Marginal Independence of INI Filtering and Test Statistics

Sepp Hochreiter

[1]Institute of Bioinformatics,
Johannes Kepler University Linz, Linz, Austria

December 15, 2010

1 Control of Type I Error Rate by I/NI Calls

In the following we show that for permutation invariant test statistics and for the t-test statistic T, the I/NI call filter applied to null hypotheses is independent of the statistic. The result is given in Theorem 1 at the end of this section. The theorem guarantees type I error rate control if first filtering by I/NI calls, then using these statistics, and finally applying correction for multiple testing.

To proof this theorem, first we need some results on summarization with Robust Multi-array Average (RMA) for Gaussian noise and for correlated probes in the probe sets. These results are given in the following lemmas.

1.1 RMA Summarization of Gaussian Probes

Robust Multi-array Average (RMA) summarizes a probe set by median polish. After removing sample median (the first RMA step), the sample effects are small and RMA basically computes the median of the probe set.

We assume a probe set with $(2m + 1)$ probes. According to Chu (1955), for $(2m + 1)$ samples drawn from a normal distribution with density $f(x) \sim \mathcal{N}(\xi, \sigma)$ and cumulative distribution function $F(x)$, the distribution of samples' median is

$$p(x) \ = \ \frac{(2m+1)!}{m!\ m!}\ \left(F(x)\ (1 - F(x)\right)^m\ f(x)\ . \qquad (1)$$

According to Chu (1955), $p(x)$ is asymptotically normal which is is formulated in following lemma.

Lemma 1 *For $2m+1$ samples randomly drawn according to a normal distribution $f(x) \sim \mathcal{N}(\xi, \sigma)$, the sample median is asymptotically normal distributed with mean ξ and variance*

$$\sigma_m^2 \ = \ \frac{1}{4\ f^2(\xi)\ (2m+1)}\ . \qquad (2)$$

Proof: This lemma is shown in Chu (1955).
Proof complete.

In Chu (1955) it is stated that the distribution of the median "tends 'rapidly' to normality." Using the bounds in Chu (1955), for a probe set of 16 probes (a standard Affymetrix probe set), the factor deviating from a normal distribution is between 0.9858317 and 1.023438.

1.2 RMA Summarization of Correlated Gaussian Probes

Now we consider summarization in the case where the probes of a probe set are correlated and driven by a hidden signal stemming from targeting the same mRNA. To introduce correlated probes, we assume a signal ξ_k for sample k, where ξ_k is the intensity of all probes. The probes of a probe set are noisy with Gaussian noise $\mathcal{N}(0, \sigma)$, therefore the median of the probes follows for fixed ξ_k the Gaussian distribution $\mathcal{N}(\xi_k, \sigma_m)$. The signal ξ_k is drawn from a Gaussian signal distribution $\mathcal{N}(\mu_s, \sigma_s)$, where (μ_s, σ_s) determine the signal strength. The probes are now correlated across samples where (μ_s, σ_s) determines the strength of correlation.

Alternatively, we could have introduced correlated probes by a linear scaled signal for each sample which is noisy observed in each probe. This is equivalent to above approach. To see this, let a multiplicative factor ρ_k, which scales the reference signal μ, follow a Gaussian $\mathcal{N}(\mu_r, \sigma_r)$. The new mean values ξ_k follow a Gaussian $\mathcal{N}(\mu\ \mu_r, \mu^2\ \sigma_r)$ which is equivalent to above approach to introduce correlation by setting $\mu_s = \mu\ \mu_r$ and $\sigma_s = \mu^2\ \sigma_r$.

Because the signal distribution determines the mean of the median distribution, the distribution of the median is the convolution of two Gaussian distributions $\mathcal{N}(\mu_s, \sigma_s)$ and $\mathcal{N}(0, \sigma_m)$.

