BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 214–221
doi:10.1093/bioinformatics/btr641

Gene expression Advance Access publication November 21, 2011

Identification of context-specific gene regulatory networks with GEMULA—gene expression modeling using LAsso

Geert Geeven[1,*], Ronald E. van Kesteren[2], August B. Smit[2] and Mathisca C. M. de Gunst[1]

[1]Department of Mathematics, Faculty of Sciences and [2]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Gene regulatory networks, in which edges between nodes describe interactions between transcriptional regulators and their target genes, determine the coordinated spatiotemporal expression of genes. Especially in higher organisms, context-specific combinatorial regulation by transcription factors (TFs) is believed to determine cellular states and fates. TF–target gene interactions can be studied using high-throughput techniques such as ChIP-chip or ChIP-Seq. These experiments are time and cost intensive, and further limited by, for instance, availability of high affinity TF antibodies. Hence, there is a practical need for methods that can predict TF–TF and TF–target gene interactions *in silico*, i.e. from gene expression and DNA sequence data alone. We propose GEMULA, a novel approach based on linear models to predict TF–gene expression associations and TF–TF interactions from experimental data. GEMULA is based on linear models, fast and considers a wide range of biologically plausible models that describe gene expression data as a function of predicted TF binding to gene promoters.

Results: We show that models inferred with GEMULA are able to explain roughly 70% of the observed variation in gene expression in the yeast heat shock response. The functional relevance of the inferred TF–TF interactions in these models are validated by different sources of independent experimental evidence. We also have applied GEMULA to an *in vitro* model of neuronal outgrowth. Our findings confirm existing knowledge on gene regulatory interactions underlying neuronal outgrowth, but importantly also generate new insights into the temporal dynamics of this gene regulatory network that can now be addressed experimentally.

Availability: The GEMULA R-package is available from http://www.few.vu.nl/~degunst/gemula_1.0.tar.gz.

Contact: g.geeven@hubrecht.eu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 27, 2011; revised on November 7, 2011; accepted on November 15, 2011

1 INTRODUCTION

Cell-type- and condition-specific interactions between transcriptional regulators and their target genes are a primary mechanism for cells to accomplish spatiotemporal changes in gene expression. Identification of such interactions is an important step in modeling transcriptional regulatory networks. Regression methods are valuable tools that can be used to address three important issues in gene regulatory network building. First, they allow identification of transcription factors (TFs) and synergistic interactions between TFs that determine observed variation in gene expression. When two TFs commonly regulate a set of target genes, the synergistic effect of the TFs on target gene expression may not be just simply the sum of the individual effects. By fitting regression models that contain interaction terms that represent synergy between TFs, and comparing these to simpler models without such interactions, regulatory interactions between TFs can be inferred. Second, regression models provide answers to the question of *how much* of the observed variation in a single gene expression condition of interest can be predicted or explained based on regulatory motifs occurring in gene promoters. Since transcriptional gene regulation is time dependent, it is of interest to get some quantitative measure of how much of the observed variation in a specific biological context can be attributed to TF activity. Finally, regression models can provide insight into spatially and temporally dynamic interactions between TFs and target genes by comparing models inferred from the sets of genes regulated at successive time-points and/or under different experimental conditions.

Pioneering work on constructing linear regression models for gene expression analysis was performed by Bussemaker *et al.* (2001). Das *et al.* (2004) subsequently suggested a non-parametric regression approach that uses Multivariate Adaptive Regression Splines [MARS, Friedman (1991)]. They demonstrated the importance of generating and selecting among competing candidate models in a systematic way in order to model interactions between synergistic TFs. Model selection, however, is a major challenge in gene expression analysis given the large number of potential model terms. When interactions between predictors are considered, the number of possible candidate predictor terms p is relatively large compared with the sample size n (typically $p \approx n$ or even $p > n$). To select and fit models that appropriately trade-off bias and variance and that do not suffer from substantial overfit is not easy in this context. If we assume that only a small subset of the candidate predictors $s << p$ is really associated to the response of interest, we can restrict ourselves to sparse models by considering penalized regression methods such as the lasso. Statistical methods that generate sparse models have already successfully been applied

[*] To whom correspondence should be addressed.

in related but different contexts. For instance, Menéndez *et al.* (2010) used the graphical lasso to infer regulatory relationships from multifactorial perturbations, and Krämer *et al.* (2009) compared different regularization methods for estimating large-scale gene association networks modeled using graphical Gaussian models. In this article, we present a new approach to identify networks of TF–target gene and TF–TF interactions that underlie variation in gene expression instead of gene–gene association or perturbation networks.

We propose GEMULA, a method based on linear models that is fast, and considers a wide range of biologically plausible models. GEMULA is a four-stage method based on the lasso (Tibshirani, 1994) and *post hoc* re-sampling (Meinshausen and Bühlmann, 2010) that can be used to identify and prioritize synergistic interactions between predictors that underlie observed variation in gene expression. On yeast data, we compare models inferred by GEMULA and MARS with different predictors by evaluating the amount of variation that can be explained *across genes*. Moreover, we demonstrate that prioritization identifies biologically important TF–TF interactions that are supported by independent sources of experimental evidence. Next, we analyzed a time course dataset of cultured neuronal cells profiled at several time-points following induction of axon growth. This identifies context-specific gene regulatory networks within the complexity of the mammalian genome. We confirm existing relationships between TFs and growth-associated gene expression, and we generate new insights into the temporal dynamics of the regulatory network underlying axon growth.

2 MODEL

2.1 GEMULA

Figure 1 contains a flowchart for GEMULA. DNA promoter sequences and gene expression for a set of n genes are assumed to be given, together with a collection of p TF binding motifs. Prior to model building, we generate $X_1,\dots,X_p$, where X_{ij} represents the *in silico* predicted affinity of TF j to bind the promoter of gene i. We then use penalized regression to fit and evaluate models that relate the observed variation in gene expression to predicted binding of TFs. These models may include interaction terms to represent the combinatorial effect of TFs on expression. An additional feature of our method is the implementation of a stability selection procedure to prioritize terms in the fitted models according to relevance. Fitted models typically contain many terms and this allows a principled selection of the most relevant terms. Below, we describe the regression model, model selection and prioritization steps in detail.

