RICE UNIVERSITY

Bayesian Models for High-Dimensional Count Data with Feature Selection

by

Qiwei Li

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Marina Vannucci, Chair
Noah Harding Professor of Statistics
Rice University

Michele Guindani
Associate Professor of Statistics
University of California, Irvine

Christopher M. Jermaine
Professor of Computer Science
Rice University

David W. Scott
Noah Harding Professor of Statistics
Rice University

Houston, Texas
October, 2016
ABSTRACT

Bayesian Models for High-Dimensional Count Data with Feature Selection

by

Qiwei Li

Modern big data analytics often involve large data sets in which the features of interest are measured as counts. My thesis considers the problem of modeling a high-dimensional matrix of count data and presents two novel Bayesian hierarchical frameworks, both of which incorporate a feature selection mechanism and account for the over-dispersion observed across samples as well as across features. For inference, I use Markov chain Monte Carlo (MCMC) sampling techniques with Metropolis-Hastings schemes employed in Bayesian feature selection.

In the first project on Bayesian nonparametric inference, I propose a zero-inflated Poisson mixture model that incorporates model-based normalization through prior distributions with mean constraints. The model further allows us to cluster the samples into homogenous groups, defined by a Dirichlet process (DP) while selecting a parsimonious set of discriminatory features simultaneously. I show how my approach improves the accuracy of the clustering with respect to more standard approaches for the analysis of count data, by means of a simulation study and an application to a bag-of-words benchmark data set, where the features are represented by the frequencies of occurrence of each word.

In the second project on Bayesian integrative analysis, I propose a negative binomial mixture regression model that integrates several characteristics. In addition to
feature selection, the model includes Markov random field (MRF) prior models that capture structural dependencies among the features. The model further allows the mixture components to depend on a set of selected covariates. The simulation studies show that employing the MRF prior improves feature selection accuracy. The proposed approach is also illustrated through an application to RNA-Seq gene expression and DNA methylation data for identifying biomarkers in breast cancer.
To Na & Thomas
I would like to express my sincere appreciation to my advisor Dr. Marina Vannucci. She has always given me tremendous support in my Ph.D. studies. Her rigorous manner towards research has encouraged me to seek the truth as well as the beauty in Statistics. I would also like to thank Dr. Michele Guindani for his guidance through my first project and Dr. Alberto Cassese for our cooperation through my second project. Without their contributions, I could have never completed this research work. I would also like to thank Dr. David W. Scott and Dr. Chrispher M. Jermaine for serving as my committee members even at hardship and providing brilliant comments and suggestions.

Last but not least, I want to thank my family for their unconditional support and trust, especially to my wife, my son, my younger brother, my parents, and my parents-in-law.
## Contents

Abstract ii  
List of Illustrations ix  
List of Tables xi  

### 1 Introduction 1  
1.1 Bayesian feature selection in regression models 2  
1.1.1 Regression models 2  
1.1.2 Bayesian feature selection methods 3  
1.2 Bayesian feature selection in mixture models 6  
1.2.1 Mixture models 6  
1.2.2 Bayesian feature selection methods 7  
1.3 Univariate count data modeling 9  
1.3.1 Over-dispersion modeling 9  
1.3.2 Zero-inflation modeling 10  
1.4 Overview of projects 12  

### 2 A Bayesian zero-inflated Poisson mixture model for clustering and feature selection 14  
2.1 Introduction 14  
2.2 Zero-inflated Poisson mixture model with feature selection 18  
2.2.1 Feature selection 18  
2.2.2 Estimation of scaling factors via mean constraints 20  
2.2.3 Clustering selected features via Dirichlet process mixtures 24
Appendix B: Detailed MCMC algorithm for Chapter 3  96
Illustrations

2.1 Hierarchical formulation of the proposed constrained Poisson mixture (CoPoM) model. .......................................................... 26
2.2 Simulated data: Heatmap of the pairwise posterior probabilities $p(z_i = z_j | ·)$. .......................................................... 31
2.3 Simulated data: Marginal posterior probabilities of inclusion $p(\gamma_j = 1 | ·)$ with the red dots indicating the truly discriminatory features. .......................................................... 32
2.4 Simulated data: Estimated mixing measures $E[d_{ij}]$ for two of the discriminatory features. .......................................................... 34
2.5 Simulated data: Receiver operating characteristic (ROC) curves, for different values of the threshold on the PPIs, obtained by CoPoM under different simulated scenarios. The bold curve corresponds to $\sigma^2 = 1$, whereas the dashed curve corresponds to $\sigma^2 = 0.5$. ............ 40
2.6 Brown Corpus data: Trace plot of the number of clusters $K$ (top); Histogram of the number of clusters $K$ (bottom). .................. 43
2.7 Brown Corpus data: Trace plot of the number of selected variables $p_\gamma$ (top); Marginal posterior probabilities of inclusion $p(\gamma_j = 1 | ·)$ (bottom). 44
2.8 Brown Corpus data: Heatmap of the pairwise posterior probabilities $p(z_i = z_j | ·)$. .......................................................... 45
2.9 Brown Corpus data: The hierarchical clustering dendrograms using the Euclidean distance measure and the complete agglomeration method achieved by CoPoM, PoiClaClu and edgeR, respectively. .... 46
3.1 Hierarchical formulation of the proposed hierarchical mixture model.

3.2 Simulation study: Empirical distributions of correlations between features in simulated data (top) and synthetic data (bottom).

3.3 Simulated data: Marginal posterior probability of inclusion $p(\gamma_j = 1|\cdot)$ with red dots indicating truly discriminatory features using the MRF prior (top) and the independent Bernoulli prior (bottom).

3.4 Simulated data: Marginal posterior probability of inclusion $p(\delta_{rj} = 1|\cdot)$ for two of the discriminatory features ($j = 74$ and $j = 1$) with red circles indicating true covariates using the MRF prior (top) and the independent Bernoulli prior (bottom).

3.5 Synthetic data: Marginal posterior probability of inclusion $p(\gamma_j = 1|\cdot)$ with red dots indicating truly discriminatory features using the MRF prior (top) and the independent Bernoulli prior (bottom).

3.6 Breast cancer data: The gene-gene network structure from KEGG database, with the red nodes indicating the selected features whose marginal posterior probability of inclusion $p(\gamma_j = 1|\cdot) \geq 0.875$ with MRF prior.

3.7 Breast cancer data: Scatter plot of marginal posterior probabilities $p(\gamma_j|\cdot)$ with vs. without MRF prior (top); The number of overlapping genes as a function of the first $p, p = 1, \ldots, 100$ genes between the models with and without MRF prior (bottom).

3.8 Breast cancer data: Marginal posterior probability of inclusion $p(\delta_{rj} = 1|\theta_{rj} = 1, \cdot)$ with red circles indicating selected probes for discriminatory genes $FZD10, PLCB1, PDGFRA, WNT16$, and $CXCL12$.  

x
## Tables

2.1 Simulated data (from zero-inflated Poisson mixtures): Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages *edgeR* and *PoiClaClu*, and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses. .................................................. 35

2.2 Simulated data (from Poisson mixtures): Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages *edgeR* and *PoiClaClu*, and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses. ........ 37

2.3 Simulated data: Average false positive rates (FPRs) and true positive rates (TPRs), achieved by the CoPoM model under different simulated scenarios, for different values of the threshold, $\tau$, on the PPIs. 39

2.4 Brown Corpus data: Pairwise probability matrix (PPM) estimates of sample allocation $\hat{z}$. ................................................................. 42

3.1 Simulated data: Sensitivity analysis on choosing the values of hyperparameters $d$ and $f$ in the MRF prior. Standard deviations are indicated in parentheses. ................................. 68
3.2 Breast cancer data: List of genes selected by the model with the MRF prior and their corresponding posterior probabilities. . . . . . . 75

3.3 Breast cancer data: Results of the enrichment analysis. . . . . . . . . . 81
Chapter 1

Introduction

In the past decades, feature selection (also known as variable selection) has attracted many statisticians from both frequentist and Bayesian fields, leading to a variety of methods. The motivation of using a feature selection mechanism in statistical models is that data, particularly high-dimensional data, contains many redundant or irrelevant features. In regression models, those unnecessary features complicate the model and increase the overfitting. In cluster analysis, the inclusion of noisy features can mask the recovery of the group structure. Therefore, implementing the process of selecting a subset of relevant features in model construction will make improvement on the model itself in many aspects. Sections 1.1 and 1.2 introduce the background of feature selection in regression models and mixture models, respectively. Several Bayesian feature selection methods corresponding to these two major statistics problems are briefly reviewed.

The novel approaches presented in this thesis deal with the data, of which features of interest are measured as counts. Continuous data are usually modeled by the normal distribution. However, analytic tools with this assumption may not be suitable for count data. For instance, data normalization on count data may result in information loss and thus lead to a biased inference. It is also very challenging to incorporate covariance structure between features measured as counts. Last but not least, the excess of zero counts and larger variance than mean are the two very common phenomena observed in count data. Section 1.3 summarizes the statistical
models designed to capture the above characteristics. In Section 1.4, I give a brief outline of how the rest of the thesis is structured.

1.1 Bayesian feature selection in regression models

In this section, I provide a brief introduction to feature selection in regression models and review two widely used Bayesian feature selection methods. I refer to the article by O’Hara et al. (2009) for a complete survey.

1.1.1 Regression models

In statistics, regression refers to an approach for modeling the relationship between a dependent variable (also known as a response variable, etc.) and one or more explanatory variables (also known as independent variables, regressors, covariates, predictors, etc.). For convenience, I take multiple linear regression models for instance, where the dependent variable is a linear combination of multiple parameters. In particular, given a vector of regression parameters \( \beta_0, \ldots, \beta_p \), the response \( y_i \) is modeled as a linear combination of the explanatory variables \( x_{ij} \), where we use \( i \) to index the sample (also known as observation). The model can be written as follows,

\[
y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{1.1}
\]

where \( \epsilon_i \sim N(0, \sigma^2) \), \( i = 1, \ldots, n \) are the normally distributed errors. Note that the covariates \( x_{ij} \)'s can be either continuous or discrete values. Generalized linear models can be further considered, assuming that \( y_i \) is a member of the exponential family of distributions. In such a model, we take the link function \( g(\cdot) \) so that the expectation of \( g(y_i) \) equals to the right-hand side of (1.1) without the error term. For instance, if the link function is logarithm and the response \( y_i \) is assumed to follow a Poisson
distribution, the model is called as a Poisson regression model and can be written as follows,

\[ y_i \sim \text{Poi}(\mu_i), \]
\[ \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \]  \hspace{1cm} (1.2)

One of the most interesting questions in regression models is deciding which explanatory variables should be used in model construction, particularly when \( p \) is large. One reason is that simplification makes researchers interpret the model more easily, and the other is that enhanced generalization can reduce overfitting. In other words, the aim of feature selection is to select a small subset of the features, which can explain a large fraction of variation in the response. This problem has led to a variety of algorithms for searching the model space and the selection criteria for choosing between competing models (see, e.g. Miller, 2002; Broman and Speed, 2002; Sillanpää and Corander, 2002).

### 1.1.2 Bayesian feature selection methods

In the Bayesian framework, feature selection is a problem of parameter estimation, rather than searching for the single optimal model based on some criterion. In general, the Bayesian feature selection procedure is seen as a way to decide which of the regression parameters \( \beta_j \)'s are equal to zero. This can be formulated by a “slab and spike” prior on \( \beta_j \), with a spike (i.e. the probability mass) be exactly at zero and a flat slab elsewhere. A convenient way of representing whether feature \( j \) is in the slab or spike part of the prior is via an auxiliary binary variable \( \gamma_j \), with \( \gamma_j = 0 \) indicates the absence of feature \( j \) in the model and \( \gamma_j = 1 \) indicates presence. We may have some \textit{a priori} knowledge which candidate features are more likely to be in the model, and
ideally such information should be taken into account in the construction of prior on \( \gamma_j \). For instance, a simple choice is to assume \( \gamma_j \)'s are independent Bernoulli random variables. Let \( \gamma \) denotes the vector of \( \gamma_1, \ldots, \gamma_p \), we can write

\[
\gamma_j \sim \text{Bern}(\omega), \quad \text{or equivalently}
\]

\[
p(\gamma) = \prod_{j=1}^{p} \omega^{\gamma_j} (1 - \omega)^{1-\gamma_j}.
\] (1.3)

The number of \( \gamma_j \)'s of which value equals to one (i.e. the number of selected features), denoted by \( p_\gamma \), thus follows a binomial distribution, and \( \omega \) can be elicited as the proportion of features expected \( \text{a priori} \) to be included in the model. Once the model has been set up, Markov chain Monte Carlo (MCMC) method is used to draw samples from the model space. Using the MCMC outputs, we can directly estimate the marginal posterior probability that feature \( j \) is included in the model. Depending on the prior definition \( p(\beta_j|\gamma_j) \), two primary Bayesian feature selection methods are summarized below.

**Gibbs variable selection (GVS)**

Gibbs variable selection (GVS) was first introduced by Dellaportas et al. (2002), extending a general idea of Carlin and Chib (1995). A mixture prior is assumed for \( \beta_j \) conditional on \( \gamma_j \),

\[
p(\beta_j|\gamma_j) = (1 - \gamma_j)N(\mu, s^2) + \gamma_jN(0, \tau^2),
\] (1.4)

where \( \mu \) and \( s^2 \) are user-defined tuning parameters, and \( \tau^2 \) is a fixed prior variance of \( \beta_j \). The intuitive idea is to use a prior for \( \beta_j|\gamma_j = 0 \) which is concentrated around the posterior density, so that when \( \gamma_j = 0 \), \( p(\beta_j|\gamma_j = 1) \) is reasonably large, and hence there is a good probability that the chain will move to \( \gamma_j = 1 \). The algorithm
does require tuning, i.e. $\mu$ and $s^2$ need to be chosen so that good values of $\beta_j$ are proposed when $\gamma_j = 0$. The data will determine which values are good without directly influencing the posterior, and hence tuning can be done to improve mixing without changing the model’s priors.

**Stochastic search variable selection (SSVS)**

Stochastic search variable selection (SSVS) is the most widely used technique that has been applied in a wide range of applications. It was first introduced by George and McCulloch (1993) and extended by Brown et al. (1998). A mixture prior for $\beta_j$ conditional on $\gamma_j$ is used,

$$p(\beta_j | \gamma_j) = (1 - \gamma_j)N(0, \tau^2) + \gamma_j N(0, g\tau^2),$$

where the spike, i.e. the first density, is safely centered around zero when choosing small $\tau^2$. The other fixed hyperparameter $g$ is usually chosen to be a large value, so that if $\gamma_j = 1$, the corresponding covariate is more likely to be included in the final model. This prior model also gives identifiability for variables $\gamma_j$ and $\beta_j$, but the algorithm requires tuning on $\tau^2$ and $g$ to obtain fast convergence. However, tuning is tricky, as $p(\beta_j | \gamma_j = 0)$ needs to be small while not too restricted around zero. Otherwise, Gibbs sampler moving between states $\gamma_j = 0$ and $\gamma_j = 1$ is not possible in practice. A similar SSVS model is the “spike and slab” prior of Mitchell and Beauchamp (1988),

$$p(\beta_j | \gamma_j) = (1 - \gamma_j)I_0(\beta_j) + \gamma_j N(0, \tau^2),$$

where $I_0$ is a point mass function on zero to indicate whether $\beta_j = 0$. 

\
1.2 Bayesian feature selection in mixture models

In this section, we provide a brief introduction to mixture models for sample clustering, especially in high-dimensional settings. Then we introduce several Bayesian approaches to cluster the samples while identifying a parsimonious set of discriminatory features.

1.2.1 Mixture models

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. One of its most important applications is model-based clustering, where the data are viewed as coming from a mixture of distributions and each distribution represents a different cluster. Let \( Y = (y_1, \ldots, y_n) \) be independent \( p \)-dimensional observations from \( K \) populations. The problem of clustering the \( n \) observations can be formulated in terms of a mixture of \( K \) underlying distributions with parameters \( \Theta = (\theta_1, \ldots, \theta_K) \),

\[
f(y_i | w, \Theta) = \sum_{k=1}^{K} w_k f(y_i | \theta_k), \tag{1.7}
\]

where \( f(y_i | \theta_k) \) is the density of observation \( i \) from component \( k \) and \( w = (w_1, \ldots, w_K) \) are the component weights. Note that \( w_k \geq 0 \) and \( \sum_{k=1}^{K} w_k = 1 \). A convenient way of representing the cluster from which each observation is drawn is via an auxiliary variable \( z_i \), where \( z_i = k \) indicates observation \( i \) comes from component \( k \). The \( z_i \)'s are assumed to be independently and identically distributed with probability mass function \( p(z_i = k) = w_k \). The sample allocation vector \( z \) is defined as the vector of \( z_1, \ldots, z_n \).
1.2.2 Bayesian feature selection methods

The practical utility of feature selection is well recognized, and several methods have been developed for regression models. However, few contributions have been made in the context of clustering. This is a more challenging problem since there is no observed response to guide the selection. Especially when $p$ is large, it is often the case that only a small number of features provide information about the group structure of the observations. The inclusion of the noisy features could be detrimental since they can obscure the recovery of the true cluster structure (Tadesse et al., 2005). A convenient way of representing the relevant features is also via an auxiliary binary variable $\gamma_j$, with $\gamma_j = 1$ defines a mixture distribution and $\gamma_j = 0$ otherwise. We use $(\gamma)$ and $(\gamma^c)$ to index the discriminating features and those favor a single homogeneous background density. By this notation, we can rewrite the mixture model with feature selection as follows,

$$f(y_i^{(\gamma)}|w, \Theta) = \sum_{k=1}^{K} w_k f(y_i^{(\gamma)}|\theta_k).$$  (1.8)

Note that this modeling approach is different from what is introduced in Bayesian feature selection for regression models in Section 1.1.2, where the latent indicator $\gamma$ is used to induce mixture priors on the regression coefficients. Generally speaking, the clustering in feature selection approaches can be categorized based on their cluster models into two groups, which both allow separate subsets of features to discriminate different groups of observations.

Distance-based methods

By formulating the problem in terms of distance-based clustering with weighted features, Friedman and Meulman (2004) proposed an algorithmic approach for clustering
observations on separate subsets of features. They used heuristic search strategies to find an optimal weighting of the features while simultaneously minimizing the clustering criterion. Their approach works in conjunction with hierarchical clustering, and hence does not provide inference on the number of clusters as well as a measure of uncertainty for the sample allocations.

Model-based methods

Model-based approaches have also recently been developed. Hoff et al. (2006) adopted a mixture of Gaussian distributions where different clusters are identified by mean shifts. The model parameters are updated using MCMC sampling techniques and then Bayes factors are computed to identify discriminating features. Tadesse et al. (2005) first put forward a feature selection method in which latent variables were introduced to identify discriminating features and the clustering was formulated in terms of a finite mixture of Gaussian distributions with an unknown number of components. They used a reversible jump MCMC technique to allow for the creation and deletion of clusters. Unlike the previous approaches, their method assumes that the same subsets of features discriminate across all components. However, Tadesse et al. (2005)’s feature selection technique has the advantage of allowing flexible inference on both joint and marginal posterior distributions of the variables. Kim et al. (2006) further built on the model of Tadesse et al. (2005) by formulating the clustering in terms of an infinite mixture of distributions via Dirichlet process (DP) mixtures. Samples from a DP are discrete with probability one and can therefore produce a number of ties, thereby forming clusters. This modeling strategy in both Tadesse et al. (2005) and Kim et al. (2006) results in a unified approach to simultaneously infer group structure in the observations while identifying the discriminatory features.
1.3 Univariate count data modeling

In statistics, count data is a data type, in which the observations can take only the non-negative integer values such as \{0, 1, 2, \ldots\}. Note that these integers arise from counting rather than from ranking. For instance, count data can be the number of monthly transactions in a bank account. An individual piece of count data is often termed a count variable. When such a variable is treated as a random variable, a Poisson distribution is usually employed to represent its distribution, denoted by

\[ y|\lambda \sim \text{Poi}(\lambda), \]

with the probability density function

\[ p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}. \]

Both the expectation \( E[y] \) and the variance \( \text{Var}(y) \) equal to the intensity parameter \( \lambda \). To model count data as a function of covariates, we can use the benchmark Poisson regression model given in (1.2).

1.3.1 Over-dispersion modeling

A key assumption of the Poisson model is that the variance equals to the mean. However, count data often exhibits over-dispersion, with a variance much larger than the mean. An approach to modeling such characteristic is to add a random effect \( \zeta \) to represent unobserved heterogeneity. Suppose then, that the conditional distribution of \( y \) given \( \zeta \) is indeed Poisson with mean and variance \( \zeta \lambda \). Let us assume that \( \zeta \) has a gamma distribution with parameters \( a = b = \phi \), where \( 1/\phi \) represents the variance of the unobservable. With this information, the marginal distribution of \( y \) can be
derived as follows,

\[
p(y|\lambda) = \int p(y, \zeta|\lambda) d\zeta \\
= \int p(y|\zeta, \lambda)p(\zeta) d\zeta \\
= \int \left(\frac{\zeta^y e^{-\zeta}}{y!} \frac{\phi^\phi}{\Gamma(\phi)} \zeta^{\phi-1} e^{-k\zeta} d\zeta\right) \\
= \frac{\Gamma(\phi + y)}{y!\Gamma(\phi)} \frac{\phi^\phi \lambda^y}{(\lambda + \phi)^{\phi+y}} \int \frac{(\lambda + \phi)^{\phi+y}}{\Gamma(\phi + y)} \zeta^{\phi+y-1} e^{-(\lambda+\phi)\zeta} d\zeta \\
= \frac{\Gamma(\phi + y)}{y!\Gamma(\phi)} \frac{\phi^\phi \lambda^y}{(\lambda + \phi)^{\phi+y}},
\]

which happens to be a negative binomial distribution, denoted by

\[
y \sim \text{NB}(\lambda, \phi).
\]

The negative binomial distribution is best known as the distribution of the number of failures before \( \phi \) successes in a series of Bernoulli trials with common probability of success \( \phi/(\lambda + \phi) \). The mean and variance are \( \lambda \) and \( \lambda + \lambda^2 \phi \), respectively. If \( \phi \) goes to infinity, the Poisson variance is obtained and thus the negative binomial distribution reduces to a Poisson distribution. To model count data as a function of covariates, we can use a negative binomial regression model, which is a generalization of Poisson regression model.

\[
y_i \sim \text{NB}(\mu_i, \phi), \\
\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.
\]

1.3.2 Zero-inflation modeling

Another common problem with count data models, including both Poisson and negative binomial models, is that data often show an excess of zeros that are more than expected. The excess zeros result from either missing of information or limitations of
the sampling effort. One approach to modeling such characteristic is to postulate that there are two latent classes of samples, the “always zero” and the “not always zero,” with the latter having a Poisson or negative binomial distribution. This so-called zero-inflated model, first introduced by Johnson and Kotz (1969), can be formulated as a mixture model, with the first kernel to be degenerated at zero,

\[ y \sim \pi I_0(y) + (1 - \pi)\text{Poi}(\lambda), \quad \text{or} \]

\[ y \sim \pi I_0(y) + (1 - \pi)\text{NB}(\lambda, \phi). \]  

(1.12)

with the weight \( \pi \in [0, 1] \). The model combines a logit model that predicts to which latent class a sample belongs. In this model, there are two kinds of zeros: some are structural zeros from the “always zero” class, and some are random zeros from the count process. The zero-inflated model is very appealing. However, its interpretation is often unclear due to the lack of clarity on the meaning of an “always zero” class.

