SubCMap: Subject and Condition Specific Effect Maps

Ender Konukoglu[a,*], Ben Glocker[b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Computer Vision Lab, ETH Zurich, Zurich, Switzerland*
[b] *Department of Computing, Imperial College London, London, United Kingdom*

Abstract

Current methods for statistical analysis of neuroimaging data identify condition related structural alterations in the human brain by detecting group differences. They construct detailed maps showing population-wide changes due to a condition of interest. Although extremely useful, methods do not provide information on the subject-specific structural alterations and they have limited diagnostic value because group assignments for each subject are required for the analysis. In this article, we propose SubCMap, a novel method to detect subject and condition specific structural alterations. SubCMap is designed to work without the group assignment information in order to provide diagnostic value. Unlike outlier detection methods, SubCMap detections are condition-specific and can be used to study the effects of various conditions or for diagnosing diseases. The method combines techniques from classification, generalization error estimation and image restoration to the identify the condition-related alterations. Experimental evaluation is performed on synthetically generated data as well as data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Results on synthetic data demonstrate the advantages of SubCMap compared to population-wide techniques and higher detection accuracy compared to outlier detection. Analysis with the ADNI dataset show that SubCMap detections on cortical thickness data well correlate with non-imaging markers of Alzheimer's Disease (AD), the Mini Mental State Examination Score and Cerebrospinal Fluid amyloid-β levels, suggesting the proposed method well captures the inter-subject variation of AD effects.

1. Introduction

Statistical analysis of neuroimaging data is instrumental in detecting condition-induced structural alterations of the human brain. It has been widely used for studying alterations due to various diseases, e.g. [1, 2, 3], lifestyle choices, e.g. [4, 5, 6], as well as genetics and inherited traits, e.g. [7, 8, 9]. The common approach in such analyses is to extract image-based anatomical measurements from structural images acquired from two groups of subjects, one showing the condition of interest and a control group without the condition, spatially normalise the measurements by mapping them to a common template and analyse the normalised data with a chosen statistical method. The types of hypotheses that can be tested and conclusions that can be deduced naturally depend on the statistical method as much as on the types of measurements.

Currently available statistical methods can process high number of measurements with complex spatial correlation structures. They allow for constructing detailed volumetric [10] and surface maps [11, 12] that highlight areas that show condition related alterations. When studying a specific condition, available techniques can be broadly divided into two categories: group analysis and predictive modeling. Despite the

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[*] Corresponding author

Email addresses: ender.konukoglu@vision.ee.ethz.ch (Ender Konukoglu), b.glocker@imperial.ac.uk (Ben Glocker)

wealth of research in both of these categories and the large number of alternative techniques, today there exist no techniques for detecting subject-specific alterations due to a condition of interest in a prospective fashion, i.e. without knowing a priori the presence of the condition.

Group analysis methods aim to detect statistical differences between groups either in a univariate fashion at each measurement indepedently [10] or in a multivariate fashion using multiple measurements simultaneously [13, 14]. Although immensely useful, detection results obtained with these techniques are population-wide, meaning they are "averages" across the population. Therefore, the identified anatomical areas not necessarily altered in all subjects who have the condition of interest. Researchers have tackled this issue in [15, 16] by performing one-versus-all group analysis. These attempts, however, do not resolve the second issue: group analysis is retrospective, i.e. condition information has to be available for each subject beforehand.

Predictive modeling approaches use machine learning tools to predict the presence of the condition of interest for a specific subject given the anatomical measurements [17]. These tools are most often multivariate and use all the available measurements simultanously. Researchers have also shown that predictive models can detect sets of measurements that hold predictive information with regards to the condition [18, 19, 20] and such detections may even yield higher reproducibility compared to conventional group analysis [21]. Detections of predictive models suffer from the same problem as group analysis: they reflect population averages and useful only retrospectively. It is important to note that although predictions are prospective and subject-specific, they only indicate that the *set of measurements as a whole show condition effect* without the possibility to localise the structural alterations.

Although not a condition specific analysis method, outlier detection is currently the only approach that can be used for detecting subject specific alterations. Techniques for outlier detection estimate normative distributions for the measurements from a population that only consists of individuals not showing the condition, i.e. controls. For a new subject, the measurements are then compared to the normative distributions and the ones with low likelihoods are determined as outliers. This has been used for detecting brain lesions and neurodegenerative changes [22, 23, 24, 25]. The main drawback is the unspecific nature of the method. Outlier detection identifies all measurements that lie outside the respective normative distribution. The resulting detections are not specific to a condition of interest, hence, cannot be easily used for studying a particular condition.

Detecting subject and condition specific alterations prospectively is important and has numerous applications. In clinical and neuroscience research, subject-specific detections can be used to identify subpopulations [26] and enable accurate stratification. In engineering research, machine learning tools are often "black-box" components. Subject-specific results can be used to analyse cases where these tools fail and facilitate model improvements and possibly identify incorrect labels in the ground truth data. For clinical practice, subject-specific detections can enable translation of predictive models into practice, where they can be used to provide reasonings to the automatic diagnosis that are otherwise produced in a fashion that leaves clinicians blind to the process.

In this article we present a novel method to detect subject and condition specific alterations in a prospective manner. The proposed method, which we refer to as "SubCMap", combines elements from predictive modeling, generalization error estimation and image restoration in a simple yet effective formulation. As common in all predictive models, SubCMap also has a training phase, where a training dataset is used to learn the model and its parameters, and a test phase, where unseen data is analysed without assuming information regarding the presence of the condition of interest. SubCMap is a univariate method by design to allow for localized interpretations, which would not be possible with multivariate approaches [27]. Lastly, it also allows for regressing out nuisance parameters which is often required to account for variations not important for a particular study.

The focus of the article is on spatial maps of image-based measurements where local measurements are extracted densely at multiple points from an anatomical structure, such as voxel-wise gray matter density [10] and surface-based cortical thickness maps [12]. SubCMap can also be applied to other types of measurements, such as volumes of multiple anatomical structures, but it is especially designed for high-dimensional measurements with spatial context and makes use of the associated correlation structure. Given the input measurement maps extracted from a new subject, the method constructs corresponding detection maps that highlight areas showing condition-related alterations.

We first describe the proposed method in Section 2 and then evaluate it in Section 3. We perform evaluations both with synthetically generated data, where ground truth information is available, and with data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), where the goal is to detect structural alterations due to Alzheimer's Disease (AD). We compare the proposed method with a univariate general linear model (GLM) [28], support vector machines (SVM) [29], random forests (RF) [30] and outlier detection. Results on synthetic data demonstrate the conceptual advantages of SubCMap compared to GLM, SVM and RF, and substantial improvements in sensitivity and specificity compared to outlier detection. Results on the ADNI dataset shows the advantages of the proposed method in studying condition-related alterations in terms of extracting subject-specific detection maps. Furthermore, our correlation analysis between global statistics derived from subject-specific detections, i.e. area of the affected region, and non-imaging AD markers, in particular Mini Mental State Examination scores (MMSE) and Cerebrosprinal Fluid amyloid-β (CSF a-β) measurements, show that global statistics can achieve as strong correlations with the non-imaging AD markers as dedicated machine learning tools using raw measurements. This suggests that SubCMap can capture inter subject variations in the structural alterations due to the condition. Finally, we conclude with discussions and conclusions in Section 4.

2. Method

The proposed method is a statistical technique to analyse measurements across individuals. In the following we will assume that measurements extracted from different individuals are spatially normalised, which means they are aligned with a common template and corresponding measurements can be directly compared. Such a normalization can be achieved for instance using publicly available tools, such as SPM and Freesurfer [2].

2.1. Notation

We represent the measurements extracted from a subject's image, i.e. the input map, with the vector $\mathbf{f} = [f_1, \dots f_d] \in \mathbb{R}^d$. In case of cortical thickness maps, each component of $\mathbf{f}$ corresponds to the thickness value at a vertex and in the case of grey-matter density maps, to the density value at a voxel. Additionally, for each subject we represent the presence of the condition of interest with a binary variable $y \in \{0, 1\}$, which we refer to as the label. For instance, when the condition of interest is a disease then the binary variable corresponds to diagnosis. When analysing the measurements of a new subject the y variable is assumed to be unknown. SubCMap takes vector $\mathbf{f}$ as input and outputs a binary vector $\mathbf{q} \in \{0, 1\}^d$, where each component indicates presence of the condition-related effect at the corresponding measurement, i.e. $q_j = 1$ means condition effect is detected at that measurement and $q_j = 0$ means no condition effect is detected. To construct vector $\mathbf{q}$, SubCMap uses a cohort of subjects where both measurements and labels, $\{\mathbf{f}_n, y_n\}_{n=1,\dots,N}$, are available, i.e. the training dataset. We assume that the training dataset contains two groups, one composed of individuals who show the condition indicated by $y_n = 1$, i.e. cases, and one with subjects wihout the condition indicated by $y_n = 0$, i.e. controls.

2.2. Previous approaches in brief

Previous methods found in the literature also generate a binary vector $\mathbf{q}$, however, they differ in the way they make use of the training data $\{\mathbf{f}_n, y_n\}_{n=1,\dots,N}$ and new measurements $\mathbf{f}$, and yield a different meaning of $\mathbf{q}$. Conventional regression based analysis only uses $\{\mathbf{f}_n, y_n\}_{n=1,\dots,N}$ and builds a linear system of equations that aims to explain the measurements in terms of labels and other nuisance variables, such as age and gender. Inference extracts coefficients of the linear model and, based on the desired significance level, constructs the vector $\mathbf{q}$ from the coefficients. As the coefficients are in a weak sense population-averages, so is the final $\mathbf{q}$ and thus, the detected areas are not subject-specific and highlight areas commonly affected

[2] see http://www.fil.ion.ucl.ac.uk/spm/ or https://freesurfer.net

across the population. Furthermore, since these models need y for each subject involved in the analysis (in the estimation of the coefficients), they cannot be used prospectively.

Predictive models on the other hand, construct mappings $C : \mathbb{R}^d \to [0,1]$ that go from measurements to labels. The mapping function is computed using the training dataset and when analysing a new subject it takes into account the entire feature vector $\mathbf{f}$ and predicts y. During the construction of C, i.e. training, some methods identify measurements that are useful for accurate prediction and these measurements are highlighted in $\mathbf{q}$. This identification is based on the training dataset and therefore, the identified features are again in a weak sense population averages. Only predicted ys are subject-specific and not the identification of predictive measurements. Prediction of y provide little direct information regarding subject-specific structural alterations, however, in its computation this information is implicitly taken into account by the predictor. The proposed SubCMap method improves on this specific point.

Lastly, outlier detection methods use only the control group in the training examples, where $y_n = 0$, and build normative distributions for $\mathbf{f}$. When analyzing a new subject, the $\mathbf{f}$ is compared to the normative distributions and the less likely regions are identified. It is clear that in this approach the identified areas are not specific to the condition of interest and therefore, might not be suitable for obtaining condition-specific insights.

