Variable Selection and Grouping with Multiple Graph Priors

Marco Signoretto **Anneleen Daemen** **Carlo Savorgnan** **Johan A. K. Suykens**

ESAT-SCD, K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Abstract

We present an approach to learn predictive models with grouping effect induced by multiple noisy estimates of some latent graph, relevant to the task of interest. Our work extends the elastic-net and recent research on network constrained selection. The aim is both to discover and model macro-variables that are highly predictive. Such problems commonly arise for example in bioinformatics, where markers for a certain disease are often found to coincide with entire groups of variables. This paper outlines a novel problem formulation, an algorithm to obtain the model coefficients and experiments illustrating the proposed ideas.

Keywords: selection, grouping effect, convex optimization, mixed-norm penalty.

1 Introduction

Graphs are a powerful device to represent and analyze information. In particular, many biological processes have been recognized to have a natural representation as large-scale sparse networks where nodes are clustered together if the corresponding variables (e.g. gene expression profiles) are functionally related. Besides, for many diseases, markers coincide with entire groups of associated variables. For these reasons there has been a growing interest in building predictive models mainly in terms of these *macro variables*. Classical statistical approaches are not entirely adequate to accomplish this task as they treat genes as independent and equally important variables. How to incorporate the prior knowledge of known interaction patterns is then largely an open challenge and requires the development of new tools. This represents the underlying motivation for this work.

Recently Li and Li proposed the NETwork-constrained regularization and variable selection (in the sequel referred to as NET) with the purpose of embedding known interactions into the model building task [10]. The approach they propose promotes a smooth profile of coefficients associated to neighboring variables on a user-provided network. This has the potential to enhance the interpretability by identifying subgroups that are relevant for the process under study.

The concept of *grouping* has been discussed in connection with the elastic-net where the emergence of blocks is driven by the empirical correlation in the sample [15]. Instead, with NET the estimated model tends to be biased towards relevant components of the provided network. This is in principle desirable as the graph synthesizes information gathered over many years of biomedical research. Nonetheless, the network topology is often the result of an involved estimation process based upon observational data[1]. In general the nodes are given and one wants to infer whether or not the corresponding variables are associated. The estimated graph is often inaccurate with a considerable amount of misspecified associations. This fact poses a serious threat on the interpretation and could decrease the performance of the model constructed with the aid of the estimated graph prior. This possibly explains why NET sometimes performs worse than expected in real life applications [4].

[1] See [9] for a statistically oriented review about models and inference of networks and e.g. [6] and reference therein for the specific case of *correlation networks* (a.k.a. covariance selection models).

Still, more than one estimated graph is often available and one can think of integrating them. In this paper we go precisely in this direction and explore how to combine multiple priors in the attempt to build a better model. Our main contribution is threefold. We first highlight the relation between the model assumptions and the spectral properties of the network prior. We then formulate Multiple NET (simply called m-NET, in the following), an optimization problem that generalizes NET to any differentiable loss function and permits to integrate multiple graph priors. As a third contribution, we propose an algorithm which is suitable to large-scale problems. The solution strategy we consider is easily adapted to any differentiable loss function and can be directly used to compute solutions to the LASSO, elastic-net, NET and m-NET.

2 Notation and Preliminaries

For any $p \in \mathbb{N}$ we use the convention of denoting the set $\{1, \dots, p\}$ by $\mathbb{N}_p$. We write x^i_j to mean the $j-$th component of $x^i \in \mathbb{R}^p$. In the following we deal with the problem of predicting a target random variable $y \in \mathcal{Y}$ from a random vector $x \in \mathcal{X}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$. It is assumed that we are given a set on n i.i.d. observations $\left\{(x^i, y_i) \,:\, i \in \mathbb{N}_n\right\} \subset \mathcal{X} \times \mathcal{Y}$ and are interested in choosing a predictor function in a certain class. Specifically, here we shall focus on linear functions $\langle x, \beta \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^p$. We define the *empirical risk* of β as $\sum_{i \in \mathbb{N}_n} l(y_i, \langle x^i, \beta \rangle)$ where $l : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ is a *loss function*. Common examples are the quadratic loss for regression: $l(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ and the logistic loss for binary classification: $l(y, \hat{y}) = \log\left(1 + e^{-y\hat{y}}\right)$. In developing our machinery we restrict ourselves to convex and differentiable loss functions.

