SF2A 2024
M. Béthermin, K. Baillié, N. Lagarde, J. Malzac, R.-M. Ouazzani, J. Richard, O. Venot, A. Siebert (eds)

DATA SCIENCE APPROACHES FOR THE EXPLORATION OF CIRCUMSTELLAR ENVIRONMENTS IN HIGH-CONTRAST AND HIGH-ANGULAR RESOLUTION IMAGING

O. Flasseur[1]

Abstract. High-contrast imaging (HCI) is critical for exoplanet detection, characterization, and exploring the circumstellar environment. We present recent advancements in data science methods for HCI, focusing on accurate modeling of nuisance statistics and their correlations. In that context, we introduce `deep PACO`, `MODEL&CO`, `PACOME` and `REXPACO ASDI` algorithms to improve detection sensitivity, characterization, and reconstruction fidelity. These techniques are well suited for data from upcoming large-aperture telescopes.

Keywords: high-contrast imaging, high angular resolution, data science, statistical methods, deep learning

1 Introduction

Exoplanet detection, atmospheric characterization, and planetary formation studies are key challenges in modern astrophysics (Currie et al. 2022). In that context, HCI is a method of choice for probing the vicinity of young nearby stars. However, such observations are challenging due to the extreme contrast and angular resolution required (Traub & Oppenheimer 2010; Follette 2023). Advanced signal processing is thus essential to separate exoplanetary and disks signals from dominant nuisance components like quasi-static speckles and stochastic noise corrupting the observations (Pueyo 2018).

Recent advances in data science have optimized astrophysical information extraction by modeling spatial, temporal, spectral, and multi-epoch correlations. These methods integrate statistical and physical modeling, deep learning, and data fusion techniques, enabling unsupervised, data-driven approaches for high-dimensional parameter estimation. Applied to real data, such as from VLT/SPHERE, these techniques significantly improve detection sensitivity, exoplanetary spectrum retrieval accuracy, and circumstellar disk reconstruction.

This paper presents recent data science advancements for HCI, combining statistical modeling and deep learning (Sect. 2), multi-epoch data fusion (Sect. 3), and inverse-problem-based disk reconstruction (Sect. 4). Future prospects, in particular in the context of upcoming 30-meter-class telescopes, are outlined in Sect. 5.

2 Hybrid approaches combining statistical modeling and deep learning

2.1 Observation-dependent model

Accurate modeling of the dominant quasi-static nuisance component is crucial, particularly in ground-based observations. Various post-processing methods have been developed, ranging from traditional image combination and subtraction techniques (Marois et al. 2006; Lafrenière et al. 2007; Lagrange et al. 2009; Soummer et al. 2012; Amara & Quanz 2012; Marois et al. 2013, 2014; Wahhaj et al. 2015; Gonzalez et al. 2016), where the nuisance is treated as a low-rank component, to statistical (Cantalloube et al. 2015; Ruffio et al. 2017; Flasseur et al. 2018, 2020b,a) and machine learning approaches (Gonzalez et al. 2018; Gebhard et al. 2022; Cantero et al. 2023; Chintarungruangchai et al. 2023; Wolf et al. 2024; Flasseur et al. 2024a; Bodrito et al. 2024).

In that context, we recently introduced the `deep PACO` algorithm (Flasseur et al. 2022a, 2023, 2024a), which integrates a statistical model with deep learning to leverage the strengths of both. `deep PACO` operates in two stages: (i) data pre-processing using `PACO`'s statistical model to capture local correlations via covariance

[1] Univ. de Lyon, Univ. Lyon1, ENS de Lyon, CNRS, Centre de Recherche Astrophysique de Lyon, Saint-Genis-Laval, France

