An improved probabilistic model for finding differential gene expression

Li Zhang   Xuejun Liu
College of Information Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China

***Abstract*—Finding differentially expressed genes is a fundamental objective of a microarray experiment. Recently proposed method, PPLR, considers the probe-level measurement error and improves accuracy in finding differential gene expression. However, PPLR uses the importance sampling procedure in the E-step of the variational EM algorithm, which leads to less computational efficiency. We modified the original PPLR to obtain an improved model for finding different gene expression. The new model, IPPLR, adds hidden variables to represent the true gene expressions and eliminates the importance sampling in original PPLR. We apply IPPLR on a spike-in data set and a mouse embryo data set. Results show that IPPLR improves accuracy and computational efficiency in finding differential gene expression.**

I. Introduction

Microarray [1] [2] are currently widely used to obtain large-scale measurements of gene expression. Finding differentially expressed (DE) genes is the most basic objective of a microarray experiment. Due to the notorious noise existing in microarray data, replicates are usually used in the experiments to deal with data variability. Moreover, some microarrays (such as Affymetrix GeneChips) contain multiple probes to interrogate gene expression profiles. This provides rich information to obtain an estimation of the technical measurement error associated with each gene expression measurement. This error information is especially significant for weakly expressed genes as these genes are often associated with high variability. Probabilistic methods provide a principle way to handle noisy data. Most of the probabilistic methods, such as the widely used methods, Cyber-T [3] and SAM [4], are based on single point estimates of gene expression values, and ignore the associated probe-level measurement error. This wastes rich information in data.

Measurement error of data points has received more and more attention in noisy data analysis [5] [6] [7] [8] in recent years. PPLR [5] considers the probe-level measurement error in finding differential gene expression. This method has been proved to be more accurate than other alternatives [5] [9]. However, PPLR uses the importance sampling procedure in the E-step of the variational EM algorithm. This leads to bad accuracy and less computational efficiency. Especially, when the experiment involves a large number of chips, PPLR is extremely time-consuming. This makes the application of PPLR difficult in reality.

In this contribution, we improve PPLR by adding hidden variables to represent the true gene expression. This eliminates the inefficient importance sampling in original PPLR. Results on a spikes-in data set and a mouse embryo data set show that the improved PPLR, IPPLR, improves accuracy and computational efficiency in finding DE genes.

II. Method

A. The Original PPLR

The original PPLR uses both gene expression measurements and probe-level measurement error to obtain more accurate results in finding DE genes [5]. For a particular gene in PPLR, the observed logged expression level for the ith replicate under the jth condition is assumed to follow a Gaussian distribution,

$$\hat{x}_{ij} \sim \mathcal{N}(\mu_j, \lambda_j^{-1} + s_{ij}^2), \tag{1}$$

where μ_j is the mean logged expression level under condition j, λ_j is the inverse of the between-replicates variance and s_{ij}^2 is the probe-level measurement error, which can be calculated from probabilistic probe-level analysis methods, such as multi-mgMOS [10].

PPLR uses a Bayesian approach for the combination of replicate measurements. It assumes that the parameters $\theta = \{\{\mu_j\}, \{\lambda_j\}\}$ are independent and λ_j^{-1} is shared across different conditions to capture the gene-specific variability. The priors of the the parameters are:

$$\begin{aligned} &\mu_j \sim \mathcal{N}(\mu_0, \eta_0^{-1}), \\ &\lambda \sim \mathrm{Ga}(\alpha, \beta), \end{aligned} \tag{2}$$

where $\phi = \{\mu_0, \eta_0, \alpha, \beta\}$ are hyperparameters. The PPLR model can be depicted in Fig.1a.

