An Aggregation Method for Sparse Logistic Regression

Zhe Liu

Abstract

We demonstrate and analyze an aggregation method for sparse logistic regression in high-dimensional settings. This approach linearly combines the estimators from various logistic models with different sparsity patterns and can balance the predictive ability and model interpretability. We also study the Kullback-Leibler risk of the aggregation estimator and show that it is comparable to the risk of the best estimator based on a single logistic regression, chosen by an oracle. Numerical performance of the estimator is also investigated using both simulated and real data.

1 Introduction

Logistic regression (LR) is a widely used classification method in many fields such as machine learning, social sciences, bioinformatics, etc. When there are a large number of parameters to be learned, logistic regression is prone to over-fitting. It is well known that $\ell_1$ regularized logistic regression is a promising approach to reduce over-fitting, and can be used for feature selection in the presence of many irrelevant features (Ng, 2004; Goodman, 2004; Lee et al., 2006). However, $\ell_1$ regularization typically selects too many variables and that so-called false positives are unavoidable.

Given a collected family of estimators, linear or convex aggregation methods are another class of techniques to address model selection problems and provide flexible ways to combine various models into a single estimator. A primary motivation for aggregating estimators is that it can improve the estimation risk, as “betting” on multiple models can provide a type of insurance against a single model being poor (Leung and Barron, 2006). Most of the work on aggregation deals with a regression learning problem. For example, exponential screening for linear models provides a form of frequentist averaging over a large model class, which enjoys strong theoretical properties (Rigollet and Tsybakov, 2011). An aggregation classifier is proposed in Lecué (2007) and an optimal rate of convex aggregation for the hinge risk is also obtained.

In this paper, we propose a novel estimating procedure for the regression coefficients in logistic models by considering a linear combination of various estimators with different sparsity patterns. A sparsity pattern is defined as a binary vector with each element indicating
whether the corresponding feature is absent or not. Given any sparsity pattern, we consider
a single logistic regression. The corresponding component weights for individual estimators
are determined to ensure a bounded risk of the aggregation estimator.

Our aggregation procedure is based on the sample-splitting: the first subsample is set to
construct the estimators and the second subsample is then used to determine the weights and
aggregate these estimators. To carry out the analysis of the aggregation step, it is enough
to work conditionally on the first subsample so that the problem reduces to aggregation
deterministic estimators (Rigollet and Tsybakov, 2012). Namely, given deterministic
estimators \( \theta_m \)'s, one can construct aggregation estimator \( \hat{\theta} \) satisfying the following oracle
inequalities

\[
\mathbb{E} R(\hat{\theta}) \leq C \min_{m \in M} R(\theta_m) + \delta_{n,p},
\]

where \( R(\cdot) \) is an empirical risk function, \( \mathbb{E} \) denotes the expectation, \( C \geq 1 \) is a constant, and
\( \delta_{n,p} \geq 0 \) is a small remainder term characterizing the performance of aggregation. Ideally,
we wish to find an aggregation estimator whose risk is as close as possible (in a probabilistic
sense) to the minimum risk of the individual estimators.

The rest of the paper is organized as follows. In Section 2, we describe the logistic
aggregation estimator in detail. Theoretical properties are given in Section 3. Numerical
experiments are presented in Sections 4 and 5. We provide further discussion in Section 6.

2 Logistic Aggregation

We consider data pairs \( \{(x_i, y_i)\}_{i=1}^{n} \) of covariates and responses, sampled from a population,
where each \( x_i \) is a \( p \)-dimensional vector, and \( y_i \in \{0, 1\} \) is a class label. Assume these pairs
satisfy a general setting that

\[
\eta(x_i) := \Pr(y_i = 1|x_i; \theta) = \sigma(x_i^T \theta) = \frac{1}{1 + \exp(-x_i^T \theta)},
\]

where \( \theta \in \mathbb{R}^p \) is a vector of parameters for regression coefficients, and \( \sigma(\cdot) \) is the sigmoid
function.

We call a sparsity pattern any binary vector \( m \in M := \{0, 1\}^p \). The \( i \)th coordinate of \( m \)
can be interpreted as indicators of presence (\( m_i = 1 \)) or absence (\( m_i = 0 \)) of the \( i \)th feature.
We denote by \( |m| \) the number of ones in \( m \).