2.3. SubCMap

The intuition behind SubCMap is that if the condition of interest affects a given measurement consistently throughout the population and that measurement is affected for the subject under analysis, then when analysed with a predictive model, i.e. a classifier, that measurement will predict the presence of the condition for the subject. If the measurement is not affected, then it will predict the absence of the condition. Therefore, based on the prediction, it is possible to detect condition effect on the measurement. If a measurement is not affected consistently throughout the population, then even if it predicts the absence or presence of the condition, this prediction is not reliable and would not provide information about the condition effect. We build SubCMap based on this intuition. We explain the formulation in detail in the following.

The building block of SubCMap is based on element-wise predictions. When analyzing a given subject's data, each element in $\mathbf{f}$ is used to make independent probabilistic predictions on whether the subject shows the condition of interest or not. These element-wise predictions are seen as noisy observations of an underlying "true" condition-effect map and the construction of the detection map is formulated as a restoration problem. This approach clearly constructs subject-specific maps as the element-wise predictions are subject-specific. It can be applied prospectively since the label information is not used when the subject's data is analysed. Lastly, the detections are condition specific because the element-wise predictions are based on condition specific classifiers.

We denote the prediction at the measurement f_j with $p_j \in [0,1]$, which denotes the probability of the subject having the condition. A value of $p_j = 1$ would mean 100% belief that the subject has the condition, $p_j = 0$ means 100% belief that the subject does not have the condition and $p_j = 0.5$ means uncertainty. We use the vector notation $\mathbf{p} \in [0,1]^d$ to denote all the element-wise predictions for all measurements in $\mathbf{f}$. The mapping $f_j \to p_j$ is the probabilistic classification step and there are multiple alternative methods one can use to construct it. In this work we choose to use a simple approach, namely a Gaussian mixture model with equal mixing rates. The distribution of the measurement values are modeled with two Gaussians, one for cases and one for controls. The prior probabilities of the classes are assumed to be equal. The probabilistic prediction is defined as the posterior distribution:

$$p_j = p(y = 1|f_j) = \frac{\mathcal{N}(f_j; \mu_1^{(j)}, \sigma_1^{(j)})}{\sum_{g=0,1} \mathcal{N}(f_j; \mu_g^{(j)}, \sigma_g^{(j)})}, \ j = 1, \ldots, d, \tag{1}$$

where $\mu_g^{(j)}$ and $\sigma_g^{(j)}$ are the mean and standard deviations for the case ($g = 1$) and the control groups ($g = 0$) at the measurement j. The proposed method estimates both of these values empirically from the training dataset during the training phase. We note that it is also possible to build classifiers that do not vary across measurements and share hyper-parameters. The restoration formulation given in the following is agnostic to the classifier choice.

As described, SubCMap views $\mathbf{p}$ as a noisy observation and models it with a logit-normal distribution with iid noise

$$\hat{\Phi}_j \triangleq \log\left(\frac{p_j}{1-p_j}\right) = \Phi_j + \epsilon,\ \epsilon \sim \mathcal{N}(0,\sigma),$$

where ϵ is zero mean Gaussian noise with σ standard deviation. In this notation, $\mathbf{\Phi} = [\Phi_1, \ldots, \Phi_d]$ is the noise-free continuous effect map in the logit domain that the method aims to restore. To retrieve the effect-map in the probability domain one simply applies the sigmoid function $1/(1+\exp(-\Phi_j))$. Given the above observation model, we formulate the restoration process as the maximum-a-posteriori (MAP) estimation:

$$\mathbf{\Phi}^* \triangleq \arg_{\mathbf{\Phi}} \max p(\mathbf{\Phi}|\hat{\mathbf{\Phi}}) = \arg_{\mathbf{\Phi}} \max p(\hat{\mathbf{\Phi}}|\mathbf{\Phi})p(\mathbf{\Phi}),$$

where $p(\mathbf{\Phi})$ is the prior distribution. Further assuming conditional independence in the observational model we write the MAP estimation as

$$\mathbf{\Phi}^* = \arg_{\mathbf{\Phi}} \max \prod_{j=1}^{d} p(\hat{\Phi}_j|\Phi_j)p(\mathbf{\Phi}), \tag{2}$$

where $p(\hat{\Phi}_j|\Phi_j) = \exp\{-(\hat{\Phi}_j - \Phi_j)^2/2\sigma^2\}/\sqrt{2\pi\sigma^2}$.

The critical element in the MAP formulation is the prior distribution. We use a Markov Random Field (MRF) model to formulate the prior distribution with a unary and a pairwise term:

$$p(\mathbf{\Phi}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\sum_{j=1}^{d} U(\Phi_j|\theta_u) + \sum_{j=1}^{d}\sum_{k\in N(j)} V(\Phi_j, \Phi_k|\theta_v)\right)\right\}, \tag{3}$$

where θ_u and θ_v are the parameters of the unary and pairwise terms, Z is the normalization constant and $N(j)$ denotes the neighborhood of the j^{th} measurement. We take two points into consideration in defining these terms: unreliable predictions and consistency across neighboring measurements. The unary term tackles the issue with the predictions coming from measurements that do not show consistent condition effect throughout the population. For instance, this would happen when the distributions of this measurement in the case and the control groups overlap. When analyzing a new subject's input map, such measurements may give p_j values that differ from 0.5, i.e. predict the presence or the absence of the condition with some level of certainty. However, these predictions most likely do not correspond to a true condition effect since the condition effect on such measurements are not consistent across the population. Therefore, the noise-free version of these predictions should be closer to 0.5. In the logit domain this means Φ_j value should be closer to 0 regardless of $\hat{\Phi}_j$. To implement this, we need prior knowledge on the consistency of the condition effect. To this end, we use the prediction accuracy (also referred to as generalization error) at each measurement, which can be estimated on the training dataset using cross validation. Let us denote the estimated prediction error at the measurement f_j with $\eta_j \in [0, 0.5]$ [3]. Using η_j we define the unary term as $U(\Phi_j|\theta_u) \triangleq g(\eta_j)\Phi_j^2$. The function $g(\eta_j)$ should ensure that when the generalization error is low the restored value should be similar to the observation and when the error is high it should be close to 0. We examine the case where $V(\cdot,\cdot)$ is zero and $\sigma = 1$ to define $g(\cdot)$. In this case the restoration result equals

$$\Phi_j^* = \hat{\Phi}_j/(1+g(\eta_j)).$$

Defining $1/(1+g(\eta_j)) = (1-2\eta_j)^\alpha$ satisfies the conditions we described above, i.e. $\eta_j = 0.5 \implies \Phi_j^* = 0$ and $\eta_j = 0 \implies \Phi_j^* = \hat{\Phi}_j$. The function $(1-2\eta_j)^\alpha$ is a smooth transition between two extremes at $\eta_j = 0$ and 0.5, and the rate of decrease in that transition is controled with the α parameter. In Figure 1 we plot

[3] We consider only one condition in this article, which means there are two groups and the classification error should be at most 0.5, i.e. $\eta_j \in [0, 0.5]$. In practice one can observe larger errors. We map all errors larger than 0.5 to 0.5

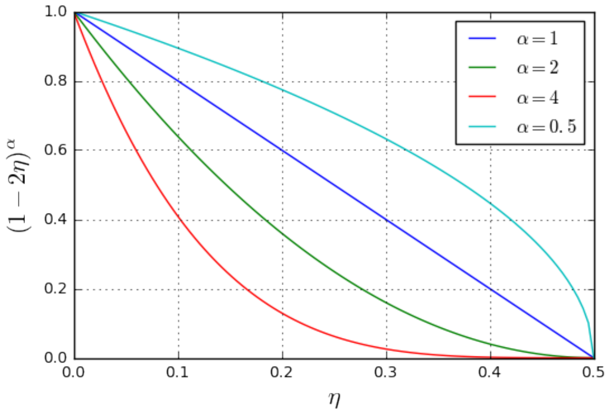


Figure 1: **Defining the unary term in the MRF.** The unary term penalizes high prediction error enforcing the corresponding Φ values towards zero, i.e. $\Phi_j = \hat{\Phi}_j(1-2\eta_j)^\alpha$

this function for different α values for demonstration. For higher α the function decreases faster which makes the model trust the observed $\hat{\Phi}$ less for the same prediction accuracy. In this work we empirically set $\alpha = 2$. Combining the above arguments, the final unary term becomes

$$U(\Phi_j|\theta_u) = \Phi_j^2 \frac{4\eta_j(1-\eta_j)}{(1-2\eta_j)^2}. \tag{4}$$

We note that this is not the only way to define the unary term and other functions can be considered.

The pairwise term enforces consistency across neighboring measurements when there is a spatial relationship between them. In most neuroimaging studies measurements are extracted from a dense set of anatomical locations. Commonly used cortical thickness maps and grey matter density maps are such examples. In these cases, measurements inherit a spatial relationship and a neighborhood structure based on the proximity of the anatomical locations they are extracted from. In the restoration model we assume consistency of the condition effect across neighboring measurements. To this end, we use the neighborhood relationship of the grid measurements are defined on, such as the surface mesh or the image grid. We define $N(j)$ in Equation 3 as the immediate neighbors of the j^{th} measurement. For surface maps $N(j)$ is the set of vertices that share a face with the vertex j and for volumetric maps it is the set of neighboring grid points based on a pre-defined image neihghborhood, such as 6 or 26 neighborhood in 3D grids.

Furthermore, the consistency requirement may vary with the location and should in theory respect the anatomical boundaries. To keep generality and implicitly take into account spatial heterogeneity, we model the consistency requirement using element-wise predictions in the training dataset. When analysing a new subject's input map, we enforce consistency between neighboring measurements if they give consistently similar predictions for the samples in the training dataset. We model the pairwise term as

$$\begin{aligned} V(\Phi_j, \Phi_k|\theta_v) &\triangleq \lambda \frac{(\Phi_j - \Phi_k)^2}{\varrho_{jk}}, \ k \in N(j) \\ \varrho_{jk} &\triangleq \mathbb{E}\left[(\Phi_j - \Phi_k)^2 | \Phi_j \Phi_k < 0\right], \end{aligned} \tag{5}$$

where λ is a trade-off parameter between the unary and the pairwise term, $N(j)$ is the neighborhood of the measurement j and the expectation is with respect to the conditional distribution $p(\Phi_j, \Phi_k|\Phi_j\Phi_k < 0)$. The ϱ_{jk} term is the expected disagreement between the j^{th} and k^{th} measurements when their predictions differ. The expectation is computed using the predictions on the training dataset as

$$\varrho_{jk} = \frac{1}{N_{\Phi_j\Phi_k<0}} \sum_{n\in\{\Phi_j\Phi_k<0\}} (\Phi_{n,j} - \Phi_{n,k})^2,$$

where we denote the set of training samples on which predictions at j^{th} and k^{th} measurements disagree with $\{\Phi_j \Phi_k < 0\}$ and its size with $N_{\Phi_j \Phi_k < 0}$. In Section 2.5, we provide further details on how to compute the predictions for the estimation of ϱ_{jk}. Low values of ϱ_{jk} indicate high prediction consistency across the neighboring measurements in the population and such a consistency is enforced when analysing a new image. High values indicate inconsistency in the population and consistency between neighbors are not enforced when analysing a new image. We note that there are various alternatives for defining the pairwise term. The advantages of the form given in Equation 5 are that it is agnostic to the type of measurements and the classifier, it does not require parcellation or anatomical segmentation and it can be directly applied to different types of maps as it only uses the neighborhood definition from the corresponding topology. We also would like to note that for measurements that do not bear a spatial relationship the pairwise term can simply be ignored.

