

## MODELING ASSOCIATION BETWEEN DNA COPY NUMBER AND GENE EXPRESSION WITH CONSTRAINED PIECEWISE LINEAR REGRESSION SPLINES

BY GWENAËL G. R. LEDAY<sup>1</sup>, AAD W. VAN DER VAART,  
WESSEL N. VAN WIERINGEN AND MARK A. VAN DE WIEL

*VU University, Leiden University and VU University Medical Center*

DNA copy number and mRNA expression are widely used data types in cancer studies, which combined provide more insight than separately. Whereas in existing literature the form of the relationship between these two types of markers is fixed a priori, in this paper we model their association. We employ piecewise linear regression splines (PLRS), which combine good interpretation with sufficient flexibility to identify any plausible type of relationship. The specification of the model leads to estimation and model selection in a constrained, nonstandard setting. We provide methodology for testing the effect of DNA on mRNA and choosing the appropriate model. Furthermore, we present a novel approach to obtain reliable confidence bands for constrained PLRS, which incorporates model uncertainty. The procedures are applied to colorectal and breast cancer data. Common assumptions are found to be potentially misleading for biologically relevant genes. More flexible models may bring more insight in the interaction between the two markers.

**1. Introduction.** The genetic material of the human cancer cells often exhibits abnormalities, of which DNA copy number aberrations are a prime example. These aberrations comprise gains and losses of chromosome pieces that are highly variable in size. Thereby, all or parts of a chromosome may have more or less than the two copies received from the parents. Abnormal DNA copy numbers (different from two) may alter expression levels of mRNA transcripts (encoding for functional proteins) that map to the aberration's genomic location. Apart from being concordant (copy number tends to correlate positively with expression level), the form of this association is not established and may even vary per gene. In this paper we use high-throughput data available for tissue-specific samples from unrelated patients to study the relationship between copy number (DNA) and gene expression (mRNA). We employ a wide class of interpretable models to reflect the biological mechanism operating between these two molecular levels and identify relevant markers that may serve as therapeutic targets.

---

Received October 2011; revised August 2012.

<sup>1</sup>Supported by the Center for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO).

*Key words and phrases.* DNA copy number, mRNA expression, regression splines, constrained inference, model selection, confidence bands.

DNA copy number aberrations are often measured by array comparative genomic hybridization (aCGH) [Pinkel and Albertson (2005)]. This measuring device is similar to expression microarrays, which measure expression levels of thousands of genes simultaneously but interrogate DNA rather than RNA. Thereby, both profiling experiments produce a continuous value for every element/probe on the array: a  $\log_2$ -value of optical fluorescence intensity. As experiments appear similar, types of information differ and so are their subsequent treatment. To understand the specific nature of these data, we include a description of their processing.

Normalization of mRNA expression profiles [Quackenbush (2002)] consists in removing experimental artifacts (such as array differences, means, scales) and yields, for every gene on each array, a continuous value (normalized  $\log_2$ -value) which represents the amount of the gene's transcript present in the sample. Preprocessing of copy number/aCGH profiles aims to characterize the genomic instability of each tumor sample and show deleted/duplicated pieces of chromosomes. Three successive steps (illustrated in Figure 1) are typically executed to recover the aberration states of all probes [van de Wiel et al. (2011)]. Through these steps, the size, genomic position and type of copy number aberrations are determined for all samples. In the first preprocessing step, the *normalization* of  $\log_2$ -values removes technical or biological artifacts (such as tumor sample contamination, GC content) and makes the data comparable across samples. Next, *segmentation* partitions the genome of each sample into segments of constant  $\log_2$ -values. These segments are

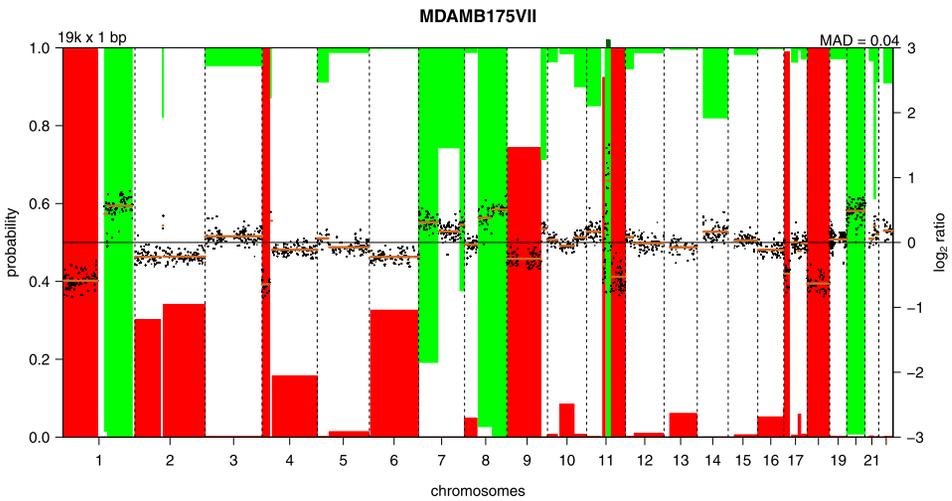


FIG. 1. Plot of a copy number/aCGH profile from the breast cancer data set [Neve et al. (2006)] showing the different preprocessing steps. Probes on the array are genomically ordered on the x-axis (only the chromosome number is displayed). Black dots and orange segments indicate the normalized and segmented  $\log_2$ -values (right y-axis), respectively. Bars represent “loss” (red) and “gain” (green and reversed) membership probabilities (left y-axis). Amplifications are indicated by tick marks on the top axis.

considered a smoothed (and thus de-noised) version of their normalized counterparts. Segmentation is motivated by the biological breakpoint process on the DNA that may cause differential copy number between neighboring locations. Finally, *calling* assigns an aberration state to each segment. Probabilistic calling, usually based on mixture models, results in a probability distribution over a set of ordered possible types of genomic aberrations (which we will refer to as states), typically comprising “loss” (<2 copies), “normal” (=2 copies), “gain” (3–4 copies) and “amplification” (>4 copies). A state is attributed to each probe using a classification rule on the membership probabilities. Nonprobabilistic calling directly assigns states to segmented values, for example, by using a threshold. Note that larger segmented values almost always correspond to a larger or equal called copy number (see Figure 1). All in all, the three steps of the preprocessing procedure provide distinct, but strongly related, data sets: (1) the normalized, (2) segmented and (3) called aCGH data. While most down-stream analyses use either segmented or called data, we use them jointly.

Current methodology for integrative genomic studies assumes rather than explores the mathematical form of the relationship between copy number and expression level. The relationship is said to be either linear or stepwise (see examples in Figure 2). A linear relationship is often assumed in combination with segmented aCGH data. For instance, the strength of the DNA-mRNA association is measured by a (modified) correlation coefficient [Lee, Kong and Park (2008), Lipson et al. (2004), Salari, Tibshirani and Pollack (2010), Schäfer et al. (2009)]. Alternatively, a linear regression approach is entertained [Asimit, Andrulis and Bull (2011), Gu, Choi and Ghosh (2008), Menezes et al. (2009)]. Recently published multivariate methods [Jörnsten et al. (2011), Peng et al. (2010), Sonesson et al. (2010), van Wieringen, Berkhof and van de Wiel (2010)] also assume linearity. A piecewise DNA-mRNA relationship is considered when using the called aCGH data for inte-

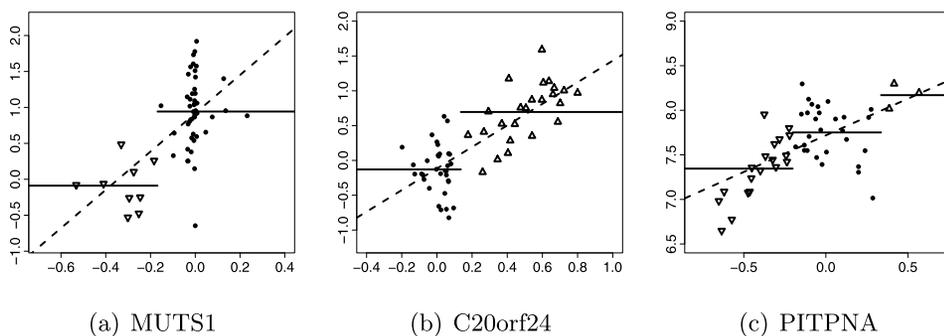


FIG. 2. Illustration of the association between DNA and mRNA for three genes in the breast cancer data set [Neve et al. (2006)] used in this study. Segmented copy number is on the x-axis, while gene expression is on the y-axis. Symbols indicate the different states, namely, loss ( $\nabla$ ), normal ( $\circ$ ) and gain ( $\triangle$ ). The dashed and “continuous” lines give the fitted linear and stepwise models, respectively.

grative analysis. van Wieringen and van de Wiel (2009) and Biciato et al. (2009) have proposed stepwise methods.

