Factor Models to Describe Linear and Non-linear Structure in High Dimensional Gene Expression Data.

by

Vinicius Diniz Mayrink

Department of Statistical Science
Duke University

Date: ______________________

Approved:

__________________________
Joseph E. Lucas, Advisor

__________________________
Mike West

__________________________
Sayan Mukherjee

__________________________
Jen-Tsan Chi

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2011
ABSTRACT
(Statistical Science)

Factor Models to Describe Linear and Non-linear Structure in High Dimensional Gene Expression Data.

by

Vinicius Diniz Mayrink

Department of Statistical Science
Duke University

Date: ______________________

Approved:

________________________  Joseph E. Lucas, Advisor

________________________  Mike West

________________________  Sayan Mukherjee

________________________  Jen-Tsan Chi

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2011
Abstract

An important problem in the analysis of gene expression data is the identification of groups of features that are coherently expressed. For example, one often wishes to know whether a group of genes, clustered because of correlation in one data set, is still highly co-expressed in another data set. For some microarray platforms there are many, relatively short, probes for each gene of interest. In this case, it is possible that a given probe is not measuring its targeted transcript, but rather a different gene with a similar region (called cross-hybridization). Similarly, the incorrect mapping of short nucleotide sequences to a target gene is a common issue related to the young technology producing RNA-Seq data. The expression pattern across samples is a valuable source of information, which can be used to address distinct problems through the application of factor models. Our first study is focused on the identification of the presence/absence status of a gene in a sample. We compare our factor model to “state of the art” detection methods; the results suggest superior performance of the factor analysis for detecting transcripts. In the second study, we apply factor models to investigate gene modules (groups of coherently expressed genes). Variation in the number of copies of regions of the genome is a well known and important feature of most cancers. Copy number alteration is detected for a group of genes in breast cancer; our goal is to examine this abnormality in the same chromosomal region for other types of tumors (Ovarian, Lung and Brain). In the third application, the expression pattern related to RNA-Seq count data is evaluated through a factor model
based on the Poisson distribution. Here, the presence/absence of coherent patterns is closely associated with the number of incorrect read mappings. The final study of this dissertation is dedicated to the analysis of multi-factor models with linear and non-linear structure of interactions between latent factors. The interaction terms can have important implications in the model; they represent relationships between genes which cannot be captured in an ordinary analysis.
To my parents
Contents

Abstract iv
List of Tables x
List of Figures xi
List of Abbreviations and Symbols xviii
Acknowledgements xxi

1 Introduction 1

2 Presence/Absence calls for gene expression 7

2.1 The data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Bayesian factor model for gene expression detection . . . . . . . . . . 13

2.3 Inference results of the Bayesian factor model . . . . . . . . . . . . . 15

2.4 Comparison of detection methods, using a simulated data set . . . . . 17

2.5 Comparison of detection methods, using a real data set . . . . . . . . . . 23

2.6 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

3 Factor model for identification of DNA copy number alteration 29

3.1 The model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

3.2 Simulated data analysis . . . . . . . . . . . . . . . . . . . . . . . . . . 33

3.3 Real data analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

3.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
# Analysis of high-throughput sequencing data

4.1 Factor model assuming Poisson distribution .................................. 44
4.2 Analysis of read mapping uncertainty in real data sets .................... 48
4.3 Evaluation of the mixture prior for factor loadings ......................... 51
4.4 Simulated study to verify inference results ..................................... 55
4.5 Conclusions ................................................................. 60

# Interactions between factors

5.1 Factor model with multiplicative interactions ................................. 63
5.2 Simulated data analysis ..................................................... 66
5.3 Real data analysis .......................................................... 72
5.4 Conclusions ................................................................. 78

# Non-linear interactions between factors

6.1 Factor model with non-linear interactions .................................. 84
6.2 Simulated data analysis ..................................................... 88
6.3 Real data analysis .......................................................... 94
6.4 Comparison between factor models with linear and non-linear structure of interactions ..................................................... 101
6.5 Conclusions ................................................................. 105

# Conclusions and future work

7.1 Future work ................................................................. 110

# Posterior computation

A Posterior computation for Chapter 2 ............................................ 115
B Posterior computation for Chapter 3 ............................................ 118
C Posterior computation for Chapter 4 ............................................ 122
C.1 Likelihood function and complete conditional posterior distributions . 122
C.2 Derivative free adaptive rejection sampling to update $\alpha_{il}$ ............... 126
D Posterior computation and additional graphs for Chapter 5 129

D.1 Likelihood function and complete conditional posterior distributions . 129
D.2 Additional results for the simulated study . . . . . . . . . . . . . . . 133
D.3 Additional results for the real data analysis . . . . . . . . . . . . . . 136

E Linear factor model to select genes in the CNA problem 138

E.1 The two-factor model . . . . . . . . . . . . . . . . . . . . . . . . . . 139
E.2 Synthetic data application . . . . . . . . . . . . . . . . . . . . . . . 140
E.3 Real application: gene selection . . . . . . . . . . . . . . . . . . . . 143

F Posterior computation and additional graphs for Chapter 6 146

F.1 Likelihood function and complete conditional posterior distributions . 146
  F.1.1 Prior and complete conditional for $\alpha_i$ and its associated pa-
        rameters . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 148
  F.1.2 Prior and complete conditional for $F$ and its associated pa-
        rameters . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
  F.1.3 Prior and complete conditional for $\lambda$ . . . . . . . . . . . . . 154
F.2 Additional results for the simulated study . . . . . . . . . . . . . . 156
F.3 Additional results for the real data analysis . . . . . . . . . . . . . . 161
F.4 Additional results for the analysis comparing the models with linear
        and non-linear structure . . . . . . . . . . . . . . . . . . . . . . . . 164

Bibliography 168

Biography 175
## List of Tables

2.1 Detection calls of probe sets containing probes matching spiked sequences. .............................................................. 26

4.1 Number of sequences where the gene identification from the RNA-Seq data is confirmed via BLAST. ................................................. 50

5.1 Pairwise intersections between the breast cancer data sets; number of common genes affected by the multiplicative interaction. In order to simplify the notation, consider the first name indicated in the citation of the related reference. ................................................................. 77

6.1 Prior specifications defining different models in the study of non-linear interactions between factors. ....................................................... 88

6.2 Prior distributions and initial values specified for the Bernoulli probabilities related to the indicators $h_{il}$ and $z_{ii}$. ......................................................... 89

6.3 Regions detected with CNA in the human genome. We apply a procedure to remove genes unrelated to the CNA factors. The number of genes before and after this removal is presented. ........................................... 95

6.4 Pairwise intersections between the breast cancer data sets; number of common genes affected by interactions. In order to simplify the notation, consider the first name indicated in the citation of the related reference. All target pairs of regions are evaluated. ............................................. 101

E.1 Regions detected with CNA in the human genome. Number of genes before and after the cleaning procedure. ........................................... 143
List of Figures

2.1 Intensities of all probes within two different probe sets (a) and (b). Samples are along the x-axis and probes are on the y-axis. White corresponds to relatively high expression levels of the corresponding probe in the corresponding sample. In both (a) and (b), the columns have been sorted so that the first principal component is monotone. Image (a) is an example displaying a strong and consistent pattern of intensities for every probe across samples. Image (b) is an example of a noise probe set with probes randomly alternating intensities across the arrays. ................................................................. 12

2.2 Synthetic data representing intensities of a probe set across microarrays. Images (a) and (b) reproduce the scenario of Figure 2.1 (a) and (b), respectively. ................................................................. 15

2.3 Real value (circle), posterior mean (x mark) and 95% credible interval (bar) for $\sigma_i^2$, $i = 1, 2, ..., m$, and each component of $\alpha$ and $\lambda$. First column associated with Figure 2.2 (a), and second column related to Figure 2.2 (b). ................................................................. 16

2.4 ROC curves comparing true positive and false positive rates for the BFM (asterisk), PANP (black lines) and MAS 5.0 P/A (gray lines). ................................................................. 20

2.5 Panel (a): True positive rates obtained via BFM for different number of samples and different choices of $\omega_1$. Panel (b): TP rates for different choices of $\omega_1$ and assuming 251 microarrays fixed. ................................................................. 20

2.6 True positive rates obtained via BFM for different choices of $\omega_1$ and assuming 251 microarrays fixed. Panel (a) for $\omega_2 = 10$ and Panel (b) for $\omega_2 = 5$. ................................................................. 22

2.7 Detection calls obtained via BFM, PANP and MAS 5.0 P/A for 42 spiked-in genes. Presence calls are displayed in gray and non-presence calls in white. We note that all 42 spiked-in genes are detected as present using BFM, which is a significant improvement over PANP or MAS 5.0 P/A. ................................................................. 25
2.8 Intensities of two probe sets across microarrays in the spike-in study. Image (a) is an example of a probe set showing a consistent and strong pattern across samples, with detection call “Present” obtained by the BFM, and detection call “Absent” obtained by both other methods. Image (b) is an example of a probe set detected as “Absent” by the BFM.

3.1 Real value (circle), posterior mean (x mark) and the 95% credible interval (bar) for $\alpha$ in Panel (a), $\lambda$ in (b) and $\sigma^2$ in (c). Panel (d) shows box plots indicating the distribution of the conditional posterior probability that $\alpha_i \neq 0$, and the heat map image of the simulated matrix $X$ with columns sorted so that the first principal component is monotone.

3.2 Factor analysis of gene expression across samples. Probe-level analysis using the model (2.1) proposed in Section 2.2: Presence (white) and Absence (black) calls are shown to the left of some images. Gene-level analysis using the model proposed in Section 3.1: Box plots representing the conditional probability (3.2) are shown in the middle of each panel. Rows 1 and 2 are associated with breast cancer; row 3 contains graphs related to lung, ovarian and brain cancer, respectively. The Genes are located between positions 108,000,000 and 112,000,000 in Chromosome 6.

3.3 Factor analysis of gene expression across samples. Probe-level analysis using the model (2.1) proposed in Section 2.2: Presence (white) and Absence (black) calls are shown to the left of some images. Gene-level analysis using the model proposed in Section 3.1: Box plots representing the conditional probability (3.2) are shown in the middle of each panel. Rows 1 and 2 are associated with breast cancer; row 3 contains graphs related to lung, ovarian and brain cancer, respectively. The Genes are located between positions 208,000,000 and 212,000,000 in Chromosome 1.

3.4 Box plots comparing the posterior distribution of $q$ for all data sets.

3.5 Posterior mean (x mark) and 95% credible interval (bar) for $\alpha$. We work with the gene list identified in Chromosome 1. Breast cancer in row 1, brain cancer in row 2, factor model of Section 2.2 (Chapter 2) in column 1, and factor model of Section 3.1 (Chapter 3) in column 2.

4.1 Count of reads for each sequence across samples. Target gene: “C6orf162” (chromosome 6) in Panel (a), “TMED4” (chr. 7) in (b), “ARL4C” (chr. 2) in (c), and “PHLDA1” (chr. 12) in (d). The original count data is displayed (columns are not sorted and rows are not standardized).
4.2 Count of reads for each sequence across samples. Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d). In order to improve the pattern visualization, the images show X with standardized rows and sorted columns so that the 1st principal component is monotone.

4.3 Box plots showing the distribution of the conditional probability (4.3). In each panel, the box plots are grouped according to the choice of $\omega_2$. Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d).

4.4 Box plots showing the posterior distribution of $q$ (probability that $\alpha_i \neq 0$ for any $i$). Each box plot is associated with a different $\omega_2$. Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d).

4.5 Simulated data representing count of reads across samples. The columns are not sorted and the rows are not standardized in each matrix X.

4.6 Box plots showing the distribution of the conditional probability (4.3), and heat map images of the simulated data sets. In order to improve the expression pattern visualization, the images show X with standardized rows and sorted columns so that the first principal component is monotone.

4.7 Real value (circle), posterior mean (x mark) and 95% credible interval (bar) for $\eta_i$, $\alpha_i$ and $\lambda_j$. Panels (a), (b), (c) and (d) in Figure 4.6 are associated with the graphs in rows 1, 2, 3 and 4 of this figure, respectively.

5.1 Synthetic data with multiplicative interaction effects ($n_f = 2$). The original data is displayed in the top. The second panel shows X with standardized rows (rows and columns are sorted so that the first principal component is monotone).

5.2 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$. Dashed lines separate the factors. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).
5.3 Images displaying the posterior mean and the true loading associated with the interaction factor. Box plots representing the distribution of the conditional probability that $h_{33} = 1$. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

5.4 Scatter plots comparing the posterior estimates with the true values. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

5.5 Scatter plot comparing the posterior estimate of $\lambda_{3j}$ and $\lambda_{1j}\lambda_{2j}$ (Assume approach 1 = Gaussian prior). Each panel represents a different breast cancer data set: Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).

5.6 Posterior mean (x mark) and the 95% credible interval (bar) for the loadings with $i \in (G_1 \cup G_2)$ (Assume approach 2 = perfect product). Intervals for $\alpha_{il}$ are computed for the component with highest probability weight a posteriori. Dashed lines separate the factors. Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).

5.7 3-dimensional surface plots representing the multiplicative interaction effect $\alpha_{i3}\lambda_{3j}$ (Assume approach 2 = perfect product). In each panel, the left graph is related to the smallest negative loading; whereas, the graph on the right is related to the largest positive loading. Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).

6.1 Synthetic matrix $X$. The original data is displayed in the top. The second panel shows $X$ with standardized rows (rows and columns are sorted so that the first principal component is monotone).

6.2 Results from Model 1: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (a), $\lambda_{ij}$ (b), $\sigma_i^2$ (c) and $F_{ij}$ (d).

6.3 Box plots representing the posterior distributions of probabilities parameters. Panel (a) is related to $\rho$ defined in the Models 3 and 4. Panel (b) is associated with $q_R$ and $p_R$ specified in Model 5.
6.4 Average absolute distance between posterior mean and the true value. Comparison involving all models and simulated data sets.

6.5 Matrix $F$ containing interaction effects: Panel 1 = full matrix (3744 genes), Panel 2 = the cases $F_i \neq 0$. These results are related to the pair of locations (2,4). In Panel 2, rows and columns are sorted so that the 1st principal components are monotone.

6.6 Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ such that $i \in (G_1 \cup G_2)$. The dashed line separates the two factors. Panels (a), (b) and (c) are related to the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005). These results are related to the pair of locations (2,4).

6.7 3-dimensional surface plot representing the estimated interaction effect in $F_{1524}$ (a) and $F_{1945}$ (c). Panels (b) and (d) contain the posterior mean (x mark), used to create the surfaces, and the corresponding 95% credible interval (bar). This result is related to the data set Chin et al. (2006) and the pair of chromosome locations (2,4).

6.8 3-dimensional surface plot representing the true interaction effect in all simulated data sets.

6.9 3-dimensional surface plot representing the estimated interaction effect. Panel (a) = Model 1, (b) = Model 2, (c) = model with approach 1 (Gaussian prior, Chapter 5), and (d) = model with approach 2 (perfect product, Chapter 5). Models 1 and 2 are fitted assuming $l_s = 0.2$.

6.10 3-dimensional surface plot representing the estimated interaction effect. Here, the models are fitted assuming $l_s = 0.3$ or 0.5.

6.11 Average absolute distance between posterior mean and the true value. Comparison involving Models 1-2 and different choices of the length-scale parameter $l_s$. Only the data set from Simulation 1 is used.

D.1 Synthetic data with multiplicative interaction effects ($n_f = 3$). The original data is displayed in the top. The second panel shows $X$ with standardized rows (rows and columns are sorted so that the first principal component is monotone).

D.2 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$. Dashed lines separate the factors. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).
D.3 Images displaying the posterior mean and the true loading associated with the interaction factor $l = 4$ (row 1), $l = 5$ (row 2) and $l = 6$ (row 3). Box plots representing the distribution of the conditional probability that $h_{il} = 1$ for $l = 4, 5, 6$. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

D.4 Scatter plots comparing the posterior estimates with the true values. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

D.5 Image of the matrix $X$ containing the data. Rows and columns are sorted so that the first principal component is monotone (rows are standardized).

D.6 Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{i3}$. . 137

D.7 Posterior mean (x mark) and the 95% credible interval (bar) for $\lambda$. . 137

E.1 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar). Panel (a) shows only the factor loadings for $i \in (G_1 \cup G_2)$, Panel (b) shows all $\alpha_{il}$. Panels (c) and (d) are related to $\lambda$ and $\sigma^2$, respectively. Intervals for $\alpha_{il}$ are computed for the component with highest probability weight a posteriori.

E.2 Distribution of the conditional probability that $h_{il} = 1$. Each bar contains the 25th, 50th and 75th percentiles. Panels (a), (b) and (c) represent the pairs (1,4), (2,4) and (3,4), respectively.

F.1 Synthetic matrix $X$. The original data is displayed in the top of each panel. The image in the bottom shows $X$ with standardized rows (rows and columns are sorted so that the first principal component is monotone).

