Sparse Polygenic Risk Score Inference with the Spike-and-Slab LASSO

Junyi Song,[1,*] Shadi Zabad,[1,*] Archer Yang[2] and Yue Li[1,§]

[1]School of Computer Science, McGill University and [2]Department of Mathematics and Statistics, McGill University
[*]Co-first authors [§]Corresponding author: yueli@cs.mcgill.ca

Abstract

Large-scale biobanks, with comprehensive phenotypic and genomic data across hundreds of thousands of samples, provide amble opportunities to elucidate the genetics of complex traits and diseases. Consequently, there is a growing demand for robust and scalable methods for disease risk prediction from genotype data. Performing inference in this setting is challenging due to the high-dimensionality of genomic data, especially when coupled with relatively smaller sample sizes. Popular Polygenic Risk Score (PRS) inference methods address this challenge by adopting sparse Bayesian priors or penalized regression techniques, such as the Least Absolute Shrinkage and Selection Operator (*LASSO*). However, the former class of methods are not as scalable and do not produce exact sparsity, while the latter may over-shrink large coefficients. In this study, we present *SSLPRS*, a novel PRS method based on the Spike-and-Slab LASSO (SSL) prior, which offers a theoretical bridge between the two frameworks. We extend previous work to derive a coordinate-ascent inference algorithm for the *SSLPRS* model that operates on GWAS summary statistics. The software implementation, which uses state-of-the-art techniques to scale PRS inference to millions of genetic variants, has an average runtime of three minutes. To illustrate the statistical properties of the new model, we conducted experiments on nine quantitative phenotypes in the UK Biobank, which showed that *SSLPRS* is competitive with state-of-the-art methods in terms of prediction accuracy. The method also produces highly sparse effect size estimates without excessive shrinkage and, on average, selects 5% fewer variants than the *LASSO*.

Key words: Polygenic risk score, GWAS, High-dimensional analysis, Penalized regression, Spike-and-slab LASSO, Sparsity

Introduction

Polygenic risk scores are emerging as an important tool to quantify the heritable component of complex traits and diseases, with an array of promising clinical applications, including genetic risk stratification and personalized medicine (Torkamani et al., 2018; Lewis and Vassos, 2020). However, inference of polygenic scores from large-scale Genome-wide Association Study (GWAS) data is practically challenging due to two main factors. First, due to privacy concerns, individual-level data is rarely publicly available, and consequently, inference has to be carried out using "GWAS summary statistics" (Pasaniuc and Price, 2017), which are the marginal association statistics per variant, coupled with Linkage-Disequilibrium (LD) matrices that record the pairwise correlations between variants in the dataset. The second factor is the ultra high-dimensional nature of genetic data, with modern biobank initiatives routinely measuring and imputing data for upwards of 20 million genetic markers (Karczewski et al., 2024). Considering that GWAS sample sizes, even for the largest meta-analyses, rarely exceed two million samples (Zhou et al., 2022), this setup calls for robust and scalable regularized or Bayesian regression frameworks.

Previous work has explored many different statistical and algorithmic approaches to meet these challenges (Jayasinghe et al., 2024). This includes sparse penalized regression methods, such as `Lassosum` (Mak et al., 2017), as well as a wide array of sparse Bayesian priors, including Horseshoe (`PRScs`) (Ge et al., 2019), Spike-and-slab (`LDPred`, `VIPRS`) (Privé et al., 2020; Zabad et al., 2023), and sparse mixture priors (`SBayesR`, `MegaPRS`) (Lloyd-Jones et al., 2019; Zhang et al., 2021). While the overall prediction accuracy of these individual methods is roughly comparable (Pain et al., 2021; Zabad et al., 2023), they can differ substantially in terms of computational resource utilization and the statistical properties of their inferred effect sizes. On typical GWAS datasets with 1 million genetic variants, the total wallclock time required can range from 15 minutes for optimization-based approaches (*VIPRS*) up to several hours for Markov-Chain Monte-Carlo (MCMC)-based Bayesian methods (`PRScs`) (Zabad et al., 2023). While Bayesian methods can provide accurate and unbiased effect sizes estimates, their coefficients are not truly sparse, which can complicate interpretation and deploying these predictors in some practical settings. On the other hand, traditional penalized regression methods such as the *LASSO* provide sparse effect sizes estimates, but they are known to over-shrink large coefficients, thus introducing a bias (James et al., 2019).

Here, we bridge the gap between sparse Bayesian methods and traditional penalized regression approaches by developing and testing a new PRS model based on the recently-proposed Spike-and-slab LASSO (SSL) prior (Ročková and George, 2018). The *SSL* prior is a mixture of two Laplace densities that together form a non-concave penalty, shrinking small coefficients to exactly zero while holding larger effect sizes steady (Ročková and George, 2018; Bai et al., 2021). In addition to fast coordinate-ascent inference algorithms, the *SSL* prior comes with many attractive theoretical properties, and, depending on the setting of its hyperparameters, can smoothly interpolate between the *LASSO* and Spike-and-slab priors (Ročková and George, 2018). Our contributions include the following: **(1)** Deriving a version of the *SSL* inference algorithms that operate on GWAS summary statistics. **(2)** Providing a fast and memory-efficient software implementation

