Vers une séparation robuste du signal cosmologique à 21 cm : modélisation structurée, régularisation et inférence bayésienne

Mehdi Latif [1]

Projet de recherche indépendant

Version initiale : 30 septembre 2026 – Version révisée : 9 novembre 2026

Résumé

La séparation du signal cosmologique à 21 cm reste particulièrement sensible à la modélisation statistique des contaminants et aux erreurs de modèle. ML-GPR permet d'introduire des modèles de covariance plus flexibles, appris à partir de simulations. La séparation reste toutefois fragile lorsque ces simulations sont insuffisamment représentatives ou lorsque les structures statistiques du signal et des contaminants deviennent trop proches. L'enjeu est ainsi de déterminer si l'introduction d'informations structurelles supplémentaires peut améliorer la robustesse de la séparation et la quantification de sa fiabilité.

Ce document présente un projet de recherche construit à partir de l'analyse de plusieurs travaux sur la séparation du signal à 21 cm par GPR et ML-GPR. Deux axes méthodologiques sont considérés : la modélisation structurée des contaminants, en étudiant notamment des covariances apprises et des informations spatio-spectrales, et la régularisation de l'apprentissage des modèles de covariance. La proposition comprend également une stratégie d'évaluation reposant sur des simulations, puis, lorsque cela est possible, sur des observations réelles. Cette évaluation s'inscrit dans un cadre probabiliste et bayésien et porte notamment sur la robustesse de la séparation, la perte de signal, les biais sur le spectre de puissance et la fiabilité des estimations obtenues.

1 Problématique scientifique et perspectives méthodologiques

Le signal à 21 cm de l'hydrogène neutre constitue une sonde des premières phases de l'évolution de l'Univers, en particulier de l'aube cosmologique (*Cosmic Dawn*) et de l'époque de réionisation (*Epoch of Reionization*, EoR).

Son exploitation reste cependant particulièrement difficile : les émissions astrophysiques d'avant-plan sont beaucoup plus intenses que le signal cosmologique recherché, tandis que les effets instrumentaux et les erreurs de calibration peuvent introduire des structures spectrales supplémentaires et rendre la séparation des composantes plus difficile (Liu et Shaw 2020 ; Mertens et al. 2018).

Cette séparation peut être vue comme un problème inverse statistique : à partir d'observations résultant de la superposition de plusieurs contributions astrophysiques et instrumentales, il s'agit d'estimer les différentes composantes qui en sont à l'origine. Leur identification repose alors sur l'exploitation de propriétés statistiques permettant de les différencier.

La régression par processus gaussiens (*Gaussian Process Regression*, GPR) fournit précisément un cadre probabiliste pour ce problème. Dans ce cadre, les contributions présentes dans les données sont modélisées comme des composantes latentes auxquelles sont associées des structures de covariance distinctes. Leur séparation repose sur ces différences statistiques, en premier lieu leurs corrélations en fréquence (Mertens et al. 2018).

La qualité des modèles de covariance joue par conséquent un rôle central. Lorsque la covariance associée à une composante décrit mal les données, ou lorsque plusieurs composantes présentent des structures statistiques proches, leur séparation devient plus difficile et peut conduire à une attribution incorrecte d'une partie du signal ou des contaminants.

Cette sensibilité au choix du modèle de covariance avait déjà été mise en évidence pour GPR-FS (*Foreground Subtraction*, FS) : Kern et al. montrent que la procédure de normalisation utilisée dans l'analyse LOFAR peut conduire à une perte de signal lorsque la covariance du signal à 21 cm est mal spécifiée (Kern et Liu 2021).

ML-GPR enrichit cette approche en remplaçant, pour le signal à 21 cm, le kernel analytique par un modèle de covariance appris à partir de simulations (Mertens et al. 2024). Dans cette approche, un *Variational Auto-Encoder* (VAE) apprend une représentation de faible dimension d'une famille de spectres de puissance simulés, utilisée ensuite pour construire le modèle de covariance du signal intégré au GPR.

Cette représentation reste toutefois dépendante de la représentativité des simulations utilisées pour son apprentissage. Les études de robustesse montrent qu'un *covariance mismatch* peut encore biaiser la récupération du signal et que la séparation devient plus difficile lorsque les structures de covariance du signal cosmologique et de certaines composantes résiduelles deviennent proches (Acharya et al. 2024).

L'application de ML-GPR aux observations LOFAR montre aussi que l'amélioration du seul modèle de covariance du signal à 21 cm ne résout pas l'ensemble des difficultés. Les contaminants restent principalement décrits à partir de kernels analytiques et les corrélations spatiales ne sont pas explicitement intégrées dans le modèle actuel (Mertens et al. 2025).

1. ORCID : 0009-0001-1480-9770

Ces résultats conduisent à considérer comme verrou principal du projet la robustesse de la séparation lorsque l'information statistique utilisée pour distinguer les composantes devient insuffisante ou imparfaitement adaptée aux données.

1.1 Question de recherche et hypothèse de travail

Le projet de recherche proposé s'organise autour de la question suivante :

> *Comment introduire dans ML-GPR des informations structurelles supplémentaires permettant d'améliorer la séparation du signal à 21 cm lorsque les contaminants sont imparfaitement modélisés ou présentent des structures statistiques proches de celles du signal cosmologique ?*

À partir de ces limitations documentées dans la littérature, nous formulons l'hypothèse de travail suivante : l'introduction d'informations structurelles supplémentaires peut améliorer la séparabilité et la robustesse des composantes. Cette hypothèse sera évaluée par comparaison au cadre GPR/ML-GPR de référence, en examinant en particulier si les développements proposés permettent de réduire les biais sur le spectre de puissance et la perte de signal, sans accroître l'attribution de contaminants à la composante cosmologique.

Dans ML-GPR, le VAE combine déjà un terme de reconstruction et une régularisation de l'espace latent fondée sur la divergence de Kullback–Leibler (Mertens et al. 2024). Le projet cherchera à déterminer si cette fonction objectif peut être complétée par des contraintes structurelles liées au problème de séparation, afin d'améliorer la robustesse de la représentation latente. L'objectif ne sera pas d'imposer *a priori* une méthode particulière, mais de déterminer quelles informations structurelles permettent effectivement d'améliorer la séparabilité et la robustesse du modèle.

Le projet étudiera comment des informations issues du traitement statistique du signal peuvent être introduites selon deux axes méthodologiques complémentaires, évalués dans un cadre probabiliste et bayésien commun :

- **Axe 1 – Modélisation structurée des composantes :** dans la représentation statistique des contaminants et des effets instrumentaux ;
- **Axe 2 – Régularisation de l'apprentissage :** au moyen de contraintes ou de régularisations guidées par le problème de séparation ;
- **Cadre transversal – Inférence bayésienne, quantification des incertitudes et évaluation des modèles :** pour l'inférence, la comparaison des hypothèses de modélisation et la quantification des incertitudes associées à la reconstruction.

Axe 1 – Modélisation structurée des composantes.

Une première direction concernera la modélisation statistique des avant-plans résiduels et des effets instrumentaux.

Les travaux de Mertens et al. (2025) suggèrent que la modélisation des contaminants par des fonctions de covariance génériques, *e.g.* issues de la famille de Matérn, peut limiter leur séparation avec le signal à 21 cm. Les analyses LOFAR font, par exemple, apparaître des avant-plans astrophysiques, des effets de *mode mixing* dont la cohérence dépend de la *baseline*, ainsi qu'une composante résiduelle d'excès.

Le projet examinera si les covariances de ces contaminants peuvent être décrites par des modèles plus représentatifs de leur variabilité que les fonctions génériques actuellement employées.

Une première piste consistera à étendre aux contaminants le principe de covariance apprise utilisé dans ML-GPR pour le signal à 21 cm. À partir de simulations décrivant diverses configurations d'avant-plans et d'effets instrumentaux, plusieurs modèles d'apprentissage pourront être étudiés afin d'apprendre une famille de covariances capable de représenter la diversité des avant-plans astrophysiques, des effets de *mode mixing* et de la composante résiduelle d'excès. Un modèle latent de type VAE constitue une possibilité, mais d'autres approches d'apprentissage pourront être envisagées si elles s'avèrent mieux adaptées à la structure des contaminants. Il s'agira de trouver un compromis entre la fidélité de la représentation des contaminants et la complexité du modèle, tout en maintenant un nombre de degrés de liberté maîtrisé. Le modèle appris pourrait ensuite être intégré au cadre ML-GPR pour évaluer si cette covariance apprise améliore la séparation par rapport aux kernels analytiques actuellement utilisés (Mertens et al. 2025).

Une seconde direction examinera l'intérêt d'exploiter conjointement des informations spectrales et spatiales.

Dans les modèles GPR actuels, la séparation repose principalement sur les différences de corrélation en fréquence. Lorsque les structures de corrélation de certaines composantes deviennent proches de celle du signal à 21 cm, cette information peut devenir insuffisante pour distinguer correctement les composantes (Acharya et al. 2024). La dépendance fréquentielle peut s'écrire, dans le cadre considéré ici (Mertens et al. 2018, *cf.* Sec. 2.2), sous la forme :

$$K_{pq} = \kappa(\nu_p, \nu_q), \text{ avec } p, q \in [\![1, N]\!],$$

où N est le nombre de canaux fréquentiels et ν_p la fréquence du p-ième canal.

Une extension spatio-spectrale consisterait à faire dépendre le modèle de covariance non seulement de la fréquence, mais aussi d'une information spatiale associée aux observations. Un tel modèle de covariance pourrait prendre en compte conjointement les corrélations spectrales et les corrélations spatiales dans le plan du ciel, contrairement au modèle actuel dans lequel les corrélations sont principalement décrites le long de la direction fréquentielle. Des structures séparables ou partiellement séparables entre dimensions spatiales et spectrales pourront par exemple être étudiées afin de conserver un modèle exploitable numériquement.

Cette direction répond à une limitation identifiée dans l'analyse LOFAR, où la prise en compte des corrélations spatiales est mentionnée comme un développement futur possible (Mertens et al. 2025).

Axe 2 – Régularisation de l'apprentissage du modèle de covariance.

Dans ML-GPR, le VAE apprend, à partir d'une famille de spectres de puissance simulés, une représentation latente utilisée pour construire une famille de covariances du signal à 21 cm, en combinant un terme de reconstruction et une régularisation fondée sur la divergence de Kullback–Leibler (Mertens et al. 2024). Cette dernière contraint la distribution latente apprise par rapport à une distribution *a priori* de référence, mais elle n'est pas spécifiquement conçue pour favoriser les propriétés utiles à la séparation du signal à 21 cm et des contaminants.

Le projet examinera si la fonction de coût du VAE peut être complétée par une régularisation guidée par le problème de séparation, que nous pouvons écrire schématiquement sous la forme :

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{VAE}} + \alpha \mathcal{R}_{\text{struct}},$$

où $\alpha \geqslant 0$ contrôle le poids de la régularisation structurelle $\mathcal{R}_{\text{struct}}$, dont la forme dépendra de la propriété que l'on cherche à favoriser.

Le poids α sera traité comme un hyperparamètre du modèle et sélectionné sur des simulations de validation indépendantes de celles utilisées pour l'apprentissage. Les valeurs testées seront comparées à la référence $\alpha = 0$ à partir de critères liés à la séparation, en particulier le biais sur le spectre de puissance, la perte de signal et l'attribution éventuelle de contaminants à la composante cosmologique.

