Group Orthogonal Matching Pursuit for Logistic Regression

Aurélie C. Lozano, Grzegorz Świrszcz, Naoki Abe
IBM Watson Research Center
Yorktown Heights, NY 10598,USA

Abstract

We consider a matching pursuit approach for variable selection and estimation in logistic regression models. Specifically, we propose Logistic Group Orthogonal Matching Pursuit (Logit-GOMP), which extends the Group-OMP procedure originally proposed for linear regression models, to select groups of variables in logistic regression models, given a predefined grouping structure within the explanatory variables. We theoretically characterize the performance of Logit-GOMP in terms of predictive accuracy, and also provide conditions under which Logit-GOMP is able to identify the correct (groups of) variables. Our results are non-asymptotic in contrast to classical consistency results for logistic regression which only apply in the asymptotic limit where the dimensionality is fixed or is restricted to grow slowly with the sample size. We conduct empirical evaluation on simulated data sets and the real world problem of splice site detection in DNA sequences. The results indicate that Logit-GOMP compares favorably to Logistic Group Lasso both in terms of variable selection and prediction accuracy. We also provide a generic version of our algorithm that applies to the wider class of generalized linear models.

1 Introduction

In many applications of statistics and machine learning, the number of exploratory variables may be very large, while only a small subset may truly be relevant in explaining the response to be modeled. In certain cases, the dimensionality of the predictor space may also exceed the number of examples. Then the only way to avoid overfitting is via some form of "capacity control" over the family of dependencies being explored. Estimation of sparse models that are supported on a small set of input variables is thus highly desirable, with the additional benefit of leading to parsimonious models, which can be used not only for predictive purposes but also to understand the effects (or lack thereof) of the candidate predictors on the response. A particularly pertinent notion in this context is that of group sparsity. In many problems a predefined grouping structure exists within the explanatory variables, and it is natural to incorporate the prior knowledge that the support of the model should be a union over some subset of these variable groups. For instance in gene expression analysis on microarrays data, genes belonging to the same functional cluster may be considered as a group; in business metric analysis, metrics belonging to a common line of business may form a natural group. There are also a number of technical settings in which variable group selection is highly desirable, for instance, when dealing with groups of dummy variables in multifactor ANOVA, or with groups of lagged variables belonging to the same time series in time series analysis.

Several methods have been proposed to address the variable group selection problem, based on minimization of a loss function penalized by a regularization term designed to encourage sparsity at the variable group level. Specifically, a number of variants of the l_1-regularized Lasso algorithm [20] have been proposed for variable group selection problem, and their properties have been extensively studied recently. First, for linear regression, Yuan & Lin [24] proposed the Group Lasso algorithm as an extension of Lasso, which minimizes the squared error penalized by the sum of l_2-norms of the group variable coefficients across groups. Here the use of l_2-norm within the groups and l_1-norm across the groups encourages sparsity at the group level. In addition, Group Lasso has been extended to logistic regression for binary classification, by replacing the squared error by the logistic error [10, 14], and several extensions thereof have been proposed [18].

A class of methods that has recently received considerable attention, as a competitive alternative to Lasso, is the class of Orthogonal Matching Pursuit techniques

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

(OMP) [12]. The basic OMP algorithm originates from the signal-processing community and is popular in the domain of compressed sensing. It is similar to boosting methods [5] in the way it does forward greedy feature selection, except that it performs re-estimation of the model parameters in each iteration, which has been shown to contribute to improved accuracy. For linear models, some strong theoretical performance guarantees and empirical support have been provided for OMP [25] and its variant for variable group selection, Group-OMP [8, 11]. A kernel version of OMP was proposed in [22]. It was shown in [26, 4] that OMP and Lasso exhibit competitive performance characteristics. It is therefore desirable to investigate extending OMP methods beyond linear regression models, and natural to ask whether such extensions may be able to improve upon those based on Lasso, for classification and for other generalized linear models. Such questions have been left open, as matching pursuit techniques have mostly been analyzed in the linear regression setting.

To satisfy the above motivations, we propose *Logistic Group Orthogonal Matching Pursuit* (Logit-GOMP), which generalizes Group-OMP to the logistic regression setting. We theoretically characterize the performance of Logit-GOMP in terms of prediction accuracy. We also provide conditions guaranteeing the correctness of variable group selection. To the best of our knowledge, such "exact recovery" conditions had not been formulated before for models other than linear regression for matching pursuit techniques, and our results provide the first *"non-asymptotic"* conditions on variable selection consistency in logistic regression. Indeed for Lasso-based methods, for instance, the results in [15] are only applicable to the case where the dimensionality of the feature space is small and the sample size goes to infinity ("small p, large n"). In addition, we conduct experiments to compare Logit-GOMP with competing methods, including Logistic Group Lasso, on simulated and real world data sets, and show that it compares favorably against them. We also provide a generic version of the algorithm that applies to the class of generalized linear models, encompassing logistic and linear regression models as special cases.

2 Group Orthogonal Matching Pursuit in Logistic Regression

2.1 Model Formulation

We begin by reviewing the logistic model for binary classification, under a pre-specified grouping structure on the predictors. Consider independent and identically distributed observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with vectors of predictors $\mathbf{x}_i \in \mathbf{R}^p$ and responses $y_i \in \{0, 1\}$. The $n \times p$ predictor matrix can be represented as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, or alternatively as $\mathbf{X} = [\mathbf{f}^1, \ldots, \mathbf{f}^p]$, with feature vectors $\mathbf{f}_j \in \mathbf{R}^n$. Let $\mathbf{y}$ denote the response vector, i.e., $\mathbf{y} = [y_1, \ldots, y_n]^T$.

For any $G \subset \{1, \ldots, p\}$ let $\mathbf{X}_G$ denote the restriction of $\mathbf{X}$ to the set of variables indexed by G, namely $\mathbf{X}_G = \{\mathbf{f}_j, j \in G\}$, where the columns $\mathbf{f}_j$ are arranged in ascending order. Let $\mathbf{x}_i^G$ denote the corresponding restriction on the individual observation $\mathbf{x}_i$. Similarly for any vector $\beta \in \mathbf{R}^p$ of regression coefficients, denote by β_G its restriction to G.

Suppose that a natural grouping structure exists within the variables consisting of J groups $\mathbf{X}_{G_1}, \ldots, \mathbf{X}_{G_J}$, where $G_i \subset \{1, \ldots, p\}$, $G_i \cap G_j = \emptyset$ for $i \neq j$ and $\mathbf{X}_{G_i} \in \mathbf{R}^{n \times p_i}$. Then the logistic regression models the class conditional probability $p_\beta(\mathbf{x}_i) = P_\beta(y = 1|\mathbf{x}_i)$ by

$$\log\left(\frac{p_\beta(\mathbf{x}_i)}{1 - p_\beta(\mathbf{x}_i)}\right) = \eta_\beta(\mathbf{x}_i),$$

where $\eta_\beta(\mathbf{x}_i)$ can be expressed in terms of the variable groups as $\eta_\beta(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^J (\mathbf{x}_i^G)^T \beta_G$, where $\beta_0 \in \mathbf{R}$ is the intercept.

