NeuroImage 109 (2015) 505–514

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application ☆

Tian Ge [a,b], Thomas E. Nichols [c], Debashis Ghosh [d], Elizabeth C. Mormino [e], Jordan W. Smoller [b,f,1], Mert R. Sabuncu [a,g,*,1], for the Alzheimer's Disease Neuroimaging Initiative

[a] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA
[b] Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA
[c] Department of Statistics & Warwick Manufacturing Group, The University of Warwick, Coventry CV4 7AL, UK
[d] Department of Statistics, The Pennsylvania State University, PA 16802, USA
[e] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA
[f] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA
[g] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Accepted 9 January 2015
Available online 16 January 2015

Keywords:
Interaction
Kernel machines
Alzheimer's disease
Cardiovascular disease
Imaging genetics

ABSTRACT

Measurements derived from neuroimaging data can serve as markers of disease and/or healthy development, are largely heritable, and have been increasingly utilized as (intermediate) phenotypes in genetic association studies. To date, imaging genetic studies have mostly focused on discovering isolated genetic effects, typically ignoring potential interactions with non-genetic variables such as disease risk factors, environmental exposures, and epigenetic markers. However, identifying significant interaction effects is critical for revealing the true relationship between genetic and phenotypic variables, and shedding light on disease mechanisms. In this paper, we present a general kernel machine based method for detecting effects of the interaction between multidimensional variable sets. This method can model the joint and epistatic effect of a collection of single nucleotide polymorphisms (SNPs), accommodate multiple factors that potentially moderate genetic influences, and test for nonlinear interactions between sets of variables in a flexible framework. As a demonstration of application, we applied the method to the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to detect the effects of the interactions between candidate Alzheimer's disease (AD) risk genes and a collection of cardiovascular disease (CVD) risk factors, on hippocampal volume measurements derived from structural brain magnetic resonance imaging (MRI) scans. Our method identified that two genes, *CR1* and *EPHA1*, demonstrate significant interactions with CVD risk factors on hippocampal volume, suggesting that *CR1* and *EPHA1* may play a role in influencing AD-related neurodegeneration in the presence of CVD risks.

Introduction

Genetic components play a significant role in most brain-related illnesses. The discovery of genetic effects can elucidate the biological pathways and processes underlying neurological disorders, and ultimately yield prevention and treatment strategies. In the field of imaging genetics, this goal is approached by using quantitative brain image derived measurements as intermediate or endophenotypes (Biffi et al., 2010; Ge et al., 2014; Gottesman and Shields, 1972; Gottesman and Gould, 2003; Meyer-Lindenberg and Weinberger, 2006; Sabuncu et al., 2012), which are biomarkers of disease, and are believed to be closer to the disease process and have a simpler genetic architecture than clinical diagnoses.

However, heritability analyses and genome-wide association studies (GWAS) (Visscher et al., 2012) of complex genetic phenotypes ranging from human height (Yang et al., 2010), body mass index, von Willebrand factor (Yang et al., 2011), and schizophrenia (Lee et al., 2012b), to various volume-, surface- or connection-based brain measurements computed from structural, functional or diffusion images (Thompson et al., 2013), indicate that phenotypic variation cannot be solely explained by genetics. The interactions between genetic and non-genetic variables such as disease risk factors, environmental exposures and epigenetic markers may play an important role in the

☆ 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 at: Athinoula A. Martinos Center for Biomedical Imaging, 149 Thirteenth Street, Suite 2301, Charlestown, Massachusetts 02129, USA.
E-mail addresses: tge1@nmr.mgh.harvard.edu (T. Ge), msabuncu@nmr.mgh.harvard.edu (M.R. Sabuncu).

[1] JWS and MRS contributed equally.

http://dx.doi.org/10.1016/j.neuroimage.2015.01.029
1053-8119/

variation of complex phenotypes (Sullivan et al., 2012), and the influence of genetic variants on the likelihood, development, and progression of a brain illness may be indirect and interactive. The presence of interactions implies that genetics can modulate the effects of various risk factors on the disease, producing variations across subjects even exposed to the same environment. Alternatively, the effect of the genotype on outcomes can depend on one or more risk factors or environmental exposures. For example, Caspi et al. (2002) reported that the effect of maltreatment of children from birth to adulthood on the development of antisocial behavior is moderated by a functional polymorphism in the *MAOA* gene. The genotype of a locus known as *5-HTTLPR* located in the promoter region of the serotonin transporter gene was found to moderate the influence of stressful life events on depression (Caspi et al., 2003). Therefore, identifying potential genetic interactions with non-genetic variables can be critical in understanding the true relationship between genotype and phenotype.

Thanks to recent advances in genotyping technology, it is now possible to investigate genetic interaction effects involving specific genetic risk factors, candidate genes, or even the entire genome, in unrelated individuals. Current statistical methods to test for interactions largely utilize multiple linear regression models with quantitative phenotypes, or logistic regression models with binary outcomes, in both the genetics community (Aschard et al., 2011; Kraft et al., 2007; Paré et al., 2010), and the imaging community (e.g., psychophysiological interactions analysis (Friston et al., 1997)). In these analyses, both main effects are typically univariate variables, and the interaction is modeled by their product. Although a number of recent papers have tried to improve the power of the classical univariate interaction test (Hsu et al., 2012; Mukherjee and Chatterjee, 2008; Murcray et al., 2011), they suffer from two main drawbacks when detecting interactions between genetic variants and non-genetic variables. First, converging evidence has shown that many complex brain disorders are polygenic and influenced by up to thousands of genetic variants with small effects (Purcell et al., 2009; Sullivan et al., 2012). Analyzing each individual locus may not identify any reliable results with a small to moderate sample size, which is typical in imaging genetic studies. And second, it is now not uncommon to collect a large number of disease risk factors, environmental variables, or epigenetic markers in a single study. The product of all possible pairs of genetic variants and non-genetic variables may be dauntingly large, which dramatically increases the burden of computation and multiple testing corrections. More critically, Lin et al. (2013) showed that if the main effects of a set of genetic variants are associated with the phenotype, testing each single genetic variant for interactions can be biased.

In this paper, inspired by Li and Cui (2012), we present a semiparametric kernel machine based method to detect interactions between multidimensional variable sets. Kernel machine based methods have been previously used in association studies between single nucleotide polymorphism (SNP) sets and complex diseases or imaging phenotypes (Kwee et al., 2008; Liu et al., 2007; Wu et al., 2010, 2011), and have been applied to voxel-wise genome-wide association studies to obtain boosted statistical power (Ge et al., 2012; Stein et al., 2010). Here, to jointly model the genetic and non-genetic variables, and their interactions, we extend the original kernel machine based method, and include three appropriately selected kernels in the model; one for genetic variants, one for non-genetic variables, and a third one, which is the Hadamard product of the genetic and non-genetic kernel, for the interaction effect. The genetic kernel provides a biologically-informed way to capture epistasis in a set of SNPs and model their joint effect on the phenotype. SNP sets can be formed by SNPs located in or near a gene, within a gene pathway or a haplotype structure; risk SNPs identified by previous studies or other a priori biological information (Wu et al., 2010). Examining the collective contribution of SNPs further opens possibilities to investigate cumulative effects of rare variants (Wu et al., 2011), and often provides improved reproducibility, biologically informed insights, and increased power relative to univariate methods. The non-genetic kernel allows for modeling the joint effect of multiple variables. By using a connection to linear mixed effects models, the interaction effect can be tested by a variance component score test (Lin, 1997; Liu et al., 2007). The proposed method thus offers a flexible framework to account for epistatic effects, multiple non-genetic factors, and test for the overall interaction effect between sets of multidimensional variables.

As a demonstration of application, we applied the proposed method to detect the interaction effects between candidate late-onset Alzheimer's disease (AD) risk genes and cardiovascular disease (CVD) risk factors including age, gender, body mass index (BMI), hypertension, current smoking status and diabetes, on hippocampal volume derived from structural brain magnetic resonance imaging (MRI) scans, which is associated with AD risk and future AD progression (Sperling et al., 2011).

AD, the most common form of dementia, is characterized by memory loss, cognitive decline, and other symptoms. The cause and progression of AD are not well understood. As a disease that often co-occurs with AD in the elderly population, vascular pathology is among the potential factors to increase the risk of AD. In particular, increasing evidence shows that many CVD risk factors including hypertension, smoking and diabetes are associated with cognitive decline and neurodegeneration, and may increase the risk and accelerate the progression of AD (Helzner et al., 2009; Kivipelto et al., 2001; Lo et al., 2012; Luchsinger et al., 2005; Purnell et al., 2009). For example, the neurovascular hypothesis of AD suggests that neurovascular dysfunction reduces the clearance of amyloid beta (Aβ) peptide across the blood–brain barrier, which could initiate a series of pathological processes and ultimately lead to neuronal injury and loss (Zlokovic, 2005). Moreover, recent studies have identified that the interaction within multiple CVD risk factors, and the interaction between CVD risk factors and the apolipoprotein E (*APOE*) polymorphism, the largest genetic determinant of late-onset AD susceptibility, may significantly influence the risk and progression of AD (Borenstein et al., 2005; Irie et al., 2008; Purnell et al., 2009; Qiu et al., 2003). We therefore hypothesized that genetic components play a role in the development and progression of AD in the presence of CVD risk factors and events. Testing for the interactions between AD risk genes and CVD risk factors on hippocampal volume may shed light on the underlying mechanisms of AD-related neurodegeneration, and suggest potential therapeutic treatment as many CVD risk factors are largely modifiable.

