2010

Supplementary material for Estimating Copy Numbers for Shared Array CGH Data: the Linear-Median Method

Yan-Xia Lin
University of Wollongong, yanxia@uow.edu.au

V. Baladandayuthapani
University of Wollongong

V. Bonato
University of Wollongong

K. A. Do
University of Wollongong

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

10-10

Supplementary material
for
Estimating Copy Numbers for Shared Array CGH Data: the Linear-Median Method

Y.-X. Lin, V. Baladandayuthapani, V. Bonato and K.-A. Do

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Supplementary material
for
Estimating Copy Numbers for Shared Array CGH Data: the Linear-Median Method
by
Y.-X. Lin, V. Baladandayuthapani, V. Bonato and K.-A. Do

Appendix A

Use Monte Carlo method to indirectly show that the value of $\frac{aE(X_p)}{\log(\frac{2+a}{2-a})m_{X_p}}$ is close to 1 for $a = 0.1, 0.2, \cdots, 1.9$ and $\pi = 0.1, 0.2, \cdots, 1$.

The simulation is conducted as follows. For each triplet (a, π, t_p), 5000 independent samples are simulated from model

$$X_p \stackrel{\triangle}{=} X_p(a,\pi) = \frac{T_p + \varepsilon}{2+\eta},$$

where random variables T_p, ε and η are independent; T_p has a distribution such that $P(T_p = t_p) = \pi$ and $P(T_p = 2) = 1-\pi$; ε and η have uniform distribution $U(-a, a)$, $a = 0.1, 0.2, 0.3, \cdots, 1.9$ and $\pi = 0.1, \cdots, 1$ with increments of 0.1 respectively; $t_p = 1, 2, \cdots, 9$ with increments of 1. The mean and median of $X_p(a,\pi)$ are estimated by its sample mean $\bar{X}_p(a,\pi)$ and sample median $median(X_p)(a,\pi)$ respectively. Then $\frac{aE(X_p(a,\pi))}{\log(\frac{2+a}{2-a})m_{X_p(a,\pi)}}$ is estimated and evaluated by

$$\frac{a\bar{X}_p(a,\pi)}{\log(\frac{2+a}{2-a})median(X_p)(a,\pi)}.$$

For each π and t_p fixed, the sample mean $m(\pi, t_p)$ and sample variance $s^2(\pi, t_p)$ of

$$\frac{a\bar{X}_p(a,\pi)}{\log(\frac{2+a}{2-a})median(X_p)(a,\pi)},\ a = 0.1, 0.2, \cdots, 1.9,$$

are calculated by the following formulae:

$$m(\pi, t_p) = \sum_{a=0.1}^{1.9} (\frac{a\bar{X}_p(a,\pi)}{\log(\frac{2+a}{2-a})median(X_p)(a,\pi)})/19,$$

$$s^2(\pi, t_p) = \sum_{a=0.1}^{1.9} (\frac{a\bar{X}_p(a,\pi)}{\log(\frac{2+a}{2-a})median(X_p)(a,\pi)} - m(\pi, t_p))^2/19,$$

and reported in Tables 1 and 2, which follow, where $s^2(\pi)$ is given within the parentheses.

Table 1: The values of $m(\pi, t_p)$ and $s^2(\pi, t_p)$ (Part A)

$\pi = 1$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9988320	0.9999922	1.0006107	0.9996400	1.0010851
(1.204417e-05)	(6.722244e-06)	(5.334313e-06)	(1.261637e-05)	(1.167334e-05)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
0.9996429	1.0007002	0.9996422	0.9995458	
(1.231912e-06)	(5.472384e-06)	(5.472957e-06)	(4.414939e-06)	
$\pi = 0.9$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
1.0234726	0.9999251	0.9945141	0.9874093	0.9815765
(7.944477e-04)	(3.122031e-06)	(7.239544e-05)	(2.153089e-04)	(3.306295e-04)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
0.9784618	0.9754699	0.9728850	0.9690245	
(4.723169e-04)	(5.685863e-04)	(5.824456e-04)	(5.801032e-04)	
$\pi = 0.8$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
1.0387630	0.9996188	0.9892445	0.9765723	0.9673282
(2.886933e-03)	(9.836780e-06)	(2.590746e-04)	(8.382101e-04)	(1.415877e-03)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
0.9586696	0.9522338	0.9460126	0.9428445	
(1.799211e-03)	(2.116493e-03)	(2.245196e-03)	(2.510819e-03)	
$\pi = 0.7$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
1.0490667	1.0001018	0.9855943	0.9663545	0.9524253 2
(5.548408e-03)	(1.432224e-05)	(5.197429e-04)	(1.709809e-03)	(2.958586e-03)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
0.9407424	0.930110	0.9227174	0.9165458	
(3.912353e-03)	(4.538055e-03)	(5.050342e-03)	(5.479345e-03)	
$\pi = 0.6$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
1.0488169	1.0010726	0.9854699	0.9623178	0.9414670
(7.753325e-03)	(6.911221e-06)	(7.413494e-04)	(2.893487e-03)	(4.946303e-03)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
0.9257995	0.9128190	0.9026656	0.8949812	
(6.682264e-03)	(8.140030e-03)	(9.115055e-03)	(1.010583e-02)	