Combining the observation model, the unary and the pairwise terms, and taking the logarithm yields the following optimization problem equivalent to the MAP estimation in Equation 2

$$\arg_{\mathbf{\Phi}} \min \sum_{j=1}^{d} \frac{(\Phi_j - \hat{\Phi}_j)^2}{2\sigma^2} + \frac{1}{2} \sum_{j=1}^{d} \Phi_j^2 \frac{4\eta_j(1-\eta_j)}{(1-2\eta_j)^2} + \frac{\lambda}{2} \sum_{j=1}^{d} \sum_{k \in N(j)} \frac{(\Phi_j - \Phi_k)^2}{\varrho_{jk}}$$

Taking the derivatives and setting to zero yields the following linear system of equations

$$\left(\sigma^2 \frac{4\eta_j(1-\eta_j)}{(1-2\eta_j)^2} + 1 \right) \Phi_j + \sigma^2 \lambda \sum_{k \in N(j)} \frac{\Phi_j - \Phi_k}{\varrho_{jk}} = \hat{\Phi}_j, \ j = 1, \ldots, d. \tag{6}$$

The solution of Equation 6 is the restored continuous effect map $\mathbf{\Phi}^*$. This system of equations can be solved efficiently even for large d when the neighborhood size $N(j)$ is small for all j using sparse matrix routines implemented in popular linear algebra packages, such as MATLAB and scipy.

2.4. *Thresholding* $\mathbf{\Phi}^*$

The MAP estimate $\mathbf{\Phi}^*$ is a continuous valued map. In order to determine the final binary effect map $\mathbf{q}$, we threshold $\mathbf{\Phi}^*$ with a global threshold,

$$q_j = \begin{cases} 0, & \Phi_j^* \leq \tau \\ 1, & \Phi_j^* > \tau \end{cases}$$

There are various alternatives for determining the threshold τ. Ideally, one would want to maximise detection accuracy on the samples in the training dataset. However, this approach would require ground truth for condition effects in these samples and such information is usually not available nor it is trivial to construct manually. In the absence of ground truth, we assume that the control group in the training dataset is composed of individuals who do not show any condition effect. In other words, $q_j = 0, \ \forall j$ for all control samples. Based on this assumption, we determine the threshold τ in order to limit the false-positive-rate (FPR) on the control group in the training dataset. Mathematically, we formulate this as

$$\tau = \min t, \text{ such that } \frac{1}{dN_{\{y=0\}}} \sum_{n \in \{y=0\}} \sum_{j} \delta(\Phi_{n,j}^* > t) \leq \tau_{\text{FPR}}, \tag{7}$$

where $\{y = 0\}$ denotes the set of control samples in the training dataset, $N_{\{y=0\}}$ its size and τ_{FPR} the desired FPR limit. Optimization given in Equation 7 is one dimensional and can be solved efficiently with golden section search algorithm. It aims to determine the minimum value of t that satisfies the τ_{FPR}, thus avoiding the trivial solution of $t = 1$.

The τ_{FPR} threshold limits the FPR over the entire set of measurements therefore, thresholding each element of $\mathbf{\Phi}^*$ with τ avoids the multiple comparisons problem. On the other hand, it yields a much more conservative threshold. Furthermore, assuming that subjects in the control group do not have any condition effect is quite strict and makes the threshold τ even more conservative. Nonetheless, in the lack of ground truth we opt for this conservative alternative for constructing the binary $\mathbf{q}$ map.

2.5. Estimating parameters with cross-validation and bootstrap

The parameters of the Gaussian mixture model, error rate estimates η_j, ϱ_{jk} of the pairwise term and the threshold τ are all estimated empirically using the training dataset. This section provides the details on how we perform these estimations. Given a training set, we estimate $\mu_{0,1}^{(j)}$ and $\sigma_{0,1}^{(j)}$ of the mixture model with the sample means and standard deviations of the respective groups. η_js are estimated as the average out-of-sample accuracy of bootstrapped samples (sampling with replacement). For each sample, a percentage of the training dataset is used to estimate the mixture model parameters and the remaining samples are used to compute a prediction accuracy. η_js are averages of many bootstrap samples (e.g. we used 10000 in our implementation) and accuracy for each measurement is estimated independently. In each bootstrap sample independent $\mu_{0,1}$ and $\sigma_{0,1}^{(j)}$ are estimated but for the final model the entire set of training samples is used for computing the mixture model parameters.

ϱ_{jk} is computed using the predictions on the training samples following the estimation of the parameters of the final mixture model. This computation could also be performed using a hold-out set or another layer of cross-validation, however, this results in longer execution times and in cases with limited data keeping a hold-out set might not be possible. Empirically, we observed that using predictions on a hold-out set instead of the training samples does not lead to substantial difference. It is worth noting that when using a complex classifier, overfitting is a danger and in this case, the difference in ϱ_{jk} can be substantial.

To estimate τ we implement a k-fold cross validation loop. For each fold, mixture model parameters and ϱ_{jk} is computed using the training portion and $\mathbf{\Phi}$ is computed for the control samples in the test portion. Completing the k-folds yields a $\mathbf{\Phi}$ prediction for each control sample in the training dataset. These predictions are used to compute a τ value based on the desired false positive rate limit. This procedure avoids contamination between estimation of τ and the other parameters.

Lastly, regarding the σ parameter, we observe in Equation 6 that σ is a multiplicative factor in front of the unary term $4\eta_j(1-\eta_j)/(1-2\eta_j)^2$ and λ. Therefore, its effect on the final $\mathbf{\Phi}$ will be through its interaction with these terms. In order to reduce the number of tuning parameters, in this work we set $\sigma = 1$.

2.6. Tuning parameters:

There are two tuning parameters of SubCMap, λ and τ_{FPR}. λ controls the strength of the consistency requirement between neighboring measurements. It is related to the smoothness of the final maps and higher values yield smoother detections. τ_{FPR} on the other hand, controls the amount of false positives a user is willing to accept in the final maps. Increase in the FPR limit yields higher number of detections at the expense of higher false detections.

2.7. Nuisance variables

The method so far does not take into account possible nuisance variables. In neuroimaging studies one often needs to account for such variables to reduce their undesired effects that may influence detections. In the proposed formulation these variables can be taken into account at the measurement-wise prediction level, in other words when computing $p(y=1|f_j)$. Different classifiers would require different strategies for the nuisance variables. In the case of the Gaussian mixture model used here, we formulate the effects of the nuisance variables with the following additive model

$$p(f_j) = p(g=1)p(f_j|\mu_1^{(j)} + \xi^T\beta^{(j)}, \sigma_1^{(j)}) + p(g=0)p(f_j|\mu_0^{(j)} + \xi^T\beta^{(j)}, \sigma_0^{(j)}), \tag{8}$$

where ξ is the vector that contains the nuisance variables appended with a one at the end for the intercept and $\beta^{(j)}$ is the vector of coefficients of the linear model. There are two points to note here. First, the $\beta^{(j)}$ vector varies across measurements. Second, the nuisance variables are subject-specific. As in the previous case, we use equal values for the prior class probabilities: $p(g=0) = p(g=1) = 0.5$. The updated class posterior distributions can then be computed as:

$$p_j = p(y=1|f_j,\xi) = \frac{\mathcal{N}(f_j;\mu_1^{(j)} + \xi^T\beta^{(j)}, \sigma_1^{(j)})}{\sum_{g=0,1}\mathcal{N}(f_j;\mu_g^{(j)} + \xi^T\beta^{(j)}, \sigma_g^{(j)})}$$

The only difference between the above equation and Equation 1 is that the means of the components have the extra additive term and the classification is based on the subject-specific ξ variable.

In this model, $\beta^{(j)}$ coefficients need to be estimated alongside $\mu_g^{(j)}$ and $\sigma_g^{(j)}$ during training. To this end, we maximise the data log likelihood on the training dataset for each measurement independently. In the absence of nuisance variables, this corresponds to estimating $\mu_g^{(j)}$ and $\sigma_g^{(j)}$ with sample mean and standard deviation. In their presence the maximization is written as:

$$\max \mathcal{L}^{(j)} = \max \left\{ - \sum_{g=0,1} \sum_{n=1}^{N_g} \frac{\left[f_{n,j} - \mu_g^{(j)} - \xi_n^T \beta^{(j)} \right]^2}{2(\sigma_g^{(j)})^2} - \ln(\sigma_g^{(j)}) N_g \right\}, \tag{9}$$

where N_g denotes the number of samples in the group g and the second summation is over the samples in that group. We are not aware of an analytical solution to this maximization problem hence, we resort to numerical optimization. Specifically, we use an iterative optimization scheme where at each iteration Newton steps are taken for $\beta^{(j)}$, and $\mu_g^{(j)}$ and $\sigma_g^{(j)}$ are determined analytically for the updated $\beta^{(j)}$ vector. Algorithm 1 summarizes the numerical approach. In our experiments, we observed empirically that the optimization routine given in Algorithm 1 converges in just a few iterations. The partial derivatives of the log likelihood and the Hessian matrix required for the Newton steps are as follows:

```
**Initialization**: $\beta^{(j,0)} = 0$,
$\mu_g^{(j,0)} = 1/N_g \sum_{n=1}^{N_g} f_{n,j}$ and
$(\sigma_g^{(j,0)})^2 = 1/N_g \sum_{n=1}^{N_g} (f_{n,j} - \mu_g^{(j,0)})^2$ for $g = 0, 1$;
**while** $|\beta^{(j,i-1)} - \beta^{(j,i-2)}| < 10^{-5}$ **do**
    $\beta^{(j,i)} = \beta^{(j,i-1)} + \gamma \mathbf{H}^{(j,i)} \frac{\partial \mathcal{L}^{(j,i-1)}}{\partial \beta^{(j,i-1)}}$;
    $\mu_g^{(j,i)} = \frac{1}{N_g} \sum_{n=1}^{N_g} f_{n,j} - \xi_n^T \beta^{(j,i)}, \ g = 0, 1$;
    $(\sigma_G^{(j,i)})^2 = \frac{1}{N_g} \sum_{n=1}^{N_g} \left( f_{n,j} - \mu_g^{(j,i)} - \xi_n^T \beta^{(j,i)} \right)^2, \ g = 0, 1$;
    i = i +1;
**end**
```

Algorithm 1: Maximization of the log likelihood given in Equation 9. The step size parameter γ is set as 0.25 for all the experiments.

$$\begin{aligned} \frac{\partial \mathcal{L}^{(j)}}{\partial \beta^{(j)}} &= \sum_{g=0,1} \frac{1}{(\sigma_g^{(j)})^2} \sum_{n=1}^{N_g} \left(f_{n,j} - \mu_g^{(j)} - \xi_n^T \beta^{(j)} \right) \\ \mathbf{H}^{(j)} &= - \sum_{g=0,1} \frac{1}{(\sigma_g^{(j)})^2} \sum_{n=1}^{N_g} \xi_n \xi_n^T \end{aligned}$$

3. Experiments

We evaluated SubCMap using synthetically generated data and a cohort of 290 subjects selected from the ADNI dataset. In the experiments with the synthetic dataset, we performed quantitative analysis assessing the detection accuracy of the proposed method and compared it with outlier detection and element-wise prediction without restoration. In the experiments with the ADNI cohort, due to the lack of ground truth information on the condition effect, we performed indirect evaluation and performed the same comparisons.

