Check for
updates

Correlation Constraints for Regression Models: Controlling Bias in Brain Age Prediction

Matthias S. Treder[1]*, Jonathan P. Shock[2,3], Dan J. Stein[4], Stéfan du Plessis[5], Soraya Seedat[5] and Kamen A. Tsvetanov[6,7]

[1] School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom, [2] Department of Mathematics and Applied Mathematics, University of Cape Town, Cape Town, South Africa, [3] National Institute for Theoretical Physics, Matieland, South Africa, [4] SA MRC Unit on Risk & Resilience in Mental Disorders, Department of Psychiatry and Neuroscience Institute, University of Cape Town, Cape Town, South Africa, [5] Department of Psychiatry, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, [6] Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom, [7] Department of Psychology, University of Cambridge, Cambridge, United Kingdom

In neuroimaging, the difference between chronological age and predicted brain age, also known as *brain age delta*, has been proposed as a pathology marker linked to a range of phenotypes. Brain age delta is estimated using regression, which involves a frequently observed bias due to a negative correlation between chronological age and brain age delta. In brain age prediction models, this correlation can manifest as an overprediction of the age of young brains and an underprediction for elderly ones. We show that this bias can be controlled for by adding correlation constraints to the model training procedure. We develop an analytical solution to this constrained optimization problem for Linear, Ridge, and Kernel Ridge regression. The solution is optimal in the least-squares sense i.e., there is no other model that satisfies the correlation constraints and has a better fit. Analyses on the PAC2019 competition data demonstrate that this approach produces optimal unbiased predictive models with a number of advantages over existing approaches. Finally, we introduce regression toolboxes for Python and MATLAB that implement our algorithm.

Keywords: age, brain, optimization, prediction, correlation, regression

1. INTRODUCTION

As the world's population ages, early detection and prevention of neurological aspects of aging, such as cognitive decline and dementia, is a public health priority and challenge. Pathological aging could be indicated by the level of deviation from the typical pattern of aging in healthy individuals (1). There has been growing interest in developing statistical approaches in order to identify individuals deviating from a healthy brain aging trajectory (2). To this end, a metric referred to as *brain age delta*, defined as the difference between brain-predicted age and chronological age, has been proposed as an index of the level of neuropathology in aging (2–4). Investigating the association between this metric with demographics, and lifestyle and cognitive variables can deepen the understanding of the processes that underpin healthy aging (5). In clinical research, brain age delta has the potential to index the severity of premature aging in patients suffering from disease. Among others, a higher delta has been associated with lower fluid intelligence and higher mortality (1), risk for developing Alzheimer's disease (6), severity of schizophrenia and depression (7).

Establishing a good estimate of brain age delta is faced with important methodological challenges. The first challenge relates to the kinds of features (biomarkers) and predictive models that are used to build a brain-age model. A number of brain metrics have been considered as features for regression models; for example, structural networks (8), cortical thickness (9), functional connectivity patterns (10, 11), and raw T1-weighted images (1, 2). Likewise, a variety of regression models has been tested, from linear regression models such as lasso and support vector regression (10, 11) to convolutional neural networks [CNNs; (2, 8)]. The quest for more accurate brain-age models lies at the heart of the Predictive Analytics Competition (PAC) 2019 upon which the eponymous Frontiers Research Topic is founded[1].

A more fundamental methodological challenge, and the starting point for this paper, is the very operationalization of the brain age delta. If we denote the chronological age for a set of participants as a vector $\mathbf{y}$, the ages predicted on the brain scans as $\hat{\mathbf{y}}$, and the residuals as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, then the negative residual $\delta = -\mathbf{e}$ (i.e., predicted brain age minus chronological age) is usually defined as the brain age delta. This metric has been shown to be problematic. A predictive bias manifesting as an overprediction of the age of young individuals and an underprediction for elderly individuals has led to much speculation and investigation (2, 3, 9, 12–15).

A useful quantification of this effect is the correlation between chronological age and delta, corr$(\mathbf{y}, \delta)$, which we will refer to as *age delta correlation (ADC)* in the rest of the paper (see **Figures 1A,B**). An analysis by (15) showed that negative ADC is ubiquitous across a range of aging datasets and regression models and independent of the age range included in the dataset. A theoretical analysis by (14) showed that this effect is an inevitable property of regression, further aggravated by regression dilution (16, 17), and hence not limited to aging data. The potential danger of non-zero ADC lies in spurious associations with other covariates: brain age delta can be trivially correlated with demographic or cognitive variables if the latter are correlated with chronological age as well. Le et al. (14) found that associations between residuals and variables obtained from clinical interviews and neuropsychological testing largely disappear when residuals are corrected for chronological age. To avoid these spurious correlations, several authors have suggested following up the regression analysis with a correction step wherein the effect of age is removed from the residuals (1, 3, 12, 14). Brain age delta is then calculated by the following two-stage approach:
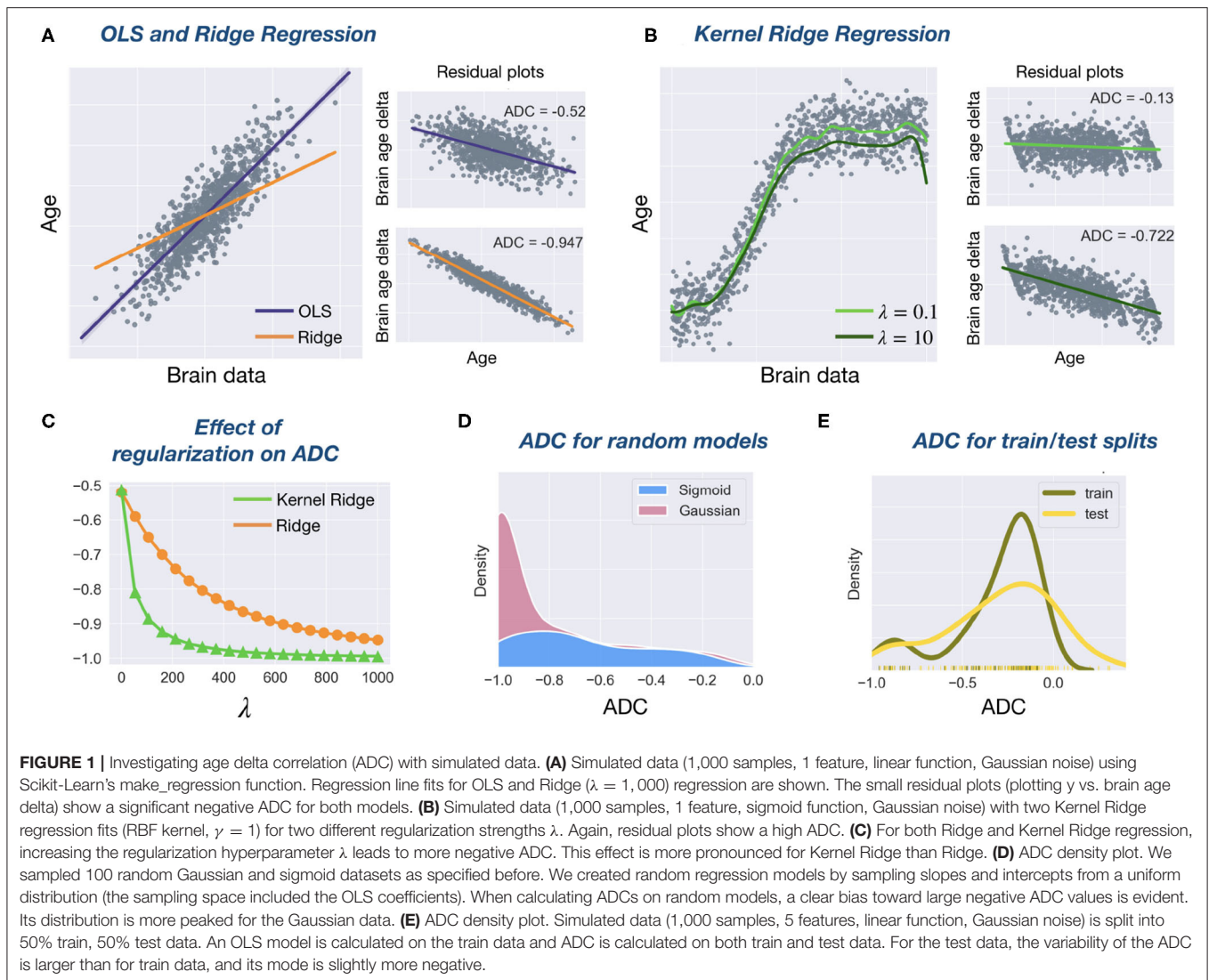
1. *Brain Age Prediction.* Train a regression model to predict age from neuroimaging data. The difference between predicted age and chronological age reflects *uncorrected brain age delta.*
2. *Correction of Brain Age Delta.* Use simple linear regression to regress delta against age. The resultant residuals are uncorrelated with age and denoted as *corrected brain age delta.*

Despite the significant methodological progress that has been made there are still concerns that warrant attention. First, the correction approach is an *ad hoc* fix because the models' predictions do not take ADC into account. Second, in a strictly sequential two-stage approach it is not clear whether the resultant brain age delta is optimal in terms of predictive accuracy. Both issues can be addressed when prediction and correction are unified within a model. Third, in predictive settings with training and test set, zeroing ADC on the training set is not of primary importance. Rather, it would be useful to have a model that is able to finely control the trade-off between ADC and predictive accuracy in order to optimize its performance on the test set.