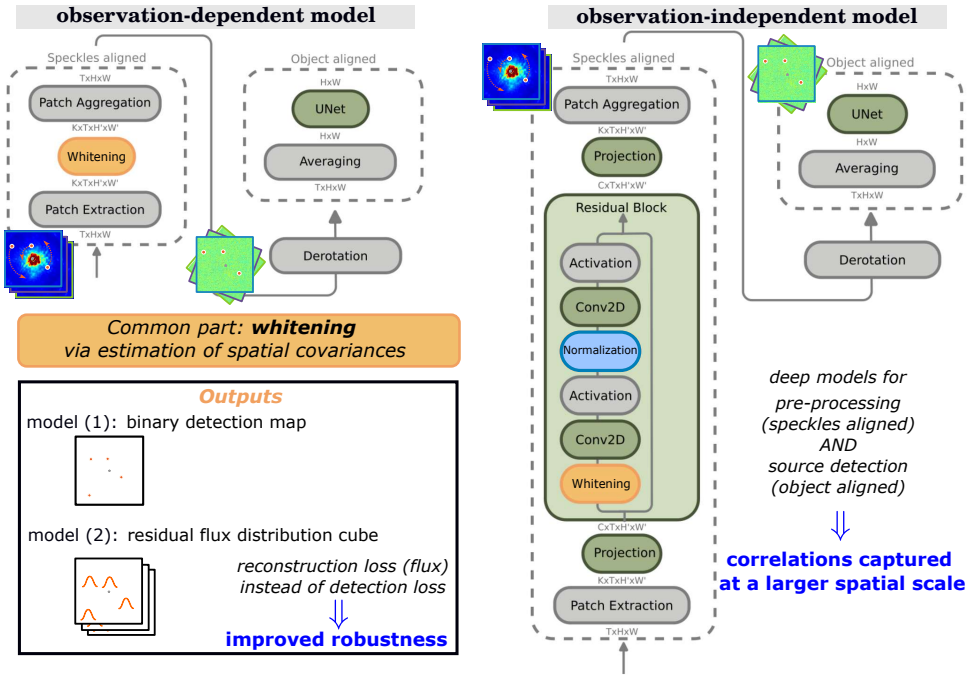

Fig. 1. Architecture of the hybrid `deep PACO` (left) and `MODEL&CO` (right) algorithms.

estimation, and (ii) detection using a deep convolutional neural network (CNN) trained on the pre-processed data to address the slight mismatch between the statistical model and the actual observations. The following section outlines the key steps of `deep PACO`.

A dataset $\boldsymbol{r} \in \mathbb{R}^{N\times T\times L}$, where N represents the number of pixels per exposure, T the number of exposures, and L the number of spectral channels, recorded through angular and spectral differential imaging (ASDI; Sparks & Ford (2002); Marois et al. (2006)) using the pupil tracking mode of the telescope and an Integral Field Spectrograph (IFS), is modeled as:

$$\boldsymbol{r} = \boldsymbol{f} + \sum_{p=1}^{P} \boldsymbol{\alpha}_p \, \boldsymbol{h}(\phi_p) \,, \tag{2.1}$$

where $\boldsymbol{f}$ is the nuisance component, and $\boldsymbol{h}(\phi_p)$ represents the contribution of point-like source p at position ϕ_p with spectral energy distribution (SED) $\boldsymbol{\alpha}_p \in \mathbb{R}^L$. The nuisance component $\boldsymbol{f}_{n,\ell}$ at location n and spectral channel ℓ is modeled locally with a multi-variate Gaussian $\mathcal{N}(\boldsymbol{m}_{n,\ell}, \mathbf{C}_n)$, where $\mathbf{C}_n$ captures correlations across patches of a few tens of pixels centered at location n. The covariance matrix is estimated through a shrinkage estimator:

$$\widehat{\mathbf{C}}_n = (1-\widehat{\rho}_n)\widehat{\mathbf{S}}_n + \widehat{\rho}_n \widehat{\mathbf{F}}_n \,, \tag{2.2}$$

where $\widehat{\mathbf{S}}_n$ is the local sample covariance, $\widehat{\mathbf{F}}_n$ is diagonal accounting for the data variances, and $\widehat{\rho}_n$ controls a bias-variance trade-off (Ledoit & Wolf 2004). Parameter $\widehat{\rho}_n$ can be estimated optimally by minimizing the estimation risk between the true (and unknown) covariance $\mathbf{C}_n$ and its shrunk estimate $\widehat{\mathbf{C}}_n$ (Chen et al. 2010; Flasseur et al. 2024c). After estimation of the parameters of the statistical model at each patch location n, the data are then centered and whitened locally via operator $\mathbf{W}_n$:

$$\widetilde{\boldsymbol{r}}_n = \mathbf{W}_n \, \boldsymbol{r}_n = \widehat{\mathbb{L}}_n^{\top}(\boldsymbol{r}_n - \widehat{\boldsymbol{m}}_n) \,, \tag{2.3}$$

where $\widehat{\mathbb{L}}_n$ is the Cholesky factorization of $\widehat{\mathbf{C}}_n^{-1}$. This step removes most of the spatial correlations and normalizes the data, improving the contrast and stationarity of the observations. After this pre-processing, the detection problem is formulated as a semantic segmentation task, where a CNN is trained in a supervised manner to detect synthetic sources. Training samples are generated by injecting synthetic sources into the pre-processed data through direct model (2.1). Beforehand, for each training sample, a shuffling of the temporal frames of $\boldsymbol{r}$ is applied. This custom data augmentation strategy allows to generate a large training set from a single spatio-temporal-spectral dataset $\boldsymbol{r}$ (i.e., the model is *observation-dependent* because it should be retrained for

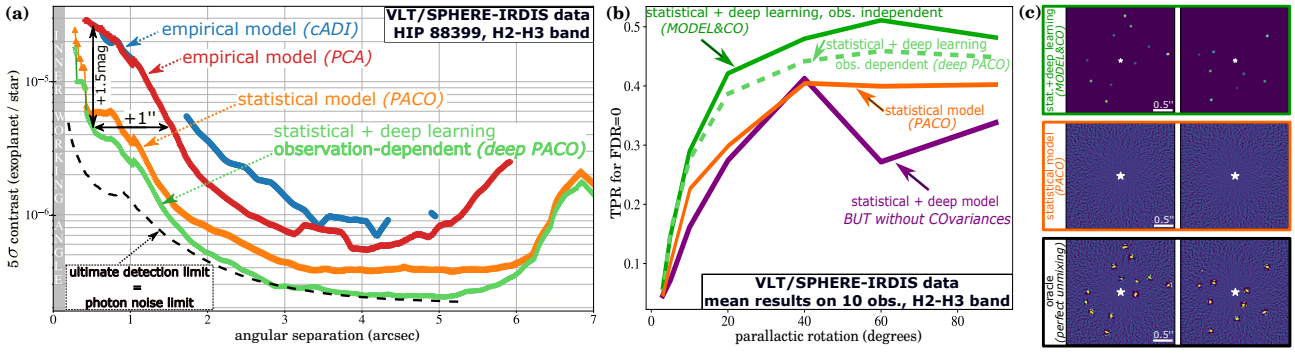

Fig. 2. Performance of hybrid `deep PACO` and `MODEL&CO` methods on VLT/SPHERE data. (a): Contrast curves comparing empirical approaches (cADI, PCA) with `PACO` and `deep PACO`. The dashed black line represents the ultimate detection sensitivity, determined by the photon-noise limit. (b): Aggregated detection score (a trade-off between precision and recall, where higher values indicate better performance) as a function of angular diversity (amplitude of parallactic rotation) for `PACO`, `deep PACO`, and `MODEL&CO`. A model ablation (purple line) of `MODEL&CO`, which ignores the covariance information in the statistical model highlights the critical importance of accounting for measurement statistics. (c): Detection maps comparing `MODEL&CO`, `PACO`, and a perfect unmixing (unachievable in practice).

each new dataset $\boldsymbol{r}$). The network leverages a U-Net architecture with a ResNet18 backbone as encoder. The CNN is trained from scratch by optimizing the Dice2 loss:

$$\mathcal{L}\left(\boldsymbol{y}, \widehat{\boldsymbol{y}}\right) = 1 - \underbrace{\frac{\sum_{m=1}^{M}(1-\boldsymbol{y}_m)(1-\widehat{\boldsymbol{y}}_m+\epsilon)}{\sum_{m=1}^{M} 2-\boldsymbol{y}_m-\widehat{\boldsymbol{y}}_m+\epsilon}}_{\text{background error}} - \underbrace{\frac{\sum_{m=1}^{M}\boldsymbol{y}_m\,\widehat{\boldsymbol{y}}_m+\epsilon}{\sum_{m=1}^{M}\boldsymbol{y}_m+\widehat{\boldsymbol{y}}_m+\epsilon}}_{\text{source error}},$$

where $\boldsymbol{y} \in \mathbb{R}^M$ is the ground truth for the synthetic sources, $\widehat{\boldsymbol{y}}$ is the predicted detection map, and ϵ is a small constant for stability.