When discussing the integration of multiple priors, we will assume that a set of m undirected graphs is available. By undirected graph we mean a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is the set of edges i.e. unordered pairs of distinct vertices. Each vertex is associated to an input variable so that we will identify $\mathcal{V}$ with $\mathbb{N}_p$. An edge between a and b is present ($a \sim b$) if the variables indexed by a and b are linked (in a sense that defines the nature of the network) and we denote by d_a the degree of a. Here each graph in the ensemble $\left\{\mathcal{G}^j = \left(\mathbb{N}_p, \mathcal{E}^j\right) \,:\, j \in \mathbb{N}_m\right\}$ is assumed to be a noisy version of the same latent graph $\mathcal{G}$. For the $j-$th graph this practically means that either a link between a and b is present when it should not (in symbols: $a \sim_j b \,\wedge\, a \not\sim b$) or vice-versa.

3 Grouping and Spectral properties of the Graph

We now recall some results connecting spectral properties of graphs with grouping. As highlighted in the introduction, we are interested in those cases where it makes sense to define macro-variables upon groups of functionally related covariates. The groups here are assumed to correspond to the connected components $c_l,\ l \in \mathbb{N}_s$ of a given graph $\mathcal{G}$ with *normalized graph Laplacian*

$$L_{ab} = \begin{cases} 1, & \text{if } a = b,\ d_a \neq 0 \\ -1/\sqrt{d_a d_b}, & \text{if } a \sim b \\ 0, & \text{otherwise .} \end{cases} \tag{1}$$

The network expresses our prior belief that the best model in the space of linear functions $f^\star$ should have a structured representation as

$$f^\star(x) = \sum_{l \in \mathbb{N}_s} \gamma^\star_l \underbrace{\sum_{a \in c_l} \sqrt{d_a} x_a}_{\text{macro variable } z_l} \tag{2}$$

for some $\gamma^\star \in \mathbb{R}^s$. In this model, variables are grouped together and within each group their contribution is further weighted according to their degree: highly connected nodes are expected to play a prominent role in the process. Define now t^l entry-wise as $t^l_a = \sqrt{d_a}$ if $a \in c_l$ and $t^l_a = 0$ otherwise. Recall that the null space of L is $\ker(L) = \{x \in \mathbb{R}^p \,:\, Lx = 0\}$ and that the number of connected components s corresponds to the number of the eigenvalues of L that are equal to zero [5]. If we assume that no isolated vertices exist then it is not difficult to see that $\left\{t^l \,:\, l \in \mathbb{N}_s\right\}$ form an orthogonal basis for $\ker(L)$. This in turns implies that any vector $\beta \in \ker(L)$ can be written as $\beta = \sum_{l \in \mathbb{N}_s} \gamma_l t^l$ for some $\gamma \in \mathbb{R}^s$ and the linear function $f(x) = \langle x, \beta \rangle$ admits a representation as in (2). We are about to use this property to penalize models which do not exhibit grouping.

4 Multiple NET

In this Section we outline the Multiple NET formulation. This optimization problem let us integrate multiple estimates of the latent network in the learning task. With reference to (2) the composite penalty that we consider further reflects the expectation that some of the original variables (and potentially entire macro-variables) might not be relevant.

4.1 Problem Formulation

We begin with associating a decision vector $\beta^j \in \mathbb{R}^p$ to each graph $\mathcal{G}^j$ in the ensemble. The vectors are collected in the *decision matrix* $B = (\beta^1, \ldots, \beta^m) \in \mathbb{R}^{p \times m}$. Our linear predictor model is now defined as $\left\langle x, \sum_{j \in \mathbb{N}_m} \beta^j \right\rangle$ and thus is obtained by averaging over the set of decision vectors. We let the empirical risk be $J(B) = \sum_{i \in \mathbb{N}_n} l\left(y_i, \left\langle x^i, \sum_{j \in \mathbb{N}_m} \beta^j \right\rangle\right)$ and introduce the mixed norm:

$$\|B\|_{2,1} = \sum_{a \in \mathbb{N}_p} \sqrt{\sum_{j \in \mathbb{N}_m} \left|\beta_a^j\right|^2} \tag{3}$$

that corresponds to the sum of the l_2 norm of the rows of B. For any $j \in \mathbb{N}_m$, let now L^j be the Laplacian associated to $\mathcal{G}^j$. We will consider the quadratic penalty:

$$\Omega_j\left(\beta^j\right) = \left\langle \beta^j, L^j \beta^j \right\rangle = \sum_{a \sim_j b} \left(\frac{\beta_a^j}{\sqrt{d_a^j}} - \frac{\beta_b^j}{\sqrt{d_b^j}} \right)^2 \tag{4}$$

