Highlights

Non-Gaussian Normative Modelling With Hierarchical Bayesian Regression

Augustijn A.A. de Boer, Seyed Mostafa Kia, Saige Rutherford, Mariam Zabihi, Charlotte Fraza, Pieter Barkema, Lars T. Westlye, Ole A. Andreassen, Max Hinne, Christian F. Beckmann, Andre Marquand

- We extended the Hierarchical Bayesian Regression framework for normative modelling
- Our extension allows modelling data with heteroskedastic skewness and kurtosis
- We developed a reparameterization of the SHASH distribution, suitable for sampling
- We provide the first implementation of the SHASH distribution in a fully Bayesian framework
- Results show that the extension outperforms current methods on various measures

Non-Gaussian Normative Modelling With Hierarchical Bayesian Regression

Augustijn A.A. de Boer[a,b,*], Seyed Mostafa Kia[a,b,c], Saige Rutherford[a,b,d], Mariam Zabihi[a,b,e], Charlotte Fraza[a,b], Pieter Barkema[a,b], Lars T. Westlye[h,i,j], Ole A. Andreassen[i,j], Max Hinne[a], Christian F. Beckmann[a,b,f], Andre Marquand[a,b,g]

[a] *Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Nijmegen, The Netherlands*
[b] *Department for Cognitive Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands*
[c] *Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands*
[d] *Department of Psychiatry, University of Michigan, Ann Arbor, MI, United States*
[e] *MRC Unit for Lifelong Health & Ageing, University College London, London, United Kingdom*
[f] *Centre for Functional MRI of the Brain, University of Oxford, Oxford, United Kingdom*
[g] *Department of Neuroimaging, Institute of Psychiatry, Psychology, & Neuroscience,King's College London, London, United Kingdom*
[h] *Department of Psychology, University of Oslo, Oslo, Norway*
[i] *Norwegian Centre for Mental Disorders Research, Division of Mental Health and Addiction, Oslo University Hospital and University of Oslo, Oslo, Norway*
[j] *KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway*

Abstract

Normative modelling is an emerging technique for parsing heterogeneity in clinical cohorts. This can be implemented in practice using hierarchical Bayesian regression, which provides an elegant probabilistic solution to handle site variation in a federated learning framework. However, applications of this method to date have employed a Gaussian assumption, which may be restrictive in some applications. We have extended the hierarchical Bayesian regression framework to flexibly model non-Gaussian data with heteroskdastic skewness and kurtosis. To this end, we employ a flexible distribution from the sinh-arcsinh (SHASH) family, and introduce a novel reparameterisation that is more suitable for Markov chain Monte Carlo sampling than existing variants. Using a large neuroimaging dataset collected at 82 different sites, we show that the results achieved with this extension are better than a warped Bayesian linear regression baseline model on most datasets. We also demonstrate that the attained flexibility is essential for accurately modelling highly nonlinear relationships between aging and imaging derived phenotypes, which shows that the extension is important for pushing the field of normative modelling forward. All methods described here are available in the open-source `pcntoolkit`.

Keywords: Normative modelling, Hierarchical Bayesian Regression, Neuroimaging, Precision Psychiatry

1. Introduction

Brain disorders affect millions of people worldwide, and have a large impact on the quality of life of the patient and the people surrounding them. Correct treatment in the form of medication or therapy can help reduce symptoms, however, studies to understand mechanisms or develop treatments for these disorders are usually performed in a case-control framework, which provides group-level inferences in which homogeneity within clinical cohorts and clinical groups is presumed. These studies aim to detect differences between the clinical and the control (i.e. healthy) groups and are therefore limited

*augustijndeboer@gmail.com

to providing inferences about group averages (e.g. the 'average patient'). However, it has been argued that the individual deviations from the groups' means contain the most important information for stratification [1, 2]. An improved understanding of the heterogeneity within clinical cohorts may ultimately lead to better diagnosis and treatment.

Normative modelling provides a framework for understanding heterogeneity within clinical cohorts [2]. A normative model aims to provide estimates of centiles of variation of a specific set of phenotypes over the full range of the clinical covariates, i.e., across the lifespan, much like growth charts in pediatric medicine. Centiles of variation of imaging derived phenotypes (IDPs) are estimated, and subjects are assigned a point somewhere along the estimated centiles. The assumption is that healthy subjects typically land in areas of higher density than non-healthy subjects. The mean or 'average' subject is thus also a meaningful concept in normative models, but contrary to case-control studies, the deviations from that mean are interpreted as anomaly in the brain structure or function, and are seen as valuable resources for further downstream processing like analysis of underlying mechanism of brain disorders.

In probabilistic modelling, a distinction is made between epistemic and aleatoric uncertainty [3]. The former is due to uncertainty about a parameter or phenomenon, and generally becomes smaller as more data is collected, the latter is uncertainty due to the variation that is naturally present in the measured phenomenon. In normative modelling, we aim to capture and separate these forms of uncertainty to the highest possible degree. Separating all sources of aleatoric and epistemic uncertainty allows adequately controlling for nuisance variation, and guiding further decisions by relevant (biological) variation only.

As in many other fields, data availability in neuroscience has increased dramatically in recent years, and there is no reason to believe that trend will not continue into the future. Medical data, however, is sometimes subject to restrictions involving sharing, due to certain permissions not being granted. There is no guarantee that all data on which a model is to be trained can be centrally located. The question of how to do normative modelling on large, distributed datasets is thus a natural one, and it comes with a number of challenges. Firstly, the method must be able to handle large amounts of data. We have seen that Gaussian processes would be viable candidates for normative modelling, if it were not for their time complexity that is scales cubic with the number of input data points [2]. Second, distributed datasets inevitably come with sources of nuisance variance, since every scan site has its own protocols, every scanner has its own unique noise characteristics, and there may be unknown sources of noise present in specific sites as well. In addition, it is well known that there are sex-related differences in neurobiological phenotypes. HBR can deal with these group effects in data by assuming a shared prior over group-specific model parameters, as we describe in Appendix B. Third, neuroimaging data are complex and provide rich information about the individual from which they were derived. These data therefore need to be treated with care. Permission to share personal data is not always available, and transportation of the data is therefore not always possible. This requires federated learning; a distributed learning approach, that does not require transportation of the data, only of the inferred parameters of model [4, 5]. In this study, we employ and extend the hierarchical Bayesian regression (HBR) approach for normative modelling [6, 7, 5] on non-Gaussian IDPs. The method involves defining a generative model, and inferring the posterior distribution over the model parameters using Markov chain Monte Carlo (MCMC) sampling. By absorbing and summarizing information contained in training data in learned posteriors, HBR supports federated learning; all information required to extend or adapt the model is contained in the model, eliminating the need of data transportation when using or even extending the model.

The current implementation of HBR assumes that the model's residuals follow a Gaussian distribution. However, the variability in many imaging phenotypes cannot be properly described by a Gaussian distribution, as they are skewed, kurtotic, or both. The issue of non-Gaussianity in neurological phenotypes has partially been addressed by Fraza et al. in [8]. This method accounts for non-Guassianity by applying a monotonic warping function to the data [9], which yields a closed form for the marginal likelihood that can be maximised using standard optimisation techniques [10]. Here,

we utilize the same warping function—the sinh-arcsinh transformation, Appendix A—but in a fully Bayesian approach, i.e., in the HBR framework, yielding full posterior distributions for all parameters. Applying the sinh-arcsinh transformation to a standard Gaussian distribution yields the SHASH distribution [11], which is detailed in section 2.1, a flexible distribution with separate parameters roughly corresponding to skew and kurtosis. Here we apply the HBR framework to the a canonical variant of SHASH distribution and a novel reparameterisation that substantially reduces dependency between the parameters, which is crucial for more efficient sampling. We assess its performance on some known problematic imaging phenotypes, and show that the HBR framework with a SHASH likelihood is able to model phenotypes in a way that was previously impossible.

The contributions of this work are; (i) a reparameterization of SHASH distribution that is suited for MCMC sampling; (ii) a thorough analysis of the HBR method with a SHASH likelihood in comparison with a warped-Bayesian linear regression as a baseline method [8]; (iii) an extensive evaluation of these methods on a large multi-site neuroimaging dataset containing estimates of cortical thickness and subcortical volume; (iv) an extension to the existing HBR implementation in the `pcntoolkit`.

The paper is structured as follows: we first provide a theoretical background of HBR, MCMC sampling, and the SHASH distribution in section 2, where the problems of using the SHASH distribution in a sampling context will also become clear. We introduce a reparameterization that addresses these problems in section 2.1.1. A framework for concisely describing common variations of HBR models is introduced in Appendix B. Then we evaluate our approach using several experiments with a large neuroimaging dataset in section 3, where we show the proposed method performs on par or better than existing methods. Then we apply the new HBR models to an even more non-linear dataset [12], where we show a clear advantage over existing methods. A discussion of the advantages and limitations of our approach follows in section 4, and we summarize all our findings in section 5.

2. Methods

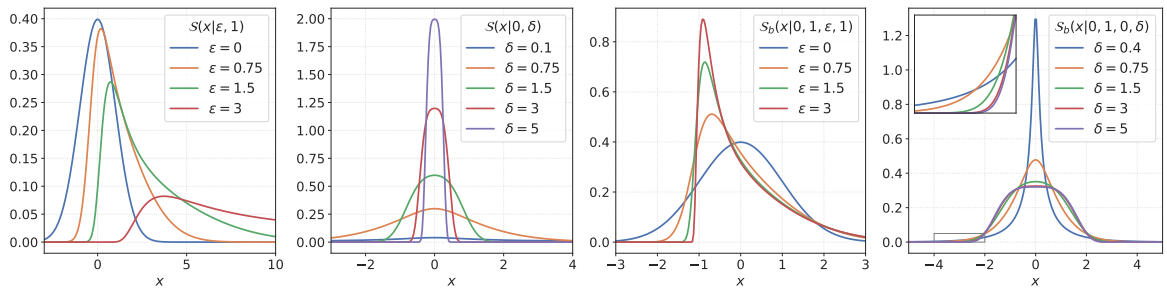
The HBR framework assumes a generative model over the response variables Y in the form of a likelihood $\mathcal{L}$ plus an independent prior distribution p for each parameter θ_i of the likelihood. A general model thus looks like this:

$$\begin{aligned} &\theta_i \sim p_{\theta_i} && i \in \{1, ..., P\} \\ &Y \sim \mathcal{L}(Y|\{\theta_i\}) \end{aligned} \tag{1}$$

where P is the number of parameters of the likelihood. The distribution over the parameters θ_i is approximated by MCMC sampling, which we discuss in Appendix C. Assuming the centiles of variation in the response variable are Gaussian distributed, we substitute the standard Gaussian $\mathcal{N}$ for $\mathcal{L}$, and our parameters θ become the mean and variance μ and σ^2. The general model described in equation 1 is fully capable to support non-Gaussian likelihoods, like $\mathcal{S}$, which is detailed in section 2.1. The way we model p_{θ_i} depends on our further assumptions about the data. Typically μ, and optionally also σ are taken to be linear functions of a set of covariates, which in this example are clinical or demographic variables. This is further detailed in Appendix B, where we also introduce a framework for model specification.

2.1. The SHASH Distribution

To accurately fit normative models on potentially non-Gaussian phenotypes, one approach is to use a family of simple but flexible, non-Gaussian distributions. This family must then contain positively as well as negatively skewed members, leptokurtic as well as platykurtic members, and preferably with smooth transitions between them. All those requirements are fulfilled by the family of SHASH distributions ($\mathcal{S}$). Figures 1a and 1b show the flexibility of the $\mathcal{S}$ in modelling various distributional forms. Samples from a $\mathcal{S}$ are obtained by applying an inverse sinh-arcsinh transformation to samples from a standard Gaussian [11]. The sinh-arcsinh transformation and its inverse are defined as:



(a) **(b)** **(c)** **(d)**

Figure 1: The effects of ϵ and δ on the shape of the SHASH and SHASH$_b$ densities. In (a) and (c), the value of ϵ was varied while δ was kept fixed at 1. In (b) and (d), the value of δ was varied while ϵ was kept fixed at 0. The densities in (c) and (d) have zero mean and unit variance. The inset in (d) corresponds to the area in the gray box in the bottom left corner of (d).

$$S_{\epsilon,\delta}(x) = \sinh(\delta \sinh^{-1}(x) - \epsilon)$$
$$S^{-1}_{\epsilon,\delta}(x) = \sinh((\sinh^{-1}(x) + \epsilon)/\delta)$$

where $\epsilon \in \mathbb{R}$ and $\delta \in \mathbb{R}^+$. So that if we let a random variable Z follow a standard Gaussian, then we can construct samples X from a SHASH distribution by applying S^{-1}:

$$Z \sim \mathcal{N}(0, 1)$$
$$X = S^{-1}_{\epsilon,\delta}(Z)$$

Because the S is a diffeomorphism[1], we can get the density of $\mathcal{S}$ by a change of variables[13]. We multiply the Gaussian density of the inverse transformed samples by the derivative of S^{-1}.

$$\begin{aligned}\mathcal{S}(x \mid \epsilon, \delta) &= \frac{\partial S_{\epsilon,\delta}}{\partial x} \mathcal{N}(S_{\epsilon,\delta} \mid 0, 1) \\ &= \frac{\delta C_{\epsilon,\delta}(x)}{\sqrt{1+x^2}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-S_{\epsilon,\delta}(x)^2}{2}\right),\end{aligned} \tag{2}$$

where $C_{\epsilon,\delta}(x) = \cosh(\delta \sinh^{-1}(x) - \epsilon)$. From Eq. 2 and the fact that $\cosh(x) > 0$ for all $x \in \mathbb{R}$, we get the constraint that $\delta \geq 0$. Fig. 1 illustrates the effect of the parameters ϵ and δ on the shape of the SHASH density. We clearly see that ϵ and δ modulate the skew and kurtosis, respectively.

