Title and abstract

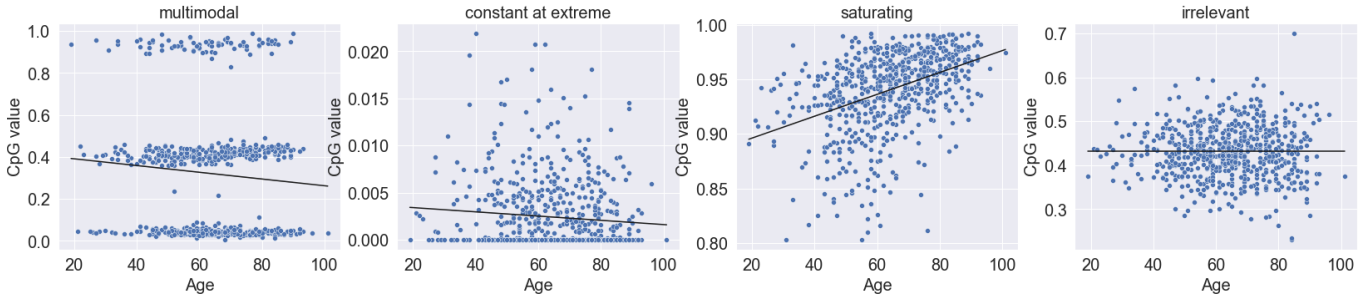
Introduction

Background

Methodology

Selecting sites linear with respect to time

In the first step, we selected only the CpG sites whose changes in methylation values can be explained by age. There is a linear relationship in between them and when age changes, so does the methylation value. This is in contrast with the sites that don't offer differentiation in between the methylation value at an early age vs. at the older age. See the figure below for examples from Generation Scotland dataset.

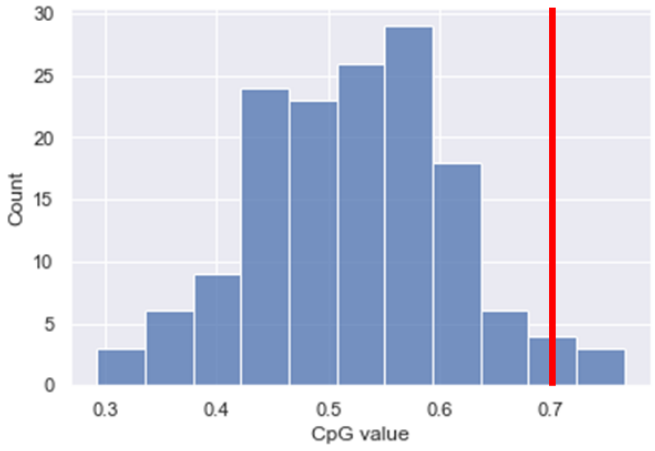


We select linear sites as a data preparation step to only use the sites that are useful in age prediction. We do this in order to prevent the problem of previous approaches, where training a model with all the sites causes a selection of sites that are disease predicting rather age predicting. See the argument in the Related Work section.

To complete this selection we train a linear regression model on all CpG sites and assess the goodness of the fit using a coefficient of determination statistic (R^2). We also explored sites that are high in variance We proceeded to set a threshold of $R^2 > 0.1$ to select linear sites. This was motivated by observing the distribution of R^2 values and how this threshold discards very constant and multimodal sites.

Measuring person likelihood

Our fundamental measure of person's biological age is the deviation of his methylation value from what is expected at his chronological age. When observing a single linear CpG site, each year is a slice in which the methylation values of the people fall into a distribution. See below for an example of person aged 60 and his value in a distribution for a given age.