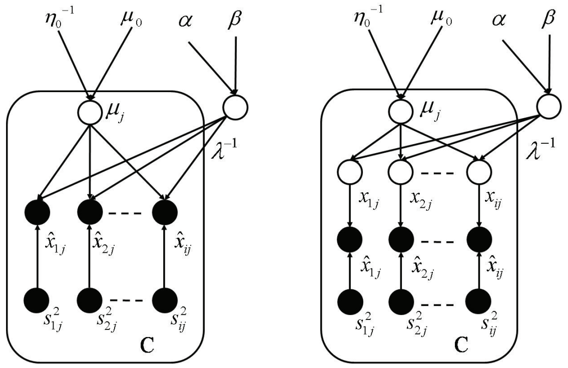
To calculate the parameters, PPLR method uses a variational Expectation-Maximization algorithm [11]. In the E-step of the EM algorithm, the distribution of λ in the t-th iteration is approximated by

$$Q(\lambda) \propto Ga(\lambda; \alpha^t, \beta^t) \prod_{ij} Ga\bigg(p_{ij}; \frac{3}{2}, \frac{1}{2}\langle(\hat{y}_{ij} - \mu_j)^2\rangle\bigg), \tag{3}$$

where $\langle\cdot\rangle$ denotes the expectation of a function with respect to $Q(\mu)$ and $p_{ij} = (\lambda^{-1} + s_{ij}^2)^{-1}$.

Since the $Q(\lambda)$ distribution has no standard form, importance sampling is therefore used to obtain the expectations of

978-1-4244-4134-1/09/$25.00 ©2009 IEEE



(a) (b)

Fig. 1: The models of (a) PPLR, and (b) IPPLR. The black circles represent the observed data and the blank circles represent the hidden parameters. C is the number of conditions.

a function $g(\lambda)$ with respect to $Q(\lambda)$,

$$\langle g(\lambda)\rangle = \sum_{k=1}^{S} w_k g(\lambda_k), \text{where} \quad w_k = \frac{f(\lambda_k)}{\sum_k f(\lambda_k)}, \tag{4}$$

$f(\lambda)$ denotes $\prod_{ij} Ga\left(p_{ij}; \frac{3}{2}, \frac{1}{2}\langle(\hat{y}_{ij} - \mu_j)^2\rangle\right)$ and S is the number of samples taking from $Ga(\lambda; \alpha^t, \beta^t)$. As the number of the chips increases, the distribution of $Q(\lambda)$ gets more and more flat. Importance sampling needs to increase S to improve the approximation of $Q(\lambda)$. This makes the computation extremely slow and unacceptable in practice.

B. IPPLR

In order to overcome the limitation of the original PPLR model, we propose an improved IPPLR model (IPPLR) to avoid the importance sampling procedure. Adopting the similar strategy in [8], the new model adds a hidden variable, which represents the true expression for each gene on each chips. The diagram plate of IPPLR is shown in Fig.1b.

IPPLR also uses both gene expression measurements and probe-level measurement error to obtain high accuracy in finding DE genes. We add a hidden variable x_{ij} into the original model, representing the true gene expression. We assume that the variable is Gaussian distributed $x_{ij} \sim \mathcal{N}(\mu_j, \lambda^{-1})$, where μ_j is the mean logged expression level under condition j and λ is the inverse of the between-replicates variance and is shared across different conditions. The measured expression level $\hat{x}_{ij}$ can be expressed as:

$$\hat{x}_{ij} \sim \mathcal{N}(x_{ij}, s_{ij}^2), \tag{5}$$

where s_{ij}^2 is the probe-level measurement error, which can be obtained from multi-mgMOS.

We make a prior assumption that μ_j and λ^{-1} are independent and put a Gaussian prior on μ_j,

$$\mu_j \sim \mathcal{N}(\mu_0, \eta_0^{-1}), \tag{6}$$

where μ_0 and η_0 are hyperparameters, on which we adopt noninformative hyperpriors. We assume a conjugate gamma prior on λ,

$$\lambda \sim \text{Ga}(\alpha, \beta). \tag{7}$$

The new hierarchical model includes the parameters $h = (\{\mu_j\}, \{x_{ij}\}, \lambda)$ and the hyperparameters $\theta = (\mu_0, \eta_0, \alpha, \beta)$.

