BIOINFORMATICS APPLICATIONS NOTE

Vol. 21 no. 9 2005, pages 2118–2122
doi:10.1093/bioinformatics/bti318

Gene expression

Applications of beta-mixture models in bioinformatics

Yuan Ji[1,*], Chunlei Wu[2], Ping Liu[1], Jing Wang[1] and Kevin R. Coombes[1]

[1]Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA and [2]3M Pharmaceuticals, St Paul, Minnesota, USA

Received on January 5, 2005; revised on January 27, 2005; accepted on February 8, 2005
Advance Access publication February 15, 2005

ABSTRACT

Summary: We propose a beta-mixture model approach to solve a variety of problems related to correlations of gene-expression levels. For example, in meta-analyses of microarray gene-expression datasets, a threshold value of correlation coefficients for gene-expression levels is used to decide whether gene-expression levels are strongly correlated across studies. *Ad hoc* threshold values such as 0.5 are often used. In this paper, we use a beta-mixture model approach to divide the correlation coefficients into several populations so that the large correlation coefficients can be identified. Another important application of the proposed method is in finding co-expressed genes. Two examples are provided to illustrate both applications. Through our analysis, we also discover that the popular model selection criteria BIC and AIC are not suitable for the beta-mixture model. To determine the number of components in the mixture model, we suggest an alternative criterion, ICL–BIC, which is shown to perform better in selecting the correct mixture model.

Contact: yuanji@mdanderson.org

Supplementary information: http://odin.mdacc.tmc.edu/∼yuanj/highcorgeneanno.html

1 INTRODUCTION

We propose a beta-mixture model approach to solve a variety of problems in bioinformatics related to a large number of correlation coefficients. The correlation coefficients could be computed for the expression levels of the same gene measured under microarrays from different studies, which is often seen in meta-analyses of multiple gene-expression experiments (Ghosh *et al*., 2003; Kuo *et al*., 2002; Parmigiani *et al*., 2004; Pusztai *et al*., 2003). Alternatively, the correlation coefficients could come from a pathway analysis related to a critical gene, (Sabatti *et al*., 2002) in which the expression levels of the gene are examined to see whether they correlate with those of other genes in the same array data. Investigators are often asked to draw conclusions based on the magnitudes of these correlations. Specifically, a threshold value is used to decide whether the gene-expression levels are strongly correlated or not. For example, if a correlation coefficient is >0.5, then the gene-expression levels are positively correlated. However, the choice of 0.5 for the threshhold value is hard to justify.

Regardless of the origin, a large number of correlation coefficients are obtained, which usually reflect certain biological behaviors of the genes under investigation, e.g. if the genes are consistently expressed across the studies or if the genes are co-expressed. Naturally, not all genes behave the same way, and as a result, the values of their correlation coefficients differ. Finite mixture models (McLachlan and Peel, 2000) are typically used to analyze data of this type. Specifically, the correlation coefficients can be considered as coming from several underlying probability distributions. Each distribution is a component of the mixture model representing a gene population with similar behavior, and all the components are combined into a comprehensive model by a mixture form. To model the correlation coefficients, we use the beta distribution, which has been known for its flexible shapes and is therefore widely used to describe data from various experiments. In our analyses, we find that a two-component beta-mixture model is usually adequate to fit the correlation coefficients, with one component representing the population of uncorrelated gene expression levels and the other component representing the population of correlated ones.

One key issue in the finite mixture model approach is to decide the number of components in the model. This has been extensively studied for the normal-mixture model. The commonly used Akaike information criterion (AIC) Akaike (1973) and Bayesian information criterion (BIC) (Schwarz, 1978) have been shown to be adequate for deciding the number of components in the normal-mixture model (Leroux, 1992; Roeder and Wasserman, 1997). However, little is known about the performance of these criteria in the case of the beta-mixture model. Our finding through both simulations and real examples is that neither criterion is suitable for the beta-mixture model. Both AIC and BIC seem to overestimate the number of components, leading to mixture models with excessive numbers of components. On the other hand, we find that the integrated classification likelihood–BIC (ICL–BIC) (Biernacki *et al*., 1998) is adequate, always selects the right model in our simulation studies and yields reasonable results when applied to real data.