An alternative approach to modeling excess zeros is to use a logit model to distinguish counts of zero from positive counts, effectively collapsing the count distribution into two categories, and then using a truncated Poisson or negative binomial model, where zero has been excluded, for the positive counts. This approach differs from the zero-inflated models because it assumes that there is only one process by which a zero can be produced while zero-inflated model assumes that there are two different processes that can produce a zero. Statisticians call such a model as the hurdle model. It was first proposed by Cragg (1971) and later developed further by Mullahy (1986). The interpretation of the logit equation is more straightforward because the binary choice is clear. However, interpretation of the Poisson equation is not so obvious because the quantity being modeled, \( \lambda \), is the mean of the entire distribution, and not the mean for those “non-zero counts,” which would be \( \lambda/(1 - e^{-\lambda}) \). For more discussions, I refer to Gurmu and Trivedi (1996) and Dalrymple et al. (2003).
1.4 Overview of projects

Chapters 2 and 3 describe two novel approaches to modeling high-dimensional count data. Both approaches incorporate a feature selection mechanism (in mixture and regression models, respectively) and account for the characteristics (zero-inflation and over-dispersion) observed in count data. I consider such data, with small to moderate sample size $n$, as a $n$-by-$p$ non-negative integer matrix, where the number of feature $p$ is relatively larger than $n$.

In Chapter 2, I develop a novel Bayesian nonparametric model to cluster the samples into homogenous groups while selecting the discriminatory features. Specifically, a binary latent vector is employed to identify a parsimonious set of discriminatory features. Simultaneously, the samples are allocated into an infinite number of components, defined by a Dirichlet process (DP). To account for zero-inflation and over-dispersion observed across samples as well as across features, I develop a zero-inflated Poisson mixture model that incorporates model-based normalization through prior distributions with stochastic mean constraints. For inference, I use MCMC sampling techniques that combine Metropolis-Hastings schemes employed in Bayesian feature selection with sampling algorithms for Bayesian nonparametric models. I first evaluate the performance of our model on simulated data and show that our approach has a competitive edge over existing standard methods for the analysis of count data, in terms of clustering accuracy (i.e. adjusted Rand index). A case study to a bag-of-words benchmark data set is then conducted, where the features are represented by the frequencies of occurrence of each word. The content of Chapter 2 corresponds to the paper “A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints,” which has been submitted to the journal *Statistical Analysis and Data Mining*. This work is co-authored with Michele Guim-
In Chapter 3, I propose a novel Bayesian integrative model to quantify the relationship between multivariate response features that are measured as counts and a large number of predictors. Simultaneously, the model selects a set of discriminatory features as well as a set of “best” predictors. A $K$-component negative binomial mixture regression model is developed, integrating several characteristics. Beyond standard feature selection mechanism, the model includes Markov random field (MRF) prior models that capture structural dependencies among the features. The model also allows the mixture components to depend on selected covariates to capture additional variance observed in the data as well as quantify the association between the response and explanatory variables. Again, I use MCMC sampling techniques with Metropolis-Hastings schemes to infer the parameters of interest. The simulation studies show that employing the MRF prior improves feature selection accuracy. The proposed integrative model is also demonstrated through an application to RNA-Seq gene expression and DNA methylation data for identifying biomarkers in breast cancer and studying the association between these genomic data. This demonstration shows that our model can help biological scientists to understand the regulation of gene expression. The content of Chapter 3 corresponds to the paper “An integrative Bayesian model with spatial feature selection on high-dimensional count data,” which is about to be submitted. This work is co-authored with Alberto Cassese (Maastricht University, Netherlands), Michele Guindani (University of California, Irvine, U.S.A.), and Marina Vannucci (Rice University, U.S.A.).

I conclude the thesis with Chapter 4, which summarizes the two projects.
Chapter 2

A Bayesian zero-inflated Poisson mixture model for clustering and feature selection

2.1 Introduction

Modern data science often involves data sets where the features of interest are measured as counts. For example, text documents are typically summarized by their word frequencies, with the size of the dictionary determining the number of features (see e.g. Griffiths and Steyvers, 2004; Airoldi et al., 2006). In those applications, clustering and feature selection techniques are often employed to provide low-dimensional summaries of the data and investigate the topical content of a text (Airoldi et al., 2010; Airoldi and Bischof, 2016). Similarly, in ecological survey data (see e.g. Gee et al., 1985), species counts are observed on a relatively large number of sites, with the objective of characterizing the different ecosystems across the sites. In biology, sequence data – e.g., SAGE data, RNA-Seq and microbiome data (Blackshaw et al., 2004; Marioni et al., 2008; Chen and Li, 2013) – are represented as a matrix of short sequence tags (e.g. taxonomic units in a microbiome experiment) and corresponding observed reads for several samples. Investigators are typically interested in discovering tags whose abundances significantly differ across samples.

One common characteristic of those studies is that the observed frequency of a feature depends on the sampling effort, or – for the analysis of text documents – the document length. This usually results in data sets characterized by two distinctive
attributes. On the one hand, the data sets contain a high percentage of zero counts. A zero count can either indicate a missing trait in the population or be due to a limited sample. Furthermore, the data sets are highly variable, both with respect to the total number of counts per sample and to the total number of counts per feature. Thus, the observed distributions of counts are typically skewed and overdispersed, since a large number of features are recorded at low frequencies whereas a few features are recorded very frequently. The amount of overdispersion may also vary sample to sample (Witten, 2011; Guindani et al., 2014).

In order to take into account the skewness and heterogeneity of the data, some type of normalization and regularization is often necessary for conducting inference on the features’ occurrence rates (Airoldi et al., 2014; Robinson and Oshlack, 2010; Robinson et al., 2010; Anders and Huber, 2010). For example, Witten (2011) proposes a Poisson log-linear model, where the Poisson intensities are robustly estimated after a normalization step which takes into account the total number of reads observed for each sample and each feature. More recently, Airoldi and Bischof (2016) have proposed a Hierarchical Poisson modeling framework where the word occurrence rate is moderated by the document’s length, and a low-dimensional representation of the data is achieved by means of a set of latent (mixture) components and sparsity-inducing shrinkage priors.

In this chapter, we propose a novel regularization approach for estimating occurrence rates in zero-inflated Poisson mixture models. Zero-inflated Poisson models have often been employed to fit count data characterized by overdispersion and a high number of zeros, both in the econometrics and the statistics literature (Lambert, 1992; Cameron and Trivedi, 2005, 2013). Our proposal is characterized by priors that employ constraints on expected values of both samples’ and features’ scaling pa-
rameters, in order to normalize the information content of each sample. Therefore, our fully Bayesian model does not rely on data-dependent plug-in estimates, typically used in this setting, which *a-priori* condition the inference over unknown parameters on the observed data counts. Our approach is similar in spirit to the recent proposal by Reich et al. (2010), who impose a stochastic constraint on the quantile functions of infinite mixtures, in order to avoid identifiability restrictions in Bayesian nonparametric approaches for quantile regression.

We further regularize the estimation problem by allowing priors that enable feature selection and discriminate samples into clusters. In the analysis of a text document, for example, our approach allows to identify a subset of words which are exclusive to particular topics, automatically balancing the influence of frequency and exclusivity of words across samples by virtue of our model-based inference (Roberts et al., 2016; Airoldi and Bischof, 2016). More specifically, to achieve feature selection we introduce latent discriminatory variables, similarly as in Tadesse et al. (2005) and Raftery and Dean (2006), who proposed the use of latent indicators for variable selection in the context of finite mixtures of Gaussian distributions. Samples are then clustered on the basis of the similarity of the resulting vectors of discriminatory features by means of an infinite Dirichlet process mixture (DPM). Dirichlet process mixtures have often been employed in Bayesian modeling for clustering purposes since they allow to estimate the number of clusters (i.e. the number of mixing components) directly from the data (Muller et al., 2015). Alternatively, one could consider a finite mixture model, and use reversible jump Markov chain Monte Carlo (RJMCMC) to determine the number of components, at the expense of increased computational cost (Richardson and Green, 1997; Tadesse et al., 2005).

In recent Bayesian nonparametric literature, it has become common to employ
an Indian Buffet Process (IBP) characterization to identify non-zero elements in a general matrix of $n$ samples $\times p$ features (Griffiths and Ghahramani, 2011). For example, IBP Compound Dirichlet process models (Williamson et al., 2010) and beta-negative binomial processes (Zhou, 2014) have been introduced for the analysis of count data in topic modeling. Differently than those, the constrained Poisson mixtures (CoPoM) approach we propose is aimed at normalizing the data through the use of mean constraints, and clustering the available samples based on the entire subsets of selected features, rather than assigning single elements of the $n \times p$ matrix to either one of the mixture components (e.g. a topic) or none.

By means of a simulation study, we show how our approach, with prior constraints, improves the accuracy of the clustering performance with respect to more standard approaches for the analysis of count data. We then present an application to a bag-of-words benchmark data set, where the features are represented by the frequencies of occurrence of each word.

The rest of the chapter is organized as follows. In Section 2, we introduce the zero-inflated hierarchical mixture model and discuss the prior formulations. In Section 3, we briefly describe the MCMC algorithm and discuss the resulting posterior inference. In Section 4, we illustrate the performance of our method on simulated data and then present an application to document clustering, based on a bag-of-words benchmark data set. Section 5 concludes the chapter with a discussion of the modeling choices, namely the use of a Poisson likelihood versus alternatives, and with future research directions.
2.2 Zero-inflated Poisson mixture model with feature selection

We consider a $n \times p$ matrix of counts, $y_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, observed on a set of $p$ features and $n$ samples. We assume that a large number of the counts is zero, either because the feature is truly missing in a subset of the population or due to limitations of the sampling effort. Thus, we start by considering a zero-inflated Poisson mixture model, i.e. a mixture model where we constrain one of the kernels to be degenerate at zero,

$$y_{ij} \sim \pi I_0(y_{ij}) + (1 - \pi) \int \text{Poi}(y_{ij}; \lambda) G(d\lambda),$$

where $\pi \in [0, 1]$, $\text{Poi}(y; \lambda)$ denotes a Poisson distribution for the random variable $y$, with expectation $\lambda$, $I_c(\cdot)$ indicates a point mass distribution on $c \in \mathbb{R}$, and $G(\cdot)$ denotes a general mixing distribution, which we use to model the overdispersion of the data. Note that if $G(\cdot)$ is Gamma distributed, then (2.1) defines a zero-inflated Negative Binomial (Cameron and Trivedi, 2013). Alternatively, we can write (2.1) also by introducing latent indicator variables $r_{ij} \sim \text{Bern}(\pi)$, such that if $r_{ij} = 1$ then $y_{ij} = 0$, whereas if $r_{ij} = 0$ then $y_{ij} \sim \int \text{Poi}(y_{ij}; \lambda) G(d\lambda)$.

2.2.1 Feature selection

We envision that only some of the features are relevant to discriminate the $n$ samples into distinct clusters. In particular, we postulate the existence of a latent binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)$, with $\gamma_j = 1$ if the counts associated to feature $j$ are relevant to discriminate among the $n$ samples, and $\gamma_j = 0$ otherwise. Let $p_\gamma = \sum_{j=1}^p \gamma_j$ denote the number of discriminatory features and, correspondingly, let $p - p_\gamma$ indicate the number of non-informative features. Our model formulation assumes that any
count, say $y_{ij}$, that maps to a discriminatory feature is drawn from a zero-inflated Poisson distribution with intensity parameter $\lambda_{ij}$, while counts that map to non-discriminatory features are drawn from a "null" model, which can be characterized as a zero-inflated Poisson distribution with intensity parameter $\lambda_{ij}^0$. After conditioning on the zero-inflation latent indicators $r_{ij}$, we can then specify model (2.1) as follows,

$$y_{ij}|r_{ij}, \gamma_j, \lambda_{ij} \overset{ind}{\sim} \begin{cases} 
\text{Poi}(y_{ij}; \lambda_{ij}) & \text{if } r_{ij} = 0 \text{ and } \gamma_j = 1 \\
\text{Poi}(y_{ij}; \lambda_{ij}^0) & \text{if } r_{ij} = 0 \text{ and } \gamma_j = 0, \\
I_0(y_{ij}) & \text{if } r_{ij} = 1,
\end{cases}$$  (2.2)

with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. A common choice for the prior on the vector $\gamma$ is to assume independent Bernoulli distributions on the individual components, with a common hyperparameter $\omega$, i.e.

$$\gamma_j|\omega \sim \text{Bern}(\omega),$$  (2.3)

which is equivalent to a Binomial prior on the number of discriminatory features, that is $p_\gamma|\omega \sim \text{Bin}(p, \omega)$. The hyperparameter $\omega$ can be elicited as the proportion of features expected a priori to be in the discriminatory set. This prior assumption can be further relaxed by formulating a $\text{Be}(a_\omega, b_\omega)$ hyperprior on $\omega$, which leads to a beta-binomial prior for $p_\gamma$ with expectation $p a_\omega / (a_\omega + b_\omega)$. A uniform prior on $\omega$, obtained by setting $a_\omega = b_\omega = 1$, leads to $E[\omega] = 0.5$. Analogously, we assume $\pi \sim \text{Be}(a_\pi, b_\pi)$ as a prior on the zero-inflation weight.

In order to account for the variability observed across samples and across features, we further parametrize the intensity parameters of the Poisson distributions as the multiplicative effect of three random effects: a) a scaling factor capturing how the sampling effort affects sample-specific occurrences across all features, denoted by $s_i$; b) a scaling factor capturing feature-specific levels across all samples, denoted by $g_j$. 


(e.g. a common usage preposition in the analysis of text documents); and c) a term capturing the occurrence rate for each count, once all the previous global effects have been accounted for, denoted by $d_{ij}$. Specifically, we parametrize the Poisson intensity in (2.2) as $\lambda_{ij} = s_i g_j d_{ij}$, if $\gamma_j = 1$, whereas we set $\lambda_{ij}^0 = s_i g_j d_0$ if $\gamma_j = 0$ in the “null” model. Here, $d_0$ is a parameter capturing homogenous background noise across samples and features. Since background noise should be the result of the variability captured by the scaling factors $s_i$ and $g_j$, and also to ensure the identifiability of all model parameters, we assume $d_0 = 1$, so that the significance of each feature is completely revealed by the values of the random effects $d_{ij}$ across the samples.

Multiplicative characterizations of the Poisson intensity parameter of model (2.2) are typical both in the frequentist as well as in the Bayesian literature to account for latent heterogeneity and overdispersion in count data (e.g. see Witten, 2011; Cameron and Trivedi, 2013; Airoldi and Bischof, 2016; Banerjee et al., 2014, the latter for examples of this specification in spatial statistics). To simplify both the prior specification and the computational algorithms, it is sometimes convenient to reparametrize model (2.2) by using the logarithmic transformations $\tilde{\lambda}_{ij} = \log\{\lambda_{ij}\}$, and, consequently, $\tilde{s}_i = \log\{s_i\}$, $\tilde{g}_j = \log\{g_j\}$ and $\tilde{d}_{ij} = \log\{d_{ij}\}$, such that $\tilde{\lambda}_{ij} = \tilde{s}_i + \tilde{g}_j + \tilde{d}_{ij}$. Under this reparametrization, the base component is characterized by $\tilde{d}_0 = 0$. In Section 2.2.2, we describe a regularizing prior specification for the scaling factors that allows flexible modeling of count data, and also avoids some limiting assumptions commonly made when estimating those models.

2.2.2 Estimation of scaling factors via mean constraints

In Poisson models with a multiplicative intensity $\lambda_{ij}$ as defined in Section 2.2.1, the inferential interest is often limited to the estimation of the occurrence rates $d_{ij}$,
whereas scaling factors are often normalized in order to regularize the inferential problem and ensure the identifiability of the relevant parameters $d_{ij}$. One typical choice is to estimate the scaling factors by means of plug-in estimators based on the observed counts. For example, in the context of the estimation of RNA-seq abundances, Marioni et al. (2008), Mortazavi et al. (2008), and Witten (2011) fix $s_i = \sum_{j=1}^{p} y_{ij} / \sum_{i=1}^{n} \sum_{j=1}^{p} y_{ij}$, so that $\sum_{i=1}^{n} s_i = 1$. Similarly, Anders and Huber (2010) propose $s_i = m_i / \sum_{i=1}^{n} m_i$, where $m_i$ is the median of the distribution of the ratios of the counts for observation $i$ to their geometric mean. As a further example, Bullard et al. (2010) propose taking $s_i = q_i / \sum_{i=1}^{n} q_i$, with $q_i$ the 75th percentile of the counts for observation $i$. Many of the above examples further fix $g_j = \sum_{i=1}^{n} y_{ij}$.

While convenient, the use of plug-in estimates for estimating $s_i$ and $g_j$ has noticeable drawbacks, since a-priori they condition the inference over unknown parameters on the observed data counts. As a consequence, the normalization to account for the sampling effort in a RNA-seq experiment could be determined more by the abilities of the research group than the limitations of the technology. Indeed, such plug-in estimates can be regarded as maximum likelihood estimators in multiple stage approaches and somewhat akin to Empirical Bayes methods, therefore relying on implicit assumptions of exchangeability of the observations, which may not be always justified in practice and can introduce bias in the estimation of posterior uncertainties (Morris, 1983; Gelman, 2008).

One of the contributions of this manuscript is to provide an alternative normalization approach, through the use of priors for the vectors $s = (s_1, \ldots, s_n)$ and $g = (g_1, \ldots, g_p)$, that respectively capture the sampling and feature heterogeneity in the count data, and avoid fixing those values a-priori. A simple choice would be to use conjugate priors, $s_i \sim \text{Ga}(a_s, b_s)$ and $g_j \sim \text{Ga}(a_g, b_g)$. However, the Poisson in-
tensity in (2.2) depends on the product \( s_i g_j \), and further constraints are necessary to allow identifiability of the parameters and simultaneous inference on the occurrence rates \( d_{ij} \). In the context of quantile regression, Reich et al. (2010) impose a stochastic constraint on the quantile functions of infinite mixtures to ensure identifiability of the parameter estimates. Here, we consider priors on \( s_i \) and \( g_j \) that impose normalizing constraints on the expected values, but still provide a flexible estimate of the posterior densities.

For computational convenience, we consider the logarithmic transformation \( \tilde{s}_i = \log \{s_i\} \) and \( \tilde{g}_j = \log \{g_j\} \), so that \( s_i g_j = \exp \{\tilde{s}_i + \tilde{g}_j\} \). We assume that the priors for \( \tilde{s}_i \) and \( \tilde{g}_j \) are mixture distributions,

\[
\tilde{s}_i | \cdot \sim \sum_{m=1}^{M} \phi_m^s f_m^s(\tilde{s}_i | \cdot), \\
\tilde{g}_j | \cdot \sim \sum_{l=1}^{L} \phi_l^g f_l^g(\tilde{g}_j | \cdot),
\]

(2.4)

with \( 0 \leq \phi_m^s, \phi_l^g \leq 1, \sum_{m=1}^{M} \phi_m^s = \sum_{l=1}^{L} \phi_l^g = 1 \), and \( M, L \) are positive integers. The use of mixture distributions allows flexible estimation of the posterior density of \( \tilde{s}_i \) and \( \tilde{g}_j \). We further build each mixture so to satisfy the desired constraint, i.e. \( E[\tilde{s}_i] = c_s \) and \( E[\tilde{g}_j] = c_g \), where \( c_s \) and \( c_g \) are some fixed values. Hoff (2003) demonstrates that any distribution with a mean constraint can be generated by an infinite sum of two-components mixture distributions, where the two-component mixtures are constrained to have the required expected value. Therefore, we assume that each \( f_m^s(\cdot) \) and \( f_l^g(\cdot) \) in (2.4) are themselves a two-component Gaussian mixture, as

\[
f_m^s(\tilde{s}_i | t_m, \eta_m) = t_m N(\eta_m, \sigma_s^2) + (1 - t_m) N \left( \frac{c_s - t_m \eta_m}{1 - t_m}, \sigma_s^2 \right), \\
f_l^g(\tilde{g}_j | q_l, \mu_l) = q_l N(\mu_l, \sigma_g^2) + (1 - q_l) N \left( \frac{c_g - q_l \mu_l}{1 - q_l}, \sigma_g^2 \right), \]

(2.5)

with \( 0 \leq t_m, q_l \leq 1 \). It is immediate to check that the densities in (2.5) satisfy the
desired constraint. Of course, one could consider other prior formulations satisfying that constraint. However, the proposed mixture-of-mixtures prior is attractive because it allows flexibility in the estimation of the unknown $s_i$ and $g_j$'s, by spanning a wide class of distributions, e.g. skewed and multi-modal densities.

Furthermore, if $M = L = \infty$, then (2.4) can be interpreted as Bayesian nonparametric infinite mixtures. In particular, Dirichlet process mixture models have been extensively used in recent literature for flexible density estimation, both for continuous and discrete data (see, e.g. Trippa and Parmigiani, 2011; Kyung et al., 2011; Taddy and Kottas, 2012). The Dirichlet process assumes that the mixing distribution can be written as a discrete random measure, $F(\cdot) = \sum_{k=1}^{\infty} \phi_k I_{\theta^*_k}(\cdot)$, where the weights $\phi_k$ are defined by the (Sethuraman, 1994) stick-breaking construction, i.e.

$\phi_1 = V_1$, $\phi_k = V_k \prod_{u=1}^{k-1} (1 - V_u)$, $V_k \sim \text{Be}(1, \alpha)$, $k = 1, 2, \ldots$, and the atoms $\theta^*_k \sim F_0$, with $F_0$ a baseline parametric model describing the prior expectation of the Dirichlet process. In symbols, we write $F \sim \text{DP}(\alpha, F_0)$. The concentration (or mass) parameter $\alpha$ provides a measure of the precision of the random measure around the baseline parametric model (see, for details on the Dirichlet process, Hjort et al., 2010; Muller et al., 2015). We note that theoretical results on large support and consistency of models based on discrete kernels have not been discussed in the literature. Indeed, given the discreteness of the support on the natural numbers, some technical issues make the derivation of such results more complex than for mixtures of continuous distributions.