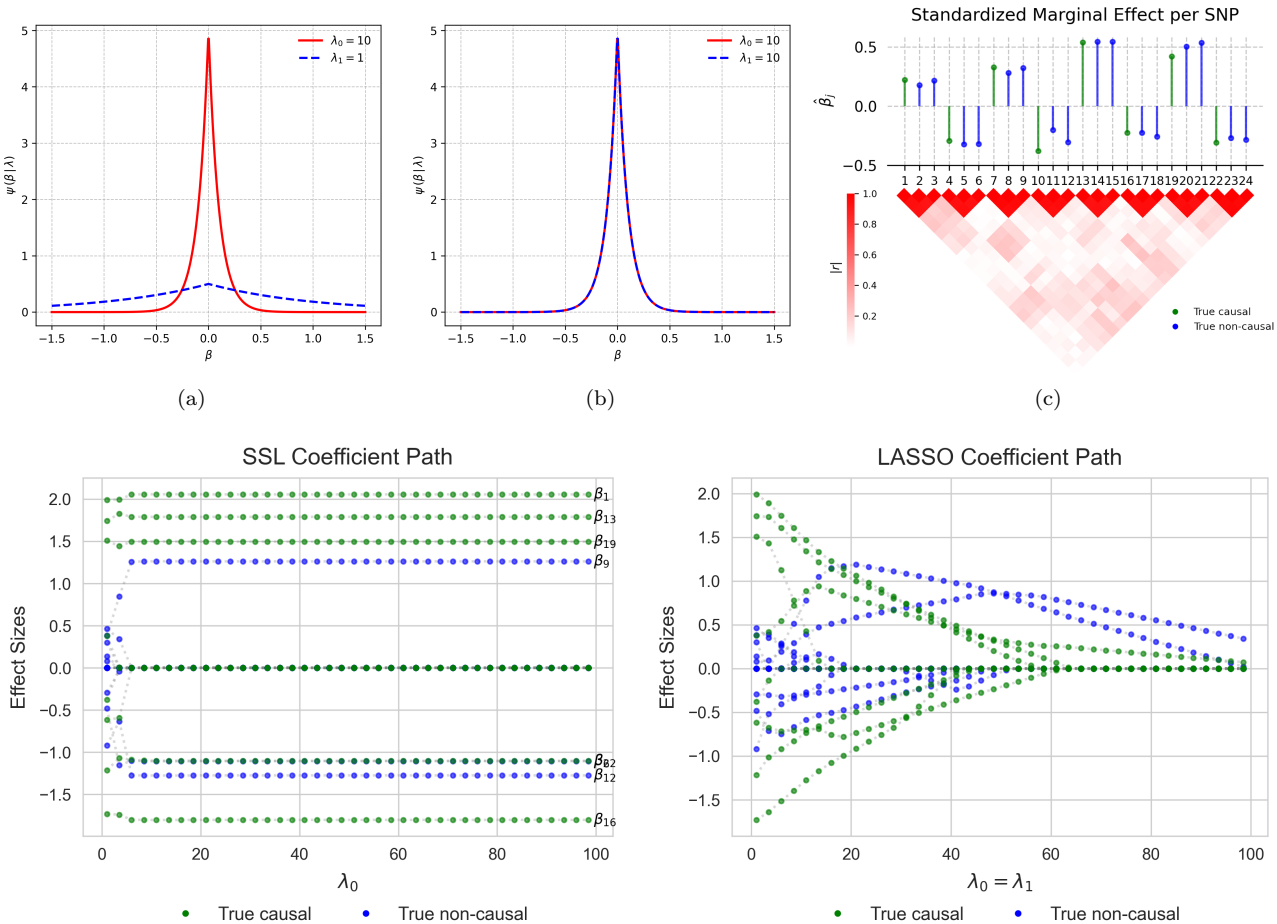


Fig. 1. Overview of the statistical properties and inference dynamics of the *SSL* and *LASSO* on simulated genetic data. (a, b) Illustration of the densities $\psi(\beta \mid \lambda_0)$, $\psi(\beta \mid \lambda_1)$ of the *SSL* prior with (a) $\lambda_0 = 10$ and $\lambda_1 = 1$ (b) $\lambda_0 = \lambda_1 = 10$, which reduces to the *LASSO* case. (c) Simulated genetic data with 24 genetic variants with block-diagonal correlation structure. In this setup, 8 of the variants are casual while 16 are non-causal. Marginal effect sizes are shown in the top panel while absolute pairwise correlations between the variants are shown in the bottom panel. (d, e) Effect size estimates of (d) SSL and (e) *LASSO* models are shown along a ladder of increasing λ_0 values. Ground truth causal variants are shown in green, while non-causal ones are shown in blue. Non-zero effect size estimates from the *SSL* fit (d) are labeled with their corresponding index.

that incorporates state-of-the-art techniques (Zabad et al., 2025) for scaling PRS inference to millions of genetic markers. Open source code is available at `https://github.com/li-lab-mcgill/penprs`. **(3)** Examining the statistical and computational performance of the *SSLPRS* model on real quantitative phenotypes in the UK Biobank (Bycroft et al., 2018). In addition, we provide a new implementation of the *LASSO* model that takes advantage of highly-optimized data structures and algorithms (Zabad et al., 2025), which improves its runtime by more than an order of magnitude compared to the original version (`Lassosum`) (Mak et al., 2017; Zabad et al., 2023).

Methods

Overview of SSLPRS

In a sample of n individuals with paired genotype and phenotype data of p variants, the standard linear model is used to parameterize the dependence of phenotype on the genotype,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypic measurements for each individual, $\boldsymbol{X}$ is an $n \times p$ genotype matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of effect sizes per variant, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector representing the residual effects of the phenotype for each individual. We adapted the *SSL* framework (Ročková and George, 2018), originally operating on individual level data, to support GWAS summary statistics for polygenic risk prediction.

The *SSL* prior primarily consists of two Laplace densities, $\psi(\beta \mid \lambda) = (\lambda/2)e^{-\lambda|\beta|}$ to model the "spike" and the "slab" respectively. This formulation facilitates the adaptive shrinkage of variants based on their covariance and the associations with

the target phenotype:

$$p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{j=1}^{p} (1-\gamma_j)\psi(\beta_j \mid \lambda_0) + \gamma_j \psi(\beta_j \mid \lambda_1),$$

$$p(\boldsymbol{\gamma} \mid \theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \tag{2}$$

where $\theta \sim Beta(a, b)$.

Typically, the scale parameters are set such that $\lambda_0 \gg \lambda_1$, so for a large λ_0, $\psi(\cdot \mid \lambda_0)$ resembles a spike that is dense around zero while a small λ_1 creates a diffuse $\psi(\cdot \mid \lambda_1)$, resembling the slab. Chosen variants that are considered to be non-causal by *SSL* are assigned to the spike for aggressive shrinkage from the large λ_0, effectively driving effect size estimates to be exactly sparse. Meanwhile, variants selected to be causal are assigned to the slab, where the effects remain significant with minimal shrinkage from the relatively much smaller λ_1 value. Here, $\theta \in (0, 1)$ denotes the mixing proportion. The prior can be parameterized by taking the expected rate of $\boldsymbol{\gamma}$:

$$p(\boldsymbol{\beta} \mid \theta) = \prod_{j=1}^{p} [(1-\theta)\psi(\beta_j \mid \lambda_0) + \theta\psi(\beta_j \mid \lambda_1)] \tag{3}$$

The selection of λ_0 and λ_1, specifically the ratio λ_0/λ_1 places *SSL* on the continuum of the *LASSO* case and the ideal spike-and-slab (Figure 1a). In the case where $\lambda_0 = \lambda_1$ or $\lambda_1 = 0$, *SSL* simplifies to the *LASSO* l_1 penalty (Figure 1b). On the other hand, choosing $\lambda_0 \to \infty$ gives a theoretical point-mass spike (Ročková and George, 2018).

SSLPRS inference

Ročková and George (2018); Moran et al. (2019) introduced an efficient coordinate ascent Maximum A Posteriori (MAP) algorithm for *SSL*. In this work, we developed an improved framework of the coordinate ascent algorithm with support for GWAS summary statistics, alongside inference optimizations. The objective function is the log posterior, where under the assumptions that both the genotype matrix $\boldsymbol{X}$ and phenotype vector $\boldsymbol{y}$ have been standardized column-wise for unit variance and zero mean for GWAS summary statistics:

$$L(\boldsymbol{\beta}) = -\frac{N}{2\sigma^2}(1 - 2\sum_{j=1}^{M}\hat{\beta}_j\beta_j + 2\sum_{j=1}^{M}\sum_{m<j} R_{jm}\beta_j\beta_m + \sum_{j=1}^{M}\beta_j^2)$$
$$- (n+2)\log\sigma + \sum_{j=1}^{p}\text{pen}(\beta_j|\theta_j) \tag{4}$$

where N is the GWAS sample size. $\hat{\beta}_j = (1/N)X_j^T\boldsymbol{y}$ denotes the standardized marginal GWAS effect size for SNP j with X_j as the j-th column of $\boldsymbol{X}$. $\boldsymbol{R} = (R_{jm})$ defines the $p \times p$ Linkage Disequilibrium (LD) matrix with $R_{jm} = (1/N)X_j^\top X_m$.

$$\text{pen}(\beta_j \mid \theta) = -\lambda_1|\beta_j| + \log\left[\frac{p_\theta^*(0)}{p_\theta^*(\beta_j)}\right] \tag{5}$$

$$p_\theta^*(\beta_j) = \frac{\theta\psi(\beta_j \mid \lambda_1)}{\theta\psi(\beta_j \mid \lambda_1) + (1-\theta)\psi(\beta_j \mid \lambda_0)} \tag{6}$$

