Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data

Seyed Mostafa Kia1,2
Andre F. Marquand1,2,3

1 Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, Netherlands
2 Donders Institute for Brain Cognition and Behaviour, Nijmegen, Netherlands
3 Department of Neuroimaging, Institute of Psychiatry, Kings College, London, United Kingdom

Abstract
Normative modeling has recently been introduced as a promising approach for modeling variation of neuroimaging measures across individuals in order to derive biomarkers of psychiatric disorders. Current implementations rely on Gaussian process regression, which provides coherent estimates of uncertainty needed for the method but also suffers from drawbacks including poor scaling to large datasets and a reliance on fixed parametric kernels. In this paper, we propose a deep normative modeling framework based on neural processes (NPs) to solve these problems. To achieve this, we define a stochastic process formulation for mixed-effect models and show how NPs can be adopted for spatially structured mixed-effect modeling of neuroimaging data. This enables us to learn optimal feature representations and covariance structure for the random-effect and noise via global latent variables. In this scheme, predictive uncertainty can be approximated by sampling from the distribution of these global latent variables. On a publicly available clinical fMRI dataset, we compare the novelty detection performance of multivariate normative models estimated by the proposed NP approach to a baseline multi-task Gaussian process regression approach and show substantial improvements for certain diagnostic problems.

Keywords: Neural Processes, Mixed-Effect Modeling, Deep Learning, Neuroimaging.

1. Introduction
Recently, there has been great interest in applying machine learning methods to neuroimaging data in order to find structural or functional biomarkers for brain disorders. Such biomarkers can potentially be used for diagnosis or predicting treatment outcome in the spirit of precision medicine (Mirnezami et al., 2012). In psychiatry, this is very challenging because clinical groups are highly heterogeneous in terms of symptoms and underlying biology (Kapur et al., 2012). However, most common analysis approaches ignore such heterogeneity and instead consider groups as distinct entities (Foulkes and Blakemore, 2018), e.g., in a case-control setting where subjects are simply labeled as ‘patients’ or ‘controls’. Supervised machine learning methods have been widely used in such settings but their accuracy is limited by the heterogeneity within each disorder (Wolfers et al., 2015).

Normative modeling (Marquand et al., 2016) is an emerging approach to address this challenge that has shown significant promise in multiple clinical settings (Wolfers et al., 2018; Zabihi et al., 2018). Intuitively, normative modeling involves estimating variation across the population in terms of mappings between clinically relevant covariates (e.g., age,
cognitive scores) and biology (e.g., fMRI activity patterns). This is analogous to the use of ‘growth charts’ in pediatric medicine to map variation in height or weight as a function of age. Currently, this is implemented using probabilistic regression methods that provide estimates of predictive uncertainty which map variation across the population. Deviations from the resulting normative model can then be interpreted as subject-specific biomarkers for brain disorders. For example, these can be used in a novelty detection setting for predicting diagnosis in an unsupervised fashion (Kia and Marquand, 2018; Kia et al., 2018).

Accurate quantification of uncertainty is crucial for normative modeling. In the original framework (Marquand et al., 2016), Gaussian process regression (Williams and Rasmussen, 1996) (GPR) was the central tool used to regress neuroimaging measures from clinical covariates. GPR is appealing because it estimates a distribution over functions, providing coherent estimates of uncertainty to map population variation. However GPR also has limitations: it is computationally prohibitive for large datasets and relies on predefined kernels with restricted functional form. Moreover, in the original implementation, brain measures were regressed independently (i.e., in a mass-univariate manner), which does not capitalize on the rich spatial structure of neuroimaging data. This last problem can be addressed by using multi-task GPR (MT-GPR) (Bonilla et al., 2008) to jointly predict multiple brain measurements. However, applying MT-GPR to neuroimaging data is very computationally demanding because of the need to invert large covariance matrices across both space and subjects. Recently, a combination of a low-rank approximations and Kronecker algebra was proposed to scale MT-GPR to whole brain neuroimaging data (Kia and Marquand, 2018; Kia et al., 2018), which reduces the computational complexity by one order of magnitude. However, this comes with restrictive assumptions that the spatial structures of the signal and noise can be expressed by sets of orthogonal basis functions.