The aggregation is based on a splitting of the sample into two independent subsamples
\( D_{n_1}^{(1)} \) and \( D_{n_2}^{(2)} \) of size \( n_1 \) and \( n_2 \), respectively, where \( n_1 + n_2 = n \). The first subsample \( D_{n_1}^{(1)} \)
is used to construct the estimators and the second subsample \( D_{n_2}^{(2)} \) is used to aggregate them.
In what follows we will denote \( D_{n_1}^{(1)} := \{(x_i^{(1)}, y_i^{(1)})\}_{i=1}^{n_1} \) and \( D_{n_2}^{(2)} := \{(x_i^{(2)}, y_i^{(2)})\}_{i=1}^{n_2} \).

For each \( m \in M \), consider a logistic regression estimator of the true regression coefficient
\( \theta \) with constraints on the sparsity pattern represented by \( m \)

\[
\hat{\theta}_m := \arg \max_{\{\theta; \theta_i = 0 \text{ for any } m_i = 0\}} \ell(\theta; D_{n_1}^{(1)}),
\]

where
where the log-likelihood

$$l(\theta; D_{n_1}^{(1)}) := \sum_{i=1}^{n_1} \left[ y_i^{(1)} \theta^T x_i^{(1)} - \log(1 + \exp(\theta^T x_i^{(1)})) \right].$$

(4)

Note that each \( \hat{\theta}_m \) is constructed using only data \( D_{n_1}^{(1)} \). The sparsity pattern \( m \) determines the zero-patterns in the estimator \( \hat{\theta}_m \) in the sense that the \( i \)th coordinate of \( \hat{\theta}_m \) is zero if \( m_i = 0 \). Working conditionally on the first subsample, we only need to consider the aggregation of deterministic estimators \( \hat{\theta}_m \).

**Definition 1.** The logistic aggregation (LA) is defined as a linear combination of estimators \( \hat{\theta}_m \)

$$\hat{\theta}^{LA} := \sum_{m \in M} w_m \hat{\theta}_m,$$

(5)

where the data-determined weights are defined as

$$w_m := \frac{\exp(l(\hat{\theta}_m; D_{n_2}^{(2)}))\pi_m}{\sum_{m' \in M} \exp(l(\hat{\theta}_{m'}; D_{n_2}^{(2)}))\pi_{m'}}.$$  

(6)

Here, \( \pi_m \) is a probability distribution on the set of sparsity pattern \( M \), defined by

$$\pi_m := \frac{1}{H} \left( \frac{|m|}{2e} \right)^{|m|},$$

(7)

where \( e \) is the base of the natural logarithm, and \( H \) is a normalization factor.

From the definition, notice that the LA estimator linearly combines a set of logistic regression estimators with various underlying sparsity patterns. Based on this data-splitting technique, the first subsample is set to format these individual estimators \( \hat{\theta}_m \), for all \( m \in M \). The second subsample is used to determine the exponential weighting for each model or estimator. In particular, each estimator is evaluated using an independent dataset, whose likelihood would serve as the weight for model aggregation. It is sometimes useful to incorporate a deterministic factor \( \pi_m \) in the model to account for model complexity or model preference (Leung and Barron, 2006). In this case, low-complexity models are favored. Finally the combined weights \( w_m \) is normalized to have unit sum.

To implement the estimating procedure, exact computation of the LA estimator requires the calculation of at most \( 2^p \) individual estimators. In many cases this number could be extremely large, and we must make a numerical approximation. Observing that the LA estimator is actually the expectation of a random variable that has a probability mass proportional to \( w_m \) on individual estimator \( \hat{\theta}_m \) for \( m \in M \), the Metropolis-Hastings algorithm can be exploited to provide such an approximation.

**Algorithm 1.** The LA estimator can be approximated by running a Metropolis-Hastings algorithm on a \( p \)-dimensional hypercube:
S1. Initialize $u_t = \{0\}^p$, $t = 0$;

S2. For each $t \geq 0$, generate $u'_t$ following an uniform distribution on the neighbors of $u_t$ in the $p$-dimensional hypercube;

S3. Generate a $[0, 1]$-uniformly distributed number $r$;

S4. Put $u_{t+1} \leftarrow u'_t$, if $r < \min\{1, w_{u'_t}/w_{u_t}\}$; otherwise, set $u_{t+1} \leftarrow u_t$;

S5. Compute $\hat{\theta}_{u_{t+1}}$. Stop if $t > T_0 + T$; otherwise, update $t \leftarrow t + 1$ and go to step S2.