In high-dimensional analysis, the posterior is likely to be multi-modal. The Karush–Kuhn–Tucker (KKT) conditions only give the necessary conditions, in the form of soft-thresholding on β_j, for $\boldsymbol{\beta^*}$ to be a global mode under the *SSL* prior (Eq. 2). Sufficient conditions require the filtering of sub-optimal local modes, which can be achieved through hard-thresholding with Δ (Moran et al., 2019; Ročková and George, 2018; Zhang and Zhang, 2012). Combining these conditions gives

Algorithm 1 SSLPRS Coordinate Ascent

```
Input: λ0, λ1, a, b
Initialize of β^(0) = 0_p, θ^(0) = 0.5, σ^2(0), M = 10
for k in 1 ... k_max do
    Update Δ^(k) by (Eq. 10)
    for j in 1 ... p do
        Update each β_j^(k) with (Eq. 7)
        if M updates of β_j^(k) have occurred then
            Update θ^(k) by (Eq. 12)
        end if
    end for
    if ∀j |β_j^(k) − β_j^(k−1)| ≤ ϵ then
        break
    end if
end for
```

$$\beta_j^* = \frac{1}{n(1+\lambda_{\min})}\left[|z_j| - \sigma^2\lambda_\theta^*(\beta_j^*)\right]_+ \text{sign}(z_j)\mathbb{I}(|z_j| > \Delta) \tag{7}$$

where

$$z_j = N(\hat{\beta}_j - \Sigma_{k\neq j}R_{jk}\beta_k) \tag{8}$$

and λ_θ^* corresponds to the partial derivative of $\text{pen}(\beta_j \mid \theta)$ with respect to $|\beta_j|$ (Ročková and George, 2018) and is expressed as:

$$\lambda_\theta^*(\beta_j) = \lambda_1 p_\theta^*(\beta_j) + \lambda_0(1 - p_\theta^*(\beta_j)) \tag{9}$$

Here, $\Delta \equiv \inf_{t>0}[nt/2 - \sigma^2\text{pen}(t\,|\theta)/t]$, which can be approximated by (Moran et al., 2019):

$$\Delta = \begin{cases} \sqrt{2n\sigma^2\log[1/p_\theta^*(0)]} + \sigma^2\lambda_1 & \text{if } g_\theta(0) > 0 \\ \sigma^2\lambda_\theta^*(0) & \text{otherwise} \end{cases} \tag{10}$$

where

$$g_\theta(\beta_j) = (\lambda_\theta^*(\beta_j) - \lambda_1)^2 + (2n/\sigma^2)\log p_\theta^*(\beta_j) \tag{11}$$

Fitting *SSLPRS* on sparsified LD matrices, which may not be positive semi-definite, may result in excessively large effect size estimates. We introduce an additional hyperparameter $\lambda_{\min}$ in Eq. (7) to ensure numerical stability by regularizing the β_j update and avoiding modes aligned with negative eigenvalues.

The global mode can subsequently be determined through an efficient coordinate ascent algorithm (Algorithm 1), applied to summary statistics, by iteratively updating the model parameters until convergence.

By Lemma 4 from Ročková and George (2018), $\theta^{(k)}$ can be approximated by

$$\theta = \frac{a + \hat{p}_\gamma}{a + b + p} \tag{12}$$

where $\hat{p}_\gamma$ is the number of non-zero coefficients.

In this work, we kept σ^2 constant as a hyperparameter. A variation of the coordinate ascent algorithm is the unknown variance case (Bai et al., 2021), which we present in full for summary statistics in Appendix 7.2.

Algorithm 2 SSLPRS Dynamic Warm-start.

Input: λ_1, a, b and ladder of $\lambda_0 = \{\lambda_0^1, \ldots, \lambda_0^L\}$
Initialize of $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p, \theta^{(0)} = 0.5, \sigma^{2(0)}, M = 10$
for l in $1, \ldots, L$ **do**
    Initialize $\lambda_0 = \lambda_0^l, \boldsymbol{\beta} = \boldsymbol{\beta}^{(l-1)}, \sigma^2 = \sigma^{2(l-1)}, \theta = \theta^{(l-1)}$

    Run coordinate ascent with Algorithm 1.

    $(\boldsymbol{\beta}^{(l)} = \boldsymbol{\beta}, \sigma^{2(l)} = \sigma^2, \theta^{(l)} = \theta)$
end for

Hyperparameter choices

The default *SSLPRS* model operates on the fixed variance case, where we set $\sigma^2 = 1$ based on the assumption that the data is standardized. For hyperparameters a, b in the θ update, we set $b = p$ and $a = \frac{0.05}{1-0.05}b$ so the beta prior mean is 0.05 (Eq. 2), reflecting an estimate of 5% causality in the dataset. The selection of penalty values λ_0, λ_1 scales with the data, both being dependent on $\lambda_{\max} = |N\hat{\beta}_j|$, which is the maximum penalty the *LASSO* case of *SSLPRS* can take before all effect sizes are shrunk to zero (Friedman et al., 2010b). To ensure that the slab is sufficiently diffuse, λ_1 is set to $\epsilon\lambda_{\max}$, where $\epsilon = 10^{-3}$. λ_0 is scaled to be $\lambda_0 = 100\lambda_1$ or 10% of $\lambda_{\max}$. $\lambda_{\min}$ is set to the minimum value for the LD matrix to remain positive semi-definite and numerically stable (Zabad et al., 2023).

Dynamic posterior exploration

Since the choice of hyperparameters λ_0 and λ_1 greatly affects the shrinkage and degree of the spike and slab formulation of *SSLPRS* on the continuum between *LASSO* and the point mass spike and slab, optimizing the performance requires a precise selection of these hyperparameters. Ročková and George (2018) detailed a dynamic posterior exploration strategy, where a ladder of gradually increasing λ_0 values $\{\lambda_0^1, \ldots, \lambda_0^L\}$ are selected while λ_1 is kept at at a sufficiently diffuse value relative to the choices of λ_0 in the ladder. The initial λ_0^1 value in the ladder is selected such that $(\lambda_1 - \lambda_0)^2 < 4$, which makes the objective (Eq. 4) convex (Bai et al., 2021). The solution to the convex problem can be used as a "warm-start" for non-convex problems, which occurs when $\lambda_0 \gg \lambda_1$. In particular, after fitting each λ_0^l on the ladder of spike penalties with Algorithm 1, the MAP estimation of $\boldsymbol{\beta}$, along with model parameters $(\Delta, \theta, \sigma^2)$ are used to "warm-start" the fit for λ_0^{l+1}.

We have also developed and optimized an efficient summary statistics version of the dynamic posterior exploration or warm-start algorithm, which we detail in Algorithm 2. Our experiments show that $\lambda_0^L = \lambda_{\max}$ or a high proportion of $\lambda_{\max}$ acts as an upper bound where *SSLPRS* effectively penalizes all selected irrelevant variants to be zero, essentially "converging" along the coefficient path of the λ_0 ladder. This also provides a strict range of λ_0 values to search in for hyperparameter tuning.

A variation of the dynamic posterior exploration approach is the path-wise fit exploration for the *LASSO* case of *SSLPRS*, where a decreasing ladder is selected instead for warm-starts. This gives a faster and more stable algorithm compared to performing a default fit on the same λ penalty (Friedman et al., 2010b).