In this paper we develop model selection for piecewise linear regression splines (PLRS) to decipher how DNA copy number abnormalities alter the mRNA gene expression level. In addition, we propose a statistical test that accounts for model uncertainty in the PLRS context to detect those genes that drive important shifts. The PLRS framework encompasses the linear and stepwise relationships, but provides flexibility, while maintaining good interpretability. In particular, it accommodates *differential* DNA-mRNA relationships across states. This is biologically plausible, because the cell has various post-transcriptional mechanisms to undo the effects of DNA aberrations. For a given gene, the efficacy of such mechanisms is likely to differ between gains and losses. For example, a gain can directly be compensated by regulatory mechanisms that cause mRNA degradation, such as methylation. On the other hand, a complete loss of both DNA copies (which is more rare than partial loss) cannot be compensated at all.

Segmented and called data are incorporated into the analysis, and biologically motivated constraints are imposed on the model parameters. As this makes model selection and inference nonstandard, we provide methodology for testing the effect of DNA on mRNA within the context of PLRS and for selecting the appropriate model. We also present a novel and computationally inexpensive method for obtaining uniform confidence bands. We apply the proposed methodology to colorectal and breast cancer data sets, where we identify many genes exhibiting nonstandard behavior.

**2. Methods.** We model the association between DNA copy number and mRNA expression by piecewise linear regression splines (PLRS), with biologically motivated constraints on the coefficients. In this section we address model selection and describe a modified Akaike criterion in this context. Further, we present a method for determining uniform confidence bands, along with a statistical test for the effect of copy number on mRNA expression.

**2.1. Model.** Consider gene expression and aCGH profiling of  $n$  independent tumor samples where for a given gene  $\{y_i, x_i, s_i\}_{i=1}^n$  are available, with  $y_i$  being the normalized mRNA expression ( $\log_2$  scale),  $x_i$  the segmented copy number ( $\log_2$  scale) and  $s_i$  the copy number state (“loss,” “normal,” “gain” and “amplification,” coded by  $-1, 0, 1$  and  $2$ ) value of the  $i$ th observation, respectively. Then, the “full” model with  $S$  states (or parts) takes the form

$$(1) \quad y_i = f_\alpha(x_i; \theta) + \varepsilon_i = \theta_0 + \theta_1 x_i + \sum_{j=1}^{S-1} \sum_{d=0}^1 \theta_{j,d} (x_i - \alpha_j)_+^d + \varepsilon_i.$$

Here  $\theta = \{\theta_0, \theta_1, \theta_{1,0}, \dots, \theta_{S-1,0}, \theta_{1,1}, \dots, \theta_{S-1,1}\}$  is a vector of  $2 \times S$  unknown parameters, the  $\varepsilon_i$  are independent random variables each normally distributed

with mean 0 and variance  $\sigma^2$ , and  $\{\alpha_j\}$  are  $S - 1$  known *knots*. The quantity  $(a)_+^d$  represents the positive part  $\max(a, 0)$  of  $a$  raised to the power  $d$ . The number of aberration states  $S$  varies across genes. In this study no more than four different aberration states are considered ( $S \leq 4$ ). Below, for the purpose of discussing model (1) we consider the general case  $S = 4$ .

Knots  $\{\alpha_j\}$  are obtained using data from the *calling* preprocessing step. Depending on the type of calling, two possibilities present themselves. First, consider nonprobabilistic calling which renders states  $\{s_i\}_{i=1}^n$ . Then,  $\alpha_j$  is taken to be the midpoint of the interval between segmented values  $x_i$  belonging to consecutive states (method I). This makes the (natural) supposition that the calling values respect the ordering of the segmented values  $x_i$  and should be reasonably precise if the between-state intervals are small, which is typical (see Figure 2). Second, consider probabilistic calling, which renders membership (or call) probabilities:  $(p_{i,-1}, p_{i,0}, p_{i,1}, p_{i,2})$ . These reflect the plausibility of the segmented value  $x_i$  to belong to the states  $s_i \in \{-1, 0, 1, 2\}$  [van de Wiel et al. (2007)]. Then for  $j \in \{1, 2, 3\}$ , we estimate  $\alpha_j$  (method II) by

$$(2) \quad \hat{\alpha}_j = \arg \max_{\alpha \in \mathbb{R}} \sum_{i=1}^n p_{i,j(i,\alpha)}, \quad j(i, \alpha) = \begin{cases} j - 2, & \text{if } x_i \leq \alpha, \\ j - 1, & \text{if } x_i > \alpha. \end{cases}$$

For instance,  $\alpha_2$  is the knot between states 0 and 1. To determine its position, we select for each sample its plausibility  $p_{i,0}$  of belonging to state 0 (when  $x_i \leq \alpha_2$ ) or  $p_{i,1}$  of belonging to state 1 (when  $x_i > \alpha_2$ ), and add over all samples. We select  $\alpha_2$  to maximize the sum. The maximum may not be unique but described by a small interval; in such a case, we use the corresponding midpoint. This method may be preferable as it accounts for the uncertainty of the calling states. The two methods taken here use data as provided by available *calling* algorithms. Proposed models for this preprocessing step typically depend on data from *all* samples, which stabilizes the estimation of  $\alpha_j$ . Furthermore, knots are to be interpreted as boundaries between the (ordered) states  $\{-1, 0, 1, 2\}$ , which gives us strong a priori knowledge as to their placing (see Figure 2). Together, these two arguments support our approach to consider knots in model (1) as being known. In Section 3 of the supplementary material (SM) [Leday et al. (2013)], a simulation shows that standard deviations of  $\hat{\alpha}_j$  are indeed very small.

Model (1) contains seven basis functions besides the intercept  $\theta_0$  and hence is quite flexible. Our approach is to select appropriate basis functions ( $2^7 = 128$  possible models) and estimate the parameters. The basis functions of degree zero  $x \mapsto (x - \alpha)_+^0$  model discontinuities, and hence allow for a different effect of copy number on expression for each state.

This framework is a natural fundament to test meaningful hypotheses. For example, the hypothesis that for a given state there is an effect of copy number on mRNA can be expressed in terms of a linear function of the parameters being zero ( $\sum_j \theta_{j,1} = 0$ ); a difference between the effects of two adjacent states corresponds

to knot deletion. The submodel consisting of piecewise constant functions [without the functions  $x \mapsto x$  and  $x \mapsto (x - \alpha)_+^1$ ] allows testing the difference in expression between states based on discrete genomic information.

To increase biological plausibility, aid interpretation and increase the stability of estimation, we impose a set of linear constraints on the parameters. As it is generally believed that direct causal effects of DNA on mRNA should be positive, we constrain all slopes to be nonnegative. More exactly, we constrain the slope corresponding to the “normal” state to be nonnegative ( $\theta_1 + \theta_{1,1} \geq 0$ ), while others are forced to be at least equal to the latter (implied by  $\theta_{1,1} \leq 0$  for losses,  $\theta_{2,1} \geq 0$  for gains and  $\theta_{2,1} + \theta_{3,1} \geq 0$  for amplifications). For the same reason we constrain jumps  $\theta_{j,0}$  from state to state to be nonnegative. Note that the restrictions adopted here force the slope of the “normal” state to be small or null and make the natural assumption that a normal copy number is not expected to affect (at least severely) gene expression.