Plusieurs formes de régularisation pourront être étudiées. Il pourra par exemple s'agir :

- de limiter la confusion entre les modèles de covariance du signal et des contaminants, *e.g.* dans le cas de la composante résiduelle d'excès (Acharya et al. 2024, *cf.* Sec. 2.4), dont la structure statistique peut devenir proche de celle du signal à 21 cm ;
- de contrôler la sensibilité de la covariance obtenue aux variations de la représentation latente *i.e.*, éviter qu'une faible variation dans l'espace latent n'entraîne une modification importante de la covariance obtenue. Cette propriété pourrait contribuer à rendre le modèle moins sensible aux écarts entre les simulations utilisées pour l'apprentissage et les données auxquelles il sera ensuite appliqué ;
- d'adapter le critère de reconstruction afin de mieux préserver les structures statistiques pertinentes pour l'estimation du spectre de puissance, *e.g.* en accordant une importance différente à certaines échelles de k selon leur pertinence pour l'analyse.

Ces pistes permettront d'étudier si l'objectif d'apprentissage doit seulement être de reproduire fidèlement les structures de covariance associées aux simulations, ou s'il peut aussi intégrer des contraintes liées à leur utilisation ultérieure pour la séparation et l'estimation du spectre de puissance.

Cadre transversal – Inférence bayésienne, quantification des incertitudes et évaluation des modèles.

Le cadre GPR est naturellement probabiliste. Les composantes latentes sont décrites par des processus gaussiens et leur distribution *a posteriori* est obtenue par conditionnement sur les observations (Mertens et al. 2018).

De même, les hyperparamètres des modèles de covariance peuvent être munis de distributions *a priori*, puis estimés à partir de leur distribution *a posteriori*. Dans ML-GPR, cette inférence est étendue aux paramètres latents du modèle appris de covariance, et les échantillons de la distribution *a posteriori* peuvent être propagés jusqu'à l'estimation du spectre de puissance (Mertens et al. 2024).

Le projet cherchera à exploiter explicitement cette articulation entre information structurelle et inférence bayésienne. Lorsque cela sera possible, les contraintes introduites pour améliorer la séparation pourront être étudiées non seulement comme des termes de régularisation, mais aussi à travers leur interprétation probabiliste sous forme d'information *a priori* portant sur les paramètres, les représentations latentes ou les structures de covariance. Cette double lecture permettra de relier des outils de régularisation issus du traitement du signal au cadre probabiliste de ML-GPR, tout en conservant une quantification explicite des incertitudes.

L'introduction de nouvelles structures ou régularisations fera apparaître plusieurs hypothèses de modélisation possibles. Leur évaluation devra porter à la fois sur leur capacité à expliquer les observations et sur la qualité effective de la séparation.

L'évidence bayésienne pourra servir à comparer plusieurs modèles après marginalisation de leurs paramètres, et pour évaluer le support que les données leur apportent (Liu et al. 2026). Cette information ne suffit cependant pas à elle seule. Des travaux récents montrent qu'un modèle bénéficiant d'une évidence globale plus élevée peut pourtant produire une reconstruction moins fidèle du signal à 21 cm (Liu et al. 2026).

L'évaluation bayésienne reposera donc sur deux niveaux complémentaires :

- la comparaison bayésienne des hypothèses de modélisation ;
- l'analyse des distributions prédictives et des incertitudes associées à la reconstruction.

Le cadre bayésien servira donc à la fois à l'inférence, à la propagation des incertitudes et à l'évaluation des modèles, et pas seulement de critère de sélection : il s'agira de mieux caractériser les conditions dans lesquelles leur utilisation reste fiable.

Ce cadre fera le lien entre les développements méthodologiques des deux axes et leur évaluation probabiliste, la comparaison des modèles et la quantification des incertitudes.

1.2 Stratégie d'évaluation et exploitation des développements

Les développements issus des deux axes méthodologiques seront évalués selon une progression commune, depuis des configurations simulées permettant de disposer d'une vérité terrain jusqu'aux observations réelles, avant d'examiner les conséquences des erreurs résiduelles sur l'exploitation scientifique du spectre de puissance.

évaluation de la robustesse sur simulations.

Les extensions proposées seront d'abord étudiées sur des données simulées permettant de disposer d'une vérité terrain et de contrôler séparément les principales sources de difficulté.

Les expériences feront varier le rapport signal sur bruit, le degré de *covariance mismatch*, la proximité entre les structures statistiques du signal et celles des contaminants, ainsi que l'écart entre les simulations utilisées pour l'apprentissage et celles utilisées pour l'évaluation. Les méthodes proposées seront comparées au cadre GPR/ML-GPR de référence.

L'évaluation portera, par comparaison au signal de référence connu dans les simulations, sur la qualité de séparation, la perte de signal (*signal loss*), les biais sur le spectre de puissance, l'incertitude associée à son estimation et la robustesse du modèle face aux écarts entre les structures de covariance supposées et celles effectivement présentes dans les données.

Dans le cas de l'extension spatio-spectrale, plusieurs configurations de dépendance spatiale et spectrale pourront être considérées. La comparaison avec le modèle purement spectral pourra être menée dans le plan $(k_\perp, k_\parallel)$ afin d'identifier les domaines dans lesquels la prise en compte des corrélations spatiales améliore ou dégrade la séparation. Des tests d'injection permettront d'évaluer la perte de signal et la capacité de la méthode à préserver une composante cosmologique injectée (Mertens et al. 2025). Des configurations sans signal cosmologique permettront parallèlement de vérifier si le modèle attribue à tort certaines structures résiduelles à la composante à 21 cm (Liu et al. 2026).

évaluation de la fiabilité sur observations réelles.

Dans un second temps, les méthodes développées pourront être évaluées sur les données du programme NenuFAR Cosmic Dawn, consacré à l'étude de l'aube cosmologique, ou du programme LOFAR-EoR, consacré à l'époque de réionisation.

Ce passage introduit une difficulté fondamentale : contrairement aux simulations, le véritable signal cosmologique présent dans les données n'est pas connu.

Les tests d'injection employés dans les analyses LOFAR fournissent un premier moyen d'évaluer la perte de signal dans cette situation (Mertens et al. 2025).

Le projet cherchera plus généralement à déterminer si les diagnostics construits sur simulations et les incertitudes issues du modèle probabiliste peuvent fournir une information exploitable sur la fiabilité de la séparation sur données réelles. Cette fiabilité pourra être étudiée en fonction du redshift et des échelles caractérisées par $k_\perp$ dans le plan du ciel et par $k_\parallel$ le long de la ligne de visée, puis après moyenne sphérique en fonction de k. L'objectif sera d'identifier les domaines dans lesquels l'estimation du spectre de puissance peut être considérée comme suffisamment contrôlée pour son exploitation scientifique (Giri et Mertens 2026).

Exploitation du spectre de puissance pour l'inférence astrophysique.

L'un des objectifs de la séparation est de préserver les propriétés statistiques du signal à 21 cm nécessaires à son exploitation scientifique, en particulier à l'estimation de son spectre de puissance. Ces propriétés statistiques seront caractérisées par le spectre de puissance sphériquement moyenné sans dimension :

$$\Delta^2_{21}(k) := \frac{k^3}{2\pi^2} P_{21}(k),$$

où k désigne le nombre d'onde associé à l'échelle considérée (Mertens et al. 2018).

L'évaluation des méthodes développées portera en priorité sur leur capacité à limiter les biais sur $\Delta^2_{21}(k)$ et à en quantifier l'incertitude pour chaque intervalle de redshift considéré.

Les travaux récents montrent par ailleurs que les erreurs résiduelles dans la séparation des composantes peuvent se propager jusqu'à l'inférence astrophysique et que quelques intervalles de redshift mal reconstruits peuvent dégrader une analyse multi-redshift (Giri et Mertens 2026). La propagation des incertitudes et des biais de séparation jusqu'aux paramètres astrophysiques pourra constituer une dernière étape de validation permettant d'évaluer l'importance scientifique des erreurs résiduelles.

Cette étape restera cependant en aval du projet : le cœur méthodologique demeure la modélisation statistique et la séparation robuste du signal à 21 cm.

1.3 Résultats attendus et portée du projet

Le projet cherchera à déterminer si des informations spatio-spectrales, des contraintes adaptées à la séparabilité ou des régularisations de l'apprentissage permettent d'améliorer la robustesse de la récupération du signal à 21 cm face aux erreurs de modèle.

L'intérêt de ces développements sera évalué depuis les simulations, où la vérité terrain permet de caractériser précisément leurs biais, jusqu'aux observations réelles, où la quantification de l'incertitude et l'évaluation de la fiabilité de la séparation deviennent essentielles.

À terme, le projet vise à fournir des modèles de covariance appris pour les contaminants, des stratégies de régularisation évaluées selon des critères de séparation, ainsi que des diagnostics permettant de mieux caractériser la fiabilité de la séparation en l'absence de vérité terrain. L'objectif final est une estimation plus robuste du spectre de puissance du signal à 21 cm et une meilleure maîtrise des incertitudes associées à son exploitation scientifique.

2 Contexte scientifique et état de l'art

Cette section présente les éléments scientifiques ayant conduit à la formulation du projet de recherche, depuis le problème de séparation du signal à 21 cm jusqu'aux développements récents du GPR et de ML-GPR, à leur validation et aux limitations qui motivent les axes de recherche proposés.

2.1 Contexte du projet

Le signal à 21 cm de l'hydrogène neutre constitue un moyen d'étudier les premières phases de l'histoire de l'Univers, en particulier l'aube cosmologique (*Cosmic Dawn*) et l'époque de réionisation (*Epoch of Reionization*, EoR). Durant ces périodes, les premières sources lumineuses modifient progressivement l'état thermique et d'ionisation du milieu intergalactique (*Intergalactic Medium*, IGM), ce qui produit des fluctuations spatiales du signal à 21 cm (Liu et Shaw 2020, Sec. 2.3).

Le signal est décrit par une *température de brillance* $T_b(\hat{\boldsymbol{r}}, \nu)$, où $\hat{\boldsymbol{r}}$ désigne une direction sur le ciel et ν la fréquence d'observation. Cette température exprime le contraste du signal par rapport au fond diffus cosmologique (*Cosmic Microwave Background*, CMB). Son signe et son amplitude dépendent en particulier de l'état thermique et de la fraction d'hydrogène neutre du milieu ; le signal peut apparaître en absorption ou en émission selon les conditions physiques considérées (Liu et Shaw 2020, Sec. 2.1–2.3).

L'expansion de l'Univers décale la raie de 21 cm vers des fréquences plus faibles que sa fréquence au repos $\nu_{21} \approx 1420$ MHz. Ce décalage spectral est caractérisé par le *redshift* z, relié à la fréquence observée ν par

$$1 + z = \frac{\nu_{21}}{\nu} \tag{1}$$

Un redshift plus élevé correspond à une fréquence observée plus faible et à une époque plus ancienne de l'Univers. Pour la raie de 21 cm, le redshift fournit aussi une information sur la distance le long de la ligne de visée. Les coordonnées angulaires sur le ciel et la fréquence permettent donc de sonder un champ tridimensionnel dont l'axe fréquentiel renseigne également sur l'évolution cosmique (Liu et Shaw 2020, Sec. 2.1).

