Biometrics 000, 000–000 DOI: 000
000 0000

A Non-parametric Bayesian Model for Estimating Spectral Densities of Resting-State EEG Twin Data

Brian Hart

Division of Biostatistics, University of Minnesota

Michele Guindani

Department of Statistics, University of California Irvine

Stephen Malone

Department of Psychology, University of Minnesota

and

Mark Fiecas

Division of Biostatistics, University of Minnesota

Summary: Electroencephalography (EEG) is a non-invasive neuroimaging modality that captures electrical brain activity many times per second. We seek to estimate power spectra from EEG data that was gathered for 557 adolescent twin pairs through the Minnesota Twin Family Study (MTFS). Typically, spectral analysis methods treat time series from each subject separately, and independent spectral densities are fit to each time series. Since the EEG data was collected on twins, it is reasonable to assume that the time series have similar underlying characteristics, so borrowing information across subjects can significantly improve estimation. We propose a Nested Bernstein Dirichlet Prior model to estimate the power spectrum of the EEG signal for each subject by smoothing periodograms within and across subjects while requiring minimal user input to tuning parameters. Furthermore, we leverage the MTFS twin study design to estimate the heritability of EEG power spectra with the hopes of establishing new endophenotypes. Through simulation studies designed to mimic the MTFS, we show our method out-performs a set of other popular methods.

Key words: Bernstein Polynomial; Heritability; Nested Dirichlet Process; Time series; Whittle likelihood.

This paper has been submitted for consideration for publication in *Biometrics*

1 Introduction

Electroencephalography (EEG) is a non-invasive neuroimaging modality that captures electrical brain activity many times per second by placing recording electrodes at various locations on the head. We seek to estimate power spectra from EEG data that was gathered for 557 adolescent twin pairs through the Minnesota Twin Family Study (MTFS) (Iacono et al., 1999). The EEG time series data collected from the twin study design of the MTFS calls for new methods that explore the nature of the twin relationships to determine the amount of shared information within and between twin pairs. We develop a novel statistical model that will identify the frequencies that drive the variations in the EEG data, an analysis approach known as *spectral analysis*. We will then consider the potential for discovering new endophenotypes from these spectral features by establishing their heritability, a quantity that describes the proportion of variation that can be attributed to genetic factors.

1.1 Spectral Analysis

Spectral analysis decomposes a time series into a set of waves oscillating at different frequencies. The primary tool for spectral analysis is the spectral density function, also known as the power spectrum. It is a density of variances that can be understood as an analysis of variance (ANOVA) where the spectral density curve shows the proportion of the total variance of a time series that is explained by waveforms oscillating at each frequency (Shumway and Stoffer, 2010). By decomposing an EEG time series in such a manner, the spectral density, assuming weak stationarity of the time series, provides a summary of the variance characteristics of the EEG signal.

Many common spectral density estimation methods use parametric forms such as autoregressive models (Shumway and Stoffer, 2010). These methods are often very fast and simple, but do not allow information sharing across multiple time series, as is desired in our data collected on twins. Other methods follow Wahba (1980) using Bayesian smoothing splines

in conjunction with the Whittle likelihood, which is an approximation of the true spectral likelihood (Whittle, 1953), or are completely nonparametric by using some form of kernel smoothing (Fiecas and Ombao, 2016; Shumway and Stoffer, 2010).

Moving beyond single time series analysis, some work has begun to address experimental design when estimating EEG spectral densities, beginning with Brillinger (1973) who assessed replicated time series as a manner to increase the signal to noise ratio. Diggle and Al Wasel (1997) and Krafty et al. (2011) expanded on these models with mixed effects models for repeated biomedical time series. More recently, Bruce et al. (2017) and Krafty et al. (2017) modeled covariate-modulated spectral densities, while work such as Fiecas and Ombao (2016) considered continuously evolving spectral densities through a learning experiment. Finally, Cadonna et al. (2018) recently developed a Bayesian method to estimate the spectral densities of multiple time series through mixture models. Each of these methods improves upon spectral analysis of a single time series by considering the experimental design. Likewise, our goal is to develop a method that takes advantage of the twin design in the MTFS to improve estimation of and inference on individual spectral densities and their features.

1.2 Endophenotypes

One motivation of spectral analysis of EEG data from a twin design is to establish *endophenotypes*, neurobiological indicators that link psychiatric disorders to genetic risk factors. Iacono et al. (2017) laid out seven different criteria for endophenotypes, one of which is that the feature must be heritable. Heritability analyses of EEG data traditionally first summarize the power spectrum to a single measure by taking the average power within a frequency band. After estimating the spectral densities using our proposed model, we develop an approach for heritability analysis that will preserve information from the entire EEG time series without having to resort to band averaging. Furthermore, we show how to calculate the heritability of the entire spectral density curve. By enabling heritability estimation of

the full spectrum and estimating its heritability using data from the MTFS, we expand the set of potential EEG endophenotypes.

1.3 Minnesota Twin Family Study

The MTFS is a population-based study of same-sex reared-together male and female twins (and their parents), the overarching goal of which is to understand genetic and environmental influences on substance abuse and related psychopathology. Data for this study consisted of resting-state EEG data from 365 monozygotic (MZ) and 192 dizygotic (DZ) twin pairs of approximately 17 years of age. More details on the MTFS data are provided in Section 4.1.

[Table 1 about here.]

Table 1 shows estimated heritability for four frequency bands: Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), and Beta (12-30 Hz) calculated from three channels, which correspond to the different electrodes placed on the scalp. While all three channels show high levels of heritability in each frequency band, the heritability varies across channel and frequency bands. Given these differences, it is easy to imagine a smooth function over frequencies that gives the heritability at each frequency, an idea we develop further in the present work.

The fact that the heritability differs between frequency bands suggests that a strict correlation structure imposed upon twin relationships may not be appropriate in this application. The importance of the twin relationship can vary from twin pair to twin pair and from one frequency band to another within a single twin pair. To learn the heterogeneity structure of the sample and allow flexible joint modeling of the spectral densities, we embed our estimation framework within a Bayesian nested Dirichlet process (DP) structure (Rodriguez et al., 2008). The method groups similar spectral densities and shrinks their estimates towards each other while allowing different groups to have potentially divergent estimates. These groupings can also be used as a way to study the similarity between two subjects in the

study. Altogether, the nested DP multi-subject spectral density model explores the twin study design of our data and allows us efficiently estimated spectral densities.

