
A Bayesian framework for knowledge driven regression model in micro-array data analysis

Rong Jin*

Department of Computer Science and Engineering,
Michigan State University, MI, USA

E-mail: rongjin@cse.msu.edu

*Corresponding author

Luo Si

Department of Computer Science,
Purdue University,

West Lafayette, IN, USA

E-mail: lsi@cs.purdue.edu

Christina Chan

Department of Chemical Engineering and Material Science,

Michigan State University, MI 48864, USA

E-mail: krischan@egr.msu.edu

Abstract: This paper addresses the sparse data problem in the linear regression model, namely the number of variables is significantly larger than the number of the data points for regression. We assume that in addition to the measured data points, the prior knowledge about the input variables may be provided in the form of pair wise similarity. We presented a full Bayesian framework to effectively exploit the similarity information of the input variables for linear regression. Empirical studies with gene expression data show that the regression errors can be reduced significantly by incorporating the similarity information derived from gene ontology.

Keywords: Bayesian analysis; knowledge driven data regression; data regression; graph Laplacian; gene expression analysis; data mining; bioinformatics.

Reference to this paper should be made as follows: Jin, R., Si, L. and Chan, C. (2008) 'A Bayesian framework for knowledge driven regression model in micro-array data analysis', *Int. J. Data Mining and Bioinformatics*, Vol. 2, No. 3, pp.250–267.

Biographical notes: Rong Jin received an PhD from School of computer Science, Carnegie Mellon University, in 2006. He began his academic career as an Assistant Professor at Michigan State University and is now an Associated Professor in the Department of Computer Science and Engineering. His research area is statistical machine learning and its application to large-scale information management. He has published over 80 technical articles, and received the NSF Career Award in 2006.

Luo Si received his PhD Degree from School of Computer Science, Carnegie Mellon University in 2006. Currently, he is an Assistant Professor of Computer Science and Statistics (by courtesy) at Purdue University.

He conducted research in information retrieval, applied machine learning and text mining. He has served on the program committees of many international conferences in computer science and related disciplines.

Christian Chan has broad experience with metabolic studies, mathematical modelling, systems biology, and bioinformatics and in applying these techniques to study cellular metabolism and function. She uses a systems approach to understand how oxidative stress and inflammation due to free fatty acids and cytokines induce lipotoxicity and cell death. She began her academic career at Michigan State University in 2002 as an Associate Professor and is now Professor in the Chemical Engineering and Materials Science Department. She holds joint appointments in the departments of Computer Science and Engineering as well as in Biochemistry and Molecular Biology.

1 Introduction

Data regression is commonly used in bioinformatics. Often, the problem is to predict the output value of a biological process for a particular biological system under certain condition. By assuming that the outputs of a biological process for a given biological system can be determined by the linear combination of the expression levels of the genes within the system, we cast the prediction problem into a Linear Regression (LR) problem. The key to any linear regression model is how to determine the regression weights assigned to each gene. In addition to helping to predict the outputs of a biological process under different conditions, the regression weights of genes are also useful in determining the importance of the genes with respect to the biological process.

A main challenge with linear regression in bioinformatics arises from the sparsity of data. Very often, the number of training data points that are available for determining the regression weights of genes is significantly smaller than the number of genes involved in the process. A number of studies have been devoted to the sparse data problem. Many of the approaches are based on the techniques of feature selection and dimension reduction. The main idea of these approaches is to reduce the number of regression weights by only selecting the subset of the genes whose gene expression levels are significantly correlated with the outputs of the biological process. The well known feature selection methods include the information gain method (Lewis and Ringuette, 1994), and the χ^2 tests Wiener et al. (1995). The well known dimension reduction methods include the Principle Component Analysis (PCA) (Bishop, 1995), the Independent Component Analysis (ICA) (Common, 1994), the Linear Discriminative Analysis (LDA) (Friedman, 1989), and the Partial Least Square (PLS) (Hskuldsson, 1988). Other approaches address the sparse data problem by regularisation. They usually introduce a penalty term into the regression problem, which will favour the solutions of sparse non-zero regression weights. The optimal weights are then obtained by minimising both the penalty term and the regression errors. Well known approaches in this category include the ridge regression (Hoerl and Kennard, 1970) and the Lasso regression (Tibshirani, 1996).

Despite intensive studies of LR in the past, the challenge of how to determine the regression weights for massive number of genes using a small number of training data

points remains an open problem. In this paper, we present a text mining approach to alleviate the sparse data problem. The key idea is to first represent a gene by a set of key words that describe its functions and properties. These extracted keywords will allow us to determine the similarity of the genes in their functions and properties. More specifically, we assume that two genes are likely to be assigned similar regression weights if their text profiles overlap significantly. Based on this assumption, we presented a full Bayesian framework for incorporating the text profiles of the genes in guiding the regression process.

Unlike many studies in bioinformatics that focused on exploiting the gene ontology information, in this study, our emphasis is on extracting the text profiles from the texts of research papers. This is important since a large portion of the genes can not be found in the existing gene ontology database. For example, in the biological system used for this study, about 15–20% genes are not present in the current gene ontology database, thus precluding the use of these genes in the regression analysis. Hence, it is important to develop text mining techniques that can produce text profiles for these genes. The key challenge in using the free text based gene profiles to guide the regression processes is that many keywords may be completely irrelevant to the biological process to be regressed. As a result, the overlap in the text profiles of genes may not accurately reflect the relationship among the genes in the biological process. In order to address this problem, we extend the proposed Bayesian framework to automatically decide the importance of words when computing the similarity of genes based on their text profiles, which greatly improved the regression analysis. Bayesian approaches (Gelman et al., 2004; Townsend and Hartl, 2002) for regression or feature selection has been studied for a long time. However, it is still an open problem for how to incorporate biological knowledge into Bayesian framework for biological applications. To achieve this goal, this paper presents a new Bayesian approach to incorporate biological ontology and text literature information into micro-array data analysis. For example, it extends the traditional Bayesian approach by automatically tuning the importance of words for computing gene similarity.

The rest of this paper is organised as follows: Section 2 presents the basic regression problem addressed in this paper. Section 3 presents the Bayesian framework for incorporating the text profiles into the regression model, and the extension that address the problem of noisy text profiles. Section 4 presents the empirical studies using microarray and metabolic data obtained for the hepatocellular system. Section 5 concludes this study. Section 6 overviews the related work.