where the last member can be derived via simple algebra (see also [5]). Finally, we let $\Omega(B) = \sum_{j \in \mathbb{N}_m} \Omega_j\left(\beta^j\right)$. Each quadratic penalty $\Omega_j\left(\beta^j\right)$ enforces smooth variation of coefficients associated to neighboring nodes in $\mathcal{G}^j$. Notice that if $\beta^j \in \ker\left(L^j\right)$ then we have $\Omega_j\left(\beta^j\right) = 0$. Thus, $\Omega_j\left(\beta^j\right)$ leaves unpenalized those linear functions that have a representation in terms of the connected components of the $j-$th graph, as in (2). We propose to find an estimate of the decision matrix by solving the following penalized empirical risk minimization problem for some $\lambda, \tau \in \mathbb{R}^+$:

$$\hat{B} = \arg\min \left\{ J(B) + \frac{\lambda}{2}\Omega(B) \ : \ \|B\|_{2,1} \leq \tau, B \in \mathbb{R}^{d \times m} \right\} . \tag{5}$$

By solving this convex problem, we perform at once model fitting and both variables selection and aggregation. On the one hand, the selection is achieved by constraining the mixed norm of the decision matrix in much the same way as the standard l_1 penalty is used for the selection in the LASSO, elastic-net and NET. On the other hand the aggregation is induced via the sum of quadratic penalties. Loosely speaking the procedure finds the subset of variables relevant for the specific task, while ensuring the agreement between different priors. We will elaborate more on this in the sequel.

Notice that the problem of incorporating grouping information has been addressed also elsewhere (besides [10] see also the approach by Composite Absolute Penalties in [14]). However, contrary to the previously proposed estimators, our approach provides a way to deal with the integration of multiple sources. This is of practical interest as multiple graphs are often available in applications (see Section 6 for a real life example).

4.2 Some Insight on the Selection Mechanism

We first remark that, when a unique network prior $\mathcal{G}^1$ is employed and the quadratic loss is used, the mixed norm boils down to the standard l_1 penalty and the entire approach coincides with NET [10]. If we further replace the Laplacian L^1 by the $p \times p$ identity matrix I_p in the penalty Ω_1, we retrieve the naïve elastic-net [15] (in the sequel referred to as e-NET). Notice that I_p is not a normalized graph Laplacian. However, if the number of nodes p goes to infinity and $\mathcal{G}^1$ is fully connected, then its Laplacian matrix L^1 tends towards I_p entry-wise. Correspondingly (4) tends to the ridge penalty, a fact that sheds some light on the uninformative prior imposed by the latter.

To provide intuition about the selection mechanism in (5), assume first that a single noisy graph is provided. Suppose that the nodes belonging to a certain connected component correspond to an

important macro-variable, which is to be included at the solution. Further assume that, due to one or more misspecified edges, a node a associated to an irrelevant variable erroneously belongs to the aforementioned component. As the quadratic penalty enforces smooth variation across nodes in the component, the coefficient associated to a could be pulled away from zero. Consider now the case where multiple graphs are available. At the solution $\hat{B}$, due to the mixed norm constraint, some rows with small contribution would shrink to zero exactly and hence all the coefficients in those rows would vanish simultaneously. Suppose the misspecified edges are randomly distributed across the networks. Then a is likely not to belong to the same connected component in most of the graphs. Hence in this scenario the $a-$th row will shrink to zero and the associated variable will not be selected.

5 Optimality Conditions and Algorithm

In this Section we first derive the optimality conditions for (5) and successively elaborate on a strategy to compute a solution. Our approach is suitable for large-scale problems and permits to compute efficiently the solution for increasing values of the parameter τ. We begin with shortening our notation by defining[2] $f_\lambda(B) = J(B) + \frac{\lambda}{2}\Omega(B)$ and $g_\tau(B) = \tau - \|B\|_{2,1}$. The Lagrangian for (5) now reads $L(B,\mu) = f_\lambda(B) - \mu g_\tau(B)$ where $\mu \geq 0$ denotes the Lagrange multiplier associated to the constraint $g_\tau(B) \geq 0$. Denote by $\partial_B g_\tau(B) \subset \mathbb{R}^{p\times m}$ the subdifferential of g_τ at B. For any $\mu \geq 0$, B minimizes $L(\cdot,\mu)$ if and only if there exists $V \in \partial_B g_\tau(B)$ such that

$$\nabla f_\lambda(B) - \mu V = 0 \tag{6}$$

where $\nabla f_\lambda(B)$ denotes the well defined gradient of f_λ at B. Define now the *active set* as:

$$\mathcal{A}(B) = \left\{ a \in \mathbb{N}_p \ : \ \sqrt{\sum_{j\in\mathbb{N}_m} \left|\beta_a^j\right|^2} \neq 0 \right\} .$$