2.1.1 The regression model Let $Y=(Y_1,\dots,Y_n)$ be a response variable, that represents some gene expression response of interest for a set of n genes, and $X_1,\dots,X_p$, all vectors of length n, a set of p predictor variables. We assume that Y and $X_1,\dots,X_p$ are related through the following regression model

$$Y=\mathbf{X}\beta+\epsilon, \tag{1}$$

where

$$\mathbf{X}=[\mathbf{1}\quad Z_1\cdots Z_d],$$

is an unknown $n\times(d+1)$ design matrix with columns $Z_j=f_j(X_1,\dots,X_p)$, for $j=1,\dots,d$ and polynomial functions $f_1,\dots,f_d$,

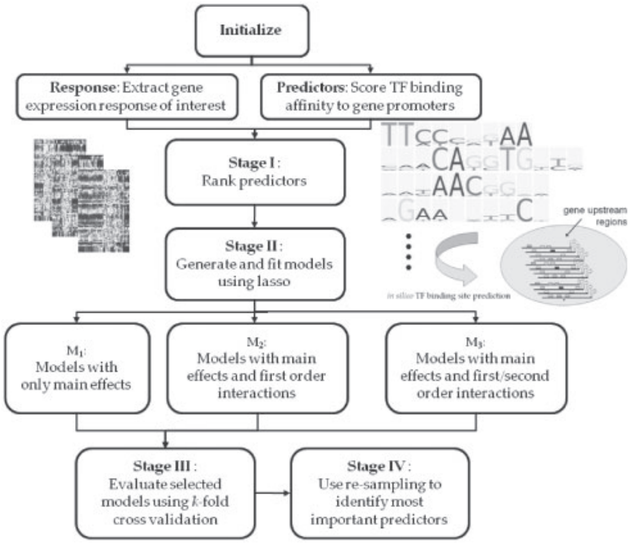
Initialize
Response: Extract gene expression response of interest
Predictors: Score TF binding affinity to gene promoters
Stage I: Rank predictors
Stage II: Generate and fit models using lasso
gene upstream regions
in silico TF binding site prediction
M1: Models with only main effects
M2: Models with main effects and first order interactions
M3: Models with main effects and first/second order interactions
Stage III: Evaluate selected models using k-fold cross validation
Stage IV: Use re-sampling to identify most important predictors

Fig. 1. Outline of the proposed method. GEMULA is a four-staged method that uses linear models to identify TF–gene and TF–TF interactions that are associated to observed variation in gene expression.

$\beta=(\beta_0,\dots,\beta_d)$ is an unknown vector of regression parameters and $\epsilon=(\epsilon_1,\dots,\epsilon_n)\sim\mathcal{N}(\mathbf{0},\sigma^2 I_n)$. GEMULA evaluates and compares many different candidate regression models M with Y as response and different subsets of the predictor variables. We identify different models M by their corresponding design matrices $\mathbf{X}_M$, where

$$\mathbf{X}_M=[\mathbf{1}\quad Z_1^M\cdots Z_{d_M}^M],$$

$Z_j^M=f_j^M(X_1,\dots,X_p)$ and $f_1^M,\dots,f_{d_M}^M$ polynomial functions. Hence, M consists of d_M candidate terms, which are polynomial functions of the input predictor variables. We let M_0 denote the model with design matrix $\mathbf{X}_{M_0}=[\mathbf{1}\quad X_1\cdots X_p]$. Given Y and $\mathbf{X}_M$, GEMULA repeatedly uses the lasso (Tibshirani, 1994) to select the predictors in model M and to estimate β_M. For $t\in\mathbb{R}^+$, the lasso estimate of β_M is determined by

$$\min_{\beta_M}\sum_{i=1}^{n}\left(Y_i-\beta_{M0}-\sum_{j=1}^{d_M}\beta_{Mj}Z_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{d_M}|\beta_{Mj}|\leq t. \tag{2}$$

We use lars (Efron *et al.*, 2004) repeatedly to solve the different penalized least squares regression problems. The parameter $t\in\mathbb{R}^+$ indexes the entire lasso path. The lars algorithm proceeds in steps, indexed by k, for $k=0,\dots,K$, where K is the number of steps needed in order to reach the unpenalized ordinary least squares solution from the intercept-only model, which roughly equals p. We identify the lasso solution at step k by $\hat{S}_M(t_k)$. We denote the entire path by $\mathcal{S}_M=\{\hat{S}_M(t):t\in\mathbb{R}^+\}$. We let $\hat{\beta}_M^k=(\hat{\beta}_{M0}^k,\dots,\hat{\beta}_{Md_M}^k)$ denote the estimate of β_M corresponding to the lasso solution at step k, $\hat{\mu}_M^k=\mathbf{X}_M\hat{\beta}_M^k$, $df(\hat{\mu}_M^k)$ the corresponding degrees of freedom, $\mathcal{B}_M^k=\{j:\beta_{Mj}^k\neq 0\}$ and $b_M^k=|\mathcal{B}_M^k|$. Initially, when $k=0$, $\mathcal{B}_M^k=\emptyset$ and $b_M^k=0$.