2 THE STATISTICAL MODEL

The beta-mixture model deals with a vector of correlation coefficients of gene-expression levels. Usually, the dimension of the vector is large, in the order of thousands. The correlation coefficients are assumed to come from multiple underlying probability distributions, in our case, beta distributions. To fit the beta distribution, for each correlation coefficient x_i, we apply a linear transformation $y_i = (x_i + 1)/2$, so that the range of the transformed values is between 0 and 1. The index i represents the gene with respect to which the correlation coefficient y is calculated.

Let $\{y_i\}, i = 1, \ldots, n$, denote the transformed correlation coefficients, where n is the total number of observations. Under a mixture

[*] To whom correspondence should be addressed.

of beta distributions,

$$y_i \sim \sum_{l=1}^{L} \pi_l f_l(y_i|\alpha_l, \beta_l), \quad l = 1, \ldots, L,$$

where

$$f_l(y|\alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}$$

denotes the density of a beta distribution and $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$ is the beta function. The quantity L is the number of components in the mixture.

We augment the data by introducing the latent indicator variable z_{il} for each gene i, where

$$z_{il} = \begin{cases} 1, & \text{if } y_i \text{ comes from population } l, \\ 0, & \text{otherwise.} \end{cases}$$

The set of values $\{z_{il}\}$ and $\{y_i\}$ is considered as the 'complete' data. We assume that each vector $z_i = (z_{i1}, \ldots, z_{iL})'$ is independent and identically distributed according to an L-category multinomial distribution with probabilities $\pi = (\pi_1, \ldots, \pi_L)'$. This is the prior distribution for z_i. Let $\theta = (\alpha_1, \beta_1, \ldots, \alpha_L, \beta_L)$ be the vector containing all the other unknown parameters α and β. The likelihood function for the complete data is given by

$$\text{Like}(\theta, \pi, z) = \prod_{i=1}^{n} \prod_{l=1}^{L} [\pi_l f_l(y_i|\alpha_l, \beta_l)]^{z_{il}}.$$

Consequently, the log-likelihood is given by

$$l(\theta, \pi, z) = \sum_{i=1}^{n} \sum_{l=1}^{L} z_{il} \left[\log \pi_l + \log f_l(y_i|\alpha_l, \beta_l)\right]. \qquad (1)$$

3 THE EXPECTATION–MAXIMIZATION ALGORITHM

We use the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977) to iteratively maximize the log-likelihood and update the conditional probability that y_i comes from the l-th component, which is defined as

$$z_{il}^* = E[z_{il}|y_i, \hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_L, \hat{\beta}_L; \hat{\pi}_1, \ldots, \hat{\pi}_L]. \qquad (2)$$

The set of parameter estimates $\{\hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_L, \hat{\beta}_L; \hat{\pi}_1, \ldots, \hat{\pi}_L\}$ is a maximizer of the log-likelihood (1), for given z^*s. We assign y_i to the component $\{l_0|z_{il_0}^* = \max_l z_{il}^*\}$. The EM algorithm iterates between an E-step where values z_{il}^* are computed from the data with the current parameter estimates, and an M-step in which the log-likelihood (1), with each z_{il} replaced by its current conditional expectation z_{il}^*, is maximized with respect to the parameters θ and π.

The detailed algorithm is given as follows:

(1) Initialize z_{il}^*: This can be done by, for example, assigning the smallest $100/L$ percentage of y_i to the first component, and the next smallest $100/L$ percentage of y_i to the second component, etc.

(2) M-step: Given z_{il}^*, maximize (1) with respect to the parameters θ and π. Specifically,

$$\hat{\pi}_l = \frac{\sum_{i=1}^{n} z_{il}^*}{n}.$$

We obtain a maximizer $\{\hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_L, \hat{\beta}_L\}$ of the log-likelihood in (1) numerically.