Then we can approximate $\hat{\theta}^\text{LA}$ by

\[
\hat{\theta}^\text{LA} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \hat{\theta}_{u_t},
\]

where $T_0 \geq 0$ and $T \geq 0$ are arbitrary integers.

The resulting Markov chain ensures the ergodicity of the chain. The Metropolis-Hastings algorithm incorporates a tradeoff between sparsity and prediction to decide whether to add or remove a feature.

Notice that the LA estimator itself would always give an estimate of $\theta$ in which all the elements are non-zero, since all the possible individual estimators are linearly mixed. However, the Metropolis-Hastings algorithm would lead to a sparse estimate as in the linear regression case (Rigollet and Tsybakov, 2011). Thus, such approximated aggregation estimator can also be used for the task of feature selection.

In high-dimensional settings where $p \gg n$, however, the Metropolis-Hastings algorithm becomes computationally unsuitable. In this case, we use $\ell_1$-penalized logistic regression to approximate the model space and select a candidate set of features. The idea is not new: a similar technique is addressed in Fraley and Percival (2013), where Bayesian model averaging is combined with Markov chain Monte Carlo model composition by using the $\ell_1$ regularization path as the model space for approximation.

To be more specific, we first partition the dataset into two parts $D^{(1)}_{n_1}$ and $D^{(2)}_{n_2}$. We identify the candidate set of features by applying the $\ell_1$ regularization logistic regression using only data $D^{(1)}_{n_1}$. Let $S$ be the set of selected features. Then our aggregation process is constructed on this subset of features.

The reduced set of sparsity patterns is defined by

\[
\mathcal{M} := \prod_{i=1}^{p} C_i,
\]

where

\[
C_i := \begin{cases} 
\{0, 1\} & \text{if the } i\text{th feature } \in S, \\
\{0\} & \text{otherwise}
\end{cases}
\]
Thus the Metropolis-Hastings algorithm introduced in Algorithm 1 will be applied to the reduced set of sparsity patterns $M'$ instead of $M = \{0, 1\}^p$.

In the following algorithm, we summarize this combination approach for applying the Metropolis-Hastings algorithm to approximate the LA estimator with the use of model space generated by $\ell_1$ regularization logistic regression.

**Algorithm 2.** Combine the Metropolis-Hastings algorithm with $\ell_1$ regularization logistic regression to approximate the LA estimator:

1. Partition the data into two sets $D_{n_1}^{(1)}$ and $D_{n_2}^{(2)}$;
2. Using $D_{n_1}^{(1)}$, fit $\ell_1$-penalized logistic regression and let $S$ denote the set of selected features;
3. Apply Algorithm 1 to the reduced set of sparsity patterns represented by $M'$.

### 3 Theoretical Properties

In this section, we study the theoretical properties and derive an oracle inequality for the logistic aggregation estimator.

For any estimator $\hat{\theta}_m$ constructed using data $D_{n_1}, m \in M$, consider the Kullback-Leibler (KL) loss given by

$$KL(\hat{\theta}_m; \{x_i^{(2)}\}) = \sum_{i=1}^{n_2} D(\eta(x_i^{(2)}))^T \sigma(\hat{\theta}_m^T x_i^{(2)}),$$

where the KL divergence of discrete probability distributions $Q$ from $P$ is defined to be

$$D(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

In the following theorem, we state the optimality of our aggregation estimator. In particular, we show that the KL risk of the logistic aggregation estimator is bounded (in a probabilistic sense) by the minimum risk of a set of individual logistic regression estimators and an extra term that quantifies the price one pays for making the aggregation estimator.