2 Model Specification and Inference

2.1 The Single Subject Model

We start by considering a model for the spectral density of a single time series to serve as the base of our multi-subject model. Suppose we observe a univariate time series, Y_t for $t = 1, \ldots, T$, and that the time series has been standardized to have mean zero and unit variance. We model the spectral density function as a mixture of probability density functions (PDFs), ensuring a strictly non-negative estimated spectral density that integrates to $\text{Var}(Y_t) = 1$. Since we will be working in the frequency domain, we use the Whittle likelihood (Whittle, 1953), given by

$$L(\boldsymbol{f} \mid \boldsymbol{Y}) \propto \prod_{\omega \in \Omega} \frac{1}{f(\omega)} \exp\Big(\frac{|d(\omega)|^2}{f(\omega)}\Big), \tag{1}$$

where $f(\omega)$ is the spectral density, $d(\omega) = T^{-0.5} \sum_{t=1}^{T} Y_t \exp(-i2\pi\omega t)$ is the discrete Fourier transform of the time series Y_t, and $\Omega = \{1/T, 2/T, \ldots, (\lfloor T/2 \rfloor - 1)/T\}$. The quantity $|d(\omega)|^2$ is known as the periodogram. Because the spectral density is considered only on an interval from zero to the Nyquist frequency (half the sampling rate), one natural model is a mixture of beta PDFs with domain scaled to lie in $(0, 0.5)$. Such a mixture of beta PDFs is commonly known as a basis of Bernstein polynomials. For ease of notation we assume all frequencies have been scaled to fall within the $(0, 0.5)$ interval. Let the spectral density be

$$f(\omega) = \sum_{\lambda=1}^{\Lambda} G\Big(\frac{\lambda - 1}{\Lambda}, \frac{\lambda}{\Lambda}\Big] \beta(\omega; \lambda, \Lambda - \lambda + 1), \tag{2}$$

where Λ is the degree of the Bernstein polynomial, $\beta(\omega; \lambda, \Lambda - \lambda + 1)$ is the beta density evaluated at ω with parameters λ and $\Lambda - \lambda + 1$, and $G\big(\frac{\lambda-1}{\Lambda}, \frac{\lambda}{\Lambda}\big] = G(\frac{\lambda}{\Lambda}) - G(\frac{\lambda-1}{\Lambda})$. We employ a Bayesian non-parametric approach and assume that G is a cumulative distribution function obtained as the realization of a Dirichlet process (Petrone, 1999a). We refer to Müller et al. (2015) for an introduction to the use of Bayesian nonparametric methods in practical

data analysis. Here it suffices to mention that the cumulative distribution function G can be represented almost surely as a discrete probability distribution, $G(x) = \sum_{l=1}^{\infty} p_l \delta_{z_l}(0, x]$, $x \in [0, 0.5]$, with atoms z_l randomly drawn from a base measure G_0, $z_l \overset{iid}{\sim} G_0$, with support on $(0, 0.5)$ and weights p_l characterized through the so-called stick-breaking representation (Sethuraman, 1994), i.e., let $p_1 = v_1$ and $p_l = v_l \prod_{s=1}^{l-1}(1 - v_s)$ where $v_l \sim beta(1, \alpha_v)$ with concentration parameter α_v. Due to the randomness of the weights and atoms, G is a random probability measure centered around the base measure since $\mathbb{E}(G(x)) = G_0(0, x]$, for all $x \in [0, 0.5]$. The concentration parameter α_v characterizes the variability of the realizations G around the base measure G_0. In symbols, we write $G \sim DP(\alpha_v, G_0)$. In the following, for computational reasons, we will consider a finite truncation approximation with truncation level L so that $v_L = 1$ and $G = \sum_{l=1}^{L} p_l \delta_{z_l}$ (Ishwaran and James, 2001). We now re-write our model as

$$f(\omega) = \sum_{\lambda=1}^{\Lambda} \sum_{l=1}^{L} p_l I\Big[\frac{\lambda - 1}{\Lambda} < z_l \leqslant \frac{\lambda}{\Lambda}\Big] \beta(\omega; \lambda, \Lambda - \lambda + 1). \tag{3}$$

This model, henceforth referred to as the Bernstein Dirichlet Prior (BDP), was introduced by Petrone (1999a) and applied to single time series spectral analysis by Choudhuri et al. (2004). These two papers along with Petrone (1999b) and Barrientos et al. (2017) demonstrated the utility and theoretical properties of the BDP model. We point out that Kakizawa (2006) extended Bernstein Polynomials to spectral density estimation in a frequentist context.

In this BDP model, Λ controls the number of beta PDF mixture components when estimating a single spectral density curve. A higher Λ corresponds to a larger number of Bernstein polynomial components, and thus the ability to capture sharper peaks in the underlying spectral density. The DP, G, then properly assigns weights to each component of the selected basis. Along these lines, Petrone (1999b) offered some nice intuition for the BDP model as a smoothed histogram, where Λ can be viewed as the number of bins in the histogram and G assigns each observation to one of the available bins.

Note that we do not specify a scalar, referred to as τ by Choudhuri et al. (2004), which only serves to multiply the spectral density by the total variance of the time series. By standardizing the time series to mean zero and unit variance, we remove the need to estimate τ while still capturing the desired information about the shape of $f(\omega)$.

2.2 The Multi-Subject Model

Having formulated the BDP model to estimate a single spectral density curve, we now consider a model to estimate many spectral densities from time series collected on a sample consisting of twin pairs. Because the twin relationships induce similarities across spectral density estimates, we propose grouping similar individual spectral densities by nesting the BDP model within a second DP. The partitions enforced by the nested DP explain the heterogeneity between subjects by assigning subjects with very different spectra to separate groups while also allowing very similar subjects to receive accordingly similar spectral density estimates by grouping these subjects together.

