

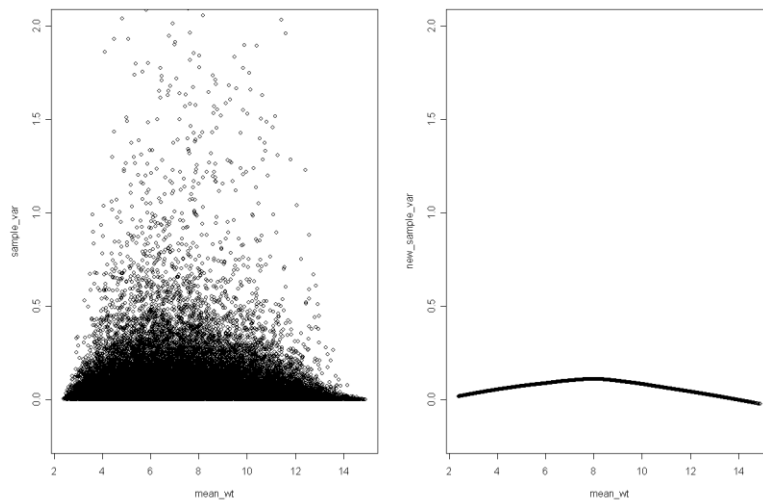
Part 2

Intensity-based hierarchical Bayes method for PPAR gene related microarray data

Yang Xu

Introduction

Generally, t test was used to judge if there is a difference between treatment and control group of gene expression level. Variance estimation is a key point in the t test, especially when sample size is small. Therefore, accurate estimation of the variability of gene expression in microarray data is crucial in identifying differently expressed genes. Usually variance was assumed as consistent throughout the whole microarray data for convenience of calculation. However, this assumption is not fair for some data (see Figure below, cited from part 1). The plot of variance vs. expression level mean displays a certain pattern, which disapproves the assumption of equal variance. A new way called intensity-base hierarchical Bayes method is introduced here to precisely estimate the variance.



Method

Variance regression

It is easy to tell that the variance of expression level is intensity dependent. There is a non-linear relationship between log-variance and log-average intensity. According to *Medvedovic et al*, this non-linear relationship can be modeled by the function

$$g(x) = p_1 e^{-0.8(x-5)} + p_2 \quad (1)$$

(x ---mean of expression level, $g(x)$ ---estimated prior variance). The following values used for p_1 and p_2 : low: $p_1 = p_2 = 0.875$, medium: $p_1 = 1.25$ and $p_2 = 0.5$, and high: $p_1 = 1.5, p_2 = 0.25$. Here, for our data, we can see the dependence of variance on expression level exists, but it's not very strong. Therefore we choose low: $p_1 = p_2 = 0.875$.

T test

$$t_{gi} = \frac{\hat{\beta}_{gi}}{\hat{S}_g \sqrt{V_{gi}}} \quad (2)$$

S_g (hat) is posterior standard deviation. V_{gi} is $1/n_1 + 1/n_2$, in our sample $1/3 + 1/3$.

Priors

S_{0g} is the estimated prior variance for each gene. It comes from the nonlinear function (1) ($g(x) = S_{0g}$).

Sample standard deviation S_g follows scaled F distribution, $S_g^2 \sim S_{0g}^2 F_{dg, d_0}$. The sample degree of freedom dg is believed as constant in our sample ($dg = n_1 + n_2 - 2 = 4$). Then we can estimate prior degree of freedom d_0 by $\psi'(d_0/2) = \text{mean}[S_g^2 - \text{pred}(S_g^2)]^2 - \text{mean}[\psi'(d_g/2)]$ (suggested by *Medvedovic et al*).

Posteriors

$$df = d_0 + d_g$$

$$\hat{S}_g^2 = \frac{d_0 S_{0g}^2 + d_g S_g^2}{d_0 + d_g}$$

}

Results:

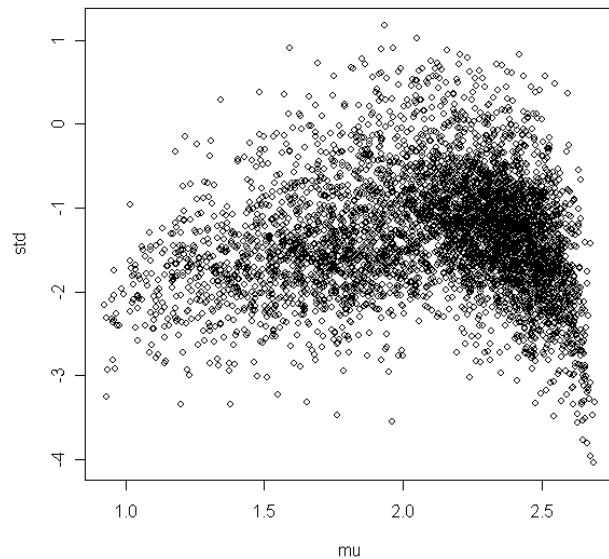
The two sample t test (formula 2) with estimates of variance and significant level $p=0.05$ was carried out by R codes showed in the R codes section below. Among the 45101 genes, there are 9472 genes (21%) expression level difference can be detected by this method.

For the 4121 genes Jinlu and Jing selected, the dependence of variance on expression level is more obvious (see the curve below). Through the intensity-based method, 2138 genes (51.88%) are identified as expressing differently.

Discussion

The intensity-based hierarchical Bayes model could work better if the dependence of variance on expression level is more obvious. As in our two examples, more of selected 4121 genes have been detected in different expression level compared than the whole dataset with 45101 genes. We made the assumption that the selected 4121 genes have the same genetic background with the whole genomes.

Therefore, consideration of intensity-dependence would produce more accurate estimates of variance, subsequently increase the possibility of detecting differently expressing genes.



R codes

R codes come from Medvedovic et al after a modification to apply to our data.

```
library("stats")
```

```
library("limma")
```

```
logVAR<-log(mdata$sigma^2)
```

```
df<-mdata$df.residual
```

```
numgenes<-length(logVAR[df>0])
```

```
df[df==0]<-NA
```

```
eg<-logVAR-digamma(df/2)+log(df/2)
```

```
egpred<-loessFit(eg,mdata$Amean,iterations=1,span=0.3)$fitted
```

```
myfct<-(eg-egpred)^2 - trigamma(df/2)
```

```
print("Local regression fit")
```

```
mean.myfct<-mean(myfct,na.rm=TRUE)
```

```
priordf<-vector(); testd0<-vector()
```

```
for (i in 1:(numgenes*10)) {
```

```

        testd0[i]<-i/10

        priordf[i]= abs(mean.myfct-trigamma(testd0[i]/2))

        if (i>2) {
            if (priordf[i-2]<priordf[i-1]) { break }
        }
    }

    d0<-testd0[match(min(priordf),priordf)]
    print("Prior degrees freedom found")

    s02<-exp(egpred + digamma(d0/2) - log(d0/2))

    post.var<- (d0*s02 + df*mdata$sigma^2)/(d0+df)
    post.df<-d0+df

    IBMTt<-mdata$coefficients[,testcol]/(mdata$stdev.unscaled[,testcol]*sqrt(post.var))
    IBMTp<-2*(1-pt(abs(IBMTt),post.df))
    print("P-values calculated")

output<-mdata
    output$IBMT.t<-IBMTt
    output$IBMT.p<-IBMTp
    output$IBMT.postvar<-post.var
    output$IBMT.priorvar<-s02
    output$IBMT.dfprior<-d0
    output
}

```

Reference

Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. M Medvedovic et al. BMC Bioinformatics 2006, 7:538