The Annals of Applied Statistics
2010, Vol. 4, No. 1, 53–77
DOI: 10.1214/09-AOAS271

REGULARIZED MULTIVARIATE REGRESSION FOR IDENTIFYING MASTER PREDICTORS WITH APPLICATION TO INTEGRATIVE GENOMICS STUDY OF BREAST CANCER

BY JIE PENG[1,2], JI ZHU[3], ANNA BERGAMASCHI, WONSHIK HAN, DONG-YOUNG NOH, JONATHAN R. POLLACK[4] AND PEI WANG[1]

University of California, Davis, University of Michigan, Ann Arbor, Rikshospitalet-Radiumhospitalet Medical Center, Seoul National University, Seoul National University, Stanford University and Fred Hutchinson Cancer Research Center

In this paper we propose a new method `remMap`—REgularized Multivariate regression for identifying MAster Predictors—for fitting multivariate response regression models under the high-dimension–low-sample-size setting. `remMap` is motivated by investigating the regulatory relationships among different biological molecules based on multiple types of high dimensional genomic data. Particularly, we are interested in studying the influence of DNA copy number alterations on RNA transcript levels. For this purpose, we model the dependence of the RNA expression levels on DNA copy numbers through multivariate linear regressions and utilize proper regularization to deal with the high dimensionality as well as to incorporate desired network structures. Criteria for selecting the tuning parameters are also discussed. The performance of the proposed method is illustrated through extensive simulation studies. Finally, `remMap` is applied to a breast cancer study, in which genome wide RNA transcript levels and DNA copy numbers were measured for 172 tumor samples. We identify a trans-hub region in cytoband 17q12-q21, whose amplification influences the RNA expression levels of more than 30 unlinked genes. These findings may lead to a better understanding of breast cancer pathology.

1. Introduction. In a few recent breast cancer cohort studies, microarray expression experiments and array CGH (comparative genomic hybridization) experiments have been conducted for more than 170 primary breast tumor specimens collected at multiple cancer centers [Sorlie et al. (2001), Sorlie et al. (2003), Zhao et al. (2004), Kapp et al. (2006), Bergamaschi et al. (2006), Langerod et

Received December 2008; revised June 2009.

[1]Supported in part by Grant 1R01GM082802 from the National Institute of General Medical Sciences.

[2]Supported in part by Grant DMS-08-06128 from the NSF.

[3]Supported in part by Grants 0705532 and 0748389 from the NSF.

[4]Supported in part by Grant CA97139 from the National Health Institute.

Key words and phrases. Sparse regression, MAP (MAster Predictor) penalty, DNA copy number alteration, RNA transcript level, v-fold cross validation.

al. (2007) and Bergamaschi et al. (2008)]. The resulting RNA transcript levels (from microarray expression experiments) and DNA copy numbers (from CGH experiments) of about $20K$ genes/clones across all the tumor samples were then used to identify useful molecular markers for potential clinical usage. While useful information has been revealed by analyzing expression arrays alone or CGH arrays alone, careful *integrative analysis* of DNA copy numbers and expression data are necessary, as these two types of data provide complimentary information in gene characterization. Specifically, RNA data give information on genes that are over/under-expressed, but do not distinguish primary changes driving cancer from secondary changes resulting from cancer, such as proliferation rates and differentiation state. On the other hand, DNA data give information on gains and losses that are drivers of cancer. Therefore, integrating DNA and RNA data helps to discern more subtle (yet biologically important) genetic regulatory relationships in cancer cells [Pollack et al. (2002)].

It is widely agreed that variations in gene copy numbers play an important role in cancer development through altering the expression levels of cancer-related genes [Albertson et al. (2003)]. This is clear for *cis-regulations*, in which a gene's DNA copy number alteration influences its own RNA transcript level [Hyman et al. (2002) and Pollack et al. (2002)]. However, DNA copy number alterations can also alter in trans the RNA transcript levels of genes from unlinked regions, for example, by directly altering the copy number and expression of transcriptional regulators, or by indirectly altering the expression or activity of transcriptional regulators, or through genome rearrangements affecting cis-regulatory elements. The functional consequences of such *trans-regulations* are much harder to establish, as such inquiries involve assessment of a large number of potential regulatory relationships. Therefore, to refine our understanding of how these genome events exert their effects, we need new analytical tools that can reveal the subtle and complicated interactions among DNA copy numbers and RNA transcript levels. Knowledge resulting from such analysis will help shed light on cancer mechanisms.

The most straightforward way to model the dependence of RNA levels on DNA copy numbers is through a multivariate response linear regression model with the RNA levels being responses and the DNA copy numbers being predictors. While the multivariate linear regression is well studied in statistical literature, the current problem bears new challenges due to (i) high-dimensionality in terms of both predictors and responses; (ii) the interest in identifying *master regulators* in genetic regulatory networks; and (iii) the complicated correlation relationships among response variables. Thus, the naive approach of regressing each response onto the predictors separately is unlikely to produce satisfactory results, as such methods often lead to high variability and over-fitting. This has been observed by many authors, for example, Breiman and Friedman (1997) show that taking into account of the relation among response variables helps to improve the overall prediction accuracy. More recently, Kim, Sohn and Xing (2009) propose a new statistical

framework to explicitly incorporate the relationships among responses by assuming the linked responses depend on the predictors in a similar way. The authors show that this approach helps to select relevant predictors when the above assumption holds.

When the number of predictors is moderate or large, model selection is often needed for prediction accuracy and/or model interpretation. Standard model selection tools in multiple regression such as AIC and forward stepwise selection have been extended to multivariate linear regression models [Bedrick and Tsai (1994), Fujikoshi and Satoh (1997) and Lutz and Bühlmann (2006)]. More recently, sparse regularization schemes have been utilized for model selection under the high dimensional multivariate regression setting. For example, Turlach, Venables and Wright (2005) propose to constrain the coefficient matrix of a multivariate regression model to lie within a suitable polyhedral region. Lutz and Bühlmann (2006) propose an L_2 multivariate boosting procedure. Obozinski, Wainwright and Jordan (2008) propose to use a ℓ_1/ℓ_2 regularization to identify the union support set in the multivariate regression. Moreover, Brown, Vannucci and Fearn (1998, 2002) and Brown, Fearn and Vannucci (1999) introduce a Bayesian framework to model the relation among the response variables when performing variable selection for multivariate regression. Another way to reduce the dimensionality is through factor analysis. Related work includes Izenman (1975), Frank and Friedman (1993), Reinsel and Velu (1998), Yuan et al. (2007) and many others.

For the problem we are interested in here, the dimensions of both predictors and responses are large (compared to the sample size). Thus, in addition to assuming that only a subset of predictors enter the model, it is also reasonable to assume that a predictor may affect only some but not all responses. Moreover, in many real applications, there often exists a subset of predictors which are more important than other predictors in terms of model building and/or scientific interest. For example, it is widely believed that genetic regulatory relationships are intrinsically sparse [Jeong et al. (2001) and Gardner et al. (2003)]. At the same time, there exist *master regulators*—network components that affect many other components, which play important roles in shaping the network functionality. Most methods mentioned above do not take into account the dimensionality of the responses and, thus, a predictor/factor influences either all or none of the responses, for example, Turlach, Venables and Wright (2005), Yuan et al. (2007), the L_2 row boosting by Lutz and Bühlmann (2006), and the ℓ_1/ℓ_2 regularization by Obozinski, Wainwright and Jordan (2008). On the other hand, other methods only impose a sparse model, but do not aim at selecting a subset of predictors, for example, the L_2 boosting by Lutz and Bühlmann (2006). In this paper we propose a novel method `remMap`—REgularized Multivariate regression for identifying MAster Predictors, which takes into account both aspects. `remMap` uses an ℓ_1 norm penalty to control the overall sparsity of the coefficient matrix of the multivariate linear regression model. In addition, `remMap` imposes a "group" sparse penalty, which in essence

is the same as the "group lasso" penalty proposed by Bakin (1999), Antoniadis and Fan (2001), Yuan and Lin (2006), Zhao, Rocha and Yu (2009) and Obozinski, Wainwright and Jordan (2008) (see more discussions in Section 2). This penalty puts a constraint on the ℓ_2 norm of regression coefficients for each predictor, which controls the total number of predictors entering the model, and consequently facilitates the detection of *master predictors*. The performance of the proposed method is illustrated through extensive simulation studies.

We apply the `remMap` method on the breast cancer data set mentioned earlier and identify a significant trans-hub region in cytoband 17q12-q21, whose amplification influences the RNA levels of more than 30 unlinked genes. These findings may shed some light on breast cancer pathology. We also want to point out that analyzing CGH arrays and expression arrays together reveals only a small portion of the regulatory relationships among genes. However, it should identify many of the important relationships, that is, those reflecting primary genetic alterations that drive cancer development and progression. While there are other mechanisms to alter the expression of master regulators, for example, by DNA mutation or methylation, in most cases one should also find corresponding DNA copy number changes in at least a subset of cancer cases. Nevertheless, because we only identify the subset explainable by copy number alterations, the words "regulatory network" ("master regulator") used in this paper will specifically refer to the subnetwork (hubs of the subnetwork) whose functions change with DNA copy number alterations, and thus can be detected by analyzing CGH arrays together with expression arrays.

