BAYESIAN SEMIPARAMETRIC MODELS FOR HETEROGENEOUS CROSS-PLATFORM DIFFERENTIAL GENE EXPRESSION

A Dissertation

by

SOMA SEKHAR DHAVALA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Statistics

BAYESIAN SEMIPARAMETRIC MODELS FOR HETEROGENEOUS CROSS-PLATFORM DIFFERENTIAL GENE EXPRESSION

A Dissertation

by

SOMA SEKHAR DHAVALA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Bani K. Mallick
Committee Members,	Raymond J. Carroll
	Jeffrey D. Hart
	Seth D. Guikema
Head of Department,	Simon J. Sheather

December, 2010

Major Subject: Statistics

ABSTRACT

Bayesian Semiparametric Models for Heterogeneous
Cross-platform Differential Gene Expression. (December 2010)

Soma Sekhar Dhavala, B.S., Andhra University;
M.S., Indian Institute of Technology, Madras
Chair of Advisory Committee: Dr. Bani K. Mallick

We are concerned with testing for differential expression and consider three different aspects of such testing procedures. First, we develop an exact ANOVA type model for discrete gene expression data, produced by technologies such as a Massively Parallel Signature Sequencing (MPSS), Serial Analysis of Gene Expression (SAGE) or other next generation sequencing technologies. We adopt two Bayesian hierarchical models—one parametric and the other semiparametric with a Dirichlet process prior that has the ability to borrow strength across related signatures, where a signature is a specific arrangement of the nucleotides. We utilize the discreteness of the Dirichlet process prior to cluster signatures that exhibit similar differential expression profiles. Tests for differential expression are carried out using non-parametric approaches, while controlling the false discovery rate. Next, we consider ways to combine expression data from different studies, possibly produced by different technologies resulting in mixed type responses, such as Microarrays and MPSS. Depending on the technology, the expression data can be continuous or discrete and can have different technology dependent noise characteristics. Adding to the difficulty, genes can have an arbitrary correlation structure both within and across studies. Performing several hypothesis tests for differential expression could also lead to false discoveries. We propose to address all the above challenges using a Hierarchical Dirichlet process with a spike-and-slab base prior on the random effects, while smoothing splines model the unknown link functions that map different technology dependent manifestations to latent processes upon which inference is based. Finally, we propose an algorithm for controlling different error measures in a Bayesian multiple testing under generic loss functions, including the widely used uniform loss function. We do not make any specific assumptions about the underlying probability model but require that indicator variables for the individual hypotheses are available as a component of the inference. Given this information, we recast multiple hypothesis testing as a combinatorial optimization problem and in particular, the 0-1 knapsack problem which can be solved efficiently using a variety of algorithms, both approximate and exact in nature.

To my parents, wife, uncle and grandmother

ACKNOWLEDGMENTS

A Ph.D. is fighting a *chosen* battle, only to win. Why wouldn't I when I am not alone, but have an army of trusted allies. Here is a chance for me to acknowledge their unflinching support.

Whenever I thought, yup, I want to solve this problem, and searched the literature, I found that it has been solved by an another, Dr. Bani Mallick. It is a pleasure to work with him, as he has an uncanny ability to envisage cutting-edge problems. As a result, I got an opportunity to work on problems of practical significance with applications in diverse areas. His wisdom on conducting research and research management has benefited my own research. I consider myself extremely fortunate to have worked under his supervision and his pleasant personality always made the work environment a home away from home.

I am grateful to Dr. Carroll for giving me a chance to work with him. Many of his earlier works from a large repository have enhanced my understanding of key statistical concepts. I am thrilled to continue working with him in my next assignment as well. The laptop provided by him beats my desktop hands-down, reminding me of his expectations, challenges and deliverables. My first formal course that is directly related to this dissertation began with Dr. Hart's course on Decision Theory. *Well begun is half done.* How true that is in my case. My collaboration with Dr. Seth started with a summer job. Even in this short duration, his style of functioning, his way of translating ideas into algorithms, his planning and scheduling work and his views of using statistics from an application point of view have greatly influenced my work. I thank all of them for being on my dissertation committee.

I could not have done this work without the support I received from our collaborators from the Veterinary department. Dr. Garry Adams, Dr. Sangeeta Khare and

Dr. Sara Lawhon answered all my questions, performed validation runs, helped me write the manuscript and painstakingly conducted the experiments to produce the data. Discussions with them have greatly improved my understanding of the problems from biologists' point of view. Certain characteristics of the data have enabled us to develop new methods. Understanding the Dirichlet process was made easy due to Dr. Sujay Datta's persistent effort to educate me and I benefited from his experience in classical hypothesis testing procedures. I also thank Dr. David Dahl for offering his views on clustering and Dr. Alan Dabney for his advice on choosing an appropriate method for normalizing and noise removal in Microarrays. I would like to thank Dr. Edsel Peña and Dr. Bob Stine for the inspiring talks presented during the Resampling Conference held at Texas A&M University. An after-presentation discussion with Dr. Peña has lead us to pursue multiple testing as a resource allocation problem presented in Chapter IV.

Any graduate student in the department agrees, in no uncertain terms, that Dr. Mike Longnecker is a role model for any aspiring teacher. He corrected me whenever I was off the track and made me realize how serious graduate education in the U.S. is. I curse myself for bunking his lectures. What a grave mistake I made, and I will regret it for the rest of my life. I particularly thank Dr. Jinhua Huang for his inspiring classroom lectures. I felt right at home in his lectures; his enthusiasm to learn and try anything new and to probe the intellect is always something that I will cherish. I would like to thank the administrative staff: Marylin Randall, Shou-Fen, Sandra, Athena, Elaine, and Joyce for treating students in a friendly manner and making the work environment very hospitable. I thank the Department Head Dr. Simon for turning the department into a professional body of excellence. I have no doubt that under his leadership, our department will rank among the top in the country in the near future. I thank the organizers of the Departmental Colloquium and Biostatistics Seminar series. The computing support team headed by Henrick in the department has always been helpful. The support system at A&M is truly remarkable. In particular, the library has been instrumental in my gaining access to a wealth of knowledge with just the click of a button. I thank my lab mate, Rajesh, and student collaborators, Lin Zhang and Alex Konomi. Other graduate students Brian Hartman, Yogesh, Beverly, and Souporno were inspiring with some new ideas and making the grad life fun-filled.

I have a long list of friends who have helped me in many ways. My apologies for an incomplete list. I would like to thank my roommates at various stages, Kaushik, Anand, Hari, Giri, Uday, Monish, and Sansi at A&M, with whom I shared many of the beautiful moments in college life. I thank Mr. & Mrs. Satish, Mr. & Mrs. Srinivas, Mr. & Mrs. Radhika, Mr. & Mrs. Tulasi, Mr. & Mrs. Murali, Mr. & Mrs. Sriram, Mr. & Mrs. Vijay, Lakshmi, Anusha, Sushmitha, and Vichika, and many others who filled my stay with a sense of completeness. I thank all the members of SPICMACY, CRY, AID, India Association, Mr. & Mrs. Prof. Akhil Datta-Gupta, Prof. Mike Greenwald, Arun Surendran, Amnaya and many others, who have helped me see the world beyond academics. I thank my friends at Iowa State, Ravi Sekhar, Balaji, Sriram A, Sriram Y, Ranga, Sundeep, Prashanth, Murali, Natarajan, and many others.

I like to thank Rama Krishna and Y. V. Kishore, my undergraduate classmates, Lakshmi, Diane at the Aggieland Credit Union, and my parents and my uncle who provided financial assistance when I was struggling financially. I acknowledge the monetary assistance offered by the department in terms of teaching assignments. It is with its support that I was able to attend the Joint Statistical Meetings several times. For the last two years, my research was completely funded by several grants from NSF and KAUST, which were possible due to Dr. Bani Mallick and others.

They have also provided monetary support to attend important conferences.

I am indebted to my parents and family for doing all that they did, many times going out of their way, absorbing peer social pressure on my behalf, and shielding me from the aftershocks. I thank my wife Deepika, without whom I could not have completed this work. Her pleasant, calm and composed nature, and simplicity in thoughts brought much needed balance and stability to my work and life. Finally, I thank the Almighty: the invisible, that explains away the unexplainable.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

A. Problem Formulation

The perpetual endeavor of mankind perambulates around the quest to understand itself and the surroundings. *How life lives?* is the fundamental question we have been asking ourselves in that process. The simplicity of the question is, perhaps, an indication of the magnitude of the complexity of the undertaking. The discoveries made in the life sciences in the 20th century, together with the technological inventions being made in the 21st century, now present a unique & challenging opportunity to understand life better than ever before. The implications of such an endeavor can result in improving the quality of life. For example, understanding *how life is encoded* in DNA and *how it regulates the functional aspects of life* can help personalize medicine, develop new cost-effective drugs to combat various infectious diseases such as HIV, Malaria, Hepatitis-B, Tuberculosis etc..(Hofmann, 2006). However, such an opportunity also poses many challenges to be overcome. In particular, vast amounts of data produced by the next generation high-throughput sequencing technologies have to be mined to uncover meaningful information encoded by the generic make-up and efficient methodologies have to be developed to leverage information that is already available in several heterogeneous formats.

Multitude of high-throughput technologies are currently used to simultaneously analyze thousands of genes of an organism, or its constituents. They can be broadly classified into analog and digital techniques based on the type of data they produce. Microarrays and its many variations produce images that can be regarded as con-

The journal model is *Journal of the American Statistical Association.*

tinuous data, while Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS) can be considered as digital technologies (Bloksberg, 2008). Digital techniques directly measure the number of cDNA or mRNA molecules present,o whereas in the case of Microarrays the image intensity is proportional to the abundance of the particular gene in the DNA sample (Seidel, 2008). Microarrays are comparatively inexpensive and as a result, they are widely used and many statistical methods have been developed to analyze them. For analyzing digital or discrete gene expression data, most of the existing methods apply some kind of transform and treat it as continuous data, thereby making the analog gene expression analysis techniques readily applicable. In this dissertation, we propose statistical methods and models to explicitly address the discrete nature of the data produced by digital technologies and attempt to develop a unified framework which allows fusion between cross-platform data. An underlying theme in our contributions to the literature is concerned with performing multiple hypothesis tests for differential expression of genes between different experimental conditions.

B. Organization

More specifically, in Chapter II, we apply Generalized Linear Model (GLM) to analyze count data obtained using Massively Parallel Signature Sequencing (MPSS). Over dispersion due to the presence of large number of zeros, as is the case with count-based technologies, is accounted by using a zero-inflated Poission likelihood. The goal is to study differential expression between different strains of Salmonella bacteria in Bovines. We show that the suggested Bayesian hierarchical model performs better than off-the-shelve techniques which model each gene independently. This is due to the hierarchical specification of the model, which borrows strength across related genes. We flag genes that are found to be differentially expressed based on Kullback-Leibler distance based hypothesis tests and we control the False Discovery Rate (FDR) to account for the large of number of simultaneous hypothesis being carried out. we extend this hierarchical model to account for dependency among genes by eliciting a Dirichlet process (DP) for the random effects. Relaxing the parametric assumption to model dependency often improves the estimates of the effects parameters in terms of variance, as will be shown. As a consequence to using DP, genes having similar expression profiles will be clustered together. This information is useful in exploring the functional relationship between genes in a given cluster.

In Chapter III, we consider methods for combining gene expression data and infer differential expression based on the combined data. We propose semiparametric methods to jointly model data arising from different studies or different technologies or a combination of the two, but assume that both studies are trying to measure the same biological tissue or organism under two experimental conditions. We treat the observed data as a manifestation of a latent process which explains the differential expression in the genes both witn-in and across studies. The observed data is linked to the latent process through a non-parametric function. An ANOVA like structure, similar in spirit to GLM in previous chapter, for the latent process forms the backbone on which inference is carried-out. More specifically, we use zero-inflated Poisson likelihood to model count data as discussed in Chapter II. Continuous data arising from Microarrays or similar technological platforms is modeled using a Normal likelihood. Canonical link functions of these exponential families link the expected means of the observed data to a transformed latent process. The transformation is accomplished using penalized cubic splines and we suppose that this monotonic transformation captures the technological differences between different platforms or studies, leaving the biological information to the latent process. We model the latent process as having an ANOVA like structure. We elicit Hierarchical Dirichlet process prior for the random effects but instead of using a normal base prior, we use a spike-n-slab base prior. The advantage of using the spike-n-slab prior is that hypothesis testing is embedded into the model and therefore, separate tests based on the Kullback-Leibler distances or related methods are not needed. The HDP allows the clusters to be shared across studies or data-sets, which means that a cluster in a study can have completely different membership in a different study, thereby allowing the genes to have discordant expression profile across studies. This is quite a relaxed assumption to considering that a gene in a study must correspond to the same cluster in a different study, referred to as concordance.

Both Chapter II and III are devoted to developing methods from a modeling point of view. A critical component in inference after modeling is, how one controls FDR or other error measures, as we are dealing with numerous multiple hypothesis tests. It is customary to assume uniform loss functions, in which case, closed-form solutions exist for controlling FDR. In Chapter IV, using coherent decision theoretic framework, we develop algorithms that optimize a chosen error measure, like FDR, under generic loss functions. Finally, in Chapter V, we summarize our findings and suggest future research directions.

CHAPTER II

SEMIPARAMETRIC MODEL FOR DISCRETE EXPRESSION DATA

A. Introduction

Transcription profiling techniques measure the expression of a gene in a biological sample by quantifying mRNA. DNA Microarray technology has dominated the fields of molecular biology and genomics by allowing researchers to measure the expression levels of thousands of genes simultaneously. However, it is now well known that the microarray technology has its own limitations. Microarray users are limited only to the probes printed on commercial/custom manufactured slides and one must have knowledge of the nucleotide sequences of the genes that are being investigated in order to create the probes. This may be a problem for genome-wide studies of higher organisms. In addition, Microarray studies are subject to variability relating to probe hybridization differences and cross-reactivity, element-to-element differences within Microarrays during spotting, and Microarray-to-Microarray differences (Audic and Claveries, 1997, Wittes and Friedman, 1999, Richmond et al, 1999).

Recently, alternative technologies such as Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS) have emerged that are capable of addressing some of the issues described above. Both SAGE and MPSS produce similar output: a list of short sequences (tags) and a frequency for each tag. However, the method of obtaining the tag list is dramatically different. SAGE uses concatenated tags that are sequenced using a traditional automated DNA sequencing method (Velculesu et al. 1995). A SAGE library may contain as few as 20,000 tags or as many as over 100,000 tags . In contrast, MPSS uses a cloning and sequencing method whereby hundreds of thousands of sequences are obtained simultaneously by sequencing off of beads using enzymatic digestion and hybridization (Brenner et al. 2000). An MPSS library may contain more than 1,200,000 tags . Both methods are capable of uniformly analyzing gene expression irrespective of mRNA abundance and without *a priori* knowledge of the transcript sequence. The data generated by these methods are count data, as opposed to the log signal intensity or log red/green ratio obtained from Microarrays.

In this chapter, we focus on gene expression of bovine ileal Peyer's patch infected with Salmonella enterica serovar Typhimurium as analyzed by MPSS. In Section B, we briefly explain the MPSS technology. In Section C, we explain the data collection process and provide a brief literature survey of the existing statistical techniques for analyzing MPSS and SAGE data. In Section D, we explain our statistical models and methodology. Section E provides some of the computational details associated with the models introduced here, as well as our analysis of the data. In Section F, we discuss the results of our analysis and Section G discusses their biological interpretation. We end with a few concluding remarks in Section H. The MCMC computational details can be found in Appendix A.

B. Review of MPSS Technology

MPSS is based on transcript counting and as a result, depends heavily on the ability to uniquely identify every mRNA in a sample. For this purpose, first a cDNA signature/tag conjugate library is constructed. Poly$(A)^+$ mRNA is extracted from the tissue of interest and from it, cDNA is synthesized. The 20 bases adjacent to a specific site upstream from the poly-A tail of each cDNA (a site that reads GATC) are captured. The 17 nucleotides including the GATC and its contiguous 13-mer form a *signature* for the mRNA they came from. These signatures are then amplified by

PCR and a unique identification tag is added to each of them. Subsequently, multiple pools of several hundred thousand signature/tag conjugates are amplified and the tags are hybridized with microbeads, each of which has on it several thousand copies of one of the anti-tags. The microbeads loaded with the signature/tag conjugates are isolated by using a florescence-activated cell sorter. A million to a million-and-a-half loaded microbeads are assembled in a flow cell and the signature sequences on those beads are determined. This process involves the parallel identification of four bases by means of hybridization to fluorescently labeled encoders, followed by the removal of those 4 bases via digestion with an endonuclease enzyme and the exposure of the next four bases, and so on.

Two separate sets of microbeads containing the same signature library are used along with two different initiating adapters for the endonuclease digestion process and for each of these, the signature identification process is independently carried out k times ($k = 2, 3$ or 4). The purpose behind these is to ensure that fewer signatures are missed, thereby increasing the resolution. The two separate runs of the endonuclease digestion process mentioned above are called a *two-stepper* process and a *four-stepper* process respectively. The k independent runs of the signature identification process within each *stepper* process are called *replicates*. The signatures corresponding to every mRNA in the tissue-sample are, therefore, identified and counted $2k$ times during MPSS. So, for every signature involved, we end up getting two sets of k counts. These counts are actually reported after standardization to a million, that is, a signature having a count of 72 among 1.5 million microbeads will be reported as having 48 TPM (transcripts per million). For example, if a two-stepper process is used along with a four-stepper and each has 4 replicates, the TPM values for a particular signature might be (5,0,9,13) and (0,3,12,20). The maximum TPM value for a signature is often called its '*selected mean*'. Once the TPM values are available, each 17-nucleotide signature, which typically matches with only *one* position in a complex genome, is associated with a proximal gene. Based on the position of the signature relative to its associated gene, each signature is categorized according to the quality of the association. For more on the MPSS technology and the associated biological details, see Reinartz et al. (2002), Stolovitzky et al. (2005), Crawford et al. (2006) and Brenner et al. (2000), among others.

C. Data Collection and Analysis

The dataset was generated by Khare et al. (2006) at the Department of Veterinary Pathobiology, Texas A & M University. Their goal was to compare the *in vivo* global gene expression responses in tissue-samples from bovine ligated ileal loops that are infected with a wild strain of the bacterium *Salmonella enterica serovar Typhimurium* (WT) or with a mutant strain of the bacterium (MUT) or are uninfected (LB). *Salmonella* is an enteric pathogen and is a major concern in food safety. *Salmonella enterica serovar Typhimurium* (*S. Typhimurium*) is among the most common *Salmonella* varieties causing salmonellosis in the U.S.A. Various animal models have been studied to understand the virulence mechanism of *S. Typhimurium*. Infection of calves, natural or experimental, with *S. Typhimurium* results in an enteric disease with clinical and pathological features that parallel the disease in humans. The invasion associated with type III secretion system encoded by Pathogenicity Island I (SPI-1) in the pathogen is required for *S. Typhimurium* colonization of the bovine small intestine and translocates various *Salmonella* effector proteins including SipA, SopA, SopB, SopD, and SopE2 to the host epithelial cell cytoplasm. These various effector molecules act in a coordinated way to induce fluid secretion and transcription of various genes associated with the pathophysiology of the disease. The mutation in the wild type strain of *Salmonella* was the deletion of the Type 3 Secretion System (T3SS) genes SipASopABDE/E2 (ZA21), which, according to recent studies, renders the bacterium defective in invasion, fluid accumulation, production of inflammatory cytokines/chemokines and transmigration of neutrophils. In this chapter, we compare the hosts' responses to the WT associated with the pathophysiology of the disease to understand the detailed effect of these virulent factors in the pathogenesis of the infection. The detailed analysis of the host response will lead us to identify gene targets for therapeutic intervention of this disease. The MPSS technique with a two-stepper procedure was used along with a three-stepper procedure, each having two replicates. The experimenters initially ran the MPSS, as described below, with more than 43000 signatures, but subsequently screened the dataset to eliminate all of the rows that did not contain a count of four or more transcripts per million in at least two of the three tissue-samples. This reduced the number of rows (signatures) to about 24000 and our analysis is based on this reduced dataset. All the replicates we obtained in this data-set are technical replicates as obtaining biological replicates was cost prohibitive.

1. Review of MPSS Data Analysis Methods

From a statistical point of view, an MPSS experiment involving m signatures produces a dataset with m rows, each containing two sets of k transcripts per million values. Suppose we have two tissue-samples, one healthy and the other diseased and we intend to discover signatures that are differentially expressed between these two samples. Reinartz et al. (2002) suggest the following procedure, in case there is only one count per signature. Let the counts be x_1 and x_2 in the two samples for a particular signature. Since a certain microbead may or may not contain that signature and a million microbeads are examined for the presence of that signature in each sample, this is like a repeating a million coin-tosses twice. Let p_1 be the expression level of this signature in sample 1 and p_2 be the level in sample 2. Then, clearly, $x_1 \sim \text{Binomial}(10^6, p_1)$ and $x_2 \sim \text{Binomial}(10^6, p_2)$ and the null hypothesis of *no differential expression* boils down to $H_0 : p_1 = p_2$. Of course, $\widehat{p_1} = x_1/10^6$, $\widehat{p_2} = x_2/10^6$ and in order to test this null hypothesis against the two-sided alternative, a normal approximation is appropriate. In other words, use the test statistic $z = (\widehat{p_1} - \widehat{p_2})/\{\widehat{p}(1-\widehat{p})(10^{-6}+10^{-6})\}^{0.5}$, where $\widehat{p} = (x_1+x_2)/(10^6+10^6)$. One problem with this approach is that it does not clarify what data are being used. If the single count per signature that is being used is actually the *selected mean*, i.e., the largest of all the counts for that signature, it would be better modeled by the *maximum* of binomial random variables. On the other hand, a normal-approximated binomial testing procedure might be appropriate when the sum of the counts for a signature is considered. Even then, such signature-by-signature testing methods incorrectly assume that the signatures are independent of each other, especially signatures corresponding to the same gene. Another drawback to this is the lack of protection against false discoveries; such protection is essential as thousands of hypothesis tests have to performed simultaneously.

Stolovitzky et al. (2005) put forth another testing procedure based on empirical modeling. Their dataset was generated by a two-stepper process along with a four-stepper, each with 4 replicates. Instead of using the TPM values directly, they use $\log_{10}(TPM)$. For the first tissue-sample, let θ_{ij} be the log-transformed TPM for the i^{th} signature in the j^{th} replicate of a stepper process, $\overline{\theta}_i$ be the mean of the θ_{ij}'s within a stepper process and s_i be their standard deviation. For the other tissue-sample, let the corresponding quantities be θ^*_{ij}, $\overline{\theta}^*_i$ and s^*_i. By plotting the s_i's against the θ_i's, Stolovitzky et al. (2005) noticed that the variation decreases as the mean increases, a phenomenon typically observed in log-transformed Poisson data, and it does so at different rates for the following three scenarios:**(i)** when none of the eight TPM counts is zero; **(ii)** when some of them are zeros but neither the sum of the four TPM counts from the two-stepper process nor that from the four-stepper process is zero; **(iii)** when exactly one of the two sums mentioned above is zero but not both.

Typically, the rate of decrease in replicate-variation as the mean increases is the highest in case **(i)**, while the other two cases are similar to each other. In view of this, they decided to standardize the replicate θ_{ij}'s for each signature in each sample by their standard deviation and model the standardized data by the curve $f(x) = \frac{1}{2}\exp\{-x^2/(0.5 + 0.6 \mid x \mid)\}$, which has slightly heavier tails than a Gaussian curve. For the i^{th} signature, they reject the null hypothesis of *no differential expression* against the two-sided alternative if the *conditional* probability of observing a greater absolute difference between the means of the standardized θ_{ij}'s from the two samples is "small", given that the average of those two means is some value Θ (say).

Although this seems to be a more sophisticated approach than the normal-approximated binomial test, it has its own drawbacks. The authors' decision to work with log-transformed counts means that they had to eliminate all of the zero counts from their analysis, except for acknowledging the effect of zero counts on the inter-replicate variations and adjusting for them. In addition, the authors do not mention false discovery rate (FDR) control despite the fact that simultaneous testing for the differential expression of several thousand signatures necessitates some protection against false discoveries.