To select a model along the path $\mathcal{S}_M$, GEMULA optionally uses either Bayesian information criterion (BIC), Akaike information criterion (AIC) or AIC$_c$ [a modified small sample version of AIC, see Sugiura (1978)]. Motivated by the results from a simulation study

(Geeven, 2010), we use the AIC$_c$ criterion when we apply GEMULA to analyze real gene expression data. Let AIC$_c(\hat{S}_M(t_k))$ denote this criterion, for $\hat{S}_M(t_k) \in \mathcal{S}_M$. Then

$$\mathrm{AIC}_c(\hat{S}_M(t_k)) = \frac{\|Y - \hat{\mu}_M^k\|^2}{n\sigma^2} + \frac{2}{n} df(\hat{\mu}_M^k) + \frac{2df(\hat{\mu}_M^k)(df(\hat{\mu}_M^k)+1)}{n - df(\hat{\mu}_M^k) - 1}. \tag{3}$$

It was shown by Zou *et al.* (2007) that the optimal model in $\mathcal{S}_M$ according to the selection criterion can be found by minimizing (3) over all t_k, $k=0,\dots,K$ and therefore we let

$$k_M^{\mathrm{AIC}_c} = \arg\min_{t_k} \mathrm{AIC}_c(\hat{S}_M(t_k)).$$

2.2 Stages of GEMULA

GEMULA is composed of four stages, i.e. a predictor pre-selection stage, a candidate model generation stage, a model evaluation stage and a stability selection stage. Although the lasso performs predictor selection, we do a pre-selection in Stage I to control computational complexity, as the number of candidate model terms grows super-exponentially as a function of the number of predictors.

GEMULA—Stage I: in Stage I, GEMULA determines the order in which the input predictors may enter the candidate models by applying the lars algorithm to M_0. Since at each step k of the algorithm, the index of exactly one predictor enters the set $\mathcal{B}_{M_0}^k$, GEMULA uses the mapping

$$r(j) = \min \quad \{k : j \in \mathcal{B}_{M_0}^k\}, \quad j \in \{1,\dots,p\},$$

and its inverse r^{-1} defined by

$$r^{-1}(s) = j \qquad \Leftrightarrow r(j) = s \qquad j \in \{1,\dots,p\}, s \in \{1,\dots,K\}$$

to define the order number s for the predictor X_j.

GEMULA—Stage II: in Stage II, GEMULA uses the lasso to generate candidate models confined to Q different candidate model subspaces. The different model subspaces are identified by 3D parameters $\gamma_q = (\gamma_{q1}, \gamma_{q2}, \gamma_{q3})$, for $q=1,\dots,Q$, where γ_{q1} represents the maximum allowed order of interactions between terms in the models, γ_{q2} the maximum power to which candidate predictors are raised in candidate terms and γ_{q3} represents the maximum number of candidate terms allowed in the model. The complete collection of models that are considered by GEMULA is $\mathcal{M} = \mathcal{M}_{\gamma_1} \cup \dots \cup \mathcal{M}_{\gamma_Q}$. For the model subspace $\mathcal{M}_{\gamma_q}$ defined by γ_q, $\mathbf{X}_{\gamma_q}$ denotes the design matrix of the model in $\mathcal{M}_{\gamma_q}$ with the largest possible number of predictors confined by the order determined in Step I. When interactions between predictors are considered, the restrictions on the maximum number of allowed terms imposed by γ_{q3} force GEMULA to limit the number of predictors. In a model with main terms and pairwise interactions between p variables, there are $p + \binom{p}{2}$ terms. When three-way interactions are also included, this increases to $p + \binom{p}{2} + \binom{p}{3}$. For instance, suppose we set $Q=3$, $\gamma_1 = (1,1,150), \gamma_2 = (1,1,150)$ and $\gamma_3 = (1,1,150)$, then

$$\begin{aligned}
\mathbf{X}_{\gamma_1} &= [\mathbf{1} \quad X_{r^{-1}(1)} \cdots X_{r^{-1}(150)}]. \\
\mathbf{X}_{\gamma_2} &= [\mathbf{1} \quad X_{r^{-1}(1)} \cdots X_{r^{-1}(16)} \\
&\qquad X_{r^{-1}(1)} X_{r^{-1}(2)} \cdots X_{r^{-1}(15)} X_{r^{-1}(16)}] \\
\mathbf{X}_{\gamma_3} &= [\mathbf{1} \quad X_{r^{-1}(1)} \cdots X_{r^{-1}(9)} \\
&\qquad X_{r^{-1}(1)} X_{r^{-1}(2)} \cdots X_{r^{-1}(8)} X_{r^{-1}(9)}] \\
&\qquad X_{r^{-1}(1)} X_{r^{-1}(2)} X_{r^{-1}(3)} \cdots X_{r^{-1}(7)} X_{r^{-1}(8)} X_{r^{-1}(9)}],
\end{aligned}$$

For each matrix $\mathbf{X}_{\gamma_q}$, GEMULA fits the entire path of lasso solutions $\mathcal{S}_{\gamma_q}$ and selects the optimal lasso-AIC shrinkage parameter $k_{\gamma_q} = k_{\gamma_q}^{\mathrm{AIC}_c}$. We denote the selected candidate model, i.e. the selected subset of model terms identified by $\mathcal{B}_{\gamma_q}^{k_{\gamma q}}$, by M_q and the corresponding fitted response values by $\hat{Y}^{M_q} = \hat{\mu}_{\gamma_q}^{k_{\gamma q}}$. For results reported in Section 3.1, we used $\gamma_1 = (1,1,500)$, $\gamma_2 = (2,1,500)$, $\gamma_3 = (3,1,500)$ and $\gamma_4 = (2,2,500)$ and in Section 3.2 we used $\gamma_1 = (1,1,500)$, $\gamma_2 = (2,1,500)$, $\gamma_3 = (3,1,500)$ for the *early responsive genes* and $\gamma_1 = (1,1,700)$, $\gamma_2 = (2,1,700)$, $\gamma_3 = (3,1,700)$ for the *late responsive genes* (which is a bigger set).

GEMULA—Stage III: in Stage III, GEMULA uses cross-validation to evaluate the fit of the Q candidate models. As goodness-of-fit measure, we use the R^2 statistic, because it has an intuitive interpretation that is of interest also biologically. Recall that for a candidate model M_q and corresponding fitted response values $\hat{Y}^{M_q}$, the R^2 is given by

$$R^2(M_q) = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}^{M_q})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

For comparison, we always use the same data splits to cross-validate the different candidate models and we used 5-fold cross-validation throughout. For each model q, in Stages I and II, all n observations are used to select and fit the candidate model. In Stage III, we randomly partition the observations into five mutually disjunct sets of (almost) equal size and quote $R^2(M_q)$'s averaged over these five validation sets as $\bar{R}^2_{cv}$, where for each set the $\hat{Y}^{M_q}$'s are determined by re-estimating the parameters on data with the observations from the corresponding validation set left out.

