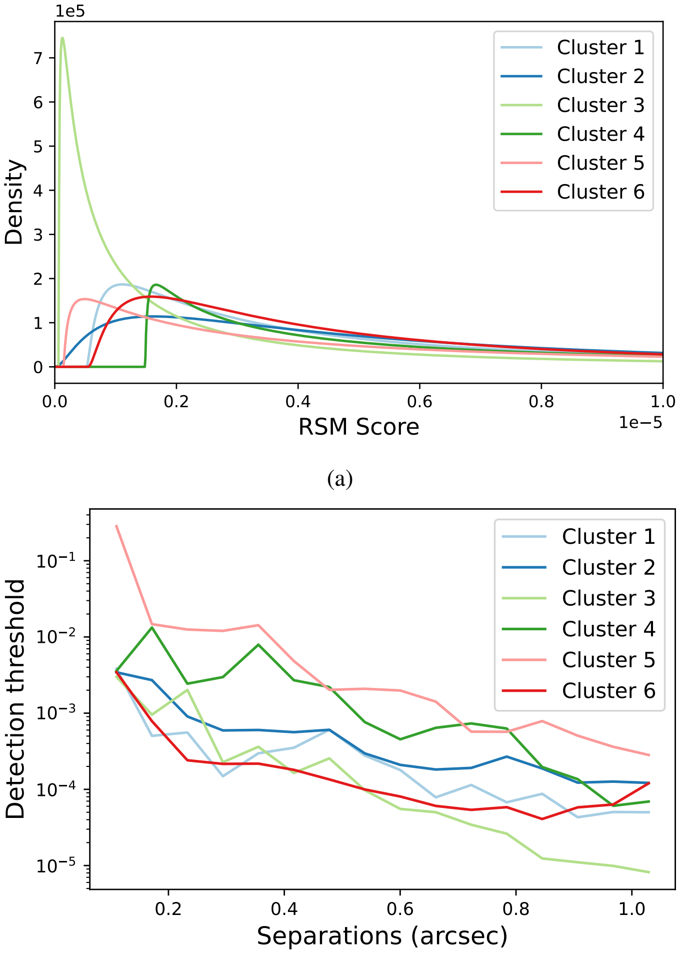
In this study, we investigate a new, consistent threshold applicable both for generating contrast curves and identifying candidate companions in RSM. Here, we examine two methods for defining a detection threshold, either (i) by fitting an empirical distribution to independent RSM noise realizations in the final RSM map (Sect. 5.1), or (ii) by balancing false positives and true positives while maximizing the F1 score, without assuming a specific RSM noise distribution (Sect. 5.2). We also explore how varying observing conditions influence the RSM noise distribution and the resulting threshold. Finally, we evaluate the strengths and limitations of each threshold within the context of RSM.

5.1. Fitting a log-normal distribution to the RSM noise

In this method, we take advantage of the large amount of data in the SHINE F150 survey to build distribution samples at each angular separation. Following Bonse et al. (2023), we treat independent pixel realizations, i.e., the central pixels of non-overlapping resolution elements, as distribution samples. Noise analysis at a given separation across the entire survey shows multiple RSM noise distributions, leading us to analyze realizations per cluster and per combination of PSF subtraction techniques in RSM, at each considered angular separation.

The final RSM maps contain only positive values, leading to a noise distribution that is not centered on zero and exhibits positive skewness. After testing multiple distributions as shown in Appendix F, we selected the log-normal distribution for its minimal parameter requirements and its consistency with the shape of the RSM noise distribution across separations and clusters. To define a detection threshold, we set a 3×10^{-7} FAP for the log-normal distribution, aligning with the conventional 5σ threshold in Gaussian statistics. To prevent bias when fitting the distribution, we exclude scores above 0.1 from the distribution samples, as these correspond to bright signals.

Figure 3 presents quantile-quantile (Q-Q) plots for cluster 1 at small (left) and large (right) separations using RSM with APCA. We compare the logarithm of independent sample values from all observations in cluster 1 at 0″.25 and 1″ with the corresponding fitted distribution values, considering only those below the 3×10^{-7} threshold. The Q-Q plots exhibit an approximately linear trend in both cases, indicating a strong fit, even though the extremities of the distribution show some outliers as it is often the case in Q-Q plots. Moreover, the coefficient of determination (R^2), representing the Pearson correlation between quantiles, is close to 1, further confirming the goodness of fit.



(a)

(b)

Fig. 4: *Top*. Log-normal distributions obtained by fitting RSM noise histograms at 0″.25 for various observing conditions (clusters) using RSM with APCA. *Bottom*. Detection thresholds derived from the lognormal distributions at 3×10^{-7} FAP, as a function of angular separation and of observing conditions.

In Figure 4, we show how varying observing conditions impact the RSM noise distribution and, consequently, the chosen detection threshold. Figure 4a presents the log-normal fits for different clusters at 0″.25 using RSM with APCA. Cluster 3, featuring a large number of frames per target, shows a narrower distribution and noise tail, while cluster 4, with mostly poor-quality observations, shows a broader tail and wider distribution, indicating a higher detection threshold. This plot highlights the varying RSM noise behaviors across observations under different conditions. Figure 4b illustrates the variations of the 3×10^{-7} FAP threshold across clusters and separations, highlighting the advantage of setting separate thresholds per separation to accommodate distinct noise properties. It also emphasizes the benefit of adjusting thresholds for observations under varying conditions. This approach, consistent with predictions, results in lower thresholds for Cluster 3 and higher ones for Cluster 4 and

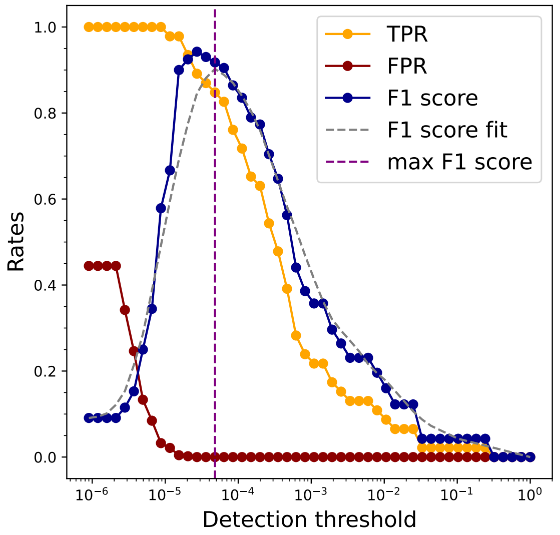


Fig. 5: Variation of the true positive rate (TPR), false positive rate (FPR), and F1 score for the cluster 1 center at $0\farcs7$. The purple dashed line marks the threshold that maximizes the fitted F1 score curve (dark blue curve), balancing high true positive (orange curve) recovery with minimal false positives (dark red curve).

5. This analysis underscores the importance of considering observing conditions when setting thresholds, as applying a single threshold across all datasets could lead to significant under- or over-estimations of the detection thresholds, as evidenced by the almost two orders of magnitude difference in thresholds between the best and worst clusters.

5.2. Threshold using maximum F1 score

In this section, we explore the possibility of defining a detection threshold without assuming a specific RSM noise distribution, an approach particularly relevant for samples containing a limited number of observations, where robust noise histograms cannot be built, especially for small angular separations. Following Daglayan et al. (2024), we define detection thresholds by empirically maximizing the F1 score at each angular separation through fake companion injections. The F1 score is defined as follows:

$$\text{F1 score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \qquad (1)$$

with TP the true positives, FP the false positives, and FN the false negatives. Maximizing the F1 score ensures an optimal balance between true positive rate and false positive rate.

To compute the false positives, all independent pixel realizations at each separation are considered as devoid of circumstellar signal (pure noise). Varying the threshold, we calculate the number of aperture values exceeding the threshold, which represent the number of false positives in the considered annulus. The false positive rate (FPR) is then defined as the ratio of these false positives to the total number of independent realizations. For true positives, we follow Mawet et al. (2014) by first estimating noise levels in the residual cube produced by APCA. We then inject synthetic point sources with fluxes between 2 and 3 σ of the measured noise. An injected source is considered recovered if its RSM score surpasses the considered detection threshold, and the number of such recoveries within the annulus defines the true positives. The optimal threshold is determined by maximizing the F1 score, ensuring a balance between high true positive recovery and minimal false positives.

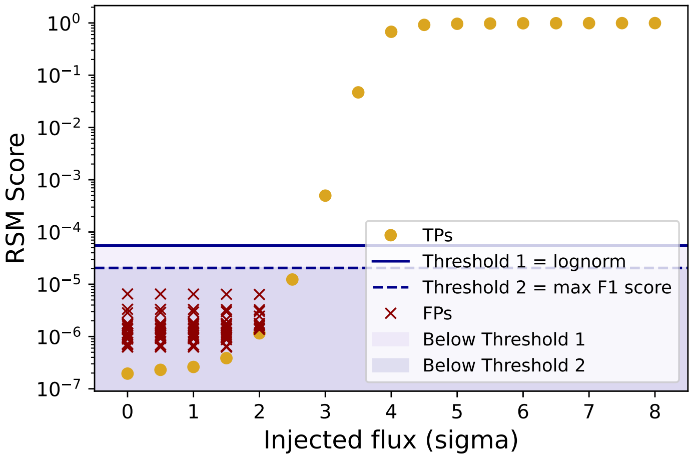


Fig. 6: Comparison of detection thresholds with false positives and injected companions for the cluster 1 center at an angular separation of $0\farcs6$, showing the RSM scores of injected fake companions (gold dots), false positives (red crosses), the maximum F1 score threshold (dashed blue), and the lognormal threshold (solid blue).

Figure 5 shows the evolution of TPR, FPR, and the F1 score for varying thresholds, applied to the cluster 1 center (corresponding to the BD-15 705 dataset) using RSM with APCA at $0\farcs7$ separation. The figure highlights the fact that the optimal F1 score favors minimal FPR and high TPR, thereby effectively distinguishing noise from potential true positives. This metric is highly dataset-specific as it relies heavily on the false positive rate and the true positive rate, the latter also depending on the flux of injected signals, which is influenced by the noise distribution. Consequently, this approach captures the unique characteristics of each dataset and aligns the threshold closely with the specific noise properties of the data. This technique has however several limitations. First, while it minimizes the false positive rate, it does not explicitly control the acceptable number of false positives per annulus. Additionally, it assumes all independent realizations are false positives, which may not hold if a real signal is present. In such cases, the maximum F1 score metric tends to exclude the true signal by minimizing false positives. One solution is to remove bright signals by injecting negative fake companions to avoid bias. Finally, this approach requires noise computation, fake companion injections, and a complete RSM processing for each separation, making it computationally expensive.