More specifically, we assume that the individual spectral density may be assigned to one of K groups, with each group characterized by a specific spectrum profile, i.e., K realizations $\{G_1^*, \ldots, G_K^*\}$ from a BDP as in Equation 3. The individual spectral density for subject n is estimated as in Equation 2. Let G_n be subject n's draw from the nested DP, then, in formulas, $G_n \sim Q$ with $Q = \sum_{k=1}^{K} \pi_k^* \delta_{G_k^*}$ and each $G_k^* = \sum_{l=1}^{L} p_{kl} \delta_{z_{kl}}$ defined as in Section 2.1 for $k = 1, \ldots, K$. The weights π_k describe the proportion of subjects assigned to group $k = 1, \ldots, K$ and we assume they are defined similarly to the p_l, through a stick-breaking construction, $\pi_k = u_k \prod_{s=1}^{k-1}(1 - u_s)$ where $u_k \sim beta(1, \alpha_u)$. We refer to our model as the nested Bernstein Dirichlet prior (NBDP) model. Let ζ_n be an allocation variable, such that $\zeta_n = k$ for $k = 1, \ldots, K$ and $n = 1, \ldots, N$, if and only if subject n is assigned to group k. Allowing Λ_{ζ_n} to differ by group, the spectral density for subject n can then be written as:

$$f_{\zeta_n}(\omega) = \sum_{\lambda=1}^{\Lambda_{\zeta_n}} \sum_{l=1}^{L} p_{\zeta_n l} I\Big[\frac{\lambda - 1}{\Lambda_{\zeta_n}} < z_{\zeta_n l} \leqslant \frac{\lambda}{\Lambda_{\zeta_n}}\Big] \beta(\omega; \lambda, \Lambda_{\zeta_n} - \lambda + 1). \tag{4}$$

[Figure 1 about here.]

Figure 1 shows a diagram representing the NBDP model. In the context of multi-subject EEG data, the nested DP models the heterogeneity of the sampled EEG spectral densities by partitioning the set of subjects into homogeneous groups. Meanwhile, the BDP fits a functional curve for each group. We use a Markov chain Monte Carlo (MCMC) sampling algorithm, where at each iteration we first assign each subject to one of the K available groups with function-level estimate G_k. These G_k BDP functional curve estimates are then updated to best fit the subjects assigned to that group. We stress that we are not concerned with the partitions induced by the model, but only employ the nested DP to account for the heterogeneity across subjects, which will improve the estimates of the power spectra. The flexible nature of the NBDP groupings allows us to capture potentially complex twin relationships that vary across twin pairs and across frequencies. Details of the MCMC algorithm can be found in the Web Appendix.

2.3 Estimating the Heritability of the Power Spectrum

Given the posterior distributions of the individual power spectra obtained using our NBDP model, we now need to estimate the heritability of features of the resting-state EEG power spectra using the twin pair relationships in the MTFS data. As previously mentioned, establishing that spectral characteristics are heritable is essential for developing endophenotypes. Falconer's formula estimates the heritability of the power spectrum at frequency ω as $h^2(\omega) = 2(r_{MZ}(\omega) - r_{DZ}(\omega))$, where $r_{MZ}(\omega)$ and $r_{DZ}(\omega)$ are the correlation in the estimated individual spectral densities at frequency ω among MZ and DZ twins respectively (Falconer, 1960). We refer to the heritability estimated across the frequency domain, Ω, as the heritability spectrum. While this estimator gives us the heritability of a single frequency, it is more scientifically useful to consider intervals of frequencies, say (ω_1, ω_2), also known as frequency bands. Findings such as those of Malone et al. (2014) and Rudo-Hutt (2015) used

frequency bands instead of interpreting each individual frequency. Because the variance of the spectral densities is not constant across Ω and heritability is the percentage of variation explained by genetics, when calculating the heritability of a frequency band we weight the heritability at each frequency by the variance of the estimated spectral densities at that frequency. Our estimator takes the form

$$h^2(\omega_1, \omega_2) = \frac{\int_{\omega_1}^{\omega_2} h^2(\omega)\text{Var}(f(\omega))d\omega}{\int_{\omega_1}^{\omega_2} \text{Var}(f(\omega))d\omega}. \tag{5}$$

We estimate $\text{Var}(f(\omega))$ by taking the sample variance at each frequency across all of the curves from our posterior distribution. We can then calculate heritability for any frequency band without first reducing the spectral densities to band power estimates. Note that we can compute the heritability of the entire spectral density, which we call the full spectrum heritability, by setting the bounds of integration in Equation (5) to $(0, 0.5)$. This full spectrum heritability could not be calculated using existing methods, and so it allows the introduction of a new set of endophenotypes based on the full spectrum of an EEG time series.

3 Simulation Study

3.1 Data Simulation Process

We assess the performance of the NBDP model compared to competitors using five different simulation scenarios. The first four simulation scenarios use autoregressive processes to simulate data that mimics the MTFS data as closely as possible and provide a means to assess the models' abilities to accurately estimate individual spectral densities. The fifth scenario was designed with more simplistic, piecewise constant spectral densities that facilitate controlled construction of the heritability spectrum. For this final scenario all methods are assessed on their ability to estimate both the individual spectral densities and group level heritability spectrum. Details of the data simulation process for each scenario can be found in the Web Appendix.

3.2 Model Comparison

For model comparisons, we chose three competitor models. The first was the generalized cross validation span selection periodogram smoother of Ombao et al. (2001). For this method, henceforth referred to as GCV, we considered spans between 3 and 100 for each subject. The second was the stationary version of Rosen et al. (2009), which uses Bayesian smoothing splines, fit to each subject separately. We considered versions of this model that used $J = 10$ and $J = 20$ spline basis components. These models will be referred to as Spline10 and Spline20, respectively. The final competitor we considered is the BDP model (Choudhuri et al., 2004), which forms the base of our NBDP model but does not allow borrowing of information across subjects. More prior specification details for the simulation study can be found in the Web Appendix. To compare the accuracy of the resulting estimated spectral density curves we calculated the integrated absolute error (IAE) for each subject. For a given estimated spectral density $\hat{f}(\omega)$ and true density $f(\omega)$, IAE $= \int_{\omega} |\hat{f}(\omega) - f(\omega)| \partial\omega$. For Scenario 5, we assess the accuracy of heritability spectrum estimation by calculating the IAE, where the truth was determined by applying Falconer's Formula to the sample of true simulated power spectra.

3.3 Simulation Study Results

[Table 2 about here.]

Table 2 contains the mean and standard deviation of the specral density mean IAE for each model and simulation scenario and the heritability spectrum IAE for Scenario 5. We also split the results into the four frequency bands commonly used in EEG analysis.

In terms of spectral density estimation, the NBDP method had the best full spectrum mean IAE in the four simulation scenarios using AR processes and was third in the final scenario. This advantage diminished as the amount of between-subject heterogeneity increased through the first four simulation scenarios, but the NBDP model maintained the lowest mean IAE

and the lowest standard deviation of IAE across all four of these scenarios. The GCV and Spline20 methods showed improved performance in Scenario 5, which has sharp jumps in spectral densities from its piecewise construction.

