Statistical 21-cm signal separation via Gaussian Process Regression analysis

Mertens, F. G.; Ghosh, A.; Koopmans, L. V. E.

Published in:
Monthly Notices of the Royal Astronomical Society

DOI:
10.1093/mnras/sty1207

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **478,** 3640–3652 (2018)
Advance Access publication 2018 May 11

doi:10.1093/mnras/sty1207

Statistical 21-cm signal separation via Gaussian Process Regression analysis

F. G. Mertens,[1]★ A. Ghosh[2,3] and L. V. E. Koopmans[1]

[1]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*
[2]*Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, Bellville 7535, South Africa*
[3]*Square Kilometre Array Radio Telescope (SKA) South Africa, The Park, Park Road, Cape Town 7405, South Africa*

Accepted 2018 May 1. Received 2018 April 26; in original form 2017 December 11

ABSTRACT

Detecting and characterizing the Epoch of Reionization (EoR) and Cosmic Dawn via the redshifted 21-cm hyperfine line of neutral hydrogen will revolutionize the study of the formation of the first stars, galaxies, black holes, and intergalactic gas in the infant Universe. The wealth of information encoded in this signal is, however, buried under foregrounds that are many orders of magnitude brighter. These must be removed accurately and precisely in order to reveal the feeble 21-cm signal. This requires not only the modelling of the Galactic and extragalactic emission, but also of the often stochastic residuals due to imperfect calibration of the data caused by ionospheric and instrumental distortions. To stochastically model these effects, we introduce a new method based on 'Gaussian Process Regression' (GPR) which is able to statistically separate the 21-cm signal from most of the foregrounds and other contaminants. Using simulated LOFAR–EoR data that include strong instrumental mode mixing, we show that this method is capable of recovering the 21-cm signal power spectrum across the entire range $k = 0.07 - 0.3\,h\,\mathrm{cMpc}^{-1}$. The GPR method is most optimal, having minimal and controllable impact on the 21-cm signal, when the foregrounds are correlated on frequency scales $\gtrsim$3 MHz and the rms of the signal has $\sigma_{21\mathrm{cm}} \gtrsim 0.1\,\sigma_{\mathrm{noise}}$. This signal separation improves the 21-cm power-spectrum sensitivity by a factor $\gtrsim$3 compared to foreground avoidance strategies and enables the sensitivity of current and future 21-cm instruments such as the *Square Kilometre Array* to be fully exploited.

Key words: methods: data analysis – methods: statistical – techniques: interferometric – dark ages, reionization, first stars – cosmology: observations.

1 INTRODUCTION

Observations of the redshifted 21-cm signal from neutral Hydrogen is the most promising method for revealing astrophysical processes occurring during the Epoch of Reionization (EoR) and the Cosmic Dawn (CD), and has great potential at independently constraining the cosmological parameters (see e.g. Furlanetto, Oh & Briggs 2006; Morales & Wyithe 2010, for reviews). Several experiments are currently underway aiming at statistically detecting the 21-cm signal from the EoR (e.g. LOFAR,[1] MWA,[2] and PAPER[3]), already achieving increasingly attractive upper limits on the 21-cm signal power spectra (Ali et al. 2015; Beardsley et al. 2016; Patil et al. 2017), and paving the way for the second generation experiments such as the SKA[4] and HERA[5] which will be capable, with their order of magnitude improvement in sensitivity, of robust power-spectra characterization and for the first time directly image the large-scale neutral hydrogen structures from EoR and CD.

A major obstacle in achieving this exciting goal is that the cosmological signal is considerably weaker than the astrophysical foregrounds. The foregrounds must be accurately and precisely removed from the observed data as any error at this stage has the ability to strongly affect the 21-cm signal extraction. While the brightest extragalactic sources can be modelled and removed after direction dependent calibration (e.g. Yatawatta et al. 2013), the remaining foregrounds, composed of extragalactic emission below the confusion noise level and diffuse and partly polarized galactic emission, are still approximately 3–4 orders of magnitude brighter than the 21-cm signal. They are nevertheless expected to be spectrally smooth

★ E-mail: mertens@astro.rug.nl
[1]Low Frequency Array, http://www.lofar.org
[2]Murchison Widefield Array, http://www.mwatelescope.org
[3]Precision Array to Probe EoR, http://eor.berkeley.edu
[4]Square Kilometre Array, http://www.skatelescope.org
[5]Hydrogen Epoch of Reionization Array, http://reionization.org

while the 21-cm signal is anticipated to be uncorrelated on frequency scales on the order of MHz or larger. This important difference is the main characteristic exploited by the many techniques that have been proposed to model and remove the foreground emission, including parametric fits (e.g. Jelić et al. 2008; Bonaldi & Brown 2015) and non-parametric methods (e.g. Harker et al. 2009; Chapman et al. 2013).

The assumption made here of a smooth foreground signal is however strongly affected by the limitations and constraints of the observational setup. Many additional contaminants have been identified related to the reality of radio interferometry, and observation in the low-frequency domain. The chromatic (i.e. wavelength dependent) response of the instrument manifests itself as a frequency dependence of both the synthesized beam, also called the point spread function (PSF), and the primary beam (PB) of a receiver station, producing chromatic side lobes from sources inside the field of view (FoV, Vedantham, Udaya Shankar & Subrahmanyan 2012; Hazelton, Morales & Sullivan 2013) and outside it (Thyagarajan et al. 2015; Mort et al. 2017; Gehlot et al. 2017). Calibration errors and mis-subtraction of sources due to imperfect sky modelling will also contribute to additional side lobe noise (Datta, Bowman & Carilli 2010; Morales et al. 2012; Trott, Wayth & Tingay 2012; Barry et al. 2016; Patil et al. 2016; Ewall-Wice et al. 2017). The rapid phase and sometime amplitudes modifications of radio waves caused by small-scale structures in the ionosphere also produce scintillation noise (Koopmans 2010; Vedantham & Koopmans 2016). These different mechanisms will all add spectral structure to the otherwise smooth astrophysical foregrounds, and are well known as 'mode-mixing' effects in the literature.

Both simulations and analytic calculations have demonstrated that these mode-mixing contaminants are essentially localized inside a wedge-like region in the two-dimensional angular ($k_\perp$) versus line-of-sight ($k_\parallel$) power spectra (see Fig. 1). This peculiar shape is explained by the fact that larger baselines (higher $k_\perp$) change length more rapidly as a function of frequency than smaller baselines, causing increasingly faster spectral fluctuations, and thus producing power into proportionally higher $k_\parallel$ modes.

Mitigating those additional foreground contaminants has proven to be extremely difficult. Increasing the degrees of freedom of a parametric fit would considerably increase the fitting error and might also suppress the 21-cm signal at the lower value k modes. Non-parametric methods are in theory not limited to smooth models, but modelling an increasingly more complex foreground often means increasing the numbers of components (without a clear understanding about what they include), which risks the leakage of 21-cm signal into the reconstructed foreground model and vice versa. In Patil et al. (2017), six to eight components of the Generalized Morphological Component Analysis (GMCA; Chapman et al. 2013) were necessary to model, even imperfectly, the foreground contaminants, reaching limits where it is increasingly more difficult to assess and be confident about the accuracy of the foreground removal process. We note that the GMCA is not based on a statistical framework but simply separates the signal in the least number of morphological components. This makes it hard to build in a-priori knowledge about the signal in any kind of signal separation.

Ideally, we would like to consistently account for every single mode-mixing contaminant that have been identified so far. Recently, Ghosh, Mertens & Koopmans (2018) have demonstrated that estimating the 21-cm power spectrum using a maximum-likelihood inversion of the spherical-wave visibility equation can considerably reduce the chromatic effects due to the frequency dependence of the PSF, effectively recovering a PSF-deconvolved sky. Vedantham et al. (2012) also proposed a new imaging technique in the attempt of decreasing visibilities gridding artefacts. Convolving the visibilities with a 'frequency independent' window function makes it easier to strongly attenuate the frequency-dependent response to the side lobes of the primary antenna pattern and Radio Frequency Interference (RFI) sources, which are mostly located on the ground (Ghosh et al. 2011). Improving the PB characterization (Thyagarajan et al. 2016), and using calibration scheme which enforce smooth gain solution in frequency (Barry et al. 2016; Yatawatta 2016), also contribute to reducing the mode mixing. Nevertheless, most of the improvements are done with the purpose of limiting the leakage of foreground contaminants outside the foreground wedge, and any foreground removal strategy will still be required to properly handle mode-mixing contaminants inside the wedge.

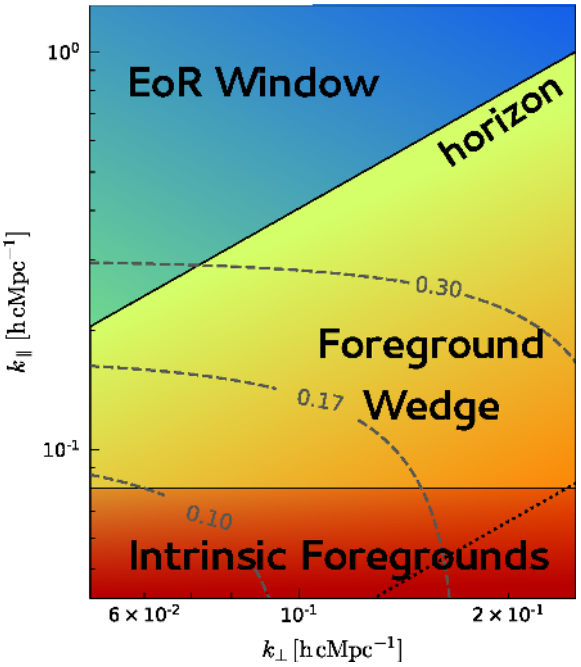

Figure 1. Schematic representation of the two-dimensional power spectra (inspired by a similar figure in Barry et al. 2016), illustrating the foreground wedge and the EoR window. Instrumental chromaticity and imperfect calibration and sky model will produce foreground mode-mixing contaminants which are mainly concentrated inside the PB FoV line (dashed line) and can leak up to the horizon line. Only modes above this line are theoretically free of foreground contaminants. Lines of equidistant $k = \sqrt{k_\perp^2 + k_\parallel}$ are overplotted in grey.