Or, les observations ne donnent pas directement accès à ce champ tridimensionnel. Il faut relier la température de brillance $T_b(\hat{\boldsymbol{r}}, \nu)$ aux quantités effectivement mesurées par l'instrument. En effet, un radio-interféromètre n'observe pas directement $T_b(\hat{\boldsymbol{r}}, \nu)$; pour une paire d'antennes séparées par un vecteur de *baseline* $\boldsymbol{b}$, la donnée mesurée est une visibilité (*visibility*) $V(\boldsymbol{b}, \nu)$. Elle dépend de la température de brillance du ciel $T_b(\hat{\boldsymbol{r}}, \nu)$, de la fréquence d'observation ν et de la réponse directionnelle de l'instrument, décrite par le faisceau primaire de puissance $A_p(\hat{\boldsymbol{r}}, \nu)$ (*primary beam*). Dans l'approximation de *ciel plat*, la direction $\hat{\boldsymbol{r}}$ repère une position angulaire dans le plan du ciel et peut être décrite localement par les coordonnées $\boldsymbol{\theta} = (\theta_x, \theta_y)$. Nous introduisons la coordonnée $\boldsymbol{u}$ comme la coordonnée de Fourier *conjuguée* de $\boldsymbol{\theta}$ (Liu et Shaw 2020, Sec. 3.1). En introduisant les coordonnées interférométriques :

$$\boldsymbol{u} = \frac{\boldsymbol{b}}{\lambda}, \text{ avec } \lambda = \frac{c}{\nu}, \tag{2}$$

la séparation physique $\boldsymbol{b}$ entre deux antennes est exprimée en unités de longueur d'onde. Pour une fréquence donnée, la coordonnée $\boldsymbol{u}$ repère une position dans l'espace de Fourier du plan du ciel. La visibilité mesurée par une *baseline* peut être interprétée, approximativement, comme une mesure d'une composante spatiale du ciel à une certaine échelle spatiale et une certaine orientation dans le domaine de Fourier. Chaque *baseline* échantillonne une composante spatiale distincte du ciel (Liu et Shaw 2020, Sec. 3.1). En résumé, $\hat{\boldsymbol{r}}$ repère une direction dans la représentation angulaire du ciel, tandis que $\boldsymbol{u}$ caractérise les modes spatiaux échantillonnés par l'interféromètre dans le domaine de Fourier.

Après discrétisation du ciel et des mesures, le processus d'observation peut être résumé par

$$\boldsymbol{v} = \mathbf{B}\boldsymbol{s} + \boldsymbol{n} \tag{3}$$

où $\boldsymbol{s}$ représente le ciel discrétisé, $\boldsymbol{v}$ le vecteur des visibilités complexes mesurées, $\mathbf{B}$ la réponse instrumentale et $\boldsymbol{n}$ le bruit. La reconstruction de $\boldsymbol{s}$ à partir de $\boldsymbol{v}$ constitue un problème d'*inversion statistique* : certains modes du ciel sont peu, ou pas, observés et l'inversion doit tenir compte du bruit en intégrant, si nécessaire, une régularisation (Liu et Shaw 2020, Sec. 10.1). À cette difficulté instrumentale s'ajoutent les émissions astrophysiques d'avant-plan (*foregrounds*), beaucoup plus intenses que le signal cosmologique recherché. Aux fréquences considérées, elles sont dominées par l'émission radio de notre Galaxie et par des sources radio extragalactiques (Liu et Shaw 2020, Sec. 7). Leur amplitude peut dépasser de plusieurs ordres de grandeur celle du signal EoR. La séparation reste néanmoins envisageable, car les avant-plans intrinsèques sont généralement plus réguliers en fréquence que le signal à 21 cm. Des effets instrumentaux et des erreurs de calibration peuvent toutefois ajouter des structures spectrales et rendre cette distinction moins nette. Le problème de séparation dépend, à la fois, des propriétés statistiques des contributions et de la manière dont la chaîne d'observation les modifie.

Une difficulté supplémentaire est le *signal loss*. Lors de la soustraction des avant-plans, une partie du signal cosmologique peut être absorbée par le modèle des contaminants et être supprimée avec eux. Le spectre de puissance reconstruit est alors biaisé vers des valeurs trop faibles. Ce biais est particulièrement problématique lorsque l'analyse ne permet pas une détection significative et conduit seulement à établir une limite supérieure sur la puissance du signal à 21 cm : si une partie du signal a été supprimée lors de la soustraction, cette limite peut elle-même être artificiellement basse. Liu et Shaw (2020) soulignent la nécessité de quantifier cette perte de signal et, lorsque cela est possible, de la corriger. Cette quantification peut reposer sur des simulations ou sur l'injection de signaux cosmologiques synthétiques dans des données réelles (Liu et Shaw 2020, Sec. 12.5).

La récupération du signal à 21 cm s'inscrit dans une chaîne d'analyse comprenant plusieurs étapes :

- la *calibration* corrige les effets de la réponse instrumentale sur les mesures ;
- le *map making* reconstruit, lorsqu'il est utilisé, une représentation du ciel à partir des visibilités, ce qui renvoie au problème d'inversion défini par l'équation (3) ;
- la *foreground mitigation* vise à limiter ou séparer la contamination des avant-plans.

Les données traitées servent ensuite à estimer le spectre de puissance du signal à 21 cm, dont l'évolution avec l'échelle et le redshift z permet de contraindre des modèles décrivant

en particulier l'état thermique et l'ionisation du milieu intergalactique, ainsi que l'influence des premières sources lumineuses.

Dans la suite, l'attention est portée sur le problème de séparation : les différentes contributions sont superposées dans les données et ne sont pas observées séparément. Leur estimation nécessite des hypothèses statistiques permettant de les distinguer. C'est dans ce cadre que s'inscrit l'utilisation de la régression par processus gaussiens.

2.2 Séparation statistique par processus gaussiens

Parmi les méthodes non paramétriques de séparation de composantes, Mertens et al. (2018) proposent d'exploiter les différences de corrélation spectrale entre les contributions au moyen de la *Gaussian Process Regression* (GPR). Le terme de *régression* désigne ici l'inférence d'une fonction *latente* à partir d'observations bruitées, tandis que le processus gaussien fournit un modèle probabiliste *a priori* sur les fonctions susceptibles de décrire cette composante.

La formulation générale du problème de régression s'écrit :

$$\boldsymbol{d} = \boldsymbol{f} + \boldsymbol{n}, \tag{4}$$

où $\boldsymbol{d}$ regroupe les N observations acquises aux points $\boldsymbol{x} = (x_1, \dots, x_N)$ et $\boldsymbol{n}$ représente le bruit de mesure. Dans l'application considérée ici, les x_i correspondent aux fréquences observées ν_i et N est le nombre de canaux fréquentiels considérés. Le vecteur $\boldsymbol{f}$ contient les valeurs de la fonction inconnue $f(\cdot)$ évaluées en $\boldsymbol{x}$:

$$\boldsymbol{f} = (f(x_1), \dots, f(x_N))^\top. \tag{5}$$

La fonction $f(\cdot)$ est dite latente, car ses valeurs $f(x_i)$ ne sont pas directement observées : seules leurs observations bruitées $\boldsymbol{d}$ sont accessibles. La régression consiste à inférer ces valeurs à partir des données, ainsi que l'incertitude associée à leur estimation.

Pour décrire les fonctions cohérentes avec les hypothèses formulées avant l'observation des données, les auteurs supposent que la fonction latente $f(\cdot)$ est modélisée *a priori* par un processus gaussien :

$$f \sim \mathcal{GP}(m, \kappa), \tag{6}$$

où $m(x)$ est sa fonction moyenne et $\kappa(x, x')$ sa fonction de covariance, ou *kernel*. Le processus gaussien $\mathcal{GP}(\cdot)$ ne fixe pas une forme paramétrique unique pour f : il définit une distribution probabiliste sur les fonctions possibles à travers leur moyenne et leur structure de covariance.

Pour l'ensemble fini de points $\boldsymbol{x}$, le vecteur $\boldsymbol{f}$ défini par (5) suit une loi normale multivariée :

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, \mathbf{K}) \text{ où } \boldsymbol{f}, \boldsymbol{m} \in \mathbb{R}^N, \quad \mathbf{K} \in \mathbb{R}^{N \times N}. \tag{7}$$

Conformément au modèle de processus gaussien (6), le vecteur moyen $\boldsymbol{m}$ est obtenu en évaluant la fonction moyenne m aux points $\boldsymbol{x}$, soit $m_p = m(x_p)$, et la matrice de covariance $\mathbf{K}$ est construite à partir du kernel κ, avec $K_{pq} = \kappa(x_p, x_q)$ pour tout $p, q \in [\![1, N]\!]$. La fonction de covariance encode les hypothèses *a priori* sur les structures que peut présenter la fonction latente (Mertens et al. 2018, Sec. 2.1).

En supposant un bruit gaussien de variance σ_n^2, les observations $\boldsymbol{d}$ et les valeurs latentes à prédire sont décrites conjointement par une loi normale multivariée. Pour un ensemble de points de prédiction $\boldsymbol{x}^\star$, nous notons $\boldsymbol{f}^\star = f(\boldsymbol{x}^\star)$ les valeurs latentes correspondantes. Le conditionnement de la loi jointe par les observations conduit à une loi *a posteriori* également gaussienne :

$$\boldsymbol{f}^\star \mid \boldsymbol{d} \sim \mathcal{N}(\boldsymbol{m}^\star, \mathbf{K}^\star)$$

La moyenne $\boldsymbol{m}^\star$ fournit une estimation des valeurs latentes, et la covariance $\mathbf{K}^\star$ quantifie l'incertitude associée à cette estimation.

La fonction de covariance κ du modèle général (6) détermine les structures que le processus gaussien est capable de représenter. Évaluée aux points d'observation, elle définit la matrice de covariance $\mathbf{K}$ de l'équation (7). Les auteurs utilisent la fonction de covariance de Matérn, où la longueur de cohérence l contrôle l'échelle sur laquelle les valeurs restent corrélées, tandis que le paramètre η contrôle la forme du kernel et, par conséquent, la régularité des fonctions décrites par le processus gaussien. Par exemple, le cas $\eta = 1/2$ correspond à un kernel exponentiel :

$$\kappa_{\eta=1/2}(x_p, x_q) = \exp\left(-\frac{|x_p - x_q|}{l}\right),$$

qui, dans l'application considérée ici, décrit une décroissance exponentielle de la covariance avec la séparation en fréquence $|\nu_p - \nu_q|$. Le choix d'une famille analytique telle que Matérn impose *a priori* une forme particulière à la matrice de covariance $\mathbf{K}$ et restreint les structures que le modèle peut représenter.

Le choix de la forme du kernel ne suffit cependant pas : pour une forme donnée, ses hyperparamètres doivent aussi être déterminés. Dans la suite, $\boldsymbol{\psi}$ désigne ces hyperparamètres, tels que les longueurs de cohérence et les variances des fonctions de covariance. Pour un ensemble de fréquences d'observation fixé, les hyperparamètres déterminent la matrice de covariance, que l'on peut écrire $\mathbf{K} \equiv \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}; \boldsymbol{\psi})$.

Il reste à déterminer les valeurs de $\boldsymbol{\psi}$ qui rendent les données observées $\boldsymbol{d}$ les plus vraisemblables. Les valeurs de la fonction latente $\boldsymbol{f}$ n'étant pas observées, elles sont marginalisées selon leur distribution *a priori*. Pour cela, les auteurs utilisent la vraisemblance marginale, ou *evidence*, définie par :

$$p(\boldsymbol{d} \mid \boldsymbol{x}, \boldsymbol{\psi}) = \int p(\boldsymbol{d} \mid \boldsymbol{f}, \boldsymbol{x}, \boldsymbol{\psi}) p(\boldsymbol{f} \mid \boldsymbol{x}, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{f},$$

où $p(\boldsymbol{f} \mid \boldsymbol{x}, \boldsymbol{\psi})$ représente le prior gaussien sur la fonction latente et $p(\boldsymbol{d} \mid \boldsymbol{f}, \boldsymbol{x}, \boldsymbol{\psi})$ le modèle d'observation *i.e.*, la vraisemblance des données conditionnellement à la fonction latente.

En supposant un bruit gaussien indépendant de covariance $\sigma_n^2 \mathbf{I}$, la marginalisation sur les valeurs latentes $\boldsymbol{f}$ peut être obtenue analytiquement. Dans le cas du processus gaussien (7) de moyenne nulle $\boldsymbol{m} = \mathbf{0}_N$, la log-vraisemblance marginale s'écrit

$$\begin{aligned} \log p(\boldsymbol{d} \mid \boldsymbol{x}, \boldsymbol{\psi}) = &-\frac{1}{2} \boldsymbol{d}^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{d} \\ &- \frac{1}{2} \log \left|\mathbf{K} + \sigma_n^2 \mathbf{I}\right| - \frac{N}{2} \log(2\pi). \end{aligned} \tag{8}$$

Pour des points d'observation $\boldsymbol{x}$ fixés, cette quantité dépend des hyperparamètres $\boldsymbol{\psi}$ à travers la matrice de covariance $\mathbf{K}$. Les valeurs de $\boldsymbol{\psi}$ peuvent alors être déterminées en maximisant la log-vraisemblance marginale.