The Spline20 method showed the worst density estimation performance in first four scenarios, as it likely over-fit the data with too many spline components, but was much improved in Scenario 5. The spline method received a considerable boost in performance when $J = 10$ basis components were used in the first four scenarios, finishing second in all scenarios behind our NBDP model. However, the Spline 10 method was the worst in Scenario 5. The number of spline components must be selected a priori in this model, whereas the BDP and NBDP models adaptively select the proper order of Bernstein polynomial to use based on the data. Comparison of the 10 and 20 component spline methods shows that using the incorrect number of components can seriously impact the performance of the model.

The BDP method showed the second worst performance of the models under comparison in all scenarios. The improvement from the BDP model to our NBDP model demonstrates the added value of a nested model that shares information between subjects. The decreasing advantage of the NBDP model from scenarios 1 to 4 is to be expected due to increased between-subject heterogeneity, and is reflected in the pattern in the Alpha band IAE.

The NBDP model showed the best performance in estimating the heritability spectrum, despite only having the third best spectral density estimation in Scenario 5. This advantage was largely driven by superior performance in the Beta band, where the other models tended to over-estimate the heritability. Interestingly, the Spline10 model showed improved heritability spectrum estimation compared to Spline20, despite having considerably worse performance in spectral density estimation. The NBDP method was the best model in terms of estimating the full heritability spectrum, demonstrating that the clustering and strong smoothing imposed by the nested DP structure did not negatively impact heritability

estimation. It is worth noting, however, that the NBDP method struggled the most in the Delta and Alpha bands, where the spectral density was highest, suggesting the higher level of smoothing may come at a cost at these peaks. Readers are directed to the Web Appendix for a comparison of estimating other power spectrum features such as peak frequency and Alpha band peak frequency.

Note that the Bayesian nature of the NBDP model makes inference on features of the spectral density curves such as peak frequency and alpha peak frequency very straightforward using functions of the posterior samples. The GCV model does not offer a similar general approach to inference on features of the estimated spectral densities and each spectral feature requires the use of large-sample results or the bootstrap.

Finally, we note that the methods compared differ in terms of computing time. For example, a single MCMC iteration while fitting the model to a single simulation scenario took 6.2 seconds on average for the NBDP model compared to 1.4 seconds for the Spline10 model. MCMC mixing is also made more difficult by fitting all participants at once in the NBDP model, but these concerns are alleviated by trace plots and MTFS results presented in the Web Appendix. While the NBDP model is computationally expensive, the improved spectral density estimates in Scenarios 1-4 and improved heritability estimation in Scenario 5 may justify the additional time required to fit the model.

4 MTFS Analysis

Our analysis of the MTFS data had two primary goals: 1) conduct spectral analysis in the resting-state EEG data to quantify the power within frequency bands of interest, and 2) examine potential for new endophenotypes by computing each band heritability as well as the heritability of the entire spectral density.

4.1 Data Description and Preprocessing

We applied our method to resting-state EEG data collected from the MTFS which constituted part of the molecular-genetic studies of Malone et al. (2014) and Smit et al. (2017). Specifically, the data used here were from the intake assessment of male and female twins from the age-17 cohort of MTFS twins and consisted of data from 3 electrodes common to both sex cohorts (Cz, O1 and O2), collected with a sampling rate of 128 Hz and 12 bits. The Cz electrode is located at the center midline on the top of the head and the O1 and O2 electrodes are located on either side of the midline at the rear of the head. Considering each channel separately, we applied our method to 8 seconds of data resulting in time series of length 1024. In total, our sample included 1116 adolescents consisting of participants from 365 distinct MZ twin pairs and 192 DZ twin pairs. The sample had approximately equal numbers of individuals of each sex with 565 females and 551 males.

While twin participants sat comfortably in a darkened room with their eyes closed, EEG signals were recorded by means of identical Grass 12 Neurodata systems, with a pass band from 1 to 30 Hz (amplifier rolloff, 6 dB/octave). Notes recorded when the data were originally collected guided identification of data that needed to be excluded because of recording problems. Subjects who reported having fallen asleep or were noted to have fallen asleep were excluded. EEG segments containing transient artifacts and excessively small or large voltage deflections were tagged for exclusion by a computer algorithm written in Matlab. Multivariate outliers across the 3 electrodes were identified using a robust version of Mahalanobis distance from the robustbase package in the R statistical programming environment and visually reviewed for contamination by high-frequency noise, other artifacts (e.g., electrocardiogram), or signs of sleepiness. Individual recording sites were excluded from analyses if fewer than 45 2-second artifact-free sweeps were available. Finally, each 8 second time series used in our analysis was centered and standardized to have mean 0 and unit variance.

4.2 MTFS Results

To assess goodness of fit of the NBDP model to the MTFS data, WAIC values were calculated for each EEG channel and each Bayesian model considered in the simulation study(Watanabe and Opper, 2010). WAIC results are shown in Table 3. While the Spline20 model demonstrated the best WAIC, the NBDP model was in line with the other models and had the lowest effective number of parameters in each case, justifying using the NBDP model on the MTFS data. More goodness of fit results can be found in the Web Appendix.

[Table 3 about here.]

Figure 2 shows, for each channel, the estimated power spectrum for each of the 1116 subjects in the MTFS. The somewhat clustered nature of our NBDP model is visible in these estimated curves, as many individual subjects' estimated curves fall nearly on top of each other, creating a darker black line. These groupings are defined by different characteristics such as the Alpha band peak visible around 10 Hz. The nested DP often detected the twin structure in the data, clustering MZ twins Cz power spectra together in 26.4% of posterior samples compared with 8.9% for DZ twins and only 4.6% for non-twins. The NBDP model's ability to discover the clustering structure without any information on twin relationships demonstrates one of it's major advantages over single subject methods such as GCV. We are interested in estimating the heritability of the power spectrum in the MTFS, taking advantage of the fact that the NBDP model enables the exploration of the population structure and similarity matrices, which may be of primary interest in a different application.

[Figure 2 about here.]