An alternative, which has been increasingly popular, is to try to avoid as much as possible the foregrounds, and only probe a triangular-shaped region in k-space where the 21-cm signal is dominant. Because most of the instrumental chromatic effects are confined inside the wedge, there exists in theory an 'EoR window' (see Fig. 1) within which one could perform statistical analyses of the 21-cm signal without significantly being affected by foreground contaminants. Liu, Parsons & Trott (2014a, b) proposed a mathematical formalism describing the wedge, allowing one to maximize the extent of the accessible EoR window. Several methods have also been developed to estimate the covariance of the foregrounds (Dillon et al. 2015; Murray, Trott & Jordan 2017) which can then be included in a power-spectra estimator (Trott et al. 2016). These foreground avoidance or suppression methods have the disadvantage, however, of considerably reducing the sensitivity of the instru-

ments, because they reduce the numbers of modes that can be probed (Furlanetto 2016). Pober et al. (2014) have estimated the impact of avoiding the foreground wedge region to be a factor ∼3 for PAPER or HERA, and even a factor ∼6 for LOFAR. It is thus not a viable alternative for experiments such as LOFAR–EoR, most sensible at $k \leq 0.3\ h\,\mathrm{cMpc}^{-1}$ with a peak sensibility at $k \sim 0.1\ h\,\mathrm{cMpc}^{-1}$, and for which very little foreground-free modes are available (see Fig. 1). Additionally, ignoring the wedge can also introduce a bias in the recovered 21-cm signal power spectra (Jensen et al. 2016) and it is also much harder to probe the redshift space distortion effects of the 21-cm signal if the foreground cleaning in the wedge region is discarded (Pober 2015).

Considering that for a successful foreground removal strategy all the foreground contaminants need to be accounted for, and that ad hoc modelling is not an option for most of them, we propose a novel non-parametric method based on Gaussian Process Regression (GPR). In this framework, the different components of the problem, including the astrophysical smooth foregrounds, mid-scale fluctuations associated with mode mixing, the noise, and a basic 21-cm signal model, are modelled with Gaussian Process (GP), allowing for a clean separation of their contributions, and a precise estimation of their uncertainty. GPR is extensively used in machine learning applications and has been successfully used in astronomy, for example to model blazar broad-band flares (Karamanavis et al. 2016), inferring stellar rotation periods (Hojjati, Kim & Linder 2013), or modelling instrumental systematics (Aigrain, Parviainen & Pope 2016). It provides flexibility and avoids having to specify an arbitrary functional form for the variations we seek to model. Implemented in a Bayesian framework, it enables us to incorporate relevant physical information in the form of covariance structure priors (spectral and possible spatial) on the various components.

We introduce the foreground modelling and removal method in Section 2. To demonstrate the ability of the technique, we perform simulations including realistic astrophysical foreground models, mid-scale frequency fluctuations, and the simulated 21-cm signal. We introduce the simulation pipeline in Section 3, before presenting the results in Section 4. Finally, we summarize the main conclusions in Section 5.

2 FORMALISM

In this section, we first introduce the GPR formalism and then proceed to describe the application of this technique to foreground modelling and removal in 21-cm signal observations.

2.1 Gaussian Process

A GP is a probability distribution over functions (Rasmussen & Williams 2005; Gelman et al. 2014). It constitutes the generalization of the Gaussian distribution of random variables or vectors, into the space of functions. A GP $f \sim \mathcal{GP}(m, \kappa)$ is fully defined by its mean m and covariance function κ (also called 'kernel') so that any set of points $\mathbf{x}$ in some continuous input space is associated with normally distributed random variables $\mathbf{f} = f(\mathbf{x})$, with mean $m(\mathbf{x})$ and where the value of κ specifies the covariance between the function values at any two points. The GP is the joint distribution of all those random variables which all share the desired covariance properties,

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})). \tag{1}$$

with $K(\mathbf{x}, \mathbf{x})$ an $n \times n$ covariance matrix with element (p, q) corresponding to $\kappa(x_\mathrm{p}, x_\mathrm{q})$.

In GPR, we seek a function $f(\mathbf{x})$ that would model our noisy observation $\mathbf{d} = f(\mathbf{x}) + \mathbf{n}$, where $\mathbf{n}$ is a Gaussian distributed noise with variance σ_n^2, observed at the data points $\mathbf{x}$. Given a GP prior $\mathcal{GP}(m, \kappa)$, the joint density distribution of the observations $\mathbf{d}$ and the predicted function values $\mathbf{f}' = f(\mathbf{x}')$ at a set of points $\mathbf{x}'$ is,

$$\begin{bmatrix} \mathbf{d} \\ \mathbf{f}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}') \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I & K(\mathbf{x}, \mathbf{x}') \\ K(\mathbf{x}', \mathbf{x}) & K(\mathbf{x}', \mathbf{x}') \end{bmatrix} \right). \tag{2}$$

where I is the identity matrix. Conditioning the joint prior distribution on the observations, we obtain the joint posterior distribution of our model at data points $\mathbf{x}'$,

$$\mathbf{f}'|\mathbf{x}, \mathbf{d}, \mathbf{x}' \sim \mathcal{N}\left(\mathrm{E}(\mathbf{f}'), \mathrm{cov}(\mathbf{f}')\right), \tag{3}$$

where E(.) and cov(.) are the standard notations for the mean and covariance, respectively, and with,

$$\begin{aligned} \mathrm{E}(\mathbf{f}') &= m(\mathbf{x}') + K(\mathbf{x}', \mathbf{x}) \left[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I\right]^{-1} (\mathbf{d} - m(\mathbf{x}')) \\ \mathrm{cov}(\mathbf{f}') &= K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', \mathbf{x}) \left[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I\right]^{-1} K(\mathbf{x}, \mathbf{x}'). \end{aligned} \tag{4}$$

The function values $\mathbf{f}'$ can then be sampled from the joint posterior distribution by evaluating the mean and covariance matrix above, the mean being the maximum a-posterior (MAP) solution. GPR can be seen as a fitting method in which we assign prior information on the function values of the model in the form of a covariance function. The results are marginalized over all functions drawn from the probability distribution function (PDF) in equation (3), unlike parametric modelling where the model family is fixed and one only marginalizes over the parameters.

While we assume here a data model with Gaussian noise, GP could be used in theory as priors associated with other likelihood functions, such as a Poisson likelihood (Diggle, Moyeed & Tawn 1998) or a Student-t likelihood (Neal 1997). Even with current Gaussian data model, the predictive mean of the posterior PDF (equation 4) is not required to be Gaussian distributed over the data points $\mathbf{x}$, enabling one to model non-Gaussian variation.

2.2 Covariance functions

The covariance function κ determines the structure that the GP will be able to model. A common class of covariance functions is the Matern class (Stein 1999). It is defined by,

$$\kappa_\mathrm{Matern}(x_\mathrm{p}, x_\mathrm{q}) = \frac{2^{1-\eta}}{\Gamma(\eta)} \left(\frac{\sqrt{2\eta} r}{l}\right)^\eta K_\eta \left(\frac{\sqrt{2\eta} r}{l}\right), \tag{5}$$

where $r = |x_\mathrm{q} - x_\mathrm{p}|$ and K_η is the modified Bessel function of the second kind. Functions obtained with this class of kernel are at least η-times differentiable. The kernel is also parametrized by the 'hyper parameter' l, which is the characteristic coherence scale. It denotes the distance in the input space after which the function values change significantly and thus defines the 'smoothness' of the function. Special cases of this class are obtained by setting η to ∞, in which case we obtain a Gaussian kernel, and by setting $\eta = 1/2$, in which case we obtain an exponential kernel. Throughout the paper, we use the functional form in equation (5) because of its flexibility. Importantly, if the observation we seek to model is composed of multiple additive sources, a GP model kernel can be the addition of their covariance functions. It is then possible to separate the contribution of the different terms.

We show in Appendix A that GPR can be formulated as a linear regression problem where one models the data $\mathbf{d}$ as $\mathbf{d} = \mathbf{H}\mathbf{f} + \mathbf{n}$,

where **f** are the weights of the basis functions and **n** is the noise contribution. In general, this is an ill-posed problem and one needs to set additional prior or constraints on **f**. Usually, in GPR, the constraint is statistical and set in the form of covariance matrix which can be modelled as a sum of covariance functions corresponding to the signals from the EoR, foregrounds, and noise.

2.3 Covariance function optimization

Model selection in the context of GPR is a twofold process. The first choice is that of the type of covariance function that could model the data, and the second is that of optimizing the 'hyper parameters' of this covariance function. Both can be done in a Bayesian framework, selecting the model that maximizes the marginal-likelihood, also called the evidence. This is the integral of the likelihood times the prior

$$p(\mathbf{d}|\mathbf{x},\theta) = \int p(\mathbf{d}|\mathbf{f},\mathbf{x},\theta)p(\mathbf{f}|\mathbf{x},\theta)\mathrm{d}\mathbf{f}, \tag{6}$$

with θ being the hyper parameters of the covariance function κ. Under the assumption of Gaussianity, we can integrate over **f** analytically, yielding the log-marginal likelihood (LML),

$$\log p(\mathbf{d}|\mathbf{x},\theta) = -\frac{1}{2}\mathbf{d}^{\intercal}(K+\sigma_{\rm n}^2 I)^{-1}\mathbf{d} - \frac{1}{2}\log|K+\sigma_{\rm n}^2 I| - \frac{n}{2}\log 2\pi \tag{7}$$

where we have used the short-hand $K \equiv K(\mathbf{x},\mathbf{x})$ and with n the number of sampled points. The posterior probability density of the hyper parameters is then found by applying Bayes' theorem:

$$\log p(\theta|\mathbf{d},\mathbf{x}) \propto \log p(\mathbf{d}|\mathbf{x},\theta) + \log p(\theta). \tag{8}$$

We may then either select the model that maximizes equation (7, maximum-likelihood estimate), or incorporate prior information on the hyper parameters and maximize equation (8, MAP estimate). The marginal likelihood does not only favour the models that fit best the data, overly complex models are also disfavoured (Rasmussen & Williams 2005). Selecting the values of θ that maximizes the LML is a non-linear optimization problem. Because the covariance function is defined analytically, it is trivial to compute the partial derivatives of the marginal likelihood with respect to the hyper parameters, which allow the use of efficient gradient-based optimization algorithm.

2.4 GPR for 21-cm signal detection

In the context of 21-cm signal detection, we are interested in modelling our data **d** observed at frequencies ν by a foreground, a 21-cm and a noise signal **n**:

$$\mathbf{d} = f_{\rm fg}(\nu) + f_{21}(\nu) + \mathbf{n}. \tag{9}$$

To separate the foreground signal from the 21-cm signal, we can exploit their different frequency behaviour: the 21-cm signal is expected to be uncorrelated on scales of a few MHz, while the foregrounds are expected to be smooth on that scale. The covariance function of our GP model can then be composed of a foreground covariance function $K_{\rm fg}$ and a 21-cm signal covariance function K_{21},

$$K = K_{\rm fg} + K_{21}. \tag{10}$$

The aim behind including explicitly a 21-cm signal component is not so much to model it but to isolate its covariance contribution from the covariance of the foregrounds. A complete model is also necessary to insure accurate estimation of the error covariance matrix. We can now write the joint probability density distribution of the observations **d** and the function values $\mathbf{f}_{\rm fg}$ of the foreground model $f_{\rm fg}$ at the same frequencies ν:

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{f}_{\rm fg}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}(K_{\rm fg}+K_{21})+\sigma_{\rm n}^2 I & K_{\rm fg}\\ K_{\rm fg} & K_{\rm fg}\end{bmatrix}\right). \tag{11}$$

Here again, we use the shorthand $K \equiv K(\nu,\nu)$. We note that we use a GP prior with a zero mean function, which is common practice in GPR (Rasmussen & Williams 2005; Gelman et al.) and allows the foregrounds to be fully defined by its covariance function. We tested the algorithm with a zero mean function and a polynomial parametric mean function and found the former to be a better choice for our application.