**Theorem 1.** Using the KL loss as a criterion, the LA estimator satisfies the inequality

$$\mathbb{E}_{D_1,D_2}\{KL(\hat{\theta}_{LA}; \{x_i^{(2)}\})\} \leq \min_{m \in M} \left\{ \mathbb{E}_{D_1}\{KL(\hat{\theta}_m; \{x_i^{(2)}\})\} + |m| \log \left( 1 + \frac{ep}{|m| \vee 1} \right) + \log 2 \right\} + \mathbb{E}_{D_1,D_2}\{ \max_{m \in M} \left| g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; D_{n_2}^{(2)}) \right| \}.$$ 

12
where \( g(\hat{\theta}_m; \{x_i^{(2)}\}) \) is defined by

\[
\sum_{i=1}^{n_2} \left[ \eta(x_i^{(2)})\hat{\theta}_m^{T}x_i^{(2)} - \log(1 + \exp(\hat{\theta}_m^{T}x_i^{(2)})) \right],
\]

(13)

and \( \mathbb{E}_{D_1,D_2} \) represents the expectation with respect to \( \{y_i^{(1)}\} \) and \( \{y_i^{(2)}\} \).

**Proof.** Notice that

\[
\text{KL}(\hat{\theta}_m; \{x_i^{(2)}\})
\]

\[
= \sum_{i=1}^{n_2} \left[ \eta(x_i^{(2)}) \log(\eta(x_i^{(2)})(1 + \exp(-\hat{\theta}_m^{T}x_i^{(2)}))) + (1 - \eta(x_i^{(2)})) \log((1 - \eta(x_i^{(2)}))(1 + \exp(\hat{\theta}_m^{T}x_i^{(2)})))) \right]
\]

\[
= \sum_{i=1}^{n_2} \left[ \eta(x_i^{(2)}) \log \eta(x_i^{(2)}) + (1 - \eta(x_i^{(2)})) \log(1 - \eta(x_i^{(2)})) - \eta(x_i^{(2)})\hat{\theta}_m^{T}x_i^{(2)} + \log(1 + \exp(\hat{\theta}_m^{T}x_i^{(2)})) \right],
\]

According to the convexity of KL loss, we obtain

\[
\text{KL}(\hat{\theta}_{LA}; \{x_i^{(2)}\}) \leq \sum_{m \in \mathcal{M}} w_m \text{KL}(\hat{\theta}_m; \{x_i^{(2)}\}).
\]

Observe that

\[
\log(w_m) = \log(\pi_m) + l(\hat{\theta}_m; D_n^{(2)}) - \log\left( \sum_{m' \in \mathcal{M}} \exp(l(\hat{\theta}_{m'}; D_n^{(2)}))\pi_{m'} \right).
\]

With \( \hat{m} \in \mathcal{M} \) being any sparsity pattern attaining

\[
\min_{m \in \mathcal{M}} \{\text{KL}(\hat{\theta}_m; \{x_i^{(2)}\}) - \log(\pi_m)\},
\]

we have

\[
\log\left( \frac{w_m}{w_{\hat{m}}} \right) = \log\left( \frac{\pi_m}{\pi_{\hat{m}}} \right) + l(\hat{\theta}_m; D_n^{(2)}) - l(\hat{\theta}_{\hat{m}}; D_n^{(2)}).
\]
Notice that
\[
KL(\hat{\theta}^{LA}; \{x_i^{(2)}\}) \leq KL(\hat{\theta}_m; \{x_i^{(2)}\}) + \sum_{m \in M} w_m (KL(\hat{\theta}_m; \{x_i^{(2)}\}) - KL(\hat{\theta}_m; \{x_i^{(2)}\}))
\]
\[
= KL(\hat{\theta}_m; \{x_i^{(2)}\}) + \sum_{m \in M} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - g(\hat{\theta}_m; \{x_i^{(2)}\}))
\]
\[
= KL(\hat{\theta}_m; \{x_i^{(2)}\}) + \sum_{m \in M} w_m (l(\hat{\theta}_m; \mathcal{D}_{n_2}) - l(\hat{\theta}_m; \mathcal{D}_{n_2})) + \sum_{m \in M} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}_{n_2})).
\]

Observe that
\[
\sum_{m \in M} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}_{n_2}))
\]
\[
= \sum_{m \in M} w_m \log \left( \frac{\pi_m}{\pi_{\hat{m}}} \right) - \sum_{m \in M} w_m \log \left( \frac{w_m}{w_{\hat{m}}} \right)
\]
\[
\leq \log \left( \frac{1}{\pi_{\hat{m}}} \right).
\]

