Bioinformatics, 34(5), 2018, 896–898
doi: 10.1093/bioinformatics/btx677
Advance Access Publication Date: 25 October 2017
Applications Note

OXFORD

Data and text mining

netReg: network-regularized linear models for biological association studies

Simon Dirmeier[1,*], Christiane Fuchs[2,3], Nikola S. Mueller[2] and Fabian J. Theis[2,3]

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland, [2]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany and [3]Department of Mathematics, Technische Universität München, 85748 Garching, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 4, 2017; revised on September 28, 2017; editorial decision on October 18, 2017; accepted on October 24, 2017

Abstract

Summary: Modelling biological associations or dependencies using linear regression is often complicated when the analyzed data-sets are high-dimensional and less observations than variables are available ($n \ll p$). For genomic data-sets penalized regression methods have been applied settling this issue. Recently proposed regression models utilize prior knowledge on dependencies, e.g. in the form of graphs, arguing that this information will lead to more reliable estimates for regression coefficients. However, none of the proposed models for multivariate genomic response variables have been implemented as a computationally efficient, freely available library. In this paper we propose netReg, a package for graph-penalized regression models that use large networks and thousands of variables. netReg incorporates a priori generated biological graph information into linear models yielding sparse or smooth solutions for regression coefficients.

Availability and implementation: netReg is implemented as both R-package and C++ commandline tool. The main computations are done in C++, where we use Armadillo for fast matrix calculations and Dlib for optimization. The R package is freely available on *Bioconductor* https://bioconductor.org/packages/netReg. The command line tool can be installed using the conda channel *Bioconda*. Installation details, issue reports, development versions, documentation and tutorials for the R and C++ versions and the R package vignette can be found on GitHub https://dirmeier.github.io/netReg/. The GitHub page also contains code for benchmarking and example datasets used in this paper.

Contact: simon.dirmeier@bsse.ethz.ch

1 Introduction

The advent of high-throughput genomic methods provided a wealth of novel biological data that allow the interpretation of previously scarcely researched genetic and epigenetic processes. Many experiments aim at establishing statistical dependencies between two different data-sets, for example finding genotype-phenotype associations such as eQTL mappings or medical cohort studies. Linear regression models are attractive for these kinds of problems since they explicitly describe the impact of a predictor onto a response. However, for problems where the number of predictors p is larger than the number of observations n unique analytical solutions for the parameters do not exist. For example in eQTL-mapping studies typically hundreds of SNPs are genotyped, but only few observations are available. Solutions for these settings have already been proposed, e.g. by Tibshirani (1996) or Zou and Hastie (2005), where penalization terms are introduced to the likelihood of the model. Recent studies suggest to incorporate further penalization terms, for example in the form of graph prior knowledge, arguing that variables may be structured and related variables might have a common or at least similar effect. For this either the regressors or regressands are mapped to biological networks. Two nodes are

connected if the variables have some biological relationship, for instance co-expression or a protein-protein interaction. Regardless of the graph used, the rationale is that biological processes might be regulated by two neighboring genes rather than by two genes far apart. With these prior networks the biological relations are directly incorporated in the objective function of the model. Consequently a better model goodness-of-fit can be achieved. Examples for network-regularized regression models include Li and Li (2008, 2010), Kim (2013) or Veríssimo (2016). Conceptually network-regularization differs from other network-based approaches, such as network enrichment (Alcaraz, 2011; Alexeyenko, 2012) or correlation analysis (Langfelder and Horvath, 2008), by making inference on the parameters of the regression of a set of dependent variables on a set of predictors, and not making inference on significant or correlated modules in a network itself. Although network-regularized models for *univariate* responses have already been efficiently implemented in R (Li, 2015a,b; Zhao, 2016), this is to our knowledge not the case for *multivariate* response variables which are however predominant in genomic studies. For many of the multivariate regression models proposed in literature, the respective software either lacks appropriate documentation, making the methods hardly usable, or the code does not compile, or in the worst case implementations are not available at all.

To our knowledge multivariate network-regularized linear models have so far not been implemented in an efficient computational framework that makes the proposed methodology usable in practice. Thus, in this paper we propose netReg, an R/C++-package that implements multivariate network-regression models, i.e. linear models with graph-penalized likelihoods. With netReg it is possible to fit linear models that employ large dense networks and use thousands of covariables. We hope to establish a common framework with implementations of different network-regularized regression models and by that unify the already proposed methodology into one easily usable, maintained software package. This should benefit the biological as well as the statistical community.