F.2 Results from Model 1: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$, $\lambda_{ij}$, $\sigma^2_i$ and $F_{ij}$. Panels in rows 1-2 = Simulation 2. Panels in rows 3-4 = Simulation 3.

F.3 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 1.
F.4 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 2. .......................................................... 159

F.5 Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 3. .......................................................... 160

F.6 Matrix $F$ containing the interaction effects. Column 1: full matrix (3744 genes); Column 2: matrix with the cases $F_i \neq 0$ (rows and columns are sorted so that the 1st principal components are monotone); Row 1: Pair of chromosome locations (1,4); Row 2: Pair of locations (3,4). .......................................................... 161

F.7 Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ ($i \in G_1 \cup G_2$). The dashed line separates the two factors. Rows 1-3 are related to the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005). Columns 1-2 correspond to the pairs of chromosome locations (1,4) and (3,4). .......................................................... 162

F.8 Column 1: 3-dimensional surface plot representing the estimated interaction effect in $F_{1859}$, $F_{3155}$, $F_{2006}$, and $F_{2719}$. Column 2: posterior mean (x mark), used to create the surfaces, and the corresponding 95% credible interval. Consider the data set Chin et al. (2006); Rows 1-2 = pair of chromosome locations (1,4), Rows 3-4 = pair (3,4). .......................................................... 163

F.9 3-dimensional surface plot representing the estimated interaction effect. According to Chapter 5: Approach 1 = model assuming $\lambda_{3j} \sim N(\lambda_{1j}^{\lambda_{2j}}, \nu)$, and Approach 2 = model assuming the equality $\lambda_{3j} = \lambda_{1j}^{\lambda_{2j}}$. .......................................................... 164

F.10 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 1. .......................................................... 165

F.11 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 2. The models are fitted assuming $l_s = 0.2$. .......................................................... 166

F.12 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 3. The models are fitted assuming $l_s = 0.2$. .......................................................... 167
Symbols

\[ p \] Probability function or density
\[ X \] Matrix containing the data
\[ \alpha \] Factor loadings
\[ \lambda \] Factor scores
\[ \epsilon \] Idiosyncratic noise
\[ \sigma^2, \omega, \phi, \nu \] Variances
\[ a, b \] Parameters in the Inverse Gamma distribution
\[ \gamma_1, \gamma_2, \beta_1, \beta_2 \] Parameters in the Beta distribution
\[ \mu, \theta \] Mean parameters
\[ \mu_{ij} \] Poisson rate
\[ \eta_i \] Subject specific mean (Poisson model)
\[ F \] Matrix of non-linear interaction effects
\[ \Sigma, K(\lambda) \] Covariance matrices
\[ l_s \] Characteristic length scale parameter
\[ q, \rho \] Bernoulli probability parameters
\[ h, z \] Indicator variable: 1 if target parameter \( \neq 0 \), 0 otherwise
\[ \psi \] Generic parameter
\[ \hat{\psi} \] Estimated value
\[ \psi^{(0)} \] Initial value for the MCMC simulation
\[ \delta_0(\psi) \] Point mass distribution at 0, i.e., \( p(\psi = 0) = 1 \)

\( m \) Number of subjects

\( n \) Number of samples

\( L \) Total number of factors

\( n_f \) Number of non-interaction factors

\( I_n \) \((n \times n)\) Identity matrix

\( 1_n \) \(n\)-dimensional row vector of ones

\( G_l, G_E \) Set of subjects (genes)

\( C \) Real-valued constant

\( \mathbb{R} \) Set of real numbers

**Abbreviations**

- **AAD** Average Absolute Distance
- **ANOVA** Analysis of Variance
- **ARS** Adaptive Rejection Sampling
- **BFM** Bayesian Factor Model
- **BLAST** Basic Local Alignment Search Tool
- **CNA** Copy Number Alteration
- **DABG** Detection Above Background method
- **DAG** Directed Acyclic Graph
- **DP** Dirichlet Process
- **DPPCA** Dual Probabilistic Principal Component Analysis
- **FP** False Positive rate
- **GCRMA** Guanine Cytosine Robust Multi-array Analysis
- **GP** Gaussian Process
- **GP-LVM** Gaussian Process Latent Variable Model
GPois  Generalized Poisson distribution
IG   Inverse Gamma distribution
MAP  Maximum a posteriori estimation
MAS 5.0 Microarray Suite version 5.0 (preprocessing technique)
MAS 5.0 P/A MAS 5.0 Presence/Absence (detection method)
MCMC Markov Chain Monte Carlo
MM   Mismatch probe
N    Normal distribution
NMF  Non-negative Matrix Factorization
NSMPs Negative Strand Matching Probe sets
PANP Presence Absence calls with Negative Probe sets
P/A  Presence/Absence
PCA  Principal Component Analysis
PLS  Partial Least Squares
PM   Perfect Match probe
RMA  Robust Multi-array Average
RNA-Seq High-throughput sequencing data
ROC  Receiver Operating Characteristic curve
SE   Squared Exponential covariance function
SSVS Stochastic Search Variable Selection
TP   True Positive rate
U    Uniform distribution
Acknowledgements

The intellectual and personal growth I have experienced at Duke University is unmatched by any other period of my life. The fulfillment of the Ph.D. requirements and the research presented here are proud achievements. For that, I thank my family for always being there and believing in me; Joe Lucas for his continued support, supervision, understanding, and willing attention; Mike West for his patience and guidance during my first year; and all faculty members who have served as instructors of the many classes I have taken at Duke. Finally, I offer my regards to all of those who supported me in any respect during my graduate school years.

This study was supported by Award Number U54CA112952 from the National Cancer Institute, by funding from the Measurement to Understand Re-Classification of Disease of Cabarrus and Kannapolis (MURDOCK) Study and by a grant from the NIH CTSA (Clinical and Translational Science Award) 1UL1RR024128-01 to Duke University.
Introduction

Multivariate statistical methods for analysis of high-dimensional data have been the topic of many publications in the past years. A challenging issue, connected with high-dimensional data, is the fact that the number of variables is much larger than the number of samples. In particular, gene expression data from DNA microarrays are characterized by measurements of many different genes, often in only a few samples. Although the number of genes is large, there may be only a few underlying cases accounting for much of the variation in the data. The analysis of high dimensional data requires special techniques such as variable selection or dimension reduction. Factor models are a flexible and powerful tool to analyze multivariate dependence and to verify patterns and relationships in the data. Several types of factor models with different constraints and computational algorithms can be found in the literature. The Principal Component Analysis (PCA) is a well known dimension reduction method used in many applications involving gene expression; see for example Yeung and Ruzzo (2001). However, the literature indicates that other techniques can be superior to reduce dimension. As an example, consider the Partial Least Squares (PLS) method compared with PCA in Nguyen and Rocke (2002), and also evaluated
in Boulesteix and Strimmer (2006). The regression response variable is taken into account in the PLS approach as opposed to PCA, and thus the PLS usually performs better than PCA in prediction problems. Another technique used to reduce the dimensionality of the data, defining a linear combination of a reduced set of factors, is called Non-negative Matrix Factorization (NMF). The method is applied to the analysis of gene expression by Brunet et al. (2004) and Kim and Tidor (2003), and within this context the factors represent groups of genes (metagenes) strongly correlated in subsets of the data. The NMF might be useful to study biological subsystems, because it can identify local and global patterns of similarities between genes. In contrast, other techniques such as PCA focus only on global patterns.

Computational advances have been critical in enabling the application of complex models for analysis of large data sets. The development of iterative MCMC simulation methods has contributed to the increasing attention devoted to the Bayesian framework as an attractive alternative to work with factor models. In recent years, numerous studies have applied factor models combined with the Bayesian approach to the analysis of gene expression data, and their results often show an improvement in the identification and estimation of metagene groups and patterns related to the underlying biology. As an example, West (2003) introduced sparse latent factor models, as a natural extension of the sparse regression modeling. The study assumes typical Bayesian variable selection priors and demonstrates the ability of latent factor models to describe pattern/signature profiling in expression genomics. Lucas et al. (2006) also apply hierarchical sparsity priors and obtain substantial improvements in terms of identification of complex patterns of covariation among genes. The paper explores the Bayesian methodology for large-scale regression, ANOVA and latent factor models. Carvalho et al. (2008) is another reference working with sparsity priors to address dimension reduction on latent factor models applied to gene expression data. Stochastic simulation and evolutionary stochastic search methods are
used in the paper to address questions of uncertainty about the number of latent factors. This same issue is also evaluated in Lopes and West (2004) via reversible jump MCMC methods.

In our study, the bimodal sparsity promoting priors are key elements in the structure of the factor models defined for different types of applications. This form of prior distribution originated in the context of Bayesian variable selection, and it has been the subject of substantial research in the past years. George and McCulloch (1993) proposed a Bayes procedure for selecting promising subsets of predictors in normal linear regression models. This procedure, called SSVS (Stochastic Search Variable Selection), entails the specification of a hierarchical Bayes mixture prior which uses the data to assign higher posterior probability to the more promising models. Each component of the mixture is modeled as having come from a scale mixture of two normal distributions with variance hyperparameters set “small and large”. Few years later, Geweke (1996) proposed an independent prior distribution for each coefficient that is a mixture of a point mass at zero (δ₀) and a possibly truncated univariate normal distribution. George and McCulloch (1997) describe and compare many hierarchical mixture prior formulations including the SSVS case and the choice with the δ₀ component. In summary, their results indicate that the prior with two Gaussian components is useful for removing predictors whose coefficients are too small to be practically significant, whereas, the prior with δ₀ eliminates only those predictors whose coefficients cannot be distinguished from 0. We apply mixture priors to evaluate the relationship between latent factors and the observed data. Depending on the application, we will use the specification with two Gaussian components or the choice with δ₀. This decision has to do with the conclusions from George and McCulloch (1997). The mixture prior with the δ₀ component (sometimes referred to as “spike and slab” prior) has been used effectively to define the sparse factor structure in West (2003), Lucas et al. (2006) and Carvalho et al. (2008).
In this work, we use factor analysis to explore different problems related to high dimensional gene expression data. In each section, we utilize simulated studies, and real data sets representing different types of cancer to exemplify the indicated methods. Two types of technologies to measure gene expression are evaluated here. In most chapters we consider the DNA microarrays, which make possible the rapid and comprehensive assessment of the transcriptional activity of a cell. The major challenge in using this technology is the analysis of its large data output. In Chapter 4, we focus on a promising new technology known as RNA-Seq or high-throughput sequencing data. All the models have a hierarchical structure including sparsity-priors for the factor loadings and other target parameters. The main aims of the studies in each chapter are listed below. This dissertation is structured as follows:

**Chapter 2** We will use a factor model assuming a single latent factor. The goal of this study is to evaluate the co-expression of probes across samples (microarrays) to determine the presence/absence classification for the corresponding gene. No other method in the literature takes advantage of this type of information to generate gene detection calls. Results from our model are compared with the results from other two detection techniques proposed in the literature.

**Chapter 3** Duplications affecting the number of transcripts in regions of human chromosomes are known to occur for breast cancer tumors. We examine a group of genes located altogether in a chromosomal region detected with copy number alteration for this type of cancer. Images of their probe sets across arrays (breast cancer data) provide visual identification of patterns suggesting copy number variation. The factor analysis can be used as a test providing a more accurate answer than the visual analysis. The main idea of this application is to use factor models to analyze the expression pattern across samples, and then investigate the presence of copy number change in the same chromoso-
mal location for other types of cancer. Many methods have been developed for identification of such chromosomal abnormality; however, none of them compare data sets representing different tumors. Further, the analysis of coherent patterns can be considered a new approach to address the identification of copy number alteration.

**Chapter 4** The introduction of high-throughput sequencing technologies has opened new doors into the field of gene expression. In this chapter, we evaluate the coherent pattern across samples of RNA-Seq data; our goal is to study the presence/absence status of genes and the corresponding read mapping uncertainty. A linear factor model based on the Poisson distribution is proposed to analyze the count data across samples. Our approach differs from all other studies in the literature dealing with this young gene expression technology.

**Chapter 5** In an ordinary factor analysis the involvement of any subject with the factors is always additive. In the context of gene expression data, the interaction between factors can have important implications in the model as a result of biological pathways defining complex structure of dependencies between genes. In this chapter, we study the expression patterns across microarrays using multi-factor models; our goal is to investigate the existence of multiplicative interactions between latent factors. The approach presented in Chapter 5 for this problem has not yet been considered in the literature. In the real data application we examine interactions effects related to two chromosomal regions detected with copy number alteration.

**Chapter 6** Here we explore a multi-factor model with a non-linear structure of interactions between factors; this formulation is more general and includes the model in Chapter 5 as a particular case. Again, the study of relationships be-
between genes motivates our analysis of interaction effects. In short, we introduce non-linearities through the specification of a mixture prior having a Gaussian component with covariance matrix determined via the Square Exponential kernel. The existing studies assuming non-linear structure of latent factors are based on optimization and deterministic approximations. Our method is novel given the following aspects: (i) it is focused on MCMC simulation to sample from the intractable posterior distributions, (ii) it is designed for the analysis of expression patterns across samples. The copy number change application developed in Chapter 5 is revisited here; also, we include a comparison between the models with linear and non-linear structure.

**Chapter 7** The final chapter summarizes the main conclusions of this dissertation. In addition, we indicate the future directions to improve the current studies.

The algorithms required to fit the proposed models are implemented using the MATLAB programming language (http://www.mathworks.com). In Chapter 2, we utilize two packages of functions written in the open source statistical language R (http://cran.r-project.org).
In gene expression analysis, for any target sample it is not likely that transcripts from all genes are present in the hybridization solution. The microarray contains probe sets whose probe sequences cannot find complementary sequences in the sample, and probe sets where transcripts are detected. As a result, an efficient method is required for determining the presence or absence of transcripts from the genes. Approaches exploring this problem have been proposed for different microarray platforms. As an example, Affymetrix implemented the detection above background (DABG) method for detection calls in GeneChip® Exon Arrays (Affymetrix, 2005). An empirical p-value is obtained for each probe by comparing the distribution of background probe intensities with identical G-C nucleic acid content. A transformation is applied to the p-values, and a Chi-square distribution is identified for their sum which allows a statistical test with null hypothesis assuming that the probes are detecting only background. As an alternative, Kapur et al. (2007) apply a method called “GeneBASE” to investigate presence/absence calls in the same type of microarray. In short, the method applies a statistical test to compare the observed probe intensities with their
background levels predicted by a specific model; a gene is classified as “absent” if the test indicates that these two quantities do not differ significantly. Li and Wong (2001) study model-based expression indices and investigate results for different tissue types and experiments involving replicate arrays. A software “dChip” is implemented for their analysis, and it includes a presence/absence method similar to the MAS 5.0 detection procedure. In brief, the authors calculate the proportion of presence calls for some probe set in a group of arrays, and use this result to evaluate how accurate are their estimates.

A popular detection method was developed for Affymetrix GeneChip® oligonucleotide arrays, and it is implemented as part of the preprocessing technique Microarray Suite version 5.0 or MAS 5.0 (Affymetrix, 2001). In this case, the detection calls can take the values absent, marginal or present. Both perfect match (PM) and mismatch (MM) probes are used to calculate the score \( R = (PM - MM)/(PM + MM) \) for each probe pair within a probe set. First, the method tests whether all the probe pairs are saturated, if so the probe set is classified as present, otherwise a one-sided Wilcoxon’s signed rank test is applied to obtain a p-value which is used to assign a detection call. McClintick and Edenberg (2006) and Liu et al. (2002) apply this presence/absence (P/A) method and show that it works relatively well in practice; however, other frameworks improving this solution can be found in the literature. As an example, Wu and Irizarry (2005) define a procedure called Half-price method. They argue that using MM data may be problematic in gene expression analysis, and then define a method based on PM probes only. Their results indicate that the half-price technique outperforms MAS 5.0 in terms of detection calls.

Another interesting method was proposed by Warren et al. (2007), and it is named Presence Absence calls with Negative Probe sets (PANP). In brief, the authors identify Affymetrix probe sets which cannot hybridize to the intended target, because they have been designed in the reverse direction against their own transcripts. These
probe sets are called Negative Strand Matching Probe sets (NSMPs) and can be obtained from chip annotation files available on Affymetrix webpage. The selected NSMPs are assumed as controls, and the empirical cumulative distribution of their intensities is used to derive a cutoff intensity; a particular probe set is classified as present if its expression value is higher than the threshold. Before applying this analysis, the data are preprocessed using any technique. In particular, the performance of PANP is evaluated in the paper with data obtained from RMA (Irizarry et al., 2003b), GCRMA (Wu et al., 2004) and MAS 5.0. Besides simplicity, the authors indicate that another advantage of PANP is the fact that it works with preprocessing techniques using (PM, MM) or PM-only probes. The paper shows that PANP combined with RMA, GCRMA and even MAS 5.0 outperforms the default detection method of MAS 5.0 in terms of several metrics of accuracy and precision. The choice of threshold in PANP is somewhat arbitrary and has a strong impact over the final result. If the threshold is slightly shifted, the detection calls of some probe sets will change.