The last inequality follows from the fact that
\[
\sum_{m \in M} w_m \log \left( \frac{w_m}{w_{\hat{m}}} \right) \geq 0,
\]
and \(\log(\pi_{\hat{m}}) \leq 0\). Thus we obtain
\[
KL(\hat{\theta}^{LA}; \{x_i^{(2)}\}) \leq KL(\hat{\theta}_m; \{x_i^{(2)}\}) + \log \left( \frac{1}{\pi_{\hat{m}}} \right) + (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}_{n_2})) - \sum_{m \in M} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}_{n_2})).
\]

Take expectations with respect to \(\mathcal{D}_{n_2}\) on both sides and notice the fact that for any \(m \in M\)
\[
\mathbb{E}_{\mathcal{D}_{n_2}} \{l(\hat{\theta}_m; \mathcal{D}_{n_2})\} = g(\hat{\theta}_m; \{x_i^{(2)}\}),
\]
thus we obtain
\[
\mathbb{E}_{\mathcal{D}_{n_2}} \{KL(\hat{\theta}^{LA}; \{x_i^{(2)}\})\}
\]
\[
\leq \min_{m \in M} \left\{ KL(\hat{\theta}_m; \{x_i^{(2)}\}) + \log \left( \frac{1}{\pi_m} \right) \right\} - \mathbb{E}_{\mathcal{D}_{n_2}} \sum_{m \in M} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}_{n_2})).
\]
Note that the LA estimator $\hat{\theta}^{LA}$ is constructed using both $\mathcal{D}^{(1)}_{n_1}$ and $\mathcal{D}^{(2)}_{n_2}$, thus the expectation of the KL risk of the LA estimator should be taken with respect to both subsamples. This is also the case for the weight $w_m$.

Taking expectations with respect to $\mathcal{D}^{(1)}_{n_1}$ on both sides and applying

$$
\mathbb{E}_{\mathcal{D}^1} \min_{m \in \mathcal{M}} \left\{ \text{KL}(\hat{\theta}_m; \{x_i^{(2)}\}) + \log\left(\frac{1}{\pi_m}\right) \right\}
\leq \min_{m \in \mathcal{M}} \left\{ \mathbb{E}_{\mathcal{D}^1} \{\text{KL}(\hat{\theta}_m; \{x_i^{(2)}\})\} + \log\left(\frac{1}{\pi_m}\right) \right\},
$$

we have

$$
\mathbb{E}_{\mathcal{D}^1, \mathcal{D}^2} \{\text{KL}(\hat{\theta}^{LA}; \{x_i^{(2)}\})\}
\leq \min_{m \in \mathcal{M}} \left\{ \mathbb{E}_{\mathcal{D}^1} \{\text{KL}(\hat{\theta}_m; \{x_i^{(2)}\})\} + \log\left(\frac{1}{\pi_m}\right) \right\} - \mathbb{E}_{\mathcal{D}^1, \mathcal{D}^2} \sum_{m \in \mathcal{M}} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}^{(2)}_{n_2})).
$$

Note that

$$
- \mathbb{E}_{\mathcal{D}^1, \mathcal{D}^2} \sum_{m \in \mathcal{M}} w_m (g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}^{(2)}_{n_2}))
\leq \mathbb{E}_{\mathcal{D}^1, \mathcal{D}^2} \left\{ \max_{m \in \mathcal{M}} \left| g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; \mathcal{D}^{(2)}_{n_2}) \right| \right\}.
$$

Also, observe that

$$
H = \sum_{k=0}^{p} \binom{p}{k} \left( \frac{k}{2ep} \right)^k \leq \sum_{k=0}^{p} \left( \frac{ep}{k} \right)^k \left( \frac{k}{2ep} \right)^k \leq 2.
$$

Then

$$
\log\left(\frac{1}{\pi_m}\right) = \log\left(\frac{2ep}{|m| \lor 1} \right)^{|m|} + \log H
\leq |m| \log\left(1 + \frac{ep}{|m| \lor 1} \right) + \log 2.
$$