The remainder of the paper is organized as follows. In the Materials and methods section, we present the kernel machine based method and the statistical test for interaction detection between multidimensional variable sets. Simulation studies are then introduced to evaluate the proposed method. In the Results section, simulation results, as well as our findings on the real data are shown, and compared to alternative interaction detection methods. The advantages and weaknesses of the method, and the implication of the findings, are summarized in the Discussion section. Some theoretical aspects of the kernel method and supplementary analyses are provided in the Appendix.

Materials and methods

Kernel methods for interaction detection

The model

We assume that there are N unrelated subjects under investigation. y_i, $i = 1, \cdots, N$, is a quantitative phenotype for the i-th subject, such as an image derived disease marker. We are interested in detecting the interaction between a collection of genetic variants and a set of non-genetic variables such as disease risk factors, environmental exposures, or epigenetic markers. In particular, let $\boldsymbol{G}_i = [G_{i,1}, \cdots, G_{i,L}]^\top$ denote the L SNP markers, where $G_{i,s}$, $s = 1, \cdots, L$, is the genotype coded to be the number of copies of the minor allele that the i-th subject possesses for the s-th SNP, and takes the values of 0 (homozygotic major alleles), 1 (heterozygote), and 2 (homozygotic minor alleles). Let $\boldsymbol{W}_i = [W_{i,1}, \cdots, W_{i,R}]^\top$ denote the R non-genetic variables for the i-th subject.

We associate the phenotype with the genetic and non-genetic variables via the following semiparametric model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f(\boldsymbol{G}_i, \boldsymbol{W}_i) + \epsilon_i,\ i = 1, \cdots, N, \tag{1}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates (e.g., age, sex) for the i-th subject, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, ϵ_i is random residual with zero-mean and homogeneous variance σ^2, f is an unknown function on the product domain $\mathcal{X} = \mathcal{X}_G \otimes \mathcal{X}_W$, with $\boldsymbol{G}_i \in \mathcal{X}_G$ and $\boldsymbol{W}_i \in \mathcal{X}_W$. According to the ANOVA decomposition of functions (Gu, 2002), f can be expanded as:

$$f(\boldsymbol{G}_i, \boldsymbol{W}_i) = h_G(\boldsymbol{G}_i) + h_W(\boldsymbol{W}_i) + h_{G\times W}(\boldsymbol{G}_i, \boldsymbol{W}_i), \tag{2}$$

where $h_G(\boldsymbol{G}_i)$ and $h_W(\boldsymbol{W}_i)$ are the main effects of genetics and non-genetic factors, respectively, and $h_{G\times W}(\boldsymbol{G}_i, \boldsymbol{W}_i)$ captures interactions. The overall mean of f can be absorbed into the intercept contained in $\boldsymbol{x}_i$, and is therefore omitted here. A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of smooth real-valued functions on $\mathcal{X}$ can be constructed (Gu and Wahba, 1993; Wahba et al., 1995). In particular, the functional space $\mathcal{H}$ has an orthogonal decomposition:

$$\mathcal{H} = \mathcal{H}_G \oplus \mathcal{H}_W \oplus \mathcal{H}_{G\times W}, \tag{3}$$

where $\mathcal{H}_G$ and $\mathcal{H}_W$ are RKHSs of functions on $\mathcal{X}_G$ and $\mathcal{X}_W$, respectively, $\mathcal{H}_{G\times W}$ is a RKHS of functions on $\mathcal{X}$, $\oplus$ denotes direct sum. Each component in Eq. (2) lies in the corresponding subspace in Eq. (3). Therefore, $\mathcal{H}$ is a RKHS with the associated reproducing kernel as the sum of the reproducing kernels of the three component subspaces. We assume that $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|_{\mathcal{H}}$.

Model estimation

The function $f \in \mathcal{H}$ can be estimated by minimizing the penalized squared-error loss function of model (1):

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\beta}, f) = \frac{1}{2}\sum_{i=1}^{N}\left[y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta} - f(\boldsymbol{G}_i, \boldsymbol{W}_i)\right]^2 + \frac{\lambda}{2}\mathcal{J}(f), \tag{4}$$

where $\mathcal{J}(\cdot) = \|\cdot\|_{\mathcal{H}}^2$ is a roughness penalty, and λ is a tuning parameter. Since the entire functional space $\mathcal{H}$ has the orthogonal decomposition (3), the penalty function $\mathcal{J}(\cdot)$ can be decomposed accordingly, and Eq. (4) can be more explicitly written as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}, \boldsymbol{\beta}, f) &= \frac{1}{2}\sum_{i=1}^{N}\left[y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta} - h_G(\boldsymbol{G}_i) - h_W(\boldsymbol{W}_i) - h_{G\times W}(\boldsymbol{G}_i, \boldsymbol{W}_i)\right]^2 \\ &\quad + \frac{\lambda_G}{2}\|h_G\|_{\mathcal{H}_G}^2 + \frac{\lambda_W}{2}\|h_W\|_{\mathcal{H}_W}^2 + \frac{\lambda_{G\times W}}{2}\|h_{G\times W}\|_{\mathcal{H}_{G\times W}}^2 \\ &= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{h}_G - \boldsymbol{h}_W - \boldsymbol{h}_{G\times W})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{h}_G - \boldsymbol{h}_W - \boldsymbol{h}_{G\times W}) \\ &\quad + \frac{\lambda_G}{2}\|h_G\|_{\mathcal{H}_G}^2 + \frac{\lambda_W}{2}\|h_W\|_{\mathcal{H}_W}^2 + \frac{\lambda_{G\times W}}{2}\|h_{G\times W}\|_{\mathcal{H}_{G\times W}}^2,\end{aligned} \tag{5}$$

where $\boldsymbol{y} = [y_1, \cdots, y_N]^\top$, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^\top$, $\boldsymbol{h}_G = [h_G(\boldsymbol{G}_1), \cdots, h_G(\boldsymbol{G}_N)]^\top$, $\boldsymbol{h}_W = [h_W(\boldsymbol{W}_1), \cdots, h_W(\boldsymbol{W}_N)]^\top$, $\boldsymbol{h}_{G\times W} = [h_{G\times W}(\boldsymbol{G}_1, \boldsymbol{W}_1), \cdots, h_{G\times W}(\boldsymbol{G}_N, \boldsymbol{W}_N)]^\top$, λ_G, λ_W, and $\lambda_{G\times W}$ are positive smoothing parameters that balance the goodness of fit and complexity of the model.

By the representer theorem (Kimeldorf and Wahba, 1971; Wahba, 1990), the functions h_G, h_W and $h_{G\times W}$ that minimize the functional (5) take the forms:

$$\begin{aligned}h_G(\boldsymbol{G}^*) &= \sum_{j=1}^{N}\alpha_{G,j}k_G\left(\boldsymbol{G}^*, \boldsymbol{G}_j\right), \\ h_W(\boldsymbol{W}^*) &= \sum_{j=1}^{N}\alpha_{W,j}k_W\left(\boldsymbol{W}^*, \boldsymbol{W}_j\right), \\ h_{G\times W}(\boldsymbol{G}^*, \boldsymbol{W}^*) &= \sum_{j=1}^{N}\alpha_{G\times W,j}k_{G\times W}\left((\boldsymbol{G}^*, \boldsymbol{W}^*), \left(\boldsymbol{G}_j, \boldsymbol{W}_j\right)\right),\end{aligned} \tag{6}$$

for arbitrary $\boldsymbol{G}^*$ and $\boldsymbol{W}^*$, where $\alpha_{G,j}$, $\alpha_{W,j}$ and $\alpha_{G\times W,j}$, $j = 1, 2, \cdots, N$, are unknown coefficients, k_G, k_W and $k_{G\times W}$ are reproducing kernel functions of the Hilbert spaces $\mathcal{H}_G$, $\mathcal{H}_W$ and $\mathcal{H}_{G\times W}$, respectively. Since the reproducing kernel of a tensor product of two RKHSs is the product of the two reproducing kernels (Aronszajn, 1950), the kernel function $k_{G\times W}$ is connected to the kernel functions k_G and k_W by:

$$k_{G\times W}\left((\boldsymbol{G}^*, \boldsymbol{W}^*), \left(\boldsymbol{G}_j, \boldsymbol{W}_j\right)\right) = k_G\left(\boldsymbol{G}^*, \boldsymbol{G}_j\right) \cdot k_W\left(\boldsymbol{W}^*, \boldsymbol{W}_j\right). \tag{7}$$

Define the $N \times N$ symmetric kernel matrices $\boldsymbol{K}_G = \{k_G(\boldsymbol{G}_i, \boldsymbol{G}_j)\}$, $\boldsymbol{K}_W = \{k_W(\boldsymbol{W}_i, \boldsymbol{W}_j)\}$ and $\boldsymbol{K}_{G\times W} = \{k_{G\times W}((\boldsymbol{G}_i, \boldsymbol{W}_i), (\boldsymbol{G}_j, \boldsymbol{W}_j))\} = \boldsymbol{K}_G \odot \boldsymbol{K}_W$, where $\odot$ is the Hadamard product (element-wise product) of two matrices. Then:

$$\boldsymbol{h}_G = \boldsymbol{K}_G\boldsymbol{\alpha}_G, \quad \boldsymbol{h}_W = \boldsymbol{K}_W\boldsymbol{\alpha}_W, \quad \boldsymbol{h}_{G\times W} = \boldsymbol{K}_{G\times W}\boldsymbol{\alpha}_{G\times W}, \tag{8}$$

where $\boldsymbol{\alpha}_G = [\alpha_{G,1}, \cdots, \alpha_{G,N}]^\top$, $\boldsymbol{\alpha}_W = [\alpha_{W,1}, \cdots, \alpha_{W,N}]^\top$ and $\boldsymbol{\alpha}_{G\times W} = [\alpha_{G\times W,1}, \cdots, \alpha_{G\times W,N}]^\top$. Substituting $\boldsymbol{h}_G$, $\boldsymbol{h}_W$ and $\boldsymbol{h}_{G\times W}$ into Eq. (5), and making use of the reproducing kernel property, we obtain:

$$\begin{aligned}&\mathcal{L}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}_G, \boldsymbol{\alpha}_W, \boldsymbol{\alpha}_{G\times W}) \\ &= \frac{1}{2}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} + \frac{\lambda_G}{2}\boldsymbol{\alpha}_G^\top\boldsymbol{K}_G\boldsymbol{\alpha}_G + \frac{\lambda_W}{2}\boldsymbol{\alpha}_W^\top\boldsymbol{K}_W\boldsymbol{\alpha}_W + \frac{\lambda_{G\times W}}{2}\boldsymbol{\alpha}_{G\times W}^T\boldsymbol{K}_{G\times W}\boldsymbol{\alpha}_{G\times W},\end{aligned} \tag{9}$$

where $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}_G\alpha_G - \boldsymbol{K}_W\alpha_W - \boldsymbol{K}_{G\times W}\alpha_{G\times W}$.