Recall that in regression, the link function is the function specifying the relationship between the class conditional expectation of the response variable and the underlying linear model, namely any function g such that $g(E[y_i|\mathbf{x}_i]) = \eta_\beta(\mathbf{x}_i)$. Note that for the logistic model $E[y_i|\mathbf{x}_i] = p_\beta(\mathbf{x}_i)$. Thus g is such that

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),\ \mu \in (0, 1)$$

and

$$g^{-1}(\eta_\beta(\mathbf{x}_i)) = \frac{\exp(\eta_\beta(\mathbf{x}_i))}{1 + \exp(\eta_\beta(\mathbf{x}_i))}.$$

Then, with a slight abuse of notation, consider $\eta_\beta(\mathbf{X}) = [\eta_\beta(\mathbf{x}_1), \ldots, \eta_\beta(\mathbf{x}_n)]^T$, and $g^{-1}(\eta_\beta(\mathbf{X})) = [g^{-1}(\eta_\beta(\mathbf{x}_1)), \ldots, g^{-1}(\eta_\beta(\mathbf{x}_n))]^T$.

We consider the problem of minimizing the negative log-likelihood, which, under the above model is expressed as

$$L(\eta_\beta) = -\sum_{i=1}^n \left[y_i \eta_\beta(\mathbf{x}_i) - \log(1 + \exp(\eta_\beta(\mathbf{x}_i)))\right].$$

Given $\beta \in \mathbf{R}^p$ let $\text{supp}(\beta) = \{j : \beta_j \neq 0\}$. For any group, or any subset of variables, G, and vector $\mathbf{v} \in \mathbf{R}^n$, denote by $\hat{\beta}_X(G, \mathbf{v})$ the coefficients resulting from applying ordinary logistic regression with non-zero coefficients restricted to G, i.e.,

$$\hat{\beta}_X(G, \mathbf{v}) = \arg\min_{\beta \in \mathbf{R}^d} -\sum_{i=1}^n \left[v_i \eta_\beta(\mathbf{x}_i) - \log\left[1 + e^{\eta_\beta(\mathbf{x}_i)}\right]\right]$$
$$\text{subject to } \text{supp}(\beta) \subseteq \text{G}.$$

2.2 Logit-GOMP

Given the notational set-up of Section 2.1, the Logit-GOMP algorithm we propose is given in Figure 1. It extends the Group-OMP [11] procedure to deal with group

selection in logistic regression. The key step in the algorithm is the greedy selection step $(*)$ in Figure 1. The vector $\mathbf{r}^{(k)} = g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}$ corresponds to the gradient of the logistic loss function evaluated at the examples $(\mathbf{x}_i, y_i)$:

$$\mathbf{r}^{(k)} = \nabla_{\eta_\beta} L(\eta_\beta) = \left(\frac{\partial L(\eta_\beta(\mathbf{x}_1), y_1)}{\partial \eta_\beta(\mathbf{x}_1)}, \ldots, \frac{\partial L(\eta_\beta(\mathbf{x}_n), y_n)}{\partial \eta_\beta(\mathbf{x}_n)} \right)$$

and the Logit-GOMP procedure picks the best group in each iteration, with respect to maximizing the projection onto the steepest descent direction. Here $\mathbf{r}^{(k)}$ can be interpreted as a "pseudo residual" vector. Hence in each iteration Logit-GOMP picks the group maximizing the projection onto the pseudo residuals (similarly to Group-OMP for linear models, where the projection is onto the usual residuals). Logit-GOMP then re-estimates the coefficients $\beta^{(k)}$, through ordinary logistic regression.

Note that to simplify the exposition, we assume that each $\mathbf{X}_{G_j}$ is orthonormalized, i.e. ${\mathbf{X}_{G_j}}^T \mathbf{X}_{G_j} = I_{p_j}$, where I_{p_j} denotes the $p_j \times p_j$ identity matrix and p_j is the number of features in group G_j. However such assumption is not required: if the groups are not orthonormalized, the criterion $(*)$ in Figure 1 should be replaced by

$$j^{(k)} = \arg\max_j \left| (\mathbf{r}^{(k)})^T \mathbf{X}_{G_j} (\mathbf{X}_{G_j}^T \mathbf{X}_{G_j})^{-1} \mathbf{X}_{G_j}^T r^{(k)} \right|.$$

Notice also that one may incorporate an intercept term by setting $\mathbf{X} = [\mathbf{1}, \mathbf{f}_1, \ldots, \mathbf{f}_d]$ and the group structure to be $G_0, G_1, \ldots, G_J$, where G_0 corresponds to the first column of the data matrix $\mathbf{X}$. If one wishes to force intercept inclusion, simply set $\mathcal{G}^{(0)} = G_0$, and $\mathbf{r}^{(0)} = \left(\frac{1}{n}\sum_{i=1}^n y_i\right)\mathbf{1} - \mathbf{y}$.

Extension to Generalized Linear Models: The procedure for logistic regression we just presented can be readily extended to apply to the wider class of generalized linear models (e.g. Poisson and multinomial logistic regression). The generic procedure for generalized lineal models is identical to that of Figure 1, expect that (i) the function g^{-1} in the greedy selection step $(*)$ corresponds to the link function for the model under consideration and (ii) the re-estimation step is performed with respect to the appropriate loss function L (See [7] for description of generalized linear models with their loss and link functions).

Input: The data matrix $\mathbf{X} = [\mathbf{f}_1, \ldots, \mathbf{f}_p] \in \mathbf{R}^{n \times p}$,
with group structure $G_1, \ldots, G_J$, such that
$\mathbf{X}_{G_j}^T \mathbf{X}_{G_j} = I_{p_j}$.
The response vector $\mathbf{y} \in \{0,1\}^n$.
Precision $\epsilon > 0$ for the stopping criterion.
Output: The selected groups $\mathcal{G}^{(k)}$,
the regression coefficients $\beta^{(k)}$.

Initialization: $\mathcal{G}^{(0)} = \emptyset$, $\beta^{(0)} = \mathbf{0}$.
For $k = 1, 2, \ldots$
Let $\mathbf{r}^{(k)} = g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}$.
Let $j^{(k)} = \arg\max_j \left\| \mathbf{X}_{G_j}^T \mathbf{r}^{(k)} \right\|_2$. $(*)$
If $\left(\left\| \mathbf{X}_{G_{j^{(k)}}}^T \mathbf{r}^{(k)} \right\|_2 \leq \epsilon \right)$ **break**
Set $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} \cup G_{j^{(k)}}$. Let $\beta^{(k)} = \hat{\beta}_X(\mathcal{G}^{(k)}, \mathbf{y})$.
End

Figure 1: Method *Logit-GOMP*

3 Theoretical Analysis

3.1 Prediction Accuracy

In this section, we characterize the performance of Logit-OMP in minimizing the negative log-likelihood or, equivalently in minimizing the empirical risk, which for a coefficient vector β is defined as $Q(\eta_\beta) = \frac{1}{n} L(\eta_\beta)$. The proofs of the theorems are provided at the end of this section.