Table 2: The values of $m(\pi, t_p)$ and $s^2(\pi, t_p)$ (Part B)

$\pi = 0.5$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9976367	1.0008558	1.0051801	1.0075940	1.0563732
(3.009497e-03)	(3.751868e-06)	(1.132037e-03)	(8.828197e-03)	(3.143084e-02)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
1.0996647	0.9949510	1.0348440	1.2778189	
(5.681301e-02)	(3.069008e-02)	(5.681301e-02)	(3.069008e-02)	
$\pi = 0.4$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9689243	1.0004657	1.0224327	1.0774329	1.1533009
(2.197164e-03)	(8.269312e-06)	(1.544194e-03)	(1.112485e-02)	(3.060863e-02)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
1.2460811	1.3484609	1.4610660	1.5757287	
(5.815739e-02)	(9.379170e-02)	(1.286891e-01)	(1.732840e-01)	
$\pi = 0.3$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9690245	1.0008585	1.0242324	1.0860159	1.1647982
(1.446965e-03)	(3.876559e-06)	(1.226318e-03)	(6.909656e-03)	(1.727091e-02)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
1.2629289	1.3679820	1.4846238	1.5987995	
(2.912726e-02)	(4.057614e-02)	(5.020234e-02)	(5.951541e-02)	
$\pi = 0.2$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9737273	1.0001785	1.0231539	1.0743057	1.1446959
(6.463392e-04)	(3.456922e-06)	(5.653691e-04)	(3.026114e-03)	(6.303903e-03)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
1.2239605	1.3104238	1.3991247	1.4862704	
(9.262918e-03)	(1.097795e-02)	(1.232603e-02)	(1.325612e-02)	
$\pi = 0.1$				
$t_p = 1$	$t_p = 2$	$t_p = 3$	$t_p = 4$	$t_p = 5$
0.9836251	0.9996371	1.0143460	1.0458579	1.0869769
(1.448035e-04)	(1.181245e-05)	(1.537200e-04)	(6.429722e-04)	(1.166530e-03)
$t_p = 6$	$t_p = 7$	$t_p = 8$	$t_p = 9$	
1.1335024	1.1808455	1.2294815	1.2743215	
(1.374409e-03)	(1.518496e-03)	(1.587512e-03)	(1.669519e-03)	

The Monte Carlo simulation results clearly show that all the sample means $m(\pi, t_p)$ are close to 1 and the sample variance $s^2(\pi, t_p)$ are close to 0. Therefore, it is reasonable to accept that $\frac{aE(X_p(a,\pi))}{\log(\frac{2+a}{2-a})m_{X_p(a,\pi)}} \approx 1$, for any $a \in (0, 2)$, $\pi \in (0, 1)$ and $t_p \in \{1, \cdots, 9\}$.

