A smoothed version of the Lassosum penalty for fitting integrated risk models

Georg Hahn, Dmitry Prokopenko, Sharon M. Lutz, Kristina Mullin, Rudolph E. Tanzi, and Christoph Lange

Abstract

Polygenic risk scores are a popular means to predict the disease risk or disease susceptibility of an individual based on its genotype information. When adding other important epidemiological covariates such as age or sex, we speak of an integrated risk model. Methodological advances for fitting more accurate integrated risk models are of immediate importance to improve the precision of risk prediction, thereby potentially identifying patients at high risk early on when they are still able to benefit from preventive steps/interventions targeted at increasing their odds of survival, or at reducing their chance of getting a disease in the first place. This article proposes a smoothed version of the ”Lassosum” penalty used to fit polygenic risk scores and integrated risk models. The smoothing allows one to obtain explicit gradients everywhere for efficient minimization of the Lassosum objective function while guaranteeing bounds on the accuracy of the fit. An experimental section demonstrates the increased accuracy of the proposed smoothed Lassosum penalty compared to the original Lassosum algorithm, allowing it to draw equal with state-of-the-art methodology such as LDpred when evaluated via the AUC (area under the ROC curve) metric.

Keywords: Integrated risk model; Lassosum; Nesterov; Polygenic Risk Scores; Smoothing.

1 Introduction

Polygenic risk scores are a statistical aggregate of risks typically associated with a set of established DNA variants. If only genotype information of an individual is used to predict its risk, we speak of a polygenic risk score. A polygenic risk score with added epidemiological covariates (such as age or sex) is called an integrated risk model (Wand et al., 2020). The goal of both polygenic risk scores and integrated risk models is to predict the disease risk of an individual, that is the susceptibility to a certain disease. Such scores are usually calibrated on large genome-wide association studies (GWAS) via high-dimensional regression of a fixed set of genetic variants (and additional covariates in case of an integrated risk model) to the outcome. In this article, we focus on the more general case of an integrated risk model.

Since the potential for broad-scale clinical use to identify people at high risk for certain diseases has been demonstrated (Khera et al., 2018), polygenic risk scores and integrated risk models have become a widespread tool for the early identification of patients who are at high risk for a certain disease and who could benefit from intervention measures. For this reason, increasing the accuracy of scores is desirable, which is the focus of the proposed smoothing approach.

Several methodological approaches have been considered in the literature to compute a polygenic risk score or an integrated risk model for a given population, and to predict a given outcome (disease status). For instance, LDpred of Vilhjálmsson et al. (2015) and LDpred2 of Privé et al. (2019) fit a Bayesian model to the effect sizes via Gibbs sampling, and obtain a score via posterior means of the fitted model. The PRS-CS approach of Ge et al. (2020) likewise utilizes a high-dimensional Bayesian regression framework in connection with a continuous shrinkage prior (hence the suffix ”CS” for continuous shrinkage) on SNP effect sizes. Fitting genotype data to a disease outcome can also be achieved by means of a simple penalized regression using the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996), for instance using the *glmnet* package on CRAN, see Friedman et al. (2010, 2020).

One popular way to fit a polygenic risk score is the ”Lassosum” approach of Mak et al. (2017). Note that in Mak et al. (2017), no integrated risk models are considered. The Lassosum method is based on a reformulation of the linear regression problem $X\beta = y$, where $X \in \mathbb{R}^{n \times p}$ denotes SNP data for n individuals and p SNP locations, and $y \in \mathbb{R}^n$ denotes a vector of outcomes. The authors start with the classic Lasso objective function $L(\beta) = \|y - X\beta\|_2 + 2\lambda\|\beta\|_1$, where β denotes the Lasso regularization parameter controlling the sparseness of the solution, and rewrite it using the SNP-wise correlation $r = X^\top y$ as

$$L(\beta) = y^\top y + (1-s)\beta^\top X_r^\top X_r \beta - 2\beta^\top r + s\beta^\top \beta + 2\lambda\|\beta\|_1, \tag{1}$$

where X_r denotes the matrix of genotype data used to derive estimates of LD (linkage disequilibrium), λ is the Lasso regularization parameter controlling the sparseness of the estimate, and $s \in (0, 1)$ is an additional regularization parameter used to ensure stability and uniqueness of the Lasso solution. Importantly, in Mak et al. (2017) the authors derive an iterative scheme to carry out the minimization of eq. (1) which only requires one column of X_r at a time, thus avoiding the costly computation of the matrix $X_r^\top X_r \in \mathbb{R}^{p \times p}$.