In both experiments measurements have an underlying spatial structure. In the case of synthetic data, measurements form an image and in the case of the ADNI cohort, we used cortical thickness maps extracted using the Freesurfer software. We first provide details for the experiments with synthetic data and then discuss the results on the ADNI cohort.

3.1. Synthetic Data

In order to evaluate the detection accuracy of SubCMap, we generate a synthetic dataset where the ground truth information for the condition effect is available for each case. We generate measurements for 200 cases, where 100 of the them belong to the group with no condition effect, i.e. control group, and the other 100 to the group with condition effect, i.e. case group. We generate an image of size 100 × 100 pixels for each case and the pixel intensities are taken as the measurements. Images for the control group contained only stationary noise with spatial covariance and no condition effect. To generate these images, we assigned samples from iid Gaussian noise with zero mean and $\sigma_n = 50$ standard deviation to each pixel, and convolved the image with a Gaussian kernel with standard deviation 2.5 pixels. The convolution yields correlation between the measurements at neighboring pixels. Example images for the control group can be seen in the top row of Figure 3.

We model the condition effect as an additive factor on top of the stationary noise. We use two types of effects for the condition in order to introduce subject level variation. The two types of effects are shown in the first row of Figure 2. For both types, the condition only affects a small set of measurements in the image, which are indicated in white. The black pixels indicate no condition effect for the corresponding measurements. The two types of effects share the central square but differ in the squares at the corners. We generate the images in the case group by first constructing a noisy image similar to the control group and then adding a constant value to the pixels indicated as white in the condition effect images. The case group consists of 50 images per effect type, 100 images in total. Examples images can be seen in the top row of Figure 3. The constant value that is added to introduce condition effect is the effect size. We experiment with different effect sizes relative to σ_n and will present results for $0.6\sigma_n, \sigma_n, 1.4\sigma_n, 2\sigma_n$ and $3\sigma_n$. We note that different variations of the effect type can also be generated. We chose to generate such a large variation for the sake of illustration.

We first applied regression analysis (GLM), Random Forest (RF) and Support Vector Machines (SVMs) to identify population-wide condition affected areas in the images. In GLM analysis we extract p-value maps and correct for multiple comparisons using Bonferroni's method [31]. The GLM analysis makes use of the entire dataset. Similarly, we trained both RF and SVM using the entire dataset and compute the feature importance measures (Gini's criteria [32]) for RF and the weights for SVM using the scikit-learn package [33]. The RF importance values and SVM weights were then converted to p-values with permutation testing similar to other works in the literature [18, 34]. We threshold the p-values at the 0.05 level. The population-wide detections of GLM, RF and SVM are shown in Figure 2. The p-values for the GLM correspond to the p-values based on Bonferroni correction. These results clearly show that detections of these methods are population averages and do not provide subject-level information nor information on the effect type variation. SubCMap on the other hand, reveals subject-level information as we demonstrate next.

We tested SubCMap using 10 randomly shuffled 5-fold cross validation (CV) experiments. During each fold the parameters of the method were estimated with a separate, inner 5-fold CV loop as described in Section 2. We use a 4-neighborhood on the image grid to define the pairwise term. Before we provide detailed quantitative analysis, in Figure 3 we show six example images from these experiments. The top row shows the images themselves, the second row shows the ground truth affected areas and the third and fourth rows show the outputs of the proposed method, i.e. subject-specific detections. We show in the third row the continuous valued $\mathbf{\Phi}^*$ and the final detections, i.e. $\mathbf{q}$ maps, in the fourth. We observe that SubCMap clearly captures subject-level information and successfully identifies the affected areas. Comparing these detections with population-wide results demonstrates the added value of the proposed method. The effect-size for these illustrative examples were $1.4\sigma_n$ and we used $\lambda = 2$ and $\tau_{\text{FPR}} = 0.01$.

For comparisons, we also applied outlier detection to identify subject-specific affected areas. We only used the control subjects in the training set of each fold to estimate a normative Gaussian distribution and used the normative distributions to compute likelihood values at each measurement for the test images. We thresholded the likelihood values to obtain the final detections. We used the same procedure as in Section 2 to determine the threshold. The fifth and sixth rows in Figure 3 show the outlier detection outputs before and after thresholding, respectively. For illustration we show 1 - likelihood maps in the fifth row. We observe

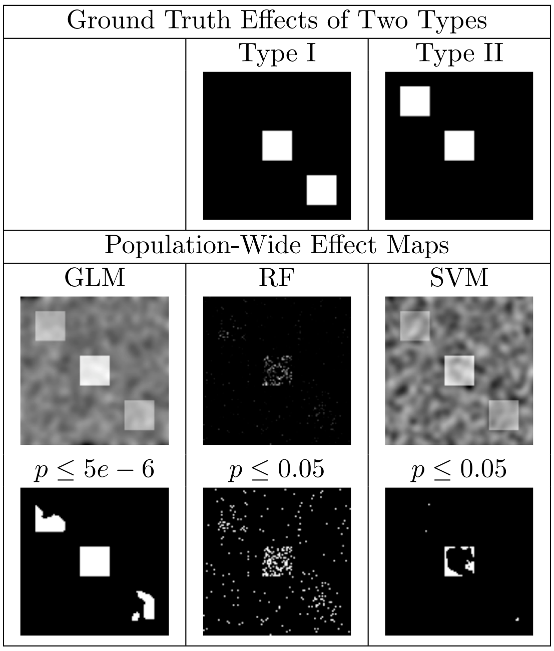


Figure 2: **Population-wide detection of condition effect with different methods.** Top row shows the ground truth for the two types of condition effects generated in the synthetic dataset. The middle row shows results of regression analysis with GLM and detection results with Random Forests (RF) and Suppport Vector Machines (SVM). The bottom row shows p-value maps corresponding to the results in the middle row thresholded at 0.05 level. The GLM map is further corrected for multiple comparisons problem using Bonferroni's method. The p-value threshold indicated on top of the GLM p-value map corresponds to the corrected value. p-value maps for RF and SVM are obtained via permutation testing but no correction is applied. It is clear that these methods provide population-wide information on the condition effect but do not provide subject-specific information nor can they provide variation in effect-type.

that the outlier detection does not achieve similar detection accuracy visually as the proposed method. This is not surprising since outlier detection is not condition-specific therefore, yields high number of false positives for all images. The thresholding procedure sets a high threshold to limit the FPR and results in low sensitivity in the final detections. The images shown in the fifth row are in accordance visually with the results of state-of-the-art outlier detection methods in the literature [25, 22].

Finally, we also provide the detection results of an alternative method that only uses element-wise predictions ignoring the restoration step of the proposed SubCMap. The last two rows in Figure 3 show the element-wise predictions and its thresholded version. The threshold, once again, is determined using the same procedure as explained in Section 2. Comparing these detections with third and fourth rows clearly demonstrates the need for the restoration formulation.

In addition to visual results, we also performed quantitative analysis. We used Dice's Similarity Coefficient (DSC) and FPR to evaluate the accuracy of subject-specific detections. We repeated the experiment for different λ values, FPR thresholds 0.01 and 0.001, and the five different effect sizes: $0.60\sigma_n$, $1.0\sigma_n$, $1.40\sigma_n$, $2.0\sigma_n$ and $3.0\sigma_n$. Graphs in Figure 4 plots the DSC and FPR for different λ values obtained with the proposed method. Both DSC and FPR are averaged over different random CV experiments. The solid lines in the graphs show the respective scores obtained by SubCMap. The dashed lines are the results of the alternative method that uses element-wise predictions as the final detections ignoring the restoration step. There are several points to make here.

- We observe that the restoration improves DSC substantially compared to element-wise predictions.
- As expected, the detection accuracy is higher for higher effect-sizes.
- Increasing λ seems to improve the results for low effect-sizes, however, it can also reduce DSC for larger effect-sizes. This is not surprising because increasing the strength of consistency in the restoration

