Multi-level Lasso for Sparse Multi-task Regression

Aurélie C. Lozano and Grzegorz Świrszcz {ACLOZANO,SWIRSZCZ}@US.IBM.COM
IBM Watson Research Center, 1101 Kitchawan Road, Yorktown Heights NY 10598,USA

Abstract

We present a flexible formulation for variable selection in multi-task regression to allow for discrepancies in the estimated sparsity patterns accross the multiple tasks, while leveraging the common structure among them. Our approach is based on an intuitive decomposition of the regression coefficients into a product between a component that is common to all tasks and another component that captures task-specificity. This decomposition yields the Multi-level Lasso objective that can be solved efficiently via alternating optimization. The analysis of the "orthonormal design" case reveals some interesting insights on the nature of the shrinkage performed by our method, compared to that of related work. Theoretical guarantees are provided on the consistency of Multi-level Lasso. Simulations and empirical study of micro-array data further demonstrate the value of our framework.

1. Introduction

We address the problem of variable selection in the settings of multiple-output and multi-task regression. Multiple-output regression extends the basic single-output regression model to one involving multiple output variables, while multi-task regression further generalizes the classical regression model to enable joint estimation of regression models for multiple tasks (each model involving one or multiple outputs). Variable selection in such settings is of significant interest due to many relevant applications in fields ranging from econometrics to computational biology. In computational biology, for instance, the fundamental problem of understanding genome associations between expression data (predictors) and phenotypic data (response) is crucial in order to identify potential biomarkers for diseases. Since many diseases such as cancer involve a variety of related phenotypes, it is desirable to perform variable selection in multiple-output regression across the related phenotypes, as information can then be shared among them. Also gene association data is often available for multiple genes within the same pathway, and it may thus be advantageous to study the associations jointly on the multiple genes, rather than performing separate studies.

A widely used approach to tackle the variable selection problem in multiple-output and multi-task regression models, is based on extending the Lasso formulation (Tibshirani, 1994) to impose block-structured regularization via the l_1-l_q norm with $q > 1$ (Turlach et al., 2005; Obozinski et al., 2006; Negahban & Wainwright, 2009; Lounici et al., 2009; Tropp et al., 2006). Specifically, the Multi-task lasso (Obozinski et al., 2006) encourages group-wise sparsity across multiple tasks. The Multi-task lasso formulation can also be applied to multiple-output regression. In this context it is often referred to as simultaneous Lasso (Turlach et al., 2005), as simultaneous feature selection is encouraged. Namely a given feature is either selected as relevant for all the outputs simultaneously, or is excluded all-together for all the outputs. These methods have been frequently used in genome-wide association studies (Puniyani et al., 2010; Zhang et al., 2010) to identify *common* mechanism of response.

The main limitation of this multi-task lasso formalism is that a common structure is imposed across the multiple regressions/tasks, as the features are selected in an "all-in-all-out" manner. Namely the set of selected features is identical across the multiple outputs/tasks, albeit allowing for different amplitude for the selected regression coefficients. However, many important problems require more flexibility. To efficiently address the above issue, this paper proposes a novel penalized regression framework, called *Multi-level Lasso*, to allow for discrepancies in support between the multiple models, while preserving the common structure among them (and thus avoiding the loss

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

of robustness if one were to estimate the models separately). Our approach is based on an intuitive decomposition of the regression coefficients into a product between a global component that is common to all tasks and another component that captures task-specificity. Such a decomposition is very natural from the standpoint of variable selection, as it is "sparsity-preserving". Namely, a specific regression coefficient is equal to zero if either of its two components is zero; furthermore the global components control the global sparsity pattern common to all tasks. We present an efficient procedure to solve the resulting optimization problem, and derive closed-form shrinkage formulae for the Multi-level Lasso in the case of orthonormal design. We also examine the shrinkage of another decomposition method recently proposed to tackle the same problem, the dirty model of (Jalali et al., 2010), which re-parameterizes the regression coefficients as a *sum* between two components (rather than a product). This reveals some interesting insights on the differences between the two methods. Another relevant model of two-level sparsity was proposed in (Dhillon et al., 2011) for an l_0 setting. As future work it would certainly be interesting to compare our method with such greedy approach. We demonstrate the strength of the Multi-level Lasso on simulated data comparing it against multitask-lasso, the dirty model, individual lasso for each task, and lasso on an aggregated dataset. Experiments on real microarray data further illustrate the usefulness of our approach.

The Multi-level Lasso objective we propose is convex with respect to each of its parameters but is not jointly convex. We show, however, that the solutions still enjoy attractive theoretical properties. In particular we characterize the asymptotic distribution of the Multi-level Lasso estimator. Our theoretical analysis together with our empirical findings are in line with the recent body of work on non-convex penalties in demonstrating their benefits as an alternative to traditional convex formulations. For instance (Fan & Li, 2001) proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty to circumvent the drawbacks of the Lasso penalty, in particular with respect to bias. Another pertinent case is that of the l_q pseudo-norm where $0 < q < 1$ which has been shown to provide sparser solutions than the Lasso and for which several theoretical guarantees exit on the accuracy of variable selection (e.g. (Fu & Knight, 2000)).

Note that our work considers general cases where there is no prior knowledge on the relatedness of the outputs. In cases where some knowledge is available, the works (Kim & Xing, 2010; Lu et al., 2009) are relevant approaches capable of accounting for discrepancies in structure across datasets. Specifically (Kim & Xing, 2010) addresses cases where outputs are related by a known tree structure so the problem can then be cast as a group lasso with groups induced by the tree. The formulation of (Lu et al., 2009) is based on (penalized) Hidden Markov Random Fields. In this setting, each dataset corresponds to a node in a relational graph, which embodies prior information on the relatedness between tasks, and is assigned a hidden state by leveraging the relational graph. Another noteworthy approach considers an adaptive multiple-output Lasso formulation, where mixture weight are introduced over the features (Lee et al., 2010) that are estimated via a Bayesian framework. Note that the formulation is presented for the multiple-output case, not the full multi-task setting.

