# Bayesian sparse logistic models in patient-level predictive studies with the R package PatientLevelPrediction

Kelly Li, University of California, Los Angeles
Jenna M. Reps, Janssen Research and Development
Marc A. Suchard, University of California, Los Angeles

## Background

L1-regularized logistic regression is a widely-used technique to help fit large-scale predictive models. It introduces a penalty term that is dependent on the absolute value of the regression coefficients, such that irrelevant covariates in the model have coefficients forced to zero. However, there are a few weaknesses to the model: it suffers from instability when there are correlated predictors, where it will arbitrarily choose one over the other. There also tends to be some bias towards small effect sizes, as L1-regularization shrinks coefficients towards zero, such that predictors with small but meaningful effects may be underestimated for a suboptimal model performance. On the other hand, Bayesian sparse regression utilizes "shrinkage" prior beliefs about the distributions of regression coefficients to estimate a resulting sparse posterior distribution. In this way, Bayesian methods inherently capture model uncertainty by yielding distributions of parameter values instead of point estimates provided by L1-regularization.

One big strength in Bayesian modelling comes in that it incorporates prior knowledge about the model parameters to fit the model. By specifying appropriate prior distributions, the issue of correlation and biasing small coefficients from L1-regularization can be mitigated appropriately. Previously, Bayesian methods to fit a sparse logistic model were of computational order $\mathcal{O}(np^3)$, where $n$ is the number of patients, $p$ is the number of predictors. Nishimura and Suchard (2022) developed a novel approach using the conjugate-gradient method, an existing optimization algorithm, to accelerate posterior simulation to approximately $\mathcal{O}(nps)$, where $s$ is the number of unshrunk predictors that have a relationship with the response variable [1]. The aim of this study is to assess the performance of Bayesian sparse logistic models within the established R package PatientLevelPrediction [2], specifically by utilizing results from a previous model to generate an informative prior that guides the subsequent model.

## Methods

Nishimura and Suchard (2022)'s accelerated Bayesian sparse logistic model is based on the Bayesian Bridge prior on the regression coefficients $\beta_j$[3]:

$$p_0\left(\beta_j|\tau\right) \propto \tau^{-1} \exp\left(-\left|\frac{\beta_j}{\tau}\right|^{\alpha}\right), \tag{1}$$

where $\alpha \in (0, 1]$ controls the shape of the prior distribution for strength of regularization in the model, and $\tau$ is a tuning parameter parameter controlling overall shrinkage applied to the coefficients [1,3].

We also entertain interest in prior distributions on $\beta_j$ that are a mixture of the Bayesian Bridge distribution and an informed normal distribution to transport information between analyses:

$$\beta_j|\tau, \mu_j, \sigma_j^2 \sim \gamma N(\mu_j, \sigma_j^2) + (1-\gamma)p_0(\beta_j|\tau), \tag{2}$$

where $\gamma \in \{0, 1\}$ is a random indicator variable with prior $\gamma \sim Bern(p)$ to determine whether each $\beta_j$ follows the Bayesian Bridge or normal distribution, and $\mu_j$ and $\sigma_j^2$ are pre-specified means and standard deviations for regression coefficients. This approach allows for the incorporation of more informative prior beliefs into the model.

A primary data set of size $10^4 \times 10^3$ will be generated in this study. The 'informed Bayesian model' will use a secondary simulated data set of the same size to estimate posterior means and standard deviations on the regression coefficients to use in an informed mixture prior for the model of interest. This will be evaluated against l1-regularized logistic regression and a Bayesian model using a Bayesian Bridge prior ('uninformed Bayesian model'). Two 'standard' models of each technique will also be fit using a combined primary and secondary data set. The primary comparison measure will be mean-squared error against the true coefficient values from the data generative process.

The same framework highlighted above will also be implemented in a real-world data example investigating hypothyroidism outcomes among patients with pharmaceutically-treated depression, and the primary comparison measures will be predictive performance and sparsity.

Computational tools for this model are implemented in the Python package `bayesbridge`. An R wrapper for the package is available with `bayesbridger` [1]. New functions are implemented in `PatientLevelPrediction` that for sparse Bayesian modelling that call theses packages.