Neural processes (NP) (Garnelo et al., 2018a,b) are latent variable models that bring all the advantages of deep learning (e.g., representation learning and computationally efficient training and prediction) to a stochastic process framework and can address the problems described above. In the NP framework, a distribution over functions is modeled by learning an approximation to a stochastic process. Here, we present an application of NP to multivariate normative modeling of clinical neuroimaging data. This provides three advantages: i) like GPR, NP provides the necessary estimates of predictive uncertainty at test time; ii) similar to MT-GPR, it provides the possibility of learning structured variation; and iii) unlike alternatives, it is computationally scalable without restrictive assumptions on the orthogonality of lower dimensional representations of data. To this end, we make four contributions: i) in a tensor Gaussian predictive process (TGPP) framework (Kia et al., 2018), we define mixed-effect models of neuroimaging data (Friston et al., 1999) as stochastic processes; ii) we show how NPs can be employed for mixed-effect modeling; iii) we use the resulting NP-based mixed-effect model to estimate a normative model of a clinical functional magnetic resonance imaging (fMRI) dataset; iv) we provide an example application of the proposed deep normative modeling for detecting psychiatric disorders in a novelty detection setting. To this end, we apply the proposed NP-based mixed-effect model to a clinical fMRI dataset (Poldrack et al., 2016) in order to jointly predict task-related fMRI brain activity from a set of impulsivity measures. Our experimental results show that the proposed method more accurately identifies ADHD patients from healthy individuals compared to the GP-based normative modeling.
2. Methods

In this text, we use respectively calligraphic capital letters, \( \mathcal{A} \), boldface capital letters, \( \bf{A} \), and capital letters, \( A \), to denote tensors, matrices, and scalar numbers. We use \( \times_n \) to denote \( n \)-mode tensor products. We denote the vertical vector which results from collapsing the entries of a tensor \( \mathcal{A} \) into a vector with \( \text{vec}(\mathcal{A}) \). Notation \( |.| \) is accordingly used to represents the determinant of a matrix or the size of a set.

2.1. Mixed-Effect Modeling of MRIs in the TGPP Framework

Consider a neuroimaging study with \( N \) subjects and let \( X \in \mathbb{R}^{N \times D} \) to denote the design matrix of \( D \) covariates of interest for \( N \) subjects. Let \( Y \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3} \) to represent a 4-order tensor of MRI data for corresponding \( N \) subjects with respectively \( T_1 \), \( T_2 \), and \( T_3 \) voxels in \( x \), \( y \), and \( z \) axes. In the normative modeling setting, we are interested in finding the function \( f : X \rightarrow Y \). Adopting the tensor Gaussian predictive process (TGPP) (Kia et al., 2018) for structured multi-way mixed-effect modeling of MRI data, we have:

\[
Y = f(X) = X \times_1 \mathcal{A} + Z + \mathcal{E},
\]

where \( \mathcal{A} \in \mathbb{R}^{D \times T_1 \times T_2 \times T_3} \) represents the fixed-effect across subjects that contains regression coefficients estimated by solving the following linear equations:

\[
\hat{Y}[:,i,j,k] = XA[:,i,j,k], \quad \text{for} \quad i = 1, \ldots, T_1; \quad j = 1, \ldots, T_2; \quad k = 1, \ldots, T_3. \tag{2}
\]

In Equation (1), \( Z \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3} \) is the random-effect that characterizes the spatially structured joint variations from the fixed-effect across individuals in different dimensions of MRIs; and \( \mathcal{E} \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3} \) is heteroscedastic noise. Assuming a tensor-variate normal distribution for \( Y \) and a zero-mean tensor-variate normal distribution for \( Z + \mathcal{E} \), we have:

\[
p(X, Y) = \mathcal{T}\mathcal{N}(\hat{Y}, \mathbf{S}) = \frac{\exp(-\frac{1}{2} \text{vec}(Y - \hat{Y})^\top S^{-1} \text{vec}(Y - \hat{Y}))}{\sqrt{(2\pi)^{NT} |S|^{NT}}}, \tag{3}
\]

where \( S \in \mathbb{R}^{NT \times NT} \) is the covariance matrix of \( Z + \mathcal{E} \). Intuitively, the distribution of the mixed-effect in the joint hypercubic space of clinical covariates and neuroimaging measures can be described as a multi-dimensional Gaussian distribution with \( \text{vec}(\hat{Y}) \) and \( \mathbf{S} \) respectively serving as its mean and covariance.