(3) E-step: Given the parameter estimates from the M-step, compute

$$\begin{aligned} z_{il}^* &= E[z_{il}|y_i, \hat{\pi}_1, \ldots, \hat{\pi}_L, \hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_L, \hat{\beta}_L] \\ &= P(z_{il} = 1|y_i, \hat{\pi}_1, \ldots, \hat{\pi}_L, \hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_L, \hat{\beta}_L) \\ &= \frac{\hat{\pi}_l f_l(y_i|\hat{\alpha}_l, \hat{\beta}_l)}{\sum_{j=1}^{L} \hat{\pi}_j f_j(y_i|\hat{\alpha}_j, \hat{\beta}_j)}. \end{aligned}$$

(4) Repeat M-step and E-step until the change in the value of the log-likelihood in Equation (1) is negligible.

Maximization with respect to α_l and β_l can only be carried out numerically. We use the 'nlm' function in R, which is freely available at http://www.r-project.org/. The EM algorithm yields the final estimated posterior probability z_{il}^*, the value of which represents the posterior probability that correlation coefficient y_i comes from component l. We assign y_i to component l_0 as previously defined, which follows a beta distribution with parameter estimates $\hat{\alpha}_{l_0}$ and $\hat{\beta}_{l_0}$, also yielded by the EM algorithm. The characteristics of the beta distribution contain information that can be used for inference about the behavior of the genes belonging to the corresponding component. For example, if the beta distribution is closely centered around a positive value, say 0.7, then the correlation coefficients in this component represent a population of strongly correlated genes.

The determination of the number of components L is challenging and is often decided by some model selection criterion, e.g. AIC, BIC or ICL–BIC. While AIC and BIC are well known, the definition of ICL–BIC is given by

$$\text{ICL–BIC} = \text{BIC} + 2EN(z^*),$$

where BIC $= -2 \log$-likelihood $+ K \log n$, in which $K = 3 * L - 1$ is the number of unknown parameters in the model. Quantity $EN(z^*) = -\sum_{i=1}^{n} \sum_{l=1}^{L} z_{il}^* \log z_{il}^*$ is the estimated entropy of the fuzzy classification matrix $\mathbf{C} = ((z_{il}))$. Typically, mixture models are fitted through the EM algorithm with different values of L, and the model with the smallest AIC, BIC or ICL–BIC is chosen. We evaluate the performance of these three criteria through a simulation study in the next section.

4 SIMULATION

To be more realistic, instead of directly simulating correlation coefficients, we simulated sets of gene-expression levels with different correlation coefficients across datasets from different experiments. Specifically, we considered two experiments in which the same set of 10 000 genes were measured. We simulated 10 expression levels for each gene in each experiment. Let R_{ik} and S_{ik} denote the k-th expression level of gene i in experiments 1 and 2, respectively, $i = 1, \ldots, 10\,000$ and $k = 1, \ldots, 10$. Assume that $(R_{ik}, S_{ik})'$ are

Table 1. Values of AIC, BIC and ICL–BIC using the proposed model with different values of L

Number of	First simulation, $L = 2$			Second simulation, $L = 3$		
components L	AIC	BIC	ICL–BIC	AIC	BIC	ICL–BIC
1	30689.55	30703.97	30703.07	38382.61	38397.03	38397.03
2	22224.89	22260.94	**25752.13**	27887.47	27923.52	32053.63
3	21599.46	21657.14	34154.75	24568.14	24625.82	**31679.98**
4	19857.84	19937.25	37157.57	**21299.82**	**21379.13**	33049.97
5	**15076.03**	**15176.97**	32043.12	23798.21	23899.15	33073.03

The bold number is the smallest value in its column. Criteria AIC and BIC choose $L = 5$ and $L = 4$ as the best models for the first and second simulations, respectively; and ICL–BIC chooses $L = 2$ and $L = 3$ as the best models, both which are correct.