2.1.1. Reparameterization

The SHASH distribution we have seen so far has two parameters, ϵ and δ, but additional location and scale parameters are necessary, because we need our distributions to be flexibily located and scaled as well. Jones et al. [11] suggest adding scale and location parameters by muliplication with σ and addition of μ, leading to the simple reparameterization known as $\mathcal{S}_o$:

$$\mathcal{S}_o(x \mid \mu, \sigma, \epsilon, \delta) = \mathcal{S}((x - \mu)\sigma^{-1} \mid \epsilon, \delta)\sigma^{-1}. \tag{3}$$

While we believe this family is flexible enough for our purposes, inferring the parameters is highly challenging because of their strong correlations. These are shown in Fig. 1a. Clearly, ϵ not only

[1] A differentiable function with a known and differentiable inverse

increases the skew, but also the mean. Similarly, in Fig. 1b we see that the δ increases the variance as well as the kurtosis. Correlations like this cause problems for inference in HBR, because of its reliance on MCMC sampling. In MCMC sampling, an algorithm traverses a path of pseudo-random points in parameter space, storing visited points as samples. Under the conditions that (i) the Markov chain has the posterior distribution as its stationary distribution, and (ii) detailed balance is satisfied, the samples that are retrieved follow the posterior distribution over model parameters, and can thus be used to approximate distributions. Correlations in the parameter space inhibit the sampler from efficiently exploring the parameter space, resulting in samples that do not follow the posterior very well, or do not converge to the target density at all. Removing the effect of ϵ and δ on the mean and variance respectively will break the correlation for a large part. All that is needed to remove that effect is a precise definition of its magnitude. Luckily, the r'th non-central moment has a known analytical form, given in Eq. 4 [11].

$$m_{\epsilon,\delta}^{(r)} = \mathbb{E}[X_{\epsilon,\delta}^r] = \frac{1}{2^r}\sum_{i=0}^{r}\binom{r}{i}(-1)^i \exp\Big((r-2i)\frac{\epsilon}{\delta}\Big)P((r-2i)/\delta), \tag{4}$$

for $r \in \mathbb{N}$, where

$$P(q) = \frac{e^{1/4}}{(8\pi)^{1/2}}\Big(K_{(q+1)/2}\big(\tfrac{1}{4}\big) + K_{(q-1)/2}\big(\tfrac{1}{4}\big)\Big), \tag{5}$$

and K is the modified Bessel function of the second kind [14]. If we now let

$$\Psi \sim \mathcal{S}_o(\mu, \sigma, \epsilon, \delta),$$

then we have the first and second moment of Ψ simply as:

$$\mathbb{E}[\Psi] = \mu + \mathbb{E}[X] \qquad \text{and} \qquad \mathbb{E}[\Psi^2] = \sigma\mathbb{E}[X^2].$$

which reveals the correlation, also illustrated by Fig. 2. The solution proposed here is a shift and scale of a standardized SHASH. Because the moments of $\mathcal{S}$ are all obtainable by Eq. 4, this is a simple operation:

$$\mathcal{S}_b(x \mid \mu, \sigma, \epsilon, \delta) = \mathcal{S}\big(((xz_{\epsilon,\delta} + m_{\epsilon,\delta}^{(1)}) - \mu)\sigma^{-1} \mid \epsilon, \delta\big)z_{\epsilon,\delta}\sigma^{-1}, \tag{6}$$

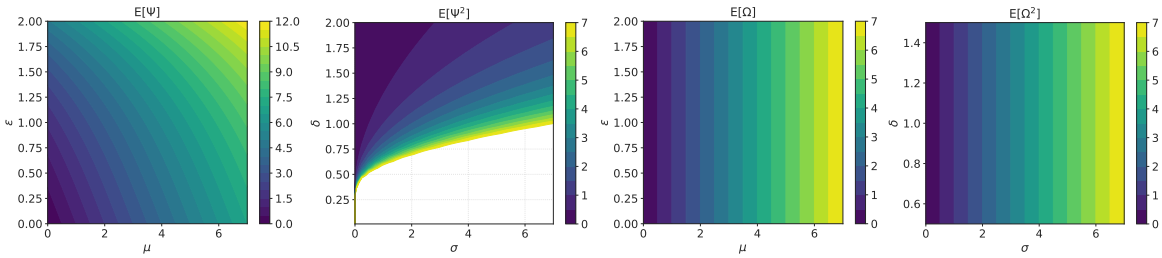
where $z_{\epsilon,\delta}$ is the central standard deviation given by $z_{\epsilon,\delta} = \left(m_{\epsilon,\delta}^{(2)} - \big(m_{\epsilon,\delta}^{(1)}\big)^2\right)^{\frac{1}{2}}$. We let $\Omega \sim \mathcal{S}_b(\mu, \sigma, \epsilon, \delta)$ and find that we now nicely have:

$$\mathbb{E}[\Omega] = \mu \qquad \text{and} \qquad \mathbb{E}[\Omega^2] = \sigma.$$

We define the newly designed transformation as $S_{b\mu\sigma\epsilon\delta}$:

$$S_{b\mu\sigma\epsilon\delta}(x) = (S_{\epsilon\delta}^{-1} - m_{\epsilon\delta}^{(1)})\Big(m_{\epsilon\delta}^{(2)}\Big)^{-\frac{1}{2}}\sigma + \mu. \tag{7}$$

The new parameterization has nicely interpretable parameters for location and scale, which the $\mathcal{S}_o$ does not. Compare figures 1a and 1b with 1c and 1d. The densities in the two rightmost figures all have zero mean and unit variance. This may not seem obvious for the density with $\delta = 0.4$ in Fig. 1d, but the tail behavior responsible for this becomes clearer from the inset. In Appendix G, the empirical moments of the $\mathcal{S}_o$ and $\mathcal{S}_b$ distributions are plotted over a range of parameters. It is important to note that the $\mathcal{S}_b$ family is isomorphic to $\mathcal{S}_o$, which means that the there is a one-on-one mapping from each member of one family to a member of the other. In simpler terms: the two families model exactly the same set of distributions.



(a) **(b)** **(c)** **(d)**

Figure 2: Correlations in the parameter space. In (a), at $\epsilon = 0$, the mean is exactly the value of μ, but when $\epsilon \neq 0$, that is not the case. Similarly, in (b), the variance is exactly σ only on those points where $\delta = 1$. Fig. (c) and (d) show the parameter space under the proposed reparameterization. The correlation between these pairs of variables is now removed.

2.1.2. Priors on $\mathcal{S}_b$ parameters

The $\mathcal{S}_b$ reparameterization greatly reduces the problem of correlation in parameter space, but it comes at a cost. Some of the reparameterization terms get extremely large quickly, and may therefore lead to numerical instability. SThis may be addressed by appropriately setting the priors for the $\mathcal{S}_b$ likelihood. Comparing Fig. 2b and Fig. 2d, the warp in the low domain of δ is apparent. The stretching effect the standardizing change of variables has is enormous in those regions, because it has to compensate for the flattening effect of δ visible in Fig. 1b. The curve in the line associated with $\delta = 0.1$ is barely visible in this range. The variance of $X_{\epsilon,\delta}$ as a function of δ is shown on a double log scale in 3. Notice that small differences in the lower domain of δ amount to substantial differences in the variance. The enormous magnitude of $\frac{d}{d\delta} m^{(2)}_{\epsilon,\delta}$ in those regions is mainly due to the behavior of the Bessel function $K_n(z)$ in Eq. 5, and to a lesser degree due to the fraction in the exponent in Eq. 4. In practical applications values of $\delta < 0.3$ are thus associated with numerical instability, and sampling in those regions should be avoided at the cost of losing some expressivity.

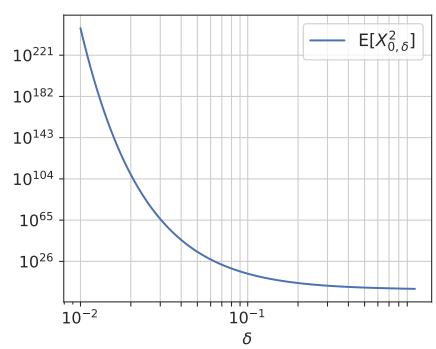


Figure 3: The variance of $X_{\epsilon,\delta}$ for $\delta < 1$

2.2. Experimental Materials

To allow a meaningful comparison with the current methods, we adapt and use the data from Rutherford et al. [15]. This is a large neuroimaging dataset collected from 82 different scanners, containing 58834 control subjects and 2925 patient subjects. Measures of cortical thickness and subcortical volume were extracted from anatomical images using Freesurfer version 6.0. Furthermore, the different UK-biobank datasets [16] were merged, because the datasets are known to be well-controlled for site effects. Counts per site and sex can be found in section Appendix E. For full

details about the data, we refer to Rutherford et al. [15]. Most data in this dataset are publicly available, but for several of the clinical samples, we do not have explicit permission for data sharing. Such permissions for sharing were not obtained from the data from the University of Michigan, and as such these data were not included in this study (n=1394 (1149 controls, 245 patients)).

The family of SHASH distributions has the flexibility to model skewed and kurtotic distributions, but also has the standard Gaussian as a central member. We pick some 'hard' (i.e. difficult) phenotypes, the residuals of which are not described well by a Gaussian, and some 'easy' phenotypes, the residuals of which are described well by a Gaussian. We aim to model both easy and hard phenotypes with the same model, using the same hyperparameters. The selected hard phenotypes are: right cerebellum white matter, because it is slightly skewed; right lateral ventricle, because it is moderately skewed and heteroskedastic, and white matter hypointensities, because it is severely skewed and heteroskedastic. For easy phenotypes, we selected the estimated total intracranial volume, the thickness of the Jensen sulcus, and the volume of the brain stem. These 'easy' phenotypes do not show a lot of skew or kurtosis, but the mean of the thickness of the Jensen sulcus shows an interesting trajectory. The distributions of the different phenotypes are shown in figure 6 and in section Appendix I below. In some instances we used the following abbreviations: Right-Cerebellum-White-Matter: RCWM, Right-Lateral-Ventricle: RLV, WM-hypointensities: WMH, EstimatedTotalIntraCranialVol: ETICV, rh_S_interm_prim-Jensen_thickness: RIPJ, Brain-Stem: BS,

A stratified 10-fold split was made using the Sklearn package [17], to which end we filtered out data from sites with less than 10 subjects of one or either sex. This ensures that the test folds contain at least one unique sample for each batch-effect. All data was standardized as a preprocessing step.

Age and sex were used as covariates, and site was used as a batch-effect. The HBR framework supports multidimensional batch-effects, allowing to use both sex and site as a batch effect. However, there is little reason to assume that the differences related to sex vary much between sites, and having this multidimensional batch effect results in a larger number of model parameters. In practice, we saw no difference in performance between models that used sex as a batch effect, and those that used sex as a covariate. (See section 4 for a discussion.)

As a second, more challenging, test of the flexibility of our model, we fit a normative model on a highly non-linear representation of brain data derived from the latent representation learned by a convolutional autoencoder. Full details are provided in Zabihi et al. [12], but briefly; data derived from the UK Biobank sample (n=20781, 47% female) was compressed to a 2-dimensional latent representation using a pipeline consisting of a convolutional autoencoder with a 100-dimensional latent space [18], followed by a UMAP dimensionality reduction [19]. The resulting representation followed a highly nonlinear distribution with a trajectory in the skew. A stratified train/test (75%/25%) split was made on the raw data level, specifically not at the 100-dimensional latent representation. This means that the train data was used for training the convolutional autoencoder as well as the UMAP embedding, and none of the test data was used. Following the rationale described above, no batch-effects were modeled. Again, we used age and sex as covariates.

2.2.1. Ethics Statement

All participants provided written informed consent for their data to be used, and all studies were approved by the institutional review boards at the respective institutions.

3. Results

In sections 3.1, the convergence of the MCMC samplers utilized by the HBR models is assessed. In section 3.2, four different types of normative models are compared. First, the warped Bayesian Linear Regression [8], which is used as a baseline, abbreviated with W-BLR. Then, three variants of the HBR method, are used. The first uses a simple Gaussian likelihood, where μ and σ are modeled exactly like μ and σ in section Appendix B.3. This model is indicated with an $\mathcal{N}$. Model $\mathcal{N}$ is the model that was previously the limit of what could be done within HBR in the `pcntoolkit`. The other

two HBR models use a $\mathcal{S}_b$ likelihood, and differ only in the way ϵ and δ are modeled. In the first model $\mathcal{S}_{b1}$, both ϵ and δ assume constant values throughout the full range of covariates. In model $\mathcal{S}_{b2}$, both ϵ and δ are linear regressions on Φ. These models are formalized in table 1. Details about the priors can be found in table 2. The models used in section 3.3 are the same as in section 3.2, just applied to different data.

Importantly, the results in section 3.2 are based on MAP estimates of z-scores (see section Appendix D). The MAP estimates were chosen because they are less computationally demanding, and because in practice, the MCMC results were only marginally better. Especially for model $\mathcal{S}_{b2}$, which has many parameters, it may take some time for the MCMC samplers to converge. It is important to note that the HBR method does not currently support federated learning on basis of the MAP estimate. In order to do that, one would need to use MCMC samples, such that those can be used to infer a factorized posterior.

$\mathcal{N}$	$\mathcal{S}_{o1}$	$\mathcal{S}_{b1}$	$\mathcal{S}_{b2}$
$\mu_n \sim \mathcal{M}_{3.1.2b}$	$\mu_n \sim \mathcal{M}_{3.1.2b}$	$\mu_n \sim \mathcal{M}_{3.1.2b}$	$\mu_n \sim \mathcal{M}_{3.1.2b}$
$\sigma_n^{\pm} \sim \mathcal{M}_{3.1.1}$	$\sigma_n^{\pm} \sim \mathcal{M}_{3.1.1}$	$\sigma_n^{\pm} \sim \mathcal{M}_{3.1.1}$	$\sigma_n^{\pm} \sim \mathcal{M}_{3.1.1}$
$\sigma_n = \text{softplus}(\sigma_n^{\pm})$	$\sigma_n = \text{softplus}(\sigma_n^{\pm})$	$\sigma_n = \text{softplus}(\sigma_n^{\pm})$	$\sigma_n = \text{softplus}(\sigma_n^{\pm})$
$y_n \sim \mathcal{N}(\mu_n, \sigma_n)$	$\epsilon \sim \mathcal{M}_1$	$\epsilon \sim \mathcal{M}_1$	$\epsilon \sim \mathcal{M}_{3.1.1}$
	$\delta^{\pm} \sim \mathcal{M}_1$	$\delta^{\pm} \sim \mathcal{M}_1$	$\delta^{\pm} \sim \mathcal{M}_{3.1.1}$
	$\delta = 0.3 + \text{softplus}(\delta^{\pm})$	$\delta = 0.3 + \text{softplus}(\delta^{\pm})$	$\delta = 0.3 + \text{softplus}(\delta^{\pm})$
	$y_n \sim \mathcal{S}_o(\mu_n, \sigma_n, \epsilon, \delta)$	$y_n \sim \mathcal{S}_b(\mu_n, \sigma_n, \epsilon, \delta)$	$y_n \sim \mathcal{S}_b(\mu_n, \sigma_n, \epsilon, \delta)$

Table 1: The generative descriptions of the models that are compared to W-BLR. For all n, $n \in \{1, ..., N\}$.

A main objective in Normative modelling is to map observations to z-scores in such a way that healthy variation within those z-sores follows a Gaussian distribution, and that patients fall in the outer quantiles of variation. Gaussianity in the outer quantiles is compared by the third and fourth moment (skew and kurtosis) of the predicted z-scores. The qq-plots, and a visual inspection of the regression curves further help assess the quantiles.

3.1. Convergence Analysis

Here a simple analysis of the samplers is performed, by analyzing the trajectory of the Gelman-Rubin, or $\hat{R}$ statistic [20], which can be thought of as the ratio between the interchain variance and the intrachain variance. The $\hat{R}$ statistic is 1 (one) when those are equal, and this is a good indicator that the sampler has converged. In practice, a rule of thumb is that an $\hat{R}$ lower than 1.1 is no reason to worry, and an $\hat{R}$ less than 1.05 indicates good convergence. To compute the $\hat{R}$ statistic, at least two chains need to be sampled. We sampled two chains of 1500 samples using the `pcntoolkit`, the first 500 of which were used for tuning and were discarded. The last 1000 samples of all chains were stored and used for further analysis.

In figure 4, the $\hat{R}$ statistics of ϵ (unbroken line) and δ (broken line) are displayed as a function of chain length. The results are averaged over the 10 folds. The clear differences between the $\mathcal{S}_{o1}$ and the $\mathcal{S}_{b1}$ models in some phenotypes are indicative of the improvement that the $\mathcal{S}_b$ reparameterization offers over $\mathcal{S}_{o1}$. Except for the `WM-hypointensities` phenotype - which is notoriously hard to model, the $\mathcal{S}_{b1}$ $\hat{R}$ converges very close to 1 within at most 300 iterations. The $\mathcal{S}_{o1}$ $\hat{R}$ lines show worse convergence, even on an easy phenotype like the `EstimatedTotalIntraCranialVol`, where $\hat{R}$ stays close to 1.2 for the entire chain length. Another plot, in which the $\hat{R}$ statistics of some parameters for $\mathcal{N}$, $\mathcal{S}_{o1}$ and $\mathcal{S}_{b1}$ are compared, is provided in section Appendix F. Overall, the convergence on the $\mathcal{S}_{b1}$ model looks better than on the $\mathcal{S}_{o1}$ model.

