sigaR: statistics for integrative genomics analyses in R

Wessel N. van Wieringen

Department of Epidemiology and Biostatistics, VU University Medical Center
P.O. Box 7075, 1007 MB Amsterdam, The Netherlands
&
Department of Mathematics, VU University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
`w.vanwieringen@vumc.nl`

Contents

1   Introduction

This vignette shows the use of the `sigaR` package. The following features are discussed in some detail:

- the matching of features from different genomic platforms of Van Wieringen *et al.* (2012a),
- the detection of the cis-effect of DNA copy number on gene expression levels as proposed by Van Wieringen and Van de Wiel (2009),
- the fitting of the random coefficients model described in Van Wieringen *et al.* (2010) and the assessment of significance of its parameters, and
- the study of genomic entropy within gene sets, as done in Van Wieringen *et al.* (2011a).

These are illustrated on a small example data set, which is introduced first.

2   Breast cancer data

The breast cancer data set of Pollack *et al.* (2002) is available at `http://www.pnas.org`. Pollack *et al.* (2002) used the same cDNA microarrays to measure both DNA copy number and gene expression of 41 primary breast tumors. Pre-processing is done as detailed in Van Wieringen and Van de Wiel (2009), where the same data set is analyzed. Here, for completeness, the preprocessing is briefly described. The pre-processing of the DNA copy number data consists of removal of clones with more than 30% missing values, imputation of remaining missing values using the k-nearest neighbor method (Troyanskaya *et al*, 2001), mode normalization, segmentation using the CBS method of Olshen *et al.* (2004), and calling using CGHcall of Van de Wiel *et al.* (2007). The gene expression data are within-array median normalized. After pre-processing, only data from chromosome 16 is maintained and included in the `sigaR`-package.

Load the full Pollack breast cancer data:

```
> library(sigaR)
> data(pollackCN16)
> data(pollackGE16)
```

The code above loads a `cghCall` and `ExpressionSet` object containing annotated DNA copy number and gene expression data, respectively.

Each of the pre-processing steps yields a different data set: normalized data, segmented data, and (hard or soft) called data. There appears to be little consensus on which should be used for down-stream (integrative) analysis. The methods, whose implementation is illustrated below, vary in the type of pre-processed DNA copy number data used. This reflects our own varying opinion on the matter. See Van Wieringen *et al.* (2007) or Van de Wiel *et al.* (2011) or for a more elaborate discussion on the type of DNA copy number data to use for downstream analysis.

3   Matching

The first step of an integrative analysis often comprises of the matching of the features of the platforms involved. For the matching of array CGH and gene expression data, the objective is to assign the appropriate DNA copy number to each feature on the gene expression array. Note this is not the same as to reproduce the matching produced by, say, Ensemble. In order to do the matching chromosome, start and end base pair information of the features of both platforms needs to be included in the `cghCall`- and ExpressionSet-object. The function `matchCGHcall2ExpressionSet` is tailor-made for the matching of DNA copy number and gene expression data stored in `cghCall`- and `ExpressionSet`-objects, and provides three types of matching. The function `matchAnn2Ann` provides other ways of matching. Details of all matching methods incorporated in the `sigaR`-package are described in Van Wieringen *et al.* (2012a).

The DNA copy number and gene expression data of the breast cancer data set included in the package have been generated on the same platform. Hence, features need not be matched, i.e., they are already matched. For the sake of illustration we will pretend they are not.

Order the `cghCall`- and `ExpressionSet`-objects genomically:

```
> pollackCN16 <- cghCall2order(pollackCN16, chr=1, bpstart=2, verbose=FALSE)
> pollackGE16 <- ExpressionSet2order(pollackGE16, chr=1, bpstart=2, verbose=FALSE)
```

Match the features of both platforms:

```
> matchIDs <- matchCGHcall2ExpressionSet(pollackCN16, pollackGE16, CNchr=1, CNbpstart=2,
+     CNbpend=3, GEchr=1, GEbpstart=2, GEbpend=3, method = "distance", verbose=FALSE)
```

Limit the cghCall and ExpressionSet-objects to the matched features:

```
> pollackCN16 <- cghCall2subset(pollackCN16, matchIDs[,1], verbose=FALSE)
> pollackGE16 <- ExpressionSet2subset(pollackGE16, matchIDs[,2], verbose=FALSE)
```

In this case (as they were already matched) the objects are unchanged.