2 Preliminaries

Consider a biological system that comprises of n genes. Let $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$ denote the expression levels of genes in the biological system under the i th condition. Let $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d)$ denote the gene expression levels of the biology system under d different conditions. By assuming that the output of a biological process under the i th condition, denoted by y_i , is a linear combination of the gene expression data under the same condition, we have

$$y_i = \sum_{j=1}^n w_j x_{i,j} = \mathbf{w}^\top \mathbf{x}_i$$

where $\mathbf{w} = (w_1, w_2, \dots, w_n)$ are the regression weights assigned to n genes. Hence, the goal of the LR problem is to find weights \mathbf{w} that minimises the regression error under all d conditions, i.e.,

$$\min_{\mathbf{w}} l_e = \sum_{i=1}^d (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \|\mathbf{y} - \mathbf{w}^\top X\|_2^2 \quad (1)$$

where vector $\mathbf{y} = (y_1, y_2, \dots, y_d)$ includes the output values of the biological process under all d conditions. The optimal solution to the above problem is

$$\mathbf{w} = (XX^\top)^\dagger X\mathbf{y} \quad (2)$$

where \dagger refers to the pseudo inverse operator. To address the sparse data problem, the ridge regression (Hoerl and Kennard, 1970) introduces the penalty term $\|\mathbf{w}\|_2^2$ into the objective function of equation (1), which leads to the following optimisation problem:

$$\min_{\mathbf{w}} l_r = \tau_e \|\mathbf{w}\|_2^2 + \sum_{i=1}^d (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \quad (3)$$

where parameter τ_e weights the importance of the penalty term against the regression error. The solution to the ridged linear regression model is

$$\mathbf{w} = (XX^\top + \tau_e I_n)^{-1} X\mathbf{y}. \quad (4)$$

3 Exploiting textual information for Linear Regression model

In this section, we will first introduce a Bayesian framework that incorporates the gene similarity information into the LR model. We will then describe the algorithm that solves the corresponding framework. At the end of this section, we will discuss the application of the proposed framework in exploiting the keyword profiles of genes for the regression problem in bioinformatics, with emphasis on how to address the problem of irrelevant keywords in text profiles.

3.1 A Bayesian framework

In this framework, we assume the prior knowledge of genes is encoded in a similarity matrix S , where each element $S_{i,j}$ expresses the similarity of two genes in terms of their properties and functions. The goal of this framework is to incorporate the gene similarity information to guide the selection of regression weights. More specifically, two genes with high similarity are likely to be assigned similar weights. In order to ensure that the assigned weights are consistent with the similarity matrix, we consider the following quantity:

$$l_g = \sum_{i,j=1}^n S_{i,j} (w_i - w_j)^2 = \mathbf{w}^\top L\mathbf{w} \quad (5)$$

where matrix L is the graph Laplacian (Chung, 1997) of similarity matrix S . It is defined as $L = D - S$ where D is a diagonal matrix and each diagonal element is

calculated as $D_{i,i} = \sum_{j=1}^n S_{i,j}$. Clearly, if the assigned weights are consistent with the similarity information, we would expect that the regression weights \mathbf{w} will minimise the quantity l_g . Hence, the above term can be used to construct a prior $\Pr(\mathbf{w})$ for the regression weights \mathbf{w} as follows:

$$\begin{aligned} \Pr(\mathbf{w}; \tau_l, \tau_e) &\sim \mathcal{N}(\mathbf{w}; \mathbf{0}_n, (\tau_l L + \tau_e I_n)^{-1}) \\ &= \sqrt{\frac{|\tau_l L + \tau_e I_n|}{(2\pi)^n}} \exp\left(-\frac{\mathbf{w}^\top (\tau_l L + \tau_e I_n) \mathbf{w}}{2}\right) \end{aligned}$$

where I_d is the identical matrix of size $n \times n$. Parameter τ_l is introduced to weight the importance of matrix L against I_n in the Gaussian distribution. Notice that we use the matrix $(\tau_l L + \tau_e I_n)^{-1}$ for covariance matrix, instead of $(\tau_l L)^{-1}$. This is because graph Laplacian L is in fact a singular matrix and therefore its inverse is not well defined. Another reason for introducing $\tau_e I_n$ into the covariance matrix is because similar to the ridge regression model, matrix $\tau_e I_n$ serves as the penalty term for the regression weights \mathbf{w} and will usually lead to sparse solutions.

Since the prior involves two parameters, τ_l and τ_e , we further introduce two prior distributions for parameters τ_l and τ_e , i.e.,

$$\begin{aligned} \tau_l &\sim \mathcal{G}(\tau_l; \alpha_l, \beta_l) \\ \tau_e &\sim \mathcal{G}(\tau_e; \alpha_e, \beta_e) \end{aligned}$$

where $\mathcal{G}(x; \alpha, \beta)$ is the Gamma distribution and defined as follows

$$\mathcal{G}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

where parameters α and β control the shape of the Gamma distribution. The complete prior for \mathbf{w} is then written as:

$$\Pr(\mathbf{w}) = \int d\tau_l d\tau_e \mathcal{G}(\tau_l; \alpha_l, \beta_l) \mathcal{G}(\tau_e; \alpha_e, \beta_e) \times \mathcal{N}(\mathbf{w}; \mathbf{0}_n, (\tau_l L + \tau_e I_n)^{-1}) \quad (6)$$

We then describe the likelihood $\Pr(\mathbf{y} | \mathbf{w}, X)$. By assuming that the output values of a biological process under different conditions are independent given the gene expression data X and the regression weights w , we have

$$\Pr(\mathbf{y} | \mathbf{w}, X) = \prod_{i=1}^d \Pr(y_i | \mathbf{w}, \mathbf{x}_i).$$

Then, according to the LR model stated in equation (1), we assume a Gaussian distribution for $\Pr(y_i | \mathbf{w}, \mathbf{x}_i)$, which is defined as follows:

$$\begin{aligned} \Pr(y_i | \mathbf{w}, \mathbf{x}_i) &\sim \mathcal{N}(y_i; \mathbf{w}^\top \mathbf{x}_i, \tau_r^{-1}) \\ &= \sqrt{\frac{\tau_r}{2\pi}} \exp\left(-\frac{\tau_r (y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2}\right) \end{aligned}$$

where parameter τ_r is introduced to control the variance in regression. Since the above probability involves an unknown parameter τ_r , the complete expression for $\Pr(\mathbf{y} | \mathbf{w}, X)$ that includes the uncertainty of parameter τ_r is written as follows:

$$\Pr(\mathbf{y} | \mathbf{w}, X) = \int d\tau_r \mathcal{G}(\tau_r; \alpha_r, \beta_r) \prod_{i=1}^d \Pr(y_i | \mathbf{w}, \mathbf{x}_i, \tau_r) \tag{7}$$

where a Gaussian prior is introduced for parameter τ_r .