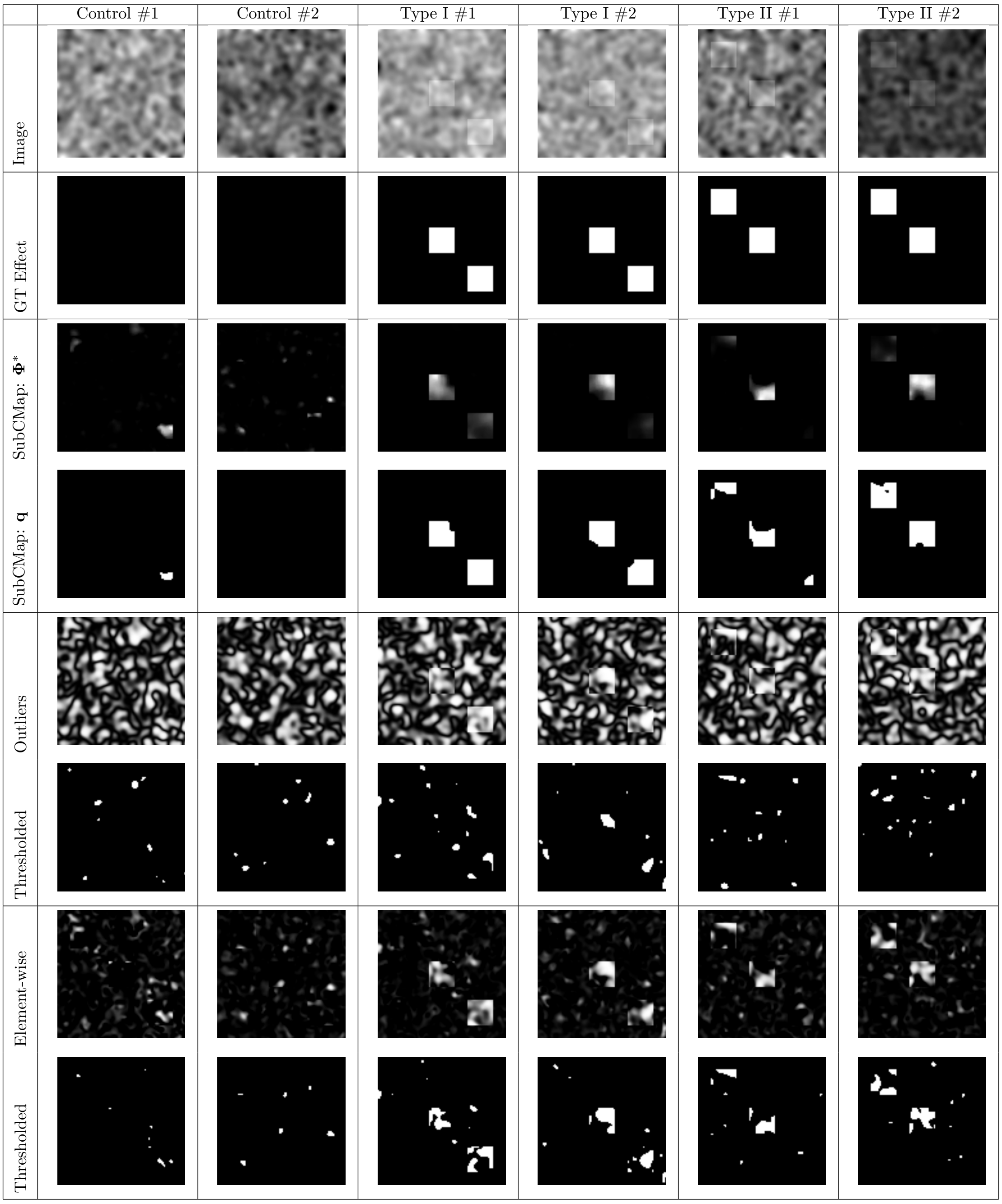


Figure 3: **Visual results for subject-specific detections in the synthetic dataset.** Six examples are shown: 2 controls, and 4 cases including 2 of each condition effect type. Top row shows the original subject images. Second row shows the ground truth condition effects. Third and fourth rows show the condition effect detections with SubCMap. Fifth and sixth rows show outlier detection scores and thresholded detections respectively. Last two rows show element-wise predictions and its thresholded version. We note that SubCMap detections achieve a much higher accuracy than the outlier detection and element-wise predictions. In these examples the noise standard deviation is $\sigma_n = 50$ and the effect size is $1.4\sigma_n = 70$. The pairwise term coefficient is chosen arbitrarily as $\lambda = 2$ and the FPR limit is set to $\tau_{\text{FPR}} = 0.01$.

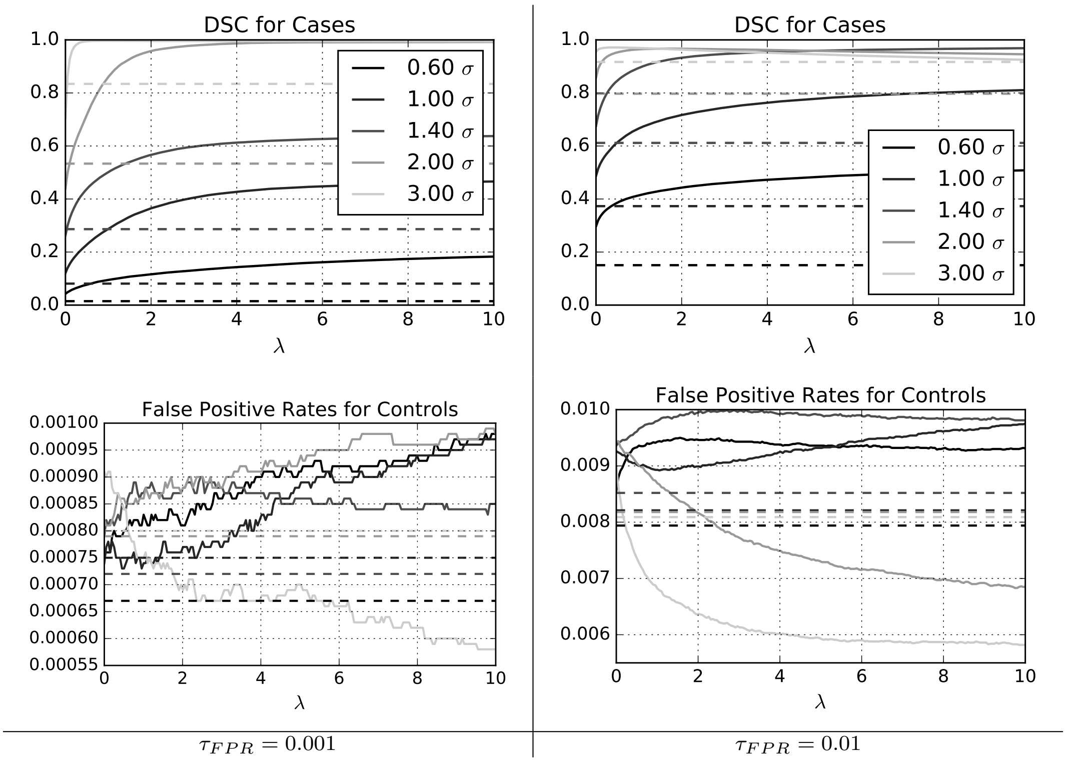


Figure 4: **Quantitative results on the synthetic dataset with SubCMap.** Graphs plot DSC and FPR obtained with the proposed method, SubCMap, on the synthetic dataset. We experiment with different λ values, shown in the x-axis, effect sizes, shown in the legend, and τ_{FPR}, shown in different columns. The solid curves are the results of SubCMap and the dashed curves are the results of detection via element-wise predictions without restoration.

increases the sizes of the detected areas and can reduce specificity for high effect sizes. For low effect sizes the improvement in sensitivity still dominates and yields higher DSC.

- DSC for lower effect-sizes are higher for higher τ_{FPR} but the reverse is observed for higher effect-sizes. Increasing τ_{FPR} allows higher number of false positives in addition to detecting effects that are subtle. As a result, for low effect sizes the detection accuracy increases but for higher effect sizes the decreasing specificity decreases the DSC.
- False positive rates stay around τ_{FPR} suggesting the proposed method to determine the threshold works well.

In Figure 5, the graphs plot the same results but obtained with outlier detection. For fairness, we also implemented a consistency criterion for the outlier detection method. We apply a restoration to the outlier detection results with an MRF model that only has a pairwise term. The pairwise term is defined as the Kullback-Leibler divergence between the normative distributions at neighboring measurements. The same neighborhood structure is used for both SubCMap and the outlier detection. As in the previous case, we tested different λ values for the pairwise term. The solid lines are the results of the restored outlier detection maps and the dashed lines are the ones without restoration. Comparing Figures 5 and 4 clearly shows that SubCMap achieves a much higher accuracy than the outlier detection method. Even a similar restoration does not increase the outlier detection accuracy to the level of the proposed method. Once again, this is not surprising since outlier detection is not condition-specific and cannot be expected to achieve similar accuracy as SubCMap.

Finally, we evaluated SubCMap in the presence of nuisance parameters using the extension presented in Section 2.7. We generated another synthetic dataset that included the effect of a nuisance parameter η

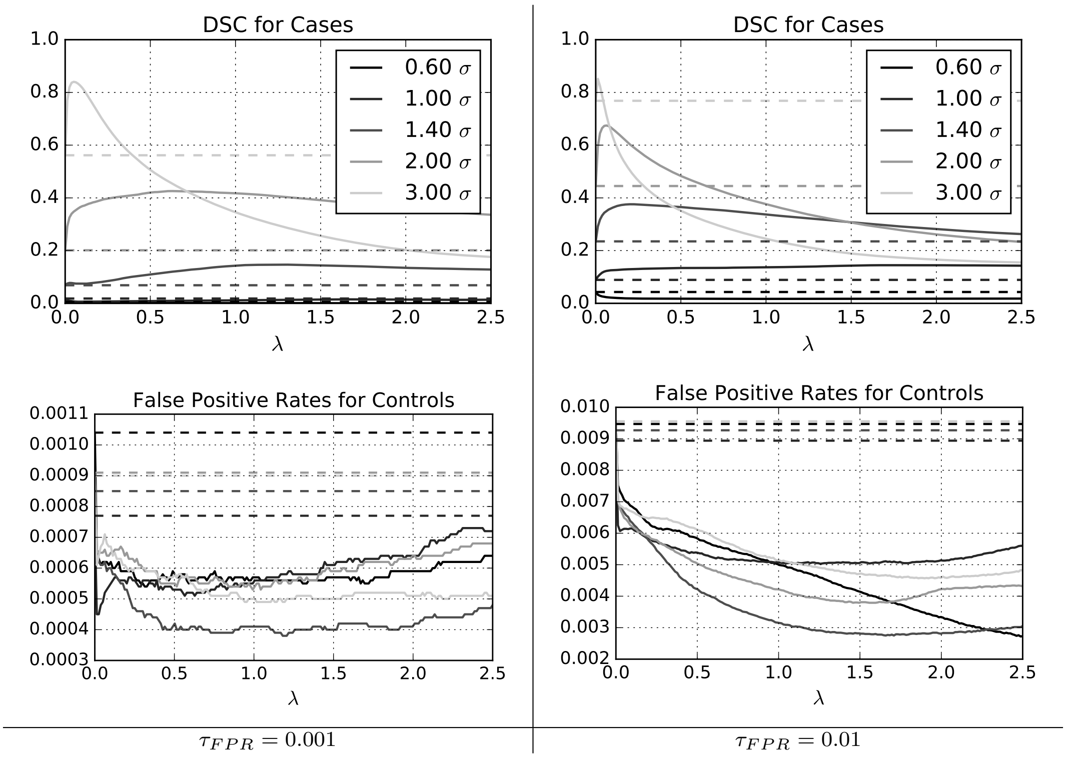


Figure 5: **Quantitative results on the synthetic dataset with outlier detection.** Graphs plot DSC and FPR outlier detection obtained on the synthetic dataset. The dashed curves are the results of pure outlier detection and the solid curves are the results after restoring the outlier detection with an MRF that enforces consistency of detection in neighboring measurements.

that was chosen randomly for each subject from a Gaussian distribution with standard deviation 10, i.e. $\mathcal{N}(0, 10)$. The effect of the nuisance parameter on a measurement depends on the $\beta^{(j)}$ value. We modeled the $\beta^{(j)}$ variable to vary across measurements smoothly. We assigned a random draw from a Gaussian distribution with standard deviation 5 to each pixel and then blurred the resulting image with a Gaussian kernel with standard deviation 2.5. This results in a smooth $\beta^{(j)}$ image. The images in the synthetic dataset were first created using the method explained above and then the factor $\beta^{(j)}\eta_n$ is added to each pixel of each image to introduce the effect of the nuisance parameter. The condition effects were identical to the previous experiment. Graphs in Figure 6 plot DSC and FPR for the detection results in the dataset with nuisance parameter. DSC in Figures 6 and 4 are very similar. The only difference is the slightly higher false positive rates, which is due to the added complexity caused by the nuisance parameter. However, even in this case the FPR remain close to τ_{FPR}, the desired limit.