Lemma 2 *If the correlation signal of probes of a probe set is drawn from a Gaussian distribution $\mathcal{N}(\mu_s, \sigma_s)$ and the noise of the probes is $\mathcal{N}(0, \sigma)$, then the median distribution is*

$$\mathcal{N}(\mu_s, \sigma_x)\ , \qquad (3)$$

where

$$\sigma_x^2 \ = \sigma_s^2 \ + \ \sigma_m^2 \ = \ \sigma_s^2 \ + \ \frac{1}{4\ f^2_{\mathcal{N}(0,\sigma)}(\xi)\ (2m+1)} \qquad (4)$$
$$= \ \sigma_s^2 \ + \ \frac{\pi\ \sigma^2}{2\ (2m+1)}\ ,$$

Proof: The lemma follows from Lemma 1 which states that the distribution of the median is $\mathcal{N}(\xi_k, \sigma_m)$ for fixed ξ_k. If ξ_k is drawn according to $\mathcal{N}(\mu_s, \sigma_s)$ then the median distribution is obtained by the convolution of $\mathcal{N}(0, \sigma_m)$ and $\mathcal{N}(\mu_s, \sigma_s)$. The distribution given in the lemma is the result of this convolution.
Proof complete.

Introducing correlations in other way would not change the results but the convolution for non-Gaussian signal distributions might be more complicated.

1.3 Independence of I/NI Filter and Test Statistic for Null Hypotheses

The Informative/NonInformative (I/NI, Talloen *et al.*, 2007) call tries to access the noise part σ^2 of the overall variance by $\mathrm{var}(z \mid \boldsymbol{x})$. Thus, the amount of signal σ_s in the probe set is estimated.

More specifically, according to Talloen *et al.* (2007) the I/NI call is

$$\mathrm{var}(z \mid \boldsymbol{x}) \ = \left(\frac{(2m+1)\ \sigma_s^2}{\sigma^2} \ + \ 1 \right)^{-1} \ < \ 0.5\ , \qquad (5)$$

where $\frac{\sigma_s^2}{\sigma^2}$ is the signal-to-noise ratio.

Probe sets containing a signal and probe sets not containing a signal, are both normal distributed. However, probes sets with a signal have larger variance because the signal variance σ_s is added to the variance of the median according to Lemma 2.

We use the notation in Bourgon *et al.* (2010) and define permutation invariance for sample size n.

Definition 1 *A test statistic U^{II} is* permutation invariant *if for fixed $\boldsymbol{Y}_i \in \mathbb{R}^n$, $i \in \mathcal{H}_0$, and Π drawn uniformly from S_n (the set of all permutations on n elements), the distribution of the test statistic $U^{II}(\boldsymbol{Y}_i)$ is equal to the distribution of $U^{II}(\Pi(\boldsymbol{Y}_i))$.*

Now we can formulate our main theorem that for permutation invariant test statistics and for the t-test statistic T, the I/NI call filter applied to null hypotheses is independent of the statistic. The theorem guarantees type I error rate control if applying correction for multiple testing.

Theorem 1 *For permutation invariant test statistics like the Wilcoxon rank sum statistic and for t-test statistic T, the I/NI call filter applied to null hypotheses is independent of the statistic.*

Proof: First we note that the I/NI call for one probe set does not dependent on another probe set as the models are independently selected for each probe set.

A) *Permutation invariant test statistics*:
For permutation invariant test statistics the statement follows directly from the permutation invariance of the I/NI call filter. The I/NI call is permutation invariant because the I/NI call model selection objective, the *a posteriori* of the parameters, is independent of the permutation of the samples. Further, the implementation of the algorithm uses only the data covariance matrix Hochreiter *et al.* (2006) which is independent of permutations of the samples.

All assumptions on the filter of the the proposition "Marginal Independence: Permutation Invariance" in Bourgon *et al.* (2010) are fulfilled. The independence between the I/NI call filter and permutation invariant test statistics is shown.

B) *t-test statistic T*:
As pointed out by Bourgon *et al.* (2010) in their supplementary, the test statistics T for the t-test is invariant to scaling and shifting of the mean. If the noise level σ is equal for each probe set, then I/NI call is equivalent to variance filtering because only the signal variance σ_s determines the overall variance. The more interesting case is where signal and noise differ at each probe set, thus variance filtering and I/NI calls yield different results.

For probe set i the signal is drawn from a Gaussian distribution $\mathcal{N}(\mu_{si}, \sigma_{si})$. According to Lemma 2 the RMA summarized data follows the Gaussian $\mathcal{N}(\mu_{si}, \sigma_{xi})$, where $\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2m+1)}})$. Let us assume that the signal strength and the noise level $(\mu_{si}, \sigma_{si}, \sigma_i)$ is drawn from some distribution $P_{(\mu_{si}, \sigma_{si}, \sigma_i)}$.