The aim of this study was to address these points by introducing modifications to three different regression models (Linear, Ridge and Kernel Ridge regression). Our models explicitly control for ADC on the training set without the need for an additional correction after model training. This was realized by formulating constrained optimization problems that incorporate age delta correlation as additional constraints. Our approach offers the following features:

- *Predictive Model.* In predictive modeling, all properties of the estimation pipeline (including correction of the residuals) should be derived from the training set and validated on a separate test set. Some of the existing approaches conflate training and test data because age prediction is based on the training data but the correction is performed on the test set. The latter introduces dependencies between test samples because the correction applied to a test sample depends on the other test samples.
- *Arbitrary Test Set Size.* An additional problem arising from conflating training and test sets is that performing correction on test data requires a sufficiently large test set. This is especially problematic with smaller datasets because less data is available for training. Since our approach estimates all parameters from the training data, it can be applied with any train/test split including leave-one-out cross-validation, when appropriate.
- *Prediction of Unlabeled Data.* Our model corrects the predictions not the residuals. Therefore, corrected predictions can be obtained even on unlabeled data if necessary.
- *Optimality.* Formulating both model training and correction as a single constrained optimization problem allowed us to show that the resultant models are optimal in terms of mean-squared error (MSE) on the training set. In other words, of all potential solutions that control ADC, our solution has the highest accuracy.
- *Correlation Bound.* Our models allow for "soft" control of the ADC by defining a correlation bound that caps the maximum permissible correlation, e.g., $|\text{corr}(\mathbf{y}, \delta)| \leq 0.1$. This is especially useful for predictive modeling, because minimizing ADC on the training set is not vital. Rather, a low bias and good predictive performance on the test set is desired. Using a correlation bound, our models allow for fine-tuning of the trade-off between ADC and predictive accuracy.
- *Interpretability.* An advantage of unifying prediction and correction in a single model is better interpretability because

---

FIGURE 1 | Investigating age delta correlation (ADC) with simulated data. **(A)** Simulated data (1,000 samples, 1 feature, linear function, Gaussian noise) using Scikit-Learn's make_regression function. Regression line fits for OLS and Ridge ($\lambda = 1,000$) regression are shown. The small residual plots (plotting y vs. brain age delta) show a significant negative ADC for both models. **(B)** Simulated data (1,000 samples, 1 feature, sigmoid function, Gaussian noise) with two Kernel Ridge regression fits (RBF kernel, $\gamma = 1$) for two different regularization strengths λ. Again, residual plots show a high ADC. **(C)** For both Ridge and Kernel Ridge regression, increasing the regularization hyperparameter λ leads to more negative ADC. This effect is more pronounced for Kernel Ridge than Ridge. **(D)** ADC density plot. We sampled 100 random Gaussian and sigmoid datasets as specified before. We created random regression models by sampling slopes and intercepts from a uniform distribution (the sampling space included the OLS coefficients). When calculating ADCs on random models, a clear bias toward large negative ADC values is evident. Its distribution is more peaked for the Gaussian data. **(E)** ADC density plot. Simulated data (1,000 samples, 5 features, linear function, Gaussian noise) is split into 50% train, 50% test data. An OLS model is calculated on the train data and ADC is calculated on both train and test data. For the test data, the variability of the ADC is larger than for train data, and its mode is slightly more negative.

the entire operation of the model is represented by its regression coefficients and quantities derived from them (18–20). Furthermore, we show in Section 2.7 that these quantities are not affected by the choice of the correlation bound (a hyperparameter in our model).

Some of the existing approaches share some of the listed features. For instance, whereas in (14) test set residuals are used for correction, Beheshti et al. (12) learns the correction parameters from the training set and applies them to the test set, in line with good practice for predictive models. However, to the best of our knowledge, this is the first study to prove optimality of the models and introduce "soft" correlation bounds for fine control of ADC.

2. METHOD

The section 2 is organized as follows. In Section 2.2, we introduce Linear regression, Ridge regression and Kernel Ridge regression.

In Section 2.3, we review existing ways to quantify brain age delta. In Section 2.4, we revisit the mathematical basis for ADC in the context of the three regression models. In Section 2.5 we develop our approach by adding correlation constraints to the model training stage that allow for a precise control of ADC. In Section 2.6 it is shown that the brain age estimates obtained with our models are closely related to existing correction approaches. In Section 2.7, we investigate how to interpret the models. In Section 2.8, we introduce corresponding toolboxes for Python and MATLAB that implement them. Finally, in sections 2.9 and 2.10, we describe our analysis of the PAC2019 competition data.

2.1. Notation

Table 1 defines the most important mathematical symbols used in the paper. Whenever the symbol represents a vector or matrix, its dimensionality is given in the second column. In general, matrices are denoted as uppercase boldface symbols (e.g., $\mathbf{X}$, $\mathbf{H}$), vectors as lowercase boldface symbols (e.g., $\boldsymbol{\beta}$, $\bar{\mathbf{x}}$), and scalars as lowercase normal face symbols (e.g., n, $\bar{y}$).

n	$\mathbb{N}$	Number of samples
p	$\mathbb{N}$	Number of features
$\boldsymbol{\beta}$	$\mathbb{R}^{p+1}$	Coefficients for standard regression model
$\boldsymbol{b}$	$\mathbb{R}^{p+1}$	Coefficients for correlation constrained model
$\mathbf{X}$	$\mathbb{R}^{n \times p}$	Brain scans (features)
$\mathbf{y}$	$\mathbb{R}^{n}$	Chronological age (targets)
$\hat{\mathbf{y}}$	$\mathbb{R}^{n}$	Predicted age ($\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$)
$\mathbf{e}$	$\mathbb{R}^{n}$	Residuals ($\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$)
$\boldsymbol{\delta}$	$\mathbb{R}^{n}$	Uncorrected brain-age delta (negative residuals $\boldsymbol{\delta} = -\mathbf{e}$)
ρ	$\mathbb{R}$	Correlation bound (hyperparameter)
$\top$		Transpose operator
$\mathbf{1}$	$\mathbb{R}^{n}$	Vector of 1's
$\bar{\mathbf{x}}$	$\mathbb{R}^{p}$	Column means of $\mathbf{X}$ given by $\frac{1}{n}\mathbf{X}^{\top}\mathbf{1}$
$\bar{y}$	$\mathbb{R}$	Mean of $\mathbf{y}$

Note that we use terminology common in the machine learning literature. In particular, *features* are also known as predictors or independent variables in the regression literature, the vector of *targets* (chronological age) is also known as response vector or dependent variable, and *training* is also known as fitting. For standard regression models, regression coefficients are denoted as $\boldsymbol{\beta}$. If an intercept is included in the model, we assume that a column of 1's is added to the matrix of features $\mathbf{X}$. Sometimes we explicitly denote the intercept as $\boldsymbol{\beta}_0$ and the non-intercept coefficients as $\boldsymbol{\beta}_{1:p}$. For models with correlation constraints, we use the notation $\boldsymbol{b}$ for the regression coefficients with $\boldsymbol{b}_0$ and $\boldsymbol{b}_{1:p}$ defined analogously.

2.2. Regression Models
2.2.1. Ordinary Least-Squares (OLS) Regression
Ordinary least-squares regression, often just called linear regression, can be formulated as a set of n equations of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \qquad i = 1, 2, ..., n$$

where y_i is the i-th response, x_{ij} is the $j-th$ predictor value in the i-th sample, the β's are regression coefficients, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is an error term. Using matrix notation, this set of equations can be written more succinctly as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (1)$$

where $\mathbf{y} = [y_1, y_2, ..., y_n]^{\top}$ comprises the responses, $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ is the matrix of features including a column of 1's for the intercept term, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the vector of regression coefficients and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_n]^{\top}$ collects all error terms. Training a model implies finding an estimate for $\boldsymbol{\beta}$ such that $\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta}$. In OLS regression this is achieved by minimizing the sum of squared errors $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$. Denoting $\hat{\mathbf{y}} := \mathbf{X}\boldsymbol{\beta}$ this can be formulated as the unconstrained optimization problem

$$\text{minimize} \quad ||\mathbf{y} - \hat{\mathbf{y}}||_2^2. \qquad (2)$$

The solution is given by

$$\boldsymbol{\beta}_{\text{ols}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}.$$

2.2.2. Ridge Regression
OLS regression suffers from high variance and does not have a unique solution if the number of the features p is larger than the number of samples n (21, 22). Ridge regression is a regularized version of OLS regression that is useful for data that suffers from multicollinearity. The model is regularized by adding an ℓ_2 penalty that shrinks the weights toward zero. For a given regularization parameter $\lambda \geq 0$, Ridge regression can be formulated as the unconstrained optimization problem

$$\text{minimize} \quad ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 + \lambda||\boldsymbol{\beta}_{1:p}||_2^2.$$

The first term is the least-squares term from Equation (2). The second term penalizes elements of $\boldsymbol{\beta}$ from becoming too large. For $\lambda = 0$ Ridge regression reduces to OLS regression. The solution is given by

$$\boldsymbol{\beta}_{\text{ridge}} = (\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{y} \qquad (3)$$

where $\mathbf{I} \in \mathbb{R}^{(p+1) \times (p+1)}$ is an identity matrix. Since the intercept term is not regularized, $\mathbf{I}$ is modified such that the 1 in the first row/column is replaced by a 0.

2.2.3. Other Linear Regression Models
Other variants of linear regression e.g., lasso (23) and elastic net (24), do not have a closed form solution but rely on iterative optimization, so they do not lend themselves to the analytical approach developed in this paper.