Experiments on synthetically generated data allowed us to quantitatively analyse the detection accuracy of SubCMap and the results demonstrate substantial advantages over outlier detection, which is the alternative approach in the literature. In the next section, we demonstrate the proposed method on real data.

3.2. ADNI Dataset

In the second experiment we applied SubCMap to analyse the effects of Alzheimer's Disease (AD) on the cortical thickness to detect subject-specific atrophy patterns. We selected a cohort of 290 subjects from the ADNI database, which consists of 145 patients with AD diagnosis and 145 age and gender matched controls. We used the structural T1-weighted magnetic resonance images of these subjects and extracted cortical thickness maps using the Freesurfer software package [12]. Maps consists of gray matter thickness

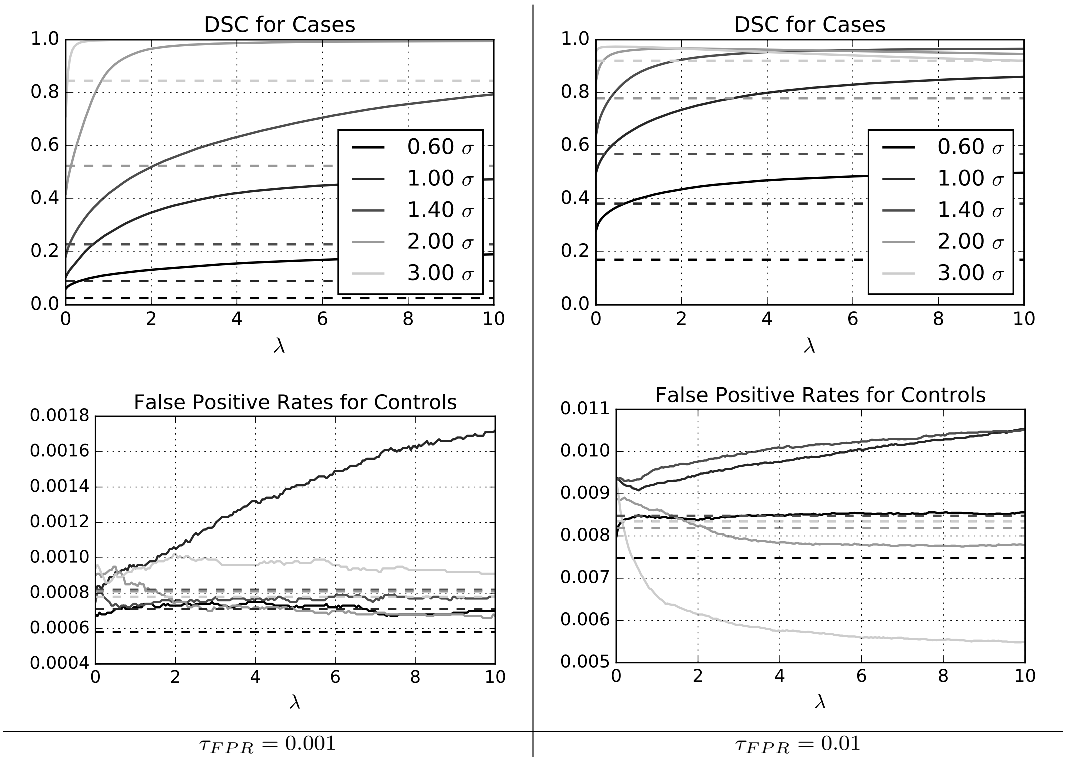


Figure 6: **Quantitative results on the synthetic dataset with nuisance parameters.** Graphs plot DSC and FPR SubCMap achieved on the synthetic dataset that included the effect of a nuisance parameter. The plots are similar to those in Figure 4 showing the method to take into account the nuisance parameters works well.

values extracted across the entire cortical mantle and discretized as a triangular surface mesh. Each vertex holds the thickness value of the underlying cortical gray matter. We aligned the cortical thickness maps of all individuals on a common reference surface mesh defined on the MNI atlas and decimated the number of vertices to 10242 using the Freesurfer for faster computation.

We performed 20 randomly shuffled 5-fold CV experiments. As in the previous section, for each fold we estimated the parameters of the proposed method using an independent, inner 5-fold CV loop. For the tuning parameters, we tested different values of $\lambda = \{0, 0.5, 1, 5\}$ and two different FPR limits $\tau_{\text{FPR}} = \{0.001, 0.01\}$. In order to define the pairwise term in the restoration formula we used the triangular surface mesh. We defined the neighbors of a vertex as the set of vertices that shared a mesh triangle with it.

The ground truth for areas affected by AD is unfortunately not available for in-vivo data. Therefore, we resort to indirect evaluation strategies. Our main hypothesis is that if the detected regions are accurate then they should contain condition related information and be statistically related to other auxillary markers of the condition. In our evaluation, we used the Mini Mental State Examination (MMSE) scores and Cerebrospinal Fluid Amyloid-β (CSF a-β) levels as these auxillary markers. We detected AD affected measurements for each subject using SubCMap and computed the number of detected measurements. In the case of cortical surface maps, since each measurement corresponds to a vertex, the number of measurements roughly corresponds to the area of the affected region on the cortical mantle. We quantified the detection quality using the Pearson's correlation coefficient between the number of detected measurements and the auxillary markers. The MMSE scores were available for all subjects in our cohort while only 147 subjects had CSF a-β measurements. There are two points we would like to note. First, we chose to use the number of measurements, or alternatively the area of the affected region, because it is conceptually a very straightforward statistic quantifying the disease load. Other statistics could also be used for the quantification.

		MMSE	CSF a-β	AUC	ACC
RF on Cortical Thickness		0.587 ± 0.019	0.303 ± 0.035	0.906 ± 0.005	0.828 ± 0.008
SubCMap $\mathbf{q}$ @ $\tau_{\text{FPR}} = 0.001$	$\lambda = 0$	-0.526 ± 0.008	-0.308 ± 0.010	0.877 ± 0.009	0.825 ± 0.009
	$\lambda = 0.5$	-0.497 ± 0.010	-0.293 ± 0.011	0.868 ± 0.008	0.808 ± 0.009
	$\lambda = 1$	-0.483 ± 0.010	-0.287 ± 0.013	0.866 ± 0.010	0.806 ± 0.006
	$\lambda = 5$	-0.439 ± 0.011	-0.261 ± 0.013	0.861 ± 0.010	0.797 ± 0.006
SubCMap $\mathbf{q}$ @ $\tau_{\text{FPR}} = 0.01$	$\lambda = 0$	-0.546 ± 0.006	-0.367 ± 0.008	0.900 ± 0.007	0.832 ± 0.012
	$\lambda = 0.5$	-0.532 ± 0.005	-0.355 ± 0.008	0.902 ± 0.005	0.833 ± 0.009
	$\lambda = 1$	-0.528 ± 0.006	-0.349 ± 0.008	0.900 ± 0.006	0.827 ± 0.011
	$\lambda = 5$	-0.526 ± 0.006	-0.349 ± 0.009	0.897 ± 0.007	0.830 ± 0.008
Element-wise @ $\tau_{\text{FPR}} = 0.001$		-0.424 ± 0.010	-0.201 ± 0.007	0.809 ± 0.018	0.748 ± 0.014
Element-wise @ $\tau_{\text{FPR}} = 0.01$		-0.490 ± 0.004	-0.300 ± 0.006	0.886 ± 0.006	0.821 ± 0.011
Outlier Detection @ $\tau_{\text{FPR}} = 0.001$		-0.149 ± 0.004	-0.013 ± 0.012	0.602 ± 0.023	0.570 ± 0.027
Outlier Detection @ $\tau_{\text{FPR}} = 0.01$		-0.287 ± 0.002	-0.138 ± 0.006	0.796 ± 0.014	0.738 ± 0.013

Table 1: **Quantitative evaluation of subject-specific detections on the ADNI database.** First two columns show correlation scores between the areas of detection regions on the cortical surface maps and auxillary markers MMSE and CSF a-β for different methods. At the very top we show baseline correlation scores for a dedicated RF algorithm trained with cortical thickness values directly. Maps constructed by SubCMap, $\mathbf{q}$, reach very similar correlation scores with the baseline for both MMSE and CSF a-β. The last two columns show classification results of RF classifier trained on cortical thickness value directly (top row) and binary detection maps extracted for each algorithm (rest of the rows). We observe that RF is able to achieve similar AUC and ACC using the binary maps SubCMap constructs.

Second, we chose to use MMSE and CSF a-β as the auxillary markers instead of say hippocampal volume because they are not extracted from the neuroimages, therefore, the correlations with the detected areas are less likely to be contaminated by spurious dependencies.

We performed the same comparative study as in the synthetic data experiments. Specifically, we used outlier detection to identify subject-specific abnormal regions and element-wise predictions for detection without the restoration procedure. We evaluated both of these alternative techniques using the same correlation coefficients between the number of detected measurements and the auxillary markers. Furthermore, in order to provide a baseline for the correlation values, we also trained separate dedicated Random Forest (RF) regressors that take aligned cortical thickness maps as input and predict MMSE and CSF a-β values. Similar to detection experiments, we performed 20 randomly shuffled 5-fold CV experiments to quantify the accuracy of RF-based regression.

Table 1 presents the results for MMSE and CSF a-β correlations in the first two columns. The correlation scores are averaged over 20 randomly shuffled CV experiments. We present the average values as well as the standard deviations. The top row presents the baseline results obtained by applying the RF regressor directly to the cortical thickness maps. The proposed method is indicated with $\mathbf{q}$ for different τ_{FPR}. We observe that for correlations with MMSE the proposed method reaches values that are close to the RF baseline in absolute value *using only the areas of the detected regions.* The correlations themselves are negative since we expect lower MMSE and larger affected areas for individuals with larger disease load. The correlation scores are higher for $\tau_{\text{FPR}} = 0.01$ than 0.001 suggesting the latter might be a conservative limit. The correlation strengths decrease slightly with increasing λ. We believe this is due to the inherent smoothness of cortical thickness maps that arises from registration and mesh decimation. Additional smoothness with increasing λ seem to have an adverse effect for the correlation scores with MMSE. Element-wise prediction without restoration yields lower correlation scores than those of the proposed method. This demonstrates the additional benefit of the restoration formulation and in particular the unary term. Lastly, outlier detection yields much weaker correlations than SubCMap, which demonstrates the value of having a condition specific method.