The theorem then follows. \qed

Let $m^* \in \mathcal{M}$ be the true sparsity pattern, then it follows from the theorem that the excess risk

$$
\mathbb{E}_{\mathcal{D}^1, \mathcal{D}^2} \{\text{KL}(\hat{\theta}^{LA}; \{x_i^{(2)}\}) - \text{KL}(\hat{\theta}_{m^*}; \{x_i^{(2)}\})\},
$$

is bounded above by the following two terms

$$
|m^*| \log\left(1 + \frac{ep}{|m^*| \lor 1} \right) + \log 2,
$$

(14)
and
\[
\mathbb{E}_{D_1,D_2}\left\{\max_{m\in M} \left| g(\hat{\theta}_m; \{x_i^{(2)}\}) - l(\hat{\theta}_m; D_n^{(2)}) \right| \right\}. \quad (16)
\]

Notice that first term is asymptotically of the same order of \(O(|m^*| \log p)\). The second term measures the maximum bias of the log-likelihood of each individual estimator, in a probabilistic sense.

### 4 Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed logistic aggregation (LA) estimator, competing with the \(\ell_1\)-penalized logistic regression (\(\ell_1\)-LR). The goal is to compare the out-of-sample prediction performance and the variable selection performance in high-dimensional settings.

In the classification problem, we generate responses \(y_i\) according to a logistic model
\[
y_i \sim \text{Bernoulli}\left(\frac{\exp(x_i^T \theta)}{1 + \exp(x_i^T \theta)}\right). \quad (17)
\]

Here, we fix a vector of true coefficients \(\theta\) with only the first \(p_0 = 5\) entries set to be nonzero, by putting \(\theta = (2, 2, 2, 2, 2, 0, \ldots, 0)^T\). The covariates \(x_i\)'s are generated by independent standard Gaussian distributions.

In each simulation, we draw \(n = 400\) or \(800\) data points on \(p = 2000, 5000,\) or \(10000\) features for training the estimators and another \(4000\) independent data points for evaluating the out-of-sample prediction performance.

The following evaluation criteria are considered in the comparisons:

- **SSE**: sum of squared error \(\|\hat{\theta} - \theta\|_2^2\), where \(\hat{\theta}\) is an estimator of \(\theta\) and \(\|\cdot\|_2\) denotes the \(\ell_2\) norm;
- **AUC**: area under the receiver operating characteristic (ROC) curve, as a measure of out-of-sample classification performance;
- **FP**: number of selected features that are actually false positive; the \(i\)th feature is considered to be selected by the estimator \(\hat{\theta}\) if \(|\hat{\theta}_i| > 1/n\);
- **FN**: number of selected features that are actually false negative;
- **Time**: computational time in seconds.

The tuning parameter \(\lambda\) for the \(\ell_1\)-LR method is chosen via cross-validations, as implemented in the R \texttt{glmnet} package (Friedman et al. 2010). For the LA estimator, half of dataset is random chosen for the construction of individual estimators and the other half is then used for aggregation. The reduced set of sparsity patterns used in the Metropolis-Hastings algorithm is also determined by \(\ell_1\)-LR with cross-validations.
We use 50 replications in such evaluations. For each criteria, results are averaged over replications and the standard error is also reported. Table 1 displays the results of simulation studies. We can see that the LA estimator has a much lower sum of squared errors for the estimates of regression coefficients than the $\ell_1$-LR estimator in all scenarios. For the number of false negatives, note that both methods successfully identify all true features. In terms of the number of false positives, however, the $\ell_1$-LR estimator has much worse selection performance, where too many false positives are present. Both methods achieve high values of AUC, although the LA estimator performs slightly better. The two methods do not differ much in the computational cost.

Figure 1 shows a typical behavior of the LA estimator for one particular realization of $n = 400$ and $p = 10000$ with $T_0 = 100$ and $T = 2100$. We can see from the top figure that the sparsity pattern is well recovered among the first 100 coordinates and the estimated values are close to the true value of 2. The right figure displays the evolution of the Metropolis-Hastings algorithm. There exits evidence that the Metropolis-Hastings algorithm converges after only 100 iterations.

![Figure 1: Typical realization for $p = 10000$ and $n = 400$. Left: values of the logistic aggregation (LA) estimator on the first 100 coordinates. Right: Number of selected features in the Metropolis-Hastings algorithm as a function of iterations, where $T_0 = 100$ and $T = 2100$.](image)

5 Analysis of Real Dataset

In this section, we apply the logistic aggregation (LA) method to a published genome-wide case-control data on Parkinson disease with approximately 540 case and control individuals and approximately 408,000 single-nucleotide polymorphisms (SNPs) (Fung et al. 2006).