The maximum likelihood estimator of the unknown vector of coefficients  $\theta$  solves the following convex optimization problem:

$$(3) \quad \underset{\theta}{\text{minimize}} (y - \mathbf{X}\theta)^T (y - \mathbf{X}\theta) \quad \text{subject to } \mathbf{C}\theta \geq 0.$$

This can be solved by quadratic programming [Boyd and Vandenberghe (2004)]. The vector  $y = \{y_1, \dots, y_n\}$  denotes the expression signature of a given gene and  $\mathbf{X}$  the associated matrix of covariates designed according to (1). The full row-rank matrix  $\mathbf{C}$  expresses the constraints that are imposed on the parameters. For the 4-state full model we define  $\mathbf{C}$  as the matrix in

$$(4) \quad \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_{1,0} \\ \theta_{1,1} \\ \theta_{2,0} \\ \theta_{2,1} \\ \theta_{3,0} \\ \theta_{3,1} \end{pmatrix} \geq 0.$$

2.2. *Model selection.* Given  $R$  competing statistical models, with log-likelihoods  $\mathcal{L}_r(\theta_r)$  based on a  $k_r \times 1$  parameter vector  $\theta_r$  and with corresponding maximum likelihood estimators (MLE)  $\hat{\theta}_r$ , the Akaike information criterion (AIC) selects as best the model that minimizes

$$(5) \quad \text{AIC}_r = -\mathcal{L}_r(\hat{\theta}_r) + k_r.$$

This information criterion consists of two parts: the negative maximized log-likelihood, which measures the lack of model fit, and a penalty for model complexity. Although AIC has found wide application, it is less suitable for models that include parameter constraints, as in our situation. It can be adapted as follows.

The original motivation for the criterion [Akaike (1973)] is to choose the model that minimizes the Kullback–Leibler (KL) divergence to the true distribution of the data. Indeed, the criterion  $AIC_r$  is (under some conditions) an asymptotically unbiased estimator of this KL divergence. The likelihood at a given parameter is an unbiased estimate of the KL divergence at this parameter, but evaluating it at the maximum likelihood estimator introduces a bias caused by “using the data twice,” which is compensated by the penalty  $k_r$  [Bozdogan (1987)]. In the constrained case (i.e., subject to  $\mathbf{C}\theta \geq 0$ ) we can follow the same motivation, but must account for a different behavior of the maximum likelihood estimator and the resulting bias. Intuitively, the penalty adjusts for an expected increase in the maximized log-likelihood when variables are added to the model, which is less likely under constraints. The likelihood of violation of the constraints must be taken into account.

Hughes and King (2003) adapted the AIC criterion using the asymptotic distribution of the Wald test statistic. In the constrained situation this statistic is not distributed as a chi-squared random variable anymore, but as a probability weighted mixture of chi-squared random variables [see Chernoff (1954), Gouriéroux, Holly and Monfort (1982), Kodde and Palm (1986) or van der Vaart (1998), Theorem 16.7]. It is of the form (partially inequality constrained Wald statistic)

$$(6) \quad \sum_{h=0}^{p_r} w(p_r, h) \chi^2(k_r - p_r + h),$$

where  $p_r$  is the number of inequality constraints and  $w(p_r, h)$  are weights [with  $\sum_h w(p_r, h) = 1$ ], which can be interpreted as the probabilities under the null hypothesis that the constrained maximum likelihood estimator  $\tilde{\theta}_r$  satisfies  $h$  out of  $p_r$  constraints.

Hughes and King (2003) propose to use the one-sided AIC (OSAIC), which is an asymptotically unbiased estimator of the KL divergence in the presence of one-sided information:

$$(7) \quad OSAIC_r = -\mathcal{L}_r(\tilde{\theta}_r) + \sum_{h=0}^{p_r} w(p_r, h)(k_r - p_r + h).$$

Calculating the weights is a combinatorial problem, which aims to determine the probability that the vector  $\tilde{\theta}_r$  lies in any face of dimension  $h$  [Grömping (2010), Kudô (1963), Shapiro (1988)]. This can be computationally intensive as the number of variables,  $k_r$ , increases [Grömping (2010)]. However, in this study the largest model has eight free parameters (because  $S \leq 4$ ). Therefore, the model selection procedure is still very fast (a couple of seconds).

2.3. *Testing.* To evaluate the effect of DNA copy number on expression, we test the hypothesis  $H_0 : \mathbf{C}\theta = 0$  against the alternative  $H_1 : \mathbf{C}\theta \neq 0, \mathbf{C}\theta \geq 0$ , that is,

we test that all inequality constraints are satisfied as equalities against the possibility that at least one of them is strict. From (4) we observe that all parameters except the intercept  $\theta_0$  are subject to inequality constraints and that the null hypothesis reduces the model to the intercept.

We employ the likelihood ratio statistic  $LR = 2(\mathcal{L}_1 - \mathcal{L}_0)$ , where  $\mathcal{L}_0$  and  $\mathcal{L}_1$  are the maximized log-likelihood under the null and alternative hypotheses, respectively. The test rejects the null hypothesis for large values of

$$(8) \quad \min_{C\theta \geq 0} (y - \mathbf{X}\theta)^T (y - \mathbf{X}\theta) - \min_{C\theta = 0} (y - \mathbf{X}\theta)^T (y - \mathbf{X}\theta).$$

This can be shown [Robertson, Wright and Dykstra (1988)] to be equivalent to rejecting for large values of

$$(9) \quad \bar{\chi}^2 = (\tilde{\theta} - \tilde{\theta}_-)^T \Sigma_{\mathbf{X}}^{-1} (\tilde{\theta} - \tilde{\theta}_-),$$

where  $\tilde{\theta}$  and  $\tilde{\theta}_-$  are the maximum likelihood estimators under the inequality and the equality constraints, respectively, and  $\Sigma_{\mathbf{X}} = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$  is the covariance matrix of the unconstrained least squares estimator. For known error variance  $\sigma^2$  the chi-bar-squared statistic  $\bar{\chi}^2$  may be employed with null distribution approximated by a weighted mixture of  $\chi^2$  distributions [Chernoff (1954), Gouriéroux, Holly and Monfort (1982)]. As  $\sigma^2$  is typically unknown, we use instead the so-called *E-bar-squared statistic* [Grömping (2010), Robertson, Wright and Dykstra (1988), Shapiro (1988), Silvapulle and Sen (2005)]

$$(10) \quad \bar{E}^2 = \frac{(\tilde{\theta} - \tilde{\theta}_-)^T \Omega_{\mathbf{X}}^{-1} (\tilde{\theta} - \tilde{\theta}_-)}{(\tilde{\theta} - \tilde{\theta}_-)^T \Omega_{\mathbf{X}}^{-1} (\tilde{\theta} - \tilde{\theta}_-) + (y - \mathbf{X}\hat{\theta})^T (y - \mathbf{X}\hat{\theta})}.$$

Here  $\Omega_{\mathbf{X}} = \mathbf{X}^T \mathbf{X}$ . The null distribution of this statistic is a weighted mixture of Beta distributions of the form

$$(11) \quad \sum_{h=0}^p w(p, h) \mathcal{B}(h/2, (n-p)/2),$$

where  $p$  is the number of parameters and  $\mathcal{B}(a, b)$  refers to a beta distribution with shape parameters  $a$  and  $b$ . The mixing weights are the same as in (6) (applied to the full model); unknown parameters are estimated by their MLEs.

Further details on these test statistics can be found in Robertson, Wright and Dykstra (1988), Shapiro (1988), Silvapulle and Sen (2005).

**2.4. Confidence bands.** Confidence bands (CBs) for the (spline) function  $\mathbf{x} \mapsto f_{\alpha}(\mathbf{x}; \theta)$  in equation (1) should take both the model selection procedure [Buckland, Burnham and Augustin (1997)] and the constraints into account.

Initially we implemented a bootstrap procedure [Grömping (2010)], accounting for model uncertainty along the lines of Burnham and Anderson (2002), who propose the construction of so-called unconditional confidence intervals where only

the selected model is considered for each bootstrap sample. Unfortunately, simulated coverage probabilities were below (and sometimes far below, e.g., 0.6 instead of 0.95) the nominal level, probably due to the presence of the inequality constraints in our model [Andrews (2000)]. We therefore developed an “exact” alternative based on the E-bar-squared statistic (10), using semidefinite programming to achieve computational efficiency. A simulation study reported in Section 3.2 shows that this approach yields accurate uniform CBs.

2.4.1. *Problem formulation.* We start by the construction of a joint confidence region for all parameters  $\theta$  in the full model, including the intercept  $\theta_0$ , by inverting the likelihood ratio test described previously. Analogously to equation (10), define

$$\bar{E}^2(\theta) = \frac{(\tilde{\theta} - \theta)^T \Omega_{\mathbf{X}}^{-1} (\tilde{\theta} - \theta)}{(\tilde{\theta} - \theta)^T \Omega_{\mathbf{X}}^{-1} (\tilde{\theta} - \theta) + (y - \mathbf{X}\hat{\theta})^T (y - \mathbf{X}\hat{\theta})}.$$

Then a  $(1 - \alpha)\%$  confidence region  $\mathcal{R}$  for  $\theta$  is

$$(12) \quad \mathcal{R} = \{\theta : \bar{E}^2(\theta) \leq \mathcal{Q}_{1-\alpha}, \mathbf{C}\theta \geq 0\},$$

where  $\mathcal{Q}_{1-\alpha}$  denotes the  $(1 - \alpha)$ -quantile of the beta mixture distribution in (11). Here we increment the first parameter of the Beta distributions to  $(h + 1)/2$ , because presently we include the intercept as a parameter, whereas before it was free under the null hypothesis. Interval estimation based on inversion of a likelihood ratio statistic is known to possess good properties [Arnold and Shavelle (1998), Brown, Cai and DasGupta (2003), Meeker and Escobar (1995)].