2 Materials and methods

2.1 Model

Multivariate linear regression models describe a dependency $f : \mathcal{X} \to \mathcal{Y}$ for a data-set $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ of n observations. Every $\mathbf{x}_i$ is a p-dimensional covariable (or feature) vector and every $\mathbf{y}_i$ is a q-dimensional response vector. For scenarios where $n \ll p$, solutions for the coefficients are, however, not unique. An attractive solution is to add an ℓ_1-penalty to the likelihood of the model yielding a sparse solution for the regression coefficients. In order to include biological graph-prior knowledge the same procedure can be applied, i.e. by extending the likelihood with penalization terms. netReg implements a modified regularization term proposed by Cheng (2014). Two prior graphs for the response and design matrices are included into an ℓ_1-regularized likelihood as:

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}} \frac{1}{2}||\mathbf{Y} - \mathbf{XB}||_2^2 + \lambda||\mathbf{B}||_1 + \frac{\phi_1}{2}\sum_{i=1}^{p}\sum_{j=1}^{q}||\boldsymbol{\beta}_{i,*} - \boldsymbol{\beta}_{j,*}||_2^2(\mathbf{S}_0)_{i,j} + \frac{\phi_2}{2}\sum_{i=1}^{p}\sum_{j=1}^{q}||\boldsymbol{\beta}_{*,i} - \boldsymbol{\beta}_{*,j}||_2^2(\mathbf{T}_0)_{i,j}, \quad (1)$$

where $\mathbf{B}$ is the matrix of coefficients, $||\mathbf{X}||_2^2$ is the squared ℓ_2-norm and $||\boldsymbol{X}||_1$ the ℓ_1-norm. Vectors $\boldsymbol{\beta}_{i,*}$ and $\boldsymbol{\beta}_{*,i}$ are the ith row or column of $\mathbf{B}$, respectively. λ, ϕ_1 and ϕ_2 are known shrinkage parameters. $\mathbf{S}_0$ and $\mathbf{T}_0$ are two non-negative adjacency matrices for $\mathbf{X}$ and $\mathbf{Y}$, encoding a biological similarity measure as described above. The prior graphs can be generated subjectively, i.e. reflecting a personal belief, from online databases or be directly estimated from a biological data-set.

Table 1. Timings of a pure R versus netReg implementation

	$n = p = 100$	$n = p = 1000$	$n = p = 10\,000$
R	2009 ms	578 s	> 3 d
netReg	25 ms	12 s	2.5 h

Note: For each setting measurements are averaged over 10 runs with $q = 10$ response variables.

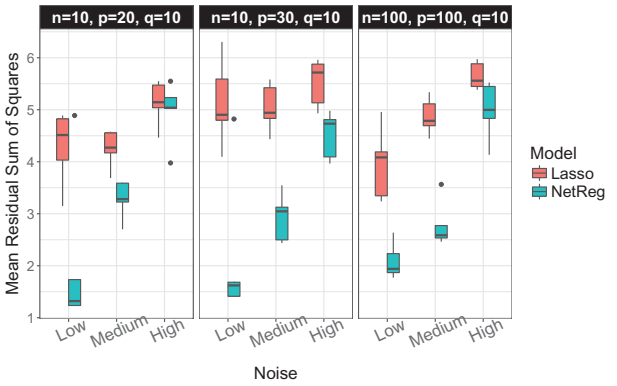

Fig. 1. Mean residual sum of squares for LASSO versus netReg [Equation (1)]. netReg outperforms the LASSO for different levels of number of observations n, covariables p and different Gaussian noise with mean 0 and variance $\sigma^2 \in \{1, 2, 5\}$ (low, medium, high) consistently. Boxes show 25, 50 and 75% quantiles

2.2 Implementation

We implemented the proposed models from Equation (1) as a freely available package written in the R and C++ programming languages. For the estimation of coefficients $\widehat{\mathbf{B}}$ we use *cyclic coordinate descent* that has recently been described elsewhere (Friedman, 2007, 2010). Since linear models require extensive computation of costly matrix multiplications, netReg uses Armadillo (Sanderson, 2010). Armadillo uses an OpenBLAS (Xianyi, 2012) or BLAS, and Lapack backend for efficient vectorized matrix-algebra, that, for modern computer architectures, enables multiple floating point operations per register. Table 1 shows the absolute speed-ups of our implementation versus a pure R implementation. netReg is considerably faster than the alternate implementation.