2. Review of SAGE Data Analysis Methods

From a data-centric view point, MPSS and SAGE technologies produce similar output: the frequency of occurrence of signatures/tags. As a result, it might be possible to use the methods developed for SAGE data analysis in order to analyze MPSS data and vice versa, as suggested in Vencio et al. (2004). For this reason, we review SAGE data analysis methods that could be applied to MPSS data as well.

Most of the off-the-shelf SAGE analysis techniques for testing differential expression use simple chi-square tests for equality of proportions or perform *t*-tests after transforming the data (Man, 2000). Even though such simplistic assumptions are easy-to-use, they do not adequately model the complexity of biological processes or account for the interdependencies among genes. Alternative approaches use hierarchical models to address these issues. Vencio et al. (2004) suggest mixture model distributions to account for within-class variability and in particular use a Beta-Binomial model. Testing differential expression is accomplished by computing the Bayes error rate, which is the area of the overlapped region of the posterior distributions. A similar approach can be found in Thygesen and Zwinderman (2006), where a gamma-Poisson model is used for analyzing the SAGE libraries, where library is a collection of expression levels of signatures/tags from a particular biological sample. They suggest that a Poisson likelihood with either a gamma prior or a log-normal prior is suitable for modeling SAGE data. A mixture Dirichlet prior is used for analyzing SAGE libraries in Morris et al. (2006). They demonstrate that such a specification leads to improved estimates of the expression levels of the signatures/tags. A key feature in the above methods is the mixture model approach to account within-class variability. However, they do not model the complex dependency among the genes or explicitly incorporate tests for differential expression across multiple samples into the model. Rather, these methods assume just exchangeability as is the case with our parametric hierarchical model and analyze one library at a time. In this article, we adopt a new approach to analyzing MPSS data that addresses these issues. In our Bayesian hierarchical model, we:

- model each signature-count by a zero-inflated Poisson (ZIP) distribution and assume a normal density for the log-transformed mean parameter of the Poisson part.

- assume the mean of the above-mentioned normal density to have a linear model structure with parameters capturing the signature effect and the treatment effect.

- start with a parametric model where these parameters are given the usual conjugate prior distributions (i.e., normal and inverse gamma).

- proceed to fit a semiparametric model where the 'treatment effect' parameter is given a Dirichlet process prior with a normal baseline distribution.

- borrow strength within each cluster of signatures, since the semiparametric model results in automatic clustering.

- use the deviance information criterion (DIC) for choosing between these two models.

- draw inference on differential expression of signatures based on the posteriors of the 'treatment effect' parameters, using symmetrized Kullback-Leibler (KL) divergences with bootstrapped cut-off values, as well as the Kruskal-Wallis test for the equality of medians. A somewhat similar modeling idea can be found in Carota and Parmigiani (2002) in the regression context. But to our knowledge, we are the first to modify and adopt it for MPSS-type count-data.

Even though our analysis and application is for the MPSS data analysis, we believe that the methods are equally applicable for analyzing SAGE data, owing to the similarities between the nature of the data.

D. Bayesian Hierarchical Model

1. Parametric Model

Let Y_{ijk} be the k^{th} replicate count observed for the i^{th} signature under the j^{th} treatment, $i = 1, \ldots, I$; $j = 1, \ldots, J$ and $k = 1, \ldots, K$. We assume that conditional on the parameters (p, λ_{ijk}), Y_{ijk} are independently distributed $ZIP(p, \lambda_{ijk})$ for $i = 1, \ldots, I$; $j = 1, \ldots, J$ and $k = 1, \ldots, K$. In other words,

$$\mathrm{P}(Y_{ijk} = y | p, \lambda_{ijk}) = pI(y = 0) + (1 - p)P(Y^*_{ijk} = y) \qquad (2.1)$$

for some $0 < p < 1$, where $Y^*_{ijk} \sim \text{Poisson}(\lambda_{ijk})$. In the next stage, we model $\log(\lambda_{ijk})$ as

$$\log(\lambda_{ijk}) = \eta_i + \beta_{ij} + \epsilon_{ijk} \ , \qquad (2.2)$$

where ϵ_{ijk} is the random residual component with $\text{Normal}(0, \sigma^2_\epsilon)$ distribution. Hence, we assume that conditional on the parameters (η_i, β_{ij}), the λ_{ijk} are independent, each with a lognormal density. The use of a residual component in the link-function specification is consistent with the belief that there may be unexplained sources of variation in the data, perhaps due to explanatory variables that were not recorded in the original study. This is particularly appropriate for Poisson data sets where over-dispersion is commonly observed. The use of residual effects within GLMs is discussed in Sun et al. (2000) and is a special case of the class of generalized linear mixed models (Zeger and Karim, 1991; Breslow and Clayton, 1993). Here, we assume that conditional on the parameters (η_i, β_{ij}), the λ_{ijk} are independent each with lognormal density so equation (2) can now be re-written as:

$$\log(\lambda_{ijk}) \sim \text{Normal}(\eta_i + \beta_{ij}, \sigma^2_\epsilon) \qquad (2.3)$$

where η_i is the effect of the i^{th} signature and β_{ij} is the effect of the j^{th} treatment nested within the i^{th} signature. We elicit conjugate priors in the hierarchical model and partially center the parameters for efficient MCMC sampling (Gelfand et al, 1995). Let $\mathcal{NIG}$ be the Normal-Inverse Gamma family of conjugate distributions in which the mean has a Normal distribution conditional on the variance and the variance marginally follows an Inverse-Gamma distribution with hyper-prior parameters u and v having the appropriate subscripts. In other words,

$$\begin{aligned} \theta, \sigma^2 &\sim \mathcal{NIG}(\theta_0, \sigma^2, u, v) \text{ implies that} \\ \theta | \sigma^2 &\sim \mathcal{N}(\theta_0, \sigma^2) \text{ and} \\ \sigma^2 &\sim \mathcal{IG}(u, v). \end{aligned}$$

With this notation in mind, we specify the priors as:

$$\begin{aligned} \beta_{ij}, \sigma^2_\beta &\sim \mathcal{NIG}(0, \sigma^2_\beta, u^{\text{pr}}_\beta, v^{\text{pr}}_\beta) \\ \eta_i, \sigma^2_\eta &\sim \mathcal{NIG}(\mu, \sigma^2_\eta, u^{\text{pr}}_\eta, v^{\text{pr}}_\eta) \\ \mu, \sigma^2_\mu &\sim \mathcal{NIG}(\mu_0, \sigma^2_\mu, u^{\text{pr}}_\mu, v^{\text{pr}}_\mu). \end{aligned}$$

However, the specification of the zero-inflation parameter makes the sampling from the (conditional) posterior distribution extremely difficult. Agarwal et al. (2002), Ghosh et al. (2006) cleverly handle the problem by introducing a latent variable. In the context of our dataset, denoting the latent variable corresponding to y_{ijk} by ζ_{ijk}, the complete likelihood of the data is

$$L(y, z \mid p, \lambda) = \prod_i \prod_j \prod_k p^{\zeta_{ijk}} \left\{ (1-p) \frac{e^{-\lambda_{ijk}} \lambda_{ijk}^{y_{ijk}}}{y_{ijk}!} \right\}^{1-\zeta_{ijk}} \tag{2.4}$$

or, equivalently,

$$L(y, \zeta \mid p, \lambda) \;\; = \;\; p^{n_0}(1-p)^{n-n_0} \prod_{y_{ijk}>0} \frac{e^{-\lambda_{ijk}} \lambda_{ijk}^{y_{ijk}}}{y_{ijk}!} \prod_{y_{ijk}=0} (e^{-\lambda_{ijk}})^{1-\zeta_{ijk}}, \qquad (2.5)$$

where $n_0 = \sum_i \sum_j \sum_k \zeta_{ijk}$ and $n = IJK$. Here, $\zeta_{ijk} = 1$ implies that the kth replicate in the jth treatment for the ith gene was not sampled. We elicit Beta(a, b) prior on p and conjugate priors for all the variance parameters.

2. Semiparametric Model

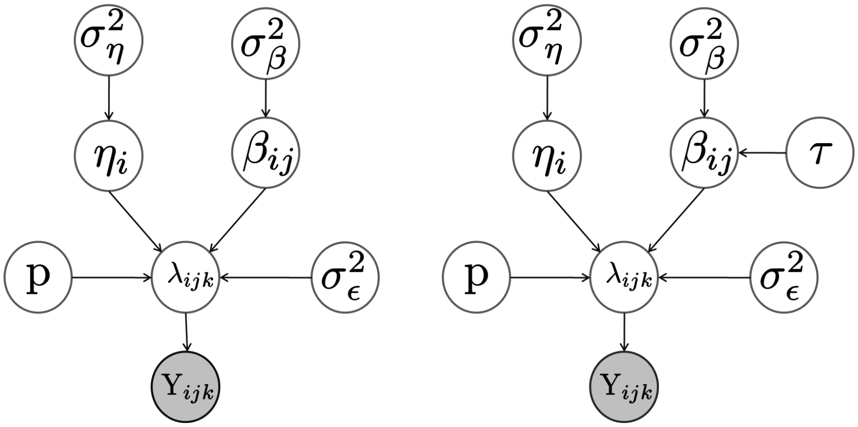
We extend the model in Section C.1 to a semiparametric setup where simultaneously we can infer the clustering of the signatures such that signatures within a cluster share a common value for their regression coefficients. Thus, similar signatures will borrow strength or shrink their regression coefficients locally rather than shrinking towards the global mean. Furthermore, clustering of the data offers insight about signatures that behave similarly in the experiment. By comparing signatures of unknown function with profiles that are similar to signatures of known functions, clues to biological function may be obtained.

We exploit the Dirichlet Process (Ferguson, 1973) prior for the regression coefficients to obtain the clusters. Assigning a Dirichlet process on the regression coefficients induce ties among them. That is, for every pair of objects $i \neq j$, there will be a positive probability that $\beta_i = \beta_j$. The clustering of the signatures encoded by the ties of the regression coefficients will simply be referred to as the clustering of the regression coefficients and, hence, clustering of the corresponding signatures. The semiparametric model is obtained by replacing prior for the treatment effect's parameter in the parametric model as:

$$\begin{aligned} \beta_{ij} &\sim \mathrm{DP}\{\tau N(0,\sigma_\beta^2)\} \\ \sigma_\beta^2 &\sim \mathcal{IG}(u_\beta^{\mathrm{pr}}, v_\beta^{\mathrm{pr}}) \end{aligned}$$

where τ is the tuning parameter and the baseline distribution is Normal$(0,\sigma_\beta^2)$. Posterior inference based on a Dirichlet process (DP) prior has been widely discussed in the literature. Ferguson (1973) introduced the Dirichlet process and Antoniak (1974) extended it to a DP mixing framework. Except in simple cases with few observations (Kuo, 1986), DP mixing was computationally intractable until Escobar and West (1995 and earlier reference therein) developed a convenient version of the Gibbs sampler (Gelfand and Smith, 1990) to handle this problem. Recent work of MacEachern and Mueller (1994), Neal (2000), Jain and Neal (2004), and Dahl (2005), among other supplied alternative simulation strategies to accommodate nonconjugate structures. All of these methods are suitable to model continuous responses; although see Mukopadhyay and Gelfand (1997) and Carota and Parmigiani (2002) for extensions to the linear model frame work. Both of these approaches obtain a nonconjugate structure and use a more complex MCMC algorithm successfully to handle that problem. In our case, the number of regression parameters is large and so these algorithms may not be very efficient.

In our modeling scheme, the introduction of the residual component ϵ makes our computation much more efficient. By adopting this Gaussian residual effect, many of the conditional distributions for the model parameters are now in the standard form, thus greatly aiding computation. To be specific, conditional on the λ-values, the model (2) is independent of y and can be written as a standard Bayes linear regression with $\log(\lambda)$ as the response and β as the regression parameters. Now using a DP prior over the β-values, this can be transformed to a conjugate problem, with the analytical form of the marginal distribution available. Hence our method enables us to use the efficient sampling scheme of Escobar and West (1998) to draw the β-values and other parameters. The details of the MCMC computation scheme are provided in the next Section. Graphical representation of both the models are shown in Fig. 1(a) and (b), respectively.



(a) parametric GLM

(b) semiparametric GLM

Fig. 1.: Graphical representation of the (a) parametric and (b) semiparametric models.

E. Details of The MCMC Computations

1. Prior Selection and Cluster Initialization

We elicited conjugate priors for all the variance parameters and set the corresponding hyper-parameters such that they are diffuse. As a result, posterior inference is insensitive to the choice of these hyper-priors. Nevertheless, a good initialization of the MCMC chain ensures stability and quicker convergence. It might also help an average user to use the algorithm in a fairly automatic manner. Further, initializing the cluster memberships greatly affects the simulation time required during burn-in period. For this reason, we suggest the following procedure:

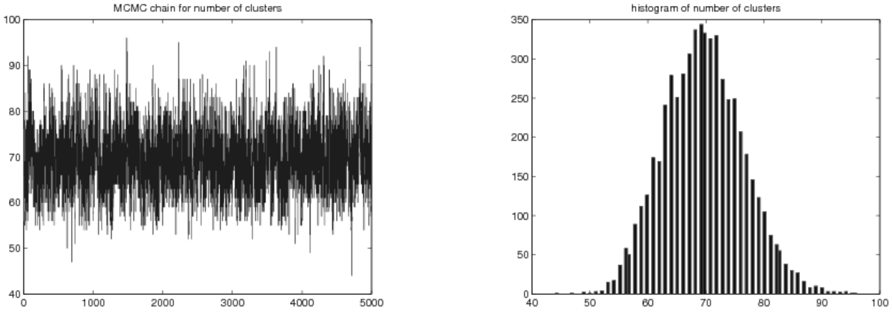
- log transform the data
- get the least-squared estimates of the treatment and signature effects parameters and their standard errors
- use k-Means or other non-parametric clustering methods to cluster the treatment effects parameters

Optionally, information obtained in the above steps can also be used to set the hyperprior parameters as well.

2. Posterior Sampling

Sampling from the posterior in the parametric case was performed using a block Gibbs sampler. All the conditional distributions except for the λ-values and the p's have conjugate forms. Using latent variables leads to sampling from the conditional distribution of the zero-inflation parameters, also from its conjugate distribution. A Metropolis-Hastings step with log-Normal proposal was used for drawing λ's. In the semiparametric case, successive sample observations are drawn using Escobar and West's (1995, 1998) Polya urn scheme. A total of 150,000 samples were drawn from the joint posterior. Of them, 30,000 were discarded as burn-in and the remaining samples were thinned down by a factor of 30 to give us reasonably less-dependent posterior samples. The DP precision parameter, τ was estimated empirically using $\sum_{n=1}^{N} \tau/(\tau + n - 1) \approx B^{-1} \sum_{b=1}^{B} N_b$ where N_b is the number of clusters in the b^{th} simulation and N is the maximum number of clusters possible (McAuliffe et al. 2006). In Figure 2, the MCMC chain of the number of clusters and its histogram are shown.

We can see that chain converged and is mixing well. A discussion on the prior elicitation on the precision parameter can be found in West (1992). Certain choices in model specification allow us to efficiently simulate the draws from the joint posterior distribution of the parameters, such as modeling $\log(\lambda)$ with a normal distribution. The full conditionals required in the MCMC simulation are given in the Appendix for both the parametric and the semiparametric models.



(a) MCMC chain

(b) Histogram

Fig. 2.: MCMC chain and histogram of N_b, the number of clusters.

3. Clustering

At each iteration of the MCMC, we obtain cluster membership information, i.e., which signatures belong to the same cluster. We can form a pairwise association or probability matrix δ based on this information, as follows: The $(i,i')^{th}$ cell is 1 if $\{\beta_{i1},\beta_{i2},\beta_{i3}\}$ and $\{\beta_{i'1},\beta_{i'2},\beta_{i'3}\}$ belong to the same cluster and is 0 otherwise. Clearly, $\delta(i,i)=1$ for each i. After M iterations in the MCMC, we can estimate the pairwise probability matrix as $\widehat{\delta}(i,i') = M^{-1}\sum_{n=1}^{M}\delta(i,i')$. Medvedovic et al (2002) used the estimated pairwise probability matrix to form the cluster structure by using $(1-\delta)$ as a distance measure. We followed this approach to form an agglomerative hierarchical cluster with complete linkage. Among the other alternatives to determine the cluster configuration, Dahl and Newton (2007) used a least-squares approach to find the best cluster within the MCMC framework. This has the advantage that the number of clusters is not needed *a priori*. Then, it is less reliable, because it is based on *a* single realization in MCMC. The advantage of using agglomerative hierarchical clustering, as opposed to choosing *a* cluster realization, is that the clusters are nested. As a result, the cluster membership does not change significantly even when the number of clusters changes. In the Bayesian framework, we can estimate the number of clusters either empirically or otherwise, as discussed before. In our case, the distribution of the *number of clusters* is unimodal and approximately symmetric with mode around 69, as shown in Fig. 2(b). We used this information to cut the tree to form the clusters in an objective manner, utilizing the cluster size information available within the MCMC.

In Figure 3, the profile plots of the treatment effects in seven representative clusters are shown. The plots in a column correspond to a specific treatment and plots in a row correspond to a particular cluster. For example, the profile plot in the 2^{nd} row and 1^{st} column corresponds to the kernel density estimates of the treatment effects parameters in the LB strain for all the signatures in the 19^{th} cluster ($\beta_{i,LB}\ i \in \mathcal{S}_{19}$). The cardinality of the cluster is also shown in the plot (34 in this case). We can tell from the profile plots that the distributions of the treatment effects' parameters are very similar when they are in the same cluster. This enables us to identify genes that are likely to be co-expressed, which may eventually lead to the discovery of pathways. As the cardinality increases, the profiles tend to be dissimilar due to the nature of the agglomerative clustering. This allows the biologists to look at different cluster configurations by cutting the dendrogram at different levels. In the present case, we chose to cut the tree by using the mean of N_b, though other subjective choices may possibly be justified too.

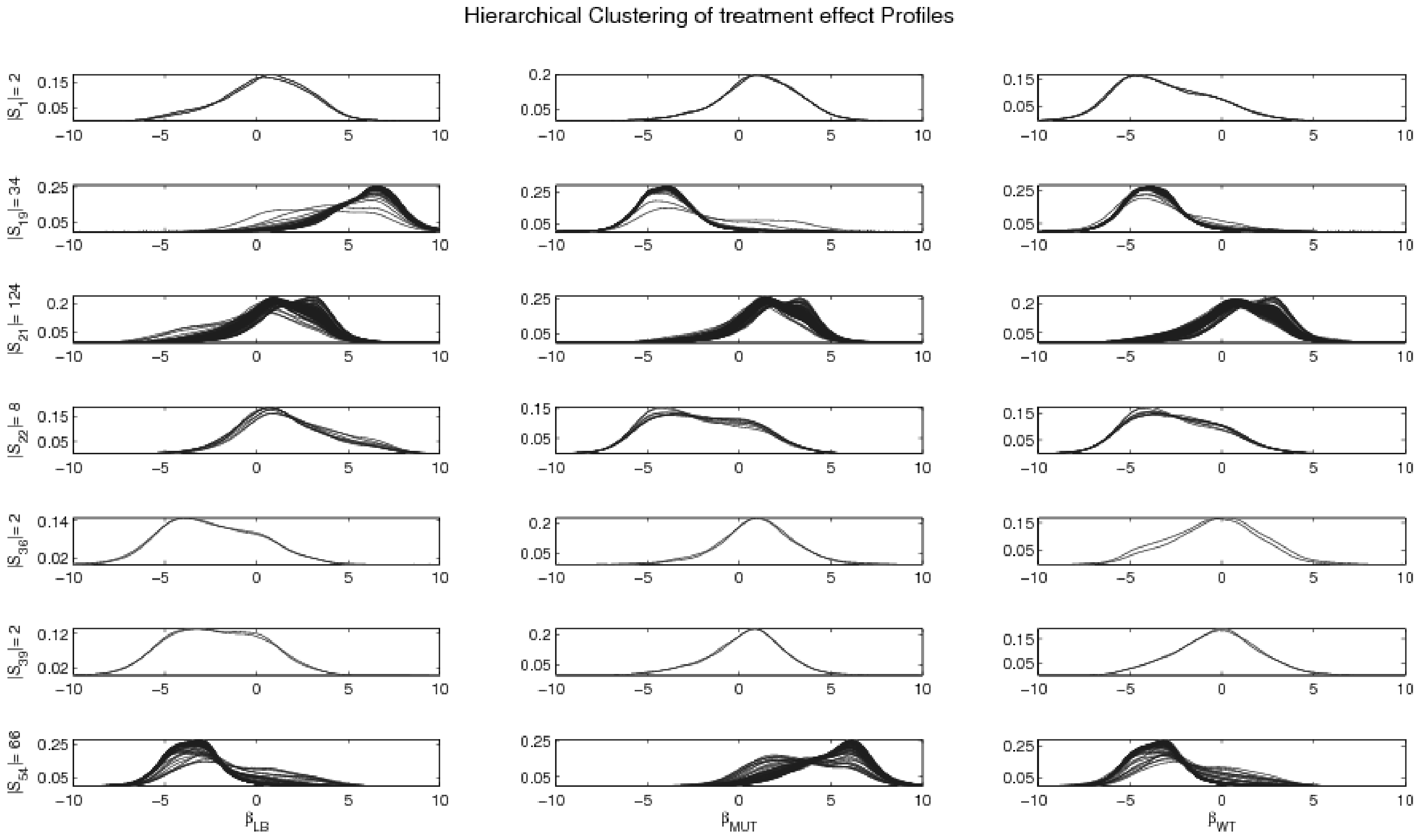


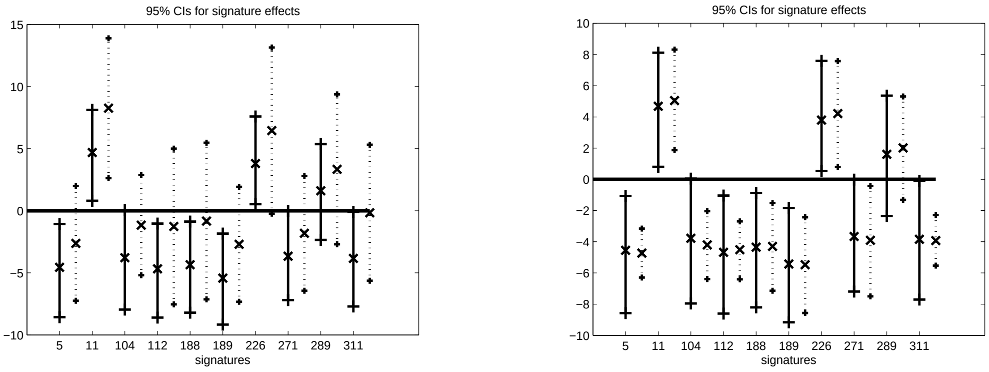
Fig. 3.: Hierarchical clustering profiles of treatments effects.

4. Model Selection

We used the Deviance Information Criteria (DIC) of Spiegelhalter et al. (2002) for model comparison. We have four models to begin with: the parametric and the semiparametric models with either regular Poisson or zero-inflated Poisson (ZIP) likelihood. We ran all four models on a small number of signatures and found that the parametric and semiparametric with ZIP likelihood were competitive and the models with ZIP likelihood have smaller deviance compared to the models based on the Poisson likelihood. In our analysis on the full dataset, the DIC for the parametric model was 29221 and it was 29091 for the semiparametric model and thus the semiparametric model was selected. The semiparametric model also leads to improved estimates, for example treatment effects and signatures effects have a smaller MSE compared to their counter-parts in the parametric model. In Figure 4(a), we plot the 95% credible intervals for some signatures η_i's, the signature effects parameters. As can be seen, the semiparametric model produced tighter intervals. We also fit the simple Poisson regression for each signature independently and plotted the corresponding credible intervals for those signatures in Fig. 4(b). It is clear that these intervals are much wider compared to the intervals corresponding to both parametric and semiparametric models. Furthermore, the remaining three models can be considered as special cases of the semiparametric model with the ZIP likelihood. It becomes evident if we note that the Poisson is a special case of ZIP with $p = 0$ and the parametric model can be obtained setting by $\tau = \infty$ in the semiparametric model.

5. Simulation Details

We have developed the software in MATLAB. We exploited the matrix representation and vector processing capabilities of MATLAB for accelerating the simulation time,



(a) 95% CIs for η_i's in parametric (solid) and independent GLM (dashed) models.

(b) 95% CIs for η_i's in parametric model (solid) and semiparametric(dashed) models.