Hyperparameter search

In the grid-search version of *SSLPRS*, hyperparameter search is conducted independently for each chromosome on λ_0 for *SSL* and λ for the *LASSO* case. Hyperparameters a, b are kept at their default setting of $a = \frac{0.05}{1-0.05}b, b = p$ and σ^2 is set to 1, similar to the default selection.

By default, the *SSL* case of *SSLPRS* performs a warm-start exploration on a 20-step λ_0 ladder selected on a $\log_2$ scale from λ_1 to $\lambda_{\max}$. The effect size estimates are chosen from the best performing λ_0 value of the ladder based on a chosen metric evaluated with a validation set.

Software and runtime optimizations

The *SSLPRS* model was implemented mainly in `Python`, except computationally expensive update equations in the coordinate ascent algorithm, which was implemented using the compiled language via the `Cython` interface for compatibility and optimal efficiency. We also employed similar optimization strategies used in *VIPRS* (Zabad et al., 2023, 2025), including q-factors to access the LD matrix only once per iteration (Details can be found in Appendix 7.3). We also introduced parallelism in the form of multi-threading across SNP updates in the main coordinate ascent loop (Algorithm 1), which improves overall inference time by multi folds. To reduce overall model peak memory usage, we introduced a low memory mode, which reduces storage requirements by more than a factor of two by storing only the upper-triangular part of the symmetric LD matrix. Further memory optimizations were achieved through floating point quantization to integer representation, minimizing memory footprint for each entry of the LD while having a minimal impact on performance (Zabad et al., 2025). In addition, quantities like $\hat{p}_\gamma$ are updated on the fly to eliminate inefficient re-computations.

A major model specific efficiency optimization we introduced leverages *SSLPRS*' ability to produce exactly sparse effect size estimates. For the subset of β_j coefficients that are identical after the update, we skip expensive computations such as q-factor updates that wouldn't have had any effects on the model parameters. (Appendix 7.3). This greatly speeds up the time per iteration of the coordinate ascent algorithm, especially for large λ_1 penalty values, where many of the effect sizes are shrunk to zero.

Baseline models and their specifications

To assess the relative performance of the *SSLPRS* model on real GWAS data, we compared it to two state-of-the-art PRS inference methods: `Lassosum` (Mak et al., 2017) and *VIPRS* (Zabad et al., 2023). Similar to the *SSL*, both methods use coordinate-wise optimization for inference, but they differ in the priors on the effect size as well as their inference strategies. The *LASSO* uses a single Laplace density, i.e. ℓ_1 penalty, on the effect sizes and performs coordinate ascent on the penalized loglikelihood to estimate the coefficients. VIPRS, on the other hand, assigns a Gaussian Spike-and-slab prior on the effect size and uses Coordinate Ascent Variational Inference (CAVI) to approximate the posterior for the effect sizes. As stated previously, theoretically, the *SSL* can interpolate between these two priors by adjusting its hyperparameters (Ročková and George, 2018), and thus comparing it to these two models in particular can shed light on its statistical and computational properties.

To ensure consistent and fair comparisons across the three models, we used our own optimized implementation of the `Lassosum` model (Mak et al., 2017), as provided in the open source software package `penprs v0.0.1`. By default, the *Lasso* model performs coordinate ascent inference using the pathwise algorithm (Friedman et al., 2010a). For all the experiments,

we used a grid of 20 values for the penalty hyperparameter λ, ranging from λ_{max} to $10^{-3}\lambda_{max}$ on a $\log_2$ scale, where λ_{max} is the minimum penalty beyond which all the coefficients would be set to zero (Friedman et al., 2010a). The coordinate-ascent inference step was parallelized across four threads.

For the *VIPRS* method (Zabad et al., 2023, 2025), we tested the standard EM algorithm that does not require any hyperparameter tuning as well as a Grid Search (*VIPRS-GS*) version that tunes the hyperparameter π, which corresponds to the prior mean on the proportion of causal variants. We used a 20 point grid for π on a $\log_{10}$ scale from 1^{-p} to $1 - 1^{-p}$, where p is the number of variants used during inference. In our experiments, we used the latest version of the software `VIPRS v0.1.2`, which includes optimized and parallel coordinate ascent inference algorithms. The software was run with four threads.

UK Biobank data

To empirically examine the performance of the *SSL* model on real complex traits and diseases, we leveraged paired genotype and phenotype data from 337 205 unrelated White British samples in the UK Biobank Bycroft et al. (2018). The data pre-processing and Quality Control (QC) pipeline were described in detail in earlier reports by the same group (Zabad et al., 2023, 2025) and we briefly summarize below.

The genotype data, which represents the features used for prediction, was extracted by applying standard quality control filters along the sample and variant dimensions. For the samples, in addition to filtering to unrelated White British individuals, we excluded samples with sex chromosome aneuploidy or missing genotype call rate exceeding 5%. The genetic variants were filtered to retain variants in the HapMap3 reference panel (Altshuler et al., 2010), while excluding markers with minor allele count (MAC) <5, minor allele frequency (MAF) <0.1%, imputation score <0.3, Hardy-Weinberg Equilibrium p-value $<10^{-10}$, or genotype missingness rate exceeding 5%. Furthermore, we removed variants with duplicated `rsIDs`, ambiguous strand, or those found in long-range LD regions (Zabad et al., 2023). This resulted in a set of 1 093 308 high quality SNPs that we used in subsequent analyses. Most of these filtering steps were done with `plink2` (Chang et al., 2015).

For the complex traits, which represent the target of our predictors, we extracted measurements for nine quantitative phenotypes for the White British samples described earlier. The phenotypes are standing height (HEIGHT), body mass index (BMI), hip circumference (HC), waist circumference (WC), birth weight (BW), Forced Vital Capacity (FVC), Forced Expiratory Volume in the first second (FEV1), High-density Lipoprotein (HDL), and Low-density Lipoprotein (LDL). Each phenotype was pre-processed by removing outliers (z-score >5), and then within each sex separately, the phenotype was adjusted for age, sex, and 10 Principal Components (PCs) of the genotype matrix. As a final step, we applied the Rank-based Inverse Normal Transform (RINT) to ensure that the distribution of the phenotype is approximately Gaussian.

To facilitate robust analyses of the predictive performance of the *SSL* and baseline models, we performed 5-fold cross-validation on each phenotype. In this instance, for each round, 80% of the data is used for training the models while the remaining 20% are used as held-out test data. Since most of the models require hyperparameter tuning, we further divided the training data of each split into 90% training and 10% validation. Since our methods use GWAS summary data to perform PRS inference, we performed genome-wide association testing within each split and generated GWAS summary statistics for the training, validation, and test sets. The 5-fold cross-validation GWAS data are available at `https://zenodo.org/records/14612130`. The association testing within each split was done with `plink2` (Chang et al., 2015).

Finally, for the Linkage-Disequilibrium (LD) matrices used to record pairwise correlations between genetic variants, we utilized LD matrices for European samples published in a recent study (Zabad et al., 2025), where LD matrices were used as input to PRS inference methods: `https://shz9.github.io/viprs/download_ld/`. The matrices were estimated from 362 446 unrelated European samples in the UK Biobank (Bycroft et al., 2018).