The gradients of $\mathcal{L}$ with respect to the parametric coefficients $\boldsymbol{\beta}$ and nonparametric coefficients $\boldsymbol{\alpha}_G$, $\boldsymbol{\alpha}_W$, and $\boldsymbol{\alpha}_{G\times W}$ are:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} &= -\boldsymbol{X}^\top\boldsymbol{\epsilon}, \qquad \frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}_G} - \boldsymbol{K}_G^\top\boldsymbol{\epsilon} + \lambda_G\boldsymbol{K}_G\boldsymbol{\alpha}_G, \\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}_W} &= -\boldsymbol{K}_W^\top\boldsymbol{\epsilon} + \lambda_W\boldsymbol{K}_W\boldsymbol{\alpha}_W, \quad \frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}_{G\times W}} = -\boldsymbol{K}_{G\times W}^\top\boldsymbol{\epsilon} + \lambda_{G\times W}\boldsymbol{K}_{G\times W}\boldsymbol{\alpha}_{G\times W}\end{aligned}. \tag{10}$$

Therefore, setting the gradients to zero, this first-order condition is given by the linear system:

$$\begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{K}_G & \boldsymbol{X}^\top\boldsymbol{K}_W & \boldsymbol{X}^\top\boldsymbol{K}_{G\times W} \\ \boldsymbol{K}_G^\top\boldsymbol{X} & \boldsymbol{K}_G^\top\boldsymbol{K}_G + \lambda_G\boldsymbol{K}_G & \boldsymbol{K}_G^\top\boldsymbol{K}_W & \boldsymbol{K}_G^\top\boldsymbol{K}_{G\times W} \\ \boldsymbol{K}_W^\top\boldsymbol{X} & \boldsymbol{K}_W^\top\boldsymbol{K}_G & \boldsymbol{K}_W^\top\boldsymbol{K}_W + \lambda_W\boldsymbol{K}_W & \boldsymbol{K}_W^\top\boldsymbol{K}_{G\times W} \\ \boldsymbol{K}_{G\times W}^\top\boldsymbol{X} & \boldsymbol{K}_{G\times W}^\top\boldsymbol{K}_G & \boldsymbol{K}_{G\times W}^\top\boldsymbol{K}_W & \boldsymbol{K}_{G\times W}^\top\boldsymbol{K}_{G\times W} + \lambda_{G\times W}\boldsymbol{K}_{G\times W}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\alpha}_G \\ \boldsymbol{\alpha}_W \\ \boldsymbol{\alpha}_{G\times W}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{y} \\ \boldsymbol{K}_G^\top\boldsymbol{y} \\ \boldsymbol{K}_W^\top\boldsymbol{y} \\ \boldsymbol{K}_{G\times W}^\top\boldsymbol{y}\end{bmatrix}. \tag{11}$$

Liu et al. (2007) showed that this first-order linear system is equivalent to the normal equation of the linear mixed effects model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h}_G + \boldsymbol{h}_W + \boldsymbol{h}_{G\times W} + \boldsymbol{\epsilon}, \tag{12}$$

where $\boldsymbol{\beta}$ is a coefficient vector of fixed effects, $\boldsymbol{h}_G$, $\boldsymbol{h}_W$ and $\boldsymbol{h}_{G\times W}$ are independent random effects, and distributed as $\boldsymbol{h}_G \sim \mathrm{N}(\boldsymbol{0}, \tau_G^2\boldsymbol{K}_G)$, $\tau_G^2 = \lambda_G^{-1}\sigma^2$, $\boldsymbol{h}_W \sim \mathrm{N}(\boldsymbol{0}, \tau_W^2\boldsymbol{K}_W)$, $\tau_W^2 = \lambda_W^{-1}\sigma^2$, $\boldsymbol{h}_{G\times W} \sim \mathrm{N}(\boldsymbol{0}, \tau_{G\times W}^2\boldsymbol{K}_{G\times W})$, $\tau_{G\times W}^2 = \lambda_{G\times W}^{-1}\sigma^2$, $\boldsymbol{\epsilon}$ is uncorrelated random effects and follows $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and $\boldsymbol{I}$ is an identity matrix. This connection indicates that the fixed effects $\boldsymbol{\beta}$, and the random effects $\boldsymbol{h}_G$, $\boldsymbol{h}_W$ and $\boldsymbol{h}_{G\times W}$, obtained by minimizing the loss function in Eq. (4), are equivalent to the best linear unbiased predictors (BLUPs) of the linear mixed effects model (12). The variance components τ_G^2, τ_W^2, $\tau_{G\times W}^2$ and σ^2 can be estimated via the restricted maximum likelihood (ReML) approach (Harville, 1977; Lindstrom and Bates, 1988) (see the Appendix for details), and the estimates of random effects $\hat{\boldsymbol{h}}_G$, $\hat{\boldsymbol{h}}_W$, $\hat{\boldsymbol{h}}_{G\times W}$ can be obtained by solving the linear system (11) and inserting the $\hat{\boldsymbol{\alpha}}$ estimates into Eq. (8).

Selection of kernels

There are a variety of choices for the kernel functions to characterize the similarity between subjects with respect to the genetic variants and non-genetic factors, as long as they are nonnegative definite (Schaid, 2010a,b). Possible candidates are the linear kernel, the polynomial kernel, the Euclidean distance (ED) kernel, the Gaussian kernel, and the identity-by-state (IBS) kernel (Kwee et al., 2008).

Here we use the IBS kernel for the genetic effect. The IBS kernel measures the similarity of the genotypes between the i-th and j-th subject by:

$$k_G\left(\boldsymbol{G}_i, \boldsymbol{G}_j\right) = \frac{1}{2L}\sum_{s=1}^{L}\left(2-\left|G_{i,s}-G_{j,s}\right|\right), \tag{13}$$

where L is the number of SNP markers to be combined. The IBS kernel is a nonparametric function of the genotypes, as it does not depend on the selection of basis or any assumption on the types of genetic interaction. Therefore, in principle, it can capture any epistatic effect between genetic variants and their nonlinear influences on the phenotypes.

We propose the linear kernel to combine multiple non-genetic factors. The linear kernel can be represented as:

$$k_W\left(\boldsymbol{W}_i, \boldsymbol{W}_j\right) = \frac{1}{R} < \boldsymbol{W}_i, \boldsymbol{W}_j > = \frac{1}{R}\sum_{s=1}^{R} W_{i,s}W_{j,s}, \tag{14}$$

where R is the number of non-genetic factors under investigation. We evaluate the performance of the two kernels by simulation studies.

Score test

We note, from the linear mixed effects model representation (12), that testing an overall genetic and non-genetic effect $\mathcal{H}_0 : h_G(\cdot) = h_W(\cdot) = h_{G\times W}(\cdot) = 0$ is equivalent to testing the variance components: $\mathcal{H}_0 : \tau_G^2 = \tau_W^2 = \tau_{G\times W}^2 = 0$. To address the issue that, under the null hypothesis, the parameters τ_G^2, τ_W^2, and $\tau_{G\times W}^2$ are on the boundary of the parameter space, Liu et al. (2007) proposed a score test based on the ReML. In particular, let $\boldsymbol{K} = \boldsymbol{K}_G + \boldsymbol{K}_W + \boldsymbol{K}_{G\times W}$, and the score test statistic is defined as:

$$\mathcal{S}\left(\sigma_0^2\right) = \frac{1}{2\sigma_0^2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_0\right)^{\top}\boldsymbol{K}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_0\right) = \frac{1}{2\sigma_0^2}\boldsymbol{y}^{\top}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{y}, \tag{15}$$

where $\hat{\boldsymbol{\beta}}_0$ is the maximum likelihood estimate (MLE) of the regression coefficients under the null model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_0$, σ_0^2 is the variance of $\boldsymbol{\epsilon}_0$, $\boldsymbol{P}_0 = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}$ is the projection matrix under the null. $\mathcal{S}\left(\sigma_0^2\right)$ is a quadratic function of $\boldsymbol{y}$ and follows a mixture of chi-squares under the null. We use the Satterthwaite method to approximate the distribution of $\mathcal{S}\left(\sigma_0^2\right)$ by a scaled chi-square distribution $\kappa\chi_\nu^2$. In practice, the unknown value of the model parameter σ_0^2 in $\mathcal{S}$ is replaced by its ReML estimate $\hat{\sigma}_0^2$ under the null model. To account for this substitution, the fitted scale parameter κ and the degrees of freedom ν are adjusted, giving $\hat{\kappa}$ and $\hat{\nu}$ (see the Appendix for details). The p-value of an observed score statistic $\mathcal{S}\left(\hat{\sigma}_0^2\right)$ is then computed using the scaled chi-square distribution $\hat{\kappa}\chi_{\hat{\nu}}^2$.