Table 4 reports the median and inter-quartile range of frequency band power and peak frequency estimates for the 1116 MTFS subjects along with the estimated sample heritability for each feature using both the NBDP and GCV models. When available, Table 4 also

reports the results from Smit et al. (2005) and Malone et al. (2014). For the frequency band heritability measurements we used Equation 5 for both the GCV and NBDP estimates. Compared to the Cz channel, the O1 and O2 channels exhibited larger Alpha peak power, Alpha band power, and Beta band power along with smaller Delta and Theta band power. The GCV method produced very similar results for the Cz channel but showed higher Alpha band power than the NBDP model in the O1 and O2 channels. Note that the full spectrum power is 1 for all subjects due to the nature of the NBDP model and the fact that the original time series were standardized to unit variance.

Table 4 shows that out Cz band heritability estimates generally match those of the GCV method, Smit et al. (2005), and Malone et al. (2014). The lower heritability of the O1 and O2 channel for the GCV and NBDP methods may be due to the weighting in our heritability estimator which emphasizes areas of the spectrum with higher variability when consider each frequency band. The differences in band heritability are largest in the Beta band, which is expected since there is relatively little variation in the power spectra above 12 Hz. Since heritability is the percentage of variation attributable to genetics, it is important to consider the overall variability of the data. Smit et al. (2005) also calculated the heritability spectrum by binning the power into 1 Hz bins and found a pattern of decreasing heritability in the Beta band, which is consistent with our findings.

Our heritability estimator also estimated full spectrum heritabilities of 0.68, 0.45, and 0.56 for the Cz, O1, and O2 channels respectively for the NBDP model and 0.75, 0.66, and 0.58 respectively for the GCV model. Our establishment of the heritability of the full spectral density curve for these three channels paves the way for future research to establish endophenotypes by connecting the spectral density curve with psychiatric disorders. The NBDP's Bayesian machinery facilitates the calculation of credible intervals for all features of the spectral densities that are not nearly as straight forward for models such as the GCV.

[Table 4 about here.]

Table 4 also shows the heritability estimates for the peak frequency and Alpha peak frequency. Both peak frequency and Alpha peak frequency showed high levels of heritability, though the Alpha peak appeared to be more heritable. While our results show that the NBDP model produces heritability estimates that are generally consistent with those found by other methods, we point out that the clustered structure of the model could cause unstable heritability estimates in other datasets with fewer participants.

We investigated the sensitivity of our results with respect to the model hyperparameters. The results are qualitatively similar to the ones presented here and can be found in the Web Appendix.

5 Discussion and Conclusions

We developed a novel Bayesian non-parametric model for estimating the spectral densities of multi-subject, resting-state EEG data and their associated heritability. Our model embeds the BDP within a nested DP to share information across subjects with similar spectral densities, which led to large improvements in estimation as we showed in our simulation study. We applied our method to resting-state EEG data from the MTFS and showed that the resulting estimated spectral densities were highly heritable, especially in the Theta and Alpha bands. Additionally, the peak frequency and Alpha peak frequency were also heritable. These findings are consistent with Malone et al. (2014) and Smit et al. (2005). We extend these works by allowing inferential statements about the features of the spectral density, such as peak frequency and power within frequency bands, using our rigorous Bayesian non-parametric framework. Proper inference on features of the spectral densities and their heritability is straightforward once the posterior samples have been collected, and such inference does not rely on asymptotic results for these features.

Furthermore, we computed the heritability of the entire power spectrum, and estimated

the full spectrum heritability for the Cz, O1, and O2 channels to be 0.68, 0.45, and 0.56, respectively. These findings alone point to the genetic underpinnings of resting-state EEG signals and their oscillatory behavior measured in the frequency domain. Our finding that a large proportion of the variability in the spectral density can be attributed to genetic factors leads to the possibility of the entire spectral density being an endophenotype.

Once a trait has been established as heritable, it still remains to show this genetic risk factor is related to the psychiatric disorder under study. Past findings have shown increased or decreased power within defined frequency bands to be associated with conditions such as alcoholism, depression, and ADHD (Rudo-Hutt, 2015). Quantitative genetics analysis approaches applied to EEG spectral densities highlight their potential for linking psychiatric disorders with genetic risk factors (Malone et al., 2014). To establish the full spectral density as an endophenotype we would need to establish its association with psychiatric conditions through techniques such as functional regression, which is beyond the scope of this work.

The nested DP allows us to simultaneously account for between subject heterogeneity and potentially complex correlations between subjects that may arise due to the twin design of the study. This allows us to model the heterogeneity in the spectral features of the data, which has potential clinical implications. For instance, Harper et al. (2018) showed the association between spectral power and behavioral disinhibition. Furthermore, Lizio et al. (2011) showed potential clinical utility for Alpha peak frequency, and Grandy et al. (2013) showed that variation in Alpha peak frequency is associated with cognitive ability.

One possible extension of our model would be to include covariate information in the BDP portion of our model through dependent Dirichlet processes (MacEachern, 1999). Barrientos et al. (2017) proposed and compared methods to make the weights and atoms of the BDP dependent on a matrix of covariates. Incorporation of covariate information, such as age or disease status, at the BDP level could, at the cost of increased computation, improve

functional curve estimation and allow more flexibility in the estimation procedure, and it would allow subjects assigned to the same group at a single MCMC iteration to have different estimated spectral densities based on their covariate values. This feature would alleviate the issue of "ties" in the data creating identical spectral densities for different subjects and may result in more stable results. Additionally, by including psychiatric disorders as covariates, a dependent DP extension of our model may be able to establish endophenotypes in a more unified framework. Bruce et al. (2017) offers a method for modeling spectral densities conditional on covariates, but not within a DP framework.

Edwards et al. (2017) formulated a B-spline prior model, which can be viewed as a generalization of the BDP model, and showed that it captured sharp peaks in spectral densities better than the BDP model at the cost of a 2-3 fold increase in computing time. There is potential for improved performance of our model by replacing the Bernstein polynomial basis with a B-spline basis within the nested DP, however, the significant increase in computational burden would make the method very cumbersome with datasets as large as the MTFS.

The nested DP has been shown to be prone to degeneracy issues (Camerlenghi et al., 2019). While we have not observed such degeneracy in our inference, we should point out that our interest is in curve fitting, and in accounting for between-subject heterogeneity, and not into properly capturing shared components at different frequencies.

References

Barrientos, A. F., Jara, A., and Quintana, F. A. (2017). Fully nonparametric regression for bounded data using dependent bernstein polynomials. *Journal of the American Statistical Association* pages 1–20.

Brillinger, D. R. (1973). The analysis of time series collected in an experimental design. In *Multivariate Analysis–III*, pages 241–256. Elsevier.