Parameter	Prior	$\mathcal{N}$	$\mathcal{S}_{o1}$	$\mathcal{S}_{b1}$	$\mathcal{S}_{b2}$
$\mathbf{w}_\mu$	$\mathcal{N}(\mathbf{w}_\mu \mid \mathbf{0}_D, 3\mathbf{I}_D)$	X	X	X	X
μ_{τ_μ}	$\mathcal{N}(\mu_{\tau_\mu} \mid 0, 1)$	X	X	X	X
σ_{τ_μ}	$\text{Cauchy}^+(\sigma_{\tau_\mu} \mid 1)$	X	X	X	X
ν_{τ_μ}	$\mathcal{N}(\nu_{\tau_\mu} \mid 0, 1)$	X	X	X	X
$\mathbf{w}_\sigma$	$\mathcal{N}(\mu_{\mathbf{w}_\sigma} \mid \mathbf{0}_D, \mathbf{I}_D)$	X	X	X	X
$\tau_\sigma^\pm$	$\mathcal{N}(\tau_\sigma^\pm \mid 1, 1)$	X	X	X	X
ϵ	$\mathcal{N}(\epsilon \mid 0, 1)$		X	X	
$\mathbf{w}_\epsilon$	$\mathcal{N}(\mathbf{w}_\epsilon \mid \mathbf{0}_D, \mathbf{I}_D)$				X
τ_ϵ	$\mathcal{N}(\tau_\epsilon \mid 0, 1)$				X
$\delta^\pm$	$\mathcal{N}(\delta^\pm \mid 1.5, 2)$		X	X	
$\mathbf{w}_\delta$	$\mathcal{N}(\mathbf{w}_\delta \mid \mathbf{0}_D, \mathbf{I}_D)$				X
τ_δ	$\mathcal{N}(\tau_\delta \mid 2, 1)$				X

Table 2: Prior settings for all HBR models used in this report. Note that parameters with a $\pm$ superscript are mapped by a softplus function post-sampling to force positivity, as displayed in table 1. The single Cauchy$^+$ distribution is the positive Cauchy distribution. $\mathbf{0}_D$ and $\mathbf{I}_D$ denote a D dimensional zero-vector and identity matrix, respectively. Central values were chosen for most priors, with δ being a notable exception. As discussed before, having a low δ leads to numerical issues, and by setting the priors up like this, we aim to push the sampler away from the low domain. Remember that we also add a constant of 0.3 to the result of applying the softplus function to $\delta^\pm$. The prior on τ_σ is centered at 1, because due to standardization we expect σ to lay close to 1 on average, and the softplus approximates an identity function for positive values, so the mode of the transformed value should still lay close to 1. We set the prior on the weights of μ a little wider than the others, because we found that improved the regression. Making the priors any wider than this did not lead to any significantly different results.

These results show that the $\mathcal{S}_b$ reparameterization is beneficial for sampling, and we will for the remainder of this report leave $\mathcal{S}_o$ out of the comparison since the inferences cannot be considered reliable if the sampler has not converged. Results acquired using a $\mathcal{S}_o$ likelihood are unreliable, and as discussed in section 2.1, the family of $\mathcal{S}_b$ is isomorphic to $\mathcal{S}_o$, meaning they can retrieve the same results under the correct priors. We have shown here that $\mathcal{S}_b$ is just a better way to get them.

3.2. Goodness-of-fit Analysis

First, we discuss how the HBR with the $\mathcal{S}_b$ likelihoods compare to the W-BLR method, and the HBR with the Gaussian likelihood. We assess (i) how well the predicted z-scores follow a Gaussian distribution, by looking at the higher-order moments of the z-scores and at qq-plots; (ii) how good the fit of the quantiles is visually, by superimposing them on the data, and (iii), how well the models have captured the batch-effects, by looking at classification performance. Specifically: for any combination of 2 sites, we computed the area-under-the-curve (AUC) metric [21] by using the z-scores as classification thresholds, and the batch-indices as labels.

Table 3 forms the first basis of our discussion. For convenience, it is useful to juxtapose figure 6 or appendix section Appendix I, which visualize the phenotypes discussed here. The skew and kurtosis are the third and fourth standardized moment respectively [22]. As the name suggests, skew describes how much the data is skewed about the mean. The kurtosis measures the tailed-ness of the distribution. A spike distribution has a kurtosis of 0, and a standard Gaussian has a kurtosis of 3. For this reason, the kurtosis values reported here are reduced by 3, which is sometimes named the *excess kurtosis*. For both measures, the ideal result is thus 0, which suggests a symmetric distribution, with tails as slim as those of a Gaussian.

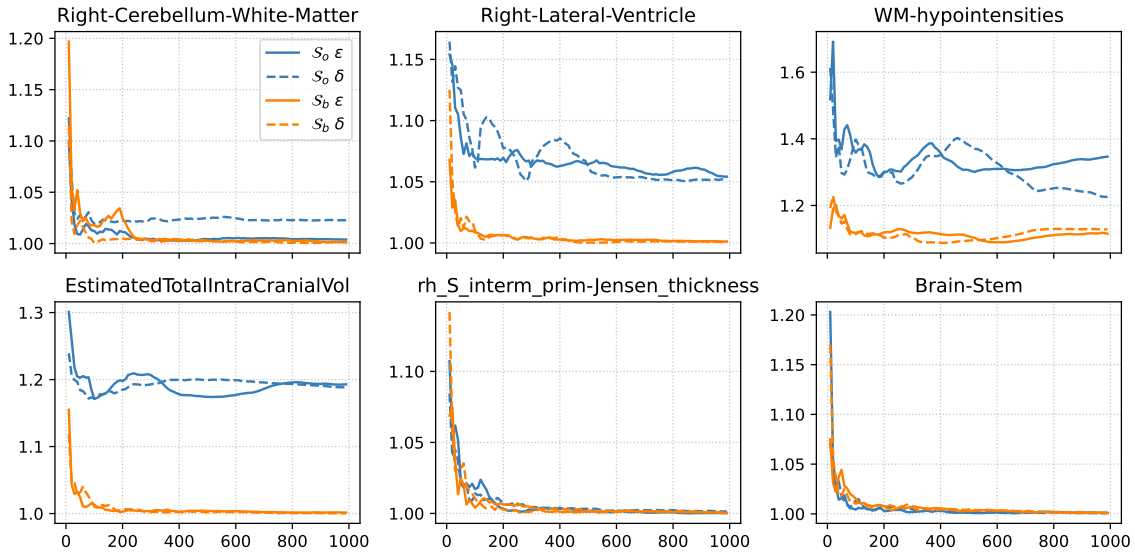


Figure 4: $\hat{R}$ of ϵ and δ for $\mathcal{S}_{o1}$ and $\mathcal{S}_{b1}$, average of 10 folds. The orange line, corresponding to the $\mathcal{S}_{b1}$ model, is typically better.

	Skew				**Kurtosis**			
	W-BLR	HBR			W-BLR	HBR		
Response variable		$\mathcal{N}$	$\mathcal{S}_{b1}$	$\mathcal{S}_{b2}$		$\mathcal{N}$	$\mathcal{S}_{b1}$	$\mathcal{S}_{b2}$
`Right-Cerebellum-White-Matter`	**0.06**	1.33	0.14	0.14	0.57	6.04	0.5	**0.5**
`Right-Lateral-Ventricle`	0.32	2.75	0.16	**0.13**	0.89	26.69	1.08	**0.58**
`WM-hypointensities`	0.34	6.38	0.33	**0.33**	**0.5**	79.82	1.52	1.39
`EstimatedTotalIntraCranialVol`	-0.04	0.04	-0.01	**-0.01**	0.71	1.33	0.32	**0.29**
`rh_S_interm_prim-Jensen_thickness`	0.03	0.29	0.0	**0.0**	0.3	1.17	0.12	**0.12**
`Brain-Stem`	**0.0**	0.31	0.02	0.03	0.52	0.91	0.25	**0.24**

Table 3: Moments of z-scores, average of 10 folds. Smaller is better.

The best results—results that are closest to 0—are printed in bold font. For most phenotypes, $\mathcal{S}_{b2}$ has the best result, but the difference with other methods is sometimes small. The two moments need to be weighted differently as well. For instance, for the `Brain-Stem` phenotype, the W-BLR has only a marginally smaller skew than $\mathcal{S}_{b2}$, but $\mathcal{S}_{b2}$ has a significantly smaller kurtosis. It is important to remember that the kurtosis is a sum of fourth powers, and the skew is a sum of third powers. Deviations thus generally have a larger effect on the kurtosis than on the skew. For the `WM-hypointensities` phenotype, the W-BLR method performs on par with the $\mathcal{S}_{b1}$ and $\mathcal{S}_{b2}$ models regarding the skew, and better when regarding the kurtosis, indicating a better fit on those phenotypes that are also most difficult to model. The W-BLR has inherent batch-effects in all aspects of the predicted quantiles, possibly explaining why it fits better to these hard phenotypes. This is further discussed in section 3.2.1. We emphasise that the comparison is not completely straightforward with the HBR models being more conservative. Despite that, the numbers indicate a fit that is overall better than W-BLR.

Figure 5 further illustrates the improvement that $\mathcal{S}_b$ model offers over the $\mathcal{N}$ model. The three hard phenotypes are not modeled well at all by the $\mathcal{N}$ model. The $\mathcal{S}_b$ models perform comparable to W-BLR, except for `WM-hypointensities`, where the W-BLR line looks slightly better.

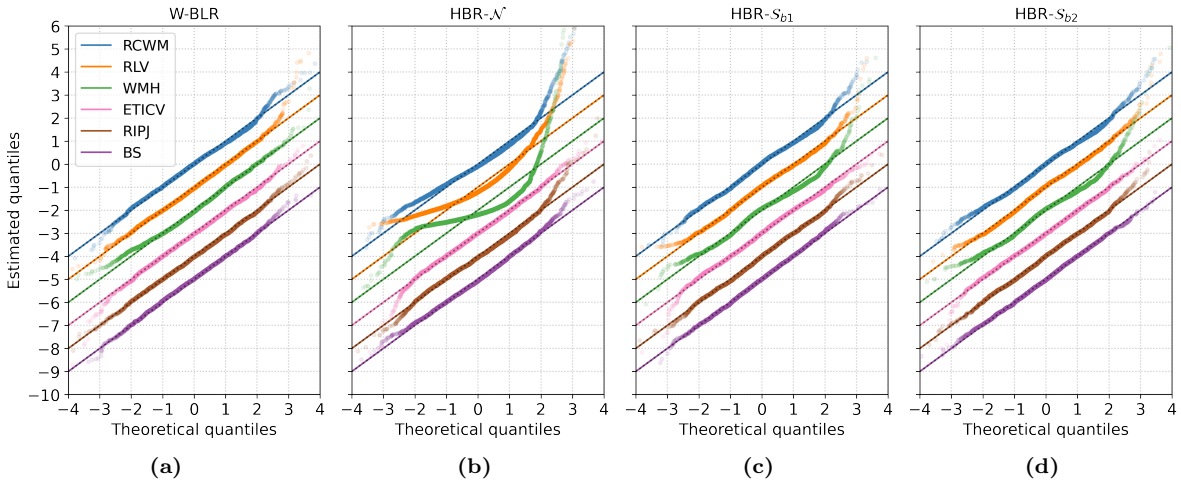


Figure 5: Qqplots of the estimated z-scores of the first test fold. Estimated quantiles of subsequent phenotypes are offset by -1 for easy comparison. For the HBR models (b), (c), (d), the z-scores are computed by MAP estimation. (a) shows very straight lines, almost perfectly covering the model line. (b) shows the fit of the HBR model with Gaussian likelihood. Those phenotypes of which the residuals follow a Gaussian distribution are modeled quite well, but the limitations become very apparent when looking at the top three qqplots. The lines are very curved, and do not reflect a good fit at all. (c) and (d) show a much improved fit over (b), with the plotted points laying mostly on top of the model line. Subtle differences can be found between (c) and (d), mostly in the outer centiles, where $|\texttt{theoretical quantiles}| > 2$. Abbreviations in the legend are specified in section 2.2.

Figure 6 visualizes the predicted centiles of the HBR models on two hard phenotypes; `Right-Lateral-Ventricle` and `WM-hypointensities`. The lines mark the median (thick line) and the first three negative and positive standard deviations in the Gaussian space, acquired by transforming the values $\{-3, ..., 3\}$ by equation 7.

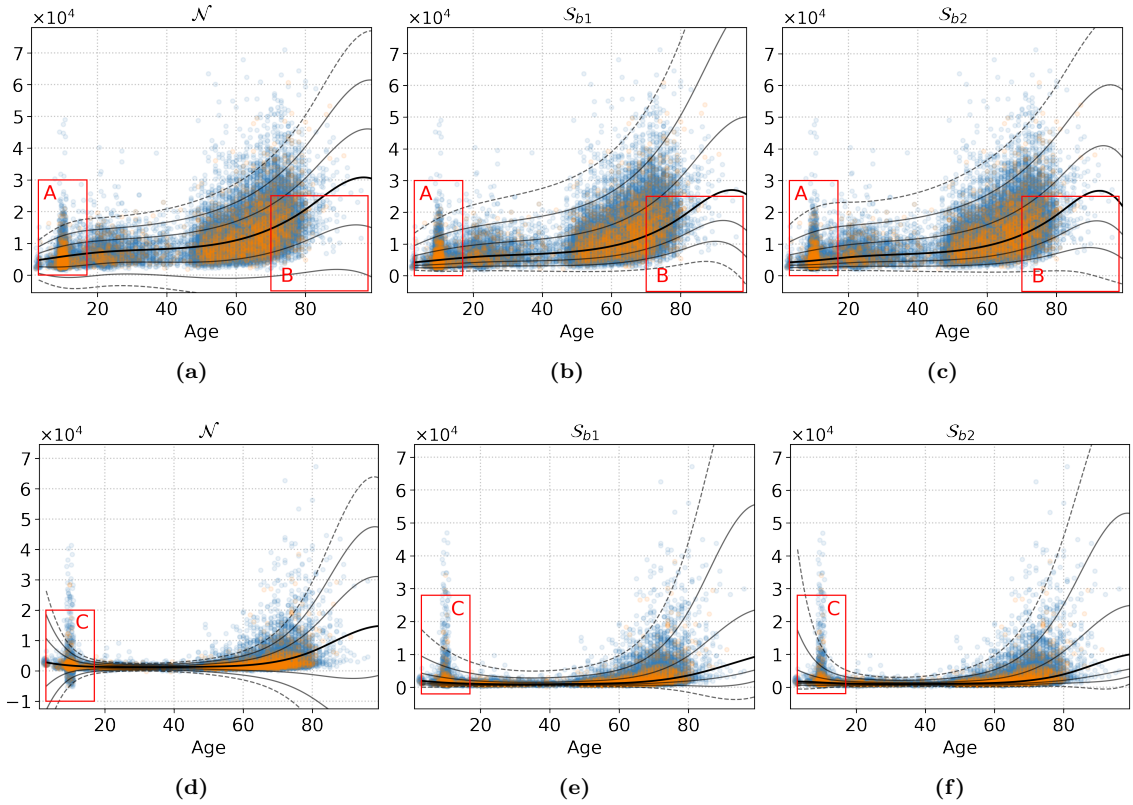


Figure 6: MAP-estimated quantiles of `Right-Lateral-Ventricle` (top row), and `WM-hypointensities` (bottom row), by three different HBR models (columns). The blue markers indicate train data, the orange dots indicate the test data. The plotted data was corrected by the offset τ_μ, learned by the corresponding model (see section Appendix B for details). Focusing on area A, the third quantile predicted by the leftmost model (a) does not extend far enough to capture a part of the upward extending variance. The prediction by model $\mathcal{S}_{b1}$ (b) is better, but a lot of data still falls outside the negative third quantile. Plot (c) shows that the more complex model fits the data a bit better, because it allows the distribution to be a little more skewed at that point. Focusing on are B, and realizing that this phenotype is a volume, and can therefore never be negative, the problem is obvious. For all models, the lowest quantile is estimated to run into the negative domain. This necessarily means that the quantiles are not modeled accurately in those regions. For the leftmost model, this effect is the most detrimental, as it happens over the entire age range. For the middle model, most of the negative third quantile is positive, but in area B it makes a sharp drop. In the right model it also drops, but the drop is less abrupt. For both effects, similar results can be observed in the bottom row. Another effect is visible in area C. On the leftmost plot, we see a lot of variance extending into the negative domain. Because it is corrected with the offset learned by the corresponding model, the data should visually follow the same curve and all lay above 0. The data does not follow the same curve, and some of it is below zero, so those offsets are not reliable or correct. This effect is not present in the other models, so the offsets learned by those models are better. These visualizations indicate that models with the $\mathcal{S}_b$ likelihood predict quantiles better and remove batch-effects better than models with a Gaussian likelihood.