To test the interaction effect, we notice that testing the null hypothesis $\mathcal{H}_0^I : h_{G\times W}(\cdot) = 0$ is equivalent to testing the variance component: $\mathcal{H}_0^I : \tau_{G\times W}^2 = 0$. Let $\boldsymbol{\Sigma} = \tau_G^2\boldsymbol{K}_G + \tau_W^2\boldsymbol{K}_W + \sigma^2\boldsymbol{I}$, where τ_G^2, τ_W^2, and σ^2 are model parameters under the null model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h}_G + \boldsymbol{h}_W + \boldsymbol{\varepsilon}$. We follow Li and Cui (2012) and design a score test statistic

$$\mathcal{S}_I\left(\tau_G^2, \tau_W^2, \sigma^2\right) = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}_I\boldsymbol{K}_{G\times W}\boldsymbol{P}_I\boldsymbol{y}, \tag{16}$$

where $\boldsymbol{P}_I = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$ is the projection matrix under the null hypothesis $\mathcal{H}_0^I$. Analogously, the Satterthwaite method is used to approximate the distribution of $\mathcal{S}_I$ by a scaled chi-square distribution $\kappa_I\chi_{\nu_I}^2$. In practice, the unknown model parameters τ_G^2, τ_W^2 and σ^2 in $\mathcal{S}_I$ are replaced by their ReML estimates $\hat{\tau}_G^2$, $\hat{\tau}_W^2$ and $\hat{\sigma}^2$ under the null model. The fitted scale parameter κ_I and the degrees of freedom ν_I are adjusted to account for this substitution, giving $\hat{\kappa}_I$ and $\hat{\nu}_I$ (see the Appendix for details). The p-value of an observed score statistic $\mathcal{S}_I\left(\hat{\tau}_G^2, \hat{\tau}_W^2, \hat{\sigma}^2\right)$ is then computed using the scaled chi-square distribution $\hat{\kappa}_I\chi_{\hat{\nu}_I}^2$.

The ADNI data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older subjects, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Data preprocessing and SNP grouping

All ADNI-1 1.5T structural brain MRI scans were processed using FreeSurfer (freesurfer.nmr.mgh.harvard.edu) (Dale et al., 1999; Fischl, 2012; Fischl et al., 1999), version 4.3. Subject specific intra-cranial volume (ICV) and bilateral hippocampal volumes were automatically computed by FreeSurfer, after skull stripping, B1 bias field correction, segmentation and labeling (Fischl et al., 2002, 2004), and passed rigorous visual quality control checks. For more details regarding the imaging processing and quality control, we refer the reader to the official website of ADNI (http://adni.loni.usc.edu).

CVD risk factors considered in the present study included age, gender, body mass index (BMI), systolic blood pressure, current smoking status and diabetes. A CVD risk score summarizing these six risk factors can be calculated using the non-laboratory, office-based cardiovascular risk profile prediction function from the Framingham Heart Study (FHS) (D'Agostino et al., 2008). The score can be treated as a continuous variable, and higher values indicate higher risks of developing individual CVD events. We use the FHS risk score as a benchmark variable to compare the results obtained with the proposed multivariate method.

We followed the ENIGMA2 1KGP cookbook (v3) (The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) consortium, http://enigma.loni.ucla.edu/wp-content/uploads/2012/07/ENIGMA2_1KGP_cookbook_v3.doc, version July 27, 2012), developed by the ENIGMA2 Genetics support team, to preprocess and impute the ADNI genome-wide SNP data. In brief, we used PLINK (Purcell et al., 2007) for preprocessing and quality control, which included sex discrepancy check, removing subjects with low genotype call rate (<95%), and filtering individual markers that contained an ambiguous strand assignment and that did not satisfy the following quality control criteria: genotype

call rate ≥95%, minor allele frequency (MAF) ≥1%, and Hardy–Weinberg equilibrium $p \geq 1 \times 10^{-6}$. We then used the MaCH software (Li et al., 2010) to impute ungenotyped SNPs based on the 1000 genomes reference (1000 Genomes Project Consortium, 2012). 697 subjects (cognitive normal controls $N = 203$, subjects with mild cognitive impairment $N = 334$, and AD patients $N = 160$) that have complete imaging and genetic data, and CVD risk factors, were included in the following analyses. Among the 334 subjects with mild cognitive impairment (MCI), 183 subjects were stable and did not convert to AD throughout the follow-up, and 151 subjects progressed to AD in at least one of the follow-up visits.

In addition to *APOE*, the major genetic risk factor for late-onset AD, a recent two-stage meta-analysis of GWAS with 74,046 individuals identified 20 susceptibility loci for late-onset AD (Lambert et al., 2013). A very recent article suggested that the *REST* gene may play a critical role in normal aging in human cortical and hippocampal neurons, and may distinguish neuroprotection from neurodegeneration (Lu et al., 2014). We therefore used these 21 genes as our candidate gene set and extracted all the SNPs on the coding regions as well as 20 kb up/downstream of each of these genes in the ADNI data set. Some of these genes, e.g., *BIN1*, *CR1* and *PICALM*, have been associated with quantitative imaging phenotypes, such as hippocampal volume, amygdala volume and entorhinal cortical thickness, in ADNI (Biffi et al., 2010; Bralten et al., 2011; Furney et al., 2010; Weiner et al., 2013). Table 1 lists the 21 genes and the final number of SNPs located on them after preprocessing and quality control.

Alternative methods

No standard method exists in the literature that can detect interactions between a collection of SNPs and a set of non-genetic variables such as CVD risk factors. Below in both simulation studies and real data analysis, we consider alternative methods based on burden tests and principal component analysis (PCA) that can summarize multiple variables into a single regressor and convert the problem into standard multiple regression analyses.

Burden tests collapse a set of variants in a genetic region into a single burden variable. They can be powerful when most variants in a region are causal and the effects are in the same direction, but suffer from dramatic power loss when these assumptions are violated (Lee et al., 2012a). Different variants of burden tests have been proposed and are mainly aimed at rare variant association tests. Here we adapt two methods to our context: (1) the rare variant test (RVT), proposed by Morris and Zeggini (2010), which calculates the proportion of minor alleles in the set of genetic variants for each subject as the burden regressor, and (2) the weighted sum test (WST) (Madsen and Browning, 2009), which calculates a genetic score as the burden variable. The genetic score is a weighted average of the count of minor alleles for each subject. Specifically, if $G_{i,s} \in \{0, 1, 2\}$ is the count of minor alleles in genetic variant s for the i-th subject, then the genetic score is $\gamma_i = \sum_{s=1}^{L} G_{i,s} w_s$, where L is the number of SNP markers, $w_s = \sqrt{N_s q_s (1-q_s)}$ is the weight, in which $q_s = (m_s + 1)/(2N_s + 2)$, N_s is the total number of subjects genotyped for variant s, and m_s is the number of minor alleles observed for variant s. Many other burden tests are similar to these two methods. We note that the underlying assumptions of these collapsing methods are that the interactions have similar effect sizes and the same direction for all the genetic variants being collapsed. The tests can be biased or have inflated type I error if these assumptions are violated.

For the second alternative method, we perform PCA on the set of SNP regressors or CVD risk factors to extract the first principal component that explains the largest possible variance of the original regressors.

After reducing the dimension of the SNP set and the CVD risk factors, we can carry out a standard multiple regression analysis, in which the interaction effect between the derived univariate SNP regressor and the CVD risk factor is modeled by their product.

Simulation studies

We conducted simulation studies to evaluate the performance of the ReML algorithm and the accuracy of the score tests. The simulation was based on real ADNI demographic information, genetic data and CVD risk factors with $N = 697$ subjects, in order to best mimic the situation of our real data application. To synthesize quantitative phenotypes, we employed the following model:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \alpha_M[h_G(\boldsymbol{G}_i) + h_W(\boldsymbol{W}_i)] + \alpha_I h_{G\times W}(\boldsymbol{G}_i, \boldsymbol{W}_i) + \sigma\varepsilon_i, \tag{17}$$

where $\boldsymbol{x}_i$ is a vector comprising an intercept, the ICV, and the education (in years) of the i-th subject, $\boldsymbol{\beta}$ is a vector of all ones, ε_i is a Gaussian distributed random error with zero mean and unit variance, σ is the standard deviation of the error and was set to 5 in our simulation studies. α_M and α_I are two free parameters. We followed (Liu et al., 2007) and designed the function h_G to have the following complex form:

$$h_G(\boldsymbol{G}_i) = 2\cos\left(G_{i,1}\right) - 3G_{i,2}^2 + 2e^{-G_{i,3}}G_{i,4} - 1.6\sin\left(G_{i,5}\right)\cos\left(G_{i,3}\right) + 4G_{i,1}G_{i,5}. \tag{18}$$

The main non-genetic effect was designed as $h_W(\boldsymbol{W}_i) = W_{i,1} + W_{i,2}$. Finally, we introduced a linear interaction effect between the genetic variants and CVD risk factors: $h_{G\times W}(\boldsymbol{G}_i, \boldsymbol{W}_i) = 3h_G(\boldsymbol{G}_i)h_W(\boldsymbol{W}_i)$.

