Joint estimation of transcription networks and expression in cancer

José Sánchez[1] **and Tatjana Pavlenko**[2] **and Rebecka Jörnsten**[1]

[1] Chalmers University of Technology and University of Gothenburg, Sweden
[2] Royal Institute of Technology, Sweden

The study of cancer at the molecular level has increased our understanding of these diseases in recent years. Yet, analysis of different cancer classes as well as integration of data types remains a sizable challenge. Network modelling has been used to construct mechanistic hypotheses and successfully derive predictions of potential drug targets or survival outcome [1,4]. Nevertheless, few attempts have been made to generate comprehensive and interpretable network models of multiple cancers [3], or to incorporate sparse estimates of the expression levels [7].

Here we propose to integrate and extend [3,7] into a joint framework to investigate a multiclass sparse inverse covariance (precision) matrix and mean vector estimation method. Our goal is to obtain estimates for transcription networks (given by the inverse covariance matrix) as well as expression levels (given by the mean vectors) for different cancer classes. To this end we propose to maximize the corresponding penalized likelihood functions using ADMM (Alternating Directions Method of Multipliers, [2]). More precisely, suppose that we have data sets $X^k \sim N(\mu^k, \Sigma^k)$, $\mu^k \in \mathbb{R}^p$, $\Sigma^k \in \mathbb{R}^{p\times p}$, for $k = 1, 2, \ldots, K$ (cancer) classes. Let $\{\mu^k\}$ and $\{\Omega^k\}$ denote the set of mean vectors and precision matrices for the K classes. The penalized log-likelihood we aim to maximize is:

$$l(\{\mu^k\}, \{\Omega^k\}) = \sum_{k=1}^{K} -\ln(|\Omega^k|) + \mathrm{tr}(\Omega^k S^k) + (\bar{x}^k - \mu^k)^T \Omega^k (\bar{x}^k - \mu^k)$$

$$+ \lambda_1 \sum_{k=1}^{K} \sum_{i} |\mu_i^k| + \lambda_2 \sum_{k<k'} \sum_{i} |\mu_i^k - \mu_i^{k'}| + \lambda_3 \sum_{k=1}^{K} \sum_{i,j} |\omega_{ij}| + \lambda_4 \sum_{k<k'} \sum_{i,j} |\omega_{ij}^k - \omega_{ij}^{k'}|,$$

which includes a *lasso* term to encourage sparsity, and a *fused* term [3] to penalize differential estimates. To estimate $\{\mu^k\}$ we optimize the profile log-likelihood function when an estimate of $\{\Omega^k\}$ is known. Similarly, to estimate $\{\Omega^k\}$ we optimize the profile log-likelihood function when an estimated of $\{\mu^k\}$ is known. In practice, given some initial values, we alternate between these two estimation procedures for $\{\mu^k\}$ and $\{\Omega^k\}$ until convergence. Each profile log-likelihood is maximized using methods that build on [6]; (i)class specific sample size corrections and (ii)a novel bootstrap procedure for estimating robust sparse and fused structures.

We perform extensive simulation studies to investigate the performance of our method. We construct mean vectors and the precision matrices with components that are either common or differential across classes, and show that our method can accurately find the true sparsity and fusing structure (of both mean vectors and precision matrices).

Since our framework is that of Gaussian graphical models, with a minor additional constraint on the precision matrices to be block diagonal, we can recast the problem into an ensemble classifier, where each block takes on the format of a linear [5] or a quadratic component of the corresponding discriminant function. Following [5], the contribution of each model component to the discriminatory power provides insight into the differences between cancer classes.

Finally, we apply our method to expression data for glioblastoma, breast and ovarian cancer from the Cancer Genome Atlas (TCGA). Results are incorporated into our analysis web tool Cancer Landscapes (available at **www.cancerlandscapes.org**) which allows biologists and other scientists to further examine the properties of the estimates and their linear and quadratic components, i.e. model components that distinguish between cancer classes based on differential expression levels only versus components differential in network structure.

Keywords: Inverse covariance estimation, lasso, fused lasso, high dimension, cancer.

References:

[**1**] Adler, A. S., Lin, M., Horlings, H., Nuyten, D. S., van de Vijver, M. J., and Chang, H. Y. (2006) Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet* 38(4), 421–30.

[**2**] Boyd, S., Parikh, N., Chu, E., Pelato, B., and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* 1–122.

[**3**] Danaher, P., Wang, P., and Witten, D. (2013) The joint graphical lasso for inverse covariance estimation across multiple classes. *Royal Stat Soc* 76(2) 373–397.

[**4**] Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol Syst Biol* 7, 486

[**5**] Pavlenko, T. Björkström, A., and Tillander, A. (2012) Covariance structure approximation via gLasso in high-dimensional supervised classification. *Applied Stat* 39(8) 1643-1666.

[**6**] Sánchez, J. (2013). Comparative network analysis of human cancer: sparse graphical models with modular constraints and sample size correction. *Technical report and liceciate thesis. Chalmers University of Technology.*

[**7**] Witten, D., Tibshirani, R. (2011) Penalized Classification Using Fisher's Linear Discriminant. *Royal Stat Soc Ser B* 73(5) 753–772.