We conclude this section by specifying the distributions of the hyper-parameters in (2.5). More specifically, we assume that the two mixtures are characterized by $\eta_m \sim \text{N}(0, \tau_\eta)$, $t_m \sim \text{Be}(a_t, b_t)$, and $\mu_l \sim \text{N}(0, \tau_\mu)$, $q_l \sim \text{Be}(a_q, b_q)$, whereas $\phi^*_m$ and $\phi^*_l$ are obtained according to the stick-breaking construction. Since the aim of this
specification is simply to achieve an automatic normalization of the scaling factors, we further assume $\sigma_s^2 = \sigma_g^2 = 1$.

### 2.2.3 Clustering selected features via Dirichlet process mixtures

In the analysis of count data across multiple samples, one common objective is to characterize and cluster the observed samples into homogenous groups on the basis of the estimated features’ occurrence rates. In this section we provide a method for clustering the $n$ samples, based on the set of selected discriminatory features from Section 2.2.1. We use the superscript $(\gamma)$ to index the set of discriminatory features, characterized by $\gamma_j = 1$ in (2.2). Similarly, $(\gamma^c)$ indicates the set of non-discriminatory features, characterized by $\gamma_j = 0$ and the “null” Poisson distribution with intensity parameter $\lambda_{ij}^0 = s_i g_j d_0$. Thus, each data sample is represented by the $1 \times p$ vector $y_i$ of observations $\{y_{ij}\}$, with $y_i^{(\gamma)}$ and $y_i^{(\gamma^c)}$ indicating the subsets of features corresponding to $\gamma_{ij} = 1$ and $\gamma_{ij} = 0$, respectively. We assume that the subset of selected features $y_i^{(\gamma)}$ can be clustered across samples, by means of a zero-inflated infinite mixture of Poisson distributed components. For that purpose, we introduce an auxiliary set of clustering allocation variables, $z = \{z_1, \ldots, z_n\}$, defined so that $z_i = k$ if and only if the vector of observations $y_i^{(\gamma)}$ belongs to cluster $k$, for some integer $k \geq 1$. Then, we can characterize the likelihood for the selected features as

$$y_{ij}|\gamma_j = 1, z_i = k, r_{ij}, s_i, g_j \xrightarrow{ind} r_{ij} I_0(y_{ij}) + (1 - r_{ij}) \text{Poi}(y_{ij}; \lambda_{ijk})$$

(2.6)

where $\lambda_{ijk} = s_i g_j d_{kj}^*$, with $d_{kj}^* \sim \text{Ga}(a, b)$ being a non-negative occurrence rate for feature $j$, common to all non-zero observations assigned to the $k$-th cluster. We further assume that each $z_i \sim \sum_{k=1}^{\infty} w_k I_{(k)}$, where the weights $w_k$ are defined by the (Sethuraman, 1994) stick-breaking construction, so that (2.6) effectively defines
a zero-inflated conjugate Dirichlet process mixture model. This modeling framework allows us to cluster the samples based on the vectors of selected features, while at the same time the number of clusters is estimated as a by-product of the usual posterior inference. The dimension of the component-specific vector \( d_k^{(\gamma)} = \{d_{k_1}^{\gamma}, \ldots, d_{kp}^{\gamma}\} \) depends on the outcome of the feature selection procedure. As a matter of fact, the joint prior probability of a given allocation of the \( n \) samples into \( K \) groups is given by

\[
p(z_1, \ldots, z_n) = \frac{\alpha^K \Gamma(\alpha) \prod_{k=1}^K \Gamma(n_k)}{\Gamma(\alpha + n)},
\]

that describes the Ewens distribution (Ewens, 1972; Antoniak, 1974; Crane et al., 2016). Our zero-inflated Poisson mixture model provides a more flexible framework for density estimation of overdispersed count data with respect to widely used Negative Binomial models, which are indeed a special case of our framework.

The proposed constrained Poisson mixture model is summarized in Figure 2.1. Since the posterior distribution of the parameters of interest is not available in closed form, we revert to Markov Chain Monte Carlo algorithms for posterior inference, as described in the next section.

### 2.3 Model fitting

We now briefly describe the MCMC algorithm for posterior inference. Our inferential strategy allows to simultaneously infer group structure in the samples while identifying the discriminatory features.
### Mixture model Likelihood:

\[
y_{ij} | \gamma_j = 1, z_i, s_i, g_j, r_{ij} \sim r_{ij} I_0(y_{ij}) + (1 - r_{ij}) \text{Poi}(y_{ij}; \lambda_{ijz_i}) \quad \text{with } \lambda_{ijz_i} = s_i g_j d_{i,j}^z
\]

\[
y_{ij} | \gamma_j = 0, s_i, g_j, r_{ij} \sim r_{ij} I_0(y_{ij}) + (1 - r_{ij}) \text{Poi}(y_{ij}; \lambda_{ij}^0) \quad \text{with } \lambda_{ij}^0 = s_i g_j
\]

### Zero-inflation latent indicator prior:

\[
r_{ij} \sim \text{Bern}(\pi), \quad \pi \sim \text{Be}(a, b)
\]

### Feature selection prior:

\[
\gamma_j \sim \text{Bern}(\omega), \quad \omega \sim \text{Be}(a, b)
\]

### Mixing distribution for selected discriminatory features:

\[
\begin{align*}
  z_i & \sim \sum_{k=1}^{\infty} w_k I_{\{k\}} \\
  d_k^{(\gamma)} & \sim \prod_{\{j: \gamma_j = 1\}} \text{Ga}(a, b) \\
  w_k & = V_k \prod_{u=1}^{k-1} (1 - V_u), \quad V_k \sim \text{Be}(1, \alpha) \\
  \alpha & \sim \text{Ga}(a, b)
\end{align*}
\]

### Priors on subject- and feature-specific scaling factors:

\[
\tilde{s}_i | \nu_i, \epsilon_i, t, \eta \sim \sum_{m=1}^{M} \phi_m^s \left[ t_m N(\eta_m, \sigma_m^2) + (1 - t_m) N \left( \frac{\epsilon_s - t_m \eta_m}{1 - t_m}, \sigma_s^2 \right) \right]
\]

\[
\phi_m^s = V_m \prod_{u=1}^{m-1} (1 - V_u), \quad V_m \sim \text{Be}(a_m, b_m)
\]

\[
\eta_m \sim N(0, \tau_\eta)
\]

\[
t_m \sim \text{Be}(a_t, b_t)
\]

\[
\tilde{g}_j | \xi_j, \psi_j, q, \mu \sim \sum_{l=1}^{L} \phi_l^g \left[ q_i N(\mu_l, \sigma_l^2) + (1 - q_i) N \left( \frac{\epsilon_g - q_l \mu_l}{1 - q_l}, \sigma_g^2 \right) \right]
\]

\[
\phi_l^g = V_l \prod_{u=1}^{l-1} (1 - V_u), \quad V_l \sim \text{Be}(a_l, b_l)
\]

\[
\mu_l \sim N(0, \tau_\mu)
\]

\[
q_i \sim \text{Be}(a_q, b_q)
\]

### Fixed hyperparameters:

\[
a, b, a_s, b_s, a_\omega, b_\omega, a_\epsilon, b_\epsilon, c_s, c_g, M, L, \sigma_s, \sigma_g, a_m, b_m, a_t, b_t, \tau_\eta, \tau_\mu, a_l, b_l, a_q, b_q
\]

Figure 2.1: Hierarchical formulation of the proposed constrained Poisson mixture (CoPoM) model.
2.3.1 MCMC algorithm

Our primary interest lies in the identification of the discriminatory features, via the vector $\gamma$, and the estimation of the sample clustering allocations, via the vector $z$. For this, we design a Markov chain Monte Carlo (MCMC) algorithm based on Metropolis search variable selection algorithms (George and McCulloch, 1997; Brown et al., 1998) and Gibbs sampling methods for Dirichlet process mixture models (Neal, 2000). We also sample the sample-specific and feature-specific scaling factors, $s$ and $g$. We give full details of our MCMC algorithm in the Appendix and report here a brief description of the most relevant updates.

**Update of $\gamma$:** This is done via an *add-delete-swap* algorithm. In this approach, a new candidate vector, say $\gamma^{\text{new}}$, is generated by randomly choosing between two types of moves. For the add/delete move, we select at random one of the elements in the current vector, say $\gamma^{\text{old}}$, and change its value from 0 to 1, or vice versa. For the swap move, we select two elements in $\gamma^{\text{old}}$ with different inclusion status and swap their values. Then, the Metropolis-Hastings ratio can be written as

$$m_{\text{MH}} = \frac{p(\gamma^{\text{new}}|z, \tilde{s}, \tilde{g}, R, X) J(\gamma^{\text{old}}|\gamma^{\text{new}})}{p(\gamma^{\text{old}}|z, \tilde{s}, \tilde{g}, R, X) J(\gamma^{\text{new}}|\gamma^{\text{old}})} ,$$

where $J(\cdot|\cdot)$ indicates the proposal probability distribution for the selected move. The move is accepted with probability $\min(1, m_{\text{MH}})$. We should notice that the feature selection and the cluster structure are determined simultaneously in the MCMC algorithm. Therefore, to improve mixing, it is necessary to allow the selection to stabilize for any visited cluster configuration. As suggested in Kim et al. (2006), we repeat the above Metropolis step $E = 20$ times within each iteration. In the applications of this chapter, no improvement in the MCMC performance was noticed beyond this value.
Update of $z$: Since we have assumed a conjugate baseline parametric distribution, $\prod_{j:y_j=1} \text{Ga}(a, b)$, we can integrate analytically over the cluster-specific parameters $d_k^{(\gamma)}$ and directly sample the cluster assignment indicators of the selected features, $z$, according to Algorithm 3 of Neal (2000). More specifically, the Gibbs sampler iteratively samples the full conditionals,

$$p(z_i|z_{-i}, \gamma, \tilde{s}, \tilde{g}, R, X) = \begin{cases} \frac{n_{k,-i}}{n-1+\alpha} f(y_i|z_i = k, z_{-i}, \gamma, \tilde{s}, \tilde{g}, R, X_{-i}) & \text{for } z_i = k; k = 1, \ldots, K_{-i}, \\ \frac{\alpha}{n-1+\alpha} f(y_i|\gamma, \tilde{s}_i, \tilde{g}_i, r_i) & \text{for } z_i = K_{-i} + 1, \end{cases}$$

(2.8)

where $z_{-i}$ denotes all the elements in $z$ excluding the $i$-th one, $n_{k,-i}$ is the size of cluster $k$ in $z_{-i}$, and $K_{-i}$ is the number of unique values in $z_{-i}$. Note that $f(y_i|z_i = k, z_{-i}, \gamma, \tilde{s}, \tilde{g}, R, X_{-i})$ is the integrated likelihood, with updated $d_k^{(\gamma)}$ based on its prior and all observations except the $i$-th one. See details in the Appendix.

Update of $s$ and $g$: The prior distribution for the scaling factors $s$ and $g$ is a Dirichlet process mixture on the log-transformed values $\tilde{s}$ and $\tilde{g}$. Since the mixture of log-normal distribution that we have assumed as baseline measure in the Dirichlet process is not conjugate to the Poisson likelihood, we consider a finite truncation of the Dirichlet process, so that $M$ and $L$ in (2.4) are large but finite (Ishwaran and James, 2001). Such a choice allows to simplify computations considerably, and still achieves flexible estimate of posterior densities, such as those commonly obtained in a Bayesian nonparametric framework. We employ Metropolis-Hastings steps for all $\tilde{s}_i$’s and $\tilde{g}_j$’s.

Update of $R$: The full conditional for the zero-inflation latent indicators takes into account that we only need to update those $r_{ij}$’s that correspond to zero counts.
For positive counts, necessarily \( r_{ij} = 0 \). We use a Metropolis-Hasting within Gibbs sampling step for updating \( r_{ij} \) and \( \pi \), after sampling \( d_{kj}^* \) for those features corresponding to \( \gamma_j = 0 \). Details are given in the Appendix.

### 2.3.2 Posterior inference

We obtain posterior inference on the parameters by post-processing of the MCMC samples after burn-in. We start by obtaining a probabilistic assessment of the cluster allocations by analyzing the MCMC samples of \( z \). One way to summarize the posterior distribution of \( z \) is via the maximum-a-posteriori (MAP) estimate that can be calculated as

\[
\hat{z}_{\text{MAP}} = \arg\max_{1 \leq b \leq B} p(z^{(b)} | \gamma^{(b)}, \mathbf{s}^{(b)}, \mathbf{g}^{(b)}, R^{(b)}, \mathbf{X})
\]

\[
= \arg\max_{1 \leq b \leq B} p(\mathbf{z}^{(b)}) \prod_{i=1}^{n} f(y_i | z_i^{(b)}, z_{-i}, \gamma^{(b)}, \mathbf{s}^{(b)}, \mathbf{g}^{(b)}, R^{(b)}, \mathbf{X}_{-i}),
\]

with \( b = 1, \ldots, B \) indicating the MCMC iterations, after burn-in, and where the marginal posterior probability that sample \( i \) is allocated to cluster \( k \) can be calculated through Equation (2.8). An alternative estimate relies on the computation of a matrix of posterior pairwise probabilities of co-clustering, i.e. the probabilities that observation \( i \) and observation \( i' \) are assigned to the same cluster, \( p_{ii'} = p(z_i = z_{i'} | \gamma, \mathbf{s}, \mathbf{g}, R, \mathbf{X}) \), as suggested by (Dahl, 2006), among others. These probabilities can be estimated by computing empirical frequencies of co-clustering based on the MCMC samples, resulting in an \( n \times n \) symmetric pairwise probability matrix (PPM). Then, a point estimate for the cluster memberships, \( \hat{z}_{\text{PPM}} \), is obtained by minimizing the sum of squared deviations of its association matrix from the PPM, i.e.

\[
\hat{z}_{\text{PPM}} = \arg\min_{z} \sum_{i < i'} [ I(z_i = z_{i'}) - p_{ii'} ]^2.
\]
As for feature selection, the MAP estimate of $\gamma$ can be obtained by enumerating all visited MCMC samples $\gamma^{(b)}$ and then considering the set of features that maximizes the posterior density. Alternatively, we can estimate the marginal posterior probability of inclusion (PPI) of single features as the proportion of MCMC iterations, after burn-in, in which the corresponding $\gamma_j$ were equal to 1, that is

$$PPI(j) = \frac{\sum_{u=1}^{B} (\gamma_j^{(b)}|z^{(b)}, s_j^{(b)}, g_j^{(b)}, r_j^{(b)}, x_{.j})}{B}. $$

A point estimate of $\gamma$ is then obtained by identifying those PPI values that exceed a given threshold. The optimal threshold is typically chosen based on a decision theoretic criterion, e.g., to maximize power under a constraint on the number of false positives (Müller et al., 2004; Guindani et al., 2009).

### 2.4 Applications

We first explore performances on simulated data and then show results on a bag-of-words benchmark data set. We also demonstrate the superiority of our constrained Poisson mixture (CoPoM) model over other widely adopted methods for the analysis of overdispersed count data.

#### 2.4.1 Simulation study

Data were generated with $n = 30$ samples and $K = 3$ clusters. We then simulated observations from a zero-inflated mixture of Poisson distributions, assuming three mixture components and $p_\gamma = 50$ discriminating features,

$$y_{ij} \sim 0.25 I_0(y_{ij}) + 0.75 \left\{ I(1 \leq i \leq 6) \text{Poi}(s_i g_j d_{i1}^*) ight\} + I(7 \leq i \leq 21) \text{Poi}(s_i g_j d_{i2}^*) + I(22 \leq i \leq 30) \text{Poi}(s_i g_j d_{i3}^*),$$

where the first six observations were drawn from the first distribution, the next 15 from the second and the last nine from the third distribution. We added 950 noisy features, which we generated from a zero-inflated Poisson model, $y_{ij} \sim 0.25 I_0(y_{ij}) + 0.75 \text{Poi}(s_i g_j)$. We simulated the $s_i$’s as independent and identically distributed $s_i \sim U(0.5, 1.5)$ and the $g_j$’s as $g_j \sim \text{Exp}(1/3)$. For the mixture components, we simulated $d_{kj}, k = 1, 2, 3$ from a standard normal distribution $N(0, \sigma^2)$ with $\sigma^2 = 1$.

As for hyperparameter settings, we used the following default settings. We set the hyperparameters that control the mixture DP prior to $a = b = 1$ and $a_\alpha = b_\alpha = 1$.

As for the Beta prior on the feature selection parameter $\omega$, we set $a_\omega = 0.2, b_\omega = 1.8$.

For the priors on $s_i$’s and $g_j$’s, we use the following default settings: $M = n/2 = 15, L = p/2 = 500, c_s = c_g = 0, \sigma_s = \sigma_g = 1, \tau_\eta = \tau_\mu = 1, a_m = b_m = 1, a_l = b_l = 1,$
Figure 2.3: Simulated data: Marginal posterior probabilities of inclusion $p(\gamma_j = 1|\cdot)$ with the red dots indicating the truly discriminatory features.

$a_t = b_t = 1$, and $a_q = b_q = 1$. Results we report below were obtained by running one MCMC chain with 10,000 iteration, discarding the first 1,000 as burn-in. We started the chain from a model with 2 randomly chosen $\gamma_j$’s set to 1 and with each observation assigned to a different cluster.

We first describe posterior inference on the relevant parameters as a result of our normalization and regularization approach. Figure 2.2 shows the heatmap of the pairwise posterior probabilities, $p(z_i = z_{i'}|\gamma, \tilde{s}, \tilde{g}, R, X)$, of allocating observations $i$ and $i'$ to the same cluster, after burn-in. It is evident from the map that the inspection of the highest posterior allocation probabilities allows to reconstruct the true allocation structure quite well.

As for the feature selection, Figure 2.3 shows the marginal posterior probability
of inclusion (PPI) of each feature \( p(\gamma_j = 1|z, \tilde{s}, g_j, r_j, x_j) \), after burn-in. The red dots indicate the truly discriminatory features. A threshold of 0.5 on the marginal probabilities results in a median model that includes 18 features, all of which are included in the set of discriminatory features used in the data generation process. Figure 2.4 shows the estimated posterior distribution of the parameter \( d_{ij} \), for two of the discriminating features. These were obtained by post-MCMC composition sampling of the parameters as in Guindani et al. (2014). Notice that the distributions are skewed, with high probability mass on low counts and small mass on extreme values, confirming that our modeling approach is able to capture the general characteristics of the simulated data.

In order to quantify the accuracy of our algorithm, and to compare its performances with other methods available in the literature, we looked at results under different scenarios in terms of sample size, \( n \), number of noisy features, \( p - p_\gamma \), and cluster dispersion, \( \sigma \). In particular, we considered \( \sigma^2 = 1, 0.5, n = 21, 99 \) (equally distributed among the \( K = 3 \) mixture components) and \( p = 200, 1000 \) (with \( p_\gamma = 50 \) discriminating features). For each of these \( 2 \times 2 \times 2 \times 50 = 400 \) scenarios, we simulated 50 data sets and ran our MCMC algorithm with the default settings described above.

For the analysis of the clustering results, we quantified performance via the adjusted Rand index (ARI, Hubert and Arabie, 1985), a variant of the Rand index (Rand, 1971). Let \( A = \sum_{i>\iota'} I(z_i = z_{\iota'})I(\hat{z}_i = \hat{z}_{\iota'}) \) be the number of pairs of observations that belong to the same group in both the true clustering and the estimated one; \( B = \sum_{i>\iota'} I(z_i = z_{\iota'})I(\hat{z}_i \neq \hat{z}_{\iota'}) \) be the number of pairs which belong to the same group in the true clustering but different groups in the estimated one; \( C = \sum_{i>\iota'} I(z_i \neq z_{\iota'})I(\hat{z}_i = \hat{z}_{\iota'}) \) be the number of pairs in different groups in the true partition but assigned to the same group in the estimated one;
Figure 2.4: Simulated data: Estimated mixing measures $E[d_{ij}]$ for two of the discriminatory features.
Table 2.1: Simulated data (from zero-inflated Poisson mixtures): Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages *edgeR* and *PoiClaClu*, and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Competing methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>n</em></td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
</tr>
<tr>
<td>99 200 1.00</td>
<td>0.412 (0.129)</td>
</tr>
<tr>
<td>99 200 0.50</td>
<td>0.257 (0.100)</td>
</tr>
<tr>
<td>99 1000 1.00</td>
<td>0.307 (0.125)</td>
</tr>
<tr>
<td>99 1000 0.50</td>
<td>0.127 (0.103)</td>
</tr>
<tr>
<td>21 200 1.00</td>
<td>0.492 (0.194)</td>
</tr>
<tr>
<td>21 200 0.50</td>
<td>0.287 (0.155)</td>
</tr>
<tr>
<td>21 1000 1.00</td>
<td>0.240 (0.131)</td>
</tr>
<tr>
<td>21 1000 0.50</td>
<td>0.135 (0.095)</td>
</tr>
</tbody>
</table>
\[ D = \sum_{i > i'} I(z_i \neq z_{i'}) I(\hat{z}_i \neq \hat{z}_{i'}) \], the number of pairs assigned to different groups in both the truth and the estimate. Then the RI is defined as

\[ RI = \frac{A + D}{A + B + C + D}, \]

and the ARI as

\[ ARI = \frac{\binom{n}{2} (A + D) - [(A + B)(A + C) + (C + D)(B + D)]}{\binom{n}{2}^2 - [(A + B)(A + C) + (C + D)(B + D)]}. \]

The Rand index yields values between 0 and 1, while the ARI can yield negative values (Santos and Embrechts, 2009). The larger the index, the more accurate the clustering result.

For comparison, we selected two commonly employed estimation methods for count data, based Poisson mixtures and implemented in the freely available R packages `edgeR` (Robinson et al., 2010) and `PoiClaClu` (Witten, 2011). Both `edgeR` and `PoiClaClu` incorporate plug-in estimates of scaling factors for normalization purposes: `edgeR` considers a negative binomial likelihood, whereas `PoiClaClu` aims at clustering count data by using a regularized Poisson log linear model. Unlike our modeling approach, which allows to directly estimate the cluster assignments, through the latent \( \hat{z} \), those methods do not provide individual allocation estimates, but rather yield a dissimilarity matrix that can be transformed into a tree via hierarchical clustering.

In order to make the comparison with our CoPoM model feasible, we considered those estimates that achieved the maximum ARI values. Table 2.1 reports results on clustering performances of all methods in terms of ARI values, averaged over the 50 replicates, under the different simulated scenarios. For our method, we report the
Table 2.2: Simulated data (from Poisson mixtures): Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages edgeR and PoiClaClu, and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Competing methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>edgeR</td>
</tr>
<tr>
<td>$n$</td>
<td>$p$</td>
</tr>
<tr>
<td>99</td>
<td>200</td>
</tr>
<tr>
<td>99</td>
<td>200</td>
</tr>
<tr>
<td>99</td>
<td>1000</td>
</tr>
<tr>
<td>99</td>
<td>1000</td>
</tr>
<tr>
<td>21</td>
<td>200</td>
</tr>
<tr>
<td>21</td>
<td>200</td>
</tr>
<tr>
<td>21</td>
<td>1000</td>
</tr>
<tr>
<td>21</td>
<td>1000</td>
</tr>
</tbody>
</table>
results obtained by using the PPM estimates for cluster allocation. The MAP estimates performed similarly (not shown). In all replicates and scenarios considered, we generated data from a zero-inflated Poisson model, fixing $\pi = 0.25$. The percentage of observed zeros was around 44%. Results show that CoPoM consistently outperforms competing methods in terms of clustering accuracy. This is to be expected, since the competing methods do not incorporate the variable selection, and one might expect the inclusion of noisy features to mask the recovery of the true clustering structure.