Fig. 4.: Comparison of 95% CIs for selected η_is (signature effects)' parameters under different models.

particularly in implementing the block-Gibbs sampler. We ran the algorithms on our shared-memory heterogeneous 64-bit Linux cluster with more than six nodes and at least one dedicated node. Typical configuration of the nodes in our cluster has 16GB RAM and eight dual core processors clocking 2.46GHz. It took nearly two-three three days to complete the simulation for 189,000 draws and for 23,000 signatures. From a computational point of view, reducing the number of genes cuts down the simulation time dramatically. For example, a simulation involving 5000 genes selected using an initial filtering method takes about 4 hours for the same number of MCMC samples.

F. Results

Here we present a summary and interpretation of the results we obtained by fitting the semiparametric ZIP model to our MPSS dataset. After obtaining the samples from the posterior densities of the β-values for the i^{th} signature, inference regarding differential expression can be drawn in a number of different ways. For example, we could use the 'area of overlap' method of Vencio et al. (2005) or the threshold-based approach of Newton et al. (2004) and Lewin et al. (2006). Other choices are Ishwaran and Rao's (2005) asymptotic approach, the marginal posterior-based method of Gottardo et al. (2006) and the posterior tail probability-based approach of Bochkina and Richardson (2006). All these approaches require that the asymptotic distribution of the test statistic be known or one has to choose the rejection region of the hypothesis test in an adhoc fashion. To avoid these difficulties, we use two nonparametric methods, one based on symmetrized Kullback-Leibler (KL) divergences and the other on the nonparametric Kruskal-Wallis test.

For computing the KL divergences for each signature, we simulated 5000 sample-observations from the posterior of each β_{ij} $(j = 1, 2, 3)$. Then we computed three pairwise KL divergences (LB vs. MUT, LB vs. WT and MUT vs. WT). We declared the signature differentially expressed if at least one of these three is 'significantly large'. In order to identify the cut-off values beyond which we will call a KL distance 'significantly large', we resorted to the bootstrapped distribution of KL divergences. For each pairwise comparison (say, between LB and MUT), recall that we have 5000 observations from each of the two corresponding posteriors. From these 5000 observation-pairs, we selected a bootstrap sample size of 1000 and computed the KL divergence between the two posteriors based on them. We repeated this process 500 times, thereby ending up with 500 bootstrapped KL divergences between those two posteriors, and computed the p-value based these bootstrapped KL-divergences. We have plotted the histograms of the p-values of tests for differential expression among all treatment pairs in Fig. 5. An examination of the histogram indicates that the distribution of the p-values in MUT vs. WT are different from the rest of the pairwise comparisons. We also note that the distributions of the posterior distributions of the treatment effects are non-normal. Thus we used a nonparametric n-way

ANOVA, the Kruskal-Wallis test for equality of the medians (KW). However, as discussed in earlier sections, there can be many false discoveries due to the $I >> K$ problem. Controlling false discoveries within the Bayesian framework is possible by eliciting a mixture distribution under the null and the alternate hypotheses for the effects parameters (see Gottardo et al. 2006). However, such a set-up depends on the number of treatments and specific hypothesis tests that are being considered. To control the false discovery rate, we used the pFDR approach developed by Storey (2002,2003), which has a Bayesian interpretation. We used the `qvalue` `R`-package developed by Alan Dabney and John Storey that is availble for download from the URL `http://genomics.princeton.edu/storeylab/qvalue/`. The significance decisions were based on these q-values at the $\alpha = 0.05$ level.

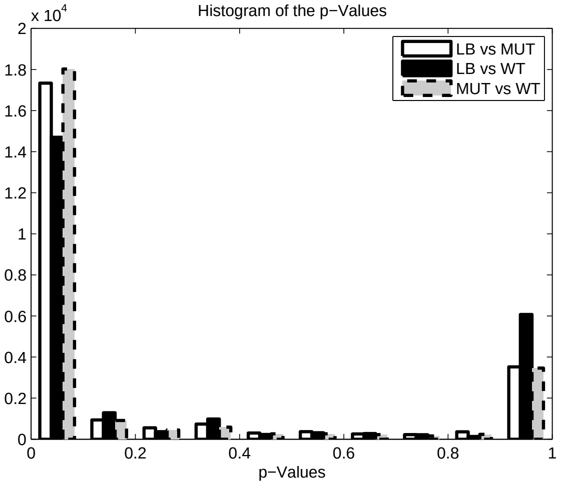


Fig. 5.: Histogram of p-values of the tests for pair-wise differential expression.

Among the numerous signatures that were detected to be differentially expressed by our inference methodology, we summarize the results for a few that spread across five important Gene Ontology categories described in the next section. Figure 6 provides the posterior densities of the treatment effects, i.e., the β_{ij} values, based on samples from those posteriors for signatures associated with genes that code for the proteins L-selectin, Ferritin, cAMP, Beta-actin, Laminin receptor, Rho-GTP, Cytokeratin-18 and MMPs. These signatures were found to be differentially expressed among the three tissue-samples by our Bayesian semiparametric method. Finally, in Table I, we show the q-values for pairwise differential expression across all pairs, along with the clustering information.

G. Biological Significance of Discoveries

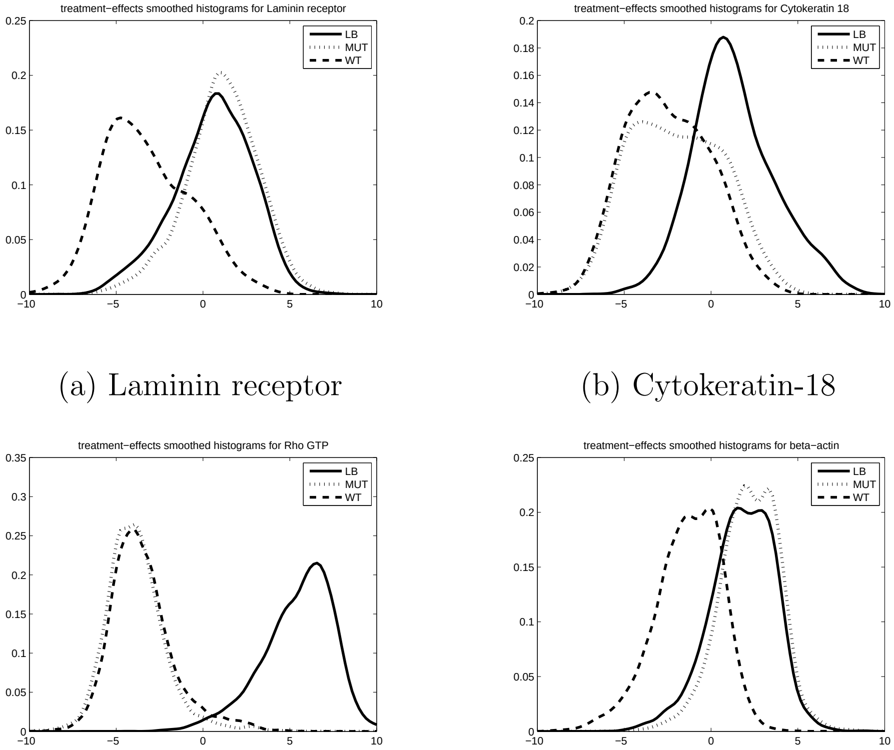
We now scrutinize the lists of differentially expressed signatures obtained through our Bayesian semiparametric analysis and discuss the biological significance of some of the corresponding genes. There are many important gene ontology (GO) groups based on the biological functions of the genes associated with the differentially expressed signatures detected by our semiparametric model, For example, we found representatives of the functional categories "actin cytoskeleton and extracellular matrix", "adhesion molecules", "ferritin-heavy polypeptide 1", "signal transduction" and "matrix metalloproteins and tissue inhibitors of metalloproteins". The detection of signatures corresponding to genes in these categories is consistent with the existing literature on the interactions between *Salmonella typhimurium* and the host-tissue proteins. What follows is a brief discussion on each of these functional categories.

Actin Cytoskeleton and Extracellular Matrix: The Type III secretion system (T3SS) encoded at *Salmonella* Pathogenicity Island I secretes effector proteins into the host intestinal/epithelial cell which bind to the actin cytoskeleton and induce the formation of ruffles in the cell membrane and Salmonella internalization (Guiney and Lesnick, 2005; Patel and Galan, 2005). The statistical methodology described

GenBank/	Cluster	net-DE	LB vs MUT		LB vs WT		MUT vs WT		
EST	Number	KW	KLD	KW	KLD	KW	KLD	KW	Annotation
BM105853	1	2.96e-05	4.18e-04	2.71e-03	1.27e-03	8.92e-04	4.70e-04	4.70e-04	telomerase associated protein
NM174379	1	2.96e-05	4.18e-04	1.01e-02	1.27e-03	8.92e-04	4.70e-04	4.70e-04	laminin receptor
X62882	16	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.48e-01	1.48e-01	L-selectin
NM174069	19	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.49e-01	1.49e-01	gap junction protein
AY156928	19	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.23e-01	1.23e-01	lectin-like receptor
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
BM435212	21	2.96e-05	1.32e-02	5.28e-03	1.42e-01	3.68e-02	4.7e-04	4.7e-04	ferritin
Be664796	21	2.96e-05	1.25e-02	2.84e-04	1.85e-03	8.92e-04	1.96e-02	1.96e-02	ATP synthase
NM176613	21	1.24e-02	4.18e-04	6.86e-02	2.73e-02	2.52e-01	1.51e-01	1.51e-01	ATP synthase
CB453188	21	2.96e-05	1.43e-02	6.54e-02	1.32e-03	8.92e-04	4.70e-04	4.70e-04	actin, beta
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
CB446386	22	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.75e-01	1.75e-01	immunoglobulin lambda chain
CB428925	22	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	3.14e-02	3.14e-02	cytokeratin 18
NM174641	22	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.18e-01	1.18e-01	guanylate cyclase
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
NM174471	36	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	1.66e-01	1.66e-01	metalloproteinase inhibitor
BF776620	36	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	4.70e-04	4.70e-04	epithelial transmembrane
BM432434	39	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	9.57e-3	9.57e-3	gelsolin
AW311904	54	2.96e-05	4.18e-04	2.84e-04	2.42e-01	6.28e-02	4.70e-04	4.70e-04	cAMP-regulated phosphoprotein
AY181987	54	2.96e-05	4.18e-04	2.84e-04	1.27e-03	8.92e-04	4.70e-04	4.70e-04	colony stimulating (macrophage)
X54183	54	2.96e-05	4.18e-04	2.84e-04	2.42e-01	2.32e-01	4.70e-04	4.70e-04	macrophage scavenger receptor
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Table I.: q-values of tests for pair-wise differential expression (LB-vs-Mut, LB-vs-WT and MUT-vs-WT) based on Kullback-Leibler distance (KLD) and Kruska-Wallis nonparametric test (KW). Reported also are the q-values for testing the hypothesis *at least one treatment has differential expression* (net-DE).

here identified MPSS signatures representing beta-actin, cytokeratin 18, Rho-GTP and laminin receptor 1 as differentially expressed between the LB and WT infected tissues; see Figures 6(a)-6(d).



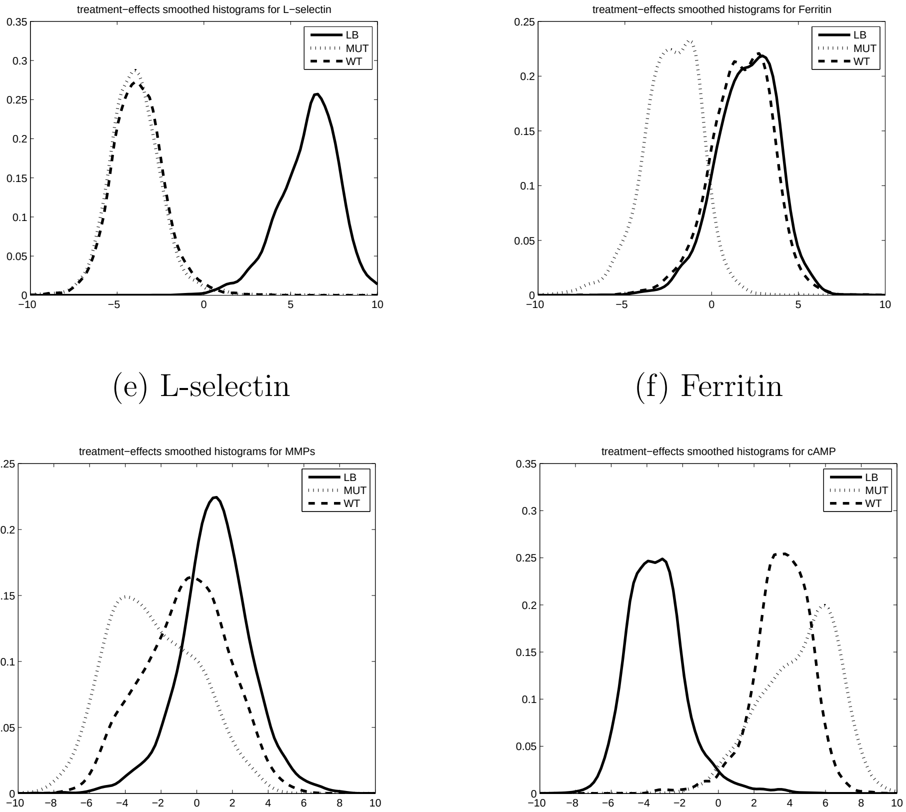
(a) Laminin receptor

(b) Cytokeratin-18

(c) Rho-GTP

(d) Beta-actin

Fig. 6.: Smoothed histograms of treatments effects (β_{ij}'s)for selected signatures: LB (solid), MUT (dots) and WT (dashes).



(e) L-selectin

(f) Ferritin

(g) MMPs

(h) cAMP

Fig. 6.: Continued.

Adhesion Molecules: L-selectin (lymphocyte adhesion molecule-1, CD62E, ELAM-1) is a transmembrane glycoprotein member of the selectin family of adhesion molecules expressed on the surface of activated leukocytes (Worthylake and Burridge, 2001). Expression of L-selectin is essential for the initial contact between leukocytes and endothelial cells required for extravasation of inflammatory cells into sites of inflammation (Barkhausen et al. 2005). Differential regulation of L-selectin in the MT infected tissue compared to LB suggests that L-selectin activation contributes to the mild tissue inflamation, see Figure 6(e).

Ferritin, Heavy Polypeptide 1: Our methodology identified MPSS signatures for heavy polypeptide 1 ferritin to be significantly differentially regulated in WT and MUT tissues compared to LB loops of ileum having Peyer's patch, see Figure (6(f). This is a biologically significant observation, because ferritins are ubiquitous iron storage proteins in plants, microorganisms and animals that play fundamental roles in soluble and cellular Fe homeostasis (Boughammoura et al. 2007). Ferritins also have a profound influence on inflammation and host resistance to pathogens (Ghio et al. 1997).

Matrix Metalloproteins: Matrix metalloproteinase (MMPs) are a group of enzymes that are capable of cleaving all components of the extracellular matrix that are involved in tissue invasion, extracellular matrix remodeling, angiogenesis and inflammation (Malemud 2006). In the present case, our methodology discovered down-regulation of MMPs in the MT infected tissue compared to the WT infected and LB tissues; see Figure 6(g).

Signal Transduction: The cAMP-regulated phosphoproteins (ARPP-16, ARPP-19, ARPP21 and DARPP32) modulate signal transduction, linking infection to host immunity, see Horiuchi et al. (1990) and Rakhilin et al. (2004). The increased expression of signal transduction molecules observed in *S. typhimurium*-infected tissues compared to LB, as discovered by our methodology, suggests that these molecules play a role in mediating some of the actions of vasoactive intestinal peptide (VIP) and cAMP-dependent protein kinases such as PKA in the intestine; see Figure 6(h).

Among the clusters detected by our semiparametric model, there are a few with special biological significance. In Table 1, we show selected signatures that were differentially expressed. Annotation information was obtained by using the Expressed Sequence Tags (ESTs) EST/ GenBank IDs at the NCBI repository. The q-values for the overall differential expression based on KW test and pairwise differential expression test across all pairs (LB Vs. MUT, LB vs WT and MUT Vs WT) using both KW and KLD are reported. The cluster number they belong to is also tabulated. Most of the genes associated with the differentially expressed signatures in cluster #19 have functions related to the immune-defense of the host tissue. Similarly, in cluster #21, most of the signatures have similar expression profiles (see Figure 3, row 3). Among the co-regulated signatures in this cluster are zinc-finger protein (not shown), ferritin and ATP-synthase related signatures. It is interesting that a colony stimulating protein and scavenger receptor are co-expressed. As is often the case, clustering in our analysis is also exploratory in nature. We have verified, though not comprehensively, the clustering information against the reference clusters available at NCBI GEO repository. We looked at the signatures that belong to a specific cluster produced by our method and searched for a cluster to which this particular signature belongs by looking at the UniGene IDs. Even though the cluster memberships are not identical, we observed similarities in terms of their biological functions and GO categories. Other clusters that we detected will be the subjects of future biological investigations.

H. Conclusion

Expression profiling techniques based on transcript counting offer a powerful alternative to conventional microarray technology and address some of its shortcomings. In the existing literature, MPSS data, or some transformation thereof, have been modeled by continuous densities. We have proposed two Bayesian hierarchical models for such count data and developed inference methodology for detecting the differential expression of a signature among the three tissue-samples. We adopted a flexible semiparametric modeling approach that enables automatic clustering and strengthborrowing within the clusters, thereby eliminating the unrealistic independence assumption among the signatures that has been exploited in the existing literature. These methods will be particularly useful when the sample size is small, because the hierarchical setup allows one to borrow strength among correlated signatures. Our model can also handle any number of experimental conditions and any number of replicates. Therefore, we believe that our method is useful in wide variety of situations.

Filtering the signatures with low counts is not necessary but it would positively impact the simulation time. Therefore, we recommend filtering the low-frequency, noisy, uninteresting signatures. In the future we shall investigate more efficient, optimal initial filtering which will accelerate the computing time.

Finally, our proposed methods can be used to analyze SAGE data, as has been discussed in section C.2. Hence, the proposed methods will be useful to analyze count data generated by deep sequencing technology which is becoming mainstream in recent biological studies.

CHAPTER III

SEMIPARAMETRIC MODEL FOR COMBINING HETEROGENEOUS GENE EXPRESSION DATA

A. Introduction

Microarrays have dominated the high-throughput genome-wide studies mainly due to cost when compared with their digital counterpart technologies such as SAGE, MPSS or other next generation sequencing technologies. More over, developing statistical methods to analyze them is relatively straightforward, owing to the normality assumption made. Consequently, there is a vast amount of literature available to analyze data-sets produced by Microarrays. For a collection of these methods, see monographs Parmigiani et al. (2003), Do et al. (2006), Mallick et al. (2009). For an exposition of analyzing digital technologies, see Chapter II of this dissertation. Often times, similar experiments are performed or similar transcript are studied, albeit at different laboratories. As complex as it gets, genome-wide studies are inherently multi-disciplinary and collaborative in nature and a logical proposition is to gain additional insights by pooling the information/knowledge available in disparate forms.

Meta-analysis is a method for combining information from multiple sources (Normand, 1999, Hedge et al. 1985). Its potential to improve the efficiency and reliability of biological investigations is being recognized; for example, meta-analysis is used to validate differential expression analysis (Daniel et al., 2000). SAGE and Microarray data were combined to discover potential biomarkers and improved detection in Nacht et al 1999. Combining data across multiple arrays can be found in Rhodes et al. 2002. A recent application of meta-analysis in Phylogenic studies can be found in Liang and Weiss (2007). Conlon et al. (2006) proposed a Bayesian hierarchical model for combining Microrray data, possibly arising from different platforms. However, many of them assume or require that genes or transcripts have similar expression profiles across studies. In other words, to apply these methods, the studies have to be reproducible. However, when considering multiple studies, the variations in the expression profiles of the genome can be both biological and technological (Irizarry et al. 2005; Consortium et al. 2006; Kerr 2007). Shcharpf et al. (2009) considered a Bayesian hierarchical model, called XDE, where genes can have either concordant or discordant differential expression across studies, accounting these differences. They have provided quite substantial evidence that their modeling framework offers better performance than many methods, see references there-in, for combining Microarray data-sets and indeed pooling can improve reliability of the biological discoveries. However, a major drawback of XDE is how strength is borrowed from different studies. They elicit multivariate normal distribution on the random effects which are of interest. While discordance/concordance is a desirable feature in modeling the random effects, it can potentially render the correlation structure non-identifiable, leading to poor mixing in the MCMC. Indeed, they also report poor mixing of these parameters, an indication that such a prior elicitation is less realistic. Further more, many of the above methods are mainly focused on combining continuous data-sets and occasionally heterogeneous data-sets but in which case, they just combine p-values, which are inefficient. In this chapter, we address the above mentioned challenges.

We let the data sources to be produced either by analog technologies like Microarrays or digital technologies like MPSS, SAGE, etc., and essentially this is Generalized linear model. Non-parametric link functions model the latent covariates with an ANOVA like structure, that is specific to each study. A Hierarchical Dirichlet process prior is elicited for the random effects, which induces ties among the genes both with-in and across studies, thereby modeling the correlation in a non-parametric fashion. In order to integrate multiple hypothesis right into the model, we employ a spike-and-slab selection prior (Ishwaran and Rao, 2005) as the base prior in the HDP. This provides a fully Bayesian approach to both modeling and hypothesis testing. As in Scharpf et al. (2009), we restrict our attention to two sample comparisons. The organization of the chapter is as follows: In Section B, we formulate the problem and develop the model. Details of prior elicitation are provided in Section C and implementation details using Markov Chain Monte Carlo are given in Section D. Several features of the model are demonstrated through simulated examples in Section E. We apply the modeling framework to analyze Salmonella infection in Bovine Illeal loop with data from MPSS and Microarrays in Section F. We conclude the chapter with a summary and discussion in Section F. Details of the MCMC computations and full conditionals are provided in Appendix B.

B. Model Formulation

Let Y_{hijk} be the k^{th} replicate observed for the i^{th} signature under the j^{th} treatment in study h. We assume exponential family for the likelihood.

$$Y_{hijk} \sim F_h(x_{hijk}) \tag{3.1}$$

We emphasize the likelihood with subscript h to explicitly suggest that it can be different for different studies, though belonging to the exponential family. More specifically, we elicit zero-inflated Poisson likelihood for discrete expression data like SAGE, MPSS, next-generation sequencing (Dhavala et al., 2010) and Normal likelihood for microarrays (Gottardo et al., 2006). That is,

$$Y_{hijk} \sim ZIP(p_h, \lambda_{hijk}), \, h = 1, 2, ..., n_d \tag{3.2}$$

where n_d is the number of discrete data-sets, p_h is the study-specific zero-inflation parameter and λ_{hijk} is the Poisson mean parameter. We use canonical link functions for linking the latent random variables which different technologies measure. For discrete-data sets,

$$\log\lambda_{hijk} \equiv y_{hijk}\, h = 1, 2, ..., n_d \tag{3.3}$$

and for continous data-sets, we use identity link function, i.e.,

$$Y_{hijk} \equiv y_{hijk},\ h = n_d + 1, n_d + 2, ..., n_d + n_c \tag{3.4}$$

where n_c is the number of continuous data-sets.

Now, each study measures study specific random effects. We use a non-parametric function to model this manifestation which is different for each study/data-set. It is typical to model unknown link functions in GLMs using non-parametric functions: Mallick and Gelfand (1999), and in measurement error models (Berry et al., 2002). That is:

$$y_{hijk} \sim \mathcal{N}(f_h(z_{hijk}), \sigma^2_{f,h}),\ h = 1, 2, ..., n_d + n_c \tag{3.5}$$

$$f_h(.) = \sum_{l=1}^{L} \alpha_{hl}\Psi_{hl}(.), \tag{3.6}$$

where z_{hijk} is the latent process which captures the differential gene-expression profiles in the h-th study and f_h is the function that maps the latent process which is then measured by which ever technology being used. Using cubic B-splines as the basis functions, Ψ_{hl}, offers flexibility and ease of implementation (Brezger and Steiner, 2008) and $\sigma^2_{f,h}$ is the study specific variance.