2. Problem Formulation: Multi-Level Lasso

We formulate the problem for the multi-task setting. The formulation for multiple-output regression follows in a straightforward manner, as it corresponds to the special case where all tasks share the same predictor matrix. Assume that there are K tasks. Let $X^{(k)} \in \mathbf{R}^{n_k \times p}$ denote the predictor matrix for the k^{th} task, whose rows are p-dimensional feature vectors for n_k training examples. Denote by $X_{ij}^{(k)}$ the j^{th} observation on the i^{th} feature. Similarly let $Y^{(k)} \in \mathbf{R}^{n_k}$ denote the response vector for the k^{th} task. For simplicity assume that data has been standardized so that we need not consider intercept terms. Consider the K-task linear regression model:

$$Y^{(k)} = X^{(k)}\bar{\beta}^{(k)} + \epsilon^{(k)}, \quad k = 1, \ldots, K,$$

where $\bar{\beta}^{(k)} \in \mathbf{R}^p$ are formed by the true regression coefficients one wishes to estimate, and $\epsilon^{(k)} \in \mathbf{R}^{n_k}$ is the error term. Let $\beta_i = (\beta_i^{(1)}, \ldots, \beta_i^{(K)})^T$, be the vector formed by the concatenation of all the coefficients of the i^{th} feature across all tasks. The multi-task Lasso estimate is the solution to the following penalized regression problem: $\min_\beta \frac{1}{2}\sum_{k=1}^{K} \|Y^{(k)} - X^{(k)}\beta^{(k)}\|_2^2 + \lambda \sum_{i=1}^{p} \|\beta_i\|_\alpha$. Typical choices for α are 2 (Obozinski et al., 2006) and ∞ (Zhang, 2006). Both encourage a common sparsity pattern across the multiple tasks. As we have stated earlier, such an "all-in-all-out" formulation is too rigid in many situations.

2.1. The Multi-Level Lasso Objective

We now motivate our multi-level approach that is based on decomposing the regression coefficients into two components: one component reflects the part that

is common across tasks, the second component accounts for the part that is task-specific.

The traditional approach in Bayesian statistics is to employ a *linear mixed effects model*, where the vector of regression coefficients for each task is rewritten as a sum between a *fixed effect* vector that is constant across tasks, and a *random effect* vector that is task-specific. Formally, for the coefficients corresponding to the i^{th} feature, $i \in \{1, \dots, p\}$, β_i^k is rewritten as $\beta_i^{(k)} = \kappa_i + \delta_i^{(k)}, \quad k = 1, \dots, K$. However, the classical linear mixed effects model is not very natural from the standpoint of variable selection. For instance if the ith feature is irrelevant for all tasks, we would need $\kappa_i = 0$ and $\delta_i^{(k)} = 0$ for all k.

A more natural setup would consist of having the "main effect" variables control the "global" sparsity. Namely it would be desirable to have the property that $\kappa_i = 0 \Rightarrow \beta_i^{(k)} = 0 \ \ \forall k$. We thus propose an alternative decomposition that satisfies the desired property by rewriting β_i^k as:

$$\beta_i^{(k)} = \theta_i \gamma_i^{(k)}, \ k = 1, \dots, K, \ i = 1, \dots, p,$$

and consider $\theta_i \geq 0$, to remove ambiguity (i.e. for model identifiability). As desired, for the i^{th} feature, the θ_i's will induce the sparsity pattern common across tasks, while the $\gamma_i^{(k)}$'s will reflect task specificity.

We then propose to address multi-task variable selection via the following optimization problem:

$$\min_{\theta_i \geq 0, \gamma_i^{(k)}} \frac{1}{2} \sum_{k=1}^{K} \|Y^{(k)} - \sum_{i=1}^{p} \theta_i \gamma_i^{(k)} X_i^{(k)}\|_2^2 + \lambda_1 \sum_{i=1}^{p} \theta_i + \lambda_2 \sum_{k=1}^{K} \sum_{i=1}^{p} |\gamma_i^{(k)}|. \quad (1)$$

Let $\theta = [\theta_1^T, \dots, \theta_p^T]^T$, $\gamma^{(k)} = [(\gamma_1^{(k)})^T, \dots, (\gamma_p^{(k)})^T]^T$, $\gamma = [(\gamma^{(1)})^T, \dots, (\gamma^{(K)})^T]^T$. With this notation, the above objective can be rewritten as:

$$\min_{\theta \geq \mathbf{0}, \gamma} \frac{1}{2} \sum_{k=1}^{K} \|Y^{(k)} - \sum_{i=1}^{p} \theta_i \gamma_i^{(k)} X_i^{(k)}\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\gamma\|_1.$$

Aside from the constraint that θ should have non-negative entries, the model penalty is the sum of two l_1 penalties, one at a global level, one a task-specific level, hence we call our formulation *Multi-level Lasso*.