In this work, we consider a different approach for minimizing eq. (1). Using the methodology of Nesterov (2005), we propose to smooth the non-differentiable L1 penalty in eq. (1), thus allowing us to compute explicit gradients of eq. (1) everywhere. This in turn allows us to efficiently minimize the Lassosum objective function using a quasi-Newton minimization algorithm such as BFGS (Broyden–Fletcher–Goldfarb–Shanno). Besides enabling a more efficient and more accurate computation of the score, our work extends the one of Mak et al. (2017) in that we do not solely consider polygenic risk scores, but the more general integrated risk models. Our approach follows as a special case from Hahn et al. (2020b,a), who propose a general framework to smooth L1 penalties in linear regression. Importantly, employing a smoothing approach has a variety of theoretical advantages following directly from Hahn et al. (2020b). Apart from obtaining explicit gradients for fast and efficient minimization, the smoothed objective is convex, thus ensuring efficient minimization, and it is guaranteed that the solution (the fitted integrated risk model) obtained by solving the smoothed Lassosum objective is never further away than a user-specified quantity from the original (unsmoothed) objective of Mak et al. (2017).

We evaluate all aforementioned approaches by computing an integrated risk model for Alzheimer's disease using the summary statistics of Kunkle et al. (2019) and Jansen et al. (2019). Our simulations demonstrate that smoothing the Lassosum objective function results in a considerably enhanced performance of the Lassosum approach, allowing it to draw equal with approaches such as LDpred or PRS-CS.

Analogously to the original Lasso of Tibshirani (1996), the L_1 penalty employed in eq. (1) causes some entries of $\arg\min_{\beta \in \mathbb{R}^p} L(\beta)$ to be shrunk to zero exactly (provided the regularization parameter λ is not too small). Therefore, Lassosum performs fitting of the polygenic risk score or integrated risk model and variable selection simultaneously. Though superior in accuracy, one drawback of our proposed smoothed version of Lassosum is that it yields dense minimizers, meaning that the variable selection property is not preserved. However, sparseness can be restored after estimation via thresholding.

This article is structured as follows. Section 2 introduces the smoothed Lassosum objective function and discusses its minimization, the theoretical guarantees it comes with, and its drawbacks. Section 3 evaluates the proposed approach, the original Lassosum approach, as well as additional state-of-the-art methods in an experimental study on Alzheimer prediction. The article concludes with a discussion in Section 4. Throughout the article, $[z, 1] \in \mathbb{R}^{q+1}$ denotes the vector obtained by concatenating $z \in \mathbb{R}^q$ and the scalar 1.

2 Methodology

The Lassosum function of eq. (1) consists of a smooth part, given by $y^\top y + (1-s)\beta^\top X_r^\top X_r \beta - 2\beta^\top r + s\beta^\top \beta$, and a non-smooth penalty, the L_1 penalty $2\lambda\|\beta\|_1$. Only the latter needs smoothing, which we achieve with the help of Nesterov smoothing introduced in Section 2.1. Section 2.2 applies the Nesterov methodology to Lassosum and introduces our proposed smoothed Lassosum objective function. The proposed smoothed Lassosum actually follows from the more general framework of Hahn et al. (2020a,b). We demonstrate this in Section 2.3, where we also state the theoretical guarantees following from the framework. The section closes with a remark on variable section and smoothing in Section 2.4.

2.1 Brief overview of Nesterov smoothing

In Nesterov (2005), the author introduces a framework to smooth a piecewise affine and convex function $f : \mathbb{R}^q \to \mathbb{R}$, where $q \in \mathbb{N}$. Since f is piecewise affine, it can be written for $z \in \mathbb{R}^q$ as

$$f(z) = \max_{i=1,\ldots,k} \left(A[z, 1]^\top \right)_i, \tag{2}$$

using $k \in \mathbb{N}$ linear pieces (components). In eq. (2), the linear coefficients of each of the k linear pieces are summarized as a matrix $A \in \mathbb{R}^{k\times(q+1)}$ (with the constant coefficients being in column $q + 1$).

The author then introduces a smoothed version of eq. (2) as

$$f^{\mu}(z) = \max_{w \in Q_k} \left\{ \langle A[z,1]^{\top}, w \rangle - \mu \rho(w) \right\}, \tag{3}$$

where $Q_k = \left\{ w \in \mathbb{R}^k : \sum_{i=1}^k w_i = 1, w_i \geq 0 \ \forall i = 1, \dots, k \right\} \subseteq \mathbb{R}^k$ is the unit simplex in k dimensions. The parameter $\mu \geq 0$ controls the smoothness of the approximation f^{μ} to f, called the Nesterov smoothing parameter. Larger values of μ result in a stronger smoothing effect, while the choice $\mu = 0$ recovers $f^0 = f$. The function ρ is called the proximity (or prox) function which is assumed to be nonnegative, continuously differentiable, and strongly convex.