Une information préalable peut également être introduite sur ces hyperparamètres à travers une distribution *a priori* $p(\boldsymbol{\psi})$. En appliquant le théorème de Bayes, leur distribution *a posteriori* s'écrit alors :

$$\log p(\boldsymbol{\psi} \mid \boldsymbol{d}, \boldsymbol{x}) \propto \log p(\boldsymbol{d} \mid \boldsymbol{x}, \boldsymbol{\psi}) + \log p(\boldsymbol{\psi}), \tag{9}$$

et les hyperparamètres peuvent être estimés par maximum *a posteriori* (MAP). La forme analytique des kernels permet en outre de calculer les dérivées de la vraisemblance marginale par rapport aux hyperparamètres, ce qui facilite leur optimisation.

Ce cadre décrit jusqu'ici la régression d'une seule fonction latente. Dans le problème du signal à 21 cm, la variable d'entrée x correspond à la fréquence ν, et le modèle est étendu à plusieurs contributions latentes superposées, auxquelles sont associées des structures de covariance distinctes.

Application à la séparation du signal à 21 cm.

Pour N canaux de fréquence, les données $\boldsymbol{d}$ du modèle (4) sont décrites par

$$\boldsymbol{d} = \boldsymbol{f}_{\rm fg} + \boldsymbol{f}_{21} + \boldsymbol{n}, \text{ où } \boldsymbol{d}, \boldsymbol{f}_{\rm fg}, \boldsymbol{f}_{21}, \boldsymbol{n} \in \mathbb{R}^N. \tag{10}$$

Les vecteurs latents $\boldsymbol{f}_{\rm fg}$ et $\boldsymbol{f}_{21}$ regroupent respectivement les contributions des avant-plans et du signal cosmologique à 21 cm aux N fréquences observées. Ces deux composantes ne sont pas observées séparément : les données $\boldsymbol{d}$ correspondent à leur mélange, auquel s'ajoute le bruit $\boldsymbol{n}$.

La séparation repose sur des différences de corrélation en fréquence. Dans leur modèle, les avant-plans astrophysiques intrinsèques sont supposés plus réguliers spectralement que le signal à 21 cm. Des effets instrumentaux et de calibration peuvent introduire une composante spectrale moins régulière, dite de *mode mixing*, tandis que le bruit thermique est modélisé comme non corrélé entre canaux. Ces différences de corrélation sont traduites par des fonctions de covariance propres à chaque composante. Dans leur formalisme, la covariance des composantes latentes s'écrit

$$\mathbf{K} = \mathbf{K}_{\rm fg} + \mathbf{K}_{21} \text{ avec } \mathbf{K}_{\rm fg}, \mathbf{K}_{21} \in \mathbb{R}^{N \times N}.$$

L'introduction de $\mathbf{K}_{21}$ vise principalement à isoler la contribution du signal à 21 cm dans la covariance totale, afin qu'elle ne soit pas attribuée aux avant-plans.

Dans l'application aux données simulées, la covariance des avant-plans est elle-même décomposée en

$$\mathbf{K}_{\rm fg} = \mathbf{K}_{\rm sky} + \mathbf{K}_{\rm mix},$$

où $\mathbf{K}_{\rm sky}$ décrit la contribution astrophysique du ciel et $\mathbf{K}_{\rm mix}$ celle associée au *mode mixing*. Le bruit est représenté séparément par $\mathbf{K}_n = \sigma_n^2 \mathbf{I}$. La covariance totale des données s'écrit :

$$\mathbf{K}_d = \mathbf{K} + \mathbf{K}_n = \mathbf{K}_{\rm fg} + \mathbf{K}_{21} + \mathbf{K}_n. \tag{11}$$

Toutes ces matrices étant évaluées sur le même ensemble de fréquences ν, les auteurs utilisent la notation abrégée $\mathbf{K} \equiv \mathbf{K}(\nu, \nu; \boldsymbol{\psi})$.

Sous l'hypothèse d'un prior gaussien de moyenne nulle, la composante d'avant-plan est entièrement caractérisée par sa fonction de covariance. Les observations $\boldsymbol{d}$ et la composante latente d'avant-plan $\boldsymbol{f}_{\rm fg}$, évaluées aux mêmes fréquences ν, suivent conjointement la loi normale multivariée

$$\begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{f}_{\rm fg} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\rm fg} + \mathbf{K}_{21} + \mathbf{K}_n & \mathbf{K}_{\rm fg} \\ \mathbf{K}_{\rm fg} & \mathbf{K}_{\rm fg} \end{bmatrix} \right).$$

Cette loi jointe décrit le modèle avant conditionnement par les observations. La covariance de $\boldsymbol{d}$ correspond à la covariance totale définie par l'équation (11), tandis que les covariances croisées sont données par $\mathbf{K}_{\rm fg}$, puisque $\boldsymbol{f}_{\rm fg}$ est précisément la contribution d'avant-plan présente dans les données.

En conditionnant cette loi jointe par les observations $\boldsymbol{d}$, nous obtenons la distribution *a posteriori* de la composante d'avant-plan $\boldsymbol{f}_{\rm fg}$. Sa moyenne et sa covariance sont données par :

$$\begin{aligned} \mathbb{E}[\boldsymbol{f}_{\rm fg} \mid \boldsymbol{d}] &= \mathbf{K}_{\rm fg} \mathbf{K}_d^{-1} \boldsymbol{d}, \\ \operatorname{cov}(\boldsymbol{f}_{\rm fg} \mid \boldsymbol{d}) &= \mathbf{K}_{\rm fg} - \mathbf{K}_{\rm fg} \mathbf{K}_d^{-1} \mathbf{K}_{\rm fg}. \end{aligned} \tag{12}$$

Le même conditionnement s'applique à toute composante latente décrite par sa propre fonction de covariance. Pour le signal cosmologique à 21 cm, nous obtenons :

$$\begin{aligned} \mathbb{E}[\boldsymbol{f}_{21} \mid \boldsymbol{d}] &= \mathbf{K}_{21} \mathbf{K}_d^{-1} \boldsymbol{d}, \\ \operatorname{cov}(\boldsymbol{f}_{21} \mid \boldsymbol{d}) &= \mathbf{K}_{21} - \mathbf{K}_{21} \mathbf{K}_d^{-1} \mathbf{K}_{21}. \end{aligned} \tag{13}$$

Ces expressions seront reprises dans les développements ultérieurs de la méthode pour l'estimation de la composante à 21 cm (Mertens et al. 2024).

Finalement, le choix d'un kernel propre à chaque contribution permet, à travers la covariance totale $\mathbf{K}_d$ (11), de répartir statistiquement les structures observées entre les composantes latentes et d'obtenir pour chacune une estimation *a posteriori*. Les auteurs exploitent cette séparation statistique pour estimer la contribution des avant-plans, dont la moyenne *a posteriori* est ensuite soustraite aux observations afin de construire les données résiduelles $\boldsymbol{d}_{\rm res}$:

$$\boldsymbol{d}_{\rm res} = \boldsymbol{d} - \mathbb{E}[\boldsymbol{f}_{\rm fg} \mid \boldsymbol{d}]. \tag{14}$$

Des visibilités résiduelles au spectre de puissance.

En reprenant la représentation des visibilités $V(\boldsymbol{u}, \nu)$ introduite en section 2.1, où $\boldsymbol{u}$ désigne la coordonnée de Fourier associée à une *baseline* et ν la fréquence (*cf.* Eq.(2)), le résidu $\boldsymbol{d}_{\rm res}$ (14) correspond aux visibilités restantes après soustraction de la contribution d'avant-plan estimée. Ces visibilités résiduelles forment le cube :

$$V_{\rm res}(\boldsymbol{u}, \nu) = V_{\rm obs}(\boldsymbol{u}, \nu) - V_{\rm fg}^{\rm rec}(\boldsymbol{u}, \nu),$$

où $V_{\rm obs}$ désigne les visibilités observées (*cf.* Mertens et al. 2018, Eq. (16)), $V_{\rm fg}^{\rm rec}$ la contribution d'avant-plan reconstruite par le GPR, et $V_{\rm res}$ les visibilités résiduelles obtenues après leur soustraction.

Le spectre de puissance (Liu et Shaw 2020, Sec. 2.2) est ensuite estimé à partir de ces visibilités résiduelles. La description en modes de Fourier introduite en section 2.1 est ici étendue aux trois dimensions du signal. Les auteurs distinguent deux composantes selon leur orientation par rapport à la ligne de visée *i.e.*, la direction d'observation :

— l'échelle transverse, caractérisée par $k_\perp$, correspond aux variations du signal perpendiculairement à la ligne de visée, dans le plan du ciel, et est reliée aux coordonnées $\boldsymbol{u}$ associées aux *baselines* ;

— l'échelle le long de la ligne de visée, caractérisée par $k_\parallel$, correspond aux variations du signal dans la direction de profondeur et est obtenue à partir de la transformation de Fourier selon la direction fréquentielle.

Cette décomposition est illustrée dans le plan $(k_\perp, k_\parallel)$ (Mertens et al. 2018, Fig. 1). Dans ce plan, une part importante de la contamination par les avant-plans, notamment celle associée aux effets de *mode mixing*, se concentre dans une région caractéristique en forme de coin appelée *foreground wedge*. La région moins affectée par ces contaminants est désignée comme l'*EoR window*.

Ces deux composantes sont ensuite regroupées pour définir le nombre d'onde total :

$$k = \sqrt{k_\perp^2 + k_\parallel^2} \tag{15}$$

Le nombre d'onde k caractérise l'échelle spatiale du mode considéré : les faibles valeurs de k sont associées aux grandes échelles spatiales, tandis que les valeurs élevées correspondent à des échelles plus fines (Liu et Shaw 2020, Sec. 3.3, Figs. 5–6). Il est ensuite utilisé pour construire le spectre de puissance sphériquement moyenné $P(k)$. La quantité sans dimension utilisée pour présenter les résultats est :

$$\Delta^2(k) = \frac{k^3}{2\pi^2} P(k).$$

Dans ce cadre, l'objectif de la séparation n'est pas nécessairement de reconstruire exactement une réalisation particulière du signal, mais d'en préserver les propriétés statistiques utiles à l'estimation du spectre de puissance. La qualité des modèles de covariance est donc centrale : si les structures attribuées aux composantes sont trop proches ou mal adaptées aux données, leur séparation devient ambiguë et l'estimation finale peut être biaisée.

Cette dépendance au modèle de covariance motive les développements ultérieurs de la méthode.

2.3 Vers des modèles de covariance appris

La méthode GPR (section 2.2) sépare les composantes en leur associant des structures de covariance distinctes. Son efficacité dépend directement de la pertinence de ces modèles de covariance.

Dans le GPR standard, la forme du kernel est choisie *a priori* parmi des familles analytiques, et leurs hyperparamètres sont estimés à partir des données. Mertens et al. (2024) proposent de conserver le cadre probabiliste du GPR tout en remplaçant, pour le signal à 21 cm, cette paramétrisation analytique par un modèle de covariance *appris* à partir de simulations. Cette évolution soulève deux questions : pourquoi les kernels analytiques peuvent-ils être insuffisants, et que permet d'apprendre une covariance à partir de simulations ? Nous traitons les deux questions dans la suite de cette section à partir des éléments présentés par les auteurs.

Pourquoi le kernel analytique du GPR standard est-il insuffisant ?