We model the latent random effects using an ANOVA like structure:

$$z_{hijk} \sim \mathcal{N}(\eta_{hi} + (2\psi_{hik} - 1)\beta_{hi}, \sigma^2_{z,\,h\psi_{hik}i}) \tag{3.7}$$

where where η_{hi} is the effect of the i-th signature/gene in the h-th study and β_{hi} is the effects parameter of the i-th signature in h-th study. $\psi_i \in [0, 1]$ indicates which of the two experimental conditional the observation belongs to. Notice that we have made the variance dependent on gene, study, and sample specific, essentially making it a heteroscadastic variance model. That completes model specification except the priors.

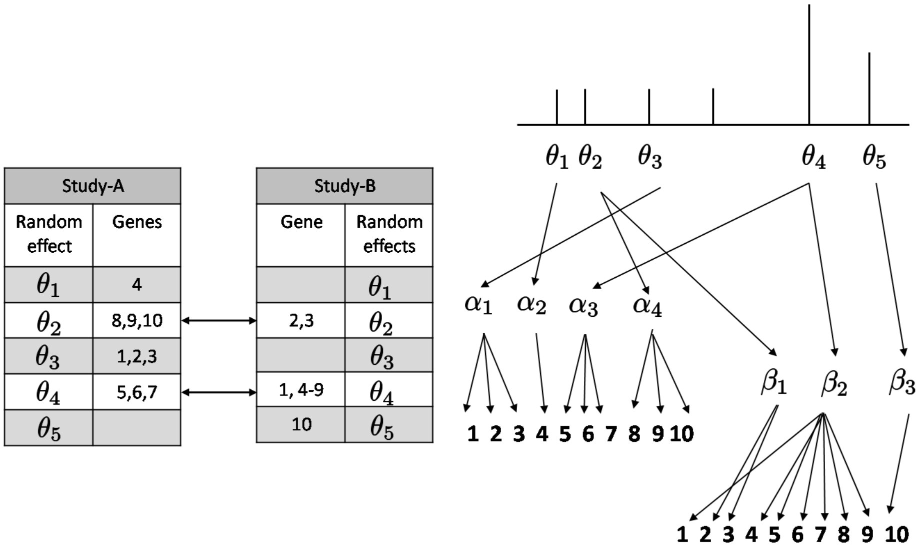
C. Prior Elicitation

For the zero-inflation parameter, we use a Beta with $a^{pr}_{p,h}$ and $b^{pr}_{p,h}$ as hyper-parameters.

$$p_{h_d} \sim \text{Beta}(a^{pr}_{p,h}, b^{pr}_{p,h}), h = 1, \ldots, n_d \tag{3.8}$$

We exploit the Hierarchical Dirichlet Process (Teh et al., 2005) prior for the regression coefficients to obtain the clusters. Assigning a HDP on the regression coefficients induce ties among them. That is, for every pair of objects $i \neq j$, there will be a positive probability that $\beta_{hi} = \beta_{hj}$. However, we do not require that $\beta_{hi} = \beta_{h^*i}$ which is typically assumed in many earlier meta-analysis approaches. If we consider the regression coefficients in a study as customers in a restaurant, then these customers will be a served a unique dish at a table. Thus, customers sitting at a table share the same dish. In regular meta-analysis models, a customer will be served the same dish in all and any restaurant. In HDP framework, a customer in a restaurant can share a different table in some other restaurant. In other words, dishes are shared across restaurants. Sharing tables across restaurants is possible because the base distribution G_h is discrete with probability one. Correlation structure and sharing of atoms is shown in Figure 7, where we consider 10 genes. Unique atoms are represented by θ, whereas α and β are the atoms in the two studies. In this example, genes $8, 9, 10$

in study-A share the same atom $\alpha_4 = \theta_2$, inducing with-in correlation. Genes $1, 2, 3$ in study-B also share the atom $\beta_1 = \theta_2$, thereby inducing both with-in and across study correlation.



Study-A	
Random effect	Genes
θ_1	4
θ_2	8,9,10
θ_3	1,2,3
θ_4	5,6,7
θ_5	

Study-B	
Gene	Random effects
	θ_1
2,3	θ_2
	θ_3
1, 4-9	θ_4
10	θ_5

θ_1 θ_2 θ_3 θ_4 θ_5
α_1 α_2 α_3 α_4
1 2 3 4 5 6 7 8 9 10
β_1 β_2 β_3
1 2 3 4 5 6 7 8 9 10

Fig. 7.: Correlation structure induced by HDP.

This is quite helpful in the context of meta-analysis because, this allows discordance between genes across studies. That is, a gene might be up-regulated in one study and can be down-regulated in some other study. If it is not done, discordant genes can wrongly borrow information and may induce false negatives or false positives depending how a gene is disagreeing across studies. We elicit Inverse-Gamma priors for the precision parameters of the HDP.

The semiparametric model is obtained by placing the HDP prior for the treatment effect's parameters:

$$\beta_{hi} \sim \mathrm{DP}\{\tau G_h\} \tag{3.9}$$

$$G_h \sim \mathrm{DP}\{\tau_0 G_0\} \tag{3.10}$$

$$G_0 = \left[\pi\delta(0) + (1-\pi)\mathcal{N}(0, \sigma_\beta^2)\right] \tag{3.11}$$

Notice that the base distribution G_o is two component mixture with a spike at with mixing probability π. The other component of the mixture is the normal distribution. Together, this is referred to as spike and slab prior. The advantage of such a specification is, multiple hypothesis testing is integrated into the modeling framework. Kim et al. (2009) used spike and slab prior as the base distribution in Dirichlet process for analyzing Microarray data-sets. Posterior inference based MCMC depends on how we interpret HDP. Just the the way DP can be represented either using Stick-breaking construction, Poly urn scheme or Chinese restaurant process, HDP can be represented with corresponding extensions. In particular, we resort to the Chinese Franchise Representation, which is simple to interpret and implement (Gerber et al., 2007). A priori, we set the probability of null hypothesis being true to π and elicit a Beta distribution to ascertain uncertainty about this parameter.

Let $\mathcal{NIG}$ be the Normal-Inverse Gamma family of conjugate distributions in which the mean has a Normal distribution conditional on the variance and the variance marginally follows an Inverse-Gamma distribution with hyper-prior parameters u and v having the appropriate subscripts. With this notation in mind, we specify the priors as:

$$\sigma_\beta^2 \sim \mathcal{IG}(u_\beta^{\mathrm{pr}}, v_\beta^{\mathrm{pr}})$$

All variance parameters are given Inverse-Gamma priors for conjugacy reasons.

$$\eta_{hi}, \sigma_\eta^2 \sim \mathcal{NIG}(\mu_h, \sigma_\eta^2, u_\eta^{\mathrm{pr}}, v_\eta^{\mathrm{pr}}) \tag{3.12}$$

$$\mu_h, \sigma_\mu^2 \sim \mathcal{NIG}(\mu_0, \sigma_\mu^2, u_\mu^{\mathrm{pr}}, v_\mu^{\mathrm{pr}}) \tag{3.13}$$

$$\sigma_{f,h}^2 \sim \mathcal{IG}(u_{f,h}^{\mathrm{pr}}, v_{f,h}^{\mathrm{pr}}) \tag{3.14}$$

$$\tag{3.15}$$

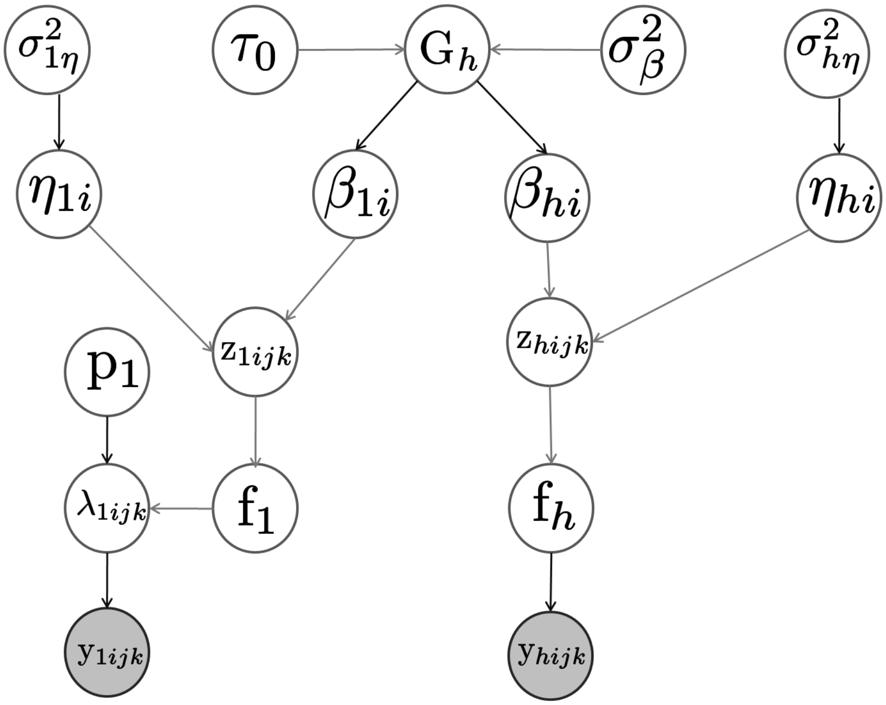


Fig. 8.: Graphical representation of the model.

A graphical representation of the model is shown in Fig. 8. Coming to the semi-parametric function, we assume that the function is smooth and can be represented by cubic B-splines (Brezger and Steiner, 2008):

$$\alpha_h \sim \mathcal{N}(0, \sigma^2_{\alpha,h}\Delta)_{I(-\infty \leq \alpha_{h,1} \leq \ldots \leq \alpha_{h,L} \leq +\infty)} \tag{3.16}$$

$$\Delta^{-1}_{L\times L} \equiv \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & 1 & & & & \\ & -1 & 2 & 1 & & & \\ & & \ddots & \ddots & & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix} \tag{3.17}$$

$$\sigma^2_{\alpha,h} \sim \mathcal{IG}(u^{\mathrm{pr}}_{\alpha,h}, v^{\mathrm{pr}}_{\alpha,h}) \tag{3.18}$$

The coefficients of the basis function are constrained to be in the increasing order. This, ensures that the function is monotonically non-decreasing. While this is not required in principle, it retains interpretability of the effects parameters.

We are using the semiparametric functions to model non-linear transformation that explain-away the technology/study specific effects and let us focus only the biological processes. However, these functions are not identifiable, by specification. For example, by scaling and shifting the latent variable, we can still obtain some other function, i.e.,

$$f(z) \quad = \quad f'(z')$$

This is the case because we do not observe z directly, which leads to non-identifiability of the function. Lack of identifiability in a Bayesian setting results in poor mixing and induces correlations among parameters. This can be resolved by constraining the functions. There are many ways to constrain the functions to make them identifiable. We constrain all the functions to be linear around the respective marginal means of the data-sets. This essentially implies that we are using the Taylor series expansion around the mean and leave the functions unconstrained at the knots far from the central knot. With out loss of generality, let us assume that the number of knots is odd. Then, we set, for all studies,

$$\alpha_{h,(L+1)/2} \quad = \quad \bar{y}_{h...} \tag{3.19}$$

$$\alpha_{h,(L+1)/2} - \alpha_{h,(L+1)/2-l} \quad = \quad d\,\sqrt{\mathrm{Var}(y_{h...})},\; l = 1, 2, \ldots, \ell \tag{3.20}$$

where d is the spacing between the equi-spaced spline knots, $\bar{y}_{h...}, \mathrm{Var}(y_{h...})$ marginal sample mean and variance of the data-set, respectively. These arbitrary constraints ensure that the latent process (z_{hijk}) has approximately zero mean and unit variance.

Here ℓ is the number of knots at which the function is constrained to be linear. Typically, we can choose ℓ such that the function is linear for about one-two standard deviations on the z_h scale. In practice, we find that the mean constraint alone is quite good enough. This completes model and prior specifications.

D. Posterior Inference

1. Prior Selection and Cluster Initialization

We elicited conjugate priors for all the variance parameters and set the corresponding hyper-parameters such that they are diffuse. As a result, posterior inference is insensitive to the choice of these hyper-priors. Nevertheless, a good initialization of the MCMC chain ensures stability and quicker convergence. As in Chapter II, we initialize the chains with empirical estimates to speed burn-in. We suggest the following procedure:

- log transform discrete data-sets and standardize all the data-sets.
- initialize the splines so that the link functions have slope corresponding to the marginal standard deviations and are centered at zero, with corresponding functional values equal to the marginal mean of the study.
- get the least-squared estimates of the treatment and signature effects parameters and their standard errors.
- use k-Means or other non-parametric clustering methods to cluster the treatment effects parameters.
- set all the random effects below a certain threshold (say, its 50th quantile) to zero, and set the latent indicator variables pointing to the null hypothesis, i.e., these genes are not differentially expressed.

Optionally, information obtained in the above steps can also be used to set the hyperprior parameters as well.

2. Sampling

We use the Gibbs sampling to approximate the posterior distribution as it is analytically intractable. Wherever the parameter sets are conditionally independent, we employ a block Gibbs sampler. Conditional distributions for λ-values and the p's, many variance parameters and the gene-specific effects are similar in form to the conditionals in Chapter II. A major difference is, we have as many sets of parameters as the number of data-sets. Specifically, a Metropolis-Hastings step with log-Normal proposal was used for drawing λ's. The smoothing spline parameters are drawn from truncated multivariate normal distributions whose truncation points ensure monotonicity of the link function. Latent process variables for each of the studies are drawn using a Metropolis-Hastings step. To sample the random effects that are given the HDP prior, augmented sampler based on Chinese Franchise representation is used. Precision parameters of the base distributions are sampled by augmenting the states so that they all have conjugate distributions. Complete details are given in Appendix B.

3. Test for Differential Expression

The MCMC sampler generates the samples for β_{hi}, the effects parameter and $\gamma_{hi} = I(\beta_{hi} \neq 0)$ from the posterior distribution. The indicator variables γ's can be readily used to estimate the posterior probability of differential expression. Let $v_{hi} = \frac{1}{B}\sum_{t=1}^{B} \gamma_{hi}^{t}$, where t is the t-th sample, and B samples are available from MCMC. Since we are testing many hypotheses simultaneously, it is necessary to control false discoveries. Under the uniform loss function, the optimal decision rule to declare a gene as differentially expressed is, given as follows:

$$\begin{aligned} &\max\nolimits_{\kappa}\ s.t && \frac{1}{\kappa}\sum_{j=1}^{\kappa} v_{(j)} \leq \alpha \\ &\text{Reject} && \kappa \text{ many genes with the smallest } v_i's \end{aligned}$$

where α is FDR set by the experimenter, typically chosen as 0.05 or 0.1 and $v_{(j)}$ is j-th ordered statistics when arranged in increasing order. In above method of testing, we have treated a gene across studies as a different entity. A much more appealing way to test for differential expression is to indicate how a particular gene is behaving across studies. For example, if a gene is up-regulated in all the studies, then we call the gene as having *concordant* differential expression. On the other hand, if a gene is differentially expressed in at least one study, and it is different at least some other study, we call the gene as having *discordant* differential expression. We define them formally as follows (Scharpf et al, 2009):

Concordant differential expression: Let $\mathcal{C}_i$ be the indicator for concordant differential expression of the i-th gene, defined as,

$$\mathcal{C}_i = \begin{cases} 1 & \text{if } N_i^{+} \times N_i^{-} = 0 \text{ and } N_i^{+} + N_i^{-} = m \\ 0 & \text{otherwise} \end{cases}$$

where N_i^{+} is the number of times the i-th gene is up-regulated, defined as $\sum_{h=1}^{N} \gamma_{hi} I(\beta_{hi} > 0)$. Similarly, N_i^{+} is the number of times the i-th gene is down-regulated, defined as $\sum_{h=1}^{N} \gamma_{hi} I(\beta_{hi} < 0)$ and m is the minimum number of studies for which the gene is differentially expressed.

Discordant differential expression: Let $\mathcal{D}_i$ be the indicator for discordant differential expression of the i-th gene, defined as,

$$\mathcal{D}_i = \begin{cases} 1 & \text{if } N_i^+ \times N_i^- \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Sample posterior means of the above indicator variables can be used in the place of γ's to control the FDR of the respective quantities. We can use these quantities to obtain the Receiver Operating Characteristic (RoC) curves to asses the performacen of the model. Area-under-RoC curve (AUC) can be used as a measure of goodness-of-fit.

4. Simulation Details

We have developed the software in MATLAB. We exploited the matrix representation and vector processing capabilities of MATLAB for accelerating the simulation time, particularly in implementing the block-Gibbs sampler. The current implementation is intended for small scale applications, for upto 500 genes and below 5 studies. The simulation time for 200 genes and 2 studies is 0.25s per MCMC iteration. As there many parameters involved, memory needed to store the intermediate results and writing them to the disk have to be carefully balanced. Holding all the parameters in chain could lead to out-of-memory problems, while writing the results to disk slows down the simulation, and potentially corrupting the disk. We devise a simple algorithm that effectively uses the memory or temporary place-holding registers. A graphical representation of updating the chains is shown in Fig. 9. Circles represent the registers that hold the states, dashed arrows show the progress of the MCMC chain and numbers in the circles show the iteration number. Contents in the dark circles are written to the file during the flush operation. In this Figure, there are four registers, and the MCMC chain is thinned by a factor of four, i.e., every fifth sample is retained excluding the boundary samples.

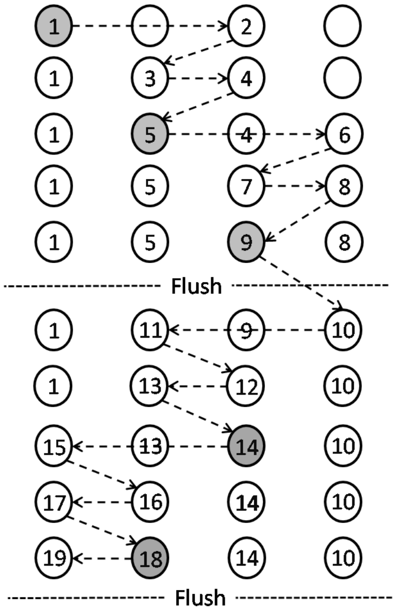


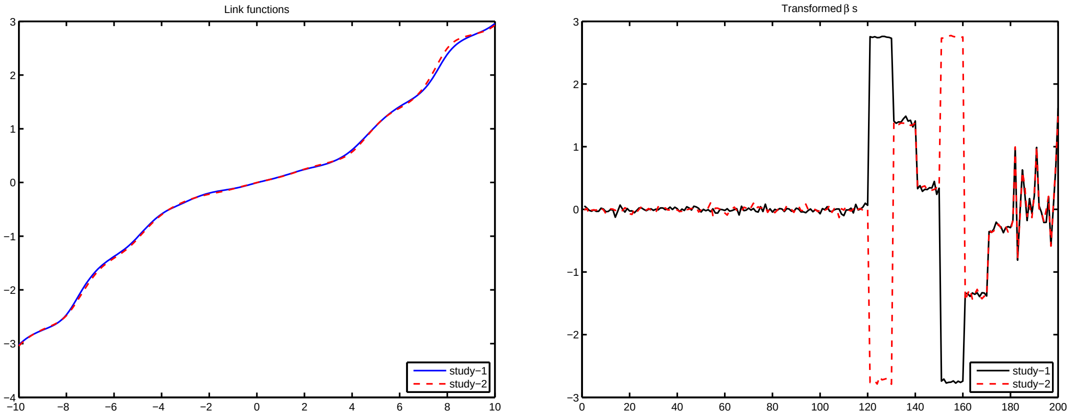
Fig. 9.: MCMC updating scheme.

E. Numerical Example

We demonstrate several features of the model using simulated a data-set, where we one study has discrete expression data and the other a continuous responses expression data, i.e., $N_d = 1$ and $N_c = 1$. While majority of the genes have concordant differential expression, some genes have discordant differential expression.

1. Simulation Settings

We generate 200 genes with 4 replicates for each gene in both of the experimental conditionals, i.e., $I = 200, K = 4$ and $J = 2$. In the first study, the true random effects are generated as follows: β_{1i} is 0 if $1 \leq i \leq 120$, 4 if $121 \leq i \leq 130$, 2 if $131 \leq i \leq 140$, 0.5 if $141 \leq i \leq 150$, -4 if $151 \leq i \leq 160$, -2 if $161 \leq i \leq 170$,-0.5 if $171 \leq i \leq 180$ and $\beta_{1i} \sim \mathcal{N}(0, 1)$ if $181 \leq i \leq 200$. In effect, we have 120 genes that belong to the null hypothesis and there are total 27 clusters. Of which, 120 genes belong to the cluster corresponding to null hypothesis. The last twenty genes are singleton clusters and the remaining six clusters are of size 10 each. For the second study, we retain the same random effects except that we swap 120th-130th effects with 150th-160th parameters. As a result, theses genes will have discordant differential expression. The gene effects η_{hi} are generated from $\mathcal{N}(0,1)$, study specific latent process, z_{hijk} from $\mathcal{N}(\eta_{[h]i} + (2\psi_{hik} - 1)\beta_{hi}, 1)$. Identity transformation is used for f_h's to generate continuous responses data $y_{hijk} \sim \mathcal{N}(f_h(z_{hijk}), 0.01)$. All the remaining variances are set to 1. At this level, we observe both continuous expression data. In Fig. 10(a), we

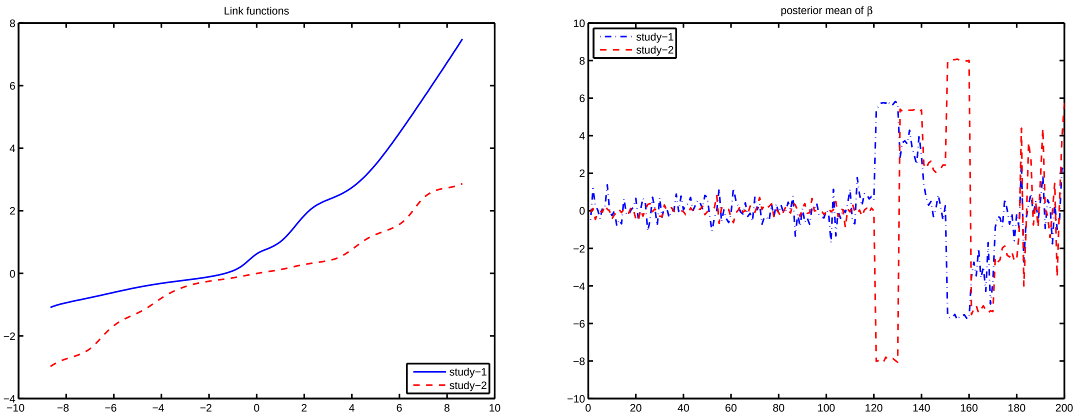


(a) f_h

(b) $f_h(\beta_{hi})$

Fig. 10.: Posterior means of the (a) link function and the (b) random effects of the genes, if both studies have continuous expression data ($N_c = 2$).

show the posterior estimate of the smoothing functions and in Fig. 10(b), we show the estimated random effects after transformation. The smoothing functions are approximately piecewise linear and the change points are approximately located where the true random effects are differing greatly (for example, near 4 and 8). We can also see that the discordant behavior in the posterior means of the random effects. In study-1, the genes 120-130 are up-regulated, while the same genes in study-2 are down-regulated. It is also important to note that, genes are sharing the clusters across studies. For example, genes 120-130 in study-1, share the same random effects with genes 150-160 in study-2. This demonstrates the how the Hierarchical Dirichlet process lets the studies share the random effects across studies and thus introduces correlation between the studies. The first study has an additional layer in the hierarchy and we generate discrete expression data as: $Y_{1,ijk} \sim ZIP(p_1, \lambda_{1,ijk})$, where $\log\lambda_{h,ijk} \equiv y_{h,ijk}$ and we set $p_1 = 0$.



(a) f_h

(b) (β_{hi})

Fig. 11.: Posterior means of the (a) link function and the (b) random effects of the genes, for mixed type expression data$(N_c = 1, N_d = 1)$.

2. Analysis

For this data-set with mixed-type responses, the observed responses appear at different levels in the hierarchy in the model specification. For example, if the expression data is discrete, there is an additional sampling variation compared to its continuous study counter part. As a result, there will be more variation in the posterior estimates of the random effects for the discrete data set. We plot the smoothing functions in

Fig. 11(a) and the estimated random effects in 11(b). It is interesting to note that the transformation function for the study-2 (continuous expression data) has lower slope compared to study-1 (discrete expression data). In other words, the random effects are shrunk so that they are shared across studies. Genes numbered 150-160 in study-1 have random effects shared with genes 160-170 in study-2. In the data-set, the true random effects are specified as integer multiples $\pm(2, 4, 8)$. Due to additional sampling variability in study-1, two-fold-changes are less likely to be detected and as a result, genes in study-1 shared random effects with genes in study-2 that are similar after transformation. Posterior mean of the random effects are transformation are shown in Fig. 12.