Theorem 1. *Assume that we run k iterations of Logit-GOMP and obtain the regression coefficient vector $\beta^{(k)}$. Then, for any $\epsilon > 0$ and coefficient vector $\bar{\beta}$ with support $\bar{\mathcal{G}}$ satisfying*

$$k \geq \frac{|\bar{\mathcal{G}}| \sum_{G_i \in \bar{\mathcal{G}}} \|\bar{\beta}_{G_i}\|_1^2}{2\epsilon},$$

where $|\bar{\mathcal{G}}|$ denotes the number of groups in $\bar{\mathcal{G}}$, we have $Q(\eta_{\beta^{(k)}}) - Q(\eta_{\bar{\beta}}) \leq \epsilon$.

An exponentially better bound on $1/\epsilon$ with respect to k can be derived under the following strong convexity assumption:

Assumption A1: *The restriction of the empirical risk on any set with support formed by no more than $k + |\bar{\mathcal{G}}|$ groups is* strongly convex *with constant ρ. Namely, for any set F with support consisting of at most $k + |\bar{\mathcal{G}}|$ groups, we have: For all vector u, v with respective supports in F,*

$$Q(\eta_v) - Q(\eta_u) - \langle \nabla_u Q(\eta_u), v - u \rangle \geq \frac{\rho}{2} \|v - u\|^2.$$

We refer the reader to [9] for a study of when the logistic loss exhibits strong convexity properties as in Assumption A1.

Theorem 2. *Assume that we run k iterations of Logit-GOMP and obtain the regression coefficient vector $\beta^{(k)}$. Then, under Assumption A1, for any $\epsilon > 0$ and coefficient vector $\bar{\beta}$ with support $\bar{\mathcal{G}}$ satisfying*

$$k \geq \frac{|\bar{\mathcal{G}}| \max_{G_i \in \bar{\mathcal{G}}} \|\bar{\beta}_{G_i}\|_0}{4\rho} \log\left(\frac{\log(2) - Q(\eta_{\bar{\beta}})}{\epsilon} \right),$$

where $|\bar{\mathcal{G}}|$ denotes the number of groups in $\bar{\mathcal{G}}$ we have $Q(\eta_{\beta^{(k)}}) - Q(\eta_{\bar{\beta}}) \leq \epsilon$.

Proofs of Theorem 1 and Theorem 2. To prove the above theorems, we use arguments similar to [19]. Due to space constraints, we refer to [19] for a few technical lemmas but

offer full details on the parts that are specific to our work. The theorems are a consequence of the following lemma.

Lemma 1. *Let $\mathcal{G}$ and $\bar{\mathcal{G}}$ be two sets of groups such that $\bar{\mathcal{G}}\backslash\mathcal{G} \neq \emptyset$. Denote by $|\bar{\mathcal{G}}\backslash\mathcal{G}|$ the number of groups in $\bar{\mathcal{G}}\backslash\mathcal{G}$. Let $\beta = \arg\min_{v:\mathrm{supp(v)}=\mathcal{G}} Q(\eta_v)$ and similarly $\bar{\beta} = \arg\min_{v:\mathrm{supp(v)}=\bar{\mathcal{G}}} Q(\eta_v)$.*
Assume that $Q(\eta_{\bar{\beta}}) - Q(\eta_\beta) - \langle \nabla_\beta Q(\eta_\beta), \bar{\beta} - \beta\rangle \geq \frac{\rho}{2}\|\bar{\beta} - \beta\|^2$. Let $\beta' = \arg\min_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G},\ \tilde{\beta}:\ \mathrm{supp}(\tilde{\beta})=\mathrm{G_i}} Q(\eta_{\beta+\tilde{\beta}})$. Then we have

$$Q(\eta_\beta) - Q(\eta_{\beta'}) \geq \frac{\left(Q(\eta_\beta) - Q(\eta_{\bar{\beta}}) + \frac{\rho}{2}\|\beta - \bar{\beta}\|_2^2\right)^2}{\frac{1}{2}|\bar{\mathcal{G}}\backslash\mathcal{G}| \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}} \|\bar{\beta}_{G_i}\|_1^2}.$$

Proof: We slightly abuse notation and let Q also denote the function such that $Q(\mathbf{u}) = -\frac{1}{n}[y_i u_i - \log(1 + \exp(u_i))]$, for a vector $\mathbf{u} \in \mathbf{R}^n$. We this abuse we have

$$\begin{aligned} Q(\eta_{\beta'}) &= \min_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G},\ \tilde{\beta} \in \mathbf{R}^{p_i}} Q(\eta_\beta(\mathbf{X}) + \mathbf{X}_{G_i}\tilde{\beta}) \\ &\leq \inf_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}, \alpha \in \mathbf{R}} Q(\eta_\beta(\mathbf{X}) + \alpha \mathbf{X}_{G_i}(\bar{\beta} - \beta)_{G_i}) \\ &\leq \frac{1}{|\bar{\mathcal{G}}\backslash\mathcal{G}|} \inf_\alpha \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}} Q(\eta_\beta(\mathbf{X}) + \alpha \mathbf{X}_{G_i}(\bar{\beta} - \beta)_{G_i}) \end{aligned}$$

The smoothness of the logistic loss implies that for any pair of coefficient vectors $\tilde{\beta}$ and $\hat{\beta}$, we have $Q(\eta_{\tilde{\beta}}(\mathbf{X}) + \eta_{\hat{\beta}}(\mathbf{X})) \leq Q(\eta_{\tilde{\beta}}) + \langle \nabla_{\tilde{\beta}} Q(\eta_{\tilde{\beta}}), \hat{\beta}\rangle + \frac{1}{8}\|\hat{\beta}\|_1^2$. (See [19], Appendix B Lemma B.1).