5.3. Comparison and reliability of the detection thresholds

To evaluate the reliability of detection thresholds, we examine their proximity to false positives for the cluster center of cluster 1. Figure 6 compares the RSM scores of fake companions injected at an angular separation of $0\farcs6$, across flux levels ranging from 0 to 8σ, with the false positives detected at that separation and the two associated detection thresholds. For each injec-

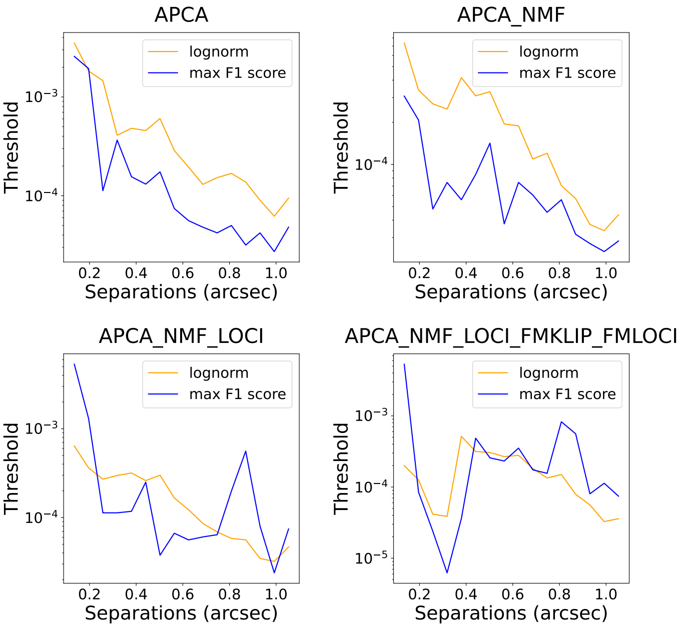


Fig. 7: Median detection thresholds for the six different clusters using the 3×10^{-7} FAP under lognormal distribution (orange), compared with the threshold derived from the maximum F1 score on the cluster centers (blue), for different combinations of PSF subtraction techniques as described in the plot titles.

tion, we computed the number of independent apertures within the considered annulus whose mean values exceeded the recovered flux, treating them as false positives. Both the maximum F1 score threshold and the log-normal threshold lie above the false positives, with the F1-score threshold closely aligning with the 2.5σ injected fake companion, near the noise-dominated region. The log-normal threshold, slightly higher at approximately 3σ, indicates a more conservative detection criterion. Figure 7 shows the median detection thresholds across all clusters, with each log-normal threshold computed individually for each cluster considering all its members, and the maximum F1 score thresholds computed for the cluster center datasets. The dip in thresholds below 0.′′4 in the bottom-right panel of Fig. 7 corresponds to the influence of forward modeling techniques (FM KLIP and FM LOCI), which are active within this separation range. This comparison highlights the differences between the two methods, with the log-normal threshold being generally more conservative. The larger variability in the detection threshold derived from the maximum F1 score across separations reflects its sensitivity to outliers in different annuli.

While it arguably provides the most relevant thresholds for a given data set, the maximum F1 score is computationally demanding to apply on each dataset of the survey, and too specific to generalize across an entire cluster (see more about that in Appendix G). Conversely, the detection thresholds derived from a log-normal distribution are more conservative, computationally efficient, and reflect the noise behavior under similar observing conditions. However, they rely on a predefined distribution, large sample sizes, and do not capture dataset-specific nuances. For the present study, focusing on a large survey, we adopt the threshold derived from the 3×10^{-7} FAP of the log-normal distribution, and derive it at each separation, for each combination of PSF subtraction technique, and for each cluster. We recommend using the F1 score threshold for smaller samples, while noting its limitations described in Sect. 5.2.

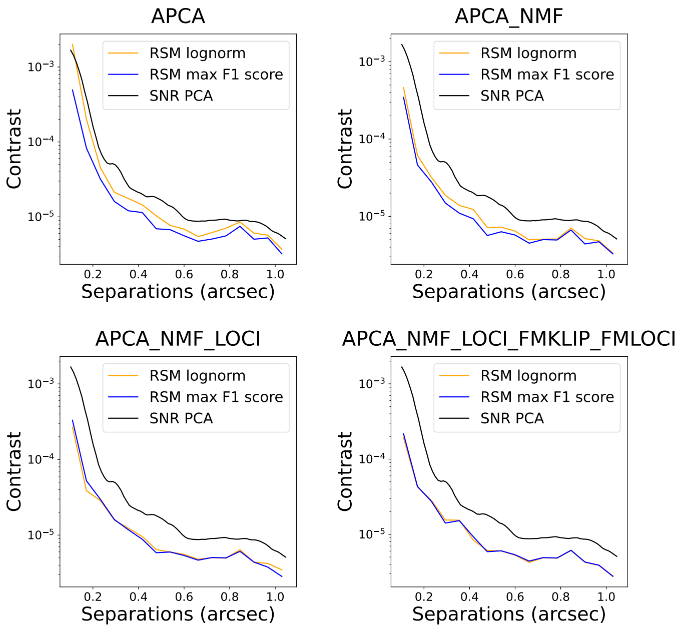


Fig. 8: Contrast comparison at 50% completeness using thresholds based on the maximum F1 score (blue), the log-normal RSM noise assumption (orange), and the 5σ APCA contrast curve (black) across different PSF subtraction techniques applied to the center of cluster 1.

6. Sensitivity limits

In this study, we adopt the 50% completeness curve described in Dahlqvist et al. (2021b), also known as contrast curves, to assess the sensitivity limits to point-like sources. Such curves indicate the flux at which 50% of injected signals are detected above the selected threshold, here the 3×10^{-7} FAP from the log-normal distribution, which corresponds to the 5σ contrast curve when assuming Gaussian noise. Here, we analyze the effects of threshold selection and PSF subtraction technique combinations on contrast curves, define the optimal RSM contrast curve, and evaluate its reliability.

6.1. Influence of threshold and PSF subtraction techniques

To evaluate how the chosen threshold impacts contrast curves, Fig. 8 presents a comparison for the center of cluster 1. It includes contrast curves derived using the maximum F1 score threshold (blue) and the lognormal threshold (orange) across four combinations of PSF subtraction techniques: APCA, APCA-NMF, APCA-NMF-LOCI, and APCA-NMF-LOCI-FMKLIP-FMLOCI. A standard 5σ contrast curve based on full-frame PCA with five principal components is shown in black for comparison. The RSM contrast curves show a good level of consistency between the two ways to define the detection threshold. This is especially the case when multi-technique combinations are used (see bottom panels in Fig. 8), while the log-normal threshold leads to slightly worse sensitivity than the F1-score one when RSM is performed with APCA only (Fig. 8 top left) due to the higher threshold in that case. Consistency was also found between the log-normal contrast curves derived here and the contrast curves based on the appearance of a first false positive proposed in Dahlqvist et al. (2022) for the SHARDDS survey (see Appendix E).

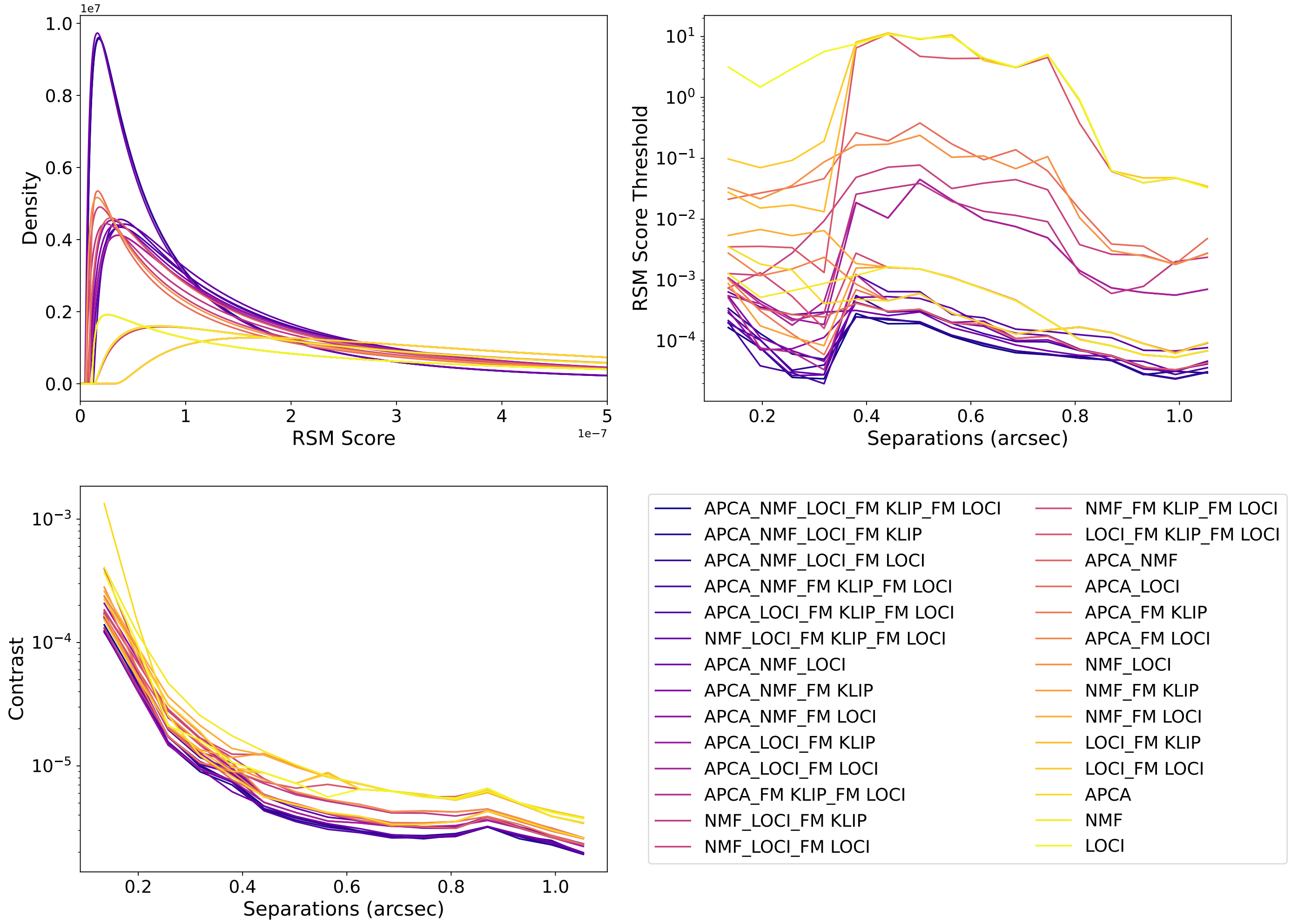


Fig. 9: Effect of different combinations of PSF subtraction techniques in RSM on the fitted lognormal distributions, detection thresholds, and contrast curves. *Top left* shows the difference of lognormal fit across different PSF subtraction techniques with colors ranging from dark violet for multi combinations to yellow for singular combinations. *Top right* shows the 3×10^{-7} FAP under lognormal distribution for the different combinations across all separations. *Bottom left* shows the contrast curve delivered by the different combinations with 50% completeness.