Le modèle GPR repose sur la covariance totale de l'équation (11). Comme dans l'équation (9), les hyperparamètres des kernels interviennent dans l'inférence à travers les matrices de covariance qu'ils paramètrent.

L'ajustement des hyperparamètres permet d'adapter une covariance au jeu de données considéré, mais uniquement à l'intérieur de la famille de kernels choisie *a priori*. Modifier $\boldsymbol{\psi}$ ne change pas cette famille elle-même. Pour la composante à 21 cm, une covariance générique, par exemple issue de la famille de Matérn, peut ne pas représenter correctement la covariance réelle du signal. Les auteurs soulignent qu'une telle inadéquation peut introduire des biais, particulièrement lorsque le rapport signal sur bruit est faible (Mertens et al. 2024, Sec. 2.1). Le problème ne consiste plus seulement à déterminer les *bonnes* valeurs des hyperparamètres d'un kernel analytique donné, mais aussi à disposer d'un modèle suffisamment flexible pour représenter la diversité des covariances *fréquence–fréquence* que peut présenter le signal à 21 cm.

Quelle information supplémentaire l'apprentissage apporte-t-il ?

Les auteurs proposent d'utiliser des simulations du signal à 21 cm pour construire un modèle de substitution (*surrogate model*) de sa covariance. En faisant varier les paramètres physiques du modèle de simulation, ils obtiennent une famille de covariances correspondant aux signaux simulés. Le modèle appris vise à représenter cette famille de covariances et à être utilisable directement au cours de l'inférence, sans avoir à recalculer une covariance à partir d'une nouvelle simulation physique.

Dans le cas étudié, les simulations sont décrites par un petit nombre de paramètres physiques. Les covariances qu'elles produisent sont donc supposées appartenir à une structure de dimension bien plus faible que leur représentation complète. Lorsqu'elles proviennent d'un modèle paramétrique, la dimension de cette structure correspond au nombre de paramètres libres du modèle. La dépendance des covariances aux paramètres physiques étant toutefois non linéaire, une réduction linéaire telle que l'analyse en composantes principales (ACP) ne fournit pas de résultats satisfaisants dans les tests des auteurs. Les auteurs se tournent donc vers des *auto-encodeurs* pour apprendre une représentation non linéaire de faible dimension de cette famille de covariances (Mertens et al. 2024, Sec. 2.2).

Un auto-encodeur comporte :

— un encodeur, qui projette les exemples d'apprentissage dans un espace latent de faible dimension ;

— un décodeur, qui reconstruit un objet dans l'espace d'origine à partir de cette représentation.

L'encodeur et le décodeur sont appris pendant la phase d'entraînement afin de reproduire au mieux les exemples du jeu d'apprentissage.

Dans ces travaux, deux constructions sont étudiées : un *Variational Auto-Encoder* (VAE), qui fournit une description probabiliste dans l'espace latent, et un *Interpolatory Auto-Encoder* (IAE), qui construit cette représentation à partir d'interpolations entre des points de référence dans l'espace latent.

L'espace latent désigne ici un espace de représentation de faible dimension, dans lequel chaque covariance est décrite par un petit nombre de variables.

Dans le VAE, la représentation latente n'est pas réduite à une valeur déterministe : elle est décrite par des distributions dont des réalisations peuvent être échantillonnées. L'entraînement équilibre la fidélité de reconstruction et la régularité de l'espace latent. Cette régularité est favorisée par le terme de divergence de Kullback–Leibler ($\mathcal{D}_{\mathrm{KL}}(\cdot)$), qui contraint les distributions latentes apprises à rester proches d'une distribution gaussienne de référence.

Dans l'IAE, un nombre fixé d'exemples est tiré du jeu d'apprentissage pour servir de points d'ancrage. Après passage par l'encodeur, leurs représentations dans l'espace latent constituent des points de référence. La représentation latente d'un autre exemple est approchée à partir des représentations de ces points d'ancrage, par combinaison de leurs positions dans l'espace latent, puis utilisée par le décodeur pour reconstruire la covariance correspondante. Autrement dit, l'IAE cherche à représenter les nouveaux exemples à partir de positions intermédiaires entre les points de référence déjà appris dans l'espace latent. Les points d'ancrage correspondent à des exemples du jeu d'apprentissage et non à de nouvelles observations.

Ces deux architectures diffèrent principalement par la manière dont elles structurent l'espace latent, mais permettent toutes deux d'apprendre une représentation de la famille de covariances simulées utilisable comme modèle du kernel du signal à 21 cm.

Retour au modèle GPR.

Une fois l'auto-encodeur appris, son décodeur peut être utilisé comme un modèle paramétrique de covariance. Les coordonnées latentes $\boldsymbol{\lambda}$ deviennent alors les paramètres de ce modèle. Pour la composante à 21 cm, la covariance $\mathbf{K}_{21}$ est fournie par le décodeur évalué aux paramètres $\boldsymbol{\lambda}$. Le modèle de covariance des données conserve la même structure que (11), soit :

$$\mathbf{K}_d = \mathbf{K}_{\mathrm{fg}} + \mathbf{K}_{21}(\boldsymbol{\lambda}) + \mathbf{K}_n.$$

Cette étape d'apprentissage du modèle de covariance introduit le *Machine Learning* dans la méthode ML-GPR. Le cadre probabiliste du GPR reste inchangé : seule la paramétrisation de $\mathbf{K}_{21}$ est remplacée par un modèle appris à partir des simulations. Dans la notation des auteurs, $\boldsymbol{\lambda}$ paramètre ce kernel appris, tandis que $\boldsymbol{\psi}$ regroupe les hyperparamètres associés aux autres composantes du modèle.

Le modèle appris permet de représenter une famille de covariances du signal à 21 cm. Les coordonnées latentes $\boldsymbol{\lambda}$ servent à paramétrer la covariance considérée au cours de l'inférence. La vraisemblance fait intervenir à la fois les hyperparamètres $\boldsymbol{\psi}$ des autres composantes et les paramètres latents $\boldsymbol{\lambda}$ du kernel appris. L'article étend l'évaluation de la distribution *a posteriori* (9) à ces deux ensembles de paramètres :

$$\log p\left(\boldsymbol{\psi}, \boldsymbol{\lambda} \mid \boldsymbol{d}, \boldsymbol{x}\right) \propto \log p\left(\boldsymbol{d} \mid \boldsymbol{x}, \boldsymbol{\psi}, \boldsymbol{\lambda}\right) + \log p(\boldsymbol{\psi}) + \log p(\boldsymbol{\lambda}),$$

en supposant des lois *a priori* indépendantes *i.e.*, $p(\boldsymbol{\psi}, \boldsymbol{\lambda}) = p(\boldsymbol{\psi})\, p(\boldsymbol{\lambda})$.

Les auteurs explorent la distribution *a posteriori* de $\boldsymbol{\psi}$ et $\boldsymbol{\lambda}$ par méthode MCMC afin d'en déterminer le maximum *a posteriori*, et de quantifier l'incertitude sur ces paramètres. La méthode MCMC ne constitue pas une nouvelle méthode de séparation : chaque échantillon fournit un jeu de paramètres qui détermine les matrices de covariance, puis le même conditionnement gaussien que dans le GPR standard est appliqué. Les échantillons de la distribution *a posteriori* peuvent ensuite être propagés jusqu'à l'estimation du spectre de puissance.

Application au signal à 21 cm.

Dans l'application au signal à 21 cm, les auteurs exploitent son isotropie et entraînent les auto-encodeurs sur des spectres de puissance simulés. Le décodeur fournit alors un spectre de puissance sphériquement moyenné et normalisé, qui est ensuite converti en matrice de covariance fréquence–fréquence $\mathbf{K}_{21}$. Cette conversion introduit une dépendance à l'échelle transverse $k_\perp$, elle-même reliée à la norme $|\boldsymbol{u}|$ des coordonnées (2) (*cf.* Mertens et al. 2024, Eqs. (17)–(19)). Le modèle de covariance peut donc varier avec $k_\perp$, contrairement au kernel générique du GPR standard.

De plus, une fois le kernel appris intégré au modèle de covariance, le conditionnement gaussien reste celui du GPR standard. Les expressions de la moyenne et de la covariance *a posteriori* des composantes conservent la même forme qu'en section 2.2. En particulier, la contribution d'avant-plan peut toujours être estimée à partir de (12), puis soustraite aux données selon la définition du résidu de (14).

Estimation du spectre de puissance.

Pour estimer le spectre de puissance du signal à 21 cm et son incertitude, les auteurs utilisent les échantillons de la distribution *a posteriori* des hyperparamètres fournis par la chaîne MCMC. Pour chaque échantillon, l'équation (13) fournit la moyenne et la covariance *a posteriori* de $\boldsymbol{f}_{21}$. Une réalisation du signal peut alors être générée sous la forme :

$$\mathbb{E}[\boldsymbol{f}_{21} \mid \boldsymbol{d}] + \boldsymbol{\delta}_{21}, \text{ avec } \boldsymbol{\delta}_{21} \sim \mathcal{N}\left(\mathbf{0}, \mathrm{cov}(\boldsymbol{f}_{21} \mid \boldsymbol{d})\right),$$

où l'ajout de $\boldsymbol{\delta}_{21}$ permet de ne pas réduire le signal reconstruit à sa seule moyenne *a posteriori*, mais de propager également l'incertitude de reconstruction jusqu'à l'estimation du spectre de puissance.

Pour chacune de ces réalisations, le spectre de puissance est calculé selon la même procédure que dans la section 2.2 : les données sont d'abord exprimées sous la forme $P(k_\perp, k_\parallel)$, puis regroupées pour obtenir le spectre sphériquement moyenné $P(k)$. Les auteurs obtiennent ainsi un ensemble de spectres correspondant aux différentes réalisations du signal. Leur médiane fournit l'estimation

retenue, tandis que leur écart-type caractérise l'incertitude à 1σ (Mertens et al. 2024, Sec. 2.3). ML-GPR ne modifie pas la définition du spectre de puissance : la différence avec le GPR standard intervient dans la modélisation de la covariance du signal et dans la propagation de l'incertitude associée jusqu'à l'estimation du spectre.

Finalement, ML-GPR conserve le cadre probabiliste du GPR tout en remplaçant, pour le signal à 21 cm, le kernel analytique par un modèle de covariance appris à partir de simulations. Néanmoins, la méthode reste dépendante de la qualité de la covariance apprise $\mathbf{K}_{21}$, et par conséquent, de la représentativité des simulations utilisées. Si celles-ci ne représentent pas correctement les structures du signal réel, une inadéquation de la covariance reste possible.

2.4 Robustesse de ML-GPR et limites du modèle appris

Acharya et al. (2024) prolongent l'étude de ML-GPR en évaluant sa robustesse au *covariance mismatch i.e.*, lorsque la covariance $\mathbf{K}_{21}$ supposée par le modèle ne correspond pas parfaitement à celle du signal effectivement présent dans les données. L'objectif est d'étudier dans quelle mesure le kernel VAE réduit les biais induits par cette inadéquation.

Modèle de données et lecture en séparation de composantes.