In none of these techniques the coherent expression is used to identify the presence of a gene. Our Bayesian factor model is proposed to take advantage of the tremendous amount of information – available from studying the co-expression of probes across the samples – about the behavior of a probe set. A consistent expression sequence for all probes clearly indicate presence, whereas, probes randomly alternating intensities across arrays indicate absence.

This chapter is organized as follows. The data and a few strategies to adjust the measurements are described in Section 2.1. Next, the Bayesian factor model (BFM) is proposed in Section 2.2. Section 2.3 shows inference results to verify the performance of the BFM in terms of parameter estimation. Section 2.4 presents the comparison between BFM and other detection methods, using a simulated data set. In Section 2.5 we develop the comparison analysis based on a real data set from a
spike-in study designed by Affymetrix. Finally, Section 2.6 presents the conclusions of this chapter.

2.1 The data

Suppose \( n \) microarrays are available for analysis; each array contains \( K \) probe sets. Miller et al. (2002) indicates that the variability in the data must be taken into account in order to develop an optimal analysis. The first step in a usual microarray analysis is to preprocess the data; the observed intensity values are transformed to remove the noise effect in an individual chip, and to adjust the information obtained from replicate microarrays. Details about three well known preprocessing techniques can be found in Irizarry et al. (2003b), Wu et al. (2004) and Affymetrix (2001); see Berger et al. (2004) for an application using MAS 5.0. Our study aims to propose a new gene detection method which does not depend on any preprocessing technique suggested in the literature; however, the technical effects affecting the data cannot be ignored. As a result, few manipulations are applied.

1. In an individual chip, there are technical variables that affect the vast majority of spots on the microarray, such as total RNA in the sample and camera exposure time. When examining expression patterns across the samples, the overwhelming signal in the raw data reflects these effects. We address the indicated problem by dividing the probe intensities by the corresponding intensity mean computed for each array.

2. Even though samples may be extracted from the same type of tissue, the distribution of intensity values has a broad range and is highly skewed. Thus it is standard practice to log-transform expression data.

3. Consider a matrix \( X \) containing intensities of probe set \( k \in \{1, \ldots, K\} \) from each chip. Each row represents an individual probe within the probe set, and each
column represents a microarray. The $n$ intensities observed for any probe are assumed to follow a Normal distribution. Different probes may be associated with different mean and variance parameters. In order to evaluate whether the pattern of expression is consistent across samples for all probes belonging to probe set $k$, the rows of the indicated $X$ matrix will be standardized.

4. Some microarrays are brighter than others due to technical effects introduced during their production (e.g. scanner setting and physical problems). In order to subtract away such effects, we compute the first principal component $(pc_1)$ based on a random subset of 500 genes selected from the population of genes in the microarray. For each row $i$ representing a probe, we subtract the component of its expression in the direction of the first principal component $(X_i - X_i pc_1 pc_1')$. See footnote 1 for further details about this step.

We recognize that there are a number of different approaches to subtracting technical data collection effects from the raw data. The factor model we utilize is general and can be applied with any data cleansing approach.

Figure 2.1 shows the image of the specified $X$ matrix for different probe sets. The previous manipulations have been applied to the original data. The data correspond to 251 oligonucleotide microarrays containing 22283 probe sets. The transcripts in the hybridization solution are extracted from breast cancer tumors. As a reference for this data set consider Miller et al. (2005) which investigates the importance of the p53 tumor suppressor functional status for predicting human breast cancer behavior. Two distinct patterns can be observed when comparing the images. Image (a) displays a consistent sequence of intensities for every probe across the arrays. The values decrease when moving from the left to the right, and this pattern is the same for all probes. In Image (a), the gene is present and the associated probes indicate such presence by exhibiting similar intensities within each array. On the other hand,
Image (b) shows that the probes randomly alternate intensities across samples. No pattern can be observed and within each array the probes disagree from each other by expressing different levels of intensities. This situation indicates absence of the gene, and the displayed values are just noise effects.

Figure 2.1: Intensities of all probes within two different probe sets (a) and (b). Samples are along the x-axis and probes are on the y-axis. White corresponds to relatively high expression levels of the corresponding probe in the corresponding sample. In both (a) and (b), the columns have been sorted so that the first principal component is monotone. Image (a) is an example displaying a strong and consistent pattern of intensities for every probe across samples. Image (b) is an example of a noise probe set with probes randomly alternating intensities across the arrays.

The existing detection methods (MAS 5.0 P/A and PANP) can provide presence/absence calls for every microarray and every gene, whereas, the factor model proposed in the next section uses the information from several microarrays contained in \( X \) to define a single detection call for the corresponding gene. In Figure 2.1 (a), the gene may be absent in a few samples; hence MAS 5.0 P/A and PANP can provide conflicting detection calls for different microarrays. The factor model indicates the presence call for this scenario based on the coherent pattern across all arrays, and this is interesting for purposes of analysis of the whole experiment.
2.2 Bayesian factor model for gene expression detection

Suppose $X_{ij}$ is the log of light intensity of the $i$-th probe within a probe set on sample $j = 1, 2, ..., n$. Assume further that the transformations indicated in Section 2.1 have been applied to the data. Let $X$ be the matrix with element $X_{ij}$ in the $i$-th row and $j$-th column. The following model\(^1\) is assumed for the expression data of $X$:

$$X = \alpha \lambda + \epsilon$$

where $\alpha$ is an $m$-dimensional column vector reflecting the strength of hybridization between the target of the probe set and each of the $m$ probes, component $\lambda$ is an $n$-dimensional row vector describing the pattern of expression of the target, and finally $\epsilon$ is the idiosyncratic noise term.

We assume a mixture prior distribution on the factor loadings.

$$\alpha \sim (1 - q)N_m(0, \Phi_1) + qN_m(0, \Phi_2)$$

with $\Phi_1 = \omega_1 I_m$ and $\Phi_2 = \omega_2 I_m$. We define $q$ as the prior probability of detecting the gene as present. A small and positive number is suggested for the scalar $\omega_1$, while $\omega_2$ is large and positive. Because we treat probe sets that have a high probability of having derived from the first component as noise, the relative sizes of $\omega_1$ and $\omega_2$ effectively define our estimated signal to noise ratio. A small $\omega_1$ is chosen for defining a Normal component centered on zero with small variability, which indicates that the factor loadings are close to zero, and then the detection call “Absent” is appropriate.

A point mass distribution at 0 has been considered as a replacement for component $N_m(0, \Phi_1)$ in the mixture prior (2.2). However, this is unsuitable for the factor

---

\(^1\) We admit there is some variation that is systematic across all probe sets. Ideally, we could address this issue by assuming the model $X_k = \beta \delta + \alpha_k \lambda_k + \epsilon$ for a probe set $k$, where $\beta \delta$ is the systematic variation. In this case, we would have to work with the full data set due to the dependence between probe sets, and this task is computationally expensive. Our normalization step 4 (Section 2.1) subtracts off an approximation of $\beta \delta$; hence, we can assume model (2.1) and work with each probe set independently.
loadings because the classification “Present” is obtained only if a completely random sequence of intensities is observed across samples. In practice, there is often structure in a probe set that has not been completely subtracted by the data cleaning techniques. Additionally, because the probes come in pairs with each pair different in only 1 of 25 locations, there is built-in correlation between subsets of the probes even in the absence of the target gene of interest.

We complete the model formulation with the conjugate priors $\lambda \sim N_n(0, I_n)$ and $\epsilon_{ij} \sim N(0, \sigma^2_i)$ with an Inverse Gamma distribution $IG(a, b)$ for $\sigma^2_i$. Denote $\sigma^2 = (\sigma^2_1, \sigma^2_2, ..., \sigma^2_m)'.$

The joint posterior distribution $p(\alpha, \lambda, \sigma^2|X)$ cannot be evaluated analytically. In order to generate a sample from this distribution we implement the Gibbs sampler algorithm; see Gamerman and Lopes (2006) for details about this Markov Chain Monte Carlo method. The complete conditional posterior distributions can be determined, and they are used to sequentially update each parameter until a sample is generated after convergence of the chains to the target distribution. The likelihood function and the complete conditional posterior distributions are shown in Appendix A. In particular, during the Gibbs sampling we draw $\alpha$ from the second mixture component with probability

$$q^* = \frac{N[0|\Phi_2^*]}{N[0|M^*_2, \Phi^*_2]} q + \frac{N[0|\Phi_1^*]}{N[0|M^*_1, \Phi^*_1]} (1 - q),$$

where $M^*_1, M^*_2, \Phi^*_1$ and $\Phi^*_2$ are specified in Appendix A. It is the average of $q^*$ across all draws from the MCMC that is used to assign presence/absence calls in the BFM.

\[2\] In Section 2.4, we define a data generating procedure that requires the specification (2.1) including a mean expression parameter $\mu$. The full model (with $\mu$) is explored in Appendix A. Let $\mu = 0$ to identify results related to model (2.1).
2.3 Inference results of the Bayesian factor model

The main aim of this section is to verify the performance of the BFM in terms of inference results for its parameters. A simulated data set is considered in this application. First, the proposed BFM is fitted to the real data and the posterior estimates of the involved parameters are assumed as real values in the process of simulating data. Next, for each microarray a random value is generated from $N(0, \sigma_i^2)$ with $i = 1, 2, ..., m$ representing a probe; this is the noise term in the BFM. The product $\alpha \lambda$ results in a $m \times n$ matrix which is added to the previous noise matrix. The pattern of a noise probe set can be simulated by letting $\alpha$ and/or $\lambda$ be a null vector.

\[ \text{(a)} \quad \text{(b)} \]

\[ \text{Figure 2.2: Synthetic data representing intensities of a probe set across microarrays. Images (a) and (b) reproduce the scenario of Figure 2.1 (a) and (b), respectively.} \]

Assume as prior specifications: $\omega_1 = 0.01$ defining small variability for the first component within the mixture prior (2.2), $\omega_2 = 100$ defining large variability for the second component, $q = 0.5$ indicating equal prior probability for both components in the mixture. In addition, consider the Inverse Gamma prior distribution with $a = 2.1$ and $b = 1.1$, which has expected value 1, mode 0.3548 and variance 10. As initial values of the chain, consider $\lambda^{(0)}$ generated from its prior $N(0, I_n)$, and a
null vector for \( \alpha^{(0)} \); the value 1 is indicated for each \((\sigma_i^2)^{(0)}\). We are aware of the identifiability problem related to the sign of \( \alpha \) and \( \lambda \). This issue does not affect \( q^* \) in (2.3); therefore, constraints are not imposed to address the problem. The Gibbs sampler is run for 2000 iterations. The first 1000 elements of the chain are considered as burn-in period, and thus removed from the analysis. Fast convergence to the limiting distribution is observed for all chains.

**Figure 2.3**: Real value (circle), posterior mean (x mark) and 95% credible interval (bar) for \( \sigma_i^2 \), \( i = 1, 2, \ldots, m \), and each component of \( \alpha \) and \( \lambda \). First column associated with Figure 2.2 \((a)\), and second column related to Figure 2.2 \((b)\).
Two scenarios are chosen to examine results, the first one is presented in Figure 2.1 (a) showing a strong pattern of expression across samples, and the second one is indicated in Figure 2.1 (b) suggesting a noise probe set. The corresponding simulated data sets are displayed in Figure 2.2. As can be seen, the simulated data are very similar to the real data in Figure 2.1.

Figure 2.3 shows the real value, posterior mean and 95% credible interval for parameters involved in the factor model. Note that in most cases the credible interval includes the real value, and the posterior mean is close to that quantity. This result indicates a good performance of the model in terms of estimating parameters. In addition, consider the fact that the panels in the first column correspond to a probe set showing strong pattern across samples, and the second column is related to a noise probe set. The intervals for $\alpha$ and $\lambda$ are centered at 0 in column 2, whereas, column 1 shows several intervals away from 0 which is expected for a non-noise probe set.

The Gibbs sampler applied to the data shown in Figure 2.2 (a) and (b) indicates that the posterior probability $q^*$ converges to 1 and 0 respectively. This result confirms the visual interpretation of the images which suggests presence of the gene for Panel (a) and absence for Panel (b).

2.4 Comparison of detection methods, using a simulated data set

In a simulated scenario the true detection calls are known, and thus the performance of the methods can be evaluated. We will now estimate the characteristics of a real data set and use those estimates to generate a simulated data set for which we know whether each gene is present or absent. First, the real data is transformed as suggested in Section 2.1. We use 500 probe sets from this data set to generate 500 different simulated probe sets. The pairwise linear correlation coefficient between each pair of rows (probes) in the $X$ matrix is computed, resulting in a $m \times m$
matrix, with \( m \) being the number of probes. An interesting aspect can be observed in the correlation matrix, probe sets with strong pattern across samples exhibit high correlations (close to 1), whereas, a noise probe set is associated with low correlations (close to 0). This characteristic can be used to rank the probe sets by computing the average correlation in the matrix. A selection of 2000 probe sets are sorted in an increasing order of average correlations. Then, the first 200 cases are selected. In addition, starting with the 201\textsuperscript{st} probe set and moving towards the case 2000, we select every 6th probe set for use in generating simulated data. This strategy ensures the selection of a wide range of cases including strong, intermediate and weak patterns across samples.

Real values of parameters are determined for each selected probe set. In the previous section, this task was accomplished by fitting the BFM to the transformed data; however, in the present case other detection methods are considered in the analysis. MAS 5.0 P/A and PANP work with original data, i.e., without the manipulations described in Section 2.1, which includes standardizing the data. Therefore, a slight modification is required in the BFM indicated in (2.1). Consider

\[
X = \mu 1_n + \alpha \lambda + \epsilon
\]  

(2.4)

where \( \mu \) is an \( m \)-dimensional column vector whose entries are mean intensities fixed for each probe, and component \( 1_n \) is a \( n \)-dimensional row vector of ones. The interpretation of other components remains the same. Assume the prior specification \( \mu \sim N_m(0, 100I_m) \). The complete conditional posterior distributions for this model are specified in Appendix A.

In summary, we consider the following steps to generate the data:

1. The factor model (2.4) is fitted to the 500 selected \( X \) matrices containing the original data (without manipulations). Again, convergence to the limiting
distribution is fast, and the posterior mean is assumed in the next step as the real value of the involved parameters.

2. As described in the previous section, for each microarray a value is randomly generated from $N(0, \sigma_i^2)$ ($i = 1, 2, ..., m$) forming a noise matrix, which is added to the matrix $\mu_1 + \alpha \lambda$. The group of 200 selected probe sets showing the weakest patterns across samples are chosen to represent noise probe sets. In other words, $\alpha$ and $\lambda$ are set to be null vectors in these cases. The remaining 300 probe sets are generated using posterior estimates of all parameters.

3. The PANP method requires Negative Strand Matching Probe sets to be used as controls. In the generated data set, 100 probe sets simulated as noise cases are assumed as NSMPs. The detection call “Absent” is automatically assigned for these cases. In addition, the generated values are preprocessed via RMA when the method PANP is used.

4. The MAS 5.0 P/A method is applied to the synthetic data obtained in step 2.

5. In the analysis of the BFM, we first apply the data cleaning procedure suggested in Section 2.1 to the generated data. Next, we fit the factor model (2.1) to the “clean” synthetic data.

Figure 2.4 displays Receiver Operating Characteristic (ROC) curves for the three detection methods. This graph plots true positive (TP) rates against false positive (FP) rates computed for different choices of a threshold parameter. In fact, two thresholds are specified in PANP and MAS 5.0 P/A, below the first value the classification is “Present”, and above the second value the classification is “Absent”. In this analysis the detection “Marginal” will be suppressed by choosing the same number for both thresholds. In order to build the ROC curve, the detection methods (PANP and MAS 5.0 P/A) are applied to the data and their p-values are computed.
Next, different choices of the single threshold are compared with the p-values, and then detection calls are defined. The true classification is known because a simulated data set is used; therefore, TP and FP rates can be calculated for each case. The BFM defines a single list of detection calls taking into account all microarrays, whereas, the other methods generate a list of calls for each chip, for this reason 251 ROC curves are plotted for PANP and MAS 5.0 P/A in Figure 2.4.

![Figure 2.4](image)

**Figure 2.4:** ROC curves comparing true positive and false positive rates for the BFM (asterisk), PANP (black lines) and MAS 5.0 P/A (gray lines).

![Figure 2.5](image)

**Figure 2.5:** Panel (a): True positive rates obtained via BFM for different number of samples and different choices of $\omega_1$. Panel (b): TP rates for different choices of $\omega_1$ and assuming 251 microarrays fixed.