The rest of the paper is organized as follows. In Section 2 we describe the `remMap` model, its implementation and criteria for tuning. In Section 3 the performance of `remMap` is examined through extensive simulation studies. In Section 4 we apply the `remMap` method on the breast cancer data set. We conclude the paper with discussions in Section 5. Technical details are provided in the supplementary material [Peng et al. (2009b)].

2. Method.

2.1. *Model.* Consider multivariate regression with Q response variables $y_1, \ldots, y_Q$ and P prediction variables $x_1, \ldots, x_P$:

$$y_q = \sum_{p=1}^{P} x_p \beta_{pq} + \epsilon_q, \qquad q = 1, \ldots, Q, \tag{2.1}$$

where the error terms $\epsilon_1, \ldots, \epsilon_Q$ have a joint distribution with mean 0 and covariance Σ_ϵ. In the above, we assume that all the response and prediction variables are standardized to have zero mean and, thus, there is no intercept term in equation (2.1). The primary goal of this paper is to identify nonzero entries in the $P \times Q$ coefficient matrix $\mathbf{B} = (\beta_{pq})$ based on N i.i.d. samples from the above

model. Under normality assumptions, β_{pq} can be interpreted as proportional to the conditional correlation $\mathrm{Cor}(y_q, x_p|x_{-(p)})$, where $x_{-(p)} := \{x_{p'} : 1 \le p' \ne p \le P\}$. In the following, we use $Y_q = (y_q^1, \ldots, y_q^N)^T$ and $X_p = (x_p^1, \ldots, x_p^N)^T$ to denote the sample of the qth response variable and that of the pth prediction variable, respectively. We also use $\mathbf{Y} = (Y_1 : \cdots : Y_Q)$ to denote the $N \times Q$ response matrix, and use $\mathbf{X} = (X_1 : \cdots : X_P)$ to denote the $N \times P$ prediction matrix.

In this paper we shall focus on the cases where both Q and P are larger than the sample size N. For example, in the breast cancer study discussed in Section 4, the sample size is 172, while the number of genes and the number of chromosomal regions are on the order of a couple of hundred (after pre-screening). When $P > N$, the ordinary least square solution is not unique, and regularization becomes indispensable. The choice of suitable regularization depends heavily on the type of data structure we envision. In recent years, ℓ_1-norm based sparsity constraints such as *lasso* [Tibshirani (1996)] have been widely used under such high-dimension-low-sample-size settings. This kind of regularization is particularly suitable for the study of genetic pathways, since genetic regulatory relationships are widely believed to be intrinsically sparse [Jeong et al. (2001) and Gardner et al. (2003)]. In this paper we impose an ℓ_1 norm penalty on the coefficient matrix $\mathbf{B}$ to control the overall sparsity of the multivariate regression model. In addition, we put constraints on the total number of predictors entering the model. This is achieved by treating the coefficients corresponding to the same predictor (one row of $\mathbf{B}$) as a group, and then penalizing its ℓ_2 norm. A predictor will not be selected into the model if the corresponding ℓ_2 norm is too small. Thus, this penalty facilitates the identification of *master predictors*—predictors which affect (relatively) many response variables. This idea is motivated by the fact that master regulators exist and are of great interest in the study of many real life networks including genetic regulatory networks. Specifically, for model (2.1), we propose the following criterion:

$$L(\mathbf{B}; \lambda_1, \lambda_2) = \frac{1}{2}\left\|\mathbf{Y} - \sum_{p=1}^{P} X_p B_p\right\|_F^2 + \lambda_1 \sum_{p=1}^{P} \|C_p \cdot B_p\|_1 + \lambda_2 \sum_{p=1}^{P} \|C_p \cdot B_p\|_2, \tag{2.2}$$

where C_p is the pth row of $\mathbf{C} = (c_{pq}) = (C_1^T : \cdots : C_P^T)^T$, which is a prespecified $P \times Q$ 0–1 matrix indicating the coefficients on which penalization is imposed; B_p is the pth row of $\mathbf{B}$; $\|\cdot\|_F$ denotes the Frobenius norm of matrices; $\|\cdot\|_1$ and $\|\cdot\|_2$ are the ℓ_1 and ℓ_2 norms for vectors, respectively; and "$\cdot$" stands for the Hadamard product (that is, entry-wise multiplication). The indicator matrix $\mathbf{C}$ is prespecified based on prior knowledge: if we know in advance that predictor x_p affects response y_q, then the corresponding regression

coefficient β_{pq} will not be penalized and we set $c_{pq} = 0$ (see Section 4 for an example). When there is no such prior information, $\mathbf{C}$ can be simply set to a constant matrix $c_{pq} \equiv 1$. Finally, an estimate of the coefficient matrix $\mathbf{B}$ is $\widehat{\mathbf{B}}(\lambda_1, \lambda_2) := \arg\min_{\mathbf{B}} L(\mathbf{B}; \lambda_1, \lambda_2)$.

In the above criterion function, the ℓ_1 penalty induces the overall sparsity of the coefficient matrix $\mathbf{B}$. The ℓ_2 penalty on the row vectors $C_p \cdot B_p$ induces row sparsity of the product matrix $\mathbf{C} \cdot \mathbf{B}$. As a result, some rows are shrunken to be entirely zero (Theorem 2.1). Consequently, predictors which affect relatively more response variables are more likely to be selected into the model. We refer to the combined penalty in equation (2.2) as the `MAP` (MAster Predictor) penalty. We also refer to the proposed estimator $\widehat{\mathbf{B}}(\lambda_1, \lambda_2)$ as the `remMap` (REgularized Multivariate regression for identifying MAster Predictors) estimator. Note that the ℓ_2 penalty is a special case (with $\alpha = 2$) of the more general penalty form, $\sum_{p=1}^{P} \|C_p \cdot B_p\|_\alpha$, where $\|v\|_\alpha := (\sum_{q=1}^{Q} |v_q|^\alpha)^{1/\alpha}$ for a vector $v \in \mathcal{R}^Q$ and $\alpha > 1$. In Turlach, Venables and Wright (2005), a penalty with $\alpha = \infty$ is used to select a common subset of prediction variables when modeling multivariate responses. In Yuan et al. (2007), a constraint with $\alpha = 2$ is applied to the loading matrix in a multivariate linear factor regression model for dimension reduction. In Obozinski, Wainwright and Jordan (2008), the same constraint is applied to identify the union support set in the multivariate regression. In the case of multiple regression, a similar penalty corresponding to $\alpha = 2$ is proposed by Bakin (1999) and by Yuan and Lin (2006) for the selection of grouped variables, which corresponds to the blockwise additive penalty in Antoniadis and Fan (2001) for wavelet shrinkage. Zhao, Rocha and Yu (2009) propose the penalty with a general $\alpha > 1$. However, none of these methods take into account the high dimensionality of response variables and, thus, predictors/factors are simultaneously selected for all responses. On the other hand, by combining the ℓ_2 penalty and the ℓ_1 penalty together in the `MAP` penalty, the `remMap` model not only selects a subset of predictors, but also limits the influence of the selected predictors to only some (but not necessarily all) response variables. Thus, it is more suitable for the cases when both the number of predictors and the number of responses are large. Last, we also want to point out a difference between the `MAP` penalty and the `ElasticNet` penalty proposed by Zou and Hastie (2005), which combines the ℓ_1 norm penalty with the squared ℓ_2 norm penalty. The `ElasticNet` penalty aims to encourage a group selection effect for highly correlated predictors under the multiple regression setting. However, the squared ℓ_2 norm itself does not induce sparsity and thus is intrinsically different from the ℓ_2 norm penalty discussed above.

In Section 3 we use extensive simulation studies to illustrate the effects of the `MAP` penalty. We compare the `remMap` method with two alternatives: (i) the `joint` method which only utilizes the ℓ_1 penalty, that is, $\lambda_2 = 0$ in (2.2); (ii) the `sep` method which performs Q separate lasso regressions. We find that if there

exist large hubs (master predictors), `remMap` performs much better than `joint` in terms of identifying the true model; otherwise, the two methods perform similarly. This suggests that the "simultaneous" variable selection enhanced by the ℓ_2 penalty pays off when there exist a small subset of "important" predictors, and it costs little when such predictors are absent. Moreover, by encouraging the selection of master predictors, the `MAP` penalty explicitly makes use of the correlations among the response variables caused by sharing a common set of predictors. We make a note that there are methods, such as Kim, Sohn and Xing (2009), that make more specific assumptions on how the correlated responses depend on common predictors. If these assumptions hold, it is possible that such methods can be more efficient in incorporating the relationships among the responses. In addition, both `remMap` and `joint` methods impose sparsity of the coefficient matrix as a whole. This helps to borrow information across different regressions corresponding to different response variables. It also amounts to a greater degree of regularization, which is usually desirable for the high-dimension–low-sample-size setting. On the other hand, the `sep` method controls sparsity for each individual regression separately and thus is subject to high variability and overfitting. As can be seen by the simulation studies (Section 3), this type of "joint" modeling greatly improves the model efficiency. This is also noted by other authors, including Turlach, Venables and Wright (2005), Lutz and Bühlmann (2006) and Obozinski, Wainwright and Jordan (2008).

2.2. *Model fitting*. In this section we propose an iterative algorithm for solving the `remMap` estimator $\widehat{\mathbf{B}}(\lambda_1, \lambda_2)$. This is a convex optimization problem when the two tuning parameters are not both zero and, thus, there exists a unique solution. We first describe how to update one row of $\mathbf{B}$, when all other rows are fixed.