Since previous work has performed extensive simulations to characterize the overall score test for the semiparametric model (Hua and Ghosh, 2014; Liu et al., 2007), we focused our simulations on testing for the interaction effect. Our major concern is to assess whether the main effects "bleed" into the interaction, yielding false positives, or "cloud" the interaction, reducing sensitivity.

In the first simulation study, we generated data under different values of α_M and α_I to evaluate the performance of the score tests. Specifically, when $\alpha_M = \alpha_I = 0$, both main and interaction effects vanish, and we studied the false positive rate of the score test for overall effect. When $\alpha_M > 0$ and $\alpha_I = 0$, there are main effects but no interaction, and we therefore assessed the power of the overall score test, and the false positive control of the score test for interaction effect. We also set α_M and α_I at a number of different values to test the power of both score tests in different situations. 1000 simulations were performed for each setting. For each run, we randomly picked a gene from Table 1 and randomly selected five adjacent SNPs on the gene, reflecting the linkage disequilibrium (LD) between genetic markers, and randomly selected two variables from the six CVD risk factors (age, gender, BMI, systolic blood pressure, smoking and diabetes). The phenotypic data were then generated using the five SNPs and two CVD risk variables following Eq. (17). We note that for all the genes the signal only comes from a very

Table 1
A list of 21 candidate risk genes for late-onset Alzheimer's disease and the final number of SNPs located on and near them.

Chr	Gene	SNP num	Chr	Gene	SNP num
19	*ABCA7*	240	6	*HLA-DRB5*	62
2	*BIN1*	301	2	*INPP5D*	495
20	*CASS4*	165	5	*MEF2C*	272
6	*CD2AP*	421	11	*MS4A6A*	63
19	*CD33*	85	11	*PICALM*	360
11	*CELF1*	97	8	*PTK2B*	419
8	*CLU*	116	4	*REST*	146
1	*CR1*	264	14	*SLC24A4*	716
18	*DSG2*	219	11	*SORL1*	233
7	*EPHA1*	115	7	*ZCWPW1*	74
14	*FERMT2*	242			

Table 2
Simulation results of the overall and interaction score tests, and the alternative methods for interaction detection based on dimension reduction and multiple regression. Nominal p-value threshold was set to 0.05. The first row corresponds to simulating the null hypothesis for both the overall and interaction effects. The second and third rows correspond to the null hypothesis of the interaction effect only. Thus, corresponding detection rates in the first three rows are desired to be below the p-value threshold of 0.05.

(α_M, α_I)	Kernel method		Alternative methods					
	Overall	Interaction	PCg × FHS	PCg × PCw	RVT × FHS	RVT × PCw	WST × FHS	WST × PCw
(0, 0)	0.048	–	0.051	0.043	0.051	0.040	0.049	0.038
(0.5, 0)	0.908	0.046	0.061	0.046	0.063	0.043	0.062	0.054
(1, 0)	1.000	0.051	0.052	0.052	0.068	0.049	0.061	0.052
(0. 0.5)	0.961	0.918	0.622	0.499	0.578	0.444	0.572	0.455
(0, 1)	0.983	0.950	0.681	0.546	0.620	0.508	0.631	0.505
(1, 0.1)	0.999	0.585	0.292	0.242	0.229	0.204	0.226	0.216
(1, 0.25)	0.995	0.865	0.506	0.405	0.432	0.324	0.433	0.325
(1, 0.5)	0.997	0.926	0.591	0.481	0.542	0.413	0.536	0.425
(0.1, 1)	0.984	0.951	0.681	0.529	0.622	0.478	0.610	0.476
(0.25, 1)	0.986	0.944	0.665	0.521	0.600	0.459	0.601	0.469
(0.5, 1)	0.983	0.951	0.654	0.517	0.612	0.472	0.587	0.446
(0.5, 0.5)	0.984	0.918	0.625	0.488	0.587	0.423	0.575	0.431
(1, 1)	0.994	0.958	0.660	0.527	0.629	0.488	0.620	0.489

PCg: first principal component of the genetic data; PCw: first principal component of the cardiovascular disease risk factors; RVT: rare variant test burden variable; WST: weighted sum test burden variable; FHS: the Framingham Heart Study vascular disease risk score.

small proportion of the SNPs. Likewise, only part of the CVD risk factors were used in producing the phenotypic data.

We then evaluated the performance of the kernel method. As a comparison, we summarized genetic variants and CVD risk factors into a single regressor respectively using different collapsing methods (for genetic data: PCA, RVT and WST; for CVD risk factors: PCA and the FHS risk score[2]), and conducted standard univariate interaction tests between all possible combinations of these univariate genetic and CVD risk variables, which amounted to six multiple regression analyses.

In the second simulation study, we fixed $\alpha_M = 1$, and for each run α_I was assigned a random number uniformly distributed on [0, 1] with a probability of 0.5, and was fixed at 0 otherwise. We then generated data following the same approach described above, and compared the Receiver Operating Characteristic (ROC) curves of the kernel method and alternative methods for interaction detection.

Real data application

As a sanity check, we started with some standard regression analyses of real data from ADNI. Specifically, we tested the association between hippocampal volume (averaged between two hemispheres) and APOE-ε4 status (carriers vs. non-carriers), after controlling for ICV, age, gender and education. We conducted multiple regression analyses to assess the main effects of the FHS CVD risk score and each CVD risk factor, and their interaction effects with APOE-ε4 status on hippocampal volume, after properly controlling for covariates. Using logistic regression, we also analyzed the association between diagnosis (AD patients vs. cognitive normal controls) and the FHS CVD risk score and each CVD risk factor.

We then applied the kernel method to detect interaction effects between each of the 21 candidate AD risk genes listed in Table 1, and the collection of six CVD risk factors, on hippocampal volume. ICV and education were included in the model as covariates. The IBS kernel was used to combine SNPs located on and near each gene, and a linear kernel was used for the CVD risk factors. All CVD risk factors were standardized (subtracting the mean and divided by the standard deviation) to transform variables measured with different units onto the same scale. Bonferroni correction was used to control the family-wise error (FWE) rate, and a gene was identified to have a significant interaction with the CVD risk factors if the p-value was smaller than $0.05/21 \approx 2.38 \times 10^{-3}$. Analogous to the simulation studies, we compared the proposed kernel machine based method to six univariate interaction tests based on different dimension reduction methods.

[2] We note that the FHS risk score was derived from real biological data. Thus the FHS risk score is likely suboptimal for detecting the simulated effects of the CVD risk factors on the phenotype.

In order to reveal the direction of a significant interaction effect, we collapsed genetic variables and CVD risk factors into scalar variables; the RVT burden variable and the FHS risk score, respectively, and defined four regimes, low genetic risk and low CVD risk, high genetic risk and low CVD risk, low genetic risk and high CVD risk, high genetic risk and high CVD risk, by splitting the data with respect to the medians of the RVT burden variable and the FHS risk score. We then averaged the estimated interaction effect $\hat{\boldsymbol{h}}_{G \times W}$ within each of the four regimes. A smaller average indicates a higher risk of the interaction effect (smaller hippocampal volume). Jackknife resampling was used to get accurate standard error estimates of these average statistics. We also compared $\hat{\boldsymbol{h}}_{G \times W}$ between the 183 stable MCI subjects (who remained MCI throughout the follow-up) and the 151 MCI subjects that progressed to AD to investigate the predictive power of the interaction effect on disease progression.

Results

Simulation results

Table 2 shows the simulation results for the overall and interaction score tests. Here we used a nominal p-value threshold of 0.05. In more than 99% of the situations, the ReML algorithm converged within 50 iterations (convergence was declared when the difference between successive log ReML likelihoods was smaller than 10^{-4}), the maximum number of iterations we set in this simulation study, and in most cases it converged very quickly within 10 iterations and a few seconds with a MATLAB implementation on a MacBook Pro with 8 GB of memory and a 2.4 GHz Intel Core i7 processor.

It can be seen that when $\alpha_M = \alpha_I = 0$, the size of the overall score test is close to the nominal p-value threshold of 0.05. When $\alpha_M > 0$ and $\alpha_I = 0$, the false positive rate of the score test for interaction effect is also well controlled. When $\alpha_I > 0$, the power of the interaction test quickly exceeds 0.90. In contrast, we observe that dimension reduction methods can have slightly inflated false positive rates and are dramatically under-powered when compared to the kernel machine based method.

Fig. 1 shows the ROC curves of the kernel method and alternative methods for interaction detection, obtained with the second simulation data. The power gain of the kernel method relative to the alternative methods is evident.

Application to ADNI data

APOE-ε4 status is significantly associated with hippocampal volume ($p = 3.97 \times 10^{-16}$), after controlling for ICV, age, gender and education.

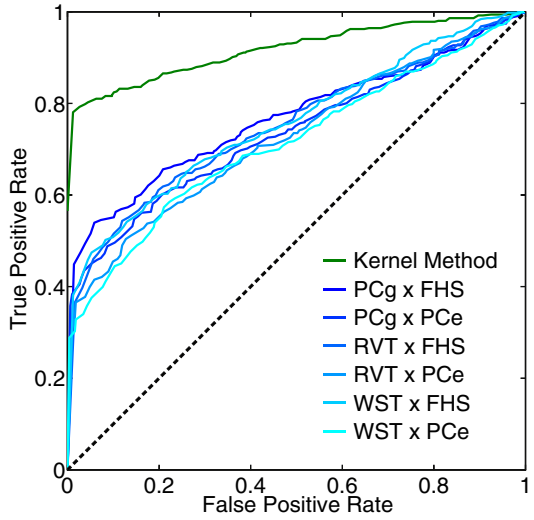

Fig. 1. Receiver operating characteristic (ROC) curves of the kernel method and alternative methods for interaction detection in the simulated data. False positive rates are plotted against true positive rates with the *p*-value threshold varying between 0 and 1 with a step size of 0.01.