2.2. Mixed-Effect Models of MRI Data as Stochastic Processes

Let \( (\Omega, \Phi, \rho) \) represent a complete probability space (see Oksendal (2003) for definitions) where \( \Omega \) is a set of neuroimaging measures for \( N \rightarrow +\infty \) subjects (i.e., \( |\Omega| = N \)) and \( \Phi \) is a \( \sigma \)-algebra on \( \Omega \) that contains all possible subsets of \( \Omega \). In other words, \( \Phi \) represents a set of all possible random neuroimaging samples drawn from the whole human population. Here, \( \rho : \Phi \rightarrow [0, 1] \) represents a probability measure that quantifies the probability of occurrence for any entry in \( \Phi \). In this setting, each mixed-effect function \( f_i \) estimated on the \( i \)th entry of \( \Phi \) is a random variable, i.e., a \( \Phi \)-measurable function from \( \Omega \) to a Borel set in \( \mathbb{R}^{N \times T_1 \times T_2 \times T_3} \). Therefore, parametrizing \( f_i \) on different subsets of \( \Omega \) and \( X \), and considering the exchangeability and consistency properties of mixed-effect models (McCullagh, 2005; Nie and Yang, 2005), \( Y_i = f_i(X_i) |_{[\Phi]} \) can be defined as stochastic processes (Garnelo et al., 2018b). As a corollary, for each entry in \( \Phi \), its joint distribution in Equation (3) can be considered as a marginal for a higher-dimensional joint distribution when \( N \rightarrow +\infty \). We
exploit this property to frame the problem of mixed-effect modeling in the neural processes framework (Garnelo et al., 2018b). To this end, given a particular realization of the mixed-effect stochastic process \( f_i \), the joint distribution in Equation (3) can be rewritten as:

\[
p(X, Y) = \sum_{i=1}^{[\Phi]} p(f_i) T N(\hat{Y} \mid f_i, S) .
\]  

(4)

In a NP paradigm, we parametrize the integration over all \( f_i(X) \) on a lower dimensional \((Q << T)\) Gaussian distributed global latent variable \( Z \in \mathbb{R}^{N \times Q} \sim \mathcal{G}(\mu, \Sigma) \) where \( f(X) = g(X, Z) \), resulting the following generative model:

\[
p(Y, Z \mid X) = p(Z) T N(\hat{Y} \mid g(X, Z), S) ,
\]  

(5)

where \( g(X, Z) \) is a deep neural network that learns the behavior of the mixed-effect model in an amortized variational inference regime. To this end, following the procedure proposed by Garnelo et al. (2018b) the first challenge is to induce stochasticity, for which we need to define ‘context’ and ‘target’ functions. The idea is to reduce the difference between the distribution of random context functions from the target function by minimizing their Kullback-Leibler (KL) divergence in the latent space. In our application in order to learn the distribution of the mixed-effect model in Equation (1), we propose to use the estimated \( \hat{Y}_C \in \mathbb{R}^{N \times M \times T_1 \times T_2 \times T_3} \) (using Equation (2)) on \( M \) randomly drawn subsets of the training set as context functions. Then, using the actual corresponding neuroimaging training samples as target functions, the following evidence lower-bound should be optimized:

\[
\log p(Y \mid X, \hat{Y}_C) \geq \mathbb{E}_{q(Z \mid X, \hat{Y}_C)} \left[ \log p(Y \mid Z, X) + \log \frac{q(Z \mid X, \hat{Y}_C)}{q(Z \mid X, Y)} \right] ,
\]  

(6)

where \( q(Z \mid X, Y) \) is the variational posterior of the global latent variable that is parametrized on an encoder \( h(X, \hat{Y}_C) \). In fact in this setting, each context function is a linear component of the target function that roughly approximates a stochastic process \( f_i \). Having enough samples of context functions, large enough \( M \), we expect the distribution of context functions to get rich enough to explain non-linear characteristics of the target function (i.e., the mixed-effect \( f_i \)). Figure 1 shows a simplified illustration of this scenario in a 2D space where fitting enough linear models on subsets of noisy observations provides an estimation of the distribution of a non-linear target function. Furthermore, by minimizing the KL term in Equation (6), it is expected that the global latent variable \( Z \) will learn characteristics of the variance structure of the random-effect and noise terms (the diagonal elements of \( S \)) from the difference between the context and target functions (recall that \( Y - \hat{Y} = Z + \mathcal{E} \)).