Correlation scores for CSF a-β show a similar behavior as MMSE. The biggest difference is that SubCMap for $\tau_{\text{FPR}} = 0.01$ reaches stronger correlations with the marker than the baseline. This result is quite surprising but it is stable across the different random experiments as can be seen from the distributions of the correlation scores. The correlation scores with MMSE and CSF a-β demonstrates:

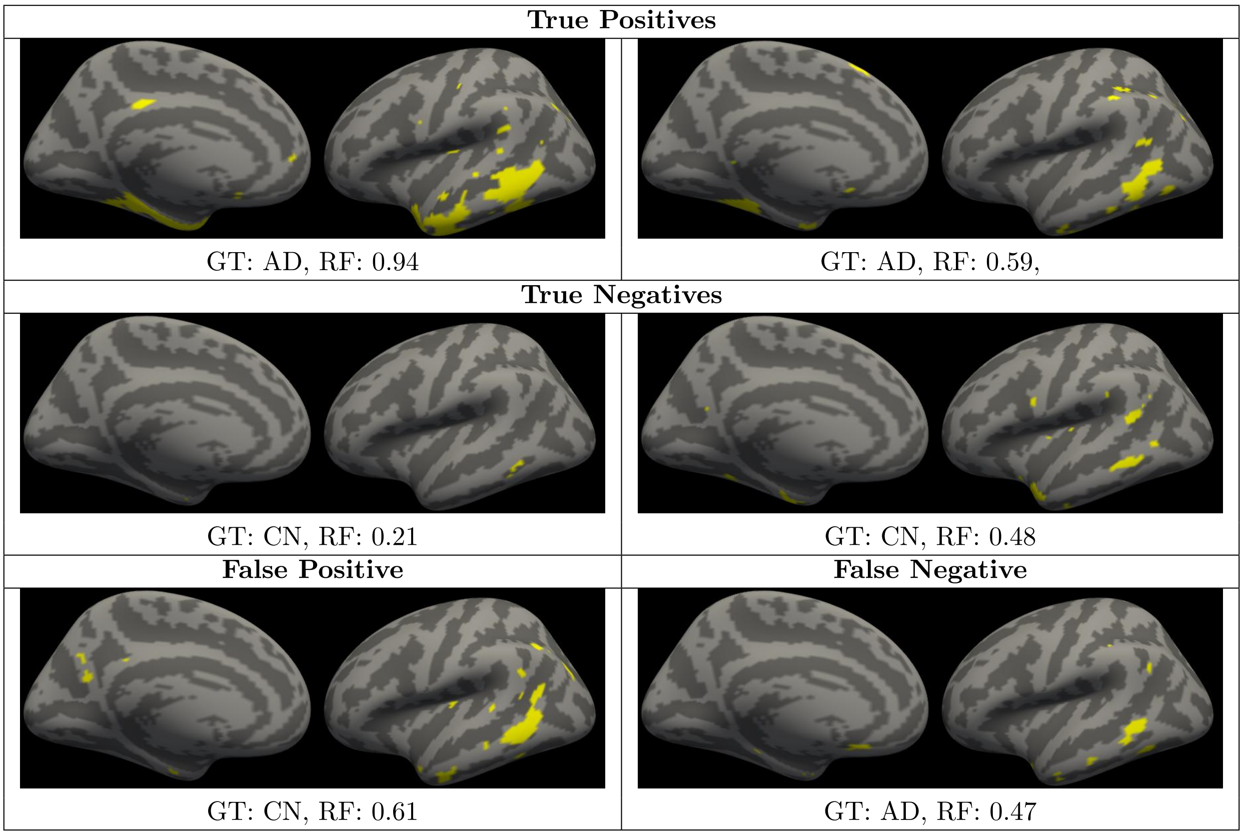
- The regions detected by SubCMap hold condition related information.
- Similarity with the baseline correlation scores suggest that the proposed method successfully captures inter subject variations of disease affected areas.
- Higher correlation strength compared to the element-wise method without restoration suggest that the restoration procedure increases the detection rate of the affected areas.
- Higher correlation strength compared to outlier detection highlights the value of SubCMap for studying specific conditions.

Lastly, the proposed method is univariate in its essence. Element-wise predictions do not consider other measurements and the detections are based on these predictions. As a result, SubCMap has the risk of missing measurements that the condition affects via multivariate interactions with other measurements. In order to determine the limitation due to the univariate nature, we performed RF based classification using cortical thickness maps, similar to [35], and in the same fashion using the binary map $\mathbf{q}$ that assigns 1 to detected measurements and 0 to everything else. If SubCMap misses substantial amount of measurements due to not modeling multivariate effects, then the classification algorithm would not be able to accurately classify subjects into control and case groups using the binary $\mathbf{q}$ maps. This can be detected as a drop in the prediction accuracy when compared to classification via the measurements directly. To assess this, we compared prediction accuracy of RF when using thickness maps and detections computed with SubCMap as well as the alternatives. We use area under the curve (AUC) and accuracy scores (ACC) for the comparison. Again, we performed 20 randomly shuffled 5-fold CV experiments and quantified ACC and AUC for outlier detection and element-wise prediction without restoration as well.

We present the results of the classification study in the last two columns of Table 1. The AUC and ACC values are averaged over 20 randomly shuffled CV experiments and we present both the mean values and the standard deviations. AUC and ACC values suggest that SubCMap might not be suffering substantially from not modeling multivariate effects. Classification results using the detections with the proposed method are very close to the results using the measurements directly. We observe a slight decrease for $\tau_{\text{FPR}} = 0.001$, which suggests once again that this limit might be too conservative. Comparing the classification results obtained using detections of the element-wise model without restoration, we observe a very slight decrease both in AUC and ACC. The bigger difference, however, is with the outlier detection method. Outlier detection yields a substantially lower AUC and ACC because it is not condition specific.

In addition to the quantitative results we also provide visual results of subject-specific detection maps. In Figure 7 we show the detection maps for six subjects from the ADNI dataset. These detections were extracted from one of the CV experiments with $\lambda = 1$ (arbitrarily chosen) and $\tau_{\text{FPR}} = 0.01$. Three of these examples are from subjects diagnosed with AD and the remaining three are from control subjects. The group assignments are indicated below each image with the tag GT, where CN denotes control group. For all these subjects we also computed RF classification probabilities, i.e. probability of being a patient, based on the cortical thickness values. These probabilities are indicated with the tag RF below each image. Comparing the RF probabilities and the ground truth diagnosis the examples fall into four categories: **true positives**, where RF probability is higher than 0.5 and the subject is an AD patient, **true negative**, RF probabilit is lower than 0.5 and the subject is a control, **false positive**, RF probability is higher than 0.5 but the subject is a control, and **false negative**, RF probability is lower than 0.5 but the subject is an AD patient. The relationship between the RF probabilities and the detected areas is noteworthy. We remark the following for the maps given in Figure 7:

- Higher RF probability is related to larger detected areas, in particular around the hippocampal formation (entorhinal cortex) and medial temporal lobe. Among true positives, the subject with larger detections also gets a larger RF probability. The same situation is also present among true negatives.
- SubCMap detects a large affected area in the medial temporal lobe for the false positive example while no effect around the entorhinal cortex. RF classifier most likely assigns a high probability due to the large condition effect in the medial temporal lobe. SubCMap is able to visualize these effects helping interpreting the classifier as well as facilitate a more informed assessment of the condition effect.



True Positives

GT: AD, RF: 0.94

GT: AD, RF: 0.59,

True Negatives

GT: CN, RF: 0.21

GT: CN, RF: 0.48

False Positive

GT: CN, RF: 0.61

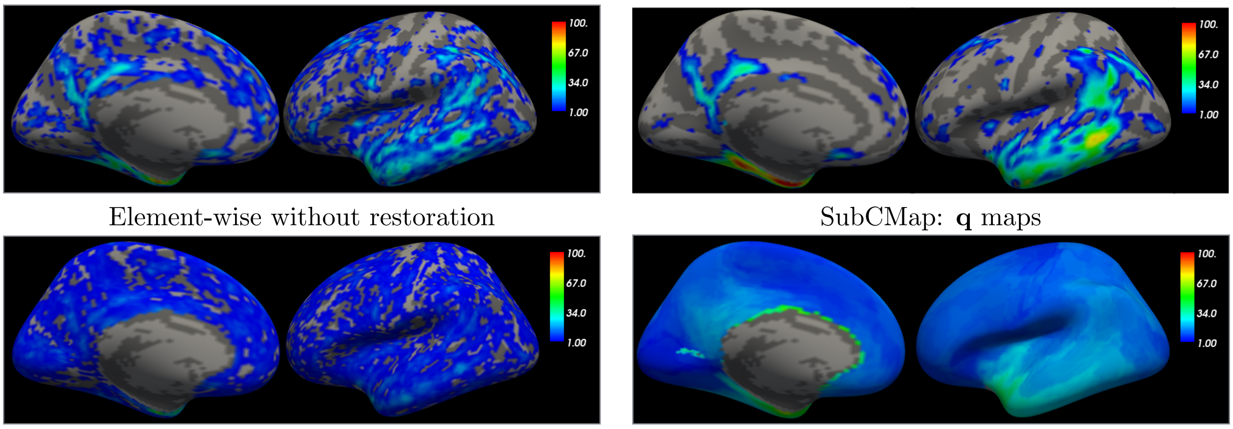
False Negative

GT: AD, RF: 0.47

Figure 7: **Examples of subject-specific maps for AD effect.** Yellow areas show regions detected by SubCMap for Alzheimer's effect. The subjects are independently analysed with a Random Forest algorithm trained on cortical thickness values. The RF based probabilities for AD is given below each image alongside the ground truth label. The extend of the detected areas shown in the subject specific maps correlate well with the RF based probabilities. The last two rows show false positive and false negative examples where the RF prediction contradicts with the ground truth label. In subject specific maps we observe that the extent of the detected areas explain the RF-based probabilities and the RF's decision becomes clearer. Such interpretation is made possible with the proposed method and can be applied to analyze the results of any machine learning technique.

- The false negative examples shows very small condition load and RF probabilities is consistent with this. The diagnosis, however, is AD showing the inconsistency between the detections, RF classifier and the ground truth. The surprising part is the difference between false positive and false negative examples. In the false positive example the area of the affected region is much larger than the false negative example. RF probabilities reflect this as well, however, the ground truth diagnosis is contradicting the detections and RF. The proposed method allows us to analyze and visualize such disagreements between black-box machine learning algorithms and ground truth labels.