Given the confidence region  $\mathcal{R}$ , we compute a confidence band by determining for each  $\mathbf{x}$  the minimum and maximum values  $f_\alpha(\mathbf{x}; \theta) = \mathbf{x}^T \theta$ . This means determining

$$\inf_{\theta \in \mathcal{R}} \mathbf{x}^T \theta \quad \text{and} \quad \sup_{\theta \in \mathcal{R}} \mathbf{x}^T \theta.$$

Thus, a simple linear function must be minimized (or maximized) subject to linear and ellipsoidal inequality constraints. In the following section, we show that this (convex) problem can be solved efficiently by semidefinite programming.

2.4.2. *Semidefinite programming.* A semidefinite program [Vandenberghe and Boyd (1996)] is concerned with the minimization of a linear objective function under the constraint that a linear combination of symmetric matrices is positive semidefinite:

$$(13) \quad \underset{y \in \mathbb{R}^m}{\text{minimize}} \quad b^T y \quad \text{subject to} \quad F(y) = F_0 + \sum_{i=1}^m y_i F_i \geq 0.$$

The vector  $b \in \mathbb{R}^m$  and the symmetric  $(n \times n)$  matrices  $F_0, \dots, F_m$  are fixed, and the expression  $F(y) \geq 0$  means that the matrix  $F(y)$  is positive semidefinite [i.e.,  $z^T F(y) z \geq 0, \forall z \in \mathbb{R}^n$ ]. Because a linear matrix inequality constraint

$F(y) \geq 0$  is convex, the program can be solved efficiently using interior-point methods [Vandenberghe and Boyd (1996)].

We may express the optimization problem of the previous section as a semidefinite program, based on two equivalences, given by Vandenberghe and Boyd (1996) and provided in Appendix B. For convenience, we replace the ellipsoidal constraint  $\bar{E}^2(\theta) \leq \mathcal{Q}_{1-\alpha}$  by  $(M\theta - M\tilde{\theta})^T(M\theta - M\tilde{\theta}) \leq \lambda$ , where  $\lambda = (y - \mathbf{X}\hat{\theta})^T(y - \mathbf{X}\hat{\theta})\mathcal{Q}_{1-\alpha}/(1 - \mathcal{Q}_{1-\alpha})$  and  $\Omega_{\mathbf{X}}^{-1} = M^T M$ . Given this, the semidefinite program is

$$(14) \quad \underset{\theta}{\text{minimize}} \mathbf{x}^T \theta \quad \text{subject to } F(\theta) = F_0 + \sum_{i=1}^p \theta_i F_i \geq 0,$$

where

$$F_0 = \begin{pmatrix} 0 & 0 \\ 0 & F_0^{(2)} \end{pmatrix}, \quad F_i = \begin{pmatrix} F_i^{(1)} & 0 \\ 0 & F_i^{(2)} \end{pmatrix}, \quad i = 1, \dots, p,$$

with the submatrices defined as

$$F_i^{(1)} = \text{diag}(c_i), \quad F_0^{(2)} = \begin{pmatrix} I & -M\tilde{\theta} \\ (-M\tilde{\theta})^T & \lambda \end{pmatrix} \quad \text{and} \quad F_i^{(2)} = \begin{pmatrix} 0 & m_i \\ m_i^T & 0 \end{pmatrix}.$$

Here  $m_i$  and  $c_i$  denote the  $i$ th column vector of the matrices  $M$  and  $C$  [the matrix of linear restrictions expressed in (3)], respectively.

The optimization procedure needs to be repeated twice in order to determine the lower and upper bounds on  $\mathbf{x}^T \theta$ . Even though this must next be repeated for every new instance  $\mathbf{x}$  to obtain a confidence band, the overall procedure is fast. For instance, for 100 new instances computation on a 2.66 GHz Intel quad-core took less than 12 s (without parallel computing).

**3. Simulation.** We conducted simulation experiments to: (1) determine the accuracy of estimates as provided by PLRS (Section 3.1); (2) examine the coverage probabilities of the method proposed in Section 2.4 (Section 3.2); and (3) evaluate the performance of the PLRS screening test in detecting associations of various functional forms (Section 3.3).

3.1. *Point estimation.* The simulation study examined the accuracy of the estimates obtained by fitting piecewise splines or a simple linear model. For simplicity, we consider a two-state model (normal and gain) and the knot was fixed to 0.5. Data were generated according to the following:

- model 1:  $y = 1 + a_2(x - 0.5)_+^1, a_2 \in \{0, 0.5, 1, 2, 5\}$
- model 2:  $y = 1 + 0.5x + (a_2 - 0.5)(x - 0.5)_+^1, a_2 \in \{0, 0.5, 1, 2, 5\}$ .

The first state (normal) has no or little effect on expression. The linear function is contained in both models, and is found for  $a_2 = 0$  and  $a_2 = 0.5$ , re-

TABLE 1

Squared bias and variance (in parentheses) of the slope estimates of the linear and piecewise spline models as a function of the true slope  $a_2$ , noise  $\sigma$  and model. In bold: setting for which the true model is linear

Model	$a_2$	$\sigma = 0.25$		$\sigma = 0.75$	
		linear	piecewise	linear	piecewise
1	<b>0</b>	<b>0.002 (0.004)</b>	<b>0.007 (0.012)</b>	<b>0.015 (0.033)</b>	<b>0.047 (0.090)</b>
	0.5	0.070 (0.011)	0.002 (0.039)	0.050 (0.060)	0.000 (0.193)
	1	0.282 (0.011)	0.004 (0.045)	0.270 (0.081)	0.008 (0.271)
	2	1.114 (0.011)	0.003 (0.045)	1.124 (0.094)	0.027 (0.339)
	5	6.962 (0.011)	0.003 (0.042)	6.908 (0.103)	0.022 (0.393)
2	0	0.060 (0.008)	0.063 (0.009)	0.075 (0.053)	0.124 (0.097)
	<b>0.5</b>	<b>0.000 (0.009)</b>	<b>0.005 (0.019)</b>	<b>0.000 (0.066)</b>	<b>0.030 (0.146)</b>
	1	0.058 (0.008)	0.000 (0.036)	0.055 (0.070)	0.006 (0.180)
	2	0.545 (0.008)	0.000 (0.041)	0.521 (0.075)	0.000 (0.289)
	5	4.782 (0.008)	0.000 (0.046)	4.857 (0.073)	0.004 (0.320)

spectively. We generated errors from a normal distribution  $\mathcal{N}(0, \sigma^2)$  where  $\sigma \in \{0.1, 0.25, 0.5, 0.75, 1\}$ . This resulted in 25 cases for each of the two models (5 values of  $a_2$  times 5 values of  $\sigma$ ). The sample size was set to 80, and the 80 values of  $x$  were generated from a uniform distribution  $\mathcal{U}(0, 1)$ .

We were interested in comparing the precision of the estimates of the slope  $a_2$  when fitting a linear or a piecewise linear model (the latter with a single knot placed at 0.5; 4 parameters). For each of the 25 cases we repeated the simulation experiment 1000 times and computed the estimator of the slope for both models. Table 1 reports the empirical squared bias and variance over the 1000 repetitions. For clarity only the results for  $\sigma = 0.25$  and  $\sigma = 0.75$  are displayed. Complementary results can be found in Section 2 of SM.

Not surprisingly, the piecewise model can capture the relationship well in all cases: the squared bias is small, and the variance never unduly large. On the other hand, the estimate of the slope given by the linear model is strongly biased for larger values of the slope  $a_2$ . As expected, the variance of the PLRS estimate is usually somewhat larger than that of the linear model estimate. However, this difference is much less prominent than for the squared bias. When the data generating process is linear, that is, when  $a_2 = 0$  in model 1 and  $a_2 = 0.5$  in model 2, the difference between the estimates from the linear and PLRS models is smaller than in the other cases.