For the matching of other platforms the function `matchAnn2Ann` can be used. Let us illustrate the use of this function on the provided breast cancer data:

```
> data(pollackCN16)
> data(pollackGE16)
> matchedIDs <- matchAnn2Ann(fData(pollackCN16)[,1], fData(pollackCN16)[,2],
+     fData(pollackCN16)[,3], fData(pollackGE16)[,1], fData(pollackGE16)[,2],
+     fData(pollackGE16)[,3], method="distance", verbose=FALSE)
```

How many gene expression features not been mapped?

```
> nrow(exprs(pollackGE16)) - length(matchedIDs)

[1] 0
```

The distribution of the number of DNA copy number features matched to a gene expression feature:

```
> table(sapply(matchedIDs, nrow, simplify=TRUE))

  1   2   3
206  32   2
```

Most gene expression features are matched to a single DNA copy number feature, but some are matched to two or more features. In the latter case, the data from those features needs to be summarized into a single DNA copy number signature for that gene expression features. This may be done by weighted averaging, but other suggestions are given in Van Wieringen *et al.* (2012a). Hereto, add offset to distances (avoids infinitely large weights):

```
> matchedIDs <- lapply(matchedIDs, function(Z, offset){ Z[,3] <- Z[,3] + offset; return(Z)},
+     offset=1)
```

Extract id's for object subsetting:

```
> matchedIDsGE <- lapply(matchedIDs, function(Z){ return(Z[, -2, drop=FALSE]) })
> matchedIDsCN <- lapply(matchedIDs, function(Z){ return(Z[, -1, drop=FALSE]) })
```

Generate matched objects:

```
> GEdata <- ExpressionSet2weightedSubset(pollackGE16, matchedIDsGE, 1, 2, 3, verbose=FALSE)
> CNdata <- cghCall2weightedSubset(pollackCN16, matchedIDsCN, 1, 2, 3, verbose=FALSE)
```

The results are matched cghCall- and ExpressionSet-objects, which are (almost) identical the matching. Almost, as the weights are chosen differently here.

4   Joint plotting

To get a overall impression of the relation between DNA copy number and gene expression data, plot the heatmaps of both molecular levels simultaneously:

```
> CNGEheatmaps(pollackCN16, pollackGE16, location = "mode", colorbreaks = "equiquantiles")
```
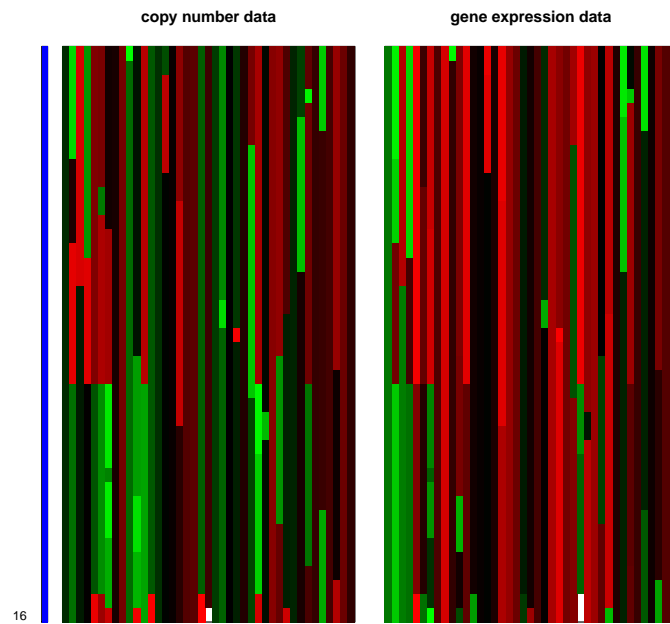

Common features in DNA copy number and gene expression data become more emphasized if, prior to simultaneous heatmap plotting, the samples both are ordered in accordance with, say, a hierarchical clustering of either of the two data sets. At all time the order of the samples should be the same for both DNA copy number and gene expression data.