Evaluation metrics and criteria

We evaluated the *SSLPRS* and baseline models based on a number of metrics and criteria for that account for both computational and statistical performance. For the first category of metrics, we examined total wallclock time (in minutes), peak memory utilization (in MB), and inference time per chromosome (in seconds). Total wallclock time includes the time to load the GWAS and LD data, perform harmonization, load software dependencies, perform inference, and finally conduct cross-validation for model selection (Zabad et al., 2025). Peak memory utilization records the maximum amount of Random Access Memory (RAM) used throughout the lifetime of the program in Megabytes. Finally, inference/fit time records the number of seconds it takes to reach convergence using the coordinate-ascent inference procedure.

For the second category, we examined statistical performance using quantitative as well as qualitative approaches. For the quantitative measures, we report prediction accuracy in terms of the proportion of variance explained (R-squared) and the percentage of selected variants for each phenotype. Given a set of inferred coefficients $\boldsymbol{\beta}^*$, the R-Squared on the held-out test set can be approximated from GWAS summary statistics via the pseudo-R^2 metric (Mak et al., 2017; Zabad et al., 2023):

$$\text{pseudo-}R^2 = 1 - 2\boldsymbol{\beta}^{*\top}\widehat{\boldsymbol{\beta}}_{test} + \boldsymbol{\beta}^{*\top}\widehat{\boldsymbol{R}}_{test}\boldsymbol{\beta}^* \tag{13}$$

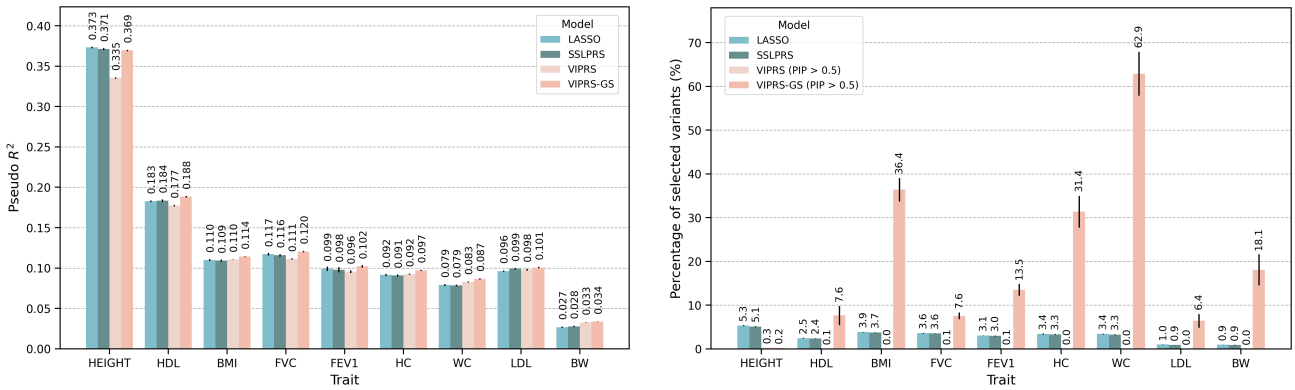
Where $\widehat{\boldsymbol{\beta}}_{test}$ are the standardized marginal GWAS effect sizes from the test set and $\widehat{\boldsymbol{R}}_{test}$ is the corresponding LD matrix. This metric is implemented in the software package `viprs v0.1.2`. For the percentage of selected variants metrics, we simply report the percentage of coefficients that are not exactly zero for the *SSLPRS* and *LASSO* models. For Bayesian models, like *VIPRS*, we simply used the median rule and select variants whose Posterior Inclusion Probability (PIP) is greater than 0.5 (Ishwaran and Rao, 2005).

Finally, we also performed a qualitative assessment to examine potential over-shrinkage in the estimation of large effect variants. By visually comparing the coefficients estimated by the three models, we highlight the statistical properties of the priors and their effects on the estimated effect sizes.

Results

Simulation study

To illustrate the properties of the SSL prior, we designed a simple simulation experiment similar to the setup presented by Bai et al. (2021) and visually illustrated in Figure 1c. The sample size was set to $N = 50$ with $p = 24$ SNPs.



(a) (b)

Fig. 2. Predictive performance of summary statistics-based PRS methods on real quantitative phenotypes in the UK Biobank. (a) Prediction accuracy (Pseudo R-Squared) on held-out test sets and (b) Percentage of selected variants in an analysis of 9 quantitative phenotypes: standing height (HEIGHT), high-density lipoprotein (HDL), body mass index (BMI), forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), hip circumference (HC), waist circumference (WC), low-density lipoprotein (LDL), and birth weight (BW). The bars represent the average model performance based on 5-fold cross-validation, with black vertical lines indicating the associated standard errors. Four summary statistics-based PRS models are shown with different colors: *LASSO*, our proposed SSLPRS, *VIPRS*, *VIPRS-GS*. All models use external validation data for hyperparameter tuning except for VIPRS. For (b), strictly non-zero effect sizes are extracted for the *LASSO* and *SSLPRS* models, and a Posterior Inclusion Probability (PIP) threshold of >0.5 is used to select variants for the *VIPRS* models.

The genotype matrix randomly drawn from a multi-variate Gaussian with block-diagonal covariance matrix $\boldsymbol{X} \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$. To approximate real-world scenarios where genetic variants are in tight LD blocks, the covariance matrix was constructed by aggregating 8 highly correlated 3x3 blocks $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{3}$, $\sigma_{ij} = 0.95$ if $i \neq j$ and $\sigma_{ii} = 1$. Once the genotype data was generated, the true coefficients were set such that 8 causal variants were distributed across the 8 tight LD blocks, with the remaining variants designated to be non-causal. All variants had an effect size of 1.5 on the phenotype, though with varying direction. Once the genotype data and coefficients were generated, the phenotype vector was simulated by following the data-generative process of the linear model: $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \boldsymbol{I})$.

By fitting the *SSLPRS* model on this simulated data with $\lambda_1 = 1$ and a increasing ladder of λ_0 penalty values, we observe from Figure 1d that the model is able to simultaneously perform variable selection and robust estimation of the effect sizes. Here, the coefficients of the non-causal variants are mostly shrunk to zero while the effect sizes of true causal variants are held steady, even at high values for the λ_0 penalty. Despite the tight correlation structure, *SSLPRS* was still able to select most of the true causal variants. In contrast, the *LASSO* demonstrates the well-known issues with excessive shrinkage as the penalty parameter increases in magnitude (Figure 1e). In fact, in this setup, most of the true causal variants are shrunk to zero before many of the non-causal ones, and, with a large enough penalty, all effect sizes are shrunk to 0.

Real phenotype analysis in the UK Biobank

To examine the predictive performance of the *SSLPRS* model on real data, we conducted a 5-fold cross-validation analysis on 9 quantitative phenotypes from the UK Biobank (Bycroft et al., 2018). For each phenotype, we performed inference on GWAS summary statistics for up to 1 million genetic variants. In these analyses, *SSLPRS* was compared to two main baselines: The *LASSO* as well as two versions of the *VIPRS* model, which is a Variational Bayesian PRS method that employs a Gaussian Spike-and-slab prior (Section 2.5). All models used the exactly the same input data and were given the same computational resources for fair comparisons.