An interesting result is observed for the BFM, the posterior probability $q^*$ in (2.3) converges to 1 for any probe set exhibiting an intermediate or strong pattern
across samples. On the other hand, this probability converges to 0 for any probe set showing weak or no pattern. These extreme probabilities determine only presence or absence calls, and thus eliminate the need of defining thresholds for a “Marginal” detection call in the BFM. As a result, the ROC curve in Figure 2.4 for the BFM is represented by a single point, i.e., for any choice of threshold in the interval (0,1), the detection calls are always the same.

The high level of certainty in presence/absence calls is in part a function of the size of the data sets in question. We have 251 observations from (depending on the probe set) a 22-40 dimensional Multivariate Normal, which offers significant evidence for the presence or absence of a non-zero mean. We evaluate the behavior of $q^*$ in a MCMC run assuming a reduced number of probes and/or a reduced number of samples. The analysis involves a random selection of 5 rows and/or 5 columns of the two matrices presented in Figure 2.1. We found that probabilities $q^*$ that are different from 0 or 1 are common if a small sample size and/or few probes are used in the analysis. However, the simulations we have presented are reflective of many publically available data sets. In particular, the number of probes in a probe set is fixed and unchanging.

According to Figure 2.4, PANP outperforms MAS 5.0 P/A for detecting gene expression. The best result would be 0% false positives and 100% true positives which is represented by the point located in the top left corner of the graph. The closer the curve is to this point, the better is the performance of the method. As can be seen, most black lines (associated with PANP) are above the gray lines (representing MAS 5.0 P/A). This finding agrees with the result obtained by Warren et al. (2007). The point representing the BFM indicates 99.67% of TP and 0% of FP, and clearly suggests that this method outperforms the other two. All probe sets simulated as noise cases are correctly classified as “Absent” by the Bayesian model.

The number of microarrays available for analysis and the choice of $\omega_1$ have an
impact over the TP rate. Figure 2.5 explores these aspects. Panel (a) shows TP rates for different numbers of samples and different choices of $\omega_1$. Note that, for a fixed $\omega_1$, the rate increases as the number of arrays increases. The TP rate is 0% if only 10 arrays and $\omega_1 \geq 0.0075$ are considered. A strong pattern across 251 samples may not be displayed on 10 samples. The FP rate is 0% for all sample sizes. A random pattern across 251 samples is still random for smaller sample sizes. For a fixed number of arrays, the TP rate seems to decrease as $\omega_1$ increases. This aspect is also explored in Panel (b) which presents the TP rates for choices of $\omega_1$ ranging from 0.05 to 0.125; all 251 samples are used in the analysis. Recall that $\omega_1$ controls the variability in the first component of the mixture prior (2.2) specified for the factor loadings. As can be seen, the largest TP rates are associated with small $\omega_1$s, and again this rate decreases as $\omega_1$ increases. If $\omega_1$ is set to be 0.11, the TP rate is 0%, i.e., the model cannot distinguish the two components of the mixture prior, and thus probe sets showing strong pattern will be incorrectly classified as “Absent”. In summary, Figure 2.5 shows that the performance of the BFM depends on the sample size, and the choice of $\omega_1$ can be used to calibrate the model by relaxing or strengthening the assumption of zero factor loadings for a noise probe set.

The BFM is insensitive to the choice of $\omega_2$, in the range of very small values for
that we are interested in. This result is expected since we work with a relatively large data set (251 microarrays) and this is sufficient to identify variance parameters. Additional graphs similar to Figure 2.5 (b) are presented in Figure 2.6, and they show how the factor model behaves, in terms of true positive rates, for other choices of $\omega_2$.

For a different type of cell, the normalization procedure defined in Section 2.1 provides transformed data with variability of intensities similar to those observed in the study of breast cancer developed here. Therefore, the choice of $\omega_1$ and $\omega_2$ does not depend on the type of cell we are examining.

2.5 Comparison of detection methods, using a real data set

In order to compare the performance of detection methods, a special data set containing information regarding the true presence/absence of some subset of genes is required in a real application. The spike-in study developed by Affymetrix for expression algorithm assessment is an interesting option involving the HG-U133A array (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). The data consist of 3 technical replicates of 14 separate hybridizations of 42 spiked transcripts in microarrays for human genome; therefore, the number of arrays available for analysis is 42. Different concentrations, ranging from 0 pM to 512 pM, are used for the spiked transcripts. Four spikes are bacterial controls, eight spikes are artificial sequences believed to be unique in the human genome, and thirty spikes correspond to cDNA clones isolated from total RNAs of a lymphoblast cell line. In other words, the chip contains sequences of 42 genes known to be absent in the human genome, and the hybridization solution contains transcripts from these special genes at different concentrations. An efficient detection method is supposed to identify those 42 spiked genes as “Present”, this is the point being evaluated in this section. Further details about this data set can be found in the Affymetrix website previously indicated. A similar data set designed for the same purpose has been
used by Irizarry et al. (2003a) to evaluate the effectiveness of expression measures produced by MAS 5.0 and RMA.

In the analysis of real data, two packages of functions written in the open source statistical language R are used. The detection method MAS 5.0 P/A is implemented in the package “affy” (Gautier et al., 2004), and PANP can be applied via “panp”. Both packages are integrated into the Bioconductor project (http://www.bioconductor.org), a collaborative effort providing softwares for computational biology and bioinformatics (Gentleman et al., 2004).

Recall that MAS 5.0 P/A and PANP require the specification of two thresholds defining regions for detection calls (present, marginal, absent). The choice of such thresholds is a crucial aspect to be considered in a study comparing three different detection methods. The number of presence/absence calls varies depending on the chosen values. As an example, assume the default thresholds (0.04 and 0.06 in MAS 5.0 P/A, 0.01 and 0.02 in PANP) to analyze the set of 42 microarrays included in the spike-in study. These techniques generate detection calls for each array, and this information is summarized in a single list of calls by selecting the most frequent classification for each probe set. MAS 5.0 P/A and PANP detect 46.95% and 32.03% of the probe sets as present, respectively. The BFM applied to the same data set identifies 0.77% of the probe sets as present. Different percentages of presence calls suggest different false positive rates for each method.

In this real application, assume again the prior specifications and MCMC configuration indicated in Section 2.3. The BFM detects 172 genes out of 22300 as “Present”. Taking the previous discussion into account, we define thresholds for PANP and MAS 5.0 P/A such that the number of presence calls is close to the result from the BFM. It is not possible to select thresholds providing exactly 172 presence calls for PANP and MAS 5.0 P/A, because some p-values are the same (precision of 4 decimal places) which implies the same classification for a group of probe sets.
Therefore, the threshold providing the smallest number of presence calls larger than 172 is chosen for the analysis (0.00180 for MAS 5.0 P/A, and 0.00015 for PANP). The first threshold is the most important, because in the study of the spike-in data interest lies in the presence calls. The second threshold is not considered here, and any probe set with p-value larger than the first threshold will be classified as “not present” combining marginal and absence calls.

**Figure 2.7:** Detection calls obtained via BFM, PANP and MAS 5.0 P/A for 42 spiked-in genes. Presence calls are displayed in gray and non-presence calls in white. We note that all 42 spiked-in genes are detected as present using BFM, which is a significant improvement over PANP or MAS 5.0 P/A.

**Figure 2.8:** Intensities of two probe sets across microarrays in the spike-in study. Image (a) is an example of a probe set showing a consistent and strong pattern across samples, with detection call “Present” obtained by the BFM, and detection call “Absent” obtained by both other methods. Image (b) is an example of a probe set detected as “Absent” by the BFM.
Table 2.1: Detection calls of probe sets containing probes matching spiked sequences.

<table>
<thead>
<tr>
<th>Spiked probe set</th>
<th>Matching probe set</th>
<th>Number of matching probes</th>
<th>Detection calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>212827_at</td>
<td>209374_s_at</td>
<td>11/11</td>
<td>P A A</td>
</tr>
<tr>
<td>205398_at</td>
<td>205397_x_at</td>
<td>5/11</td>
<td>P A A</td>
</tr>
<tr>
<td>206060_s_at</td>
<td>208010_s_at</td>
<td>9/11</td>
<td>P A A</td>
</tr>
</tbody>
</table>

Figure 2.7 compares detection calls for the 42 spiked-in genes. As can be seen, the BFM correctly identifies all 42 cases as “Present”, whereas, PANP and MAS 5.0 P/A indicate 16 and 14 presence calls, respectively. This result based on real data reinforces the conclusion of the simulated study in the previous section, where the BFM outperforms the other two methods. Figure 2.8 displays images of two probe sets across samples in the spike-in data. Panel (a) shows a consistent and strong pattern for each probe across samples. The corresponding probe set belongs to the group of 42 spiked-in cases, and it was detected as “Present” by the BFM and as “Absent” by the other techniques. Panel (b) presents a probe set whose probes randomly alternate intensities across samples. This is a typical “Absent” case correctly identified by the BFM.

Three probe sets in the array are known to contain probes that exactly match some spiked clone sequences. As a result, these probe sets are potentially affected by cross-hybridization, and thus their detection calls should be “Present”. Table 2.1 indicates the number of matching probes within each case and also presents the detection calls obtained by the three methods. Note that the BFM classifies all cases as “Present” while the other techniques indicate “Absence”. This result shows once more that the BFM can be more efficient than PANP and MAS 5.0 P/A for detecting transcripts.
2.6 Conclusions

In this chapter, we proposed a Bayesian factor model for detection of gene expression in oligonucleotide microarrays. The BFM evaluates whether the probes within a probe set exhibit a consistent pattern of intensities across samples. First, we fitted the BFM to synthetic data sets and evaluated the posterior estimates of the involved parameters. The BFM can estimate very well the chosen real values which indicates a good performance in terms of inference. Next, a second simulated study was developed to compare the BFM and other two detection techniques suggested in the literature. The main conclusion was obtained from ROC curves comparing the methods in terms of true positive and false positive rates. The BFM provides the best combination of high true positive and low false positive rates. However, the performance of the proposed factor model depends on the number of microarrays available for analysis. The smaller the sample size, the lower the TP rate. The study suggests that the BFM is preferred to PANP and MAS 5.0 P/A, particularly if a large number of samples is available. A real data set related to a spike-in study was also used to compare the methods. In the designed experiment, transcripts from 42 spiked genes were known to be present in the arrays. The BFM correctly detected the presence of all spikes, whereas, PANP and MAS 5.0 P/A indicated absence call for some cases.

Extreme probabilities (near 0 or 1) are observed for \( q^* \) in (2.3) defining the two detection calls in the BFM. Non-extreme values are observed only if very few samples and/or probes (e.g. 5) are used in the analysis. A straightforward extension to the proposed factor model would be the inclusion of a Beta distribution to express our prior knowledge about \( q \) in (2.2). Define \( h \) as a binary indicator latent variable and
rewrite the expression (2.2) as

$$\alpha \sim (1 - h)N_m(0, \Phi_1) + hN_m(0, \Phi_2) \quad \text{with} \quad (2.5)$$

$$(h|q) \sim \text{Bernoulli}(q) \quad \text{and} \quad q \sim \text{Beta}(\gamma_1, \gamma_2).$$

As a result, the posterior distribution is $(q|h) \sim \text{Beta}(\gamma_1 + h, \gamma_2 - h + 1)$. If we assume the $U(0, 1)$ prior, the posterior will be $\text{Beta}(1, 2)$ or $\text{Beta}(2, 1)$. Note that a single observation of $h$ is used to update $q$ in this model. One can argue that we could define a mixture prior for each loading within $\alpha$ and then update $q$ based on $m$ observations of binary variables $h$, with $m$ being the number of probes. However, this assumption implies that different probes within the same probe set might exhibit distinct patterns across samples, and this behavior is not expected for this type of data. In summary, the inclusion of a Beta distribution for $q$ in (2.2) does not improve much the Presence/Absence detection obtained via $q^*$ in (2.3).

In the next chapter we develop a different application of factor models to analyze probe set expression data, as opposed to the probe level data considered in this chapter. In that case, the assumption of a mixture prior for each component of $\alpha$ is reasonable.
The number of copies of a gene in a chromosome can be modified as a consequence of problems during cell division. These alterations play an important role in human cancer. Several methods have been developed for identification of such chromosomal abnormalities; see for example Lai et al. (2005), Diskin et al. (2006), Wieringen et al. (2006) and Rueda and Uriarte (2007). However, none of these methods compare data sets representing different tumors to investigate whether a gene list, known to be representative of copy number variation in one cancer type, is also representative of copy number alteration in other types of tumors. This chapter illustrates a different application of factor models to examine the gene expression pattern across samples.

Consider seven data sets evaluated in Chin et al. (2006), Miller et al. (2005), Sotiriou et al. (2006), Wang et al. (2005), Bild et al. (2006), Marks et al. (1991) and Freije et al. (2004). The first four studies involve microarrays for breast cancer, and the remaining references are associated with lung, ovarian and brain cancer, respectively. In addition, assume a collection of genes known to be coherently expressed.
These co-expressed genes are located altogether in the chromosome, and their locations are known. An annotation file identifying the chromosome location of each probe set can be obtained from the Affymetrix webpage, and it can be used to determine the gene lists investigated in this section. We generate the lists by collecting all genes within a fixed range, defined around some central point representing the location of the group.

The selected genes have been shown to be overexpressed in certain breast cancers due to duplications of their DNA segment; see Pollack et al. (2002) and Lucas et al. (2010). Extras copies of the DNA leads directly to a higher concentration of mRNA for those genes through a dosing effect, and thus the measurements in the microarrays are affected. The question driving the present study is: does the same chromosomal duplication occur in other types of cancer cells? In other words, given a chromosomal region that is known to exhibit copy number abnormalities (CNA) in breast cancer, can we measure that abnormality in breast cancer gene expression and can we assess whether that same region exhibits CNA in other tumor tissue types? Our study focuses on the effects of CNA over the gene expression pattern across samples. The factor model can be used to statistically assess the presence/absence of copy number alteration based on expression patterns.

In this chapter, $X$ represents the matrix of data with probe sets in each row, and microarrays in each column. We are no longer looking at the expression of probes. Instead, we have utilized RMA to summarize each probe set into a single expression vector. In the scenario without copy number variation, we expect the expression of probe sets across samples to randomly alternate values due to the activities of pathways that are relevant to each gene individually. On the other hand, in the scenario with chromosomal duplications, the pathway activity is swamped by the impact of CNA, and coherent patterns across all samples can be detected. If a strong and consistent pattern is observed, the gene list should be classified as relevant for
the corresponding tissue.

The outline of this chapter is as follows. The new version of the factor model, dealing with gene-level instead of probe-level data, is presented in Section 3.1. In Section 3.2, we develop a simulated data analysis to verify the performance of the model in terms of inference. Section 3.3 shows the analysis involving seven different real data sets. Section 3.4 presents the conclusions.

3.1 The model

Let \( X_{ij} \) be the RMA output representing the expression of gene \( i \) in sample \( j \). Denote \( X \) as the \((m \times n)\) matrix containing these data. The value \( m \) represents the total number of genes being investigated, and \( n \) is the sample size. Two different rows of matrix \( X \) can exhibit different mean expressions so that the heat map image would indicate one row brighter than the other. We choose to standardize the rows of matrix \( X \) in any analysis involving the factor model proposed in this section. As a result, we can consider a more parsimonious model without a mean expression parameter \( \mu \). The factor analyses with and without \( \mu \) provide very similar results which makes the parsimonious version more attractive. The heat map image of \( X \) with standardized rows shows more visual evidence of patterns across samples.

Consider the factor model: \( X = \alpha \lambda + \epsilon \) where \( \alpha \) is the \( m \)-dimensional column vector of factor loadings, \( \lambda \) is the \( n \)-dimensional row vector of factor scores, and \( \epsilon \) is the idiosyncratic noise. Note that the structure of this factor model is the same as the structure of the model (2.1) in the previous chapter. Only 1 factor is included in the analysis for two reasons. First, the previous results indicate that the one-factor model performs well in the analysis of expression patterns across samples. Second, the vector representation of \( \alpha \) is an attractive strategy to evaluate the overall probability of genes displaying coherent patterns in \( X \).

The model defined in Chapter 1 and the current one differ in terms of the mixture
prior distribution specified for the factor loadings.

\[ \alpha_i \sim (1 - h_i)\delta_0(\alpha_i) + h_iN(0, \omega), \quad (3.1) \]

\[ h_i \sim \text{Bernoulli}(q) \quad \text{and} \quad q \sim \text{Beta}(\gamma_1, \gamma_2). \]

Because each row of \( X \) represents a different gene and some genes may exhibit distinct patterns, the mixture prior is specified for each loading \( i = 1, 2, ..., m \). Note that \( h_i \) is a binary latent variable indicating whether \( \alpha_i \) equals 0 or not. The probability \( q \) measures the overall level of sparsity in the factor loadings, and we express our uncertainty about this parameter through the Beta distribution. The posterior estimate of \( q \) is an interesting measure of coherent pattern in \( X \).