The selection of a covariance function for the 21-cm signal can be done by comparison to a range of 21-cm signal simulations. In Fig. 2, we show the covariance as a function of frequency difference $\Delta\nu$ of a 21-cm signal, calculated with 21CMFAST (Mesinger, Furlanetto & Cen 2011) when compared to the Matern $\eta = 1/2$ covariance functions for various values of the frequency coherence-scale l. For this particular set of simulations, the 21-cm signal can be well modelled using an exponential ($\eta = 1/2$) kernel with a frequency coherence scale ranging between 0.3 and 1.2 MHz depending on the reionization stage. The foregrounds need to be modelled by a smoother function. The Gaussian kernel ($\eta = \infty$) yields very smooth models which might be unrealistic for modelling physical processes and a better alternative may be a Matern kernel with $\eta = 5/2$ or 3/2. Ultimately, the choice of the foreground covariance function is driven by the data in a Bayesian sense, by selecting the one that maximizes the evidence. Because the 21-cm signal is faint compared to the foregrounds and the noise, finding the correct hyper parameters of the 21-cm signal would be close to impossible if this were done on each spatial line of sight individually. We therefore first optimize the LML for the full set of visibilities, assuming the frequency coherence scale is spatially invariant. This determines the covariance matrix structure that we then use to model the data for each spatial line of sight separately. This way we find that it is possible to perform much deeper modelling and reach the level of the 21-cm signal.

After GPR, we retrieve the foregrounds part of the model:

$$E(\mathbf{f}_{\rm fg}) = K_{\rm fg}\left[K+\sigma_{\rm n}^2 I\right]^{-1}\mathbf{d} \tag{12}$$

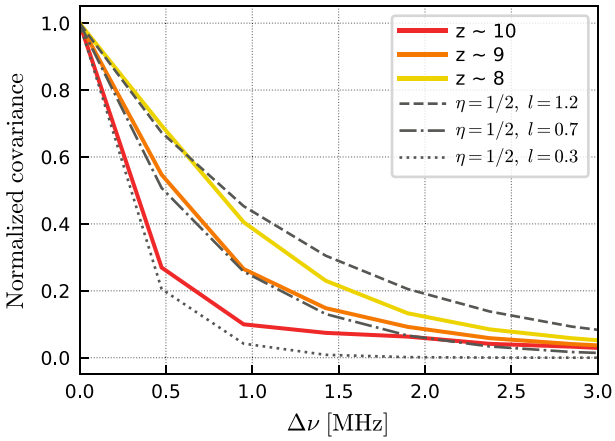

Figure 2. Exponential covariance functions for different values of the coherence-scale l (grey lines), compared to the covariance of a simulated 21-cm EoR signal at different redshift (coloured lines).

$$\mathrm{cov}(\mathbf{f}_{\rm fg}) = K_{\rm fg} - K_{\rm fg}\left[K + \sigma_{\rm n}^2 I\right]^{-1} K_{\rm fg}. \quad (13)$$

We are interested in estimating the residual after foregrounds are subtracted,

$$\mathbf{d}_{\rm res} = \mathbf{d} - E(\mathbf{f}_{\rm fg}). \quad (14)$$

3 SIMULATION

In this section, we describe the simulated astrophysical diffuse foregrounds, 21-cm EoR signal, instrumental mode-mixing contaminants and noise that are used to test the performance of the GPR foregrounds. Bright unresolved sources are not included in the simulation, assuming they can be properly modelled and subtracted from the data.

3.1 21-cm EoR signal

We use the semi-analytic code 21cmFAST (Mesinger & Furlanetto 2007; Mesinger et al. 2011) to simulate 21-cm signal corresponding to the FoV of one LOFAR–HBA station beam. The code treats physical processes with approximate methods, and it is therefore computationally much less expensive than full radiative transfer simulations. The semi-analytic codes generally agree well with hydrodynamical simulations for comoving scales >1 Mpc. We use the same 21-cm signal simulation as described in Chapman et al. (2012) and further used in Ghosh et al. (2015, 2018), which was initialized with 1800^3 dark matter particles at $z = 300$. The velocity fields were calculated on a grid of 450^3 which was used to perturb the initial conditions and the simulation boxes of the 21-cm brightness temperature fluctuations. A minimum virial mass of $10^9\,\mathrm{M}_\odot$ was defined for the haloes contributing to ionizing photons. Once the evolved density, velocity, and ionization fields have been obtained, 21cmFAST computes the $\delta T_{\rm b}$ fluctuations at each redshift. For further details of the simulation, we refer the reader to Chapman et al. (2012).

Fig. 2 shows that to first order the 21-cm signal can be approximated and modelled by a GP with an exponential covariance function, and that the frequency coherence scale is a function of redshift i.e. of the stage of reionization. The coherence scale of fluctuations in frequency of the mode-mixing contaminants and of the 21-cm signal can affect the GPR method. To test this, we also generate 21-cm signal via a GP with an exponential kernel for which we vary the frequency coherence-scale l_{21} between 0.3 and 1.2 MHz. This range should cover a wide range of possible 21-cm signal models during the EoR.

3.2 Astrophysical diffuse foregrounds

We use the foreground simulation from Jelić et al. (2008, 2010). The Galactic foregrounds have three main contributions:

(i) The largest contribution (70 per cent around 100–200 MHz) comes from the Galactic diffuse synchrotron emission (GDSE) due to the interaction of cosmic ray electrons with the galactic magnetic field.

(ii) The next contribution is coming from synchrotron emission from extended sources, mostly supernova remnants.

(iii) The final component is the free–free radio emission from diffuse ionized gas which contributes roughly 1 per cent to the total Galactic foreground emission.

The individual Galactic foreground components are modelled as Gaussian random fields. The GDSE is modelled as a power law as a function of frequency with a spectral index of -2.55 ± 0.1 (Shaver et al. 1999) and -2.15 for the free–free emission. We have not included polarization of the foregrounds in our simulation. We also assume that point sources brighter than 0.1 mJy can be identified and accurately removed from the maps and therefore these sources are not included in the current diffuse foreground simulation (Jelić et al. 2008). Unresolved extragalactic sources were added to the simulation based on radio source counts at 151 MHz (Jackson 2005). The simulated radio galaxies are clustered using a random walk algorithm.

3.3 Instrumental mode-mixing contaminants

The source of mode-mixing contaminants are manifold (Section 1). In essence, they are due to the combination of the instrument chromaticity and imperfect calibration. In the present paper, we will not attempt to simulate those effects, and we defer that to a future publication. Instead, we will simulate them using a GP. This treatment is motivated by the analysis of LOFAR data which shows that these medium-scale fluctuations can be well modelled by a GP with a Matern covariance function, $\eta = 3/2$ and a coherence-scale $l_{\rm mix} \sim 2$ MHz.[6] In Section 4.4, we will test GPR against others methods to generate mode-mixing contaminants using random polynomials and Matern kernel with different hyper parameters for the different baselines.

The mode mixing is usually confined to a wedge-like structure in k space (Datta et al. 2010; Morales et al. 2012) (see Fig. 1). In the present publication, we do not simulate the $k_\perp$ dependence of the wedge and also defer this to future work. In fact, current assessments of the mode-mixing contaminants in LOFAR data tend to favour a baseline independent 'brick' effect observed in Ewall-Wice et al. (2017), which probably comes mainly from transferring the gain errors from longer to shorter baselines (Barry et al. 2016; Patil et al. 2016). For the purpose of testing the impact of the 'brick' extent, we simulate instrumental mode-mixing contaminants with frequency coherence-scale $l_{\rm mix}$ varying between 1 and 8 MHz.

3.4 Noise

In order to obtain realistic simulations of the noise, we first compute weights maps $W(u, v, \nu)$ which reflect the baseline distribution in the gridded uv plane. A noise visibility cube is created by filling it with random Gaussian noise for the real and imaginary parts of the visibility separately with a noise standard deviation,

$$\sigma(u, v, \nu) = \frac{1}{\sqrt{W(u, v, \nu)}} \frac{\mathrm{SEFD}(\nu)}{\sqrt{2\,\Delta\nu\,\Delta t}}, \quad (15)$$

where $\Delta\nu$ and Δt are the frequency bandwidth and integration time, respectively, and the SEFD is the system equivalent flux density. We note that the SEFD is generally frequency dependent and varies across the sky. The SEFD depends largely on the sky temperature ($\mathrm{T}_{\rm sky} \propto \nu^{-2.55}$) of the total sky brightness and the effective area of the LOFAR array ($A_{\rm eff}$). Here, we assume a constant SEFD $\sim$4000 Jy (van Haarlem et al. 2013) over the simulated bandwidth, and assume a LOFAR–HBA data set of about 100 nights of 12 h long observations.

[6] A more detailed description of LOFAR–HBA mode-mixing modelling will be given in a forthcoming publication. We refer the reader to Patil et al. (2016, 2017) for a recent analysis of this contaminants.

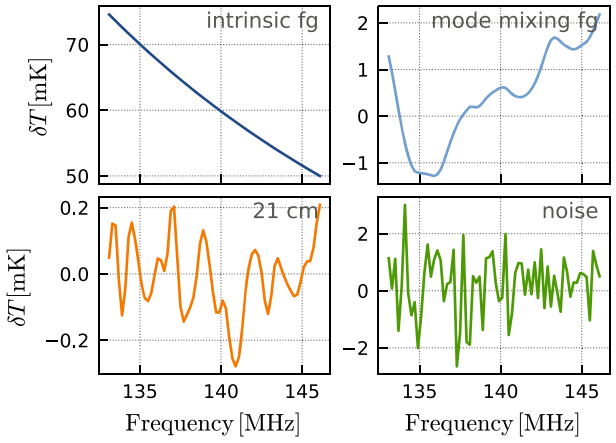

Figure 3. The different components of the simulated signal. The astrophysical diffuse emission (top left panel), instrumental mode-mixing contaminants (top right panel), 21-cm signal (bottom left panel), and noise component (bottom right panel) of a randomly selected visibility from the simulated cube is plotted as a function of frequency.

3.5 Simulation cube

The simulation spans a frequency range of 132–148 MHz with a spectral resolution of 0.2 MHz, i.e. a bandwidth of 16 MHz and 80 sub-bands, from which 12 MHz are used for power-spectra calculation centred around a redshift of $z\sim 9.1$. The maps cover a FoV of 6 deg with a pixel size of 1.17 arcminute. The mean value of the brightness temperature is subtracted to mimic a typical interferometric observation. The intrinsic foreground, mode-mixing, 21-cm signal and noise, respectively, of the simulation are converted into visibilities via a Fourier transform and added together to create an observation cube:

$$V_{\rm obs}(\boldsymbol{u},\nu) = V_{\rm sky}(\boldsymbol{u},\nu) + V_{\rm mix}(\boldsymbol{u},\nu) + V_{21}(\boldsymbol{u},\nu) + V_{\rm n}(\boldsymbol{u},\nu), \quad (16)$$

where $\boldsymbol{u} = (u, v)$ is the vector representing the coordinates in wavelength in the uv plane and ν is the observing frequency. We restrict our analysis to the baseline range $50 - 250\,\lambda$ currently used by LOFAR (Patil et al. 2017). An example of these components are shown in Fig. 3 as a function of frequency. The distinct frequency correlation is the characteristic exploited in the GPR method to separate these signals. We note that the signal separation (in this case foreground) method could be applied equally well to visibilities, image pixels, or spherical harmonics coefficients (Ghosh et al. 2018).