The choice of which PSF subtraction techniques to combine is an important aspect of the RSM algorithm, which is not straightforward to optimize. Instead of relying on a single combination identified as optimal by the RSM framework, we propose to systematically evaluate the performance of various combinations of PSF subtraction techniques. We analyze the differences in noise distributions in the final frames, the corresponding thresholds, and the contrast curves achievable with each combination. Figure 9 examines these effects across the 28 possible combinations of PSF subtraction techniques, ranging from single PSF subtraction techniques (in yellow) to multi-combinations (in dark violet). Figure 9 (top left) displays the median lognormal fits for each combination, clearly showing narrower noise distributions for multi-combinations and broader distributions for single techniques, indicating improved performance with combined techniques. Figure 9 (top right) presents the thresholds derived from these fits across all combinations and separations. Noise thresholds, on average, range from approximately 10^{-4} for multi-combinations to nearly 10 for LOCI – exceeding RSM's maximum score. This renders LOCI non-physical and unreliable when used alone, a consequence of its high false positive rate in RSM for many datasets. Figure 9 (bottom left) illustrates the contrast curves for the various combinations, showing an improvement by a factor of 3 for multi-combinations. Contrast values with lognormal thresholds exceeding 0.1 in RSM were excluded from the bottom-left panel due to difficulties in achieving convergence. This further demonstrates that incorporating multiple PSF subtraction techniques in RSM significantly enhances detection limits. Moreover, it highlights RSM's ability to leverage the diverse noise patterns of different PSF subtraction methods, effectively reducing the misclassification of transient speckles as planetary signals. RSM's efficiency in integrating additional information – either by combining different PSF subtraction techniques or by increasing the number of frames – is evident through narrower noise distributions, improved thresholds, and enhanced detection limits (see also Figure 4a).

6.2. Optimal usage of RSM

Although combining multiple techniques yields the best RSM performance, the optimal contrast at each separation does not always result from the same combination. Figure 10 presents a histogram illustrating the contributions of various RSM combinations to the best contrast values across all separations, based

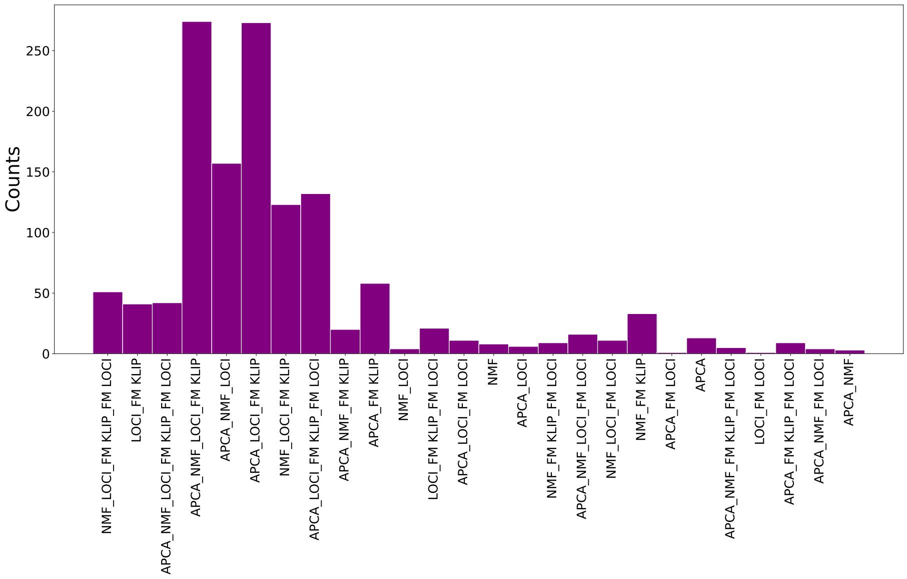


Fig. 10. Histogram of RSM combination contributions to the best contrast values across all separations, based on all datasets

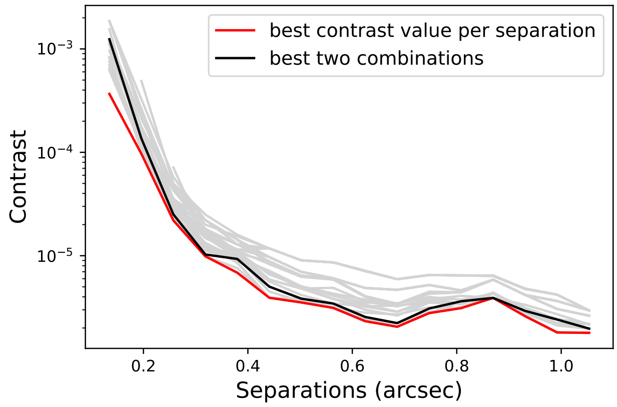


Fig. 11: Optimal contrast curve achieved by different combinations of PSF subtraction techniques in RSM. Gray curves show the median contrast performance for each combination, computed across all datasets in cluster 1. The red curve highlights the best achievable performance at each separation, given by the median of the minimum contrast values across all combinations. The black curve shows the median of the maximum contrast values delivered by the two best-performing combinations: APCA-NMF-LOCI-FMKLIP and APCA-LOCI-FMKLIP.

on all datasets. We identify the best contrast at each separation, shown as the red curve in Fig. 11. Figure 10 indicates that approximately 26 combinations out of the 28 possible combinations of PSF subtraction techniques contribute to the red curve of Fig. 11, with two combinations standing out due to their significantly higher contributions compared to others: APCA-NMF-LOCI-FMKLIP, and APCA-LOCI-FMKLIP. Since tracking candidates across 26 different RSM maps is complex, we opted to use only the two combinations with the highest contributions. Contrast curves were then computed using the maximum (= worst) contrast value from these two combinations at each separation, and point source candidates were identified if they appeared in both of the combinations above their respective thresholds. This contrast resulting from the aggregation of the best two combinations is represented by the black curve in Fig. 11. We recommend RSM users adopt the median contrast of these combinations to derive reliable contrast curves and use the corresponding maps to identify candidates. This metric is employed to determine detection limits in Sect. 6.4.

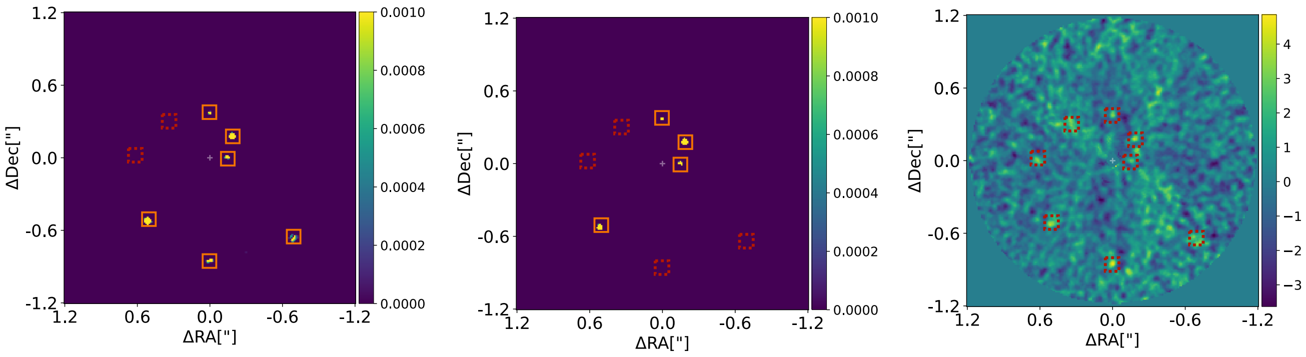
6.3. Reliability of the contrast curves

Evaluating the adopted contrast curve metric through fake companion injections helps verify the consistency of its core assumptions, including the applied noise threshold, the definition of the median contrast curve, and the expected 50% recovery rate of injected signals. We inject fake point sources at the predicted sensitivity limit using the VIP package on a cluster center and examine their recovery across different RSM maps. Figure 12 compares the recovery of injected signals using RSM maps against the signal-to-noise ratio (S/N) map produced by PCA with five principal components, for the center of cluster 1. The flux levels of the injected companions have been chosen to be just above the expected RSM sensitivity limits as illustrated in Fig. 13, which also compares the detection limits of RSM with the full-frame PCA 5σ contrast curve. The injections are performed at multiple separations, with one injection at each separation and varying position angles.

As expected, a significant fraction of the injections are successfully detected above the threshold in RSM maps, achieving approximately 50% recovery consistent with the definition of the contrast curve. Conversely, the same injections all appear well below the 5σ level in the S/N PCA map. This highlights RSM's ability to deliver deeper detection limits and reliable contrast curves, surpassing conventional post-processing methods in HCI.

6.4. Improvement to the SHINE survey

To conclude this analysis of sensitivity limits, we present the detection limits for the F150 SHINE sample derived from 213 observations using the detection threshold and modifications outlined in this study. Figure 14 shows all RSM contrast curves (shaded blue) against full-frame PCA ADI curves (shaded pink). For consistency within the survey, we do not use a high-pass filter at large separations, although we acknowledge that it could deliver better results in these cases. The PCA-ADI curves may



(a)

(b)

(c)

Fig. 12: Comparison of recovered fake point sources in cluster center 1 using RSM maps with different PSF subtraction technique combinations: (a) APCA-NMF-LOCI-FMKLIP and (b) APCA-LOCI-FMKLIP. These are compared to the signal-to-noise ratio PCA map with five principal components (c). The RSM maps shown in this figure are thresholded (RSM map minus the separation-dependent threshold). Injected signals recovered above the detection threshold are marked with yellow squares, while non-recovered signals appear in red squares.