Our results show that across all the phenotypes examined, *SSLPRS* shows competitive prediction accuracy compared to the baseline models (Figure 2). Small, but significant differences between the models are observed for some of the phenotypes, such as standing height, where models that use grid search to tune their hyperparameters outperform the EM version of *VIPRS* by up to 10%. Interestingly, the only phenotype where *SSLPRS* significantly outperforms the *LASSO* is for LDL cholesterol (Figure 2), a trait that is known to be modulated by some large effect variants (Graham et al., 2021). This may be a case where bias due to excessive shrinkage hurts the performance of the *LASSO*, as discussed earlier.

Sparsity of effect size estimates

Even though the PRS models examined in our UK Biobank analysis seem hardly distinguishable in terms of their prediction accuracy, Figure 2b shows that they display relatively large difference in terms of their variable selection performance. Notably, *SSLPRS* and the *LASSO* tend to select similar proportion of variants, though the former selects 5% fewer variants on average across the 9 phenotypes. This corresponds to approximately 1000 variants on datasets of this scale. The Variational Bayesian Spike-and-slab models, on the other hand, do not provide exact sparsity and their estimates have to be thresholded post hoc (Ročková and George, 2018). Here, we use the median rule (Ishwaran and Rao, 2005) and select variants with Posterior Inclusion Probability (PIP) greater than 0.5. This threshold results in highly sparse estimates for the default *VIPRS* model, yet using those variants alone may not match the full model's prediction accuracy. This sparsity likely arises from the Variational algorithm's tendency to place most probability mass on one or a few causal variants in tight LD blocks (Carbonetto and Stephens, 2012). The *VIPRS-GS* model seems to select a lot more variants than the other baselines, up to 63% in some cases, without providing proportional improvement in prediction accuracy (Figure 2a).

Inference dynamics and statistical properties of inferred coefficients

Next, we examined the inference dynamics and statistical properties of the inferred effect sizes on subsets of the real GWAS data discussed above. In Figure 3a, we show the coefficient path from a warm-start fit on GWAS summary statistics for standing height on chromosome 5. The path follows a ladder of λ_0 values ranging from $\lambda_0^1 = 1$ to $\lambda_0^L = \lambda_{\max}$, the minimum penalty beyond which the *LASSO* assigns all effect sizes to zero (Friedman et al., 2010b). We observe that *SSLPRS*' behavior on real phenotype data (Figure 3a) is consistent with the trend on the simulated data (Figure 1). As the strength of the spike penalty λ_0 increases along the ladder, most variants are driven towards zero, while the large-effect variants by *SSLPRS* retained by the model. The coefficients eventually stabilize, with minimal changes beyond a certain λ_0 penalty.

To illustrate the SSL prior's ability to perform variable selection without excessive shrinkage, we compared its effect size estimates for the same variants on chromosome 5 to those same estimates from the baseline models (Figure 3bb-c). Our analysis demonstrates *SSLPRS*' ability in mitigating the bias inherent to the *LASSO* through it's selective shrinkage mechanism (James et al., 2019; Ročková and George, 2018): Small effects are shrunk to zero whereas large coefficients are barely affected. In these panels, we see that the *LASSO* systematically over-shrinks the large coefficients whereas the SSL and Gaussian Spike-and-Slab models (*VIPRS*) consistently capture their effects on the phenotype. Curiously, panel (c) of Figure 3b also shows that, compared to *VIPRS*, the SSL model more consistently shrinks coefficients with small but non-zero effects.

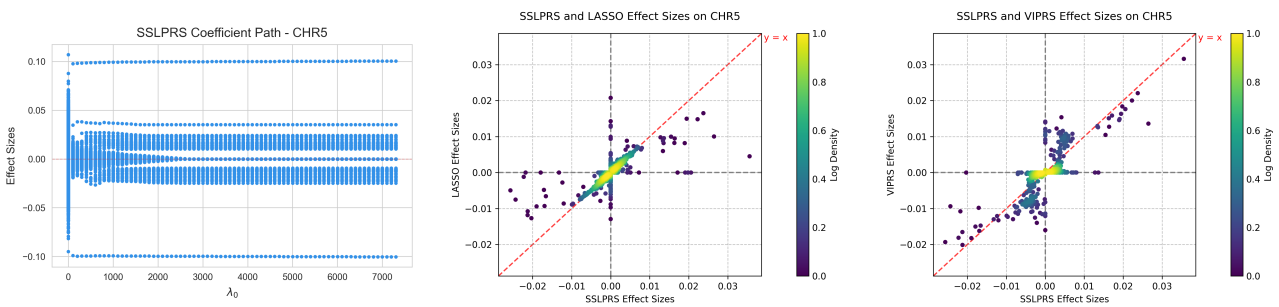
Scalability and computational performance

Throughout the UK Biobank experiments discussed in earlier sections and in Figure 2, we monitored and recorded important model performance characteristics to evaluate their computational performance. These metrics are displayed and summarized in Figure 4. The first metric includes inference time per chromosome, which shows the time it takes for the inference algorithm to reach convergence on GWAS data for a particular chromosome. Figure 4a shows that the runtime scales with the number of variants and that, while the models are overall comparable, the *LASSO* and default *VIPRS* models are the most efficient, reaching convergence in under 5 seconds for all chromosomes. Due to the warm-start fitting strategy, where all variants are systematically updated in earlier iterations, the *SSLPRS* model can be 1-2 slower than the baseline models (Figure 4a).

The other two performance metrics are aggregated over the entire lifetime of the program. Panel (b) of Figure 4 shows the total wallclock time (in minutes), which includes the time to read and harmonize the input data, perform inference, and select models based on cross-validation performance. Compared to state-of-the-art methods, all models displayed here are highly efficient, with average runtime below 5 minutes. *SSLPRS* computational performance is highly competitive given the overall landscape of PRS inference methods, though it is 2-3 times slower than the fastest models tested in this work, which, on average, require less than 90 seconds to reach convergence genome-wide. This impressive speed is combined with modest usage of memory resources, with none of the methods exceeding a Gigabyte at their peak (Figure 4c). The default *VIPRS* uses the least memory since it does not perform model selection. For comparison, most previously published Bayesian and penalized regression methods require 30-120 minutes of wallclock time and several Gigabytes of memory (Jayasinghe et al., 2024; Zabad et al., 2023, 2025).

Discussion

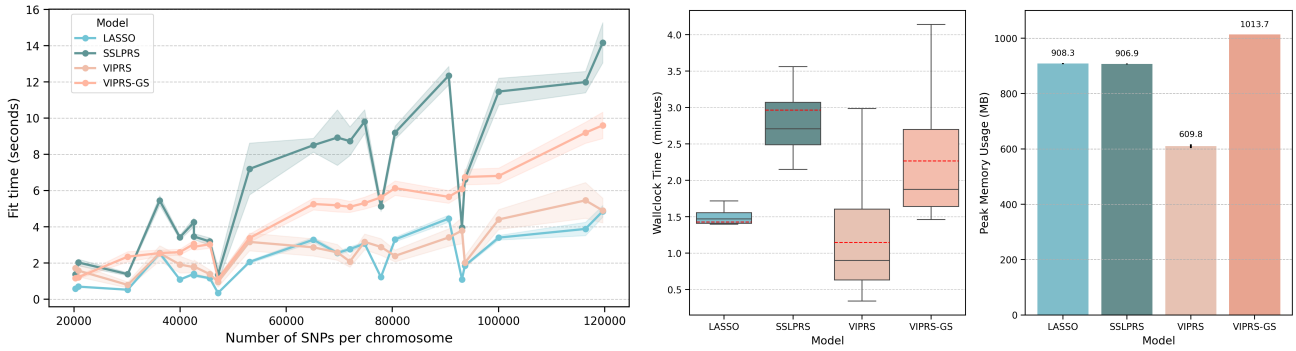
In this work, we presented the novel PRS method *SSLPRS* for sparse and accurate prediction of complex traits from GWAS summary data. Our experiments on real and simulated traits highlight *SSLPRS*' competitive performance against state-of-the-art methods such as the *LASSO* and *VIPRS*. This predictive performance is complemented by its ability to achieve truly sparse effect size estimates without excessive shrinkage, as well as the availability of fast, memory-efficient, and highly scalable algorithms for genome-wide inference. Overall, we believe that *SSLPRS* provides effective analytical tools for predicting and understanding the genetic underpinnings of complex traits and diseases.