Thus we have $Q(\eta_\beta(\mathbf{X}) + \alpha \mathbf{X}_{G_i}(\bar{\beta} - \beta)_{G_i}) \leq Q(\eta_\beta) + \alpha\langle(\nabla_\beta Q(\eta_\beta))_{G_i}, (\bar{\beta} - \beta)_{G_i}\rangle + \frac{1}{8}\alpha^2 \|(\bar{\beta} - \beta)_{G_i}\|_1^2$. Thus, noting that $\beta_{G_i} = \mathbf{0}$ for $G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}$, we have

$$\begin{aligned} Q(\eta_{\beta'}) &\leq Q(\eta_\beta) + \tfrac{1}{|\bar{\mathcal{G}}\backslash\mathcal{G}|} \inf_\alpha \{\alpha \langle \nabla_\beta Q(\eta_\beta), \bar{\beta} - \beta\rangle + \\ &\quad \tfrac{1}{8}\alpha^2 \textstyle\sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}} \|\bar{\beta}_{G_i}\|_1^2\} \\ &\leq Q(\eta_\beta) + \tfrac{1}{|\bar{\mathcal{G}}\backslash\mathcal{G}|} \inf_\alpha \{\alpha[Q(\eta_{\bar{\beta}}) - Q(\eta_\beta) - \\ &\quad \tfrac{\rho}{2}\|\beta - \bar{\beta}\|_2^2] + \tfrac{1}{8}\alpha^2 \textstyle\sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}} \|\bar{\beta}_{G_i}\|_1^2\}. \end{aligned}$$

Optimizing for α, we obtain the lemma. □

Let $\epsilon_k = Q(\eta_{\beta^{(k)}}) - Q(\eta_{\bar{\beta}})$. Theorem 1 follows by noting that Lemma 1 with $\rho = 0$ (simple convexity assumption which holds by convexity of the logistic loss) leads to $\epsilon_{k+1} \leq \epsilon_k - \frac{\epsilon_k^2}{\frac{1}{2}|\bar{\mathcal{G}}\backslash\mathcal{G}|\| \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}} \|\bar{\beta}_{G_i}\|_1^2}$, and combining this with the fact that if there exists $r > 0$ such that for all t, $\epsilon_{t+1} \leq \epsilon_t - r\epsilon_t^2$, then for $\epsilon > 0$ and $k \geq \lceil \frac{1}{r\epsilon} \rceil$, there holds $\epsilon_k \leq \epsilon$ (see [19] Appendix B Lemma B.2.).

Theorem 2 follows by noting that Lemma 1 with $\mathcal{G} = \mathcal{G}^{(k)}$ implies

$$\begin{aligned} \epsilon_k - \epsilon_{k+1} &\geq \frac{(\epsilon_k + \frac{\rho}{2}\|\beta - \bar{\beta}\|_2^2)^2}{\frac{1}{2}|\bar{\mathcal{G}}\backslash\mathcal{G}^{(k)}| \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}^{(k)}} \|\bar{\beta}_{G_i}\|_1^2} \\ &\geq \frac{4\epsilon_k \frac{\rho}{2} \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}^{(k)}} \|\bar{\beta}_{G_i}\|_2^2}{\frac{1}{2}|\bar{\mathcal{G}}\backslash\mathcal{G}^{(k)}| \sum_{G_i \in \bar{\mathcal{G}}\backslash\mathcal{G}^{(k)}} \|\bar{\beta}_{G_i}\|_1^2} \\ &\geq \frac{4\rho\epsilon_k}{|\bar{\mathcal{G}}| \max_{G_i \in \bar{\mathcal{G}}} \|\bar{\beta}_{G_i}\|_0}. \end{aligned}$$

By recursion we obtain $\epsilon_k \leq \epsilon_0(1 - \frac{4\rho}{|\bar{\mathcal{G}}| \max_{G_i \in \bar{\mathcal{G}}} \|\bar{\beta}_{G_i}\|_0})^k$. The theorem follows by using $1 - x \leq \exp(-x)$. □

3.2 Variable Selection Accuracy

In this section, we identify conditions which guarantee that the Logit-GOMP algorithm does not select any wrong groups. The case of Logit-GOMP differs significantly from the "regular" Group-OMP due to the strongly non-linear characteristics of the operators involved, namely, due to the effects introduced by the link function g^{-1}. As a result, one should expect a non-linear set of conditions to be necessary to guarantee that the algorithm selects the correct feature groups and this is the case indeed. Nevertheless the set of conditions presented here have an elegant geometric character.

Let $\mathcal{G}_{\mathrm{good}}$ denote the set of all the groups included in the true model, and let $\mathcal{G}_{\mathrm{bad}}$ denote the set of all the groups which are not included. Similarly denote by $\mathrm{g}_{\mathrm{good}}$ the set of feature indices for the features in the true model, and by $\mathrm{g}_{\mathrm{bad}}$ the set of feature indices for the features that are not in the true model. In this notation $\mathrm{supp}(\bar{\beta}) \subseteq \mathrm{g}_{\mathrm{good}}$, where $\bar{\beta}$ denote the true model coefficients.
Define $\Theta(\mathbf{y})$ as

$$\Theta(\mathbf{y}) = \left\{\tilde{\mathbf{y}} \ :\ \|\tilde{\mathbf{y}}\| = 1 \wedge \forall_{i \in \{1,\ldots,n\}} \mathrm{sign}(\tilde{y}_i) = \mathrm{sign}(y_i - \tfrac{1}{2})\right\},$$

that is the intersection of an $(n-1)$-dimensional unit sphere in $\mathbf{R}^n$ with the n-dimensional orthant containing the vector $[\mathrm{sign}(y_1 - \frac{1}{2}), \mathrm{sign}(y_2 - \frac{1}{2}), \ldots, \mathrm{sign}(y_n - \frac{1}{2})]^T$.

Analogously, for a set of vectors $\mathcal{G}$ let

$$\Lambda(\mathcal{G}) = \{\tilde{\mathbf{y}} \ :\ \|\tilde{\mathbf{y}}\| = 1 \wedge \forall_{i \in \{1,\ldots,n\}} \mathrm{sign}(\tilde{y}_i) = \mathrm{sign}(v_i), v \in \mathrm{span}\mathcal{G}\},$$

An example for $\Theta(\mathbf{y})$ is depicted in Figure 3.2.

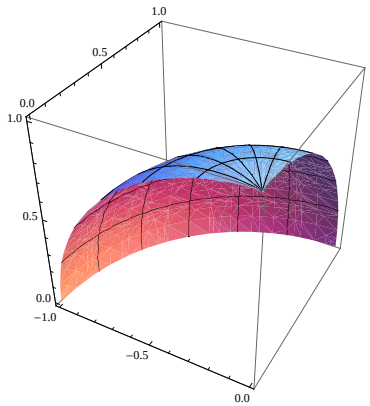

Figure 2: The set $\Theta(\mathbf{y})$ in 3 dimensions for $\mathbf{y} = [1, 0, 1]^T$

Theorem 3. *Suppose that the design matrix $\mathbf{X}$ and the response vector $\mathbf{y}$ satisfy the following condition:*

$$\sup_{\substack{G_j \in \mathcal{G}_{\mathrm{bad}} \\ \tilde{\mathbf{y}} \in \Theta(\mathbf{y})}} \|\mathbf{X}_{G_j}^T \tilde{\mathbf{y}}\|_2 < \sup_{\substack{G_i \in \mathcal{G}_{\mathrm{good}} \\ \tilde{\mathbf{y}} \in \Theta(\mathbf{y})}} \|\mathbf{X}_{G_i}^T \tilde{\mathbf{y}}\|_2. \qquad (1)$$

Then Logit-GOMP will never make a mistake.