By combining the likelihood in equation (7) with the prior in equation (6), we have the conditional probability $\Pr(\mathbf{y} | X)$ expressed as

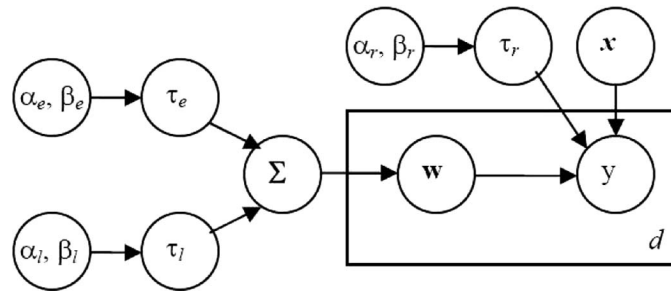
$$\begin{aligned} \Pr(\mathbf{y} | X) &= \int d\mathbf{w} \Pr(\mathbf{y} | \mathbf{w}, X) \Pr(\mathbf{w}) \\ &= \int d\tau_l d\tau_e d\tau_r d\mathbf{w} \mathcal{G}(\tau_l; \alpha_l, \beta_l) \mathcal{G}(\tau_e; \alpha_e, \beta_e) \mathcal{G}(\tau_r; \alpha_r, \beta_r) \\ &\quad \times \mathcal{N}(\mathbf{w}; \mathbf{0}_n, (\tau_l L + \tau_e I_n)^{-1}) \prod_{i=1}^d \mathcal{N}(y_i; \mathbf{w}^\top \mathbf{x}_i, \tau_r^{-1}). \end{aligned} \tag{8}$$

In summary, the stochastic process of generating the output values y_i for a biological process from the gene expression data \mathbf{x}_i under the i th condition is described by the following steps:

- sample τ_e and τ_l from priors $\mathcal{G}(\tau_e; \alpha_e, \beta_e)$ and $\mathcal{G}(\tau_l; \alpha_l, \beta_l)$, respectively
- construct the covariance matrix as $\Sigma = (\tau_e I_n + \tau_l L)^{-1}$
- sample the regression weights \mathbf{w} from the distribution $\mathcal{N}(\mathbf{w}; \mathbf{0}_n, \Sigma)$
- sample precision τ_r from the distribution $\mathcal{G}(\tau_r; \alpha_r, \beta_r)$
- sample each y_i in \mathbf{y} from the distribution $\mathcal{N}(y_i; \mathbf{w}^\top \mathbf{x}_i, \tau_r^{-1})$.

Figure 1 shows a graphical representation for the above stochastic process.

Figure 1 The graphical representation for the Bayesian framework



3.2 Variational algorithm

Directly inferring the posterior distribution $\Pr(\mathbf{w} | X, \mathbf{y})$ from equation (8) is computationally expensive. On the other side, approximate inference algorithms like

the variational approach (Jordan et al., 1999) has been used for inferring posterior distributions. In this subsection, we will present an efficient algorithm that is based on the variational approach.

First, we write the conditional probability $\Pr(\mathbf{y} | X)$ in the logarithm form, i.e.,

$$\begin{aligned} \log \Pr(\mathbf{y} | X) = & \log \left(\int d\tau_l d\tau_e d\mathbf{w} \left[\mathcal{G}(\tau_l; \alpha_l, \beta_l) \mathcal{G}(\tau_e; \alpha_e, \beta_e) \times \right. \right. \\ & \left. \left. \mathcal{N}(\mathbf{w}; \mathbf{0}_n, (\tau_l L + \tau_e I_n)^{-1}) \right] \right. \\ & \left. \times \int d\tau_r \mathcal{G}(\tau_r; \alpha_r, \beta_r) \prod_{i=1}^d \mathcal{N}(y_i; \mathbf{w}^\top \mathbf{x}_i + b, \tau_r^{-1}) \right). \end{aligned}$$

To infer the posterior probability $\Pr(\mathbf{w} | \mathbf{y}, X)$, we introduce the variational distributions $\phi_l(\tau_l)$, $\phi_e(\tau_e)$, $\phi_r(\tau_r)$ and $\phi_w(\mathbf{w})$ for variables τ_l , τ_e , τ_r and \mathbf{w} , respectively. They are:

$$\begin{aligned} \phi_l(\tau_l) & \sim \mathcal{G}(\tau_l; a_l, b_l) \\ \phi_e(\tau_e) & \sim \mathcal{G}(\tau_e; a_e, b_e) \\ \phi_r(\tau_r) & \sim \mathcal{G}(\tau_r; a_r, b_r) \\ \phi_w(\mathbf{w}) & \sim \mathcal{N}(\mathbf{w}; \bar{\mathbf{w}}, S_w). \end{aligned}$$

The reasons for introducing variational distributions are twofold:

- these variational distributions will allow us to lower bound $\log \Pr(\mathbf{y} | X)$, which will decouple the correlation among the different parameters and therefore allow efficient computation
- these variational distribution will also serve as approximations of the posterior distribution for parameters τ_e , τ_l , τ_r , and \mathbf{w} .