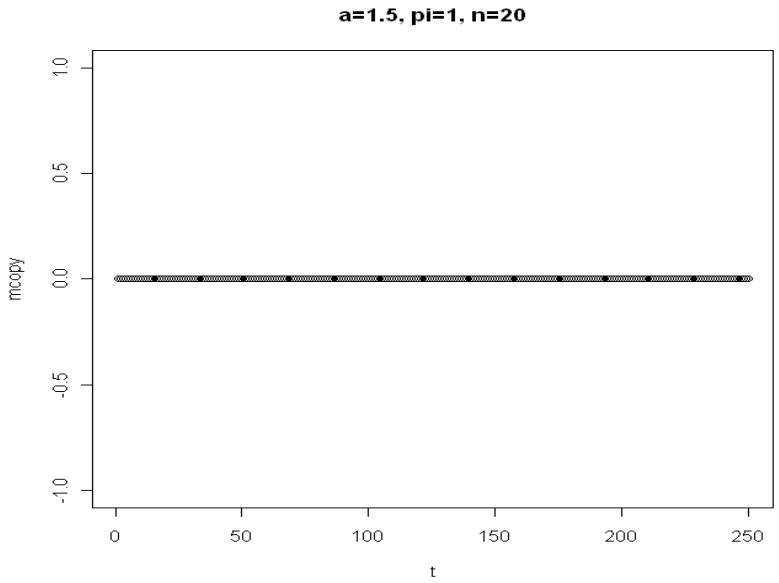

Figure 1: The plot of the mean of gains/losses obtained at each probe position using the cghMCR method.

Appendix B

The locations of the genes of NR/U, CR and LCR in non-small cell adenocarcinoma (NA) and related references.

- Probe positions from 1 to 295: A total of 200 genes are found in this region, 28 of them (14%) are genes related to cancer phenotype while 3 (1.5%) are related to lung cancer phenotype. All LCR genes are located in chromosomal regions identified as losses by both methods (LM and cghMCR). The LCR genes located at this region are PSIP1, CDKN2A, and TUSC1. PSIP1 and CDKN2A, a well-known lung cancer suppressor[1] are both located in a region frequently found deleted in lung cancer patients[2]. In addition, TUSC1 is found mutated and silent in nonsmall cell lung carcinoma cell lines[3].

- Probe positions from 296 to 331: A total of 12 NR/U genes are found in this region.

- Probe positions from 332 to 341: Only 3 genes are located in this region with

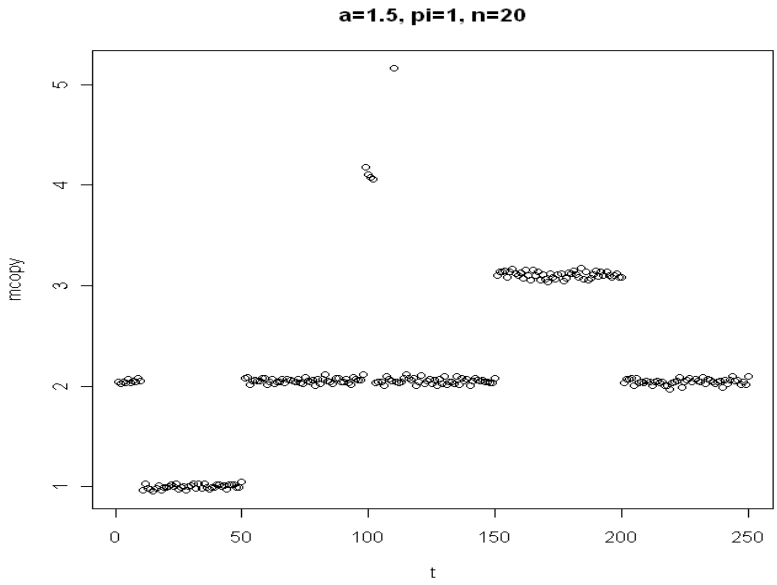

Figure 2: The plot of the mean of copy numbers obtained at each probe position using the linear-median method.

one of them being classified as CR (ACO1). Both methods identify the region where this gene is located as loss.

- Probe positions from 342 to 375: A total of 113 genes are located in this regions with 14 of them being classified as CR.

- Probe positions from 376 to 500: A total of 171 genes are located in this region. Four of them are CR and only one (IGFBPL1, classified as loss by both methods) is classified as LCR. IGFBPL1 has already been shown to be downregulated in lung tumor samples[4].