Again, the conjugate priors are specified for the remaining parameters: \( \lambda_j \sim N(0, 1) \) which is a standard choice in factor models to avoid scale problems between factor loadings and scores, \( \epsilon_{ij} \sim N(0, \sigma_i^2) \) and \( \sigma_i^2 \sim IG(a, b) \).

We cannot evaluate analytically the joint posterior distribution \( p(\alpha, \lambda, \sigma^2, h, q|X) \). The Gibbs sampler algorithm is implemented to generate values from this distribution; once again, we recommend Gamerman and Lopes (2006) for details about this MCMC method. The derivation of the likelihood function and the complete conditional posterior distributions are presented in Appendix B\(^1\). The conditional probability that \( \alpha_i \neq 0 \) is an interesting element in the full conditional posterior distribution of \( \alpha_i \).

\[ p(h_i = 1 \mid \omega, \lambda, \sigma^2, q, X) = \frac{q}{q + (1 - q)\frac{N[0,M_\alpha,V_\alpha]}{N[0,0,\omega]}}, \quad (3.2) \]

where \( M_\alpha \) and \( V_\alpha \) are expressions defined in Appendix B. We can use this probability to measure how strong is the expression pattern of each gene \( i \).

---

\(^1\) In Appendix B, we specify a more general model with \( L \) factors and including a mean expression parameter \( \mu \). Let \( \mu = 0 \) and \( L = 1 \) to identify expressions related to the model proposed in Section 3.1.
3.2 Simulated data analysis

The main aim of the study developed in this section is to verify the performance of the model proposed in Section 3.1 using a synthetic data set. In terms of prior specifications, we assume $\omega = 10$, $\gamma_1 = \gamma_2 = 1$ in (3.1). Note that we choose the $U(0,1)$ to express high uncertainty about $\alpha_i \neq 0$. In addition, consider $a = 2.1$ and $b = 1.1$ as the parameters of the Inverse Gamma distribution defined for $\sigma_i^2$, which gives mean = 1, mode 0.3548 and variance = 10. The MCMC algorithm is set to perform 2000 iterations with the first 1000 values assumed as burn-in period. The initial values of the chains are $\alpha_i^{(0)} = 0$, $(\sigma^2_i)^{(0)} = 1$, $q_i^{(0)} = 0.5$, $h_i^{(0)} \sim \text{Bernoulli}(0.5)$ and $\lambda_j^{(0)}$ is generated from its prior distribution. Fast convergence to the target distribution is observed in all applications of the algorithm.

The first step of the analysis is to generate the data. The procedure is simple, we select a real data set, associated with breast cancer (Miller et al., 2005), representing the expressions of 23 genes across 251 samples. The factor model defined in Section 3.1 is fitted to the real data and the posterior estimates of $\alpha$, $\lambda$ and $\sigma^2 = (\sigma^2_1, \ldots, \sigma^2_m)'$ are assumed as the real values to generate the synthetic data. Note that we are choosing the true values of parameters based on the model fit for real data, and thus we do not have to guess which values are more realistic for representing these parameters. In short, our procedure to simulate data is a reconstruction of matrix $X$ with its most important features. We generate $\epsilon_{ij}$ from $N(0, \sigma^2_i)$ and compute $\alpha \lambda + \epsilon$ to obtain the simulated $X$.

Figure 3.1 shows the results obtained from the factor model fitted to the simulated data. Panel (a) displays results for the factor loadings. The conditional posterior distribution of $\alpha_i$ is a mixture distribution of $\delta_0(\alpha_i)$ and a Normal component. The 95% credible intervals indicated in Panel (a) are computed for the component with the highest probability weight in the mixture. For this reason, the graph shows some
\( \alpha_i \) with intervals represented by a point at zero. As can be seen, Panels (a), (b) and (c) suggest a good performance of the proposed model, where in most cases the true value is located very close to the posterior mean and inside the 95\% interval.

![Figure 3.1: Real value (circle), posterior mean (x mark) and the 95\% credible interval (bar) for \( \alpha \) in Panel (a), \( \lambda \) in (b) and \( \sigma^2 \) in (c). Panel (d) shows box plots indicating the distribution of the conditional posterior probability that \( \alpha_i \neq 0 \), and the heat map image of the simulated matrix \( X \) with columns sorted so that the first principal component is monotone.](image)

Panel (d) in Figure 3.1 presents the heat map image of the simulated data and box plots representing the posterior distribution of the probabilities (3.2). Note that the box plots concentrate the probability mass around 1 for the genes located at the bottom of the image, which means that these genes have a high posterior probability that \( \alpha_i \neq 0 \), and it suggests a coherent pattern across samples.
the other hand, the genes located at the top of the image have box plots more
dispersed in the unit interval. In particular, the box plots in rows 4-8 indicate
probability mass concentrated below 0.5, which suggests genes displaying random
patterns. The expression pattern observed in the heat map image seems to agree
with the interpretation obtained from the box plots.

3.3 Real data analysis

In this section, we investigate the CNA effect over the gene expression across samples
of 4 different types of cancer. Again, we work with seven different data sets, four of
them are related to breast cancer and the remaining three correspond to ovarian, lung
and brain cancer. The main idea is that CNA is known to occur in some regions of
chromosomes extracted from breast tumors. We identify the group of genes located
in those affected regions for breast cancer, and then examine the expressions of these
genes for other types of cancer.

The same prior specifications and MCMC configurations described in Section 3.2
are applied in this study, and once again fast convergence to the target distribution
is observed for all chains. The main results are shown in Figures 3.2, 3.3, 3.4 and
3.5.

Figures 3.2 and 3.3 show images of matrices $X$. The columns, representing sam-
ples, have been sorted so that the first principal component is monotone. Each row
of $X$ corresponds to a gene, and we evaluate the same gene list in each panel within
the same figure. The genes analyzed in Figure 3.2 belong to Chromosome 6, and the
genes evaluated in Figure 3.3 are from Chromosome 1. We apply the factor model
described in Section 3.1 to each data set, and generate box plots representing the
distribution of the conditional probability (3.2) for each gene $i$. These graphs are
shown in the middle of each panel. In addition, we perform the factor analysis for
probe-level data using the model (2.1) described in Section 2.2 of the previous chap-
ter. The presence/absence detection calls for each gene are presented in the color bar displayed to the left of some panels; the probe-level data are not available for one breast cancer and the lung cancer data sets.

**Figure 3.2:** Factor analysis of gene expression across samples. Probe-level analysis using the model (2.1) proposed in Section 2.2: Presence (white) and Absence (black) calls are shown to the left of some images. Gene-level analysis using the model proposed in Section 3.1: Box plots representing the conditional probability (3.2) are shown in the middle of each panel. Rows 1 and 2 are associated with breast cancer; row 3 contains graphs related to lung, ovarian and brain cancer, respectively. The Genes are located between positions 108,000,000 and 112,000,000 in Chromosome 6.

Note that very few genes are detect as “absent” in the breast cancer data sets. This result is expected since the genes were selected in a region of the genome where the CNA is known to occur for breast cancer. In the ovarian tumor, most genes are detected as “Present”; however, the number of absence calls increases a bit as
compared to the breast cancer. The brain cancer case indicates distinct results, i.e., a majority of presence calls for the gene list of Chromosome 6 (Figure 3.2) and a majority of absence calls for the gene list of Chromosome 1 (Figure 3.3). Another interesting result is the association between the expression pattern and the gene detection call. In most cases, presence calls seem associated with a row displaying increasing or decreasing patterns in the image, and absence calls seem associated with random patterns.

**Figure 3.3:** Factor analysis of gene expression across samples. Probe-level analysis using the model (2.1) proposed in Section 2.2: Presence (white) and Absence (black) calls are shown to the left of some images. Gene-level analysis using the model proposed in Section 3.1: Box plots representing the conditional probability (3.2) are shown in the middle of each panel. Rows 1 and 2 are associated with breast cancer; row 3 contains graphs related to lung, ovarian and brain cancer, respectively. The Genes are located between positions 208,000,000 and 212,000,000 in Chromosome 1.
The box plots in Figures 3.2 and 3.3 show how the conditional probability (3.2) is distributed in the interval 0-1 for each gene \( i \). As can be seen, the probability mass is concentrated above 0.5 for most genes in almost all panels. The brain cancer panel in Figure 3.3 is the case showing the largest number of box plots not concentrated around 1. Another interesting observation is the association between the pattern exhibited in each row of the image graph and the distribution suggested by the box plot. Note that coherent patterns are related to box plots located above 0.5, and random pattern correspond to dispersed box plots centered below 0.5.

Figure 3.4: Box plots comparing the posterior distribution of \( q \) for all data sets.

Figure 3.4 shows box plots representing the posterior distribution of the probability \( q \) specified in the mixture prior (3.1). As can be seen, the graph for the Brain cancer (genes from Chromosome 1) is the only one suggesting probability mass concentrated below 0.5. This result indicates that the sparsity level in the factor loadings vector is high, and thus the copy number alteration can be considered absent in that case. All other data sets have box plots suggesting: low sparsity level in \( \alpha \), strong coherent pattern for most genes, and presence of copy number alteration.

Figure 3.5 shows the posterior mean and the 95% credible interval for \( \alpha \). The results refer to the gene list selected from Chromosome 1. The panels in the first row correspond to breast cancer, and the second row are related to the brain cancer
data. In addition, we compare the behavior of the factor models defined in Section 2.2 (Chapter 2) and Section 3.1 of the present chapter. As can be seen, the posterior estimates for the breast cancer data are similar for both models; several intervals do not contain the value 0, and the posterior means are located in similar positions.

According to Figure 3.4, the genes from Chromosome 1 are not affected by CNA in brain cancer. The posterior estimates for $\alpha$ reflect this result. Note that the model in column 1 provides intervals centered around zero, and the model in column 2 estimates several $\alpha_i$ as zero. The main difference between the models being investigated is the mixture prior formulation, indicated in (2.2) and (3.1), for the factor loadings. Again, it is important to highlight that we compute the 95% interval for the component of the posterior mixture with highest probability weight.

Figure 3.5: Posterior mean (x mark) and 95% credible interval (bar) for $\alpha$. We work with the gene list identified in Chromosome 1. Breast cancer in row 1, brain cancer in row 2, factor model of Section 2.2 (Chapter 2) in column 1, and factor model of Section 3.1 (Chapter 3) in column 2.

We have applied the model specified in Section 2.2 (Chapter 2) to the data sets $X$
shown in Figures 3.2 and 3.3. The results are in accordance with the interpretation of the box plots in Figure 3.4, i.e., the only case associated with a small probability $q^*$ in (2.3) is the brain cancer with genes from Chromosome 1. A small $q^*$ suggests no CNA effect in that case.

3.4 Conclusions

This chapter shows an application involving factor models for the analysis of probe set expression data instead of probe level data. Duplications affecting the number of transcripts of genes located in a specific region of the chromosome are known to occur in breast cancer tumors. The main idea of this application was to examine copy number changes for the same group of genes in other types of cancer. Images of probe sets across arrays can be inspected for visual identification of patterns suggesting copy number variation; however, we can use the factor analysis as a test providing a more accurate answer than the visual analysis.

The rows of $X$ represent genes, and different genes may exhibit distinct expression patterns; therefore, we have proposed in Section 3.1 a factor model with a mixture prior for each loading $\alpha_i$. We induce sparsity assuming $\delta_0(\alpha_i)$ as one of the components of the mixture, and a Beta distribution is used to express our uncertainty about the probability of $\alpha_i \neq 0$. In the first application, we have studied the performance of the model in terms of inference results, using a simulated data set. We have concluded that the model can estimate well the true values of the parameters. In the second application, we have analyzed real data for different types of cancer. The main conclusions are obtained from box plots representing the posterior distributions of $p(h_i = 1 \mid \omega, \lambda, \sigma^2, q, X)$ and $q$. We have noted an agreement between the expression pattern, displayed in the heat map image, and the position of the box plot for the conditional probability that $h_i = 1$ or $\alpha_i \neq 0$. The probability mass is concentrated above 0.5 for coherent patterns, and below 0.5 for random patterns. The
posterior estimate of $q$ indicates the overall level of sparsity in the factor loadings. We use this result to measure our posterior uncertainty about the presence/absence of CNA. Presence of CNA was detected for almost all data sets, the only exception was the brain cancer with genes from Chromosome 1.
RNA-Seq is a promising new technology to measure gene expression. Its main steps are: (i) RNA’s are isolated from a sample and converted to cDNA fragments, (ii) a high-throughput sequencer is used to generate millions of reads (short nucleotide sequences) from the cDNA fragments, (iii) an alignment tool is used to map the reads to a reference genome, and (iv) counts of reads mapped to each gene are used to estimate expression levels. Because the outputs of RNA-Seq are counts, they are referred to as “digital” gene expression, as opposed to the “analog” fluorescence intensities from microarrays. Although the technology is still young, its output data have been analyzed in several scientific publications (e.g. Marioni et al., 2008; Mortazavi et al., 2008; Wang et al., 2009). RNA-Seq data have some advantages over microarrays, such as low background noise, an ability to detect novel transcripts, and the requirement of less RNA samples. On the other hand, some experimental challenges must be addressed, such as read mapping uncertainty. The reads are much shorter than the transcripts from which they are derived, and there is the possibility that a single read may map to multiple genes, complicating the expression analysis. Different approaches have been proposed in the literature to deal with this problem;
one might discard reads that map to multiple locations Marioni et al. (2008), allocate the reads to genes heuristically Faulkner et al. (2008), assume a statistical model with latent variables representing the true mappings Li et al. (2010).

In this section we apply factor models to analyze the expression pattern across samples of RNA-Seq data, and then determine the presence/absence status of the corresponding gene. This strategy can be an interesting approach for the read mapping uncertainty issue, where a random pattern is potentially observed for a gene with a large number of incorrectly mapped sequences. On the other hand, a large number of nucleic acid sequences targeting the correct gene will contribute to a strong pattern across samples. The expression data is obtained via the high-throughput sequencing system Illumina Genome Analyzer 2. This sequencer generates images of size 1Tb, approximately. The data is then processed and passed through a quality control where bad reads with a chastity score lower than 0.2 are discarded. Next, the remaining reads with a quality score larger than 15 are consolidated into a frequency. Finally, the reads are mapped to the human genome wherever possible. The data set is composed of 32 samples related to Ovarian tumors (source: Harvard Medical School; access: The Cancer Genome Atlas data portal, http://tcga-data.nci.nih.gov/tcga/).

For each sample, the data set contains a list of sequences (17 nucleotides in length), the total number of reads per sequence and the associated gene symbol. Each nucleic acid sequence has a single entry in the list, and more than one sequence may target the same gene. The number of nucleotide sequences and the number of identified genes may vary between samples. The total number of genes in the union of all samples is 44,320; however, only 16,082 cases can be found in the intersection. Our study is focused on those genes in the intersection. For each gene, any row of matrix $X$ represents a DNA sequence targeting that gene in at least one sample. In general, it is possible that, as a consequence of read mapping uncertainty, the
sequence belongs to another gene.

Of the 16,082 genes that we consider, 852 are represented by a single nucleotide sequence (i.e., \( X \) is a \( n \)-dimensional row vector). In addition, 1,634 genes are associated with more than 100 sequences, most of which are identified in only 1 sample. Neither of these situations provide significant information regarding coherent expression. With this in mind, we filter the 16,082 genes by selecting those cases with corresponding matrix \( X \) having more than 20 rows and very few rows full of zeros (criterion: at least 70% of the DNA sequences detected in at least 5 samples). It is important to highlight that the focus of this application is to show how the analysis of expression patterns across samples can be useful to handle the presence/absence detection problem for a subset of genes identified in RNA-Seq data.

This chapter is organized as follows. In Section 4.1, the factor model is defined for count data; we consider the Poisson distribution. In Section 4.2, we investigate the read mapping uncertainty associated with 4 real data sets displaying coherent or random patterns. In Section 4.3, we evaluate the behavior of the model for different choices of a variance parameter defined in the mixture prior assigned for factor loadings. Section 4.4 shows a simulated study, where the goal is to verify the performance of the Poisson linear factor model in terms of inference results. Finally, Section 4.5 presents the conclusions of the chapter.

4.1 Factor model assuming Poisson distribution

Direct applications of standard factor models to the analysis of count data are not very common in the literature. This fact might be explained by the computational difficulties associated with this type model, even though developments in simulation methods for estimation enable such analyses. Wedel et al. (2003) is an example of a paper applying factor models for multivariate count data. The parameters in their Poisson model are estimated via simulated maximum likelihood. A matrix of
weights is specified and its rank defines the number of factors. We are focused on the analysis of RNA-Seq count data, and applications of factor models to this new gene expression technology are relatively rare. Langmead et al. (2010) proposed a generalized linear model to examine RNA-Seq read counts for each gene. Their model can be specified with a Poisson distribution, and it generates a p-value to test for differential expression between groups; each group represents a factor in their analysis. Our model, described in this section, can be considered a new approach combining the Bayesian framework, factor model for count data, and the analysis of expression patterns in sequencing data.