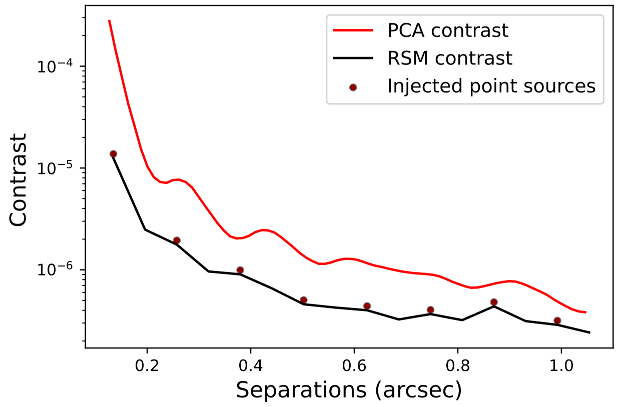


Fig. 13: Injected fake point sources compared to the optimal RSM contrast curve (black) and PCA 5-component contrast curve (red) for cluster center 1

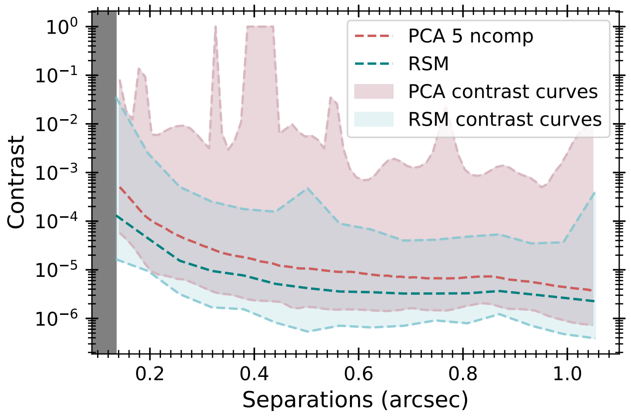


Fig. 14: Comparison of RSM and full-frame PCA contrast curves for the F150 SHINE sample: A visualization of all RSM contrast curves (shaded blue) and PCA contrast curves (shaded pink), including their median values, highlighting RSM's significant improvement in detection limits at both small and large separations.

show slight differences from those published in Langlois et al. (2021), reflecting variations in the data reduction process. Nevertheless, our PCA-ADI results align well with those reported in Chomez et al. (2023), as both are based on the same reduction pipeline. The PCA contrast curves display outliers due to bright companions and background star signals. The median contrast curves indicate RSM's substantial improvement, achieving a factor 2 enhancement at 800 mas and a factor 4 to 5 at 135 mas compared to the five-component PCA results.

This improvement is further highlighted in Fig. 15, where the magnitude difference between RSM and full-frame PCA sensitivity limits shows a median improvement of 1.8 mag at 0.″135, emphasizing RSM's superior detection at small separations and its ability to distinguish planetary signals amidst stellar residual noise. The improvement closely aligns with that achieved using PACO (Chomez et al. 2023), with both methods delivering a factor ∼ 5 improvement at small separations and a factor ∼ 2 overall, compared to full-frame PCA. These results emphasize the consistent advancements in sensitivity limits across different post-processing techniques. Figure 15 shows that a few datasets exhibit slightly better performance with PCA than with RSM.

This occurs primarily for observations located near the boundaries of their respective clusters, where the adopted RSM parameters may deviate from their true local optima, or in cases where the generalized noise model across the cluster is more restrictive than the dataset's actual RSM noise properties. These exceptions are rare and do not affect the overall improvement trends observed across the survey, although they do highlight certain limitations of parameter and noise generalization for a small subset of datasets.

This study also allows us to evaluate RSM's performance under varying observational conditions. Figure 16 highlights the improvements across distinct clusters in the SHINE survey, demonstrating consistent enhancements across different datasets.

This underscores RSM's robustness in handling diverse observational scenarios and its ability to improve point source detection across the entire survey.

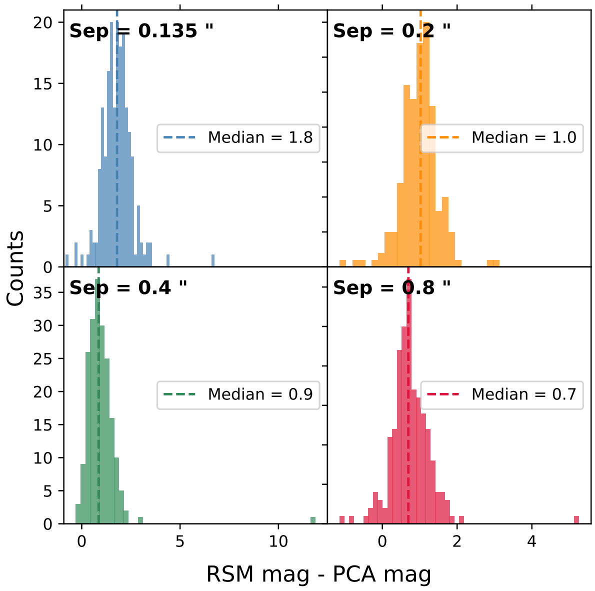


Fig. 15: Comparison of RSM and full-frame PCA improvements in magnitude at separations of 0.″135, 0.″2, 0.″4, and 0.″8. The median difference is shown as a dashed line, while histograms represent the full dataset.

7. Identification of point sources

The application of the RSM algorithm to the F150 SHINE sample resulted in the identification of 87 signals across 59 different observations, while 154 observations from the F150 sample showed no detectable signal within the 1.″1 field of view. As previously described, a signal is considered significant when it lies above the detection threshold in all four RSM detection maps, considering the two selected algorithm combinations and the parallel observations performed by SPHERE-IRDIS in the H2 and H3 filters. A dedicated tool was developed to cross-match detections between maps within half a FWHM, ensuring consistent identification of point sources across filter combinations. The photometric and astrometric measurements were derived following the procedure detailed in Dahlqvist et al. (2022). The performance of these measurements was assessed by (Cantalloube et al. 2024) during the Exoplanet Imaging Data Challenge (phase II); this analysis is not discussed further here, as it lies beyond the scope of the present work.

Among the detected signals, 49 were already recovered by Chomez et al. (2025), while 38 were not identified with PACO and required a dedicated analysis to distinguish genuine astrophysical sources from false positives. This counting of false positives does not include some artifacts that tend to appear in RSM maps in the presence of bright companions, at the same angular distance as those companions. This section presents a detailed comparison between the signals detected with RSM and PACO in Sect. 7.1, followed by an analysis of the new point sources recovered exclusively with RSM in Sect. 7.2 using proper motion and color-magnitude diagram tests.

7.1. Comparison with PACO on the F150 sample

Given the comparable performance of RSM and PACO, a direct comparison of their detected signals provides valuable insight into the relative sensitivity of both methods, the reliability of faint detections, and potential missed detections on either side.

To ensure a fair comparison between PACO and RSM, only the common observations in the archival point-source detections reported in Chomez et al. (2025) were considered. These correspond to common H2/H3-band observations covering fields of view between 0.″11 and 1.″1. Within this sample, PACO reported no detection in 181 observations. Among these, 29 revealed a detection in RSM, while 152 are classified as non-detections in both methods. These 29 observations contain 38 signal detections that will be examined in more detail in the next section. In addition to these 181 observations classified as non-detections, PACO recovered 52 signals in the remaining 32 observations, including 24 confirmed planets (observed at different epochs), 24 background stars, and 4 ambiguous detections. Here, we compare these 52 PACO detections with our RSM results.

Out of the 52 PACO detections, 49 were also recovered by RSM based on our four-map detection criterion. Figure 17 presents the contrast values of each PACO-detected signal, with colors indicating the recovery status in RSM; dark blue points represent signals fully recovered in four maps, while other colors correspond to non-recovered sources. As shown in the figure, most signals detected by PACO were also recovered by RSM, with only a few exceptions.

The non-recovered signals correspond to the confirmed planet HD 95086 b (in two observations), and a background star around HD 151726, although these sources were recovered in at least one of the RSM maps. The primary reason for this discrepancy lies in the different detection criteria employed by the two algorithms. The RSM method requires recovery above the detection threshold in both H2 and H3 filters, whereas PACO derives a single signal-to-noise ratio (S/N) that combines information from both spectral channels. Consequently, sources with marginal S/N in one band but a stronger signal in the other may be recovered by PACO but not by RSM.

This behavior is clearly illustrated by the case of HD 95086 b, observed on 2015-05-05 and 2015-05-11, with PACO S/N values of 6.8 and 5.0, respectively. Desgrange et al. (2022) showed that the planet/star contrast for this target is twice as large in H3 compared to H2 (see their Fig. 5), consistent with the Chomez et al. (2025) database, where the H2 contrast errors are nearly as large as the measured contrast values themselves.

7.2. Identification of new point sources

The application of the RSM algorithm to the F150 SHINE sample revealed 38 newly detected point sources across 29 different observations for 27 stars. Among these targets, 10 stars have additional epochs within the F150 sample, while the remaining 17 correspond to single-epoch observations.

Multi-epoch data enable the differentiation between false positives and genuine astrophysical sources through proper motion tests.

Out of the ten targets that have additional epochs, two exhibit detections in more than one epoch, while the remaining eight show detections in only a single epoch despite the availability of multiple observations.

For detections found in only one epoch among several available observations, both the expected position of the source within the field of view (FOV) and the observing conditions of

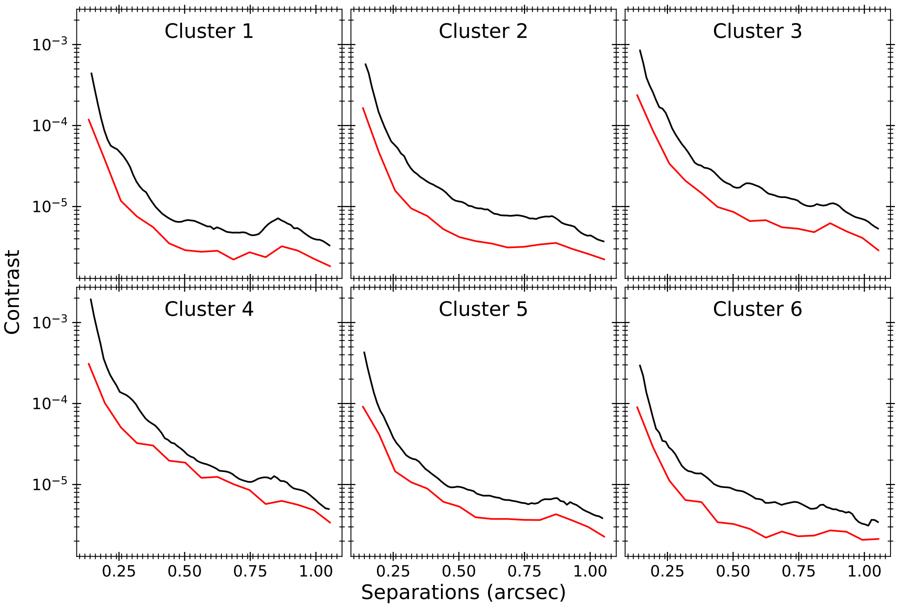


Fig. 16. Contrast performance improvements of RSM across various observing conditions in the SHINE survey clusters: median contrast curves are shown in red for RSM and black for PCA, demonstrating RSM's superior efficiency overall.