- Probe positions from 501 to 1234: A total of 744 genes are located in this region, 90 of them being classified as CR, and 9 as LCR. The cghMCR method does not identify any region containing LCR as altered. On the other hand, the LM method identifies five of the LCR genes in chromosomal regions of loss (TLE1, FRMD3, DAPK1, MIRLET7A1, PTPN3) and, consequently, are expected to have lower expression in lung tumor samples. In fact, TLE1 is frequently found altered in squamous cell carcinomas and adenocarcinomas[5] while FRMD3 expression is usually silenced in primary nonsmall cell lung carcinomas[6]. Likewise, mouse lung carcinoma clones characterized by highly aggressive metastatic behavior did not express Dapk1[7]. Also, MIRLET7A1

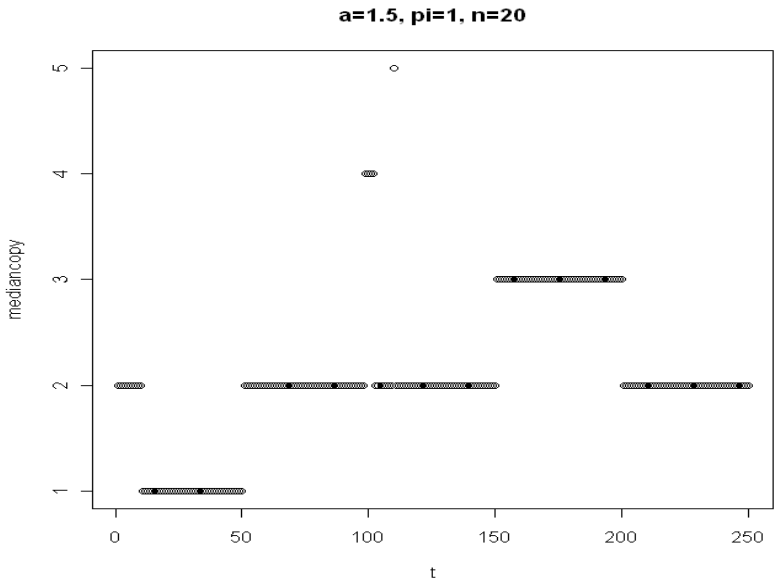

Figure 3: The plot of the median of copy numbers obtained at each probe position using the linear-median method.

and PTPN3 expressions are downregulated in lung cancer[8,9]. The LM indetifies one gene located in a gain region (GAS1), and therefore, it is expected to be overexpressed in lung cancer samples. Surprisingly, Gas1 expression is known by its capacity of suppressing metastasis in lung[10], therefore, we hypothesize that the this gene might be regulated epigenetically or it is a false positive identified by the LM method. Again, the cghMCR method does identifies this region as neutral. In addition, 3 genes are found by both methods in neutral regions (PHF19, DAB2IP, RPL12) and, therefore, we believe that their regulation is being performed by epigenetic factors. In fact, PHF19 mRNA is known to be overexpressed in lung cancers[9] as well as methylation of the promoter of DAB2IP is associated with the lung cancer phenotype[11]. Likewise, RPL12 splice variant are frequently found in human lung carcinoma cell[12].

- Probe positions from 1235 to 1249: A total of 17 genes are located in this regions with only one of them (ABL1) being classified as CR and identified as a gain by both methods.

Appendix C

R code for the linear-median function

$\sharp$ x is an $n \times T$ matrix, the elements of y are aCGH observations in linear format

♯ n denotes the number of independent samples

♯ T denotes the size of each individual sample

♯ At any probe position p, if the true **shared** copy number is not 2, the probability of having copy number changed is "prob".

♯ Function "Linear_Median" gives the estimate of **shared** copy number at each probe position.

```
Linear_Median=function(x,n,T,prob){
medianx=c()
for (i in 1:T){
medianx[i]= median(x[i,])
}
justx=c()
justx=2*(medianx-1+prob)/prob

xx=c()
xx=floor(justx)

for(i in 1:T){
if (justx[i]>=xx[i]+0.5)
xx[i]=xx[i]+1
}
xx
}
```

References

[1] Kamb, A., Gruis, N. A., Weaver-Feldhaus, J., Liu, Q., Harshman, K., Tavtigian, S. V., Stockert, E., Day, R. S., III, Johnson, B. E., Skolnick, M. H. (1994). A cell cycle regulator potentially involved in genesis of many tumor types. *Science*, **264**, 436-440.

[2] Singh, D. P., Kimura, A., Chylack, L. T., Jr., Shinohara, T. (2000). Lens epithelium-derived growth factor (LEDGF/p75) and p52 are derived from a single gene by alternative splicing. *Gene* **242**, 265-273.