THEOREM 2.1. *Given* $\{B_p\}_{p\neq p_0}$ *in* (2.2), *the solution for* $\min_{B_{p_0}} L(\mathbf{B}; \lambda_1, \lambda_2)$ *is given by* $\widehat{B}_{p_0} = (\widehat{\beta}_{p_0,1}, \ldots, \widehat{\beta}_{p_0,Q})$, *which satisfies, for* $1 \le q \le Q$,

(i) *If* $c_{p_0,q} = 0$, $\widehat{\beta}_{p_0,q} = X_{p_0}^T \widetilde{Y}_q / \|X_{p_0}\|_2^2$ (*OLS*), *where* $\widetilde{Y}_q = Y_q - \sum_{p\neq p_0} X_p \beta_{pq}$;
(ii) *If* $c_{p_0,q} = 1$,

$$\widehat{\beta}_{p_0,q} = \begin{cases} 0, & \text{if } \|\widehat{B}_{p_0}^{\text{lasso}}\|_{2,C} = 0, \\ \left(1 - \dfrac{\lambda_2}{\|\widehat{\mathbf{B}}_{p_0}^{\text{lasso}}\|_{2,C} \cdot \|X_{p_0}\|_2^2}\right)_+ \widehat{\beta}_{p_0,q}^{\text{lasso}}, & \text{otherwise}, \end{cases} \tag{2.3}$$

where

$$\|\widehat{B}_{p_0}^{\text{lasso}}\|_{2,C} := \left\{ \sum_{q=1}^{Q} c_{p_0,q} (\widehat{\beta}_{p_0,q}^{\text{lasso}})^2 \right\}^{1/2},$$

and

$$\widehat{\beta}_{p_0,q}^{\text{lasso}} = \begin{cases} X_{p_0}^T \widetilde{Y}_q / \|X_{p_0}\|_2^2, & \text{if } c_{p_0,q} = 0, \\ (|X_{p_0}^T \widetilde{Y}_q| - \lambda_1)_+ \dfrac{\operatorname{sign}(X_{p_0}^T \widetilde{Y}_q)}{\|X_{p_0}\|_2^2}, & \text{if } c_{p_0,q} = 1. \end{cases} \tag{2.4}$$

The proof of Theorem 2.1 is given in Supplement A [Peng et al. (2009b)].

Theorem 2.1 says that, when estimating the p_0th row of the coefficient matrix $\mathbf{B}$ with all other rows fixed, if there is a prespecified relationship between the p_0th predictor and the qth response (i.e., $c_{p_0,q} = 0$), the corresponding coefficient $\beta_{p_0,q}$ is estimated by the (univariate) ordinary least square solution (OLS) using current responses $\widetilde{Y}_q$; otherwise, we first obtain the lasso solution $\widehat{\beta}_{p_0,q}^{\text{lasso}}$ by the (univariate) soft shrinkage of the OLS solution [equation (2.4)], and then conduct a group shrinkage of the lasso solution [equation (2.3)]. From Theorem 2.1, it is easy to see that, when the design matrix $\mathbf{X}$ is orthonormal, $\mathbf{X}^T\mathbf{X} = I_p$ and $\lambda_1 = 0$, the `remMap` method amounts to selecting variables according to the ℓ_2 norm of the OLS estimates of the corresponding coefficients.

Theorem 2.1 naturally leads to an algorithm which updates the rows of $\mathbf{B}$ iteratively until convergence. In particular, we adopt the `active-shooting` idea proposed by Peng et al. (2009a) and Friedman, Hastie and Tibshirani (2008), which is a modification of the `shooting` algorithm proposed by Fu (1998) and also Friedman, Hastie and Tibshirani (2007), among others. The algorithm proceeds as follows:

1. Initial step: for $p = 1, \ldots, P$; $q = 1, \ldots, Q$,

$$\widehat{\beta}_{p,q}^0 = \begin{cases} X_p^T Y_q / \|X_p\|_2^2, & \text{if } c_{p,q} = 0, \\ (|X_p^T Y_q| - \lambda_1)_+ \dfrac{\operatorname{sign}(X_p^T Y_q)}{\|X_p\|_2^2}, & \text{if } c_{p,q} = 1. \end{cases} \tag{2.5}$$

2. Define the current *active-row set* $\Lambda = \{p : \text{current } \|\widehat{B}_p\|_{2,C} \neq 0\}$.
   (2.1) For each $p \in \Lambda$, update $\widehat{B}_p$ with all other rows of $\mathbf{B}$ fixed at their current values according to Theorem 2.1.
   (2.2) Repeat (2.1) until convergence is achieved on the current active-row set.
3. For $p = 1$ to P, update $\widehat{B}_p$ with all other rows of $\mathbf{B}$ fixed at their current values according to Theorem 2.1. If no $\widehat{B}_p$ changes during this process, return the current $\widehat{\mathbf{B}}$ as the final estimate. Otherwise, go back to step 2.

It is clear that the computational cost of the above algorithm is in the order of $O(NPQ)$.

2.3. *Tuning*. In this section we discuss the selection of the tuning parameters (λ_1, λ_2) by v-fold cross validation. To perform the v-fold cross validation,

we first partition the whole data set into V nonoverlapping subsets, each consisting of approximately $1/V$ fraction of the total samples. Denote the ith subset as $D^{(i)} = (\mathbf{Y}^{(i)}, \mathbf{X}^{(i)})$, and its complement as $D^{-(i)} = (\mathbf{Y}^{-(i)}, \mathbf{X}^{-(i)})$. For a given (λ_1, λ_2), we obtain the `remMap` estimate: $\widehat{\mathbf{B}}^{(i)}(\lambda_1, \lambda_2) = (\widehat{\beta}^{(i)}_{pq})$ based on the ith training set $D^{-(i)}$. We then obtain the *ordinary least square estimates* $\widehat{\mathbf{B}}^{(i)}_{\mathrm{ols}}(\lambda_1, \lambda_2) = (\widehat{\beta}^{(i)}_{\mathrm{ols},pq})$ as follows: for $1 \le q \le Q$, define $S_q = \{p : 1 \le p \le P, \widehat{\beta}^{(i)}_{pq} \ne 0\}$. Then set $\widehat{\beta}^{(i)}_{\mathrm{ols},pq} = 0$ if $p \notin S_q$; otherwise, define $\{\widehat{\beta}^{(i)}_{\mathrm{ols},pq} : p \in S_q\}$ as the ordinary least square estimates by regressing $Y_q^{-(i)}$ onto $\{X_p^{-(i)} : p \in S_q\}$. Finally, prediction error is calculated on the test set $D^{(i)}$:

$$\texttt{remMap.cv}_i(\lambda_1, \lambda_2) := \|\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\widehat{\mathbf{B}}^{(i)}_{\mathrm{ols}}(\lambda_1, \lambda_2)\|_2^2. \tag{2.6}$$

The v-fold cross validation score is then defined as

$$\texttt{remMap.cv}(\lambda_1, \lambda_2) = \sum_{i=1}^{V} \texttt{remMap.cv}_i(\lambda_1, \lambda_2). \tag{2.7}$$

The reason for using OLS estimates in calculating the prediction error is because the true model is assumed to be sparse. As noted by Efron et al. (2004), when there are many noise variables, using shrunken estimates in the cross validation criterion often results in overfitting. Similar results are observed in our simulation studies: if in (2.6) and (2.7), the shrunken estimates are used, the selected models are all very big, which result in large numbers of false positive findings. In addition, we also try AIC and GCV for tuning and both criteria result in overfitting as well. These results are not reported in the next section due to space limitation.

In order to further control the false positive findings, we propose a method called `cv.vote`. The idea is to treat the training data from each cross-validation fold as a "bootstrap" sample. Then variables being consistently selected by many cross validation folds should be more likely to appear in the true model than the variables being selected only by few cross validation folds. Specifically, for $1 \le p \le P$ and $1 \le q \le Q$, define

$$s_{pq}(\lambda_1, \lambda_2) = \begin{cases} 1, & \text{if } \sum_{i=1}^{V} I\big(\widehat{\beta}^{(i)}_{pq}(\lambda_1, \lambda_2) \ne 0\big)/V > V_a, \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

where V_a is a prespecified proportion. We then select edge (p, q) if $s_{pq}(\lambda_1, \lambda_2) = 1$. In the next section we use $V_a = 0.5$ and, thus, `cv.vote` amounts to a "majority vote" procedure. Simulation studies in Section 3 suggest that `cv.vote` can effectively decrease the number of false positive findings while only slightly increasing the number of false negatives.

An alternative tuning method is by a BIC criterion. Compared to v-fold cross validation, BIC is computationally cheaper. However, it requires many more assumptions. In particular, the BIC method uses the degrees of freedom of each

remMap model which is difficult to estimate in general. In Supplement B [Peng et al. (2009b)], we derive an unbiased estimator for the degrees of freedom of the remMap models when the predictor matrix $\mathbf{X}$ has orthonormal columns. In Section 3 we show by extensive simulation studies that, when the correlations among the predictors are complicated, this estimator tends to select very small models. For more details, see Supplement B [Peng et al. (2009b)].