Le modèle de données de l'article étend celui de l'équation (10) par l'ajout d'une composante résiduelle corrélée $\boldsymbol{f}_{\rm ex}$:

$$\boldsymbol{d} = \boldsymbol{f}_{\rm sky} + \boldsymbol{f}_{\rm mix} + \boldsymbol{f}_{\rm ex} + \boldsymbol{f}_{21} + \boldsymbol{n}.$$

Les termes $\boldsymbol{f}_{\rm sky}$ et $\boldsymbol{f}_{\rm mix}$ correspondent à la décomposition de $\boldsymbol{f}_{\rm fg}$ en deux composantes présentant des structures de corrélation en fréquence différentes : $\boldsymbol{f}_{\rm sky}$ décrit les émissions astrophysiques, tandis que $\boldsymbol{f}_{\rm mix}$ regroupe principalement les effets instrumentaux. La composante $\boldsymbol{f}_{\rm ex}$ provient de l'analyse LOFAR de Mertens et al. (2020). Après application d'un modèle GPR comprenant les avant-plans, le signal à 21 cm et le bruit, un résidu spectralement corrélé subsistait dans les données. Mertens et al. (2020, Sec. 4.2.1) l'introduisent alors dans le modèle sous la forme d'une covariance supplémentaire $\mathbf{K}_{\rm ex}$, décrite par un kernel de Matérn caractérisé, entre autres, par une longueur de cohérence $l_{\rm ex}$ et une variance $\sigma^2_{\rm ex}$ (Mertens et al. 2020, Sec. 4.2.1). Cette composante est modélisée séparément afin de rendre compte de cette puissance résiduelle corrélée dans le modèle GPR. La covariance totale de l'équation (11) se réécrit alors :

$$\mathbf{K}_d = \mathbf{K}_{\rm sky} + \mathbf{K}_{\rm mix} + \mathbf{K}_{\rm ex} + \mathbf{K}_{21} + \mathbf{K}_n.$$

Le conditionnement gaussien reste celui de la section 2.2. L'équation (13) montre que la part des données attribuée au signal à 21 cm dépend du modèle de covariance qui lui est associé. Une erreur sur $\mathbf{K}_{21}$ peut modifier sa récupération.

De même, lorsque deux composantes présentent des structures de covariance *proches*, leur séparation devient plus difficile, car leurs covariances fournissent moins d'information pour les distinguer.

Covariance mismatch et intérêt du kernel appris.

Le problème central étudié par les auteurs est celui du *covariance mismatch i.e.*, d'une différence entre la covariance supposée pour le signal à 21 cm et sa covariance réelle.

Cette difficulté doit être distinguée d'un simple problème de bruit : une augmentation du temps d'observation réduit la contribution de $\mathbf{K}_n$, mais ne corrige pas nécessairement un mauvais choix de $\mathbf{K}_{21}$.

Pour étudier cet effet dans un cadre contrôlé, les auteurs construisent des jeux de données simulés inspirés de la configuration LOFAR de Mertens et al. (2020), auxquels un signal à 21 cm connu est ajouté. Sa récupération après GPR peut être comparée directement au signal d'entrée, suivant la logique des tests sur données simulées (Mertens et al. 2020, Annexe A). Ils comparent le kernel VAE appris à partir de simulations `GRIZZLY` [2] à plusieurs kernels analytiques de la famille de Matérn, pour plusieurs niveaux de bruit. Le kernel appris est également testé sur des signaux produits avec `CRASH`, qui ne fait pas partie de son cadre d'apprentissage. Dans les configurations étudiées, les kernels analytiques peuvent conserver un biais lorsque leur forme de covariance ne correspond pas à celle du signal d'entrée, y compris lorsque le niveau de bruit diminue. Le kernel VAE se montre plus robuste et récupère mieux le signal issu de `CRASH` (Acharya et al. 2024, Sec. 3.3).

Ces résultats montrent qu'une amélioration du rapport signal sur bruit ne suffit pas à supprimer le biais dû à un mauvais choix de $\mathbf{K}_{21}$.

Les auteurs testent aussi le kernel VAE à plusieurs redshifts (*cf.* Eq. (1)) et montrent que ses performances restent dépendantes du rapport signal sur bruit (Acharya et al. 2024, Secs. 3.4 et 4.2). Lorsque celui-ci devient trop faible, la récupération du signal n'est plus systématique dans les intervalles à 2σ et conduit plutôt à des limites supérieures (Acharya et al. 2024, Sec. 4.2).

En résumé, le kernel VAE améliore la robustesse de la séparation face au *covariance mismatch*, sans supprimer pour autant les limitations liées à un faible rapport signal sur bruit.

Rôle de la composante d'excès.

La composante d'excès $\boldsymbol{f}_{\rm ex}$ fournit un second test de robustesse. Selon les auteurs, cette composante était mal contrainte dans l'analyse de Mertens et al. (2020), de sorte qu'elle avait été récupérée conjointement avec le signal à 21 cm, leur séparation n'étant pas alors statistiquement justifiable.

Les auteurs étudient plus directement les conditions dans lesquelles elles peuvent être distinguées (Acharya et al. 2024, Sec. 4.1). Pour ce faire, ils font varier la longueur de cohérence $l_{\rm ex}$ et la variance $\sigma^2_{\rm ex}$ de cette composante, puis examinent leur effet sur la récupération de $\boldsymbol{f}_{21}$. Lorsque la puissance et la forme spectrale de la composante d'excès deviennent proches de celles du signal à

2. `GRIZZLY` et `CRASH` sont deux codes de simulation de la réionisation utilisés dans ces travaux pour générer des signaux à 21 cm.

21 cm, la séparation devient plus difficile (Acharya et al. 2024, Sec. 4.1).

Du point de vue du traitement statistique du signal, ce comportement suggère une difficulté de séparabilité lorsque $\mathbf{K}_{\text{ex}}$ et $\mathbf{K}_{21}$ deviennent proches. Il est cependant préférable de ne pas conclure à une non-identifiabilité théorique générale : l'article met en évidence ce comportement à travers des expériences numériques, sans établir de condition générale d'identifiabilité du modèle.

Correction de biais et limites de la validation.

Nous pouvons distinguer la correction de biais de l'estimateur du spectre de puissance de la question du *covariance mismatch*.

Mertens et al. (2020, Sec. 3.3.2) introduisent une correction afin d'obtenir un estimateur non biaisé de la covariance résiduelle. Cette dérivation suppose cependant que le modèle de covariance adopté décrit correctement les données. En revanche, une mauvaise spécification de $\mathbf{K}_{21}$ constitue un problème différent : elle peut modifier la séparation elle-même avant l'estimation finale du spectre.

Dans ce contexte, les auteurs examinent si une meilleure modélisation de la covariance du signal à 21 cm permet de réduire cette sensibilité. Leur étude conserve les choix de normalisation et de correction de biais employés par Mertens et al. (2020). Les auteurs indiquent toutefois que d'autres schémas de normalisation et de correction de biais pourraient réduire la dépendance au prior de covariance, une piste qu'ils laissent à des travaux futurs (Acharya et al. 2024, Sec. 4.3).

Leurs résultats sont obtenus sur des données simulées, pour lesquelles le signal à 21 cm d'entrée est connu. Ils permettent d'évaluer directement les biais de récupération, mais ne répondent pas entièrement à la question de la robustesse sur des observations réelles, où le signal recherché est inconnu.

Ces résultats conduisent à s'interroger sur la sensibilité de la séparation au modèle de covariance adopté et sur les moyens d'évaluer sa fiabilité en l'absence de vérité terrain.

Questions de recherche soulevées par cette analyse.

Les résultats et limitations précédents conduisent à plusieurs questions qui devront être confrontées aux travaux suivants. Elles ne sont pas toutes formulées explicitement sous cette forme par les auteurs, mais découlent directement des difficultés mises en évidence dans l'article :

- comment déterminer si un kernel appris reste représentatif des covariances rencontrées dans les données, en particulier lorsque celles-ci s'écartent des simulations utilisées pour l'apprentissage ;
- comment quantifier la sensibilité de $\mathbb{E}[\boldsymbol{f}_{21} \mid \boldsymbol{d}]$ à une erreur sur $\mathbf{K}_{21}$ et propager cette erreur jusqu'au spectre de puissance ;
- comment caractériser la dégradation de la séparation entre $\boldsymbol{f}_{\text{ex}}$ et $\boldsymbol{f}_{21}$ lorsque leurs structures de covariance deviennent proches ;
- comment réduire conjointement les biais introduits par le modèle de covariance et ceux provenant de l'estimation du spectre ;
- comment évaluer la robustesse de la séparation sur des observations réelles lorsque le signal cosmologique vrai est inconnu.

Ces questions fournissent une base d'analyse pour les travaux observationnels étudiés ensuite. Leur intérêt est précisément de permettre de distinguer ce qui relève d'une limitation déjà traitée dans les analyses récentes de ce qui peut encore constituer une difficulté méthodologique ouverte.

L'analyse de la bibliographie fait également apparaître l'étude de Kern et Liu (2021), qui reformule la soustraction des avant-plans par GPR dans le cadre des estimateurs quadratiques et met en évidence la sensibilité de certaines procédures d'estimation du spectre de puissance à une mauvaise spécification de la covariance du signal à 21 cm.

2.5 Application de ML-GPR aux observations LOFAR

Après les études réalisées sur données simulées, Mertens et al. (2025) appliquent ML-GPR à des observations réelles du radiotélescope LOFAR (*Low-Frequency Array*).

L'analyse porte sur trois intervalles de redshift, autour de $z \simeq \{8.3, 9.1, 10.1\}$, et vise à établir des limites supérieures sur le spectre de puissance du signal à 21 cm. En l'absence de détection significative, ces contraintes peuvent être écrites sous la forme :

$$\Delta^2_{21}(k) < \Delta^2_{\text{max}}(k),$$

où $\Delta^2_{\text{max}}(k)$ représente la limite supérieure compatible avec les observations à l'échelle k définie par (15).

Cette application introduit une difficulté absente des simulations précédentes : *le signal cosmologique effectivement présent dans les données n'est pas connu*. La qualité de la séparation ne peut plus être évaluée par comparaison directe à une vérité terrain.

Application aux observations.

ML-GPR intervient après plusieurs étapes de traitement des données, dont la calibration, la soustraction du modèle du ciel et la mitigation des interférences radiofréquences (*cf.* Mertens et al. 2025, Fig. 1). À ce stade de la chaîne, les données restent dominées par les avant-plans résiduels, environ trois ordres de grandeur plus brillants que le signal à 21 cm. Leur modélisation distingue une composante de ciel et une composante de *mode mixing*, auxquelles s'ajoutent une composante d'excès, le bruit et le signal recherché (Mertens et al. 2025, Eq. (11)).

Le signal à 21 cm est modélisé à l'aide du kernel VAE de la section 2.3. Les avant-plans restent décrits par des kernels analytiques de la famille de Matérn, mais leur modélisation est adaptée aux observations. En particulier, la longueur de cohérence associée au *mode mixing* dépend de la longueur des *baselines*, afin de mieux représenter la structure du *foreground wedge* (Mertens et al. 2025, Eq. (10)). Cette adaptation illustre le rôle du modèle de covariance des contaminants dans la séparation.

Dans les données analysées, la composante à 21 cm reste cependant largement non contrainte : sa variance est

significativement inférieure à celle du bruit, et les auteurs n'en obtiennent qu'une limite supérieure. Ils établissent ces limites supérieures sur le spectre de puissance aux trois redshifts considérés (Mertens et al. 2025, Secs. 4.1 et 4.4).

Validation par injection de signaux.

L'absence de vérité terrain pose donc la question soulevée dans la section précédente : *comment vérifier que la séparation ne supprime pas artificiellement une partie du signal recherché ?*

Pour l'étudier, les auteurs réalisent des tests d'injection. Des signaux à 21 cm synthétiques et connus sont ajoutés aux données avant l'application de ML-GPR. Le spectre récupéré après séparation peut alors être comparé au spectre injecté afin d'évaluer une éventuelle perte ou déformation du signal (Mertens et al. 2025, Sec. 5.3).

Les tests portent sur différentes formes du spectre à 21 cm et sur plusieurs amplitudes du signal injecté relativement au niveau de bruit. Ils incluent aussi des signaux issus de simulations `21cmFAST`[3], alors que le kernel utilisé par ML-GPR a été appris à partir de simulations `GRIZZLY`.

La récupération est donc testée sur des signaux qui ne proviennent pas du modèle d'apprentissage. Dans les configurations étudiées, les signaux injectés sont généralement récupérés de manière compatible avec leurs valeurs d'entrée. Le *signal loss* induit par ML-GPR reste faible, de même que le biais résiduel sur le spectre de puissance à 21 cm. Les cas les plus difficiles correspondent à des spectres dont la structure devient plus proche de celle du *mode mixing* (Mertens et al. 2025, Sec. 5.3).