The simulation cube is parametrized by four main parameters:

$\sigma_{21}/\sigma_{\rm n}$: The ratio between the standard deviation of the 21-cm signal cube and the standard deviation of the noise cube, for the $50 - 250\,\lambda$ baselines range. This allows to test different reionization scenario while keeping the same noise level, and vice versa.

l_{21}: The frequency coherence scale of the exponential covariance kernel in the case when a GP is used to simulate the 21-cm signal. This parameter is ignored when 21cmFAST is used instead.

$\sigma_{\rm mix}/\sigma_{\rm n}$: The ratio between the standard deviation of the instrumental mode-mixing contaminants cube and the standard deviation of the noise cube, for the $50 - 250\,\lambda$ baselines range.

$l_{\rm mix}$: The frequency coherence scale of the Matern covariance kernel used to simulate the instrumental mode-mixing contaminants.

4 RESULTS

In the following section, the GPR procedure described in Section 2 is applied to the simulated data sets described in Section 3, in order to model and remove the foreground components, and subsequently compute the power spectrum of the 21-cm signal. Specifically, we apply the method on simulated cubes which reproduce the level of noise, mode-mixing contaminants, and foregrounds diffuse emission that we currently or theoretically can achieve with LOFAR, and subsequently explore various values of simulation parameters.

4.1 Recovering the 21-cm signal power spectra

4.1.1 Foregrounds modelling and removal

The simulated foregrounds cube is composed of a frequency smooth sky signal and less smooth mode-mixing contaminants. We build this property into our GP covariance function by decomposing our foregrounds covariance into two separate parts,

$$K_{\rm fg} = K_{\rm sky} + K_{\rm mix} \quad (17)$$

with 'sky' denoting the intrinsic sky and 'mix' denoting the mode-mixing contaminants. We use a Matern covariance function for all components of our data GP model. A Matern kernel has three hyper parameters, l, σ, and η. The function becomes especially simple when η is half integer (Rasmussen & Williams 2005), which is why only discrete values of η are used, $\eta \in (1/2, 3/2, 5/2, 7/2)$, choosing the best value based on the LML. This reduces the numbers of hyper parameters to be optimized to six (two for each of the intrinsic sky, mode-mixing and 21-cm components of the GP model). We use the PYTHON package GPY[7] to do the optimization using the full set of visibilities. This is done in two steps. We first use a uniform prior on the hyper parameters and test different values of η, selecting the model that maximizes the evidence. A final run is then done with a more restricted range for the hyper parameters. The foreground subtracted visibility is then obtained by computing the residual:

$$V_{\rm res}(\boldsymbol{u},\nu) = V_{\rm obs}(\boldsymbol{u},\nu) - V_{\rm fg}^{\rm rec}(\boldsymbol{u},\nu), \quad (18)$$

where $V_{\rm fg}^{\rm rec}(\boldsymbol{u},\nu)$ is the MAP GPR foregrounds model.

We recollect that for this particular set of simulations, the 21-cm signal was modelled using an exponential ($\eta = 1/2$) kernel with a frequency coherence scale ranging between 0.3 and 1.2 MHz. For foregrounds, we choose a Matern kernel with $\eta = 5/2$ or 3/2. Ultimately, the choice of the foreground covariance function is driven by the data in a Bayesian sense, by selecting the one that maximizes the evidence. Because the 21-cm signal is faint compared to the foregrounds and the noise, we therefore first optimize the LML for the full set of visibilities, assuming the frequency coherence scale is spatially invariant. In this way, we determine the covariance matrix structure that we then use to model the data for each spatial line of sight separately. In GPR, we retrieve the foregrounds part of the model first using equation (12) and the residuals were subsequently calculated using equation (14).

4.1.2 Power-spectrum estimation

Next, we determine the power spectra to quantify the scale-dependent second moment of the signal by taking the Fourier transform of the various visibility cubes $V(\boldsymbol{u},\nu)$ in the frequency direction. We define the cylindrically averaged power spectrum as

[7] https://sheffieldml.github.io/GPy/

(Parsons et al. 2012):

$$P(k_\perp, k_\parallel) = \frac{X^2 Y}{\Omega_{\rm PB} B} \left\langle \left| \hat{V}(\boldsymbol{u}, \tau) \right|^2 \right\rangle, \tag{19}$$

where $\hat{V}(\boldsymbol{u}, \tau)$ is the Fourier transform in the frequency direction, B is the frequency bandwidth, $\Omega_{\rm PB}$ is the PB FoV, X and Y are conversion factors from angle and frequency to comoving distance, and $< .. >$ denote the averaging over baselines. The Fourier modes are in units of inverse comoving distance and are given by (Morales, Bowman & Hewitt 2006; Trott, Wayth & Tingay 2012):

$$k_\perp = \frac{2\pi |\boldsymbol{u}|}{D_{\rm M}(z)}, \tag{20}$$

$$k_\parallel = \frac{2\pi H_0 \nu_{21} E(z)}{c(1+z)^2} \tau, \tag{21}$$

$$k = \sqrt{k_\perp^2 + k_\parallel^2}, \tag{22}$$

where $D_{\rm M}(z)$ is the transverse comoving distance, H_0 is the Hubble constant, ν_{21} is the frequency of the hyperfine transition, and $E(z)$ is the dimensionless Hubble parameter (Hogg 1999). Finally, we average the power spectrum in spherical shells and define the spherically averaged dimensionless power spectrum as,

$$\Delta^2(k) = \frac{k^3}{2\pi^2} P(k). \tag{23}$$

The recovered 21-cm signal power spectrum is obtained by subtracting the noise bias from the residual power spectra, derived from the residuals in equation (18). In general, the noise bias can be estimated with reasonable accuracy from the Stokes V image cube (circularly polarized sky), or by taking the difference between Stokes I data separated by a small frequency or time interval. The sky is only weakly circularly polarized and the Stokes V image cube is expected to provide a good estimator of the thermal noise. In our simulation, the noise bias is estimated using the same noise cube used to generate the simulation cube. This ensures that the variance in the recovered 21-cm signal that we estimate are inherent to GPR and not due to thermal noise sampling variance limitations.

4.2 Application on the reference simulation

Our reference simulation is representative of the capability of LOFAR–HBA based on current observation of the noise and the level of mode-mixing errors. Specifically, the foregrounds data cube is composed of diffuse emission foreground and instrumental mode-mixing contaminants simulated using a Matern $\eta_{\rm mix}= 3/2$ covariance function with frequency coherence scale of $l_{\rm mix} = 2$ MHz and a variance $(\sigma_{\rm mix}/\sigma_{\rm n})^2 = 2$. The 21-cm signal is simulated from 21cmFAST with a variance $(\sigma_{21}/\sigma_{\rm n})^2= 0.007$. The noise realization corresponds to 1200 h of LOFAR-HBA observations and an SEFD = 4000 K. The input parameters of the reference simulation are summarized in Table 1.

4.2.1 *Power-spectrum results*

We generate a total of 200 simulations, each with different noise and instrumental mode-mixing contaminants realizations, but with exactly the same astrophysical foregrounds and 21-cm signal. The power spectra of the different components are shown in Fig. 4. The top panel shows the spherically averaged power spectra. The intrinsic foregrounds are orders of magnitude brighter than the 21-cm signal on large scales (small k), but drop below the 21-cm signal at $k > 0.3\ h\,{\rm cMpc}^{-1}$. While the mode-mixing component is only a small percent of the total power, it occupies a wider range of k modes. This is better understood when looking at the cylindrically averaged power spectra as a function of $k_{\shortparallel}$ (bottom panel in Fig. 4); while most of the power of the intrinsic foregrounds is concentrated at low $k_{\shortparallel}$, the mode-mixing components still dominate the 21-cm signal at large $k_{\shortparallel}$, due to their smaller coherence in the frequency direction. This illustrates the importance of adding mode mixing to any foreground removal strategy. We note that the k mode at which the foreground power steeply decreases depends on the maximum baseline considered for the analysis. For this baseline configuration, we also note that a characterization of the power spectra is theoretically possible for $k \leq 0.3$ h cMpc^{-1} assuming perfect foreground removal and considering only the thermal noise uncertainty on the 21-cm signals (see also Fig. 6).

Table 1. Summary of the input parameters of the reference simulation and estimate on the median and confidence interval of their respective GP model hyper parameters obtained using an MCMC method. The input intrinsic sky is simulated using astrophysical foreground simulation from Jelić et al. (2008), while the 21-cm signal is simulated from 21cmFAST (Mesinger et al. 2011).

	Input	Prior	Estimate
$\sigma_{\rm sky}/\sigma_{\rm n}$	–	$\mathcal{U}(30, 45)$	$37.4^{+0.4}_{-0.4}$
$l_{\rm sky}$ (MHz)	–	$\mathcal{U}(60, 100)$	$80.1^{+1.2}_{-1.2}$
$\sigma_{\rm mix}/\sigma_{\rm n}$	1.478	$\mathcal{U}(1, 2)$	$1.47^{+0.01}_{-0.01}$
$l_{\rm mix}$ (MHz)	2	$\mathcal{U}(1.5, 2.5)$	$2.01^{+0.02}_{-0.02}$
$\sigma_{21}/\sigma_{\rm n}$	0.083	$\mathcal{U}(0.002, 0.25)$	$0.11^{+0.03}_{-0.04}$
l_{21} (MHz)	–	$\Gamma(3.6, 4.2)$	$0.90^{+0.05}_{-0.04}$

The initial GPR runs with uniform priors on all hyper parameters reveal that, in about 40 per cent of the cases, the 21-cm coherence-scale hyper-parameter l_{21} converges to the prior higher bound. A more informative prior can be used to solve this issue and better constrain l_{21}. Fig. 2 shows that the simulated 21-cm signal coherence scales range between about 0.3 and 1.2 MHz. A gamma distribution prior, thus honoring the positivity of the hyper parameter, can then be used instead of the uniform prior with a variance broad enough such that it includes all probable values. The PDF of the gamma distribution $\Gamma(\alpha, \beta)$, parametrized by the shape α and rate β, is defined as,

$$P(x|\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} {\rm e}^{-\beta x}}{\gamma(\alpha)}, \tag{24}$$

where $\gamma(\alpha)$ is the gamma function. For the hyper-parameter l_{21}, we use the $\Gamma(3.6, 4.2)$ prior which is characterized by an expectation value of 0.85, a median value of 0.77, a 16th percentile value of 0.42 and an 84th percentile value of 1.29. To test the impact of this prior on the recovery of the 21-cm signal, we perform simulations similar to the ones described above but with the 21-cm signal simulated from a GP for which we know the true value of l_{21}. We then compare the input value of l_{21} and the value estimated from the GPR. This shows that in case of a uniform prior, the values of l_{21} are not well estimated while, using a $\Gamma(3.6, 4.2)$ prior, the estimated values of l_{21} are significantly less biased and have an uncertainty of $\sim$0.2. We found that using this prior is only necessary because the reference simulation is characterized by a low signal-to-noise ratio (S/N) of the 21-cm signal and a low-frequency coherence scale of the mode-