GEMULA—Stage IV: we give a brief outline of the method of Meinshausen and Bühlmann (2010) that we implemented to assess the significance of terms in candidate models generated by GEMULA. Let a model M with candidate model predictors $Z_1,\dots,Z_{d_M}$, generated by GEMULA in the first two stages, be given. The lasso produces the estimate $\hat{\beta}_M^t = (\hat{\beta}_0, \hat{\beta}_1^t, \dots, \hat{\beta}_{d_M}^t)$ based on a shrinkage parameter $t = k_M^{\mathrm{AIC}_c}$, but provides no formal inference. The stability selection procedure is based on resampling and allows us to compute 'selection probabilities' for the candidate predictors. We iteratively resample for $b=1,\dots,B$, uniformly at random, subsets of $\{1,\dots,n\}$ of size $[n/2]$, and recompute the regularization paths and corresponding coefficients on the subsample. For $t \in \mathbb{R}^+$, let $\hat{\beta}_M^{tb^*} = (\hat{\beta}_0^{b^*}, \hat{\beta}_1^{tb^*}, \dots, \hat{\beta}_{d_M}^{tb^*})$ denote the lasso solution obtained in iteration b. Then, for each candidate predictor j, its selection probability can be computed as the proportion $\hat{\pi}_j^t$ of the B iterations in which the coefficient of Z_j does not equal zero:

$$\hat{\pi}_j^t = \frac{1}{B}\sum_{b=1}^{B} I_{\{\hat{\beta}_j^{tb*}\neq 0\}}. \tag{4}$$

For any $t \in \mathbb{R}^+$, the $\hat{\pi}_j^t$ given by (4) define a ranking of the variables according to selection probability. We use the selection probabilities to prioritize the predictor terms in models generated by GEMULA, for a fixed $t = k_M^{\mathrm{AIC}_c}$.

Multivariate-responses: the univariate model presented above, does not quantify explained variation across experiments. For general applicability, we suggest an extension to the case where gene expression is obtained under a set of k different conditions and Y_{ij} represents the expression of gene i in condition k. We equip GEMULA with a feature extraction procedure based on principal component analysis (PCA). This allows for the extraction of strong signals across experiments [which may be modeled using a common set of predictors, see e.g. (Bonneau, 2007)] without having to deal with a mathematically and computationally more complex regression problem (Obozinski *et al.*, 2011). Let $\mathbf{Y} = \mathbf{V}\Lambda\mathbf{W}'$ be a singular value decomposition, where $\mathbf{V} = [v_1 \cdots v_k]$ is a matrix of score vectors, Λ a diagonal matrix and $\mathbf{W}$ the $k \times k$ matrix containing the loadings of the k principal vectors. Score vectors v_j can be used as input to GEMULA to discover interesting regulatory interactions *across* experiments.

Outer cross-validation: another option in our GEMULA-package is the use of an outer loop of cross-validation to reduce selection bias in the $\bar{R}^2_{cv}$ estimates in Stage III, which will cause these to be overly optimistic. Since this will at the same time increase the variability (and reduce the power) of the whole fitting procedure, this is only recommended when selection bias is expected to be a serious problem, e.g. when p is large with respect to n. We provide a comparison of the results obtained with both ordinary and nested CV in Supplementary Table S3.

3 RESULTS

3.1 Yeast response to environmental stress

To compare the competitive predictive strength of different sets of predictor variables and to compare GEMULA to MARS on data from a well-characterized biological system, we applied GEMULA to gene expression data from the yeast *Saccharomyces cerevisiae*. Gasch *et al.* (2000) published genome-wide expression patterns in yeast cells that were exposed to various changes in environmental conditions, including heat shock, nitrogen depletion and amino acid starvation. Approximately 900 genes show a comparable and strong transcriptional response to almost all stress conditions examined (Gasch *et al.*, 2000). The term 'environmental stress response' (ESR) was coined to describe this phenomenon. The heat shock expression dataset contains time course profiles of all ESR genes at different different time-points, ranging from 5 to 80 mins post-shock. Here, we consider the observed gene expression at 20 mins as our response variable Y. Analysis of the other time-points yielded similar results.

We consider two different sets of predictor variables. From the experimentally derived DNA binding sites published by Macisaac *et al.* (2006), we extracted 123 different position frequency matrices (PFMs) representing models of the DNA sequences bound by yeast TFs. We used the TRAP (TRanscription factor Affinity Prediction) method developed by Roider *et al.* (2007) to calculate TF–DNA binding affinities for binding of TFs to the genomic DNA sequences from 1 to 1000 bp directly upstream of yeast open reading frames. The resulting predictors are referred to as MRM.

Table 1. Comparison of selected models fitted using GEMULA and MARS on yeast heat shock gene expression data with different sets of predictors

Pred	Model	GEMULA				MARS	
		N.P	N.T	$\bar{R}^2_{cv}$	95% C.I	$\bar{R}^2_{cv}$	95% CI
MRM	M1	42	42	0.39	0.34–0.44	0.27	0.17–0.32
MRM	M2	31	71	0.48	0.43–0.53	0.11	0.00–0.19
MRM	M3	14	16	0.39	0.33–0.44	0.05	0.00–0.15
NUC	M1	18	18	0.51	0.34–0.57	0.64	0.63–0.65
NUC	M2	19	49	0.64	0.56–0.68	0.56	0.02–0.65
NUC	M3	14	91	0.61	0.40–0.69	0.49	0.01–0.64
Both	M1	41	41	0.61	0.51–0.65	0.55	0.45–0.61
Both	M2	31	61	0.65	0.42–0.72	0.31	0.00–0.50
Both	M3	14	63	0.70	0.66–0.73	0.15	0.00–0.39

The M_1 models contain only main effects, whereas M_2, and M_3 also contain interactions (Section 2.2). The column 'N.P' lists the number of predictors in the fitted GEMULA models and 'N.T' denotes the number of model terms. $\bar{R}^2_{cv}$ is the goodness-of-fit criterion of the selected model (Section 2.2), together with a 95% bootstrap confidence interval.