2.2.4. Kernel Ridge Regression (KRR)
A non-linear version of Ridge regression can be developed by applying a non-linear transformation to the features and then performing Ridge regression on these transformed features (25). Let this transformation be represented by a map $\phi : \mathbb{R}^p \rightarrow \mathcal{F}$ from input space to a higher-dimensional Reproducing Kernel Hilbert Space and $\Phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_n)]^{\top}$ (26, 27). The solution is given by replacing $\mathbf{X}$ by $\Phi(\mathbf{X})$ in Equation (3),

$$\boldsymbol{\beta}_{\text{krr}} = (\Phi(\mathbf{X})^{\top}\Phi(\mathbf{X}) + \lambda\mathbf{I})^{-1}\Phi(\mathbf{X})^{\top}\mathbf{y}.$$

This solution, also known as primal solution, is of limited practical use, since the feature space is often too high-dimensional to represent $\boldsymbol{\beta}_{\text{krr}}$ and $\Phi(\mathbf{X})$. As an alternative, the convex optimization problem can be rewritten into its dual Lagrangian form first (28). The resultant dual solution is given by

$$\boldsymbol{\beta}_{\text{krr}} = \Phi(\mathbf{X})^{\top}(\Phi(\mathbf{X})\Phi(\mathbf{X})^{\top} + \lambda\mathbf{I})^{-1}\mathbf{y}. \qquad (4)$$

The equivalence between the primal and dual solution can be verified by left-multiplying both solutions with $(\Phi(\mathbf{X})\Phi(\mathbf{X})^{\top} + $

$\lambda \mathbf{I}$). Since $\boldsymbol{\beta}_{\mathrm{krr}}$ cannot be represented directly, we instead calculate a vector of dual weights $\boldsymbol{\alpha} \in \mathbb{R}^n$. To this end, define $\mathbf{K} = \Phi(\mathbf{X})\,\Phi(\mathbf{X})^\top$ as the kernel matrix with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for a kernel function k. Then the vector of dual weights is given by the latter part of Equation (4),

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\,\mathbf{y}. \qquad (5)$$

Using $\boldsymbol{\alpha}$ the predicted response to a test sample $\mathbf{x}$ can be rewritten in terms of kernel evaluations:

$$f(\mathbf{x}) = \boldsymbol{\beta}_\phi^\top\,\phi(\mathbf{x}) = \boldsymbol{\alpha}^\top\Phi(\mathbf{X})\,\phi(\mathbf{x}) = \sum_{i=1}^n \alpha_i\,k(\mathbf{x}_i, \mathbf{x}). \qquad (6)$$

2.3. Calculating the Brain-Age Delta

The regression of age on brain features often leads to a biased model that manifests as an overprediction of the age of younger individuals and an underprediction of the age of elderly ones. This effect can be quantified as a negative age delta correlation (ADC) denoted as $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta})$. In the literature, ADC has been set to zero by adding a second stage to the analysis wherein the regression predictions from the first stage are corrected. Hence, brain-age delta prediction can be formulated as the following two-stage approach:

(a) *Brain Age Prediction.* Train a regression model f to predict age such that $\mathbf{y} \approx f(\mathbf{X})$. The negative residuals, denoted as

$$\boldsymbol{\delta} = f(\mathbf{X}) - \mathbf{y} \qquad (7)$$

represent the uncorrected brain age delta.

(b) *Correction of Brain Age Delta.* A number of authors proposed correction procedures to rid $\boldsymbol{\delta}$ of ADC (1, 3, 4, 12, 14). Many of these approaches are mathematically equivalent. They boil down to two approaches that yield two differently corrected residuals $\boldsymbol{\delta}_1$ (approach 1) and $\boldsymbol{\delta}_2$ (approach 2). These two approaches are discussed in detail in the following two subsections.

2.3.1. Approach 1: Scale Down y (Chronological age)

Approach 1 has been proposed by a number of authors (3, 12, 14) and boils down to the following operation: Train a simple regression model $\boldsymbol{\delta} \approx \mathbf{y}\beta_1 + \beta_0$ to remove the linear effect of age from the delta estimate. The new estimate

$$\boldsymbol{\delta}_1 = \boldsymbol{\delta} - \mathbf{y}\beta_1 - \beta_0 \qquad (8)$$

represents corrected brain age delta which is uncorrelated with age. We can inspect the value of β_1 by taking the simple linear regression formula: $\beta_1 = r_{\delta y}\frac{s_\delta}{s_y}$, where $r_{\delta y}$ is age delta correlation (ADC) which is negative (Section 2.4). s_δ is the standard deviation of the residuals and unnormalized square root of the residual sum of squares (RSS), $s_{\mathbf{y}}$ is the standard deviation of the responses and the unnormalized square root of the total sum of squares (TSS). In OLS, Ridge, and Kernel Ridge

regression, we have TSS $\geq$ RSS. Hence, $\beta_1 \in [-1, 0]$. Combining Equations (8) and (7) the entire model can be written in one equation as

$$\boldsymbol{\delta}_1 = f(\mathbf{X}) - \mathbf{y}\,(1 + \beta_1) - \beta_0 \qquad (9)$$

where $(1 + \beta_1) \in [0, 1]$. From Equation (9), we can see that the correction does not affect the predictions $f(\mathbf{X})$. Instead, it implies shrinking $\mathbf{y}$.

This approach is invalid in a predictive modeling framework because it *corrects the data* $\mathbf{y}$, *not the predictions* $f(\mathbf{X})$. Beheshti et al. (12) report a lower error and a larger R^2 value compared to approach 2 introduced in the next section. However, since this effect is obtained by shrinking the data it can be considered as an artifact of this approach.

2.3.2. Approach 2: Scale up $f(\mathbf{X})$ (Predicted age)

As an alternative approach, it has been suggested that $\hat{\mathbf{y}}$ instead of $\boldsymbol{\delta}$ should be used in the regression (1, 4). To this end, train a simple regression model $\hat{\mathbf{y}} \approx \mathbf{y}\beta_1 + \beta_0$. Then define corrected predictions $\hat{\mathbf{y}}_2$ as

$$\hat{\mathbf{y}}_2 = f(\mathbf{X})\,\beta_1^{-1} - \beta_0\,\beta_1^{-1} \qquad (10)$$

with corresponding brain age delta $\boldsymbol{\delta}_2 = \hat{\mathbf{y}}_2 - \mathbf{y}$. This modified age delta estimate is again uncorrelated with age. As in approach 1, the correction is performed using simple linear regression and we have $\beta_1 = r_{\hat{y}y}\frac{s_{\hat{y}}}{s_y}$. $s_{\hat{y}}$ is the standard deviation of the predictions and the unnormalized square root of the explained sum of squares (ESS). In OLS, Ridge, and Kernel Ridge regression, we have TSS $\geq$ ESS and $r_{\hat{y}y} \in [0, 1]$. This implies that $\beta_1 \in [0, 1]$. Combining Equations (10) and (7) this can be written in one equation as

$$\boldsymbol{\delta}_2 = f(\mathbf{X})\,\beta_1^{-1} - \mathbf{y} - \frac{\beta_0}{\beta_1} \qquad (11)$$

with $\beta_1^{-1} \in [1, \infty)$. Comparing Equations (9) and (11), we see that in approach 1 the data vector $\mathbf{y}$ is scaled down whereas in approach 2 the predictions $f(\mathbf{X})$ are scaled up. In Section 2.6, we show that the two types of corrected residuals are actually identical up to scaling and therefore $\mathrm{corr}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = 1$. Consequently, they perform equally well on secondary analyses e.g., relating brain age delta to cognition. They are further closely related to our zero correlation constraint (Section 2.5.1). In a predictive modeling framework, we consider approach 2 as preferable since corrections should be applied to the model not to the data.

2.4. Negative Age Delta Correlation (ADC)

The theoretical basis for negative ADC has already been discussed in (14). In particular, the authors highlighted that ADC ≤ 0 for any sensible regression model. Here, we discuss ADC more specifically for the three regression models introduced above. We start with OLS regression. Let us expand the age delta correlation term as

$$\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}) = \frac{\mathbf{y}^\top (\hat{\mathbf{y}} - \mathbf{y})}{||\mathbf{y}|| \, ||\hat{\mathbf{y}} - \mathbf{y}||} \qquad (12)$$

where to simplify the notation we assume that $\mathbf{X}$ and $\mathbf{y}$ have been centered. The sign of the correlation is determined by the numerator. Defining $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ and writing $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ we obtain

$$-\mathbf{y}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{H} \mathbf{y} = -\mathbf{y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{y} \leq 0 \qquad (13)$$

where the inequality follows from the fact that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent and therefore positive semi-definite (29). This implies that the ADC is always non-positive in OLS regression. Ridge regression coincides with OLS regression for $\lambda = 0$. As λ increases, $\boldsymbol{\beta}$ tends to zero due to the shrinkage effect of the regularization (21). This implies that $\hat{\mathbf{y}} \to 0$ and therefore $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}) \to -1$ as λ increases. This is illustrated empirically in **Figure 1C**. The same argumentation holds for Kernel Ridge, one only has to replace $\mathbf{X}$ by $\Phi(\mathbf{X})$. Often, Kernel Ridge models will have a smaller prediction bias because their higher complexity allows for a better fit to the data. Furthermore, **Figure 1D** shows a clear bias toward large negative ADC values when regression coefficients are randomly sampled. **Figure 1E** shows that the bias persists in both train and test sets. Together, these results suggest that a negative ADC is inevitable and that regularization further exacerbates this effect, in line with previous work (3, 14).