Table 3 shows the main effects of the CVD risk factors, and their interaction effects with APOE-ε4 status on hippocampal volume obtained by conventional interaction analyses, as well as the association between diagnosis (AD patients vs. cognitive normal controls) and each CVD risk factor obtained by logistic regression analyses. As expected, the association between age and hippocampal volume is highly significant, indicating the reduction in the size of the hippocampus over time. The FHS CVD risk score is also significantly associated with hippocampal volume. Specifically, higher CVD risk scores suggest smaller hippocampal volumes. Age also shows a suggestive significant interaction with APOE-ε4 status but did not survive a Bonferroni correction for the total number of statistical tests performed here.

Table 4 lists the ReML estimates of τ_G^2, τ_W^2, $\tau_{G\times W}^2$ and σ^2, and the *p*-values for the interaction effects between each of the 21 candidate AD risk genes and the CVD risk factors on hippocampal volume. Two genes, *CR1* ($p = 4.85 \times 10^{-4}$) and *EPHA1* ($p = 5.64 \times 10^{-4}$), are identified to have significant interaction with the CVD risk factors.

Fig. 2 shows the average of the estimated interaction effect $\hat{\boldsymbol{h}}_{G\times W}$ within each of the four regimes (low genetic risk and low CVD risk, high genetic risk and low CVD risk, low genetic risk and high CVD risk, high genetic risk and high CVD risk) for the two genes *CR1* and *EPHA1* that show a significant interaction effect. For both genes, CVD risks largely dominate the interaction effect with higher CVD risk associated with higher risk of interaction and vice versa. The genetic risk appears to have an opposite effect of its marginal effect on interaction under high CVD risk, i.e., high genetic risk reduces the interaction effect in the presence of high CVD risk. One interpretation of this interaction pattern is that under low genetic risk, CVD risk factors have a more detrimental effect.

Two-sample *t*-tests showed that subjects with stable MCI have significantly larger $\hat{\boldsymbol{h}}_{G\times W}$ (lower risk) than subjects that progressed to AD for both genes (*CR1*, $p = 0.049$; *EPHA1*, $p = 0.044$), suggesting that disease progression is predicted by the interaction effect.

Comparison to alternative methods

Table 4 also shows the *p*-values for the alternative methods to test the interaction effect. PCA on both the genetic data and CVD risk factors, followed by multiple linear regression analyses, also identified *CR1* with a FWE-corrected significant *p*-value, but failed to find *EPHA1*. Other alternative methods did not identify any significant interaction effect.

Discussion

In this paper, we have proposed a kernel machine based method to test for interactions between multidimensional variable sets. Compared to traditional collapsing and PCA-based methods, the proposed method provides a more flexible and biological plausible way to model epistasis between genetic variants, accommodates multiple factors that potentially moderate genetic effects, and can test for complex interaction effects between multidimensional variable sets. Although multivariate methods typically produce more powerful and reproducible results, which can also be biologically more insightful, the interpretation of model parameters is often challenging. In this paper, we made some preliminary attempts to reveal the direction of interaction between multidimensional variable sets and investigate the prediction of disease progression by interaction effects. Further improvement of model interpretation would be facilitated by incorporating more biological information when a better understanding of the underlying mechanisms is achieved.

One particular case where model interpretation might be straightforward is when we use a linear kernel, as we did to model non-genetic effects. In our analyses, the non-genetic effect $\boldsymbol{h}_W$ can be represented as a linear combination of the CVD risk factors: $\boldsymbol{h}_W = \boldsymbol{W}\boldsymbol{\beta}_W$, where $\boldsymbol{W} = [\boldsymbol{W}_1, \cdots, \boldsymbol{W}_N]^T$ are the individual CVD risk factors. The linear coefficients $\boldsymbol{\beta}_W$ reflect the influence of each variable on the phenotype. The covariance matrix of the coefficient estimates can be computed as

$$\mathrm{cov}\left(\hat{\boldsymbol{\beta}}_W - \boldsymbol{\beta}_W\right) = \left(\tau_W^2/R\right)\boldsymbol{I} - \left(\tau_W^2/R\right)^2 \boldsymbol{W}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{W}, \tag{19}$$

where R is the number of non-genetic variables, $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$, and $\boldsymbol{V} = \tau_G^2\boldsymbol{K}_G + \tau_W^2\boldsymbol{K}_W + \tau_{G\times W}^2\boldsymbol{K}_{G\times W} + \sigma^2\boldsymbol{I}$. An estimate of this covariance structure can be obtained by inserting the ReML estimates of the variance component parameters τ_G^2, τ_W^2, $\tau_{G\times W}^2$ and σ^2 into Eq. (19), assuming that the error of the ReML estimation can be ignored. Supplementary Table S1 presents the point estimates

Table 3
Results of standard regression analyses with two different outcomes: hippocampal volume and Alzheimer's disease (AD) diagnosis (AD vs. control). The *p*-values for the main effects of the Framingham Heart Study (FHS) cardiovascular disease (CVD) risk score and each CVD risk factor, and their interaction effects with APOE-ε4 status on hippocampal volume obtained by conventional interaction analyses are presented. The *p*-values for the association between diagnosis (AD patients vs. cognitive normal controls) and each CVD risk factor obtained by logistic regression analyses are shown. Significant associations, with Bonferroni corrected *p*-values smaller than 0.05, are highlighted in bold.

Risk factor	Covariates adjusted	Hippocampal volume (linear regression)		AD vs. control (logistic regression)
		Main effect	Interaction with APOE-ε4	
FHS risk score	ICV, edu	$\mathbf{5.12 \times 10^{-4}}$	0.132	0.761
Age	ICV, edu, gender	$\mathbf{3.99 \times 10^{-18}}$	4.03×10^{-3}	0.301
Gender	ICV, edu, age	0.103	0.982	0.467
Body mass index (BMI)	ICV, edu, age, gender	2.37×10^{-3}	0.227	0.011
Systolic blood pressure	ICV, edu, age, gender	0.832	0.591	0.077
Smoking status	ICV, edu, age, gender	0.062	0.112	0.974
Diabetes	ICV, edu, age, gender	0.609	0.541	0.247

Table 4
Results of the multivariate interaction analyses. The restricted maximum likelihood (ReML) estimates of τ_G^2, τ_W^2, $\tau_{G\times W}^2$ and σ^2, and the *p*-values for the interaction effects between each of the 21 candidate Alzheimer's disease (AD) risk genes and the cardiovascular disease (CVD) risk factors on hippocampal volume, using the kernel method and the alternative methods based on dimension reduction and multiple regression, are shown. *p*-values that survive multiple testing corrections are highlighted in bold.

Gene	Kernel method					Alternative methods					
	τ_G^2	τ_W^2	$\tau_{G\times W}^2$	σ^2	*p*-value	PCg × FHS	PCg × PCw	RVT × FHS	RVT × PCw	WST × FHS	WST × PCw
ABCA7	3.44E-7	3.24E-2	9.03E-4	0.286	0.479	0.138	0.253	0.215	0.183	0.291	0.258
BIN1	6.41E-4	1.66E-2	2.03E-2	0.279	0.167	0.861	0.615	0.165	0.296	0.049	0.441
CASS4	2.17E-3	3.31E-2	3.44E-7	0.284	0.492	0.522	0.889	0.971	0.565	0.742	0.838
CD2AP	3.44E-7	4.31E-2	3.44E-7	0.290	0.954	0.150	0.737	0.145	0.618	0.229	0.728
CD33	7.76E-5	4.39E-2	3.44E-7	0.289	0.845	0.196	0.245	0.807	0.543	0.734	0.807
CELF1	3.44E-7	2.74E-2	6.52E-3	0.284	0.284	0.591	0.626	0.102	0.115	0.230	0.216
CLU	1.62E-3	3.60E-2	3.44E-7	0.286	0.602	0.920	0.998	0.140	0.479	0.154	0.563
CR1	3.44E-7	7.21E-3	4.68E-2	0.273	**4.85E-4**	0.559	**4.08E-4**	0.268	0.023	0.354	0.109
DSG2	3.44E-7	3.51E-2	3.44E-7	0.286	0.566	0.232	0.135	0.823	0.367	0.617	0.202
EPHA1	3.44E-7	3.44E-7	5.12E-2	0.273	**5.64E-4**	0.323	0.133	0.182	0.270	0.939	0.979
FERMT2	2.05E-2	3.17E-2	5.93E-3	0.278	0.342	0.556	0.817	0.767	0.794	0.852	0.876
HLA-DRB5	3.44E-7	3.80E-2	3.44E-7	0.286	0.759	0.764	0.388	0.763	0.388	0.763	0.388
INPP5D	6.18E-4	3.44E-2	3.44E-7	0.285	0.527	0.910	0.395	0.267	0.022	0.375	0.156
MEF2C	1.74E-3	2.08E-2	1.77E-2	0.279	0.058	0.125	0.408	0.085	0.436	0.067	0.418
MS4A6A	9.32E-4	3.77E-2	3.44E-7	0.287	0.902	0.769	0.969	0.779	0.945	0.781	0.926
PICALM	1.89E-2	3.91E-2	3.44E-7	0.281	0.679	0.473	0.295	0.580	0.390	0.922	0.294
PTK2B	1.68E-3	3.25E-2	9.00E-4	0.284	0.460	0.779	0.615	0.630	0.406	0.043	0.048
REST	3.66E-2	2.11E-2	1.95E-2	0.273	0.189	0.732	0.497	0.800	0.578	0.694	0.671
SLC24A4	8.22E-3	1.64E-2	2.35E-2	0.277	0.211	0.345	0.029	0.634	0.222	0.942	0.268
SORL1	2.30E-3	4.77E-2	3.44E-7	0.290	0.919	0.663	0.521	0.129	0.481	0.126	0.455
ZCWPW1	3.44E-7	2.85E-2	3.80E-3	0.284	0.265	0.634	0.373	0.600	0.422	0.574	0.437