# Code

The following is some sample code for a L1-regularized logistic regression and a Bayesian Bridge regression in `PatientLevelPrediction`:

```r
lr <- setLassoLogisticRegression()
res.lr <- runPlp(plpData = plpData,
                 outcomeId = 2,
                 analysisId = "L1-regularized logistic regression",
                 populationSettings = createStudyPopulationSettings(),
                 splitSettings = createDefaultSplitSetting(),
                 modelSettings = lr)

bayes <- setBayesBridge(n_iter = 1000,
                        n_burnin = 100,
                        bridge_exponent = 0.5,
                        coef_sampler_type = "cg")
res.bayes <- runPlp(plpData = plpData,
                    outcomeId = 2,
                    analysisId = "L1-regularized logistic regression",
                    populationSettings = createStudyPopulationSettings(),
                    splitSettings = createDefaultSplitSetting(),
                    modelSettings = bayes)
```

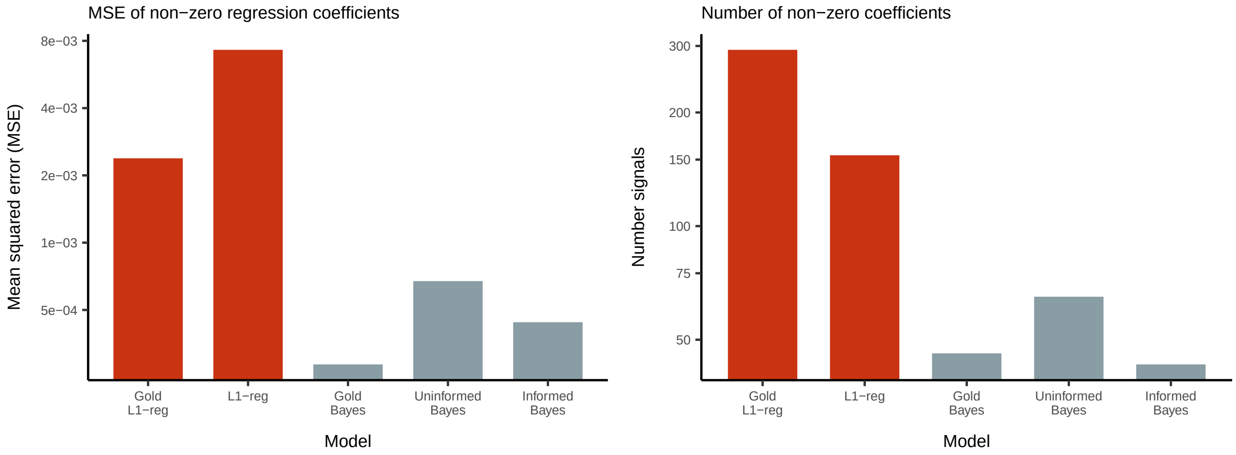
# Results

## Simulated Data Example

We apply `bayesbridger` [1] and `Cyclops` [4] to fit logistic regression models using two simulated data sets of equal size (data sets 1 and 2), and one that combines data sets 1 and 2 together (combined data set).

Regression coefficients are known, with the first 25 coefficients being repeats of the vector $[2.5, 1.5, 1, 0.5, 0.25]$, all else 0.

Figure 1. Mean squared error (MSE) and number of non-zero regression coefficients.