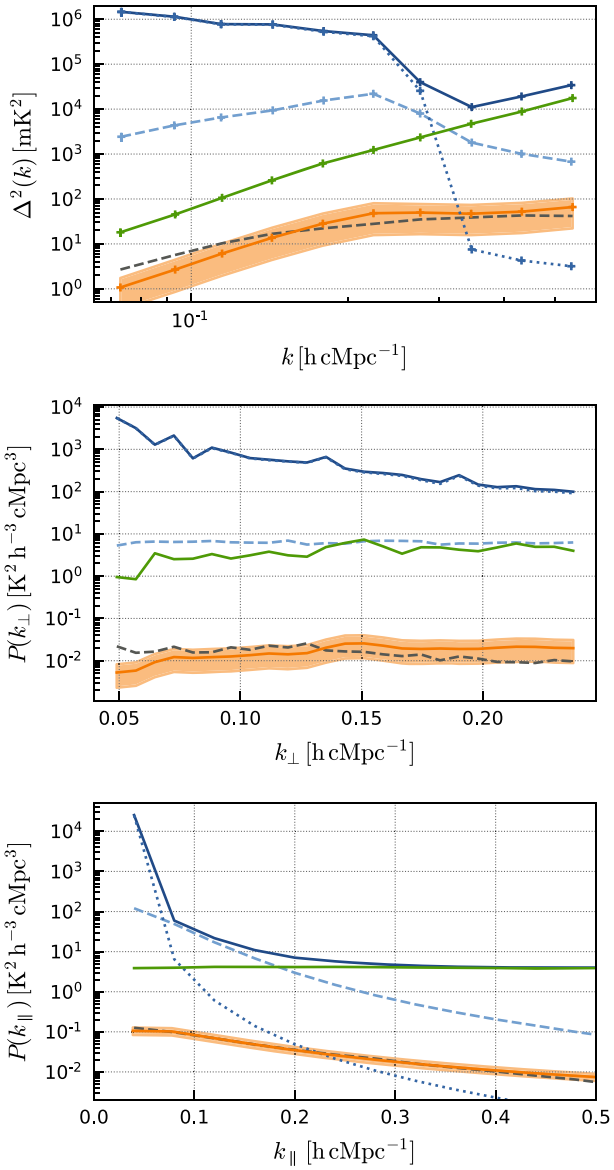

Figure 4. Detection of the EoR signal with the reference simulation. The top panel shows the spherically averaged power spectra. The central and bottom panels show the cylindrically averaged power spectra, averaged over $k_{\shortparallel}$ and over $k_\perp$ respectively. The simulated observed signal (dark blue) is composed of intrinsic astrophysical foregrounds (dotted dark blue), instrumental mode-mixing contaminants (dashed light blue), noise (green), and a simulated 21-cm signal (dashed grey). Using our GPR method to model and remove the foregrounds from the simulated cube, the 21-cm signal (orange) is well recovered with limited bias. The orange filled region represents the standard deviation of the recovered 21-cm signal over 200 simulated cubes.

mixing component. The gamma prior helps in better separating the contributions from the mode-mixing and 21-cm signal.

The initial GPR runs are also used to set the values of η of the Matern covariance function for the different GP components. We find that the evidence is maximized using $\eta_{\rm sky} = 5/2$, $\eta_{\rm mix} = 3/2$, and $\eta_{21} = 1/2$.

Having found the most probable settings of GP model and hyper-parameter priors, we perform a final GPR on each of the simulated cubes. Fig. 4 shows the power spectra of the recovered 21-cm signal compared to the input cosmological signal power spectra. The orange filled region represents the standard deviation of the recovered signal over the 200 simulated cubes, the line corresponding to the mean. This provides an estimate respectively of the variance and the bias of the method. The bias is overall limited but is more pronounced at low k modes. It is maximum at $k = 0.073\ h\,\mathrm{cMpc}^{-1}$ where we have a bias equal to 86 per cent of the uncertainty. The variance is almost always similar or below, on the k modes probed, the thermal noise limit. We however find it to be 30 per cent greater at $k = 0.18\ h\,\mathrm{cMpc}^{-1}$. We recall that the noise bias is estimated using the same noise cube used in the simulated cube. Hence, the variance that we estimate is inherent to GPR and does not include thermal noise sampling variance.

Investigating the cylindrically averaged power spectra reveals that most of the bias of the current implementation of the GPR method is introduced because of the one-dimensional fit to the data in the frequency direction. The power spectra as a function of $k_{\shortparallel}$ (bottom panel of Fig. 4) show an excellent correspondence between the input and recovered signal with small uncertainty. On the contrary, the power spectra as a function of $k_\perp$ (central panel of Fig. 4) show a much larger bias and uncertainty. The method is capable of retaining the correct variance in the frequency direction but not so well in the baseline direction. This is explained by the fact that the regression is currently only done in the frequency direction and assumes that the frequency coherence scale of the different components is spatially invariant.

In Section 5, we explore various improvements to the method that may be implemented to reduce the bias and uncertainty. Nevertheless, current results already demonstrate that the approach is able to achieve a reliable first measurement of the 21-cm signal and an initial characterization of its power spectra in 1200 h of LOFAR observations.

4.2.2 *Estimating the model hyper-parameter uncertainties*

The MAP solution of the model hyper parameters is evaluated through an optimization algorithm, using the analytically defined likelihood function (equation8). However, to fully sample the posterior distribution of the hyper parameter, characterize its topology, and analyse the correlations between parameters, we resort to Monte Carlo Markov Chain (MCMC).

An MCMC method samples the posterior probability distribution of the model parameters given the observed data. We use an ensemble sampler algorithm based on the affine-invariant sampling algorithm (Goodman & Weare 2010), as implemented in the EMCEE PYTHON package[8] (Foreman-Mackey et al. 2013). Fig. 5 shows the resulting posterior probability distribution of the GP model hyper parameters. We find that the input values are always inside the 68 per cent confidence interval. The hyper parameters of the mode-mixing covariance function are very well constrained. The confidence interval on the 21-cm signal kernel hyper parameters are relatively larger, because in this particular simulation, the 21-cm signal is an order of magnitude fainter than the noise. The parameter estimates and confidence intervals are summarized in Table 1, along with their input values and associated priors. We note that for this setup the 21-cm signal has no input l_{21} because it was simulated using 21CMFAST.

[8] http://dfm.io/emcee/current/

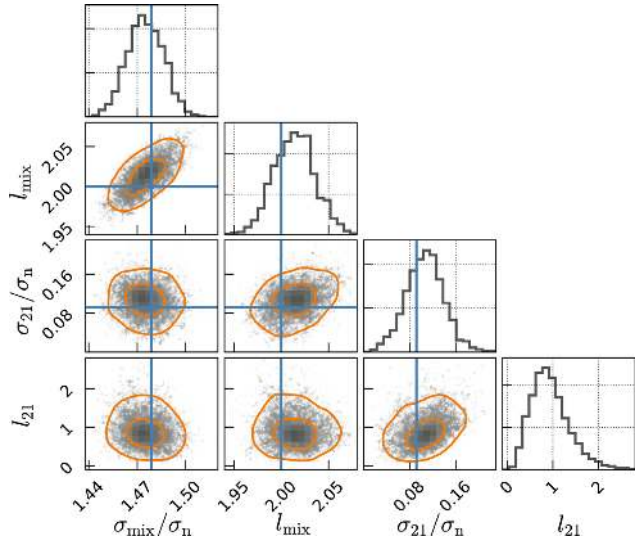

Figure 5. Posterior probability distributions of the GP model hyper parameters for the reference simulation. We show here the coherence scale and strength of the EoR covariance function (l_{21} in MHz and σ_{21}), and the coherence scale and strength of the mode-mixing foreground kernel ($l_{\rm mix}$ in MHz and $\sigma_{\rm mix}$). The input parameters of the simulation are marked in blue. The orange contours show the 68 per cent and 95 per cent confidence interval. We note that the PDFs are all narrower than their priors.

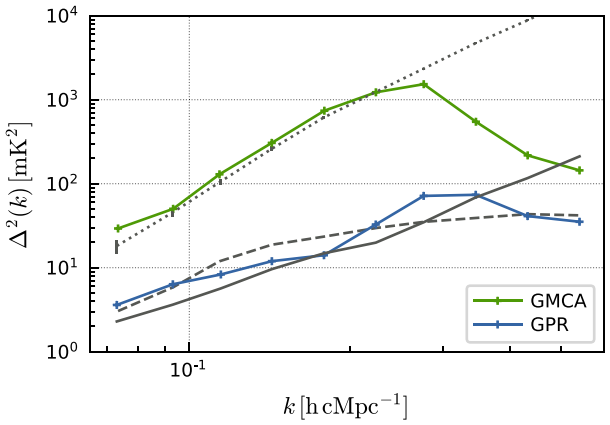

Figure 6. Spherically averaged power spectra of the foreground modelling error using the GPR and GMCA method. With GPR (blue line), the foreground error is at the level of the 21-cm signal (dashed black line) and is close to the thermal noise uncertainty (plain black line) which is the inherent statistical error level we could achieve, while the foreground error with GMCA is at the level of the noise (dotted black line).

4.2.3 Comparison between GPR and GMCA

Next, we compare GPR to another well-tested foreground removal method. From the currently available algorithms, the GMCA (Bobin et al. 2008; Chapman et al. 2013) is the one that has demonstrated the best results (Chapman et al. 2015; Ghosh et al. 2018). We use the PYTHON based toolbox PYGMCALAB[9] and run the algorithm on our simulated cubes. We model the foregrounds by the minimum numbers of components that minimize the overall fitting error. An optimal eight components are used to represent the foregrounds. We then compare the power spectra of the foreground modelling error when using GPR and GMCA. Fig. 6 shows that GMCA has difficulty to correctly model the complex mode-mixing contaminants and does not reach a level of modelling error better than the noise for $k \leq 0.3\ h\,{\rm cMpc}^{-1}$. Using GPR, we improve these results by an order of magnitude, and this allows us to achieve an error in the foreground power spectra that is at or below the 21-cm signal power spectrum. We also note that this level is similar to the thermal noise uncertainty which is the ultimate error level we can achieve.