Proof. The proof is by induction on the number of iterations k. Obviously, the theorem statement is true for $k = 0$ (no groups are selected, so all selected groups are good). Now assume that the theorem statement holds for $k-1$, that is, only good groups have been selected in steps $1, \ldots, k-1$. We have: $y_i \in \{0, 1\}$ and $g^{-1}(\mathbf{x_i}\beta^{(k-1)}) \in (0, 1)$, so it holds that

$$\frac{1}{\|g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}\|_2}\left[g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}\right] \in \Theta(\mathbf{y}).$$

Therefore if $j \in \mathrm{g_{bad}}$ then

$$\begin{aligned} &\left\|\mathbf{X}_{G_j}^T\left[g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}\right]\right\|_2 \\ < &\sup_{i \in \mathrm{g_{good}}}\left\|\mathbf{X}_{G_i}^T\left[g^{-1}(\mathbf{X}\beta^{(k-1)}) - \mathbf{y}\right]\right\|_2 \end{aligned}$$

and a good group will be chosen in the step k as well. □

Corollary 1. *Suppose that the design matrix* $\mathbf{X}$ *satisfies the following condition:*

$$\sup_{\substack{G_j \in \mathcal{G}_{\mathrm{bad}} \\ \tilde{\mathbf{y}} \in \Lambda(\bigcup \mathcal{G}_{\mathrm{good}})}} \|\mathbf{X}_{G_j}^T \tilde{\mathbf{y}}\|_2 < \sup_{\substack{G_i \in \mathcal{G}_{\mathrm{good}} \\ \tilde{\mathbf{y}} \in \Lambda(\bigcup \mathcal{G}_{\mathrm{good}})}} \|\mathbf{X}_{G_i}^T \tilde{\mathbf{y}}\|_2. \quad (2)$$

Then Logit-GOMP will never make a mistake.

Proof. It is sufficient to show that $\Theta(y) \subseteq \Lambda(\bigcup \mathcal{G}_{\mathrm{good}})$. By definition of good groups there exists $\bar{\beta}$ such that $y_i = 0$ if and only if $g^{-1}(\eta_{\bar{\beta}}(x_i)) < \frac{1}{2}$ and $y_i = 1$ if and only if $g^{-1}(\eta_{\bar{\beta}}(x_i)) > \frac{1}{2}$, where $\eta_{\bar{\beta}}(x) = \sum_{G \in \mathcal{G}_{\mathrm{good}}} x_G \bar{\beta}_G$. But $g^{-1}(x) > \frac{1}{2}$ if and only if $x > 0$ and $g^{-1}(x) < \frac{1}{2}$ if and only if $x < 0$ and the corollary follows. □

Intuitively, $\Theta(\mathbf{y})$ and $\Lambda(\bigcup \mathcal{G}_{\mathrm{good}})$ capture the sub-space in which the (pseudo) residual vectors reside, given that the algorithm has not selected a wrong group up to that stage. For linear regression, the equivalent of $\Theta(\mathbf{y})$ happened to be the span of the "good" feature vectors, but here (for the non-linear case) they no longer coincide. Hence, the need arises to geometrically characterize a region in which the residuals will fall into, as long as the algorithm does not make any mistake, and make sure that the maximum projection onto this region is achieved by a "good" group.

We remark that our results can be seen as the counterpart of Theorem 3.1 in [21] for OMP, and Corollary 1 in [11] for Group-OMP. Note that the conditions of (1) and (2) are similar in character to the (linear) condition for the Group-OMP (see definition and condition on quantity $\mu_X(\mathcal{G}_{\mathrm{good}})$ in Corollary 1 of [11]). Sets Θ and Λ are nonlinear objects (as opposed to "regular" norms in cases of OMP and Group-OMP), which reflects the nature of the logistic setting. To the best of our knowledge, such "exact recovery" conditions had not been formulated before for matching pursuit techniques with models others than linear regression. An important direction for future research is a probabilistic analysis to bound the probability of observing a random sample for which the conditions are violated.

4 Experiments

4.1 Simulation Results

We empirically evaluate the performance of the proposed Logit-GOMP method against a number of comparison methods. The goal of Experiments 1 and 2 is to compare Logit-GOMP, Logit-OMP, Logistic Group Lasso, Logistic Lasso and Ordinary Logistic Regression (denoted by OLR). Logistic Group Lasso is included as a representative method of variable group selection, while comparison with Logit-OMP and Logistic Lasso will test the effect of "grouping". Note that by Logit-OMP we denote the special case of Logit-GOMP with groups of size one. In Experiment 3, we focus more closely on the grouped methods, namely Logit-GOMP and Logistic Group Lasso, by comparing their performance on models involving two-way interactions and a much larger parameter space. We test how their performance changes when we vary the correlation between the predictors and the Bayes risk (in other words the difficulty of the classification task).

In each experiment, we compare the performance of the competing methods in terms of the accuracy of variable selection, variable *group* selection and prediction accuracy. As measure of variable (group) selection accuracy we use the F_1 measure, which is defined as $F_1 = \frac{2PR}{P+R}$, where P denotes the precision and R denotes the recall. For computing variable group F_1 for a variable selection method, we consider a group to be selected if *any* of the variables in the group is selected. The measure of F_1 for the variable *group* selection accuracy is analogously defined, where precision and recall are measured with respect to the variable groups, instead of individual variables. As measure of prediction accuracy, we use the test set negative log-likelihood.

Recall that Logistic Group Lasso solves $\mathrm{argmin}_\beta -\sum_{i=1}^n \Big[y_i \eta_\beta(\mathbf{x}_i) - \log(1 + e^{\eta_\beta(\mathbf{x}_i)})\Big] + \lambda \sum_{j=1}^J \|\beta_{G_j}\|_2$, and Logistic Lasso solves the same problem under the special case where groups are individual features. So the tuning parameter for Logistic Lasso and Logistic Group Lasso is the penalty parameter λ. For Logit-GOMP and Logit-OMP, rather than parameterizing the models according to precision ϵ, we use the iteration number as tuning parameter (i.e. a stopping point). Then for all methods we consider the "holdout validated estimate", which is obtained by selecting the tuning parameter that minimizes the negative log-likelihood on on a validation set.

We now describe the experimental setup. Recall that the empirical Bayes risk for a model is

$$r_b = \frac{1}{n}\sum_{i=1}^n \min\{p_\beta(\mathbf{x}_i), 1 - p_\beta(\mathbf{x}_i)\},$$

where $p_\beta(\mathbf{x}_i)$ denote the class conditional probability and n is large. For each observation $\mathbf{x}_i$, the response y_i is simulated according to a Bernoulli distribution with probability $p_\beta(\mathbf{x}_i)$, where $p_\beta(\mathbf{x}_i)$ is defined in each experiment. For each experiment, we ran 100 runs, each with a training set

of size $n_{\text{train}} = 500$, and validation set (for picking the tuning parameter for each method) of size $n_{\text{val}} = 500$, and a test set of size $n_{\text{test}} = 500$.