Variation in rates of transcription can result from factors other than TF binding. For instance, several studies have pointed out that there is a clear relationship between patterns of histone acetylation and observed gene expression [Karlić *et al.* (2010), Markowetz *et al.* (2010), Yuan *et al.* (2006)]. Ramsey *et al.* (2010) have shown that indeed such data can be used to improve prediction of transcription factor binding sites (TFBS)s. Therefore, we included experimental data from the genome-wide map of nucleosome acetylation and methylation (Pokholok *et al.*, 2005) as an additional set of predictors. This second set of predictors consist of 19 extra variables, including eight sets of histone modifications, measured under both normal conditions and following oxidative stress. We refer to these predictors as NUC.

Fitted GEMULA models contain TFBS motifs of TFs that are well known to be crucial for the transcriptional regulation of heat shock responsive genes. For instance, the GEMULA M2 model fitted using the MRM + NUC predictors contains the predictors MSN2, MSN4, ROX1, HSF1 and CST6. These 5 are among the 12 TFs that regulate gene regulatory modules of heat shock genes identified by Wu and Li (2008). Furthermore, we find that GEMULA models containing interactions between predictors outperform models with only main effects (Table 1, models M_2,M_3). In contrast, $\bar{R}^2_{cv}$ of MARS models decreases when interactions are included, likely due to overfitting. Another noteworthy conclusion is that inclusion of the additional NUC predictors consistently leads to models with higher $\bar{R}^2_{cv}$s. This supports the hypothesis that both chromatin structure dynamics and TF binding play an important role in the regulation of gene expression.

3.1.1 Identification of TF–TF interactions An additional powerful feature of GEMULA is that we can include interaction terms and model context-specific interactions between TFs. The experiments performed by Gasch *et al.* include several different stress conditions. Here, we consider three of these: Amino

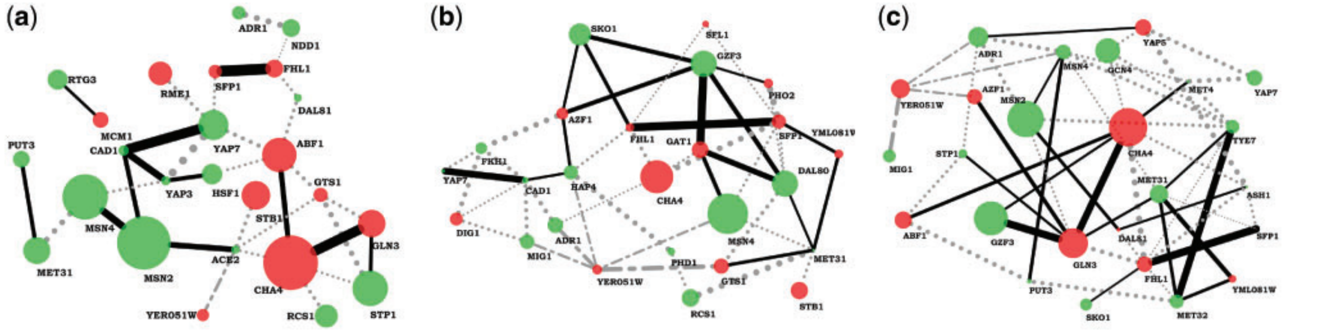

Fig. 2. Graphical representation of condition-specific TF regulatory networks inferred by GEMULA for yeast gene expression in response to different stress conditions: (**a**) Hydrogen peroxide treatment (**b**) Amino acid starvation (**c**) Nitrogen depletion. Green nodes correspond to TFs with a positive effect on expression, whereas red nodes indicate repressors. Node size is proportional to regression coefficients. Edge width is determined by selection probability (Section 2.2). Edges that correspond to TF–TF pairs with a CC > 1 are colored black. The CCs, taken from an independent study on combinatorial regulation in the literature Balaji *et al.* (2006), measure the significance of overlap in target genes and hence high CCs correspond to functionally interacting TFs. Dotted edges correspond to TF–TF pairs not present in the study by Balaji *et al.* (2006).

acid starvation, Nitrogen depletion and Hydrogen peroxide treatment. We fit models using GEMULA and focus on the selected TFs and their interactions in the inferred models.

Results are presented as context-specific TF regulatory networks in Figure 2. GEMULA correctly predicts important roles for the major regulators of the general stress response MSN2 and MSN4, the regulator of ribosomal protein genes FHL1, the osmotic and oxidative stress response TF SKO1 and SFP1, a TF that controls expression of ribosome biogenesis genes in response to nutrients and stress. Of particular interest are condition specific changes in TF activity suggested by the networks in Figure 2. For instance, GEMULA predicts interactions between GAT1, DAL80 and GZF3, all TFs known to mediate nitrogen-responsive expression, in the amino acid starvation network.

To assess the functional relevance of the TF–TF interactions identified by GEMULA, we use results of an independent study on combinatorial regulation from the literature. Balaji *et al.* (2006) computed co-regulatory coefficients (CCs) of 5622 pairs of TFs in the yeast transcriptional network. The CC measures the significance of overlap in target genes shared between pairs of TFs and thus high CC values correspond to pairs of TFs that co-regulate common sets of targets. We tested whether TF–TF interactions with high selection probability (Section 2.2) have higher CCs on average, using a Wilcoxon rank sum test. Hence, we test for a significant change in location of the distribution of CCs for TF–TF pairs with high selection probability relative to all candidate TF–TF pairs considered by GEMULA. Table 2 contains the results and shows that, on average, the interactions with a high selection probability correspond to TF–TF pairs with significantly higher CCs and hence suggests their functional relevance.