[9] http://www.cosmostat.org/software/gmcalab

4.3 Performance of the GPR method

4.3.1 Exploring the input parameter space

The efficiency of a foreground removal algorithm depends on the characteristics of the foregrounds and of the 21-cm signal. To explore the performance of GPR in terms of bias and variance, we explore the input parameters of the simulated cube, varying one parameter at a time. As a quality criterion, we use the fractional bias of the recovered spherically averaged 21-cm signal power spectra,

$$r_{\rm rec}(k) = \frac{\Delta^2_{\rm rec}(k) - \Delta^2_{21}(k)}{\Delta^2_{21}(k)}. \tag{25}$$

where $\Delta^2_{\rm rec}(k)$ is the GPR recovered power spectrum, and $\Delta^2_{21}(k)$ is the power spectrum of the input 21-cm signal.

For these tests, we build simulation cubes with central parameters $\sigma_{\rm mix} = 1.478\sigma_{\rm n}$, $l_{\rm mix} = 3$ MHz, $\sigma_{21} = 0.12\sigma_{\rm n}$, and $l_{21} = 0.75$ MHz around which we vary the parameters. We use a GP with an exponential covariance function (see Section 3.1) to generate 21-cm signals such that we can control the frequency correlation of the signal (i.e. l_{21}). A total of 3000 simulations with different realizations of the noise, 21-cm signal, and mode-mixing contaminants are generated. We determine the relative difference between recovered and input power spectra for different k bins and compute its mean and standard deviation[10] over the full set of simulated cubes (Fig. 7). This provide us with an estimate of the fractional bias and uncertainty introduced by the method. We also compare the later to the minimal uncertainty due to thermal noise.

By varying the strength of the 21-cm signal, we find that the bias is limited (below 35 per cent) for the full range of the investigated values and falls below 20 per cent for $\sigma_{21} \geq 0.12\sigma_{\rm n}$. The uncertainty and bias increase with lower S/N as expected, and we find it to be significantly higher than the thermal noise uncertainty for $\sigma_{21} \lesssim 0.1\sigma_{\rm n}$. Varying the frequency coherence scale of the mode-mixing contaminants, we also find limited bias and a small increase of the uncertainty at low $l_{\rm mix}$. As $l_{\rm mix}$ approaches that of l_{21}, it becomes increasingly more difficult to statistically differentiate the two signals. This is the reason why the uncertainty increases for values $l_{\rm mix} < 3$ MHz. A decrease in the value of $l_{\rm mix}$ also corresponds to increasing the extent of the foreground wedge (or 'brick'), and equivalently reducing the EoR window. Varying the frequency coherence scale of the 21-cm signal, we find that some bias is introduced at small and large l_{21}, related to the use of a Gamma prior to this GP hyper parameters.

Overall, GPR is limited in situation of very low S/N and/or when the foregrounds start to mix with the 21-cm signal. In most situations, it performs relatively well, with limited bias and uncertainty level on par with the thermal noise uncertainty.

[10] We note that the distribution of $r_{\rm rec}$ is actually not Gaussian, being the ratio of two distributions, but the mean and standard deviation were found to be appropriate enough to characterize this distribution.

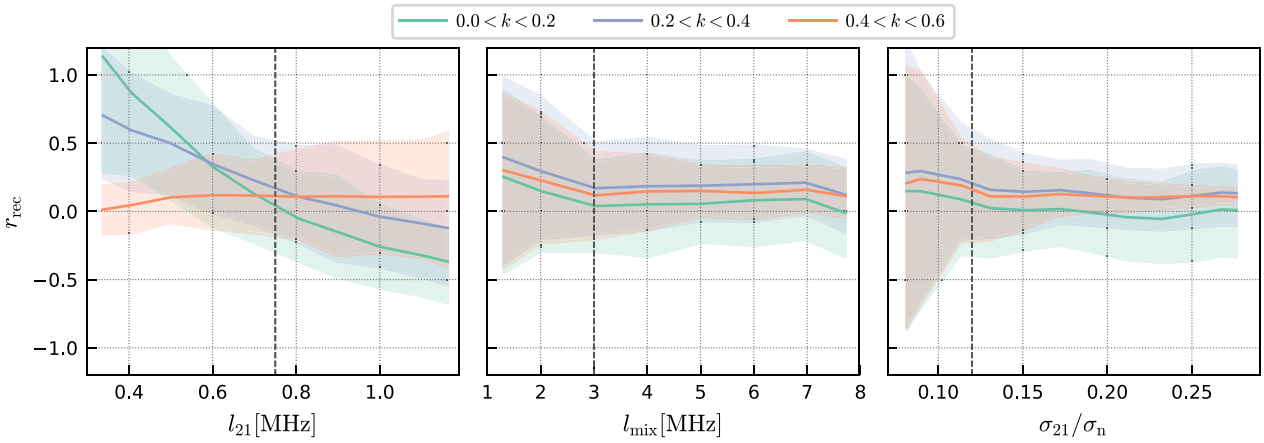

Figure 7. Fractional bias of the recovered 21-cm signal ($r_{\rm rec}$) with varying coherence scale of the 21-cm signal (l_{21}, left-hand panel), coherence scale of the mode-mixing contaminants ($l_{\rm mix}$, central panel), and strength of the 21-cm signal ($\sigma_{21}/\sigma_{\rm n}$, right-hand panel), for different k ranges. We show the mean (plain line) and standard deviation (filled area) of the fractional bias calculated from a total of 3000 simulations, giving an estimate of the bias and uncertainty, respectively, introduced by the method. GPR performance is optimal for $l_{\rm mix} > 3.0$ MHz, $\sigma_{21} \gtrsim 0.1\sigma_{\rm n}$, and for 0.6 MHz $< l_{21} <$ 1 MHz. The vertical dashed lines represent the nominal values around which we vary the parameters.

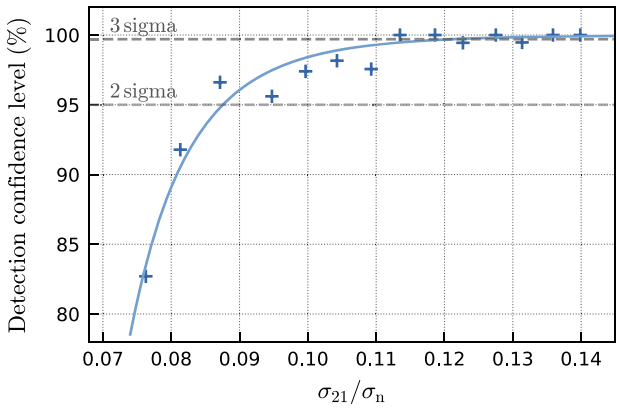

Figure 8. Detection confidence interval for the reference simulation as a function of the S/N of the 21-cm signal $\sigma_{21}/\sigma_{\rm n}$. The measured confidence levels (blue points) are fitted using an inverse function (blue line). The dashed grey lines show the 95 per cent and 99.7 per cent confidence level.

4.3.2 Detection confidence level

We define the detection confidence level as the probability that the model is preferred (i.e. the evidence is maximal) if it contains a 21-cm signal component compared to one that does not. In GPR, the evidence as a function of the hyper-parameters θ is analytically defined (equation 7) and can be efficiently estimated for the optimal values of θ. We note that comparing this maximum evidence for two different covariance structures parametrized by different numbers of hyper parameters does not usually provide definitive answer on which kernel is the most suitable to model the data, especially if the difference of the evidences is small (Rasmussen & Williams 2005; Fischer et al. 2016). Nevertheless, this criterion is fast to compute and can still provide informative approximation on the confidence level of the detection. To determine it as a function of S/N of the 21-cm signal, we generate new reference simulations, varying now the input 21-cm signal strength σ_{21}. We use equation (7) to compute the evidence for the optimal values of the hyper-parameters θ. In Fig. 8, we show the detection confidence level as a function of the input 21-cm signal $\sigma_{21}/\sigma_{\rm n}$, calculated using a total of 3000 simulations. A 95 per cent and 99.7 per cent detection confidence level is observed for $\sigma_{21} \gtrsim 0.09\,\sigma_{\rm n}$ and $\sigma_{21} \gtrsim 0.12\sigma_{\rm n}$ respectively, rapidly increasing with S/N.

The above calculation is obtained using the expression of the evidence from equation (7) which is a function of the hyper-parameters θ. A more robust way to compare the models is to estimate the evidence values integrated over the hyper parameters and take their ratio, also called the Bayes factor. This is generally much more computationally expensive, and we only perform this test, as a confirmation of the above results, for a limited number of cases. We compute the evidence with an implementation of the nested sampling algorithm of Mukherjee, Parkinson & Liddle (2006). For $\sigma_{21} = 0.083\,\sigma_{\rm n}$ (i.e. the reference simulation), we obtain Bayes factors ranging between 3.8 and 19 corresponding to a 'substantial' to 'strong' strength of evidence according to the scale of Jeffreys (1961). For $\sigma_{21} = 0.12\,\sigma_{\rm n}$, we obtain Bayes factors ranging between 5.2 and 55 corresponding to a 'substantial' to 'very strong' strength of evidence. Finally, for $\sigma_{21} = 0.2\,\sigma_{\rm n}$, we obtain a Bayes factors ranging between 328 and 1.9×10^4 corresponding to a 'decisive' strength of evidence.

We note that these estimates are only for a single frequency bandwidth of 12 MHz, and that usually several redshift bins are combined which will increase the overall confidence level on the detection of the 21-cm signal.

4.4 Testing different methods of simulating mode mixing

In this sub-section, we test the versatility of GPR against alternative form of mode-mixing contaminants. In previous simulations, we used a Matern kernel with fixed coherence scale. We now perform similar simulation with three others methods to generate the instrumental mode-mixing components. The simulation cubes are generated with parameters $\sigma_{\rm mix} = 1.478\sigma_{\rm n}$, $\sigma_{21} = 0.12\sigma_{\rm n}$, and $l_{21} = 0.75$ MHz.