3. Simulation. In this section we investigate the performance of the remMap model and compare it with two alternatives: (i) the joint model with $\lambda_2 = 0$ in (2.2); (ii) the sep model which performs Q separate lasso regressions. For each model, we consider three tuning strategies, which results in nine methods in total:

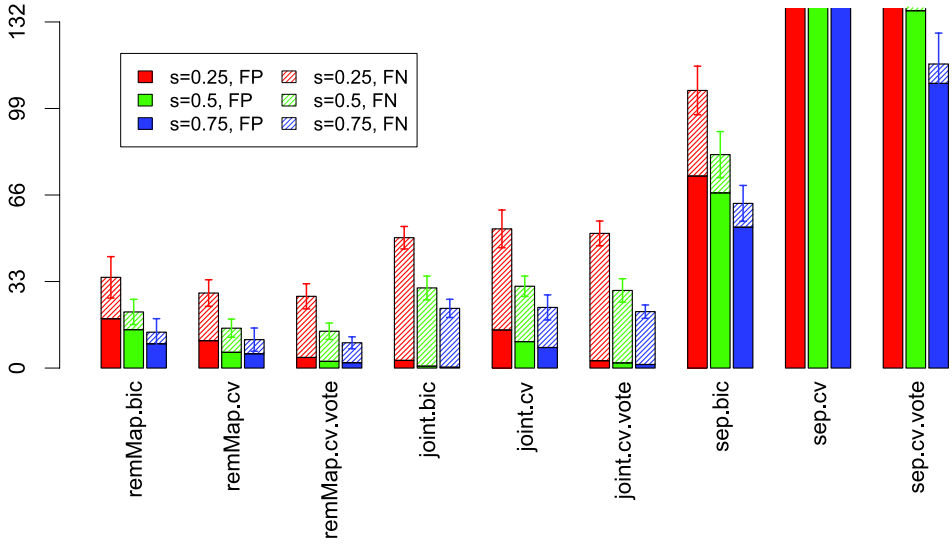
1. remMap.cv, joint.cv, sep.cv: The tuning parameters are selected through 10-fold cross validation;
2. remMap.cv.vote, joint.cv.vote, sep.cv.vote: The cv.vote procedure with $V_a = 0.5$ is applied to the models resulted from the corresponding $*$.cv approaches;
3. remMap.bic, joint.bic, sep.bic: The tuning parameters are selected by a BIC criterion. For remMap.bic and joint.bic, the degrees of freedom are estimated according to equation (S-6) in Supplement B [Peng et al. (2009b)]; for sep.bic, the degrees of freedom of each regression are estimated by the total number of selected predictors [Zou, Hastie and Tibshirani (2007)].

We simulate data as follows. Given (N, P, Q), we first generate the predictors $(x_1, \ldots, x_P)^T \sim \text{Normal}_P(0, \Sigma_X)$, where Σ_X is the predictor covariance matrix [for simulations 1 and 2, $\Sigma_X(p, p') := \rho_x^{|p-p'|}$]. Next, we simulate a $P \times Q$ 0–1 adjacency matrix $\mathbf{A}$, which specifies the topology of the network between predictors and responses, with $\mathbf{A}(p, q) = 1$ meaning that x_p influences y_q or, equivalently, $\beta_{pq} \neq 0$. In all simulations, we set $P = Q$ and the diagonals of $\mathbf{A}$ equal to one, which is viewed as prior information (thus, the diagonals of $\mathbf{C}$ are set to zero). This aims to mimic cis-regulations of DNA copy number alternations on its own expression levels. We then simulate the $P \times Q$ regression coefficient matrix $\mathbf{B} = (\beta_{pq})$ by setting $\beta_{pq} = 0$, if $\mathbf{A}(p, q) = 0$; and $\beta_{pq} \sim \text{Uniform}([-5, -1] \cup [1, 5])$, if $\mathbf{A}(p, q) = 1$. After that, we generate the residuals $(\epsilon_1, \ldots, \epsilon_Q)^T \sim \text{Normal}_Q(0, \Sigma_\epsilon)$, where $\Sigma_\epsilon(q, q') = \sigma_\epsilon^2 \rho_\epsilon^{|q-q'|}$. The residual variance σ_ϵ^2 is chosen such that the average signal to noise ratio equals a prespecified level s. Finally, the responses $(y_1, \ldots, y_Q)^T$ are generated according to model (2.1). Each data set consists of N i.i.d. samples of such generated predictors and responses. For all methods, predictors and responses are standardized to have (sample) mean zero and standard deviation one before model fitting. Results reported for each simulation setting are averaged over 25 independent data sets.

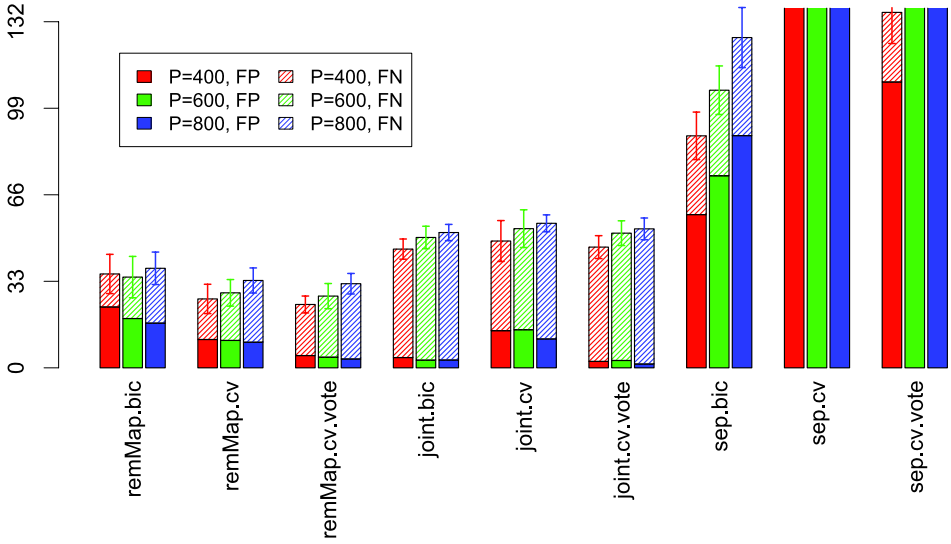
For all simulation settings, $\mathbf{C} = (c_{pq})$ is taken to be $c_{pq} = 0$, if $p = q$; and $c_{pq} = 1$, otherwise. Our primary goal is to identify the `trans-edges`—the predictor-response pairs (x_p, y_q) with $\mathbf{A}(p, q) = 1$ and $\mathbf{C}(p, q) = 1$, that is, the edges that are not prespecified by the indicator matrix $\mathbf{C}$. Thus, in the following, we report the number of false positive detections of `trans-edges` (FP) and the number of false negative detections of `trans-edges` (FN) for each method. We also examine these methods in terms of predictor selection. Specifically, a predictor is called a `cis-predictor` if it does not have any `trans-edges`; otherwise, it is called a `trans-predictor`. Moreover, we say a false positive trans-predictor (FPP) occurs if a `cis-predictor` is incorrectly identified as a `trans-predictor`; we say a false negative trans-predictor (FNP) occurs if it is the other way around.

SIMULATION I. We first assess the performances of the nine methods under various combinations of model parameters. Specifically, we consider the following: $P = Q = 400, 600, 800$; $s = 0.25, 0.5, 0.75$; $\rho_x = 0, 0.4, 0.8$; and $\rho_\epsilon = 0, 0.4, 0.8$. For all settings, the sample size N is fixed at 200. The networks (adjacency matrices $\mathbf{A}$) are generated with 5 master predictors (hubs), each influencing $20 \sim 40$ responses; and all other predictors are `cis-predictors`. We set the total number of `tran-edges` to be 132 for all networks. Results on `trans-edge` detection are summarized in Figures 1 and 2. From these figures, it is clear that `remMap.cv` and `remMap.cv.vote` perform the best in terms of the total number of false detections (FP+FN), followed by `remMap.bic`. The three `sep` methods result in too many false positives (especially `sep.cv`). This is expected since there are in total Q tuning parameters selected separately, and the relations among responses are not utilized at all. This leads to high variability and overfitting. The three `joint` methods perform reasonably well, though they have considerably larger numbers of false negative detections compared to `remMap` methods. This is because the `joint` methods incorporate less information about the relations among the responses caused by the master predictors. Finally, comparing `cv.vote` to `cv`, we can see that the `cv.vote` procedure effectively decreases the false positive detections and only slightly inflates the false negative counts.