Table 2.2 shows the performances of our algorithm when data were generated from a Poisson mixture with three components (i.e. $\pi = 0$). Also in this setting, our zero-inflated Poisson model performs favorably or similarly with respect to the other methods in all cases. The widely used edgeR method shows the worst performances, especially for weakening signal strength, i.e. decreasing $\sigma$ or increasing $p$, whereas our CoPoM method and PoiClaClu show the best performances.

Since our CoPoM model also allows the selection of a subset of discriminatory features, we quantified its performances in terms of averaged false positive rate (FPR) and true positive rate (TPR) achieved under the different simulated scenarios, for different values of the threshold on the PPIs. Results are reported in Table 2.3, and the corresponding receiver operating characteristic (ROC) curves are shown in Figure 2.5. Each sub-figure, corresponding to different values of $n$ and $p$, shows that the estimate becomes more accurate with increasing separation between the clusters, captured by the between-cluster variability parameter $\sigma$.

2.4.2 Application to the analysis of bag-of-words data

Bag-of-words data sets report the frequencies of occurrence of each word in a text document. Clustering of documents on the basis of a subset of relevant words is
Table 2.3: Simulated data: Average false positive rates (FPRs) and true positive rates (TPRs), achieved by the CoPoM model under different simulated scenarios, for different values of the threshold, $\tau$, on the PPIs.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\tau = 0.2$</th>
<th>$\tau = 0.4$</th>
<th>$\tau = 0.6$</th>
<th>$\tau = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CoPoM (PPI)</td>
<td>CoPoM (PPI)</td>
<td>CoPoM (PPI)</td>
<td>CoPoM (PPI)</td>
</tr>
<tr>
<td>$n$</td>
<td>$p$</td>
<td>$\sigma^2$</td>
<td>FPR</td>
<td>TPR</td>
</tr>
<tr>
<td>99</td>
<td>200</td>
<td>1.00</td>
<td>0.075</td>
<td>0.888</td>
</tr>
<tr>
<td>99</td>
<td>200</td>
<td>0.50</td>
<td>0.057</td>
<td>0.803</td>
</tr>
<tr>
<td>99</td>
<td>1000</td>
<td>1.00</td>
<td>0.008</td>
<td>0.803</td>
</tr>
<tr>
<td>99</td>
<td>1000</td>
<td>0.50</td>
<td>0.006</td>
<td>0.677</td>
</tr>
<tr>
<td>21</td>
<td>200</td>
<td>1.00</td>
<td>0.354</td>
<td>0.872</td>
</tr>
<tr>
<td>21</td>
<td>200</td>
<td>0.50</td>
<td>0.269</td>
<td>0.706</td>
</tr>
<tr>
<td>21</td>
<td>1000</td>
<td>1.00</td>
<td>0.054</td>
<td>0.579</td>
</tr>
<tr>
<td>21</td>
<td>1000</td>
<td>0.50</td>
<td>0.044</td>
<td>0.431</td>
</tr>
</tbody>
</table>
Figure 2.5: Simulated data: Receiver operating characteristic (ROC) curves, for different values of the threshold on the PPIs, obtained by CoPoM under different simulated scenarios. The bold curve corresponds to $\sigma^2 = 1$, whereas the dashed curve corresponds to $\sigma^2 = 0.5$. 
one of the most common tasks when analyzing bag-of-words data. Here we illustrate the performance of the CoPoM approach using the widely employed Brown Corpus (Kucera and Francis, 1967). In linguistics, a corpus defines a large and structured collection of texts, which are often used for conducting statistical analyses and tests of hypotheses about a linguistic variety or particular characteristics of a language. The Brown Corpus consists of 500 samples, distributed across 15 genres. Each sample begins at a random sentence-boundary in the article and continues up to the first sentence boundary after 2,000 words. The total vocabulary is about 50,000 words and half of them occur equal or less than once in the corpus. Thus, the data are typically characterized by an excessive number of zeros and overdispersion.

For the application of this chapter, we selected a subset of the Brown Corpus, composed of five sports reportage, three society reportage, seven spot news, and seven editorials, whose length range from 2,200 to 2,374 words, for a total of 22 texts and 8,826 features. The data is quite sparse, as about 90% of the counts are zeros.

We report results obtained by running the MCMC algorithm described in Section 2.3.1 with the same default settings used in the simulations. Figure 2.6 shows the posterior inference on the number of clusters, more specifically the MCMC trace plot across iterations (top) and the resulting estimate of the posterior distribution of $K$, after burn-in (bottom). Figure 2.7 reports the results of the feature selection. More specifically, Figure 2.6 (top) reports the trace plot for the total number of included features, whereas Figure 2.6 (bottom) shows the estimated marginal posterior probabilities of inclusion (PPIs) of each single feature, $p(\gamma_j = 1|z, \tilde{s}, g_j, r_j, x_j)$, after burn-in. A threshold of 0.5 on the marginal probabilities results in a median model that includes 203 features (2.3% of the total). Finally, Figure 2.8 shows the heatmap of the pairwise posterior probabilities of co-clustering, $p(z_i = z_j|\gamma, \tilde{s}, \tilde{g}, R, X)$. Table
Table 2.4: Brown Corpus data: Pairwise probability matrix (PPM) estimates of sample allocation \( \hat{z} \).

<table>
<thead>
<tr>
<th>( \hat{z}_{\text{PPM}} )</th>
<th>Sports</th>
<th>Society</th>
<th>Spot news</th>
<th>Editorial</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1 1</td>
<td>2 2 2</td>
<td>3 3 4 4 5</td>
<td>3 5</td>
<td>6 6 6 6 6 6</td>
<td>0.788</td>
</tr>
</tbody>
</table>

2.4 summarizes the posterior sample allocations \( \hat{z} \), based on the PPM estimate, and the corresponding ARI value. Our results suggest that the proposed CoPoM model is able to roughly recover distinctive features of the 4 broad categories.

For comparison, we already pointed out that the R packages edgeR and PoiClaClu do not provide a single point estimate, \( \hat{z} \), but yield a dissimilarity matrix that can be used as input in a hierarchical clustering algorithm. Within our approach, the squared Euclidean distance between each pair of observations, based on the selected subset of discriminatory features, can be defined as

\[
d(\mathbf{y}_i^{(\gamma)}, \mathbf{x}_{p'}^{(\gamma)}) = \sqrt{\sum_{\{i:j=1\}} \left( d^*_{z_{ij}} - d^*_{z'_{ij}} \right)^2},
\]

and an estimate of \( d^*_{kj} \) can be computed as

\[
\hat{d}^*_{kj} = \frac{\sum_{\{i:z_i=k, r_{ij}=0\}} y_{ij}}{\sum_{\{i:z_i=k, r_{ij}=0\}} \hat{s}_i},
\]

with \( \hat{z}_{\text{PPM}} \) and \( \hat{s} \) and \( \hat{g} \) the MAP estimates of the parameters. Note that the distance between each pair of observations within the same cluster is zero. Figure 2.9 shows the dendrogram of the hierarchical cluster analysis for the CoPoM model and the two R packages edgeR and PoiClaClu, employing the Euclidean distance and the complete agglomeration method. Our CoPoM model and Witten’s methods appear to be the most efficient to separate the four genres.
Figure 2.6: Brown Corpus data: Trace plot of the number of clusters $K$ (top); Histogram of the number of clusters $K$ (bottom).
Figure 2.7: Brown Corpus data: Trace plot of the number of selected variables $p_\gamma$ (top); Marginal posterior probabilities of inclusion $p(\gamma_j = 1 | \cdot)$ (bottom).
Figure 2.8: Brown Corpus data: Heatmap of the pairwise posterior probabilities $p(z_i = z_j | \cdot)$. 
Figure 2.9: Brown Corpus data: The hierarchical clustering dendrograms using the Euclidean distance measure and the complete agglomeration method achieved by CoPoM, PoiClaClu and edgeR, respectively.
2.5 Conclusion

In this chapter, we have introduced a zero-inflated Constrained Poisson Mixture (CoPoM) model for the analysis of count data, where multiple features are observed as counts over a number of samples. We assume that the data are characterized by an excessive number of zeros and by overdispersion, which is a large number of features do not get frequently observed, whereas a few features are characterized by large occurrence rates in one or more samples. We assume that the amount of overdispersion may also vary from sample to sample. Our constrained Poisson mixture (CoPoM) approach proposes to regularize the estimation of the feature occurrence rates by means of a model-based normalization approach, which employs prior distributions with mean constraints that ensure the identifiability of all model parameters, and simultaneously, by incorporating a feature selection mechanism to identify a parsimonious set of discriminatory features and infer group structures in the samples.

When applied to simulated data, the CoPoM model has shown improved accuracy of the clustering performance with respect to more standard approaches. We have also presented an application to a bag-of-words benchmark data set, where the data matrix is represented by the frequencies of word occurrences in multiple documents. We have shown the good performance of our Poisson mixture model with mean constraints in terms of feature selection and the clustering of the samples into larger topical groups.

Various alternative models with respect to Poisson mixtures have recently been proposed for the analysis of overdispersed count data. For example, in the Bayesian nonparametric literature, rounded Gaussian kernel models have been introduced as a more flexible and robust choice for analyzing both underdispersed and overdispersed count data (Canale and Dunson, 2011; Canale and Prünster, 2016). To date those methods have been developed mostly for density estimation. In our manuscript,
instead, we are concerned with the estimation and comparison of features’ occurrence rates across multiple samples, and feature selection, which require some degree of regularization due to the nature of the data generating mechanism. Future work will aim at exploring how our inferential aims can be comprised into the domain of those more recent approaches.

Although here we have focused on the analysis of text documents, our methodology is quite general. In particular, the class of prior distributions with mean constraints that we propose can be successfully employed on other types of high-dimensional count data sets, such as those encountered in ecology, genomics and spatial statistics, where the normalization of Poisson intensities is commonly employed to account for the overdispersion and heterogeneity observed across samples and across features. Future work will explore how the approaches presented here can be extended to the analysis of multivariate vectors of dependent count data, observed spatially or longitudinally in time.
Chapter 3

A Bayesian negative binomial mixture regression model with feature selection

3.1 Introduction

In recent years, RNA sequencing (RNA-Seq, also known as high throughput or next-generation sequencing) has emerged as a powerful biotechnology for quantifying gene expression (see, e.g. Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm and Landry, 2009; Wang et al., 2009; Pepke et al., 2009). RNA-Seq data involve non-negative counts in the form of numbers of reads observed in the corresponding region of interest (e.g. gene or exon) after genome mapping. Compared to its predecessor, the microarray, RNA-Seq does not suffer from cross-hybridization and poor quantification of low- and high-expressed genes (Kukurba and Montgomery, 2015). Unlike microarray data which are continuous and can be easily modeled or normalized, RNA-Seq count data has led to a need for specialized methods for the analysis of skewness and heterogeneity. Generally speaking, the research interests in RNA-Seq data analysis include data normalization (Mortazavi et al., 2008; Bullard et al., 2010; Hansen et al., 2012), differentially expressed features detection (Anders and Huber, 2010; Robinson and Oshlack, 2010; Wu et al., 2013), sample classification and clustering (Berninger et al., 2008; Witten, 2011), and feature clustering (Viallefont et al., 2002; Cai et al., 2004; Rau et al., 2011). To the best of our knowledge, there exist few rigorous integrative modeling approaches to study the relationship between RNA-Seq
and other related genomic data, such as DNA methylation data.

Required for embryonic development (Li et al., 1992), DNA methylation is a process by which methyl groups are added to DNA. It typically acts to suppress gene transcription when located in a gene promoter. In recent years, more and more genetic studies have revealed that DNA methylation plays an important role in cancer since it can influence gene expression (Egger et al., 2004). Several studies (Jaenisch and Bird, 2003; Murrell et al., 2005) have investigated the causal relationship between DNA methylation and gene regulation. Song et al. (2005) and Kitamura et al. (2007) showed the correlation between more than 150 tissue-specific differentially methylated regions and the expression of genes associated with these regions. Futscher et al. (2002) and Ferrón et al. (2011) linked DNA methylation to tissue-specific gene expression by studying the promoter regions of a small subset of imprinted genes in mice. Xie et al. (2011) first conducted an integrative analysis by correlating DNA methylation variation between human tissues with gene expression levels. Their results indicate that DNA methylation influences tissue differentiation via regulating gene expression. As the field of epigenomics expands to study multiple diverse normal and pathological processes, it becomes increasingly significant to understand the role that global genome-wide DNA methylation patterns play in influencing global RNA-Seq gene expression. However, understanding of DNA methylation in regulating gene expression in a quantitative way is still limited.

Motivated by an integrative analysis of RNA-Seq and DNA methylation data in breast cancer study, in this chapter, we develop a hierarchical mixture model for studying the association between multivariate response features, measured as counts, and a set of covariates. Our model includes several innovative characteristics. First, it incorporates the selection of features that discriminate samples into different groups.
Second, the feature selection process further incorporates structural dependencies via Markov random field (MRF) prior models. Take RNA-Seq data as an example, the relationships among the features can be described by an existing gene-gene network. Third, the model allows the mixture components to depend on selected covariates. Applied to the data in a case study on breast cancer, the model tends to include only a small number of DNA probes to explain the variation of RNA-Seq gene expressions. Last but not least, although this project is motivated by a genome-wide association study, the proposed method is more generally applicable. For instance, our model can also be used to model the effect of weather on insurance claims as in Scheel et al. (2013).

The remainder of the chapter is organized as follows. In section 2, we introduce our modeling framework and discuss the MRF prior formulation. In section 3, we briefly describe the MCMC algorithm and discuss the resulting posterior inference. In Section 4, we first assess performances of our proposed model on both simulated and synthetic data and then investigate results on data from a case study on breast cancer. Section 5 concludes the chapter with some remarks on future research directions.

### 3.2 Model

The primary goal of the proposed hierarchical mixture model is to identify a small subset of features that can best distinguish groups while selecting the associated covariates that can best explain each feature. Without loss of generality, we first depict the three required inputs of our model as follows. We consider the high-dimensional count data $\mathbf{Y}$, observed on a set of $n$ observations and $p$ features (e.g. RNA-Seq genes), as a $n \times p$ matrix of counts, with each element denoted by $y_{ij}$. Their associated covariate matrix $\mathbf{X}$ is denoted by a $n \times R$ matrix of numerical values, with
each element denoted by $x_{ir}$. That is, there are $R$ possibly explanatory regressors (e.g. DNA probes) for each observation. The sample allocation (e.g. cancer types or stages) is also available and encoded in a $n \times 1$ latent vector $z$, with $z_i = k, k = 1, \ldots, K$ indicates that observation $i$ belongs to group $k$. The major components of our model are described below and the hierarchical formulation is summarized at the end of this section.

3.2.1 A negative binomial mixture model with feature selection

We start to model the over-dispersed counts by a negative binomial model with feature selection,

$$y_{ij} | z_i = k, \gamma_j \sim \begin{cases} 
\text{NB}(y_{ij}; \lambda_{ijk}, \phi_j) & \text{if } \gamma_j = 1 \\
\text{NB}(y_{ij}; \lambda_{i0}, \phi_j) & \text{if } \gamma_j = 0,
\end{cases} \quad (3.1)$$

where $\text{NB}(y; \lambda, \phi)$ denotes a negative binomial distribution for the random variable $y$, with expectation $\lambda$ and dispersion $1/\phi$. With this parametrization, the variance can be written as $\lambda + \lambda^2/\phi$. Note that increase of $\phi$ towards infinity corresponds to a Poisson distribution with both expectation and variance equal to $\lambda$. We envision that some of the features could discriminate the $n$ samples into $K$ distinct groups. With this assumption, a $p \times 1$ binary latent variable $\gamma$ is introduced to indicate the discriminatory features, with $\gamma_j = 1$ if the count associated with feature $j$ contributes to the classification of the $n$ samples, and $\gamma_j = 0$ otherwise. We denote with $p_\gamma = \sum_{j=1}^p \gamma_j$ and $p - p_\gamma$ the number of discriminatory and non-informative features, respectively. Our model formulation assumes that any count that maps to a discriminatory feature and belongs to group $k$ is drawn from a negative binomial distribution with mean $\lambda_{ijk}$, while counts that map to non-discriminatory features are drawn from a negative
binomial distribution with mean $\lambda_{ij0}$, which is the “null” model. Following the independent assumption among the features (Tadesse et al., 2005), we define the full data likelihood that allows to separate the discriminating features from the noisy ones as

$$f(Y|\gamma, z) = \prod_{\{j: \gamma_j = 0\}} \prod_{i=1}^n \text{NB}(y_{ij}; \lambda_{ij0}, \phi_j) \prod_{\{j: \gamma_j = 1\}} \prod_{k=1}^K \prod_{\{i: z_i = k\}} \text{NB}(y_{ij}; \lambda_{ijk}, \phi_j).$$

(3.2)

A common choice for the prior on the vector $\gamma$ is to assume independent Bernoulli distributions on the individual components, with a common hyperparameter $\omega_\gamma$, i.e. $\gamma_j | \omega_\gamma \sim \text{Bern}(\omega_\gamma)$. It is equivalent to a binomial prior on the number of discriminatory features, that is $p_\gamma | \omega_\gamma \sim \text{Bin}(p_\gamma, \omega_\gamma)$. The hyperparameter $\omega_\gamma$ can be elicited as the proportion of features expected a priori to be in the discriminatory set. We can further relax this prior assumption by formulating a $\text{Be}(a_\omega, b_\omega)$ hyperprior on $\omega_\gamma$, which leads to a beta-binomial prior for $p_\gamma$ with expectation $pa_\omega/(a_\omega + b_\omega)$. However, such a prior model do not take into account correlation among features that may contribute to the discriminant analysis.

The major difficulty in modeling multivariate count data is to allow for covariance structure in the data. Here, we incorporate the dependency structure into the prior on $\gamma$ since the prior model is more flexible (Stingo and Vannucci, 2011). We consider a Markov random field (MRF) prior that takes into account spatial dependencies among the features, rather than an independent beta-Bernoulli prior. An MRF is a graphical model in which the distribution of a set of random variables follows Markov properties that can be described by an undirected graph. Specifically, let $\Psi$ be a $p \times p$ symmetric matrix, with each binary element $\psi_{jj'}$ indicates whether features $j$ and $j'$ are neighbors. For instance, the undirected graph $\Psi$ could be defined by an existing gene-gene network in KEGG database, where genes are represented by nodes
and neighborhood relations between them by edges. Then, we hypothesize that two neighbor features are more likely to have a joint effect on the relevant process. This assumption is incorporated into an MRF selection prior,

\[ p(\gamma_j | \gamma_{-j}, \Psi) = \frac{\exp \left( \gamma_j (d + f \sum_{j' \in N_j} \gamma_{j'}) \right)}{1 + \exp \left( d + f \sum_{j' \in N_j} \gamma_{j'} \right)}, \]

(3.3)

where \( d \) and \( f \) are fixed hyperparameters to be chosen (to avoid phase transition problem), \( \gamma_{-j} \) denotes the vector of \( \gamma \) excluding the \( j \)-th element, and \( N_j \) denotes the set of direct neighbors of feature \( j \), as defined by \( \Psi \). Here the hyperparameter \( d \) controls the sparsity of the prior model, while \( f \) effects the probability of selection of a variable according to its neighbor values. Note that if a feature does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter \( \omega_j = \exp(d)/(1 + \exp(d)) \), which is a logistic transformation of \( d \). We conclude this subsection by specifying the prior distribution for \( \phi_j \) as

\[ \phi_j \sim \text{Ga}(a_{\phi}, b_{\phi}). \]

(3.4)

### 3.2.2 A negative binomial regression model with covariate selection

We describe the mean of the negative binomial mixture model using a log link with three additive components (two components for the "null" model),

\[
\begin{align*}
\log \lambda_{ij0} &= \alpha_{0j} + x_i^T \beta_j \\
\log \lambda_{ijk} &= \alpha_{0j} + \alpha_{kj} + x_i^T \beta_j \quad \text{if } z_i = k.
\end{align*}
\]

(3.5)

The first component \( \alpha_{0j} \) is a baseline process for feature \( j \) shared by all observations. Note that \( \exp(\alpha_{0j}), j = 1, \ldots, p \) can also be considered as the scaling factor capturing feature-specific levels across all observations, introduced in Chapter 2. The second component \( \alpha_{kj} \) is component-specific. It is the shift to the baseline shared by all
observations that belong to group $k$ when feature $j$ is discriminatory; otherwise, it is set to zero. To avoid identifiability problem arises from the sum of components, we fix the mean shifts in the reference group, which is usually the first group, to zero, i.e. $\alpha_{1j} = 0, j = 1, \ldots, p$. The vector $(\alpha_{1j}, \ldots, \alpha_{Kj})$ is denoted by $\mathbf{\alpha}_j$. The third component is subject-specified. The $R \times 1$ coefficient vector $\mathbf{\beta}_j$, which is the $j$-th column of the $R \times p$ association matrix $\mathbf{B}$, describes the effect of the covariates $\mathbf{x}_i$ on the observed count $y_{ij}$. Note that our model assumes all features have relationships with their associated covariates.

Following Stingo et al. (2013), we allow different covariates to affect the individual mixture components, because not all regressors contribute to the response. We use a latent variable $\delta_{rj}$ to specify a mixture of a normal density and a point mass at zero for the prior on each $\beta_{rj}$, i.e.

$$
\beta_{rj}|\delta_{rj}, \sigma_{\beta j}^2 \sim (1 - \delta_{rj})I_0(\beta_{rj}) + \delta_{rj}N(0, \sigma_{\beta j}^2),
$$

(3.6)

where $I_0$ is a point mass function to indicate whether $\beta_{rj} = 0$. This prior is also called spike and slab prior. If $\delta_{rj} = 1$, then covariate $r$ is considered relevant to explain the observed measurements in feature $j$. In this case, we allow the corresponding regression coefficient $\beta_{rj}$ to be sampled from a normal distribution. If $\delta_{rj} = 0$, then covariate $r$ does not affect the response data and the corresponding regression coefficient $\beta_{rj}$ is set to 0.