2.2 Observation-independent model

In HCI, the detection challenge arises from the overwhelming stellar glare and the small angular separation between the planet and star. The `deep PACO` approach (see Sect. 2.1) introduces a model that captures local correlations in the nuisance component, combining it with deep learning techniques for improved detection sensitivity. However, `deep PACO` builds the nuisance model from the observations $\boldsymbol{r}$ themselves, which inherently limits its effectiveness at small angular separations due to the lack of angular diversity in the data. In that context, we recently proposed `MODEL&CO` (Bodrito et al. 2024) extending the `deep PACO` framework by addressing this limitation through the use of an external archive of multiple observations $\boldsymbol{r}$ to construct the nuisance model. The methodology strongly differs from reference differential imaging (RDI; see e.g. Ruane et al. (2019)) in the sense that a highly non-linear model is learned from the data. In addition, contrary to `deep PACO`, a learnable module is introduced within two complementary representations of the data having either speckles or signals of the sought objects spatially co-aligned, see Fig. 1 comparing the architecture of `deep PACO` and `MODEL&CO`. This allows the model to capture more general and diverse nuisance structures, leading to improved sensitivity, particularly in scenarios with reduced angular diversity, as illustrated by Fig. 2.

3 Joint exoplanet's detection and characterization by multi-epoch fusion

Beyond optimal post-processing of individual observations, fusing multiple observations of the same star taken over different epochs can significantly improve the detection sensitivity. The key challenge in this approach lies in accounting for both the nuisance statistics and the orbital motion of the exoplanet across epochs. To address this, we recently introduced `PACOME` (`PACO` Multi-Epoch; Dallant et al. (2022, 2023b,a)), which builds upon the statistical modeling of the nuisance component described in Sect. 2.1. The by-products of `PACO` from each epoch provide sufficient statistics that can be optimally combined using `PACOME`, while efficiently exploring the Keplerian motion of exoplanets through a Hamiltonian Monte Carlo algorithm. This multi-epoch

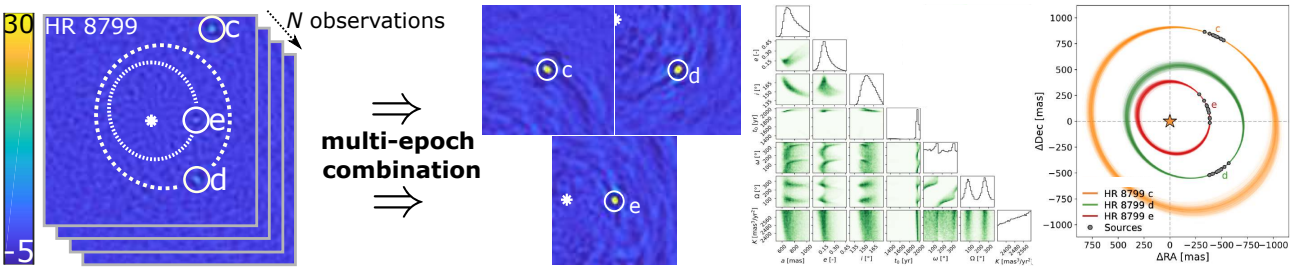

Fig. 3. Principle of the multi-epoch fusing algorithm `PACOME` on VLT/SPHERE data. Left: N_{obs} mono-epoch results serving as inputs. Right: Outputs produced by `PACOME`, i.e. optimally combined S/N maps (contrast improved by $\sqrt{N_{\text{obs}}}$ at controlled false alarm rate) and corner plots (joint and marginal distributions) on the estimated orbital parameters.