Using the variational distributions, $\log \Pr(\mathbf{y} | X)$ is lower bounded by the following expression:

$$\begin{aligned} \log \Pr(\mathbf{y} | X) \geq & \langle \log \Pr(\mathbf{w}; \tau_l, \tau_e) \rangle \\ & + H(\phi_w) \langle \log \Pr(\tau_l) \rangle + \langle \log \Pr(\tau_e) \rangle + H(\phi_l) + H(\phi_e) \\ & + \langle \log \Pr(\tau_r) \rangle + H(\phi_r) + \sum_{i=1}^d \langle \log \Pr(y_i | \mathbf{w}, \mathbf{x}_i) \rangle \end{aligned} \quad (9)$$

where the operator $\langle \cdot \rangle$ stands for expectation and $H(\cdot)$ is an entropy function. We then lower bound the expectation of $\log \Pr(\mathbf{w}; \tau_l, \tau_e)$ as follows:

$$\begin{aligned} & \langle \log \Pr(\mathbf{w}; \tau_l, \tau_e) \rangle \\ & \geq \frac{1}{2} \sum_{k=1}^n (\gamma_k \langle \log(\tau_l \lambda_k) \rangle + (1 - \gamma_k) \langle \log \tau_e \rangle + H(\gamma_k)) \\ & \quad - \frac{1}{2} \text{tr}(\langle (\tau_l) L + \langle \tau_e \rangle I_n \rangle \langle \mathbf{w} \mathbf{w}^\top \rangle) - \frac{d}{2} \log \pi \end{aligned} \quad (10)$$

where $\lambda_i, i = 1, \dots, n$ are the eigenvectors of the graph Laplacian L . In the above, we insert the variational distributions γ_k s in order to decompose the correlation between τ_e and τ_l . γ_k s are chosen to maximise the expression in equation (10). They are:

$$\gamma_k = \frac{\lambda_k b_e \exp(\psi(a_l))}{\lambda_k b_e \exp(\psi(a_l)) + b_l \exp(\psi(a_e))} \tag{11}$$

where $\psi(x)$ is the digamma function defined as $\psi(x) = d \log \Gamma(x)/dx$. More detailed derivation of γ_k can be found in the appendix.

Putting the lower bounds in equations (9) and (10) together, we have the lower bounds for $\log \Pr(\mathbf{y} | X)$ finally written as:

$$\begin{aligned} F = & \langle \log \Pr(\tau_e) \rangle + H(\phi_l) + H(\phi_e) \\ & + \langle \log \Pr(\tau_r) \rangle + H(\phi_r) + \sum_{i=1}^d \langle \log \Pr(y_i | \mathbf{w}, \mathbf{x}_i) \rangle \\ & - \frac{1}{2} \text{tr}(\langle \tau_l \rangle L + \langle \tau_e \rangle I_n \langle \mathbf{w} \mathbf{w}^\top \rangle) \\ & + H(\phi_w) \frac{1}{2} \sum_{k=1}^n (\gamma_k \langle \log(\tau_l \lambda_k) \rangle + (1 - \gamma_k) \langle \log \tau_e \rangle + H(\gamma_k)). \end{aligned}$$

By optimising the variational distributions with respect to the lower bound of $\log \Pr(\mathbf{y} | X)$, we have the following equation for updating the parameters in the variational distributions $\phi_l(\tau_l)$, $\phi_e(\tau_e)$, $\phi_r(\tau_r)$, and $\phi_w(\tau_w)$:

$$a_l = \alpha_l + \frac{1}{2} \sum_{k=1}^n \gamma_k \tag{12}$$

$$b_l = \beta_l + \frac{1}{2} \text{tr}(L(\bar{\mathbf{w}} \bar{\mathbf{w}}^\top + S_w)) \tag{13}$$

$$a_e = \alpha_e + \frac{1}{2} \sum_{k=1}^n (1 - \gamma_k) \tag{14}$$

$$b_e = \beta_e + \frac{1}{2} \text{tr}(\bar{\mathbf{w}} \bar{\mathbf{w}}^\top + S_w) \tag{15}$$

$$a_r = \alpha_r + \frac{n}{2} \tag{16}$$

$$b_r = \beta_r + \frac{1}{2} (\|\mathbf{y} - X^\top \bar{\mathbf{w}}\|_2^2 + S_w) \tag{17}$$

$$S_w = \left(\frac{a_l}{b_l} L + \frac{a_e}{b_e} I_n + \frac{a_r}{b_r} X X^\top \right)^{-1} \tag{18}$$

$$\bar{\mathbf{w}} = S_w X \mathbf{y}. \tag{19}$$

Figure 2 shows the detailed steps for computing the mean and variance for the regression weights \mathbf{w} .

Figure 2 Procedures for computing means and variance of regression weights \mathbf{w}

<p>Preprocessing:</p> <ul style="list-style-type: none"> • Construct the graph Laplacian L from the similarity matrix S. • Compute the eigenvalues of L, λ_i, $i = 1, 2, \dots, n$. <p>Initialization:</p> <ul style="list-style-type: none"> • Randomly initialize the mean of regression weights $\bar{\mathbf{w}}$. • Set covariance matrix $S_w = I_n$. • Set $\gamma_k = 0.5$, $k = 1, 2, \dots, n$ <p>Repeat</p> <ul style="list-style-type: none"> • parameters γ_k is calculated by using Eqn. (11) • Computing parameters $a_l, b_l, a_e, b_e, a_r, b_r, S_w$, and $\bar{\mathbf{w}}$ using Eqn. (12) to (19) • Until the difference in $\bar{\mathbf{w}}$ between two consecutive iterations is smaller than a threshold ϵ
--

3.3 Constructing gene similarity matrix

The key to the proposed model is how to determine the similarity matrix of genes. To this end, we will first extract a text profile for each gene, and then estimate the similarity of two genes by comparing their text profile. Our hypothesis is that two genes are likely to have similar function in the biological process if their text profiles overlap heavily. In this section, we consider two different ways of extracting text profiles for the genes: representing a gene by its gene ontology codes, and representing a gene by the common keywords that are found in the research papers that are relevant to the gene.

3.4 Gene profiles by GO codes

For each gene g_i , we extract the set of gene ontology codes $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \dots, t_{i,n_i})$ that describe the properties and functions of gene g_i . Here n_i is the number of codes for gene g_i , and $t_{i,j}$ is a gene ontology code. To compute the similarity of two gene profiles, we will first calculate the similarity of two gene ontology codes. To this end, we assume that two gene ontology codes are similar if their positions in the gene ontology are relatively close to each other. Based on this assumption, we estimate the similarity of two gene ontology codes based on the overlap between their paths from the root. More specifically, the similarity between two gene ontology codes t_i and t_j , denoted by $f_t(t_i, t_j)$, is calculated as follows:

$$f_t(t_i, t_j) = \frac{2|\{e | e \in \mathcal{Z}_i \wedge e \in \mathcal{Z}_j\}|}{|\mathcal{Z}_i| + |\mathcal{Z}_j|} \quad (20)$$

where $\mathcal{Z}_i = \{z_{i,1}, z_{i,2}, \dots, z_{i,m_i}\}$ represents the set of nodes that constitute the path from the root to the i th gene. Given the similarity function f_t of gene ontology codes,

the similarity measure between two genes g_i and g_j is calculated as the aggregation of similarity among the gene ontology codes:

$$\text{sim}(g_i, g_j) = \frac{1}{n_i} \sum_{k=1}^{n_i} \max_{1 \leq l \leq n_j} f_t(t_{i,l}, t_{j,k}). \quad (21)$$

The similarity matrix S is finally generated by setting $S_{i,j} = \text{sim}(g_i, g_j)$.