Figure 4.1: Count of reads for each sequence across samples. Target gene: “C6orf162” (chromosome 6) in Panel (a), “TMED4” (chr. 7) in (b), “ARL4C” (chr. 2) in (c), and “PHLDA1” (chr. 12) in (d). The original count data is displayed (columns are not sorted and rows are not standardized).

Suppose $X_{ij}$ is the read count value observed for the nucleic acid sequence $i =$
\{1, ..., m\} in sample \(j = \{1, ..., n\}\). Denote \(X\) as the \((m \times n)\) matrix containing the data associated with a gene; \(m\) is the total number of nucleotide sequences targeting that gene, and \(n\) is the number of samples. In contrast with microarray data, the RNA-Seq data is not significantly affected by background noise; therefore, a normalization procedure is not required to remove such noise effect from \(X\).

Figure 4.1 shows heat maps images representing matrices \(X\) for 4 different genes; the real data is displayed. As can be seen, the magnitude of the count values can be different between rows. Note that the image in Panel (b) indicates rows 1, 6, 20 and 24 containing the largest values in that matrix. Another important feature is the fact that some few rows of \(X\) contain several zeros, i.e., the nucleotide sequence does not appear in many samples. As an example, consider rows 5 and 6 in Panel (a), and row 3 in Panel (c). A factor model proposed for this type of data must take into account these characteristics of matrix \(X\).

Since we are investigating the pattern of count data across samples, it seems reasonable to assume a model based on the Poisson distribution.

\[
X_{ij} \sim \text{Poisson}(\mu_{ij}) \quad \text{with} \quad \log(\mu_{ij}) = \eta_i + \alpha_i \lambda_j. \tag{4.1}
\]

We introduce the parameter \(\eta_i\) to account for differences between rows in terms of magnitude of the count values. Assume the hierarchical prior specification \((\eta_i \mid \theta_i, \sigma^2_i) \sim N(\theta_i, \sigma^2_i), \theta_i \sim N(\theta_0, \phi)\) and \(\sigma^2_i \sim IG(a,b)\). If row \(i\) is associated with large count values, then \(\eta_i\) tend to be large defining a large Poisson rate \(\mu_{ij}\). On the other hand, if row \(i\) has several zero counts, then \(\eta_i\) should be small, possibly negative, such that \(\exp(\eta_i)\) is a small value defining the rate \(\mu_{ij}\).

Let \(\alpha = (\alpha_1, ..., \alpha_m)'\) be the column vector representing the factor loadings. Again, we assume a model with a single factor for two reasons: parsimony and the fact that a single \(\alpha_i\) for each nucleotide sequence allows us to compute an overall probability of rows displaying coherent patterns in \(X\). Consider the mixture prior
formulation for \( \alpha_i \) shown below.

\[
\alpha_i \sim (1 - h_i)N(0, \omega_1) + h_iN(0, \omega_2), \quad (4.2)
\]

\[ h_i \sim \text{Bernoulli}(q) \quad \text{and} \quad q \sim \text{Beta}(\gamma_1, \gamma_2). \]

Here, the “small and large” configuration is set for \( \omega_1 < \omega_2 \). Some nucleotide sequences can be incorrectly mapped to the gene represented by matrix \( X \); as a result, different rows of \( X \) may exhibit distinct patterns. A coherent pattern would be observed for those sequences correctly associated with the gene, whereas, a random pattern would be detected for the sequences targeting the wrong gene. This characteristic of the data is the reason why we specify the mixture prior for each loading \( i = 1, \ldots, m \).

Note that \( h_i \) is a binary indicator of \( \alpha_i \neq 0 \), and the probability \( q \) measures the overall level of sparsity in the factor loadings. The posterior estimate of \( q \) can be used as a measure of coherent pattern in \( X \). Another interesting element of the model is the conditional probability that \( \alpha_i \neq 0 \), i.e., \( h_i = 1 \).

\[
p(h_i = 1 \mid \omega_1, \omega_2, \alpha_i, q) = \frac{N[\alpha_i|0, \omega_2] q}{N[\alpha_i|0, \omega_2] q + N[\alpha_i|0, \omega_1] (1 - q)}. \quad (4.3)
\]

This expression is used to update the indicator \( h_i \) in the MCMC algorithm implemented for the Poisson linear factor model. The full conditional posterior distribution of \( \alpha_i \) cannot be evaluated analytically; therefore, the conditional probability (4.3) is not obtained from the derivation of that distribution. Note that expression (4.3) depends on \( \alpha_i, q \) and the Normal density functions specified \textit{a priori} in (4.2). Further details regarding (4.3) are shown in Appendix C.

Denote \( \lambda = (\lambda_1, \ldots, \lambda_n) \) as the row vector of factor scores. We also assume the standard conjugate prior \( \lambda_j \sim N(0,1) \) in this factor analysis. This specification completes the list of prior distributions defined for the Poisson factor model. The
derivation of the likelihood function and the complete conditional posterior distributions are presented in Appendix C.1.

We use the structure of the Gibbs sampler algorithm to generate values from the joint posterior distribution of the parameters involved in this model. As can be seen in Appendix C, the conjugate analysis allows us to compute the closed form of the full conditionals for $\theta_i$ and $\sigma^2_i$. The Metropolis-Hastings is required, within the Gibbs Sampling implementation, to sample from the full conditionals of $\eta_i$ and $\lambda_j$. Given the mixture prior formulation for $\alpha_i$, the application of the Metropolis-Hastings algorithm to sample from the corresponding full conditional would be problematic. As an alternative, we consider the derivative free version of the Adaptive Rejection Sampling algorithm based on secants.

The Adaptive Rejection Sampling (ARS) algorithm for Gibbs Sampling was introduced in Gilks and Wild (1992). In brief, the algorithm is used to sample from any univariate log-concave probability density $f(x)$ known only up to a constant of integration. The method is called “Adaptive” because it defines functions forming an upper and lower bounds to $f(x)$, which converge to the density $f(x)$ as sampling proceeds. One of the functions defining the bounds to $f(x)$ is known as “envelope” and it can be constructed based on secant lines such that the evaluation of derivatives is not required, which reduces the computational cost. The algorithm implementing this strategy is known as derivative free ARS; see Gilks (1992). Further details regarding the application of ARS in this study are presented in Appendix C.

4.2 Analysis of read mapping uncertainty in real data sets

In this section, we examine the read mapping uncertainty related to 4 real data sets. The Basic Local Alignment Search Tool (BLAST) is an interesting program

---

1 In Appendix C, we specify a more general model with $L$ factors. Let $L = 1$ to identify expressions related to the model proposed in Section 4.1.
that can be used to identify nucleotide sequences; see Altschul et al. (1990) and (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The algorithm verifies whether an input sequence has similarities with other sequences in a public database. In fact, BLAST is a group of programs available for different types of query sequences, and in this application, we are interested in the BLAST search for short DNA sequences (17 nucleotides) associated with each row of matrix $X$. Therefore, we consider the “blastn” program to search a nucleotide database using a nucleotide query. The goal of this analysis is to verify whether the RNA-Seq data and BLAST identify the same gene for the sequences represented in matrix $X$. The BLAST output for a particular sequence is a list of genes containing nucleotide sequences that resemble the input sequence above certain threshold. This is a computationally expensive, but very fidelity technique for mapping short nucleotide sequences to the transcriptome.

Figure 4.2: Count of reads for each sequence across samples. Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d). In order to improve the pattern visualization, the images show $X$ with standardized rows and sorted columns so that the 1st principal component is monotone.
Figure 4.2 shows heat map images representing the matrices $X$. In order to improve the visualization of patterns, these images display $X$ with standardized rows and sorted columns. The same data sets without modifications are presented in Figure 4.1 located in the previous section. As can be seen, the visual inspection of heat map images in Figure 4.2 suggests scenarios of coherent or random patterns.

Consider the following short analysis exploring the model formulation (2.4) in Section 2.4 (Chapter 2). We set the prior distribution $\mu \sim N_m(\hat{\mu}, 100I_m)$ with $\hat{\mu}$ being a $m$-dimensional column vector containing the minimum value of each row of $X$. Further, we assume the same prior specifications defined in Section 2.3 (Chapter 2) for $\alpha$, $\lambda$ and $\sigma^2_i$. The MCMC algorithm is set to perform 2000 iterations (burn-in period $= 1000$). The initial values of parameters are the same as those defined in Section 2.3; we also set $\mu^{(0)} = 1_m'$. We fit this factor model to the four $X$ matrices in Figure 4.1 assuming that the count data in each row follows the Gaussian distribution. This assumption may be reasonable when the sample size is large. The result confirms the visual interpretation of Figure 4.2, i.e., the “Absent” call is obtained for the genes represented in Panels (a) and (b), and the “Present” call is determined for the genes in Panels (c) and (d).

Table 4.1: Number of sequences where the gene identification from the RNA-Seq data is confirmed via BLAST.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Visual pattern</th>
<th>P/A call</th>
<th># of agreements</th>
<th># of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>C6orf162</td>
<td>Random</td>
<td>Absent</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>TMED4</td>
<td>Random</td>
<td>Absent</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>ARL4C</td>
<td>Coherent</td>
<td>Present</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>PHLDA1</td>
<td>Coherent</td>
<td>Present</td>
<td>25</td>
<td>28</td>
</tr>
</tbody>
</table>

We performed the BLAST search for every sequence (row) of the matrices presented in Figure 4.2; the results are reported in Table 4.1. Note that the genes exhibiting coherent patterns have a large number of sequences showing agreement.
between BLAST and the RNA-Seq gene identification. In particular, this agreement result is observed for all sequences related to gene “ARL4C” in Panel (c) and 25 of 28 for Panel (d). On the other hand, the genes displaying random patterns indicate a large number of disagreement cases, i.e., the gene suggested in the data set was not found in the BLAST search. The disagreement between BLAST and the data set gene identification can be considered an indication of incorrect sequence mapping. The analysis of the coherent pattern across samples via factor models can be an interesting strategy for detection of genes containing a relatively large number of sequences with incorrect mapping. We explore this idea in the next two sections.

4.3 Evaluation of the mixture prior for factor loadings

The mixture prior (4.2) involves two Normal components with mean 0 and variances $\omega_1 < \omega_2$. The choice of $\omega_1$ and $\omega_2$ plays an important role in the Poisson model; for example, the conditional probability (4.3) is directly affected by these quantities. As a result, it is important to study the behavior of the model under different choices of these parameters. Our goal in this application is to evaluate how sensible is the model in terms of pattern classification via probabilities $q$ and (4.3). We will use the conclusions of this analysis to choose $\omega_2$.

In terms of prior specifications, we choose $\gamma_1 = \gamma_2 = 1$ in (4.2) defining the $U(0, 1)$ a priori for the probability that any $\alpha_i \neq 0$. In addition, consider a Inverse Gamma prior distribution with parameters $a = 2.1$ and $b = 1.1$ for $\sigma_i^2$. Finally, we set $\theta_0$ as the log of the mean of the $i$-th row of $X$, and $\phi = 10$ as parameters in the prior distribution for $\theta_i$. Note that we borrow information from the data to define $\theta_0$. We could set a fixed value for $\theta_0 \forall i$; however, different rows of $X$ can have distinct magnitude of count values. Our choice of $\theta_0$ seems a more logical option to account for the differences in magnitude.

The MCMC algorithm is set to perform 2000 iterations with burn-in period in-
volving the first 1000 cases. The initial values of the chains are \( \eta_i^{(0)} = 0, \theta_i^{(0)} = 0, \)
\( (\sigma_i^2)^{(0)} = 1, \alpha_i^{(0)} = 0, q^{(0)} = 0.5, h_i^{(0)} \sim \text{Bernoulli}(0.5) \) and \( \lambda_j^{(0)} \) is randomly generated from \( N(0, 1) \). Convergence to the target distribution occurs before the end of the burn-in period for all chains and all model configurations studied in this section. The acceptance rate in the Metropolis-Hastings algorithm for \( \eta_i \) is 64% on average, and it can reach 22% for the chains of \( \lambda_j \). We observed 93.68% as the average percentage of MCMC iterations where the ARS algorithm required a single evaluation step to accept the candidate \( \alpha_i \). This average represents the mean of percentages computed for each \( \alpha_i \). In our tested MCMC run, the ARS performed at most 7 iterations.

![Box plots showing the distribution of the conditional probability (4.3). In each panel, the box plots are grouped according to the choice of \( \omega_2 \). Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d).](image)

**Figure 4.3:** Box plots showing the distribution of the conditional probability (4.3). In each panel, the box plots are grouped according to the choice of \( \omega_2 \). Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d).

The variance in the first Normal component of the mixture (4.2) is supposed to be a small positive number suggesting that the factor loading is close to zero.
We fix $\omega_1 = 0.01$ for this purpose. Different choices of $\omega_2$ will be tested; assume $\omega_2 = \{1, 2, 5, 10, 25, 50, 100\}$. All other prior specifications are fixed.

Figure 4.3 shows a comparison between box plots representing the distribution of the conditional probability (4.3) for different choices of $\omega_2$. Consider again the fact that Panels (c) and (d) are associated with data sets displaying coherent patterns in Figure 4.2; whereas, Panels (a) and (b) are related to random patterns. Denote the conditional probability (4.3) as $q_i^*$ for sequence $i$. As can be seen, Panel (c) suggests that $q_i^*$ converges to 1 for most sequences, and the magnitude of $\omega_2$ does not affect this behavior. In particular, $q_3^*$ converges near 1 when $\omega_2 = 1$, and the uncertainty about this probability seems higher when $\omega_2$ is large. On the other hand, $q_{12}^*$, $q_{19}^*$, $q_{21}^*$ and $q_{22}^*$ converge around 0.5 when $\omega_2 = 1$, and their variability seems to decrease as $\omega_2$ increases. The row 3 of matrix $X$ has 31 zeros and a single sample indicating read count 1; this configuration suggests a low Poisson rate for row 3. $\lambda$ is potentially different across samples to express strong pattern, and thus $\alpha_3$ is probably near zero to satisfy the idea of low Poisson rate. The overall coherent pattern implies that probability $q$ is large, so the model gives more weight to a non-zero $\alpha_i$. If $\omega_2 = 1$, the two components of the mixture are close in terms of small variability, and then the model can identify a small non-zero $\alpha_3$ satisfying the aspect “low Poisson rate of row 3”. Rows 12, 19, 21 and 22 are sequences containing non-zero read counts for every sample. The model indicates that these sequences do not follow the same coherent pattern of the other rows. In this situation, a large $\omega_2$ seems helpful to reduce uncertainty about $q_i^*$ estimated near zero. The interpretations of Panels (a) and (b) are similar. Note that $q_i^*$ does not converge to 1 for most sequences in these data sets. If $\omega_2$ is small, the two components of the mixture are close in terms of variability, and then the uncertainty related to $q_i^*$ is higher because it is difficult to distinguish the Normal components. If $\omega_2$ increases, the uncertainty about $q_i^*$ estimated near zero decreases. In Panel (d), most sequences in $X$ have a large
number of zero counts; therefore, the Poisson rate is probably low, and thus $\alpha_i$ tends to be around zero. The coherent pattern classification is favored when $\omega_2 = 1$. In this case, the Normal components are similar in terms of small variability, and the model determines a small (low Poisson rate) non-zero (coherent pattern) $\alpha_i$. Increasing $\omega_2$, increases the uncertainty about $q_i^*$ for most sequences. If $\omega_2 = 100$, the second component of the mixture suggests that $\alpha_i$ can be distant from zero, and then the model cannot conciliate the two ideas “low Poisson rate” (small $\alpha_i$) and “coherent pattern” (non-zero $\alpha_i$).

Figure 4.4: Box plots showing the posterior distribution of $q$ (probability that $\alpha_i \neq 0$ for any $i$). Each box plot is associated with a different $\omega_2$. Target gene: “C6orf162” in Panel (a), “TMED4” in (b), “ARL4C” in (c), and “PHLDA1” in (d).

Figure 4.4 presents box plots showing the posterior distribution of $q$ in (4.2) for any sequence $i$. The corresponding visual pattern for each panel can be observed in Figure 4.2. As can be seen, Panel (c) indicates that $\omega_2$ does not seem to affect $q$; the
coherent pattern is correctly detected ($q > 0.5$) for all $\omega_2$. In Panel (d), we can see that $q$ decreases as $\omega_2$ increases. The posterior uncertainty is high when $\omega_2 \geq 25$. The classification “coherent pattern”, suggested in Figure 4.2, seems more evident when $\omega_2 \leq 10$. Panels (a) and (b) suggest that $q$ decreases as $\omega_2$ increases. Note that the posterior distribution is centered around 0.5 with a large variability when $\omega_2 = 1$. If $\omega_2 \geq 10$, the 75-th percentile is smaller than 0.5, indicating high probability mass below 0.5. The classification “random pattern”, observed in Figure 4.2, seems more clear for $\omega_2 \geq 10$. In conclusion, the posterior distribution of $q$ indicates a classification coherent/random in accordance with the visual interpretation when: $\omega_2 \geq 10$ (random pattern) and $\omega_2 \leq 10$ (coherent pattern). Therefore, $\omega_2 = 10$ seems to be an appropriate specification when fitting the model for a matrix $X$ with unknown classification.