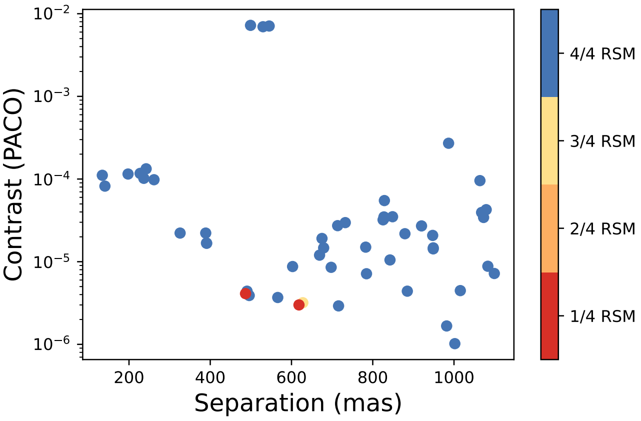


Fig. 17: Contrast distribution and separation of the 52 PACO-detected signals and their recovery by RSM. Dark blue points represent signals recovered in all four RSM maps, other colors correspond to sources not recovered by RSM, or recovered in only some RSM maps.

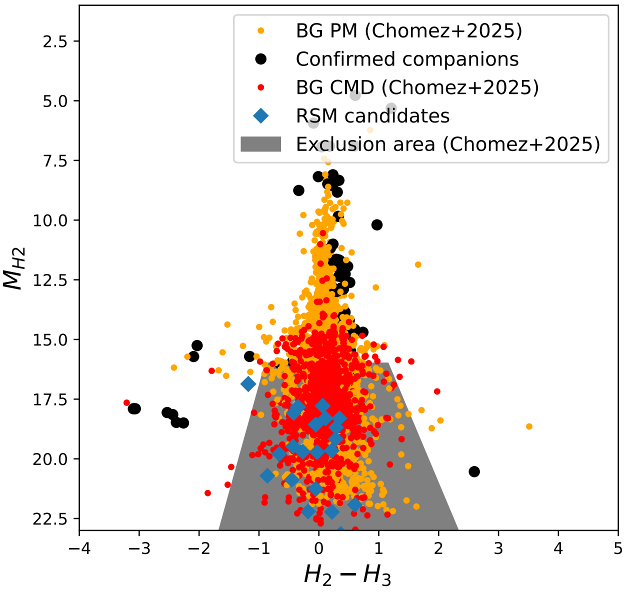


Fig. 18: Color–magnitude diagram showing the new RSM detections in the F150 SHINE sample, plotted alongside the signals reported in Chomez et al. (2025) for confirmed exoplanets, background stars identified through proper motion, and background stars classified via CMD analysis. The gray shaded area indicates the exclusion region for background stars defined by Chomez et al. (2025).

the non-detection epochs were examined. If the expected position in other epochs laid outside the FOV or if the observing conditions were less favorable, these signals were retained for further analysis; otherwise, they were classified as false positives. This resulted in one detection being retained for further analysis, while 12 signals were classified as false positives. These false positives are primarily associated with observations exhibiting low Strehl ratios and limited field rotation.

Regarding the multi-epoch detections, RSM successfully recovered PZ Tel in an observation from 2014, which was not detected by PACO, unlike other epochs where the signal was identified by both methods. RSM also recovered a background star around HD 164249 identified through the proper motion test between the 2015 and 2016 epochs. These signals were not reported by Chomez et al. (2025) but were previously mentioned by Langlois et al. (2021).

Verifying the nature of a signal for single-epoch targets is more challenging. In these cases, following Chomez et al. (2025), a color-magnitude diagram (CMD) analysis is employed to assess the likelihood of each detection being consistent with a planetary companion or a background star. The 17 targets that had no additional epochs within the SHINE F150 survey generated a total of 22 significant signals in our RSM maps. To these 22 signals, we add the ambiguous candidate identified in the previous paragraph, and perform a CMD analysis to evaluate the astrophysical relevance of each signal.

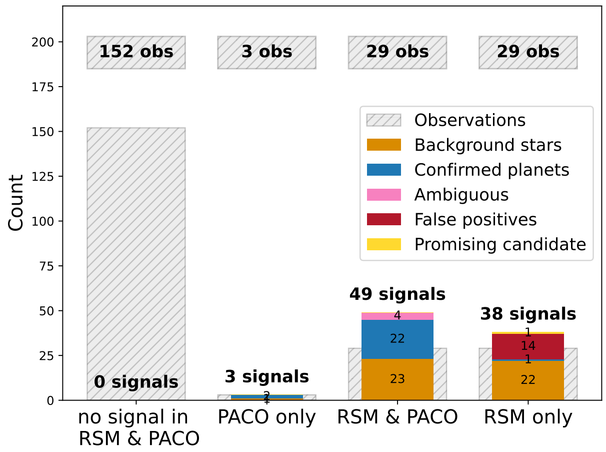


Fig. 19: Summary of the detection statistics for the F150 sample, comparing the outcomes of RSM and PACO across four cases: no signal in either method, signal in PACO only, signal recovered by both methods, and signal recovered by RSM only. The dashed gray bars indicate the number of observations in each category, with the corresponding counts labeled above them. The colored stacked bars represent the number of detected signals, with colors denoting their classification: confirmed planets, background stars, ambiguous sources, false positives, and the newly identified promising candidate.

The CMD is shown in Fig. 18 for all the detections reported in Chomez et al. (2025), supplemented by the new RSM candidates. As discussed by Chomez et al. (2025), background stars occupy a well-defined region in the diagram, and we adopt their background-star region, referred to as the exclusion area in Fig. 18.

Most of the RSM candidates fall within this exclusion region, indicating that they are likely background stars. Two signals, falling outside the plot range of Fig. 18, display photometric properties inconsistent with an astrophysical source and are therefore classified as false positives. One last source lies just outside the exclusion region, exhibiting photometric characteristics consistent with an exoplanet-like nature, and is thus considered a promising planet candidate. Appendix H presents several examples of newly identified background-star detections and compares their corresponding PCA signal-to-noise ratios with the RSM scores.

Figure 19 summarizes the detection statistics for the F150 sample, including the number of observations showing no signal, the signals recovered by RSM, by PACO, or by both methods, and the classification of each detected source. The figure also serves as a comprehensive visual summary of all numerical values discussed in the paper.

8. Conclusions

This study introduces new methods for determining detection thresholds in high-contrast imaging, utilizing the Regime Switching Model (RSM) applied to the SHINE F150 survey. Clustering based on simple environmental parameters using k-means effectively classified observations into meaningful groups, such as those with large number of frames, poor conditions, or specific effects like wind driven halo and low-wind effect. Spearman correlation tests confirmed the distinct noise characteristics in each cluster and validated the cluster representatives in terms of environmental and noise parameters. We have additionally highlighted how observational conditions influence RSM noise distribution in final RSM maps. By fitting a lognormal distribution at each separation, detection thresholds with a 3×10^{-7} FAP revealed notable variations across different clusters, emphasizing the importance of adapting thresholds to observational conditions. We also compared lognormal thresholds with those derived from maximizing the F1 score. While both approaches are compatible, the F1 score method better addresses individual bright false positives, making it suitable for single observations. Conversely, the lognormal fit reflects the overall noise behavior and provides a more generalizable solution for large surveys.

Leveraging the diverse noise behavior across frames in the science cubes and applying different PSF subtraction techniques further enhanced RSM's ability to distinguish planetary signals from speckles. This led to narrower RSM noise distributions, reduced thresholds, and improved detection limits, achieving a fivefold improvement at 135 mas and a twofold overall enhancement at larger separations compared to Langlois et al. (2021) for the F150 SHINE sample. Compared to Chomez et al. (2025), who used the PACO algorithm, applying RSM to the F150 sample achieved comparable performance. Our study led to the identification of 87 signals, 49 of which were previously reported by Chomez et al. (2025). Among the 38 new detections, 22 were classified as background stars, 14 as false positives, one turned out to be a known low-mass companion, and one was identified as a new exoplanet candidate for which follow-up observations are needed. This result highlights RSM's robustness and potential for further applications.

Finally, the methodology developed in this work can be directly extended to the full SHINE survey and adapted to other ground-based high-contrast imaging programs, such as the Gemini-GPIES (Nielsen et al. 2020), SPHERE-BEAST(Janson et al. 2019, 2021), or NACO-ISPY (Launhardt et al. 2020) surveys. This framework provides a robust and versatile foundation for statistically consistent planet detection and performance characterization across a wide range of observing conditions.

Data Availability

The RSM algorithm is publicly available on GitHub as a Python package[1]. The raw and reduced data used in this study can be requested through the High-Contrast Data Centre (HC-DC)[2]. The code developed for the analysis presented in this work is openly accessible on GitHub[3]. The full point-source candidate list will be presented in a forthcoming publication, together with an NA-SODINN analysis of the SHINE F150 sample (Cantero et al., in prep). The table for the 213 F150 SHINE observations used in this study, along with their observing conditions, is only available in electronic form at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via `http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/`.

Acknowledgements. This work has made use of the High Contrast Data Centre, jointly operated by OSUG/IPAG (Grenoble), PYTHEAS/LAM/CeSAM (Marseille), OCA/Lagrange (Nice), Observatoire de Paris/LESIA (Paris), and Observatoire de Lyon/CRAL, and supported by a grant from Labex OSUG@2020 (Investissements d'avenir – ANR10 LABX56). This project received funding from the European Research Council (ERC) under the European Union's Horizon

[1] `https://github.com/chdahlqvist/RSMmap`
[2] `https://hc-dc.cnrs.fr`
[3] `https://github.com/marisabalbal/RSM_f150SHINE`

2020 research and innovation program (grant agreement No 819155). Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and the Walloon Region.
We extend our gratitude to Faustine Cantalloube, Valentin Christiaens, and Sandrine Juillard for their insightful discussions.