Bruce, S. A., Hall, M. H., Buysse, D. J., and Krafty, R. T. (2017). Conditional adaptive bayesian spectral analysis of nonstationary biomedical time series. *Biometrics* .

Cadonna, A., Kottas, A., and Prado, R. (2018). Bayesian spectral modeling for multiple time series. *Journal of the American Statistical Association* **0,** 1–38.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., Rodríguez, A., et al. (2019). Latent nested nonparametric priors. *Bayesian Analysis* .

Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association* **99,** 1050–1059.

Diggle, P. J. and Al Wasel, I. (1997). Spectral analysis of replicated biomedical time series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46,** 31–71.

Edwards, M. C., Meyer, R., and Christensen, N. (2017). Bayesian nonparametric spectral density estimation using b-spline priors. *Statistics and Computing* pages 1–12.

Falconer, D. S. (1960). *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London.

Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association* **111,** 1440–1453.

Grandy, T. H., Werkle-Bergner, M., Chicherio, C., Lövdén, M., Schmiedek, F., and Lindenberger, U. (2013). Individual alpha peak frequency is related to latent factors of general cognitive abilities. *Neuroimage* **79,** 10–18.

Harper, J., Malone, S. M., and Iacono, W. G. (2018). Impact of alcohol use on eeg dynamics of response inhibition: a cotwin control analysis. *Addiction biology* **23,** 256–267.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., and McGue, M. (1999). Behavioral disinhibition and the development of substance-use disorders: findings from the minnesota twin family study. *Development and psychopathology* **11,** 869–900.

Iacono, W. G., Malone, S. M., and Vrieze, S. I. (2017). Endophenotype best practices. *International Journal of Psychophysiology* **111,** 115–144.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96,** 161–173.

Kakizawa, Y. (2006). Bernstein polynomial estimation of a spectral density. *Journal of Time Series Analysis* **27,** 253–287.

Krafty, R. T., Hall, M., and Guo, W. (2011). Functional mixed effects spectral analysis. *Biometrika* **98,** 583–598.

Krafty, R. T., Rosen, O., Stoffer, D. S., Buysse, D. J., and Hall, M. H. (2017). Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. *Journal of the American Statistical Association* **112,** 1405–1416.

Lizio, R., Vecchio, F., Frisoni, G. B., Ferri, R., Rodriguez, G., and Babiloni, C. (2011). Electroencephalographic rhythms in alzheimers disease. *International Journal of Alzheimers disease* **2011,**.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.

Malone, S. M., Burwell, S. J., Vaidyanathan, U., Miller, M. B., McGue, M., and Iacono, W. G. (2014). Heritability and molecular-genetic basis of resting eeg activity: A genome-wide association study. *Psychophysiology* **51,** 1225–1245.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*, volume 18. Springer.

Ombao, H. C., Raz, J. A., Strawderman, R. L., and Von Sachs, R. (2001). A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika* **88,** 1186–1192.

Petrone, S. (1999a). Bayesian density estimation using bernstein polynomials. *Canadian Journal of Statistics* **27,** 105–126.

Petrone, S. (1999b). Random bernstein polynomials. *Scandinavian Journal of Statistics* **26,** 373–393.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested dirichlet process. *Journal of the American Statistical Association* **103,** 1131–1154.

Rosen, O., Stoffer, D. S., and Wood, S. (2009). Local spectral analysis via a bayesian mixture of smoothing splines. *Journal of the American Statistical Association* **104,** 249–262.

Rudo-Hutt, A. S. (2015). Electroencephalography and externalizing behavior: a meta-analysis. *Biological psychology* **105,** 1–19.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica* pages 639–650.

Shumway, R. H. and Stoffer, D. S. (2010). *Time Series Analysis and its Applications: with R Examples*. Springer Science & Business Media.

Smit, D., Posthuma, D., Boomsma, D., and Geus, E. d. (2005). Heritability of background eeg across the power spectrum. *Psychophysiology* **42,** 691–697.

Smit, D. J., Wright, M. J., Meyers, J., Martin, N., Ho, Y. Y., Malone, S. M., Zhang, J., Burwell, S. J., Chorlian, D. B., de Geus, E. J., et al. (2017). Genome-wide association analysis links multiple psychiatric liability genes to oscillatory brain activity. *bioRxiv* page 232330.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association* **75,** 122–132.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research* **11,**.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik* **2,** 423–434.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supporting Information

Web Appendices, Tables, and Figures (limit to the types of material actually in question) referenced in Sections (give section numbers) are available with this paper at the Biometrics website on Wiley Online Library." If applicable, please also describe the availability of data/code in this Supporting Information section.

Acknowledgements

Computational resources for this work were provided by the Minnesota Supercomputing Institute at the University of Minnesota. This work was funded in part by the University of Minnesota Informatics Institute.