[3] Shan, Z., Parker, T., Wiest, J. S.(2004). Identifying novel homozygous deletions by microsatellite analysis and characterization of tumor suppressor candidate 1 gene, TUSC1, on chromosome 9p in human lung cancer. *Oncogene*, **23**, 6612-6620.

[4] Cai, Z., Chen, H. T., Boyle, B., Rupp, F., Funk, W. D., Dedera, D. A. (2005). Identification of a novel insulin-like growth factor binding protein gene homologue with tumor suppressor like properties. *Biochem. Biophys. Res. Commun.*, **331**, 261-266.

[5] Allen, T., van Tuyl, M., Iyengar, P., Jothy, S., Post, M., Tsao, M.-S., Lobe, C. G. (2006). Grg1 acts as a lung-specific oncogene in a transgenic mouse model. *Cancer Res.*, **66**, 1294-1301.

[6] Haase, D., Meister, M., Muley, T., Hess, J., Teurich, S., Schnabel, P., Hartenstein, B., Angel, P. (2007). FRMD3, a novel putative tumour suppressor in NSCLC. *Oncogene*, **26**, 4464-4468.

[7] Inbal, B., Cohen, O., Polak-Charcon, S., Kopolovic, J., Vadai, E., Eisenbach, L., Kimchi, A. (1997). DAP kinase links the control of apoptosis to metastasis. *Nature*, **390**, 180-184.

[8] Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A, *et al.* (2005). RAS Is Regulated by the let-7 MicroRNA Family. *Cell*, **120**, 635C647.

[9] Gobeil, S., Zhu, X., Doillon, C. J., and Green1, M. R. (2008). A genome-wide shRNA screen identifies GAS1 as a novel melanoma metastasis suppressor gene. *Genes Dev.*, **22**, 2932-2940.

[10] Wang, Z., Shen, D., Parsons, D. W., Bardelli, A., Sager, J., Szabo, S., Ptak, J., Silliman, N., Peters, B. A., van der Heijden, M. S., Parmigiani, G., Yan, H., Wang, T.-L., Riggins, G., Powell, S. M., Willson, J. K. V., Markowitz, S., Kinzler, K. W., Vogelstein, B., Velculescu, V. E. (2004). Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science*, **304**, 1164-1166.

[11] Yano, M., Toyooka, S., Tsukuda, K., Dote, H., Ouchida, M., Hanabata, T., Aoe, M., Date, H., Gazdar, A. F., and Shimizu, N. (2005). Aberrant promoter methylation of human DAB2 interactive protein (hDAB2IP) gene in lung cancers. *Int. J. Cancer*, **113**, 59-66

[12] Cuccurese, M., Russo, G., Russo, A. and Pietropaolo, C. (2005). Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Research*, **33**, 5965-5977.

Table 3: The true positive (TP) rates and false positive (FP) rates for the linear-median method and the cghMCR method where $n = 20$.

$n =$ 20	L-M	cgh MCR	L-M	cgh MCR	L-M	cgh MCR
π	$a = 0.5$		$a = 1$		$a = 1.5$	
0.2						
TP	0.6382	0.0714	0.7568	0.0024	0.8096	0
	(0.0496)	(0.1101)	(0.0406)	(0.0188)	(0.0414)	(0)
FP	0.3785	0.0040	0.6549	2.83e-04	0.7657	0
	(0.0384)	(0.0154)	(0.0384)	(0.0041)	(0.0357)	(0)
0.4						
TP	0.7849	0.6760	0.7696	0.0308	0.7616	0
	(0.0429)	(0.1830)	(0.0413)	(0.0779)	(0.0453)	(0)
FP	0.0861	0.0415	0.3827	0.0011	0.5611	0
	(0.0248)	(0.0302)	(0.0402)	(0.0081)	(0.0408)	(0)
0.6						
TP	0.9503	0.9075	0.8708	0.2759	0.8013	0
	(0.0227)	(0.0224)	(0.0359)	(0.1129)	(0.0410)	(0)
FP	0.0122	2.58e-05	0.2000	0.0023	0.3905	0
	(0.0090)	(0.0004)	(0.0310)	(0.0114)	(0.0419)	(0)
0.8						
TP	0.9966	0.9030	0.9451	0.3877	0.8677	0
	(0.0060)	(0.0206)	(0.0204)	(0.1308)	(0.0331)	(0)
FP	0.0013	0	0.0238	0	0.2617	0
	(0.0028)	(0)	(0.0917)	(0)	(0.0358)	(0)
1						
TP	1	0.9490	0.9817	0.6542	0.9237	0
	(0)	(0.0147)	(0.0147)	(0.1561)	(0.0287)	(0)
FP	7.74e-05	0	0.04026	0	0.1667	0
	(0.0007)	(0)	(0.0154)	(0)	(0.0314)	(0)

Table 4: The true positive (TP) rates and false positive (FP) rates for the linear-median method and the cghMCR method, where $n = 50$.