PCg: first principal component of the genetic data; PCw: first principal component of the CVD risk factors; RVT: rare variant test burden variable; WST: weighted sum test burden variable; FHS: the Framingham Heart Study vascular disease risk score.

and standard errors for each element of β_W, for the ADNI analyses corresponding to each one of the 21 candidate AD risk genes. The above strategy does not apply to nonlinear kernels, but individual subjects can be examined by inspecting the estimated main and interaction effects $\hat{\boldsymbol{h}}_G$, $\hat{\boldsymbol{h}}_W$, $\hat{\boldsymbol{h}}_{G\times W}$, and their variabilities. More specifically, $\mathrm{cov}(\hat{\boldsymbol{h}}_{G\times W} - \boldsymbol{h}_{G\times W}) = \tau_{G\times W}^2\boldsymbol{K}_{G\times W} - (\tau_{G\times W}^2\boldsymbol{K}_{G\times W})\boldsymbol{P}(\tau_{G\times W}^2\boldsymbol{K}_{G\times W})$, and the variability of $\hat{\boldsymbol{h}}_G$ and $\hat{\boldsymbol{h}}_W$ can be quantified analogously. Analyses of individual subjects may provide additional information about the model, but we consider this beyond the scope of the present paper.

Due to the moderate sample size in the present study, we constrained our analysis to a list of candidate late-onset AD risk genes. However, the proposed method can be applied to genome-wide interaction studies. In particular, we note that when testing for the overall genetic and non-genetic effect, the variance component parameters τ_G^2, τ_W^2, and $\tau_{G\times W}^2$ need not be estimated. Therefore, the overall score test offers an efficient and non-iterative approach to screen the whole genome for genetic variants that might show significant contribution to the phenotypic variation. Fitting the full model, estimating the variance components, and testing for interactions can then focus on genetic components with significant overall effect, which will dramatically reduce computational burden. A similar argument applies to voxel-/vertex-wise interaction studies.

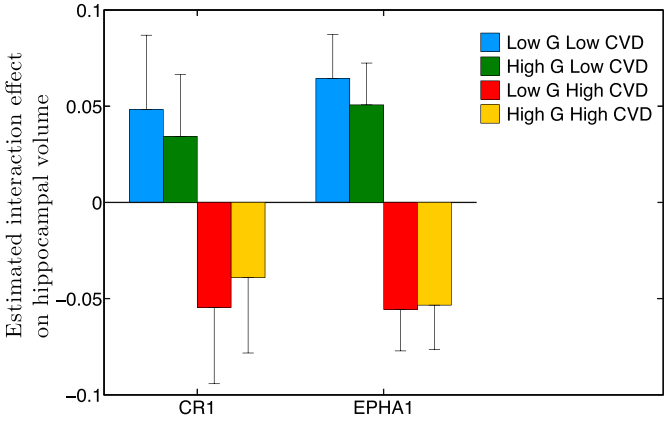

Fig. 2. Direction of significant interaction effects. For genes *CR1* and *EPHA1* that show significant interaction effect, genetic variables and cardiovascular disease (CVD) risk factors were collapsed into scalar variables; the RVT burden variable and the FHS risk score, respectively. The average of the estimated interaction effect $\hat{\boldsymbol{h}}_{G\times W}$ within each of the four regimes (low genetic risk and low CVD risk, high genetic risk and low CVD risk, low genetic risk and high CVD risk, high genetic risk and high CVD risk) is shown with a standard error estimate obtained by Jackknife resampling. A smaller average indicates a higher risk of the interaction effect (smaller hippocampal volume).

We would like to note that most of the CVD risk factors we employed in our ADNI analyses are largely endogenous and thus are, to some extent, under genetic control. Although, this might make the interpretation of the results difficult, this challenge, we believe, exists in many interaction effects probed and detected in the genetics literature. Furthermore, even though the non-genetic variables we used are collectively associated with cardiovascular risk, and thus our interpretation of the detected interaction effects as genetic influences modified by cardiovascular risk is highly likely, alternative explanations that do not involve cardiovascular mechanisms are also possible. Finally, while hippocampal volume is a sensitive biomarker of AD, it is not solely related to this condition. In fact, we conducted additional analyses with entorhinal cortex thickness and volume (also MRI markers of late-onset AD) as alternative outcome variables. These analyses (not included here) did not reveal any statistically significant interaction effect. Although our presented results demonstrated that the detected interaction effects with hippocampal volume predict future MCI-to-AD conversion, one possibility is that these associations might not be specific to AD. Elucidating these issues is beyond the scope of this paper and will require careful follow-up studies that will consider all alternative possibilities.

Another potential concern in the present study is that we took the coding regions and 20 kb up/downstream of the 21 candidate genes as units for interaction detection. Although Lambert et al. (2013) examined all SNPs that have strong associations with the top SNPs to confirm the relevance of these genes, we are aware that they are likely not the causative genes. Also, the size of the regulatory region of different genes may vary substantially. Therefore, an alternative strategy is to group SNPs in high LD with the most associated SNP, whether or not they are in or close to the nearest gene.

The choice of kernels may have an impact on the validity and power of the method too. In the present study, we employed an IBS kernel for

the genetic data and a linear kernel for the CVD risk factors, as both kernels are parameter free and can in principle capture complex epistatic effects between genetic variants and model the joint effect of multiple non-genetic variables. We found, through simulation studies, that the proposed selection of kernels appears to work well in our setting, both in terms of false positive rate control and statistical power. Using other kernel functions, e.g., the Gaussian kernel for combining non-genetic factors, is certainly possible, but might require preselecting or estimating additional parameters. Our preliminary implementation (results not shown) suggests that incorporating the estimation of the spread parameter in the Gaussian kernel into the ReML algorithm might lead to unstable estimates and failure of convergence. The performance of various kernel functions in different data structures, and the optimal selection of kernels, deserve future investigation.

Although we illustrated the proposed method using a univariate quantitative image derived phenotype, genes as units to group SNPs, and CVD risk factors as non-genetic variables, the modeling framework is general and can be applied to other types of phenotypic and genetic data, and to detecting other types of interactions such as genotype-by-environment interactions. Our method can also be extended to accommodate binary outcomes, and thus has potential wide applications to case–control studies. Recently, a series of papers have been published on the proper modeling of longitudinal and time-to-event data in neuroimaging studies (Bernal-Rusiel et al., 2013a,b; Sabuncu et al., 2014). Incorporating genetic components and interactions in longitudinal and survival models, and investigating the genetic contributions to the progression of a brain-related illness and the timing of a clinical event of interest, seem promising directions for future research.

Two genes, *CR1* and *EPHA1*, were identified to have significant interaction effects with the CVD risk factors in the present study. The associations between the two genes and AD have been identified and replicated by a number of independent studies (see e.g., Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011), in addition to Lambert et al. (2013), and their potential contributions to the mechanism of AD have been under active investigation (Biffi et al., 2012; Chibnik et al., 2011; Thambisetty et al., 2013). Moreover, recent studies show that many of the AD risk genes have potential roles in relationship with CVD risk factors, such as hypertension, hypercholesterolemia, and obesity (Guerreiro et al., 2012; Liu et al., 2014). In particular, excess adiposity may act as an enhanced substrate for *CR1*-related inflammatory events (Guerreiro et al., 2012). Our findings indicate that genetic components may contribute to the etiology of late-onset AD in the presence of CVD risks, and warrant further investigations.

Acknowledgments

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, and the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; 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; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of 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.

This research was carried out in whole or in part at Athinoula A. Martinos Center for Biomedical Imaging at Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institutes of Health (NIH).

This research was also funded in part by NIH grants R01 EB015611-01 and U54 MH091657-03 (TEN), R01 NS083534, and R01 NS070963, and NIH NIBIB 1K25EB013649-01 (MRS), K24MH094614 and R01 MH101486 (JWS), Wellcome Trust grants 100309/Z/12/Z and 098369/Z/12/Z (TEN), and a BrightFocus Foundation grant AHAF-A2012333 (MRS). Two additional R01 grants from the National Institute on Aging (R01 AG008122 and R01 AG016495) provided partial support for this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.01.029.