Similar to PPLR, we also use the EM algorithm combined with a variational method [12] [13] to work out the model parameters. We maximize the following function with respect to $Q(h)$ and θ iteratively.

$$\mathcal{L}(h,\theta) = \int dhQ(h)\log P(D|h,\theta)P(h|\theta) - \int dhQ(h)\log Q(h), \tag{8}$$

where D represents the observed data. Meanwhile, we assume that the parameters are independent, so $Q(h)$ can be factorized as:

$$\begin{aligned} Q(h) &= Q(\{x_{ij}\})Q(\{\mu_j\})Q(\lambda) \\ &= \prod_{ij} Q(x_{ij}) \prod_{j} Q(\mu_j)Q(\lambda). \end{aligned} \tag{9}$$

We use (9) instead of the $Q(h)$ to maximize (8) with respect to $Q(x_{ij})$, $Q(\mu_j)$ and $Q(\lambda)$.

The procedure of IPPLR is similar to the original PPLR, except for the E-step, where the calculations of $Q(x_{ij})$,$Q(\mu_j)$ and $Q(\lambda)$ are all tractable,

$$Q(x_{ij}) \propto \mathcal{N}\left(x_{ij}; \frac{\hat{x}_{ij} + s_{ij}^2\langle\mu_j\rangle\langle\lambda\rangle}{1 + s_{ij}^2\langle\lambda\rangle}, \frac{s_{ij}^2}{1 + s_{ij}^2\langle\lambda\rangle}\right), \tag{10}$$

$$Q(\mu_j) \propto \mathcal{N}\left(\mu_j; \frac{\mu_0^t\eta_0^t + \langle\lambda\rangle\sum_i^r\langle x_{ij}\rangle}{\eta_0^t + \sum_i^r\langle\lambda\rangle}, (\eta_0^t + \sum_i^r\langle\lambda\rangle)^{-1}\right), \tag{11}$$

$$Q(\lambda) \propto Ga\left(\lambda; \alpha^t + \sum_{ij}\frac{1}{2}, \beta^t + \sum_{ij} C_{ij}\right). \tag{12}$$

In the above equations, $C_{ij} = \frac{1}{2}\langle(x_{ij} - \mu_j)^2\rangle$ and $\langle\cdot\rangle$ denotes the expectation of a function with respect to $Q(x_{ij})$, $Q(\mu_j)$ or $Q(\lambda)$. The three distributions are all in the standard forms, so IPPLR avoids the inefficient importance sampling procedure. We find that the EM algorithm may need more iterations to converge for a small number of genes. We control the convergence by setting a maximum number of iterations, N.

IPPLR and PPLR use the same method to compute the probability of positive log-ratio (PPLR) to represent the significance of differential expression between any two conditions. The obtained significance values are judged by a level of confidence, like α-level in the conventional statistical test. For more information please refer to [5].

The IPPLR model has been implemented in an R package, ipplr, which is currently available from http://parnec.nuaa.edu.cn/liux/zhangl.

III. Results and Discussion

We compare our method with the original PPLR algorithm and other previous methods to show the improvement by averting the importance sampling procedure. Two data sets are used to evaluate the performance of IPPLR. One is a wholly defined spike-in data set, called the Golden Spike-in data set

TABLE I: AUC values from IPPLR, PPLR and other methods under 1-sided test of up-regulation and 2-sided DE test. FC is the simple fold change method in finding DE genes. GCRMA, DFW, CP, FARMS and multi-mgMOS are different summarization methods in probe-level analysis of microarray data. Bold numbers means the best values on each case.

	All fold change		Low fold change		1.2 fold change	
Method	1-sided	2-sided	1-sided	2-sided	1-sided	2-sided
FC+GCRMA	-	-	-	-	0.708	-
FC+DFW	-	-	-	-	-	0.530
Cyber-T+CP	-	-	0.886	0.825	-	-
Cyber-T+multi-mgMOS	0.949	0.919	-	-	-	-
PPLR+FARMS	-	-	-	-	-	0.538
PPLR+multi-mgMOS	0.951	0.922	0.888	0.821	0.704	-
IPPLR+multi-mgMOS	**0.953**	**0.928**	**0.899**	**0.849**	**0.727**	**0.599**

[14], which is a validated benchmark in finding differential gene expression [9]. The other is a real-world mouse embryo data set [15], which shows that the Oct4-regulated gene set is changed at the 1- to 2-cell stages in early mouse embryos and is validated by q-PCR data.