2.5. Correlation Constraints for Regression

The regression problems defined in Section 2.2 can be cast as unconstrained optimization problems. The optimization involves the minimization of a *loss function* $\mathcal{L}$ which measures the amount of discrepancy between the true responses $\mathbf{y}$ and the model predictions $\hat{\mathbf{y}}$:

$$\text{minimize} \quad \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}). \qquad (14)$$

In OLS regression the loss function is the squared loss $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = ||\mathbf{y} - \hat{\mathbf{y}}||_2^2$ whereas it is $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = ||\mathbf{y} - \hat{\mathbf{y}}||_2^2 + \lambda ||\boldsymbol{\beta}_{1:p}||_2^2$ in Ridge and Kernel Ridge regression. To control for age delta correlation in the training data, we can add a correlation constraint that caps the permitted magnitude of correlation between the brain age delta and age. To this end, consider the constrained optimization problem

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) \\ \text{subject to} \quad & |\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta})| \leq \rho. \end{aligned} \qquad (15)$$

The same loss function as before is minimized. However, the set of feasible solutions is limited to solutions for which the absolute value of the correlation does not exceed ρ, where $\rho \geq 0$ is the correlation bound selected by the user. As a special case of Equation (15), we can consider the case $\rho = 0$, that is, the responses have to be perfectly uncorrelated with the residuals:

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) \\ \text{subject to} \quad & \mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}) = 0. \end{aligned} \qquad (16)$$

We will address the latter case first and see that it leads to a simple solution. In the following, we will assume that $\mathbf{X}$ and $\mathbf{y}$ have been centered and the model contains no intercept since this simplifies the equations. This does not limit the generality of our results. As shown in the **Supplementary Material (Section 1)**, a model with centered data and without intercept yields the same regression coefficients as a model with intercept. In other words, we can first calculate $\boldsymbol{\beta}_{1:p}$ on the centered data and subsequently calculate the intercept β_0 to obtain the model for non-centered data.

2.5.1. Zero Correlation Constraint

A hard correlation constraint can be set that requires the correlation between the residuals and the response values to be zero, that is $|\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta})| = 0$. In the **Supplementary Material (Section 2)** the optimal solution, $\boldsymbol{b}$, is derived for OLS, Ridge, and Kernel Ridge regression. It is given as a scaled version of the standard, unconstrained solution

$$\boldsymbol{b}_{1:p} = \theta_0 \, \boldsymbol{\beta}_{1:p} \qquad (17)$$

where $\boldsymbol{\beta}$ is the standard OLS, Ridge, or Kernel Ridge solution and it is assumed that $\mathbf{X}$ and $\mathbf{y}$ have been centered. Using Equation (6), we can see that for Kernel Ridge regression this translates into a scaling of the dual weights

$$\boldsymbol{\alpha}_\rho = \theta_0 \, \boldsymbol{\alpha}. \qquad (18)$$

The scaling factor θ_0 is given by

$$\theta_0 = \frac{||\mathbf{y}||^2}{\mathbf{y}^\top \mathbf{H} \mathbf{y}} \qquad (19)$$

with model-specific hat matrices $\mathbf{H}$:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \quad \text{(OLS)}$$
$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \quad \text{(Ridge)}$$
$$\mathbf{H} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} \quad \text{(Kernel Ridge)}.$$

Intercept term: If the data has not been centered and the model includes an intercept term, $\mathbf{y}$ and $\mathbf{y}^\top \mathbf{H} \mathbf{y}$ need to be centered before calculating θ. The intercept $\boldsymbol{b}_0$ can be obtained from the equation

$$\begin{aligned} (\mathbf{y} - \mathbf{1}\bar{y}) &= (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^\top) \, \boldsymbol{b}_{1:p} \\ \Leftrightarrow \quad \mathbf{y} &= \mathbf{X}\boldsymbol{b}_{1:p} + \mathbf{1}(\bar{y} - \bar{\mathbf{x}}^\top \boldsymbol{b}_{1:p}) \end{aligned}$$

from which it follows that $\boldsymbol{b}_0 = \bar{y} - \bar{\mathbf{x}}^\top \boldsymbol{b}_{1:p}$. The full correlation constrained model with intercept term is then given by the concatenation of the coefficients $\boldsymbol{b} = [\boldsymbol{b}_0, \boldsymbol{b}_{1:p}]$.

2.5.2. Bounded Correlation Constraint

The zero correlation solution successfully removes the correlation between residuals and responses. However, it does so at the cost of goodness of fit to the training data. Furthermore, in predictive modeling, perfect control of ADC on the training set is less important than good predictive performance and low bias on the test set. This suggests the need for a softer constrained optimization solution wherein the equality constraint is replaced by an inequality constraint. In the **Supplementary Material (Section 3)** it is shown that the optimal solution is again given by scaling, $\boldsymbol{b}_{1:p} = \theta_\rho \, \boldsymbol{\beta}_{1:p}$, where there are now two possible solutions for the scaling factor,

$$
\begin{aligned}
\theta_{1,2} = \; & \|\mathbf{y}\|^2 \, \mathbf{y}^\top \hat{\mathbf{y}} \, (1 - \rho^2)/c \\
& \pm \frac{\|\mathbf{y}\|^2}{|c|} \sqrt{\rho^2 \, (1 - \rho^2) \, (\|\mathbf{y}\|^2 \, \|\hat{\mathbf{y}}\|^2 - (\mathbf{y}^\top \hat{\mathbf{y}})^2)}
\end{aligned}
$$

where $c = (\mathbf{y}^\top \hat{\mathbf{y}})^2 - \rho^2 \|\mathbf{y}\|^2 \|\hat{\mathbf{y}}\|^2$ and $\hat{\mathbf{y}}$ is the predictions under the unconstrained OLS, Ridge, or Kernel Ridge model. The two solutions define an interval $[\theta_1, \theta_2]$. Setting θ to any value within this interval will guarantee $-\rho \leq \mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}_{\mathrm{cc}}) \leq \rho$, where $\boldsymbol{\delta}_{\mathrm{cc}}$ is the brain age delta under the correlation constrained models. Setting $\theta = \theta_1$ or $\theta = \theta_2$ will set $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}_{\mathrm{cc}}) = -\rho$ or $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}_{\mathrm{cc}}) = \rho$, respectively. From this we can derive the following algorithm for correlation constrained models with an inequality constraint:

1. Calculate the standard, unconstrained solution for the model (OLS, Ridge, or Kernel Ridge). If $|\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta})| \leq \rho$, the unconstrained solution does not violate the correlation constraints so we are done.
2. If $|\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta})| > \rho$, calculate $\theta_{1,2}$ and set θ to the value that smaller in absolute value. This will assure that $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}_{\mathrm{cc}}) = -\rho$ if $\mathrm{corr}(\mathbf{y}, \boldsymbol{\delta}) < -\rho$.

Figure 2 depicts the geometrical intuition underlying the correlation constraints. Without constraints, the solution is the minimum of a quadratic function (**Figure 2A**). For a zero correlation constraint, the set of feasible solutions is reduced to a line within this space (**Figure 2B**). For a bounded correlation constraint, the set of feasible solutions is the space between two paraboloids (**Figure 2C**). In both cases, the correlation constraints lead to a larger slope for the regression coefficients as compared to the unconstrained model (**Figure 2D**).

2.6. Relationship Between Zero Correlation Constraint and Existing Correction Approaches

In Section 2.3, we reviewed the two main approaches for correcting brain age delta used in the literature. Here, we investigate their mutual relationship as well as their relationship to our approach. Without loss of generality we assume that $\mathbf{y}$ and $\mathbf{X}$ have been centered (see **Supplementary Material, Section 1**). Let us start from Equation (8) corresponding to approach 1. The regression slope β_1 for a simple linear regression model is given by

$$
\beta_1 = \mathrm{corr}(\boldsymbol{\delta}, \mathbf{y}) \, \frac{\|\boldsymbol{\delta}\|}{\|\mathbf{y}\|}.
$$

Writing $\boldsymbol{\delta} = \hat{\mathbf{y}} - \mathbf{y}$ and expanding the correlation term as in Equation (S5) (**Supplementary Material**), we find that

$$
\beta_1 = \frac{\mathbf{y}^\top \mathbf{H} \mathbf{y}}{\|\mathbf{y}\|^2} - 1 = \theta_0^{-1} - 1
$$

with θ_0 as defined in Equation (19). Therefore, the solution to Equation (8) is given by

$$
\boldsymbol{\delta}_1 = \hat{\mathbf{y}} - \theta_0^{-1} \mathbf{y}. \tag{20}
$$

Alternatively, using approach 2 (correction of predictions) and starting from Equation (10) we perform a regression of $\hat{\mathbf{y}}$ on $\mathbf{y}$. Again, this is a simple linear regression model whose slope is given by

$$
\beta_1 = \mathrm{corr}(\hat{\mathbf{y}}, \mathbf{y}) \, \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{y}\|} = \theta_0^{-1}
$$

yielding the corrected predictions $\hat{\mathbf{y}}_2 = \hat{\mathbf{y}}/\beta_1 = \theta_0 \, \hat{\mathbf{y}}$ with corresponding brain age delta