2.3. Deep Normative Modeling using Neural Processes

Using NPs in the TGPP framework brings all the advantages of deep learning methods (e.g., representation learning from structured data and computational efficiency) for modeling the multi-way structured variation in neuroimaging data. It has been shown that modeling such structured variation provides the possibility of accurate unsupervised stratification of psychiatric patients in the normative modeling paradigm (Kia and Marquand, 2018; Kia et al., 2018). To this end, here we introduce deep normative modeling, which utilizes an NP-based mixed-effect modeling and involves following three steps:
Neural Processes for Deep Normative Modeling

Figure 1: A schematic illustration on approximating the distribution of a non-linear target function (red curve), e.g., a mixed-effect, from the distribution of linear context functions (blue lines), e.g., fixed-effects, which are fitted on $M$ random subsets of noisy observations (circles).

1. **Encoding phase:** where an encoder $h(X, \hat{Y}_C)$ is learned to transfer the covariates, $X$, and the estimated fixed-effects on $M$ randomly drawn samples from the training set, $\hat{Y}_C$, to the parameters of the global latent variable $Z$. Here, to preserve the 3D MRIs structure in the TGPP framework, we propose to use 3D convolutional neural network (3D-CNN) layers to first transfer the $\hat{Y}_C$ to a lower dimensional representation of neuroimages $R_{\hat{Y}} \in \mathbb{R}^{N \times T'}$. Then, fully connected (FC) layers can be used to derive a latent representation in the joint space of clinical covariates ($X$) and neuroimages, $R \in \mathbb{R}^{N \times T''}$. It is worthwhile to emphasize that in this architecture, the aggregation across $M$ context functions is implicitly done by the 3D-CNN layers as they are considered as $M$ input channels to the CNN. Finally, two separate FC layers are used to transfer $R$ to the means ($\mu_Z \in \mathbb{R}^{N \times Q}$) and standard deviations ($\sigma_Z \in \mathbb{R}^{N \times Q}$) of $Z$.

2. **Decoding phase:** where a decoder $g(X, Z)$ is learned to transfer back the joint covariates-latent space to the neuroimaging data $Y$. Fully connected and 3D inverse CNN (3D-ICNN) layers can be accordingly used to reconstruct MRIs in the original space.

3. **Normative modeling:** let $Y^* \in \mathbb{R}^{N^* \times T_1 \times T_2 \times T_3}$ to represent the reconstructed neuroimaging data by the decoder $g(X^*, Z)$ on $N^*$ test samples. Following Marquand et al. (2016), we then compute statistical maps describing the deviation for each individual subject from the normative model, referred to as normative probability maps (NPMs), denoted by $N \in \mathbb{R}^{N^* \times T_1 \times T_2 \times T_3}$ where $N = (Y - Y^*)/\sqrt{S}$. Here, $S$ represents the sum of epistemic and aleatoric uncertainties, which respectively describe uncertainty about the true model parameters and inherent variation in the data (Kendall and Gal, 2017). To be able to calculate the epistemic uncertainty in our NP model, we keep the dropout layers active at test time (Gal and Ghahramani, 2016). In the context of mixed-effect modeling of neuroimaging data (in Equation (1)), the aleatoric uncertainty is the byproduct of two factors: i) the across-subject variability which is captured via the covariance of the random-effect $Z$; and ii) noise in the data which is captured via covariance of $E$. In the proposed NP framework, these two sources of uncertainties are learned from data and are summarized in the distribution of the global latent variable $Z$. Therefore, given a test example of clinical covariates $x^* \in X^*$, we calculate the associated aleatoric uncertainty by sampling from the distribution of $Z$.

3. **Experimental Material and Setups**

In our experiments, we use the response inhibition (i.e., ‘stop signal’) task from the UCLA Consortium for Neuropsychiatric Phenomics dataset (Poldrack et al., 2016). Specifically,
The data consist of 119 healthy subjects; and 49, 39, and 48 individuals with schizophrenia (SCHZ), attention deficit hyperactivity disorder (ADHD), and bipolar disorder (BIPL), respectively. We cropped the volumes to the minimal bounding-box of $49 \times 61 \times 40$ voxels ($T_1 = 49, T_2 = 61, T_3 = 40, T = 119560$). In order to accommodate the optimization scheme in Equation (6) for fMRI data, the values of voxels are independently projected to the uniform $[0, 1]$ interval using a robust quantile transformation. For clinical covariates, we use 11 factors of Barratt impulsiveness scores (Patton et al., 1995) ($D = 11$) as impulsivity is a well-known feature for multiple psychiatric disorders and is implicated in response inhibition (Moeller et al., 2001).