A. Golden Spike-in Dataset

The golden spike-in data is recently used for the validation of algorithms in finding DE genes [9]. The golden spike-in data set includes two conditions each of which has three replicates chips. This data set contains a large number of differentially expressed genes with known fold change, 1.2 to 4, and provides enough true positives to obtain adequate statistics. The data set contains 14010 genes, which are composed of 1331 up-regulated genes and 12679 invariant genes. For more information please refer to [14]. This data set has been intensively investigated on the validation of different approaches in finding DE genes [9]. The extensive investigation in [9] provides an evaluation tool, AffyDEComp, to compare DE detection methods combined with various gene expression summarization approaches.

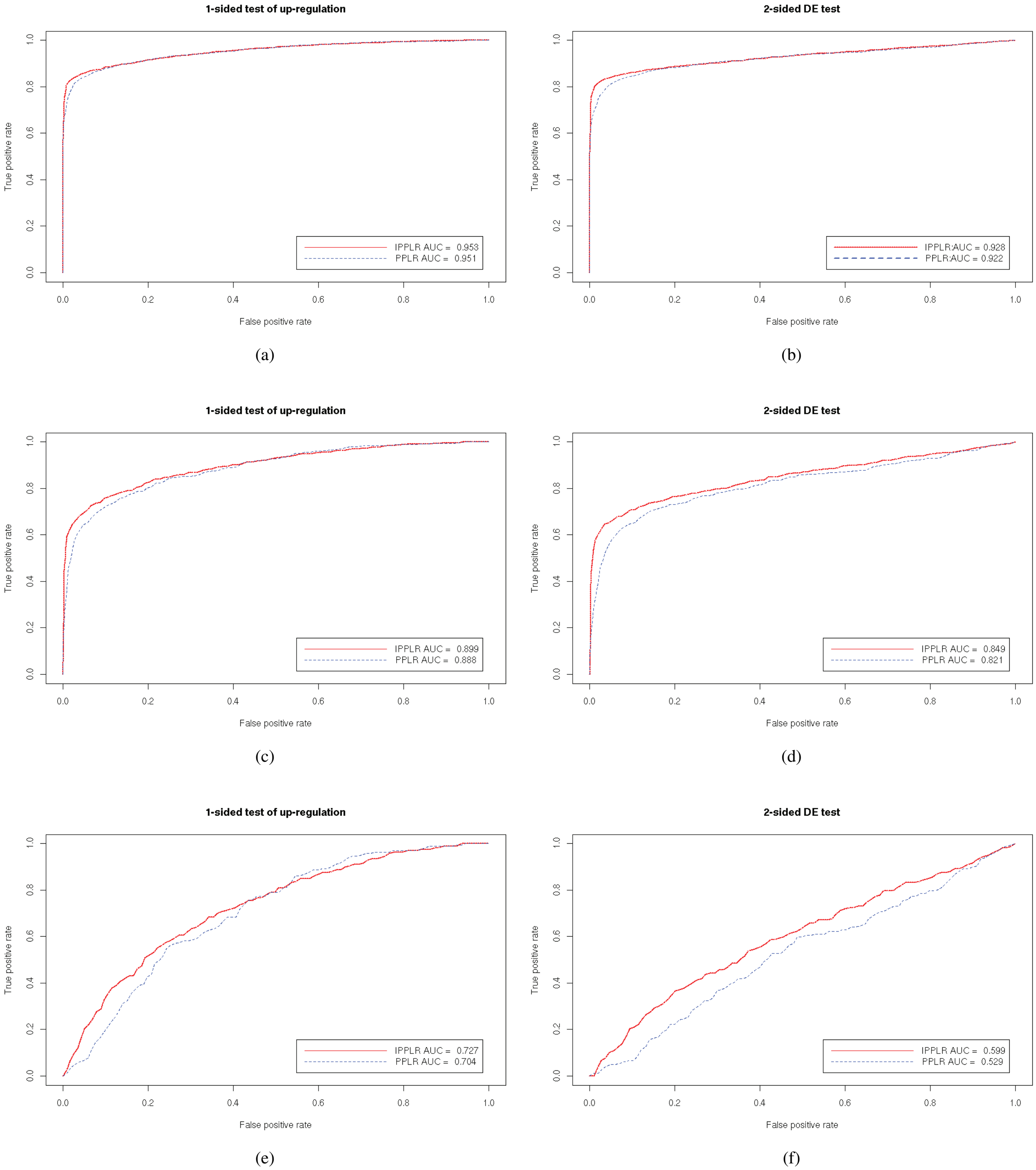
We use AffyDEComp to plot Receiver Operator Characteristic (ROC) curves and calculate Area Under ROC curves (AUC) on the golden spike-in data set, and show the accuracy of different methods. We show the ROC curves of PPLR and IPPLR in Fig.2. The ROC curves of other approaches are included in [9]. We set the maximum number of N in IPPLR, as 200. We find that the accuracy of IPPLR can not be improved obviously for the values of N larger than 200 in this data set. Considering all DE genes, the top two AUC values obtained in [9] are from PPLR and Cyber-T, both combined with multi-mgMOS. The first two columns in Table I shows the AUC values from Cyber-T, PPLR and IPPLR under 1-sided test of up-regulation and 2-sided DE test. IPPLR uses the gene expression values obtained from multi-mgMOS. On these two test cases, IPPLR obtains the highest AUC values, 0.953 and 0.928. We find that on these two tests, IPPLR outperforms both PPLR and Cyber-T.