Importantly, f^{μ} is both smooth for any $\mu > 0$ and uniformly close to f, that is the approximation error is uniformly bounded as

$$\sup_{z \in \mathbb{R}^q} |f(z) - f^{\mu}(z)| \leq \mu \sup_{w \in Q_k} \rho(w) = O(\mu),$$

see (Nesterov, 2005, Theorem 1). Though several choice of the prox function ρ are considered in Nesterov (2005), we fix one particular choice (called the entropy prox function) in the remainder of the article for the following reasons: (a) The different prox functions are equivalent in that all choices yield the same theoretical guarantee and performance; and (b) the entropy prox function leads to a closed-form expression of eq. (3) given by

$$f_e^{\mu}(z) = \mu \log \left(\frac{1}{k} \sum_{i=1}^{k} e^{\frac{\left(A[z,1]^{\top}\right)_i}{\mu}} \right), \tag{4}$$

which satisfies the uniform bound

$$\sup_{z \in \mathbb{R}^q} |f(z) - f_e^{\mu}(z)| \leq \mu \log(k), \tag{5}$$

see Nesterov (2005) and Hahn et al. (2020a,b).

2.2 A smoothed version of the Lassosum objective function

As observed at the beginning of Section 2, it suffices to smooth the non-differentiable penalty $2\lambda \|\beta\|_1$ of the Lassosum objective function, where $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$. To this end, we apply Nesterov smoothing to each absolute value independently.

We observe that the absolute value can be expressed as piecewise affine function with $k = 2$ components, given by $f(z) = \max\{-z, z\} = \max_{i=1,2}\left(A[z,1]^\top\right)_i$, where

$$A = \begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix}$$

and $z \in \mathbb{R}$ is a scalar. Substituting this specific choice of A into eq. (4) leads to a smoothed approximation of the absolute value given by

$$f_e^\mu(z) = \mu \log\left(\frac{1}{2}e^{-z/\mu} + \frac{1}{2}e^{z/\mu}\right). \tag{6}$$

Substituting the absolute value in the L_1 norm in eq. (1) by the entropy prox approximation in eq. (6) results in a *smoothed version of the Lassosum objective function*, given by

$$L^\mu(\beta) = y^\top y + (1-s)\beta^\top X^\top X\beta - 2\beta^\top r + s\beta^\top\beta + 2\lambda\sum_{i=1}^{p} f_e^\mu(\beta_i). \tag{7}$$

The first derivative of f_e^μ is explicitly given by

$$\frac{\partial}{\partial z} f_e^\mu(z) = \frac{-e^{-z/\mu} + e^{z/\mu}}{e^{-z/\mu} + e^{z/\mu}} =: g_e^\mu(z),$$

see also Hahn et al. (2020a,b), from which the closed-form gradient of the smoothed Lassosum objective function of eq. (7) immediately follows as

$$\frac{\partial}{\partial\beta} L^\mu = (1-s)2(X^\top X)\beta - 2r + 2s\beta + 2\lambda\sum_{i=1}^{p} g_e^\mu(\beta_i).$$

Using the smoothed version of the Lassosum objective function, given byL^μ, and its explicit gradient $\frac{\partial}{\partial\beta}L^\mu$, an integrated risk model can easily be computed by minimizing L^μ using a quasi-Newton method such as BFGS (Broyden–Fletcher–Goldfarb–Shanno), implemented in the funcition *optim* in *R* (R Core Team, 2014).

In eq. (7), the quantity X is not limited to contain only genotype information. Any data on the individuals (including additional epidemiological covariates) to compute the integrated risk model can be summarized in X. The other quantities in eq. (7) are the outcome y (either binary/discrete or continuous), the correlations $r = X^\top y$, and the additional regularization parameter $s \in (0,1)$ introduced by Mak et al. (2017) used to ensure stability and uniqueness of the Lasso solution.

2.3 Theoretical guarantees

Using the fact that the absolute value can be expressed as a piecewise affine function with $k = 2$, see Section 2.2, the error bound of eq. (5) simplifies to

$$\sup_{z \in \mathbb{R}} |f(z) - f_e^\mu(z)| \leq \mu \log(2). \tag{8}$$

Since in our proposed smoothed version of eq. (7), only the non-smooth L_1 contribution of the original Lassosum objective function of eq. (1) has been replaced, the bound of eq. (8) immediately carries over to a bound on the smoothed Lassosum. In particular,

$$\sup_{\beta \in \mathbb{R}^p} |L(\beta) - L_e^\mu(\beta)| \leq \sup_{\beta \in \mathbb{R}^p} 2\lambda \left| \sum_{i=1}^p |\beta_i| - \sum_{i=1}^p g_e^\mu(\beta_i) \right| \leq 2\lambda p \mu \log(2). \tag{9}$$

For a given computation of an integrated risk model, the Lasso parameter $\lambda > 0$ and the dimension p are fixed by the problem specification. According to eq. (9), this allows one to make the approximation error of our proposed smoothed Lassosum to the original Lassosum arbitrarily small as the smoothing parameter $\mu \to 0$.

