Chapter 4
The group Lasso

Abstract In many applications, the high-dimensional parameter vector carries a structure. Among the simplest is a group structure where the parameter is partitioned into disjoint pieces. This occurs when dealing with factor variables or in connection with basis expansions in high-dimensional additive models as discussed in Chapters 5 and 8. The goal is high-dimensional estimation in linear or generalized linear models being sparse with respect to whole groups. The group Lasso, proposed by Yuan and Lin (2006) achieves such group sparsity. We discuss in this chapter methodological aspects, and we develop the details for efficient computational algorithms which are more subtle than for non-group problems.

4.1 Organization of the chapter

We present in this chapter the group Lasso penalty and its use for linear and generalized linear models. The exposition is primarily from a methodological point of view but some theoretical results are loosely described to support methodology and practice. After an introduction in Section 4.2 with the definition of the group Lasso penalty, we present in Section 4.3 the important case with factor variables including a specific example. In Section 4.4 we sketch the statistical properties of the group Lasso estimator while a mathematically rigorous treatment is presented later in Chapter 8. In Section 4.5 we discuss a slight generalization of the group Lasso penalty which is more flexible and we explain more about parametrizations and their invariances. In Section 4.7 we give a detailed treatment of computational algorithms for the Group Lasso. Thereby, the case with squared error loss is substantially simpler than for non-squared error losses as arising in generalized linear models.

P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, DOI 10.1007/978-3-642-20192-9_4,

4.2 Introduction and preliminaries

In some applications, a high-dimensional parameter vector β in a regression model is structured into groups $\mathcal{G}_1, \ldots, \mathcal{G}_q$ which build a partition of the index set $\{1, \ldots, p\}$. That is, $\cup_{j=1}^{q} \mathcal{G}_j = \{1, \ldots, p\}$ and $\mathcal{G}_j \cap \mathcal{G}_k = \emptyset$ $(j \neq k)$. The parameter vector β then carries the structure

$$\beta = (\beta_{\mathcal{G}_1}, \ldots, \beta_{\mathcal{G}_q}), \;\; \beta_{\mathcal{G}_j} = \{\beta_r;\ r \in \mathcal{G}_j\}. \tag{4.1}$$

An important class of examples where some group structure occurs is in connection with factor variables. For example, consider a real-valued response variable Y and p categorical covariates $X^{(1)}, \ldots, X^{(p)}$ where each $X^{(j)} \in \{0,1,2,3\}$ has 4 levels encoded with the labels $0,1,2,3$. Then, for encoding a main effect describing the deviation from the overall mean, we need 3 parameters, encoding a first-order interaction requires 9 parameters and so on. Having chosen such a parametrization, e.g., with sum contrasts, the group structure is as follows. The main effect of $X^{(1)}$ corresponds to $\beta_{\mathcal{G}_1}$ with $|\mathcal{G}_1| = 3$; and likewise, the main effect of all other factors $X^{(j)}$ corresponds to $\beta_{\mathcal{G}_j}$ with $|\mathcal{G}_j| = 3$ for all $j = 1, \ldots, p$. Furthermore, a first-order interaction of $X^{(1)}$ and $X^{(2)}$ corresponds to $\beta_{\mathcal{G}_{p+1}}$ with $|\mathcal{G}_{p+1}| = 9$, and so on. More details are described in Section 4.3.

Another example is a nonparametric additive regression model where the groups $\mathcal{G}_j$ correspond to basis expansions for the jth additive function of the jth covariate $X^{(j)}$. A detailed treatment is given in Chapter 5.

4.2.1 The group Lasso penalty

When estimating models with a group structure for the parameter vector, we often want to encourage sparsity on the group-level. Either all entries of $\hat{\beta}_{\mathcal{G}_j}$ should be zero or all of them non-zero. This can be achieved with the group Lasso penalty

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2, \tag{4.2}$$

where $\|\beta_{\mathcal{G}_j}\|_2$ denotes the standard Euclidean norm. The multiplier m_j serves for balancing cases where the groups are of very different sizes. Typically we would choose

$$m_j = \sqrt{T_j},$$

where T_j denotes the cardinality $|\mathcal{G}_j|$.