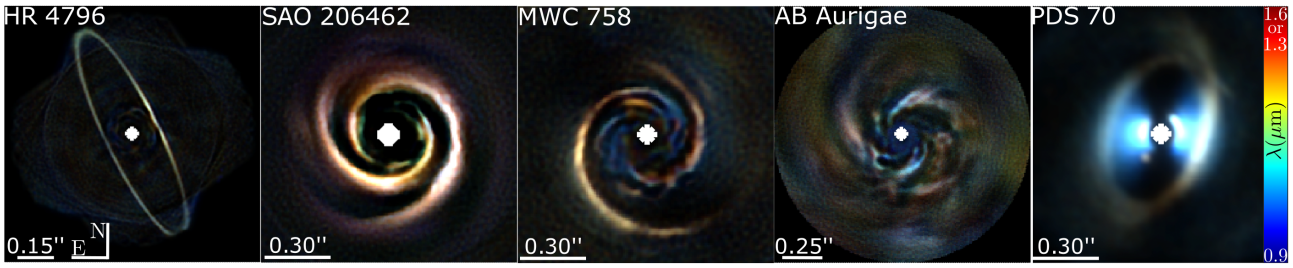

Fig. 4. `REXPACO ASDI` reconstructions on VLT/SPHERE-IFS datasets. Fake colors encode for the wavelength.

strategy yields a combined detection score that is directly interpretable as a measure of detection confidence. In addition to improving sensitivity, `PACOME` enables the estimation of orbital parameters, along with their joint and marginal distributions. This approach is especially well-suited for future large-aperture telescopes, where achieving the required contrast will require long total exposure times on the same star, making a multi-epoch strategy essential. Figure 3 illustrates the principle of `PACOME` and the typical outputs it generates.

4 Reconstruction of the circumstellar environment

Beyond the detection of point-like sources, reconstructing the spatio-spectral flux distribution of circumstellar environments is crucial for advancing our understanding of exoplanetary system formation, evolution, and diversity. To address this challenge, we recently introduced the `REXPACO ASDI` algorithm (Flasseur et al. 2022b, 2024b), designed for joint unmixing and deconvolution of ASDI data, with the aim of reconstructing circumstellar disks at high contrast. The algorithm models the stellar light and noise by a multi-variate Gaussian distribution within spatio-spectral patches (see Sect. 2.1), parametrized by spectral mean and separable spatio-spectral covariances. This reconstruction method simultaneously estimates the objects of interest (i.e., the disk) and the nuisance statistics using an inverse problem approach. The hyper-parameters are optimally estimated directly from the data by minimizing quantitative metrics, without the need for synthetic disk injection. `REXPACO ASDI` provides a significant improvement in reconstruction fidelity compared to state-of-the-art methods like PCA ASDI. Figure 4 presents examples of iconic circumstellar disks reconstructed with `REXPACO ASDI`.

5 Conclusions and prospects

HCI main goals such as exoplanet detection, characterization, and circumstellar environment reconstruction highlight the crucial need for accurately modeling measurement statistics, particularly their multiple correlations. Ignoring these correlations severely limits detection sensitivity, strongly degrades the accuracy of astro-photometric estimates, and reduces the fidelity of flux distribution reconstructions in circumstellar disks. To address these challenges, we have developed tailored data science techniques allowing optimal signal extraction and unsupervised tuning of algorithm hyper-parameters, without the need for simulated astrophysical scenes. In this paper, we present our recent developments in data science methods for high-contrast imaging and demonstrate their key benefits.

The next generation of thirty-meter-class telescopes, such as the ELT, offers the potential to greatly extend our exploration of the inner regions around solar-type stars. However, several data science challenges must still be overcome: (i) achieving optimal signal extraction to reach the performance limits of these instruments, (ii) accurately modeling spatially structured noise with significant fluctuations, and (iii) mitigating the limitations of A(S)DI, which becomes less effective at small angular separations. These challenges are particularly relevant for upcoming instruments like ELT/HARMONI, where recent simulations have demonstrated the benefits of advanced data-driven methodologies.

The author thanks the SF2A committee and the Action Spécifique Haute Résolution Angulaire (ASHRA).