2.2. Algorithm for the Multi-level Lasso

In this section we present a procedure to efficiently solve the Multi-level Lasso problem. We adopt an alternating optimization approach, where we iteratively solve for either γ or θ, while fixing the other. Minimizing the Multi-level Lasso objective with respect to γ while fixing θ boils down to solving a classical Lasso problem, which can be efficiently solved (e.g. using (Efron et al., 2004) or (Friedman et al., 2007)). Minimizing the objective with respect to θ while fixing γ reduces to solving a classical non-negative Garrote objective (Breiman, 1995), which can also be solved efficiently (e.g. via (Yuan & Lin, 2007), (Cantoni et al., 2006)). The alternating optimization procedure is stated as Algorithm 1. Note that each step of the algorithm decreases the original objective (1). Hence the procedure necessarily converges.

Algorithm 1 Alternate Optimization Algorithm for the Multi-level Lasso

Input: Standardized training data $X^{(k)}, Y^{(k)}, k = 1, \dots, K$, parameters $\lambda_1, \lambda_2, \epsilon > 0$.
Initialize $m = 0$, $\theta_i(0) = 1$, $i = 1, \dots, p$ and $\gamma^{(k)}(0) = \hat{\gamma}^{(k)}$, $k = 1, \dots, K$, where $\hat{\gamma}^k$ is an initial estimate (e.g. the ordinary least estimate for task k or the estimate from a ridge regression). Set $\beta_j^{(k)}(0) = \theta_j(0)\gamma_j^{(k)}(0)$ $\forall k$.
for $m = 1...$ **do**
  //Solve for γ
  Let $W_i^{(k)} = \theta_i(m-1)X_i^{(k)}$, $i = 1, \dots, p$, $k = 1, \dots, K$.
  Solve the Lasso problem (P_1): $\gamma(m) = \arg\min_{\gamma} \frac{1}{2} \sum_{k=1}^{K} \|Y^{(k)} - \sum_{i=1}^{p} \gamma_i^{(k)} W_i^{(k)}\|_2^2 + \lambda_2 \|\gamma\|_1$
  //Solve for θ
  Let $Z_i = [(\gamma_i^{(1)}(m)X_i^{(1)})^T, \dots, (\gamma_i^{(K)}(m)X_i^{(K)})^T]^T$, $i = 1, \dots, p$, $Y = [(Y^{(1)})^T, \dots, (Y^{(K)})^T]^T$.
  Solve the non-negative Garrote problem (P_2): $\theta(m) = \arg\min_{\theta \geq \mathbf{0}} \frac{1}{2} \|Y - \sum_{i=1}^{p} \theta^{(i)} Z^{(i)}\|_2^2 + \lambda_1 \|\theta\|_1$
  //Update β
  Set $\beta_i^{(k)}(m) = \theta_i(m)\gamma_i^{(k)}(m)$
  If $R(\beta(m-1)) - R(\beta(m)) \leq \epsilon$ **break**, where $R(\beta)$ denotes the squared loss over all tasks.
end for

2.3. Orthonormal Design Case and Relationship with the Dirty Model decomposition

As advocated in (Tibshirani, 1994), inspecting the special case of an orthonormal design sheds light on the nature of the shrinkage.

Shrinkage for the Multi-Level Lasso: In the orthonormal design case, closed-form solutions are readily available for each step of the alternating optimization algorithm 1, as it is well know that both lasso and non-negative garrote estimators are equivalent to

soft-thresholding operators in the orthonormal case. Let $\hat{\beta}_i^{(k)} = (X_i^{(k)})^T Y^{(k)}\ \forall i,k$. Namely $\hat{\beta}$ if the ordinary least square estimate. Similarly as in (Tibshirani, 1994) we obtain for the step (P_1) of Algorithm 1:

$$(P_1')\quad \gamma_i^{(k)}(m) \;=\; \mathbf{I}[\theta_i(m-1)>0]\mathrm{sign}(\hat{\beta}_i^{(k)}) \times \left(\frac{|\hat{\beta}_i^{(k)}|}{\theta_i(m-1)} - \frac{\lambda_2}{[\theta_i(m-1)]^2}\right)^+,$$

and for the step (P_2) of Algorithm 1:

$$(P_2')\quad \theta_i(m) = \mathbf{I}[\|\gamma_i(m)\|_1 > 0] \times \left(\hat{\theta}_i(\gamma(m),\hat{\beta}_i) - \frac{\lambda_1}{(\sum_{k=1}^K(\gamma_i^{(k)}(m))^2)^{3/2}(\hat{\theta}_i(\gamma(m),\hat{\beta}_i))}\right)^+,$$

where $\hat{\theta}_i(\gamma(m),\hat{\beta}_i) = \frac{\sum_{k=1}^K \gamma_i^{(k)}(m)\hat{\beta}_i^{(k)}}{\sum_{k=1}^K(\gamma_i{}^{(k)}(m))^2}$.

Focusing on (P_1'), first note that if $\theta_i(m-1) = 0$ for the i^{th} feature, then all the task-specific coefficients $\gamma_i^{(k)}(m)$ for that feature are also set to zero, which is desirable. Noting that $\gamma_i^{(k)}$ is "comparable" to $\frac{\hat{\beta}_i^{(k)}}{\theta_i}$, we can see that (P_1') has the effect of shrinking $\gamma_i^{(k)}$ towards zero, by a quantity proportional to $1/[\theta_i(m-1)]^2$. Intuitively, this reflects the fact that since a large value for the estimated θ_i indicates that the ith feature is considered to be significant, the task-specific coefficients for that feature should not be shrunk by a large amount.