Ces résultats prolongent les observations faites sur données simulées : la difficulté de la séparation ne dépend pas uniquement du niveau du signal relativement au bruit, mais également de la proximité des structures statistiques des composantes.

Limites et questions soulevées.

L'application aux observations LOFAR montre que l'amélioration du modèle de covariance du signal à 21 cm ne suffit pas, à elle seule, à supprimer les difficultés de séparation. Les auteurs indiquent que la séparation entre les avant-plans résiduels et le signal à 21 cm reste difficile. Alors que $\mathbf{K}_{21}$ est décrit par un modèle appris à partir de simulations, les covariances des avant-plans reposent principalement sur des formes analytiques. Ils envisagent donc des modèles de covariance mieux adaptés aux effets instrumentaux présents dans les données, éventuellement construits à partir de simulations (Mertens et al. 2025, Sec. 6.3). De plus, les auteurs mentionnent que le modèle actuel n'intègre pas les corrélations spatiales dont la prise en compte est laissée à un développement futur (Mertens et al. 2025, Sec. 3.7).

Ces résultats suggèrent plusieurs questions qui pourront être examinées à la lumière des travaux suivants :

— comment construire des modèles de covariance des contaminants mieux adaptés aux structures effectivement présentes dans les observations ;

— dans quelle mesure l'exploitation conjointe des corrélations fréquentielles et spatiales peut améliorer la séparation des composantes ;

— comment évaluer la robustesse de la séparation lorsque les données dépendent aussi des différentes étapes de calibration et de traitement instrumental.

2.6 Comparaison bayésienne et validation de modèles GPR

Après l'introduction du GPR et de ML-GPR, puis l'étude de leur robustesse sur simulations et sur observations réelles, Liu et al. (2026) s'intéressent à une question complémentaire : *dans quelle mesure la récupération du signal à 21 cm dépend-elle du modèle statistique adopté pour décrire les différentes composantes ?*

L'étude repose sur des observations simulées du radiotélescope SKA-Low, composante basse fréquence du *Square Kilometre Array* (SKA), et compare plusieurs paramétrisations du modèle GPR. L'objectif n'est plus seulement d'estimer les paramètres d'un modèle de covariance donné, mais aussi de comparer plusieurs modèles possibles et d'évaluer la fiabilité de la séparation qu'ils produisent.

Retour au modèle GPR.

Les auteurs reprennent le formalisme GPR de la section 2.2. Dans l'application au signal à 21 cm, ils conservent le modèle additif de l'équation (10), et la structure de covariance totale donnée par (11). La covariance des avant-plans est à nouveau décomposée entre une contribution associée aux émissions astrophysiques et une contribution associée au *mode mixing* (Liu et al. 2026, Sec. 4.1). La reconstruction de la composante à 21 cm repose sur le même conditionnement gaussien que celui de (13).

La nouveauté proposée par les auteurs ne réside pas dans le mécanisme de reconstruction, mais dans la comparaison bayésienne de plusieurs *modèles* GPR et dans la validation de la séparation qu'ils produisent.

Les modèles candidats se distinguent principalement par les hypothèses adoptées pour décrire la dépendance du *mode mixing* à la longueur des *baselines*, par l'introduction éventuelle d'un facteur de correction de la variance du bruit, et par le choix du kernel associé à $\mathbf{K}_{21}$, appris par VAE ou défini analytiquement (Liu et al. 2026, Tab. 6). Ces paramétrisations fournissent m modèles candidats auxquels les auteurs appliquent la même procédure d'inférence et de comparaison bayésienne.

Expressions de la vraisemblance, du prior et de la distribution *a posteriori* des paramètres.

Pour un modèle donné $\mathcal{M}$, les auteurs regroupent dans le vecteur $\boldsymbol{\psi}_{\mathcal{M}}$ l'ensemble de ses paramètres, y compris ceux associés aux kernels. Selon le modèle considéré, $\boldsymbol{\psi}_{\mathcal{M}}$ peut contenir, par exemple, les paramètres du kernel $\boldsymbol{\lambda}$ appris par VAE.

La vraisemblance $p(\boldsymbol{d} \mid \boldsymbol{\psi}_{\mathcal{M}}, \mathcal{M})$ décrit la probabilité des données observées pour un jeu donné de paramètres du

3. `21cmFAST` est un code semi-numérique permettant de générer rapidement des réalisations tridimensionnelles du signal cosmologique à 21 cm.

modèle. Dans le cadre du GPR, elle correspond à la vraisemblance marginale déjà introduite dans l'équation (8), obtenue après marginalisation des fonctions latentes. Le prior $p(\boldsymbol{\psi}_{\mathcal{M}} \mid \mathcal{M})$ décrit, quant à lui, l'information attribuée aux paramètres avant l'observation des données.

L'application du théorème de Bayes permet d'obtenir la distribution *a posteriori* des paramètres du modèle :

$$p(\boldsymbol{\psi}_{\mathcal{M}} \mid \boldsymbol{d}, \mathcal{M}) = \frac{p(\boldsymbol{d} \mid \boldsymbol{\psi}_{\mathcal{M}}, \mathcal{M}) p(\boldsymbol{\psi}_{\mathcal{M}} \mid \mathcal{M})}{p(\boldsymbol{d} \mid \mathcal{M})}. \tag{16}$$

Cette expression prolonge la distribution *a posteriori* des hyperparamètres déjà introduite dans l'équation (9), et fait désormais apparaître explicitement le modèle $\mathcal{M}$ et le terme $p(\boldsymbol{d} \mid \mathcal{M})$ (Liu et al. 2026, Eq. (31)) décrit dans le paragraphe suivant.

évidence bayésienne.

Le terme $p(\boldsymbol{d} \mid \mathcal{M})$ de l'équation (16) correspond à l'*évidence bayésienne* globale du modèle :

$$\begin{aligned} \mathcal{Z}_{\mathcal{M}} &\equiv p(\boldsymbol{d} \mid \mathcal{M}) \\ &= \int p(\boldsymbol{d} \mid \boldsymbol{\psi}_{\mathcal{M}}, \mathcal{M}) p(\boldsymbol{\psi}_{\mathcal{M}} \mid \mathcal{M}) \mathrm{d}\boldsymbol{\psi}_{\mathcal{M}}. \end{aligned} \tag{17}$$

L'intégration porte ici sur l'ensemble des valeurs possibles des paramètres $\boldsymbol{\psi}_{\mathcal{M}}$, pondérées par leur distribution *a priori* (Liu et al. 2026, Eq. (30)).

Il convient ici de distinguer deux niveaux de marginalisation :

- la vraisemblance marginale du GPR de l'équation (8), appelée *evidence* par Mertens et al. (2018), résulte de la marginalisation des fonctions latentes pour un jeu fixé de paramètres ;
- l'évidence bayésienne globale $\mathcal{Z}_{\mathcal{M}}$ définie par (17) marginalise à son tour les paramètres du modèle.

L'évidence globale permet de comparer plusieurs modèles après marginalisation de leurs paramètres respectifs.

Les auteurs estiment ensuite les distributions *a posteriori* des paramètres ainsi que les évidences des modèles candidats. Ces distributions fournissent également les estimations MAP des paramètres et permettent d'en caractériser l'incertitude (Liu et al. 2026, Sec. 4.5).

Comparaison bayésienne des modèles.

Pour comparer deux modèles $\mathcal{M}_i$ et $\mathcal{M}_j$, les auteurs utilisent le facteur de Bayes $B_{ij} := \mathcal{Z}_i / \mathcal{Z}_j$ sous forme logarithmique :

$$\Delta \log \mathcal{Z}_{\mathcal{M}} := \log B_{ij} = \log \mathcal{Z}_i - \log \mathcal{Z}_j. \tag{18}$$

Ces quantités mesurent le support apporté par les données à chaque paramétrisation du GPR, l'évidence fournissant une mesure globale de leur adéquation aux données tout en tenant compte de leur complexité (Liu et al. 2026, Sec. 4.5).

En outre, les auteurs évaluent le support apporté spécifiquement à l'introduction d'une composante à 21 cm. Pour chaque modèle $\mathcal{M}$ contenant $\mathbf{K}_{21}$, ils construisent une version $\mathcal{M}_{\text{no-21}}$ dans laquelle cette composante est retirée. En notant respectivement $\mathcal{Z}_M$ et $\mathcal{Z}_N$ leurs évidences, ils considèrent la quantité :

$$\Delta \log \mathcal{Z}_{21} = \log \mathcal{Z}_M - \log \mathcal{Z}_N. \tag{19}$$

Une valeur positive indique que l'ajout d'une composante dédiée au signal à 21 cm augmente le support apporté au modèle par les données. Notons que dans les simulations considérées, $\Delta \log \mathcal{Z}_{21}$ est positif pour les cinq paramétrisations étudiées (Liu et al. 2026, Sec. 5.1).

Il faut toutefois distinguer cette quantité de $\Delta \log \mathcal{Z}_{\mathcal{M}}$ donnée par (18) : cette dernière compare deux paramétrisations différentes du modèle GPR, tandis que $\Delta \log \mathcal{Z}_{21}$ compare un modèle avec et sans composante $\mathbf{K}_{21}$.

Propagation de l'incertitude et qualité de reconstruction.

La distribution *a posteriori* (16) ne fournit pas uniquement une estimation des paramètres. Les auteurs utilisent aussi les échantillons de la distribution *a posteriori* afin de propager leur incertitude jusqu'à la reconstruction du signal à 21 cm puis à l'estimation de son spectre de puissance.

Pour chaque jeu de paramètres échantillonné, la moyenne et la covariance prédictives de la composante à 21 cm sont calculées à partir du conditionnement déjà donné dans l'équation (13). Des réalisations du signal sont ensuite générées et utilisées pour construire une distribution prédictive du spectre de puissance (Liu et al. 2026, Sec. 5.3). Les données étant simulées, le véritable signal injecté est connu. Les auteurs peuvent donc comparer directement le spectre reconstruit au spectre injecté, au moyen des résidus, de la RMSE et de z-scores tenant compte de l'incertitude prédictive. Cette comparaison fournit une information différente de l'évidence bayésienne : elle porte spécifiquement sur la qualité de reconstruction de la composante à 21 cm.

évidence globale et qualité de la séparation.

La confrontation de ces métriques met en évidence une limite importante de l'évidence globale (17). Cette quantité caractérise la capacité du modèle complet à décrire les données après marginalisation de ses paramètres, mais ne renseigne pas directement sur la qualité de reconstruction de chacune des composantes. Le z-score (Liu et al. 2026, Eq. (34)) complète cette information en rapportant l'écart entre le spectre reconstruit et le spectre injecté à l'incertitude prédictive associée ; une valeur faible indique un écart réduit au regard de cette incertitude.

Les auteurs montrent que l'évidence globale, la RMSE et les z-scores ne conduisent pas nécessairement au même classement des modèles. Autrement dit, un modèle qui décrit bien les données dans leur ensemble peut néanmoins fournir une reconstruction moins fidèle de la composante à 21 cm. Cela suggère que, pour certains modèles étudiés, certaines structures résiduelles des avant-plans ou du bruit peuvent être partiellement attribuées à la composante cosmologique, sans que cette situation soit révélée par l'évidence globale (Liu et al. 2026, Secs. 5.4–5.5).

Test nul bayésien.

Cette divergence entre métriques motive l'utilisation d'un test nul bayésien. Les auteurs construisent pour cela un jeu de validation dans lequel le signal à 21 cm est absent :

$$\boldsymbol{d}_{\mathrm{no}-21} = \boldsymbol{f}_{\mathrm{fg}} + \boldsymbol{n}.$$

Pour chaque paramétrisation, le modèle $\mathcal{M}$ et sa version $\mathcal{M}_{\text{no-21}}$ sont évalués sur ce même jeu de données.