Alternatively, one may be interested in the relation between DNA copy number and gene expression levels within an individual sample. This is visualized by plotting the profiles of two samples on top each other. This plotting may be limited to a particular chromosome of interest via the `chr` parameter.

```
> profilesPlot(pollackCN16, pollackGE16, 23)
```
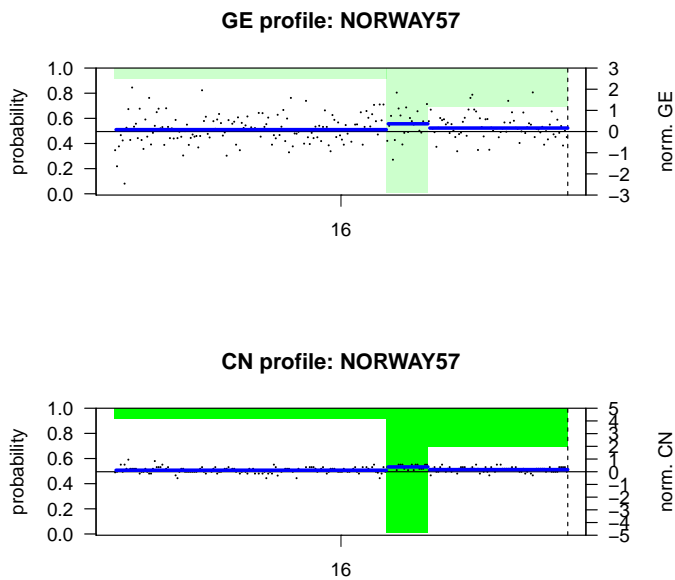

The color coding in the background indicate the aberration call probabilities as produced by CGH-call (Van de Wiel *et al.*, 2007).

5 Statistical unit

Before to engage in any integrative analysis it is important to identify the statistical unit of interest. The statistical unit refers to the biological entity upon which the integrative analysis is supposed to shed light w.r.t. the relationship between the molecular levels involved. The following statistical units are discerned (and illustrated in Figure 1):

- *Gene*: The individual transcripts interrogated by the expression array.
- *Region*: This is a set of contiguous genes with the same DNA copy number signature. Extreme cases of a region are the chromosomes, or even the whole genome. Regions are often determined by the data of the samples in the study. This implies that their definition may vary between data sets, even though they have been generated on the same platform.
- *Pathway*: This is a set of genes from all over the genome. A pathway is determined by knowledge from previous experiments that has been casked in repositories like GO (Gene Ontology Consortium, 2000) and KEGG (Ogata *et al.*, 1999). Also the presence of genes on the expression array determines the actual constitution of the set.

A gene is a limiting case of either a region or a pathway. Similarly, a region is a special case of a pathway.
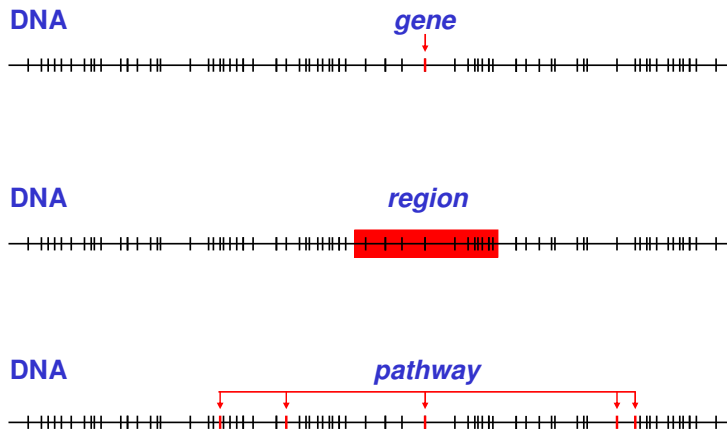

Figure 1: The three statistical units discerned.

6 Gene-wise analysis

We illustrate the integration of DNA copy number and mRNA gene expression data. The two are linked through the central dogma of molecular biology. The dogma suggests that a(n) decrease/increase in copy number of a particular genomic segment leads to a(n) decrease/increase in the expression of genes that map to that segment. This proportional relationship will be a leading principle in our integrative approaches.

A sensible starting point is a univariate integrative analysis, i.e. an analysis at the level of the individual gene Van Wieringen and Van de Wiel (2009). Such approaches aim to detect genes whose expression levels are positively associated with copy number changes. Such genes are candidate cancer genes. The detection of cancer genes is performed within a model relating the two molecular levels. The model enables the estimation of the amount of differential expression due to copy number changes and the employment of a statistical test to assess the significance of the association.

The method of Van Wieringen and Van de Wiel (2009) comes out second in a comparison of genomic *cis*-effect detection methods (Louhimo *et al.*, 2012). That is, second after the method developed by the authors of the same comparison.

Note that the method for *cis*-effect detection of Van Wieringen and Van de Wiel (2009) uses the call probabilities of the preprocessed DNA copy number data. Would one prefer to use the segmentated DNA copy number data, a good alternative is the method of Leday *et al.* (2012). The method of Leday *et al.* (2012) models the *cis*-effect of DNA copy number on gene expression levels by means of piecewise linear regression splines. The method of Leday *et al.* (2012) is implemented in the `plrs`-package.