3.5 Gene profiles by free texts

In the second approach, we generate gene profiles by identifying the key concepts from the related research papers. With these profiles, the similarity between the genes can be calculated to guide the regression model. More specifically, to find the key concepts related to a gene, our approach first uses all the name variations of the given gene as query words to retrieve the research papers that are likely to provide information about the gene. Key concepts are then extracted from the returned documents to build the profile of this gene. Finally, each gene is seen as a vector represented in the space of key concepts and the similarity scores between two genes by their vector dot product. The document collection used in our experiment consists of a sub-collection from PubMed for the last ten years (1993–2003), which includes roughly three millions biomedical abstracts.

There are three issues that are important to the above approach of calculating similarity of gene based on the free texts of research papers:

Retrieval algorithms. The first issue is which retrieval algorithm is more effective for identifying the documents that are relevant to given genes. Two approaches are considered in our study: the *Boolean* retrieval approach that retrieves documents that exactly match with the textual queries, and the well known *Okapi* method (Robertson and Walker, 1999) that weights the key concepts based on their Term Frequency (TF), Inverse Document Frequency (IDF), and the document length. More specifically, if $\mathbf{d} = (d_1, d_2, \dots, d_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$ denote the term vectors of a document and a query, respectively, the Okapi method computes the similarity between \mathbf{d} and \mathbf{q} as follows:

$$\text{sim}_t(\mathbf{d}, \mathbf{q}) = \sum_{k=1}^m \frac{d_k q_k \log \left(\frac{N+0.5}{n_k+0.5} \right)}{0.5 + 1.5|\mathbf{d}|/|\bar{\mathbf{d}}| + d_k}$$

where N is the number of documents in the collection, n_k is the number of documents that contain the k th concept, and $\bar{\mathbf{d}}$ is the average document length for the entire collection. To further enhance the retrieval accuracy of the Okapi method, the pseudo-relevance feedback technique is used. In particular, the top 20 documents retrieved by the Okapi method are analysed and the ten most common words of the top ranked documents are used to expand the queries. Then, the Okapi retrieval algorithm is further applied to the expanded query, and the top 100 documents are used to generate the gene profiles.

Key concepts. The second issue is what type of key concepts should be extracted from the retrieved documents to form the profiles of genes. One natural choice is to use the common words among the abstracts of the retrieved papers. However, since many of

the extracted words may have nothing to do with the biological functions of the genes, the resulting similarity may not reflect the true relationship among the genes in the biological process to be regressed. In order to reduce the number of irrelevant words, we use the keywords provided by the Medical Subject Headings (MeSH) ontology to construct the text profiles of genes. More specifically, only the keywords of the MeSH codes that belong to the Mesh G category (i.e., biological science) are considered. From the profiles of all genes, we rank the all the concepts (i.e., MeSH codes in the G category) by their occurrences. Some most common concepts (e.g., Amino Acid Sequence; G06.184.603.060) are found not informative with respect to biological processes. These concepts are removed from the profiles of the genes to reduce the data noise. Finally, there are a total of 2742 unique Mesh G codes that remain as the key concepts.

Similarity. The third issue is how to compute the similarity of two genes in the vector space of 2752 concepts. In particular, there are a number of MeSH G codes that are irrelevant in determining the similarity among genes. Hence, we need to weight each concept appropriately such that the resulting similarity of genes can reflect the true association of genes in their roles in the biological processes. To this end, we consider two different approaches for term weighting: the *TF.IDF* term weighting method and the *automatic* term weighting method. The TF.IDF term weighting is commonly used in the study of information retrieval (Salton and Buckley, 1988). Let the text profiles of gene g_i and g_j be represented by $\mathbf{z}_i = (z_{i,1}, z_{i,2}, \dots, z_{i,c})$ and $\mathbf{z}_j = (z_{j,1}, z_{j,2}, \dots, z_{j,c})$, where c is the number of concepts used to represent the genes. Then, the TF.IDF method computes the similarity of gene g_i and g_j as follows:

$$\text{sim}(g_i, g_j) = \sum_{k=1}^c z_{i,k} z_{j,k} \log \left(\frac{N + 0.5}{n_k + 0.5} \right)$$

where N is the number of documents within the entire collection, n_k is the number of documents that contains the k th concept, and c is the number of concepts.

The TF.IDF based similarity measurement is based on the assumption that two genes are more likely to share similar biological functions when the rare words/concepts are shared by their text profiles than when the common words/concepts are shared by their text profiles. One problem with the TF.IDF term weighting method is that such an assumption may not necessarily be valid in computing the similarity between genes in their biological functions. The second problem with the TF.IDF method is that the weights assigned to the key concepts are completely independent from the regression model, and therefore may not be optimal for the given regression process. To resolve these two problems, we propose the automatic term weighting method that exploits the information within the regression process for the assignment of term weights. In the following, we give the mathematical details of how to connect the regression model with the assignment of term weights.