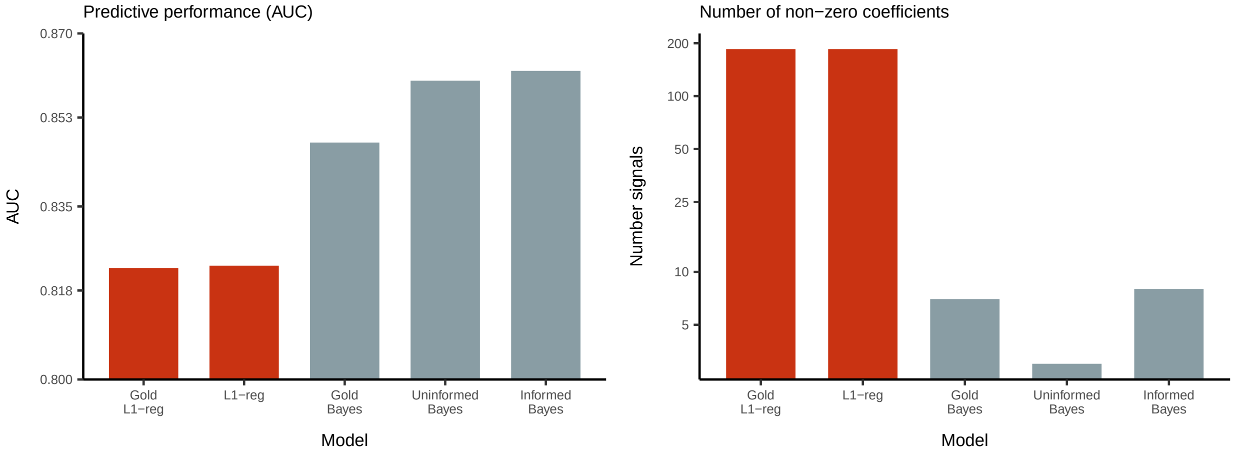


Five different models were fit and results for L1-regularization logistic (red) and Bayesian Bridge logistic regression (grey) are compared. "Gold L1-reg" uses L1-regularized logistic regression on the combined data set. "L1-reg" uses L1-regularized logistic regression on data set 2. "Gold Bayes" uses Bayesian logistic regression with the Bayesian Bridge prior for all regression coefficients on the combined data set. "Uninformed Bayes" uses Bayesian logistic regression with the Bayesian Bridge prior for all regression coefficients on data set 2. "Informed Bayes" uses the posterior means and standard deviations of the regression coefficients from an "Uninformed Bayes" analysis on data set 1 to specify an informed mixture prior on the regression coefficients for Bayesian regression on data set 2. The left figure shows mean-squared error of the non-zero regression coefficients, and the right figure shows the number of non-zero regression coefficients, where non-zero for the Bayesian models is computed by those with more than 50% sampling probability to yield a coefficient with absolute magnitude greater than 0.05. All 5 models yielded similar predictive performance.

## Real world example

We use the R package `PatientLevelPrediction` to fit the same logistic regression models on the following predictive problem: "Amongst patients with pharmaceutically-treated depression, which patients will develop hypothyroidism following the start of the depression episode?" The data set was randomly split into two partitions of equal size, with the first used in the uninformed model and results of the uninformed model used with the second partition in the informed model. The "gold" models were fit using the entire data set. Data was obtained from the IBM MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) mapped to the OMOP common data model.

Figure 2. Predictive performance and number of non-zero regression coefficients.



Five different models were fit with the same framework as the simulated data. The left figure shows predictive performance of the models measured by area under the curve (AUC), and the right figure shows the number of non-zero regression coefficients (signals). Number of signals was determined the same way as for the simulated data.

# Conclusions

Bayesian models consistently exhibit lower mean squared error (MSE) compared to L1-regularized models. Furthermore, incorporating an informed mixture prior derived from a previous fit causes the MSE to approach the performance achieved when both data sets are used simultaneously, as compared to when only the Bayesian Bridge prior is used. Notably, the Bayesian analyses across the board yield sparser models, with the informed model specifying the largest number of non-zero regression coefficients relative to other Bayesian models, but still nearly 5 times fewer than in L1-regularization. These findings support the conclusion that in our data examples, Bayesian regression exhibits reduced bias compared to L1-regularization, and informing the model with the mixture prior further aids in reducing bias. Regardless of the prior chosen, Bayesian analyses yield much sparser models than L1-regularization.

The `PatientLevelPrediction` package now has the ability to fit large-scale Bayesian sparse logistic models in an observational database. These models have strengths over previously implemented L1-regularzied logistic regression in that prior knowledge can be incorporated to help guide the model. However, it is important to note these results come with a higher computational time; However, by using Nishimura and Suchard's (2022) conjugate-gradient based sampler, we are able to successfully implement a scalable Bayesian model to reliably estimate regression coefficients on the log-odds of the probability of an outcome in patient-level predictive models.

There are implications of the results in information transportation in models between data sources. In smaller data sources, a results of a Bayesian Bridge model from a similar data source can be used to pretrain a Bayesian mixture model. It may be of interest the impact of this style of translational learning on parameter estimation, generalization, and data efficiency.

# References

[1] Nishimura A, & Suchard MA (2022). Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in “large n, large p” Bayesian sparse regression. Journal of the American Statistical Association, 1–14. doi:10.1080/01621459.2022.2057859

[2] Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek P (2018). “Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data.” Journal of the American Medical Informatics Association, 25(8), 969-975. doi:10.1093/jamia/ocy032

[3] Polson NG, Scott JG, Windle J (2014). The Bayesian bridge. Journal of the Royal Statistical Society, Series B, 76, 713-733. doi:10.1111/rssb.12042

[4] Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D (2013). “Massive parallelization of serial inference algorithms for complex generalized linear models.” ACM Transactions on Modeling and Computer Simulation, 23, 10. doi:10.1145/2414416.2414791