To make the stochastic search in $\mathbf{\delta}_j = (\delta_{1j}, \ldots, \delta_{Rj})$ more efficiently in our case study on genomic data, we only consider the local associations that are in cis and are opposed to associations in trans, defined by expression quantitative trait loci (eQTLs). eQTLs are genomic loci that contribute to variation in expression levels of mRNAs. eQTLs that map to the approximate location of their gene-of-origin are referred to as cis while those that map far from the location of their gene of origin,
often on different chromosomes, are referred to as trans. Although more than half of
the genetically explained variance in gene expression is due to trans acting variants
(Price et al., 2011; Pai et al., 2015), many eQTLs studies have focused on cis eQTLs
because reliable detection of trans eQTLs has been challenging in humans. This is due
to the smaller effect size of trans-acting variants and to the higher statistical penalty
on multiple testing (Westra et al., 2013; Battle et al., 2013; Pai et al., 2015). Following
this spirit, we use a $R \times p$ binary matrix $\Theta$ to indicate whether gene symbol associated
with probe $r$ corresponds to gene $j$ and then impose $\delta_{rj} = 0$ when the corresponding
$\theta_{rj} = 0$.

We impose an independent Bernoulli distribution over $\delta_{rj}$ that corresponds to
$\theta_{rj} = 1$, with a common hyperparameter $\omega_{\delta}$, i.e. $\delta_{rj}|\theta_{rj} = 1, \omega_{\delta} \sim \text{Bern}(\omega_{\delta})$. The
hyperparameter $\omega_{\delta}$ can be elicited as the proportion of features expected a priori
to be in the selected covariate set. We can further relax this prior assumption by
formulating a $\text{Be}(a_{\omega}, b_{\omega})$ hyperprior on $\omega_{\delta}$, which leads to a beta-Bernoulli prior.
After all the above formulation, the prior distribution of $\delta_{rj}$ can be described as a
spike and slab prior,

$$p(\delta_{rj}|\theta_{rj}) = (1 - \theta_{rj})I_0(\delta_{rj}) + \theta_{rj}\frac{\Gamma(a_{\omega} + b_{\omega}) \Gamma(a_{\omega} + \delta_{rj}) \Gamma(b_{\omega} + 1 - \delta_{rj})}{\Gamma(a_{\omega}) \Gamma(b_{\omega}) \Gamma(a_{\omega} + b_{\omega} + 1)}. \quad (3.7)$$

We complete the model by specifying priors for the remaining parameters $\alpha_{0j}$,
$\alpha_{kj}$, and $\beta_{rj}$. We impose normal hyperpriors on them, i.e. $\alpha_{0j} \sim N(0, \sigma_{0j}^2)$, $\alpha_{kj} \sim
N(0, \sigma_{\alpha j}^2)$, and $\beta_{rj} \sim N(0, \sigma_{\beta j}^2)$. We further formulate the hyperparameters $\sigma_{0j}^2$, $\sigma_{\alpha j}^2$,
and $\sigma_{\beta j}^2$ by using an inverse-gamma distribution with form $\text{IG}(a, b)$, respectively. This
formulation leads to a non-standardized Student’s t-distribution for them,

\[
\alpha_{0j} \sim t_{2a_0}(0, b_0/a_0), \\
\alpha_{jk} \sim t_{2a_\alpha}(0, b_\alpha/a_\alpha), \\
\beta_{rj} \sim t_{2a_\beta}(0, b_\beta/a_\beta).
\] (3.8)

The proposed hierarchical mixture model is summarized in Figure 3.1. Since the posterior distribution of the parameters of interest is not available in closed form, we revert to Markov chain Monte Carlo algorithms for posterior inference, as described in the next section.

### 3.3 Model Fitting

In this section, we briefly describe the MCMC algorithm for posterior inference. Our inferential strategy allows to simultaneously identify the discriminatory features while selecting the covariates that can best predict each feature. We give full details of our MCMC algorithm in the Appendix.

#### 3.3.1 MCMC algorithm

Our primary interest lies in the identification of the discriminatory features, via the vector \( \gamma \), and the selection of the “best” covariates for each feature, via the vector \( \delta_j, j = 1, \ldots, p \). We design a Markov chain Monte Carlo (MCMC) algorithm based on Metropolis search variable selection algorithms (George and McCulloch, 1997; Brown et al., 1998) to search the model space that consists of \((\gamma, \phi, \alpha_0, A, B, \Delta)\).

**Update of \( \phi \):** We update each \( \phi_j, j = 1, \ldots, p \) sequentially by using independent Metropolis-Hastings algorithm. We first propose a new \( \phi^*_j \) from Ga\((a_\phi, b_\phi)\) and then
### Parameters:

- $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_j, \ldots, \gamma_p]$
- $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_j, \ldots, \phi_p]$
- $\boldsymbol{\alpha}_0 = [\alpha_{01}, \ldots, \alpha_{0j}, \ldots, \alpha_{0p}]$
- $\boldsymbol{A} = [\alpha_1, \ldots, \alpha_j, \ldots, \alpha_p]$
- $\boldsymbol{B} = [\beta_1, \ldots, \beta_j, \ldots, \beta_p]$
- $\boldsymbol{\Delta} = [\delta_1, \ldots, \delta_j, \ldots, \delta_p]$

### Mixture model likelihood:

- $y_{ij}|z_i = k, \gamma_j = 1 \overset{iid}{\sim} \text{NB}(y_{ij}; \lambda_{ijk}, \phi_j)$ with $\lambda_{ijk} = \alpha_{0j} + \alpha_{kj} + x_i^T \beta_j$
- $y_{ij}|\gamma_j = 0 \overset{iid}{\sim} \text{NB}(y_{ij}; \lambda_{ij0}, \phi_j)$ with $\lambda_{ij0} = \alpha_{0j} + x_i^T \beta_j$

### Feature selection prior:

- $\gamma_j|\gamma_{-j}, \Psi \sim \text{MRF}_\varphi(d, f)$

### Dispersion prior:

- $\phi_j \sim \text{Ga}(a_\phi, b_\phi)$

### Covariate-dependent characterization priors:

- $\alpha_{0j}|\sigma_{0j} \sim \text{N}(0, \sigma_{0j}^2)$, $\sigma_{0j}^2 \sim \text{IG}(a_0, b_0) \Rightarrow \alpha_{0j} \sim t_{2a_0}(0, b_0/a_0)$
- $\alpha_{kj}|\sigma_{kj} \sim \text{N}(0, \sigma_{kj}^2)$, $\sigma_{kj}^2 \sim \text{IG}(a_\alpha, b_\alpha) \Rightarrow \alpha_{kj} \sim t_{2a_\alpha}(0, b_\alpha/a_\alpha)$
- $\beta_{rj}|\delta_{rj}, \sigma_{\beta j}^2 \sim (1 - \delta_{rj})\text{I}_0(\beta_{rj}) + \delta_{rj}N(0, \sigma_{\beta j}^2)$, $\sigma_{\beta j}^2 \sim \text{IG}(a_\beta, b_\beta) \Rightarrow$

\[ \beta_{rj}|\delta_{rj} \sim (1 - \delta_{rj})\text{I}_0(\beta_{rj}) + \delta_{rj}t_{2a_\beta}(0, b_\beta/a_\beta) \]

### Covariate selection prior:

- $\delta_{rj}|\theta_{rj} = 1, \omega_\delta \sim (1 - \theta_{rj})\text{I}_0(\delta_{rj}) + \theta_{rj}\text{Bern}(\omega_\delta)$, $\omega_\delta \sim \text{Be}(a_\omega, b_\omega) \Rightarrow$
- $p(\delta_{rj}|\theta_{rj}) = \frac{(1 - \theta_{rj})\text{I}_0(\delta_{rj}) + \theta_{rj}\Gamma(a_\omega + \delta_{rj})\Gamma(b_\omega + \delta_{rj})\Gamma(b_\omega + 1 - \delta_{rj})}{\Gamma(a_\omega + b_\omega + 1)}$

### Fixed hyperparameters:

- $d, f, a_\phi, b_\phi, a_0, b_0, a_\alpha, b_\alpha, a_\beta, b_\beta, a_\omega, b_\omega$

---

Figure 3.1: Hierarchical formulation of the proposed hierarchical mixture model.
accept the proposed value with probability $\min(1, m_{\text{MH}})$, where

$$m_{\text{MH}} = \frac{f(y_j | \gamma_j, \phi_j^*, \alpha_{0j}, \alpha_j, \beta_j, \delta_j) \ p(\phi_j^*) \ J(\phi_j^{\text{old}}, \phi_j^*)}{f(y_j | \gamma_j, \phi_j^{\text{old}}, \alpha_{0j}, \alpha_j, \beta_j, \delta_j) \ p(\phi_j^{\text{old}}) \ J(\phi_j^{\text{old}}, \phi_j^*)}. \quad (3.9)$$

Note that the last two ratios can be cancelled out for this independent Metropolis update.

**Update of $\alpha_0$:** We update each $\alpha_{0j}, j = 1, \ldots, p$ sequentially by using Metropolis-Hastings random walk algorithm. We first propose a new $\alpha_{0j}^*$ from $N(\alpha_{0j}^{\text{old}}, \tau_\alpha^2)$ and then accept the proposed value with probability $\min(1, m_{\text{MH}})$, where

$$m_{\text{MH}} = \frac{f(y_j | \gamma_j, \phi_j, \alpha_{0j}^*, \alpha_j, \beta_j, \delta_j) \ p(\alpha_{0j}^*) \ J(\alpha_{0j}^{\text{old}}, \alpha_{0j}^*)}{f(y_j | \gamma_j, \phi_j, \alpha_{0j}^{\text{old}}, \alpha_j, \beta_j, \delta_j) \ p(\alpha_{0j}^{\text{old}}) \ J(\alpha_{0j}^{\text{old}}, \alpha_{0j}^*)}. \quad (3.10)$$

Note that the proposal density ratio equals to 1 for this random walk Metropolis update.

**Joint update of $\gamma$ and $A$:** We perform a between-model step to update these two groups of parameters jointly. This is done via an *add-delete* algorithm. We first select a $j$ from $\{1, \ldots, p\}$ at random and then change its value. For *add* case, i.e. $(\gamma_j^{\text{old}} = 0 \rightarrow \gamma_j^* = 1)$, we further propose a new $\alpha_{kj}^*, k = 2, \ldots, K$ from $N(0, \tau_\alpha^2)$. For *delete* case, i.e. $(\gamma_j^{\text{old}} = 1 \rightarrow \gamma_j^* = 0)$, we set all $\alpha_{kj}^*, k = 2, \ldots, K$ equal to 0. We finally accept the proposed values with probability $\min(1, m_{\text{MH}})$. The Hastings ratio is written as,

$$m_{\text{MH}} = \frac{f(y_j | \gamma_j^*, \phi_j, \alpha_{0j}^*, \alpha_j^*, \beta_j, \delta_j) \ p(\alpha_j^* | \gamma^*) \ p(\gamma^*)}{f(y_j | \gamma_j^{\text{old}}, \phi_j, \alpha_{0j}^{\text{old}}, \alpha_j^{\text{old}}, \beta_j, \delta_j) \ p(\alpha_j^{\text{old}} | \gamma^{\text{old}}) \ p(\gamma^{\text{old}}) \ J(\alpha_j^{\text{old}}, \alpha_j^* | \gamma^{\text{old}}, \gamma^*) \ J(\gamma^{\text{old}}, \gamma^*)}{J(\alpha_j^*; \alpha_j^{\text{old}} | \gamma^*; \gamma^{\text{old}}) \ J(\gamma^*; \gamma^{\text{old}})}, \quad (3.11)$$

where the last proposal density ratio equals to 1 and the second last proposal density ratio equals to

$$
\begin{cases}
1/ \prod_{k=2}^K N(\alpha_{kj}^*; 0, \tau_\alpha^2) & \text{if add} \\
\prod_{k=2}^K N(\alpha_{kj}^{\text{old}}; 0, \tau_\alpha^2) & \text{if delete.}
\end{cases}
$$
Update of $A$: We perform a within-model step to update each $\alpha_{kj}, k = 2, \ldots, K, j = 1, \ldots, p$ by using Metropolis-Hastings random walk algorithm. We first propose a new $\alpha_{kj}^*$ from $N(\alpha_{kj}^{old}, (\tau_{\alpha}/10)^2)$ and then accept the proposed value with probability $\min(1, m_{MH})$, where

$$m_{MH} = \frac{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^*, \beta_{kj}^*, \delta_{kj}) p(\alpha_{kj}^*) J(\alpha_{kj}^{old}, \alpha_{kj}^*)}{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^{old}, \beta_{kj}, \delta_{kj}) p(\alpha_{kj}^{old}) J(\alpha_{kj}^{old}, \alpha_{kj}^{old})}.$$  (3.12)

Note that the proposal density ratio equals to 1 for this random walk Metropolis update.

Joint update of $\Delta$ and $B$: We perform a between-model step again to update each $(\delta_{kj}, \beta_{kj}), j = 1, \ldots, p$, jointly. We first propose a new value of $\delta_{kj}^*$ by either an add or delete step. Then, we propose a new $\beta_{kj}^*$ from $N(0, \tau_{\beta}^2)$ for add step or set $\beta_{kj}^* = 0$ for delete step. We finally accept the proposed values with probability $\min(1, m_{MH})$. The Hastings ratio is written as,

$$m_{MH} = \frac{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^*, \beta_{kj}^*, \delta_{kj}^*) p(\beta_{kj}^*) J(\beta_{kj}^{old}, \beta_{kj}^*)}{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^{old}, \beta_{kj}, \delta_{kj}) p(\beta_{kj}^{old}) J(\beta_{kj}^{old}, \beta_{kj}^{old})} \frac{p(\delta_{kj}^*) J(\delta_{kj}^{old}, \delta_{kj}^*)}{p(\delta_{kj}^{old}) J(\delta_{kj}^{old}, \delta_{kj}^{old})},$$  (3.13)

where that the last proposal density ratio equal to 1 and the second last proposal density ratio equals to

$$\begin{cases} 
1/N(\beta_{kj}^*; 0, \tau_{\beta}^2) & \text{if add} \\
N(\beta_{kj}^{old}; 0, \tau_{\beta}^2) & \text{if delete}. 
\end{cases}$$

Update of $B$: We perform a within-model again to update each $\beta_{kj}$ that corresponds to $\delta_{kj} = 1$. We first propose a new value of $\beta_{kj}^*$ from $N(\beta_{kj}^{old}, (\tau_{\beta}/10)^2)$ and then accept the proposed value with probability $\min(1, m_{MH})$, where

$$m_{MH} = \frac{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^*, \beta_{kj}^*, \delta_{kj}) p(\beta_{kj}^*) J(\beta_{kj}^{old}, \beta_{kj}^*)}{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{kj}^{old}, \beta_{kj}, \delta_{kj}) p(\beta_{kj}^{old}) J(\beta_{kj}^{old}, \beta_{kj}^{old})}.$$  (3.14)
Note that the proposal density ratio equal to 1 for this random walk Metropolis update. In addition, we can update $\beta^{(d_j)}_j$ as a group. In particular, we first propose a new value of $\beta^{(d_j)}_j$ from $\text{MN}(\beta^{(d_j)}_j^{\text{old}}, \Sigma)$, where $\Sigma = \hat{\sigma}^2 \left((X^{(d_j)})^T X^{(d_j)}\right)^{-1}$ and $\hat{\sigma}^2$ is the sample variance of $y_j$. Then, we accept the proposed value with probability $\min(1, m_{\text{MH}})$, where

\[ m_{\text{MH}} = \frac{f(y_j | \gamma_j, \phi_j, \alpha_0, \alpha_j, \beta^{(d_j)}_j, \delta_j)}{f(y_j | \gamma_j, \phi_j, \alpha_0, \alpha_j, \beta^{(d_k)}_j, \delta_j)} \frac{p(\beta^{(d_j)}_j^*) J(\beta^{(d_j)}_j^*, \beta^{(d_k)}_j)}{p(\beta^{(d_j)}_j^{\text{old}}) J(\beta^{(d_j)}_j, \beta^{(d_k)}_j^{\text{old}})}. \] (3.15)

Note that the proposal density ratio equal to 1 for this random walk Metropolis update.

### 3.3.2 Posterior estimation

We obtain posterior inference on $\gamma$ and $\delta_{-j}$, $j = 1, \ldots, p$ by post-processing of the MCMC samples after burn-in. One way to summarize the posterior distribution of the parameters of interest is via the maximum-a-posteriori (MAP) estimate that can be obtained by

\[ \hat{\gamma}^{\text{MAP}} = \arg\max_{1 \leq b \leq B} p(\gamma^{(b)}) \prod_{j=1}^{p} f(y_j | \gamma_j^{(b)}, \phi_j^{(b)}, \alpha_0, \alpha_j, \beta_j^{(b)}, \delta^{(b)}_j) \] (3.16)

and

\[ \hat{\delta}_{j}^{\text{MAP}} = \arg\max_{1 \leq b \leq B} p(\delta^{(b)}_j) f(y_j | \gamma_j^{(b)}, \phi_j^{(b)}, \alpha_0, \alpha_j, \beta_j^{(b)}, \delta^{(b)}_j), \] (3.17)

respectively, with $b = 1, \ldots, B$ indexing the MCMC iterations, after burn-in. An alternative estimate relies on the marginal posterior probability of inclusion (PPI) of single features or covariates as the proportion of MCMC iterations, after burn-in, in which the corresponding $\gamma_j$ or $\delta_{-j}$ were equal to 1, that is

\[ \text{PPI}_j^\gamma = \sum_{b=1}^{B} \left( \gamma_j^{(b)} | \phi_j^{(b)}, \alpha_0, \alpha_j, \beta_j^{(b)}, \delta^{(b)}_j, y_j \right) / B \] (3.18)
and

\[
\text{PPI}_{r,j} = \sum_{b=1}^{B} \left( \phi_{rj}^{(b)} | \gamma_j^{(b)}, \delta_j^{(b)}, \alpha_{0j}, \alpha_j^{(b)}, \beta_j^{(b)}, y_{j} \right) / B, \tag{3.19}
\]

respectively. A point estimate of \( \gamma \) or \( \delta_j \) is then obtained by identifying those PPI values that exceed a given cutoff \( c \). The cutoff is typically chosen based on a decision theoretic criterion. Here, we follow Newton et al. (2004) so that an expected rate of false detection (i.e. Bayesian FDR) smaller than a fixed threshold \( t \) can be guaranteed.

### 3.4 Simulation

Two experiments are designed to validate our model. The first experiment on fully simulated data, where each element of the covariate matrix \( X \) is independently and identically generated from a beta distribution, explores the convergence property and the statistical performance of the model. The second experiment on synthetic data, where a real highly-correlated DNA methylation data is used as the covariate matrix \( X \), shows that our model can work under this challenging scenario.

#### 3.4.1 Evaluation with simulated data

We generated simulated data with a relatively small sample size \( (n = 78) \) with respect to the number of features \( (p = 172) \). To test the ability of our method to discover relevant features in the presence of a good amount of noise, we focused on the situation where only a few of features are truly discriminatory \( (p_\gamma = 20) \). For each observation, we first simulated the counts that map to the set of 20 discriminating features from a Poisson distribution \( y_{ij} | \gamma_j = 1 \sim \text{Poi}(\lambda_{ij}) \). To mimic the characteristic that features exhibit correlation structure, we generated \( \lambda_{ij}^{(\gamma)} \), where \( (\gamma) \) is used to index the discriminating features, from a mixture of four \( (K = 4) \) multivariate normal densities
plus the effect of the covariates \( x_i \),

\[
\chi_i^{(\gamma)} \sim (1 \leq i \leq 19) \text{MN}(\mu_1, \Sigma_1) + I(20 \leq i \leq 46) \text{MN}(\mu_2, \Sigma_2) + I(47 \leq i \leq 61) \text{MN}(\mu_3, \Sigma_3) + I(62 \leq i \leq 78) \text{MN}(\mu_4, \Sigma_4) + x_i^T B^{(\gamma)},
\]

where the first 19 observations were drawn from the first distribution, the second 27 from the second distribution, the next 15 from the third and the last 17 from the forth. We set the means of the normal distributions equal to 

\[
\mu_1 = 3 \times 1_{p_y}, \quad \mu_2 = 5 \times 1_{p_y},
\]

\[
\mu_3 = 4.5 \times 1_{p_y}, \quad \mu_4 = 3.5 \times 1_{p_y},
\]

where \( 1_{p_y} \) is a unit vector of dimension \( p_y = 20 \). We constructed the four covariance matrices of the 20 discriminatory features, with the diagonal elements set to \( \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2 \) and \( \sigma_4^2 = 3 \). The correlation structure of the 20 features was represented by \( 5 \times 4 \) grids, with the corresponding off-diagonal elements of the four covariance matrices set to 0.5 if two features are connected and 0 otherwise. In this \( 5 \times 4 \) lattice, there are 4 features connected to two other features, 10 features connected to three other features, and 6 features connected to four other features. This generating mechanism creates correlation also between discriminating features not directly connected in the lattice. We then added 152 noisy features, whose underlying mean from a multivariate normal distribution centered at four, with the diagonal and the off-diagonal elements of the covariance matrix set to 1 and 0.1, respectively, i.e. 

\[
\chi_i^{(\gamma)} \sim \text{MN}(4 \times 1_{p_y-p_n}, \Sigma_0) + x_i^T B^{(\gamma)}.
\]

Here we use \( (\gamma^c) \) to denote the set indexed by \( j \) whose \( \gamma_j = 0 \). As for the coefficient matrix \( B \), we first simulated each \( \delta_{rj} \) whose \( \theta_{rj} = 1 \) from a Bernoulli distribution with parameter \( \omega_\delta = 0.1 \), then we generated each \( \beta_{rj} \) whose \( \delta_{rj} = 1 \) from a mixture of two uniform distributions \( 0.7 \times U(1, 2) + 0.3 \times U(-2, -1) \). We borrowed the indicator matrix \( \Theta \) from the case study demonstrated in the next section. Each element in the covariate matrix \( X \) was independently and identically generated from a beta distribution \( \text{Be}(0.4, 0.6) \). The correlations between the covariates were in the range \((-0.55, 0.55)\). Figure 3.2 (top)
shows the empirical distribution of correlations between the features $y_j$'s obtained by the generative model described above.

We set the hyperparameters that control the MRF prior to $d = -4$ and $f = 1$, which means that if a feature does not have any neighbor in the network, its prior distribution reduces to an independent Bernoulli distribution with parameters $\omega = \frac{\exp(-4)}{1 + \exp(-4)} = 0.018$. As for the beta prior on the covariate selection parameters $\omega_\delta$, we set $a_\omega = 0.2$ and $b_\omega = 1.8$, which lead to a vague prior for $p_\gamma$ with expectation equals to $p/10$. For the priors on $\sigma^2_{0j}$, $\sigma^2_{\alpha j}$, and $\sigma^2_{\beta j}$, we used the same flat hyperprior $\text{IG}(2, 1)$. For the prior that controls the dispersion of negative binomial model, we set $\phi_j \sim \text{Ga}(a_\phi = 1, b_\phi = 0.01)$. Results we report below were obtained by running the MCMC chain with 100,000 iterations, discarding the first 10% as burn in. We started the chain from a model with 10% randomly chosen $\gamma_j$’s and $\delta_{rj} | \theta_{rj} = 1$’s set to 1. All experiments were implemented in R with $\text{Rcpp}$ package on a Mac PC with 2.30GHz CPU and 16GB memory. In our implementation, the MCMC algorithm ran in 50 - 60 minutes.