En notant leurs évidences respectives sur ces données $\mathcal{Z}'_M$ et $\mathcal{Z}'_N$, les auteurs définissent la quantité :

$$\Delta \log \mathcal{Z}_{\mathrm{null}} = \log \mathcal{Z}'_N - \log \mathcal{Z}'_M. \tag{20}$$

Une valeur positive, $\Delta \log \mathcal{Z}_{\mathrm{null}} > 0$, indique que les données, qui ne contiennent aucun signal à 21 cm par construction, favorisent le modèle sans $\mathbf{K}_{21}$.

Dans les simulations étudiées, certains des modèles considérés satisfont ce critère, tandis que d'autres conduisent à des valeurs négatives de $\Delta \log \mathcal{Z}_{\mathrm{null}}$ (Liu et al. 2026, Fig. 9). Pour ces derniers, le modèle incluant la composante à 21 cm obtient une évidence plus élevée alors même que le signal est absent. Les auteurs interprètent ce comportement comme pouvant résulter de l'absorption de structures résiduelles du bruit ou des avant-plans par la composante attribuée au signal cosmologique (Liu et al. 2026, Sec. 5.5).

Questions de recherche soulevées par cette analyse.

En résumé, l'approche de Liu et al. (2026) introduit plusieurs niveaux complémentaires d'inférence.

Pour un modèle donné, la distribution *a posteriori* (16) décrit les paramètres compatibles avec les données et leur incertitude. L'évidence (17) marginalise ces paramètres et permet de comparer plusieurs modèles. Le facteur de Bayes (18) quantifie le support relatif, tandis que $\Delta \log \mathcal{Z}_{21}$ (*cf.* Eq. (19)) quantifie le support apporté par les données à la présence d'une composante dédiée au signal à 21 cm.

Enfin, le test nul (20) apporte une information différente : il permet d'évaluer si un modèle favorise l'introduction d'une composante cosmologique lorsque celle-ci est absente.

Les résultats suggèrent ainsi que la sélection bayésienne d'un modèle et l'évaluation de la qualité de la séparation correspondent à deux aspects distincts de l'analyse. Une évidence élevée ne suffit pas nécessairement, à elle seule, à établir que les composantes sont correctement séparées.

Ces résultats prolongent plusieurs questions apparues dans les travaux précédents et dans les axes considérés pour la suite :

- comment construire des modèles de covariance suffisamment adaptés aux contaminants pour limiter leur attribution au signal cosmologique ;
- comment construire des diagnostics permettant d'évaluer la fiabilité de la séparation lorsque le véritable signal à 21 cm n'est pas connu ;
- dans quelle mesure les incertitudes et les biais de séparation se répercutent sur l'estimation du spectre de puissance et sur l'inférence physique.

Les travaux de Liu et al. (2026) montrent comment une approche bayésienne permet de comparer et de valider plusieurs modélisations GPR. Dans la perspective des questions précédentes, ce cadre pourrait être étendu à l'étude de la robustesse des modèles de covariance, à la quantification des incertitudes de séparation et à l'évaluation de la fiabilité de la séparation.

2.7 Impact de la mitigation sur l'inférence astrophysique

Les travaux de Giri et Mertens (2026) prolongent l'étude de la séparation du signal à 21 cm en considérant ses conséquences sur l'interprétation astrophysique. Les auteurs étudient dans quelle mesure les choix effectués lors de la mitigation des avant-plans peuvent affecter les paramètres physiques déduits du signal.

Deux stratégies de mitigation des avant-plans.

Les travaux comparent pour cela deux stratégies de mitigation des avant-plans :

- *foreground avoidance* : l'analyse est restreinte à l'*EoR window* dans le plan $(k_\perp, k_\parallel)$, ce qui exclut les modes situés dans le *foreground wedge* et entraîne une perte d'information aux faibles valeurs de k, correspondant aux plus grandes échelles spatiales ;
- *foreground removal* : les avant-plans sont modélisés puis soustraits afin de récupérer des modes supplémentaires, en particulier à l'intérieur du *foreground wedge*. Dans cette étude, cette stratégie repose sur ML-GPR.

Dans le cas du *foreground removal*, ML-GPR exploite les différences de cohérence spectrale entre les avant-plans et le signal à 21 cm pour réaliser leur séparation statistique (Giri et Mertens 2026, Sec. 3). Le compromis est donc le suivant : le *foreground avoidance* évite cette séparation statistique, mais au prix d'une perte d'information sur les plus grandes échelles spatiales (Giri et Mertens 2026, Sec. 3.1), tandis que le *foreground removal* permet de récupérer une partie de ces modes, au risque d'introduire des biais systématiques si le modèle statistique est mal spécifié (Giri et Mertens 2026, Sec. 3.2).

Conséquences sur l'inférence astrophysique.

Après avoir comparé les spectres de puissance reconstruits, les auteurs utilisent ces spectres pour réaliser une inférence bayésienne des paramètres astrophysiques (Giri et Mertens 2026, Sec. 4). Cette analyse permet d'examiner comment les différences introduites lors de la mitigation des avant-plans se répercutent sur les contraintes astrophysiques obtenues. Les données étant simulées, les valeurs des paramètres utilisées pour les générer sont connues et peuvent servir de référence (Giri et Mertens 2026, Sec. 2.1).

Dans le scénario utilisant l'ensemble des intervalles de redshift, les deux stratégies présentent des biais sur certains paramètres astrophysiques. Les distributions *a posteriori* obtenues avec le *foreground avoidance* sont systématiquement plus larges que celles obtenues avec le *foreground removal*, en raison de la perte d'information liée

aux modes exclus du *foreground wedge* (Giri et Mertens 2026, Sec. 4.1).

Dans le cas du *foreground removal*, les auteurs observent des biais plus marqués pour certains paramètres, qu'ils attribuent à la puissance sous-estimée aux grandes échelles dans les spectres reconstruits (Giri et Mertens 2026, Sec. 4.1).

De ce fait, les deux approches de mitigation aboutissent à des contraintes astrophysiques distinctes, aussi bien en termes de largeur que de biais observés.

Rôle des échelles et du redshift.

Afin d'identifier l'origine des biais observés, les auteurs examinent deux effets sur l'inférence astrophysique : celui des modes k supplémentaires récupérés par le *foreground removal*, et celui des intervalles de redshift les moins bien reconstruits (Giri et Mertens 2026, Sec. 4.2).

Lorsque le *foreground removal* est restreint aux mêmes modes k que le *foreground avoidance*, les distributions *a posteriori* obtenues avec les deux stratégies se situent dans des régions similaires. Celles obtenues avec le *foreground avoidance* restent toutefois plus larges, en raison de la variance plus élevée du spectre de puissance associée au nombre plus faible de visibilités utilisées dans l'*EoR window*.

Les auteurs examinent ensuite l'effet des intervalles de redshift les moins bien reconstruits. Lorsque les deux intervalles les plus contaminés sont exclus de l'inférence, les biais de plusieurs paramètres sont réduits dans le cas du *foreground removal*.

Les auteurs concluent que des intervalles de redshift mal reconstruits peuvent dégrader une inférence conjointe sur plusieurs redshifts (Giri et Mertens 2026, Sec. 5).

Limites pour les données réelles et ouverture.

L'analyse proposée par les auteurs repose sur des simulations, pour lesquelles le signal de référence est connu. Cette connaissance permet d'identifier les intervalles de redshift dont la reconstruction est moins fidèle. Une telle comparaison directe n'est toutefois plus possible sur des observations réelles. Les auteurs soulignent que l'identification des intervalles mal reconstruits nécessite des critères de qualité indépendants avant leur utilisation dans une inférence multi-redshift (Giri et Mertens 2026, Sec. 5). Les résultats précédents montrent par ailleurs que les erreurs de mitigation peuvent affecter les spectres de puissance reconstruits et les contraintes astrophysiques qui en sont déduites (Giri et Mertens 2026, Secs. 4.1–5).

Les auteurs envisagent une autre piste méthodologique : intégrer directement les modèles d'avant-plan dans le modèle d'inférence, afin d'estimer conjointement les paramètres astrophysiques et ceux décrivant les avant-plans (Giri et Mertens 2026, Sec. 5).

Ces résultats font apparaître plusieurs questions de recherche :

- comment construire, en l'absence de vérité terrain, des diagnostics permettant d'évaluer la fiabilité de la séparation en fonction du redshift et des modes k ;
- comment utiliser ces diagnostics pour identifier les données dont la reconstruction est suffisamment fiable pour contribuer à une inférence multi-redshift ;
- dans quelle mesure l'intégration directe des modèles d'avant-plan dans le modèle d'inférence pourrait réduire les biais introduits par leur mitigation préalable.

Ces questions prolongent les problématiques de robustesse et de validation introduites dans les sections 2.4 à 2.6, en les reliant aux conséquences de la séparation sur l'inférence astrophysique.

2.8 Synthèse des limitations et positionnement de la proposition

Les travaux examinés conduisent à distinguer trois difficultés complémentaires.

La première concerne la sensibilité de la séparation à la spécification des covariances, mise en évidence dans les études de GPR et de ML-GPR (Acharya et al. 2024 ; Mertens et al. 2024, 2018).

La deuxième porte sur la représentation des contaminants et de leurs structures spatiales et spectrales dans les analyses observationnelles (Mertens et al. 2025).

La troisième concerne l'évaluation de la fiabilité de la séparation et de ses conséquences sur le spectre de puissance et l'inférence astrophysique (Giri et Mertens 2026 ; Liu et al. 2026).

Cette analyse a conduit à formuler une proposition de recherche articulée autour de deux axes méthodologiques : la modélisation structurée des contaminants et la régularisation de l'apprentissage des modèles de covariance.

Cette proposition s'accompagne d'une stratégie d'évaluation reposant sur la quantification des incertitudes, la comparaison des modèles et l'analyse de la robustesse de la séparation dans un cadre probabiliste et bayésien.

Références

Acharya, A. et al. (2024). "21-cm signal from the Epoch of Reionization : a machine learning upgrade to foreground removal with Gaussian process regression". Dans : *Monthly Notices of the Royal Astronomical Society*.

Giri, S. K. et F. Mertens (2026). "Impact of 21-cm foreground mitigation strategies on reionization power spectrum constraints". Dans : *Monthly Notices of the Royal Astronomical Society*.

Kern, N. S. et A. Liu (2021). "Gaussian process foreground subtraction and power spectrum estimation for 21 cm cosmology". Dans : *Monthly Notices of the Royal Astronomical Society*.

Liu, A. et J. R. Shaw (2020). "Data analysis for precision 21 cm cosmology". Dans : *Publications of the Astronomical Society of the Pacific*.

Liu, Y., E. de Lera Acedo et P. H. Sims (2026). "Bayesian model comparison and validation with Gaussian Process Regression for interferometric 21-cm signal recovery". Dans : *Monthly Notices of the Royal Astronomical Society*.

Mertens, F., J. Bobin et I. P. Carucci (2024). “Retrieving the 21-cm signal from the Epoch of Reionization with learnt Gaussian process kernels”. Dans : *Monthly Notices of the Royal Astronomical Society*.

Mertens, F., A. Ghosh et L. Koopmans (2018). “Statistical 21-cm signal separation via Gaussian Process Regression analysis”. Dans : *Monthly Notices of the Royal Astronomical Society*.

Mertens, F. et al. (2020). “Improved upper limits on the 21 cm signal power spectrum of neutral hydrogen at z $\approx$ 9.1 from LOFAR”. Dans : *Monthly Notices of the Royal Astronomical Society*.

Mertens, F. et al. (2025). “Deeper multi-redshift upper limits on the epoch of reionisation 21 cm signal power spectrum from LOFAR between z= 8.3 and z= 10.1”. Dans : *Astronomy & Astrophysics*.