The study suggests that, when estimating or testing the effect of DNA copy number on mRNA expression, there is potentially more to lose than to gain (due to misspecification versus overspecification of the model) by applying the linear instead of the piecewise linear spline model.

TABLE 2  
*Simulated coverage probability for different sample sizes, noise levels  
and significance levels*

	$\sigma = 0.5$		$\sigma = 1$	
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
$n = 20$	0.953	0.898	0.968	0.922
$n = 40$	0.952	0.883	0.967	0.926
$n = 80$	0.939	0.863	0.960	0.915

3.2. *Uniform CBs.* To study the coverage probabilities of the method proposed in Section 2.4, we simulated data according to the model  $y = 1 + (x - 0.5)_+^0 + (x - 0.5)_+^1$ , with  $x$ -values drawn from a uniform distribution  $\mathcal{U}(0, 1)$ . Gaussian errors of standard deviation  $\sigma \in \{0.5, 1\}$ , and three sample sizes  $n \in \{20, 40, 80\}$ . For a given data set we computed the confidence band on a grid of 10 equidistant values, for two different significance levels  $\alpha \in \{0.05, 0.1\}$ , and checked whether the 10 corresponding values of the function in the display fall simultaneously into the estimated confidence band. (For computational reasons the simulation was limited to 10 values; we believe that using the continuous range would not have altered the findings.) Table 2 shows the empirical coverage probabilities over 10,000 data sets for each situation.

The simulated coverage probabilities are close to their corresponding nominal values. Even though the coverage procedure is motivated by asymptotic approximations, this is true even when the sample size is small, in agreement with previous literature on likelihood-based interval estimation.

3.3. *PLRS screening test.* We evaluated the performance of the PLRS testing procedure in detecting associations of various functional shapes. PLRS was compared to the LM test (see Section 4.2), Spearman's correlation test and the test proposed by van Wieringen and van de Wiel (2009). SM Figures 2 to 11 show partial ROC curves (sensitivity versus type I error  $\alpha$ , where  $\alpha \leq 0.2$ ) and partial AUC. Details are provided in SM Section 4. Here, we summarize the results.

The PLRS test yielded good performance in detecting various types of associations. It achieved the highest AUC in 68 out of the 90 simulation cases (against 23 for LM). When the true effect is linear, PLRS performed reasonably well. In other cases, it always produced a high, if not the highest, AUC. In particular, PLRS presented a clear advantage over others in detecting partial effects on gene expression, that is, when only one abnormal state (among others) affects expression. In all, results suggest that PLRS accommodates well both continuous and discrete genomic information and, unlike others, is able to detect various types of association.

**4. Application.** The proposed framework was applied to two data sets. The first data set [Carvalho et al. (2009); available at [ncbi.nlm.nih.gov/geo](http://ncbi.nlm.nih.gov/geo); accession number GSE8067] consists of copy number and gene expression values for 57 samples of colorectal cancer tissue. These were generated with BAC/PAC and Human Release 2.0 oligonucleotide arrays, respectively. Normalization is as in Carvalho et al. (2009). aCGH data were segmented with the CBS algorithm of Olshen et al. (2004) and discretized with CGHcall [van de Wiel et al. (2007)]. Matching of mRNA and aCGH features was based on minimizing the distance between the midpoints of the genomic locations of the array elements. The final data set comprises 25,869 matched features. The second data set [Neve et al. (2006); available from Bioconductor] consists of copy number and expression data for 50 samples (cell lines) of breast cancer, profiled with OncoBAC and Affymetrix HG-U133A arrays. Preprocessing of mRNA expression is described in Neve et al. (2006). aCGH data were segmented and called as above. The resulting data set contains 19,224 matched features. For the colorectal and breast cancer data sets, knots of the PLRS model were estimated using methods I and II, respectively.

We first present some global results on model selection, and next consider testing the association between DNA and mRNA. Finally, some relevant relationships are illustrated.

4.1. *Model selection with the OSAIC procedure.* Table 3 reports the number of genes for which our procedure (column OSAIC) selects a certain type of model, for both data sets. Clearly, both the piecewise linear model and the piecewise level model are selected a large number of times. Different procedures such as AIC and BIC,  $BIC_r = -2 \cdot \mathcal{L}_r(\tilde{\theta}_r) + \log(n) \cdot k_r$ , which put bigger penalties on larger models (too large given the constraints), still often prefer piecewise splines. This gives strong evidence on the inadequacy of both the simple linear and piecewise constant models for many genes. In Section 1 of SM, an overlap comparison of the three procedures shows differences induced by the different penalty functions.

TABLE 3

*The number of times a model is selected by type of model, by three model selection procedures, for the two data sets*

Type of model	Carvalho et al. (2009)			Neve et al. (2006)		
	OSAIC	AIC	BIC	OSAIC	AIC	BIC
Intercept	<b>14,720</b>	18,083	21,700	<b>5081</b>	6968	9379
Simple linear	<b>4916</b>	3674	2043	<b>5262</b>	6689	6345
Piecewise level	<b>2667</b>	1977	992	<b>2761</b>	2477	1608
Piecewise linear	<b>3566</b>	2135	1134	<b>6120</b>	3090	1892

TABLE 4  
*Number of associations with an estimated false discovery rate below 0.1 for different model comparisons*

$H_0$	$H_a$	Carvalho et al. (2009)	Neve et al. (2006)
Intercept	linear	1726	9783
Intercept	full	1554	9105

4.2. *Testing the effect of DNA on mRNA.* The hypothesis that DNA copy number has no effect on mRNA expression corresponds to model (1) with only the intercept parameter  $\theta_0$  nonzero. We tested this as the null model both versus the full model (1) (test “PLRS”) and versus the linear submodel (test “LM”), with the purpose to compare these two screening models in their effectiveness to detect an association. A third possibility would be to test the null model versus the model selected by the OSAIC procedure. However, because this would naively suggest that the form of the relationship is known a priori, we did not pursue this option. For the PLRS test a minimum number of five observations (the default being three) per state was imposed.

Table 4 gives the number of associations with a  $q$ -value below 0.1 [based on the Benjamini and Hochberg (1995) FDR]. The LM test is seen to detect slightly more associations as being significant than the PLRS test. This may be a consequence of the fact that the linear model involves fewer parameters. However, closer inspection shows that the sets of detected genes are not nested, and the PLRS test is able to detect biologically meaningful genes that are not detected by the LM test. To illustrate, three DNA-mRNA relationships are plotted in Figure 3. The first corresponds to an association detected as significant with the LM test, but not with the PLRS test. Reciprocally, the last two associations (genes PDE3B and CLIP1) are detected with the PLRS test but not with the LM test. The figure shows that the PLRS test is able to detect relationships for which an effect is present for only a few samples (but at least five). Identifying the last two genes may be more important than the first, as they are more interesting potential targets for studying individual effects.

The first gene in Figure 3 also illustrates that the testing procedures may differ considerably in  $q$ -values, even though the estimated regression function found by the two models is the same. This is partly explained by the difference in complexity between the alternative models. However, we note that  $q$ -values for a single gene are not directly comparable, since they also depend on  $p$ -values of other genes. In Appendix A, we provide, for selected genes,  $p$ - and  $q$ -values for the different types of tests.

4.3. *Results for selected genes.* In this section we show the estimated relationships for selected genes. The selection is based on the Cancer Gene Census

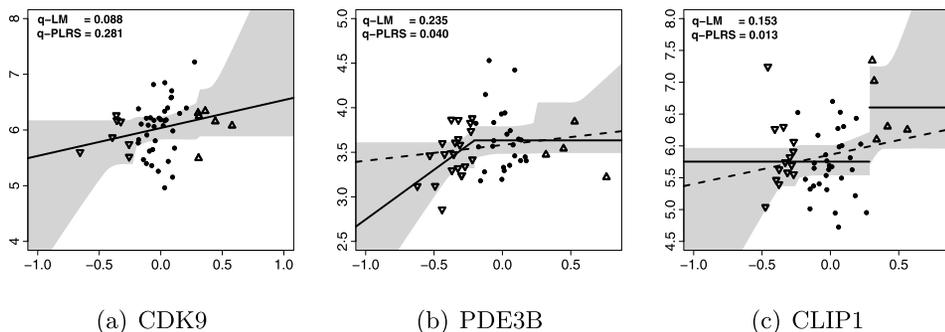


FIG. 3. Association between DNA and mRNA for different genes in the breast cancer data set [Neve et al. (2006)]. Segmented copy number is on the x-axis, while gene expression is on the y-axis. Symbols indicate the different states, namely, loss ( $\nabla$ ), normal ( $\circ$ ) and gain ( $\triangle$ ). Grey surfaces correspond to 95% uniform CBs. The top left values correspond to  $q$ -values of tests LM and PLRS, respectively. The dashed line gives the fitted LM model; the “continuous” spline is the fitted PLRS model.

list (available at [www.sanger.ac.uk/genetics/CGP/Census/](http://www.sanger.ac.uk/genetics/CGP/Census/)) and on our observation that some associations are atypical. Also, we show results for genes C20orf24, TCFL5 and TH1L, which were reported in Carvalho et al. (2009) as important for colorectal cancer progression.