References

Absil, O., Milli, J., Mawet, D., et al. 2013, A&A, 559, L12
Amara, A. & Quanz, S. P. 2012, MNRAS, 427, 948
Beuzit, J.-L., Vigan, A., Mouillet, D., et al. 2019, A&A, 631, A155
Bonse, M. J., Garvin, E. O., Gebhard, T. D., et al. 2023, AJ, 166, 71
Bonse, M. J., Gebhard, T. D., Dannert, F. A., et al. 2025, AJ, 169, 194
Cantalloube, F., Christiaens, V., Cantero, C., et al. 2024, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 13097, Adaptive Optics Systems IX, ed. K. J. Jackson, D. Schmidt, & E. Vernet, 1309713
Cantalloube, F., Dohlen, K., Milli, J., Brandner, W., & Vigan, A. 2019, The Messenger, 176, 25
Cantalloube, F., Farley, O. J. D., Milli, J., et al. 2020, A&A, 638, A98
Cantalloube, F., Gomez-Gonzalez, C., Absil, O., et al. 2020, in Adaptive Optics Systems VII, ed. L. Schreiber, D. Schmidt, & E. Vernet, Vol. 11448, International Society for Optics and Photonics (SPIE), 114485A
Cantalloube, F., Mouillet, D., Mugnier, L. M., et al. 2015, A&A, 582, A89
Cantero, C., Absil, O., Dahlqvist, C. H., & Van Droogenbroeck, M. 2023, A&A, 680, A86
Chauvin, G., Desidera, S., Lagrange, A. M., et al. 2017a, in SF2A-2017: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. C. Reylé, P. Di Matteo, F. Herpin, E. Lagadec, A. Lançon, Z. Meliani, & F. Royer, Di
Chauvin, G., Desidera, S., Lagrange, A. M., et al. 2017b, A&A, 605, L9
Chomez, A., Delorme, P., Lagrange, A.-M., et al. 2025, A&A, 697, A99
Chomez, A., Lagrange, A.-M., Delorme, P., et al. 2023, A&A, 675, A205
Christiaens, V., Gonzalez, C., Farkas, R., et al. 2023, The Journal of Open Source Software, 8, 4774
Claudi, R. U., Turatto, M., Giro, E., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III, ed. I. S. McLean, S. K. Ramsay, & H. Takami, 77350V
Currie, T., Biller, B., Lagrange, A., et al. 2023, in Astronomical Society of the Pacific Conference Series, Vol. 534, Protostars and Planets VII, ed. S. Inutsuka, Y. Aikawa, T. Muto, K. Tomida, & M. Tamura, 799
Daglayan, H., Vary, S., Absil, O., et al. 2024, A&A, 692, A126
Dahlqvist, C. H., Cantalloube, F., & Absil, O. 2020, A&A, 633, A95
Dahlqvist, C. H., Cantalloube, F., & Absil, O. 2021a, A&A, 656, A54
Dahlqvist, C. H., Louppe, G., & Absil, O. 2021b, A&A, 646, A49
Dahlqvist, C. H., Milli, J., Absil, O., et al. 2022, A&A, 666, A33
de Boer, J., Salter, G., Benisty, M., et al. 2016, A&A, 595, A114
Delorme, P., Meunier, N., Albert, D., et al. 2017, in SF2A-2017: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. C. Reylé, P. Di Matteo, F. Herpin, E. Lagadec, A. Lançon, Z. Meliani, & F. Royer, 237
Desgrange, C., Chauvin, G., Christiaens, V., et al. 2022, A&A, 664, A139
Desidera, S., Chauvin, G., Bonavita, M., et al. 2021, A&A, 651, A70
Dohlen, K., Langlois, M., Saisse, M., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7014, Ground-based and Airborne Instrumentation for Astronomy II, ed. I. S. McLean & M. M. Casali, 70143L
Fedrigo, E., Donaldson, R., Soenke, C., et al. 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 6272, Advances in Adaptive Optics II, ed. B. L. Ellerbroek & D. Bonaccini Calia, 627210
Feldt, M., Olofsson, J., Boccaletti, A., et al. 2017, A&A, 601, A7
Flasseur, O., Bodrito, T., Mairal, J., et al. 2024, MNRAS, 527, 1534
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2018, A&A, 618, A138
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2020, A&A, 637, A9
Galicher, R., Boccaletti, A., Mesa, D., et al. 2018, A&A, 615, A92
Gomez Gonzalez, C. A., Absil, O., & Van Droogenbroeck, M. 2018, A&A, 613, A71
Gomez Gonzalez, C. A., Wertz, O., Absil, O., et al. 2017, AJ, 154, 7
Janson, M., Asensio-Torres, R., André, D., et al. 2019, A&A, 626, A99
Janson, M., Squicciarini, V., Delorme, P., et al. 2021, A&A, 646, A164
Jensen-Clem, R., Mawet, D., Gomez Gonzalez, C. A., et al. 2018, AJ, 155, 19
Juillard, S., Christiaens, V., Absil, O., Stasevic, S., & Milli, J. 2024, A&A, 688, A185
Lafrenière, D., Marois, C., Doyon, R., Nadeau, D., & Artigau, É. 2007, ApJ, 660, 770
Lagrange, A. M., Langlois, M., Gratton, R., et al. 2016, A&A, 586, L8
Langlois, M., Gratton, R., Lagrange, A. M., et al. 2021, A&A, 651, A71
Launhardt, R., Henning, T., Quirrenbach, A., et al. 2020, A&A, 635, A162
Maire, A.-L., Langlois, M., Dohlen, K., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, 990834
Marois, C., Correia, C., Galicher, R., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9148, Adaptive Optics Systems IV, ed. E. Marchetti, L. M. Close, & J.-P. Vran, 91480U
Marois, C., Lafrenière, D., Doyon, R., Macintosh, B., & Nadeau, D. 2006, ApJ, 641, 556
Mawet, D., Milli, J., Wahhaj, Z., et al. 2014, ApJ, 792, 97
Milli, J., Kasper, M., Bourget, P., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10703, Adaptive Optics Systems VI, ed. L. M. Close, L. Schreiber, & D. Schmidt, 107032A
Milli, J., Mouillet, D., Fusco, T., et al. 2017, in Adaptive Optics for Extremely Large Telescopes 5, 0034
Müller, A., Keppler, M., Henning, T., et al. 2018, A&A, 617, L2
Nielsen, E. L., De Rosa, R. J., Wang, J. J., et al. 2020, AJ, 159, 71
Pairet, B., Cantalloube, F., Gomez Gonzalez, C. A., Absil, O., & Jacques, L. 2019, MNRAS, 487, 2262
Pueyo, L. 2016, ApJ, 824, 117
Ren, B., Pueyo, L., Zhu, G. B., Debes, J., & Duchêne, G. 2018, ApJ, 852, 104
Sarazin, M. & Roddier, F. 1990, A&A, 227, 294
Soummer, R., Pueyo, L., & Larkin, J. 2012, ApJ, 755, L28
Vigan, A., Fontanive, C., Meyer, M., et al. 2021, A&A, 651, A72

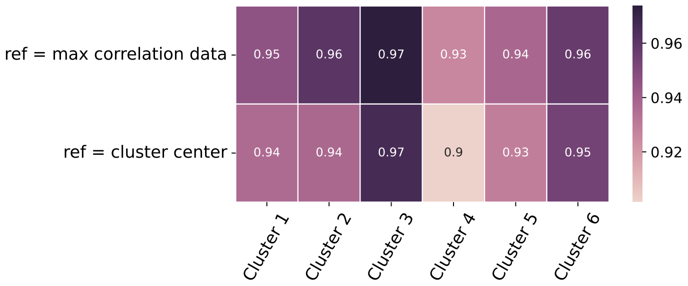


Fig. A.1: Spearman rank correlation comparison between cluster centers and most correlated datasets in each cluster.

Appendix A: Pixel correlation between datasets

To assess how well the chosen cluster centers reflect the noise behavior of other datasets within the same cluster, we performed a pixel correlation test. This test computes the correlation between the selected cluster center and the other datasets in the cluster. Additionally, we compare this correlation to that of the dataset exhibiting the highest correlation within the cluster.

To study the correlation between datasets, we adopted the approach used in Juillard et al. (2024), using a correlation metric from the VIP package based on the Spearman rank correlation coefficient. This coefficient was chosen to avoid the normality assumption required by the Pearson coefficient and to circumvent the dependence of the Structural Similarity Index (SSIM) on the image dynamic range, which will be normalized later in the post-processing. For each dataset, we identified the frame within the science cube that exhibits the highest correlation with other frames, designating it as the representative frame of that dataset. We then computed the pixel correlation between this representative frame and the representative frames of all other datasets in the cluster. This process was repeated for all datasets within one cluster as well as for all clusters. The dataset exhibiting the highest correlation with the other datasets in its cluster is expected to share the most prominent noise features and thus be the most representative. We then compared the Spearman correlation of the selected cluster centers, chosen based on environmental conditions, to that of the most correlated dataset in each cluster. The results are shown in Figure A.1.

The high correlation values for the cluster centers confirm that these selections reliably represent their respective clusters. This correlation metric proves less effective for clusters affected by strong wind-driven halo effects (clusters 4 and 5). This limitation arises from the rotation of wind-driven speckles with the parallactic angle (Cantalloube et al. 2020), which introduces additional complexity. In such cases, high correlation coefficients tend to identify datasets with less pronounced wind effects rather than true feature similarity, reducing the effectiveness of correlation as a metric in these specific scenarios.

Appendix B: Clustering illustration using principal components

To analyze the distinctions between clusters, we projected the environmental datasets onto their first three principal components. Figure B.1 (top) shows that most clusters are well-separated, although clusters 2 and 6 overlap. This overlap is resolved when examining the third principal component, as shown in Plot B.1 (bottom).

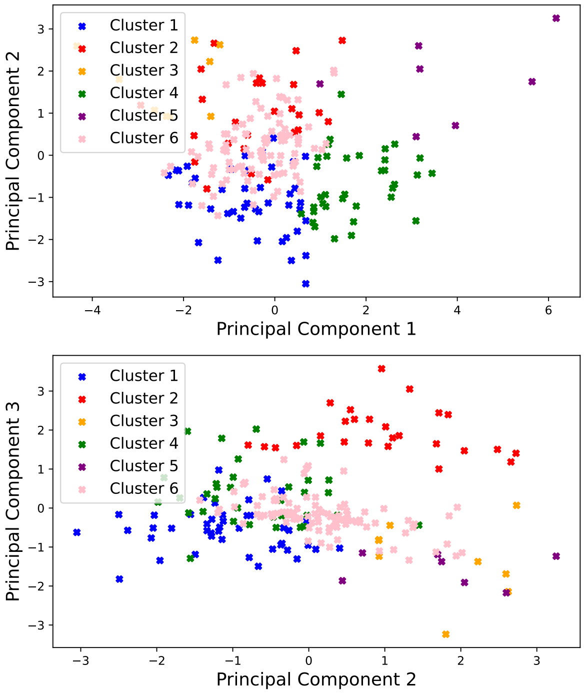


Fig. B.1: Visualization of clusters using principal component analysis. *Top* shows the first vs second component projection. *Bottom* shows second vs third component projection.