Another important aspect of the HBR method is the way it can deal with batch effects. Unlike the W-BLR method, where one-hot encoded batch vectors are appended to the design matrix (effectively treating site as a fixed effect), causing the batch effects to directly or indirectly affect the mean, variance, skew, and kurtosis, in HBR the batch effects only affect aspects of the likelihood if so intended. By having a batch-specific offset in μ, sampled as a deviation from a group mean, the HBR models' predictions are ultimately indistinguishable between batches. To assess to which degree the predictions are indistinguishable between batches, we compare the Area Under the Reciever-Operator-Curve (AUC) scores [21], which vary between 0 and 1, in a 1v1 setting. Only site was used as a batch effect in our models, so the batch-effects correspond exactly to the site-ids. For each combination of

two sites, we compute the AUC score using the `sklearn.metrics.roc_auc_score` method, where we use the binary class labels as `y_true`, and the z-score predictions for `y_pred`. Because an AUC score of 0.5 would indicate perfect indistinguishibility, we plot the absolute deviation from the 0.5 mark in figure 7.

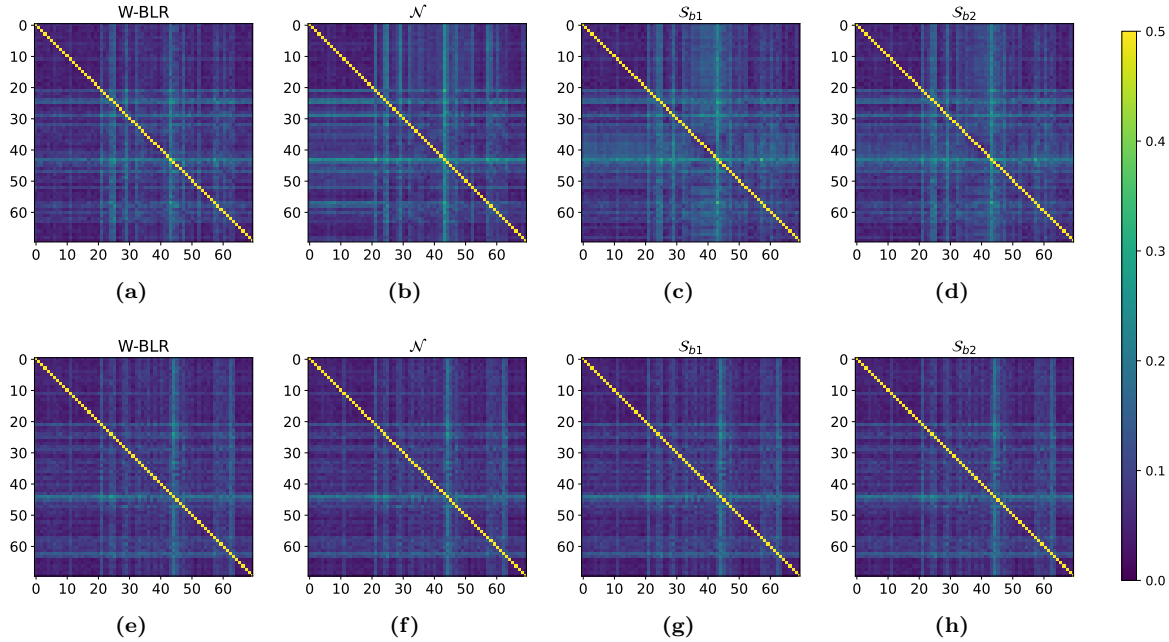


Figure 7: AUC scores of 1v1 site classification based on z-scores, averaged over 10 folds. The color of the x-y'th pixel represents $|a - 0.5|$, where a is the AUC of the binary classification problem of site x and y. For off-diagonal pixels, values closer to 0 are better. The top row is on z-scores of the `WM-hypointensities` phenotype, the bottom row on `rh_s_interm_prim-Jensen_thickness`. For all models, the off-diagonal terms are somewhat homogeneously close to 0. W-BLR on `WM-hypointensities` (a) seems to perform a bit better than (c) and (d), with a less distinct stripe pattern. See section 3.2.1 for a possible explanation. All other results are listed in the appendix section Appendix J, but most look similar to the plots in the bottom row; very little difference between the models is observed.

3.2.1. A Subtle Difference Between HBR and W-BLR

The comparison between HBR and BLR seems to indicate the BLR should be preferred in some cases where the modeled response variable takes on an extreme shape. However, the W-BLR method and the HBR differ in the order in which they apply certain transformations, leading to a subtle difference.

The BLR method applies the sinh-arcsinh transformation after applying the affine transformation for mean and variance, and the HBR method applies the sinh-arcsinh transformation before applying the affine transformation for μ and σ. Because μ and σ potentially contain batch-effects, and because in W-BLR, μ influences the shape of the distribution (as illustrated in Fig. 8), the random-effects in BLR indirectly influence the skew and the kurtosis. The better fit of BLR on the `Right-Lateral-Ventricle` and the `WM-hypointensities` data in table 3 may partly be due to this flexibility. In most phenotypes, batch effects are not expected in the variance, skew and kurtosis as much as they are in the mean. Phenotypes derived from images from a single specific scanner may be overestimated by some constant amount, but it is less likely that the shape of the trajectory of the feature is dependent on the scan location itself. Nonetheless, a latent correlation between the site and the measurement could be present. In contrast, the HBR method provides better control because it allows modelling these features optionally, but not necessarily (see section Appendix B). With BLR,

	mean	variance	skew	kurtosis
$\mathcal{S}_o(\mu, \sigma, \epsilon, \delta)$	-0.24	2.4	-0.96	0.9
$\mathcal{S}_o(\mu + 0.5, \sigma, \epsilon, \delta)$	0.26	2.4	-0.96	0.9
$S^{-1}_{\epsilon,\delta}(\mathcal{N}(\mu, \sigma))$	-0.05	1.27	-1.34	2.76
$S^{-1}_{\epsilon,\delta}(\mathcal{N}(\mu + 0.5, \sigma))$	0.37	0.91	-1.31	3.37

Table 4: Moments of the samples used in figure 8. The first two rows differ only in the mean, by the exact amount by which μ was increased, whereas the bottom two rows differ entirely. This illustrates that in W-BLR, a change in μ changes the entire shape of the distribution.

the flexibility is inherent to the method, and can not be turned off. The results in table 3 should be seen in light of this subtle difference. The BLR model may outperform HBR on some phenotypes, but the results listed under HBR are made by more conservative models, which can be controlled better.

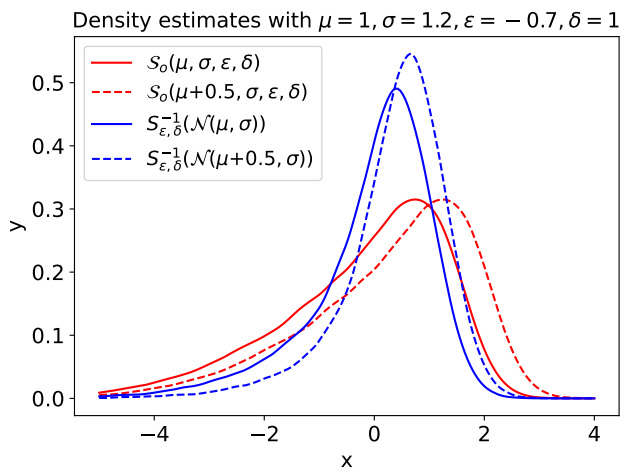


Figure 8: Density estimates made with 100000 samples from a standard Gaussian. For the red lines, the mean and variance (μ and σ) were added after applying the sinh-arcsinh transformation to the samples. Thus, the red lines represent a SHASH$_o$ distribution. For the blue lines, the mean and variance were added to the Gaussian samples before applying the sinh-arcsinh transformation. The broken lines represent the result on exactly the same data, but with a slightly larger μ. The difference between the two red lines is limited to a shift to the right, just like one would expect when increasing μ. The difference between the blue lines is more than just a shift to the right, The broken blue line has a larger mode, and the general shape of the line has also changed slightly. Table 4 further illustrates this. BLR uses the transformation used here to create the blue line, and hence the shape of the distribution can be controlled indirectly by Φ (the design matrix, containing a basis expansion of covariates, along with the one-hot encoded batch labels in the case of BLR) through μ and σ, causing it to fit better to some datasets.

For instance, the ABCD dataset constitutes a large portion of the data, but all observations fall into a small age range. The ABCD samples therefore necessarily have a larger variance than sites with fewer samples. For those small sites, the W-BLR can predict a smaller variance, subsequently affecting the skew and kurtosis on a site-level, constituting to a better fit. The HBR models used here do not necessarily have that flexibility in the variance, skew or kurtosis. The HBR models in this report were designed to only contain batch-effects in the mean, but more flexibility is possible.

3.3. Fitting Highly Nonlinear Data With $\mathcal{S}_b$

To demonstrate the importance of the flexibility of model $\mathcal{S}_{b2}$ likelihood, we take as an example a recent study in which problematic data emerged from a dimensionality reduction.

In [12], fMRI data was compressed using a autoencoder [18], and then further compressed using UMAP dimensionality reduction [19]. The result was a 2d latent representation of brain activation patterns, which followed a somewhat problematic, highly nonlinear, distribution. Postively skewed on the lower end, and negatively skewed on the higher end of the age range, this representation can not be modeled by HBR with a Gaussian likelihood, and W-BLR is also not designed to deal with these kinds of effects. If we take $\mathcal{S}_{b1}$, we find that we run into the limits of that model as well, because it assumes a fixed skew and kurtosis throughout the age range. With model $\mathcal{S}_{b2}$ we have the flexbility to model heterogeneous skew and kurtosis, and Fig. 9 shows that this flexibility is absolutely essential for modelling this phenotype.

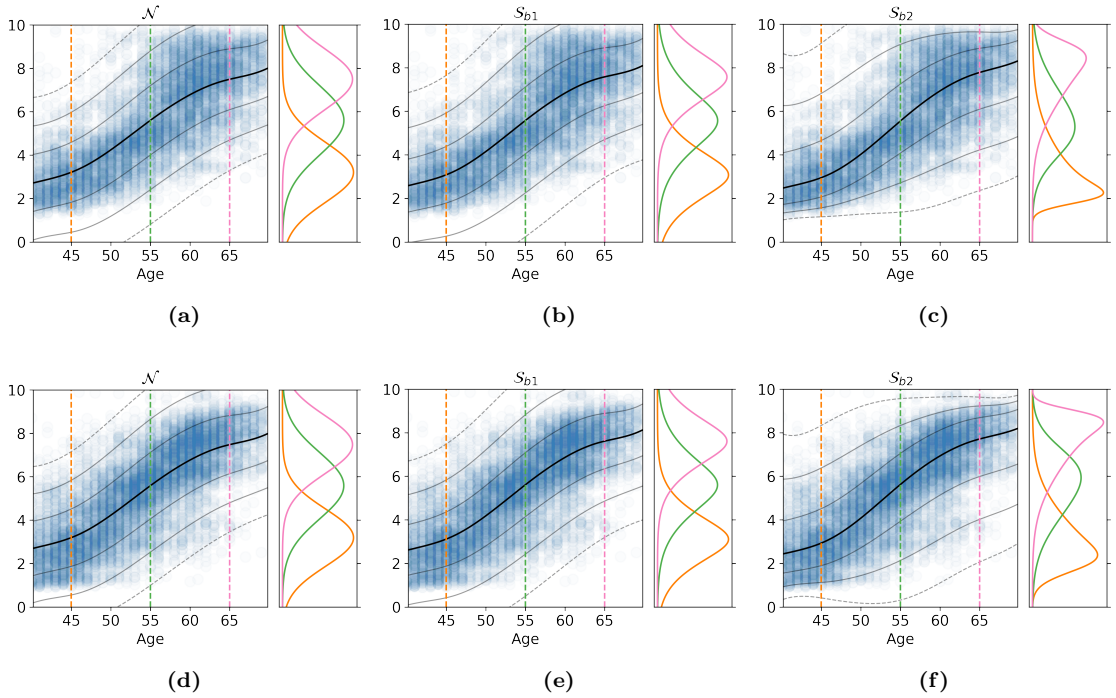


Figure 9: More demanding data from Zabihi et al. [12], modeled by three different HBR models. The top row is female data, the bottom row is male data. The orange, yellow, and green densities plotted on the right panels are the conditional distributions at ages 45, 55, and 65. We see that although the middle models can model for skewness and kurtosis, there is barely any difference from the Gaussian likelihoods on the left. Because the data contains roughly equal amounts of positive skew and negative skew, the best bet for a point estimate is right in the middle, at zero skew. This is reflected in the conditionals, which reflect an almost perfectly Gaussian posterior. For the models on the right, which have more flexibility in the skew and kurtosis, the conditionals change shape as they move throughout the age spectrum, truely following the data distribution.

4. Discussion

In this work, we proposed a flexible method for modelling non-Gaussianity in normative models based on the SHASH distribution. We proposed a novel reparameterisation that enabled us to develop an MCMC sampling approach for these models. We first applied this method to a large neuroimaging dataset and show competitive performance with respect to competing methods, whilst providing full Bayesian inference and full distributions over all model parameters. Then, we applied this method to a highly challenging problem where we aim to model a set of learned parameters from a highly non-linear model (derived from an autoencoder). In this case the classical methods are not able to satisfactorily model the distribution of the brain phenotype, while the proposed method can.

Normative modelling aims to find centiles of healthy variation in neuroimaging phenotypes, much like growth charts in pediatric medicine. Many of those phenotypes follow complex nonlinear, asymmetric distributions, so for a normative model to be generally applicable, it has to be able to model features like skewness and kurtosis. Our extension builds upon the HBR framework for Normative modelling introduced by Kia et al. [5], which retrieves a distribution over model parameters given a prior distribution over the parameters and a likelihood over the data. The original paper reported only on results where a Gaussian likelihood was used, but the framework is flexible enough to support any parameterized likelihood. Here we adapted the HBR framework to make use of a different one, namely the SHASH likelihood, which has additional parameters roughly modelling for skew and kurtosis.

The basic SHASH distribution 2 has no parameters for location and scale, but those can be added resulting in the SHASH_o ($\mathcal{S}_o$) distribution, as in equation 3. Unfortunately, that reparameterization introduces severe correlations in parameter space, which make the distribution unsuited for sampling, an essential component of HBR. We introduce a reparameterization called the SHASH_b ($\mathcal{S}_b$) distribution, which aims to remove the severe correlations by applying an affine tranformation for location and scale to a standardized SHASH distribution.

Our results on a large neuroimaging dataset show that the $\mathcal{S}_b$ reparameterization indeed converges better and more reliably than the SHASH_o distribution. Since the $\mathcal{S}_b$ distribution is isomorphic to $\mathcal{S}_o$—the two can model exactly the same set of distributions—there is no argument for prefering $\mathcal{S}_o$ over $\mathcal{S}_b$ in the context of HBR.

We showed that HBR with a $\mathcal{S}_b$ likelihood performs better than Warped Bayesian Linear Regression (W-BLR) [8] on most datasets, in terms of Gaussianity of the predicted deviation scores. W-BLR predictions are better on a specific pathological phenotypes. We argue that the W-BLR method fits better to that data due to some inherent flexibility, which is not always desired. The W-BLR method has no option to disable this flexibility, as it is inherent to the method. This is why we argue that the HBR method should be preferred as it supports enabling or disabling that flexibility as a modelling choice.

We have shown that HBR with $\mathcal{S}_b$ likelihood is able to model highly non-linear distributions with intricate trajectories in several parameters in a way that is impossible to replicate with existing techniques. This indicates that HBR will possibly play an important role in pushing the field of Normative modelling forward.

The computational complexity of the likelihood may be the most important limitation. Computing the derivatives of sinh, arcsinh, and the Bessel function for every sample is quite expensive, so much so that we resorted to using the MAP in many of our analyses. It must be noted that the computation time is significantly reduced if fewer batch-effects are present. We hope that future releases of software packages like `pymc` will help reduce the required computational time.

Lastly, all our work is freely available to be used and extended, in the open-source `pcntoolkit` package.