The golden spike-in data set has some DE genes with low fold change, ranging from 1.2 to 1.7. These genes are always difficult to detected in previous methods. We especially compare IPPLR with the methods which obtain the top two AUC values in [9] on these low fold change cases. The last four columns in Table I show the AUC values from IPPLR and the previous top two AUC values under 1-sided test of up-regulation and 2-sided DE test for low fold change case (1.2 $\sim$1.7) and 1.2 fold change case. For low fold change case, the original top two methods are Cyber-T and PPLR which are combined with CP [9] and multi-mgMOS respectively. In the 1-sided test of up-regulation for 1.2 fold change case, the original top two methods are fold change (FC) and PPLR which combined with GCRMA [16] and multi-mgMOS respectively. In 2-sided DE test for 1.2 fold change case, FC and PPLR, which are combined with DFW [17] and FARMS [18] respectively, obtain the original top two AUC values. From results in Table I, IPPLR combined with multi-mgMOS goes beyond these methods with obvious higher AUC values for these low fold change cases.

B. Mouse Embryo Dataset

In order to further evaluate our method, we apply IPPLR on a real-world mouse embryo data set [15]. The mouse embryo data set is used to discriminate the Octa4-regulated gene set at the 1- to 2-cell stages of early embryos.The data set includes four conditions each of which has three replicates. The two experiment conditions are Oct4-MO-injected and Ccna2-MO-injected, and the others are two control conditions. Under Oct4-MO-injected and the corresponding control condition, 42 genes are selected to represent transcriptional, post-transcriptional and signalling function for q-PCR assays. After removing 3 genes for which there were technical difficulties, 34 genes show altered expression level in Oct4 knockdown in the expected directions, while 5 genes, did not change. 21 of 34 genes, show statistically significant differential expression by q-PCR at $p \leq 0.05$ or less. Please refer to [15] for more information.