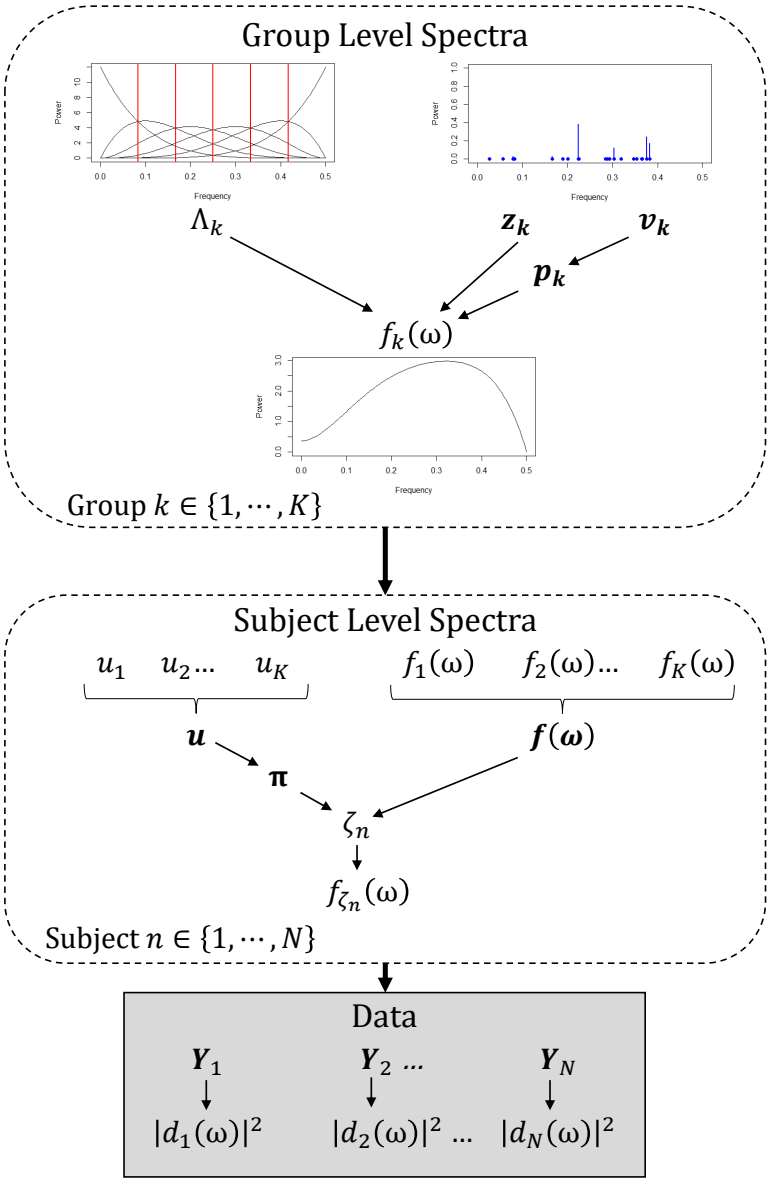

Figure 1. A schematic representation of the NBDP model. The BDP is used to estimate a group level spectra for each of the K different groups. The nested DP then assigns each of the N subjects to one of the K group spectral densities.

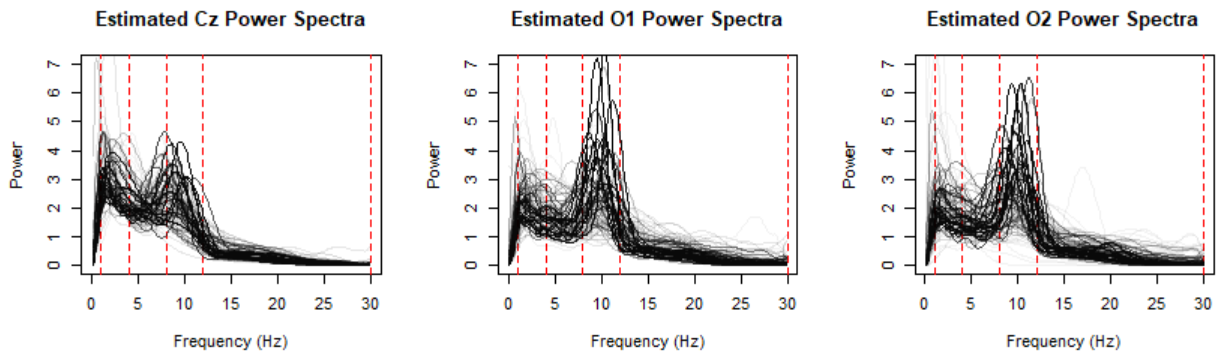

Figure 2. The NBDP estimated spectral density curves for each channel and each of the 1116 participants in the MTFS. Each line represents a single subject. The red vertical dashed lines represent the boundaries of the four frequency bands, namely, the delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), and beta (12-30 Hz) frequency bands. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Channel	Frequency Band	Heritability
Cz	Delta (1-4 Hz)	0.67
	Theta (4-8 Hz)	1.00
	Alpha (8-12 Hz)	0.87
	Beta (12-30 Hz)	0.72
O1	Delta (1-4 Hz)	0.41
	Theta (4-8 Hz)	0.70
	Alpha (8-12 Hz)	0.60
	Beta (12-30 Hz)	0.36
O2	Delta (1-4 Hz)	0.49
	Theta (4-8 Hz)	0.70
	Alpha (8-12 Hz)	0.45
	Beta (12-30 Hz)	0.36

Table 1

The estimated heritability for each of the four frequency bands and three EEG recording channels. Heritability estimates were calculated using generalized cross-validation (GCV) smoothed periodograms and Falconer's formula.

Scenario	Method	Delta	Theta	Alpha	Beta	Full Spectrum
Scenario 1: Power Spectrum Estimation	**GCV**	0.36 (0.20)	0.38 (0.17)	0.36 (0.19)	0.23 (0.10)	1.44 (0.40)
	Spline10	0.37 (0.22)	0.37 (0.19)	0.33 (0.17)	0.20 (0.09)	1.38 (0.40)
	Spline20	0.42 (0.22)	0.45 (0.19)	0.45 (0.21)	0.27 (0.10)	1.70 (0.41)
	BDP	0.35 (0.19)	0.52 (0.23)	0.40 (0.19)	0.20 (0.09)	1.62 (0.36)
	NBDP	**0.28 (0.16)**	**0.32 (0.18)**	**0.33 (0.16)**	**0.17 (0.07)**	**1.22 (0.32)**
Scenario 2: Power Spectrum Estimation	**GCV**	0.36 (0.21)	0.38 (0.17)	0.35 (0.17)	0.22 (0.10)	1.42 (0.38)
	Spline10	0.37 (0.22)	0.36 (0.18)	**0.33 (0.17)**	0.20 (0.09)	1.36 (0.39)
	Spline20	0.42 (0.23)	0.45 (0.20)	0.44 (0.22)	0.27 (0.10)	1.69 (0.41)
	BDP	0.35 (0.19)	0.52 (0.23)	0.39 (0.19)	0.19 (0.09)	1.60 (0.36)
	NBDP	**0.28 (0.17)**	**0.34 (0.20)**	0.36 (0.18)	**0.17 (0.07)**	**1.28 (0.33)**
Scenario 3: Power Spectrum Estimation	**GCV**	0.33 (0.16)	0.35 (0.15)	0.43 (0.18)	0.25 (0.10)	1.45 (0.35)
	Spline10	0.34 (0.17)	0.35 (0.17)	0.42 (0.19)	0.21 (0.09)	1.41 (0.38)
	Spline20	0.39 (0.19)	0.43 (0.18)	0.51 (0.21)	0.26 (0.10)	1.69 (0.38)
	BDP	0.33 (0.16)	0.51 (0.24)	0.49 (0.22)	0.20 (0.08)	1.67 (0.44)
	NBDP	**0.30 (0.15)**	**0.34 (0.15)**	**0.39 (0.20)**	**0.18 (0.06)**	**1.35 (0.35)**
Scenario 4: Power Spectrum Estimation	**GCV**	0.33 (0.16)	0.36 (0.15)	0.45 (0.19)	0.25 (0.10)	1.48 (0.37)
	Spline10	0.34 (0.17)	0.36 (0.17)	0.44 (0.19)	0.21 (0.09)	1.44 (0.38)
	Spline20	0.39 (0.19)	0.43 (0.18)	0.52 (0.22)	0.26 (0.09)	1.71 (0.42)
	BDP	0.33 (0.16)	0.50 (0.24)	0.51 (0.22)	**0.20 (0.08)**	1.68 (0.43)
	NBDP	**0.30 (0.15)**	**0.33 (0.15)**	**0.42 (0.20)**	0.20 (0.07)	**1.37 (0.35)**
Scenario 5: Power Spectrum Estimation	**GCV**	0.34 (0.20)	0.42 (0.21)	**0.50 (0.23)**	0.37 (0.14)	**1.74 (0.42)**
	Spline10	**0.32 (0.20)**	0.80 (0.40)	0.81 (0.34)	0.34 (0.15)	2.37 (0.78)
	Spline20	0.35 (0.25)	0.47 (0.38)	0.76 (0.35)	**0.25 (0.11)**	1.95 (0.62)
	BDP	0.37 (0.22)	**0.39 (0.25)**	0.58 (0.28)	0.80 (0.27)	2.34 (0.45)
	NBDP	0.41 (0.31)	0.44 (0.35)	0.56 (0.34)	0.69 (0.25)	2.28 (0.66)
Scenario 5: Heritability Estimation	**GCV**	4.0	5.4	6.1	7.4	23.2
	Spline10	**2.9**	5.7	**3.7**	11.3	24.7
	Spline20	5.9	5.8	4.3	12.0	28.6
	BDP	3.6	**4.6**	4.6	12.2	25.6
	NBDP	6.0	5.5	6.3	**2.6**	**21.5**