4.4 Simulated study to verify inference results

The goal of this section is to evaluate the performance of the Poisson linear factor model using simulated data. Assume the same prior specifications indicated at the beginning of the previous section, and include $\omega_2 = 10$. The MCMC configuration is also the same, which involves the number of iterations (2000), the burn-in period (1000) and the initial values of the chains. Once again, convergence to the target distribution is observed for all chains before the end of the burn-in period.

In order to generate the data, we first fit the Poisson factor model (Section 4.1) to the real data; consider the data sets represented in Figure 4.1. The posterior estimates of $\eta = (\eta_1, ..., \eta_m)'$, $\alpha$ and $\lambda$ are assumed as the true values, and we use them to compute the $(m \times n)$ matrix of Poisson rates $\mu = \exp\{\eta_1 + \alpha \lambda\}$. Next, the observation $X_{ij}$ is randomly generated from Poisson($\mu_{ij}$), with $i = 1, ..., m$ and $j = 1, ..., n$. In brief, our procedure to simulate data is a reconstruction of matrix $X$ based on important features captured by the factor model applied to the real data.
Figure 4.5: Simulated data representing count of reads across samples. The columns are not sorted and the rows are not standardized in each matrix $X$.

Figure 4.5 displays the images of the four simulated data sets. As can be seen, the Poisson linear factor model can capture relatively well the characteristics of the real data presented in Figure 4.1. Note that the brightest rows in each panel are the same in both figures, which suggests that the model correctly identifies the differences between rows in terms of magnitude of count values. Panel (a) seems to show the best approximation among the four cases.

The RNA-Seq data can have an erratic behavior due the presence of outliers. The count value of some samples can be much bigger or smaller than the overall magnitude of values observed in the row. In order to allow a more detailed comparison between images in Figures 4.1 and 4.5, we set the same maximum value in their color maps; as a result, we avoid dark images in Figure 4.1. We truncate the color map, but we
analyze the data without this modification. The count of reads can reach 40, 348, 1243 and 1396 in the real data represented in Panels (a), (b), (c) and (d), respectively. Panel (a) represents a data set without outliers. In this case, the maximum count is not bigger than 40, and we can see that the synthetic X recreates well the behavior of the count values across samples. In particular, consider the row 2 containing values near 0 or 40 depending on the sample. The Poisson linear model does not deal very well with the outliers in the real data of Panels (b), (c) and (d). Figure 4.5 indicates that the brightest rows behave as if the Poisson rate is the same across all samples. In the real data, the count values of a bright row can vary a lot; for example, the mean of row 20 (Panel (b), Figure 4.1) is 96.62, but we observe 3 and 303 for samples 10 and 12, respectively. The fact that we assume a model with \( \eta_i \forall j = 1, ..., n \) may explain this issue. The estimate of \( \eta_i \) captures the overall behavior of the data in row \( i \), which means that we are ignoring the presence of any outlier. We plan to address this problem in a future work; Chapter 7 describes a possible solution. In the rest of this chapter, we continue to use the model proposed in Section 4.1 relying on a data set without outliers as seen in Panel (a).

Figure 4.6 presents the images of the simulated data sets with standardized rows and sorted columns. The expression pattern is more evident in these graphs, and we can see that the coherent/random behavior is recreated well in the synthetic data. The Ovarian cancer data set studied in this chapter is composed by 32 samples, and we may generate only zeros to represent read counts in a row associated with a low Poisson rate. This event is explicit in rows 5 and 6 of Panel (a) for example. The box plots showing the distribution of the conditional probability (4.3) are presented on the left of each image. Note that most graphs concentrate the probability mass above 0.5 in Panels (c) and (d) related to coherent patterns; whereas, most graphs are dispersed or concentrate the probability mass below 0.5 in Panels (a) and (b) related to random patterns.
Figure 4.6: Box plots showing the distribution of the conditional probability (4.3), and heat map images of the simulated data sets. In order to improve the expression pattern visualization, the images show $X$ with standardized rows and sorted columns so that the first principal component is monotone.

Figure 4.7 shows the 95% credible interval, the posterior mean and the true value for $\eta$, $\alpha$ and $\lambda$. Each row of Panels is associated with one of the four data sets studied in Figure 4.6. As can be seen, if the $i$-th row of $X$ has too many zeros, the uncertainty about $\eta_i$ is higher. In this case, the Poisson rate is low with $\exp\{\eta_i\}$ being probably small, which implies that $\eta_i$ tends to be a negative number. A wide range of negative values, representing $\eta_i$, can define a low Poisson rate associated with a row $i$ full of zero counts. On the other hand, we observe low posterior uncertainty in the intervals for $\eta_i$ related to rows of $X$ containing several non-zero counts. In this case, the intervals tend to be centered around a positive number. Note that most intervals for $\alpha$ are centered around zero in the graphs related to images displaying
random patterns (Panels in column 2, rows 1 and 2). In contrast, the posterior estimates for factor loadings tend to differ from 0 for most sequences, in a matrix $X$ displaying coherent pattern. The comparison between posterior estimates and real values indicates an overall good performance of the model, i.e., several cases show the true value inside the 95% interval and pretty close to the posterior mean.

Figure 4.7: Real value (circle), posterior mean (x mark) and 95% credible interval (bar) for $\eta_i$, $\alpha_i$ and $\lambda_j$. Panels (a), (b), (c) and (d) in Figure 4.6 are associated with the graphs in rows 1, 2, 3 and 4 of this figure, respectively.
4.5 Conclusions

In this chapter, we have developed a factor analysis for high-throughput sequencing data. The coherent pattern across samples is an important feature which we can use to evaluate the frequency of incorrect sequence mapping for the target gene. We have applied the BLAST search to each sequence targeting gene “G” in the RNA-Seq data. As a result, we noted that if the matrix $X$, for gene “G”, displays a coherent pattern, then very few BLAST outputs do not indicate “G” as a potential target for the tested sequences. On the other hand, if gene “G” has a matrix $X$ showing a random pattern, then several BLAST outputs do not indicate “G” as a target. Only four genes were analyzed in Section 4.2, but we have obtained similar conclusions from other cases not described in this study.

A linear factor model based on the Poisson distribution was proposed to analyze the count data across samples. A mixture prior with two Normal components was specified for the factor loadings, and the classification coherent/random pattern was obtained from the probability weight in the posterior mixture. The model was implemented using the Gibbs Sampling structure, but the Metropolis-Hastings was required to sample from the full conditionals of $\eta_i$ and $\lambda_j$, while the Adaptive Rejection Sampling was necessary for the factor loadings. In order to set appropriate choices of variance parameters in the mixture prior, we have evaluated the model fit assuming different values for the largest variance. This analysis have resulted in the configuration $\omega_1 = 0.01$ and $\omega_2 = 10$. In the last section, we have simulated four data sets and studied the Poisson linear factor model in terms of inference results. In summary, the model performs relatively well providing posterior estimates close to the true values, and the classification coherent/random pattern is in accordance with the visual interpretation of the expression across samples.
Interactions between factors

In the previous sections, we have considered applications of one-factor models to analyze the gene expression patterns across samples. Sparse latent multi-factor models have been applied in many exploratory and predictive problems with high-dimensional multivariate observations; consider for example West (2003), Lucas et al. (2006), Carvalho et al. (2008) and Lucas et al. (2009). None of these publications work with a factor model allowing for interactions between factors. As it stands now, the involvement of any gene with the factors is always additive, but this idea may not be necessarily true in some situations. As an example, suppose that Copy Number Alteration (CNA) occurs in two chromosomal regions, and a gene may be affected by the modification in one region or both regions. We can define one factor for each region, and it might be the case that an interaction between these factors have an effect on the observed over-expression. Biological pathways, establishing dependences between a group of genes, may also be a reason to support the idea of a factor model with interaction effects.

In statistics, interactions between variables are most commonly studied in the context of regression analysis. The general analytic strategy of constructing multiple
regression equations to test interactions effects were first proposed by Cohen (1968) in the social sciences. According to this strategy, any combination of categorical and continuous variables can be analyzed within a multiple regression framework simply through the appropriate dummy coding of the categorical variables. Interactions can be represented as product terms, and curvilinear relationships can be represented through higher order terms in the regression equation. The model is designed to describe a situation in which the simultaneous influence of two covariates on the response variable is not additive. Several multiplicative interaction variables can be constructed in a model with more than two covariates; pairwise products represent pairwise interactions and higher order products represent higher order interactions. The presence of interactions can have important implications for the interpretation of the regression model. Interactions are also important elements in the analysis of data arising from experiments with factorial design. In this case, the levels of the predictor variables are discrete rather than continuous. The experimental units take on all possible combinations of these levels across all factors/predictors. A factorial experiment can be analyzed using ANOVA models, which allows studying the effect of each factor on the response variable, as well as the effects of interactions between factors. These models can also be equally well represented using the multiple regression approach.

Arminger and Muthen (1998) consider latent variable models including quadratic forms or higher order polynomial terms, and interactions of latent regressor variables. Two groups of observed variables are used in their analysis: $y$ is the response vector and $x$ is a vector of covariates. Their model specifies two equations; the first one expresses $y$ as a linear combination of polynomial terms and/or interactions of elements in the latent vector $\lambda$. The second equation defines a factor model without interaction terms, where $\lambda$ is the factor score and $x$ is the target data. Because the model includes components representing functions of latent variables in the first
equation, the authors denote the formulation as non-linear. They use the Bayesian framework with conjugate priors to estimate the parameters. Sparsity priors are not considered in their analysis.

We are interested in the study of multi-factor models developed for the analysis of matrices representing gene expression patterns across samples. Our goal is to investigate the existence of multiplicative interactions between pairs of factors defined in the model. In order to test the significance of interaction terms, the spike and slab mixture prior is assumed for the factor loadings to allow for sparsity. In this section, we continue to explore factor models with a linear structure defining the association between latent factors and the expression data; non-linear effects are investigated in the next chapter.

The outline of this chapter is as follows. Section 5.1 presents the factor model to study the expression patterns. Two approaches are defined to introduce the multiplicative interaction terms. In Section 5.2 we explore synthetic data to evaluate the performance of the proposed model under different scenarios. Section 5.3 shows the results of applications involving real data. In Section 5.4 we summarize the main conclusions of Chapter 5.

5.1 Factor model with multiplicative interactions

Reconsider the gene expression data originated from oligonucleotide microarrays. Assume $X$ as the $(m \times n)$ matrix with $X_{ij}$ representing the RMA output for gene $i$ in the sample $j$. We propose the model:

$$X = \mu 1_n + \alpha \lambda + \epsilon,$$  \hspace{1cm} (5.1)

where $\mu$ is the mean expression represented by an $m$-dimensional column vector, $1_n$ is an $n$-dimensional row vector of ones, $\alpha$ is the $(m \times L)$ matrix of factor loadings, $\lambda$ is the $(L \times n)$ matrix of factor scores, and $\epsilon$ is the $(m \times n)$ noise matrix with
\( \epsilon_{ij} \sim N(0, \sigma^2_i) \).

Note that we include the mean expression parameter \( \mu \) in the model; therefore, we can work with matrix \( X \) without standardizing its rows. One could prefer to standardize the rows of \( X \) and remove \( \mu \) to define a more parsimonious model. We do not explore the second option, but it is an interesting alternative worth mentioning.

The interaction terms are defined in \( \lambda \). The matrix of factor scores is divided in two parts; assume \( 2 \leq n_f < L \). The first \( n_f \) rows represent the target latent factors, and the last \( L - n_f \) rows are associated with the product of two target factors. Consider the following configuration: \( \lambda_{(n_f+1)j}, \lambda_{(n_f+2)j}, \ldots, \lambda_{Lj} \) are related to the products \( \lambda_{1j}\lambda_{2j}, \lambda_{1j}\lambda_{3j}, \ldots, \lambda_{(n_f-1)j}\lambda_{nj} \), respectively. Note that \( L = n_f + \frac{n_f!}{(n_f-2)!2!} \).

In terms of prior distributions, we assume the conjugate specifications \( \sigma^2_i \sim IG(a,b) \) and \( \mu \sim N_m(\mu_0, \Sigma) \). The spike and slab mixture prior is defined for the factor loadings to allow for sparsity and to test whether the factors/interactions have significant effect. Consider:

\[
\alpha_{il} \sim (1 - h_{il})\delta_0(\alpha_{il}) + h_{il}N(0, \omega), \\
h_{il} \sim \text{Bernoulli}(q_{il}) \text{ and } q_{il} \sim \text{Beta}(\gamma_1, \gamma_2).
\]

The typical conjugate prior \( \lambda_{lj} \sim N(0, 1) \) is defined for \( l = 1, \ldots, n_f \). For the other factors, \( l = n_f + 1, \ldots, L \), we consider two approaches to introduce the corresponding multiplicative interaction term; they are enumerated below.

1. Introduce the interaction via Gaussian prior: \( \lambda_{lj} \sim N(\lambda_{l_1j}\lambda_{l_2j}, \nu) \).

2. Assume the product with probability 1: \( \lambda_{lj} = \lambda_{l_1j}\lambda_{l_2j} \).

In the cases above, let \( l_1 < l_2 \in \{1, \ldots, n_f\} \) be the indices of factors involved in the product term related to \( \lambda_{lj} \) where \( l \in \{n_f + 1, \ldots, L\} \).

In the version number 1, we specify the product \( \lambda_{l_1j}\lambda_{l_2j} \) as the mean parameter of the Gaussian distribution. This approach can be generalized with the specification
of any function \( f(\lambda_{l_1j}, \lambda_{l_2j}) \), which makes it possible to investigate other types of relationships between factors. The variance \( \nu \) is another important element to be considered in approach 1. This parameter must have a small value; otherwise, we are indicating a weak association between \( \lambda_{lj} \) and \( \lambda_{l_1j}\lambda_{l_2j} \); the multiplicative effect is lost and the interaction factor is just another factor in the model. If the number of genes is large, the variability in the posterior distribution can be very small due to the large amount of data. In this case, \( \nu \) is difficult to set and only extremely small values will ensure that \( \lambda_{lj} \) is associated with \( \lambda_{l_1j}\lambda_{l_2j} \). The target posterior in approach 1 is \( p(\mu, \alpha, \lambda, \sigma^2 | X) \).

In the approach number 2, we force the perfect association between the interaction factor and the corresponding product term; this strategy is convenient to deal with large data sets. Here, the target posterior distribution is \( p(\mu, \alpha, \{\lambda_{1j}, ..., \lambda_{(n_f)j}\}, \sigma^2 | X) \). Note that \( \lambda_{lj} \), for \( l = n_f + 1, ..., L \), is regarded as fixed variables; \( \lambda_{lj} \) is the product of the observed values \( \lambda_{l_1j} \) and \( \lambda_{l_2j} \).

We cannot evaluate analytically the joint posterior distribution of the parameters involved in the proposed models. The Gibbs sampler algorithm is applied to generate observations from this target distribution. Appendix D presents the derivations of the associated complete conditional distributions.

Identifiability is an important issue to be addressed in factor models. If no constraint is imposed, the loadings and the corresponding factor scores can switch their signs. Further, when two or more factors are involved, the rows of \( \lambda \) and the columns of \( \alpha \) can change positions accordingly. In many applications of factor analysis, \( \alpha \) is constrained so that it is a block lower triangular matrix with strictly positive diagonal elements. This strategy is used, for example, in Aguilar and West (2000) and Lopes and West (2004), and provides the identification of the model. Our study is focused on applications involving groups of genes associated with different factors. We can take advantage of such information to constrain \( \alpha \) in the model (5.1). The
next section describes in details our strategy to deal with the identifiability problem.

5.2 Simulated data analysis

The goal of the analysis developed in this section is to evaluate the performance of the factor model (5.1) using synthetic data representing different scenarios. With respect to the prior specifications, we assume $\omega = 10$ in (5.2), $a = 2.1$ and $b = 1.1$ defining the Inverse Gamma for $\sigma_i^2$, and $\mu_0 = 0$ and $\Sigma = 10I_m$ as the mean and covariance matrix of the Multivariate Normal prior for $\mu$. In addition, we set $\nu = 0.01$ in the applications of the model with Gaussian prior introducing the multiplicative interaction.

Suppose that a group of genes is known to be associated with some activity or event in an organism; for example, genes with copy number alteration located altogether in a specific region of a chromosome. Now suppose that 2 or more groups are available for analysis; each group is associated with a different activity/event. In our factor model, we will use this information to specify the Beta prior distribution for $q_{il}$ and then address the identifiability issue.