Removing two altered genes, Eif3s10 and Dppa5, which do not have corresponding probesets in microarray data, we find 93 probesets related to the remaining 32 altered genes in mouse embryo data set. We use the q-PCR results of these 32 altered genes as golden standard for finding DE genes in microarray data. The study in [15] uses dChip [19] to normalize the raw data and a method, which is based on fold change, logged fold change and unpaired t-statistic [20], is used to detect the DE genes. 1254 probe-sets, corresponding to



(a) (b)

(c) (d)

(e) (f)

Fig. 2: Comparison of IPPLR with PPLR, ROC charts of Golden Spike data using a 1-sided test of up-regulation and 2-sided DE test. We use the same processing procedure as in [9]. Only the equal spike-ins are used as true negative. The different fold change spike-ins are used as true positives. A post-summarization loess normalization base on the equal-value spike-ins was used. (a) probesets selected using a 1-sided test of up-regulation, (b) probesets selected using a 2-sided DE test, (c) the spike-in genes with 1.2<FC<1.7 are used as true positives under 1-sided test of up-regulation, (d) the spike-in genes with 1.2<FC<1.7 are used as true positives under 2-sided DE test, (e) the spike-in genes with FC=1.2 are used as true positives under 1-sided test of up-regulation, and (f) the spike-in genes with FC=1.2 are used as true positives under 2-sided DE test.

TABLE II: Finding differential gene expression among the 32 q-PCR validated genes (related to 93 probe-sets) in the mouse embryo data set. The method in [15] is denoted as UPT-test. [15] does not provide the α-level used in UPT-test. IPPLR and PPLR both calculate the probability of positive log-ratio of the significant genes. Note that the α-level shown for IPPLR and PPLR has different meaning from the significant level in a conventional statistical test. For different N used in IPPLR, we choose different α-level values, 0.00234, 0.0037 and 0.004817 for 200, 500, and 1000 respectively. Bold numbers means the best values on each case.

	UPT-test	PPLR	IPPLR_200	IPPLR_500	IPPLR_1000
α -level	-	0.000977	0.00234	0.0037	0.004817
Consistent probe-sets to q-PCR	43	38	40	43	**44**
Consistent rate to q-PCR	0.462	0.409	0.430	0.462	**0.473**
Number of significant genes	1254	1254	1254	1254	1254

DE genes, were identified by a threshold of 5 percent median false discovery rate (FDR).

We compare IPPLR with PPLR and the method in [15]. We set N in IPPLR as 200, 500 and 1000 respectively. For values of N larger than 1000, the accuracy of IPPLR are not obviously improved. In order to make different methods comparable, we select different credibility levels so that the methods obtain the similar number of significant genes to 1254 obtained in [15]. The number of significant genes at the different credibility levels for different methods are shown in the Table II. Among the 93 probesets related to 32 q-PCR validated DE genes, the consistent rate to q-PCR results obtained by the method in [9] is 0.462, and PPLR is 0.409. For our method, the consistent rates for different N, 200, 500 and 1000, are 0.430, 0.462 and 0.473, respectively. We find that IPPLR performs better than PPLR using different values of N. If we select a proper value of N (500 and 1000), we can obtain the same or better results than the method in [15]. However, as the value of N increases, the computation of IPPLR gets more intensive. This is to be discussed in the next section.

C. Computational Efficiency

We compare IPPLR with PPLR in terms of computational efficiency on the golden spike-in data set and mouse embryo data set. Computation time for the two methods are shown in the Table III.

TABLE III: The computation time (in hours) of PPLR and IPPLR on the spike-in data set and the mouse embryo dataset. Computation time is obtained on a 1.6GHz Intel machine with 1G RAM.

Data	PPLR	IPPLR_200	IPPLR_500	IPPLR_1000
Golden	1.7	1.4	1.5	1.5
Mouse	20	5	7.5	14.5

From Table III, the improvement of computational efficiency for IPPLR on the spike-in data set is not obvious compared with PPLR, but the improvement is obvious on the mouse embryo data set. This is caused by the different numbers of chips used in these two data sets. As the number of chips increases, the importance sampling used in the E-step in PPLR needs a larger number of samples to obtain a reasonable approximation of $Q(\lambda)$. For a small number of samples (default 1000), the EM algorithm needs more iterations to reach convergence. However, the E-step in IPPLR is tractable and the size of the data set has less effects on the computational efficiency of IPPLR than PPLR.