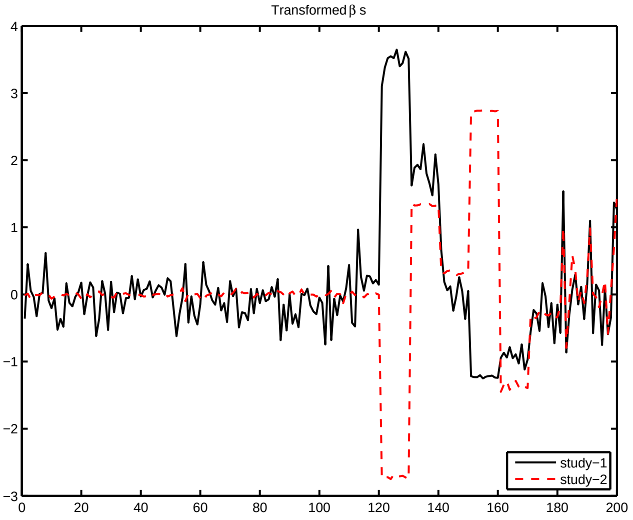
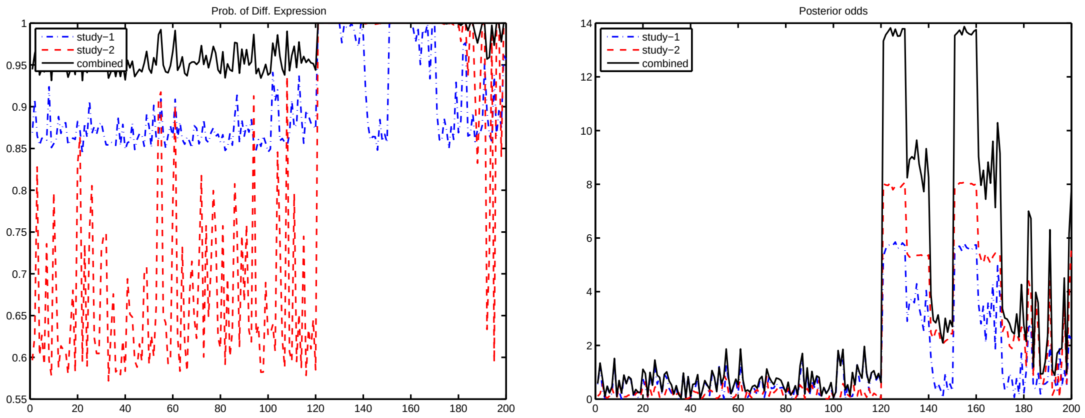


Fig. 12.: Posterior means of the random effects of the genes after transformation, for mixed type expression data ($N_c = 1, N_d = 1$).

A gene is differentially expressed either if it has discordant or concordant differential expression. The posterior probability of overall differential expression is shown in Fig. 13(a) for this example.

The posterior probability is relatively large for the genes that are simulated under



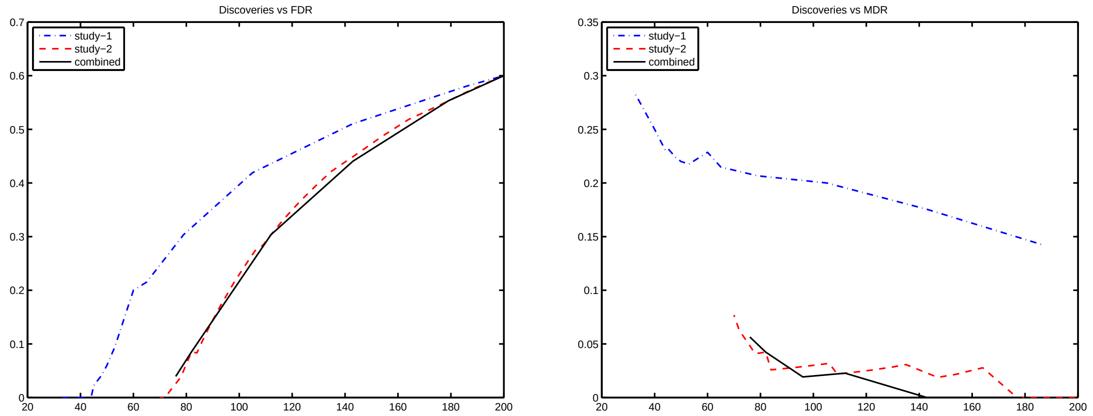
(a) 0-1 loss

(b) generic loss

Fig. 13.: Score functions to be used in Multiple hypothesis testing: (a) Probability of differential expression (b) Ratio of Posterior mean over Probability of differential expression.

the null hypothesis. This phenomena is observed in Kim et al. (2009) and Dunson et al. (2008) when using a spike-and-slab prior as a base-prior in the Dirichlet Process. They circumvent the problem by specifying informative priors, which the former term as *super sparse* prior to reduce bias in estimating FDR. However, we believe that this problem can be addressed by using the posterior means or other summaries in the loss function, while performing hypothesis testing in coherent decision theoretic set-up. For example, an approximate algorithm uses the posterior odds ratio of profits to costs (weight/value) to control FDR, shown in Fig. 13(b). Posterior mean of the random effects is the reward for discovering a gene while probability of overall differential expression is the cost (weight) in declaring the gene as differentially expressed. We pursue this problem in the next chapter in more detail. We provide several summaries of the decision process using the 0-1 loss function. As shown in Fig. 14(a) and (b), FDR and MDR (Mean Discovery Rate, defined similar to FDR; see Scharphf et al.,

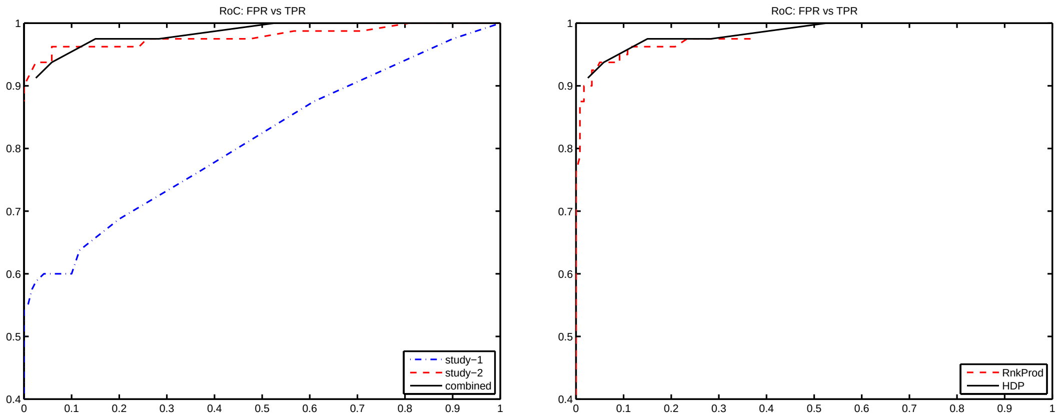
2009) are improved when the both studies are combined. The AUC for the combined data is marginally better than the AUCs of the individual studies, as evidenced in Fig. 15(a). We also report another non-parametrc method for meta-analysis in the literature, the `RankProd` (Breitling, 2004) and our method marginally performs better than the existing method; see Fig. 15(b). We point that, our estimates can be improved if we consider the generic loss functions (Fig. 13(b)). However, we are not aware of any algorithms whihc can accomplish this task and we undertake this challenge in the next chapter.



(a) Number of discoveries vs FDR

(b) Number of discoveries vs MDR

Fig. 14.: Hypothesis testing performance summaries (a) Number of discoveries vs FDR (b) Number of discoveries vs MDR.



(a) HDP

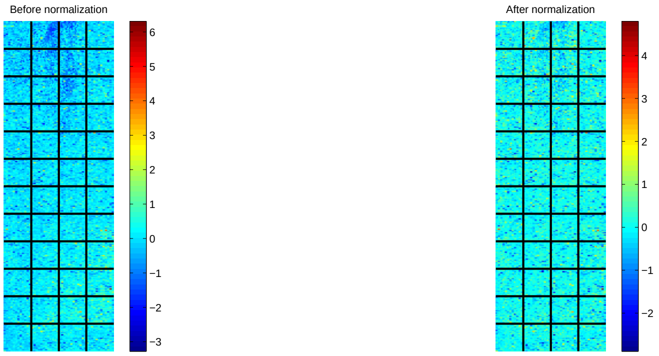
(b) HDP vs RnkProd

Fig. 15.: Receiver operating Characteristics.

F. Bovine Salmonella Microarray Data-set

The data-set was generated by Lawhon et al at Texas A&M university. Same bacterial strains and culture discussed in the previous chapter were considered, that is, derivatives of Salmonella enterica serotype Typhimurium strain ATCC 14028, IR715 and ZA21, were used in this study. Subjects were matched and the same biological samples were collected. At each of seven time points (15 min, 30 min, 1, 2, 4, 8, and 12 hours) three loops (one LB inoculated control loop, one wild type inoculated loop and one mutant inoculated loop) were excised. From each of these loops, samples were collected for histopathology, bacteriology, electron microscopy, frozen sections, and RNA extraction. Total RNA was extracted after dissection of the ileal loops obtained at 15 min and 30 min and at 1, 2, 4, 8, and 12 hours post-infection. A custom, bovine cDNA array consisting of 13,257 unique 70-mer oligonucleotides representing 12,220 cattle ORFs was designed from normalized and subtracted cattle cDNA libraries.

Microarrays were used to examine the transcriptional profiles of bovine intestinal epithelia (control and wild type or mutant infected) across seven time points (15 min, 30 min, 1, 2, 4, 8, and 12 hours). Experiments were performed in quadruplicate, generating a total of 84 arrays. cDNA from bovine experimental samples (i.e. from infected and control loops) and cDNA generated from the bovine reference RNA sample were co-hybridized to the previously described custom 13K bovine 70-mer oligoarray. The cDNA was reverse-transcribed using Superscript III reverse transcriptase and are labeled with amino-allyl-UTP. The slides were scanned using a commercial laser scanner (GenePix 4100; Axon Instruments Inc., Foster City, CA). The spots representing genes on the arrays were adjusted for background and normalized to internal controls using image analysis software (GenePixPro 4.0; Axon Instruments Inc.). Spots with fluorescent signal values below background were disregarded in all analyses. Samples were normalized against the bovine reference RNA signals across slides and within each slide (across duplicate spots). Lowess print-tip normalization was performed followed by quantile normalization (Bolsted, 2006). The effect of normalization for a specific slide is shown in Fig. 16.



(a) before

(b) after

Fig. 16.: Lowess print-tip normalization followed by quantile normalization.

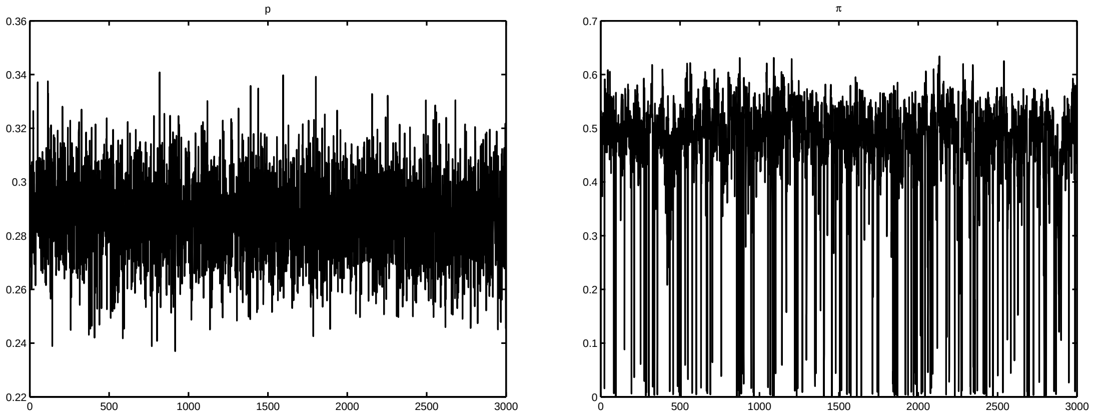
1. Merging the MPSS and Microarray data-sets

A critical step in the integrated analysis of cross-platform data-sets is to merge the data-sets and in general there is no consensus as to how to merge data-sets (Water et al 2006). For example, each platform will have different annotation updation schedules leading to discrepancy in the annotations and each manufacturer may have a different reference identifiers and sequence data bases. Adding to the complexity, in our case, both MPSS and Microarray measured mRNA much differently. In the MPSS data-set, mRNA was pooled at 30min, 1h and 4h, whereas Microarray data was collected at seven different time points including 30min, 1h and 4h. Upon consulting our collaborating biologists, we have aggregated the Microarray data 30min, 1h and 4h after appropriate normalization. Thus, a major difference between the two data sets is, MPSS measured pooled mRNA for abundance, while Microarray has pooled measurements at different time-points but not the mRNA. We conjecture that, this allows us for merging the data sets than considering individual time points. Further, Microarray probes are 70 base-pairs long, while MPSS is only 17 bp long. In order to match the tags in both the data-sets, we have used a two-level approach. First, we selected the tags (genes) that have the same probe annotation information. We subsequently performed a sequence alignment of the tags using BLAST, a sequence alignment search algorithm (Altschul et al., 1990). We selected the subset genes whose tags are perfectly aligned (matched). That is, the 17bp of a tag in MPSS has to appear exactly in the Microarray data-set. After this merging process, we have a matched data set that has 200 tags, under two experimental conditions with four replicates in each gene, under each experimental condition. Our analysis is based on this new merged data-set. Profiling the two studies in this merged data revealed that none of the genes in the Microarray are differentially expressed, while some genes are differentially expressed in the MPSS data-set. This indicates that we have discordant differential expression and both the studies are not reproducible. Nevertheless, we caution that, both studies use different technologies and we can not rule out the sensitivity of the technologies for this lack of agreement, besides many biological reasons.

2. Analysis

It is typical of MPSS data to have large number of zeros and this is the reason a zero inflation model is used for the likelihood. Marginally, the proportion of zeros in study-1 is approximately 40% and MCMC chain for the zero-inflation parameter in Fig. 17(a) is reflective of this observation. As mentioned before, Mcroarry data has no differentiallly expressed genes. In Fig. 17(b), we plot the MCMC chain for the prortion of the null hypothesises. Around 50% of genes belong to the null hypothesis and we can safely say that all of them are from the Microarray data-set (study-2). The effect of the Metropolis update step for this parameter, described in the Appendix B, can also be seen from the Figure. Occassionally, the state moves to alternate hypothesis, even though the posterior means of the random effects are practially not different from zero.

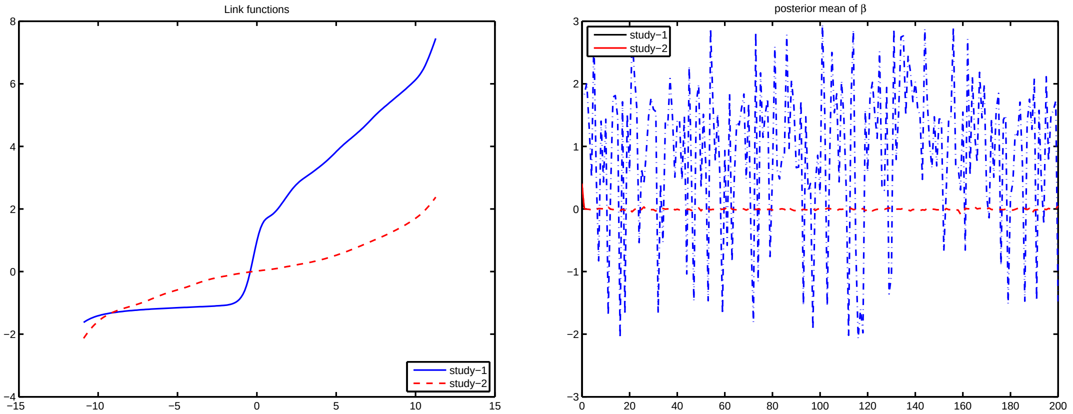
The estimated link functions are shown in Fig. 18(a). A flat segment can be seen (study-1) which is indicative of the large number of zeros in the data. Most of the interesting genes which are expressed correspond to the piecewise linear component with slope different from zero. Compared to the MPSS, the link function for the Microarray data-set is smooth. Posterior means of the random effects are shown in Fig. 18(b). It suggests that MPSS is more responsive than Microarray. It is reflected in the probability differential expression for both the studies shown in Fig. 19. While, there are no differentially expressed genes in study-1, most of the genes in Study-2



(a) p

(b) π

Fig. 17.: MCMC chain for (a) p (b) π.



(a) f_h

(b) β_{hi}

Fig. 18.: Posterior means of the (a) link function and the (b) random effects of the genes, for the Bovine Salmonella mixted-type data.

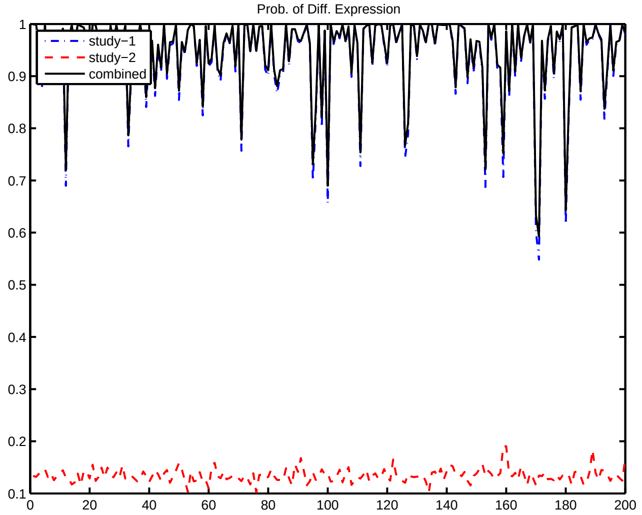


Fig. 19.: Posterior means of Probability of differential expression.

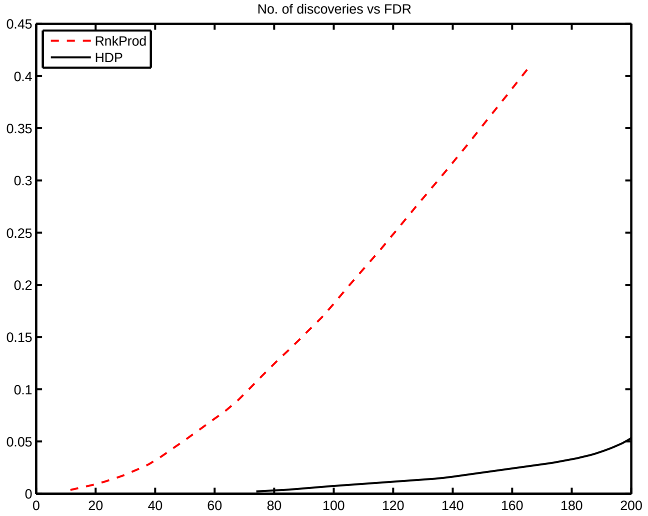


Fig. 20.: No. of Discoveries Vs FDR.

are differentially expressed. We compare the FDR using our proposed method and using the RankProd in Fig. 20. Our method produces more number of discoveries for a given FDR. However, we caution that there could be severe bias in the estimation procedure (Dudoit and van der Laan, 2008).

G. Summary

We addressed the problem of detecting differentially expressed genes from combined expression data produced by different technologies such as next-generation sequencing, SAGE, MPSS, Microarrays etc.. We had several challenges in developing a model for fusing the information from cross-platform data. Depending on the technology, the expression data can be continuous or discrete and can have different technology dependent noise characteristics. Adding to the difficulty, gene expression analysis alone poses several challenges. Notable among them is the dependency of the genes among themselves, having arbitrary correlation structure with-in and across studies. Performing several hypothesis tests for differential expression could also lead to false discoveries. Our model proposes to address all the above challenges.

We modeled the observed data using either a Poisson likelihood for count data (number of mRNA molecules) or Normal likelihood for continuous data (image intensities). Canonical link functions modeled the manifestation of latent random effects. The platform/study specific manifestation is modeled using smoothing splines. A Hierarchical Dirichlet process with Spike and Slab prior is elicited for the random effects, which models the correlation among the genes in a non-parametric fashion. Inference is carried-out based on the MCMC. Gibbs sampling forms the back-bone of the computation. We applied the model to analyze matched MPSS and Microarray data-sets obtained for understanding Bovine Salmonella infection.

CHAPTER IV

DYNAMIC PROGRAMMING APPROACH TO FALSE DISCOVERY RATE CONTROL

A. Introduction

Recent advances microarrays and other genomics technologies have spurred the interest in multiple testing procedures. Due to the nature of the problem, one has to perform potentially thousands of test simultaneously with relatively small sample sizes, often referred to as the *large p, small n* problem. It is imperative that, under these settings, having control on the number of incorrect decisions being made is of paramount importance. Naturally, many methods have been proposed to control different errors measures such as the Family-Wise Error Rate (FWER), False Discovery Rate (FDR) and positive FDR (pFDR) among others. In large-scale multiple hypothesis testing problems, the FWER is severely conservative. Benjamani and Hotcheberg, in their seminal paper (Benjamini and Hochberg, 1995), introduced FDR, defined as the expected proportion of false positives among the rejected hypothesis, which offered a practical alternative to controlling decision errors that is less conservative than FWER. Following this work, many attempts have been made to improve/extend/generalize FDR under different conditions; see for example; Genovese and Wasserman (2002), Storey (2002, 2003), Sun and Cai (2007), Tang and Zhang (2007), Effron (2007, 2008), , Ferkingstad et al. (2008), Sarkar et al (2008), Roquain and van de Wiel (2009), Chen et al. (2009). An an optimal decision process (S_{ODP}) was proposed in (Storey, 2007), which maximizes expected true positives counts while minimizing expected false positives counts, emphasizing the need to consider compound decision processes as opposed considering individual tests in isolation to others. Decision theoretic approach were also used in Mueller et al. (2004, 2007), Scott and Berger (2003), and Pena et al. (2010) to offer different insights into controlling FDR.

Guindani et al. (2009) showed that their Bayesian version of the ODP (B_{ODP}) approximates S_{ODP} closely, under a non-parametric set-up. Based on this observation, they determine the optimal decision rules by using the procedures developed for S_{ODP} in Storey (2007). They also suggest several extensions to the models by considering different loss functions, motivated from biological studies. However, while the probability model is set-up in the Bayesian framework, inference is carried in the frequentest framework. In particular, they use cluster configuration induced by the Dirichlet Process as a part of the Markov Chain Monte-Carlo (MCMC), and plug-in pooled maximum likelihood estimates in S_{ODP}. It is unclear what their motivation is to follow this circuitous route, but we suspect that, unavailability of efficient solutions to determine optimal decision rules could be a reason. This is one of the motivations for us to suggest algorithms to find optimal decision rules in a fully Bayesian way without having to bootstrap or perform grid-based searches as described in Mueller et al. (2004). Secondly, as demonstrated in Mueller et al . (2007), loss functions provide flexibility in controlling FDR, with the possibility of improving FDR in specific cases, similar in spirit to test-specific power-functions being used in the Neyman-Pearson schema of Pena et al. (2010) and Foster and Stine (2008). However, we believe that lack of efficient algorithms for completing the inference in Bayesian multiple hypothesis testing is a hurdle to be overcome, which we attempt in this chapter.

The rest of the chapter is organized as follows: In Section B, we lay down the inferential goals for multiple hypothesis testing from a Bayesian standpoint, utilizing decision theoretic framework. In the next section, we discuss the 0-1 knapsack problem, one of the well-studied combinatorial optimization problems. We recast the multiple hypothesis testing problem as an the 0-1 knapsack problem by making suitable modifications in Section D. We provide simulation examples along with a discussion in the next section. We conclude with a discussion of the proposed algorithms in Section F. Pseudocode for the FDR control is provided in the Appendix C.