Focusing on (P_2'), note that if all the task-specific coefficients for the i^{th} feature are zero, then so is θ_i, as desired. Also, noting that θ_i is "comparable" to $\hat{\theta}_i(\gamma(m),\hat{\beta}_i)$ (since the latter is a weighted average of $\hat{\beta}_i^{(1)}/\gamma_i^{(1)}(m),\ldots,\hat{\beta}_i^{(K)}/\gamma_i^{(K)}(m)$), we can see that the effect of (P_2') is to shrink the estimate for θ_i towards 0 by a quantity whose dependence on $\gamma_i^{(k)}$ is of the order of $1/(\gamma_i^{(k)})^2$. Again this is consistent with the intuition that the larger the task-specific coefficients for the ith feature, the lesser shrinkage should be applied to θ_i.

Shrinkage for the dirty model: We now examine the nature of the shrinkage for the dirty block sparse method of (Jalali et al., 2010), which re-parameterizes the regression coefficients as a sum of two coefficients: $\beta_i^{(k)} = s_i^{(k)} + b_i^{(k)}$, and imposes different regularization to the resulting vectors $s^{(k)}$ and $b^{(k)}$. While the former are encouraged to have task-specific sparsity patterns, the latter are encouraged to exhibit identical sparsity patterns. Notice that our model involves $(p+1)K$ parameters, whereas the dirty model involves $2pK$ parameters, as both sets of coefficients depend on task and feature. Formally the dirty block sparse method solves

$$\min_{s_i^{(k)},b_i^{(k)}} \tfrac{1}{2}\sum_{k=1}^K \|Y^{(k)} - \sum_{i=1}^p (s_i^{(k)} + b_i^{(k)})X_i^{(k)}\|^2 + \lambda_1\|S\|_{1,1} + \lambda_2\|B\|_{1,\infty},$$

where S and B are the matrices formed by the coefficients $s_i^{(k)}$ and $b_i^{(k)}$ respectively, $\|S\|_{1,1} = \sum_{i,k}|s_i^{(k)}|$, and $\|B\|_{1,\infty} = \sum_i \max_k |b_i^{(k)}|$. The objective can be solved by alternating minimization fixing B and S respectively. We now examine the method's behavior under orthonormal design. Unfortunately there is no closed-form shrinkage formula for the $l_{1,\infty}$ norm. To shed light on the nature of the shrinkage we use the $l_{1,2}$ norm as a proxy and consider instead

$$\min_{s_i^{(k)},b_i^{(k)}} \tfrac{1}{2}\sum_{k=1}^K \|Y^{(k)} - \sum_{i=1}^p (s_i^{(k)} + b_i^{(k)})X_i^{(k)}\|^2 + \lambda_1\|S\|_{1,1} + \lambda_2\|B\|_{1,2},$$

where $\|B\|_{1,2} = \sum_i(\sum_k {b_i^{(k)}}^2|)^{1/2}$.
At iteration m, fixing B, we get:
$(D_1)\ s_i^{(k)}(m) = \mathrm{sign}(\hat{\beta}_i^{(k)} - b_i^{(k)}(m-1))(|\hat{\beta}_i^{(k)} - b_i^{(k)}(m-1)| - \lambda_1)^+$.
Then, fixing S we get:
$(D_2)\ b_i(m) = \left(1 - \frac{\lambda_2}{\|\hat{\beta}_i - s_i(m)\|_2}\right)^+ (\hat{\beta}_i - s_i(m))$.
Noting that $s^{(k)}$ is comparable to $\hat{\beta}^{(k)} - b^{(k)}$, we can see that (D_1) is comparable to $s_i^{(k)}(m) = \mathrm{sign}(s_i^{(k)}(m-1))(|s_i^{(k)}(m-1)| - \lambda_1)^+$, and hence the amount of shrinkage applied to $s_i^{(k)}$ is similar across features and tasks. Similarly noting that $b^{(k)}$ is comparable to setting $\hat{\beta}^{(k)} - s^{(k)}$, we can see that (D_2) is comparable to setting $b_i(m) = \left(1 - \frac{\lambda_2}{\|b_i(m-1)\|_2}\right)^+ (b_i(m-1))$, and hence the amount of shrinkage applied to $b_i^{(k)}$ is similar across tasks.

Comparing (P_1') and (P_2') to (D_1) and (D_2) reveals that the shrinkage of the global and task-specific coefficients are more tightly coupled for our multi-level lasso model than for the dirty model. The product decomposition used in our model is more natural from the standpoint of variable selection, whereas the dirty model employs an additive decomposition, which similarly to the linear mixed effects model is not sparsity preserving: under the dirty model the ith feature is excluded from task k, if both of its component coefficients $s_i^{(k)}$ and $b_i^{(k)}$ are equal to zero, or if $s_i^{(k)} = -b_i^{(k)}$. An additional advantage of our multi-level Lasso model is that its global coefficients have a useful interpretation with respect to "sure screening" as the sparsity pattern of the global coefficients induces the removal of the features irrelevant to all tasks.

2.4. Extensions for Variable Grouping and Multi-task Multiple-output Regression

Our formulation can be readily extended to incorporate input variable grouping. Consider G groups