Figure 3.3 shows the marginal posterior probability of inclusion (PPI) of single features when using the MRF($d = -4, f = 1$) prior (top) and the independent Bernoulli prior $\text{Bern}(\omega = 0.018)$ (bottom). The red dots indicate the truly discriminatory features. A threshold corresponding to an expected false discovery rate (Bayesian FDR) $t = 5\%$ on the marginal probabilities was chosen for the two scenarios, separately. It resulted in a model that included 23 features with MRF prior, 19 of which were in the set of truly discriminatory features while resulting in a model included 17 features without MRF prior, 15 of which were in the set of truly discriminatory features. With MRF prior, a threshold of $c = 0.76$ led to an almost perfect feature selection with only one false negative and one true negative while without MRF prior, the same
Figure 3.2: Simulation study: Empirical distributions of correlations between features in simulated data (top) and synthetic data (bottom).
threshold led to a poor feature selection with two false negatives and six true negatives. Besides, the posterior probabilities of the discriminatory features are shown to be higher when the MRF prior was used. All the above findings suggest employing MRF prior helps in the selection of the correct features.

We also conducted a sensitivity analysis by choosing different values of MRF prior hyperparameters $d$ and $f$. In particular, we independently generated 25 data sets. For each of the 25 data sets, we implemented the algorithm with seven different MRF settings and then computed their individual precision, recall, and F1-score based on the PPI estimate of $\gamma$. Here, precision is defined as the number of truly discriminatory features that are correctly estimated, divided by the number of all included features in the model. Recall is defined as the number of truly discriminatory features that are correctly estimated, divided by the number of all truly discriminatory features. F1-score, which takes values between 0 and 1, is defined as

$$F1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$  

The larger the F1-score is, the better the model should be. The detailed results are given in Table 3.1, which suggest the model is insensitive to the hyperparameters $d$ and $f$, unless either $d$ or $f$ approaches to 0 which leads to a less sparse model or a model with an independent Bernoulli prior, respectively.

As for the covariate selection, Figure 3.4 shows the marginal posterior probability of inclusion of single covariates associated with one truly discriminatory feature $j = 74$ and one truly non-informative feature $j = 1$ when using the MRF prior (top) and the independent Bernoulli prior (bottom). As we can see in the figure, the performance difference between both models is neglectable. We also computed the precision and recall for each PPI estimate of $\delta_j$. The threshold we chose corresponds to an expected Bayesian FDR $t = 5\%$. Whichever prior was used, we obtained averaged precision
Figure 3.3: Simulated data: Marginal posterior probability of inclusion $p(\gamma_j = 1 | \cdot)$ with red dots indicating truly discriminatory features using the MRF prior (top) and the independent Bernoulli prior (bottom).
Table 3.1: Simulated data: Sensitivity analysis on choosing the values of hyperparameters $d$ and $f$ in the MRF prior. Standard deviations are indicated in parentheses.

<table>
<thead>
<tr>
<th>Scenario $(d, f)$</th>
<th>$(−1, 1)$</th>
<th>$(−2, 1)$</th>
<th>$(−3, 1)$</th>
<th>$(−4, 0)$</th>
<th>$(−4, 0.5)$</th>
<th>$(−4, 1)$</th>
<th>$(−4, 2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.624</td>
<td>0.712</td>
<td>0.783</td>
<td>0.930</td>
<td>0.902</td>
<td>0.831</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.042)</td>
<td>(0.034)</td>
<td>(0.059)</td>
<td>(0.055)</td>
<td>(0.030)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Recall</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.748</td>
<td>0.966</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.113)</td>
<td>(0.043)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.768</td>
<td>0.831</td>
<td>0.878</td>
<td>0.823</td>
<td>0.931</td>
<td>0.907</td>
<td>0.872</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.029)</td>
<td>(0.021)</td>
<td>(0.068)</td>
<td>(0.024)</td>
<td>(0.018)</td>
<td>(0.021)</td>
</tr>
</tbody>
</table>

around 0.60 and averaged recall around 0.85.

### 3.4.2 Evaluation with synthetic data

In this experiment, the generative model is as the same as the previous experiment on simulated data, except: a) the covariate matrix $X$ is from a real DNA methylation data, which will also be used in the case study and it is highly-correlated, where the correlations between the covariates are in the range $(-0.92, 0.99)$; b) the structure on noisy features has been changed in the following way. We added 152 noisy features, with each counts independently and identically distributed from $y_{ij} | \gamma_j = 0 \sim \text{Poi}(\alpha_{0j} + \mathbf{x}_i^T \beta_j)$, where we generated $\alpha_{0j} \sim \text{U}(2, 10)$. The algorithm setting was the same as in Section 3.4.1. Figure 3.2 (bottom) shows the empirical distribution of correlations between the features $\mathbf{y}_j$’s obtained by the above generating mechanism.
Figure 3.4: Simulated data: Marginal posterior probability of inclusion $p(\delta_{rj} = 1|\cdot)$ for two of the discriminatory features ($j = 74$ and $j = 1$) with red circles indicating true covariates using the MRF prior (top) and the independent Bernoulli prior (bottom).
Figure 3.5 shows the marginal posterior probability of inclusion (PPI) of single features when using the MRF\( (d = -4, f = 1) \) prior (top) and the independent Bernoulli prior Bern\( (\omega = 0.018) \) (bottom). A threshold of Bayesian FDR \( t = 0.05 \) was used. It resulted in a model that included 18 features with MRF prior, all of which were in the set of truly discriminatory features while resulting in a model that included 13 features without MRF prior, all of which were in the set of truly discriminatory features. With MRF prior, a threshold of \( c = 0.40 \) led to a perfect feature selection. Without MRF prior, applying the same threshold led to a poor feature selection with seven true negatives. Besides, the posterior probabilities of the discriminatory features are shown to be higher when the MRF prior was used. All the above findings again suggest employing MRF prior helps in the selection of the correct features, even if the covariate matrix \( X \) is highly-correlated. As for the covariate selection, with or without MRF prior, the performances were almost the same, with averaged precision around 0.50 and averaged recall around 0.65.

### 3.5 Case study on breast cancer

In this section, we use a real RNA-Seq gene expression data and its associated DNA methylation data to conduct a case study on breast cancer. We demonstrate the integration of existing biological knowledge into the MRF prior results in an increased ability to identify genes with strong discriminatory power.

#### 3.5.1 Introduction to the real data set

Breast cancer, occurred in women and rarely in men, is a malignant breast tumor characterized by uncontrolled cell growth in tissues of the breast (e.g. ducts and lobules). It is the second most commonly diagnosed cancer in women in the United States. The
Figure 3.5: Synthetic data: Marginal posterior probability of inclusion $p(\gamma_j = 1|\cdot)$ with red dots indicating truly discriminatory features using the MRF prior (top) and the independent Bernoulli prior (bottom).
American Cancer Society estimates that there are about 230,000 new cases of breast cancer each year. In the past decades, the identification of biomarkers in the breast cancer diagnosis has attracted many researchers, especially after new biotechnology (i.e. RNA-Seq) for more accurate measurements on bioactivities released. However, to the best of our knowledge, there is no attempt at a rigorously quantitative analysis for selection of discriminatory RNA-Seq genes and their associated DNA methylation probes in breast cancer. Here we use publicly available data from the TCGA data portal to identify those plausibly discriminatory genes in breast cancer and study their associations with DNA methylation levels. In particular, RNA-Seq count data, as $Y$, has been selected as those genes ($p = 172$) belonging to KEGG cancer pathway. The corresponding DNA methylation data, as $X$, has been mapped to the above selected genes and the total number of probes is $R = 3,528$. The samples were collected on $n = 78$ breast cancer patients. $n_1 = 19$ of them are in stage I, $n_2 = 27$ are in stage II, $n_3 = 15$ are in stage III, and $n_4 = 17$ are in stage IV, categorized by tumor size.

### 3.5.2 Results

We report the results by specifying an MRF prior on $\gamma$ that uses the gene-gene network structure from KEGG pathway database. In particular, the network structure, as shown in Figure 3.6, was obtained using the R package `KEGGgraph` of Zhang and Wiemann (2009). Note that some of the genes do not have neighbors. We used the same hyperparameter setting in Section 3.4, that is, $a_{\phi} = 1, b_{\phi} = 0.01, a_0 = a_{\alpha} = a_{\beta} = 2, b_0 = b_{\alpha} = b_{\beta} = 1, a_{\omega} = 0.2, b_{\omega} = 1.8$, and $d = -4, f = 1$. We ran 100,000 iterations with 10,000 sweeps as burn-in. To assess the concordance of the results, we ran four independent MCMC chains and looked at the correlations between the marginal posterior probabilities for features $p(\gamma_j = 1 | \cdot)$ and for covariates.
These indicated substantial agreement between the four MCMC chains, with pairwise correlation coefficients ranging from 0.989 to 0.993 for features and from 0.959 to 0.969 for covariates. We also used the Gelman and Rubin’s convergence diagnostics (Gelman and Rubin, 1992) to assess the convergence of the parameters $\phi_j$’s to their posterior distributions. Those statistics were below 1.2, ranging from 1.000 to 1.190, clearly indicating that the MCMC chains were run for a satisfactory number of iterations.

Results we report here were obtained by pooling the outputs from the four chains together. Table 3.2 lists 12 genes whose marginal posterior probabilities of inclusion $p(\gamma_j | \cdot)$ are beyond 0.875 when using MRF prior. This threshold corresponds to an expected Bayesian FDR of $t = 5\%$. As shown in Figure 3.6, we also highlight these 12 genes in red in the gene-gene network. To compare the results without MRF prior, we show the scatter plot of marginal posterior probabilities of single genes, with MRF prior against without MRF prior, in Figure 3.7 (top). The magnitudes of posterior probabilities of inclusion increase significantly when using MRF prior, as most of the points are beyond the $y = x$ equality ratio line. Meanwhile, the number of overlapping genes between models with and without MRF prior as a function of the first $p, p = 1, \ldots, 100$ genes sorted by marginal posterior probabilities $p(\gamma_j | \cdot)$ are given in Figure 3.7 (bottom). As expected, most of the genes were shared by both models. For instance, there were 14, 35, 54, and 78 genes co-occurred in the list of top 25, 50, 75, and 100 genes obtained by the two models, respectively. All these findings suggest the incorporation of the gene-gene network into the model results in an increased ability to identify genes with strong discriminatory power. Figure 3.8 shows the marginal posterior probabilities $p(\delta_{rj} | \theta_{rj}, \cdot)$ associated with five of the 12 genes, with the red circle indicating the selected probes when choosing an expected
Figure 3.6: Breast cancer data: The gene-gene network structure from KEGG database, with the red nodes indicating the selected features whose marginal posterior probability of inclusion $p(\gamma_j = 1|\cdot) \geq 0.875$ with MRF prior.
Table 3.2: Breast cancer data: List of genes selected by the model with the MRF prior and their corresponding posterior probabilities.

| Gene   | Description                                      | $p(\gamma_j = 1|\cdot)$ |
|--------|--------------------------------------------------|--------------------------|
|        |                                                  | w/ MRF  | w/o MRF    |
| FZD10  | frizzled class receptor 10                       | 1.000   | 1.000      |
| PLCB1  | phospholipase C beta 1                           | 0.994   | 0.003      |
| PDGFRA | platelet derived growth factor receptor alpha     | 0.993   | 0.013      |
| FH     | fumarate hydratase                               | 0.968   | 0.000      |
| FLT3LG | fms related tyrosine kinase 3 ligand             | 0.953   | 0.500      |
| PIAS2  | protein inhibitor of activated STAT 2            | 0.947   | 0.014      |
| MAPK1  | mitogen-activated protein kinase 1               | 0.945   | 0.000      |
| CSF1R  | colony stimulating factor 1 receptor             | 0.937   | 0.036      |
| SHH    | sonic hedgehog                                   | 0.927   | 0.949      |
| WNT16  | Wnt family member 16                             | 0.894   | 0.645      |
| CXCL12 | C-X-C motif chemokine ligand 12                  | 0.879   | 0.043      |
| FIGF   | vascular endothelial growth factor D             | 0.875   | 0.291      |

Bayesian FDR of $t = 5\%$ as the threshold.

### 3.5.3 Biological findings

In order to assess the biological relevance of our findings, we performed a literature search on the list of genes shown in Table 3.2, which revealed interesting results. FDZ10, a member of the Wnt signal receptor family, plays critical roles in cell survival and growth, which both are highly related to cancer progression. It is also highly up-
Figure 3.7: Breast cancer data: Scatter plot of marginal posterior probabilities $p(\gamma_j|\cdot)$ with vs. without MRF prior (top); The number of overlapping genes as a function of the first $p, p = 1, \ldots, 100$ genes between the models with and without MRF prior (bottom).
Figure 3.8: Breast cancer data: Marginal posterior probability of inclusion \( p(\delta_{rj} = 1|\theta_{rj} = 1, \cdot) \) with red circles indicating selected probes for discriminatory genes FZD10, PLCB1, PDGFRA, WNT16, and CXCL12.
regulated and may have a molecular signaling mechanism in synovial sarcoma, a rare form of soft tissue cancer (Fukukawa et al., 2009). FDZ10 epigenetic silencing plays a mediating role, through HDAC1 recruitment and histone H3K9 deacetylation, on BRMS1L inhibition of breast cancer cells migration and invasion (Gong et al., 2014). PLCB1 is upregulated in highly metastatic breast cancer cells, and its increased expression associates with human metastatic relapse, as was found in Sengelaub et al. (2016). Also, Molinari et al. (2012) reported copy number aberrations of this gene in breast cancer. An association between over-expression of PDGFRA, platelet-derived growth factor receptor alpha, in breast cancer with tumour progression was found in Carvalho et al. (2005) and an alteration of this gene in bone marrow mesenchymal stroma cells from advanced untreated lung and breast cancer patients was found in Hofer et al. (2005). MAPK1 has been found associated with proliferation and migration of breast cancer cells (Yu et al., 2007; You et al., 2009), and may play a crucial network role in breast invasive carcinoma (Yellapu et al., 2016). CSF1R is a breast cancer promoter (Flick et al., 2002; Woo et al., 2009) and is associated with bone metastases of breast cancer (Hiraga and Nakamura, 2009). WNT16 is part of the Wnt signaling pathway, a very complex pathway which includes numerous ligands, receptors and transcription effectors. It has been found downregulated in a study on 6 breast cancer cell lines (Benhaj et al., 2006). CXCL12 is well known for breast cancer cell migration (Kang et al., 2005; Lee et al., 2005), growth (Zhao et al., 2008), invasiveness (Orimo et al., 2005) and metastasis (Wendt et al., 2008; Yang et al., 2008). FIGF is a growth factor active in angiogenesis, lymphangiogenesis and endothelial cell growth, stimulating their proliferation and migration (Teramoto et al., 2008; Akahane et al., 2006). It is also a prognostic, survival and metastasis factor for breast cancer (Nakamura et al., 2003).
Our literature search brought to less interesting findings on FH, FLT3LG, PIAS2 and SHH. For example, germline mutations in FH occur in uterine fibroids, skin leiomyomata and papillary renal cell cancer (Tomlinson et al., 2002). The activity of FLT3LG, a growth factor for hematopoietic progenitors, on one lineage of dendritic cells (DC1) may have implications for the development of cancer immunotherapy strategies (Mosca et al., 2002). PIAS2 is a transcriptional coregulator in various cellular pathways, including the STAT pathway, the p53 pathway and the steroid hormone signaling pathway. SHH (sonic hedgehog) has been found associated with the development of prostate and pancreatic cancer (Thayer et al., 2003; Sanchez et al., 2004), and to motility and invasiveness of gastric cancer cells (Yoo et al., 2008).

Finally, we enrich our results by employing the database for annotation visualization and integrated discovery (DAVID) tool (Dennis et al., 2003). Specifically, we focused on an enrichment analysis performed on the genes whose PPIs were higher than 0.5. Table 3.3 shows the significant terms at a level of 0.05, applied on corrected p-values (Benjamini). We found significant terms in gene ontology (GO) category for biological process (BP) and molecular functions (MF). In addition, we found a significant term for the SWISS-PROT protein information resource (SP-PIR) keywords category. Interestingly, among the results in the GO biological process, we found enrichment for processes related to cell proliferation (regulation of cell proliferation, positive regulation of cell proliferation and cell proliferation), to cell differentiation and development (mesenchyme development, mesenchymal cell differentiation, mesenchymal cell development, neural crest cell differentiation and neural crest cell development), to cell motion and migration (regulation of cell motion, regulation of cell migration, regulation of locomotion, positive regulation of cell migration, positive regulation of locomotion, positive regulation of cell motion and ameboidal cell migration) and cell
death (regulation of cell death, regulation of apoptosis, regulation of programmed cell death, positive regulation of cell death, positive regulation of programmed cell death and positive regulation of apoptosis). Also, tube development, tube morphogenesis, branching morphogenesis of a tube and morphogenesis of a branching structure are related to the process of tube generation and organization. Epithelial and endothelial tubes transport gases, liquids and cells from one site to another. We also identified many processes related to external stimulus, such as response to estrogen, steroid hormone, estradiol, hormone, endogenous stimulus, and response to oxygen levels, hypoxia and nutrient. Cell surface receptor linked signal transduction takes part in regulation of downstream cellular processes, such as transcription. We also found enrichment for morphogenesis, development, growth and sprouting of blood vessels giving rise to the organized vascular system (blood vessel morphogenesis, patterning of blood vessels and blood vessel development). Reproductive developmental process involves the progressive change in the state of some part of an organism, specifically contributing to its ability to form offspring. Positive regulation of macromolecule metabolic process increases the frequency, rate or extent of the chemical reactions and pathways involving macromolecule. In the GO molecular functions we found enrichment for the protein homodimerization activity, which refers to the selective and non-covalent interaction of a protein with an identical protein with the result of forming a homodimer. In addition, identical protein binding refers to the selective and non-covalent interaction with one or more identical proteins. The protein dimerization activity refers to the formation of a protein dimer. A protein dimer is a macromolecular structure which consists of two non-covalently associated identical or nonidentical sub-units. As for the SP-PIR keywords, not surprisingly we found enrichment for the proto-oncogene term. Proto-oncogene are those proteins whose
normal cellular gene can be converted into cancer promoting oncogenes via activating mutations, chromosomal translocation or DNA amplification.

Overall our findings suggest that the genes identified by our method are not only related to breast cancer in general, but with the differentiation induced by breast cancer stage. Indeed, many of them have been found related to cancer cell proliferation, migration, invasion, and metastasis, as well as cancer progression.

Table 3.3: Breast cancer data: Results of the enrichment analysis.

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0042127</td>
<td>regulation of cell proliferation</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0008284</td>
<td>positive regulation of cell proliferation</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0043627</td>
<td>response to estrogen stimulus</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0035295</td>
<td>tube development</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0008283</td>
<td>cell proliferation</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0051270</td>
<td>regulation of cell motion</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0048545</td>
<td>response to steroid hormone stimulus</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0035239</td>
<td>tube morphogenesis</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0060485</td>
<td>mesenchyme development</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0032355</td>
<td>response to estradiol stimulus</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0070482</td>
<td>response to oxygen levels</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0048762</td>
<td>mesenchymal cell differentiation</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0014031</td>
<td>mesenchymal cell development</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0007166</td>
<td>cell surface receptor linked signal transduction</td>
</tr>
<tr>
<td>Category</td>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0033273</td>
<td>response to vitamin</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0030334</td>
<td>regulation of cell migration</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0010033</td>
<td>response to organic substance</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0048754</td>
<td>branching morphogenesis of a tube</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0001763</td>
<td>morphogenesis of a branching structure</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0009725</td>
<td>response to hormone stimulus</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0040012</td>
<td>regulation of locomotion</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0010941</td>
<td>regulation of cell death</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0042981</td>
<td>regulation of apoptosis</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0043067</td>
<td>regulation of programmed cell death</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0048514</td>
<td>blood vessel morphogenesis</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0030335</td>
<td>positive regulation of cell migration</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0010604</td>
<td>positive regulation of macromolecule metabolic process</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0009719</td>
<td>response to endogenous stimulus</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0040017</td>
<td>positive regulation of locomotion</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0051272</td>
<td>positive regulation of cell motion</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0007423</td>
<td>sensory organ development</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0001569</td>
<td>patterning of blood vessels</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0010942</td>
<td>positive regulation of cell death</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0043068</td>
<td>positive regulation of programmed cell death</td>
</tr>
</tbody>
</table>

Continued on next page
Table 3.3 – continued from previous page

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0001568</td>
<td>GO:0001568</td>
<td>blood vessel development</td>
</tr>
<tr>
<td>GO:0001944</td>
<td>GO:0001944</td>
<td>vasculature development</td>
</tr>
<tr>
<td>GO:0043065</td>
<td>GO:0043065</td>
<td>positive regulation of apoptosis</td>
</tr>
<tr>
<td>GO:0003006</td>
<td>GO:0003006</td>
<td>reproductive developmental process</td>
</tr>
<tr>
<td>GO:0014033</td>
<td>GO:0014033</td>
<td>neural crest cell differentiation</td>
</tr>
<tr>
<td>GO:0014032</td>
<td>GO:0014032</td>
<td>neural crest cell development</td>
</tr>
<tr>
<td>GO:0001666</td>
<td>GO:0001666</td>
<td>response to hypoxia</td>
</tr>
<tr>
<td>GO:0001657</td>
<td>GO:0001657</td>
<td>ureteric bud development</td>
</tr>
<tr>
<td>GO:0007584</td>
<td>GO:0007584</td>
<td>response to nutrient</td>
</tr>
<tr>
<td>GO:0048565</td>
<td>GO:0048565</td>
<td>gut development</td>
</tr>
<tr>
<td>GO:0001667</td>
<td>GO:0001667</td>
<td>ameboidal cell migration</td>
</tr>
<tr>
<td>GO:0042803</td>
<td>GO:0042803</td>
<td>protein homodimerization activity</td>
</tr>
<tr>
<td>GO:0042802</td>
<td>GO:0042802</td>
<td>identical protein binding</td>
</tr>
<tr>
<td>GO:0046983</td>
<td>GO:0046983</td>
<td>protein dimerization activity</td>
</tr>
<tr>
<td>PIR:KEYWORDS</td>
<td>Proto-oncogene</td>
<td></td>
</tr>
</tbody>
</table>

3.6 Conclusion

We have presented a hierarchical modeling framework for the analysis of data that arise in a genome-wide association study. Our model can identify the discriminatory response features among distinct groups while selecting the covariates that explain the associated feature. We have demonstrated its ability to identify the differentially expressed genes in breast cancer and quantify their association with DNA methyla-
The hierarchical model we have developed has several innovative characteristics. It is integrative. As demonstrated in the case study, it not only links two high-dimensional genomic data, RNA-Seq and DNA methylation data, but also makes good use of KEGG gene pathway. It achieves the simultaneous selection of a set of discriminatory features and relevant covariates. It employs spatially based selection process priors that capture available graphical information, so that related features are more likely to be selected together.

