# Grouped and Hierarchical Model Selection through Composite Absolute Penalties

Peng Zhao, University of California, Berkeley
pengzhao@stat.berkeley.edu

Guilherme V. Rocha, University of California, Berkeley
gvrocha@stat.berkeley.edu

Bin Yu, University of California, Berkeley
binyu@stat.berkeley.edu

November 23, 2005

**Abstract**

Recently much attention has been devoted to model selection through regularization methods in regression and classification where the penalty is capable of selecting some of the regressors (e.g. Lasso in Tibshirani,1996). While the resulting sparsity leads to more interpretable models, one may want to further incorporate natural groupings or a particular hierarchical structure present within the predictors.

Natural grouping arises in many situations. For gene expression data analysis, genes belonging to the same pathway might be viewed as a group. In ANOVA factor analysis, the dummy variables corresponding to the same factor form a natural grouping in that one wants these variables to be excluded and included in the estimated model as a group. If interaction terms are to be considered, a natural hierarchy exists as the interaction term between two factors should only be included after the corresponding main effects. A natural hierarchy also exists in multiresolution models, such as wavelet regression, where it is desirable that a lower resolution base be included before any higher resolution base in the same region.

The proposed Composite Absolute Penalties (CAP) method allows both grouping and hierarchical structures to be enforced on the estimated coefficients for the predictors, and the groups could overlap. Assume we have data $(Y_i, X_i); i = 1, ..., n$, with $X_i \in R^p$ as the predictors and $Y_i \in R$ a binary (classification) or continuous response (regression) variable. For given parameters $\beta = (\beta_1, ..., \beta_p)^T \in R^p$ and a convex loss function $L(Z; \beta)$, we propose the CAP estimator to minimize:

$$\hat{\beta} \quad = \quad L(Y, X; \beta) + \lambda \| \left\{ \|\beta_{G_1}\|_{\gamma_1}, \|\beta_{G_2}\|_{\gamma_2}, \ldots, \|\beta_{G_k}\|_{\gamma_k} \right\} \|_{\gamma_0}$$

where $\|.\|_{\gamma_i}$ is the $L_{\gamma_i}$ norm. $G_i$'s are possibly overlapping sets of indices corresponding to the $i$-th pre-defined groups with $\beta_{G_i}$ the corresponding vector of coefficients.

Setting $\gamma_0 = 1$ and $\gamma_i > 1, \forall i \geq 1$ makes the selection to operate across the groups as in the LASSO while, within a group, either all coefficients are zero or nonzero together which enforces the natural grouping (the case $\gamma_i = 2, \forall i \geq 1$ is discussed in Yuan and Lin, 2005). Furthermore, by allowing the groups to overlap, this property also provides a mechanism for expressing hierarchical relationships between the fitted coefficients. For instance, if the parameter $\beta_2$ is included in every group where $\beta_1$ is present, then $\beta_1$ can only be included in the estimated model if $\beta_2$ is included.

When CAP with $L_1$ and $L_\infty$ norms is used for least squares regression, the exact regularization frontier can be computed efficiently by exploiting its piecewise linearity (similarly to LARS, Efron et. al 2004). For general cases, the BLasso algorithm of Zhao & Yu (2004) is used. Simulations are carried out to illustrate the method in terms of sparsity and prediction performance relative to Lasso.