As stated in (Mak et al., 2017, Section 2.1), the Lassosum objective of eq. (1) is equivalent to a Lasso problem, in particular its convexity is preserved. According to (Hahn et al., 2020b, Proposition 2), the smooth approximation of eq. (7) obtained via Nesterov smoothing is strictly convex. Since strictly convex functions have one unique minimum, and since a closed-form gradient $\frac{\partial}{\partial \beta} L^\mu$ of L^μ is available (see Section 2.2), this makes the minimization of our proposed smoothed Lassosum in lieu of the original Lassosum very appealing.

Furthermore, two additional properties of eq. (7) can be derived from (Hahn et al., 2020b, Section 4.3). First, the $\arg\min_{\beta \in \mathbb{R}^p} L^\mu(\beta)$ is continuous with respect to the supremum norm (Hahn et al., 2020b, Proposition 4), which implies that the minimum of our proposed smoothed Lassosum L^μ converges to the one of the original Lassosum as $\mu \to 0$. Second, in addition to this qualitative statement, the error between the minimizers of the smoothed and original Lassosum function can be quantified a priori (Hahn et al., 2020b, Proposition 5).

2.4 A note on variable selection after smoothing

Using an L_1 penalty in eq. (1) has the advantage that, in analogy to the original Lasso of Tibshirani (1996), computing $\arg\min_{\beta \in \mathbb{R}^p} L(\beta)$ performs both regression of the polygenic risk score or integrated risk model and variable selection simultaneously. The smoothed version of Lassosum L^μ we propose typically yields dense minimizers, meaning that the variable selection property is not preserved. As an easy fix, sparseness can be imposed after estimation by thresholding.

3 Experimental studies

We benchmark the performance of our proposed smoothed Lassosum approach of Section 2.2 (referred to as ”smoothed Lassosum”) against the original Lassosum of Mak et al. (2017) (referred to as ”Lassosum”), implemented in the R package *lassosum* (Mak et al., 2020), as well as against the following state-of-the-art approaches:

1. LDpred and LDpred-2 (Vilhjálmsson et al., 2015; Privé et al., 2019) compute a polygenic risk score (but not integrated risk model) by inferring the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel. To run LDpred-2, we employed the implementation in the R package *bigsnpr* on CRAN (Privé et al., 2020). We will refer to this algorithm as ”LDpred-Gibbs”.

2. PRScs of Ge et al. (2019) utilizes a high-dimensional Bayesian regression framework which places a continuous shrinkage prior on SNP effect sizes, an innovation which the authors claim is robust to varying genetic architectures. We use the implementation of Ge et al. (2020) and will refer to it as ”PRScs”.

3. We employ a simple penalized regression using the Lasso of Tibshirani (1996) to fit the genotype data to disease outcome. We employ the *glmnet* package on CRAN, see Friedman et al. (2010, 2020). We will refer to this method as ”Glmnet”.

4. We employ the unsmooted and smoothed Lasso of Hahn et al. (2020a), implemented in the R package *smoothedLasso* on CRAN (Hahn et al., 2020c). We will refer to these methods as ”Lasso” and ”smoothed Lasso”.

5. We train a neural network with the *Keras* interface (Falbel et al., 2020a) to the *Tensorflow* machine learning platform (Falbel et al., 2020b). Before training the neural network, we perform a PCA on the genomic input data (i.e., exclusive the environmental covariates) and summarize the data using 20 principal components. We then train a network with four layers, having 20, 8, 4 and 2 nodes. We employ the leaky relu activation function, a dropout rate of 0.1, a validation splitting rate of 0.1, the *he_normal* truncated normal distribution for kernel initialization, and kernel, bias and activity regularization with L_1 penalty. The last layer employs the sigmoid activation function. The model is compiled for binary crossentropy loss using the Adam optimizer, and evaluated with the AUC metrics using 800 epochs. We will refer to the neural network as ”NN”.

The original Lassosum, LDpred, and PRScs algorithms are only designed to fit polygenic risk scores, but not integrated risk models. To include other epidemiological covariates for these methods, we first perform a linear regression of the epidemiological covariates to the outcome, and then run the aforementioned methods on the residuals. We perform similarly for Glmnet, as well as the unsmoothed and smoothed Lasso as this turned out to be advantageous in our simulations. We always use a Lasso regularization parameter of $\lambda = 2^{-3}$, set the Lassosum regularization parameter s in eq. (1) to $s = 0.5$ (this parameter is used to ensure stability and uniqueness of solution), and employ our proposed smoothed Lassosum of Section 2.2 with the smoothing parameter $\mu = 0.1$.