4.4.1 Random polynomial

We generate mode-mixing visibilities using polynomial functions of random order taken in the range 3 –$o_{\rm max}$ and random coefficients. Applying GPR to this simulation shows that this component is best modelled (i.e. the evidence is maximized) using a Matern covariance

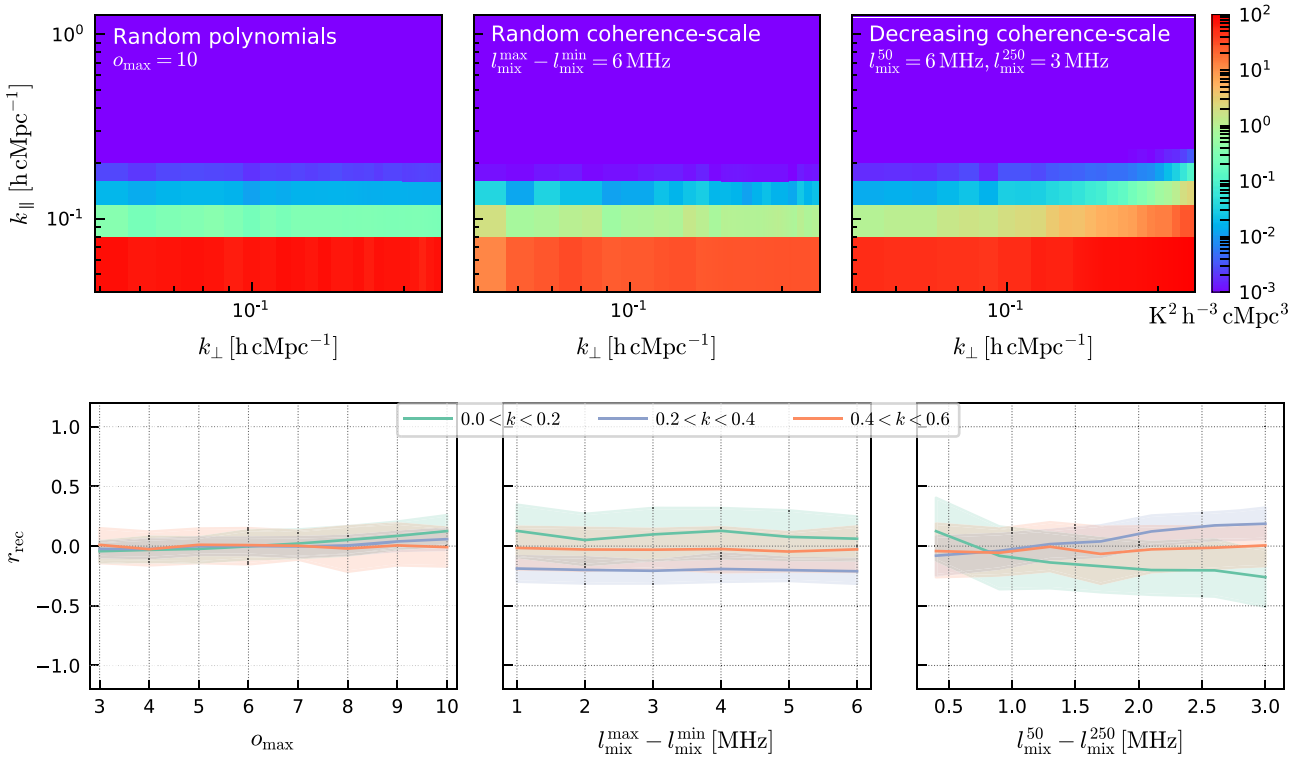

Figure 9. Fractional bias of the recovered 21-cm signal (r_{rec}) for mode-mixing contaminants generated using random polynomials with maximum order $o_{\max}$ (left), GP with random coherence scale selected in the range $l_{\mathrm{mix}}^{\max} - l_{\mathrm{mix}}^{\min}$ with $l_{\mathrm{mix}}^{\min} = 3$ (middle), and GP with decreasing coherence scale as function of baseline length with $l_{\mathrm{mix}}^{50} = 6\,\mathrm{MHz}$ (right). The top panel shows the two-dimensional cylindrically averaged power spectra of the different mode-mixing contaminants for the most extreme tested scenarios (the axis are in log scale). The bottom panel show the mean (plain line) and standard deviation (filled area) of the fractional bias calculated from a total of 1000 simulations.

function with $\nu = \infty$ (equivalent to a Gaussian covariance kernel). The results of this test are shown in the left-hand panel of Fig. 9. The measured bias is minimal for all tested cases.

4.4.2 *Random coherence scale*

We now generate mode-mixing visibilities using a Matern kernel and randomly selected coherence-scale l_{mix} in the range $3 - l_{\mathrm{mix}}^{\max}$ MHz for each different visibilities modes $\boldsymbol{u}$. For this test, we set $\nu = \infty$. Running GPR on this simulation shows that the mode-mixing component is best modelled by a Rational Quadratic covariance function which is defined as:

$$\kappa_{\mathrm{RQ}}(x_{\mathrm{p}}, x_{\mathrm{q}}) = \left(1 + \frac{|x_{\mathrm{q}} - x_{\mathrm{p}}|^2}{2\alpha l}\right)^{-\alpha}, \tag{26}$$

and can be seen as an infinite sum of Gaussian covariance functions with different characteristic coherence scales (Rasmussen & Williams 2005). The results of this test is shown in the middle panel of Fig. 9. We again find limited bias which is also independent of the range of coherence scales.

4.4.3 *Decreasing coherence scale*

A wedge-like feature can be simulated by generating mode-mixing visibilities with decreasing coherence scale as a function of baseline. For this test, we use a Matern kernel with $\nu = \infty$, and a coherence scale that is linearly decreasing as a function of baseline with the coherence scale of the 50 lambda baselines $l_{\mathrm{mix}}^{50} = 6\,\mathrm{MHz}$, and the coherence scale of the 250 lambda baselines l_{mix}^{250} taken in the range $3 - 5.5\,\mathrm{MHz}$. The result of this test is shown in the right-hand panel of Fig. 9. It shows that an increase of the bias with increasing range of coherence scales. The maximum bias is nevertheless limited to about 30 per cent.

In the future, we will implement the ability of GPR to perform a fit of the hyper parameters with different coherence scale for different baselines ranges, which should reduce further this bias.

5 DISCUSSION AND CONCLUSION

In this paper, we have introduced a novel signal separation method for EoR and CD experiments. The method uses GPR to model various mixed components of the observed signal, including the spectrally smooth sky, mode mixing associated with the instrument chromaticity and imperfect calibration, and a 21-cm signal model. Including covariance functions for each of these components in the GPR ensures a relatively unbiased separation of their contribution and accurate uncertainty estimation, even in very low signal-to-noise observations.

In building the GP model, we make use of prior information about the different components of the signal. This makes the method very useful in the initial diagnostic and analysis stage of the data processing as it allows one to get a better insight into the data in terms of potential contaminants (i.e. mode mixing but also the ionosphere). Additionally, GPR is flexible, and the GP model can be easily adapted to integrate new systematics. Cable reflexion, for example, could be easily modelled in this framework, adding a periodic covariance function component to the model.

GPR is shown to accurately model the foreground contaminants including instrumental mode mixing which have proven to be an Achilles heel of current foreground removal algorithms. When applied to simulation data sets, equivalent to LOFAR 1200 h of observations and based on its current assessment of noise and systematic errors, GPR limits biasing the 21-cm signal, and recovers the input power spectrum well across the whole k range $0.07 - 0.3\,h$ cMpc^{-1}. When compared to GMCA, we find that GPR decreases the uncertainty on the recovered 21-cm signal power spectra by an order of magnitude, in the presence of mode mixing. Exploring the performance of GPR using a range of different foregrounds and EoR signal, we find an optimal recovery for $l_{\rm mix} \geq 3$ MHz and $\sigma_{21} \gtrsim 0.12\sigma_{\rm n}$, with fractional bias below 20 per cent and with at least a 3σ confidence level on the detection. Outside this range, the detectability of the signal is still adequate, but with larger bias and larger uncertainty. These values hold for a single-frequency bandwidth of 12 MHZ, and combining several redshift bins will improve the confidence limit on the detection. They partially depend as well on the observation configuration, such as *uv*-coverage, the lower and upper baseline limits, the FoV, and so they are most representatives for LOFAR–HBA in 1200 h of observations.

The fundamental improvement of GPR resides in its complete statistical description of all components contributing to the observed signal. In its current implementation,[11] we use a generic model for the 21-cm signal and mode-mixing components which only make use of our prior knowledge on the frequency dependence of the signals. While this treatment may be sufficient for a detection of the 21-cm signal and its characterization with LOFAR, an improved model may be built for future experiments with e.g. the more sensitive SKA. The mode-mixing model for example can be improved by integrating the $k_\perp$ dependency of the foreground wedge and folding into the model the analytic work describing the effect on the signal of the instrumental chromaticity, calibration errors and sky-model incompleteness. Exploiting the isotropic nature of the 21-cm signal and its evolution at different redshift bins will also ensure a more sensitive and accurate modelling. Finally, in the course of determining the physical 21-cm signal parameters from the 21-cm signal power spectra using, for example, an MCMC sampler (Greig & Mesinger 2015; Kern et al. 2017), the GPR bias could be determined and integrated at each MCMC steps.