B. Bayesian Multiple Hypothesis Testing

Let γ_i be the indicator variable associated with the i-th gene, with $\gamma_i = 0$ when the null hypothesis is true (for example, a gene is not differentially expressed). We use the term gene purely for historical reasons. The set-up is equally applicable to other scenarios such as testing edges in graphical models. We do not make any specific assumptions about the underlying probablity model. For example, the indicator variable used could be specified in the product form as in Scharpf et al. (2009) or as a familiar two-component mixture model under parametric setting as in Gottardo et al. (2006), or under semiparametric settings as in Kim et al. (2009). We only require some mild requirements described in Sec. 3 of Mueller et al. (2007) to carry out inferences from the posterior distribution. Our goal is to test the P hypothesis of the following form:

$$H_{0i} : \gamma_i = 0 \text{ Vs } H_{1i} : \gamma_i = 1 \ \forall i = 1, 2, \ldots, P$$

Let $d_i = 1$ ($d_i = 0$) be decition to reject (fail to reject) the null. Then the outcomes from the above hypothesis test can be summarized in the following table:

	Null true	Alternate true	Total
Accept null	U	T	P-R
Reject null	V	S	R $= \sum d_i$
Total	$P - m_1$	$m_1 = \sum \gamma_i$	P

Clearly, we need to control the number of incorrect decisions. Some commonly used measures are:

$$\begin{aligned} \text{FWER} &= E(V > 1) & (4.1) \\ \text{FDR} &= E(V/R|R > 1) & (4.2) \\ \text{pFDR} &= E(V/R)E(R > 1) & (4.3) \\ \text{FNR} &= E(T/P - R) & (4.4) \end{aligned}$$

among others and the overall goal is to simultaneously minimize a function of V and T, and maximize a function of U and S. A reasonable way to trade-off these conflicting goals is to, maximize U(or S) while keeping V(or T) within manageable limits specified in terms of the error measures defined above. Of course, we do not know them in reality, so we work with their expected values instead. In the above error measures, we penalized incorrect decisions and rewarded correct decisions in every test the same way, without regard to the type of decision or the test-specific marginal posterior summaries. This implies that, we assumed a uniform loss/reward function for the decisions. In the Bayesian context, we can formalize these ideas using the decision theoretic approaches. Under general settings,

$$L_0(.|d_i = 0) = \begin{cases} f_{00}(.) & \text{if Null true} \\ f_{01}(.) & \text{if Alternative true} \end{cases} \tag{4.5}$$

$$L_1(.|d_i = 1) = \begin{cases} f_{10}(.) & \text{if Null true} \\ f_{11}(.) & \text{if Alternative true} \end{cases} \tag{4.6}$$

where L_0 is the loss when the decision is to fail to reject the null and L_1 is the loss incurred when the null is rejected. The specification of the loss functions gives the flexibility in rewarding some genes differently than others. Typically, one assumes a uniform loss function in the absence of any specific information. That is, if we choose, $f_{00} = f_{01} = 0$ and $f_{10} = \lambda, f_{11} = -1$, we are essentially maximizing the true positives, while keeping the false negatives below certain value. The expected posterior loss in this case is given as:

$$-\sum_{i=1}^{I} d_i v_i + \lambda \sum_{i=1}^{I} d_i(1 - v_i) \tag{4.7}$$

where $v_i = E[\gamma_i]$. The optimal decision sequence d^* minimizes the expected posterior loss and λ shall be caliberated so that the estimated FDR is below the user-specified bound. Under the above settings, one can find optimal solution easily, given by:

$$\max_\kappa s.t \quad \frac{1}{\kappa}\sum_{j=1}^{\kappa}(1 - v_{(j)}) \leq \alpha \tag{4.8}$$

$$\text{Reject} \quad \kappa \text{ many genes with the largest } v_i's \tag{4.9}$$

The solution can determined easily because, the odds ratio $\frac{v_i}{1-v_i}$ is monotonic in the weights. That is, if $v_i > v_{i'}$, then $\frac{v_i}{1-v_i} > \frac{v_{i'}}{1-v_{i'}}$. Instead, if one chooses gene dependent loss functions which reward true positives in a non-uniform fashion, then the posterior expected loss is given by:

$$-\sum_{i=1}^{I} d_i v_i f_i^* + \lambda \sum_{i=1}^{I} d_i(1 - v_i) \tag{4.10}$$

where $f_i^* = E_i(f_{11})$, is the expectation with respect to the posterior marginal distribution of the i-th gene. We can make the dependence of the decisions on the false discovery rate (or some other measure) explicit by re-expressing the above objective function as:

$$\arg\max_{d*} \quad \sum_{i=1}^{I} d_i v_i f_i^* \tag{4.11}$$

$$s.t \quad \sum_{i=1}^{I} d_i(1 - v_i) \leq \sum_{i=1}^{I} d_i \alpha \tag{4.12}$$

where, $\sum_{i=1}^{I} \ d_i(1 - v_i)$ is the expected false positives and $\sum_{i=1}^{I} \ d_i v_i$ is the expected true positives and α is the desired FDR. By the way, we still did not specify what f_i^* should be but just said that f_i^* is not a constant anymore. Consequently, we may loose monotonicity of the odds ratio, which in this case is, $\frac{v_i f_i^*}{1-v_i}$. As a result, the solution for the uniform loss function is no longer optimal. One could perform grid-based searches Mueller et al. (2004) or choose the thresholds Scott and Berger (2003). Our question is, can we obtain the optimal solutions in this case? What we mean by a solution is, finding the optimal decision sequence without resorting to bootstrapping or grid-searches or such approximation techniques. One could pretend that the odds ratio is monotonic and employ the global thresholding techniques which are greedy, but we only find a suboptimal solution. In the next section, we briefly discuss the knapsack problem that has striking similarities to the problem at hand.

C. The 0-1 Knapsack Problem

Consider again P genes with the i-th gene having a cost w_i (a positive integer) and profit v_i(a non-negative number). Let C (a non-negative integer) be the capacity of the knapsack. Our goal is to fill a knapsack with as many genes as possible maximizing profit but not fill the knapsack beyond its capacity, thus keeping the cost below a threshold. Stated formally, the objective is to (Kellerer et al. (2004):

$$\arg\max_{d*} \quad \sum_{i=1}^{I} d_i v_i \tag{4.13}$$

$$\text{s.t} \quad \sum_{i=1}^{I} d_i w_i \leq C \tag{4.14}$$

The formulation above is, in spirit, the same as α-investing in Foster and Stine (2008) and ideas in Pena et al. (2010), who consider multiple hypothesis testing as a resource allocation problem. We want to maximize profits (true positives) while keeping the costs (false positives) down. These ideas have lead us to consider multiple hypothesis testing from an Operations Research (OR) perspective. Since our decision (or action) space is the collection of all P-tuples, it is essentially a combinatorial optimization problem (Korte and Vygen, 2008). For large P, obtaining the optimal solution by enumerating 2^P possible solutions is NP-hard and therefore is non-trivial. In the above problem, suppose that

$$\frac{v_1}{w_1} < \frac{v_2}{w_2} < ... < \frac{v_P}{w_P}. \tag{4.15}$$

That is, the profit/cost is an increasing function w.r.t the cost, then the optimal strategy to fill the knapsack is to simply pack all genes with the lowest cost first, until the capacity is reached . This is in fact the same solution we got in the previous section for the multiple hypothesis testing problem with uniform loss function, where we have monotonicity for the odds ratio. This version of the solution is called a greedy method in the OR literature. Even when the profit/cost does not have any specific pattern, the knapsack problem has some recurring substructures whose optimal solutions can obtained efficiently. Key features of the knapsack problem are:

- optimal substructure: an optimal solution to the problem contains within it optimal solutions to subproblem.

- overlapping substructures: some subproblems will be visited again and again.

More specifically, let `KP(i,c)` denote the optimal solution for the above problem. Then,

- If $d_P = 0$, that is if we do not place P-th gene in the knapsack, then, $d_1, d_2, \ldots, d_{P-1}$ must be the optimal solution for the problem `KP(P-1,c)`
- If $d_P = 1$, then, $d_1, d_2, \ldots, d_{P-1}$ must be the optimal solution for the problem `KP(P-1,c-`w_M`)`

More specifically,

$$\texttt{KP[i,c]} \;=\; \max(\texttt{KP[i-1,c]}, \texttt{KP[i-1,c-}w_i\texttt{]} + v_i) \tag{4.16}$$

The table `KP[,]` contains all the information to determine the optimal decisions for any given capacity not exceeding C. Pseudo code for the completing the table is given in the Appendix C. Optimal state with the maximum profit for the given capacity is obtained by traversing the table in a specific manner. Suppose we set i=P and c=C-1 in Algorithm-2 given in the Appendix, we would get the optimal decision sequence with capacity bounded by C-1. Dynamic Programming (DP) principles help us to generate uniformly optimal sequential decisions for all capacities. In other words, if we we have the table `KP` computed for `KP(P,C-1)`, we only need to update the table by adding a column to it, without changing the first $C-1$ columns. This version of the knapsack is known as all-capacity knapsack and DP solves the all-capacity knapsack using the same resources (time and memory) as it takes for the knapsack with the largest capacity among them. This feature is particularly helpful to report decisions, profits and costs for a range of capacities. For a comprehensive review of knapsack problems, refer Kellerer et al. (2004), Korte and Vygen (2008). In the next section, we apply the above principles to the Bayesian multiple hypothesis testing.

D. Multiple Hypothesis Testing as a 0-1 Knapsack Problem

For illustration purposes, consider the uniform loss case. By setting

$$\begin{aligned} v_i &= P(\gamma_i = 1) \\ w_i &= P(\gamma_i = 0) = 1 - v_i \end{aligned}$$

in the knapsack problem, we can obtain the optimal decision region but there is technical difficulty. First, we do not know $P(\gamma_i = 1)$, so we replace it by its posterior estimate, i.e.

$$\begin{aligned} v_i &= \hat{P}(\gamma_i = 1) \\ w_i &= \hat{P}(\gamma_i = 0) = 1 - v_i \end{aligned}$$

However, we require the costs to be positive integers and profits be non-negative. If we have genes with zero costs, the optimal strategy is to always pack them. Therefore, we will declare all genes with zero costs, irrespective of the profits, as differentially expressed. To apply the knapsack framework for the remaining genes, recognize that these estimated proportions ($\hat{P}(\gamma_i = 0)$) are rational numbers because we use B number of samples obtained using MCMC or some other inference engine. In other words, $\hat{P}(\gamma_i = 0) = \frac{1}{B}\sum_b \gamma_i^{(b)}$ and we actually have the costs specified as positive integers by construction:

Null true	Alternate true	
$B - X_i$	$X_i = \sum_b \gamma_i^{(b)}$	B

Therefore, if we set,

$$v_i = X_i, w_i = B - X_i$$

we can use DP to find the optimal decisions. False discovery rate can be estimated at the optimal decisions as:

$$\hat{\alpha} = \frac{1}{\sum_i d_i} \sum_i d_i(1 - v_i) = \frac{1}{B \sum_i d_i} \sum_i d_i w_i \tag{4.17}$$

It is nothing but the average cost of the genes in the knapsack scaled by the number of posterior samples used. We state the connection between Bayesian multiple hypothesis with generic loss functions and the 0-1 knapsack problem as follows:

Proposition:. *For any loss functions f_{01}, f_{10} with non-negative integer valued posterior expectations penalizing incorrect decisions and for any loss functions f_{00}, f_{11} with non-negative posterior expectations rewarding correct decisions, an optimal decision sequence can be obtained that maximizes the profts not the exceeding given loss bounded by C, using dynamic programming, with worst case complexity $O(CP)$.*

The requirements layed out in the above proposition are not as restrictive as they appear to be. One could simply set $f_{00} = 0, f_{01} = 0, f_{10} = 1$ and bring flexibility be altering $f_{11} > 0$. This allows one to control an aspect of the true positives but keeps a bound on the false positives. It is based on the same reasoning for constraining the costs and profits in the knapsack problem to be non-negative. For example, if both are negative, we can consider them as positive but invert the decisions. These technical requirements ensure that the algorithm is simple to manage and maintain. It does not take away flexibility in designing sensible loss functions. By no means, the framework is restricted to controlling FDR. For example, consider $f_{00} = -\lambda, f_{01} = 1, f_{10} = 0, f_{11} = 0$, then we would be maximizing true negatives while keeping a bound on the false negatives. A complete algorithm in the form of pseudo code is given in Appendix C which reports the optimal decision for a given FDR.

E. Simulation Examples and Discussion

Example-1: Let us begin with a version of the simple example considered in pp. 16, Kellerer et al. (2004). We have seven genes and the knapsack's capacity is 0.9. Costs and profits of the genes are given below:

i	1	2	3	4	5	6	7
v_i	0.6	0.5	0.3	0.6	0.8	0.9	0.7
w_i	0.2	0.3	0.4	0.5	0.6	0.7	0.9
$\frac{v_i}{w_i}$	3	1.67	0.75	1.2	1.33	1.29	0.78

When using the greedy approach, we pick the gene with largest profit/cost first and repeat this until the knapsack reaches its capacity or we exhaust the genes. If a gene does not fit into knapsack, we simply skip and go to the next gene. So, we pick the 1st, 2nd and the 3rd genes that exactly fill the knapsack. Using DP, we pick the 1st and the 6th genes. For the given capacity, the optimal solution had total value 1.5, while the greedy had 1.4. The greedy solution would have been the optimal solution, had the weights been ordered according to profit/cost.

Example-2: We simulate P=100 genes. To avoid any model specific assumptions and inferential goals, we generate γ_i i.i.d Beta$(3, 1)$ with mean 0.75. Then, for each gene i, we generate B=100 Bernouli random variables with success probability γ_i, i.e., $\gamma_i^{(b)}$ i.i.d Bin$(1, \gamma_i)$. The costs assigned to the genes in the knapsack are $w_i = B - X_i = B - \sum_b \gamma_i^{(b)}$ with corresponding profits $v_i = \frac{X_i}{B}$. In this case, the greedy solution and the dynamic approach should produce identical results because the odds ratio or profit/cost is monotonic as shown in Fig. 21 (solid).

We plot FDR vs the number of discoveries made (total number genes in the knapsack) in Fig. 22(a),. Both DP and greedy algorithms produce identical results. Estimated FDR is plotted against the estimated TDR (true discovery rate, which is

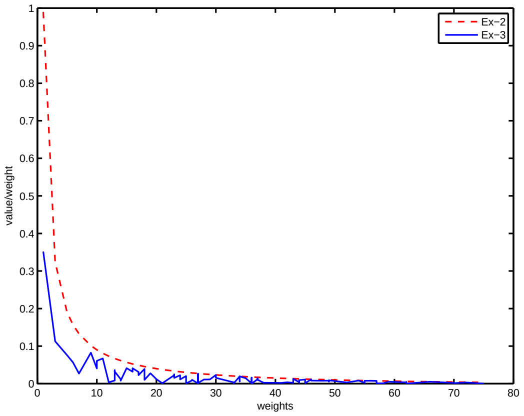
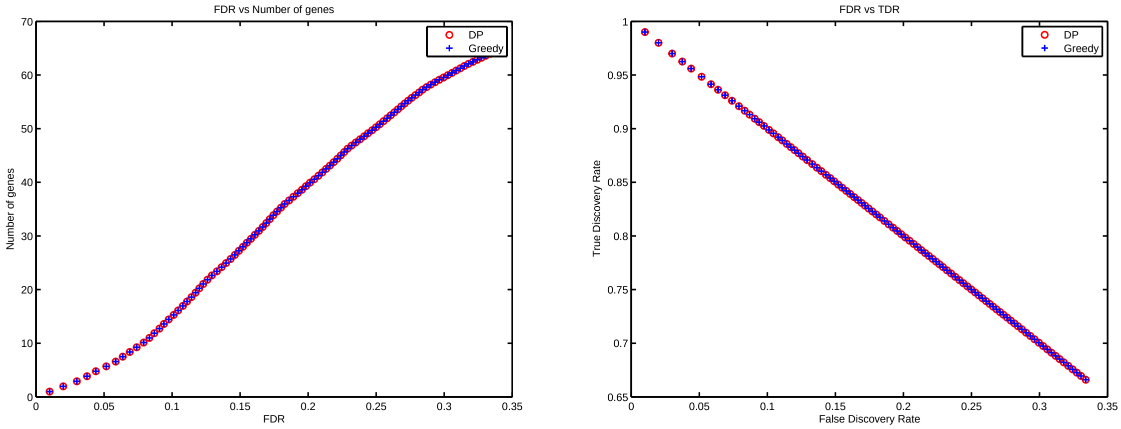


Fig. 21.: Profit/cost in Example-2 (solid), Example-3(dashed).

the average profit of genes in the knapsack) in Fig. 22(b). It is a decreasing function of FDR because as we increase the capacity, we add more genes having smaller profits which bring down the average profit. If profit/cost is monotonically decreasing with cost, then we expect TDR to drop down monotonically as well. Total capacity vs total profit of the genes in the knapsack is shown in Fig. 23(a) and as expected profit increases with capacity. We see from Fig. 23(b) that number of genes in the knapsack increase as a the capacity is increased.

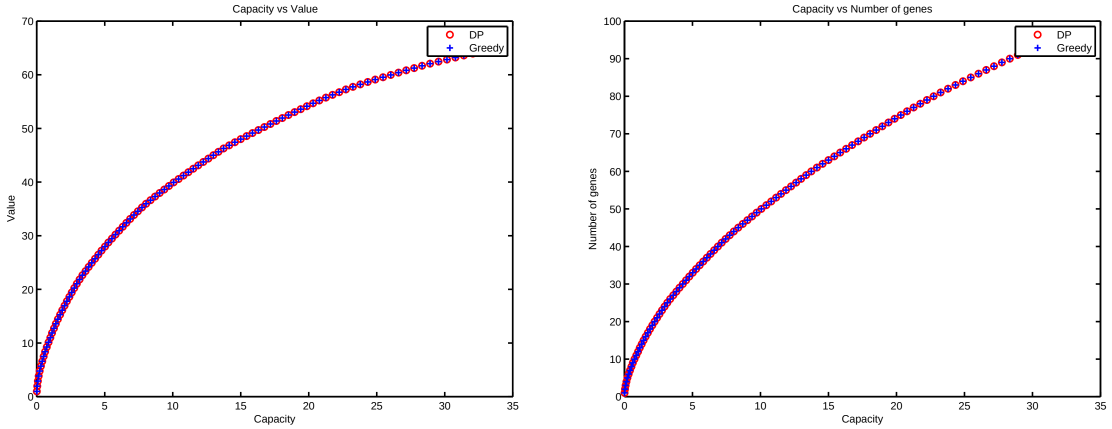
Example-3: To simulate a more complex scenario that could be applicable with generic loss functions, we break the monotonicity of profit/cost by perturbing the profits randomly. The new profits are generated by multiplying the previously assigned profits with random weights drawn from U(0,1), i.e., $v_i^* = u_i * v_i, u_i \sim \mathrm{U}(0,1)$. The resulting profit/cost ratio is shown Fig. 21 (dotted). We again run the greedy algorithm and DP algorithms with v_i^* and w_is. In Fig. 24(a), FDR vs the number of discoveries is plotted. Both DP and Greedy algorithms produce similar results. But looking at the estimated FDR vs the estimated TDR plot in Fig. 24(b), it is clear that greedy algorithm has lower TDR for a given FDR than the DP solution. That is, the DP knapsack has higher average profit than the greedy knapsack. It is worth



(a) FDR vs Number of discoveries

(b) FDR vs TDR

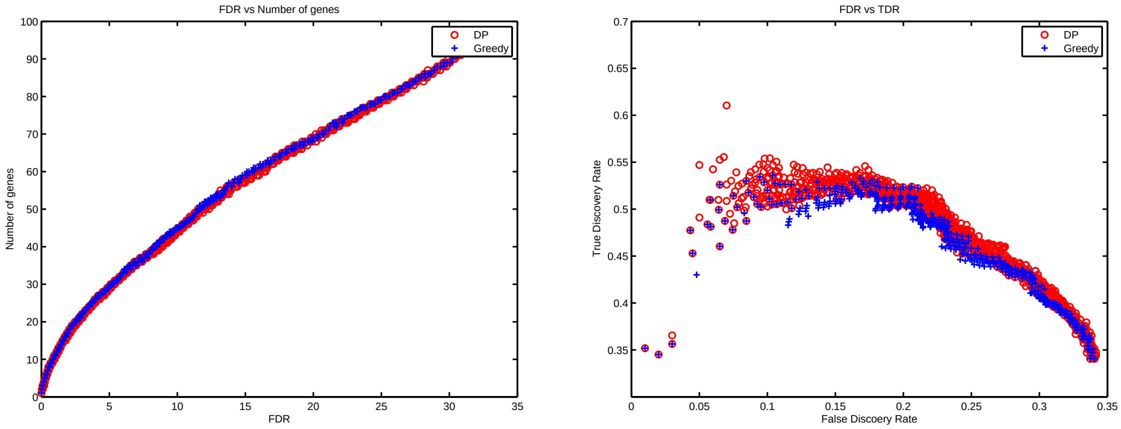
Fig. 22.: FDR, TDR summaries for genes in Example-2 using DP(o) and Greedy(+) algorithms.



(a) Total cost vs Total profit

(b) Total capacity vs Number of genes

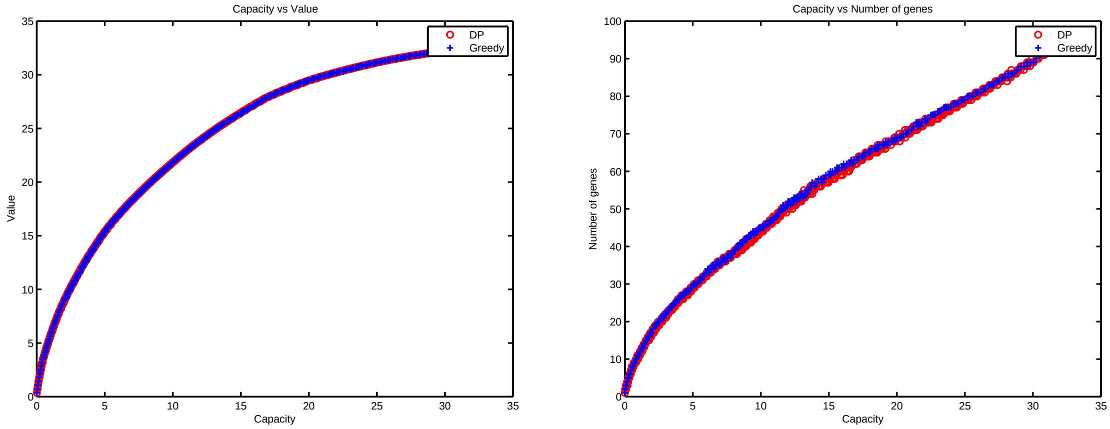
Fig. 23.: Cost, Capacity and Profit summaries of genes the knapsack for Example-2 using DP(o) and Greedy(+) algorithms.



(a) FDR vs Number of discoveries

(b) FDR vs TDR

Fig. 24.: FDR, TDR summaries for genes in Example-3 using DP(o) and Greedy(+) algorithms.



(a) Total cost vs Total profit

(b) Total capacity vs Number of genes

Fig. 25.: Cost, Capacity and Profit summaries of genes the knapsack for Example-3 using DP(o) and Greedy(+) algorithms.

noting that, TDR increases in the beginning with FDR and then falls-off. This is a consequence of breaking the profit/cost ordering. Total capacity vs total profit of the genes in the knapsack is shown in Fig. 25(a), which exhibits characteristics similar to Example-2. Likewise, Total capacity increases with as the number of genes in the knapsack increase, as shown Fig. 25(b). Both the simulation examples show that DP framework can be used to solve multiple hypothesis testing problems under different scenarios.

Discussion: An important component of these combinatorial optimization problems is the time and space complexity. The 0-1 knapsack problem is a psedu-polynomial time algorithm in the sense that the complexity depends on capacity and it not bounded asymptotically. In the multiple hypothesis testing framework, this translates to how many posterior are samples needed to compute the optimal solution in a reasonable amount of time. A conservative estimate for B can be $\left\lceil \sqrt{\frac{1}{4e}} \right\rceil$, where e is the Monte-Carlo standard error for estimating the posterior probabilities $P(\gamma_i)$ assuming i.i.d samples. A loose bound for the capacity can now be given as $C \leq P \sum X_i$ (note that $X_i \leq B$) and the worst case complexity for the computing the table KP is $O(P^2B)$. Often in practice, the desired FDR will be attained at a much lower capacity and finding tighter bounds is a topic for future research (Martello et al. 1999). We point that computing the table has lower complexity than determining the optimal decision sequence. This makes computing the FDR at every capacity, which is based optimal decision sequence at that capacity, impractical. However, it is possible to reduce this time by employing a leap-and-bound strategy (not implemented). That is, we estimate the FDRs at certain increments of the capacity and periodically check if the FDR has exceeded. While we are not recommending any particular strategies, there are many choices for customizing the algorithm for the problem at hand and this is an active area of research on it own (Martello et al., 2000). Several hard to solve knapsack problems are discussed in Pisinger, (2003).