4.1. Future Work

4.1.1. Transfer Learning

In the current work, we have not applied the method in a transfer learning setting, but the methods for doing so are described in [5]. For transfering a learned model to a new site, one would have to create a factorized approximation of the posterior by using the MCMC samples retrieved earlier. Restricting ourselves to a known distributional form, this can easily be done by standard optimization techniques on the log likelihood. The factorized posterior is then fixed, and can be used as an informed prior for MCMC sampling with data from a new site. The samples retrieved in this manner can then be used for approximation of z-scores or any other downstream analysis.

4.1.2. Variational Inference

We have shown the value of the $\mathcal{S}_b$ likelihood in the context of MCMC sampling. The reparameterization removes much of the correlation in the posterior that is problematic for sampling methods,

leading to more stable results. Nonetheless, the MCMC sampling procedure is still very computationally demanding, especially on large datasets with many batch-effects. The computation time needed on the 10-fold cross validation of our dataset led us to fall back to a MAP estimate instead of the more principled MCMC estimate. With the $\mathcal{S}_b$ reparameterization in our tool belt, a variational approach may now be a viable option. Variational inference [23] aims to approximate a latent distribution with a known, simpler family of distributions. The distribution used for the approximation (the *inference model*) is fitted by performing a gradient ascent the evidence lower bound (ELBO), which is, as the name suggests, a lower bound on the evidence. Maximizing the ELBO is therefore equivalent to minimizing the Kullback-Leibler (KL) divergence, which expresses how dissimilar two distributions are. Variational inference is a general method; any distribution of which the gradient with respect to the model parameters can be computed can be used as an inference model. The accuracy of the results will only be limited by the degree to which the inference model is able to fit the true distribution, so having a true distribution that behaves somewhat nicely is beneficial. The $\mathcal{S}_b$ reparameterization has removed dependencies between the parameters, such that the posterior can be approximated by a factorized distribution. Variational inference is a fast approximation, and HBR is a slow but (asymptotically) exact method. We believe that in some cases a fast approximation suffices, so variational approximation could be a suitable extension to the present work.

4.1.3. Generalisations to more expressive noise distributions

The extension presented here is just the tip of the iceberg of Hierarchical Bayesian modelling. This extension was developed for modelling skewness and kurtosis, and while we showed that this can fit a wide range of distributions, the flexibility of the SHASH distribution is still limited. Specifically, the limitations in tail thickness are easily read from plot 1d. Mixture models could further help increase the expressive power of the HBR method. One could design a mixture of a single SHASH_b and two Gamma distributions, one left-tailed and one right-tailed, conveniently located as to act as surrogate tails. The central $\mathcal{S}$ component would be dedicated to learning the shape of the bulk of the data, while the Gamma distributions absorb the largest outliers. With the current version of the `pcntoolkit`, this extension is now within reach, but we leave this for future work.

4.1.4. Batch Effects

Multidimensional batch effects are currently implemented as a D dimensional matrix in the `pcntoolkit`, containing one value per combination of batch effects. The number of values grows exponentially with the number of batch effects. An alternative approach is to model a separate vector for each batch-effect, and to learn parameters linking the vectors together. So instead of having $\theta_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma)$, one could have $\theta_{i,j} \sim \mathcal{N}(\alpha\theta_i^1 + \beta\theta_j^2 + \gamma\theta_i^1\theta_j^2)$. The number of required linking parameters grows as $n(n+1)/2 \in \mathcal{O}(n^2)$, much better than the current $\mathcal{O}(d^n)$, where d is the number of unique values for a batch effect. We expect that the correlation between batch-effects is not sufficient to justify the approach, and we believe the most important effects in the batches can be modeled this way.

5. Conclusion

We have extended a method for federated normative modelling of neuroimaging data to support non-Gaussian attributes such as skew and kurtosis. To this end, we have introduced a 4-parameter reparameterization of the SHASH distribution. We have demonstrated that our method outperforms a baseline method on simple and complex data, thereby showing that this method can push the field of normative modelling forward.

References

[1] A. F. Marquand, T. Wolfers, M. Mennes, J. Buitelaar, C. F. Beckmann, Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders, Biological

Psychiatry: Cognitive Neuroscience and Neuroimaging 1 (2016) 433–447. doi:10.1016/j.bpsc.2016.04.002.

[2] A. F. Marquand, I. Rezek, J. Buitelaar, C. F. Beckmann, Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies, Biological Psychiatry 80 (2016) 552–561. doi:10.1016/j.biopsych.2015.12.023.

[3] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Advances in neural information processing systems 30 (2017).

[4] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Processing Magazine 37 (3) (2020) 50–60.

[5] S. M. Kia, H. Huijsdens, S. Rutherford, R. Dinga, T. Wolfers, M. Mennes, O. A. Andreassen, L. T. Westlye, C. F. Beckmann, A. F. Marquand, Federated multi-site normative modeling using hierarchical bayesian regression, bioRxiv (2021). doi:10.1101/2021.05.28.446120.

[6] A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences, Statistical Science 7 (1992) 457–472.

[7] S. M. Kia, H. Huijsdens, R. Dinga, T. Wolfers, M. Mennes, O. A. Andreassen, L. T. Westlye, C. F. Beckmann, A. F. Marquand, Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data, International Conference on Medical Image Computing and Computer-Assisted Intervention (2020) 699–709.
URL http://arxiv.org/abs/2005.12055

[8] C. J. Fraza, R. Dinga, C. F. Beckmann, A. F. Marquand, Warped bayesian linear regression for normative modelling of big data, Neuroimage 245 (2021). doi:10.1101/2021.04.05.438429.
URL https://doi.org/10.1101/2021.04.05.438429

[9] E. Snelson, C. E. Rasmussen, Z. Ghahramani, Warped gaussian processes, Advances in neural information processing systems 16 (2003).

[10] R. Dinga, C. J. Fraza, J. M. M. Bayer, S. M. Kia, C. F. Beckmann, A. F. Marquand, Normative modeling of neuroimaging data using generalized additive models of location scale and shape, bioRxiv (2021). doi:10.1101/2021.06.14.448106.
URL https://doi.org/10.1101/2021.06.14.448106

[11] M. C. Jones, A. Pewsey, Sinh-arcsinh distributions, Biometrika 96 (2009) 761–780. doi:10.1093/biomet/asp053.

[12] M. Zabihi, S. M. Kia, T. Wolfers, S. de Boer, C. Fraza, R. Dinga, A. L. Arenas, D. Bzdok, C. F. Beckmann, A. Marquand, Explanatory latent representation of heterogeneous spatial maps of task-fmri in large-scale datasets, bioRxiv (2022). doi:10.1101/2021.03.10.434856.

[13] C. M. Bishop, Pattern Recognition and Machine Learning, 4th Edition, Vol. 4, Springer, 2006.
URL https://link.springer.com/book/9780387310732papers2://publication/uuid/05A8B4CF-0248-4692-8B1D-DCC065B79465

[14] F. Bowman, Introduction to Bessel Functions, Courier Corporation, 1939.

[15] S. Rutherford, C. Fraza, R. Dinga, S. M. Kia, T. Wolfers, M. Zabihi, P. Berthet, A. Worker, D. Andrews, L. Han, J. Bayer, P. Dazzan, P. McGuire, R. T. Mocking, A. Schene, B. W. Pennix, C. Sripada, I. F. Tso, E. R. Duval, S.-E. Chang, M. Heitzeg, S. A. Burt, L. Hyde, D. Amaral, C. W. Nordahl, L. T. Westlye, R. Zahn, H. G. Ruhe, C. Beckmann, A. F. Marquand, Charting brain growth and aging at high spatial precision, eLife 11 (2022). doi:10.1101/2021.08.08.455487.

[16] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al., Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS medicine 12 (3) (2015) e1001779.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[18] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (12 2013).
URL http://arxiv.org/abs/1312.6114

[19] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2 2018).
URL http://arxiv.org/abs/1802.03426

[20] S. P. Brooks, A. Gelman, General methods for monitoring convergence of iterative simulations, Journal of Computational and Graphical Statistics 7 (1998) 434–455.

[21] T. Hastie, R. Tibshirani, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Vol. 2, Springer, 2009.

[22] G. Casella, R. L. Berger, Statistical inference, Cengage Learning, 2021.

[23] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American Statistical Association 112 (5 2017). doi:10.1080/01621459.2017.1285773.
URL http://arxiv.org/abs/1601.00670http://dx.doi.org/10.1080/01621459.2017.1285773

[24] M. Betancourt, M. Girolami, Hamiltonian monte carlo for hierarchical models, Current Trends in Bayesian Methodology with Applications 79 (2015) 2–4. doi:10.1201/b18502-5.

[25] C. D. Boor, A practical guide to splines, Vol. 27, Springer, 1978.

Appendix A. Sinh-Arcsinh transformation

$$S_{\epsilon,\delta}(x) = \sinh(\delta \sinh^{-1}(x) - \epsilon) \tag{A.1}$$
$$S^{-1}_{\epsilon,\delta}(x) = \sinh((\sinh^{-1}(x) + \epsilon)/\delta) \tag{A.2}$$
$$= \sinh(\sinh^{-1}(x)\delta^{-1} + \epsilon\delta^{-1}) \tag{A.3}$$
$$= S_{\delta^{-1},\epsilon\delta^{-1}}(x) \tag{A.4}$$

Appendix B. Model Specification

Every phenotype has its own defining characteristics. Some may be severely skewed, while others appear to be nicely Gaussian. We may see strong correlations between age and variance (heteroskedasticity) in one, while another seems to retain a consistent variance throughout the entire age spectrum. Batch effects like sex-specific differences, or site-specific noise characteristics may or may not be present in the data. Characteristics like these can be learned by a sufficiently flexible model. However, Occam dictates that we should prefer a simple model over a complex model, if they perform equally well. A model is therefore optimally defined if it has enough flexibility to capture the parameters of interest, whilst being constrained in every other way. An ideal modelling framework would allow specifying constraints on all aspects of the model, be flexible enough to support a wide range of models, while still remaining concise. Here a model definition framework is proposed that allows us to concisely write down a wide range of different models. This framework will simplify the later sections of this report. In addition, the framework introduced here aligns with the implementation of the source code in the `pcntoolkit` software package. The graphical notation used here is slightly adapted from [13] chapter 8, where here we indicate the range and name of the iterator variables more explicitly in the plates, i.e. N becomes $n \in \{1, ..., N\}$. For the remainder of this section we will uphold the nomenclature in Tab. B.5. Without loss of generality, let us now focus on modelling a single parameter, θ. For more

Variable	Interpretation
$N \in \mathbb{N}$	The number of observations, subjects, rows in the table
d	The number of clinical covariates
$X \in \mathbb{R}^{N \times d}$	Clinical covariates
$Y \in \mathbb{R}^{N \times 1}$	Response variable
B	The number of unique combinations of batch effects. So if we model 2 sexes and 7 sites, this will be $2 \times 7 = 14$
$Z \in \mathbb{N}^{N \times 1}$	A list of N indices, corresponding to the combination of batch effects that apply to the data. So every $z_n \in \{1, ..., B\}$. Not to be confused with the z-statistics discussed below.
D	The dimensionality of the basis expansion
$\Phi \in \mathbb{R}^{N \times dD}$	A design matrix derived from X. This can be a polynomial basis expansion, a b-spline basis expansion, or simply X.

Table B.5: modelling nomenclature

complex distributions that take more parameters, like almost any distribution, we just apply what follows to every parameter separately.

Appendix B.1. Constant θ

Here we discuss simple models where θ is constant for the whole spectrum of X, save batch effects.

Model $\mathcal{M}_{\mathbf{1}}$. Let $\mathcal{M}_1$ represent the case where θ does not depend on X or Z, i.e. it takes a single value for all y_n. The only dependency of θ in $\mathcal{M}_1$ is on the (set of) hyperparameters α through its prior distribution p:

$$\theta \sim p_\theta(\alpha) \tag{B.1}$$

and Fig. B.10 shows the graphical model.

Figure B.10: Graphical model for $\mathcal{M}_1$

Model $\mathcal{M}_{\mathbf{2a}}$. Combinations of site, sex, and possible other factors define distinct groups called *batches*. Here we use the term *batch-effects* to indicate possibly unknown differences between known and distinct batches present in the data. HBR can deal with these batch-effects in a principled way by modelling batch-specific characteristics as deviations from a learned group mean. This allows information from all batches to flow into the estimated shared parameters, which in turn influence the estimated individual parameters for each batch. The next simplest case is model $\mathcal{M}_{2a}$, in which θ does depend on Z but not on X. This encodes the belief that the parameter is constant for all x, but that it may be influenced by batch effects. We sample the batch-effects as Gaussian deviations from a group mean with a group variance, such that the sites can learn from each other. Additionally, we have that under the appropriate prior, the group variance acts as a regularizer.

$$\begin{aligned}
&\mu_\theta \sim p_\mu(\alpha_{\mu_\theta}) && \\
&\sigma_\theta \sim p_\sigma(\alpha_{\sigma_\theta}) && \\
&\theta_b \sim \mathcal{N}(\mu_\theta, \sigma_\theta) && b \in \{1, \ldots, B\} \\
&\theta_n = \theta_{z_n} && n \in \{1, \ldots, N\}
\end{aligned} \tag{B.2}$$

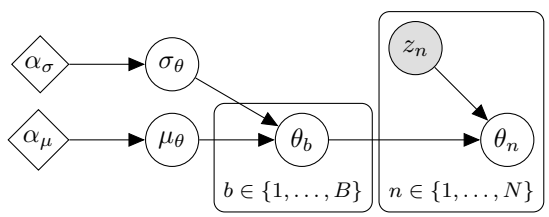


Figure B.11: Graphical model for $\mathcal{M}_{2a}$

Model $\mathcal{M}_{\mathbf{2b}}$. The phenomenon known as the funnel is an issue that commonly afflicts hierarchical models, and it may occur in model $\mathcal{M}_{2a}$. To give a brief explanation, when the parameter σ_θ is small, but we can not determine θ with enough confidence, the posterior distribution attains a wide support in the domain of σ_θ, including very small values. If we sample in those regions where σ_θ is very small, the values θ all collapse to μ_θ, and sampling of θ becomes very hard. Most samples of θ will be rejected, since the proposal distribution is tuned for a larger step size. For a more thorough explanation see [24]. In those situations where we fear a funnel may occur, we apply the following reparameterization as a remedy; we sample the group magnitude of the deviations from a positive distribution, and then multiply with an *offset* sampled from a central member of a symmetric distribution—like the standard

Gaussian—for each $b \in \{1, \ldots, B\}$. Sampling like this is sometimes called a non-centered sampling approach, as opposed to the centered approach in model $\mathcal{M}_{2a}$. Here we propose model $\mathcal{M}_{2b}$:

$$
\begin{aligned}
&\mu_\theta \sim p_\mu(\alpha_{\mu_\theta}) && \\
&\sigma_\theta \sim p_\sigma(\alpha_{\sigma_\theta}) && \\
&\nu_b \sim \mathcal{N}(0,1) && b \in \{1, \ldots, B\} \\
&\theta_b = \mu_\theta + \sigma_\theta \nu_b && b \in \{1, \ldots, B\} \\
&\theta_n = \theta_{z_n} && n \in \{1, \ldots, N\}
\end{aligned}
\tag{B.3}
$$

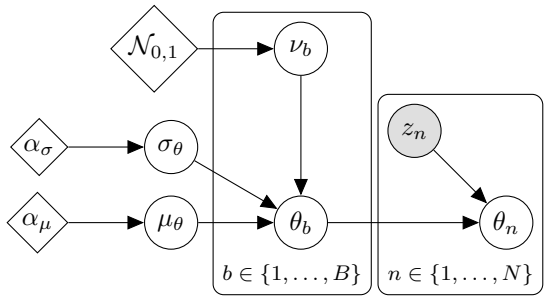


Figure B.12: Graphical model for $\mathcal{M}_{2b}$

Appendix B.2. θ as a function of X

More complex models can be defined if we let θ follow a linear relationship with the basis expansion $\Phi(X)$ through a set of weights $\mathbf{w} \in \mathbb{R}^{dD}$ and an intercept $\tau \in \mathbb{R}$. We sample $\mathbf{w}$ and τ from their own respective distributions, which is where the definitions from section Appendix B.1 are re-used.