We use three layers of 3D-CNNs followed by an FC layer to project $\hat{Y}_C$ to $\mathbf{R}\hat{Y}$. In each CNN layer, we alternate a 3D-convolutional layer, a batch normalization layer (Ioffe and Szegedy, 2015), an average pooling layer, and a leaky ReLU activation function (Xu et al., 2015) (with negative slope of $0.01$). Then, two FC layers are used to transfer the merged $\mathbf{R}\hat{Y}$ and $\mathbf{X}$ to the middle joint representation $\mathbf{R}$. A similar reverse architecture is used for the decoder $g(\mathbf{X}, \mathbf{Z})$ to transfer back the $\mathbf{Z}$ to $\mathbf{Y}$ space. Figure 2 depicts a schematic of the employed NP architecture with detailed hyperparameter descriptions. Due to the small sample size and illustrative purpose of our experiments, we did not optimize the architecture and its hyperparameters (e.g., number of layers, number and the size of filters, number of neurons, etc.). The ADAM optimizer (Kingma and Ba, 2014) with decreasing learning rate (from $10^{-2}$ to $10^{-5}$) is used for optimization in 100 epochs.

We compare the normative models derived by NP and scalable multi-task Gaussian process tensor regression (sMT-GPTR) (Kia et al., 2018), in terms of their accuracy in detecting healthy subjects from patients. In the sMT-GPTR case, we set the number of basis functions across xyz dimensions of data 5, 10 and 3, 5 for the signal and noise, respectively (as they produced the best results in the original study). We evaluate normative modeling

---

1. Available at https://openfmri.org/dataset/ds000030/.
Figure 3: Comparison between novelty detection performances of normative models derived by sMT-GPTR (with different number of bases for signal and noise) and NPs (with different $M$).

accuracy in a novelty detection scenario where we first train a model on a random subset of subjects (75 healthy, 5 SCHZ, 5 ADHD, and 5 BIPL) and then calculate NPMs on a test set of remaining healthy subjects and patients. We emphasize that the model has no access to the diagnostic labels during the training phase and thus our novelty detection approach is completely unsupervised. As in Marquand et al. (2016), we use extreme value statistics to provide a statistical model for the deviations. Specifically, we use a block-maximum approach on the top 1% values in NPMs and fit these to a generalized extreme value distribution (GEVD) (Davison and Huser, 2015). Then for a given test sample and given the shape parameter of GEVD, we compute the value of the cumulative distribution function of GEVD as the probability of that sample being an abnormal sample (Roberts, 2000). Given these probabilities and actual labels, we evaluate the area under the ROC curve (AUC) to measure the performance of the model. All steps (random sampling, modeling, and evaluation) are repeated 10 times in order to estimate the fluctuations of models trained on different training sets. In all these experiments, ordinary least squares are used to estimate the fixed-effect (Equation (2)) on bootstrapped subsets of the training set.\(^2\)

4. Results

Figure 3 compares the AUC of normative models derived by sMT-GPTR and NP. While sMT-GPTR shows slightly better performance in detecting SCHZ patients, NP provides substantially higher accuracy for ADHD cases. The methods perform similarly for BIPL. Considering the fact that these differences in performance are consistent across different model parameters and repetitions, it can be concluded that sMT-GPTR and NP are capturing different characteristics of the underlying biology of impulsivity. Furthermore, the above chance-level detection rates of NP models in ADHD and BIPL confirm a successful application of the proposed NP-based mixed-effect modeling in unsupervised diagnostic prediction. Another important observation in NP models is the ascending trend of the detection performance as the number of samples from the fixed-effect (i.e., $M$) increases. This is compatible with the consistency property of mixed-effects as stochastic processes.