Figures 4 and 5 show nine DNA-mRNA associations for each of the two data sets. Each plot displays the fit of the linear model and of the PLRS model chosen by the OSAIC criterion. Uniform 95% confidence bands (that account for model selection uncertainty) are also plotted. (Some curious shapes result from the fact that pointwise variation bursts near the boundaries and around knots.)

Both figures show a diverse set of forms of associations. Fitted models with jumps reveal that discrete copy number states can, by themselves, explain variation in expression. This is even more true when a piecewise level relationship is identified (as for genes APC and MTUS1 in Figure 4). More generally, piecewise linear models capture effects that differ for losses, gains and/or amplifications. Statistically speaking, this has the advantage of giving more accurate estimates of slope(s), as is clearly observed for genes ATMIN, PTPNA and PTEN in Figure 5. Having a better estimator, we may expect a better test. From a biological point of view, the ability to distinguish effects between states may help the detection of onco and tumor-suppressor genes. Moreover, genes for which these effects concern only a few samples may also be interesting to biologists for studying individual effects.

The simple linear model is observed to be a tight template for modeling. As a matter of fact, it is potentially misleading when the relationship really depends on the underlying copy number state. This happens to be the case for known cancer genes (see FGFR1, PAK1 and PTEN in Figure 5). As a result, when testing the effect of DNA on mRNA with the LM and PLRS tests (see Section 4.2), one may

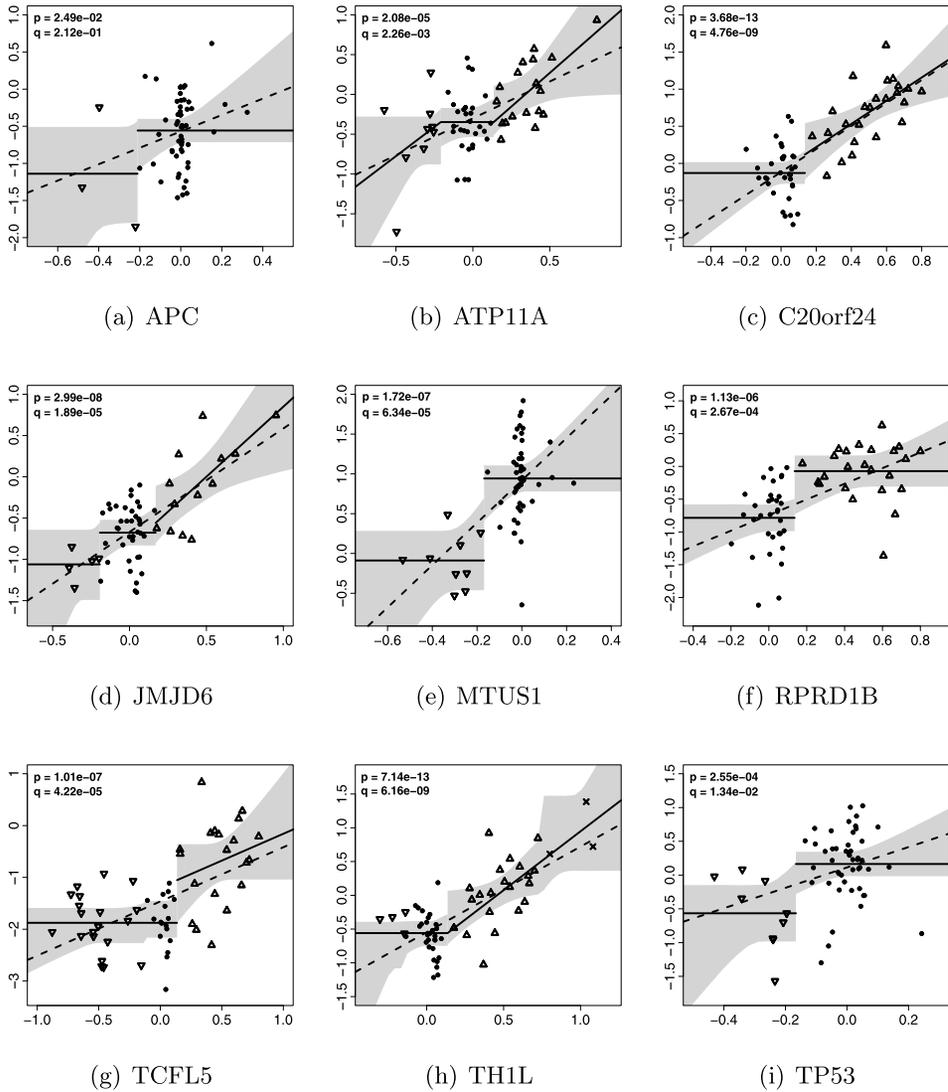


FIG. 4. Association between DNA and mRNA for different genes in the colorectal cancer data set. Segmented copy number is on the x-axis, while gene expression is on the y-axis. States are indicated by different symbols: loss ( $\nabla$ ), normal ( $\circ$ ), gain ( $\triangle$ ) and amplification ( $\times$ ). Grey surfaces correspond to 95% uniform CBs. In all cases the piecewise linear model is preferred to the simple linear one (dashed line). The top left values correspond to the  $p$ - and  $q$ -values of the PLRS test.

obtain a considerable difference between the  $p$ -values, and hence  $q$ -values (see Appendix A). For this reason the proposed framework may improve the detection of (highly) significant associations and their ranking.