Let $\theta = (\theta_1, \theta_2, \dots, \theta_c)$ denote the weights assigned to all the words used by the text profiles. Then, the similarity $\text{sim}_a(g_i, g_j)$ between gene g_i and g_j is computed as

$$\text{sim}_a(g_i, g_j) = \sum_{k=1}^c z_{i,k} z_{j,k} \theta_k.$$

Using the weighted similarity function $\text{sim}_a(g_i, g_j)$ for $S_{i,j}$, we can rewrite the entire similarity matrix S as the linear combination of similarity matrix of each individual word, i.e.,

$$S = \sum_{k=1}^c \theta_k S^k$$

where S^k stands for the gene similarity matrix based on the k th word. More specifically, element $S_{i,j}^k$ is 1 only when both gene g_i and g_j have the k th word in their text profiles. Similarly, we can decompose the graph Laplacian as

$$L = \sum_{k=1}^c \theta_k L^k \tag{22}$$

where L^k is the graph Laplacian that is constructed based on the information of the k th word. Like most Bayesian approaches, we can then introduce a prior for each weight θ_k and extend the variational method described before to estimate the posterior distribution $\Pr(\mathbf{w} | \mathbf{y}, X)$ of the regression weights \mathbf{w} . However, this approach could be computationally expensive when the number of words used in the text profiles is large. In our empirical study, over 2000 unique keywords are used for the gene text profiles. In order to significantly reduce the computational cost, we first construct the matrix $U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$ for the text profiles of all the genes. We then apply the Singular Value Decomposition (SVD) to the matrix U to extract its first m principle components. Let $(\lambda_i, \mathbf{v}_i), i = 1, \dots, K$ be the top K eigenvalues and eigenvectors of matrix $U^T U$. Similar to equation (22), we rewrite the graph Laplacian L in the following linear combination form:

$$L = \sum_{i=1}^K \theta_i \lambda_i \mathbf{v}_i \mathbf{v}_i^T.$$

Notice that in the above expression, we use parameters θ s to represent the weights of the principle eigenvectors, instead of the weights of the words. By choosing a relatively small number of eigenvectors, we are able to reduce the number of parameters θ significantly. Similar to the Bayesian framework described before, we introduce a Gamma distribution $\mathcal{G}(\theta_i; \alpha_i, \beta_i)$ as the prior $\Pr(\theta_i)$ for each weight θ_i , and the likelihood $\Pr(\mathbf{w} | \theta, \tau_e)$ becomes

$$\Pr(\mathbf{w} | \theta, \tau_e) \sim \mathcal{N}\left(\mathbf{w}; \mathbf{0}_n, \left(\tau_e I_n + \sum_{i=1}^K \theta_i \lambda_i \mathbf{v}_i \mathbf{v}_i^T\right)^{-1}\right).$$

By putting the above distribution together, we then have the probability $\Pr(\mathbf{y} | X)$ expressed as:

$$\Pr(\mathbf{w} | \mathbf{y}, X) = \int d\theta_1 d\theta_2 \dots d\theta_K d\tau_e d\tau_r \Pr(\tau_e) \Pr(\tau_r) \prod_{i=1}^K \Pr(\theta_i) \Pr(\mathbf{w} | \theta, \tau_e) \Pr(\tau_r) \prod_{i=1}^d \Pr(y_i; \mathbf{w}, \mathbf{x}_i, \tau_r)$$

Using the variational approach, we introduce the variational distributions $\{\phi_i(\theta_i)\}_{i=1}^K$, $\phi_e(\tau_e)$, $\phi_r(\tau_r)$, and $\phi_w(\mathbf{w})$, which are

$$\begin{aligned}\phi_i(\theta_i) &\sim \mathcal{G}(\theta_i; a_l, b_l) \\ \phi_e(\tau_e) &\sim \mathcal{G}(\tau_e; a_e, b_e) \\ \phi_r(\tau_r) &\sim \mathcal{G}(\tau_r; a_r, b_r) \\ \phi_w(\mathbf{w}) &\sim \mathcal{N}(\mathbf{w}; \bar{\mathbf{w}}, S_w).\end{aligned}$$

We have the updating equations that are similar to equation (12) to (19) except the equations for γ_k , a_k , and b_k . They are:

$$\begin{aligned}\gamma_k &= \frac{\lambda_k b_e \exp(\psi(a_k))}{\lambda_k b_e \exp(\psi(a_k)) + b_k \exp(\psi(a_e))} \\ a_k &= \alpha_l + \frac{1}{2} \gamma_k \\ b_k &= \beta_l + \frac{1}{2} (\mathbf{v}_k^\top \mathbf{w})^2.\end{aligned}$$

4 Experiments

4.1 Experimental data

cdNA microarray and metabolic data. Microarray gene expression and metabolic data are obtained for HepG2 cells exposed to Free Fatty Acids (FFAs) and Tumor Necrosis Factor (TNF- α). The biological experiment are designed and employed to identify the effects of individual treatments as well as the interaction effects of FFA and TNF- α in the development of hepatic disorders. In total eight different conditions are used in the biological experiment. For each condition, we obtained cDNA microarray gene expression and metabolic data. The data consisted initially of 19458 genes and 64 metabolic functions or processes. Of the 830 genes identified by the ANOVA analysis, the top 100 genes identified by the GA/PLS analysis to be important for the LDH release were selected for further analysis. Performing the Gene Ontology Tree Machine (GOTM) analysis on the 100 genes identified that five different GO-groups were significantly enriched ($p < 0.01$). The 40 genes belonging to these gene groups were then selected for further analysis (Li et al., 2007).

In this regression problem, we assumed that the value of each metabolic function can be approximated by a linear combination of the expression levels of the selected genes. We further assumed that there is a different LR model for each metabolic function or process and the regression model is independent of the condition used for obtaining the gene expression and metabolic data. Thus, for each regression model, there are 40 regression weights to be determined with only eight data points (corresponding to the number of conditions).

4.2 Baselines and evaluation

We evaluate the quality of regression models using the normalised regression error, which is defined as follows:

$$err = (y - \hat{y})^2 / \sigma_y^2$$

where y and \hat{y} is the true and the estimated output value, respectively. σ_y^2 is the variance of y , which is estimated from the metabolic data and represent different metabolic process, are obtained under different conditions. We use the leave one out cross validation to evaluate the proposed regression models: for each condition, a separate regression model is trained on the metabolic data that are measured on the other seven conditions, and the metabolic data of the given condition is predicted using the trained regression model. We measure the normalised regression error for each condition and average the errors across eight conditions and 64 metabolic processes, which is used as the final indicator of the quality of regression.

Two baseline models are used in this study. The first baseline model is the straightforward regression that is already described in Section 2. The second baseline model is the ridge regression model that is also described in Section 2. To determine the optimal value of τ_e in the regularised regression model, we further apply the leave one out cross validation to the training sets that consist of seven conditions.