6.1 Pre-test and tuning

The method of Van Wieringen and Van de Wiel (2009) exploits the census of cancer genes (Futreal *et al.*, 2004), which distinguishes between proto-onco and tumor-suppressor genes associated with gain and loss, respectively. This gain (or loss) of a particular genomic segment is, through the central dogma of biology, likely to result in increased (or decreased) transcription levels of the genes on the segment. Motived by Figure 1b of Beroukhim *et al.* (2010), it is assumed that, within cancer of a particular tissue, a gene cannot be a proto-onco gene as well as tumor-suppressor gene for that tissue.

Unfortunately it is unknown for every gene whether it is a proto-onco or tumor-suppressor gene. Consequently, one does not know whether to compare the gene expression between samples with a normal and gain, or between those with a loss and normal. This is decided by the array CGH data: e.g., if, for a particular gene, the call probability mass (as measured over the samples) of a gain exceeds that of a loss, the loss and normal call probability masses will be merged and the 'no-gain vs. gain' comparison is carried out for this gene.

Also prior to testing, it is recommendable to discard genes beforehand. This benefits the overall (FDR) power of the testing procedure. Exclusion of genes is done:

1. On the basis of the sum of a gene's marginal call probabilities of loss and gain. If it is smaller than `minCallProbMass`, the gene is discarded from further analysis. Effectively, this ensures identifiability of the copy number effect on expression levels.

2. On the basis of a metric, calculated from the DNA copy number data only, which aims to identify two situations for which the test is likely to have low power.

- The first situation occurs when there is an unbalance between the expected call probabilities, as assessed *over* all samples. For instance, genes with

$$\sum_{i=1}^{n} P(\text{sample } i \text{ has a loss at the location of gene } j) = 0.001$$

and

$$\sum_{i=1}^{n} P(\text{sample } i \text{ has no aberration at the location of gene } j) = 0.999$$

have an unbalanced call probability distribution. A priori one expects that the proposed tests may not be powerful to detect a shift for such genes.

- The second situation occurs when many samples individually (i.e. *within* sample) have a uniformly distributed call probability mass: $P(\text{sample } i \text{ has loss at the location of gene } j) = \frac{1}{2}$ and $P(\text{sample } i \text{ has an aberration at the location of gene } j) = \frac{1}{2}$. This indecision on the call is propagated into the test, resulting in low power.

The cut-off for this metric is chosen in such a way that the expected number of true discoveries is maximized.

The following command line performs the pre-testing and tuning:

```
> genes2test <- cisEffectTune(pollackCN16, pollackGE16, "wmw", nGenes = 100,
    nPerm = 250, minCallProbMass = 0.10)
```

To obtain the number of excluded genes:

```
> nrow(pollackGE16) - length(genes2test)
```

The genes2test object is a vector of the genes that are passed on for testing.

The number of excluded genes depends among others on the DNA copy number profiles. If these are 'wild', exhibiting many aberrations all over the genome, we expect most genes to have a reasonably balanced expected (over the samples) call probability distribution. If, however, there are only few genomic regions aberrated, the contrary is expected, and more genes are expected to be excluded. The number of excluded genes also depends upon the number of genes whose expression is affected by copy number changes. This, in combination with an FDR rule, increases the probability of detecting shifts for genes with unbalanced or imprecise call probability mass.

6.2 Testing

We are now ready to test for DNA copy number induced differential gene expression on the set of selected genes:

```
> cisTestResults <- cisEffectTest(pollackCN16, pollackGE16, genes2test, 1,
    "univariate", "wmw", nPerm = 10000, lowCiThres = 0.10)
```

The number of significant genes, obtained through:

```
> fdrCutoff <- 0.10
> sum(cisTestResults@adjP.values < fdrCutoff)
```

equals 11 at a FDR significance level of 0.05 and 16 at 0.10. Hence, approximately 10% of the genes included in the test (114) are declared significant. This is somewhat lower than the roughly 20% found in the analysis of the full data set (Van Wieringen and Van de Wiel, 2009), and may be due to the fact that fewer genomic aberrations occur on this chromosome compared to other in the data set. Irrespectively, such large percentages of significant genes are in line with 'major direct role' of DNA copy number alterations in the transcriptional program as claimed by Pollack *et al.* (2002), but forces us to look not only at statistical significance, but also at biological relevance. Gene prioritization (ranking) could be done by using the effect size and/or the coefficient of determination.