$$
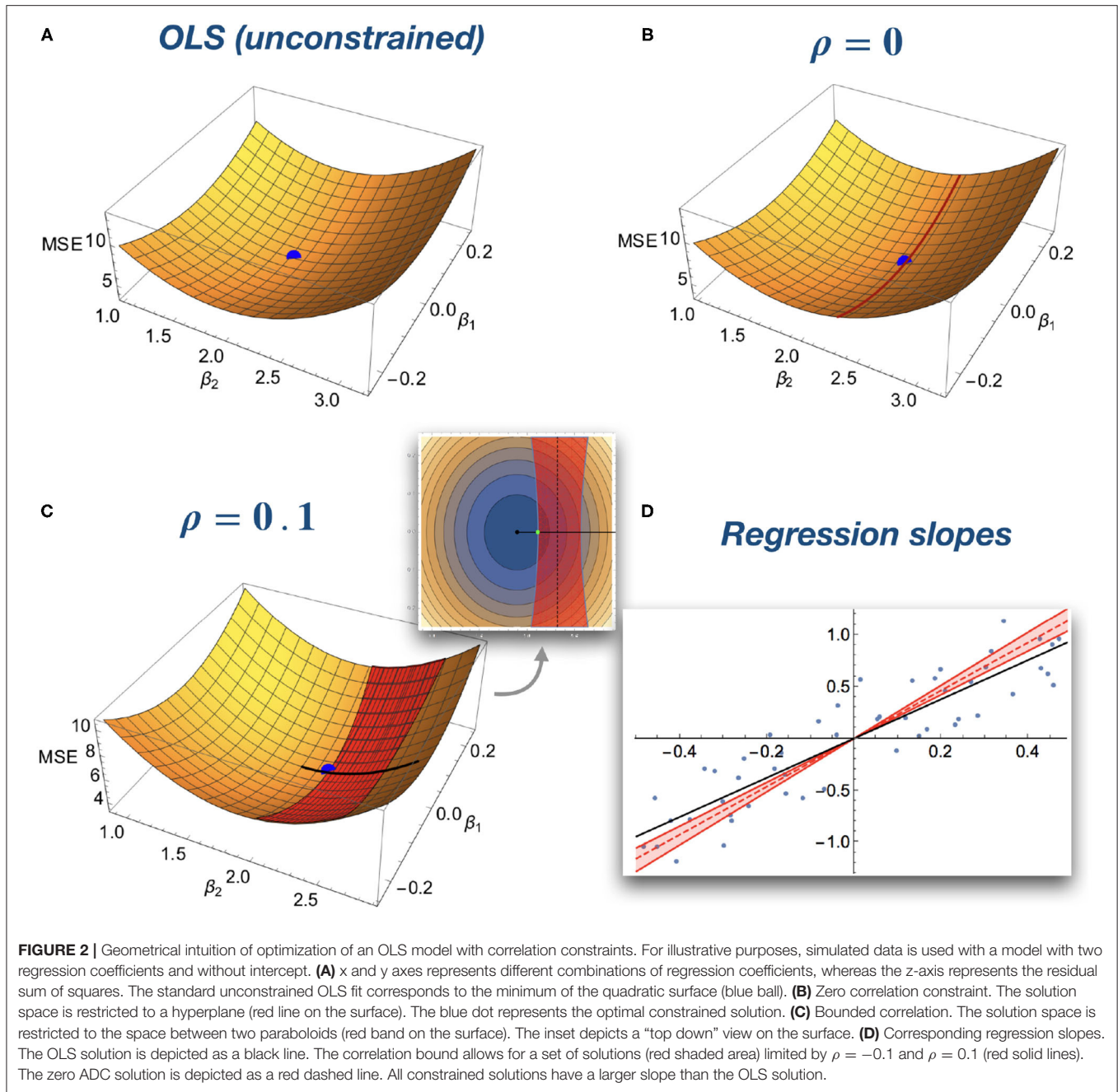\boldsymbol{\delta}_2 = \theta_0 \, \hat{\mathbf{y}} - \mathbf{y}. \tag{21}
$$

This solution uses a scaling of $\hat{\mathbf{y}}$ by θ_0 and is thus equivalent to our zero correlation solution when an OLS model is used. Furthermore, comparing Equations (20) and (21) we see that both solutions are proportional to each other. Their relationship is given by

$$
\boldsymbol{\delta}_2 = \theta_0 \, \boldsymbol{\delta}_1 \tag{22}
$$

and therefore $\mathrm{corr}(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = 1$. In other words, the brain age delta estimates from the two approaches used in the literature are identical up to a scaling factor of θ_0.

2.7. Interpretability

Maximizing predictive performance is the primary objective when optimizing statistical models. To empirical researchers, understanding *what* the regression model learns from the data is useful, too. Linear regression models such as OLS and Ridge can be interpreted in terms of their $\boldsymbol{\beta}$ coefficients. To keep the notation simple, we will assume that the data has been demeaned and the model contains no intercept term. If the data is standardized, large components of the $\boldsymbol{\beta}$ can be interpreted as features that are relevant to the regression task. Since our models combine prediction and correction into a single task, the $\boldsymbol{\beta}$'s capture the entire operation of the model. Importantly, the choice of the correlation bound ρ does not change the interpretation. Since the regression coefficients in our models are just scaled versions of the

FIGURE 2 | Geometrical intuition of optimization of an OLS model with correlation constraints. For illustrative purposes, simulated data is used with a model with two regression coefficients and without intercept. **(A)** x and y axes represents different combinations of regression coefficients, whereas the z-axis represents the residual sum of squares. The standard unconstrained OLS fit corresponds to the minimum of the quadratic surface (blue ball). **(B)** Zero correlation constraint. The solution space is restricted to a hyperplane (red line on the surface). The blue dot represents the optimal constrained solution. **(C)** Bounded correlation. The solution space is restricted to the space between two paraboloids (red band on the surface). The inset depicts a "top down" view on the surface. **(D)** Corresponding regression slopes. The OLS solution is depicted as a black line. The correlation bound allows for a set of solutions (red shaded area) limited by $\rho = -0.1$ and $\rho = 0.1$ (red solid lines). The zero ADC solution is depicted as a red dashed line. All constrained solutions have a larger slope than the OLS solution.

original regression coefficients, $\boldsymbol{b} = \theta\,\boldsymbol{\beta}$ for some $\theta \in \mathbb{R}$, the choice of ρ does not affect the ratio between any pair of coefficients.

For collinear data, coefficients can become uninterpretable with large weights for features that are not related to the target variable. In this case, *structure coefficients* (18) and *activation patterns* (19) have been proposed as alternatives metrics. An activation pattern is given by

$$\boldsymbol{a}_\rho = \boldsymbol{\Sigma}\,\boldsymbol{b}$$

where $\boldsymbol{a}_\rho \in \mathbb{R}^p$ is the activation pattern and $\boldsymbol{\Sigma}$ is the data covariance matrix. Let $\boldsymbol{a}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}\,\boldsymbol{\beta}$ be the activation pattern for

the standard (uncorrected) model. Then setting ρ merely scales the activation pattern by θ_ρ since $\boldsymbol{a}_\rho = \boldsymbol{\Sigma}\,\boldsymbol{b} = \boldsymbol{\Sigma}\,\theta_\rho\boldsymbol{\beta} = \theta_\rho\,\boldsymbol{a}_{\boldsymbol{\beta}}$. An example for an activation pattern is depicted in **Figure 5** and discussed in Section 3.3.

Structure coefficients are given by the vector of Pearson correlations between $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{b}$ and each column of $\mathbf{X}$, where the i-th column is denoted as $\mathbf{X}_i$. Since the correlation coefficient is invariant to constant shifts and scaling, we have corr($\mathbf{X}_i, \mathbf{X}\boldsymbol{b}$) = corr($\mathbf{X}_i, \mathbf{X}\boldsymbol{\beta}$), that is, structure coefficients are invariant to the choice of ρ.

For kernel methods, these coefficients are generally not available. As an alternative, we considered the vector of *partial*

derivatives of regression model with the respect to each of the features,

$$\partial_i f(\mathbf{x}) = (\partial_i k(\mathbf{x}))^\top \boldsymbol{\alpha},$$

where k is a kernel function (20) and i refers to the index of the feature. Using Equation (18), we see that for a correlation constrained Kernel Ridge model f_ρ these partial derivatives are again just scaled versions of the unconstrained version $\partial_i f_\rho(\mathbf{x}) = \theta_\rho \, \partial_i f(\mathbf{x})$.

To summarize, our models are interpretable because the regression coefficients capture the whole operation of the model. Furthermore, the choice of the correlation bound does not affect the interpretation of the model: structure coefficients are invariant to the choice of ρ, whereas regression coefficients, activation patterns and partial derivatives are merely scaled by a constant factor.

2.8. Toolbox

The Linear, Ridge, and Kernel Ridge regression models with correlation constraints presented in this paper have been implemented in Python and MATLAB. For Python, the models are available on GitHub (github.com/treder/correlation-constrained-regression). They inherit from and are fully compatible with the Scikit-Learn framework (30). The models extend Scikit-Learn's LinearRegression, Ridge, and KernelRidge models with an additional parameter `correlation_bound` that corresponds to ρ in Equations (15) and (16). Setting the parameter to 0 enforces a zero correlation constraint whereas setting it to a positive value bounds the correlation accordingly. For MATLAB, the models have been integrated into MVPA-Light (31), an open-source machine learning toolbox. By setting the hyperparameter `correlation_bound`, ADC can be controlled in the same way as for the Python-based models. Code examples for both Python and MATLAB can be found on the GitHub page.

2.9. Neuroimaging Data

Neuroimaging data supplied within the Predictive Analytics Competition were fully pre-processed T1-weighted MRI scans from 2,640 training set and 660 validation set subjects as described previously (2). All normalized 3D maps of gray matter (GM) and white matter (WM) volume were used to create group GM and WM masks. Each GM and WM image was smoothed using an 8-mm Gaussian kernel, masked and concatenated into a vector of 153,237 and 86,143 voxels, respectively.

Concatenated GM and WM images were intensity normalized and submitted to Independent Component Analysis using the Group ICA of fMRI toolbox [https://trendscenter.org/software/gift/; (32)]. The optimal number of components of the ICA decomposition (72 and 99 for GM and WM images, respectively) was determined using Principal Component Analysis (PCA) with minimum description length (MDL) model order selection criteria (33). Normalized features in the training model, based on 2,640 participants, included scores for all GM and WM components ($N = 171$). Additional covariates included total GM, total WM, gender and dummy coding for 17 scanning sites.

We also included low-order interaction terms (such as bivariate interaction) between total GM, total WM, gender, PC1 and PC2 scores.

2.10. Brain Age Prediction

We performed brain age prediction using Python with Scikit-Learn and our custom extensions. The models were tested in three different conditions: *Unconstrained* (using Scikit-Learn's models without correlation constraints), *zero correlation* ($\rho = 0$) (using our extensions with a correlation constraint of 0), *bounded correlation* ($\rho = 0.1, 0.2, 0.3$) (using our extensions with a correlation bound of 0.1, 0.2, 0.3), and approaches 1 and 2 from the literature introduced in Section 2.3. To obtain both in-sample and out-of-sample statistics, models were applied to both training and test data. To estimate the variability of predictive performance, we performed 100 iterations. In every iteration, training data was randomly sampled from the training set using bootstrapping. The 171 Independent Components were used as features.