Table 2. Assessment of functional relevance of context specific TF–TF interactions identified by GEMULA in three different stress conditions in yeast on independent literature evidence

Condition	TF pairs	Mean CC	*P*-value
Hydrogen peroxide treatment	All	1.59	NA
Hydrogen peroxide treatment	$\hat{\pi}_j > 0.5$	4.01	0.010
Hydrogen peroxide treatment	$\hat{\pi}_j > 0.75$	6.03	0.004
Amino acid starvation	All	2.08	NA
Amino acid starvation	$\hat{\pi}_j > 0.5$	2.16	0.542
Amino acid starvation	$\hat{\pi}_j > 0.75$	4.11	0.001
Nitrogen depletion	All	3.06	NA
Nitrogen depletion	$\hat{\pi}_j > 0.5$	4.25	0.019
Nitrogen depletion	$\hat{\pi}_j > 0.75$	5.60	0.021

On average, TF–TF interactions j with a high selection probability $\hat{\pi}_j$ have a significantly higher CCs, based on a Wilcoxon rank sum test. NA, not applicable.

3.2 TF regulatory networks underlying axon growth

We next used GEMULA to infer TF–TF interactions in the gene regulatory network underlying neuronal outgrowth. As a cellular model, we used F11 cells (Platika *et al.*, 1985). Upon stimulation with Forskolin, F11 cells acquire a neuronal phenotype, which results in the outgrowth of neurites (Ghil *et al.*, 2000). We reanalyzed previously published genome-wide gene expression time course profiles of F11 cells measured at four time-points following Forskolin stimulation (MacGillavry *et al.*, 2011). Using Bayesian Analysis of Time Series [BATS, Angelini *et al.* (2008)], we identified a set of *Forskolin responsive* genes, i.e. genes that show differential expression in Forskolin stimulated F11 cells compared with unstimulated control cells. BATS is a Bayesian method for analysis of gene expression data, specifically tailored to handle short, replicated time series.

Initially, we analyzed the *entire* group of Forskolin responsive genes at all four time-points separately (Supplementary Table S2). We observed rather low $\bar{R}^2_{cv}$s for the fitted models, but interestingly, the amount of 'explained' expression variation seemed to vary in time. We then further examined the expression profiles of the regulated genes using PCA and noticed clear differences between the expression at the first two time-points immediately following Forskolin stimulation and the two 'late' time-points. We performed PCA on the gene expression matrix to see if the observed variation in expression could be further divided into biologically meaningful patterns. For each gene, we used the signs of the coefficients corresponding to the first and second principal component to define four gene sets of interest (Supplementary Table S1). The first

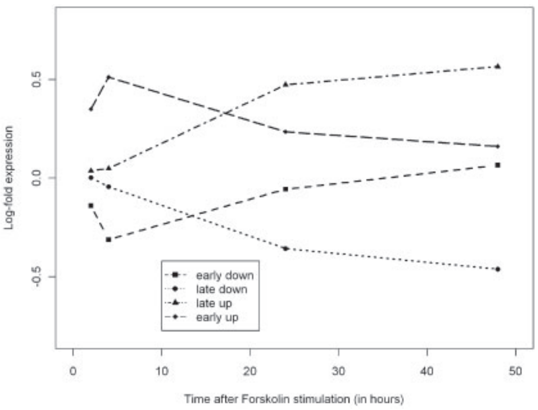

Fig. 3. Plot of average gene expression of clusters of Forskolin-responsive genes in F11 cells at several time-points following Forskolin stimulation.

principal component seems to identify the main direction of the induced gene expression changes and can be used to distinguish between predominantly up- versus predominantly downregulated genes, whereas the second principal component discriminates between early genes and late genes. Based on these observations, we named these four expression clusters 'early up', 'late up', 'late down' and 'early down'. Figure 3 shows a plot of the average expression of genes in these clusters.

The definition of these gene sets allowed us to investigate the possible underlying time-dependent activity of TFs and the presence of an 'early' and a 'late' transcriptional response. This separation into early and late genes makes sense from a computational point of view, because inclusion of background genes, i.e. genes that are not actually regulated under the experimental condition of interest, may adversely affect the overall fit of the models, as the measured gene expression for such genes constitutes only noise (Das *et al.*, 2004). From a biological point of view, it makes sense to distinguish between 'early' and 'late' genes, because interactions between TFs and their target genes are expected to be condition-specific and time-dependent. The yeast ESR set that we analyzed in Section 3.1 consists of genes that show a remarkably consistent transcriptional response to different stress conditions. In contrast, the set of Forskolin-responsive genes displays strong time-heterogeneity, which is why a further subdivision is necessary. We applied GEMULA again, but this time we distinguished between early responsive genes, consisting of the union of genes in the 'early up' and 'early down' sets, and late responsive genes, consisting of all genes in the 'late up' and 'late down' sets. If the early genes are transcriptionally regulated at the early, *but not* the late time-points, we expect the models fitted at the late time-points using data from the early responsive genes to have considerably lower $\bar{R}^2_{cv}$s and vice versa. We indeed find that this is the case (Table 3). Moreover, this time the fitted models have notably higher $\bar{R}^2_{cv}$s than the models fitted on the *entire* group of genes (Supplementary Table S2). These results suggest that indeed the experimental data arise from two separate waves of transcriptional changes. Of particular interest biologically are the sets of TFs that are associated to these early and late gene expression changes, the interactions between them and their effects on axon outgrowth. To this end, we made a subselection of genes based on informative GO-terms (Supplementary Material) and calculated selection probabilities for candidate terms in the models generated by GEMULA. This allowed us to prioritize terms in the models and focus on the most significant predictors. The resulting TF networks are shown in Figure 4. Note that the network inferred for the *early*, but not the *late* changes, contains the predictors V.CREB.01 and V.CREBATF.Q6, among others. These motifs represent the TFBSs for CREB, a cAMP-inducible TF. Activation of CREB is known to be induced by Forskolin stimulation of F11 cells (MacGillavry *et al.*, 2009) and Gao *et al.* (2004) have shown that activated CREB is sufficient to promote spinal axon regeneration.