By taking into account the closed form of $\partial_B g_\tau(B)$ it is not difficult to see that if B is a minimizer of $L(\cdot,\mu)$ and $r_\lambda(B,a) = \left(\partial_{\beta_a^1} f_\lambda(B), \ldots, \partial_{\beta_a^m} f_\lambda(B)\right)$ then (6) implies

$$\|r_\lambda(B,a)\|_2 \begin{cases} = & \mu, \quad \text{for } a \in \mathcal{A}(B) \\ \leq & \mu, \quad \text{for } a \notin \mathcal{A}(B) \end{cases} .$$

Thus, any minimizer B of $L(\cdot,\mu)$ must satisfy $\mu = \max_{a\in\mathbb{N}_p} \|r_\lambda(B,a)\|_2$ which relates the Lagrange multiplier μ to the entries of the gradient $\nabla f_\lambda(B)$.

5.1 Active Set Strategy

We exploit this fact and compute a global solution for (5) via an active set strategy similar to the one proposed for the LASSO in [12]. The same idea has also been exploited in [13] for the group-LASSO. The method is based on the following simple observation. Suppose that an algorithm has found the solution $\bar{B}$ for the restricted problem defined upon the current guess $\mathcal{A}$ as:

$$\min \left\{ f_\lambda(B) \ : \ g_\tau(B) \geq 0, \ B \in \mathbb{R}^{p\times m} \text{ s.t. } \beta_a^j = 0 \ \forall a \notin \mathcal{A}, \ \forall j \in \mathbb{N}_m \right\} . \tag{7}$$

If $\bar{\mu}$ is the Lagrange multiplier associated to $g_\tau\left(\bar{B}\right)$ then we can test the optimality of $\bar{B}$ for the original problem by verifying whether $\bar{\mu} \geq \left\|r_\lambda\left(\bar{B},a\right)\right\|_2$ for all $a \in \mathbb{N}_p$. In Table 1 we sketch the full active set approach that, in turns, makes use of the algorithm presented in the next Subsection. We remark that, although the restricted problem was stated in terms of the full decision matrix, finding a solution of (7) involves only those rows indexed by elements of $\mathcal{A}$. Access to the full matrix is only required to check the optimality of $\bar{B}$ for the full problem.

5.2 Optimal First Order Scheme For Large-scale Problems

It remains to illustrate how a solution for the restricted problem (7) can be obtained. In the context of the group-LASSO, it was proposed in [8] to seek for a solution via a gradient projection strategy

[2] We have used a subscript here to emphasize the dependency upon the parameters $\lambda, \ t$.

Active Set Algorithm $\mathcal{AS}(B_0, \mathcal{A}_0, \lambda, t)$

Initialization: $k = 1$.

REPEAT Call $\mathcal{AC}(B_{k-1}, \mathcal{A}_{k-1}, \lambda, t)$ to get B_k and $\mathcal{A}_k$. Set $\mu = \max\{\|r_\lambda(B_k, a)\| \ : \ a \in \mathcal{A}_k\}$.
IF $\max\{\|r_\lambda(B_k, a)\| \ : \ a \notin \mathcal{A}_k\} > \mu$ THEN Set $\mathcal{A}_k = \mathcal{A}_k \cup \arg\max\{\|r_\lambda(B_k, a)\| \ : \ a \notin \mathcal{A}_k\}$.
UNTIL $\max\{\|r_\lambda(B_k, a)\| \ : \ a \notin \mathcal{A}_k\} \leq \mu$.

Table 1: Active set method.