independent for different values of i and k and

$$\begin{pmatrix} R_{ik} \\ S_{ik} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} 1.5\mu_i^2 & \rho_i \times 1.5\mu_i^2 \\ \rho_i \times 1.5\mu_i^2 & 1.5\mu_i^2 \end{pmatrix}\right), \qquad (3)$$

where μ_i is the mean expression level of gene i and ρ_i is the correlation of the expression levels across two experiments. We let the variance expression level be proportional to the square mean expression level, which is often observed in practice. In the first simulation, we let $\rho_i = 0.7$ for the first 3000 genes and $\rho_i = 0.0$ for the last 7000 genes. We sampled each μ_i independently from a normal distribution with mean 0 and variance 4. In the second simulation, the μ_i were sampled in the same way. However, we let $\rho = 0.7$ for the first 3000 genes, $\rho = 0.0$ for the next 4000 genes and $\rho = -0.7$ for the last 3000 genes. We sampled $(R_{ik}, S_{ik})'$ from the distribution in Equation (3) and computed the sample correlation of the vectors $(R_{i1}, \ldots, R_{i10})$ and $(S_{i1}, \ldots, S_{i10})$ for each gene i. We thus obtained 10 000 sample correlation coefficients, which came from two underlying distributions for the first simulation and three for the second simulation. We computed the values of AIC, BIC and ICL–BIC for each simulation after implementing the proposed beta-mixture model with different values of L. The results are presented in Table 1. Both AIC and BIC prefer models with an excessive number of components. The criterion ICL–BIC, however, correctly chooses the right model for both simulations.

5 TWO EXAMPLES

The proposed method is illustrated with two real bioinformatics studies. The first study requires combining gene expression from different types of microarrays. The second deals with the identification of co-expressed genes.

5.1 Study I: combining gene expression levels

This study was a meta-analysis involving a total of 33 patients with diagnosed stages I–III breast cancer. Two microarray technologies were used in this study: Affymetrix Human Genome U133 chip sets (HG-U133A and HG-U133B), and radioactively labeled cDNA nylon membrane microarrays printed by Millennium Pharmaceuticals Inc. (Cambridge, MA). There were 9285 pairs of genes that were common to both array platforms. One of the study objectives was to determine which pairs were strongly correlated in their expression levels. For each gene on each platform, we had 33 measurements corresponding to the 33 patients. Using these 33 measurements on each platform, we computed correlation coefficients for each gene, which resulted in 9285 correlation coefficients. We applied the beta-mixture model to these correlation coefficients, with the number of components ranging from 1 to 5. ICL–BIC chose $L = 2$, while AIC and BIC both preferred the largest model with $L = 5$ (Table 2). Figure 1 displays the fitted densities of the beta-mixture distributions with L ranging from 2 to 5. It seems that when $L = 2$, the fitting is already adequate; the improvement in the fit for the case when $L > 2$ is negligible.

Table 2. Values of AIC, BIC and ICL–BIC using the proposed model with different values of L

Number of	Model selection criteria		
components L	AIC	BIC	ICL–BIC
1	23333.90	23348.17	23348.17
2	16757.83	16793.51	**22684.72**
3	13651.47	13708.56	24613.63
4	10645.44	10723.94	23882.83
5	**8758.72**	**8558.63**	23422.24

The bold number is the smallest value in its column. Criteria AIC and BIC choose $L = 5$ as the best model; and ICL–BIC chooses $L = 2$ as the best model.

We proceeded by fitting a two-component beta-mixture model to the correlation coefficients. The means of the fitted beta distributions respectively equaled 0.54 and 0.76, which corresponded to the values 0.09 and 0.53 on the correlation scale. The standard deviations of both beta distributions equaled 0.1. Therefore, one component (with mean correlation 0.09) corresponded to the genes with weak or no correlations and the other (with mean correlation 0.53) to the genes with strong correlations. The cut-off value based on the posterior probability was 0.31, i.e. if a correlation coefficient was >0.31, it would be considered as coming from the component with strong correlations.