As to the impact of different model parameters, signal size s plays an important role for all methods: the larger the signal size, the better these methods perform [Figure 1(a)]. Dimensionality (P, Q) also shows consistent impacts on these methods: the larger the dimension, the more false negative detections [Figure 1(b)]. With increasing predictor correlation ρ_x, both `remMAP.bic` and `joint.bic` tend to select smaller models, and consequently result in less false positives and more false negatives [Figure 2(a)]. This is because when the design matrix $\mathbf{X}$ is further away from orthogonality, (S-6) in Supplement B [Peng et al. (2009b)] tends to overestimate the degrees of freedom and consequently



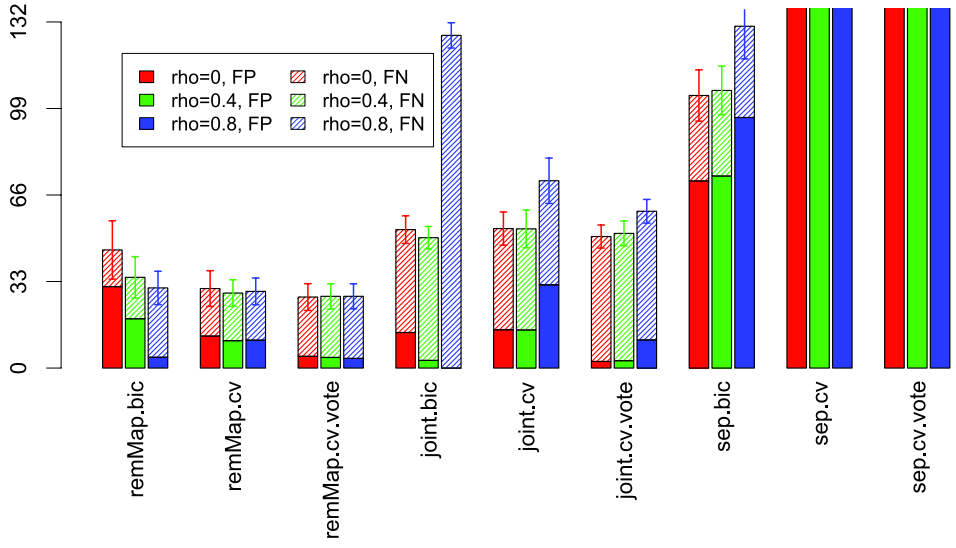
(a) Impact of signal size s. $P = Q = 600$, $N = 200$; $\rho_x = 0.4$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.



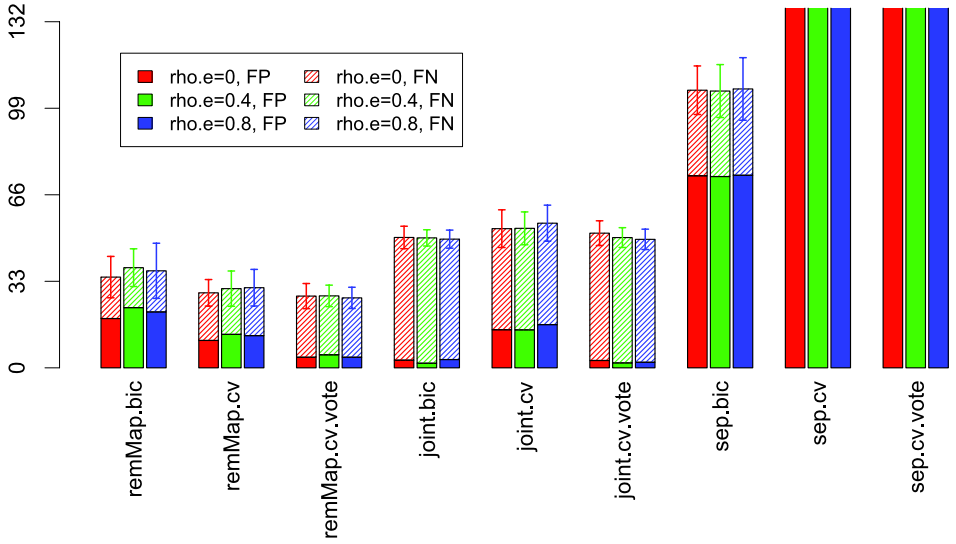
(b) Impact of predictor and response dimensionality P ($Q = P$). $N = 200$; $s = 0.25$; $\rho_x = 0.4$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.

FIG. 1. *Impact of signal size and dimensionality. Heights of solid bars represent numbers of false positive detections of* `trans-edges` *(FP); heights of shaded bars represent numbers of false negative detections of* `trans-edges` *(FN). All bars are truncated at height* = 132.

smaller models are selected. The residual correlation ρ_ϵ seems to have little impact on `joint` and `sep`, and some (though rather small) impacts on `remMap`



(a) Impact of predictor correlation ρ_x. $P = Q = 600$, $N = 200$; $s = 0.25$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.



(b) Impact of residual correlation ρ_ε. $P = Q = 600$, $N = 200$; $s = 0.25$; $\rho_x = 0.4$; the total number of `trans-edges` is 132.

FIG. 2. *Impact of correlations. Heights of solid bars represent numbers of false positive detections of* `trans-edges` *(FP); heights of shaded bars represent numbers of false negative detections of* `trans-edges` *(FN). All bars are truncated at height* = 132.

[Figure 2(b)]. Moreover, `remMap` performs much better than `joint` and `sep` on master predictor selection, especially in terms of the number of false pos-

TABLE 1
Simulation II. Network topology: *uniform network with* 151 `trans-edges` *and* 60 `trans-predictors`. $P = Q = 600$, $N = 200$; $s = 0.25$; $\rho_x = 0.4$; $\rho_\epsilon = 0$

Method	FP	FN	TF	FPP	FNP
`remMap.bic`	4.72 (2.81)	45.88 (4.5)	50.6 (4.22)	1.36 (1.63)	11 (1.94)
`remMap.cv`	18.32 (11.45)	40.56 (5.35)	58.88 (9.01)	6.52 (5.07)	9.2 (2)
`remMap.cv.vote`	2.8 (2.92)	50.32 (5.38)	53.12 (3.94)	0.88 (1.26)	12.08 (1.89)
`joint.bic`	5.04 (2.68)	52.92 (3.6)	57.96 (4.32)	4.72 (2.64)	9.52 (1.66)
`joint.cv`	16.96 (10.26)	46.6 (5.33)	63.56 (7.93)	15.36 (8.84)	7.64 (2.12)
`joint.cv.vote`	2.8 (2.88)	56.28 (5.35)	59.08 (4.04)	2.64 (2.92)	10.40 (2.08)
`sep.bic`	78.92 (8.99)	37.44 (3.99)	116.36 (9.15)	67.2 (8.38)	5.12 (1.72)
`sep.cv`	240.48 (29.93)	32.4 (3.89)	272.88 (30.18)	179.12 (18.48)	2.96 (1.51)
`sep.cv.vote`	171.00 (20.46)	33.04 (3.89)	204.04 (20.99)	134.24 (14.7)	3.6 (1.50)

FP: false positive; FN: false negative; TF: total false; FPP: false positive trans-predictor; FNP: false negative trans-predictor. Numbers in the parentheses are standard deviations.

itive `trans-predictors` (results not shown). This is because the ℓ_2 norm penalty is more effective than the ℓ_1 norm penalty in excluding irrelevant predictors.

SIMULATION II. In this simulation we study the performance of these methods on a network without big hubs. The data are generated similarly as before with $P = Q = 600$, $N = 200$, $s = 0.25$, $\rho_x = 0.4$, and $\rho_\epsilon = 0$. The network consists of 540 `cis-predictors`, and 60 `trans-predictors` with $1 \sim 4$ `trans-edges`. This leads to 151 `trans-edges` in total. As can be seen from Table 1, `remMap` methods and `joint` methods now perform very similarly and both are considerably better than the `sep` methods. Indeed, under this setting, λ_2 is selected (either by `cv` or `bic`) to be small in the `remMap` model, making it very close to the `joint` model.

SIMULATION III. In this simulation we try to mimic the true predictor covariance and network topology in the real data discussed in the next section. We observe that, for chromosomal regions on the same chromosome, the corresponding copy numbers are usually positively correlated, and the magnitude of the correlation decays slowly with genetic distance. On the other hand, if two regions are on different chromosomes, the correlation between their copy numbers could be either positive or negative and, in general, the magnitude is much smaller than that of the regions on the same chromosome. Thus, in this simulation, we first partition the P predictors into 23 distinct blocks, with the size of the ith block proportional to the number of CNAI (copy number alteration intervals) on the ith

chromosome of the real data (see Section 4 for the definition of CNAI). Denote the predictors within the ith block as $x_{i1}, \ldots, x_{ig_i}$, where g_i is the size of the ith block. We then define the *within-block* correlation as $\mathrm{Corr}(x_{ij}, x_{il}) = \rho_{\mathrm{wb}}^{0.5|j-l|}$ for $1 \le j, l \le g_i$; and define the *between-block* correlation as $\mathrm{Corr}(x_{ij}, x_{kl}) \equiv \rho_{ik}$ for $1 \le j \le g_i, 1 \le l \le g_k$ and $1 \le i \ne k \le 23$. Here, ρ_{ik} is determined in the following way: its sign is randomly generated from $\{-1, 1\}$; its magnitude is randomly generated from $\{\rho_{\mathrm{bb}}, \rho_{\mathrm{bb}}^2, \ldots, \rho_{\mathrm{bb}}^{23}\}$. In this simulation we set $\rho_{\mathrm{wb}} = 0.9$, $\rho_{\mathrm{bb}} = 0.25$ and use $P = Q = 600$, $N = 200$, $s = 0.5$, and $\rho_\epsilon = 0.4$. The heatmaps of the (sample) correlation matrix of the predictors in the simulated data and that in the real data are given by Figure S-2 in the Supplement [Peng et al. (2009b)]. The network is generated with five large hub predictors each having $14 \sim 26$ `trans-edges`; five small hub predictors each having $3 \sim 4$ `trans-edges`; 20 predictors having $1 \sim 2$ `trans-edges`; and all other predictors being `cis-predictors`.