Finally, a global view of the effect of DNA copy number on gene expression levels is provided by a histogram of the effect sizes for all selected genes, which may be obtained through:

```
> hist(cisTestResults@effectSize, n=50, col="blue", border="lightblue",
    xlab="effect size", main="histogram of effect size")
```
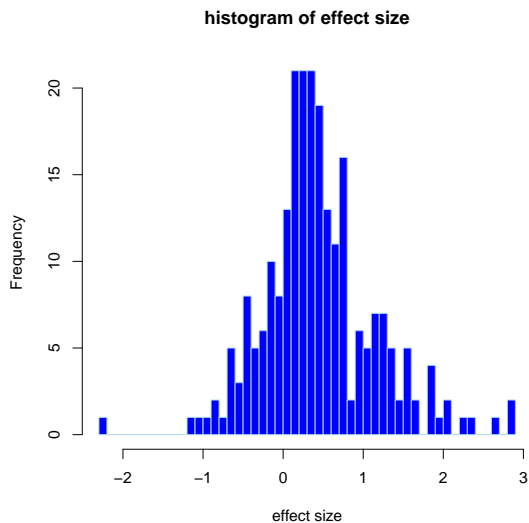

For the Pollack chromosome 16 data the histogram shows an effect size distribution that is clearly shifted away from zero, indicating that many genes have affected expression levels, in turn confirming the aforementioned 'major direct role'.

The top ten of most significant genes are obtained as follows:

```
>   cisEffectTable(cisTestResults, number=10, sort.by="p.value")
            ... geneId comparison av.prob1 av.prob2 effectSize      R2 p.value adjP.value
IMAGE:366728 ...    174          1   0.1110   0.8888     1.1840 0.3138  0.0001 0.00570000
IMAGE:51320  ...    237          2   0.8970   0.1029     1.3465 0.4165  0.0002 0.00570000
IMAGE:625683 ...    229          2   0.8977   0.1022     1.5268 0.3378  0.0002 0.00570000
IMAGE:825335 ...    239          2   0.8970   0.1029     1.1664 0.1830  0.0002 0.00570000
IMAGE:897774 ...    236          2   0.8970   0.1029     1.8775 0.3143  0.0004 0.00912000
IMAGE:845419 ...    238          2   0.8970   0.1029     1.6652 0.4841  0.0011 0.01646667
IMAGE:52226  ...     90          2   0.8911   0.1087     2.6185 0.2609  0.0012 0.01646667
IMAGE:261971 ...    240          2   0.8970   0.1029     1.0780 0.3931  0.0013 0.01646667
IMAGE:279152 ...     55          2   0.8764   0.1234     0.6754 0.1778  0.0013 0.01646667
IMAGE:487831 ...     89          2   0.8911   0.1087     2.0120 0.2403  0.0019 0.02166000
```

The most significant gene is interrogated by clone IMAGE:366728. It is lost in approximately 11% of the samples. The estimated *cis*-effect size of DNA copy number aberrations on the expression levels of this transcripts equals 1.1840. With an $R^2 = 0.31$, the genomic aberrations explain explains 31% of the variation in the expression of the transcript interratogated by IMAGE:366728. The multiplicity corrected p-value of the proposed test equals IMAGE:366728 is depicted in the figure below.

IMAGE:51320__1
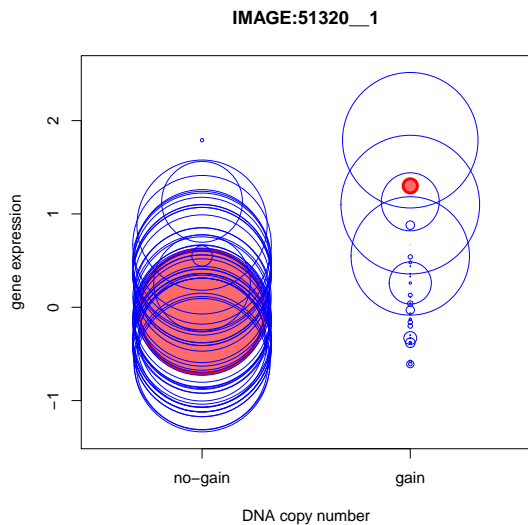
6.3   Regional analysis

The breakpoint nature of the copy number data implies that neighboring genes share the same copy number signature. One expects that their expression levels are affected in a similar fashion. Indeed, it has been observed that co-expressed neighborhoods, neighborhoods of contiguous genes showing markedly similar expression patterns, appear throughout the cancer genome and often coincide with the location of well-known recurrent copy number aberrations. This suggests that CNAs (Copy Number Aberrations) not only affect the expression of key proto-onco or tumor-suppressor genes, but may also alter the expression levels of many other genes in the cancer genome. In particular, whole chromosome aberrations have been shown to affect expression levels of many genes are affected in accordance with the gene dosage.