4.3 Experiment (I): Effectiveness of the proposed Bayesian framework

In this experiment, we compare the results of the proposed Bayesian framework to the straightforward LR model and the ridge regression model that is commonly used for the sparse data problem. The similarity matrix used by the Bayesian approach is constructed based on the gene ontology database. The detailed description of using gene ontology information for similarity measurement can be found in Section 3.4. The normalised regression errors for the straightforward LR and the proposed framework using different similarity matrices are presented in Table 1.

First, we see that the Bayesian approach is able to reduce the regression errors substantially from 12.9–7.9%. This implies that the prior knowledge extracted from the gene ontology is useful for guiding the selection of regression weights. Second, it is surprising to see that the RLR performs slightly worse than the the straightforward LR. Since the key parameter in the ridge regression is τ_e and is chosen by leave one cross validation, we vary the value of the parameter τ_e from 1 to 20. The normalised regression errors of RLR using different τ_e are presented in Table 2. Clearly, none of these values lead to a normalised regression error that is noticeably better than the original LR. Based on the above observation, we conclude that the Bayesian approach is effective in incorporating the gene ontology information into the LR model.

Table 1 Normalised regression errors for Linear Regression (LR), Regularized Linear Regression (RLR), and the Bayesian framework using the gene ontology for computing gene similarity (Bayesian (GO))

LR (%)	Bayesian (GO) (%)	RLR (%)
12.9	7.9	14.5

Table 2 Normalised regression errors for the Regularized Linear Regression model using different τ_e

τ_e	1 (%)	3 (%)	5 (%)	7 (%)	10 (%)	20 (%)
	12.9	12.8	12.8	12.8	13.0	13.6

4.4 Experiment (II): Effectiveness of free text profiles for Linear Regression

In this experiment, we examine the effectiveness of using the gene profiles based on the free texts on the LR model. In particular, we will examine the impact of different retrieval algorithms and different weighting schemes on the regression accuracy. Two retrieval algorithms, namely the Boolean retrieval and the Okapi retrieval method, and two term weighting algorithms, namely the TF.IDF term weighting and the automatic term weighting, are used in the study. The normalised regression errors for the four possible combinations between the two retrieval algorithms and the two weighting schemes are summarised in Table 3.

Table 3 Normalised regression errors for the Bayesian framework using two retrieval algorithms (i.e., the Boolean retrieval algorithm (Boolean) and the Okapi retrieval algorithm (Okapi)) and two term weighting schemes (i.e., the TF.IDF term weighting (TF.IDF) and the automatic term weighting (Automatic))

	TF.IDF (%)	Automatic (%)
Boolean	14.3	10.8
Okapi	14.7	8.8

First, we observed that the similarity measurement based on the TF.IDF term weighting is unable to reduce the regression error regardless of the retrieval algorithms that are used to find relevant documents. In fact, the normalised regression error is slightly increased, from 12.9% to over 14.0%, when the gene similarity based on the TF.IDF term weighting is used in the regression model. On the other hand, the normalised regression error is reduced noticeably when the automatic term weighting is used. In particular, the regression error is reduced from 12.9% to 10.8% when the Boolean retrieval algorithm is used, and is reduced to 8.8% when the Okapi retrieval algorithm is used. This result indicates that term weighting schemes are more important in determining the similarity of genes than the retrieval algorithms. We believe that this is due to the fact that many of the 2742 concepts are irrelevant to the similarity measure of the genes. In particular, this result implies that the underlying assumption behind the TF.IDF term weighting method may not be appropriate. For example, the MeSH keyword ‘Frameshift Mutation’ appears in the text profile of the gene G13.920.590.300. Although this keyword appears in 1654 research papers out of 3 million documents, it is not related to a specific biological function of the gene. Thus, the co-occurrence of this word in the profiles of two genes does not imply similarity of the two genes specific to an actual cellular function.

Second, comparing the regression errors of using the automatic term weighting, we see that the Okapi method is more effective than the Boolean retrieval algorithm. This is consistent with most of the previous studies in information retrieval, which usually reveals a significant advantage of using the Okapi method than the simple Boolean retrieval algorithm. Finally, comparing the normalised regression error of using free text retrieval based on the automatic weighting scheme to that of using the gene ontology, we observed that the two approaches achieved similar performance. This result indicates that the free text retrieval based approach can be as effective as the gene ontology based approach if appropriate weighting scheme is applied.

5 Conclusion

In this paper, we studied the problem of exploiting text mining methods to improve the quality of the LR models. The main idea of this paper is to represent genes in a biological system by their text profiles, which include keywords that describe the properties and functions of the genes. These text profiles are used to estimate the similarity of genes, which is applied to guide the selection of the regression weights that are assigned to the genes. A Bayesian framework is presented to incorporate the text information into the regression process. Unlike most of the previous studies that focus on using the gene ontology information, one of the key contribution of this paper is to explore the approach of representing genes by keywords that are extracted from the free text retrieval. This is significant given that a large portion of genes are not included in the gene ontology database. In order to address the problem of irrelevant keywords in the gene profiles, an automatic method is presented to weight the keywords appropriately. Our empirical study showed that the proposed Bayesian framework is effective in incorporating the gene similarity information into the regression model. Our empirical study also showed that the automatic weighting approach is effective in reducing the effect of irrelevant keywords. Furthermore, the experimental results indicate that it is better to use the GO code information for calculating similarity measures between genes when the genes can be found in existing gene ontology database since this approach is faster and more accurate than the text mining approach.

6 Relate work

There has been considerable previous research on information retrieval for biomedical documents. The Genomics track of NIST's Text REtrieval Conferences (TREC) was established in 2003 with the aim to encourage researcher in information retrieval for biological text application by providing a large test collection including large corpora with a sufficient set of queries and relevance judgements (Hersh and Bhupatiraju, 2003; Hersh et al., 2004). A wide range of information retrieval and text categorisation methods has been studied in the TREC Genomics track. Studies from several groups have shown the effectiveness of the Okapi retrieval method and the pseudo relevance feedback approach for biological document retrieval and classification (Seki and Mostafa, 2005; Buttcher et al., 2005; Huang et al., 2005).