Being able to extract subject-specific maps also allows us to construct population level condition effect maps similar to conventional regression analysis. However, the main difference is that the values we can assign to each measurement becomes more interpretable. Figure 8 shows frequency maps of AD effect. At each vertex we show the number of patients in the cohort that shows AD effect at the corresponding thickness measurement. We compute these maps simply by adding the subject-specific maps. Figure 8 shows results for element-wise prediction without restoration, the proposed method and outlier detection with and without MRF restoration. For all the methods we used $\tau_{\text{FPR}} = 0.01$, for the proposed method we used $\lambda = 1$ and for outlier detection with MRF restoration we used $\lambda = 0.1$ based on the results of synthetic experiments. SubCMap is able to extract population-average maps that are visually similar to the ones in



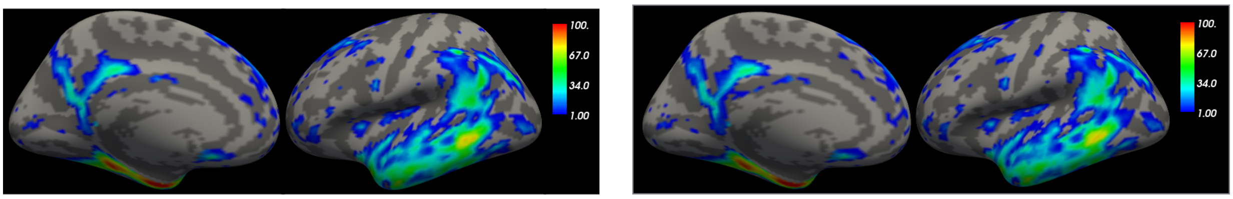
Element-wise without restoration

SubCMap: $\mathbf{q}$ maps

Outlier detection

Outlier detection after restoration

Figure 8: **Frequency of condition effect maps.** Meshes display for each vertex the number of subjects in the AD group where effect was detected in the corresponding measurement. The maps were constructed by summing the subject-specific maps constructed with the technique indicated under the images. For each image the threshold for the false positive rate was set as $\tau_{FPR} = 0.01$. We observe that outlier detection results are dispersed throughout the cortex not displaying any structure. This is due to the unspecific nature of the outlier detection methodology. SubCMap's detections show similar structure as the AD effect maps presented in previous works using regression analysis and neurohistopathology [36**?**]. The critical advantage of the maps here is that at each vertex the value corresponds to the number of subjects and for each location ids of the subjects who show alteration at that point are known. Conventional regression analysis on the other hand, only provides correlation scores with no links to individual subjects.



$\mathbf{q}$ maps

$\mathbf{q}$ maps after regressing out age and gender

Figure 9: **Frequency of condition effect maps with and without taking into account the nuisance parameters age and gender.** The images show the frequency maps similar to Figure 8. On the left we show the frequency map without taking into account the nuisance parameters and on the right with. We observe no substantial difference between these maps for this dataset.

the literature [36] but with an important difference. While values on the maps presented in the literature are statistical quantities that cannot be easily translated to numbers of subjects or links to individuals, the numbers in the maps in Figure 8 are directly the numbers of subjects with explicit links to individuals. The maps SubCMap yields also coincides with the neuropathology results of AD staging Braak and Braak presented in [37]. Outlier detection does not show high frequency in the areas that are known to be affected by AD, such as entorhinal cortex and medial temporal lobe. Compared to the element-wise method without restoration, SubCMap is able to assign higher frequency values to AD related regions.

As a last experiment, we constructed population average maps with SubCMap by taking into account age and gender as nuisance parameters. Figure 9 shows maps constructed with and without taking into account the nuisance parameters side by side. We observe that the maps are quite similar, which is not very surprising since the data was age and gender matched by construction.

4. Conclusions

This article proposed SubCMap, a novel method for detecting subject and condition specific structural alterations in neuroimaging data. The method is a simple yet effective formulation that combines elements from classification and image restoration. Experimental analysis demonstrated the advantages of SubCMap compared to population-wide methods, such as SVM, RF and GLM, as well as compared to outlier detection. Results from the analysis with the ADNI dataset showed that areas of the regions detected by SubCMap correlates with non-imaging markers of AD as well as dedicated machine learning algorithm that uses the raw measurements. This suggested that SubCMap is able to capture inter-subject variation of condition-related structural alterations.

SubCMap was presented for analyzing effects of one condition via element-wise binary classification. However, it is possible to extend the method to analyzing multiple conditions simultaneously as well as continuous valued response variables, such as age, instead of categorial groups. Such extensions is one of the avenues for our future research. Furthermore, in the present method we only used one measurement, gray-level intensity for the synthetic experiments and cortical thickness measurements for the ADNI experiments. It is however, possible to use multiple measurements without modifying the method extensively. The element-wise predictors for instance, can take multiple measurements as input.

In its current form SubCMap is a univariate method in its essence. This was by design to allow for localised interpretations. Multivariate extensions of the method is possible but they might come at the cost of losing the ability to make local interpretations of structural alterations. Even so, we believe it is an interesting avenue to extend the method to multivariate setting, where measurements at multiple locations can be analysed simultaneously.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.;NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes Rev December 5 , 2013 Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] P. M. Thompson, et al., Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas, Cerebral Cortex 11 (1) (2001) 1–16.

[2] H. Rosas, et al., Regional and progressive thinning of the cortical ribbon in Huntingtons disease, Neurology 58 (5) (2002) 695–701.

[3] E. J. Burton, et al., Cerebral atrophy in Parkinsons disease with and without dementia: a comparison with Alzheimers disease, dementia with Lewy bodies and controls, Brain 127 (4) (2004) 791–800.

[4] E. Garrido, A. Castello, J. Ventura, A. Capdevila, F. Rodriguez, Cortical atrophy and other brain magnetic resonance imaging (MRI) changes after extremely high-altitude climbs without oxygen, International journal of sports medicine 14 (04) (1993) 232–234.

[5] D. Miller, J. OCallaghan, Effects of aging and stress on hippocampal structure and function, Metabolism 52 (2003) 17–21.

[6] R. Kanai, G. Rees, The structural basis of inter-individual differences in human behaviour and cognition, Nature Reviews Neuroscience 12 (4) (2011) 231–242.

[7] K. E. Watkins, F. Vargha-Khadem, J. Ashburner, R. E. Passingham, A. Connelly, K. J. Friston, R. S. Frackowiak, M. Mishkin, D. G. Gadian, MRI analysis of an inherited speech and language disorder: structural brain abnormalities, Brain 125 (3) (2002) 465–478.

[8] J. S. Peper, R. M. Brouwer, D. I. Boomsma, R. S. Kahn, H. Pol, E. Hilleke, Genetic influences on human brain structure: a review of brain imaging studies in twins, Human brain mapping 28 (6) (2007) 464–473.

[9] P. M. Thompson, T. D. Cannon, K. L. Narr, T. Van Erp, V.-P. Poutanen, M. Huttunen, J. Lönnqvist, C.-G. Standertskjöld-Nordenstam, J. Kaprio, M. Khaledy, et al., Genetic influences on brain structure, Nature neuroscience 4 (12) (2001) 1253–1258.

[10] J. Ashburner, K. J. Friston, Why voxel-based morphometry should be used, Neuroimage 14 (6) (2001) 1238–1243.

[11] D. N. Greve, An absolute beginner's guide to surface-and voxel-based morphometric analysis, in: Proc Intl Soc Mag Reson Med, vol. 19, 2011.

[12] B. Fischl, FreeSurfer, Neuroimage 62 (2) (2012) 774–781.

[13] A. Krishnan, et al., Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review, Neuroimage 56 (2) (2011) 455–475.

[14] K. J. Worsley, et al., Characterizing the response of PET and fMRI data using multivariate linear models, NeuroImage 6 (4) (1997) 305–319.

[15] C. Maumet, P. Maurel, J.-C. Ferré, B. Carsin, C. Barillot, Patient-specific detection of perfusion abnormalities combining within-subject and between-subject variances in Arterial Spin Labeling, Neuroimage 81 (2013) 121–130.

[16] C. Maumet, P. Maurel, J.-C. Ferré, C. Barillot, An a contrario approach for the detection of patient-specific brain perfusion abnormalities with arterial spin labelling, Neuroimage 134 (2016) 424–433.

[17] M. R. Arbabshirani, S. Plis, J. Sui, V. D. Calhoun, Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls, NeuroImage .

[18] B. Gaonkar, C. Davatzikos, Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification, NeuroImage 78 (2013) 270–283.

[19] B. Mwangi, T. S. Tian, J. C. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics 12 (2) (2014) 229–244.

[20] M. Rahim, et al., Integrating Multimodal Priors in Predictive Models for the Functional Characterization of Alzheimers Disease, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Springer, 207–214, 2015.

[21] M. Ganz, et al., Relevant feature set estimation with a knock-out strategy and random forests, NeuroImage 122 (2015) 131–148.

[22] X. Tomas-Fernandez, S. K. Warfield, A Model of Population and Subject (MOPS) Intensities With Application to Multiple Sclerosis Lesion Segmentation, Medical Imaging, IEEE Transactions on 34 (6) (2015) 1349–1361.

[23] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens, Automated segmentation of multiple sclerosis lesions by model outlier detection, IEEE Transactions on Medical Imaging 20 (8) (2001) 677–688.

[24] M. Prastawa, A brain tumor segmentation framework based on outlier detection*1, Medical Image Analysis 8 (3) (2004) 275–283.

[25] K. Zeng, G. Erus, A. Sotiras, R. T. Shinohara, C. Davatzikos, Abnormality detection via iterative deformable registration and basis-pursuit decomposition, Medical Imaging, IEEE Transactions on PP (99) (2016) 1–1.

[26] K. o. Iqbal, Subgroups of Alzheimer's disease based on cerebrospinal fluid molecular markers, Annals of neurology 58 (5) (2005) 748–757.

[27] K. J. Friston, C. D. Frith, R. S. J. Frackowiak, R. Turner, Characterizing Dynamic Brain Responses with fMRI: A Multivariate Approach, Neuroimage 2 (2, Part A) (1995) 166–172.

[28] S. Kiebel, . Holmes, The general linear model, Academic Press, 2003.

[29] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[30] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[31] C. E. Bonferroni, Il calcolo delle assicurazioni su gruppi di teste, Tipografia del Senato, 1935.

[32] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[33] F. Pedregosa, et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[34] P. I. Good, Permutation, parametric and bootstrap tests of hypotheses, Springer Science+ Business Media, 2005.

[35] M. R. Sabuncu, E. Konukoglu, Clinical prediction from structural brain MRI scans: a large-scale empirical study, Neuroinformatics 13 (1) (2015) 31–46.

[36] B. C. Dickerson, et al., The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals, Cerebral cortex 19 (3) (2009) 497–510.

[37] H. Braak, E. Braak, Neuropathological stageing of Alzheimer-related changes, Acta neuropathologica 82 (4) (1991) 239–259.