F. Conclusion

We considered multiple hypothesis in the Bayesian context without making any specific assumptions about the underlying probability model. Assuming that there is an efficient mechanism to generate posterior samples for the indicator variables for the individual tests, we showed how the knapsack problem, a combinatorial optimization framework, can be adopted for the problem at hand. We provided an algorithm for maximizing true discoveries while keeping the FDR below the bound. Our approach solves the problem in the cohesive decision theoretic setup. We believe that discrete optimization is a feasible solution in a variety of multiple hypothesis testing scenarios, which can be solved exactly. We highlighted several features of the knapsack framework that could lend insights into the multiple hypothesis testing problem. In the event that capacity bounds are enormously large, we can leverage several approximate algorithms available in the knapsack problem literature. We hope that our contribution stimulates research to find better solutions.

CHAPTER V

CONCLUSIONS AND SCOPE FOR FUTURE RESEARCH

A. Summary

We have considered several models for analyzing mixed type expression data. Our primary goal in all the approaches is to discover differentially expressed genes. There exist numerous statistical methods for analyzing Microarray data. Partly because, Microarray is inexpensive relative to next generation sequencing technologies. In the first two chapters, we focused on developing methods for analyzing count data.

In Chapter II, we reviewed literature for analyzing discrete expression data and demonstrated that Bayesian hierarchical models work better than traditional tests which model each gene individually. We extend the idea by relaxing the parametric assumptions for the random effects by using Dirichlet process prior which clusters genes that share similar differential expression profiles. We assumed homoscedastic variance for the measurement error for all genes. Over dispersion, typically in discrete expression data due the presence of large number of zeros, is addressed by using a zero-inflated Poisson likelihood. We used the q-value approach to control false discovery rate on the p-values obtained by performing Kullback-Liebler distance based hypothesis test for differential expression of the random effects.

In the next chapter, we focused on developing methods for combining data from a multitude of sources. The goal is to fuse information to make better predictions or estimate the parameters more accurately. Meta-analysis in genome-wide studies is still a nascent area and there do not exist any methods that combine mixed-type expression data. We devised a model that combines mixed type data. We do not require that the studies being pooled are are reproducible. In other words, if a gene is up-regulated in one study, we do not expect it to be up-regulated with similar magnitude in the remaining studies as well. We use non-parametric functions to account for technological differences and/or study dependent normalizations. Model developed in Chapter II forms the backbone of the inference engine but instead of using a Dirichlet process with normal base prior, we use hierarchical Dirichlet process prior with spike-n-slab base prior. Thereby testing for differential expression is integrated into the model. We also allow the variance to be gene dependent to make the model more flexible.

In Chapter IV, we formulated multiple hypothesis testing as a resource allocation problem. When indicator variables denoting the state of the hypothesis are available as component of the inference, multiple hypothesis testing is akin to a combinatorial optimization problem. This formulation provides optimal solutions for generic loss functions including the widely used uniform loss function. We give several pointers which could be helpful in expanding this new approach to FDR control. Due to the coherent decision theoretic set-up used, the same formulation can be used to control not just FDR but similar error measures.

B. Future Research

In all the models we developed, we assumed that the sample size is the same across experimental conditions and across studies. The models can be easily extended to handle unbalanced case, but with some involved algebra. We did not use any covariate information in the ANOVA specification. The proposed models can be extended to incorporate covariate information with out much difficulty. Another extension possible is to cluster variances as well. So far in our models, we either assumed homoscedastic variance or considered them to be different for each gene/study/sample. However, they can also estimated accurately by using Dirichlet Process prior. The above mentioned extensions do not involve methodological research but just extensions to suit the real world problems.

In Chapter II, we considered saturate ANOVA representation to primarily consider pairwise comparisons when two or more experimental conditions exist. The model can be simplified and can design efficient samplers if there are only two experimental conditions or if one knows that one of the experimental condition is a reference/control treatment group. Further, it would not be very difficult place a spike-n-slab prior in place of the normal base prior. We used partial hierarchical centering for prior specification. The sampling efficiency can be improved by using more complex algorithms like adaptive MCMC type algorithms or covariance adjusted sampler to reduce posterior correlation and improve mixing.

On the contrary, we have considered two-sample comparisons in Chapter III mainly to expose the features of the model. It would be interesting to extend the model to handle more than two treatment groups, but nevertheless, the algebra is little more involved. A main relaxation we made in this model is that studies need not be reproducible. But in cases where studies are reproducible, our modeling looses power compared to a method which coerces positive correlation among the random effects across studies. The following modeling might help in this situation:

Let Y_{hijk} be the k^{th} replicate observed for the i^{th} signature under the j^{th} treatment in study h. We assume exponential family for the likelihood.

$$Y_{hijk} \sim F_{[h]}(x_{hijk})$$

We elicit zero-inflated Poisson likelihood for discrete expression data like SAGE, MPSS, next-generation sequencing etc.. and Normal likelhood for microarrays. That is,

$$Y_{[h],ijk} \sim ZIP(p_h, \lambda_{h,ijk}),\, h = 1, 2, ..., n_d$$

where n_d is the number of discrete data-sets, p_h is the study-specific zero-inflation parameter and $\lambda_{h,ijk}$ is the Poisson mean parameter. We use canonical link functions for linking the latent random variables which different technologies measure. For discrete-data sets,

$$\log\lambda_{h,ijk} \equiv y_{h,ijk}\, h = 1, 2, ..., n_d$$

and for continous data-sets, we use identity link function, i.e.,

$$Y_{hijk} \equiv y_{hijk},\, h = n_d + 1, n_d + 2, ..., n_d + n_c$$

where n_c is the number of continous data-sets.

Now, each study measures a version of the latent random effects. We use a semiparametric function to model this manifestation which is different for each study/data-set.

$$\begin{aligned} y_{hijk} &\sim \mathcal{N}(f_{[h]}(z_{ijk}), \sigma^2_{f,h}),\, h = 1, 2, ..., n_d + n_c \\ f_{[h]}(.) &= \sum_{l=1}^{L} \alpha_{h,l}\Psi_{h,l}(.) \end{aligned}$$

Here, z_{ijk} is the latent process which captures the differential gene-expression profiles and $f_{[h]}$ is the function that maps the latent process which is then measured by different technologies, $\Psi_{h,l}$ are the basis functions which we choose as cubic B-splines in this case and $\sigma^2_{f,h}$ is the study specific variance.

We model the latent random effects in the following fashion,

$$z_{ijk} \sim \mathcal{N}(\eta_i + \beta_{ij}, \sigma^2_{z,i})$$

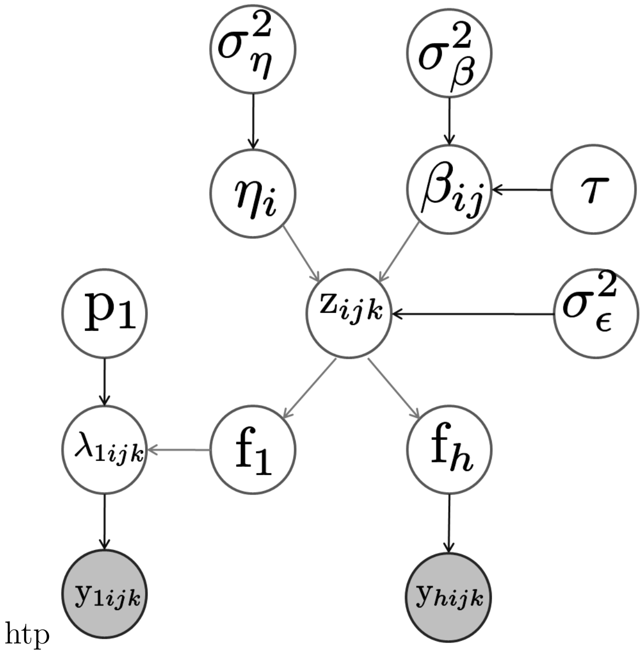


Fig. 26.: Graphical Representation of the model.

where where η_i is the effect of the i^{th} signature and β_{ij} is the effect of the j^{th} treatment nested within the i^{th} signature. That completes model specification except the priors. A graphical representation is given the following Fig. 26.

We elicit Dirichlet process for on the random effects which are common for all studies as:

$$\beta_{ij} \quad \sim \quad \mathrm{DP}\{\tau G_0\}$$

where τ is the tuning parameter and the baseline distribution is G_0 with

$$G_0 \quad = \quad \left[\pi \delta_{\forall j}(\beta_{ij} = \beta^{\mathrm{null}})\mathcal{N}(0, \sigma^2_{\beta,\mathrm{null}}) + (1-\pi)\prod_{j=1}^{J}\mathcal{N}(0, \sigma^2_{\beta,j})\right]$$

This means that, under the null hypothesis, all treatment effect parameters are identical. Apriori, we set the probability of null hypothesis being true to π and elicit a Beta distribution to ascertain uncertainty about this parameter. Note that here we did not restrict ourselves to two sample case, but more than two experimental conditions. We are working on this model.

In Chapter IV, we established the connection between Bayesian multiple hypothesis testing and the 0-1 knapsack problem, but have not taken advantage of this approach. It would be interesting to consider, gene specific loss functions in the hope to offer better control over FDR or other error measures.

REFERENCES

Agarwal, D. K., Gelfand, A., and Citron-Pousty, S. (2002), "Zero-inflated models with applications to count data," *Environmental and Ecological Statistics*, 9, 341-355.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990), "Basic local alignment search tool," *Journal of Molecular Biololgy*, 215, 403-410.

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to nonparametric problems," *Annals of Statistics*, 2, 1152-1174.

Audic, S., and Claverie, J. (1997), "The significance of digital gene expression profiles," *Genome Research.*, 7, 986-995.

Barkhausen, T., Krettek, C., and van Griensven, M. (2005), "L-selectin: adhesion, signalling and its importance in pathologic post-traumatic endotoxemia and nonseptic inflammation," *Experimental and Toxicologic Pathology*, 57, 39-52.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, Series B, 57, 289-300.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160-169.

Bochkina, N., and Richardson, S. (2006), "Tail posterior probability for inference in pairwise and multiclass gene expression data," *Biometrics*, 63, 1117-1125.

Bolstad, B. M. (2006), "Pre-processing Microarray Data," in *Fundamentals of Data Mining for Genomics and Proteomics*, eds. W. Dubitzky, M. Granzow, and D. P. Berrar, Springer, New York, NY.

Boughammoura, A., Franza, T., Dellagi A., Roux, C., R., Berthold, M.-M., and Expert, D. (2007), "Ferritins, bacterial virulence and plant defense," *Biometals*, 20, 347-353.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004), "Rank Products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Letter*, 573, 83-92.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. J., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao. J., and Corcoran, K. (2000), "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays," *Nature Biotechnology*, 18, 630-634.

Breslow, N., and Clayton, D. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9-25.

Brezger, A., and Steiner, W. J. (2008), "Monotonic Regression Based on Bayesian P-splines: An Application to Estimating Price Response Functions From Store-Level Scanner Data," *Journal of Business and Economic Statistics*, 26, 90-104.

Carota, C., and Parmigiani, G. (2002), "Semiparametric regression for count-data," *Biometrika*, 89, 251-263.

Chen, C., Cohen, A., and Sackrowitz, H. B. (2009), "Admissible, consistent multiple testing with applications including variable selection," *Electronic Journal of Statistics*, 3, 633-650.

Conlon, E. M., Song, J. J., and Liu, J. (2006), "Bayesian Models for Pooling Microarray Studies With Multiple Sources of Replications," *BMC Bioinformatics*, 7, 247.

Consortium, M. A. Q. C., Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao,W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Hui Fan, X., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q.-Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W. (2006), The MicroArray Quality Control (MAQC) Project Shows Interand Intraplatform Reproducibility of Gene Expression Measurements," *Nature Biotechnology*, 24, 1151-1161.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, D. B., Denise, T., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Jasicek, T., Daly, M. J., Wolfsberg, T. G., and Collins, F. S. (2006), "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)," *Genome Research*, 16, 123-131.

Dahl, D. B. (2005), "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," Technical Report 1086, University of Wisconsin-Madison.

Dahl, D. B., and Newton, M. A. (2007), "Multiple Hypothesis Testing by Clustering Treatment Effects," *Journal of the American Statistical Association*, 102, 517-526.

Dudoit, S., and van der Laan, M. J. (2008), "Multiple testing procedures with applications to genomics," *Springer Series in Statistics*, Springer, New York, NY.

Dunson, D. B., Herring, A. H., and Engel, S. A. (2008), "Bayesian Selection and Clustering of Polymorphisms in Functionally-Related gene," *Journal of the American Statistical Association*, 103, 534-546.

Escobar, M. D. and West. M. (1995), “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, 90, 577-588.

—- (1998), “Computing nonparametric hierarchical models,” in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. P. Mueller, D. Dey, and D. Sinha, Springer-Verlag, New York, NY.

Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, 1, 209-230.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008), “Unsupervised empirical Bayesian multiple testing with external covariates,” *The Annals of Applied Statistics*, 2, 714-735.

Foster, D. P., and Stine, R. A. (2008), “α-investing: a procedure for sequential control of expected false discoveries,” *Journal of the Royal Statistical Society*, Series B, 70, 429-444.

Gelfand, A., Sahu, S. K., and Carlin, B. P. (1995), “Efficient parametrisations for Normal Linear Mixed Models,” *Biometrika*, 82, 479-488.

Gelfand, A., Smith, A. F. M. (1990), “Sampling Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85, 398-409.

Genovese, C., and Wasserman, L. (2002), “Operating characteristic and extensions of the false discovery rate procedure,” *Journal of the Royal Statistical Society*, Series B 64, 499-517.

Gerber, G. K., Dowell, R.D . Jaakkola, T. S., and Gifford, D. K. (2007), “Hierarchical Dirichlet Process-Based Models For Discovery of Cross-species Mammalian Gene Expression,” Technical Report MIT-CSAIL-TR-2007-037, M.I.T.

Ghio, A. J., Piantadosi, C. A., and Crumbliss, A. L. (1997), “Hypothesis: iron chelation plays a vital role in neutrophilic inflammation,” *Biometals*, 10, 135-142.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. (2006), “Bayesian analysis of zero-inflated regression models,” *Journal of Statistical Planning and Inference*, 136, 1360-1375.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006), “Bayesian robust inference for differential gene expression in microarrays with multiple samples,” *Biometrics*, 62, 10-18.

Guindani, M., Zhang, S., and Mueller, P.M. (2009). “A Bayesian discovery procedure,” *Journal of the Royal Statistical Society*, Series B, 71, 905-925.

Guiney, D. G., and Lesnick, M. (2005), “Targeting of the actin cytoskeleton during infection by Salmonella strains,” *Clinical Immunology*, 114, 248-255.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martnez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., and Yu, W. (2005), “Multiple-Laboratory Comparison of Microarray Platforms,” *Nature Methods*, 2, 345-350.

Ishwaran, H., and Rao, J. S. (2005), “Spike and slab gene selection for multigroup microarray data,” *Journal of the American Statistical Association*, 100, 764-780.

Hedges, L. V., and Ingram, O. (1985), *Statistical Methods for Meta-Analysis*, Academic Press Inc., Orlando, FL.

Hofmann, W-K. (2006), *Gene expression profiling by microarrays: clinical implications*, ed., Cambridge University press, Cambridge, U.K.

Hong, F., Brietling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006), "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, 22, 2825-2827.

Horiuchi, A., Williams, K. R., Kurihara, T., Nairn, A. C., and Greengard, P. (1990), "Purification and cDNA cloning of ARPP-16, a cAMP-regulated phosphoprotein enriched in basal ganglia, and of a related phosphoprotein, ARPP-19," *Journal of Biological Chemistry*, 265, 9476-9484.

Jain, S., and Neal, R. M.(2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158-182.

Janitz, M., and Hofmann, W-K. (2008), *Next-generation genome sequencing : towards personalized medicine* eds., Wiley-VCH, Weinheim, Germany.

Kellerer, H., Pferschy, U., and Pisinger, D. (2004), *Knapsack problems*,. Springer-Verlag, Berlin, Germany.

Kerr, K. (2007), "Extended Analysis of Benchmark Datasets for Agilent Two-Color Microarrays," *BMC Bioinformatics*, 8, 371.

Khare, S., Lawhon, S. D. Adams, L. G. (2006). "MPSS analysis of *in vivo* host sense/antisense gene expression expands the molecular basis for *Salmonella typhimurium* induced enteritis,", Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station (personal communication).

Kim, S., Dahl, D. B., Vannucci, M. (2009), "Spike and Slab Dirichlet Process Prior for Multiple Hypothesis Testing in Random Effects Models," *Bayesian Analysis*, 4, 707-732.

Kortem B. H., and Vygen, J. (2008), *Combinatorial optimization: theory and algorithms*, Springer-Verlag, Berlin, Germany.

Kuo, L. (1986), "Computations of mixtures of Dirichlet Processes," *SIAM Journal on Scientific and Statistical Computing*, 7, 60-71.

Lewin, A., Richardson, S., Marshall C., Glazier, A., and Aitman, T. (2006), "Bayesian Modelling of Differential Gene Expression," *Biometrics*, 62, 10-18.

Linag, L-L., and Weiss, R. E. (2007), "A Heirarchical Semiparametric Regression model for Combining HIV-1 Phylogenic Analysis Using Iterative Reweighing Algorithms," *Biometrics*, 63, 733-741.

MacEachern, S. N., and Mueller, P. (1994), "Estimating mixture of Dirichlet processes," *Journal of Computational and Graphical Statistics*, 7, 223-238.

Malemud, C. J. (2006), "Matrix metalloproteinases (MMPs) in health and disease: an overview," *Frontiers in Bioscience*, 11, 1696-1701.

Mallick, B. K., Gold, D., and Baladandayuthapani, V. (2009), *Bayesian Analysis of Gene Expression Data.* Wiley and Sons Ltd., U.K.

Man, M. (2005), “Statistical Analysis and Modeling of SAGE Transcriptome,” in *SAGE: Current Technologies and Applications.* ed. S. M. Wang, 181-188, Horizon Bioscience, Norfolk, U.K.

Martello, S., Pisinger, D. and Toth, P. (1999), “Dynamic programming and strong bounds for the 0-1 knapsack problem,” *Management Science*, 45, 414-424.

Martello, S., Pisinger, D. and Toth, P. (2000), “New trends in exact algorithms for the 0-1 knapsack problem,” *European Journal of Operational Research*, 123, 325-332.

McAuliffe, J., Blei, D., and Jordan, M. (2006), “Nonparametric empirical Bayes for the Dirichlet process mixture model,” *Statistics and Computing*, 16, 5-14.

Medvedovic, M., and Sivaganesan, S. (2002), “Bayesian infinite mixture model based clustering of gene expression profiles,” *Statistics and Computing*, 18, 1194-1206.

Morris, J. S., Baggerly, K. A., and Coombes, K. R. (2006), “Shrinkage Estimation for SAGE Data using a Mixture Dirichlet Prior,” in *Bayesian Inference for Gene Expression and Proteomics.*, eds. K. A. Do, P. Mueller, M. Vannucci, 254-267, Cambridge University Press, New York, NY.

Mueller, P., Parmigiani, G., and Rice, K. (2007), “FDR and bayesian multiple comparisons rules,” in *Bayesian Statistics 8*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 349-370, Oxford University Press, Oxford, U.K.

Mueller, P., Parmigiani, G., Robert, C. P, and Rousseau, J. (2004), “Optimal sample size for multiple testing: the case of gene expression microarrays,” *Journal of the American Statistical Association*, 99, 999-1001.

Nacht, M., Ferguson, A. T., Zhang, W., Ptroziello, J.M., Cook, B. P., Gao, Y. H., Maguire, S., Riley, D., Coppala, G., Landes, G. M., Madden, S. L., and Sukumar, S. (1999), “Combining Serial Analysis of Gene Expression and Array technologies to identify Differentially expressed breast-cancer,” *Cancer Research*, 59, 5464-5470.

Neal, R. M. (2000), Markov chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, 9, 249-265.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), “Detecting differential gene expression with a semiparametric hierarchical mixture method,” *Biostatistics*, 5, 155-176.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. (2003), *The Analysis of Gene Expression Data: Methods and Software*, eds. Springer, New York, NY.

Patel, J. C., and Galan, J. E. (2005), “Manipulation of the host actin cytoskeleton by Salmonella–all in the name of entry,” *Current Opinion in Microbiology*, 8, 10-15.

Peña, A. E., Habiger, J. D., and Wu, W. (2010), “Power-enhanced multiple decision functions controlling family-wise error and false discovery rates,” `arXiv:0908.1767v2 [math.ST]`.

Pisinger, D (2003), “Where are the hard knapsack problems?,” *Computers & Operations Research*, 32, 2271-2284. Code available at `http://www.diku.dk/hjemmesider/ansatte/pisinger/codes.html`.

Rakhilin, S. V., Olson, P. A., Starkova, N. N., Fienberg, A. A., Nairn, A. C., Surmeier,

D. J., and Greengard, P. (2004), "A network of control mediated by regulator of calcium/calmodulin-dependent signaling," *Science*, 306, 698-701.

Reinartz, J., Bruyns, E., Lin, J. Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002), "Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms," *Briefings in Functional Genomics and Proteomics*, 1, 95-104.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A.M. (2002), "Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer," *Cancer Research*, 62, 4427-4433.

Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R. (1999), "Genome-wide expression profiling in Escherichia coli K-12," *Nucleic Acids Research*, 27, 3821-3835.

Robert, C. P. (1995), "Simulation of truncated normal variables," *Statistics and Computing*, 5, 121-125.

Roquain, E., and van de Wiel, M. A. (2009), "Optimal weighting for false discovery rate control," *Electronic Journal of Statistics*, 3, 678-711.

Sarkar, S. K., Zhou, T., and Ghosh, D. (2008), "A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective," *Statistica Sinica*, 18, 925-945.

Scharpf, R.B, Tjelmeland, H., Parmigiani, G., and Nobel, A. (2009), "A Bayesian model for cross-study differential gene expression," *Journal of the American Statistical Association*, 104, 1295-1310.

Scott, J., and Berger, J. (2003), "An exploration of aspects of Bayesian multiple testing," *Journal of Statistical Planning and Inference*, 136, 2144-2162.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, 64, 583-639 (with discussion).

Stolovitzky, G. A., Kundaje, A., Held, G. A., Duggar, K. H., Haudenschild, C. D., Zhou, D. , Vasicek, T. J., Smith, K. D., Aderem, A., and Roach, J. C. (2005), "Statistical analysis of MPSS measurements: Application to the study of LPS-activated macrophage gene expression," *Proceedings of the National Academy of Sciences*, 102, 1402-1407.

Storey, J. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society*, Series B, 64, 479-498.

Storey, J. (2003), "The positive false discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, 31, 2012-2035.

Storey, J. (2007), "The optimal discovery procedure: a new approach to simultaneous significance testing," *Journal of the Royal Statistical Society*, Series B, 69, 347-368.

Storey, J., and Tibshirani, R. (2003a), "Statistical methods for identifying differentially expressed genes in DNA microarrays," *Methods in Molecular Biology*, 224, 149-158.

—- (2003b), "Statistical significance for genomewide studies," *Proceedings of the Na-*

tional Academy of Sciences, 100, 9440-9445.

Sun, D., Speckman, P. L., and Tsutakawa, R. K. (2000), "Random effects in generalized linear mixed models (GLMMs)," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, B. K. Mallick, and S. Ghosh, 23-40, Marcel Dekker Inc., New York, NY.

Sun, W. and Cai, T. (2007). "Oracle and adaptive compound decision rules for false discovery rate control," *Journal of the American Statistical Association*, 102, 901-912.

Tang, W., and Zhang, C-H. (2007), "Empirical Bayes Methods for Controlling the False Discovery Rate with Dependent Data," in *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, eds. R. Liu, W. Strawderman, and C.-H. Zhang, 54, 151-160, IMS, Beachwood, OH.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566-1581.