Interestingly, many TFs binding to the DNA binding motifs in Figure 4a such as V.CEBPDELTA.Q6, V.PPARA.02 and V.PBX.Q3 were previously found to have a significant effect on axon growth upon knockdown in F11 cells (Geeven *et al.*, 2011). In contrast, the network that was inferred from the late time-points (Fig. 4b) contains no known transcriptional regulators of neurite outgrowth, suggesting that these cells indeed have entered a different stage of differentiation. Together, these observations show that GEMULA can be used to detect important context-specific regulatory interactions in gene networks that underlie transitional changes in cell fates.

4 DISCUSSION

The success of a regression-based approach to modeling gene expression and DNA sequence data depends on appropriate choices for the type of model and the predictors used as input. This was first demonstrated by Das *et al.* (2004) who proposed a strategy that uses the non-parametric MARS method as core regression routine (Das *et al.*, 2004, 2006). Das *et al.* (2004) claim that their MARSMOTIF algorithm, which allows modeling of synergistic interactions between predictors, is approximately 1.5 to 3.5 times more accurate than the method of Bussemaker *et al.* (2001), which is based on a linear model. The comparison is based on an R^2-like $\Delta\chi^2$ statistic and no cross-validation was performed. A comprehensive comparison is lacking. The results we present in this article show that similar synergistic interactions as in Das *et al.* (2004, 2006) can be modeled using linear models. In fact, GEMULA produces biologically plausible models with superior fit compared with MARS when applied to the same data. Typically, models contain a rather large number of predictors, whereas not all of these are equally relevant for the underlying biological processes. We demonstrate that GEMULA identifies synergistic pairs of TFs that are likely to be functionally relevant, i.e. on average the overlap in target genes of such identified TF–TF pairs is significantly higher than expected by chance.

In order to build models that are biologically useful and interpretable, the availability of relevant biological predictors used as input are crucial. The TRAP predictors we consider in this article represent *in silico* predicted binding affinities of TFs. Using yeast data, we showed that GEMULA in combination with TRAP predictors successfully identifies interactions between known heat shock regulating TFs such as MSN2, MSN4, HSF1, CST6 and ROX1 that underlie the observed variation in expression of yeast genes in response to hypothermic shock. The TRAP predictors are real-valued and have an interpretation that is closer to experimentally measured TF–DNA binding profiles as obtained, for instance, with ChIP-chip assays than other motif representations, which use exact words. Models that use TRAP predictors have a clear

Table 3. Comparison of models fitted using GEMULA and MARS for early and late Forskolin-responsive genes in F11 cells at all four time-points

Time (h)	Model	Early responsive genes						Late responsive genes					
				GEMULA		MARS				GEMULA		MARS	
		N.P	N.T	$\bar{R}^2_{cv}$	95% CI	$\bar{R}^2_{cv}$	95% CI	N.P	N.T	$\bar{R}^2_{cv}$	95% CI	$\bar{R}^2_{cv}$	95% CI
2	M1	30	30	0.19	0.12–0.27	0.08	0.05–0.11	8	8	0.04	0.02–0.06	0.02	0.01–0.03
2	M2	31	72	0.33	0.23–0.44	0.01	0.00–0.04	36	53	0.11	0.07–0.16	0.01	0.00–0.08
2	M3	14	47	0.25	0.12–0.38	0.00	0.00–0.03	16	27	0.09	0.04–0.15	0.00	0.00–0.01
4	M1	16	16	0.11	0.06–0.18	0.06	0.04–0.10	0	0	0	0–0	0.02	0.01–0.03
4	M2	31	75	0.21	0.13–0.32	0.01	0.00–0.04	36	53	0.04	0.01–0.07	0.00	0.00–0.01
4	M3	14	39	0.13	0.05–0.24	0.00	0.00–0.02	16	16	0.02	0.00–0.04	0.00	0.00–0.01
24	M1	50	50	0.06	0.02–0.13	0.03	0.01–0.05	39	39	0.25	0.20–0.29	0.13	0.10–0.15
24	M2	31	80	0.08	0.02–0.19	0.00	0.00–0.02	36	63	0.25	0.21–0.30	0.04	0.00–0.09
24	M3	14	20	0.05	0.01–0.09	0.00	0.00–0.01	15	27	0.22	0.18–0.27	0.02	0.00–0.06
48	M1	5	5	0.02	0.00–0.05	0.05	0.03–0.08	44	44	0.23	0.19–0.28	0.16	0.14–0.18
48	M2	31	85	0.14	0.06–0.27	0.00	0.00–0.02	35	52	0.25	0.20–0.30	0.06	0.00–0.11
48	M3	14	60	0.09	0.03–0.21	0.00	0.00–0.01	16	37	0.22	0.18–0.27	0.03	0.00–0.08

Columns 3–10 correspond to models fitted for the early responsive genes and columns 11–18 to models for the late responsive genes.The column 'N.P' lists the number of predictors in the fitted GEMULA models and 'N.T' denotes the number of model terms. $\bar{R}^2_{cv}$ is the goodness-of-fit criterion of the selected model (Section 2.2), together with a 95% bootstrap confidence interval.

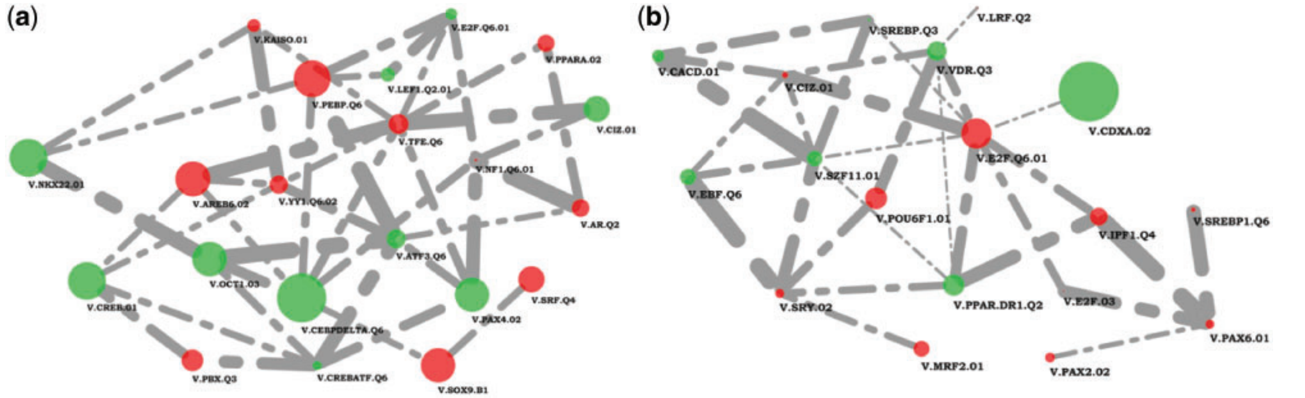

Fig. 4. Graphical representation of the network of TF–TF interactions that are associated to early (**a**) and late (**b**) changes in gene expression in F11 cells after Forskolin stimulation. Node size is proportional to regression coefficients. The edges represent interactions between predictors in models inferred by GEMULA with a selection probability >0.25. Thicker edges indicate higher selection probabilities.