Table 2

Mean (sd) mean integrated absolute error (MIAE) in spectral density estimation across the 1,000 simulated subjects for each simulation scenario. The final section is the IAE of the heritability spectrum of each method for Scenario 5. ***Bold*** *numbers indicate the best performing model in each scenario and frequency band.*

Channel	**Cz**		**O1**		**O2**	
Method	WAIC	p_{WAIC}	WAIC	p_{WAIC}	WAIC	p_{WAIC}
Spline10	-189	9,928	-181	9,275	-181	9,380
Spline20	-192	13,667	-185	13,231	-185	13,449
BDP	-189	6,642	-183	7,617	-183	7,501
NBDP	-188	2,208	-182	2,401	-182	1,944

Table 3

WAIC and effective number of parameters for each Bayesian model and EEG channel for the MTFS data. All WAIC values are scaled by 0.0001 for readability.

Channel	Feature	NBDP Median (IQR)	GCV Median (IQR)	NBDP Heritability (CI)	GCV Heritability	Malone 2014 Heritability	Smit 2005 Heritability
Cz	Delta Band	0.25 (0.23 - 0.32)	0.26 (0.20 - 0.33)	0.43 (0.33, 0.53)	0.48	0.49	0.79
	Theta Band	0.28 (0.25 - 0.31)	0.28 (0.22 - 0.33)	0.88 (0.84, 0.92)	0.88	0.69	0.87
	Alpha Band	0.29 (0.23 - 0.33)	0.29 (0.20 - 0.39)	0.77 (0.69, 0.87)	0.86	0.84	0.93
	Beta Band	0.13 (0.12 - 0.18)	0.08 (0.06 - 0.11)	0.65 (0.55, 0.77)	0.83	0.85	0.88
	Full Spectrum	1.00	1.00	0.68 (0.62, 0.74)	0.75	-	-
	Peak Frequency	2.25 (1.63 - 8.63)	7.38 (2.50 - 9.38	0.51 (0.35, 0.66))	0.48	-	-
	Alpha Peak	8.63 (8.13 - 9.63)	9.00 (8.25 - 9.88)	0.81 (0.70, 0.91)	0.67	-	-
O1	Delta Band	0.18 (0.15 - 0.21)	0.15 (0.11 - 0.21)	0.30 (0.18, 0.40)	0.35	-	0.60
	Theta Band	0.21 (0.16 - 0.24)	0.19 (0.13 - 0.25)	0.49 (0.43, 0.57)	0.63	-	0.82
	Alpha Band	0.39 (0.30 - 0.45)	0.46 (0.33 - 0.59)	0.48 (0.44, 0.54)	0.71	0.78	0.85
	Beta Band	0.19 (0.14 - 0.26)	0.12 (0.08 - 0.20)	0.39 (0.32, 0.46)	0.57	-	0.83
	Full Spectrum	1.00	1.00	0.45 (0.41, 0.50)	0.66	-	-
	Peak Frequency	9.50 (8.63 - 10.25)	9.50 (8.59 - 10.25)	0.37 (0.21, 0.54)	0.55	-	-
	Alpha Peak	9.56 (9.00 - 10.25)	9.75 (9.00 - 10.38)	0.52 (0.35, 0.67)	0.95	0.83	-
O2	Delta Band	0.17 (0.15 - 0.21)	0.14 (0.10 - 0.20)	0.40 (0.32, 0.47)	0.41	-	0.66
	Theta Band	0.21 (0.17 - 0.26)	0.19 (0.13 - 0.26)	0.67 (0.60, 0.73)	0.62	-	0.85
	Alpha Band	0.38 (0.30 - 0.45)	0.46 (0.32 - 0.57)	0.58 (0.53, 0.63)	0.60	0.78	0.85
	Beta Band	0.21 (0.15 - 0.26)	0.13 (0.08 - 0.20)	0.56 (0.50, 0.60)	0.72	-	0.83
	Full Spectrum	1.00	1.00	0.56 (0.52, 0.60)	0.58	-	-
	Peak Frequency	9.63 (8.50 - 10.38)	9.50 (8.38 - 10.38)	0.25 (0.13, 0.38)	0.13	-	-
	Alpha Peak	9.63 (8.88 - 10.38)	9.63 (8.88 - 10.38)	0.86 (0.69, 1.00)	0.79	0.83	-

Table 4

The median and inter-quartile range of different spectral density features across the 1116 MTFS subjects along with the heritability calculated from the sample for each feature and channel for both the NBDP and GCV models.