2.3 Model selection

In order to select the optimal shrinkage parameters λ, ϕ_1 and ϕ_2 we use the BOBYQA-algorithm (Powell, 2009), a gradient-free convex optimization method, implemented in Dlib-Ml (King, 2009). To assess the current set of shrinkage parameters we apply 10-fold cross-validation. The mean residual sum of squares of 10 cross-validation runs is computed and used as minimization criterion for the BOBYQA algorithm yielding an optimal solution for the shrinkage parameters λ, ϕ_1 and ϕ_2.

2.4 Application

Figure 1 shows a benchmark of the LASSO (ℓ_1-penalization) versus network-based regularization [Equation (1)]. For variable number of observations n, covariables p and noise variance σ^2 the network-based regularization outperforms the LASSO consistently. Due to the integration of biological prior graphs, the mean sum of errors is considerably lower than in the version that uses ℓ_1-penalization only.

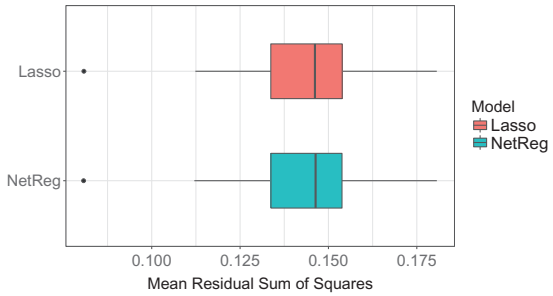

Fig. 2. Mean residual sum of squares for LASSO versus netReg [Equation (1)]. netReg and the LASSO have similar estimates for coefficients

We applied netReg on a yeast eQTL data-set of 112 yeast segregants [Brem (2005); Storey (2005); similarly to Cheng (2014)]. The filtered dataset consists of 112 observations of 500 genetic markers and 231 expression profiles. We excluded genes that had a node degree of less than 10 from a yeast protein-protein interaction network (BioGRID https://thebiogrid.org/). Figure 2 shows that the LASSO and netReg have almost identical estimates for the coefficients. This either means that the graph was non-informative or the mapping from SNPs to eQTLs contains little signal. In either case the model selection converges to the same result such that the netReg solution can only improve model fits but not worsen them.

3 Outlook

So far the library implements a single graph-regularized likelihood for linear models with normally distributed responses. Next versions of the package will include models for binomial, Poisson or categorial variables and Cox-proportional hazard models; and other proposed regularizations, such as in Li and Li (2010) and Kim (2013). Furthermore, so far netReg excels on large, dense networks with high node degrees. For sparse (scale-free) matrices, as they are common in biology, speedups can be gained by working with adjacency lists instead of full graphs.

Acknowledgements

The authors thank David Seifert for fruitful discussions.

Conflict of Interest: none declared.

References

Alcaraz,N. *et al.* (2011) KeyPathwayMiner: detecting case-specific biological pathways using expression data. *Internet Math.*, **7**, 299–313.

Alexeyenko,A. *et al.* (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinf.*, **13**, 226.

Brem,R.B. *et al.* (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, **436**, 701.

Cheng,W. *et al.* (2014) Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics*, **30**, i139–i148.

Friedman,J. *et al.* (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**, 302–332.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1.

Kim,S. *et al.* (2013) Network-based penalized regression with application to genomic data. *Biometrics*, **69**, 582–593.

King,D.E. (2009) Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, **10**, 1755–1758.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.*, **9**, 559.

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

Li,C. and Li,H. (2010) Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.*, **4**, 1498.

Li,X. *et al.* (2015a) ADMMnet: Regularized Model with Selecting the Number of Non-Zeros. *R package version 0.1*.

Li,X. *et al.* (2015b) Coxnet: Regularized Cox Model. *R package version 0.2*.

Powell,M.J. (2009) The BOBYQA algorithm for bound constrained optimization without derivatives. In: *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge.

Sanderson,C. (2010) Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments.

Storey,J.D. *et al.* (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.*, **3**, e267.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Veríssimo,A. *et al.* (2016) DegreeCox – a network-based regularization method for survival analysis. *BMC Bioinf.*, **17**, 449.

Xianyi,Z. *et al.* (2012) Model-driven level 3 BLAS performance optimization on Loongson 3A processor. In: *2012 IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, pp. 684–691.

Zhao,S. (2016) Grace: Graph-Constrained Estimation and Hypothesis Tests. *R package version 0.5.3*.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.