REFERENCES

Aigrain S., Parviainen H., Pope B. J. S., 2016, MNRAS, 459, 2408
Ali Z. S. et al., 2015, ApJ, 809, 61
Barry N., Hazelton B., Sullivan I., Morales M. F., Pober J. C., 2016, MNRAS, 461, 3135
Beardsley A. P. et al., 2016, ApJ, 833, 102
Bobin J., Moudden Y., Starck J.-L., Fadili J., Aghanim N., 2008, Stat. Methodol., 5, 307
Bonaldi A., Brown M. L., 2015, MNRAS, 447, 1973
Chapman E. et al., 2012, MNRAS, 423, 2518
Chapman E. et al., 2013, MNRAS, 429, 165
Chapman E. et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), Giardini Naxos, Italy, p. 5
Datta A., Bowman J. D., Carilli C. L., 2010, ApJ, 724, 526
Diggle P., Moyeed R. A., Tawn J. A., 1998, Appl. Stat., 47, 299
Dillon J. S. et al., 2015, Phys. Rev. D, 91, 123011
Ewall-Wice A., Dillon J. S., Liu A., Hewitt J., 2017, MNRAS, 470, 1849
Fischer B., Gorbach N., Bauer S., Bian Y., Buhmann J. M., 2016, preprint (arXiv:1610.00907)
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Furlanetto S. R., 2016, in Mesinger A., ed., Astrophysics and Space Science Library Vol. 423, Understanding the Epoch of Cosmic Reionization: Challenges and Progress, Springer International Publishing, Switzerland, p. 247
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gehlot B. K. et al., 2017, preprint (arXiv:1709.07727)
Gelman A., Carlin J., Stern H., Dunson D., Vehtari A., Rubin D., 2014, Bayesian Data Analysis, 3rd edn. (Chapman & Hall/CRC Texts in Statistical Science). Chapman and Hall/CRC. UK
Ghosh A., Bharadwaj S., Ali S. S., Chengalur J. N., 2011, MNRAS, 418, 2584
Ghosh A., Koopmans L. V. E., Chapman E., Jelić V., 2015, MNRAS, 452, 1587
Ghosh A., Mertens F. G., Koopmans L. V. E., 2018, MNRAS, 474, 4552
Goodman J., Weare J., 2010, Commun. Appl. Math. Comput. Sci., 5, 65
Greig B., Mesinger A., 2015, MNRAS, 449, 4246
Harker G. et al., 2009, MNRAS, 397, 1138
Hazelton B. J., Morales M. F., Sullivan I. S., 2013, ApJ, 770, 156
Hogg D. W., 1999, preprint (arXiv:9905116)
Hojjati A., Kim A. G., Linder E. V., 2013, Phys. Rev. D, 87, 123512
Jackson C., 2005, PASA, 22, 36
Jeffreys H., 1961, Theory of Probability, 3rd edn. Clarendon Press, Oxford
Jelić V. et al., 2008, MNRAS, 389, 1319
Jelić V., Zaroubi S., Labropoulos P., Bernardi G., de Bruyn A. G., Koopmans L. V. E., 2010, MNRAS, 409, 1647
Jensen H., Majumdar S., Mellema G., Lidz A., Iliev I. T., Dixon K. L., 2016, MNRAS, 456, 66
Karamanavis V. et al., 2016, A&A, 590, A48
Kern N. S., Liu A., Parsons A. R., Mesinger A., Greig B., 2017, ApJ, 848, 23
Koopmans L. V. E., 2010, ApJ, 718, 963
Liu A., Parsons A. R., Trott C. M., 2014a, Phys. Rev. D, 90, 023018
Liu A., Parsons A. R., Trott C. M., 2014b, Phys. Rev. D, 90, 023019
Mesinger A., Furlanetto S., 2007, ApJ, 669, 663
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Morales M. F., Wyithe J. S. B., 2010, ARA&A, 48, 127
Morales M. F., Bowman J. D., Hewitt J. N., 2006, ApJ, 648, 767
Morales M. F., Hazelton B., Sullivan I., Beardsley A., 2012, ApJ, 752, 137
Mort B., Dulwich F., Razavi-Ghods N., de Lera Acedo E., Grainge K., 2017, MNRAS, 465, 3680
Mukherjee P., Parkinson D., Liddle A. R., 2006, ApJ, 638, L51
Murray S. G., Trott C. M., Jordan C. H., 2017, ApJ, 845, 7
Neal R. M., 1997, preprint (arXiv)
Parsons A., Pober J., McQuinn M., Jacobs D., Aguirre J., 2012, ApJ, 753, 81
Patil A. H. et al., 2016, MNRAS, 463, 4317
Patil A. H. et al., 2017, ApJ, 838, 65
Pober J. C., 2015, MNRAS, 447, 1705
Pober J. C. et al., 2014, ApJ, 782, 66
Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge
Shaver P. A., Windhorst R. A., Madau P., de Bruyn A. G., 1999, A&A, 345, 380
Stein M., 1999, Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics, Springer, New York
Thyagarajan N. et al., 2015, ApJ, 804, 14
Thyagarajan N., Parsons A. R., DeBoer D. R., Bowman J. D., Ewall-Wice A. M., Neben A. R., Patra N., 2016, ApJ, 825, 9
Trott C. M., Wayth R. B., Tingay S. J., 2012, ApJ, 757, 101
Trott C. M. et al., 2016, ApJ, 818, 139
van Haarlem M. P. et al., 2013, A&A, 556, A2
Vedantham H. K., Koopmans L. V. E., 2016, MNRAS, 458, 3099
Vedantham H., Udaya Shankar N., Subrahmanyan R., 2012, ApJ, 752, 137

[11]The code implementing the algorithm described in this paper is freely available at https://gitlab.com/flomertens/ps_eor

Yatawatta S., 2016, preprint (arXiv:1605.09219)
Yatawatta S. et al., 2013, A&A, 550, A136

APPENDIX A: GPR AS A LINEAR REGRESSION PROBLEM

In a linear regression problem, one models the data $\mathbf{d}$ as,

$$\mathbf{d} = \mathbf{H}\mathbf{f} + \mathbf{n} \tag{A1}$$

where f are the weights of the basis functions that form the columns of matrix $\mathbf{H}$. The noise on the data is $\mathbf{n}$ with a covariance matrix $\Sigma_\mathbf{n} = \langle \mathbf{n}\mathbf{n}^T \rangle$. In GPR often no basis functions are chosen, such that $\mathbf{H} = \mathbf{I}$ and $\mathbf{f}$ are the true function values where the data was taken. Hence,

$$\mathbf{d} = \mathbf{f} + \mathbf{n} \tag{A2}$$

We note that equation (A2) is ill-posed and additional constraints need to be set on $\mathbf{f}$. In GPR, this constraint is statistical and set in the form of a covariance matrix $\Sigma_\mathbf{f} = \langle \mathbf{f}\mathbf{f}^T \rangle$. In other words, the values in $\mathbf{f}$ should follow a particular covariance structure, which is set by some simple functional form, such as e.g. a Matern Kernel. If we assume that both $\mathbf{n}$ and $\mathbf{f}$ are Gaussian distributed between any two values in $\mathbf{n}$ or $\mathbf{f}$, we have a GP. We show this as follows. We can rewrite equation (A2) in matrix notation as,

$$\mathbf{z} = \begin{bmatrix} \mathbf{d} \\ \mathbf{n} \end{bmatrix} = \begin{bmatrix} \mathbf{I}\mathbf{I} \\ \mathbf{0}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{n} \end{bmatrix} \tag{A3}$$

Here, $\mathbf{z}$ is also a Gaussian random variable because it is a linear combination of two Gaussian random variables $\mathbf{f}$ and $\mathbf{n}$. The covariance matrix of z then becomes,

$$\Sigma_\mathbf{z} = \langle \mathbf{z}\mathbf{z}^T \rangle = \begin{bmatrix} \Sigma_\mathbf{d} & \Sigma_\mathbf{f} \\ \Sigma_\mathbf{f} & \Sigma_\mathbf{f} \end{bmatrix} \tag{A4}$$

where $\Sigma_\mathbf{d} \equiv \Sigma_\mathbf{f} + \Sigma_\mathbf{n}$. Now, given this covariance structure, we have the joint PDF as,

$$\mathbf{P}(\mathbf{z}) = \frac{1}{\sqrt{\det(2\pi\mathbf{C}_\mathbf{z})}} \mathrm{e}^{(-\frac{1}{2}\mathbf{z}^T \mathbf{C}_\mathbf{z}^{-1}\mathbf{z})} \tag{A5}$$

We can think of this as a multivariate Gaussian PDF with correlations between $\mathbf{d}$ and $\mathbf{f}$, where $\mathbf{d}$ is a noisy version of $\mathbf{f}$ ($\mathbf{d} = \mathbf{f} + \mathbf{n}$). We note that we actually know $\mathbf{d}$ and hence this PDF is a conditional PDF. The conditional PDF is another Gaussian with an expectation value,

$$\langle \mathbf{f}|\mathbf{d} \rangle = \langle \mathbf{d} \rangle + \Sigma_\mathbf{f} {\Sigma_\mathbf{d}}^{-1} (\mathbf{d} - \langle \mathbf{d} \rangle) \tag{A6}$$

Now, if $\langle \mathbf{f} \rangle = \mathbf{0}$ and $\langle \mathbf{n} \rangle = \mathbf{0}$ as often assumed then we have $\langle \mathbf{d} \rangle = \mathbf{0}$ and equation (A6) becomes,

$$\langle \mathbf{f}|\mathbf{d} \rangle = \Sigma_\mathbf{f} \left(\Sigma_\mathbf{f} + \Sigma_\mathbf{n}\right)^{-1} \mathbf{d}. \tag{A7}$$

With a little more linear algebra, it can be shown that the covariance of this expectation value is given by,

$$\Sigma_{\langle \mathbf{f}|\mathbf{d} \rangle} = \Sigma_\mathbf{f} - \Sigma_\mathbf{f} \left(\Sigma_\mathbf{f} + \Sigma_\mathbf{n}\right)^{-1} \Sigma_\mathbf{f}. \tag{A8}$$

We note these sets of equations (A7) and (A8) are exactly similar to mean and covariance quoted in Section 2, equation (4). On the other hand, the posterior probability of the data given $\mathbf{f}$ times a prior on $\mathbf{f}$, can be written as,

$$\begin{aligned} \mathbf{logP}(\mathbf{f}|\mathbf{d}) &= -\frac{1}{2}(\mathbf{f} - \mathbf{d})^T {\Sigma_\mathbf{n}}^{-1} (\mathbf{f} - \mathbf{d}) - \frac{1}{2}\mathbf{f}^T {\Sigma_\mathbf{f}}^{-1}\mathbf{f} + \text{constant} \\ &= -\frac{1}{2}\mathbf{f}^T {\Sigma_\mathbf{n}}^{-1}\mathbf{f} - \frac{1}{2}\mathbf{f}^T {\Sigma_\mathbf{f}}^{-1}\mathbf{f} + \frac{1}{2}\mathbf{f}^T {\Sigma_\mathbf{n}}^{-1}\mathbf{d} \\ &\quad + \frac{1}{2}\mathbf{d}^T {\Sigma_\mathbf{n}}^{-1}\mathbf{f} + \text{constant} \end{aligned} \tag{A9}$$

Now, maximizing equation (A9) with respect to the functional values $\mathbf{f}$, we can find the MAP solution,

$$\langle \mathbf{f} \rangle = ({\Sigma_\mathbf{f}}^{-1} + {\Sigma_\mathbf{n}}^{-1})^{-1} {\Sigma_\mathbf{n}}^{-1}\mathbf{d} = \Sigma_\mathbf{f}(\Sigma_\mathbf{f} + \Sigma_\mathbf{n})^{-1}\mathbf{d} \tag{A10}$$

here, we used the Searle identity

$$({\Sigma_\mathbf{f}}^{-1} + {\Sigma_\mathbf{n}}^{-1})^{-1} {\Sigma_\mathbf{n}}^{-1} = \Sigma_\mathbf{f}(\Sigma_\mathbf{f} + \Sigma_\mathbf{n})^{-1}\Sigma_\mathbf{n}{\Sigma_\mathbf{n}}^{-1}$$

with $\Sigma_\mathbf{n}{\Sigma_\mathbf{n}}^{-1} = \mathbf{I}$. Hence, $\langle \mathbf{f} \rangle$ is the MAP solution of $\mathbf{d} = \mathbf{f} + \mathbf{n}$ with $\mathbf{n} \sim \mathbf{N}(\mathbf{0}, \Sigma_\mathbf{n})$ and $\mathbf{f} \sim \mathbf{N}(\mathbf{0}, \Sigma_\mathbf{f})$. In conclusion, equations (A7) and (A10) show that GPR is fully equivalent to the usual linear regression $\mathbf{d} = \mathbf{H}\mathbf{f} + \mathbf{n}$, where $\mathbf{d}$ is the data, $\mathbf{H} = \mathbf{I}$ is assumed the identity matrix, $\mathbf{f}$ are the inferred functional value, and $\mathbf{n}$ is the (Gaussian) noise. If one then assumes $\mathbf{n} \sim \mathbf{N}(\mathbf{0}, \Sigma_\mathbf{n})$ and $\mathbf{f} \sim \mathbf{N}(\mathbf{0}, \Sigma_\mathbf{f})$, where the Σ's are the covariance matrices of the noise and the functional values, with the former used in the likelihood function and the latter in a prior, in the usual Bayesian sense, then one arrives exactly at the GPR equations (for $\mathbf{x} = \mathbf{x}'$ in equation A7) in Section 2 (we assume as in the paper that mean = 0 for the Gaussian PDFs).

This paper has been typeset from a TEX/LATEX file prepared by the author.