The data $\boldsymbol{Y}_i$ can be generated by first drawing n samples from a standard normal distribution giving $\boldsymbol{X}_i \in \mathbb{R}^n$, where $P_{\boldsymbol{X}_i} \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ with $\boldsymbol{0}$ as the n-dimensional zero vector and $\boldsymbol{I}_n$ as the n-dimensional identity matrix. Then $\boldsymbol{X}_i$ is scaled by $\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2m+1)}}$ and shifted componentwise by μ_{si}. The shifting and scaling values are drawn from $P_{(\mu_{si}, \sigma_{si}, \sigma_i)}$ which is independent from $P_{\boldsymbol{X}_i}$.

For the null hypothesis $i \in \mathcal{H}_0$, we assume that both distributions $P_{\boldsymbol{X}_i}$ and $P_{(\mu_{si}, \sigma_{si}, \sigma_i)}$ are independent of the conditions $\boldsymbol{C}$.

For showing the independence of filtering U^I and test statistic U^{II}, we are interested in the probability of the event $\{U_i^I \in \mathcal{A}, U_i^{II} \in \mathcal{B}\}$. Here we define $U_i^I(\boldsymbol{Y}) = \mathrm{var}(z \mid \boldsymbol{x})(\boldsymbol{Y})$ with $\mathcal{A} = \{u \mid u < 0.5\}$ and $U_i^{II}(\boldsymbol{Y}) = T(\boldsymbol{Y}, \boldsymbol{C})$ for t-test statistic T, conditions $\boldsymbol{C}$, and $\mathcal{B} = \{u \mid u > \theta\}$. Let $\delta_{\mathcal{A}}$ and $\delta_{\mathcal{B}}$ be indicator functions for $\mathcal{A}$ and $\mathcal{B}$.

We consider a probe set $\boldsymbol{Y}_i$ for which $i \in \mathcal{H}_0$ (a true null hypothesis).

$$
\begin{aligned}
& P\big(U_i^I \in \mathcal{A}, U_i^{II} \in \mathcal{B}\big) \qquad (6)\\
&= \int \delta_{\mathcal{A}}\big(U^I(\boldsymbol{Y}_i)\big)\ \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{Y}_i)\big)\ dP_{\boldsymbol{Y}_i}\\
&= \int\int \delta_{\mathcal{A}}\big(U^I(\mu_{si}\ \boldsymbol{1}\ +\ \boldsymbol{X}_i\ \sigma_{xi})\big)\\
&\quad \delta_{\mathcal{B}}\big(U^{II}(\mu_{si}\ \boldsymbol{1}\ +\ \boldsymbol{X}_i\ \sigma_{xi})\big)\ dP_{\boldsymbol{X}_i}\ dP_{(\mu_{si}, \sigma_{si}, \sigma_i)}\\
&= \int\int \delta_{\mathcal{A}}\big(U^I(\sigma_{si}, \sigma_i)\ \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{X}_i)\ dP_{\boldsymbol{X}_i}\ dP_{(\mu_{si}, \sigma_{si}, \sigma_i)}\\
&= \int \delta_{\mathcal{A}}\big(U^I(\sigma_{si}, \sigma_i)\ dP_{(\mu_{si}, \sigma_{si}, \sigma_i)} \int \delta_{\mathcal{B}}\big(U^{II}(\boldsymbol{X}_i)\big)\ dP_{\boldsymbol{X}_i}\\
&= P\big(U_i^I \in \mathcal{A}\big)\ P\big(U_i^{II} \in \mathcal{B}\big)\ ,
\end{aligned}
$$

where

$$
\sigma_{xi} = \sqrt{\sigma_{si}^2 + \frac{\pi\ \sigma_i^2}{2\ (2m+1)}}) \qquad (7)
$$

and $\boldsymbol{1}$ is the vector of ones with length n. The equality of the 3rd/4th line to the 5th line is obtained by the shift and scale invariance of U^{II} and the fact that U^I depends only on σ_{si} and σ_i.
Proof complete.

Note, that for equal noise level σ on each probe set, the I/NI call is equivalent to variance filtering. Also for a low noise level relative to the signal, I/NI call is similar to variance filtering.

References

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiment. *Proc Natl Acad Sci U S A*, **107**(2), 9546–9551.

Chu, J. T. (1955). On the distribution of the sample median. *Ann. Math. Statist.*, **26**(1), 112–116.

Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**(8), 943–949.

Talloen, W., Clevert, D.-A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., and Göhlmann, H. W. H. (2007). I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**(21), 2897–2902.