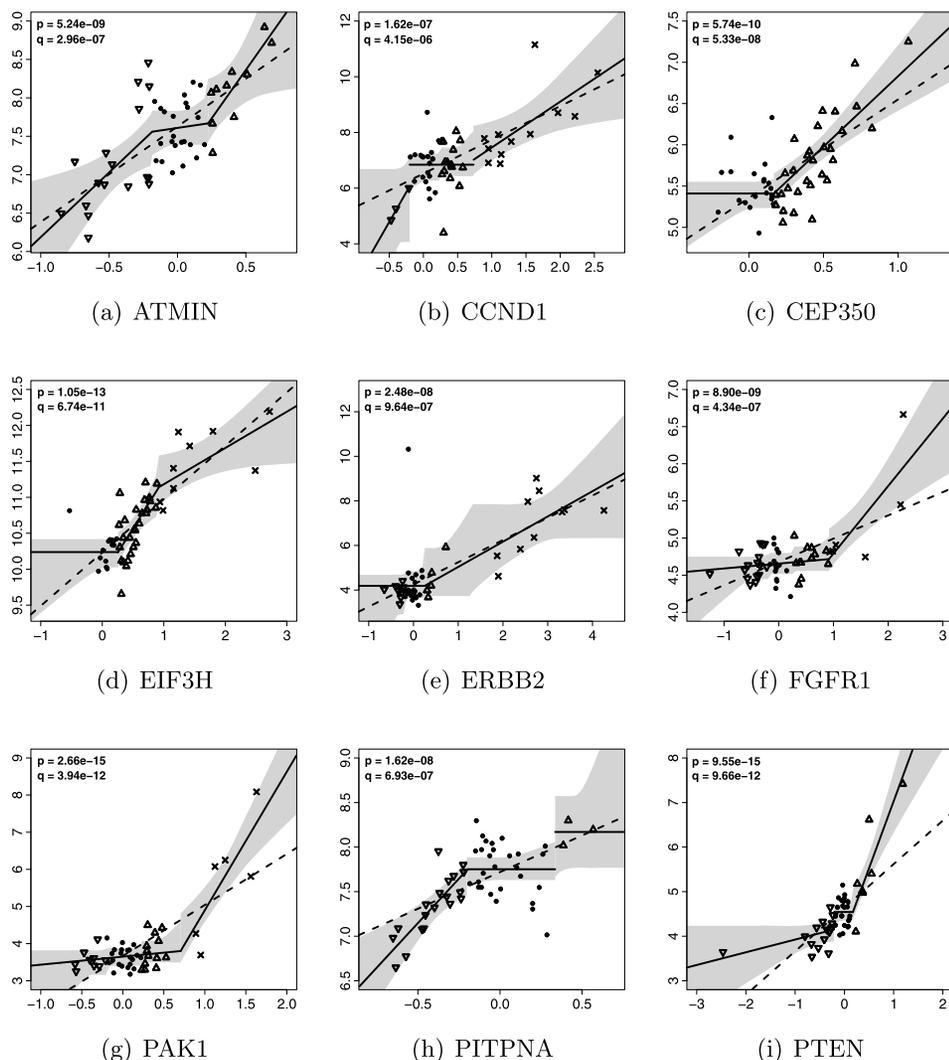


FIG. 5. Association between DNA and mRNA for different genes in the breast cancer data set. Segmented copy number is on the x-axis, while gene expression is on the y-axis. States are indicated by different symbols: loss ( $\nabla$ ), normal ( $\circ$ ), gain ( $\triangle$ ) and amplification ( $\times$ ). Grey surfaces correspond to 95% uniform CBs. In all cases the piecewise linear model is preferred to the simple linear one (dashed line). The top left values correspond to the  $p$ - and  $q$ -values of the PLRS test.

Finally, we dwell on the notion of effect in itself. The notion of “association” is broad, and can be expressed both by an intercept and a slope. This can imply a clear difference in interpretation with respect to the linear model. Consider the simple example of gene *MTUS1* in Figure 4, where a piecewise level model is preferable. Here intuition clearly tells us that one is more interested in assessing the difference in expression level between samples presenting loss and normal aberrations than

an overall trend. Therefore, a linear model may focus on the wrong quantity of interest, whereas the PLRS procedure may yield meaningful interpretation.

We concentrated on comparing our results with those of the linear model. However, it is clear from Figures 4 and 5 that also the other alternative, the piecewise level model (which allows only horizontal lines per state), is often not adequate (see TH1L and PITPNA).

**5. Conclusion.** We proposed a statistical framework for the integrative analysis of DNA copy number and mRNA expression, which incorporates segmented and called aCGH data. By using discrete aCGH data we improved model flexibility and interpretability. The form of the relationship is allowed to vary per gene. Model interpretation is ameliorated with biologically motivated constraints on the parameters. This complicates the statistical procedures for identifying and inferring the relationship between the markers, but we provided methods for model selection, interval estimation and testing the strength of the association. We applied the methodology to two real data sets. Many (reported) genes exhibited interesting behavior.

A novelty of this work is the combined use of segmented and called aCGH data. Which of the two data types is more suitable is a matter of debate in the aCGH community, and may depend on the type of downstream analysis [van Wieringen, van de Wiel and Ylstra (2007)]. Our method provides a compromise that uses both characteristics of the data.

The form of association between copy number and expression in breast cancer is also explored in the recent paper Solvang et al. (2011) (which we received after completion of this paper). This interesting paper distinguishes (only) between linear and quadratic types of effect and uses (only) two types of aberrations, without distinguishing gains from amplifications. The interpretation of the coefficients in our model seems to be simpler.

The proposed methodology is also applicable to the joint analysis of copy number and microRNA expression. This class of noncoding RNA was shown to play an important role in tumor development. Our method may be particularly suitable for these data, because microRNA transcripts are often expressed in part of the samples only.

Next generation sequencing data will impose new challenges, which will be taken up in future work. This type of data provides higher resolution than microarrays, while reducing biases, in particular, at the lower end of the spectrum. Because expression levels are measured as counts rather than intensities, the distribution of the response variable cannot be assumed to be Gaussian and, hence, a different noise model is needed.

In short, we provide methodology for statistical inference and model selection in the framework of constrained PLRS, and show that this is able to reveal interesting DNA-mRNA relationships for cancer genes. The method is implemented in R and available as a package from [www.few.vu.nl/~mavdwiel/software.html](http://www.few.vu.nl/~mavdwiel/software.html).

TABLE 5

*p* and *q*-values of the test when under the alternative hypothesis  $H_a$  the linear, OSAIC-selected and the full models are successively considered. The top and bottom parts correspond, respectively, to the selected genes from the colorectal and breast data

	Linear		OSAIC		Full	
	<i>p</i>	<i>q</i>	<i>p</i>	<i>q</i>	<i>p</i>	<i>q</i>
APC	2.49e-02	2.04e-01	2.26e-02	6.38e-02	2.49e-02	2.12e-01
ATP11A	7.34e-06	9.79e-04	5.88e-06	2.98e-04	2.08e-05	2.26e-03
C20orf24	1.71e-12	2.21e-08	3.06e-13	1.14e-09	3.68e-13	4.76e-09
JMJD6	5.44e-09	4.85e-06	1.78e-08	4.31e-06	2.99e-08	1.89e-05
MTUS1	6.83e-07	1.77e-04	6.38e-08	1.06e-05	1.72e-07	6.34e-05
RPRD1B	3.18e-06	5.45e-04	5.17e-07	5.02e-05	1.13e-06	2.67e-04
TCFL5	6.49e-06	8.88e-04	1.75e-08	4.31e-06	1.01e-07	4.22e-05
TH1L	1.06e-10	3.25e-07	2.72e-13	1.14e-09	7.14e-13	6.16e-09
TP53	6.54e-03	9.87e-02	9.42e-05	2.25e-03	2.55e-04	1.34e-02
ATMIN	1.12e-09	6.45e-08	1.13e-09	4.56e-08	5.24e-09	2.96e-07
CCND1	1.91e-08	5.71e-07	3.56e-08	6.88e-07	1.62e-07	4.15e-06
CEP350	8.55e-08	1.93e-06	3.07e-10	1.69e-08	5.74e-10	5.33e-08
EIF3H	1.70e-12	4.88e-10	8.22e-15	3.75e-12	1.05e-13	6.74e-11
ERBB2	4.46e-10	3.18e-08	4.34e-10	2.15e-08	2.48e-08	9.64e-07
FGFR1	1.62e-06	2.03e-05	3.99e-10	2.02e-08	8.90e-09	4.34e-07
PAK1	1.15e-10	1.21e-08	<2.2e-16	<2.2e-16	2.66e-15	3.94e-12
PITPNA	1.85e-06	2.25e-05	8.40e-10	3.66e-08	1.62e-08	6.93e-07
PTEN	7.27e-09	2.64e-07	9.10e-15	4.02e-12	9.55e-15	9.66e-12

## APPENDIX A: TESTING

See Table 5.

## APPENDIX B: SEMIDEFINITE PROGRAMMING

Here, we provide the two equivalence relationships from Vandenberghe and Boyd (1996) that are necessary to express the semidefinite program. We recall that a linear matrix inequality (LMI) type of constraint includes, among others, linear and convex quadratic inequalities. These are the two types of constraints we are interested in. To express them as two LMIs, we make use of the following equivalences.

A linear inequality constraint  $Ax + b \geq 0$ , where  $A = [a_1 \cdots a_k]$  and  $x \in \mathbb{R}^n$ , is equivalent to the following LMI:

$$F(x) = F_0 + \sum_{i=1}^k x_i F_i \geq 0,$$

where  $F_0 = \text{diag}(b)$ ,  $F_i = \text{diag}(a_i)$ ,  $i = 1, \dots, k$ .  $\text{diag}(v)$  represents the diagonal matrix with the vector  $v$  on its diagonal.

A convex quadratic constraint  $(Ax + b)^T(Ax + b) - c^T x - d \leq 0$ , where  $A = [a_1 \cdots a_k]$  and  $x \in \mathbb{R}^n$ , is equivalent to the following LMI:

$$F(x) = F_0 + \sum_{i=1}^k x_i F_i \succeq 0,$$

where

$$F_0 = \begin{pmatrix} I & b \\ b^T & d \end{pmatrix}, \quad F_i = \begin{pmatrix} 0 & a_i \\ a_i^T & c_i \end{pmatrix}, \quad i = 1, \dots, k.$$

Multiple LMIs can be expressed as a single one using block diagonal matrices [VanAntwerp (2000)].