The data we consider are summary statistics for Alzheimer’s disease (AD). The summary statistics are matched to genotype data for chromosomes 1–22 of 1204 patients available in the Partners Biobank (Partners, 2020). Specifically, we use two types of datasets: the one of clinically defined AD cases of Kunkle et al. (2019), and the one of AD-by-proxy phenotypes of Jansen et al. (2019). The dataset of Kunkle et al. (2019) contains a total of 11480632 summary statistics, given by p-value, effect size (beta), and standard deviation of the effect size. Each entry is characterized by its chromosome number, position on the chromosome, as well as the effect allele and non-effect allele. The dataset of Jansen et al. (2019) contains a total of 13367299 summary statistics in the same format as the one of Kunkle et al. (2019). As epidemiological covariates, we consider age, sex, and apoe status (with classes: none (0), single e4 (1), or e4/e4 (1)).

Partners Biobank (Partners, 2020) is a hospital-based cohort from the MassGeneral Brigham (MGB) hospitals. This cohort includes collected DNA from consented subjects linked to electronic health records. We have obtained a subset in April 2019, which included AD cases and controls. Cases were defined as subjects who were diagnosed with AD (ICD-10). Controls were selected as individuals of age 60 and greater, who had no family history of AD, no diagnosed disease of nervous system (G00-G99) and no mental, behavioral and neurodevelopmental disorders (F01-F99), and had a Charlson Age-Comorbidity Index of 2, 3, or 4. We obtained data, imputed on the Haplotype Reference Consortium (HRC) and performed the following quality control steps. Relatedness was assessed with KING and population structure was assessed with principal components. Principal components were calculated on a pruned subset (PLINK2 parameters: --indep-pairwise 50 5 0.05) of common variants (MAF$>$ 0.1). We excluded subjects which had a KING kinship coefficient $>$ 0.0438 and which were at least 5 standard deviations (sd) away from the mean value of the inbreeding coefficient. We kept only self-reported non-hispanic white (NHW) individuals. There was a total of 1204 subjects (348 cases) left for analysis.

To compare performance across both datasets, we determined the set of variants which are found in both datasets, as well as in the genotype data of the Partners Biobank. We randomly select 20000 loci with the --*thincount* option in PLINK2 (Purcell and Chang, 2019) and made sure that the region from $45000000--46000000$bp on chromosome 19 is excluded while the two apoe loci 19:45411941:T:C and 19:45412079:C:T are kept in the data. This leaves 18055 loci.

In the following simulation study, we considered the datasets of Kunkle et al. (2019) and Jansen et al. (2019) separately and extracted SNP weights based on corresponding effect sizes. Next, we withhold a proportion $p \in \{0.1, \ldots, 0.9\}$ of the pool of Partners genotyped subjects as a validation dataset to fit an integrated risk model with the aforementioned methods, or to tune the hyperparameters of the neural network. Finally, we evaluated their performance on the unseen proportion of the data $(1-p)$. We report the sum of the absolute values of residuals (that is, the difference between the predicted AD outcome and the known outcome in the unseen data), the sum of squared residuals, the AUC (area under the ROC curve), and the vector correlation between predicted outcome and known indicator vector.