The results are summarized in Table 2. Among the nine methods, `remMap.cv.vote` performs the best in terms of both edge detection and master predictor prediction. `remMAP.bic` and `joint.bic` result in very small models due to the complicated correlation structure among the predictors. While all three cross-validation based methods have large numbers of false positive findings, the three `cv.vote` methods have many reduced false positive counts and only slightly increased false negative counts. These findings again suggest that `cv.vote` is an effective procedure in controlling false positive rates while not sacrificing too much in terms of power.

We also carried out an additional simulation where some columns of the coefficient matrix B are related, and the results are reported in Table S-1 of Supple-

TABLE 2

Simulation III. Network topology: five large hubs and five small hubs with 151 `trans-edges` *and* 30 `trans-predictors`. $P = Q = 600$, $N = 200$; $s = 0.5$; $\rho_{\mathrm{wb}} = 0.9$, $\rho_{\mathrm{bb}} = 0.25$; $\rho_\epsilon = 0.4$

Method	**FP**	**FN**	**TF**	**FPP**	**FNP**
`remMap.bic`	0 (0)	150.24 (2.11)	150.24 (2.11)	0 (0)	29.88 (0.33)
`remMap.cv`	93.48 (31.1)	20.4 (3.35)	113.88 (30.33)	15.12 (6.58)	3.88 (1.76)
`remMap.cv.vote`	48.04 (17.85)	27.52 (3.91)	75.56 (17.67)	9.16 (4.13)	5.20 (1.91)
`joint.bic`	7.68 (2.38)	104.16 (3.02)	111.84 (3.62)	7 (2.18)	10.72 (1.31)
`joint.cv`	107.12 (13.14)	39.04 (3.56)	146.16 (13.61)	66.92 (8.88)	1.88 (1.2)
`joint.cv.vote`	63.80 (8.98)	47.44 (3.90)	111.24 (10.63)	41.68 (6.29)	2.88 (1.30)
`sep.bic`	104.96 (10.63)	38.96 (3.48)	143.92 (11.76)	64.84 (6.29)	1.88 (1.17)
`sep.cv`	105.36 (11.51)	37.28 (4.31)	142.64 (12.26)	70.76 (7.52)	1.92 (1.08)
`sep.cv.vote`	84.04 (10.47)	41.44 (4.31)	125.48 (12.37)	57.76 (6.20)	2.4 (1.32)

FP: false positive; FN: false negative; TF: total false; FPP: false positive trans-predictor; FNP: false negative trans-predictor. Numbers in parentheses are standard deviations.

ment C [Peng et al. (2009b)]. The overall picture of the performances of different methods remains similar as other simulations.

4. Real application. In this section we apply the proposed `remMap` method to the breast cancer study mentioned earlier. Our goal is to search for genome regions whose copy number alterations have significant impacts on RNA expression levels, especially on those of the unlinked genes, that is, genes not falling into the same genome region. The findings resulting from this analysis may help to cast light on the complicated interactions among DNA copy numbers and RNA expression levels.

4.1. *Data preprocessing*. The 172 tumor samples were analyzed using cDNA expression microarray and CGH array experiments as described in Sorlie et al. (2001, 2003), Zhao et al. (2004), Kapp et al. (2006), Bergamaschi et al. (2006, 2008) and Langerod et al. (2007). Below, we outline the data preprocessing steps. More details are provided in Supplement D [Peng et al. (2009b)].

Each CGH array contains measurements ($\log_2$ ratios) on about $17K$ mapped human genes. A positive (negative) measurement suggests a possible copy number gain (loss). After proper normalization, `cghFLasso` [Tibshirani and Wang (2008)] is used to estimate the DNA copy numbers based on array outputs. Then, we derive *copy number alteration intervals* (CNAIs)—basic CNA units (genome regions) in which genes tend to be amplified or deleted at the same time within one sample—by employing the Fixed-Order Clustering (FOC) method [Wang (2004)]. In the end, for each CNAI in each sample, we calculate the mean value of the estimated copy numbers of the genes falling into this CNAI. This results in a 172 (samples) by 384 (CNAIs) numeric matrix.

Each expression array contains measurements for about $18K$ mapped human genes. After global normalization for each array, we also standardize each gene's measurements across 172 samples to median $= 0$ and MAD (median absolute deviation) $= 1$. Then we focus on a set of 654 breast cancer related genes, which is derived based on 7 published breast cancer gene lists [Sorlie et al. (2003), van de Vijver et al. (2002), Chang et al. (2004), Paik et al. (2004), Wang et al. (2005), Sotiriou et al. (2006) and Saal et al. (2007)]. This results in a 172 (samples) by 654 (genes) numeric matrix.

When the copy number change of one CNAI affects the RNA level of an unlinked gene, there are two possibilities: (i) the copy number change directly affects the RNA level of the unlinked gene; (ii) the copy number change first affects the RNA level of an intermediate gene (either linked or unlinked), and then the RNA level of this intermediate gene affects that of the unlinked gene. Figure 3 gives an illustration of these two scenarios. In this study we are more interested in finding the relationships of the first type. Therefore, we first characterize the interactions among RNA levels and then account for these relationships in our model so that we can better infer direct interactions. For this purpose, we apply the `space` (Sparse

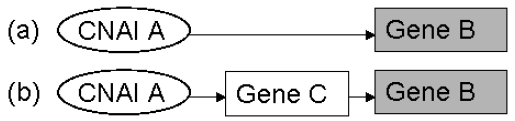


FIG. 3. (a) *Direct interaction between CNAI A and the expression of gene B*; (b) *indirect interaction between CNAI A and the expression of Gene B through one intermediate gene*.

PArtial Correlation Estimation) method to search for associated RNA pairs through identifying nonzero partial correlations [Peng et al. (2009a)]. The estimated (concentration) network (referred to as *Exp.Net.664* hereafter) has in total 664 edges—664 pairs of genes whose RNA levels significantly correlate with each other after accounting for the expression levels of other genes.

Another important factor one needs to consider when studying breast cancer is the existence of distinct tumor subtypes. Population stratification due to these distinct subtypes might confound our detection of associations between CNAIs and gene expressions. Therefore, we introduce a set of subtype indicator variables, which later on is used as additional predictors in the `remMap` model. Specifically, following Sorlie et al. (2003), we divide the 172 patients into 5 distinct groups based on their expression patterns. These groups correspond to the same 5 subtypes suggested by Sorlie et al. (2003)—Luminal Subtype A, Luminal Subtype B, ERBB2-overexpressing Subtype, Basal Subtype and Normal Breast-like Subtype.

4.2. *Interactions between CNAIs and RNA expressions*. We then apply the `remMap` method to study the interactions between CNAIs and RNA transcript levels. For each of the 654 breast cancer genes, we regress its expression level on three sets of predictors: (i) expression levels of other genes that are connected to the target gene (the current response variable) in *Exp.Net.664*; (ii) the five subtype indicator variables derived in the previous section; and (iii) the copy numbers of all 384 CNAIs. We are interested in whether any unlinked CNAIs are selected into this regression model (i.e., the corresponding regression coefficients are nonzero). This suggests potential trans-regulations (`trans-edges`) between the selected CNAIs and the target gene expression. The coefficients of the linked CNAI of the target gene are not included in the `MAP` penalty (this corresponds to $c_{pq} = 0$; see Section 2 for details). This is because the DNA copy number changes of one gene often influence its own expression level, and we are less interested in this kind of cis-regulatory relationships (`cis-edges`) here. Furthermore, based on *Exp.Net.664*, no penalties are imposed on the expression levels of connected genes either. In other words, we view the cis-regulations between CNAIs and their linked expression levels, as well as the inferred RNA interaction network, as "prior knowledge" in our study.

Note that, different response variables (gene expressions) now have different sets of predictors, as their neighborhoods in Exp.Net.664 are different. However,

the remMap model can still be fitted with a slight modification. The idea is to treat all CNAI (384 in total), all gene expressions (654 in total), as well as the subtype indicators, as nominal predictors. Then, for each target gene, we force the coefficients of those gene expressions that do not link to it in *Exp.Net.664* to be zero. We can easily achieve this by setting those coefficients to zero without updating them throughout the iterative fitting procedure.