Experiment 1: categorical variables. We use an additive model with categorical variables obtained by adapting model I in [24] to the logistic regression setting. Consider variables $z^1, \ldots, z^{15}$, where $z^j \sim \mathcal{N}(0,1)(j = 1, \ldots, 15)$ and $\text{cov}(z^j, z^k) = 0.5^{|j-k|}$. Let $w^1, \ldots, w^{15}$ be such that

$$w^j = \begin{cases} 0 & \text{if } z^j < \Phi^{-1}(1/3) \\ 1 & \text{if } z^j > \Phi^{-1}(2/3) \\ 2 & \text{if } \Phi^{-1}(1/3) \leq z^j \leq \Phi^{-1}(2/3) \end{cases},$$

where Φ^{-1} is the quantile function for the normal distribution. The true underlying model is

$$\eta_\beta(\mathbf{w}) = 1.8I(w^1 = 1) - 1.2I(w^1 = 0) + I(w^3 = 1) + 0.5I(w^3 = 0) + I(w^5 = 1) + I(w^5 = 0),$$

where I denote the indicator function. Then we reparameterize the model and let $(x^{2(j-1)+1}, x^{2j}) = \big(I(w^j = 1), I(w^j = 0)\big)$, which are the variables used as explanatory variables by the estimation method, with the following variable groups: $G_j = \{2j-1, 2j\}(j = 1, \ldots, 15)$. The empirical Bayes risk is $r_b = 0.23$.

Experiment 2: continuous variables with polynomial expansion. We use an additive model with continuous variables obtained by modifying model III in [24] to the logistic regression setting. In this model, the groups correspond to the expansion of each variable into a third-order polynomial. . Consider variables $z^1, \ldots, z^{17}$, with z^j i.i.d. $\sim \mathcal{N}(0,1)$ $(j = 1, \ldots, 17)$. Let $w^1, \ldots, w^{16}$ be defined as $w^j = (z^i + z^{17})/\sqrt{2}$. The true underlying model is $\eta_\beta(\mathbf{w}) = (w^3)^3 + (w^3)^2 + w^3 + \frac{1}{3}(w^6)^3 - (w^6)^2 + \frac{2}{3}w^6$. Then let the explanatory variables be

$$(x^{3(j-1)+1}, x^{3(j-1)+2}, x^{3j}) = \big((w^j)^3, (w^j)^2, w^j\big)$$

with the variable groups $G_j = \{3(j-1)+1, 3(j-1)+2, 3j\}(j = 1, \ldots, 16)$. The Bayes risk for this model is $r_b = 0.20$.

Experiment 3: Categorical variables with two-way interaction.

We use the same simulation scheme as that of [14]. Consider variables $z^1, \ldots, z^9$, where $z^i \sim \mathcal{N}(0,1)$ and $\text{cov}(z^i, z^j) = \rho^{|i-j|}$, where ρ will be specified. Let $w^1, \ldots, w^9$ be such that

$$w^j = \begin{cases} 0 & \text{if } z^j < \Phi^{-1}(1/4) \\ 1 & \text{if } \Phi^{-1}(1/4) \leq z^j < \Phi^{-1}(1/2) \\ 2 & \text{if } \Phi^{-1}(1/2) \leq z^j < \Phi^{-1}(3/4) \\ 3 & \text{if } \Phi^{-1}(3/4) \leq z^j \end{cases},$$

where Φ^{-1} is the quantile function for the normal distribution. The encoding scheme for such categorical variables is dummy encoding with sum constraint [17]. Recall that a categorical predictor taking k possible values has $k-1$ degrees of freedom. The vector β_g corresponding the encoding of a predictor with df_g degrees of freedom is setup as follows. Coefficients $\tilde{\beta}_{g,1}, \ldots, \tilde{\beta}_{g,df+1}$ are sampled from the distribution $\mathcal{N}(0,1)$. These are then transformed as follows.

$\beta_{g,j} = \tilde{\beta}_{g,j} - \frac{1}{df_g+1}\sum_{k=1}^{df_g+1} \tilde{\beta}_{g,k}$,for $j \in \{1, \ldots, df_g\}$.

The intercept β_0 is set to 0. The whole vector β is subsequently rescaled according to a desired value of the empirical Bayes risk. For each model below and values of ρ and r_b, the true model coefficient vector β is then held fixed and data is simulated accordingly for each simulation run. We consider the following underlying models:

Model A The true model is made from the main effects and two-way interaction between the first two factors x^1 and x^2. Thus it involves $J = 4$ terms (Intercept, x^1, x^2, x^1x^2) implying a dummy encoding of $p = 16$ parameters.

Model B The true model is made from the main effects and two-way interaction between the first five factors $x^1, \ldots, x^5$. Thus it involves $J = 16$ terms (Intercept, x^1, x^2, x^3, x^4, x^5, $x^1x^2, x^1x^3, x^1x^4, x^1x^5$, x^2x^3, x^2x^4, x^2x^5, x^3x^4, x^3x^5, x^4x^5) implying a dummy encoding of $p = 106$ parameters.

The candidate models are those with the main effects and two-way interactions for all the 9 variables, which involves $J = 46$ terms or $p = 352$ parameters. For each model, we study $r_b \in \{0.15, 0.25\}$ and $\rho \in \{0.20, 0.50\}$.

The results of Experiments 1 and 2 are presented in Table 1. The results of Experiments 3, Model A are presented in Table 2, and those for Model B in Table 3. We note that F_1 (Var) and F_1 (Group) are identical for the grouped methods for Experiments 1, 2 since in these the groups have equal size. Overall, Logit-GOMP performs consistently better than all the comparison methods, with respect to all measures considered. In particular, Logit-GOMP does better than Logit-OMP not only for variable group selection, but also for variable selection and predictive accuracy. Consistently to what was noted in [14], we observe that Logistic Group Lasso has a tendency to over-select (groups of) variables, leading to poorer F_1 measure, while Logit-GOMP is more parsimonious and accurate. This becomes more pronounced when the true model gets sparser (e.g. Experiment 3, Model B vs. Experiment 3, Model A). Also Logit-GOMP is at least competitive to Logistic Group Lasso with respect to negative log-likelihood minimization. We remark that the negative log-likelihood values we got for Logistic Group Lasso in Experiments 3 are consistent with what have been reported in the literature [14].

4.2 Experiments on Splice Site Detection

In this section, we compare Logit-GOMP and Logistic Group Lasso using the *MEMset donor* data set [23] on splice site detection (available at http://genes.mit.edu/burgelab/maxent/ssdata/).