**Acknowledgments.** We wish to thank Thang V. Pham for helpful discussions on optimization.

## SUPPLEMENTARY MATERIAL

**Complementary results and simulations** (DOI: [10.1214/12-AOAS605SUPP](https://doi.org/10.1214/12-AOAS605SUPP); .pdf). We present a simulation study which compares the performance of the PLRS testing procedure in detecting associations of various functional shapes with that of other procedures. Additionally, we provide an overlap comparison of model selection procedures, complementary results for the simulation on point estimation and a description of the simulation on the precision of knots.

## REFERENCES

- AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* (B. N. Petrov and F. Csaki, eds.) 267–281. Akadémiai Kiadó, Budapest. [MR0483125](#)
- ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. [MR1748009](#)
- ARNOLD, B. C. and SHAVELLE, R. M. (1998). Joint confidence sets for the mean and variance of a normal distribution. *Amer. Statist.* **52** 133–140. [MR1628439](#)
- ASIMIT, J. L., ANDRULIS, I. L. and BULL, S. B. (2011). Regression models, scan statistics and reappearance probabilities to detect regions of association between gene expression and copy number. *Stat. Med.* **30** 1157–1178. [MR2767848](#)
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- BICCIATO, S., SPINELLI, R., ZAMPIERI, M., MANGANO, E., FERRARI, F., BELTRAME, L., CIFOLA, I., PEANO, C., SOLARI, A. and BATTAGLIA, C. (2009). A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res.* **37** 5057–5070.

- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#)
- BOZDOGAN, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. [MR0914460](#)
- BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2003). Interval estimation in exponential families. *Statist. Sinica* **13** 19–49. [MR1963918](#)
- BUCKLAND, S. T., BURNHAM, K. P. and AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53** 603–618.
- BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. [MR1919620](#)
- CARVALHO, B., POSTMA, C., MONGERA, S., HOPMANS, E., DISKIN, S., VAN DE WIEL, M. A., VAN CRIEKINGE, W., THAS, O., MATTHÄI, A., CUESTA, M. A., DROSTE, J. S. T. S., CRAANEN, M., SCHRÖCK, E., YLSTRA, B. and MEIJER, G. A. (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* **58** 79–89.
- CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statistics* **25** 573–578. [MR0065087](#)
- GOURIÉROUX, C., HOLLY, A. and MONFORT, A. (1982). Likelihood ratio test, Wald test, and Kuhn–Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* **50** 63–80. [MR0640166](#)
- GRÖMPING, U. (2010). Inference with linear equality and inequality constraints using R: The package `ic.infer`. *J. Stat. Softw.* **33** 1–31.
- GU, W., CHOI, H. and GHOSH, D. (2008). Global associations between copy number and transcript mRNA microarray data: An empirical study. *Cancer Inform.* **6** 17–23.
- HUGHES, A. W. and KING, M. L. (2003). Model selection using AIC in the presence of one-sided information. *J. Statist. Plann. Inference* **115** 397–411. [MR1985874](#)
- JÖRNSTEN, R., ABENIUS, T., KLING, T., SCHMIDT, L., JOHANSSON, E., NORDLING, T. E. M., NORDLANDER, B., SANDER, C., GENNEMARK, P., FUNA, K., NILSSON, B., LINDAHL, L. and NELANDER, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.* **7** 486.
- KODDE, D. A. and PALM, F. C. (1986). Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* **54** 1243–1248. [MR0859464](#)
- KUDÔ, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50** 403–418. [MR0163386](#)
- LEDAY, G. G. R., VAN DER VAART, A. W., VAN WIERINGEN, W. N. and VAN DE WIEL, M. A. (2013). Supplement to “Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines.” DOI:[10.1214/12-AOAS605SUPP](#).
- LEE, H., KONG, S. W. and PARK, P. J. (2008). Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics* **24** 889–896.
- LIPSON, D., BEN-DOR, A., DEHAN, E. and YAKHINI, Z. (2004). Joint analysis of DNA copy numbers and gene expression levels. In *Algorithms in Bioinformatics. Lecture Notes in Computer Science* **3240** 135–146. Springer, Berlin. [MR2155600](#)
- MEEKER, W. Q. and ESCOBAR, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *Amer. Statist.* **49** 48–53.
- MENEZES, R., BOETZER, M., SIESWERDA, M., VAN OMMEN, G. J. and BOER, J. (2009). Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics* **10** 203+.

- NEVE, R. M., CHIN, K., FRIDLYAND, J., YEH, J., BAEHNER, F. L., FEVR, T., CLARK, L., BAYANI, N., COPPE, J.-P. P., TONG, F., SPEED, T., SPELLMAN, P. T., DEVRIES, S., LAPUK, A., WANG, N. J., KUO, W.-L. L., STILWELL, J. L., PINKEL, D., ALBERTSON, D. G., WALDMAN, F. M., MCCORMICK, F., DICKSON, R. B., JOHNSON, M. D., LIPPMAN, M., ETHIER, S., GAZDAR, A. and GRAY, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10** 515–527.
- OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.
- PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. [MR2758084](#)
- PINKEL, D. and ALBERTSON, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* **37** Suppl. S11–S17.
- QUACKENBUSH, J. (2002). Microarray data normalization and transformation. *Nat. Genet.* **32** Suppl. 496–501.
- ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. [MR0961262](#)
- SALARI, K., TIBSHIRANI, R. and POLLACK, J. R. (2010). DR-Integrator: A new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics* **26** 414–416.
- SCHÄFER, M., SCHWENDER, H., MERK, S., HAFERLACH, C., ICKSTADT, K. and DUGAS, M. (2009). Integrated analysis of copy number alterations and gene expression: A bivariate assessment of equally directed abnormalities. *Bioinformatics* **25** 3228–3235.
- SHAPIRO, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.* **56** 49–62. [MR0963140](#)
- SILVAPULLE, M. J. and SEN, P. K. (2005). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Wiley, Hoboken, NJ. [MR2099529](#)
- SOLVANG, H. K., LINGJÆRDE, O. C., FRIGESSI, A., BØRRESEN-DALE, A.-L. and KRISTENSEN, V. N. (2011). Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics* **12** 197.
- SONESON, C., LILLJEBJÖRN, H., FIORETOS, T. and FONTES, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11** 191.
- VAN DE WIEL, M. A., KIM, K. I., VOSSE, S. J., VAN WIERINGEN, W. N., WILTING, S. M. and YLSTRA, B. (2007). CGHcall: Calling aberrations for array CGH tumor profiles. *Bioinformatics* **23** 892–894.
- VAN DE WIEL, M. A., PICARD, F., VAN WIERINGEN, W. N. and YLSTRA, B. (2011). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief. Bioinformatics* **12** 10–21.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#)
- VAN WIERINGEN, W. N., BERKHOF, J. and VAN DE WIEL, M. A. (2010). A random coefficients model for regional co-expression associated with DNA copy number. *Stat. Appl. Genet. Mol. Biol.* **9** 30. [MR2721705](#)
- VAN WIERINGEN, W. N., VAN DE WIEL, M. A. and YLSTRA, B. (2007). Normalized, segmented or called aCGH data? *Cancer Inform.* **3** 321–327.
- VAN WIERINGEN, W. N. and VAN DE WIEL, M. A. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics* **65** 19–29. [MR2665842](#)
- VANANTWERP, J. (2000). A tutorial on linear and bilinear matrix inequalities. *J. Process Contr.* **10** 363–385.

VANDENBERGHE, L. and BOYD, S. (1996). Semidefinite programming. *SIAM Rev.* **38** 49–95.  
[MR1379041](#)

G. G. R. LEDAY  
DEPARTMENT OF MATHEMATICS  
VU UNIVERSITY  
DE BOELELAAN 1081A  
1081 HV AMSTERDAM  
THE NETHERLANDS  
E-MAIL: [gleday@few.vu.nl](mailto:gleday@few.vu.nl)

A. W. VAN DER VAART  
MATHEMATICAL INSTITUTE  
LEIDEN UNIVERSITY  
PO BOX 9512  
2300 RA LEIDEN  
THE NETHERLANDS  
E-MAIL: [avdvaart@math.leidenuniv.nl](mailto:avdvaart@math.leidenuniv.nl)

W. N. VAN WIERINGEN  
M. A. VAN DE WIEL  
DEPARTMENT OF EPIDEMIOLOGY AND BIostatISTICS  
VU UNIVERSITY MEDICAL CENTER  
PO BOX 7057  
1007 MB AMSTERDAM  
THE NETHERLANDS  
E-MAIL: [w.vanwieringen@vumc.nl](mailto:w.vanwieringen@vumc.nl)  
[mark.vdwiel@vumc.nl](mailto:mark.vdwiel@vumc.nl)