References

Amara, A. & Quanz, S. P. 2012, Monthly Notices of the Royal Astronomical Society, 427, 948
Bodrito, T., Flasseur, O., Mairal, J., et al. 2024, Monthly Notices of the Royal Astronomical Society, 534, 1569
Cantalloube, F., Mouillet, D., Mugnier, L., et al. 2015, Astronomy & Astrophysics, 582, A89
Cantero, C., Absil, O., Dahlqvist, C. H., & Van Droogenbroeck, M. 2023, A&A, 680, A86
Chen, Y., Wiesel, A., Eldar, Y. C., & Hero, A. O. 2010, IEEE Transactions on Signal Processing, 58, 5016
Chintarungruangchai, P., Jiang, G., Hashimoto, J., Komatsu, Y., & Konishi, M. 2023, New Astronomy, 100, 101997
Currie, T., Biller, B., Lagrange, A.-M., et al. 2022, arXiv preprint arXiv:2205.05696
Dallant, J., Langlois, M., Flasseur, O., & Thiébaut, É. 2023a, Astronomy & Astrophysics, 679, A38
Dallant, J., Langlois, M., Thiébaut, É., & Flasseur, O. 2022, in Adaptive Optics Systems VIII, Vol. 12185, SPIE, 1015–1021
Dallant, J., Langlois, M., Thiébaut, E., & Flasseur, O. 2023b, in Adaptive Optics for Extremely Large Telescopes 7th Edition
Flasseur, O., Bodrito, T., Mairal, J., et al. 2022a, in Adaptive Optics Systems VIII, Vol. 12185, SPIE, 1154–1167
Flasseur, O., Bodrito, T., Mairal, J., et al. 2024a, Monthly Notices of the Royal Astronomical Society, 527, 1534
Flasseur, O., Bodrito, T., Mairal, J., et al. 2023, in 2023 31st European Signal Processing Conference (EUSIPCO), IEEE, 1723–1727
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2018, Astronomy & Astrophysics, 618, A138
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2020a, Astronomy & Astrophysics, 637, A9
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2020b, Astronomy & Astrophysics, 634, A2
Flasseur, O., Denis, L., Thiébaut, É., & Langlois, M. 2024b, Monthly Notices of the Royal Astronomical Society, stae2291
Flasseur, O., Denis, L., Thiébaut, É., Langlois, M., et al. 2022b, in Adaptive Optics Systems VIII, Vol. 12185, SPIE, 1175–1189
Flasseur, O., Thiébaut, E., Denis, L., & Langlois, M. 2024c, accepted in EUSIPCO, arXiv preprint arXiv:2403.07104
Follette, K. B. 2023, Publications of the Astronomical Society of the Pacific, 135, 093001
Gebhard, T. D., Bonse, M. J., Quanz, S. P., & Schölkopf, B. 2022, Astronomy & Astrophysics, 666, A9
Gonzalez, C., Absil, O., & Van Droogenbroeck, M. 2018, Astronomy & Astrophysics, 613, A71
Gonzalez, C. G., Absil, O., Absil, P.-A., et al. 2016, Astronomy & Astrophysics, 589, A54
Lafrenière, D., Marois, C., Doyon, R., Nadeau, D., & Artigau, E. 2007, The Astrophysical Journal, 660, 770
Lagrange, A.-M., Gratadour, D., Chauvin, G., et al. 2009, Astronomy & Astrophysics, 493, L21
Ledoit, O. & Wolf, M. 2004, Journal of multivariate analysis, 88, 365
Marois, C., Correia, C., Galicher, R., et al. 2014, in SPIE Astronomical Intrumentation + Telescopes, Vol. 9148, International Society for Optics and Photonics, 91480U
Marois, C., Correia, C., Véran, J.-P., & Currie, T. 2013, International Astronomical Union, 8, 48
Marois, C., Lafrenière, D., Doyon, R., Macintosh, B., & Nadeau, D. 2006, The Astrophysical Journal, 641, 556
Pueyo, L. 2018, Handbook of Exoplanets, 705
Ruane, G., Ngo, H., Mawet, D., et al. 2019, The Astronomical Journal, 157, 118
Ruffio, J.-B., Macintosh, B., Wang, J. J., et al. 2017, The Astrophysical Journal, 842, 14
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
Traub, W. A. & Oppenheimer, B. R. 2010, Exoplanets, 111
Wahhaj, Z., Cieza, L. A., Mawet, D., et al. 2015, Astronomy & Astrophysics, 581, A24
Wolf, T. N., Jones, B. A., & Bowler, B. P. 2024, The Astronomical Journal, 167, 92