Model $\mathcal{M}_{\mathbf{3}.*.*}$. Since $\mathbf{w}$ and τ are both constant parameters with possible batch-effects, we can model them exactly like we did θ in $\mathcal{M}_1$, or θ_{z_n} in models $\mathcal{M}_{2a}$ and $\mathcal{M}_{2b}$. The definitions there are general enough to allow modelling $\mathbf{w}$ as distributed according to a multidimensional density.

The subscript of these linear models consist of three terms, separated by periods; $\mathcal{M}_{3.*.*}$. The first term is a 3, signifying that this is a linear model. The second and third terms indicate the model types that are used for $\mathbf{w}$ and τ, respectively. For instance: $\mathcal{M}_{3.1.1}$ is a model where θ is a linear function of Φ through $\mathbf{w}$ and τ, which are both constants like θ in $\mathcal{M}_1$. Model $\mathcal{M}_{3.2a.2b}$ encodes for the model with a linear θ, where $\mathbf{w}$ and τ have unique values for each batch-effect, but the former is modeled by the centered approach, and the latter by the non-centered approach explained on page 29. Both $\mathbf{w}$ and τ can thus be modeled in three distinct ways, allowing us to define a total of 9 different models with linear θ in this way. θ can thus be modeled in a total of 12 ways. 3 in which it is fixed, and 9 in which it has a linear dependence on Φ.

$$
\begin{aligned}
\tau &\sim \text{ like Eq. B.1, B.2, or B.3} \\
\mathbf{w} &\sim \text{ like Eq. B.1, B.2, or B.3} \\
\theta &= \phi_n^T \mathbf{w} + \tau && n \in \{1, ..., N\}
\end{aligned}
$$

Appendix B.3. Hypothetical application

The proposed framework allows us to concisely specify how parameters are modeled.

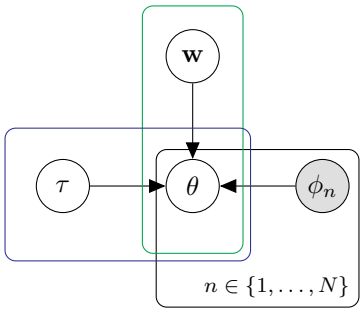


Figure B.13: Graphical model for $\mathcal{M}_{3.*.*}$. The green and blue box are added to emphasize that $\mathbf{w}$ and τ may be modeled with batch-effects, resulting in multiple values for θ.

Scenario. A pattern that is commonly observed in neurological phenotypes is the following: As the age of the participant (X) increases, the phenotype (Y) increases first (until around 25), and then starts to decrease. However, the variance in the phenotype increases with age as well. The data shows some positive skew, and may also be slightly kurtotic. Furthermore, the data is collected from various sites, so site-effects (Z_1) are expected, and sex-related differences (Z_2) need to be accounted for as well.

modelling. Clearly, the parameter that controls the mean should be a function of age, but the non-monotonic behavior of the mean guides us away from a directly linear approach. Rather, we would use a polynomial basis expansion, or a regression spline [25]. We usually prefer the regression spline basis over a simple polynomial basis, because it does not induce global curvature [8]. A Gaussian distribution would not be able to capture the observed skew, so the decision for a likelihood function falls on the SHASH_b, parameterized by μ, σ, ϵ, and δ. Model $\mathcal{M}_3$ is a fitting choice for both μ and σ, as they appear to be related with age. For μ, we expect that the batch-effects do not affect the trajectory of the phenotype, but mainly influence the magnitude. Therefore, the weights $\mathbf{w}$ do not need to be modeled with batch-effects, but τ does. The expected batch-effects are not very large, so the non-central modelling approach (page 29) is a good choice. Model $\mathcal{M}_{3.1.2b}$ is decided for μ. We do not expect batch-effects in σ, but it does clearly have a relation with age. We can safely choose model $M_{3.1.1}$ for σ. The last two parameters, ϵ and δ, are both assumed to be constant across sites and age. So they are both modeled by $\mathcal{M}_1$. To enforce positivity of both σ and δ, we apply a softplus transformation to the outputs of the corresponding models, and to fulfill the constraint discussed in section 2.1.2, we add a constant to δ.

We now have:

$$
\begin{aligned}
\mu_n &\sim \mathcal{M}_{3.1.2b} && n \in \{1, \dots, N\} \\
\sigma_n^{\pm} &\sim \mathcal{M}_{3.1.1} && n \in \{1, \dots, N\} \\
\sigma_n &= \text{softplus}(\sigma_n^{\pm}) && n \in \{1, \dots, N\} \\
\epsilon &\sim \mathcal{M}_1 \\
\delta^{\pm} &\sim \mathcal{M}_1 \\
\delta &= 0.3 + \text{softplus}(\delta^{\pm}) \\
y_n &\sim \mathcal{S}_b(\mu_n, \sigma_n, \epsilon, \delta) && n \in \{1, \dots, N\}
\end{aligned}
$$

Appendix C. MCMC sampling in a nutshell

Uncertainty quantification ideally accounts for all uncertainty in the model parameters, but is often hard to quantify exactly. MCMC approximates this uncertainty by sampling from the desired

distribution. Given a generative model like above, Bayes' rule gives us an expression for the posterior over our model parameters:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \tag{C.1}$$

which we then use to compute the expectation of a function f:

$$\mathbb{E}_{\theta\sim p(\theta|D)}[f(\theta)] = \int_{\theta} f(\theta)p(\theta|D)d\theta \tag{C.2}$$

For instance, in normative modelling, f could return the deviation score of some data point. The expectation above gives us an estimate in which all the uncertainty in the model parameters is captured. Unfortunately, the integral in Eq. C.2 is intractable, and the denominator in Eq. C.1, $p(D)$, contains an intractable integral as well:

$$p(D) = \int_{\theta'} p(D|\theta')p(\theta')d\theta' \tag{C.3}$$

In practice it is impossible to determine $p(\theta|D)$ exactly, unless the prior is conjugate to the likelihood. MCMC samplers use a number of different methods to retrieve pseudo-random samples from $p(\theta|D)$, without ever needing to analyze the integral in Eq. C.3. With enough of these samples in hand, we use the following result:

$$\theta \sim p(\theta|D),$$
$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(\theta_i) = \int_{\theta} f(\theta)p(\theta|D)d\theta = \mathbb{E}_{\theta\sim p(\theta|D)}[f(\theta)]$$

to approximate our desired expectation.

Appendix D. Maximum-a-posteriori (MAP) approximation

Instead of doing an MCMC estimation, which is quite computationally expensive, another method is to find the MAP, and to compute the desired function based on that point.

$$\theta^{MAP} = \arg\max_{\theta} p(\theta|D)$$

So in other words, the MAP estimate is just that point in parameter space that maximizes the posterior. It can be found by for example a gradient based method. Using the MAP-estimate is not the most principled method, which can easily be imagined thinking of multimodality, or of large flat optima in the posterior. However, the computational gains may be worth the loss of a representation of uncertainty.

Appendix E. Data counts

Site	# Females	# Males	Mean age
ABCD_01	199	189	9.89
ABCD_02	252	290	10.07
ABCD_03	268	301	9.88
ABCD_04	307	324	9.81
ABCD_05	178	167	9.89
ABCD_06	286	278	9.94
ABCD_07	153	172	9.87
ABCD_08	160	176	9.95
ABCD_09	200	207	9.96
ABCD_10	280	295	9.86
ABCD_11	206	208	9.81
ABCD_12	77	84	9.88
ABCD_13	277	278	9.81
ABCD_14	267	316	10.19
ABCD_15	178	217	9.90
ABCD_16	414	507	9.90
ABCD_17	267	290	9.81
ABCD_18	160	181	9.90
ABCD_19	272	262	10.06
ABCD_20	320	320	10.05
ABCD_21	222	269	9.91
ABIDE_GU	27	27	10.42
ABIDE_KKI	65	122	10.29
ABIDE_NYU	28	107	14.40
ABIDE_USM	3	56	22.07
ADD200_KKI	27	34	10.25
ADD200_NYU	20	18	10.21
AOMIC_1000	483	445	22.84
AOMIC_PIPO2	120	89	22.18
ATV	17	60	22.67
CIN	22	44	47.89
CMI_CBIC	73	126	11.19
CMI_RU	125	251	10.53
CMI_SI	38	68	11.01
CNP-35343.0	43	47	31.94
CNP-35426.0	11	9	29.50
COI	78	46	51.86
HCP_A_MGH	86	85	59.75
HCP_A_UCLA	71	53	53.32
HCP_A_UM	120	84	61.59
HCP_A_WU	112	66	58.83
HCP_D_MGH	109	107	13.78
HCP_D_UCLA	62	65	14.13
HCP_D_UM	85	71	13.26
HCP_D_WU	75	79	13.97
HCP_EP_BWH	2	6	23.09
HCP_EP_IU	11	14	23.90
HCP_EP_MGH	3	8	27.61
HCP_EP_McL	4	9	25.15
HKH	17	12	45.41
HRC	36	13	41.69
HUH	38	29	34.74
KCL	25	16	34.12
KTT	41	87	31.03
KUT	66	93	36.50
Oasis2	127	58	76.88
Oasis3	631	921	69.78
SWA	15	85	28.48
SWU_SLIM_ses1	318	230	20.07
UCDavis	62	74	3.12
UTO	106	96	35.19
cam	329	318	54.19
delta	18	31	50.44
ds001734	60	48	25.54
ds002236	38	48	11.49
ds002330	37	29	26.62
ds002345	131	76	21.69
ds002731	28	31	21.25
ds002837	42	44	26.73
hcp_ya	606	507	28.80
ixi	313	245	48.71
nki	308	174	42.63
pnc	701	677	14.21
top	134	158	34.57
ukb-11025.0	12944	12042	62.98
ukb-11027.0	5438	4534	64.39

Table E.6: Control data, by site

Site	# Females	# Males	Mean age
ABIDE_GU	8	39	10.94
ABIDE_KKI	19	58	10.22
ABIDE_NYU	16	111	12.83
ABIDE_USM	2	72	21.60
ADD200_KKI	10	12	10.21
ADD200_NYU	14	38	10.05
CIN	28	45	56.80
CNP-35343.0	30	33	35.33
CNP-35426.0	15	44	34.62
COI	40	29	45.07
HCP_EP_BWH	8	15	22.38
HCP_EP_IU	22	37	22.96
HCP_EP_MGH	3	7	20.31
HCP_EP_McL	15	16	23.56
HKH	13	20	44.81
HRC	10	6	40.50
HUH	25	32	43.33
KCL	75	29	31.44
KTT	21	26	37.89
KUT	30	31	41.70
Oasis2	82	97	76.94
Oasis3	158	123	75.37
SWA	19	115	33.61
UTO	54	95	35.61
delta	142	160	41.25
top	254	277	32.42

Table E.7: Patient data, by site

Group	# Females	# Males	Mean age
AD	142	94	75.75
ADHD	40	71	17.60
ASD	61	389	19.08
BD	130	106	34.67
Converted	24	13	79.75
Demented	58	84	76.21
MCI	16	29	73.40
MDD	260	239	40.26
MDD_motar	30	21	35.50
Other	15	26	34.21
Pain	25	39	55.28
Patient	72	98	30.01
Patient_motar	32	38	40.20
SZ	131	225	34.85
Stroke	3	6	67.66
otherpsychosis	32	52	29.76
schizoaffective	23	10	35.69
schizophreniform	11	12	25.43