After the two populations of genes were discovered, efforts were taken to explain why some pairs were poorly correlated.

Those corresponding genes with strong correlations were used for further analysis.

5.2 Study II: gene co-expression

The second example used the dataset from the study conducted by Beer *et al*. (2002). In this study, the authors identified a list of

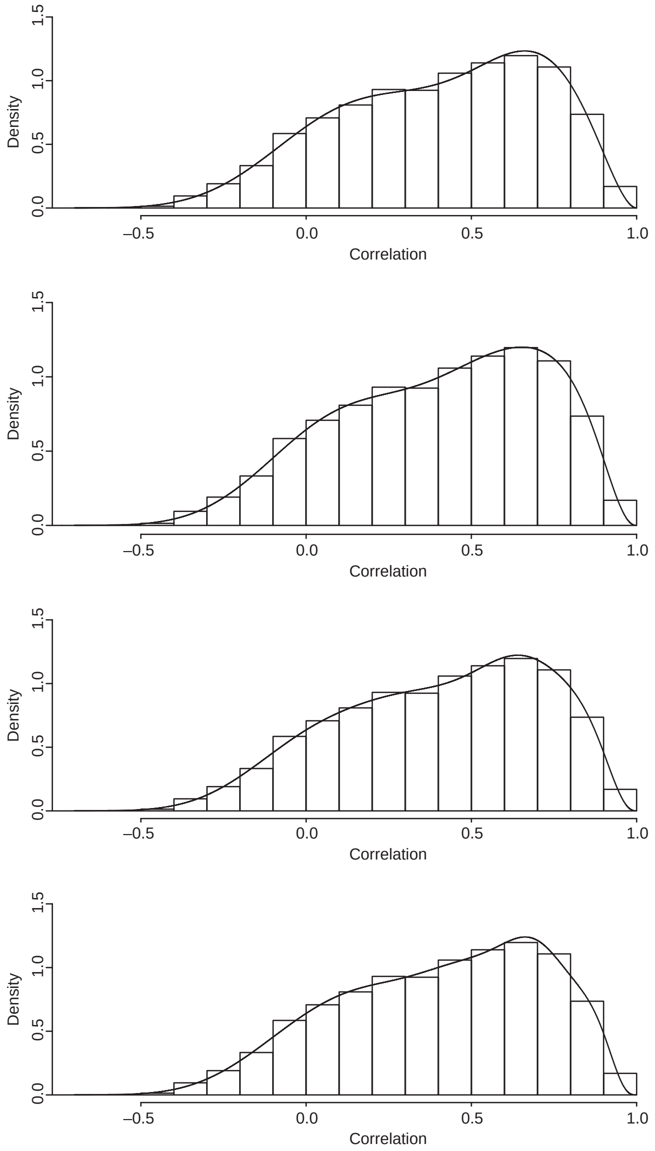

Fig. 1. From top to bottom are fitted densities of the beta-mixture model with two, three, four and five components.

important genes that are related to the survival of patients with adenocarcinoma. We arbitrarily selected one gene, tyrosine hydroxylase (TH), from the original list and computed the correlation coefficients of its expression levels to the remaining genes in the data. We obtained 7128 correlation coefficients. We repeated the same analysis as in the last section, and found that again ICL–BIC seemed to select the right mixture model with two components while AIC and BIC seemed to overestimate the number of components. We fitted a two-component beta-mixture model suggested by ICL–BIC, the density of which is given in Figure 2. The means of the two beta distributions equaled 0.46 and 0.71, corresponding to the values -0.07 and 0.43, respectively, on the correlation scale. The standard deviations of both distributions equaled 0.1. Therefore, we have two gene populations, one with weak correlations and the other strong. The ones in the strongly correlated population are potentially co-expressed genes, a possibility that needs to be validated by further analysis.