Thygesen, H. H., and Zwinderman, A. H. (2006), "Modeling Sage data with a truncated gamma-Poisson model," *BMC Bioinformatics*, 7, 157.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995), "Serial analysis of gene expression," *Science*, 270, 484-487.

Vencio, R. Z. N., Brentani, H., Patrao, D. F., Pereira, C. A. (2004), "Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE)," *BMC Bioinformatics*, 5, 119.

Water, K. M., Pounds, J. G., and Thrall, B. D. (2006), "Data merging for integrated microarray and proteomic analysis," *Briefings in functional genomics and proteomics*, 5, 261-272.

West, M. (1992), "Hyperparameter estimation in Dirichlet process mixture models," *ISDS Discussion paper*, 92-A03, Duke University.

Wittes, J., and Friedman, H. P. (1999), "Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data," *Journal of the National Cancer Institute*, 91, 400-401.

Worthylake, R. A., and Burridge, K. (2001), "Leukocyte transendothelial migration: orchestrating the underlying molecular machinery," *Current Opinion in Cell Bioloлgy*, 13, 569-577.

Zeger, S., and Karim, L. (1991), "Generalized linear models with random effects: a Gibbs sampling approach," *Journal of the American Statistical Association*, 86, 79-86.

APPENDIX A

MCMC SAMPLER: FULL CONDITIONALS FOR THE PARAMETRIC AND THE SEMIPARAMETRIC MODELS (CHAPTER I)

In the parametric model, Gibbs sampling has standard conditionals given by:

- ζ_{ijk}, the latent variables as Bernoulli trials:

$$\zeta_{ijk,/Y_{ijk}=0,\lambda_{ijk}} \sim \text{Bernoulli}\left\{\frac{p}{p+(1-p)\exp(-\lambda_{ijk})}\right\}.$$

- p from its conjugate posterior (using the latent variables):

$$p \quad \sim \quad \text{Beta}(a+n_0, b+n-n_0),\ \text{where } n = IJK \text{ and } n_0 = \sum \zeta_{ijk}$$

- λ_{ijk} using a Metropolis-Hastings step with a random-walk log-$\mathcal{N}$ proposal density:

$$P(\lambda_{ijk}) \quad \propto \quad \left[I(y_{ijk}=0)\{p+(1-p)e^{-\lambda_{ijk}}\} + I(y_{ijk}\neq 0)\left\{(1-p)\frac{e^{-\lambda_{ijk}}\lambda_{ijk}^{y_{ijk}}}{y!_{ijk}}\right\}\right].$$
$$\text{log-}\mathcal{N}_{\lambda_{ijk}}(\eta_i+\beta_{ij},\sigma_\epsilon^2)$$

- σ_ϵ^2 from its conjugate posterior $\sigma_\epsilon^2 \sim \mathcal{IG}(u_{\epsilon,p}, v_{\epsilon,p})$ where $\nu_{\epsilon,p} = \nu_{\epsilon,\pi} + \frac{1}{2}IJK$, $v_{\epsilon,p} = v_{\epsilon,\pi} + \frac{1}{2}\sum_{ijk}\gamma_{ijk}^2$ and $\gamma_{ijk} = \log\lambda_{ijk} - (\eta_i + \beta_{ij})$.

- $\beta_i \equiv \{\beta_{i1}, \beta_{i2}, \beta_{i3}\}$ and σ_β from

$$\beta_i \sim \prod_{j=1}^{J}\mathcal{N}_{\beta_{ij}}\left\{\theta_{ij}, (\tau_1+\tau_2)^{-1}\right\} \text{ and } \sigma_\beta^2 \sim \mathcal{IG}(u_{\beta,p}, v_{\beta,p}) \text{ where}$$

  $\theta_{ij} = (\tau_1+\tau_2)^{-1}(\gamma_{ij}\tau_1 + 0\tau_2)$, $\gamma_{ij} = K^{-1}\sum_k(\log\lambda_{ijk} - \eta_i)$, $u_{\beta,p} = u_{\beta,\pi} + \frac{1}{2}IJ$ and $v_{\beta,p} = v_{\beta,\pi} + \frac{1}{2}\sum_{ij}(\beta_{ij} - 0)^2$

- η_i and σ_η from $\eta_i \sim \mathcal{N}_{\eta_i}\left\{\theta_i, (\tau_1+\tau_2)^{-1}\right\}$ and $\sigma_\eta^2 \sim \mathcal{IG}(u_{\eta,p}, v_{\eta,p})$ where $\theta_i = (\tau_1+\tau_2)^{-1}(\gamma_i\tau_1 + \mu\tau_2)$, $\gamma_i = (JK)^{-1}\sum_{jk}(\log\lambda_{ijk} - \beta_{ij})$, $\tau_1 = JK/\sigma_\epsilon^2, \tau_2 = 1/\sigma_\eta^2$, $u_{\eta,p} = u_{\eta,\pi} + \frac{1}{2}I$ and $v_{\eta,p} = v_{\eta,\pi} + \frac{1}{2}\sum_i(\eta_i - \mu)^2$

- μ and σ_μ from $\mu \sim \mathcal{N}_\mu\left\{\theta, (\tau_1+\tau_2)^{-1}\right\}$ and $\sigma_\mu^2 \sim \mathcal{IG}(u_{\mu,p}, v_{\mu,p})$ where $\theta = (\tau_1+\tau_2)^{-1}(\gamma\tau_1 + 0\tau_2)$, $\gamma = I^{-1}\sum_i \eta_i$, $\tau_1 = I/\sigma_\eta^2$, $\tau_2 = 1/\sigma_\mu^2$, $u_{\mu,p} = u_{\mu,\pi} + \frac{1}{2}$ and $v_{\mu,p} = v_{\mu,\pi} + \frac{1}{2}(\mu - \mu_0)^2$

In the semiparametric case, all of the above conditional distributions remain the same except for the βs. The polya-urn scheme based method gives us the following conditionals (Escobar and West 1998):

- sample $\beta_i \equiv \{\beta_{i1}, \beta_{i2}, \beta_{i3}\}$ as vector using the Polya-urn scheme:

$$\begin{aligned} P(\beta_i) &\sim q_0 G_b + \sum q_k I_{\beta_i}(\beta_k), \text{ where} \\ q_0 &\propto \tau \prod_{j=1}^{J} \phi\left(\frac{\gamma_{ij}}{\sqrt{\sigma_\epsilon^2/K + \sigma_\beta^2}}\right) \\ q_k &\propto \prod_{j=1}^{J} \phi\left(\frac{\gamma_{ij} - \beta_{ij}}{\sqrt{\sigma_\epsilon^2/K}}\right) \text{ such that } \Sigma_k q_k + q_0 = 1 \\ G_b &\sim \prod_{j=1}^{J} \mathcal{N}_{\beta_{ij}}\left\{\theta_{ij}, (\tau_1+\tau_2)^{-1}\right\}, \end{aligned}$$

  and where $\theta_{ij} = (\tau_1+\tau_2)^{-1}(\gamma_{ij}\tau_1 + 0\tau_2)$, $\gamma_{ij} = K^{-1}\sum_k(\log\lambda_{ijk} - \eta_i)$ and $\tau_1 = K/\sigma_\epsilon^2, \tau_2 = 1/\sigma_\beta^2$.

- draw cluster representatives as: $\beta_i^* \sim \prod_{j=1}^{J} \mathcal{N}_{\beta_{ij}^*}\left\{\theta_{ij}, (\tau_1+\tau_2)^{-1}\right\}$ where $\theta = (\tau_1+\tau_2)^{-1}(\bar{\gamma}_{ij}\tau_1 + 0\tau_2)$, $\bar{\gamma}_{ij} = (n_i K)^{-1}\sum_{i'\in\mathcal{S}_\rangle}\sum_k(\log\lambda_{i'jk} - \eta_{i'})$ $\tau_1 = n_i K/\sigma_\epsilon^2$, $\tau_2 = 1/\sigma_\beta^2$. Here $\mathcal{S}_i$ is the set of all (signature) indexes in cluster i and n_i is $|\mathcal{S}_i|$, the cardinality of the set $\mathcal{S}_i$. After drawing the cluster representatives, update the cluster membership, number of clusters and τ.

APPENDIX B

MCMC SAMPLER: FULL CONDITIONALS FOR THE SEMIPARAMETRIC MODEL (CHAPTER II)

Let $g = h_d = 1, 2, ..., n_d$ for simplicity. For each of g (discrete data-set)

- Sample $\zeta_{g,ijk}$, the latent variables as Bernoulli trials:

$$\zeta_{g,ijk} | x_{g,ijk} = 0, \lambda_{g,ijk} \sim \text{Bernouli} \left(\frac{p_g}{p_g + (1 - p_g)\exp(-\lambda_{g,ijk})} \right).$$

- Using the latent variables drawn above, sample p_g from its conjugate posterior: $p_g \sim \text{Beta}(a^{\text{pst}}_{p,g}, b^{\text{pst}}_{p,g})$ where $a^{\text{pst}}_{p,g} = a^{\text{pr}}_{p,g} + n_{0,g}$, $b^{\text{pst}}_{p,g} = b^{\text{pr}}_{p,g} + n - n_{0,g}$ $n_{0,g} = \sum \zeta_{g,ijk}$ and $n = IJK$.

- Using a Metropolis-Hastings step with a random-walk log-$\mathcal{N}$ proposal density, sample $\lambda_{g,ijk}$ from:

$$\begin{aligned} P(\lambda_{g,ijk}) \quad \propto \quad & \left[I(Y_{g,ijk=0})(p_g + (1 - p_g)e^{-\lambda_{g,ijk}}) + I(Y_{g,ijk \neq 0}) \left((1 - p_g) \frac{e^{-\lambda_{g,ijk}} \lambda_{g,ijk}^{x_{g,ijk}}}{x_{g,ijk}} \right) \right] \\ & \text{log-}\mathcal{N}_{\lambda_{g,ijk}}(f_g(z_{ijk}), \sigma^2_{f,g}) \end{aligned}$$

- The link function measurement error variances $\sigma^2_{f,h}, h = 1, 2, ..., n_c + n_d$ can be sampled from: $\sigma^2_{f,h} \sim \mathcal{IG}(u^{\text{pst}}_{f,h}, v^{\text{pst}}_{f,h})$ where $u^{\text{pst}}_{g,h} = u^{\text{pr}}_{f,h} + \frac{1}{2} IJK$, $v^{\text{pst}}_{f,h} = v^{\text{pr}}_{f,h} + \frac{1}{2} \sum_{ijk} \left[y_{h,ijk} - f_h(z_{hijk}) \right]^2$

- Genes/signature effects and their variances, η_{hi} and $\sigma^2_{\eta,h}$, have the following conditional distributions:

  - draw η_{hi}
    $\eta_{hi} \sim \mathcal{N} \left(\theta_{hi}, (\sum_j \tau_{hij,1} + \tau_{h,2})^{-1} \right)$, where $\theta_{hi} = \frac{\sum_j \bar{z}_{hij} \tau_{hij,1} + \mu_h \tau_{h,2}}{\sum_j \tau_{hij,1} + \tau_2}$, $\bar{z}_{hij} = K^{-1} \sum_k (z_{hijk} - (2\psi_{hij} - 1)\beta_{hi})$ $\tau_{hij,1} = (\sigma^2_{z,hij}/K)^{-1}, \tau_{h,2} = (\sigma^2_{\eta,h})^{-1}$

- draw $\sigma^2_{\eta,h}$

  $\sigma^2_{\eta,h} \sim \mathcal{IG}(u^{\text{pst}}_{\eta,h}, v^{\text{pst}}_{\eta,h})$, where $u^{\text{pst}}_{\eta,h} = u^{\text{pr}}_{\eta,h} + \frac{1}{2}I$, $v^{\text{pst}}_{\eta,h} = v^{\text{pr}}_{\eta,h} + \frac{1}{2}\sum_i (\eta_{hi} - \mu_h)^2$

- Study specific means and the corresponding variances, conditional for μ_h and σ^2_μ, can be sampled from:

  - draw μ_h

    $\mu_h \sim \mathcal{N}(\theta, (\tau_1 + \tau_2)^{-1})$, where $\theta = (\tau_1 + \tau_2)^{-1}(\gamma\tau_1 + \mu_0\tau_2)$ $\gamma = S^{-1}\sum_i \eta_{hi}$ $\tau_1 = (\sigma^2_{\eta,h}/S)^{-1}, \tau_2 = (\sigma^2_\mu)^{-1}$

  - draw σ^2_μ

    $\sigma^2_\mu \sim \mathcal{IG}(u^{\text{pst}}_\mu, v^{\text{pst}}_\mu)$ where $u^{\text{pst}}_\mu = u^{\text{pr}}_\mu + \frac{S}{2}$ $v^{\text{pst}}_\mu = v^{\text{pr}}_\mu + \frac{1}{2}\sum_h (\mu_h - \mu_0)^2$

- The posterior distribution of the spline parameters α_h and $\sigma^2_{\alpha,h}$ takes the form of constrained multivariate normal distribution whose conditional are truncated univariate normal densities, i.e.,

  - draw α_h, the spline coefficients

    $[\alpha_h] \propto \mathcal{N}(\theta, \Sigma)|_{-\infty \leq \alpha_{h,1} \leq \cdots \leq \alpha_{h,L} \leq +\infty}$ where $\theta = \Sigma^{-1}X^TY/\sigma^2_{y,l}$ and $\Sigma^{-1} = \frac{1}{\sigma^2_{f,h}}X^TX + \frac{1}{\sigma^2_{\alpha,h}}\Delta^{-1}$. And X is the design matrix obtained by evaluating the splines at z_{hijk}'s and $Y_{IJK\times 1}$ is the vectorized $y_{h,ijk}$. The number and locations of the knots is fixed across all simulations. There is additional Gibbs sampling which involves univariate truncated normal distributions (Robert, 1995).

  - draw $\sigma^2_{\alpha,h}$

    $\sigma^2_{\alpha,h} \sim \mathcal{IG}(u^{\text{pst}}_{\delta,p}, v^{\text{pst}}_{\delta,p})$, where $u^{\text{pst}}_{\alpha,h} = u^{\text{pr}}_{\alpha,h} + \frac{1}{2}L$ and $v^{\text{pst}}_{\alpha,h} = v^{\text{pr}}_{\alpha,h} + \frac{1}{2}\alpha^T_h\Delta^{-1}\alpha_h$

- Conditional distribution for $\sigma^2_{z,hij}$ $(j = 0, 1)$ is given by:

  $\sigma^2_{z,hij} \sim \mathcal{IG}(u^{\text{pst}}_{z,hij}, v^{\text{pst}}_{z,hij})$ where $u^{\text{pst}}_{z,hij} = u^{\text{pr}}_{z,hij} + \frac{1}{2}K$ and $v^{\text{pst}}_{z,hij} = v^{\text{pr}}_{z,hij} + \frac{1}{2}\sum_k [z_{hijk} - (\eta_{hi} + (2\psi_{hij} - 1)\beta_{hi})]^2$

- Conditional distribution for z_{hijk} is given as:

$$z_{hijk} \propto \mathcal{N}(\eta_{hi} + (2\psi_{hij} - 1)\beta_{hi}, \sigma^2_{z,hi\psi_{hi}}) \exp[-\frac{1}{2\sigma^2_{f,h}}\{y_{h,ijk} - f_h(z_{hijk})\}^2]$$

- The random effects β_{hi} are sampled based on Chinese Franchise representation (CFR) with augmented sampling (Teh et al., 2006; Gerber et al, 2006). Gathering the terms involving β_{hi} in the hierarchical model, excluding the prior part and expressing them in terms of the sufficient statistics, we:

$$x_{hi} \propto \mathcal{N}(\beta_{hi}, \tau_{hi}^{-1})$$

where $x_{hi} = \tau_{hi}^{-1} \sum_j \bar{z}_{hij}\tau_{hij}$, $\bar{z}_{hij} = K^{-1}\sum_k(z_{hijk} - \eta_{hi})(2\psi_{hij} - 1)$ $\tau_{hij} = (\sigma^2_{z,hij}/K)^{-1}$, $\tau_{hi} = \sum_j \tau_{hij}$ Let β^0_q be the q-th unique cluster representative. There can, at most be, SI many clusters. In the CFR parlance, β^0_q is the q-th dish being served with probability w^0_q, h- indexes the restaurant and i- indexes the customer and β_{hi} is the dish being served to the i-th customer in the j-th restaurant. Further, let n_{hq} represent the number of customers in the h-th restaurant being served the dish β^0_q. In the same vein, n^{-i}_{hq} excludes the i-th customer in the h-th restaurant from counting.

  - Sample the cluster configuration:

$$\begin{aligned} P(\beta_{hi} = \beta^0_q) &\propto (\tau w^0_q + n^{-i}_{hq})\phi([x_{hi} - \beta^0_q]/\tau_{hi}) \\ P(\beta_{hi} \neq \beta^0_q\ \forall q) &\propto \tau w^0_* [\pi\phi(x_{hi}/\tau_{hi,1}) + (1-\pi)\phi(x_{hi}/\tau_{hi,2})] \end{aligned}$$

    where $\phi(x)$ is the standard normal probability density evaluated at x, $\tau_{hi,1} = \tau_{hi}$, $\tau_{hi,2} = \tau_{hi} + (\sigma^2_\beta)^{-1}$ and $w^0_* = 1 - \sum_{q=1}^{Q} w^0_q$, Q is the existing number of unique dishes. When a new dish is created (or new cluster is created), $\zeta_{hi} = 0$ if β_{hi} is drawn from point mass at 0 with probability $\frac{\pi\phi(x_{hi}/\tau_{hi,1})}{\pi\phi(x_{hi}/\tau_{hi,1})+(1-\pi)\phi(x_{hi}/\tau_{hi,2})}$ and 1 otherwise. In the event that the current customer is served an existing dish, then γ_{hi} is inherited. Thus, γ is the indicator variable which denotes whether the effects belong to the null or the alternate. When indeed $\zeta_{hi} = 1$, a new β^0_{Q+1} is drawn from $\mathcal{N}({\tau_{hi,2}}^{-1}\tau_{hi,1}x_{hi}, {\tau_{hi,2}}^{-1})$ and set $Q \leftarrow Q + 1$. A Metropolis-Hasting update step can often improve mixing of the latent indicator variables γ_{hi}. Let p_{birth} be the probability of switching $\gamma_{hi} = 0$ to $\gamma_{hi} = 1$ and the posterior odds defined as:

$$r \quad = \quad \frac{\phi(x_{hi}/\tau_{hi,2})p_{\text{birth}}}{\phi(x_{hi}/\tau_{hi,1})(1 - p_{\text{birth}})}.$$

When $\gamma = 0$, with probability p_{birth}, we propose $\gamma = 1$ and accept this proposal with probability r.

– Sample the auxiliary variable, m_{hq}, the number of customers eating dish q in restaurant h:

$$P(m_{hq} = m) \quad \propto \quad \text{Stirling}(n_{hq}, m)(\tau w^0_q)^m$$

where Stirling(n, m) are the unsigned Stirling numbers of first kind, that count the permutations of n objects having m permutations.

– Sample the unique dish weights w^0_q

$$P(w^0) \quad \propto \quad \text{Dirichlet}(\sum_h m_{h1}, \ldots, \sum_h m_{hQ}, \tau_0)$$

where Dirichlet$(a_1, \ldots, a_{Q+1})$ is a $Q + 1$-variate Dirichlet distribution.

– Sample the precision parameter τ in the bottom hierarchy of the Dirichlet Process (based on Auxiliary sampling)

(a) Draw w_h an auxiliary variable

$$P(w_h) \;\propto\; w_h^{\tau}(1-w_h)^{N_h-1}$$

where $N_h = I$ is the total number of customers in the h-th restaurant.

(b) Draw b_h another auxiliary variable

$$P(b_h) \;\propto\; \left(\frac{N_h}{\tau}\right)^{b_h}$$

(c) Draw τ, the precision parameter

$$P(\tau) \;\propto\; \text{Gamma}(u_\tau^{\text{pst}}, v_\tau^{\text{pst}})$$

$v_\tau^{\text{pst}} = v_\tau^{\text{pr}} + \sum_h (M_h - b_h)$,$u_\tau^{\text{pst}} = u_\tau^{\text{pr}} - \sum_h \log w_h$

where $M_h = \sum_i m_{hi}$ and repeat the process for 10-20 iterations for local convergence.

- Sample the precision parameter τ_0 in the top hierarchy of the Dirichlet Process, from:

$$P(\tau) \;\propto\; \text{Stirling}(\sum_h M_h, Q)\tau^Q \frac{\Gamma(\tau)}{\Gamma(\tau + \sum_h M_h)}$$

as in Escobar and West (1995).

- Conditional distribution for σ_β^2 is given by:
  $\sigma_\beta^2 \sim \mathcal{IG}(u_\beta^{\text{pst}}, v_\beta^{\text{pst}})$ where $u_\beta^{\text{pst}} = u_\beta^{\text{pr}} + \frac{1}{2}Q^*$ and $v_\beta^{\text{pst}} = v_\beta^{\text{pr}} + \frac{1}{2}\sum_q [\beta_q^0 - 0]^2$ where Q^* is the unique number of dishes that are different from zero.

- The null hypothesis proportion π is sampled from:
  $\pi \sim \text{Beta}(u_\pi^{\text{pst}}, u_\pi^{\text{pst}})$ where $u_\pi^{\text{pst}} = u_{p,g}^{\text{pr}} + n_0$, $v_\pi^{\text{pst}} = v_\pi^{\text{pr}} + n - n_0$ $n_0 = \sum 1 - \gamma_{hi}$ and $n = SI$.

APPENDIX C

ALGORITHM FOR FDR CONTROL BASED ON DYNAMIC PROGRAMMING

Algorithm 1: Computing the table

```
for c = 0 to C, $K[0,c] = 0$

for i = 0 to I, $K[i,0] = 0$

for i = 1 to I

    for c = 1 to C

        if $w_i >$ c

            K[i,c]=K[i-1,c]

        else

            if $v_i$+K[i-1 ,c-$w_i$] $>$ K[i-1,c], K[i,c] = $v_i$ + K[i-1,c-$w_i$]

            else K[i,c] = K[i-1,c]

        end

    end

end
```

Algorithm 2: Finding items in the knapsack

```
set i=I,c=C.

do untill i=0

    if K[i,c] $\neq$ K[i-1,c]

        mark the i-th item as in the knapsack
        i=i-1, c= c-$w_i$
    else
        i=i-1
    end
end
```

Algorithm 3: FDR control

```
for c = 0 to C, $K[0,c] = 0$
for i = 0 to I, $K[i,0] = 0$
set c = 1, fdr[0] = 0
do until fdr[c] > $\alpha$ or c = C+1
    for i = 1 to I
        if $w_i >$ c
            K[i,c]=K[i-1,c]
        else
            if $v_i$+K[i-1 ,c-$w_i$] $>$ K[i-1,c], K[i,c] = $v_i$ + K[i-1,c-$w_i$]
            else K[i,c] = K[i-1,c]
        end
    set j = c, $d_i = 0 \forall i$
    do until i=0
        if K[i,j] $\neq$ K[i-1,j]
            set $d_i$ = 1
            i=i-1, j= j-$w_i$
            else
               i=i-1
            end
         end
         fdr[c] = $(B\sum_i d_i)^{-1}\sum_i d_i w_i$
         M[c,i] = $d_i \forall i$
         c = c+1
     end
     report fdr[c-1] and the decisions $d_i^* = M[c-1,i]\ \forall i$
```

VITA

Soma Sekhar Dhavala graduated from Andhra University, Bhimavaram , India in August 1997 with a Bachelor of Engineering (B.E) in electronics and communications engineering. In March 2000, he received a Master of Science degree in electrical engineering from Indian Institute of Technology, Madras. Before returning to school, he worked at the General Electric Global Research Center, in Bangalore, India until 2002. After spending two years at Iowa State University, he joined the statistics department at Texas A&M University, in Fall'05, where he received his Ph.D. in December, 2010. He will work as post-doc with Distinguished Professor Raymond J. Carroll and Professor Bani K. Mallick, at the same university. He can be contacted at:

Soma S. Dhavala
Bloc 437
Department of Statistics
Texas A&M University
3143, TAMU
College Station, TX, 77843-3143
email: `soma@stat.tamu.edu`

The typist for this dissertation was Soma S. Dhavala.