The valuable knowledge from biomedical text retrieval/mining has been successfully applied to the analysis of gene expression data (Tanabe et al., 1999) and gene clustering (Shatkay et al., 2000; Speer et al., 2004). Although a number of studies have been devoted to exploring the gene ontology information for the prediction of protein functions (Jensen et al., 2003; Lu et al., 2004), none to date have considered using GO in regression problems. The goal of most regression problems in bioinformatics is to predict the activity level of an entire biology system, which is considerably more challenging than the classification problems that only predict the properties and functions of individual genes. Unlike the clustering problems in bioinformatics that only consider the gene expression data, the regression problems have to take in account the gene expression data as well as the outputs of a biological system. Hence, it is usually more difficult to incorporate the prior knowledge into the regression models than into the clustering models.

Acknowledgement

This research was partially supported by NSF grant (IIS0610784), NIH grant (R01 GM079688-01) and Michigan Universities Commercialisation Initiative (MUCI) Challenge Fund. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors, and do not necessarily reflect those of the sponsors.

References

- Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.
- Buttcher, S.L., Clarke, C.L.A. and Cormack, G.V. (2005) 'Domain-specific synonym expansion and validation for biomedical information retrieval', *Proc. 13th Text Retrieval Conference*, Gaithersburg, MD.
- Casella, G. and Berger, R. (2001) *Statistical Inference*, Duxbury Resource Center, Brooks-Cole, CA.
- Chung, F.R.K. (1997) *Spectral Graph Theory*, AMS Press, Providence, RI.
- Common, P. (1994) 'Independent component analysis, a new concept?', *Singal Processing*, Vol. 36, pp.287–314.
- Friedman, J. (1989) 'Regularized discriminative analysis', *J. Amer. Stat. Assoc.*, Vol. 84, pp.165–175.
- Gelman, A., Carlin, B.J. and Rubin, B.D. (2004) *Bayesian Data Analysis*, 2nd ed., Chapman & Hall/CRC, London, UK.
- Hersh, W. and Bhupatiraju, R. (2003) 'TREC genomics track overview', *Proceedings of the 12th Text Retrieval Conference*, Gaithersburg, MD.
- Hersh, W., Bhupatiraju, R.T., Rose, L., Johnson, P., Cohen, A.M. and Kraemer, D.F. (2004) 'TREC 2004 genomics track overview', *Proc. 13th Text Retrieval Conference*, Gaithersburg, MD.
- Hoerl, A. and Kennard, R. (1970) 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics*, Vol. 12, pp.55–67.
- Hskuldsson, A. (1988) 'PLS regression methods', *J. Chemometrics*, Vol. 2, No. 3, pp.211–228.
- Huang, J., Zhong, M. and Si, L. (2005) 'York University at TREC 2005: genomic track', *Proc. 14th Text REtrieval Conference*, Gaithersburg, MD.
- Jensen, L., Gupta, R., Staefeldt, H.H. and Brunak, S. (2003) 'Prediction of human protein function according to gene ontology categories', *Bioinformatics*, Vol. 19, No. 5, pp.635–642.
- Jordan, M.J., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K. (1999) 'An introduction to variational methods for graphical models', *Machine Learning*, Vol. 37, No. 2, pp.183–233.
- Lewis, D.D. and Ringuette, M. (1994) 'Comparison of two learning algorithms for text categorization', *Proc. Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, Las Vegas, US, pp.81–93.
- Li, Z., Srivastava, S., Yang, S., Mittal, S., Norton, P., Resau, J. and Chan, C. (2007) 'A hierarchical approach to identify pathways that confer cytotoxicity in HepG2 cells from metabolic and gene expression profiles', *BMC Systems Biology*, May 11, Vol. 1, No. 1, p.21.
- Lu, X., Zhai, C., Gopalakrishnan, V. and Buchanan, B.G. (2004) 'Automatic annotation of protein motif function with gene ontology terms', *BMC Bioinformatics*, Vol. 5, p.122.
- Robertson, S.E. and Walker, S. (1999) 'Okapi/Keenbow at TREC-8', *Proc. Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg, MD.

- Salton, G. and Buckley, C. (1988) ‘Term-weighting approaches in automatic text retrieval’, *Information Processing and Management*, Vol. 24, No. 5, pp.513–523.
- Seki, K. and Mostafa, J. (2005) ‘An application of text categorization methods to gene ontology annotation’, *SIGIR’05*, pp.138–145.
- Shatkay, H., Edwards, S., Wilbur, W.J. and Boguski, M. (2000) ‘Genes, themes and microarrays, using information retrieval for large-scale gene analysis’, *Proc. ISMB 2000*, La Jolla, California, pp.317–328.
- Speer, N., Spieth, C. and Zell, A. (2004) ‘A memetic clustering algorithm for the functional partition of genes based on the gene ontology’, *Proc. CIBCB 2004*, La Jolla, California, pp.252–259.
- Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. (1999) ‘MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling’, *BioTechniques*, Vol. 27, pp.1210–1217.
- Tibshirani, R. (1996) ‘Regression shrinkage and selection via the lasso’, *J. Royal. Statist. Soc. B.*, Vol. 58, pp.267–288.
- Townsend, J. and Hartl, D. (2002) ‘Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments’, *Genome Biology*, Vol. 3, No. 12, pp.research0071.1–research0071.16.
- Wiener, E., Pedersen, J.O. and Weigend, A.S. (1995) ‘A neural network approach to topic spotting’, *Proc. SDAIR’95*, Las Vegas, NV, pp.317–332.

Appendix

In the appendix, we show a brief derivation of calculating the variational distributions γ_k in equation (11).

Each variational distribution γ_k is chosen to maximise the value of $\gamma_k \langle \log(\tau_l \lambda_k) \rangle + (1 - \gamma_k) \langle \log \tau_e \rangle + H(\gamma_k)$. We can calculate the derivative of this value with respect of γ_k . Furthermore, we can set the derivative to 0 and solve the value of γ_k as follows:

$$\gamma_k = \frac{\exp(\langle \log(\tau_l \lambda_k) \rangle)}{\exp(\langle \log \tau_l \lambda_k \rangle) + \exp(\langle \log \tau_e \rangle)}. \quad (23)$$

Since gamma distributions belong to the exponential family, the expectations of sufficient statistics $\langle \log \tau_l \rangle$ and $\langle \log \tau_e \rangle$ can be calculated by differentiating the normalisation factor (Casella and Berger, 2001). By inserting the expectations into the above function, we obtain:

$$\gamma_k = \frac{\lambda_k b_e \exp(\psi(a_l))}{\lambda_k b_e \exp(\psi(a_l)) + b_l \exp(\psi(a_e))}. \quad (24)$$