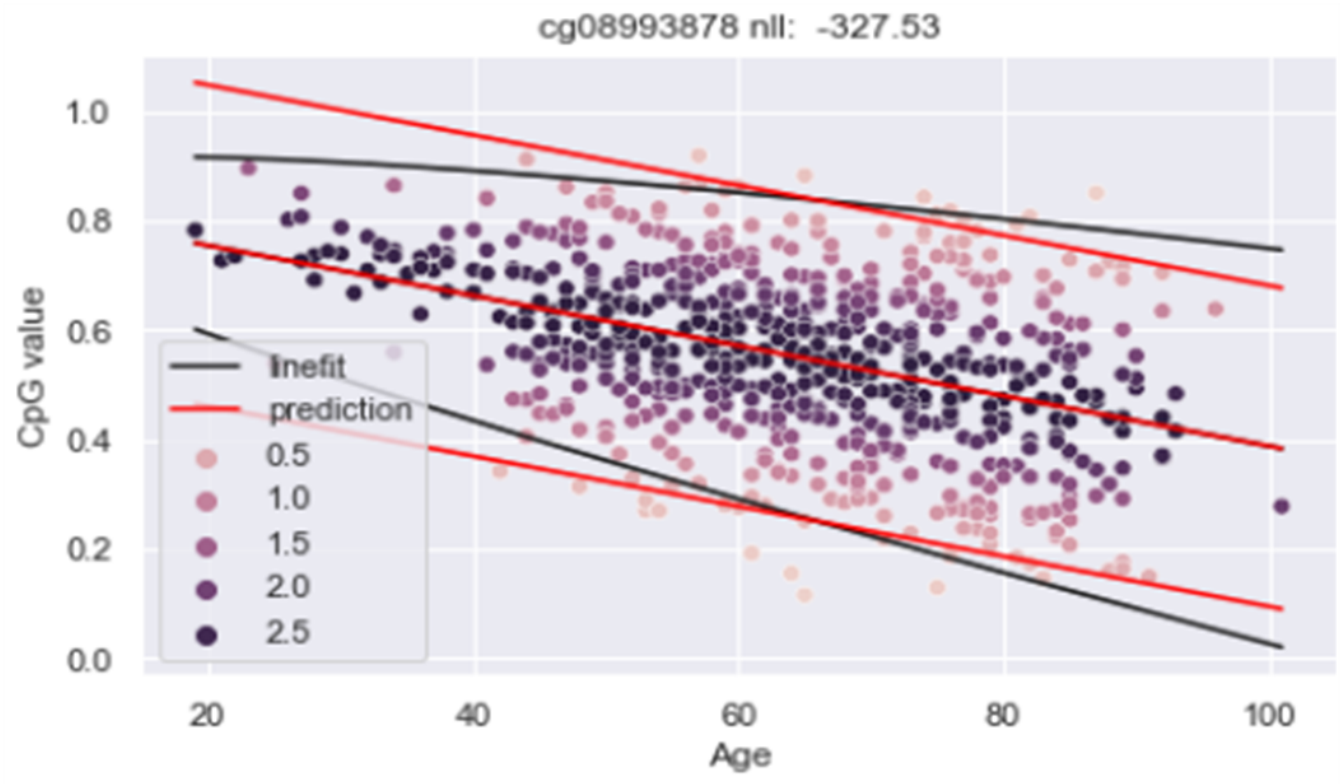
This distribution can be approximated by a normal distribution. Since the mean of the year's distribution changes as the time passes, we can express it using a linear equation $at + b$. The variance is to be constant c. When randomly sampling from the distribution, we are more likely to observe a person closer to the mean; person likeliness is high. This is the probabilistic view of ordinary linear regression, given by an equation.

$$\mathcal{N}(\mu = at + b, \quad \sigma = c)$$

We express person's likelihood in a distribution instead of measuring the distance from the mean because it gives us more information. Like in the Related Work discussion, the distance from the mean does not carry it's meaning when observed in a different distribution. Additional benefits of this approach is the ability to experiment with various models and distribution and understanding which explain the data better.

To measure person's likelihood, we compared two models for each site. The first model was an ordinary linear regression, where the complete probability of a site i, given mean slope a_i and intercept b_i and constant variance c_i, is a product of probability density function for each person methylation value m. Importantly, this model assumes variance that is constant with time. See the figure below for a site plot where each person is coloured by the intensity of its likelihood and its total likelihood expressed with negative log likelihood.

$$P(m|t^j, a_i, b_i, c_i) = \prod_j \mathcal{N}_{pdf}(m;\ a_i t^j + b,\ c_i)$$



The second model we tested was modifying the assumption of constant variance, allowing for variance parameter c_i to change with time.