(a) (b) (c)

Fig. 3. Inference dynamics and properties of the inferred effect sizes for variants on chromosome 5 for standing height in the UK Biobank.

(a) Coefficient path of *SSLPRS* on a increasing ladder of λ_0 penalties from $\lambda_0^1 = 1$ to $\lambda_0^L = \lambda_{\max}$ with a step size of 100. (b, c) Comparing effect size estimates of the *SSLPRS* model (x-axis) against (b) *LASSO* (c) default *VIPRS* on the y-axis. Coefficients with magnitudes $< 5 * 10^{-4}$ were excluded to reduce visual clutter. Log-density of the points is shown in color and the reference line $y = x$ is shown in red.



(a) (b) (c)

Fig. 4. Computational performance of summary statistics-based PRS methods on real quantitative phenotypes in the UK Biobank with up to 1 million genetic variants. All metrics are shown as aggregate statistics across 5 folds and 9 quantitative phenotypes, for a total of 45 experiments. Panel (a) shows inference time per chromosome (in seconds). Chromosomes are represented on the x-axis in terms of the number of variants, dots show the average runtime and shaded regions show interpolated standard errors. Models are differentiated by color. (b) Boxplot of total *Wallclock* time (in minutes) for each of the four models. The thick horizontal block line shows the median and the red dotted line shows the mean wallclock time for each model. (c) Average *Peak Memory* utilization (in Megabytes) for each of the four models. Small vertical block lines on top of the bars show standard errors.

Competing interests

None declared.

Acknowledgments

We thank members of the Li lab for their feedback and comments on earlier iterations of this work. Y.L. is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare (CRC-2021-00547) and Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2016-05174).

References

David M. Altshuler, Richard A. Gibbs, Leena Peltonen, Stephen F. Schaffner, and et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 2010. ISSN 14764687. doi: 10.1038/nature09298.

Ray Bai, Veronika Ročková, and Edward I George. Spike-and-slab meets lasso: A review of the spike-and-slab lasso. *Handbook of Bayesian variable selection*, pages 81–108, 2021.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0579-z.

Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7, 2012. ISSN 19360975. doi: 10.1214/12-BA703.

Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation plink: Rising to the challenge of larger and richer datasets. *GigaScience*, 4, 2015. ISSN 2047217X. doi: 10.1186/s13742-015-0047-8.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–22, 2010a. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v033i01`.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–22, 2010b. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v033i01`.

Tian Ge, Chia Yen Chen, Yang Ni, Yen Chen Anne Feng, and Jordan W. Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications*, 10, 2019. ISSN 20411723. doi: 10.1038/s41467-019-09718-5.

Sarah E. Graham, Shoa L. Clarke, Kuan-Han H. Wu, Stavroula Kanoni, Greg J. M. Zajac, and Shweta et al. Ramdas. The power of genetic diversity in genome-wide association studies of lipids. *Nature*, 600(7890):675–679, 2021. doi: 10.1038/s41586-021-04064-3. URL `https://doi.org/10.1038/s41586-021-04064-3`.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *Introduction to Statistical Learning with Applications in R*, volume 11. 2019.

Dovini Jayasinghe, Setegn Eshetie, Kerri Beckmann, Beben Benyamin, and S. Hong Lee. Advancements and limitations in polygenic risk score methods for genomic prediction: a scoping review. *Human Genetics*, 2024. doi: 10.1007/s00439-024-02716-8. URL `https://doi.org/10.1007/s00439-024-02716-8`.

Konrad J. Karczewski, Rahul Gupta, Masahiro Kanai, and et al. Pan-uk biobank gwas improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. *medRxiv*, 2024. doi: 10.1101/2024.03.13.24303864. URL `https://www.medrxiv.org/content/early/2024/03/15/2024.03.13.24303864`.

Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: From research tools to clinical instruments. *Genome*

Medicine, 12, 2020. ISSN 1756994X. doi: 10.1186/s13073-020-00742-5.

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang, and Peter M. Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications*, 10, 2019. ISSN 20411723. doi: 10.1038/s41467-019-12653-0.

Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469—480, September 2017. ISSN 0741-0395. doi: 10.1002/gepi.22050. URL https://www.biorxiv.org/content/biorxiv/early/2017/03/22/058214.full.pdf.

Gemma E. Moran, Veronika Ročková, and Edward I. George. Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Analysis*, 14(4):1091–1119, December 2019. doi: 10.1214/19-BA1149.

Oliver Pain, Kylie P. Glanville, Saskia P. Hagenaars, Saskia Selzam, Anna E. Fürtjes, Héléna A. Gaspar, Jonathan R. I. Coleman, Kaili Rimfeld, Gerome Breen, Robert Plomin, Lasse Folkersen, and Cathryn M. Lewis. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLOS Genetics*, 17(5):1–22, 05 2021. doi: 10.1371/journal.pgen.1009021. URL https://doi.org/10.1371/journal.pgen.1009021.

Bogdan Pasaniuc and Alkes L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18, 2017. ISSN 14710064. doi: 10.1038/nrg.2016.142.

Florian Privé, Julyan Arbel, and Bjarni J. Vilhjálmsson. Ldpred2: Better, faster, stronger. *Bioinformatics*, 36, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btaa1029.

Veronika Ročková and Edward I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018. doi: 10.1080/01621459.2016.1260469.

Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19, 2018. ISSN 14710064. doi: 10.1038/s41576-018-0018-x.

Shadi Zabad, Simon Gravel, and Yue Li. Fast and accurate bayesian polygenic risk modeling with variational inference. *The American Journal of Human Genetics*, 110(5):741–761, 2023. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2023.03.009. URL https://www.sciencedirect.com/science/article/pii/S0002929723000939.

Shadi Zabad, Chirayu Anant Haryan, Simon Gravel, Sanchit Misra, and Yue Li. Towards whole-genome inference of polygenic scores with fast and memory-efficient algorithms. *bioRxiv*, 2025. doi: 10.1101/2025.01.19.633789. URL https://www.biorxiv.org/content/early/2025/01/22/2025.01.19.633789.

Cun-Hui Zhang and Tong Zhang. A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems. *Statistical Science*, 27(4):576 – 593, 2012. doi: 10.1214/12-STS399. URL https://doi.org/10.1214/12-STS399.

Qianqian Zhang, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications*, 12(1):4192, 2021. doi: 10.1038/s41467-021-24485-y. URL https://doi.org/10.1038/s41467-021-24485-y.

Wei Zhou, Masahiro Kanai, Kuan-Han H. Wu, Humaira Rasheed, Kristin Tsuo, and et al. Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10):100192, 2022. ISSN 2666-979X. doi: https://doi.org/10.1016/j.xgen.2022.100192. URL https://www.sciencedirect.com/science/article/pii/S2666979X22001410.