Table E.8: Patient data, by group

Appendix F. $\hat{R}$ on parameters of mean and variance

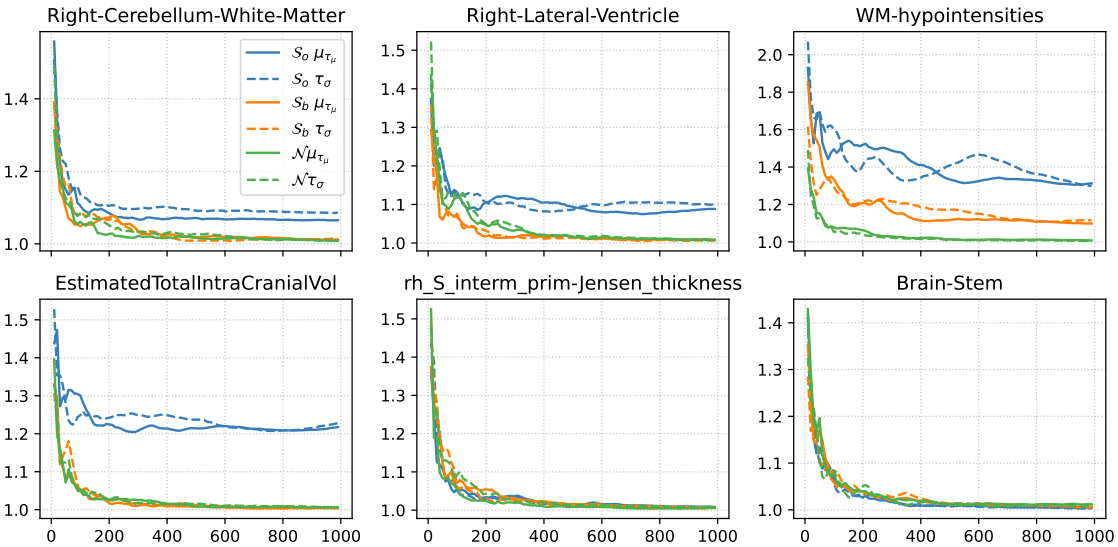


Figure F.14: $\hat{R}$ statistics of parameters of mean and variance

Appendix G. Empirical moments of SHASHo

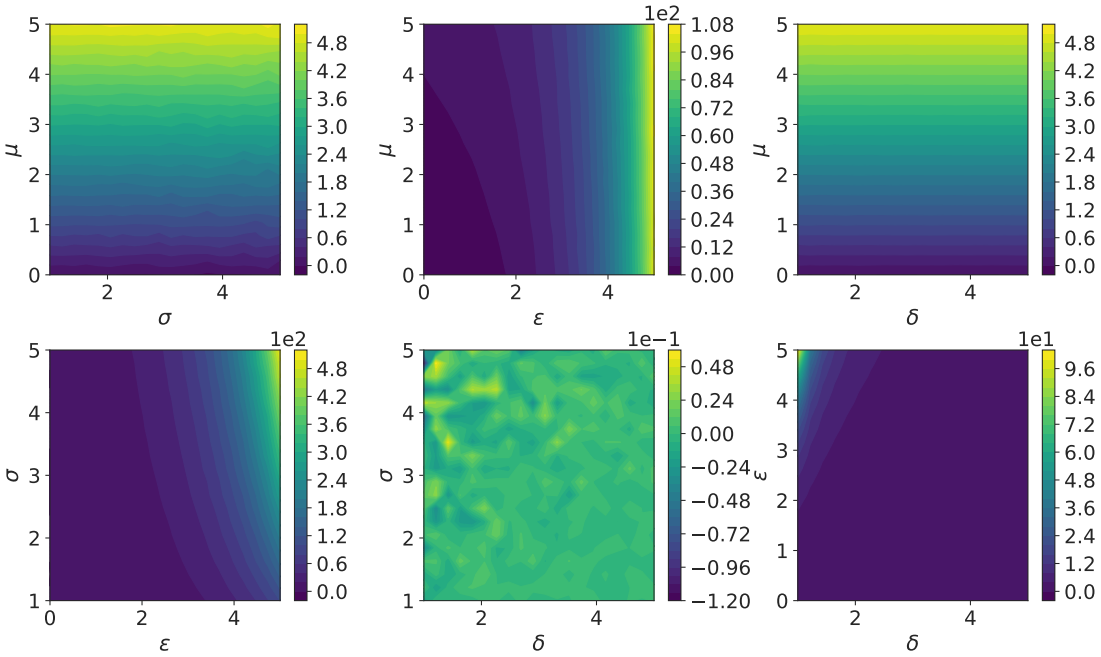


Figure G.15: SHASH$_o$ first moment $E[x]$

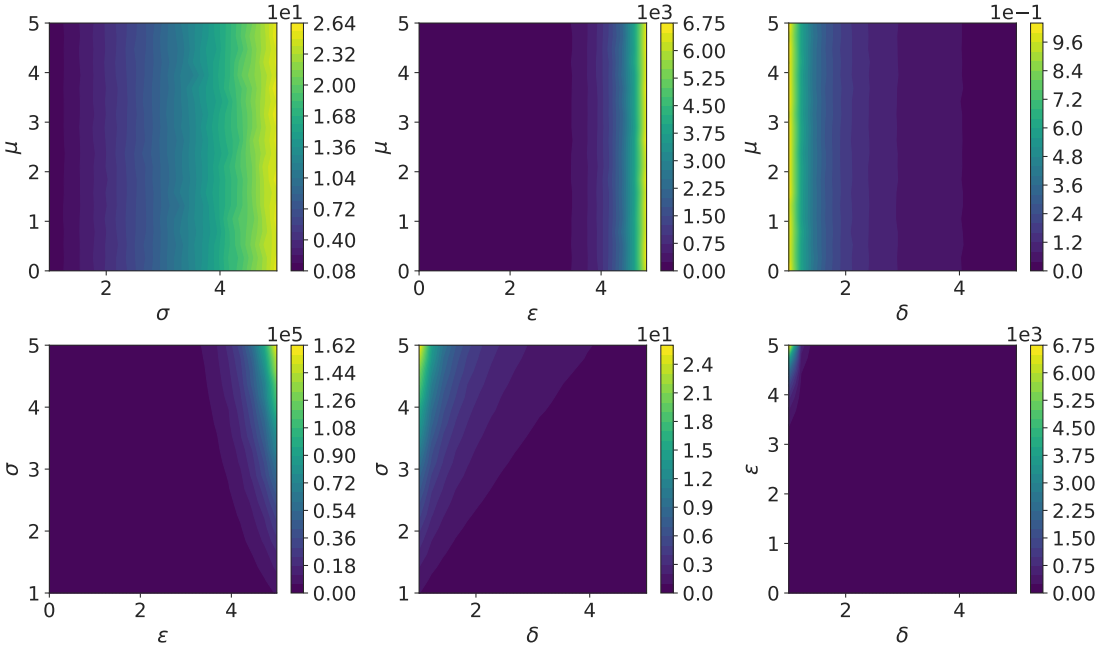


Figure G.16: SHASH$_o$ second moment $E[(x-\mu)^2]$

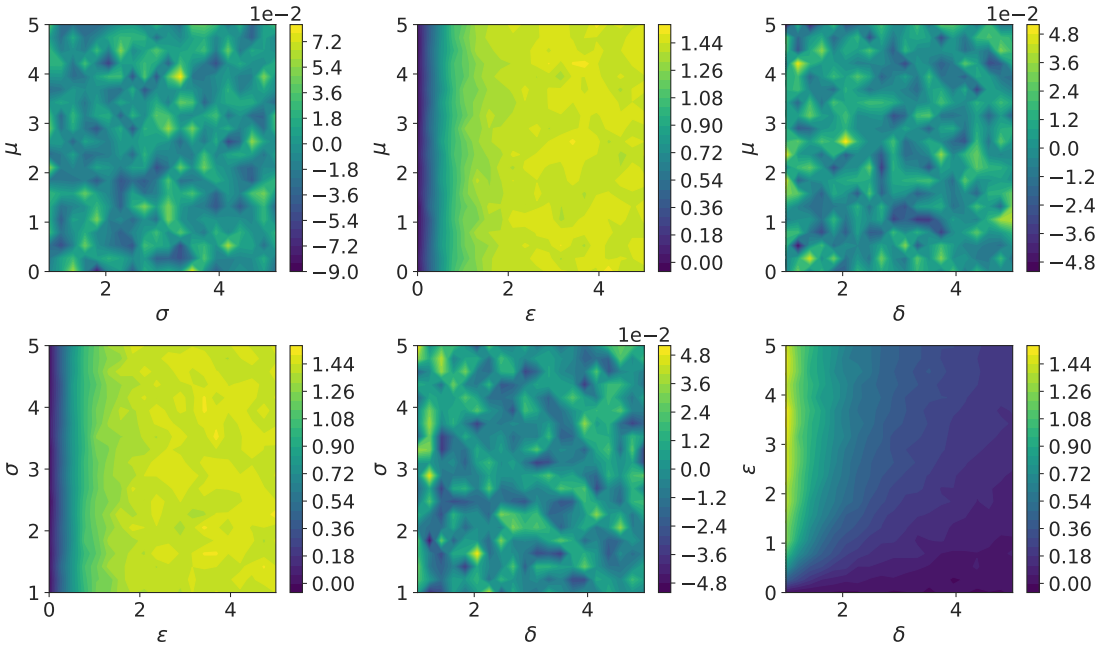


Figure G.17: SHASH$_o$ third moment $E[((x-\mu)/\sigma)^3]$

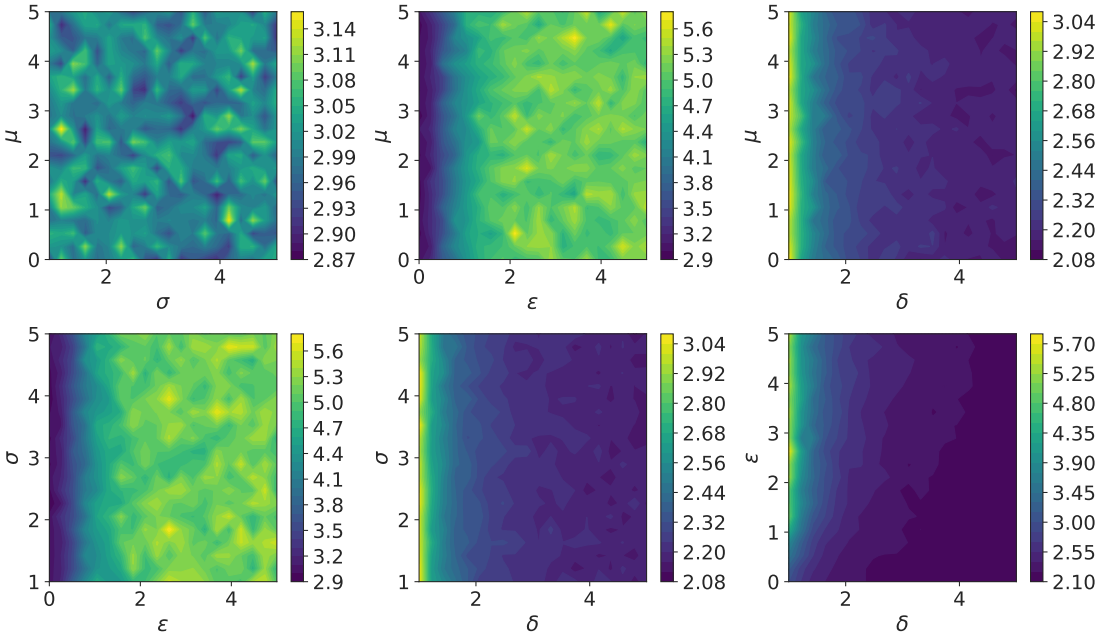


Figure G.18: SHASH$_o$ fourth moment $E[((x-\mu)/\sigma)^4]$

Appendix H. Empirical moments of SHASHb

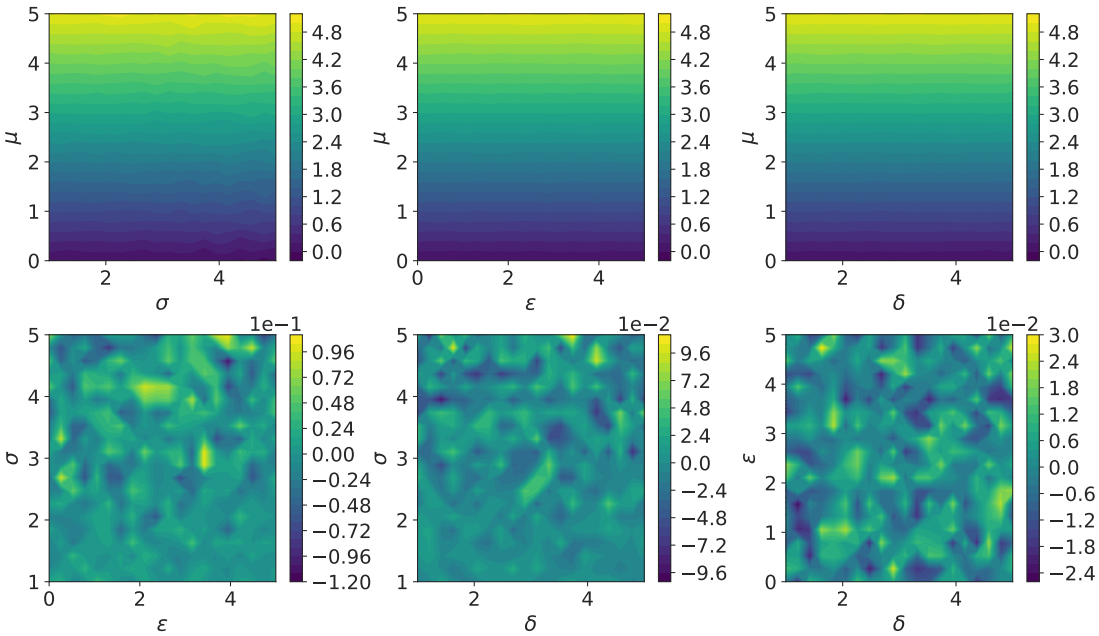


Figure H.19: SHASH$_b$ first moment $E[x]$

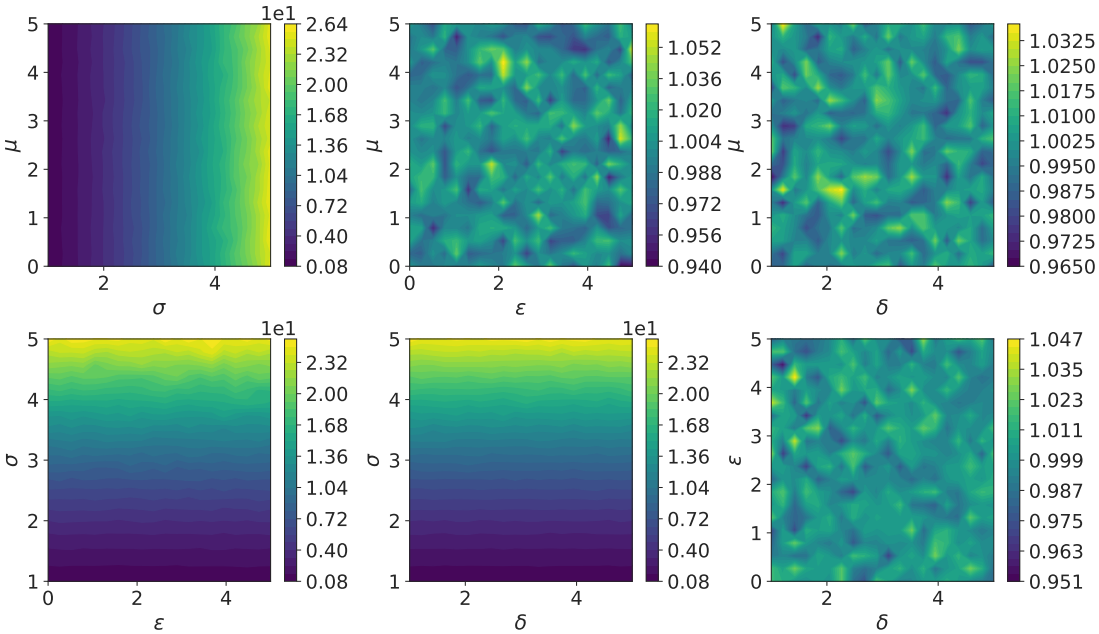


Figure H.20: SHASH$_b$ second moment $E[(x-\mu)^2]$

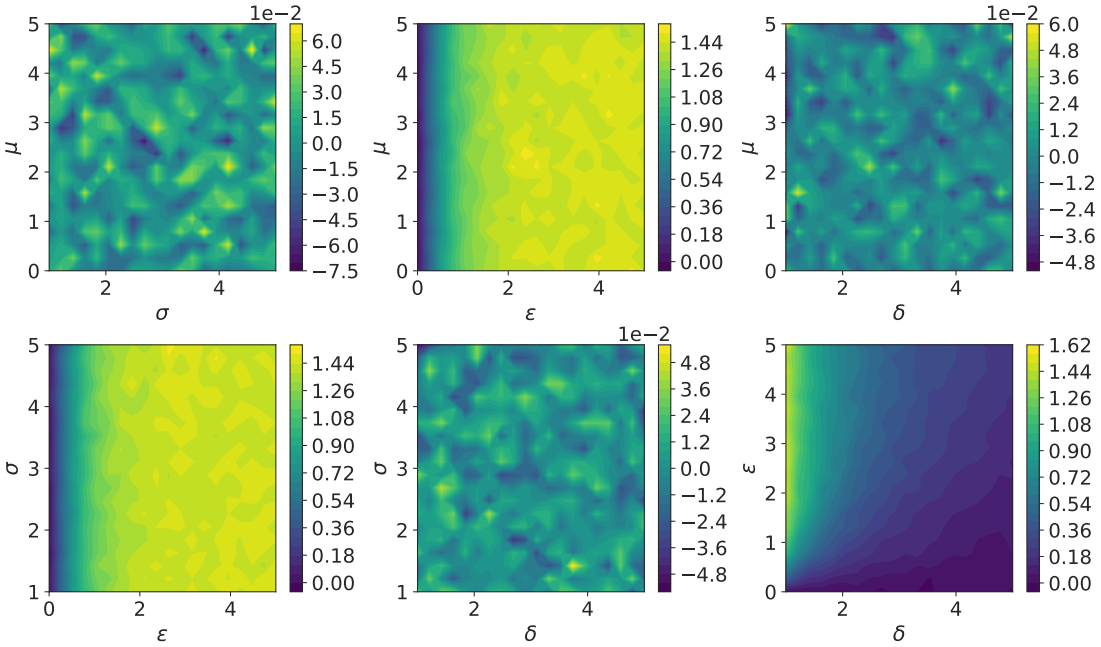


Figure H.21: SHASH$_b$ third moment $E[((x-\mu)/\sigma)^3]$

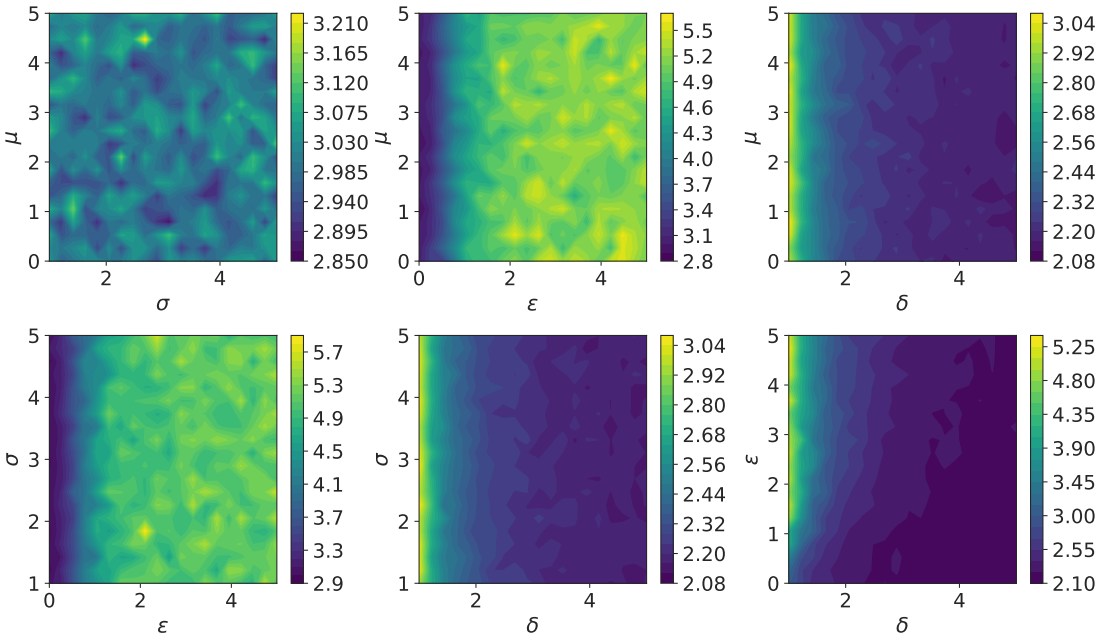
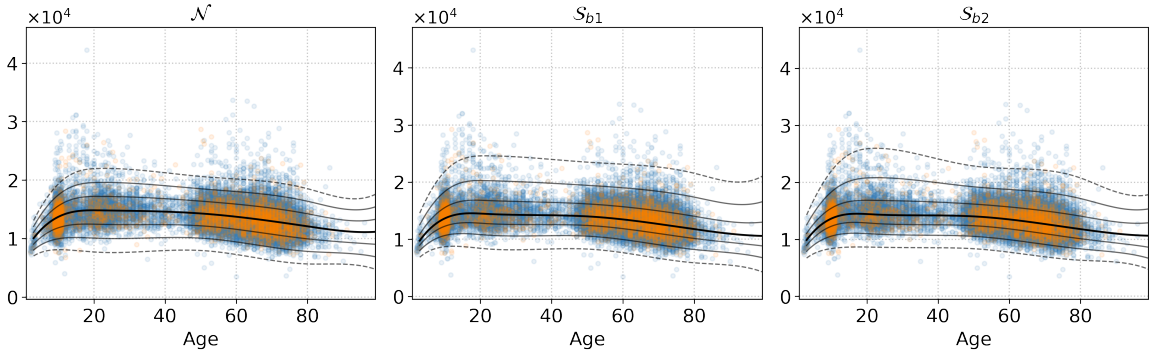