Three regression models were considered, OLS, Ridge regression, and Kernel Ridge regression with a RBF kernel. For both Ridge and Kernel Ridge regression, hyperparameters were tuned using a grid search with Scikit-Learn's GridSearchCV and five-fold cross-validation. For Ridge regression, the regularization parameter was tuned using candidate values $\alpha = (10^{-3}, 10^{-2}, 10^{-1}, 1, 10)$. For Kernel Ridge with a RBF kernel, kernel width $\gamma = (100, 10, 1, 10^{-1})$ and $\alpha = (10^{-3}, 10^{-2}, 10^{-1}, 1, 10)$ were tuned. The resultant best model was used to calculate in- and out-of-sample metrics. Mean absolute error (MAE) and age delta correlation (ADC) served as metrics. Denoting train and test sets as $\mathcal{TR}$ and $\mathcal{TE}$, MAE was estimated as

$$\begin{aligned} \mathrm{MAE}_{\mathrm{train}} &= \frac{1}{|\mathcal{TR}|} \sum_{i \in \mathcal{TR}} |\mathbf{y}_i - f(\mathbf{x}_i)| \\ \mathrm{MAE}_{\mathrm{test}} &= \frac{1}{|\mathcal{TE}|} \sum_{i \in \mathcal{TE}} |\mathbf{y}_i - f(\mathbf{x}_i)| \end{aligned} \quad (23)$$

where f is a regression model trained on the training data and $|\mathcal{TR}|$ and $|\mathcal{TE}|$ are the sizes of the train and test sets, respectively. Similarly, ADC was calculated separately for the predictions in the train and test sets. All analyses were performed on a Desktop computer with an Intel Core i7-6700 @ 3.40 GHz x 8 CPU with 64 GB RAM running on Ubuntu 18.04. The analysis code is available as part of the toolbox[2].

3. RESULTS

3.1. Brain Age Prediction

Figure 3 depicts the MAE and ADC results on the PAC data for train and test sets separately, comparing the three regression models (OLS, Ridge, and Kernel Ridge regression) and different constraints on the age delta correlation (ADC): unconstrained

[2]https://github.com/treder/correlation-constrained-regression/blob/main/run_regression_analysis.py
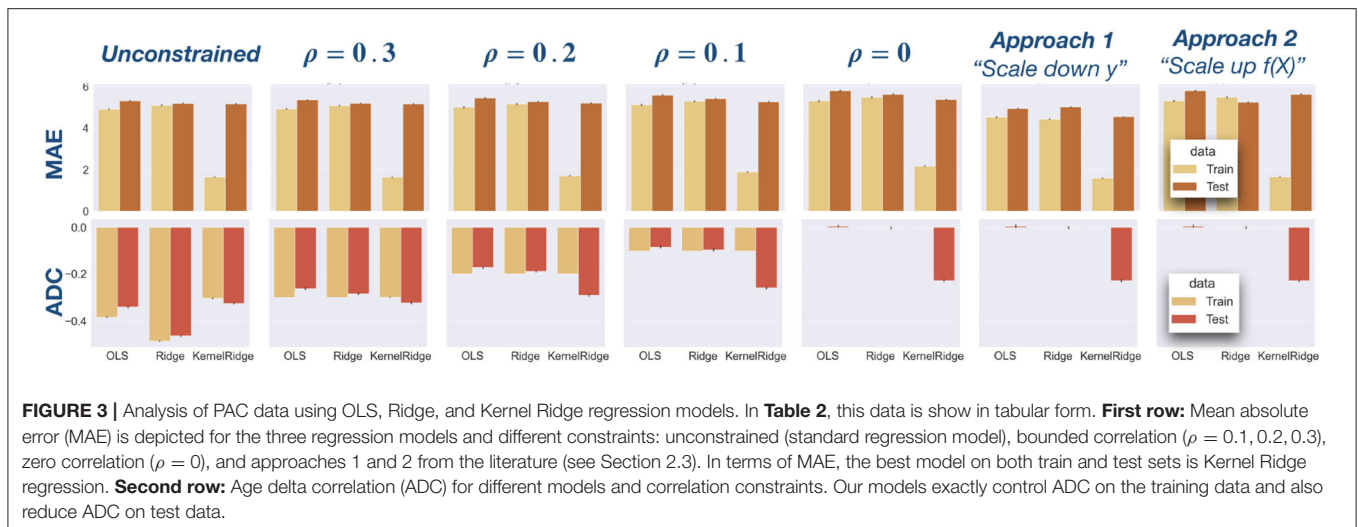
FIGURE 3 | Analysis of PAC data using OLS, Ridge, and Kernel Ridge regression models. In **Table 2**, this data is show in tabular form. **First row:** Mean absolute error (MAE) is depicted for the three regression models and different constraints: unconstrained (standard regression model), bounded correlation ($\rho = 0.1, 0.2, 0.3$), zero correlation ($\rho = 0$), and approaches 1 and 2 from the literature (see Section 2.3). In terms of MAE, the best model on both train and test sets is Kernel Ridge regression. **Second row:** Age delta correlation (ADC) for different models and correlation constraints. Our models exactly control ADC on the training data and also reduce ADC on test data.

(standard regression model), bounded correlation constraint ($\rho = 0.1, 0.2, 0.3$), zero correlation constraint ($\rho = 0$), and approaches 1 and 2 from the literature introduced in Section 2.3. The same data is presented in tabular form in **Table 2**. Bonferroni correction was used in case of multiple comparisons.

With respect to mean absolute error (MAE), a significantly lower error was found on the train data compared to test data (Wilcoxon signed-rank test, $W = 62,449, p < 0.0001$). For the constrained models, we performed a regression analysis of MAE on ρ. We found significant negative slopes for all models and train and test sets separately (all $p < 0.0001$), indicating that MAE decreases significantly as ρ increases. On the test data, we used Wilcoxon signed-rank tests to compare MAE for the unconstrained model with MAE for each of the constrained models. A significantly higher MAE was obtained for most unconstrained models (all $p < 0.0001$). The only exception was approach 1 ("scale down **y**") wherein the relationship is reversed: MAE decreases after correction. This is an artifact of the fact that the correction is applied to the data, not the predictions (see Section 2.3).

With respect to age delta correlation (ADC), our models perfectly controlled for ADC on the training data. ADC was significantly larger in magnitude for the test data compared to the train data ($W = 877,894, p < 0.0001$). Linear regression of ADC on ρ showed a significant negative slope for all models on both train and test sets (all $p < 0.001$). On the test set, ADC was significantly larger in magnitude for the unconstrained model than for the constrained models (all $p < 0.0001$), suggesting that the constrained models also control ADC on the test set.

3.2. ADC-MAE Trade-Off

To better characterize how the choice of ρ mediates the trade-off between ADC and MAE, we repeated the prediction analysis. This time the correlation bound was varied in small steps of 0.02. Results averaged across 100 bootstrap iterations are depicted in **Figure 4**. In line with the results above, MAE generally decreases with increasing ρ and eventually flattens off. ADC decreases

roughly linearly with ρ. It flattens off at a value of ρ that corresponds to the ADC value of the uncorrected model, since no correction needs to be applied when $\rho > |\text{ADC}|$. Furthermore, MAE and ADC change similarly for train and test set, albeit with different slopes. This is a useful observation since it suggests that the hyperparameter ρ can be optimized on the training set alone, in line with good practice in predictive modeling.

3.3. Interpretability via Activation Patterns

Figure 5 shows an activation pattern for an OLS model with $\rho = 0$ trained on the whole training set, using gray matter Independent Components (ICs) only. To create the maps, a vector of activation patterns **a** was calculated for the gray matter ICs (see Section 3.3). Since each IC corresponds to a brain map, we multiplied each entry of **a** with its corresponding map and added up the maps. The figure depicts the resultant summed map indicating an age-related decrease in intensity values in deep cortical areas. An age-related increase in intensity values was observed at the boundaries between gray matter and other tissue types, likely reflecting CSF signals in older adults (34). Crucially, the map is independent of the choice of ρ, only its scaling is affected by ρ. This illustrates that the interpretation of the models is not affected by a change of ρ.

4. DISCUSSION

A predictive bias manifesting as an overprediction of the age of young individuals and an underprediction of the age of elderly individuals has been consistently reported in the brain age literature (2, 3, 14, 15). It can be quantified as age delta correlation (ADC), that is, the correlation between brain age delta (predicted age minus chronological age) and chronological age. We introduced modifications to three popular regression models, OLS, Ridge and Kernel Ridge regression, that effectively control ADC. To this end, we introduced a hyperparameter ρ that can be set by the user. It represents a correlation bound that controls the maximum permissible ADC. The resultant models are optimal

TABLE 2 | Mean absolute error (MAE) and age delta correlation (ADC) for different types of correlation constraints and regression models.