We select tuning parameters (λ_1, λ_2) in the remMap model through a 10-fold cross validation as described in Section 2.3. The optimal (λ_1, λ_2) corresponding to the smallest CV score from a grid search is $(355.1, 266.7)$. The resulting model contains 56 trans-regulations in total. In order to further control false positive findings, we apply the cv.vote procedure with $V_a = 0.5$, and filter away 13 out of these 56 trans-edges which have not been consistently selected across different CV folds. The remaining 43 trans-edges correspond to three contiguous CNAIs on chromosome 17 and 31 distinct (unlinked) RNAs. Figure 4 illustrates the topology of the estimated regulatory relationships. The detailed annotations of the three CNAIs and 31 RNAs are provided in Tables 3 and 4. Moreover, the

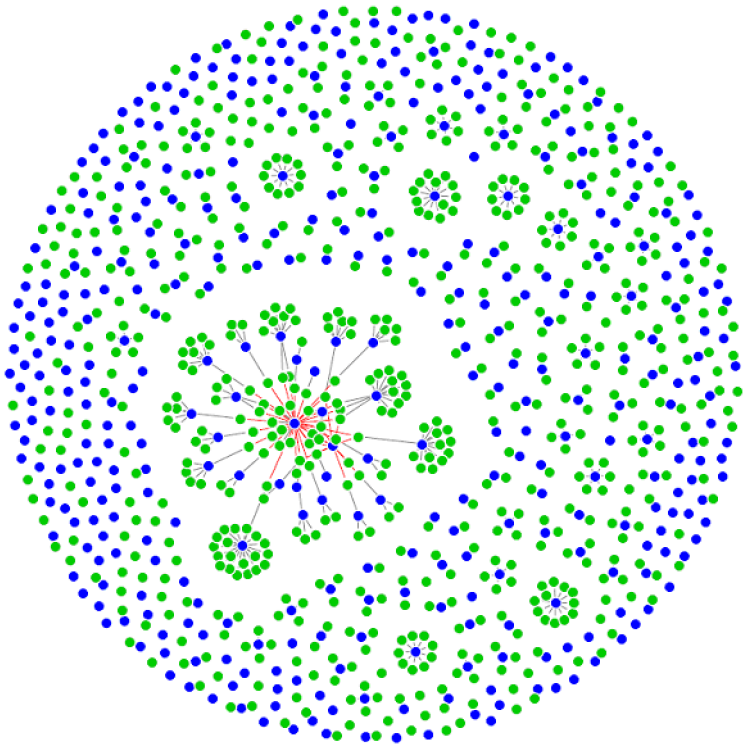


FIG. 4. *Network of the estimated regulatory relationships between the copy numbers of the* 384 *CNAIs and the expressions of the* 654 *breast cancer related genes. Each blue node stands for one CNAI, and each green node stands for one gene. Red edges represent inferred trans-regulations* (43 *in total*). *Grey edges represent cis-regulations.*

TABLE 3
Genome locations of the three CNAIs having (estimated) trans-regulations

Index	Cytoband	Begin[1]	End[1]	# of clones[2]	# of Trans-Reg[3]
1	17q12-17q12	34811630	34811630	1	12
2	17q12-17q12	34944071	35154416	9	30
3	17q21.1-17q21.2	35493689	35699243	7	1

1. Nucleotide position (bp).
2. Number of genes/clones on the array falling into the CNAI.
3. Number of unlinked genes whose expressions are estimated to be regulated by the CNAI.

Pearson-correlations between the DNA copy numbers of CNAIs and the expression levels of the regulated genes/clones (including both `cis-regulation` and `trans-regulation`) across the 172 samples are reported in Table 4. As expected, all the cis-regulations have much higher correlations than the potential trans-regulations. In addition, none of the subtype indicator variables are selected into the final model. We also apply the `remMap` model while forcing these indicators in the model (i.e., not imposing the `MAP` penalty on these variables). Even though this results in a slightly different network, the hub CNAIs remain the same as before. These imply that the three hub CNAIs are unlikely due to the stratification of tumor subtypes.

The three CNAIs being identified as trans-regulators sit closely on chromosome 17, spanning from 34811630*bp* to 35699243*bp* and falling into cytoband 17q12-q21.2. This region (referred to as CNAI-17q12 hereafter) contains 24 known genes, including the famous breast cancer oncogene ERBB2, and the growth factor receptor-bound protein 7 (GRB7). The overexpression of GRB7 plays pivotal roles in activating signal transduction and promoting tumor growth in breast cancer cells with chromosome 17q11-21 amplification [Bai and Louh (2008)]. In this study CNAI-17q12 is highly amplified (normalized $\log_2$ ratio> 5) in 33 (19%) out of the 172 tumor samples. Among the 654 genes/clones considered in the above analysis, 8 clones (corresponding to six genes, including ERBB2, GRB7 and MED24) fall into this region. The expressions of these 8 clones are all up-regulated by the amplification of CNAI-17q12 (see Table 4 for more details), which is consistent with results reported in the literature [Kao and Pollack (2006)]. More importantly, as suggested by the result of the `remMap` model, the amplification of CNAI-17q12 also influences the expression levels of 31 unlinked genes/clones. This implies that CNAI-17q12 may harbor transcriptional factors whose activities closely relate to breast cancer. Indeed, there are 4 transcription factors (NEUROD2, IKZF3, THRA, NR1D1) and 2 transcriptional co-activators (MED1, MED24) in CNAI-17q12. It is possible that the amplification of CNAI-17q12 results in the overexpression of one or more transcription factors/co-activators in this region, which then influence the expressions of the

TABLE 4
RNAs[1] *being influenced by the amplifications of the three CNAIs in Table 3*

Clone ID	Gene symbol	Cytoband	Correlation
753692	ABLIM1	10q25	0.199
896962	ACADS	12q22-qter	−0.22
753400	ACTL6A	3q26.33	0.155
472185	ADAMTS1	21q21.2	0.214
210687	AGTR1	3q21-q25	−0.182
856519	ALDH3A2	17p11.2	−0.244
270535	BM466581	19	0.03
238907	CABC1	1q42.13	−0.174
773301	CDH3	16q22.1	0.118
505576	CORIN	4p13-p12	0.196
223350	CP	3q23-q25	0.184
810463	DHRS7B	17p12	−0.151
50582	FLJ25076	5p15.31	0.086
669443	HSF2	6q22.31	0.207
743220	JMJD4	1q42.13	−0.19
43977	KIAA0182	16q24.1	0.259
810891	LAMA5	20q13.2-q13.3	0.269
247230	MARVELD2	5q13.2	−0.214
812088	NLN	5q12.3	0.093
257197	NRBF2	10q21.2	0.275
782449	PCBP2	12q13.12-q13.13	−0.079
796398	PEG3	19q13.4	0.169
293950	PIP5K1A	1q22-q24	−0.242
128302	PTMS	12p13	−0.248
146123	PTPRK	6q22.2-q22.3	0.218
811066	RNF41	12q13.2	−0.247
773344	SLC16A2	Xq13.2	0.24
1031045	SLC4A3	2q36	0.179
141972	STT3A	11q23.3	0.182
454083	TMPO	12q22	0.175
825451	USO1	4q21.1	0.204
68400	BM455010	17	0.748
756253, 365147	ERBB2	17q11.2-q12–17q21.1	0.589
510318, 236059	GRB7	17q12	0.675
245198	MED24	17q21.1	0.367
825577	STARD3	17q11-q12	0.664
782756[2]	TBPL1	6q22.1-q22.3	0.658

1. The first part of the table lists the inferred trans-regulated genes. The second part of the table lists cis-regulated genes.
2. This cDNA sequence probe is annotated with `TBPL1`, but actually maps to one of the 17q21.2 genes.

unlinked 31 genes/clones. In addition, some of the 31 genes/clones have been reported to have functions directly related to cancer and may serve as potential drug targets (see Supplement D.5 [Peng et al. (2009b)] for more details). In the end, we want to point out that, besides RNA interactions and subtype stratification, there could be other unaccounted confounding factors. Therefore, caution must be applied when one tries to interpret these results.

5. Discussion. In this paper we propose the `remMap` method for fitting multivariate regression models under the large P, Q setting. We focus on model selection, that is, the identification of relevant predictors for each response variable. `remMap` is motivated by the rising needs to investigate the regulatory relationships between different biological molecules based on multiple types of high dimensional omics data. Such genetic regulatory networks are usually intrinsically sparse and harbor hub structures. Identifying the hub regulators (master regulators) is of particular interest, as they play crucial roles in shaping network functionality. To tackle these challenges, `remMap` utilizes a `MAP` penalty, which consists of an ℓ_1 norm part for controlling the overall sparsity of the network, and an ℓ_2 norm part for further imposing a row-sparsity of the coefficient matrix, which facilitates the detection of master predictors (regulators). This combined regularization takes into account both model interpretability and computational tractability. Since the `MAP` penalty is imposed on the coefficient matrix as a whole, it helps to borrow information across different regressions. As illustrated in Section 3, this type of "joint" modeling greatly improves model efficiency. Also, the combined ℓ_1 and ℓ_2 norm penalty further enhances the performance on both edge detection and master predictor identification. We also propose a `cv.vote` procedure to make better use of the cross validation results. As suggested by the simulation study, this procedure is very effective in decreasing the number of false positives while only slightly increasing the number of false negatives. Moreover, `cv.vote` can be applied to a broad range of model selection problems when cross validation is employed. In the real application, we apply the `remMap` method on a breast cancer data set. The resulting model suggests the existence of a trans-hub region on cytoband 17q12-q21. This region harbors the oncogene ERBB2 and may also harbor other important transcriptional factors. While our findings are intriguing, clearly additional investigation is warranted. One way to verify the above conjecture is through a sequence analysis to search for common motifs in the upstream regions of the 31 RNA transcripts, which remains as our future work.

Besides the above application, the `remMap` model can be applied to investigate the regulatory relationships between other types of biological molecules. For example, it is of great interest to understand the influence of single nucleotide polymorphism (SNP) on RNA transcript levels, as well as the influence of RNA transcript levels on protein expression levels. Such investigation will improve our understanding of related biological systems as well as disease pathology. In addition, we can utilize the `remMap` idea to other models. For example, when selecting a group of variables in a multiple regression model, we can impose both

the ℓ_2 penalty (that is, the `group lasso` penalty), as well as an ℓ_1 penalty, to encourage within group sparsity. Similarly, the `remMap` idea can also be applied to vector autoregressive models and generalized linear models.