and denote by $X_g^{(k)}$ the columns of $X^{(k)}$ corresponding to the g^{th} group. Consider the decomposition $\beta_g^{(k)} = \theta_g \gamma_g^{(k)}$, where $\beta_g^{(k)}$ and $\gamma_g^{(k)}$ are coefficient vectors for the g^{th} group, and θ_g is a scalar controlling the group-sparsity across tasks. Then the multi-level group lasso objective is: $\min_{\theta_i \geq 0, \gamma_i^{(k)}} \frac{1}{2} \sum_{k=1}^K \|Y^{(k)} - \sum_{g=1}^G \theta_g X_g^{(k)} \gamma_g^{(k)}\|_2^2 + \lambda_1 \sum_{g=1}^G \theta_g + \lambda_2 \sum_{k=1}^K \sum_{g=1}^G \|\gamma_g^{(k)}\|_2$.
Algorithm 1 can be extended in a straightforward manner: Solving for $\gamma_g^{(k)}$ boils down to a classical group-lasso problem, while solving for θ_g can still be reduced to a non-negative garrote problem by considering $Z_g = [(X_g^{(1)}\gamma_g^{(1)})^T, \ldots, (X_g^{(K)}\gamma_g^{(K)})^T]^T$, $g = 1, \ldots, G$. In addition, note that out Multi-level Lasso objective can also be generalized in a straightforward manner to handle the case where each task involves a multiple-output regression. Namely for each k, $Y^{(k)}$ is an $n_k \times q$ matrix and $\beta^{(k)}$ is a $p \times q$ regression coefficient matrix. A similar re-parametrization is used in (Guo et al., 2011) for learning multiple graphical models, but the procedure and algorithm depart significantly from ours, as the objective cannot be solved directly and local linear approximation of the penalty is performed.

3. Theoretical Guarantees

The Multi-level objective of (1) is convex with respect to each of its parameters θ and γ individually but it is not jointly convex. However the local solutions still enjoy attractive theoretical properties. In this section we characterize the asymptotic distribution of the Multi-level Lasso estimator. Details of the proofs are skipped due to space constraint, but will be provided in a longer version of this manuscript. The first step is to show that the Multi-level Lasso objective can be reformulated as an alternate optimization problem as follows.

Proposition 1 *Solving the Multi-level Lasso problem of (1) is equivalent to solving*

$$\min \frac{1}{2n} \sum_{k=1}^K \|Y^{(k)} - X^{(k)}\beta^{(k)}\|_2^2 + \lambda T(\boldsymbol{\beta}), \qquad (2)$$

with $\lambda = \frac{2\sqrt{\lambda_1 \lambda_2}}{n}$ *and where* T *is defined as*

$$T(\boldsymbol{\beta}) = \sum_{i=1}^p \sqrt{\|\beta_i\|_1}, \text{ with } \|\beta_i\|_1 = \sum_{k=1}^K |\beta_i^{(k)}|. \qquad (3)$$

The proof of the proposition follows a reasoning similar to to (Lin & Zhang, 2006).

We will state the convergence theorem in terms of the equivalent objective (2). Before we do so we need to introduce the following relevant quantities. Let $R(\beta) = \frac{1}{2} \sum_{k=1}^K \mathbf{E}\|Y^{(k)} - \beta^{(k)}X(k)\|_2^2$, be the risk corresponding to the (unpenalized) squared loss $L(\beta) = \frac{1}{2} \sum_{k=1}^K \|Y^{(k)} - \beta^{(k)}X(k)\|_2^2$. Let $\bar{\beta}$ be the minimizer of the risk: $\bar{\beta} = \arg\min_\beta R(\beta)$. Let $\hat{\beta}(\lambda)$ be a solution of (2). Let

$$\begin{aligned} &J(\bar{\beta}) \\ &= \mathbf{E}[\nabla_\beta L(\bar{\beta}) \nabla_\beta L(\bar{\beta})^T] \\ &= \mathbf{E}\left[\begin{pmatrix} (Y^{(1)} - X^{(1)}\bar{\beta}^{(1)})^T X^{(1)} \\ \vdots \\ (Y^{(K)} - X^{(K)}\bar{\beta}^{(K)})^T X^{(K)} \end{pmatrix} \begin{pmatrix} (Y^{(1)} - X^{(1)}\bar{\beta}^{(1)})^T X^{(1)} \\ \vdots \\ (Y^{(K)} - X^{(K)}\bar{\beta}^{(K)})^T X^{(K)} \end{pmatrix}^T \right]. \qquad (4) \end{aligned}$$

Let H be the Hessian of the risk R. Since we are dealing with the squared loss, the Hessian is constant and does not depend on β. We have

$$H = \text{diag}(\mathbf{E}X^{(1)^T}X^{(1)}, \ldots, \mathbf{E}X^{(K)^T}X^{(K)}). \qquad (5)$$

Let $G_{\bar{\beta}}(u) = \lim_{h \to 0} \frac{T(\bar{\beta} + hu) - T(\bar{\beta})}{h}$, where T is defined in (3). Denote by $\mathcal{I}$ the set of indices i for which $\bar{\beta}_i$ is not all zero. We obtain

$$\begin{aligned} G_{\bar{\beta}}(u) \quad = \quad & \frac{1}{2} \sum_{i \in \mathcal{I}} \|\bar{\beta}_i\|_1^{-1/2} \sum_{k=1}^K u_i^{(k)} \text{sign}(\bar{\beta}_i^{(k)}) \mathbf{I}(\bar{\beta}_i^{(k)} \neq 0) \\ & + |u_i^{(k)}| \mathbf{I}(\bar{\beta}_i^{(k)} = 0). \qquad (6) \end{aligned}$$

The following theorem characterizes the asymptotic distribution of the Multi-level Lasso estimator and shows that it is $\sqrt{n}$ consistent.