Data	Constraint	OLS	Ridge	Kernel ridge
MAE (Train)	Unconstrained	4.91 ± 0.09	5.1 ± 0.08	1.65 ± 0.04
	$\rho = 0.3$	4.93 ± 0.09	5.09 ± 0.08	1.65 ± 0.03
	$\rho = 0.2$	5.0 ± 0.09	5.16 ± 0.1	1.7 ± 0.04
	$\rho = 0.1$	5.12 ± 0.1	5.29 ± 0.1	1.89 ± 0.05
	$\rho = 0$	5.3 ± 0.11	5.48 ± 0.11	2.17 ± 0.06
	Approach 1	4.52 ± 0.08	4.43 ± 0.07	1.59 ± 0.03
	Approach 2	5.3 ± 0.1	5.48 ± 0.11	1.67 ± 0.04
MAE (Test)	Unconstrained	5.31 ± 0.09	5.18 ± 0.06	5.16 ± 0.09
	$\rho = 0.3$	5.35 ± 0.09	5.19 ± 0.07	5.16 ± 0.09
	$\rho = 0.2$	5.44 ± 0.1	5.27 ± 0.07	5.19 ± 0.09
	$\rho = 0.1$	5.59 ± 0.1	5.42 ± 0.08	5.26 ± 0.1
	$\rho = 0$	5.79 ± 0.11	5.62 ± 0.09	5.37 ± 0.11
	Approach 1	4.95 ± 0.07	4.99 ± 0.09	4.54 ± 0.06
	Approach 2	5.82 ± 0.09	5.22 ± 0.1	5.64 ± 0.08
ADC (Train)	Unconstrained	−0.384 ± 0.007	−0.486 ± 0.007	−0.305 ± 0.006
	$\rho = 0.3$	−0.3 ± 0.0	−0.3 ± 0.0	−0.299 ± 0.002
	$\rho = 0.2$	−0.2 ± 0.0	−0.2 ± 0.0	−0.2 ± 0.0
	$\rho = 0.1$	−0.1 ± 0.0	−0.1 ± 0.0	−0.1 ± 0.0
	$\rho = 0$	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	Approach 1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	Approach 2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
ADC (Test)	Unconstrained	−0.341 ± 0.021	−0.464 ± 0.018	−0.325 ± 0.022
	$\rho = 0.3$	−0.263 ± 0.021	−0.284 ± 0.021	−0.324 ± 0.023
	$\rho = 0.2$	−0.173 ± 0.022	−0.19 ± 0.022	−0.291 ± 0.023
	$\rho = 0.1$	−0.083 ± 0.023	−0.095 ± 0.022	−0.26 ± 0.024
	$\rho = 0$	0.005 ± 0.023	−0.001 ± 0.022	−0.228 ± 0.024
	Approach 1	0.005 ± 0.023	−0.001 ± 0.022	−0.228 ± 0.024
	Approach 2	0.005 ± 0.023	−0.001 ± 0.022	−0.228 ± 0.024

*In **Figure 3**, this data is depicted as bar graphs.*

in that they give the lowest mean-squared error on the training set (among all models from the same class) while controlling for ADC.

Our models were tested on the PAC competition data using different values for ρ. The models not only perfectly controlled ADC on the training data, they also approximately controlled ADC on unseen test data (see **Figure 3**). For all constrained models, ADC on the test set was lower than for the unconstrained models. The flip side of this was an increase of mean absolute error (MAE) for our constrained models as compared to the unconstrained model, but often this increase was modest. For instance, for an OLS model MAE increased from 5.31 for the unconstrained model to 5.35 for the model with $\rho = 0.3$, an increase of only 0.8%. Across all models in the test set, we found that an increase in ρ led to a decrease in MAE. This suggests that ρ can be used as a lever to finely control the trade-off between predictive performance (MAE) and age delta correlation (ADC).

In the same analysis, we included the two existing correction methods used in the literature, denoted as approach 1 (3, 12, 14) and approach 2 (1, 4) discussed in detail in Section 2.3. For

the special case of using a OLS model with a zero correlation constraint, $\rho = 0$, our models' brain age deltas are equivalent to approach 2. In Section 2.6, we furthermore show that the brain age deltas in approaches 1 and 2 are actually identical up to scaling. They differ only by the scaling factor θ_0 defined in Equation (19). In particular, in approach 1 chronological age is scaled down by this factor before calculating brain age delta, whereas in approach 2 $f(\mathbf{X})$ (predicted age) is scaled up by the same amount. The downscaling in approach 1 leads to a lower MAE which is even smaller than for an uncorrected model. We would like to stress that this is because approach 1 corrects the data, not the predictions. This is not permissible in predictive modeling and approach 2 should be preferred.

The hyperparameter ρ controlling ADC has to be selected by the user. **Figure 4** shows how the choice of ρ affects both MAE and ADC. A possible selection criterion would involve defining a maximum permissible MAE or ADC value and choosing ρ accordingly. A more comprehensive analysis would take into account any follow-up analyses. For instance, brain age delta is often correlated with cognitive variables in a second step. An
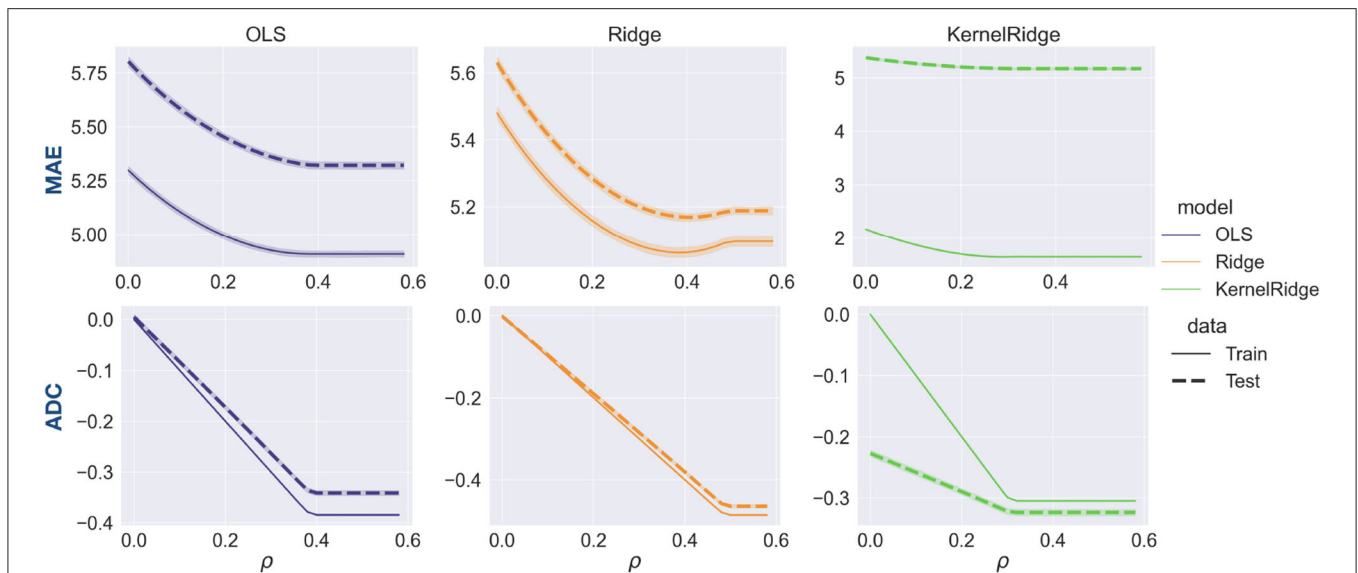
FIGURE 4 | Trade-off between MAE (row 1) and ADC (row 2) as a function of the hyperparameter ρ that represents the correlation bound in our models. The shaded area around the lines represents standard deviation. Increasing ρ leads to a lower MAE but this comes at the expense of ADC increasing in magnitude. MAE and ADC change in a similar way on training and test sets, suggesting that the training set is a good proxy for test set performance.
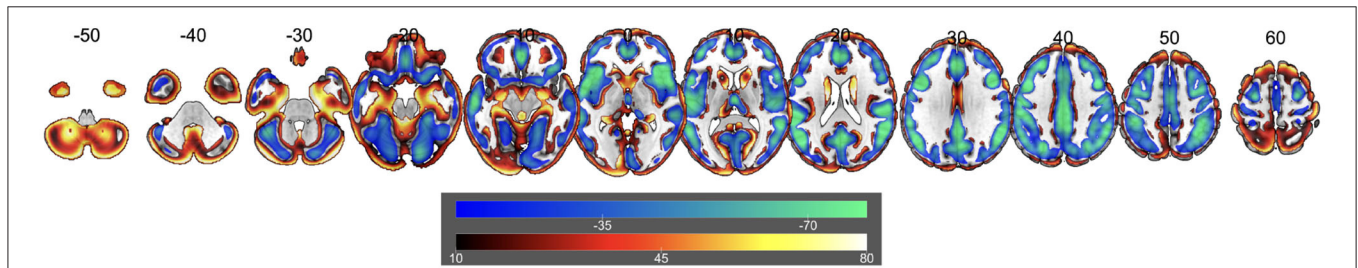

FIGURE 5 | Activation pattern for an OLS model trained on the gray matter Independent Components (ICs). Warms colors correspond to positive values and cold colors to negative values. The brain map has been produced by multiplying each entry of the activation pattern with the map corresponding to each IC, and then summing up all the maps. The choice of ρ does not affect the map in relative terms, it only affects its scaling.

optimal selection of ρ could involve e.g., the regression slope or p-value of the association on the training set. A detailed exploration of how the choice of ρ affects follow-up analyses requires a dataset including cognitive test scores and is left for future work.

In terms of interpretability, our models offer a greater degree of transparency than the traditional two approaches because the model coefficients capture the entire brain age prediction pipeline (i.e., both prediction and correction). Moreover, in Section 2.7 we show that the interpretation is not affected by changes in ρ, a hyperparameter in our models. Structure coefficients (18) are invariant to changes in ρ, whereas the other metrics are simply scaled by a constant value.