Appendix C: Optimal parameters for cluster centers

Appendix D: Variation of optimal RSM parameters across cluster centers and individual datasets

This section examines the impact of clustering on optimal parameter selection in RSM, emphasizing how different observing conditions influence the choice of PSF subtraction technique parameters and assessing the reliability of generalizing cluster center parameters to all datasets within a cluster. To quantify these effects, we measure the normalized distance between parameters. For numerical parameters, we compute their normalized distances, while for categorical parameters (e.g., the noise estimation methods), we use the dissimilarity index, a normalized measure of how different two RSM parameter sets are, with 0 indicating identical configurations and 1 indicating distinct ones. The mean normalized distance across all parameters is then calculated for each cluster center pair. Figure D.1 (top) presents these distances, revealing substantial variability in PSF subtraction parameters, with a median distance of 0.47, while RSM parameters exhibit lower variability, with a median distance of 0.1. This indicates significant differences in optimal parameter choices between cluster centers for PSF subtraction techniques. Overall, different PSF subtraction methods exhibit varying median distances across the clusters, with APCA having the lowest median distance of 0.33 and LOCI the highest at 0.68.

To assess the validity of generalizing parameters from a single representative dataset to an entire cluster, we conducted the same analysis within a small but diverse cluster (Cluster 1). Figure D.1 (bottom) illustrates the normalized distances between

Table C.1: Optimal parameters for PSF subtraction techniques and RSM for each cluster center.

Parameters	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
APCA components	20	14	18	8	7	12
APCA segments	4	3	3	3	3	4
APCA FOV rotation	0.42	0.29	0.26	0.31	0.5	0.76
NMF components	9	10	10	22	11	6
LOCI tolerance	0.0038	0.0016	0.0012	0.0057	0.0096	0.002
LOCI FOV rotation	0.49	0.14	0.48	0.1	0.13	0.73
FM KLIP components	9	19	8	12	5	5
FM KLIP FOV rotation	0.51	0.27	0.27	0.34	0.25	0.81
FM LOCI tolerance	0.0042	0.0074	0.0072	0.0073	0.0099	0.0016
FM LOCI FOV rotation	0.11	0.13	0.13	0.11	0.2	0.61
APCA δ	5	5	5	5	5	5
NMF δ	5	5	5	5	5	5
LOCI δ	5	5	5	5	5	5
FM KLIP δ	5	5	5	5	5	5
FM LOCI δ	5	5	5	5	5	5
APCA crop size	4	4	4	4	4	4
NMF crop size	4	4	4	4	4	4
LOCI crop size	4	4	4	4	4	4
FM KLIP crop size	7	7	7	7	7	7
FM LOCI crop size	7	7	7	7	7	7
APCA variance	SM	ST	SM	SM	SM	FR
NMF variance	SM	SM	SM	SM	SM	SM
LOCI variance	FR	ST	ST	SM	SM	ST
FM KLIP variance	FR	ST	ST	SM	ST	SM
FM LOCI variance	SM	SM	ST	ST	ST	ST

Notes. The RSM parameters include the planetary flux multiplicative factor (δ), defined as a multiple of the standard deviation, the crop size, and the variance term used to determine the noise estimation region within the annulus. The noise estimation methods are denoted as follows: "SM" (Segment with mask-based estimation), "ST" (Spatio-Temporal estimation), and "FR" (Frame-based estimation) (for further details, see Dahlqvist et al. 2021a).

20 datasets and the cluster center, showing a reduced median distance of 0.33 for PSF subtraction parameters. One dataset (HD 109573) failed to converge during optimization due to the presence of bright disk emission, explaining the use of 20 instead of 21 datasets. The slightly larger distance of 0.15 observed for RSM parameters primarily stems from variations in the noise estimation region parameter, which has minimal impact on the final RSM maps. In contrast, PSF subtraction parameters have a much stronger influence on the results. The reduction in parameter distance suggests that datasets with similar observing conditions tend to converge toward more comparable parameter choices.

Overall, this analysis underscores the substantial differences in optimal parameter selection between different cluster centers, highlighting the need to adapt parameter choices to specific datasets across a survey. At the same time, it demonstrates that datasets with similar observing conditions generally exhibit closer parameter distances. However, some datasets within a single cluster still show notable variations, reflecting dataset-specific noise properties. Despite these differences, we adopt a cluster-based parameter optimization approach in this study, as optimizing RSM parameters for each individual dataset remains computationally costly.

Appendix E: Comparison with the Dahlqvist contrast curve definition

This section compares the contrast curve methodology used in Dahlqvist et al. (2022) for the SHARDDS survey with the approach applied in our study. Dahlqvist et al. (2022) define the detection threshold based on the first false positive within the field of view and considers 95% recovery of injected signals. In contrast, our approach employs a detection threshold corresponding to an FAP of 3×10^{-7} under a lognormal distribution at a given separation and evaluates contrast at the 50% injection recovery level. Figure E.1 presents these contrast curves for two datasets: one without a potential signal (Fig. E.1a) and another with a potential signal (Fig. E.1b). In Fig. E.1a, the contrast curves derived from different detection thresholds remain consistent at the same completeness levels. However, this trend does not hold in Fig. E.1b, where the contrast curve based on the lognormal detection threshold outperforms that derived from the first false positive at both completeness levels. In both figures, there is only a minor gap between the contrast levels required for 50% and 95% signal recovery. This result highlights a key limitation of the approach in Dahlqvist et al. (2022), where the contrast is influenced by the brightest signal in the field of view without determining whether it originates from noise or a potential candidate —a limitation that also applies to thresholds based on the maximum F1 score. Moreover, this method defines the detection threshold based on the entire field of view rather than at each separation, limiting its ability to assess contrast variations across different separations.

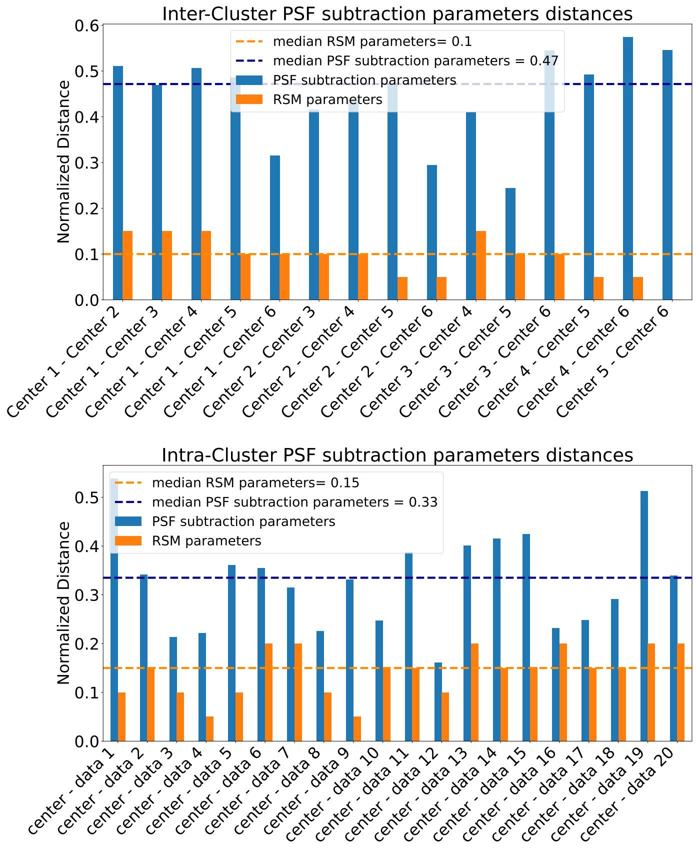


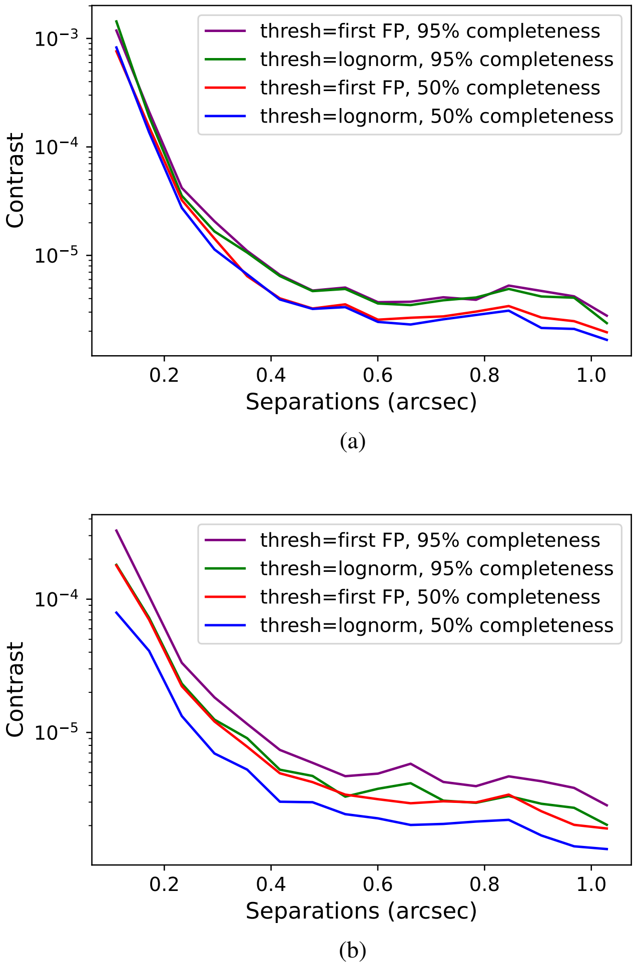
Fig. D.1: Normalized Distances for PSF Subtraction (blue) and RSM Parameters (orange) between cluster centers (top) and datasets within a single cluster (bottom).

Table F.1: Kolmogorov-Smirnov test for different distributions fits for cluster 1 RSM noise realization samples: p-value at $0\farcs25$ and $1''$using RSM APCA.

Distributions	sep = $0\farcs25$	sep = $1''$
Log of Lognorm	0.5	0.09
Norm	0.72	0.00
Generalized norm	0.43	0.71
Power norm	0.56	0.17
Laplace	0.04	0.00

Appendix F: Comparing RSM noise distribution models

The distribution of independent RSM noise realizations, as discussed in Section 5.1, exhibits a near-normal profile when viewed on a logarithmic scale. Here, we investigate various normal-like distributions to determine the best fit for the datasets. Specifically, we compare the log of lognormal distribution (three parameters), the normal distribution (two parameters), the power-normal distribution (three parameters), the generalized normal distribution (three parameters), along with the Laplace distribution (two parameters). Figure F.1 presents the distribution of independent RSM noise realizations using APCA in the RSM context for cluster 1 at two separations, along with the fitted distributions. It is important to note that the data used in this section fall below the 3×10^{-7} threshold. The histograms demonstrate a clear normal-like profile, with all tested distributions providing a reasonable fit, with the laplace being the least.