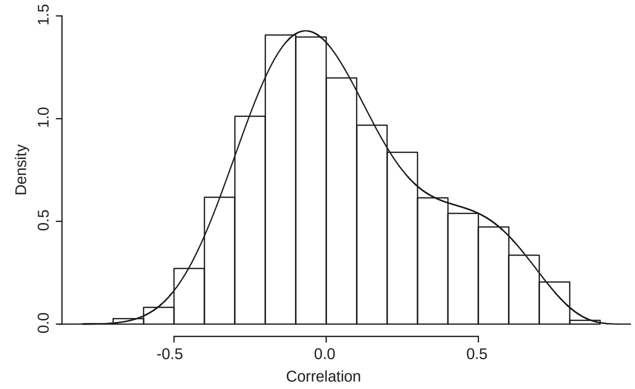

Fig. 2. The histogram contains the original correlation coefficients for all the gene-expression levels. The solid line is the density curve of the fitted beta-mixture distribution.

Among the top 20 genes highly correlated with the target gene *TH*, which is related to tyrosine metabolism, we found at least 9 genes that are functionally related to TH. Many of the highly correlated genes are related to muscle development and contraction, including genes *KCNQ1*, *MMP16*, *TNNI1*, *SMPD1*, *GPR68*, *HTR2C*, *CUL5*, *EMD* and *OXT*. Tyrosine hydroxylase is essential in tyrosine metabolism, and tyrosine kinase, which requires tyrosine for its activity, is related to the function of muscle tissue. Du *et al*. (1994) reported that the expression level of TH is related to the function of muscle. Patients with adenocarcinoma might have a higher level of TH activity than healthy people because the smooth muscle of their lungs needs to do more work to compensate for the lung tissue that is compromised by disease. An annotation table of the top 20 highly correlated genes is provided in the Supplemental information section.

6 DISCUSSION

We make two contributions in this paper. First, we provide a statistical tool, the beta-mixture model, for analyzing a large number of correlation coefficients, which is often required in bioinformatics. To our knowledge, there is no objective method for performing the analysis so far. Second, we show that the model selection criteria AIC and BIC do not work for the beta-mixture model, although they have been proved to be adequate for normal-mixture models. Specifically, both criteria seem to overestimate the number of components in the beta-mixture model, a sign of lack of penalties for the model size. We further find that the criterion ICL–BIC seems to work well and selects the right model in our simulations. ICL–BIC adds more penalties for a larger model and, therefore, performs better.

As theoretical justification is needed for further work, we hope that the findings provided in this paper will generate more research in this area.

ACKNOWLEDGEMENTS

This research was partly supported by the National Institutes of Health through the University of Texas M.D. Anderson Cancer Center SPORE in Lung Cancer grant CA070907 and the Prostate Cancer grant CA90270.

REFERENCES

Akaike,H. (1973) Information theory and the extension of the maximum likelihood principle. In Petrov,V. and Csaki,F. (eds), *Proceedings of the Second International Symposium on Information Theory*, Akailseoniai-kiudo, Budapest, pp. 267–281.

Beer,D. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Biernacki,C. *et al*. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report No. 3521. INRIA, Rhone-Alpes.

Dempster,A. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.

Ghosh,D. *et al*. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics*, **4**, 180–188.

Kuo,W. *et al*. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.

Leroux,B (1992). Consistent estimation of a mixing distribution. *Ann. Stat.*, **20**, 1350–1360.

McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. John Wiley and Sons, NY.

Parmigiani,G. *et al*. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.*, **10**, 2922–2927.

Pusztai,L. *et al*. (2003) Gene expression profiles obtained from single passage fine needle aspirations (FNA) of breast cancer reliably identify prognostic/predictive markers such as estrogen (er) and HER-2 receptor status and reveal large scale molecular differences between ER-negative and ER-positive tumors. *Clin. Cancer Res.*, **9**, 2406–2415.

Roeder,K. and Wasserman,L. (1997) Practical density estimation using mixtures of normals. *J. Am. Stat. Assoc.*, **92**, 894–902.

Sabatti,C. *et al*. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **13**, 2886–2893.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann Stat.*, **6**, 461–464.