Several extensions of our model are worth investigating. First, we can extend our model to more complex mixture models for clustering the observations (i.e. make inferences on \( z \)). For instance, infinite mixture models can be fit based on the Dirichlet process, which is similar in spirit to Chapter 2. Second, we can take into account correlation among features by modeling the baseline parameter \( \alpha_0 \) as a multivariate normal distribution, in addition to using the MRF prior. Stingo et al. (2013) proposed a similar model to study the association between continuous neuroimaging responses and genetic covariates. Last but not least, we can learn about the fixed hyperparameters such as \( d \) and \( f \) in the MRF prior by formulating hyperpriors (see, e.g. Liang, 2010; Stingo and Vannucci, 2011).
Chapter 4

Conclusion

In this thesis, I have introduced two novel approaches for modeling high-dimensional count data, which both incorporate a feature selection mechanism and account for the characteristics observed in count data. For posterior inference, I have used Markov chain Monte Carlo (MCMC) sampling techniques that combine both Gibbs sampler and Metropolis-Hastings algorithm to search the model space.

The goal of the first Bayesian nonparametric model has been to cluster the samples with a fully unsupervised approach while selecting the discriminatory variables. To do so in a unified way, I have introduced a binary latent vector to identify discriminating features and used Dirichlet process zero-inflated Poisson mixture models to define the cluster structure. Furthermore, prior distributions that appropriately account for identifiability constraints on the model parameters have been developed. Both simulation and application on real bag-of-words data have shown that the methodology has a competitive edge over existing methods.

The goal of the second Bayesian integrative model has been to identify the features that best distinguish groups with known labels while exploring the association between the features and a large number of observed regressors. Specifically, I have specified a negative binomial regression model for each feature and used a spike and slab prior to select the covariates that have more influence on the corresponding response. To increase the ability to identify discriminatory genes, the model includes Markov random field (MRF) prior models that can capture structural dependencies
among the features. The proposed integrative model has been demonstrated through an application to RNA-Seq gene expression and DNA methylation data for identifying biomarkers in breast cancer and studying the association between gene expressions and DNA methylation levels.

Although I have focused on the analysis of text documents in the first project and genomic data in the second project, my methodologies are quite general and can be extended to the analysis of many other high-dimensional count data.
Appendix A: Detailed MCMC algorithm for Chapter 2

We start by writing the marginal likelihood for each sample $i$, $i = 1, \ldots, n$, after integrating out the parameters $d^*_{kj}$.

$$f(y_i | \gamma, \tilde{s}_i, \tilde{g}, r_i)$$

$$= \int f(y_i | z_i = k, \gamma, d^*_{kj}, \tilde{s}_i, \tilde{g}, r_i) p(d^*_k) \, dd^*_k.$$  

$$= \int \prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \text{Poi}(y_{ij}; s_i g_j d^*_{kj}) \prod_{\{j : \gamma_j = 0, r_{ij} = 0\}} \text{Poi}(y_{ij}; s_i g_j) \prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \text{Ga}(d^*_{kj}; a, b) \, dd^*_k$$

$$= \prod_{\{j : r_{ij} = 0\}} \left( s_i g_j \right)^{y_{ij}} \frac{1}{y_{ij}!} \exp \left\{ -s_i \sum_{\{j : \gamma_j = 0, r_{ij} = 0\}} g_j \right\} \prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + y_{ij})}{(b + s_i g_j)^{a+y_{ij}}}.$$  

where $s_i = \exp\{\tilde{s}_i\}$, $g_j = \exp\{\tilde{g}_j\}$, and $d^*_{kj} = \exp\{\tilde{d}^*_{kj}\}$

In terms of each feature $j$, $j = 1, \ldots, p$, we write

$$f(y_{ij} | z, \gamma_j = 0, \tilde{s}, \tilde{g}_j, r_j) = \prod_{\{i : r_{ij} = 0\}} \left( s_i g_j \right)^{y_{ij}} \frac{1}{y_{ij}!} \exp \left\{ -g_j \sum_{\{i : r_{ij} = 0\}} s_i \right\}$$

and

$$f(y_{ij} | z, \gamma_j = 1, \tilde{s}, \tilde{g}_j, r_j) = \prod_{\{i : r_{ij} = 0\}} \left( s_i g_j \right)^{y_{ij}} \frac{1}{y_{ij}!} \left( \frac{b^a}{\Gamma(a)} \right)^{\frac{K}{K}} \times \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{\{i : z_i = k, r_{ij} = 0\}} y_{ij})}{(b + g_j \sum_{\{i : z_i = k, r_{ij} = 0\}} s_i)^{a+\sum_{\{i : z_i = k, r_{ij} = 0\}} y_{ij}}}.$$  

At each MCMC iteration, we perform the following steps:

**Update of the set of discriminatory features $\gamma$:** We randomly perform an *add-delete-swap* step. We repeat this step 20 times to ensure validation of the feature selection for each given cluster assignment. The general Hasting ratio can be written...
as

$$r = \frac{p(\gamma^* | z, \tilde{s}, \tilde{g}_j, R, \omega, y) J(\gamma^{(b-1)} | \gamma^*)}{p(\gamma^{(b-1)} | z, \tilde{s}, \tilde{g}_j, R, \omega, y) J(\gamma^* | \gamma^{(b-1)})}$$

$$= \frac{f(y | z, \gamma^*, \tilde{s}, \tilde{g}_j, R)}{f(y | z, \gamma^{(b-1)}, \tilde{s}, \tilde{g}_j, R)} \frac{p(\gamma^* | \omega)}{p(\gamma^{(b-1)} | \omega)} J(\gamma^{(b-1)} | \gamma^*).$$

For both add/delete and swap steps, the proposal density ratio equals to 1.

More specifically, for the add-delete step, we randomly sample a feature $j$, $j = 1, \ldots, p$ and propose to change the value of $\gamma_j$ with probability $\min(1, m_{\text{MH}})$, where

$$m_{\text{MH}}^{\text{add}} = \frac{ \left( \frac{b^*}{\Gamma(a)} \right)^K \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{i:k,r_{ij}=0} y_{ij})}{(b + g_j \sum_{i:k,r_{ij}=0} s_i)^a \sum_{i:k,r_{ij}=0} y_{ij}} \omega \exp \left( -g_j \sum_{i:r_{ij}=0} s_i \right) 1 - \omega}{\left( \frac{b^*/\Gamma(a)}{\Gamma(\alpha)} \right)^K \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{i:k,r_{ij}=0} y_{ij})}{(b + g_j \sum_{i:k,r_{ij}=0} s_i)^a \sum_{i:k,r_{ij}=0} y_{ij}} \omega \exp \left( -g_j \sum_{i:r_{ij}=0} s_i \right) 1 - \omega}.$$ 

For the swap step, we randomly swap two features if applicable, i.e. change the value of a feature $\gamma_{j_1}$ from 1 to 0, and, correspondingly, of another currently non-discriminatory feature $j_2$, i.e. change the value of $\gamma_{j_2}$ from 0 to 1, with probability $\min(1, r_{\text{MH}}^{\text{swp}})$, where

$$m_{\text{MH}}^{\text{swp}} = \frac{ \exp \left( -g_{j_1} \sum_{i:r_{ij}=0} s_i \right) \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{i:k,r_{ij}=0} y_{ij})}{(b + g_{j_1} \sum_{i:k,r_{ij}=0} s_i)^a \sum_{i:k,r_{ij}=0} y_{ij}} \times \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{i:k,r_{ij}=0} y_{ij})}{(b + g_{j_2} \sum_{i:k,r_{ij}=0} s_i)^a \sum_{i:k,r_{ij}=0} y_{ij}} \omega \exp \left( -g_{j_2} \sum_{i:r_{ij}=0} s_i \right) 1 - \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega \omega}.$$ 

Finally, we update the hyperparameter $\omega$ for inclusion probability,

$$\omega | \gamma, a_\omega, b_\omega \sim \text{Be}(a_\omega + p_\gamma, b_\omega + n - p_\gamma).$$

**Update of the cluster allocation for the selected features $z$:** At each iteration, we perform a Gibbs sampling to update $z_i$ sequentially from observation 1 to $n$
Then, we propose to either:

\[
f(y_i | z_i = k, z_{-i}, \gamma, \bar{s}, \bar{g}, R, y_{-i})
= \int f(y_i | z_i = k, \gamma, d_k^*, \bar{s}, \bar{g}, r_i) p(d_k^* | z_{-i}, \gamma, \bar{s}_{-i}, \bar{g}, R_{-i}, y_{-i}) \, dd_k^*
= \int \prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \text{Poi}(y_{ij}; s_i g_j d_{kj}^*) \prod_{\{j : \gamma_j = 0, r_{ij} = 0\}} \text{Poi}(y_{ij}; s_i g_j)
\prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \text{Ga}\left(d_{kj}^*; a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j}, b + g_j \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} s_i\right) \, dd_k^*
= \prod_{\{j : r_{ij} = 0\}} \frac{(s_i g_j)^{y_{ij}}}{y_{ij}!} \exp\left\{-s_i \sum_{\{j : \gamma_j = 0, r_{ij} = 0\}} g_j\right\}
\prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \frac{\left(b + g_j \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} s_i\right)^{y_{i'j}}}{\Gamma(a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j})} \, b + g_j \sum_{\{i' : z_{i'} = k, r_{i'j} = 0\}} s_i
\prod_{\{j : \gamma_j = 1, r_{ij} = 0\}} \frac{\left(a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j}\right)}{\Gamma\left(a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j}\right)}
\cdot \frac{\Gamma \left(a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j}\right)}{\Gamma(a + \sum_{\{i' : z_{i'} = k, r_{i'j} = 0, i' \neq i\}} y_{i'j})}.
\]

Then, we propose to either:

- form a new cluster (change the value of \(z_i\) to \(K_{-i} + 1\)) with probability \(p_{K_{-i} + 1}\),
  where
  \[
p_{K_{-i} + 1} = \frac{\alpha}{n - 1 + \alpha} f(y_i | \gamma, \bar{s}, \bar{g}, r_i)
  \]
  \[
  + \sum_{k=1}^{K_{-i} + 1} \frac{n_{k-1}}{n - 1 + \alpha} f(y_i | z_i = k, z_{-i}, \gamma, \bar{s}, \bar{g}, R, y_{-i}).
  \]

- allocate the observation to any of the existing clusters, say cluster \(k \in \{1, \ldots, K_{-i} + 1\}\), with probability \(p_k\), where
  \[
p_k = \frac{n_k}{n - 1 + \alpha} f(y_i | z_i = k, z_{-i}, \gamma, \bar{s}, \bar{g}, R, y_{-i})
  \]
  \[
  + \sum_{k=1}^{K_{-i} + 1} \frac{n_{k-1}}{n - 1 + \alpha} f(y_i | z_i = k, z_{-i}, \gamma, \bar{s}, \bar{g}, R, y_{-i}).
  \]
Finally, we update the concentration parameter $\alpha$, by following the algorithm in Escobar and West (1995), that is we generate an auxiliary variable $\eta_\alpha|\alpha \sim \text{Be}(\alpha + 1, n)$ and then we sample $\alpha$ from a mixture of two gamma densities,

$$\alpha|\eta_\alpha, z \sim \pi_\eta \text{Ga}(a_\alpha + K, b - \log(\eta_\alpha)) + (1 - \pi_\eta) \text{Ga}(a_\alpha + K - 1, b_\alpha - \log(\eta_\alpha)),$$

with the weights $\pi_\eta$ defined by $\pi_\eta/(1 - \pi_\eta) = (a_\alpha + K - 1)/(n(b_\alpha - \log(\eta_\alpha))).$

**Update of the scaling factors $s$ and $g$:** We can rewrite the prior distribution in (2.4)–(2.5) by introducing latent auxiliary variables, that specify how the $\tilde{s}_i$ and $\tilde{g}_j$ are assigned to any of the inner and outer mixture components. More specifically, we can introduce a $n \times 1$ vector of assignment indicators, $\nu$, with $\nu_i = m$ indicating that $\tilde{s}_i$ is a sample from $f^s_m(\tilde{s}_i|t_m, \eta_m)$. The weights $\phi^s_m$ determine the probability of each value $\nu_i = m$, with $m = 1, 2, \ldots$. Correspondingly, we can consider a $n \times 1$ vector $\epsilon$ of binary elements $\epsilon_i$, where $\epsilon_i = 1$ indicates that, given $\nu_i = m$, $\tilde{s}_i$ is drawn from a $\text{N}(\eta_m, \sigma^2_s)$ with probability $t_m$, and $\epsilon_i = 0$ indicates that $\tilde{s}_i$ is drawn from the right component of $f^s_m(\tilde{s}_i|t_m, \eta_m)$, i.e. $\text{N}\left(\frac{c_s - t_m \eta_m}{1 - t_m}, \sigma^2_s\right)$, with probability $1 - t_m$. Similarly, we can introduce a $p \times 1$ vector $\xi$, with $\xi_j = l$ indicating that $\tilde{g}_j$ is sampled from $f^g_l(\tilde{g}_j|q_l, \mu_l)$, $l = 1, 2, \ldots$, and $\phi^g_l = p(\xi = l)$. Correspondingly, given the assignments obtained in the vector $\xi$, we can define a $p \times 1$ vector $\psi$ of binary elements $\psi_j$, where $\psi_j = 1$ indicates that, given $\xi_j = l$, then $\tilde{g}_j$ is drawn from $\text{N}(\mu_l, \sigma^2_g)$, whereas $\psi_j = 0$ indicates that $\tilde{g}_j$ is from $\text{N}\left(\frac{c_g - q_l \mu_l}{1 - q_l}, \sigma^2_g\right)$. Thus, the prior model (2.4)–(2.5) can be rewritten as

$$\tilde{s}_i|\nu_i, \epsilon_i, t, \eta \sim \text{N}\left(\epsilon_i \eta_{\nu_i} + (1 - \epsilon_i) \frac{c_s - t_{\nu_i} \eta_{\nu_i}}{1 - t_{\nu_i}}, \sigma^2_s\right), \quad (1)$$

and

$$\tilde{g}_j|\xi_j, \psi_j, q, \mu \sim \text{N}\left(\psi_j \mu_{\xi_j} + (1 - \psi_j) \frac{c_g - q_{\xi_j} \mu_{\xi_j}}{1 - q_{\xi_j}}, \sigma^2_g\right), \quad (2)$$
where \( t, \eta, q, \) and \( \mu \) denote the vectors of \( t_m, \eta_m, q_t, \) and \( \mu_t, \) respectively. Therefore, the update of the sample and feature specific scaling factors \( s_i \) and \( g_j \) can proceed as follows, after logarithmic transformation:

**a) Update of the \( \tilde{s}_i \)'s:** We perform Metropolis sampling to update \( \tilde{s}_i, \) where \( \tilde{s}_i = \log\{s_i\}, \) sequentially from observation 1 to \( n. \) We propose a new \( \tilde{s}_i^* \) from \( N(\tilde{s}_i^{(b-1)}, \tau_s^2) \) and accept it with probability \( \min(1, m_{\text{MH}}) \), where

\[
m_{\text{MH}} = \frac{p(s_i^* | \gamma, \tilde{g}, \tilde{r}_i, \nu_i, \epsilon_i, t, \eta, y_i)}{p(s_i^{(b-1)} | \gamma, \tilde{g}, \tilde{r}_i, \nu_i, \epsilon_i, t, \eta, y_i)} \frac{J(s_i^{(b-1)} | \tilde{s}_i^*)}{J(s_i^{(b-1)} | \tilde{s}_i)}
\]

\[
= \frac{f(y_i | \gamma, \tilde{s}_i^*, \tilde{g}, \tilde{r}_i)}{f(y_i | \gamma, \tilde{s}_i^{(b-1)}, \tilde{g}, \tilde{r}_i)} \frac{p(s_i^{(b-1)} | \nu_i, \epsilon_i, t, \eta)}{p(s_i^{(b-1)} | \nu_i, \epsilon_i, t, \eta)}
\]

\[
= \frac{s_i \exp \left( -s_i \sum_{j: \gamma_j = 0, r_{ij} = 0} y_{ij} \right) \exp \left( -s_i \sum_{j: \gamma_j = 0, r_{ij} = 0} g_j \right)}{s_i^{(b-1)} \exp \left( -s_i^{(b-1)} \sum_{j: \gamma_j = 0, r_{ij} = 0} g_j \right)} \prod_{j: \gamma_j = 0, r_{ij} = 0} (b + s_i^{(b-1)} g_j)^{-a-y_{ij}} \prod_{j: \gamma_j = 0, r_{ij} = 0} (b + s_i^{(b-1)} g_j)^{-a-y_{ij}}
\]

\[
\times N \left( \tilde{s}_i^*; \epsilon_i \eta_i \nu_i + (1 - \epsilon_i) \frac{c_s - t_m \eta_m}{1-t_m}, \sigma_s^2 \right)
\]

\[
\times N \left( \tilde{s}_i^{(b-1)}; \epsilon_i \eta_i \nu_i + (1 - \epsilon_i) \frac{c_s - t_m \eta_m}{1-t_m}, \sigma_s^2 \right).
\]

Since \( \nu, \epsilon, t, \) and \( \eta \) have conjugate full conditionals, we use Gibbs sampling to update them one after another

- **Gibbs sampling for updating \( \nu_i, i = 1, \ldots, n: \)**

\[
\pi(\nu_i | \nu_{-i}, \epsilon_i, t, \eta, \tilde{s}_i) \propto \phi_m^2 N \left( \tilde{s}_i; \epsilon_i \eta_i \nu_i + (1 - \epsilon_i) \frac{c_s - t_m \eta_m}{1-t_m}, \sigma_s^2 \right).
\]

- **Gibbs sampling for updating \( \epsilon_i, i = 1, \ldots, n: \)**

\[
\pi(\epsilon_i | \nu_i = m, \epsilon_{-i}, t, \eta, \tilde{s}_i) \propto \begin{cases} 
(1 - t_m) N \left( \tilde{s}_i; \frac{c_s - t_m \eta_m}{1-t_m}, \sigma_s^2 \right) & \text{if } \epsilon_i = 0 \\
N \left( \tilde{s}_i; \eta_m, \sigma_s^2 \right) & \text{if } \epsilon_i = 1
\end{cases}.
\]
• Gibbs sampling for updating $t_m, m = 1, \ldots, M$:

$$t_m|\nu, \epsilon \sim \text{Be} \left( a_t + \sum_{i=1}^{n} I(\nu_i = m)I(\epsilon_i = 1), b_t + \sum_{i=1}^{n} I(\nu_i = m)I(\epsilon_i = 0) \right) .$$

• Gibbs sampling for updating $\eta_m, m = 1, \ldots, M$:

$$\eta_m|\nu, \epsilon, t, \tilde{s} \sim N \left( \frac{c_m/\sigma_s^2}{e_m/\sigma_s^2 + 1/\tau_q^2}, \frac{1}{e_m/\sigma_s^2 + 1/\tau_q^2} \right) ,$$

where $c_m = \sum_{i: \nu_i = m, \epsilon_i = 1} \tilde{s}_i - \frac{t_m}{1 - t_m} \sum_{i: \nu_i = m, \epsilon_i = 0} \left( \tilde{s}_i - \frac{c_m}{1 - t_m} \right)$ and $e_m = \sum_{i=1}^{n} I(\nu_i = m)I(\epsilon_i = 1) + \sum_{i: \nu_i = m, \epsilon_i = 0} \left( \frac{t_m}{1 - t_m} \right)^2$.