Theorem 1 *Consider a sequence* $\lambda_n = \frac{2\sqrt{\lambda_{1,n}\lambda_{2,n}}}{n}$, *with* $n = \sum_{k=1}^K n_k$ *such that* $\lambda_n n^{-1/2} \to \lambda \geq 0$ *as* $n \to \infty$. *Let* $V_{\bar{\beta}}(w, u) = u^T H u + w^T u + \lambda G_{\bar{\beta}}(u)$,*where* $u \in \mathcal{R}^{pK}$ *and* $v \in \mathcal{R}^{pK}$, H *and* $G_{\bar{\beta}}(u)$ *are defined in (5) and (6) respectively. There exists a random vector* $W \sim \mathcal{N}(0, J(\bar{\beta}))$, *where* J *is defined in (4), such that*

$$\sqrt{n}(\hat{\beta}_n(\lambda_n) - \bar{\beta}) \xrightarrow{d} \arg\min_u V_{\bar{\beta}}(W, u)$$

In particular if $\lambda_n \to 0$ *and the error terms are i.i.d. with mean zero and variance* $(\sigma^{(k)})^2$, $k = 1, \ldots, K$ *we get*

$$\sqrt{n}(\hat{\beta}_n(\lambda_n) - \bar{\beta}) \xrightarrow{d} -H^{-1}W \sim \mathcal{N}(0, \Sigma^2 H^{-1}),$$

where $\Sigma = diag(\sigma^{(1)} I_p, \ldots, \sigma^{(K)} I_p)$.

The proof follows easily from Theorem 4 in (Rocha et al., 2009). As remarked in (Rocha et al., 2009), local minima may exist for finite sample, yet asymptotically the penalty is negligible compared to the squared

loss and the minimizer is unique. Though the penalty in (3) is concave, computation of the Hessian of the full objective (risk + penalty) reveals that there are regions where the objective is locally convex. We plan to characterize these regions as future work (in the spirit of (Breheny & Huang, 2011)).

4. Experimental Results

4.1. Synthetic Data

We evaluate the performance of Multi-level Lasso against Multi-task Lasso (abbreviated as "Multi-task"), the dirty model estimator, running Lasso independently for each task (referred to as "indLasso"), and running Lasso on the aggregated dataset formed by combining data for all task (referred to as "allLasso"). As a measure of variable selection accuracy, we use the F_1 measure, which is the harmonic mean of precision and recall (The F_1 measure is between 0 and 1; the larger F_1, the higher the accuracy). For all methods, we consider the "holdout validated" estimates, namely we select the penalty parameters that minimize the average squared error on a validation set. We remark that for the multi-level Lasso, the coefficient vectors γ and θ for the multi-level Lasso are combined multiplicatively. Therefore only one of the two regularization parameter is necessary, since one can always multiply γ by a constant and divide θ by the same constant. Thus, one does not need to search over a 2-D grid of regularization parameters, while such a search cannot be avoided for the dirty model. We consider the K-tasks regression model $Y^{(k)} = X^{(k)}\bar{\beta}^{(k)} + \epsilon^{(k)}, k = 1, \ldots, K$. For each task we generate a $n \times p$ predictor matrix $X^{(k)}$, where the rows are generated independently according to $N_p(0, \mathbf{S})$, with $\mathbf{S}_{i,j} = 0.7^{|i-j|}$. The noise vector for each task is generated according to $N(0, 1)$. The true regression coefficients for each task are generated as a $p \times K$ matrix $\bar{B}$, where $\bar{B}_{i,k} = \bar{\beta}_i^{(k)}$. We consider two setups. The first setup is as follows. All the entries of $\bar{B}$ are first set to zero. Next we generate a row-wise sparsity pattern that will determine which features are irrelevant to all tasks. Specifically, the row-wise sparsity is determined by selecting $\lfloor p \cdot \rho_p \rfloor$ rows at random to contain non-zero coefficients (the remaining rows are 0 for all tasks), where ρ_p is a simulation parameter. For each of the selected rows, we introduce some amount of disagreement between tasks with respect to the within-row sparsity pattern. We do so by randomly selecting $\lfloor K \cdot \rho_K \rfloor$ entries to be set to 0, where ρ_K is a simulation parameter. Then for each non-zero entry of $\bar{B}$, independently, we set its value according to $N(0, 1)$. Note that ρ_K=0 corresponds to the case where the relevant predictors are common to all tasks, a setting that should be favorable to Multitask Lasso. Our second setup is for the extreme case where the sparsity pattern is arbitrary (no task sharing), which should be favorable to Lasso. For that setup, we randomly select $\rho_{\text{arb}} pK$ entries of $\bar{B}$ to contain non-zero coefficients, where ρ_{arb} is a simulation parameter. For each setting, we ran 50 runs. We set training and evaluation sample sizes to $n_{\text{train}} = n_{\text{eval}} = 50$, feature size to $p = 20$. We considered various combinations for the values of $(K, \rho_p, \rho_K, \rho_{\text{arb}})$, so as to enforce more or less discrepancies in the sparsity pattern across tasks, so as to consider models of varying sparsity, and varying ratios between feature dimensionality and number of tasks. The results are presented in Table 1.

Overall, Multi-level Lasso performs better than all the comparison methods in multitask settings with some amount of discrepancy across tasks. In addition, our method is remarkably competitive with Lasso and Multitask Lasso for the extreme cases of "no task-sharing" (red rows in Table 1) and "full-sharing" (blue rows in Table 1) respectively.

4.2. Application to Microarray Data Analysis

We apply our method to the analysis of gene expression data using a microarray dataset pertaining to isoprenoid biosynthesis in Arabidopsis thaliana (A. thaliana) provided by (Wille et al., 2004). A. thaliana is a small flowering plant widely used as a model organism for studies in genetics and molecular biology. Isoprenoids play a key role in major plant processes including photosynthesis, respiration and defense against pathogens. They are also important components in a variety of drugs (e.g. against cancer and malaria), fragrances (e.g. menthol) and food colorants (carotenoids). Understanding the mechasnisms of isoprenoid synthesis is thus highly relevant to a large spectrum of applications. Of particular relevance is to develop an understanding of the crosstalks between the two isoprenoid pathways: the mevalonate pathway an the plastidial pathway. In the dataset considered the predictors are the expression levels of 21 genes in the mevalonate pathway, the responses are the expression levels of 19 genes in the plastidial pathway. There are 131 samples. All variables are log transformed. The predictors are centered and standardized to unit variance.