The above motivates the modification of the univariate approach. In Van Wieringen and Van de Wiel (2009) this is done by borrowing information across the genes within each region (defined as a series of adjacent clones with the same DNA copy number signature), but test for DNA copy number induced differential expression per gene. This is done by shrinking the test statistics within the region. In order to perform such a 'regional analysis' change the `analysisType` parameter:

```
> cisTestResults <- cisEffectTest(pollackCN16, pollackGE16, genes2test, 1,
    "regional", "wcvm", nPerm = 10000, lowCiThres = 0.10)
```

Compare this to the results of the univariate analysis.

7 Region analysis

A more formaly extension of the univariate approach would describe the relation between the two molecular levels in a set of neighboring genes with identical copy number aberration patterns explicitly. In Van Wieringen *et al.* (2010) we proposed a multivariate random coefficients model which addresses regional co-expression through the incorporation of fixed parameters for the joint copy number effect on the expression levels of all genes in the region, with the inclusion of random coefficients for possible individual gene effects. In addition, co-expression non-attributable to copy number changes is accounted for by the correlation structure of the residual gene expression (caused by, e.g., epigenetic effects or a common transcription factor). This random coefficients model facilitates a global analysis of CNA associated regional co-expression at the level of the region (rather than its genes). It allows to assess a) whether there is a shared CNA effect on the expression levels of the genes within the region, and b) whether the CNA effect is identical for all genes. The model parameters are estimated from high-throughput data. In order to deal with the data's high-dimensionality ($p > n$), we have optimized the estimation procedure in terms of computational speed and memory use. Hypotheses of interest regarding copy number induced co-expression model parameters are evaluated by re-sampling. The prior knowledge on the direction of the effect of copy number changes on gene expression is incorporated in the estimation. Two real data examples illustrate how the proposed methodology may be utilized to study regional co-expression associated with DNA copy number aberrations. The proposed random coefficients model may also be applied to expression data of other products that are transcribed from the DNA, such as microRNAs, that share the same copy number signature. The application of the random coefficients model of Van Wieringen *et al.* (2010) is illustrated on these Pollack data.

The data are first put in the appropriate format. Select feature of interest:

```
> featureNo <- 240
```

Determine features having the same DNA copy number signature:

```
> ids <- getSegFeatures(featureNo, pollackCN16)
```

```
perform input checks...
```

Extract copy number and expression data of features comprising the region:

```
> Y <- exprs(pollackGE16)[ids,]
```

Transpose **Y**, the traditional data matrix representation for regression

```
> Y <- t(Y)
```

Extract copy number profile of the region (segmented log2-ratios):

```
> X <- segmented(pollackCN16)[featureNo,]
```

Put **X** in the right format:

```
> X <- matrix(as.numeric(X), ncol=1)
```

To fit the random coefficients model to the gene expression data $\mathbf{Y}$ and DNA copy number data $\mathbf{X}$ of the selected region, first center the expression data of each gene around zero (to avoid having to fit an intercept), and make the linear parameter constraints matrix $\mathbf{R}$:

```
> Y <- sweep(Y, 2, apply(Y, 2, mean))
> R <- matrix(1, ncol=1)
```

The regression parameter $\bar{\beta}$ represents the DNA copy number effect on expression levels, and is assumed to be non-negative as the relationship between the two molecular levels is believed to be concordant.

Now fit the random coefficients model to the data:

```
> RCMresult <- RCMestimation(Y, X, R)
```

To display the results of the model fit:

```
> summary(RCMresult)

Coefficients:
      [,1]
[1,] 1.624
Random effects (as variances):
      [,1]
[1,] 0.228
Variance (average): 0.29
Correlation (unif): 0.487
Shrinkage: 0
Log-likelihood: -124.607
```

This analysis reveals that there is a non-zero shared copy number effect on the expression levels of the genes in the region:

```
> RCMresult@betas

[1] 1.62361
```

In addition, the analysis indicates that expression levels of the genes are not affected in a heterogeneous manner (there is no random effect) by the gene dosage:

```
> RCMresult@tau2s

[1] 0.2283116
```

Also noteworthy is the estimate of the 'residual co-expression' ρ, which is rather high:

```
> RCMresult@rho

[1] 0.48688
```

14

This suggests that other factors (like a common transcription factor or methylation) may play a role in the co-expression of the region.