(a)

(b)

Fig. E.1: Comparison of contrast curves for two datasets, one with no potential companion (a) and one with a potential companion (b): 95% completeness using the first false positive threshold and lognormal threshold (purple and green), and 50% completeness for both thresholds (red and blue)

To assess the goodness of fit, we perform a Kolmogorov-Smirnov (KS) test and compute the p-values for each distribution at different separations. Table F.1 presents the results for Cluster 1, where RSM noise realizations were generated using APCA in the RSM detection map. The table shows that the generalized normal, the logarithm of the lognormal, and the power-normal distributions achieve p-values above 0.05, indicating a good fit. While the generalized normal and power-normal distributions yield high p-values, they do not consistently perform well across all clusters (e.g., failing for Cluster 6). Given its strong overall fit across different clusters and PSF subtraction technique combinations, we adopt the lognormal distribution as the preferred model.

Appendix G: Maximum F1 score threshold generalized across a cluster

This section evaluates the generalization of thresholds determined using the maximum F1 score from a given dataset across other datasets with similar observing conditions. The maximum F1 score is highly sensitive to false positives, often influenced by

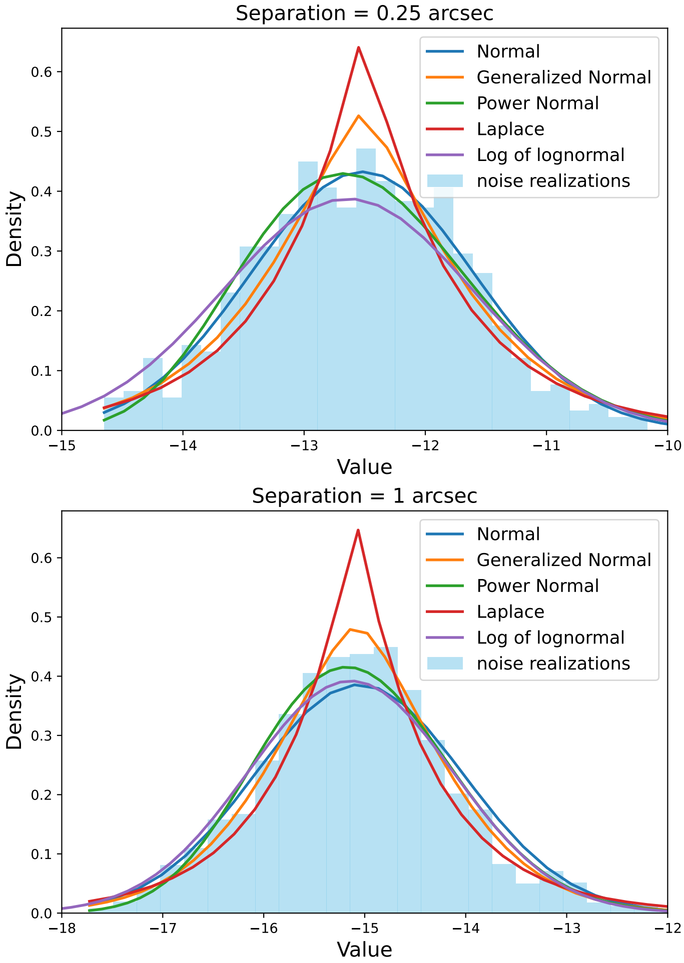


Fig. F.1: Histograms of independent RSM noise realizations (log scale) in the RSM maps of Cluster 1, obtained using APCA RSM, at 0.″25 (top) and at 1″ (bottom). The fitted distributions include normal (blue), generalized normal (orange), power-normal (green), Laplace (red), and log of lognormal (purple).

the brightest speckle in the representative dataset rather than the overall noise profile. Generalizing such thresholds can lead to inconsistencies, as the presence of these speckles reflects specific environmental parameters and their interaction with adaptive optics. These factors are inherently unique to individual observations and cannot be reliably generalized, even under broadly similar observing conditions. Figure G.1 illustrates this issue by generalizing the maximum F1 score threshold from the cluster center BD-15 705 to the dataset HIP 92680 in cluster 1. The threshold derived from the cluster center (pink) shows a factor of 10 difference compared to the self-computed maximum F1 score (cyan) and the lognormal threshold (dark blue). This significant discrepancy leads to the generalized threshold accepting false positives that would otherwise be excluded, underscoring the difficulties in applying overly tuned thresholds across datasets.

Appendix H: Comparison between RSM maps and PCA Signal-to-Noise Ratio maps

This section presents a comparison between the RSM maps and the corresponding PCA signal-to-noise ratio (S/N) maps for several newly identified background-star detections. Specifically, we compare the optimal RSM maps obtained in the H2 and H3 filters using the combined APCA–NMF–LOCI–FMKLIP configuration with the five-component PCA S/N maps.

The distinction between the two approaches is evident: in the PCA S/N maps, many signals are barely distinguishable from the surrounding noise, whereas in the RSM maps these same sources stand out clearly with significantly higher RSM scores. Based on the color–magnitude diagram (CMD) analysis, these signals are classified as background stars.

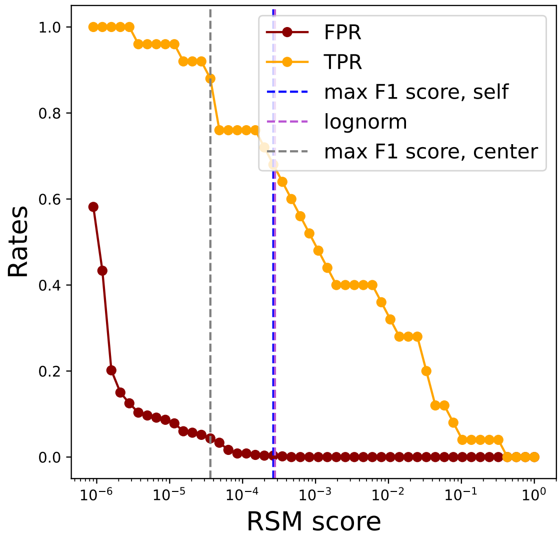
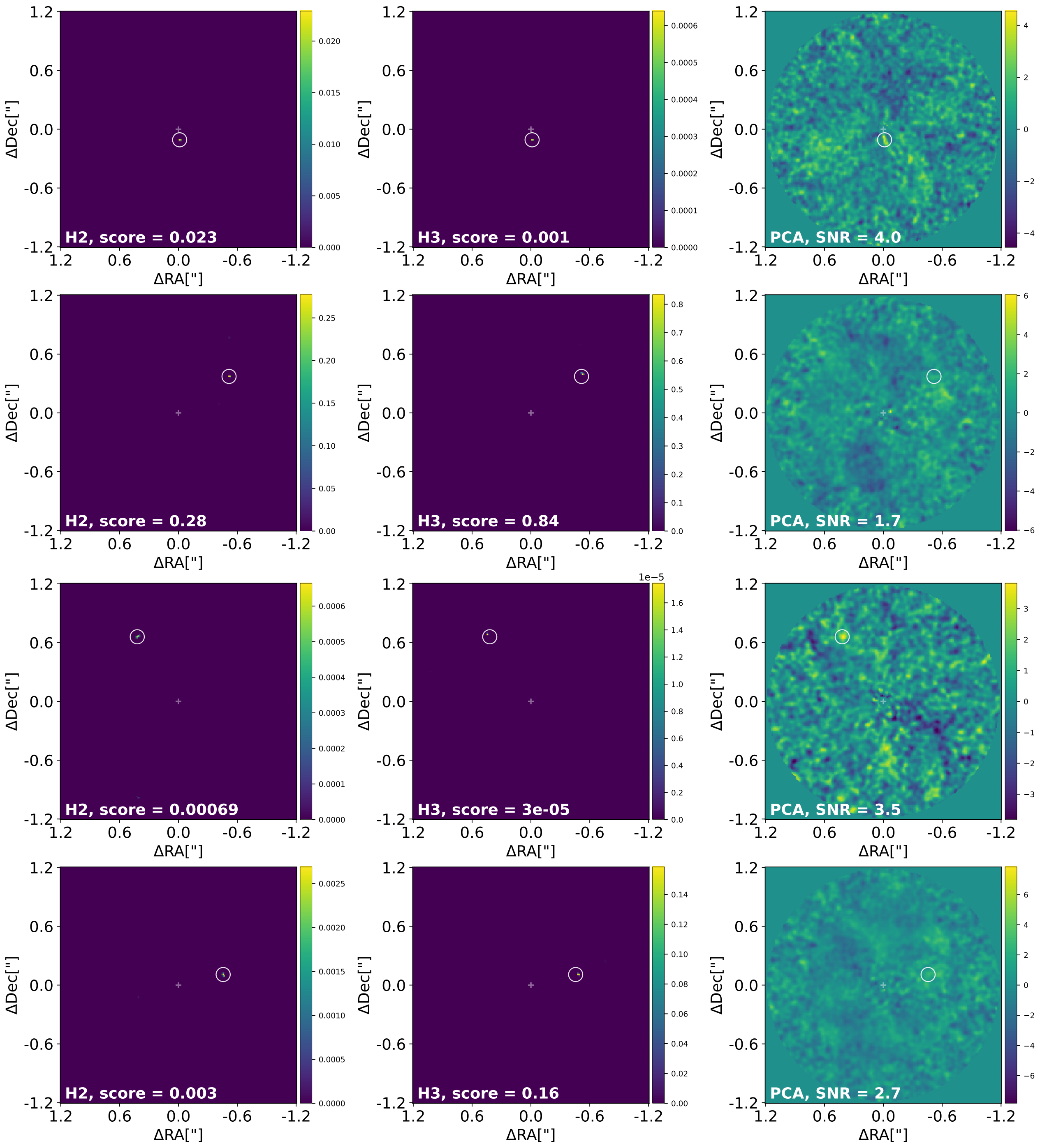


Fig. G.1: Maximum F1 score threshold for the cluster center 1 BD-15 705 (in pink) applied on the HIP 92680 dataset, in comparison to the maximum F1 score computed by the dataset HIP 92680 itself denoted as "self" (in cyan) and the lognormal threshold (in dark blue) done by the cluster 1, at separation 0.″55.

Fig. H.1: Comparison of RSM and PCA S/N maps for several newly identified background stars. The left panels show the RSM maps generated with the APCA–NMF–LOCI–FMKLIP combination in the H2 and H3 filters, along with their corresponding RSM scores. The right panels display the PCA S/N maps computed with 5 components.


H2, score = 0.023
H3, score = 0.001
PCA, SNR = 4.0
H2, score = 0.28
H3, score = 0.84
PCA, SNR = 1.7
H2, score = 0.00069
H3, score = 3e-05
PCA, SNR = 3.5
H2, score = 0.003
H3, score = 0.16
PCA, SNR = 2.7
ΔDec["]
ΔRA["]
1e−5