References

1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 4910 (7422), 56–65.
Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Am. Math. Soc. 680 (3), 337–404.
Aschard, H., Hancock, D.B., London, S.J., Kraft, P., 2011. Genome-wide meta-analysis of joint tests for genetic and gene–environment interaction effects. Hum. Hered. 700 (4), 292–300.
Bernal-Rusiel, L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., 2013a. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. NeuroImage 66, 249–260.
Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R., 2013b. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81, 358–370.
Biffi, A., Anderson, C.D., Desikan, R.S., Sabuncu, M., Cortellini, L., et al., 2010. Genetic variation and neuroimaging measures in Alzheimer disease. Arch. Neurol. 670 (6), 677–685.
Biffi, A., Shulman, J.M., Jagiella, J.M., Cortellini, L., Ayres, A.M., et al., 2012. Genetic variation at CR1 increases risk of cerebral amyloid angiopathy. Neurology 780 (5), 334–341.
Borenstein, A.R., Wu, Y., Mortimer, J.A., Schellenberg, G.D., McCormick, W.C., et al., 2005. Developmental and vascular risk factors for Alzheimer's disease. Neurobiol. Aging 260 (3), 325–334.
Bralten, J., Franke, B., Arias-Vásquez, A., Heister, A., Brunner, H.G., et al., 2011. CR1 genotype is associated with entorhinal cortex volume in young healthy adults. Neurobiol. Aging 320 (11), 2106-e7 –2106-e11.
Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., et al., 2002. Role of genotype in the cycle of violence in maltreated children. Science 2970 (5582), 851–854.
Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., et al., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 3010 (5631), 386–389.
Chibnik, L.B., Shulman, J.M., Leurgans, S.E., Schneider, J.A., Wilson, R.S., et al., 2011. CR1 is associated with amyloid plaque burden and age-related cognitive decline. Ann. Neurol. 690 (3), 560–569.
D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., et al., 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 1170 (6), 743–753.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage 90 (2), 179–194.
Fischl, B., 2012. Freesurfer. NeuroImage 620 (2), 774–781.
Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. NeuroImage 90 (2), 195–207.
Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 330 (3), 341–355.
Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., et al., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 140 (1), 11–22.
Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 60 (3), 218–229.
Furney, S.J., Simmons, A., Breen, G., Pedroso, I., Lunnon, K., et al., 2010. Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease. Mol. Psychiatry 160 (11), 1130–1138.
Ge, T., Feng, J., Hibar, D.P., Thompson, P.M., Nichols, T.E., 2012. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. NeuroImage 63, 858–873.
Ge, T., Schumann, G., Feng, J., 2014. Imaging genetics — towards discovery neuroscience. Quant. Biol. 10 (4), 1–19.

Gottesman, I.I., Gould, T.D., 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. Am. J. Psychiatr. 1600 (4), 636.
Gottesman, I.I., Shields, J., 1972. Schizophrenia genetics: a twin study vantage point. Academic Press, New York.
Gu, C., 2002. Smoothing Spline ANOVA Models. Springer.
Gu, C., Wahba, G., 1993. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". J. Comput. Graph. Stat. 20 (1), 97–117.
Guerreiro, R.J., Gustafson, D.R., Hardy, J., 2012. The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. Neurobiol. Aging 330 (3), 437–456.
Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., et al., 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. 41, 1088–1093.
Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 720 (358), 320–338.
Helzner, E.P., Luchsinger, J.A., Scarmeas, N., Cosentino, S., Brickman, A.M., et al., 2009. Contribution of vascular risk factors to the progression in Alzheimer disease. Arch. Neurol. 660 (3), 343–348.
Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J., et al., 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. 430 (5), 429–435.
Hsu, L., Jiao, S., Dai, J.Y., Hutter, C., Peters, U., et al., 2012. Powerful cocktail methods for detecting genome-wide gene-environment interaction. Genet. Epidemiol. 360 (3), 183–194.
Hua, W., Ghosh, D., 2014. Equivalence of kernel machine regression and kernel distance covariance for multidimensional trait association studies. ArXiv (preprint arXiv: 1402.2679).
Irie, F., Fitzpatrick, A.L., Lopez, O.L., Kuller, L.H., Peila, R., et al., 2008. Enhanced risk for Alzheimer disease in persons with type 2 diabetes and APOE ε4: the Cardiovascular Health Study Cognition Study. Arch. Neurol. 650 (1), 89–93.
Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 330 (1), 82–95.
Kivipelto, M., Helkala, E., Laakso, M.P., Hänninen, T., Hallikainen, M., et al., 2001. Midlife vascular risk factors and Alzheimer's disease in 'later life: longitudinal, population based study. BMJ 3220 (7300), 1447–1451.
Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J., Gauderman, W.J., 2007. Exploiting gene-environment interaction to detect genetic associations. Hum. Hered. 630 (2), 111–119.
Kwee, L.C., Liu, D., Lin, X., Ghosh, D., Epstein, M.P., 2008. A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. 820 (2), 386–397.
Lambert, J.C., Heath, S., Even, G., Campion, D., Sleegers, K., et al., 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat. Genet. 41, 1094–1099.
Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., et al., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. 45, 1452–1458.
Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., et al., 2012a. Optimal unified approach for rare-variant association testing with application to small-sample case–control whole-exome sequencing studies. Am. J. Hum. Genet. 910 (2), 224–237.
Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., et al., 2012b. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. 440 (3), 247–250.
Li, S., Cui, Y., 2012. Gene-centric gene–gene interaction: a model-based kernel machine method. Ann. Appl. Stat. 60 (3), 1134–1161.
Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R., 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 340 (8), 816–834.
Lin, X., 1997. Variance component testing in generalised linear models with random effects. Biometrika 840 (2), 309–326.
Lin, X., Lee, S., Christiani, D.C., Lin, X., 2013. Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics 140 (4), 667–681.
Lindstrom, M.J., Bates, D.M., 1988. Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J. Am. Stat. Assoc. 830 (404), 1014–1022.
Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 630 (4), 1079–1088.
Liu, G., Yao, L., Liu, J., Jiang, Y., Ma, G., et al., 2014. Cardiovascular disease contributes to Alzheimer's disease: evidence from large-scale genome-wide association studies. Neurobiol. Aging 350 (4), 786–792.
Lo, R.Y., Jagust, W.J., Weiner, M., Aisen, P., Petersen, R., et al., 2012. Vascular burden and Alzheimer disease pathologic progression. Neurology 790 (13), 1349–1355.
Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., et al., 2014. REST and stress resistance in ageing and Alzheimer's disease. Nature 5070 (7493), 448–454.
Luchsinger, J.A., Reitz, C., Honig, L.S., Tang, M.X., Shea, S., et al., 2005. Aggregation of vascular risk factors and risk of incident Alzheimer disease. Neurology 650 (4), 545–551.
Madsen, B.E., Browning, S.R., 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 50 (2), e1000384.
Meyer-Lindenberg, A., Weinberger, D.R., 2006. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nat. Rev. Neurosci. 70 (10), 818–827.
Morris, A.P., Zeggini, E., 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 340 (2), 188–193.
Mukherjee, B., Chatterjee, N., 2008. Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 640 (3), 685–694.
Murcray, C.E., Lewinger, J.P., Conti, D.V., Thomas, D.C., Gauderman, W.J., 2011. Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet. Epidemiol. 350 (3), 201–210.
Naj, A.C., Jun, G., Beecham, G.W., Wang, L., Vardarajan, B.N., et al., 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. 430 (5), 436–441.
Paré, G., Cook, N.R., Ridker, P.M., Chasman, D.I., 2010. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet. 60 (6), e1000981.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 810 (3), 559–575.
Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., et al., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 4600 (7256), 748–752.
Purnell, C., Gao, S., Callahan, C.M., Hendrie, H.C., 2009. Cardiovascular risk factors and incident Alzheimer disease: a systematic review of the literature. Alzheimer Dis. Assoc. Disord. 230 (1), 1.
Qiu, C., Winblad, B., Fastbom, J., Fratiglioni, L., 2003. Combined effects of APOE genotype, blood pressure, and antihypertensive drug use on incident AD. Neurology 610 (5), 655–660.
Sabuncu, M.R., Buckner, R.L., Smoller, J.W., Lee, P.H., Fischl, B., et al., 2012. The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects. Cereb. Cortex 220 (11), 2653–2661.
Sabuncu, M.R., Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., 2014. Event time analysis of longitudinal neuroimage data. NeuroImage 97, 9–18.
Schaid, D.J., 2010a. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum. Hered. 700 (2), 109–131.
Schaid, D.J., 2010b. Genomic similarity and kernel methods II: methods for genomic information. Hum. Hered. 700 (2), 132–140.
Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., et al., 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 70 (3), 280–292.
Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., et al., 2010. Voxelwise genome-wide association study (vGWAS). NeuroImage 530 (3), 1160–1174.
Sullivan, P.F., Daly, M.J., O'Donovan, M., 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. 130 (8), 537–551.
Thambisetty, M., An, Y., Nalls, M., Sojkova, J., Swaminathan, S., et al., 2013. Effect of complement CR1 on brain amyloid burden during aging and its modification by APOE genotype. Biol. Psychiatry 730 (5), 422–428.
Thompson, P.M., Ge, T., Glahn, D.C., Jahanshad, N., Nichols, T.E., 2013. Genetics of the connectome. NeuroImage 80, 475–488.
Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery. Am. J. Hum. Genet. 900 (1), 7–24.
Wahba, G., 1990. Spline models for observational data. SIAM.
Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B., 1995. Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. Ann. Stat. 230 (6), 1865–1895.
Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., et al., 2013. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. 90 (5), e111–e194.
Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., et al., 2010. Powerful SNP-set analysis for case–control genome-wide association studies. Am. J. Hum. Genet. 860 (6), 929–942.
Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., et al., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 890 (1), 82–93.
Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., et al., 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 420 (7), 565–569.
Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., et al., 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 430 (6), 519–525.
Zlokovic, B.V., 2005. Neurovascular mechanisms of Alzheimer's neurodegeneration. Trends Neurosci. 280 (4), 202–208.