The goal of genome-wide association study (GWAS) is to examine many common genetic variants in different individuals to see if any variant is associated with a trait, like major
human diseases. In this investigation, we would like to identify a set of causal SNPs associated with Parkinson disease.

Initially we compute the $p$-value for each SNP using a single-locus logistic regression with the binary response coding as 1 := case and 0 := control, that is, testing each SNP one at a time, where an additive model on the genotype is assumed. To make the scale
of the dataset manageable, we excluded those SNPs whose $p$-values (without correction for multiply testing) exceeded 0.001; 401 SNPs remained in the dataset for the further analysis.

Then we apply our method to this smaller dataset. In a random split, half of the data was used in the first stage for constructing estimators, while the other half was used in the second stage for aggregation. For comparisons, we also apply the $\ell_1$ regularization logistic regression to this data and use cross-validations to determine the tuning parameter.

Figure 2 displays the estimated values of regression coefficients for the LA estimator and the evolution of Metropolis-Hastings algorithm where $T_0 = 500$ and $T = 2500$. We can see that the Metropolis-Hastings algorithm converges after 500 iterations.

The LA estimator selected 98 SNPs, while the $\ell_1$ regularization logistic regression selected 212 SNPs. Among them, 93 SNPs were selected by both methods.

![Figure 2](image_url)

**Figure 2:** Analysis results on Parkinson disease data. Left: values of the logistic aggregation (LA) estimator on the remaining features after marginal screenings. Right: Number of selected features in the Metropolis-Hastings algorithm as a function of iterations, where $T_0 = 500$ and $T = 2500$.

### 6 Discussion

In this paper, we proposed an aggregation algorithm for sparse logistic regression and demonstrated that this estimator can give comparable or better results than the $\ell_1$-penalized logistic regression. We show that this method could be a promising toolset in practical applications.

Our aggregation method is based on a sample-splitting procedure: the first subsample is set to construct the estimators and the second subsample is used to determine the component weights and aggregate these estimators. In high-dimensional settings, the Metropolis-Hastings algorithm becomes computationally unsuitable, since it takes the form of a random walk over the hypercube of the $2^p$ all-subsets models. Instead, we use $\ell_1$-penalized logistic regression to approximate the model space and select a candidate set of sparsity patterns.
There are several areas of potential improvements in this method. First, we sacrifice half of sample in either stage, which may become a problem especially when sample size is small. Second, the Metropolis-Hastings algorithm relies on the reduced set of sparsity patterns selected by the $\ell_1$-penalized logistic regression. Other good approximation algorithm for the LA estimator is still needed.

References

[1] Chris Fraley and Daniel Percival (2013). Model-averaged $\ell_1$ regularization using Markov chain Monte Carlo model composition. *Journal of Statistical Computation and Simulation*.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22.

[3] Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez, Dena Hernandez, Angela Britton, J Rapha Gibb, Carl Langefeld, Matt L Stiegert, Jennifer Schymick, Michael S Okun, Ronald J Mandel, Hubert H Fernandez, Kelly D Foote, Ramón L Rodriguez, Elizabeth Peckham, Fabienne Wavrant De Vrieze, Katrina Gwinn-Hardy, John A Hardy, and Andrew Singleton (2006). Genome-wide genotyping in Parkinson’s disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurology*, 5, 911–916.

[4] Joshua T. Goodman (2004). Exponential priors for maximum entropy models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[5] Guillaume Lecué (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4), 1000–1022.

[6] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng (2006). Efficient $\ell_1$ regularized logistic regression. *Proceedings of the 21st National Conference on Artificial Intelligence*.

[7] Gilbert Leung and Andrew R. Barron (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 5, 3396–3410.

[8] Andrew Y. Ng (2004). Feature selection, $\ell_1$ vs. $\ell_2$ regularization, and rotational invariance. *International Conference on Machine Learning*.

[9] Philippe Rigollet and Alexandre B. Tsybakov (2011). Exponential Screening and optimal rates of sparse estimation. *Annals of Statistics*, 39(2), 731–771.

[10] Philippe Rigollet and Alexandre B. Tsybakov (2012). Sparse Estimation by Exponential Weighting. *Statistical Science*, 27(4), 558–575.