[2]. Adapted to our setting, this idea would prescribe to iterate between taking a gradient step via $B^+ = B - \frac{1}{\gamma}\nabla f_\lambda(B)$, for a suitable $\gamma \in \mathbb{R}^+$, and projecting B^+ onto the feasible set $g_\tau(B) \geq 0$. As for the group-LASSO, the latter step can be computed efficiently by exploiting the specific geometry of our problem. Although appealing due to the low complexity of each iteration, an algorithm of this type would have a rate of convergence of $O\left(\frac{1}{k}\right)$ where k represents the number of iterations. Instead, for the solution of (7) we adopt an accelerated scheme that provably achieves a convergence rate of $O\left(\frac{1}{k^2}\right)$, optimal in the class of first order methods [11]. Our first order optimal scheme is reported in Table 2. It encompasses the problem-specific procedure necessary to compute the projection. The reader is referred to Section 2.2.4 of [11] for an account on general optimal schemes. We conclude by remarking that, in the algorithm, γ should be set equal to the Lipschitz constant of ∇f_λ. Depending on the loss function used in f_λ, the Lipschitz constant can be generally computed in closed form as a function of the data. This is the case, in particular, for quadratic loss and logistic regression.

Accelerated First Order Algorithm $\mathcal{AC}(B_0, \mathcal{A}_0, \lambda, t)$

Initialization: $\alpha_0 \in (0, 1)$, $Y_0 = B_0$. Set γ equal to the Lipschitz constant of ∇f_λ.

Iteration $k \geq 0$

I. Compute $\nabla f_\lambda(Y_k)$. Set $\tilde{Y} = Y_k - \frac{1}{\gamma}\nabla f_\lambda(Y_k)$ and $\tilde{\mathcal{A}} = \mathcal{A}_k$.

REPEAT

II. a. For $a \in \tilde{\mathcal{A}}$ set $M_a = \sqrt{\sum_{j \in \mathbb{N}_m} \tilde{Y}_{aj}^2} + \frac{1}{|\tilde{\mathcal{A}}|}\left(t - \sum_{i \in \tilde{\mathcal{A}}} \sqrt{\sum_{j \in \mathbb{N}_m} \tilde{Y}_{ij}^2}\right)$.

b. IF $M_a < 0$ for some $a \in \tilde{\mathcal{A}}$ THEN Set $\tilde{\mathcal{A}} = \{a \ : \ M_a > 0\}$.

UNTIL $M_a \geq 0$ for all $a \in \tilde{\mathcal{A}}$.

III. Set $\mathcal{A}_{k+1} = \tilde{\mathcal{A}}$.

IV. For $j \in \mathbb{N}_m$ set $(B_{k+1})_{aj} = M_a \tilde{Y}_{aj} / \sqrt{\sum_{j \in \mathbb{N}_m} \tilde{Y}_{aj}^2}$ if $a \in \mathcal{A}_{k+1}$ and $(B_{k+1})_{aj} = 0$ if $a \notin \mathcal{A}_{k+1}$.

V. Set $\alpha_{k+1} = \frac{1}{2}\left(-\alpha_k^2 + \sqrt{\alpha_k^4 + 4\alpha_k^2}\right)$ and $\beta = (\alpha_k(1 - \alpha_k)) \ / \left(\alpha_k^2 + \alpha_{k+1}\right)$.

VI. Set $Y_{k+1} = B_{k+1} + \beta\left(B_{k+1} - B_k\right)$.

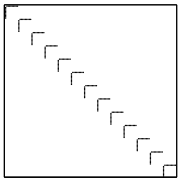
Table 2: The algorithm to compute a solution for the restricted problem (7).

6 Experiments

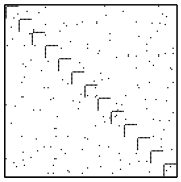
This final Section is devoted to demonstrate the advantage of combining multiple graph priors. In all cases we compute solutions corresponding to a grid of values for $\lambda, \tau \in \mathbb{R}^+$. In particular, for different fixed values of λ, we approximate the solution path in τ on a tight grid of values $0 < \tau_{(0)} < \tau_{(1)} < \cdots < \tau_{(N)}$. This can be done efficiently by warm starting: given the solution corresponding to $\tau_{(i)}$ only few further iterations in $\mathcal{AS}\left(B_{(i)}, \mathcal{A}_{(i)}, \lambda, \tau_{(i+1)}\right)$ are typically required.