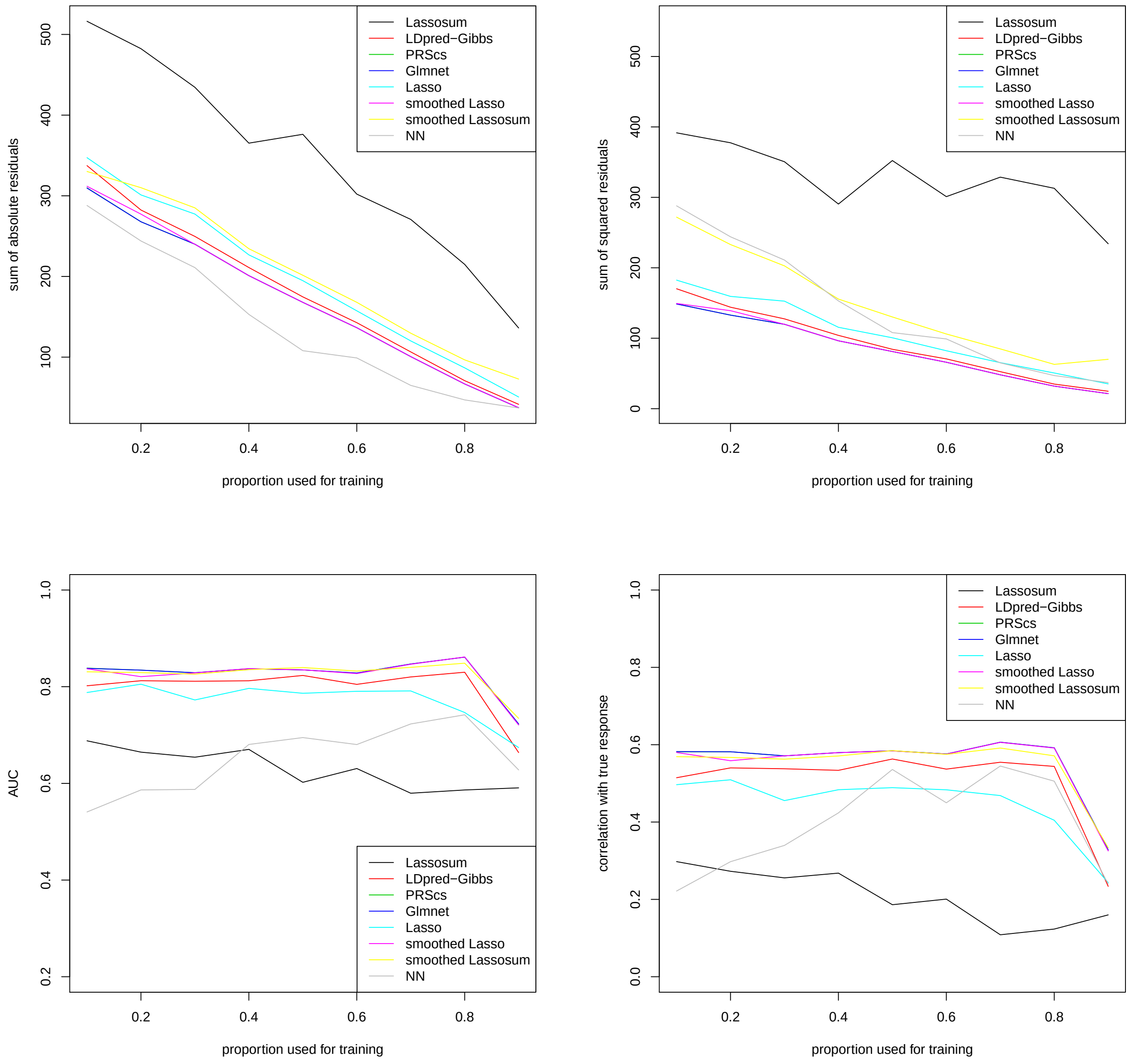

Figure 1: Dataset of clinically defined AD cases of Kunkle et al. (2019). Absolute residual sums (top left), sum of squared residuals (top right), AUC (bottom left), and correlation to the true response (bottom right) as a function of the proportion of data used for training.

Figure 1 shows results for the dataset of Kunkle et al. (2019). A series of observations are noteworthy. First, the absolute residual sum and the sum of squared residuals decreases with an increasing proportion of the data used for training, as expected.

Second, the AUC is very high (around 0.80 to 0.85) for all methods apart from the original Lassosum of Mak et al. (2017) and the neural net approach. Interestingly, it is much less affected than the residuals by the proportion of data used for training and stays almost constant at above 0.8. The reason for the reduced performance of the original Lassosum method is not fully understood. However, it is likely related to the fact that the original Lassosum method is not designed to incorporate epidemiological covariates. To apply Lassosum with epidemiological covariates, we first perform a linear regression of the epidemiological covariates to the outcome, and then run Lassosum on the residuals. To this end, we recompute the SNP-wise correlation $r = X^\top y$ as in eq. (1) using the residuals in place of y. This is a valid approach, and an admissible input to the Lassosum objective function, though the distribution of residuals is different from the one of the original binary response (without regressing out the covariates), which might cause a suboptimal behavior of the original Lassosum algorithm.

Third, our proposed smoothed version of Lassosum considerably improves upon the original Lassosum of Mak et al. (2017), now drawing equal with or even slightly outperforming state-of-the-art methodology such as LDpred with respect to the AUC measure. The correlation of prediction and truth is equally high for all methods apart from the original Lassosum and the neural net. Moreover, our proposed smoothed Lassosum achieves a considerably improved absolute and squared sum of residuals compared to the original Lassosum, achieving a performance that is similar to other state-of-the-art methods. The neural net we trained achieves a very low sum of absolute residuals, though the AUC somewhat lacks behind the other methods for low proportions of training data. The neural net does manage to reach the performance of the state-of-the-art methodology for high proportions of training data (in both the AUC metric and with respect to the correlation of prediction and truth). This is sensible, as neural nets traditionally need large amounts of data to be trained on.

The results for the dataset of Jansen et al. (2019), reported in Figure 2, are qualitatively similar and confirm the aforementioned conclusions.