Significance of either the shared or heterogeneous (or jointly) DNA copy number effect is assessed through the parametric bootstrap. This is illustrated for the shared DNA copy number effect for the selected region. To test the hypothesis of no DNA copy effect vs. the hypothesis of a shared effect:

```
> RCMtestResult <- RCMtest(Y, X, R, testType="II")
```

Display the results:

```
> summary(RCMtestResult)
```

```
Coefficients:
      [,1]
[1,] 1.265
Random effects (as variances):
      [,1]
[1,]    0
Variance (average): 0.29
Correlation (unif): 0.487
Shrinkage: 0
Log-likelihood: -125.948
```

```
Test statistic: 13.165,  p-value: 0
Remarks: none
```

The test for a DNA copy number effect (both shared and random) on the expression levels is significant at the 0.05 level.

The results of this analysis may be visualized. This visualization should also provide us with an impression of the variation of the DNA copy number-gene expression relationship over the genes. To that end, we sample from the random coefficient distribution, and calculate corresponding expected expression values:

```
> GEpred <- numeric()
> for (u in 1:1000){
+         slope <- rnorm(1, mean=RCMresult@betas[1], sd=sqrt(RCMresult@tau2s[1]))
+         slope[slope < 0] <- 0
+         GEpred <- rbind(GEpred, as.numeric(slope * X[,1]))
+ }
> verts <- rbind(apply(GEpred, 2, min), apply(GEpred, 2, max))
```

Now plot the result:

The pink area indicates where we would expect – on the basis of the fitted random coeffients model – the regression lines of the individual genes. Indeed, they fall inside the pink area. The red line represents the shared DNA copy number effect on gene expression levels within the region.

```
> plot(lm(Y[,1] ~ X[,1])$fitted.values ~ X[,1], type="l", ylim=c(-1.0, 2.2),
+      ylab="gene expression", xlab="DNA copy number")
> polygon(x=c(X[order(X[,1]), 1], X[order(X[,1], decreasing = TRUE), 1]),
+      y=c(verts[1, order(X[,1])], verts[2, order(X[,1], decreasing = TRUE)]),
+      col="pink", border="pink")
> for (j in 1:ncol(Y)){
+         lines(X[,1], lm(Y[,j] ~ X[,1])$fitted.values)
+ }
> lines(X[,1], RCMresult@betas[1] * X[,1], type="l", col="red", lwd=4)
```
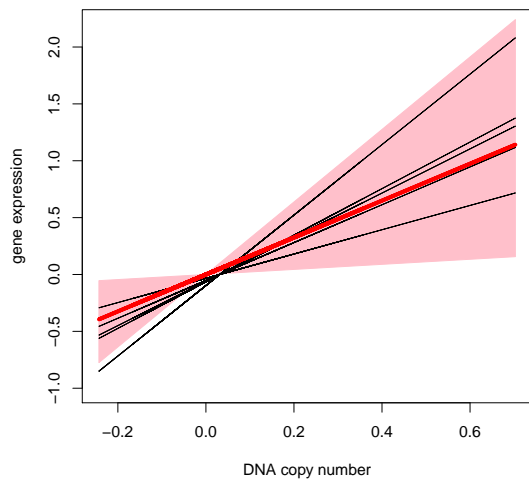
8 Pathway analysis

The random coefficients model of Van Wieringen *et al.* (2010) analyzes regions, a gene set that comprises of contiguous genes. As such, the relationship between DNA copy number and gene expression is investigated locally at the genome. Alternatively, one may wish the study gene sets that constitue of genes originating from all over the genome, and that together form a pathway. The presence of a DNA copy number effect shared by all genes in the pathway – as is modeled by the random coefficients model of Van Wieringen *et al.* (2010) – is unlikely, and we resort to the methodology discussed in Van Wieringen *et al.* (2011a).

Van Wieringen *et al.* (2011a) jointly analyze DNA copy number and gene expression data within pathways rather then regions. In particular, a pathway may comprise the whole genome. To this end, Van Wieringen *et al.* (2011a) use the information theoretic concepts *entropy* (a multivariate measure of spread) and *mutual information* (a multivariate measure of correlation). The estimation of and testing procedures related to these concepts using high-dimensional genomics data of Van Wieringen *et al.* (2011a) have been implemented in the `sigaR`-package.