Appendix

Appendix A. Derivations of main SSL equations

Objective function. (Eq. 4) We maximize

$$\log p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \theta) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) + p(\boldsymbol{\beta}|\theta) + \text{const}$$
$$= -\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{j=1}^{p} \log p(\beta_j|\theta) + \text{const}$$

First, prior can be marginalized to $p(\beta_j \mid \theta)$ (Eq. 3), which is

$$p(\beta_j \mid \theta) = \sum_{\gamma_j \in \{0,1\}} p(\beta_j, \gamma_j \mid \theta)$$
$$= \sum_{\gamma_j \in \{0,1\}} p(\beta_j \mid \gamma_j) p(\gamma_j \mid \theta)$$
$$= \sum_{\gamma_j \in \{0,1\}} \psi(\beta_j \mid \lambda_{\gamma_j}) \theta^{\gamma_j} (1-\theta)^{1-\gamma_j}$$
$$= (1-\theta)\psi(\beta_j \mid \lambda_0) + \theta\psi(\beta_j \mid \lambda_1)$$

and

$$\log p(\beta_j \mid \theta) = -\lambda_1|\beta_j| + \log\left(\frac{(1-\theta)\,\lambda_0}{\theta\lambda_1} e^{-(\lambda_0-\lambda_1)|\beta_j|} + 1\right) + \text{const}$$
$$= -\lambda_1|\beta_j| + \log\left(\frac{1}{p_\theta^\star(\beta_j)}\right) + \text{const}$$
$$= -\lambda_1|\beta_j| + \log\left(\frac{p_\theta^\star(0)}{p_\theta^\star(\beta_j)}\right) + \text{const}$$
$$= \text{pen}(\beta_j \mid \theta) + \text{const}$$

which gives

$$\log p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \theta) = -\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{j=1}^{p} \text{pen}(\beta_j \mid \theta) + \text{const}$$

Update equations in terms of summary statistics

In the following section, with the assumption for GWAS summary statistics, the following equations hold:

$$X_j^\top X_j = N$$
$$X_j^\top X_m = N R_{jm}$$
$$X_j^T \boldsymbol{y} = N\hat{\beta}_j$$
$$\boldsymbol{y}^T\boldsymbol{y} = N$$

where N is the GWAS sample size, $\boldsymbol{R}$ is the $p \times p$ Linkage Disequilibrium (LD) matrix, and $\hat{\beta}_j$ is the standardized marginal GWAS effect size for SNP j. Then,

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 = \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}$$
$$= N(1 - 2\sum_{j=1}^{M}\hat{\beta}_j\beta_j + 2\sum_{j=1}^{M}\sum_{m<j} R_{jm}\beta_j\beta_m + \sum_{j=1}^{M}\beta_j^2)$$

which gives (Eq. 4, 14) in terms of summary statistics, and Eq.2.2 can be computed in terms of summary statistics by

$$z_j = X_j^\top(\boldsymbol{y} - \sum_{k\neq j} X_k\beta_k)$$
$$= N(\hat{\beta}_j - \sum_{k\neq j} R_{jk}\beta_k)$$

Appendix B. Unknown variance case

$\sigma^{2(k)}$ can be updated in terms of summary statistics by

$$\sigma^2 = \frac{N}{N+2}(1 - 2\sum_{j=1}^{M}\hat{\beta}_j\beta_j + 2\sum_{j=1}^{M}\sum_{m<j} R_{jm}\beta_j\beta_m + \sum_{j=1}^{M}\beta_j^2) \tag{14}$$

and the main coordinate ascent and dynamic posterior exploration algorithms (3, 4) can be adopted to the unknown variance for summary statistics.

Algorithm 3 SSLPRS Coordinate Ascent Unknown Variance.

Input: $\lambda_0, \lambda_1, a, b$
Initialize of $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p, \theta^{(0)} = 0.5, \sigma^{2(0)}, M = 10$
for k in $1 \ldots k_{max}$ **do**
    Update $\Delta^{(k)}$ by (10)
    for j in $1 \ldots p$ **do**
        Update each $\beta_j^{(k)}$ with (7)
        if M updates of $\beta_j^{(k)}$ have occurred **then**
            Update $\theta^{(k)}$ by (12)
            Update $\sigma^{2(k)}$ by (14)
        end if
    end for
    if $\forall j \ |\beta_j^{(k)} - \beta_j^{(k-1)}| \leq \epsilon$ **then**
        break
    end if
end for

Algorithm 4 SSLPRS Warm-start. Unknown Variance.

Input: λ_1, a, b and ladder of $\lambda_0 = \{\lambda_0^1, \ldots, \lambda_0^L\}$
Initialize of $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p, \theta^{(0)} = 0.5, \sigma^{2(0)}, M = 10$
for l in $1, \ldots, L$ **do**
    Initialize $\lambda_0 = \lambda_0^l, \boldsymbol{\beta} = \boldsymbol{\beta}^{(l-1)}, \sigma^2 = \sigma^{2(l-1)}, \theta = \theta^{(l-1)}$
    Initialize $\sigma^2_{\text{update}} = \text{False}$
    if $l \neq 1$ and $k^{(l-1)} < 100$ **then**
        $\sigma^2_{\text{update}} = \text{True}$
    end if

    Run coordinate ascent with Algorithm 3.

    $(\boldsymbol{\beta}^{(l)} = \boldsymbol{\beta}, \sigma^{2(l)} = \sigma^2, \theta^{(l)} = \theta)$
end for

Appendix C. q-factor optimization

To minimize the number of LD accesses per iteration, which can induce a bottleneck in the main coordinate ascent updates, we introduce a q-factor quantity for each SNP in a similar fashion in Zabad et al. (2023). The q-factor, for a SNP j is defined as

$$q_j = \sum_{k \neq j} \beta_k R_{jk}$$

and when β_j is updated to a new estimate $\beta_j^{(new)}$, a partial update to the q factors of SNPs belonging to the neighborhood of SNP j can be done by:

$$q_k \leftarrow q_k + R_{jk}(\beta_j^{(new)} - \beta_j^{(old)}) \quad \forall k \neq j$$

When $\beta_j^{(new)} = \beta_j^{(old)}$, then the q factor for SNP j remains the same. Thus, the q-factor update can be skipped for all neighbors of SNP j, which is the main optimization we perform that takes advantage of *SSLPRS*' exactly sparse estimates. Previously, the LD-matrix R had to be accessed in 3 separate updates and computations of the algorithm, specifically Eq. 2.2, 14 and 4, which has now been reduced to one with the following equations utilizing q-factors:

$$z_j = N(\hat{\beta}_j - q_j)$$

$$\sigma^2 = \frac{N}{N+2}(1 - 2\sum_{j=1}^{M} \hat{\beta}_j \beta_j + \sum_{j=1}^{M} \beta_j q_j + \sum_{j=1}^{M} \beta_j^2)$$

and the objective can be updated as

$$L(\boldsymbol{\beta}) = -\frac{N}{2\sigma^2}(1 - 2\sum_{j=1}^{M} \hat{\beta}_j \beta_j + \sum_{j=1}^{M} \beta_j q_j + \sum_{j=1}^{M} \beta_j^2)$$
$$- (n+2)\log\sigma + \sum_{j=1}^{p} \text{pen}(\beta_j | \theta_j) \tag{15}$$