R package `remMap` is publicly available through CRAN (http://cran.r-project.org/).

Acknowledgments. We are grateful to two anonymous reviewers for their valuable comments.

SUPPLEMENTARY MATERIAL

Supplement A–D [Peng et al. (2009b)] (DOI: 10.1214/09-AOAS271SUPP; .pdf).

Supplement A: This section provides the detailed proof of Theorem 2.1.

Supplement B: In this section we describe the BIC criterion for selecting $(\lambda_1; \lambda_2)$. We also derive an unbiased estimator of the degrees of freedom of the `remMap` estimator under orthogonal design.

Supplement C: This section contains one simulation study. We consider the scenario where some columns of the coefficient matrix B are dependent.

Supplement D: In this section we describe the preprocessing analysis before fitting the `remMap` model on the real data set.

REFERENCES

ALBERTSON, D. G., COLLINS, C., MCCORMICK, F. and GRAY, J. W. (2003). Chromosome aberrations in solid tumors. *Nature Genetics* **34** 369–376.

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

BAI, T. and LUOH, S. W. (2008). GRB-7 facilitates HER-2/Neu-mediated signal transduction and tumor formation. *Carcinogenesis* **29** 473–479.

BAKIN, S. (1999). Adaptive regression and model selection in data mining problems. Ph.D. thesis, Australian National Univ., Canberra.

BEDRICK, E. and TSAI, C. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50** 226–231.

BERGAMASCHI, A., KIM, Y. H., WANG, P., SORLIE, T., HERNANDEZ-BOUSSARD, T., LONNING, P. E., TIBSHIRANI, R., BORRESEN-DALE, A. L. and POLLACK, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* **45** 1033–1040.

BERGAMASCHI, A., KIM, Y. H., KWEI, K. A., CHOI, Y. L., BOCANEGRA, M., LANGEROD, A., HAN, W., NOH, D. Y., HUNTSMAN, D. G., JEFFREY, S. S., BORRESEN-DALE, A. L. and POLLACK, J. R. (2008). CAMK1D amplification implicated in epithelial-mesenchymal transition in basal-like breast cancer. *Mol. Oncol.* **2** 327–339.

BREIMAN, L. and FRIEDMAN, J. H. (1997). Predicting multivariate responses in multiple linear regression (with discussion). *J. Roy Statist. Soc. Ser. B* **59** 3–54. MR1436554

BROWN, P., FEARN, T. and VANNUCCI, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika* **86** 635–648. MR1723783

BROWN, P., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B* **60** 627–641. MR1626005

BROWN, P., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. Roy. Statist. Soc. Ser. B* **64** 519–536. MR1924304

CHANG, H. Y., SNEDDON, J. B., ALIZADEH, A. A., SOOD, R., WEST, R. B., MONTGOMERY, K., CHI, J. T., VAN DE RIJN, M., BOTSTEIN, D. and BROWN, P. O. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biol.* **2**.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FRANK, I. and FRIEDMAN, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

FU, W. (1998). Penalized regressions: The bridge vs the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Regularized paths for generalized linear models via coordinate descent. Technical report, Dept. Statistics, Stanford Univ.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. MR2415737

FUJIKOSHI, Y. and SATOH, K. (1997). Modified AIC and Cp in multivariate linear regression. *Biometrika* **84** 707–716. MR1603952

GARDNER, T. S., DI BERNARDO, D., LORENZ, D. and COLLINS, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301** 102–105.

HYMAN, E., KAURANIEMI, P., HAUTANIEMI, S., WOLF, M., MOUSSES, S., ROZENBLUM, E., RINGNER, M., SAUTER, G., MONNI, O., ELKAHLOUN, A., KALLIONIEMI, O.-P. and KALLIONIEMI, A. (2002). Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res.* **62** 6240–6245.

IZENMAN, A. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264. MR0373179

JEONG, H., MASON, S. P., BARABASI, A. L. and OLTVAI, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411** 41–42.

KAPP, A. V., JEFFREY, S. S., LANGEROD, A., BORRESEN-DALE, A. L., HAN, W., NOH, D. Y., BUKHOLM, I. R., NICOLAU, M., BROWN, P. O. and TIBSHIRANI, R. (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics* **7** 231.

KAO, J. and POLLACK, J. R. (2006). RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes. *Genes Chromosomes Cancer* **45** 761–769.

KIM, S., SOHN, K.-A. and XING, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* **25** 204–212.

LANGEROD, A., ZHAO, H., BORGAN, O., NESLAND, J. M., BUKHOLM, I. R., IKDAHL, T., KARESEN, R., BORRESEN-DALE, A. L. and JEFFREY, S. S. (2007). TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res.* **9** R30.

LUTZ, R. and BÜHLMANN, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statist. Sinica* **16** 471–494. MR2267246

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Union support recovery in high-dimensional multivariate regression. Available at http://arxiv.org/abs/0808.0711.

PAIK, S., SHAK, S., TANG, G., KIM, C., BAKER, J., CRONIN, M., BAEHNER, F. L. WALKER, M. G., WATSON, D., PARK, T., HILLER, W., FISHER, E. R., WICKERHAM, D. L., BRYANT, J. and WOLMARK, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England J. of Medicine* **351** 2817–2826.

PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009a). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746.

PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D. Y., POLLACK J. R. and WANG, P. (2009b). Supplement to "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer." DOI: 10.1214/09-AOAS271SUPP.

POLLACK, J., SRLIE, T., PEROU, C., REES, C., JEFFREY, S., LONNING, P., TIBSHIRANI, R., BOTSTEIN, D., BRRESEN-DALE, A. and BROWN, P. (2002). Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99** 12963–12968.

REINSEL, G. and VELU, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York. MR1719704

SAAL, L. H., JOHANSSON, P., HOLM, K., GRUVBERGER-SAAL, S. K., SHE, Q. B., MAURER, M., KOUJAK, S., FERRANDO, A. A., MALMSTRÖM, P., MEMEO, L., ISOLA, J., BENDAHL, P., ROSEN, N., HIBSHOOSH, H., RINGNER, M., BORG, A. and PARSONS, R. (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl. Acad. Sci. USA* **104** 7564–7569.

SORLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LØNNING P. E. and BØRRESEN-DALE, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98** 10869–10874.

SORLIE, T., TIBSHIRANI, R., PARKER, J., HASTIE, T., MARRON, J. S., NOBEL, A., DENG, S., JOHNSEN, H., PESICH, R., GEISLER, S., DEMETER, J., PEROU, C. M., LØNNING, P. E., BROWN, P. O., BØRRESEN-DALE, A.-L. and BOTSTEIN, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100** 8418–8423.

SOTIRIOU, C., WIRAPATI, P., LOI, S., HARRIS, A., FOX, S., SMEDS, J., NORDGREN, H., FARMER, P., PRAZ, V., HAIBE-KAINS, B., DESMEDT, C., LARSIMONT, D., CARDOSO, F., PETERSE, H., NUYTEN, D., BUYSE, M., VAN DE VIJVER, M. J., BERGH, J., PICCART, M. and DELORENZI, M. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer. Inst.* **98** 262–272.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* **9** 18–29.

TURLACH, B., VENABLES, W. and WRIGHT, S. (2005). Simultaneous variable selection. *Technometrics* **47** 349–363. MR2164706

WANG, P. (2004). Statistical methods for CGH array analysis. Ph.D. thesis, Stanford Univ.

WANG, Y., KLIJN, J. G., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. and Yu, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365** 671–679.

VAN DE VIJVER, M. J., HE, Y. D., VAN'T VEER, L. J., DAI, H., HART, A. A., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C., MARTON, M. J., PARRISH, M., ATSMA, D., WITTEVEEN, A., GLAS, A., DELAHAYE, L., VAN DER VELDE, T., BARTELINK, H., RODENHUIS, S., RUTGERS, E. T., FRIEND, S. H. and BERNARDS, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England J. of Medicine* **347** 1999–2009.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67. MR2212574

YUAN, M., EKICI, A., LU, Z. and MONTERIO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Statist. Soc. Ser. B* **69** 329–346. MR2323756

ZHAO, H., LANGEROD, A., JI, Y., NOWELS, K. W., NESLAND, J. M., TIBSHIRANI, R., BUKHOLM, I. K., KARESEN, R., BOTSTEIN, D., BORRESEN-DALE, A. L. and JEFFREY, S. S. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell.* **15** 2523–2536.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

J. PENG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA
USA

A. BERGAMASCHI
DEPARTMENT OF GENETICS
INSTITUTE FOR CANCER RESEARCH
RIKSHOSPITALET-RADIUMHOSPITALET
MEDICAL CENTER
OSLO
NORWAY

J. R. POLLACK
DEPARTMENT OF PATHOLOGY
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
USA

J. ZHU
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN
USA

W. HAN
D.-Y. NOH
CANCER RESEARCH INSTITUTE AND
DEPARTMENT OF SURGERY
SEOUL NATIONAL UNIVERSITY
COLLEGE OF MEDICINE
SEOUL
SOUTH KOREA

P. WANG
1100 FAIRVIEW AVE N.
M2-B500
FRED HUTCHINSON CANCER
RESEARCH CENTER
SEATTLE, WASHINGTON 98109
USA
E-MAIL: pwang@fhcrc.org