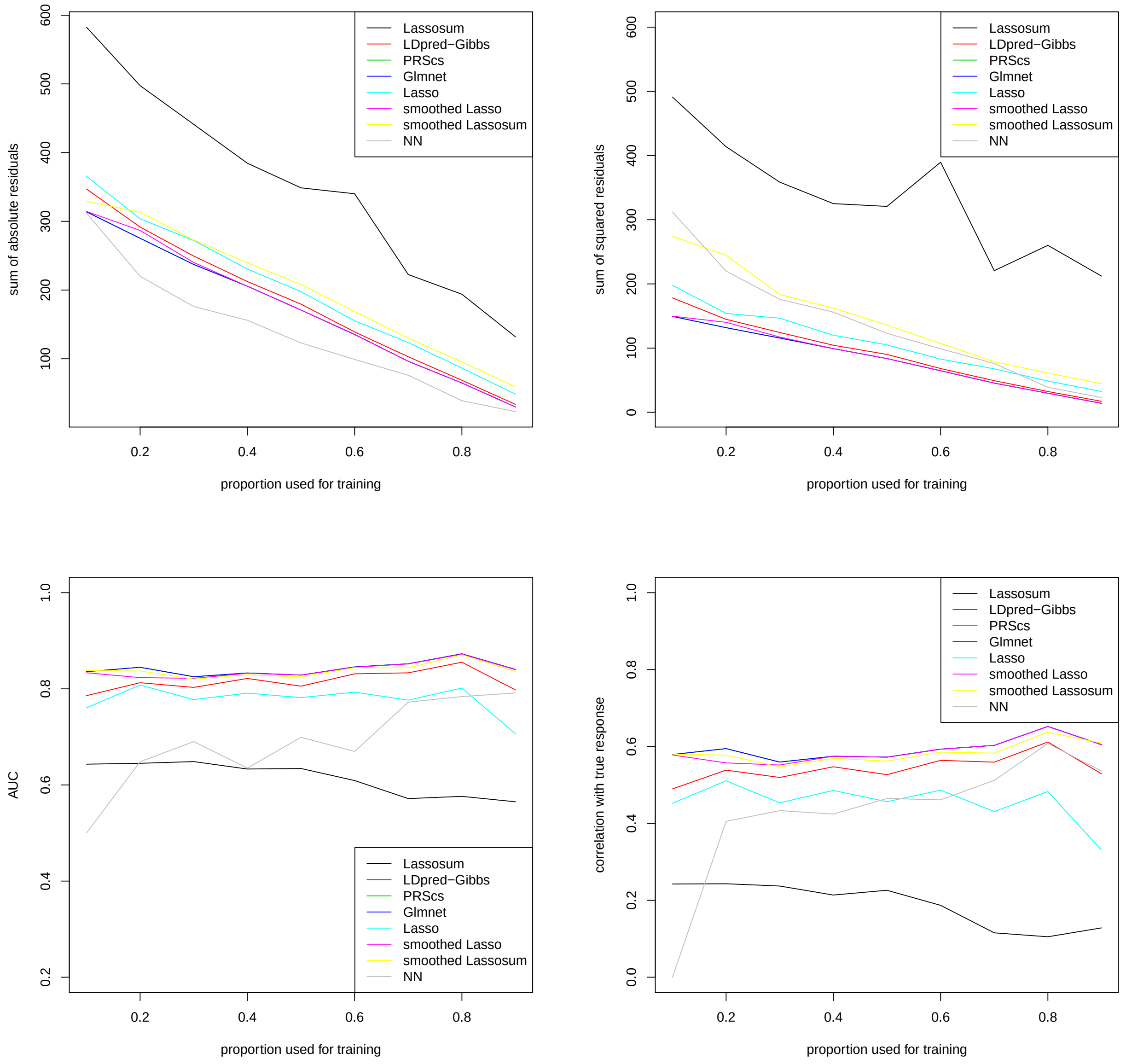

Figure 2: Dataset of AD-by-proxy phenotypes of Jansen et al. (2019). Absolute residual sums (top left), sum of squared residuals (top right), AUC (bottom left), and correlation to the true response (bottom right) as a function of the proportion of data used for training.

4 Discussion

In this article, we considered the calculation of an integrated risk model by means of the Lassosum objective function (see eq. (1)) introduced in Mak et al. (2017). However, in contrast to the originally proposed iterative scheme to minimize eq. (1), we circumvent the non-differentiability of the Lassosum approach caused by the L_1 norm in eq. (1) by using Nesterov smoothing (Nesterov, 2005). Our smoothing approach has many desirable properties.

We show that the smoothed Lassosum objective is differentiable everywhere, and we derive its closed-form gradient. The latter enables us to efficiently compute integrated risk models using quasi-Newton algorithms. We also show that the smoothed objective is strictly convex, thus guaranteeing a unique minimum. To assess the approximation error that is attributable to the Nesterov smoothing of the L_1 norm in the objective function, we prove that the smoothed Lassosum objective function can be made arbitrarily close to the original one proposed in Mak et al. (2017). From a theoretical point of view, these features make the minimization of the smoothed Lassosum in lieu of the original approach very appealing.