$n =$ 50	L-M	cgh MCR	L-M	cgh MCR	L-M	cgh MCR
π	$a = 0.5$		$a = 1$		$a = 1.5$	
0.2						
TP	0.6309	0.02442	0.7147	0	0.7521	0
	(0.0547)	(0.0626)	(0.0499)	(0)	(0.0455)	(0)
FP	0.1712	6.45e-04	0.4866	0	0.6425	0
	(0.0346)	(0.0065)	(0.0488)	(0)	(0.0437)	(0)
0.4						
TP	0.8895	0.6643	0.8574	0.0019	0.7975	0
	(0.0347)	(0.1542)	(0.0357)	(0.0109)	(0.0420)	(0)
FP	0.0089	0.0439	0.1737	0	0.3603	0
	(0.0070)	(0.0297)	(0.0365)	(0)	(0.0416)	(0)
0.6						
TP	0.9949	0.9046	0.9581	0.2762	0.8926	0
	(0.0072)	(0.0149)	(0.0212)	(0.0842)	(0.0358)	(0)
FP	6.45e-05	0	0.0482	0	0.1814	0
	(0.0006)	(0)	(0.0189)	(0)	(0.0364)	(0)
0.8						
TP	1	0.8962	0.9912	0.3384	0.9545	0
	(0)	(0.0118)	(0.0100)	(0.0416)	(0.0209)	(0)
FP	0	0	0.0100	0	0.0826	0
	(0)	(0)	(0.0082)	(0)	(0.0238)	(0)
1						
TP	1	0.9207	0.9992	0.4155	0.9848	0
	(0)	(0.0154)	(0.0029)	(0.1679)	(0.0107)	(0)
FP	0	0	0.0023	0	0.0348	0
	(0)	(0)	(0.0038)	(0)	(0.0153)	(0)

Table 5: The true positive (TP) rates and false positive (FP) rates for the linear-median method and the cghMCR method, where $n = 100$.

$n =$ 100	L-M	cgh MCR	L-M	cgh MCR	L-M	cgh MCR
π	$a = 0.5$		$a = 1$		$a = 1.5$	
0.2						
TP	0.6771	0.0048	0.7438	0	0.7335	0
	(0.0539)	(0.0203)	(0.0505)	(0)	(0.0461)	(0)
FP	0.0561	0	0.3266	0	0.5146	0
	(0.0187)	(0)	(0.0412)	(0)	(0.0381)	(0)
0.4						
TP	0.9566	0.6650	0.9299	0.0004	0.8718	0
	(0.0233)	(0.1317)	(0.02653)	(0.0030)	(0.0341)	(0)
FP	0.0003	0.0455	0.0578	0	0.2012	0
	(0.0013)	(0.0270)	(0.0196)	(0)	(0.0340)	(0)
0.6						
TP	0.9998	0.9033	0.9920	02804	0.9556	0
	(0.0015)	(0.0108)	(0.01010)	(0.0706)	(0.0239)	(0)
FP	0	0	0.0065	0	0.0621	0
	(0)	(0)	(0.0075)	(0)	(0.0224)	(0)
0.8						
TP	1	0.8956	0.9998	0.3345	0.9922	0
	(0)	(0.0087)	(0.0015)	(0.0193)	(0.0099)	(0)
FP	0	0	0.0003	0	0.0167	0
	(0)	(0)	(0.0013)	(0)	(0.0125)	(0)
1						
TP	1	0.8971	0.9998	0.2263	0.9983	0
	(0)	(0.0177)	(0.0015)	(0.1156)	(0.0044)	(0)
FP	0	0	0.0003	0	0.0043	0
	(0)	(0)	(0.0013)	(0)	(0.0056)	(0)