The Pollack breast cancer data is used to illustrate how these concepts may be employed for the integrative analysis of DNA copy number and gene expression. To assess whether there is an assocation between the DNA copy number and gene expression of chromosome 16 in breast cancer, we analyze the mutual information between the two molecular levels. By studying the mutual information between $\mathbf{Y}$ and $\mathbf{X}$, we compare the unconditional entropy of the gene expression to its conditional counterpart, conditional on DNA copy number.

Again, extract the DNA copy number (normalized log2 ratios) and gene expression data:

```
> X <- copynumber(pollackCN16)
> Y <- exprs(pollackGE16)
```

Transpose both data matrices to the traditional regression representation:

```
> Y <- t(Y)
> X <- t(X)
```

Calculate the MI (mutual information):

```
> hdMI(Y, X, method="knn")

[1] 1.103678
```

Test whether the MI is equal to zero:

```
> MItestResults <- mutInfTest(Y, X, nPerm=100, method="knn", verbose=FALSE)
> summary(MItestResults)

Mutual information test:
Test statistic: 1.104,  p-value: 0
Remarks: none
```

The *p*-value is smaller the 0.01. Using a significance level of 0.05, this implies that there is a significant association between the two molecular levels. This 'mutual information' test can be

used as a general purpose gene set test to investigate the association between two molecular levels within pathways.

This association between the two molecular levels as suggested by the mutual information test, may be visualized. The k-NN entropy statistic is composed of the entropies at each observation. Each sample's contribution to the k-th nearest neighbor genomic entropy estimate may then be plotted against its contribution to the k-th nearest neighbor transcriptomic entropy estimate. If indeed the entropies of the two molecular levels are closely related, we expect the 'marginal' entropies at each observation to be positively associated. Below we plot these 'marginal' entropies of both molecular levels against other:

```
> plot(isoreg(hdEntropy(Y, method="knn") ~ hdEntropy(X, method="knn")),
+     lwd=2, pch=20, main="", ylab="marginal transcriptomic entropy",
+     xlab="marginal genomic entropy")
```
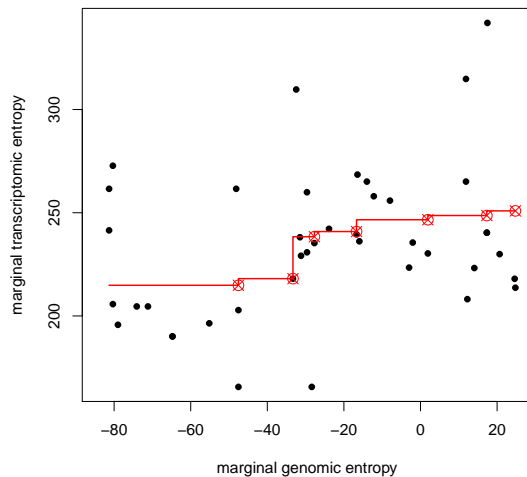

There is a small positive association visible. To emphasize this we have added the isotonic regression curve. This may also be assessed by Spearman's correlation coefficient:

```
> cor(hdEntropy(Y, method="knn"), hdEntropy(X, method="knn"), m="s")
```

```
[1] 0.2332665
```

The correlation between the 'marginal' entropies of the two molecular levels too reveals a positive association.

References

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., McHenry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y. J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M. S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**(7283), 1899–905.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, **4**, 177–183.

Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Leday, G. G. R., Van der Vaart, A. W., Van Wieringen, W. N. and Van de Wiel, M. A. (2012). Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. *Submitted*.

Louhimo, R., Lepikhova, T., Monni, O. and Hautaniemi, S. (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nature Methods*, **9**, 351–355.

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.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Research*, **27**, 29–34.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, **99**, 12963–12968.

Troyanskaya, H., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

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. *Briefings in Bioinformatics*, **12**(1), 10–21.

Van Wieringen, W. N., Van de Wiel, M. A. and Ylstra, B. (2007). Normalized, segmented or called aCGH data? *Cancer Informatics*, **3**, 331–337.

Van Wieringen, W. N. and Van de Wiel, M. A. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**(1), 19–29.

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 aberrations. *Statistical Applications in Genetics and Molecular Biology*, **9**(1), 1–28.

Van Wieringen, W. N., and Van der Vaart, A. W. (2011a). Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics*, **27**(4), 556–563.

Van Wieringen, W. N., Unger, K., Leday, G. G. R, De Menezes, R. X., Ylstra, B., and Van de Wiel, M. A. (2012). Matching of array CGH and gene expression microarray features for the purpose of integrative genomics analysis. *BMC Bioinformatics*.