We also notice that the computation of IPPLR gets slower as the value of N increases, but the accuracy gets higher as shown in section 3.2. Although there is a tradeoff between the computational efficiency for IPPLR, IPPLR still obtains improved performance compared with PPLR.

IV. Conclusion

We propose an approach to lift the limitation of the original PPLR, which is mainly coursed by the importance sampling procedure in the EM algorithm. The tractable E-step in our new method, IPPLR, leads to fast EM iterations, especially for experiments involving a large number of chips. Results from the golden spike-in data set and a mouse embryo data set show that IPPLR improves the accuracy and computation efficiency in finding DE genes. IPPLR has been implemented in an R package, *ipplr*, for public use of our method.

Acknowledgment

The authors acknowledge the support from the National Science Foundation of China (60703016), the Jiangsu Science Foundation (BK2007589), Specialized Research Fund for the Doctoral Program of Higher Education (20070287051) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] Lockhart. D.J, Dong. H, Byrne. M.C, Follettie. M.T, Gallo. M.V, Chee. M.S, Mittmann. M, Wang. C, Kobayashi. M, Horton. H. and Brown. E.L, *Expression monitoring by hybridization to high-density oligonucleotide arrays*. Nat.Biotechnol, 1996.

[2] Schena. M, Shalon. D, Davis. R.W, and Brown. P.O, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*. Science, 1995.

[3] Baldi. P and Long. A.D, *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*. Bioinformatics, 2001.

[4] Tusher. VG, Tibshirani. R and Chu. G, *Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci USA, 2001.

[5] Liu. X, Milo. M, Lawrence. N.D and Rattray. M, *Probe-level measurement error improves accuracy in detecting differential gene expression*. Bioinformatics, 2006.

[6] Sanguinetti. G, Milo, M, Rattray. M, and Lawrence. N.D, *Account- ing for probe-level noise in principal component analysis of microarray data*. Bioinformatics, 2005.

[7] Liu. X, Lin. K.K Andersen. B and Rattray. M *Including probe-level uncertainty in model-based gene expression clustering*. BMC Bioinformatics, 2007.

[8] Sun. J, Ata. Kaban, and Raychaudhury. S, *Robust Mixtures in the Presence of Measurement Errors*. International Conference on Machine Learning 24t , 2007.

[9] Richard. D.Pearson, *A comprehensive re-analysis of the Golden Spike data: Towards a benchmark for differential expression methods*. BMC Bioinformatice, 2008.

[10] Liu. X, Milo. M, Lawrence. N.D and Rattray. M, *A tractable probabilistic model for Affymetrix probel- level analysis across multiple chips*. Bioinformatics, 2005.

[11] Dempster. A.P, Laird. N.M and Rubin. D.B, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society , 1977.

[12] Jordan. M.I, Ghahramani. Z, Jaakkola. T.S. and Saul. L.K, *An introduction to variational methods for graphical models*. Machine Learning , 1999.

[13] Beal. M, *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London , 2003.

[14] Choe. SE, Boutros. M, Michelson. AM, Church. GM and Halfon. MS, *Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset*. Genome Biol, 2005.

[15] Foygel. K, Choi. B, Jun. S, Leong. DE, Lee. A, et al, *A novel and critical role for Oct4 as a regulator of the maternal-embryonic transition*. PloS ONE, 2008.

[16] Wu. Z, Irizarry. RA, Gentleman. R, Maritinze-Murillo. F and Spencer. F, *A model-based background adjustment for oligonucleotide expression arrays*. Journal of the American Statistical Association, 2004.

[17] Chen. Z, McGee. M, Liu. Q and Scheuermann. RH, *A distribution free summarization method for Affymetrix GeneChip arrays.*. Bioinformatics, 2007.

[18] Hochreiter. S, Clevert. DA and Obermayer. K, *A new summarization method for Affymetrix probe level data.*. Bioinformatics, 2006.

[19] Li. C and Wong. WH, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection*. Proc Natl Acad Sci USA, 2001.

[20] Johnson. N and W.H.W., *Combining scientific and statistical significance in gene ranking*. Unpublished, 2007.