A limitation of our study is that it only covers OLS, Ridge, and Kernel Ridge regression. In the age of deep neural networks, the focus on linear and kernel methods may seem very limiting. Therefore, we would like to emphasize that linear regression models lie at the heart of many non-linear approaches including Convolutional Neural Networks. As illustrated in **Figure 6**, non-linear approaches can often be conceived of as linear regression

operating on non-linearly extracted features. Linear models can uniquely combine predictive power with computational efficiency and interpretability. In line with this, (35) found that kernel regression was as performant as deep neural networks when predicting phenotypes from functional connectivity data. Nevertheless, future work could address the incorporation of correlation constraints into other models classes such as Lasso (36), Support Vector Regression (37) or CNNs. Since these models use iterative optimization, a possible approach could be adding the correlation term $-\text{corr}^2(\mathbf{y}, \boldsymbol{\delta})$ directly to the loss function. Alternatively, since $\mathbf{y}$ is constant this can be simplified to the quantity $\mathbf{y}^\top \hat{\mathbf{y}} / ||\boldsymbol{\delta}||^2$.

On a more speculative note, future development of the brain age delta metric might benefit from work on errors-in-variables models (38–40) or measurement error models (16). Standard linear regression models assume that chronological age has been measured with an error whereas the brain data is noisefree. It is more likely that the opposite is true: chronological age can be measured with high accuracy but

FIGURE 6 | Coupled with non-linear feature extraction techniques, linear regression becomes a powerful model that is implicitly a part of many regression models. **(A)** Linear regression using three features to predict age. **(B)** Kernel regression can be conceived of as a projection of the data into a higher-dimensional, Reproducing Kernel Hilbert Space (RKHS), followed by linear regression in this space. **(C)** MRIs can be transformed into components using ICA. For each MRI, its loadings on the different Independent Components can be used as features for linear regression. **(D)** In Convolutional Neural Networks (CNN) for regression, the last layer is typically linear. The preceding layers can be conceived of as feature extraction layers, and the last layer performs linear regression on non-linear features.

there is noise and individual variability in the brain scans. Not accounting for measurement error in the features leads to regression dilution which in OLS regression manifests as an underestimation of the regression coefficients. This phenomenon is known in the brain age literature (3, 14). Our scaling factor θ inflates the regression coefficients and therefore un-dilutes the model, but it is not clear to the authors whether there is a more formal relationship between correction of the residuals and measurement error. Unfortunately, estimating measurement error in brain scans requires repeated sampling which is often not available.

Concluding, without accurate control for ADC, the use of brain age delta can lead to false associations with other phenotypes and limit our understanding of the processes that underpin brain aging. We highlighted the importance of estimating brain age delta and controlling for age delta correlation within a given model, as we introduced a novel class of regression models that allow for fine control of ADC. Our solution is optimal on the training set and shows approximate control of ADC on the test set. In an era of "big data" predictive modeling, this approach nicely dovetails with strategies to develop reliable models that generalize to independent test sets for use in personalized and precision medicine (41).

DATA AVAILABILITY STATEMENT

The neuroimaging dataset analyzed in this study has been provided by the PAC 2019 team. All the program code and

functions can be found in the following repository: https://github.com/treder/correlation-constrained-regression.

REFERENCES

1. Cole JH, Ritchie SJ, Bastin ME, Valdés Hernández MC, Muñoz Maniega S, Royle N, et al. Brain age predicts mortality. *Mol Psychiatry*. (2018) 23:1385–92. doi: 10.1038/mp.2017.62
2. Cole JH, Poudel RPK, Tsagkrasoulis D, Caan MWA, Steves C, Spector TD, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*. (2017) 163:115–24. doi: 10.1016/j.neuroimage.2017.07.059

3. Smith SM, Vidaurre D, Alfaro-Almagro F, Nichols TE, Miller KL. Estimation of brain age delta from brain imaging. *Neuroimage*. (2019) 200:528–39. doi: 10.1016/j.neuroimage.2019.06.017
4. de Lange AMG, Cole JH. Commentary: correction procedures in brain-age prediction. *Neuroimage: Clin.* (2020) 26:102229. doi: 10.1016/j.nicl.2020.102229
5. Borgeest GS, Henson RN, Shafto M, Samu D, Kievit RA. Greater lifestyle engagement is associated with better age-adjusted cognitive abilities. *PLoS ONE*. (2020) 15:e0230077. doi: 10.1371/journal.pone.0230077

6. Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PLoS ONE*. (2013) 8:e67346. doi: 10.1371/journal.pone.0067346

7. Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, Bottlender R, Frodl T, et al. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophrenia Bull*. (2014) 40:1140–53. doi: 10.1093/schbul/sbt142

8. Jiang H, Lu N, Chen K, Yao L, Li K, Zhang J, et al. Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. *Front Neurol*. (2020) 10:1346. doi: 10.3389/fneur.2019.01346

9. Aycheh HM, Seong JK, Shin JH, Na DL, Kang B, Seo SW, et al. Biological brain age prediction using cortical thickness data: a large scale cohort study. *Front Aging Neurosci*. (2018) 10:252. doi: 10.3389/fnagi.2018.00252

10. Zhai J, Li K. Predicting brain age based on spatial and temporal features of human brain functional networks. *Front Hum Neurosci*. (2019) 13:62. doi: 10.3389/fnhum.2019.00062

11. Monti RP, Gibberd A, Roy S, Nunes M, Lorenz R, Leech R, et al. Interpretable brain age prediction using linear latent variable models of functional connectivity. *PLoS ONE*. (2020) 15:e0232296. doi: 10.1371/journal.pone.0232296

12. Beheshti I, Nugent S, Potvin O, Duchesne S. Bias-adjustment in neuroimaging-based brain age frameworks: a robust scheme. *Neuroimage: Clin*. (2019) 24:102063. doi: 10.1016/j.nicl.2019.102063

13. Cole JH, Leech R, Sharp DJ. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Ann Neurol*. (2015) 77:571–81. doi: 10.1002/ana.24367

14. Le TT, Kuplicki RT, McKinney BA, Yeh HW, Thompson WK, Paulus MP. A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Front Aging Neurosci*. (2018) 10:317. doi: 10.1101/377648

15. Liang H, Zhang F, Niu X. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Hum Brain Mapp*. (2019) 40:3143–52. doi: 10.1002/hbm.24588

16. Fuller WA. *Measurement Error Models*. Hoboken, NJ: John Wiley & Sons, Inc. (1987). doi: 10.1002/9780470316665

17. MacMahon S, Peto R, Collins R, Godwin J, MacMahon S, Cutler J, et al. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*. (1990) 335:765–74. doi: 10.1016/0140-6736(90)90878-9

18. Thompson B, Borrello GM. The importance of structure coefficients in regression research. *Educ Psychol Meas*. (1985) 45:203–9. doi: 10.1177/001316448504500202

19. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*. (2014) 87:96–110. doi: 10.1016/j.neuroimage.2013.10.067

20. Emmanuel Johnson J, Laparra V, Pérez-Suay A, Mahecha MD, Camps-Valls G. Kernel methods and their derivatives: concept and perspectives for the earth system sciences. *PLoS ONE*. (2020) 15:e0235885. doi: 10.1371/journal.pone.0235885

21. Marquardt DW, Snee RD. Ridge regression in practice. *Am Stat*. (1975) 29:3–20. doi: 10.1080/00031305.1975.10479105

22. van Wieringen WN. Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*. (2015).

23. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x

25. Hainmueller J, Hazlett C, Alvarez RM. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Anal*. (2014) 22:143–68. doi: 10.1093/pan/mpt019

26. Schölkopf B, Smola AJ. A short introduction to learning with kernels. In: Mendelson S, Smola AJ, editors. *Advanced Lectures on Machine Learning*. Berlin; Heidelberg: Springer (2003). p. 41–64. doi: 10.1007/3-540-36434-X_2

27. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *Ann Stat*. (2008) 36:1171–220. doi: 10.1214/009053607000000677

28. Bishop CM. *Pattern Recognition and Machine Learning*. Springer: Berlin, Heidelberg (2006).

29. Draper NR, Smith H. *Applied Regression Analysis*. New York, NY: Wiley (1966).

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.5555/1953048.2078195

31. Treder MS. MVPA-light: a classification and regression toolbox for multi-dimensional data. *Front Neurosci*. (2020) 14:289. doi: 10.3389/fnins.2020.00289

32. Calhoun VD, Adali T, Pearlson GD, Pekar JJ. A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp*. (2001) 14:140–51. doi: 10.1002/hbm.1048

33. Hui M, Li J, Wen X, Yao L, Long Z. An empirical comparison of information-theoretic criteria in estimating the number of independent components of fMRI data. *PLoS ONE*. (2011) 6:e29274. doi: 10.1371/journal.pone.0029274

34. Tsvetanov KA, Henson RNA, Jones PS, Mutsaerts H, Fuhrmann D, Tyler LK, et al. The effects of age on resting-state BOLD signal variability is explained by cardiovascular and cerebrovascular factors. *Psychophysiology*. (2020). doi: 10.1111/psyp.13714. [Epub ahead of print].

35. He T, Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*. (2020) 206:116276. doi: 10.1016/j.neuroimage.2019.116276

36. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Ser B*. (2011) 73:273–82. doi: 10.1111/j.1467-9868.2011.00771.x

37. Müller KR, Smola AJ, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V. Predicting time series with support vector machines. In: *Proceedings of the 7th International Conference on Artificial Neural Networks*. Berlin; Heidelberg. (1997). p. 999–1004. doi: 10.1007/BFb0020283

38. Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable on JSTOR. *J R Stat Soc Ser A*. (2000) 163:173–89. doi: 10.1111/1467-985X.00164

39. Berglund L. Regression dilution bias: tools for correction methods and sample size calculation. *Upsala J Med Sci*. (2012) 117:279–83. doi: 10.3109/03009734.2012.668143

40. Gleser LJ. Estimation in a multivariate "errors in variables" regression model: large sample results. *Ann Stat*. (1981) 9:24–44. doi: 10.1214/aos/1176345330

41. Bzdok D, Varoquaux G, Steyerberg EW. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry*. (2020). doi: 10.1001/jamapsychiatry.2020.2549. [Epub ahead of print].