Splice sites detection is a critical prerequisite to gene iden-

$\mathbf{F_1}$ (Var)	Exp 1	Exp 2
Ordinary Logistic Regression	0.333 ± 0	0.222 ± 0
Logistic Lasso	0.547 ± 0.032	0.371 ± 0.011
Logit-OMP	0.851 ± 0.027	0.629 ± 0.0251
Logistic Group Lasso	0.494 ± 0.035	0.312 ± 0.0181
Logit-GOMP	$\mathbf{0.896 \pm 0.037}$	$\mathbf{0.990 \pm 0.010}$

$\mathbf{F_1}$ (Group)	Exp 1	Exp 2
Ordinary Logistic Regression	0.333 ± 0	0.222 ± 0
Logistic Lasso	0.463 ± 0.033	0.311 ± 0.019
Logit-OMP	0.845 ± 0.038	0.873 ± 0.031
Logistic Group Lasso	0.494 ± 0.035	0.312 ± 0.0181
Logit-GOMP	$\mathbf{0.896 \pm 0.037}$	$\mathbf{0.990 \pm 0.010}$

Negative log-likelihood	Exp 1	Exp 2
Ordinary Logistic Regression	248.38 ± 2.42	237.23 ± 4.64
Logistic Lasso	237.17 ± 2.09	207.87 ± 2.86
Logit-OMP	237.08 ± 2.45	207.26 ± 3.44
Logistic Group Lasso	236.25 ± 2.08	207.70 ± 2.86
Logit-GOMP	$\mathbf{236.06 \pm 2.40}$	$\mathbf{196.73 \pm 2.96}$

Table 1: Results of Experiment 1 and Experiment 2: Average F_1 score at the variable level and group level (higher value is better), and negative log-likelihood (summed over test set; lower value is better) for the models output by Ordinary Logistic Regression, Logistic Lasso, Logit-OMP, Logistic Group Lasso, and Logit-GOMP.

$\mathbf{F_1}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	0.247 ± 0.078	0.318 ± 0.033	0.76 ± 0.019	0.307 ± 0.021
Logit-GOMP	$\mathbf{0.878 \pm 0.011}$	$\mathbf{0.890 \pm 0.021}$	$\mathbf{0.880 \pm 0.014}$	$\mathbf{0.890 \pm 0.022}$

$\mathbf{F_1}$ (Group)	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	0.355 ± 0.010	0.406 ± 0.036	0.309 ± 0.0233	0.383 ± 0.0222
Logit-GOMP	$\mathbf{0.793 \pm 0.023}$	$\mathbf{0.800 \pm 0.006}$	$\mathbf{0.793 \pm 0.026}$	$\mathbf{0.938 \pm 0.013}$

Negative log-likelihood	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	258.76 ± 9.42	184.58 ± 2.12	274.81 ± 2.18	$\mathbf{187.58 \pm 2.45}$
Logit-GOMP	$\mathbf{249.75 \pm 9.49}$	$\mathbf{181.92 \pm 5.46}$	$\mathbf{272.91 \pm 2.48}$	193.246 ± 3.53

Table 2: Results of Experiment 3 Model A: Average F_1 score at the variable level and group level (higher value is better), and negative log-likelihood (summed over the test set; lower value is better) for the models output by Logistic Group Lasso and Logit-GOMP.

$\mathbf{F_1}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	$\mathbf{0.571 \pm 0.007}$	0.553 ± 0.008	0.577 ± 0.010	0.546 ± 0.006
Logit-GOMP	0.534 ± 0.033	$\mathbf{0.633 \pm 0.021}$	$\mathbf{0.600 \pm 0.030}$	$\mathbf{0.686 \pm 0.024}$

$\mathbf{F_1}$ (Group)	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	$\mathbf{0.598 \pm 0.007}$	0.5856 ± 0.008	0.615 ± 0.010	0.576 ± 0.006
Logit-GOMP	0.496 ± 0.014	$\mathbf{0.5934 \pm 0.020}$	$\mathbf{0.695 \pm 0.026}$	$\mathbf{0.578 \pm 0.021}$

Negative log-likelihood	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.5}, \mathbf{r_b} = \mathbf{0.15}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.25}$	$\rho = \mathbf{0.20}, \mathbf{r_b} = \mathbf{0.15}$
Logistic Group Lasso	294.78 ± 1.76	$\mathbf{235.16 \pm 4.08}$	308.70 ± 2.07	234.73 ± 2.45
Logit-GOMP	$\mathbf{293.06 \pm 2.40}$	236.92 ± 6.14	$\mathbf{307.48 \pm 3.02}$	$\mathbf{234.20 \pm 2.31}$

Table 3: Results of Experiment 3 Model B: Average F_1 score at the variable level and group level (higher value is better), and negative log-likelihood for the models (summed over the test set; lower value is better) output by Logistic Group Lasso and Logit-GOMP.

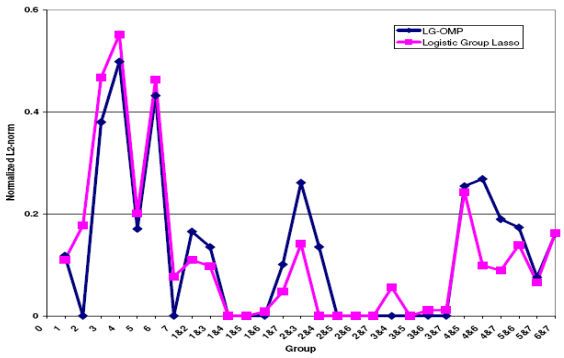

Figure 3: Normalized l_2-norms $\|\beta_{G_j}\|/\|\beta\|$, $j = 1, \ldots, J$ for the estimated models produced by Logit-GOMP and Logistic Group Lasso for the splice site detection problem.

tification in genomic DNA. These sites are the regions between exons (coding) and introns (non-coding) DNA segments. The 5' boundary of an intron, also called donor splice site starts with the dinucleotide 'GT', and the 3' boundary of an exon, also called acceptor splice site contains the dinucleotide 'AG' (see [3] for more details). Several computational methods have been developed for detecting splice sites [2, 23, 16].

The MEMset donor data set consists of "real splice site" and "false splice site" instances. A "real splice site" consists of the last three bases of the exon and the first six bases of the intron, and hence, contains 'GT' at positions 4 and 5. A "fast splice site" also contains those letters at positions 4 and 5 but is not an actual splice site. After removal of the common letters 'GT' at positions 4 and 5, the predictor variables are thus sequences of 7 bases, where each base takes value in $\{A, C, G, T\}$ while the target is binary: $Y = 1$ for "false splice sites" and $Y = 0$ for true ones.

We follow the experimental setup of [14]. Namely, we build a balanced training set with 5610 positive and 5610 negative instances, using the original training set, as well as an unbalanced validation set with 2850 positive and 59804 negative instances (with same class ratio as that in the test set). The instances are randomly sampled without replacement, and hence training and evaluation sets do not intersect. We use the original test set as is. We perform intercept correction to account for the difference in class-ratio between training and validation sets as:

$$\beta_0^{\text{adj}} = \beta_0 + \log\left(\frac{\text{Number of true sites in validation set}}{\text{Number of false sites in validation set}}\right).$$

The stopping point of Logit-GOMP and the penalty parameter for Logistic Group Lasso are chosen so as to minimize the negative log-likelihood on the validation set using estimated models with corrected intercept.