• Gibbs Sampling for updating $\phi_m^s, m = 1, \ldots, M$ by stick-breaking process Ishwaran and James (2001):

$$\phi_1^s = v_1,$$

$$\phi_2^s = (1 - v_1)v_2,$$

$$\vdots$$

$$\phi_M^s = (1 - v_1) \cdots (1 - v_{M-1})v_M ,$$

where $\nu_m|\nu \sim \text{Be} \left( a_m + \sum_{i=1}^{n} I(\nu_i = m), b_m + \sum_{i=1}^{n} I(\nu_i > m) \right)$.

b) Update of the $g_j$’s: We perform Metropolis sampling to update $\tilde{g}_j$, where $\tilde{g}_j = \log\{g_j\}$, sequentially from feature 1 to $p$. We propose a new $\tilde{g}_j^*$ from $N(\tilde{g}_j^{(b-1)}, \tau_{g}^2)$.
and accept it with probability $\min(1, m_{\text{MH}})$, where

$$
m_{\text{MH}} = \frac{p(\tilde{g}_j^{*} | z, \gamma_j, \tilde{s}, r_j, \xi_j, \psi_j, q, \mu, g_j)}{p(g_j^{(b-1)} | z, \gamma_j, \tilde{s}, r_j, \xi_j, \psi_j, q, \mu, g_j)} \cdot \frac{J(\tilde{g}_j^{*} | \tilde{g}_j^{(b-1)})}{J(g_j^{(b-1)} | \tilde{g}_j^{*})}
$$

$$
= \frac{f(y_j | z, \gamma_j, \tilde{s}, g_j^{(b-1)}, r_j)}{f(y_j | z, \gamma_j, \tilde{s}, \tilde{g}_j^{*}, r_j)} \cdot \frac{p(\tilde{g}_j^{*} | \xi_j, \psi_j, q, \mu)}{p(\tilde{g}_j^{(b-1)} | \xi_j, \psi_j, q, \mu)}
$$

$$
= \begin{cases} 
\frac{g_j^{*} \sum_{(i,r_{ij}=0)} y_{ij}}{g_j^{(b-1)} \sum_{(i,r_{ij}=0)} y_{ij}} \exp\left\{-g_j^{*} \sum_{(i,r_{ij}=0)} s_i\right\} \cdot \frac{\exp\left\{-g_j^{(b-1)} \sum_{(i,r_{ij}=0)} s_i\right\}}{g_j^{(b-1)} \sum_{(i,r_{ij}=0)} y_{ij}} & \text{if } \gamma_j = 0 \\
\frac{g_j^{*} \sum_{(i,r_{ij}=0)} \Pi_{k=1}^{b} (1 + g_j^{*} \sum_{(i,z_k=k,r_{ij}=0)} s_i) \cdot \frac{1}{\sum_{(i,z_k=k,r_{ij}=0)} y_{ij}}}{g_j^{(b-1)} \sum_{(i,r_{ij}=0)} y_{ij}} \cdot \frac{\exp\left\{-g_j^{*} \sum_{(i,r_{ij}=0)} s_i\right\}}{g_j^{(b-1)} \sum_{(i,r_{ij}=0)} y_{ij}} & \text{if } \gamma_j = 1
\end{cases}
$$

Since $\xi$, $\psi$, $q$, and $\mu$ have conjugate full conditionals, we use Gibbs sampling to update them one after another:

- Gibbs sampling for updating $\xi_j, j = 1, \ldots, p$:
  $$
p(\xi_j = l | \xi_{-j}, \psi_j, q, \mu, \tilde{g}_j) \propto \phi^2 N\left( \tilde{g}_j; \psi_j \mu_l + (1 - \psi_j) \frac{c_q - q l \mu_l}{1 - q_l}, \sigma_g^2 \right).
$$

- Gibbs sampling for updating $\psi_j, j = 1, \ldots, p$:
  $$
\pi(\psi_j | \xi_j = l, \psi_{-j}, q, \mu, \tilde{g}_j) \propto \begin{cases} 
(1 - q_l) N\left( \tilde{g}_j; \frac{c_q - q l \mu_l}{1 - q_l}, \sigma_g^2 \right) & \text{if } \psi_j = 0 \\
q_l N\left( \tilde{g}_j; \mu_l, \sigma_g^2 \right) & \text{if } \psi_j = 1
\end{cases}
$$

- Gibbs sampling for updating $q_l, l = 1, \ldots, L$:
  $$
q_l | \xi, \psi \sim \text{Be}\left(a_q + \sum_{j=1}^{p} I(\xi_j = l) I(\psi_j = 1), b_q + \sum_{j=1}^{p} I(\xi_j = l) I(\psi_j = 0)\right).
$$

- Gibbs sampling for updating $\mu_l, l = 1, \ldots, L$:
  $$
\mu_l | \xi, \psi, q, \tilde{g} \sim N\left( \frac{c_l / \sigma_g^2}{c_l / \sigma_g^2 + 1/\tau_\mu^2}, \frac{1}{c_l / \sigma_g^2 + 1/\tau_\mu^2} \right).
$$
where \( c_l = \sum_{j: \xi_j = l, \psi_j = 1} \tilde{g}_j - 1_q \sum_{j: \xi_j = l, \psi_j = 0} \left( \tilde{g}_j - \frac{c_q}{1-q} \right) \) and \( \epsilon_l = \sum_{j=1}^p I(\xi_j = l)I(\psi_j = 1) + \sum_{j: \xi_j = l, \psi_j = 0} \left( \frac{q}{1-q} \right)^2. \)

- **Gibbs Sampling for updating \( \phi_l^q \), \( l = 1, \ldots, L \) by stick-breaking process Ishwaran and James (2001):**

\[
\begin{align*}
\phi_l^1 &= v_1, \\
\phi_l^2 &= (1 - v_1)v_2, \\
& \vdots \\
\phi_l^L &= (1 - v_1) \cdots (1 - v_{L-1})v_L,
\end{align*}
\]

where \( v_l | \xi_l \sim \text{Be} \left( 1 + \sum_{j=1}^p I(\xi_j = l), 1 + \sum_{j=1}^p I(\xi_j > l) \right). \)

**Update of zero-inflation latent indicator \( R \):** The full conditionals of the \( r_{ij} \)'s can be obtained after considering that we need to consider only those cases for which \( y_{ij} = 0 \), and

\[
p(r_{ij} \mid y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^* , \pi) \propto f(y_{ij} = 0 \mid z_i = k, \gamma_j, s_i, g_j, d_{kj}^* , r_{ij}) p(r_{ij} \mid \pi)
= \left( \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\} \right)^{1-r_{ij}} \pi^{r_{ij}} (1 - \pi)^{1-r_{ij}}
= \pi^{r_{ij}} \left[ (1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\} \right]^{1-r_{ij}}.
\]

It is equivalent to

\[
\begin{align*}
p(r_{ij} = 1 \mid y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^* , \pi) &= \frac{\pi}{\pi + (1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\}} \\
p(r_{ij} = 0 \mid y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^* , \pi) &= \frac{(1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\}}{\pi + (1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\}}, \text{ or}
\end{align*}
\]

\[
\begin{align*}
p(r_{ij} \mid y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^* , \pi) &= \frac{\pi^{r_{ij}} \left[ (1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\} \right]^{1-r_{ij}}}{\pi + (1 - \pi) \exp \left\{ -s_i g_j d_{kj}^* \gamma \right\}}.
\end{align*}
\]
In order to sample from the above full conditional, we proceed with a Metropolis Hasting within Gibbs approach, where within each iteration we first propose a value \( d_{kj}^* \) for which \( \gamma_j = 0 \) by sampling

\[
d_{kj}^* \sim \text{Ga} \left( a + \sum_{i:z_i=k,r_{ij}=0} y_{ij}, b + g_j \sum_{i:z_i=k,r_{ij}=0} s_i \right),
\]

and then propose a value for \( r_{ij} \) by sampling

\[
p(r_{ij}|y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^*, \pi) = \begin{cases} 
\pi^r_{ij} \left[ (1-\pi) e^{-s_i g_j} \right]^{1-r_{ij}} \\
\pi^r_{ij} \left[ (1-\pi) \exp\left(-s_i g_j\right) \right]^{1-r_{ij}} 
\end{cases} \quad \text{if } \gamma_j = 0,
\]

\[
p(r_{ij}|y_{ij} = 0, z_i = k, \gamma_j, s_i, g_j, d_{kj}^*, \pi) = \begin{cases} 
\pi^r_{ij} \left[ (1-\pi) e^{-s_i g_j d_{kj}^*} \right]^{1-r_{ij}} \\
\pi^r_{ij} \left[ (1-\pi) \exp\left(-s_i g_j d_{kj}^*\right) \right]^{1-r_{ij}} 
\end{cases} \quad \text{if } \gamma_j = 1.
\]

After the Metropolis Hasting step, we use a Gibbs sampling step to update \( \pi \), as

\[
\pi|R, a_\pi, b_\pi \sim \text{Be} \left( a_\pi + \sum_{i=1}^{n} \sum_{j=1}^{p} r_{ij}, b_\pi + \sum_{i=1}^{n} \sum_{j=1}^{p} I(y_{ij} = 0) - \sum_{i=1}^{n} \sum_{j=1}^{p} r_{ij} \right).
\]
Appendix B: Detailed MCMC algorithm for Chapter 3

We start by writing the likelihood for each feature $j$, $j = 1, \ldots, p$, depending on the value of $\gamma_j$.

\[
f(y_j | \gamma_j = 0, \phi_j, \alpha_{0j}, \alpha_j, \beta_j, \delta_j) = \prod_{i=1}^{n} \text{NB} \left( y_{ij}; \alpha_{0j} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)}, \phi_j \right)
\]

\[
= \left( \frac{\phi_j^{\gamma_j}}{\Gamma(\phi_j)} \right) \prod_{i=1}^{n} \frac{\Gamma(\phi_j + y_{ij})}{y_{ij}!} \left( \frac{\alpha_{0j} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)}}{\alpha_{0j} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)} + \phi_j} \right)^{\phi_j + y_{ij}}
\]

and

\[
f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_j, \beta_j, \delta_j) = \prod_{k=1}^{K} \prod_{\{ i: z_i = k \}} \text{NB} \left( y_{ij}; \alpha_{0j} + \alpha_{kj} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)}, \phi_j \right)
\]

\[
= \left( \frac{\phi_j^{\gamma_j}}{\Gamma(\phi_j)} \right) \prod_{i=1}^{n} \frac{\Gamma(\phi_j + y_{ij})}{y_{ij}!} \prod_{k=1}^{K} \prod_{\{ i: z_i = k \}} \left( \frac{\alpha_{0j} + \alpha_{kj} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)}}{\alpha_{0j} + \alpha_{kj} + \left( x^{(\delta_j)}_i \right)^T \beta_j^{(\delta_j)} + \phi_j} \right)^{\phi_j + y_{ij}}.
\]

At each MCMC iteration, we perform the following steps:

**Update of over-dispersion parameter $\phi$:** We perform independent Metropolis sampling to update each $\phi_j, j = 1, \ldots, p$ sequentially. We propose a new $\phi_j^*$ from
\[ m_{\text{MH}} = \frac{f(y_j | \gamma_j, \phi_j^*, \alpha_0, \alpha_j, \beta_j, \delta_j) \ p(\phi_j^*) \ J(\phi_j^{(b-1)}; \phi_j^*)}{f(y_j | \gamma_j, \phi_j^{(b-1)}, \alpha_0, \alpha_j, \beta_j, \delta_j) \ p(\phi_j^{(b-1)}) \ J(\phi_j^{(b-1)}; \phi_j^{(b-1)})} \]

\[
\begin{cases}
\left( \frac{\phi_j^{(b-1)} \Gamma(\phi_j^*)}{\Gamma(\phi_j^{(b-1)})} \right)^n \prod_{k=1}^n \Gamma(\phi_j^* + y_{ij}) \left( \alpha_0 + \left( x_i^{(dj)} \right)^T \beta_j + \phi_j^* \right)^{-\phi_j^* - y_{ij}} & \text{if } \gamma_j = 0 \\
\left( \frac{\phi_j^{(b-1)} \Gamma(\phi_j^*)}{\Gamma(\phi_j^{(b-1)})} \right)^n \prod_{k=1}^n \Gamma(\phi_j^* + y_{ij}) \prod_{i \neq i_k} \left( \alpha_0 + \left( x_i^{(dj)} \right)^T \beta_j + \phi_j^* \right)^{-\phi_j^* - y_{ij}} & \text{if } \gamma_j = 1
\end{cases}
\]

**Update of baseline parameter \( \alpha_0 \):** We perform Metropolis sampling to update each \( \alpha_0, j = 1, \ldots, p \) sequentially. We propose a new \( \alpha_0^* \) from \( N(\alpha_0^{(b-1)}, \tau_0^2) \) and then accept the proposed value with probability \( \min(1, m_{\text{MH}}) \), where

\[ m_{\text{MH}} = \frac{f(y_j | \gamma_j, \phi_j^*, \alpha_0^*, \alpha_j, \beta_j, \delta_j) \ p(\alpha_0^*) \ J(\alpha_0^{(b-1)}; \alpha_0^*)}{f(y_j | \gamma_j, \phi_j^*, \alpha_0, \alpha_j, \beta_j, \delta_j) \ p(\alpha_0^{(b-1)}) \ J(\alpha_0^{(b-1)}; \alpha_0^{(b-1)})} \]

\[
\begin{cases}
\prod_{i=1}^n \left( \alpha_0^* + \left( x_i^{(dj)} \right)^T \beta_j + \phi_j^* \right)^{-\phi_j^* - y_{ij}} \left( \frac{b_0 + \alpha_0^{(b-1)} - \alpha_0^*/2}{b_0 + \alpha_0^{(b-1)} - \alpha_0^*/2} \right)^{a_0 + 1/2} & \text{if } \gamma_j = 0 \\
\prod_{i=1}^n \left( \alpha_0^{(b-1)} + \left( x_i^{(dj)} \right)^T \beta_j + \phi_j \right)^{-\phi_j - y_{ij}} \left( \frac{b_0 + \alpha_0^{(b-1)} - \alpha_0^*/2}{b_0 + \alpha_0^{(b-1)} - \alpha_0^*/2} \right)^{a_0 + 1/2} & \text{if } \gamma_j = 1
\end{cases}
\]

**Joint update of the set of discriminatory feature \( \gamma \) and \( A \):** We perform a joint update here since \( A \) depends on \( \gamma \). This is done via an *add-delete* algorithm.

We repeat this step 20 times to ensure validation of the feature selection. The general
Hasting ratio can be written as

\[
\text{min}_{\text{MH}} = \frac{f(y_j | \gamma_j^*, \phi_j, \alpha_{0j}, \alpha_{k_j}^*, \beta_j, \delta_j) p(\alpha_j^* | \gamma^*) p(\gamma^*)}{f(y_j | \gamma_j^{(b-1)}, \phi_j, \alpha_{0j}, \alpha_{k_j}^{(b-1)}, \beta_j, \delta_j) p(\alpha_j^{(b-1)} | \gamma^{(b-1)}) p(\gamma^{(b-1)})}
\]

\[
J(\alpha_j^{(b-1)}; \alpha_j^* | \gamma^{(b-1)}; \gamma^*) J(\gamma^{(b-1)}; \gamma^*)
\]

More specifically, we first select a \( j \) from \( \{1, \ldots, p\} \) at random and then change its value. For add case, i.e. \( (\gamma_j^{(b-1)} = 0 \rightarrow \gamma_j^* = 1) \), we further propose a new \( \alpha_{k_j}^*, k = 2, \ldots, K \) from \( N(0, \tau_\alpha^2) \). For delete case, i.e. \( (\gamma_j^{(b-1)} = 1 \rightarrow \gamma_j^* = 0) \), we set all \( \alpha_{k_j}^*, k = 2, \ldots, K \) equal to 0. We finally accept the proposed values with probability \( \min(1, \text{min}_{\text{MH}}) \). The Hastings ratio is written as,

\[
\text{min}_{\text{MH}}^{\text{add}} = \frac{\prod_{k=1}^K \prod_{i: z_i = k} \left( \alpha_{0j} + \alpha_{k_j}^* + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)} + \phi_j \right)^{-\phi_j - y_{ij}}}{\prod_{i=1}^n \left( \alpha_{0j} + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)} + \phi_j \right)^{-\phi_j - y_{ij}}}
\]

\[
\prod_{k=2}^K \left( b_\alpha + \alpha_{k_j}^* \cdot \alpha_{k_j}^*/2 \right)^{-\alpha_\alpha - 1/2} \exp \left( -\alpha_{k_j}^* \cdot \alpha_{k_j}^*/(2\tau_\alpha^2) \right) \exp \left( d + f \sum_{j \in N_j} \gamma_j^{(b-1)} \right),
\]

\[
\text{min}_{\text{MH}}^{\text{del}} = \frac{\prod_{k=1}^K \prod_{i: z_i = k} \left( \alpha_{0j} + \alpha_{k_j}^{(b-1)} + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)} + \phi_j \right)^{-\phi_j - y_{ij}}}{\prod_{k=1}^K \prod_{i: z_i = k} \left( \alpha_{0j} + \alpha_{k_j}^{(b-1)} + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)} + \phi_j \right)^{-\phi_j - y_{ij}}}
\]

\[
\prod_{k=2}^K \left( b_\alpha + \alpha_{k_j}^{(b-1)} \cdot \alpha_{k_j}^{(b-1)}/2 \right)^{-\alpha_\alpha - 1/2} \exp \left( -\alpha_{k_j}^{(b-1)} \cdot \alpha_{k_j}^{(b-1)}/(2\tau_\alpha^2) \right) \exp \left( -d - f \sum_{j \in N_j} \gamma_j^{(b-1)} \right).
\]

**Update of mean shift parameter \( A \):** We perform Metroplis sampling to update each \( \alpha_{k_j}, k = 2, \ldots, K, \{j : \gamma_j = 1\} \). We propose a new \( \alpha_{k_j}^* \) from \( N(\alpha_{k_j}^{(b-1)}, (\tau_\alpha/10)^2) \)
and accept the proposed value with probability \( \min(1, \text{m}_{\text{MH}}) \), where

\[
\text{m}_{\text{MH}} = \frac{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha^{*}_{j}, \beta_{j}, \delta_{j})}{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha^{(b-1)}_{j}, \beta_{j}, \delta_{j})} \frac{p(\alpha^{*}_{kj})}{p(\alpha^{(b-1)}_{kj})} \frac{J(\alpha_{kj}^{(b-1)}; \alpha_{kj}^{*})}{J(\alpha_{kj}^{(b-1)}; \alpha_{kj}^{*})} \frac{\text{min}(1, \text{m}_{\text{MH}})}{\text{min}(1, \text{m}_{\text{MH}})}
\]

\[
= \prod_{k=1}^{K} \prod_{i: z_i = k} \left( \alpha_{0j} + \alpha_{kj}^{*} \right) + \left( x_{i}^{(\delta_{j})} \right)^T \beta_{j}^{*} + \phi_{j} \right)^{-\phi_{j} - y_{ij}}
\]

\[
\prod_{k=1}^{K} \prod_{i: z_i = k} \left( \alpha_{0j} + \alpha_{kj}^{*} \right) + \left( x_{i}^{(\delta_{j})} \right)^T \beta_{j}^{*} + \phi_{j} \right)^{-\phi_{j} - y_{ij}}
\]

\[
\left( b_{\alpha} + \alpha_{kj}^{*} / 2 \right) a_{\alpha} + 1/2
\]

Joint update of the set of predictable covariates \( \Delta \) and \( B \): We perform a joint update again to update each pair of \( (\delta_{j}, \beta_{j}), j = 1, \ldots, p \). This is done via an add-delete algorithm. We repeat time step 20 times to ensure validation of the covariate selection. The general Hastings ratio can be written as

\[
\text{m}_{\text{MH}} = \frac{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{j}, \beta_{j}, \delta_{j})}{f(y_j | \gamma_j = 1, \phi_j, \alpha_{0j}, \alpha_{j}, \beta_{j}, \delta_{j})} \frac{p(\beta^{*}_{j})}{p(\beta^{(b-1)}_{j})} \frac{p(\delta^{*}_{j})}{p(\delta^{(b-1)}_{j})} \frac{J(\beta^{(b-1)}_{j}; \beta^{*}_{j}, \delta^{(b-1)}_{j})}{J(\beta^{(b-1)}_{j}; \beta^{*}_{j}, \delta^{(b-1)}_{j})} \frac{\text{min}(1, \text{m}_{\text{MH}})}{\text{min}(1, \text{m}_{\text{MH}})}
\]

We first propose a new value of \( \delta_{j}^{*} \) by either an add or delete step. Then, we propose a new \( \beta^{*}_{rj} \) from \( N(0, \tau^{2}_{j}) \) for add step or set \( \beta^{*}_{rj} = 0 \) for delete step. We finally accept the proposed values with probability \( \min(1, \text{m}_{\text{MH}}) \). The specific Hastings ratio
is written as,

\[
\begin{cases}
\Pi_n^{\text{add}} & \left( \alpha_{ij} + \alpha_{kj} + \left( x_i \right)^T \beta_j^{(\delta_{ij})^*} + \phi_j \right)^{-\phi_j - y_{ij}} \\
\Pi_n^{\text{del}} & \left( \alpha_{ij} + \alpha_{kj} + \left( x_i \right)^T \beta_j^{(\delta_{ij})} + \phi_j \right)^{-\phi_j - y_{ij}}
\end{cases}
\]

\[
\frac{\exp(-\beta_{ij}^* \cdot \beta_{ij}^*/(2\tau^2_{ij})) \Gamma(a_{ij}+1)\Gamma(b_{ij})}{(b_{ij}+\beta_{ij}^* \cdot \beta_{ij}^*/2)^{-a_{ij}-1/2} \Gamma(a_{ij})\Gamma(b_{ij})+1}
\]

\[
\Pi_k^{\text{del}} \Pi_{i \in i = k} \left( \alpha_{ij} + \alpha_{kj} + \left( x_i \right)^T \beta_j^{(\delta_{ij})} + \phi_j \right)^{-\phi_j - y_{ij}}
\]

\[
\frac{\exp(-\beta_{ij}^* \cdot \beta_{ij}^*/(2\tau^2_{ij})) \Gamma(a_{ij}+1)\Gamma(b_{ij})}{(b_{ij}+\beta_{ij}^* \cdot \beta_{ij}^*/2)^{-a_{ij}-1/2} \Gamma(a_{ij})\Gamma(b_{ij})+1}
\]

**Update of coefficient \( B \):** We perform Metropolis sampling to update each \( \beta_{ij} \) that corresponds to \( \delta_{ij} = 1 \). We first propose a new value of \( \beta_{ij}^* \) from \( \text{N}(\beta_{ij}^{(b-1)}, (\tau_{ij}/10)^2) \).
and then accept the proposed value with probability \( \min(1, m_{\text{MH}}) \), where

\[
\begin{align*}
m_{\text{MH}} &= \frac{f(y_j | \gamma_j, \phi_j, \alpha_{0j}, \alpha_j, \beta_j^*, \delta_j)}{f(y_j | \gamma_j, \phi_j, \alpha_{0j}, \alpha_j, \beta_j^{(b-1)}, \delta_j)} \frac{p(\beta_j^*)}{p(\beta_j^{(b-1)})} \frac{J(\beta_j^{(b-1)}, \beta_j^*)}{J(\beta_j^{(b-1)}, \beta_j^*)} \\
&= \begin{cases}
\Pi_{i=1}^n \left( a_{0j} + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)^*} + \phi_j \right) \frac{1}{\gamma_j - \nu_{ij}} \left( \frac{b_\beta + \beta_j^{(b-1)}, \beta_j^*}{b_\beta + \beta_j^{(b-1)}, \beta_j^*/2} \right)^{a_\beta + 1/2} & \text{if } \gamma_j = 0 \\
\Pi_{i=1}^n \left( a_{0j} + \phi_j \right) \frac{1}{\gamma_j - \nu_{ij}} \left( \frac{b_\beta + \beta_j^{(b-1)}, \beta_j^*}{b_\beta + \beta_j^{(b-1)}, \beta_j^*/2} \right)^{a_\beta + 1/2} & \text{if } \gamma_j = 1
\end{cases}
\end{align*}
\]

In addition, we perform Metropolis sampling to update \( \beta_j^{(\delta_j)} \) as a group. In particular, we propose a new value of \( \beta_j^{(\delta_j)^*} \) from \( \text{MN}(\beta_j^{(\delta_j)^*}, \Sigma) \), where \( \Sigma = \hat{\sigma}^2 \left( (X^{(\delta_j)})^T X^{(\delta_j)} \right)^{-1} \) and \( \hat{\sigma}^2 \) is the sample variance of \( y_j \). Then, we accept the proposed value with probability \( \min(1, r) \), where

\[
\begin{align*}
m_{\text{MH}} &= \frac{f(y_j | \gamma_j, \phi_j, \alpha_{0j}, \alpha_j, \beta_j^*, \delta_j)}{f(y_j | \gamma_j, \phi_j, \alpha_{0j}, \alpha_j, \beta_j^{(b-1)}, \delta_j)} \frac{p(\beta_j^*)}{p(\beta_j^{(b-1)})} \frac{J(\beta_j^{(b-1)}, \beta_j^*)}{J(\beta_j^{(b-1)}, \beta_j^*)} \\
&= \begin{cases}
\Pi_{i=1}^n \left( a_{0j} + \left( x_i^{(\delta_j)} \right)^T \beta_j^{(\delta_j)^*} + \phi_j \right) \frac{1}{\gamma_j - \nu_{ij}} \left( \frac{b_\beta + \beta_j^{(b-1)}, \beta_j^*}{b_\beta + \beta_j^{(b-1)}, \beta_j^*/2} \right)^{a_\beta + 1/2} & \text{if } \gamma_j = 0 \\
\Pi_{i=1}^n \left( a_{0j} + \phi_j \right) \frac{1}{\gamma_j - \nu_{ij}} \left( \frac{b_\beta + \beta_j^{(b-1)}, \beta_j^*}{b_\beta + \beta_j^{(b-1)}, \beta_j^*/2} \right)^{a_\beta + 1/2} & \text{if } \gamma_j = 1
\end{cases}
\end{align*}
\]
Bibliography


Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008), ‘Map-


Zhou, M. (2014), Beta-negative binomial process and exchangeable random partitions