Parametric Linear Programming Package for Regularization Methods

Yonggang Yao
The Ohio State University, USA

Software Description lpRegPath is an R package designed for solving a family of regularization problems that satisfy certain conditions on their loss and penalty. By incorporating the additive penalty, regularization methods modify many commonly used statistical procedures to deal with both ill-posedness and over-fitting. For instance, LASSO is a modified version of linear regression (Tibshirani; 1996). Such modification may ensure that the solution is unique and continuous in the data. In general, a regularization method involves a tuning parameter which controls the trade-off between goodness of fit and model complexity. Instead of solving the regularization problem for a fixed value of the tuning parameter, lpRegPath implements an algorithm that generates the entire solution path as a function of the tuning parameter. Such a solution path offers rich information about how constrained models evolve with features and what features are persistent in the data, which are of general interest in data analysis. In addition to facilitating computation and tuning, the path-finding algorithm for feature selection can equip the data analyst with a useful tool for visualizing a path of the fitted models instead of a single one. With the aids of risk measures, such a path can portray a full spectrum of potentially good models for selection and averaging.

The regularization methods for feature selection that the package can currently handle include \( l_1 \)-norm Support Vector Machine (SVM) (Zhu et al.; 2004), \( l_1 \)-norm Quantile Regression (Li and Zhu; 2005), and functional component selection for kernel methods (e.g., \( \theta \)-step of multi-category SVM (Lee et al.; 2006)). For each of the regularization methods, lpRegPath generates the entire solution path, empirically evaluates the performance of all the models in the path, selects the most plausible models or features, outputs numerical as well as graphical summaries, and make predictions for new data sets.

The family of the regularization methods in consideration can be rephrased as parametric linear programming (LP) problems as noted in Yao and Lee (2007). To take advantages of the commonality in the problems for computational efficiency, the package has the core module that implements a tailored tableau-simplex method for generating the solution path of the parametric LP. In addition, for each regularization method, it has a shell function that identifies the corresponding components in the standard form of the LP and calls the same core module for computation. Due to this structural division of the core module and shell functions, other regularization procedures with convex and piecewise linearity can be easily incorporated in the package. For example, the path-finding algorithm can be extended to Dantzig selector (Candes and Tao; 2007) and Support Vector regression with \( \epsilon \)-insensitive loss (Vapnik; 1998) by writing relevant shell functions.

Methodology To further illustrate the core-shell relationship, we mathematically describe the connections between the parametric LP and regularization methods. Let \( \mathcal{L}(y, f(x)) \) denote a convex loss function for the prediction error and \( J(f) \) be a convex penalty functional that measures the model complexity. The empirical risk with respect to \( \mathcal{L} \) is defined by \( L(Y, f(X)) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) \), where \( Y := (y_1, \cdots, y_n)' \) and \( X := (x_1', \cdots, x_n') \). Formally, the regularization problems are defined to be

\[
\min \ L(Y, f) + \lambda \cdot J(f) \quad \text{or} \quad \min \ L(Y, f) \quad \text{subject to} \quad J(f) \leq s \quad \text{with respect to} \ f \in \mathcal{F},
\]

where \( \lambda \) and \( s \) are the pre-specified nonnegative regularization parameters, and \( \mathcal{F} \) denotes a model space. Its solution can be viewed as a function of \( \lambda \) or \( s \), called the solution path.

Consider the model space \( \mathcal{F} := \{ f(x; \beta) : \beta \in D \} \) with \( \beta \) and \( D \) respectively being the model parameter and the parameter space. The regularization problem in (1) can be transformed into a parametric LP problem, if both of \( L \) and \( J \) are convex piecewise linear functions with respect to \( \beta \) and \( D \) is a polyhedron. An

†Address for correspondence: Yonggang Yao, Department of Statistics, The Ohio State University, 1958 Neil Ave, Columbus, OH 43210, USA.
E-mail: yao@stat.osu.edu
example is the $l_1$-norm SVM which can be written as:
\[
\min_{\beta_0, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i^+ + \lambda \|\beta\|_1 \right\} \quad \text{s.t. } \xi_i = 1 - y_i(\beta_0 + \beta'x_i) \text{ for } i = 1, \cdots, n.
\]

The involved LP formulations are respectively

\[
\begin{cases}
\min_{z \in \mathbb{R}^N (c + \lambda a)'z} \\
\text{subject to } A z = b \\
z \geq 0,
\end{cases}
\]

\[
\begin{cases}
\min_{z \in \mathbb{R}^N} c'z \\
\text{subject to } A z = b \\
a'z \leq s \\
z \geq 0,
\end{cases}
\]

where $z$ is an $N$-vector of variables, $c$ and $a$ are fixed $N$-vectors, $b$ is a fixed $M$-vector, and $A$ is an $M \times N$ fixed matrix. The transformation is implemented by shell functions in \texttt{lpRegPath}. The former in (2) is the standard form of the parametric-cost LP, whose solution is a piecewise constant function. And the latter can be viewed as a special case of the parametric right-hand-side LP, whose solution is a piecewise linear function. To characterize the solutions completely, it is sufficient to identify the joint solutions at a finite sequence of $\lambda_l$. By the correspondence between the two formulations in (2), solving one of them will produce the solution to the other.

Tableau-simplex algorithm is a reliable and efficient way to generate the solution path for a parametric-cost LP problem (Murty; 1983; Bertsimas and Tsitsiklis; 1997). By using the commonality in the structure of $A$ (that is $A = [A^*, I, -I]$) for the regularization methods in consideration, the tableau-simplex algorithm can be tailored into a faster algorithm. Finding each joint solution, the computational complexity of the tailored algorithm is less than $O(M(N - 2M))$ compared with $O(MN)$ for the original one. Such a simplification is the theoretical basis of the core program in \texttt{lpRegPath}.

By connecting regularization with parametric LP, the package proposes a broad and unified paradigm that can be adopted to solve a wide family of regularization problems through the same core program. To the best of my knowledge, none of the current commercial software packages deals with generic parametric LP problems, although many LP algorithms (e.g., \texttt{lpSolve} in R) have been developed for different computing platforms (Fourer, 2007). Therefore, \texttt{lpRegPath} is a practical supplement to both regularization methods and parametric LP algorithms.

References