We use as candidate model a logistic regression model with all main effects and two way interactions (it was noted in [14] that considering up to three way interactions gives similar results). We thus consider $J = 28$ groups, and $p = 211$ variables. As measure of performance, we used the "maximum coefficient" $\rho_{\max}$ defined as: $\rho_{\max} = \max_{\alpha \in (0,1)}\{\sum_{i=1}^{n_{\text{test}}}(y_i(I(p_\beta(\mathbf{x}_i) > \alpha)))\}$.

Method	Maximal correlation coefficient
Logistic Group Lasso	0.6588
Logit-GOMP	**0.6591**

Table 4: Maximal correlation coefficients for Logistic Group OMP and Logistic Group Lasso on the splice site detection data.

The results are provided in Table 4. Figure 4 depicts the normalized l_2-norm of the regression coefficients for the groups, as estimated by the two competing methods. Notice from the figure that Logit-GOMP and Logistic Group Lasso grant similar importance to a majority of groups and that the model output by Logit-GOMP is sparser.

In view of these results, we conclude that on this high dimensional problem, Logit-GOMP is at least competitive with Logistic Group Lasso, as well as with the results of [23], where $\rho_{\max} = 0.6589$ was achieved.

5 Concluding Remarks and Perspectives

By proposing a generic matching pursuit method for variable group selection in generalized linear models, and demonstrating its competitive performance for logistic regression, we are opening up a research agenda on the consideration and analysis of matching pursuit techniques as competitive alternatives to l_1 penalized regression methods, for a variety of models beyond linear regression.

Relevant directions for future research include extending our theoretical analysis to the stochastic setting, and proving performance guarantees for other instances of our generic algorithm under generalized linear models.

In view of the recent results of [4, 26], the discrepancies between the theoretical guarantees for Lasso and OMP obtained so far are very narrow and subtle. A very pertinent direction for future investigation is thus to theoretically and experimentally characterize precise settings where one type of method would perform better than the other.

We also plan to apply the proposed method in a variety of problems in which variable group selection involving both continuous and categorical variables is important, such as modeling from time series data with mixed data types.

Acknowledgements

Research supported through participation in the Measurement Science for Cloud Computing, sponsored by the National Institute of Standards and Technology (NIST) under Agreement Number 60NANB10D003.

References

[1] BACH, F.R., *Consistency of the Group Lasso and Multiple Kernel Learning*, J. Mach. Learn. Res., **9**, 1179-1225, 2008.

[2] BURGE, C. AND KARLIN, S., *Prediction of complete gene structures in human genomic DNA.*, J. Molec. Biol., 268, 7894, 1997.

[3] BURGE, C., *Modeling dependencies in pre-mrna splicing signals.* In Computational Methods in Molecular Biology (eds S. Salzberg, D. Searls and S. Kasif), ch. 8, pp. 129164. New York: Elsevier Science. 1998.

[4] FLETCHER A., RANGAN S., *Orthogonal Matching Pursuit From Noisy Random Measurements: A New Analysis,* in proc. NIPS'09, 2009.

[5] FREUND. Y., SCHAPIRE, R. E., *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and system Sciences, Vol. 55(1), 119-139, 1997.

[6] FRIEDMAN J., HASTIE T., TIBSHIRANI R., *Additive Logistic Regression: a Statistical View of Boosting*. Technical report, Department of Statistics, Stanford University, 1998.

[7] HASTIE T., TIBSHIRANI R., FRIEDMAN J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Springer, 2003.

[8] HUANG J., ZHANG T., METAXAS D., *Learning with Structured Sparsity,* in ICML'09, 2009.

[9] KAKADE S., SHAMIR O., SRIDHARAN K., TEWARI S., *Learning Exponential Families in High Dimensions: Strong Convexity and Sparsity,* AISTATS 2010.

[10] KIM Y., KIM J., KIM Y., *Blockwise sparse regression,* Statistica Sinica, 16, 375–390, 2006.

[11] LOZANO A.C., SWIRSZCZ G., ABE N., *Grouped Orthogonal Matching Pursuit for Variable Selection and Prediction*,in proc. NIPS'09, 2009.

[12] MALLAT S., ZHANG Z., *Matching pursuits with time-frequency dictionaries*, IEEE Transactions on Signal Processing, **41**, 3397-3415, 1993.

[13] MASON L., BAXTER J., BARTLETT P., FREAN M., *Boosting Algorithms as Gradient Descent,* in Neural Information Processing Systems, Vol. 12, pp. 512518, 2000.

[14] MEIER L., VAN DE GEER, S., BÜHLMANN P., *The group lasso for logistic regression*, J. Royal Statistical Society: Series B, **70(1)**, 53-71, 2008.

[15] ROCHA G., WANG X., YU B., *Asymptotic distribution and sparsistency for l_1 penalized parametric M-estimators, with applications to linear SVM and logistic regression*, Technical report, 2009

[16] M. PERTEA , X. LIN , S. L. SALZBERG, *GeneSplicer : a new computational method for splice site prediction* Nucleic Acids Res, **29(5)**, 1185-90, 2001

[17] QUINN, GERALD PETER, AND KEOUGH, MICHAEL, *J. Experimental design and data analysis for biologists*, Cambridge University Press, 2002

[18] ROTH V., FISCHER B. *The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms.* In Proceedings of the 25th ICML conference (ICML'08), **307**, 848-855, 2008.

[19] SHALEV-SHWARTZ S., SREBRO N., ZHANG T., *Trading Accuracy for Sparsity*, Technical Report TTIC-TR-2009-3, May 2009.

[20] TIBSHIRANI, R., *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B., **58**(1), 267-288, 1996.

[21] TROPP J.A., *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Info. Theory, **50**(10), 2231-2242, 2004.

[22] VINCENT P., BENGIO Y., *Kernel matching pursuit,* Machine Learning, **48**. 165–187, 2002.

[23] YEO, G. W. AND BURGE, C. B., *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J. Computnl Biol., 11, 475494, 2004.

[24] YUAN, M., LIN, Y., *Model selection and estimation in regression with grouped variables*, J. R. Statist. Soc. B, **68**, 4967, 2006.

[25] ZHANG, T., *On the consistency of feature selection using greedy least squares regression*, J. Machine Learning Research, 2008.

[26] ZHANG, T., *Sparse Recovery with Orthogonal Matching Pursuit under RIP*, Tech Report arXiv:1005.2249, 2010.

[27] ZHAO, P, ROCHA, G. AND YU, B., *Grouped and hierarchical model selection through composite absolute penalties,* Manuscript, 2006.

[28] ZOU, H., HASTIE T., *Regularization and variable selection via the Elastic Net.*, J. R. Statist. Soc. B, **67**(2) 301-320, 2005.