interpretation, which facilitates the step toward biological validation. Our analysis of yeast stress response gene expression data shows that GEMULA can identify context-specific TF–TF interactions that underly observed variation in gene expression under different physiological conditions.

We demonstrate that our method can also be used to analyze *mammalian* gene expression data. We applied GEMULA to identify TFs associated to observed variations in early and late gene expression changes in F11 cells in response to Forskolin stimulation. The observed fit in terms of $\bar{R}^2_{cv}$ of the resulting models is considerably less than for GEMULA models inferred from yeast heat shock data where additional predictors are available. However, we show that GEMULA identifies several crucial TFs with a well-established role in the regulation of neuronal outgrowth-associated gene expression and additionally provides new insights into the temporal dynamics of the regulatory network underlying axon growth. The strength of our approach is the combination of l_1-penalization and re-sampling methods to fit and select appropriately regularized models. A current limitation is that the linear models we discuss in this article are not capable of quantifying contributions of TFs to variation in gene expression *across experiments*. To extend the penalized regression framework to the multivariate case of multiple experiments would require fitting a model that includes parameters representing TF activity in each condition separately [see e.g. Bonneau (2007)], resulting in a computationally more complex problem. How to induce the right amount of sparsity and solve the corresponding optimization problem would be an interesting direction for further research.

5 CONCLUSION

Linear models are valuable tools for inference of transcriptional gene regulatory interactions and synergistic pairwise interactions between predictors that underlie observed changes in gene expression under a given condition of interest. We believe that in the near future, as more accurate experimental data based on ultra high-throughput RNA sequencing (RNA-Seq) technology will become available, `GEMULA` will prove to be an even more useful method for modeling transcriptional networks of complex mammalian systems.

Funding: Financial support from the Netherlands Organization for Scientific Research (NWO; CLS grant 635.100.008 to M.C.M.dG.). R.E.vK. and A.B.S received funding from the Center for Medical Systems Biology (CMSB) and from the EU FP7 framework Health program SynSys (grant no. 242167).

Conflicts of interest: None declared.

REFERENCES

Angelini,C. *et al.* (2008) BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, **9**:415.

Balaji,S. *et al.* (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.

Bonneau,R. (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell*, **131**, 1354–1365.

Bussemaker,H. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Das,D. *et al.* (2006) Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.*, **2**:2006.0029.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Friedman,J. (1991) Multivariate adaptive regression splines. *Ann. Stat.*, **19**, 1–67.

Gao,Y. *et al.* (2004) Activated creb is sufficient to overcome inhibitors in myelin and promote spinal axon regeneration in vivo. *Neuron*, **44**, 609–621.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Geeven,G. (2010) *Computational Statistics for Identification of Transcriptional Gene Regulatory Interactions*. PhD Thesis, Vrije Universiteit, Amsterdam.

Geeven,G. *et al.* (2011) LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data. *Nucleic Acids Res*, **39**.

Ghil,S.-H. *et al.* (2000) Neurite outgrowth induced by cyclic amp can be modulated by the a subunit of go. *J. Neurochem.*, **74**, 151–158.

Karlić,R. *et al.* (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.

Krämer,N. *et al.* (2009) Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, **10**, 384.

MacGillavry,H.D. *et al.* (2009) NFIL3 and cAMP response element-binding protein form a transcriptional feedforward loop that controls neuronal regeneration-associated gene expression. *J. Neurosci.*, **29**, 15542–15550.

MacGillavry,H.D. *et al.* (2011) Genome-wide gene expression and promoter binding analysis identifies nfil3 as a repressor of c/ebp target genes in neuronal outgrowth. *Mol. Cell. Neurosci.*, **46**, 460–468.

Macisaac,K. *et al.* (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**:113.

Markowetz,F. *et al.* (2010) Mapping dynamic histone acetylation patterns to gene expression in nanog-depleted murine embryonic stem cells. *PLoS Comput. Biol.*, **6**, e1001034.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B* , **72**, 417–473.

Menéndez,P. *et al.* (2010) Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. *PLoS One*, **5**, e14147.

Obozinski,G. *et al.* (2011) Support union recovery in high-dimensional multivariate regression. *Ann. Stat.*, **39**, 1–47.

Platika,D. *et al.* (1985) Neuronal traits of clonal cell lines derived by fusion of dorsal root ganglia neurons with neuroblastoma cells. *Proc. Natl Acad. Sci. USA*, **82**, 3499–3503.

Pokholok,D.K. *et al.* (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**, 517–527.

Ramsey,S.A. *et al.* (2010) Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, **26**, 2071–2075.

Roider,H.G. *et al.* (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

Sugiura,N. (1978) Further analysis of the data by akaike's information criterion and the finite corrections. *Commun. Stat. Theor. Methods*, **7**, 13–26.

Tibshirani,R. (1994) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Wu,W.S. and Li,W.H. (2008) Identifying gene regulatory modules of heat shock response in yeast. *BMC Genomics*, **9**:439.

Yuan,G.-C. *et al.* (2006) Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Genome Biol.*, **7**, R70.

Zou,H. *et al.* (2007) On the 'degrees of freedom' of the lasso. *Ann. Stat.*, **35**, 2173–2192.