To understand the difference between both approaches in practice, we evaluate the proposed smoothed Lassosum together with the original Lassosum of Mak et al. (2017) and other state-of-the-art approaches in an experimental study on Alzheimer's disease, based on the summary statistics of Kunkle et al. (2019) and Jansen et al. (2019). We demonstrate that the smoothing approach improves upon the original formulation of the Lassosum approach (measured with respect to the sum of absolute residuals or the AUC), thus making it draw equal in accuracy with other state-of-the-art approaches.

Acknowledgements

This work involved the use of the Enterprise Research Infrastructure & Services (ERIS) at Massachusetts General Hospital. We thank the MGB/Partners HealthCare Biobank for providing genomic and health information data.

Data Availability Statement

The genotype data used in the simulations is available from the Partners Biobank (Partners, 2020). The summary statistics of Kunkle et al. (2019) and Jansen et al. (2019) used in the simulations are available online, see NIAGADS (2016) and CTG Lab (2021).

Conflict of Interest

The authors declare no conflict of interest.

Funding

The methodology work in this paper was funded by Cure Alzheimer's Fund.

References

CTG Lab (2021). Summary statistics for Alzheimer's dementia from Iris Jansen et al., 2019. https://ctg.cncr.nl/software/summary_statistics.

Falbel, D., Allaire, J., Chollet, F., RStudio, Google, Tang, Y., Bijl, W. V. D., Studer, M., and Keydana, S. (2020a). keras: R Interface to 'Keras'. R-package version 2.3.0.0: https://cran.r-project.org/package=keras.

Falbel, D., Allaire, J., RStudio, Tang, Y., Eddelbuettel, D., Golding, N., Kalinowski, T., and Google (2020b). tensorflow: R Interface to 'TensorFlow'. R-package version 2.2.0: https://cran.r-project.org/package=tensorflow.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2020). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R-package version 4.0: https://cran.r-project.org/package=glmnet.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*, 10:1776.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2020). PRS-CS: a polygenic prediction method that infers posterior SNP effect sizes under continuous shrinkage (CS) priors using GWAS summary statistics and an external LD reference panel. `https://github.com/getian107/PRScs`.

Hahn, G., Lutz, S., Laha, N., Cho, M., Silverman, E., and Lange, C. (2020a). A fast and efficient smoothing approach to LASSO regression and an application in statistical genetics: polygenic risk scores for Chronic obstructive pulmonary disease (COPD). *bioRxiv:2020.03.06.980953*, pages 1–20.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020b). A framework to efficiently smooth l1 penalties for linear regression. *bioRxiv:2020.09.17.301788*, pages 1–35.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020c). smoothedLasso: Smoothed LASSO Regression via Nesterov Smoothing. R-package version 1.5: `https://cran.r-project.org/package=smoothedLasso`.

Jansen, I. E. et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genet*, 51:404–413.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224.

Kunkle, B. W. et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*, 51:414–430.

Mak, T., Porsch, R., Choi, S., Zhou, X., and Sham, P. (2017). Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*, 41(6):469–480.

Mak, T., Porsch, R., Choi, S., Zhou, X., and Sham, P. (2020). Lassosum: a method for computing LASSO/Elastic Net estimates of a linear regression problem given summary statistics from GWAS and Genome-wide meta-analyses. `https://github.com/tshmak/lassosum`.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program. Ser. A*, 103:127–152.

NIAGADS (2016). NG00075 - IGAP Rare Variant Summary Statistics - Kunkle et al. (2019). `https://www.niagads.org/datasets/ng00075`.

Partners (2020). Partners healthcare biobank. `https://biobank.partners.org`.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2019). LDpred2: better, faster, stronger. *Bioinformatics*, btaa1029.

Privé, F., Blum, M., and Aschard, H. (2020). bigsnpr: Analysis of Massive SNP Arrays. R-package version 1.5.2: `https://cran.r-project.org/package=bigsnpr`.

Purcell, S. and Chang, C. (2019). PLINK2.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Stat Comp, Vienna, Austria.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J Roy Stat Soc B Met*, 58(1):267–288.

Vilhjálmsson, B. J. et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*, 97(4):576–592.

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. V., Chatterjee, N., Kooperberg, C., Edwards, K., Vlessis, K., Kinnear, K., Danesh, J. N., Parkinson, H., Ramos, E. M., Roberts, M. C., Ormond, K. E., Khoury, M. J., Janssens, A. C. J., Goddard, K. A., Kraft, P., MacArthur, J. A. L., Inouye, M., and Wojcik, G. (2020). Improving reporting standards for polygenic scores in risk prediction studies. *bioRxiv:2020.04.23.20077099*, pages 1–19.