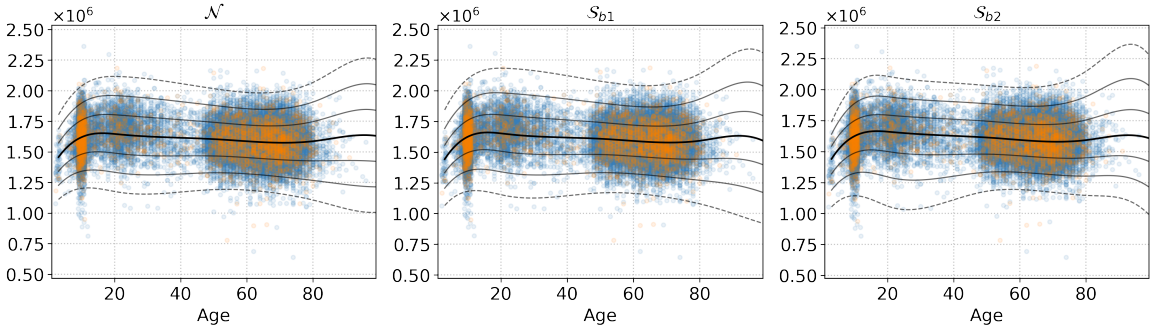
Figure H.22: SHASH$_b$ fourth moment $E[((x-\mu)/\sigma)^4]$

Appendix I. Quantiles of all features, by all HBR models



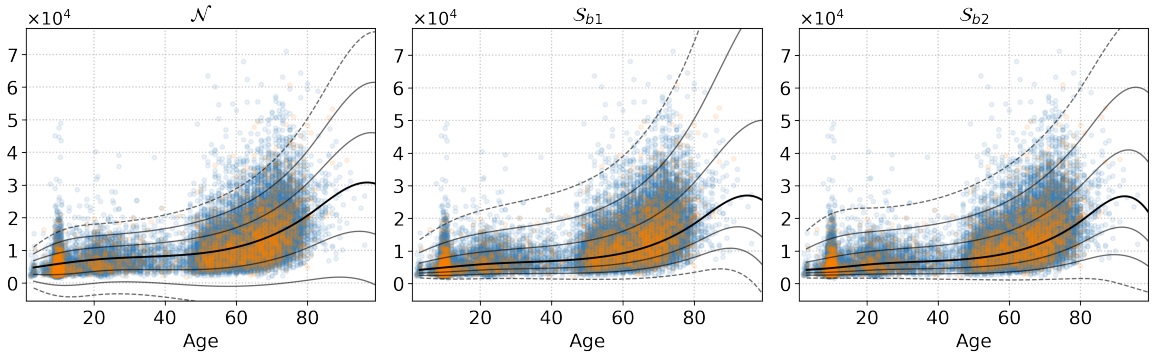
(a) (b) (c)

Figure I.23: `Right-Cerebellum-White-Matter`



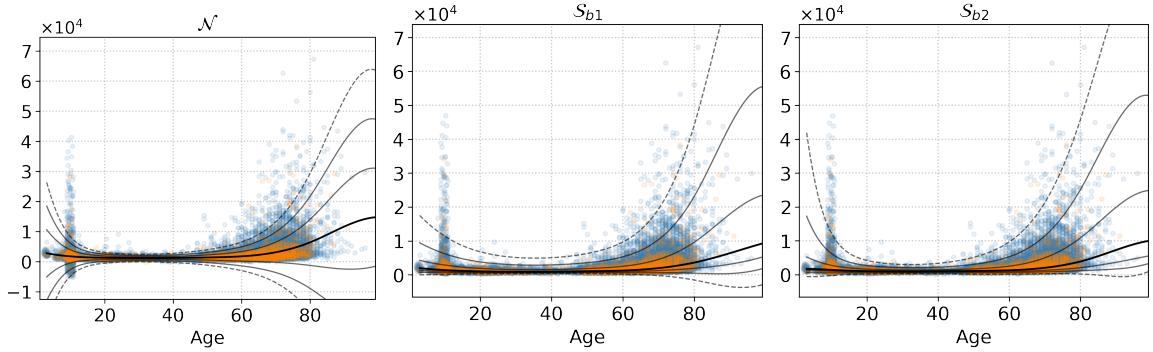
(a) (b) (c)

Figure I.24: `EstimatedTotalIntraCranialVol`



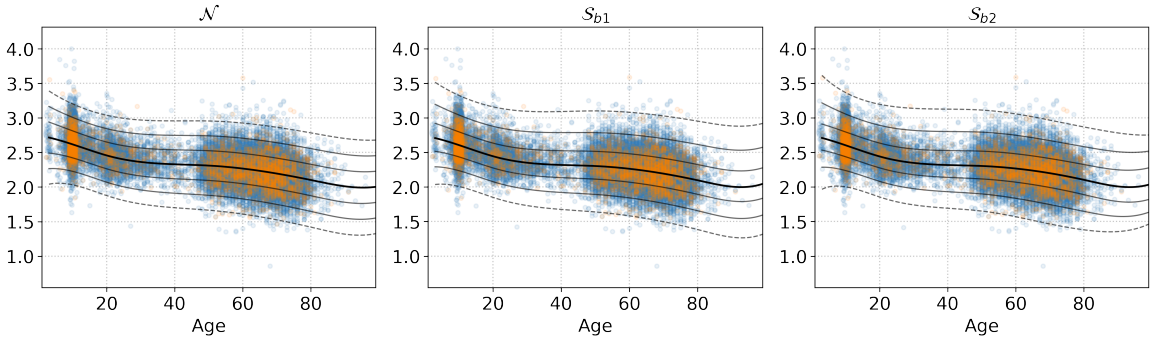
(a) (b) (c)

Figure I.25: `Right-Lateral-Ventricle`



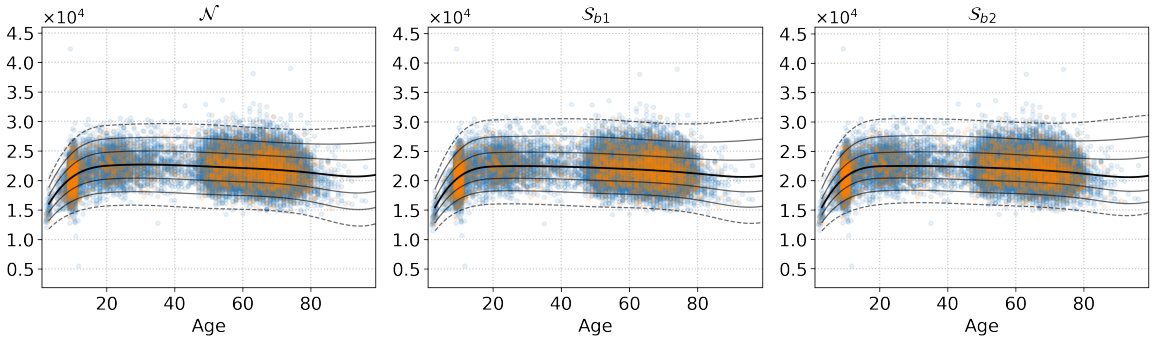
(a) **(b)** **(c)**

Figure I.26: `WM-hypointensities`



(a) **(b)** **(c)**

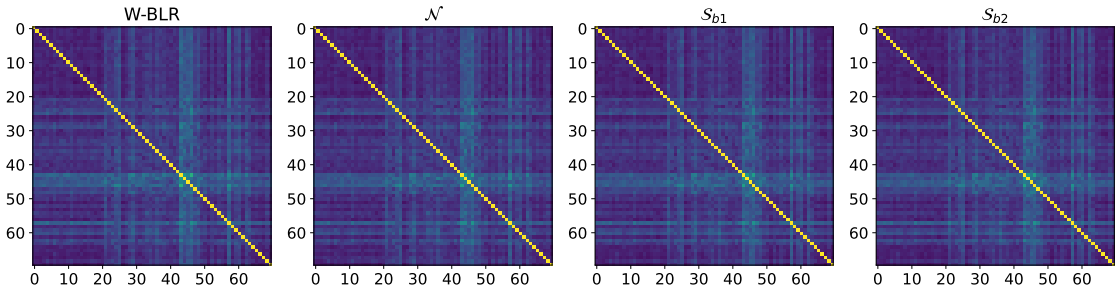
Figure I.27: `rh_S_interm_prim-Jensen_thickness`



(a) **(b)** **(c)**

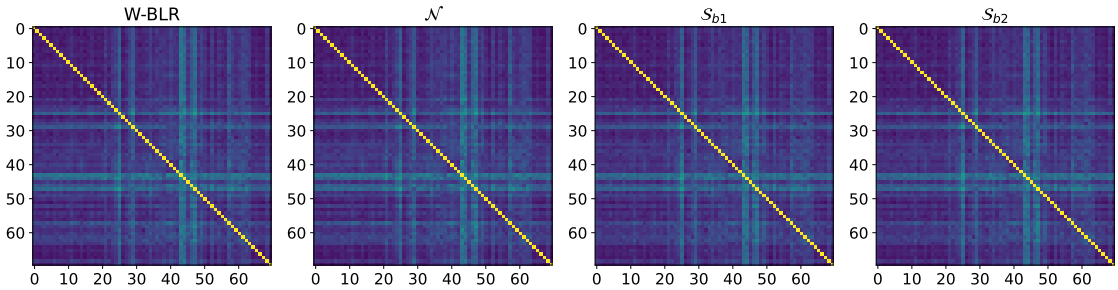
Figure I.28: `Brain-Stem`

Appendix J. AUC scores on all phenotypes



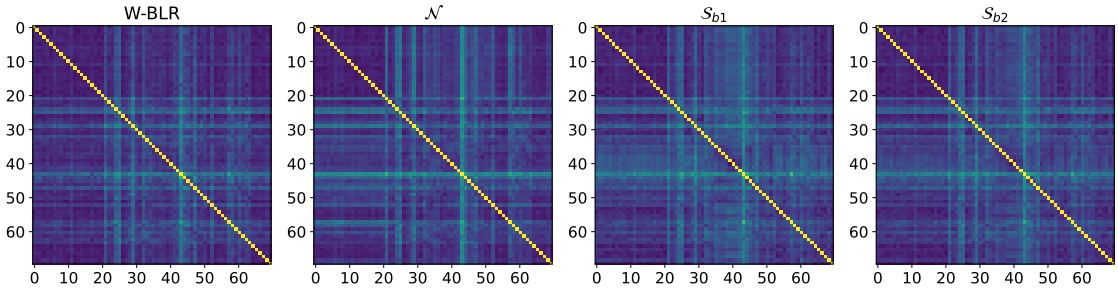
(a) **(b)** **(c)** **(d)**

Figure J.29: AUC scores of the Right-Cerebellum-White-Matter phenotype



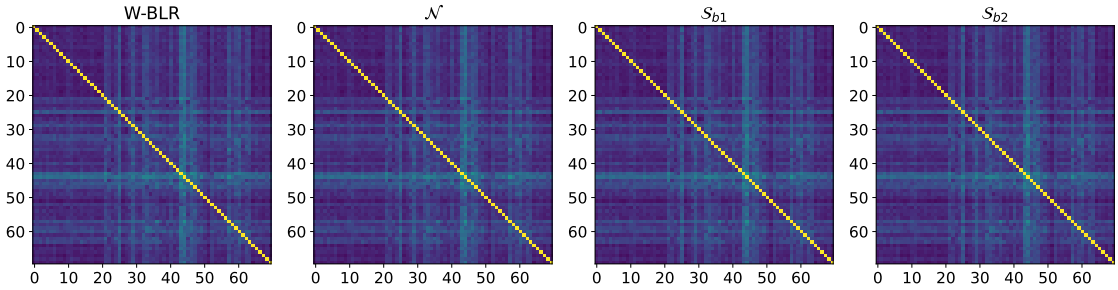
(a) **(b)** **(c)** **(d)**

Figure J.30: AUC scores of the Right-Lateral-Ventricle phenotype



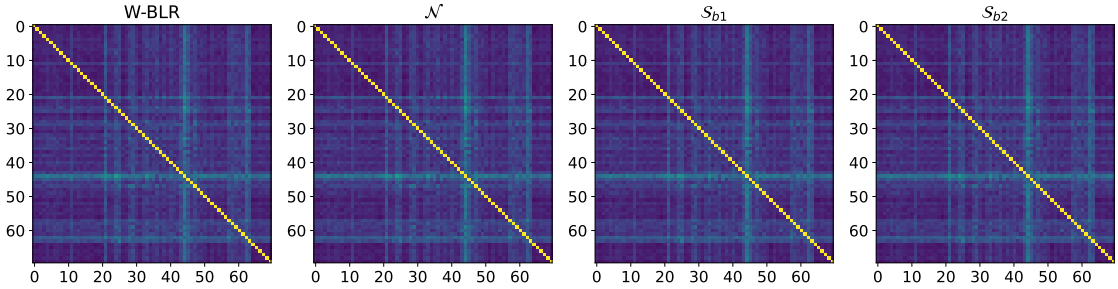
(a) **(b)** **(c)** **(d)**

Figure J.31: AUC scores of the WM-hypointensities phenotype



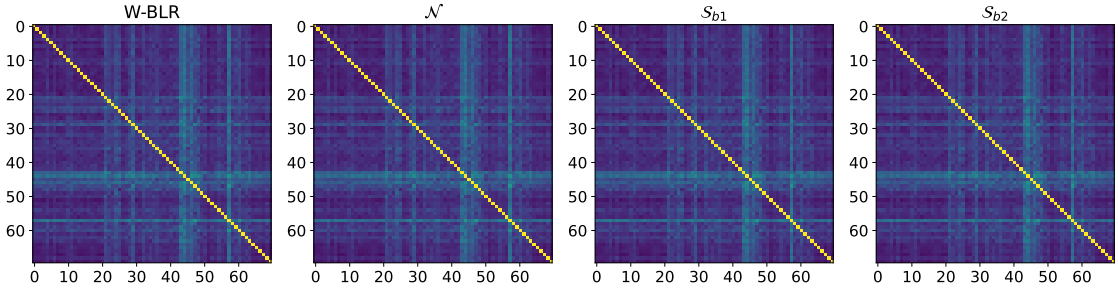
(a) **(b)** **(c)** **(d)**

Figure J.32: AUC scores of the EstimatedTotalIntraCranialVol phenotype



(a) **(b)** **(c)** **(d)**

Figure J.33: AUC scores of the rh_S_interm_prim-Jensen_thickness phenotype



(a) **(b)** **(c)** **(d)**

Figure J.34: AUC scores of the Brain-Stem phenotype