Figure 4(a) depicts the average difference in NPMs of patient groups from the healthy population for NP(20) model (see Appendix A for supplementary results). Different patterns of deviations from one diagnosis to another shed light on their different underlying biological causes. For example, the sign of deviations changes from SCHZ to ADHD patients in many regions. To further explore the link between these deviations and the level of impulsivity, we computed the coefficient of determination ($R^2$) between the average NPMs

\(^2\) The scripts for experiments are available at https://github.com/smkia/DNM.
in 9 anatomical brain areas and the first principal component of covariates across different diagnostic groups (see Figure 4(b)). The results show significantly (Bonferroni corrected F-test p-values) greater association between impulsivity and deviations in temporal lobes in ADHD and SCHZ patients compared to healthy individuals. This observation is compatible with previous research on the structural and functional engagement of temporal lobes in SCHZ and ADHD (Suddath et al., 1989; Kobel et al., 2010).

5. Related Work

Rad et al. (2018) used a convolutional autoencoder for unimodal deep normative modeling of human movements recorded by wearable sensors. They used dropout technique in order to evaluate the parameter uncertainty of the model. Our proposed NP-based approach extends their effort in applying deep architectures to normative modeling from two perspectives: i) it provides the possibility of bimodal normative modeling. This is more appropriate for clinical usages where we are generally interested in interpreting the association between clinical covariates and biological measures (Marquand et al., 2016); ii) using the fully probabilistic NP regime, we are also capable to evaluate aleatoric uncertainties resulting from individual differences and noise in addition to the epistemic parameter uncertainty.

6. Summary and Discussion

In this paper, we proposed a principled approach for estimating spatially structured mixed-effects in neuroimaging data using neural processes. We demonstrated normative modeling as a possible target application for NP-based mixed-effect modeling. Even though the main focus in this study was on neuroimaging data, our contribution in framing the popular mixed-effect modeling as stochastic processes is quite general and opens the door for a wide range of NP applications in different research areas. Moreover, the presented application of NP for deep normative modeling of clinical neuroimaging data brings the advantages of deep neural networks in representation learning to the applications in precision psychiatry. Finally, the computational efficiency of NP in the training and evaluation phases (provided by its reliance on the variational inference) overcomes the lack of computational tractability of the GP-based normative modeling approaches especially when applied to large cohorts of high-dimensional neuroimaging data. For a possible future direction, we consider applying the proposed deep normative modeling approach to a large clinical neuroimaging cohort.
References

Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task Gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2008.

Anthony C Davison and Raphaël Huser. Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235, 2015. doi: 10.1146/annurev-statistics-010814-020133. URL https://doi.org/10.1146/annurev-statistics-010814-020133.

Lucy Foulkes and Sarah-Jayne Blakemore. Studying individual differences in human adolescent brain development. *Nature neuroscience*, page 1, 2018.

Karl J Friston, Andrew P Holmes, CJ Price, C Büchel, and KJ Worsley. Multisubject fMRI Studies and Conjunction Analyses. *NeuroImage*, 10(4):385 – 396, 1999. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.1999.0484. URL http://www.sciencedirect.com/science/article/pii/S1053811999904846.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/gal16.html.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018a. PMLR. URL http://proceedings.mlr.press/v80/garnelo18a.html.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Krzysztof J Gorgolewski, Joke Durnez, and Russell A Poldrack. Preprocessed consortium for neuropsychiatric phenomics dataset [version 2; referees: 2 approved]. *F1000Research*, 6(1262), 2017. doi: 10.12688/f1000research.11964.2.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/ioffe15.html.

Shitij Kapur, Anthony G Phillips, and Thomas R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry*, 17(12):1174, 2012.
Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems, pages 5580–5590, 2017.

Seyed Mostafa Kia and Andre Marquand. Normative modeling of neuroimaging data using scalable multi-task gaussian processes. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pages 127–135, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00931-1.

Seyed Mostafa Kia, Christian F. Backmann, and Andre F. Marquand. Scalable multi-task gaussian process tensor regression for normative modeling of structured variation in neuroimaging data. arXiv preprint arXiv:1808.00036, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Maja Kobel, Nina Bechtel, Karsten Specht, Markus Klarhfer, Peter Weber, Klaus Scheffler, Klaus Opwis, and Iris-Katharina Penner. Structural and functional imaging approaches in attention deficit/hyperactivity disorder: Does the temporal lobe play a key role? Psychiatry Research: Neuroimaging, 183(3):230 – 236, 2010. ISSN 0925-4927. doi: https://doi.org/10.1016/j.pscychresns.2010.03.010. URL http://www.sciencedirect.com/science/article/pii/S0925492710001137.