Synthetic Example As a first example, we simulated a latent graph with a block-wise star topology, representative of Transcription Factor (TF, in the following) Networks. We considered $p = 1000$ nodes and $s = 100$ connected components with identical structure. Each component has vertices $a_1, \ldots, a_{10}$ with edges connecting a_1 to $a_2, \ldots, a_{10}$ (Figure 1(a)). We generated correlated input data by letting $x \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{p \times p}$ and $\Sigma_{ab} = 0.5^{|a-b|}$. With reference to (2), we set $y = 5z_1 + 10z_2 + 10z_3 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and σ was set so that the signal to noise ratio is 6. Finally, we generated multiple graph priors with edges having a probability 0.01 to be misspecified, see Figure 1(b). We compared performances of ridge regression (RR), LASSO, e-NET, NET (1 graph prior) and m-NET (10 graph priors). We also considered true-NET corresponding to the use of the true network as a single graph prior. We trained the models with quadratic loss with respectively $n = 50$, $n = 100$ and $n = 200$ input-output pairs. In all the cases the values for

λ and τ were selected based upon a hold-out sample of 400 pairs. We measured the accuracy[3] in terms of root mean square error (RMSE) computed over 1000 test pairs. We ran 30 simulations and report in Table 3 the average RMSE with standard errors (in parentheses). The best performance - disregarding true-NET which is expectedly superior - is highlighted. Even with very low probability of misspecification the advantage of combining sources is sensible.



(a) True network



(b) A noisy graph prior

Test Accuracy (RMSE)

	$n = 50$	$n = 100$	$n = 200$
RR	$85.42(2.20)$	$81.71(2.45)$	$75.00(2.02)$
LASSO	$79.51(6.06)$	$62.11(5.82)$	$48.34(2.40)$
e-NET	$78.51(5.69)$	$62.09(5.80)$	$49.34(2.41)$
NET	$77.89(5.64)$	$61.94(5.86)$	$49.30(2.42)$
m-NET	$\mathbf{77.57(5.50)}$	$\mathbf{60.66(6.10)}$	$\mathbf{46.37(2.83)}$
true-NET	$66.85(6.37)$	$54.63(4.25)$	$44.58(1.42)$

Figure 1 & Table 3: A part of adjacency matrices and accuracy for the synthetic example.

Real Life Example As real-life case study, we considered a microarray data set on advanced stage serous ovarian cancer [1]. The binarized survival status was available for 53 patients: survival of 29 patients did not exceed 3 years, while 24 patients survived longer than 7 years. As input data we considered, per patient, the expression profiles of 3000 genes. Graph priors were here TF Networks that are believed to play a role in human cancers. The two graphs considered were obtained based on (I) miRNAmap [7] and (II) microRNA.org [3]. They were considered both independently and combined via m-NET. We used the logistic loss and tried to predict the binarized survival status. The accuracy was measured via $5-$fold cross-validation estimate of the area under the ROC curve. We report performance averaged over 100 randomized simulations (CV-AUC, Table 4).

ROC performance (CV-AUC)

LASSO	e-NET	NET(I)	NET(II)	m-NET
$0.76(0.05)$	$0.78(0.04)$	$0.80(0.05)$	$0.80(0.05)$	$\mathbf{0.82(0.05)}$

Table 4: Accuracy for the real-life example.

Acknowledgments

Research supported by Research Council KUL: GOA AMBioRICS, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM; Flemish Government: FWO: PhD/postdoc grants, G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0302.07 (SVM/Kernel); IWT: PhD Grants, McKnow-E, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, POM Belgian Federal Science Policy Office: IUAP P6/04.

References

[1] A. Berchuck *et al.* Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clinical cancer research*, 11(10):3686–3696, 2005.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[3] D. Betel *et al.* The microRNA. org resource: targets and expression. *Nucleic acids research*, 36:D149–D153, 2007.

[4] H. Binder and M. Schumacher. Comment on 'network-constrained regularization and variable selection for analysis of genomic data'. *Bioinformatics*, 24(21):2566–2568, 2008.

[5] F.R.K. Chung. *Spectral graph theory*. Amer Mathematical Society, 1997.

[6] N. El Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. of Stat.*, 36(6):2717–2756, 2008.

[7] S.D. Hsu *et al.* miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, 36:D165–D169, 2008.

[8] Y. Kim *et al.* Blockwise sparse regression. *Statistica Sinica*, 16(2):375, 2006.

[9] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[10] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[11] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Pub, 2003.

[12] M.R. Osborne *et al.* On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[13] V. Roth and B. Fischer. The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM New York, NY, USA, 2008.

[14] P. Zhao *et al.* Grouped and hierarchical model selection through composite absolute penalties. *Ann. of Stat.*, 37:3468–3497, 2009.

[15] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B*, 67(2):301–320, 2005.

[3] We measured the performance of the optimized models, we did not apply any correction for double shrinkage (see [15] for details about double shrinkage correction).