We first evaluated the predictive accuracy of our method and the comparison methods by randomly partitioning the data into training and test sets, using 90 observations for training and the remainder for testing. The tuning parameters were selected via 5-fold

K	ρ_{P}	ρ_{K}	ρ_{arb}	Multi-level	Dirty model	Multi-task	indLasso	allLasso
5	0.5	0.5	N/A	$\mathbf{0.809 \pm 0.007}$	0.731 ± 0.008	0.645 ± 0.006	0.685 ± 0.009	0.457 ± 0.010
5	0.2	0.5	N/A	$\mathbf{0.811 \pm 0.005}$	0.765 ± 0.005	0.697 ± 0.009	0.623 ± 0.007	0.392 ± 0.008
5	0.5	0.2	N/A	$\mathbf{0.855 \pm 0.003}$	0.833 ± 0.006	0.771 ± 0.005	0.689 ± 0.009	0.572 ± 0.009
5	0.2	0.2	N/A	$\mathbf{0.908 \pm 0.004}$	0.845 ± 0.003	0.813 ± 0.005	0.636 ± 0.003	0.557 ± 0.004
5	0.5	0	N/A	0.883 ± 0.004	0.856 ± 0.005	$\mathbf{0.913 \pm 0.003}$	0.693 ± 0.004	0.660 ± 0.004
5	0.2	0	N/A	0.901 ± 0.004	0.880 ± 0.006	$\mathbf{0.938 \pm 0.005}$	0.655 ± 0.004	0.524 ± 0.004
5	N/A	N/A	0.5	0.753 ± 0.002	0.693 ± 0.003	0.668 ± 0.001	$\mathbf{0.754 \pm 0.002}$	0.666 ± 0.002
5	N/A	N/A	0.2	0.676 ± 0.005	0.676 ± 0.003	0.430 ± 0.005	$\mathbf{0.681 \pm 0.006}$	0.335 ± 0.007
10	0.5	0.5	N/A	$\mathbf{0.801 \pm 0.008}$	0.730 ± 0.007	0.622 ± 0.006	0.650 ± 0.008	0.421 ± 0.008
10	0.2	0.5	N/A	$\mathbf{0.825 \pm 0.003}$	0.727 ± 0.004	0.667 ± 0.008	0.587 ± 0.005	0.372 ± 0.004
10	0.5	0.2	N/A	$\mathbf{0.889 \pm 0.002}$	0.836 ± 0.008	0.845 ± 0.008	0.679 ± 0.004	0.587 ± 0.005
10	0.2	0.2	N/A	0.913 ± 0.007	$\mathbf{0.926 \pm 0.005}$	0.834 ± 0.006	0.600 ± 0.007	0.525 ± 0.007
10	0.5	0	N/A	0.871 ± 0.003	0.852 ± 0.004	$\mathbf{0.938 \pm 0.005}$	0.702 ± 0.006	0.664 ± 0.006
10	0.2	0	N/A	0.932 ± 0.003	0.913 ± 0.005	$\mathbf{1.000 \pm 0.006}$	0.659 ± 0.004	0.601 ± 0.004
10	N/A	N/A	0.5	$\mathbf{0.733 \pm 0.001}$	0.687 ± 0.001	0.667 ± 0.002	0.708 ± 0.005	0.663 ± 0.008
10	N/A	N/A	0.2	0.628 ± 0.004	$\mathbf{0.636 \pm 0.006}$	0.372 ± 0.004	0.625 ± 0.005	0.333 ± 0.009
20	0.5	0.5	N/A	$\mathbf{0.799 \pm 0.003}$	0.793 ± 0.002	0.653 ± 0.007	0.628 ± 0.004	0.400 ± 0.003
20	0.2	0.5	N/A	$\mathbf{0.821 \pm 0.006}$	0.739 ± 0.005	0.667 ± 0.009	0.593 ± 0.003	0.388 ± 0.003
20	0.5	0.2	N/A	$\mathbf{0.886 \pm 0.004}$	0.799 ± 0.004	0.875 ± 0.009	0.703 ± 0.006	0.601 ± 0.008
20	0.2	0.2	N/A	$\mathbf{0.898 \pm 0.005}$	0.849 ± 0.008	0.889 ± 0.003	0.638 ± 0.007	0.518 ± 0.016
20	0.2	0	N/A	0.917 ± 0.005	0.846 ± 0.007	$\mathbf{1.000 \pm 0.004}$	0.644 ± 0.009	0.575 ± 0.011
20	0.5	0	N/A	0.838 ± 0.001	0.797 ± 0.004	$\mathbf{0.995 \pm 0.003}$	0.704 ± 0.007	0.665 ± 0.012
20	N/A	N/A	0.5	$\mathbf{0.743 \pm 0.004}$	0.683 ± 0.004	0.667 ± 0.004	0.703 ± 0.001	0.667 ± 0.018
20	N/A	N/A	0.2	0.641 ± 0.003	0.579 ± 0.004	0.349 ± 0.007	$\mathbf{0.632 \pm 0.003}$	0.333 ± 0.009

Table 1. Average F_1 score for the models output by Multi-level Lasso and representative comparison methods on simulated data. (Larger values indicate higher accuracy)

Method	MSE
Multi-level	$\mathbf{0.22 \pm 0.01}$
Dirty Model	0.35 ± 0.03
Multi-task	0.64 ± 0.05
indLasso	0.36 ± 0.08
all Lasso	0.93 ± 0.06

Table 2. Test MSEs under different methods based on 100 random partitions of the microarray dataset into training and test sets. (Smaller values indicate higher predictive accuracy.

cross-validation. We computed the prediction MSE for the testing set. The average MSEs based on 100 random partitions are presented in Table 2. We can see that overall the predictive performance of the Multilevel Lasso is superior to the other methods.