Andre F Marquand, Iead Rezek, Jan Buitelaar, and Christian F Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biological psychiatry, 80(7):552–561, 2016.

Peter McCullagh. Exchangeability and regression models. Oxford Statistical Science Series, 33:89, 2005.

Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for Precision Medicine. New England Journal of Medicine, 366(6):489–491, 2012. doi: 10.1056/NEJMp1114866. PMID: 22256780.

F. Gerard Moeller, Ernest S. Barratt, Donald M. Dougherty, Joy M. Schmitz, and Alan C. Swann. Psychiatric aspects of impulsivity. American Journal of Psychiatry, 158(11):1783–1793, 2001. doi: 10.1176/appi.ajp.158.11.1783. URL https://doi.org/10.1176/appi.ajp.158.11.1783. PMID: 11691682.

Lei Nie and Min Yang. Strong consistency of mle in nonlinear mixed-effects models with large cluster size. Sankhyā: The Indian Journal of Statistics, pages 736–763, 2005.

Bernt Oksendal. Stochastic differential equations: an introduction with applications, 2003.

Jim H. Patton, Matthew S. Stanford, and Ernest S. Barratt. Factor structure of the barratt impulsiveness scale. Journal of Clinical Psychology, 51(6):768–774, 1995. doi: 10.1002/1097-4679%28199511%29513A6%3C768%3A%3AAID-JCLP2270510607%3E3.0.CO;2-1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-4679%28199511%29513A6%3C768%3A%3AAID-JCLP2270510607%3E3.0.CO;2-1.
Russell A Poldrack, Eliza Congdon, William Triplett, KJ Gorgolewski, KH Karlgodt, JA Mumford, FW Sabb, NB Freimer, ED London, TD Cannon, et al. A phenome-wide examination of neural and cognitive function. *Scientific data*, 3:160110, 2016.

Nastaran Mohammadian Rad, Twan van Laarhoven, Cesare Furlanello, and Elena Marchiori. Novelty detection using deep normative modeling for imu-based abnormal movement monitoring in parkinsons disease and autism spectrum disorders. *Sensors*, 18(10), 2018. ISSN 1424-8220. doi: 10.3390/s18103533. URL http://www.mdpi.com/1424-8220/18/10/3533.

S.J. Roberts. Extreme value statistics for novelty detection in biomedical data processing. *IEE Proceedings - Science, Measurement and Technology*, 147:363–367(4), November 2000. ISSN 1350-2344. URL http://digital-library.theiet.org/content/journals/10.1049/ip-smt_20000841.

Richard L Suddath, Manuel F Casanova, Terry E Goldberg, David G Daniel, John R Kelsoe Jr, and Daniel R Weinberger. Temporal lobe pathology in schizophrenia: a quantitative magnetic resonance imaging study. *American Journal of Psychiatry*, 146(4):464–472, 1989. doi: 10.1176/ajp.146.4.464. URL https://doi.org/10.1176/ajp.146.4.464. PMID: 2929746.

Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

Thomas Wolters, Jan K. Buitelaar, Christian F. Beckmann, Barbara Franke, and Andre F. Marquand. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews*, 57:328 – 349, 2015. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2015.08.001. URL http://www.sciencedirect.com/science/article/pii/S0149763415002018.

Thomas Wolters, Nhat Trung Doan, Tobias Kaufmann, and et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry*, 75(11):1146–1155, 2018. doi: 10.1001/jamapsychiatry.2018.2467. URL +http://dx.doi.org/10.1001/jamapsychiatry.2018.2467.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Mariam Zabihi, Marianne Oldehinkel, Thomas Wolters, Vincent Frouin, David Goyard, Eva Loth, Tony Charman, Julian Tilman, Tobias Banaschewski, Guillaume Dumas, Rosemary Holt, Simon Baron-Cohen, Sarah Durston, Sven Bolte, Declan Murphy, Christine Ecker, Jan K. Buitelaar, Christian F. Beckmann, and Andre F. Marquand. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *bioRxiv*, 2018. doi: 10.1101/477596. URL https://www.biorxiv.org/content/early/2018/11/27/477596.
Appendix A. Supplementary Results

Figure 5: The average difference between NPMs of healthy subjects and patients for NP models with different $M$.

Figure 6: The average difference between NPMs of healthy subjects and patients for sMT-GPTR models with different number of basis functions for the signal and noise.