We now proceed with an analysis of the associations identified by our method between genes from the mevalonate pathways (predictors) and those from the plastidial pathway (responses), using the full dataset. We apply bootstrap resampling to determine the statistical confidence of the associations identified. The associations identified using the full dataset that also appear more than 70 percent of the time in the bootstrap datasets are depicted in Figure 1. We note that several of our findings are consistent with findings from the biological literature. For instance, connections between genes *MK* and *GGPPS* 6 & 12, between genes *FPPS2* and *IPPL1*, and between genes *DPPS2* and *PPDS1* have also been reported in (Wille et al., 2004). The absence of connections stemming from genes *GGPPS*1,3,4,5,8,9 is also consistent with findings in (Wille et al., 2004). The above insights illustrate the value of our approach for gene association discovery. We plan to apply our method on a variety of datasets in this domain, and hope to shed light on important aspects of the regulatory mechanisms.

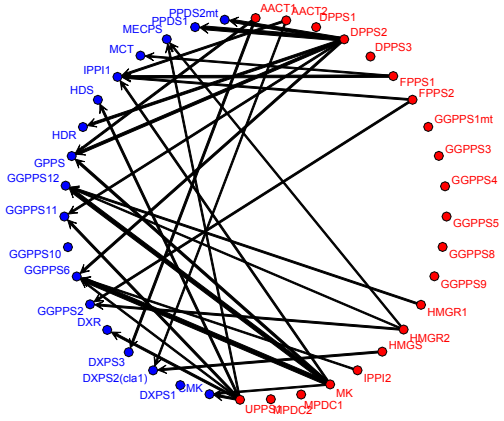

Figure 1. Associations identified by the Multi-level Lasso method between genes from the mevalonate isoprenoid pathway (in blue) and those from the plastidial pathway (in red).

References

Breheny, P. and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *ArXiv e-prints*, 2011.

Breiman, L. Better subset regression using the nonneg-

ative garrote. *Technometrics*, 37(4):373–384, 1995.

Cantoni, E., Mills Flemming, J., and Ronchetti, E. Variable selection in additive models by non-negative garrote. Technical report, Département d'Econométrie, Université de Genève, March 2006.

Dhillon, Paramveer S., Foster, Dean, and Ungar, Lyle. Minimum description length penalization for group and multi-task sparse learning. *Journal of Machine Learning Research (JMLR)*, 12:525–564, February 2011. ISSN 1532-4435.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32: 407–499, 2004.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, pp. 1348–1360, 2001.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. Technical report, Annals of Applied Statistics, 2007.

Fu, W. and Knight, K. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28:1356–1378, 2000.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Jalali, A., Ravikumar, P., Sanghavi, A., and Ruan, C. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems 23*, pp. 964–972. 2010.

Kim, S. and Xing, E.P. Tree-guided group lasso for multi-task regression with structured sparsity. In *International Conference on Machine Learning*, pp. 543–550, 2010.

Lee, S., Zhu, J., and Xing, E. Adaptive multi-task lasso: with application to eqtl detection. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 1306–1314. 2010.

Lin, Y and Zhang, H. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34:2272–2297, 2006.

Lounici, K., Tsybakov, A.B., Pontil, M., and Van De Geer, S. A. Taking advantage of sparsity in multi-task learning. In *22nd Conferecnce on Learning Theory (COLT)*. 2009.

Lu, Y., Rosenfeld, R., Nau, G. J., and Bar-Joseph, Z. Cross species expression analysis of innate immune response. In *Proc. of the 13th Annual International Conference on Research in Computational Molecular Biology*, 2009.

Negahban, S. and Wainwright, M. Phase transitions for high-dimensional joint support recovery. In *Advances in Neural Information Processing Systems 21*, pp. 1161–1168. 2009.

Obozinski, G., Taskar, B., and Jordan, M. Multi-task feature selection. Technical report, Department of Statistics, University of California, Berkeley, 2006.

Puniyani, S., Kim, S., and Xing, E.P. Multi-population gwa mapping via multi-task regularized regression. *Bioinformatics [ISMB]*, 26(12):208–216, 2010.

Rocha, G. V., Wang, X., and Yu, B. Asymptotic distribution and sparsistency for l1-penalized parametric M-estimators with applications to linear SVM and logistic regression. *ArXiv e-prints*, 2009.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Tropp, J.A., Gilbert, A.C., and Strauss, M.J. Algorithms for simultaneous sparse approximation. In *Signal Processing, Special issue on "Sparse approximations in signal and image processing*, volume 86, 2006.

Turlach, B.A., Venables, W. N., and Wright, S. J. Simultaneous variable selection. *Technometrics*, 47, 2005.

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Buhlmann, P. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5, 2004. doi: 10.1186/gb-2004-5-11-r92.

Yuan, M. and Lin, Y. On the non-negative garrotte estimator. *Journal Of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

Zhang, J. *A probabilistic framework for multi-task learning*. PhD thesis, Carnegie Mellon University, 2006.

Zhang, K., Gray, J.W., and Parvin, B. Sparse multi-task regression for identifying common mechanism of response to therapeutic targets. *Bioinformatics*, 15:i97–105, 2010.