Let $G_1$ and $G_2$ be two disjoint sets of genes known to be associated with the events 1 and 2, respectively. Suppose that the genes in $G_1$ are not related to event 2, and the genes in $G_2$ are not related to event 1. Given this information we can chose the following configuration of priors for $q_{il}$:

- Beta(2, 1) for $(i \in G_1, l = 1)$ or $(i \in G_2, l = 2)$;
- Beta(1, 2) for $(i \in G_1, l = 2)$ or $(i \in G_2, l = 1)$;
- Beta(1, 1) for $i \in (G_1 \cup G_2)^C$ and $l \in \{1, 2\}$.

$(G_1 \cup G_2)^C$ is the set of genes represented in matrix $X$ which are not members of $G_1$ or $G_2$. According to the model presented in Section 5.1, the matrix $\alpha$ has a third column containing the loadings for the interaction terms. It seems reasonable to assume
that few genes in matrix \( X \) are affected by interactions between the two factors; therefore, we specify \( q_{il} \sim \text{Beta}(1, 2) \) for \( l = 3 \) and \( \forall i \). Note that the probability mass is concentrated around 0 for the Beta(1, 2), around 1 for the Beta(2, 1) and uniformly distributed between 0 and 1 for the Beta(1, 1). As will be seen later, these choices work very well for the small data sets used in this section. When a large number of genes is represented in \( X \), stronger priors should be considered to induce the identification of the model. This is the case for applications involving real data, which we explore in the next section.

The example above describes a situation where a two-factor model is applied. We can easily extend that notation to accommodate more factors. As an example, a three-factor model would include the set \( G_3 \) containing genes related to factor 3. In relation to the prior distribution for \( q_{il} \), we have the configuration:

- Beta(2, 1) for \( (i \in G_1, l = 1) \) or \( (i \in G_2, l = 2) \) or \( (i \in G_3, l = 3) \);
- Beta(1, 2) for \( (i \in G_1, l \in \{2, 3\}) \) or \( (i \in G_2, l \in \{1, 3\}) \) or \( (i \in G_3, l \in \{1, 2\}) \);
- Beta(1, 1) for \( i \in (G_1 \cup G_2 \cup G_3)^C \) and \( l \in \{1, 2, 3\} \).

In this case, the matrix \( \alpha \) has 6 columns. The priors above are related to the loadings in the first three columns. The remaining columns correspond to interaction terms, and we set the Beta(1,2) to induce sparsity. Note that, we do not need to consider the different groups of genes and choose different prior specifications for the loadings in columns 4, 5 and 6. If the first three factors are correctly identified, the position of the other three is also correct, because each one is related to a specific multiplication.

The sign switching between \( \alpha_l \) and \( \lambda_l \) does not affect the inference involving \( q_{il} \) or the conditional probability that \( h_{il} = 1 \). If the posterior estimate of \( \alpha_l \) and its real value have opposite signs, we simply multiply the estimates of \( \alpha_l \) and \( \lambda_l \) by -1 to determine the correct values.

The MCMC algorithm is set to perform 2000 iterations with the first 1000 cases representing the burn-in period. As initial values for the chains, we set \( \mu_i^{(0)} = 0 \),
\( \alpha_{il}^{(0)} = 0, (\sigma_i^2)^{(0)} = 1, \lambda_{ij}^{(0)} \sim N(0,1) \) for \( l = 1, \ldots, n_f \) and \( \lambda_{ij}^{(0)} = \lambda_{il}^{(0)} \lambda_{lj}^{(0)} \) for \( l > n_f \) and \( l_1 < l_2 \in \{1, \ldots, n_f\} \). We also assume a specific configuration of initial values for \( q_{il} \) to induce the identification of the model. Without loss of generality, let \( n_f = 2 \) and consider:

- 0.9999 for \((i \in G_1, l = 1)\) or \((i \in G_2, l = 2)\);
- 0.0001 for \((i \in G_1, l = 2)\) or \((i \in G_2, l = 1)\);
- 0.5 for \(i \in (G_1 \cup G_2)^C\) and \(l \in \{1, 2\}\);
- 0.05 for any \(i\) and \(l = 3\).

The initial value of the indicator variable \( h_{il} \) is generated from \( \text{Bernoulli}(q_{il}^{(0)}) \). Fast convergence to the target distribution is observed in all applications of the MCMC algorithm.

Assume a scenario with \( n_f = 2 \) factors, and let \( G_E = (G_1 \cup G_2)^C \). In order to generate the synthetic data, consider the steps enumerated below. Once again, the notation can be easily modified for cases where \( n_f > 2 \).

1. Set \( m = 30 \) subjects, \( n = 100 \) samples and \( n_f = 2 \) factors;
2. Assume \( G_1 \) and \( G_2 \) with 5 genes each (\( G_E \) has 20 genes);
3. Generate the mean expression \( \mu_i \) from \( N(5, 0.5) \);
4. Set \( q_{il} = 1 \) for \((i \in G_1, l = 1)\) and \((i \in G_2, l = 2)\),
   \[ q_{il} = 0 \] for \((i \in G_1, l \neq 1)\) and \((i \in G_2, l \neq 2)\),
   \[ q_{il} = 0.5 \] for \((i \in G_E, l = \{1, 2\})\) and 0.25 for \((i \in G_E, l = 3)\);
5. Generate \( h_{il} \) and \( \alpha_{il} \) using the prior (5.2) with \( \omega = 1 \) and the values in Step 4;
6. For \( i \in G_l \), set \( \alpha_{il} \) with the same sign. The subjects in \( G_l \) are expected to show similar correlation between \( X_i \) and \( \lambda_i \);
7. Generate \( \lambda_{ij} \sim N(0,1) \) for \( l \in \{1, 2\} \), and then compute \( \lambda_{3j} = \lambda_{1j} \lambda_{2j} \);
8. Generate \( \sigma_i^2 \) from an \( IG \) distribution with high probability mass in 0-1;
9. Simulate \( \epsilon_{ij} \sim N(0, \sigma_i^2) \), and compute \( X \) using expression (5.1).
Figure 5.1 shows the matrix $X$ generated in a scenario with $n_f = 2$ factors. The top panel represents the data without modifications, and the groups $G_1$ and $G_2$ are located in the first 10 rows. Figure 5.2 indicates the 95% credible intervals calculated for the component of the posterior mixture distribution with highest probability weight. Dashed lines are included to separate the loadings in terms of factors. Panel (a) explores the version 1 of the factor model, introducing the product effect via Gaussian prior. Panel (b) is related to the second version, where the perfect association ($\lambda_{3j} = \lambda_{1j}\lambda_{2j}$) is defined. Note that most intervals include the real value and both models seem to approximate well the true loadings. As indicated in Step 6 of the simulation procedure, the first 5 loadings ($G_1$) related to factor 1 are all positive, whereas, the loadings 6-10 ($G_2$) related to factor 2 are all negative. The multiplicative interactions are located in the third box of each graph; most of them are zero and the model correctly identifies the non-zero cases.

**Figure 5.1:** Synthetic data with multiplicative interaction effects ($n_f = 2$). The original data is displayed in the top. The second panel shows $X$ with standardized rows (rows and columns are sorted so that the first principal component is monotone).
Figure 5.2: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{it}$. Dashed lines separate the factors. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

Figure 5.3: Images displaying the posterior mean and the true loading associated with the interaction factor. Box plots representing the distribution of the conditional probability that $h_{i3} = 1$. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

Figure 5.3 shows images comparing the true value and the posterior mean of the loadings in $\alpha_{i3}$, which are related to the interaction term. Once again, we can see the good approximation obtained from both versions of the factor model. The box plots represent the distribution of the conditional probability that $h_{i3} = 1$ or $\alpha_{i3} \neq 0$. As can be seen, most real values are 0 and their corresponding box plots are
located below 0.5. The non-zero cases are associated with graphs located above 0.5. In particular, consider the gene 15 which has a small loading. The model correctly identifies this non-zero value, but its proximity to zero determines a large variability in the box plot.

\[(a)\]

\[(b)\]

**Figure 5.4**: Scatter plots comparing the posterior estimates with the true values. Consider the two versions of the factor model defined in Section 5.1. Panel \((a)\) shows results for the approach 1 (Gaussian prior), and Panel \((b)\) is related to approach 2 (perfect product).

The scatter plots presented in Figure 5.4 indicate that the model can estimate very well all parameters. These graphs compare the true value versus the posterior mean, and the strong linear behavior of the cloud of points suggests the good approximation.
The same type of analysis was developed for a synthetic data generated in a scenario with $n_f = 3$ factors. The graphs are presented in Section D.2 of Appendix D. In short, those results indicate a good performance of both versions of the factor model. The posterior estimates are close to the real values, the loadings related to interaction terms are well estimated, and no identifiability issue is observed.

5.3 Real data analysis

In this section, we visit again the Copy Number Alteration (CNA) problem discussed in Chapter 3. Results presented in the literature identify collections of genes indicating over-expression due to duplications of their DNA segment; see Pollack et al. (2002) and Lucas et al. (2010). These results are associated with breast cancer data, and each collection of genes is located in a different region of the human chromosomes. The locations suggesting CNA are known, and an annotation file identifying the chromosome position for each probe set can be obtained from the Affymetrix website. In order to define which genes are members of the over-expressed group, we consider a range (2,000,000 to the left and right) around the central position where the CNA seems to occur. We explore four different breast cancer data sets in the real analysis developed in this section: Chin et al. (2006), Miller et al. (2005), Sotiriou et al. (2006) and Wang et al. (2005).

We choose to investigate the results for two groups of over-expressed genes. The first one has central position 35,152,961 in the Chromosome 22; we denote this group as $G_1$. The second collection of genes is located around the central point 68,771,985 in the Chromosome 16; let $G_2$ represent this group. We will fit a factor model with $n_f = 2$ latent factors describing the expression pattern, of the genes in $G_1$ and $G_2$, across samples. The model includes a third factor representing the multiplicative interaction between the first two. Our goal is to identify the genes affected by the interaction factor.
The group \( G_1 \) has 50 genes, and \( G_2 \) contains 42 elements. As described above, the selection of these genes is based on an interval specified around a position in the genome. This strategy can lead to the inclusion of cases unrelated to the CNA detected for the studied region. In order to remove the unrelated cases from the current gene lists, we fit a two-factor model (without interaction terms) to the \((92 \times n)\) matrix \( X \). The following configuration is expected for the estimated \( \alpha : \{ \alpha_{i1} : i \in G_1 \} \) with the same sign, \( \{ \alpha_{i2} : i \in G_2 \} \) with the same sign, and \( \alpha_{il} = 0 \) for all other cases. The genes in \((G_1 \cup G_2)\) violating this assumption are considered problematic, and thus removed from the analysis. This cleaning procedure involving \( G_1 \) and \( G_2 \) is described with more details in Appendix E. The procedure defines 22 genes in \( G_1 \) and 18 in \( G_2 \).

Let \( G_E \) represent a group of extra genes to be included in the analysis; \( G_1, G_2 \) and \( G_E \) are disjoint sets. The microarrays selected for this application have 22283 genes, and each breast cancer data set has more than 100 samples available for analysis. As a result, the MCMC algorithm can be rather slow to handle this large amount of data. As an alternative to reduce the computational cost, we implement a gene selection procedure to eliminate the cases which might not be affected by interactions. The full description of the selection process is given in Appendix E. In short, we fit a two-factor model (without interaction terms) to the \((22283 \times n)\) matrix \( X \) assuming 22 genes in \( G_1 \), 18 genes in \( G_2 \) and 22243 genes in \( G_E \). The distribution of the conditional probability \( p(h_{il} = 1|\cdots) \) is evaluated to accept or reject \( \alpha_{il} \neq 0 \). It seems reasonable to assume that the genes affected by both factors are more likely to be affected by interactions; therefore, the final result includes only the cases satisfying this requirement. This selection process yields 3704 genes in the updated \( G_E \).

Consider again the prior specifications: \( \omega = 10 \) in the mixture prior for \( \alpha_{il} \), \( \sigma_i^2 \sim IG(a = 2.1, b = 1.1) \), \( \mu \sim N_m(0, 10I_m) \). This time, our goal is to fit the factor
model with multiplicative interaction effects (using approach 1 = Gaussian prior) to the real data having 22 genes in $G_1$, 18 genes in $G_2$ and 3704 genes in $G_E$. The heat map images for all data sets are shown in Section D.3 of Appendix D. Given the large amount of genes, we need to set strong priors for $q_{il}$ to impose our assumptions related to $G_1$ and $G_2$, and assure the identification of the model. The following configuration is used:

- $p(q_{il} = 1) = 1$ for $(i \in G_1, l = 1)$ or $(i \in G_2, l = 2)$;
- $p(q_{il} = 0) = 1$ for $(i \in G_1, l \in \{2, 3\})$ or $(i \in G_2, l \in \{1, 3\})$;
- Beta$(1, 1)$ for $(i \in G_E, l \in \{1, 2\})$, and Beta$(1, 10)$ for $(i \in G_E, l = 3)$.

Degenerated priors are assumed to impose our assumptions regarding the gene-factor relationship for the cases in $G_1$ and $G_2$. This strategy is important to retain the CNA interpretation of factors 1 and 2; otherwise, the target association can be overwhelmed by the large amount of information in $G_E$. Note that, we assume no interaction affecting the genes in $(G_1 \cup G_2)$. The Beta$(1, 10)$ is specified to induce sparsity in the loadings $(i \in G_E)$ related to the interaction factor. Finally, the $U(0, 1)$ is indicated for all other cases.

The MCMC algorithm performs 600 iterations with the first 400 cases representing the burn-in period. In terms of initial values of the chains, consider the same choices defined in the previous section for $\mu_i^{(0)}$, $\alpha_{il}^{(0)}$, $(\sigma_i^2)^{(0)}$, and $\lambda_{ij}^{(0)}$. The probability $q_{il}$ is initialized with 1 if $(i \in G_1, l = 1)$ or $(i \in G_2, l = 2)$, 0 if $(i \in G_1, l \in \{2, 3\})$ or $(i \in G_2, l \in \{1, 3\})$, 0.01 if $(i \in G_E, l = 3)$ and 0.1 otherwise. The indicator variable $h_{il}^{(0)}$ is generated from Bernoulli$(q_{il}^{(0)})$. The chains seem to converge in all applications of the MCMC algorithm.

The model assuming the prior $\lambda_{3j} \sim N(\lambda_{1j}, \lambda_{2j}, \nu)$ (approach 1) is the focus of the first application in the current section. As previously discussed, the variance parameter $\nu$ must be small to guarantee the target multiplicative effect. The real data
set contains a large number of genes, and thus the posterior variance is expected to be small. In this case, only extremely small values for $\nu$ will ensure that $\lambda_{3j}$ and $\lambda_{1j}\lambda_{2j}$ are correlated. Figure 5.5 shows scatter plots comparing the posterior estimates of $\lambda_{3j}$ and the product $\lambda_{1j}\lambda_{2j}$. Here, the factor model is fitted with $\nu = 10^{-5}$. Note that the model fit for the data set in Panel (c) is the only one indicating correlated results. In the other applications, the multiplicative effect is lost and the interaction factor is just another factor.

Figure 5.5: Scatter plot comparing the posterior estimate of $\lambda_{3j}$ and $\lambda_{1j}\lambda_{2j}$ (Assume approach 1 = Gaussian prior). Each panel represents a different breast cancer data set: Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).
Figure 5.6: Posterior mean (x mark) and the 95% credible interval (bar) for the loadings with $i \in (G_1 \cup G_2)$ (Assume approach 2 = perfect product). Intervals for $\alpha_{il}$ are computed for the component with highest probability weight a posteriori. Dashed lines separate the factors. Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).

Given the difficulty to set $\nu$, no further analysis is developed for the factor model (approach 1) fitted to real data sets. Our next step is to investigate the model defined as approach 2, where we force the perfect association $\lambda_{3j} = \lambda_{1j}\lambda_{2j}$. Consider the same breast cancer data sets, configuration of prior distributions, initial values and MCMC setup defined in the previous application. Because we impose the equality between $\lambda_{3j}$ and $\lambda_{1j}\lambda_{2j}$, the scatter plots comparing their values indicate correlation 1. Figure 5.6 shows the 95% credible interval and the posterior mean for $\alpha_{il}$ such that $i \in (G_1 \cup G_2)$. Note that most non-zero loadings, related to the same factor, indicate posterior estimates with the same sign. This fact is observed for all data
sets, and it supports the CNA interpretation for factors 1 and 2. Recall that the zero estimates are imposed via prior distribution to satisfy our assumptions for this group of genes.

Table 5.1 indicates (main diagonal) the number of genes affected by multiplicative interactions in each real data application. Note that the majority of subjects are free from interactions effects. The elements off diagonal are the number of common genes belonging to the intersection between the groups of affected genes. As can be seen, at least 17 genes can be found in the intersections involving different data sets. This result can be used as an argument against the idea that the model might be identifying interactions for a random set of genes. The intersections involving three data sets have 3-5 elements. Only 1 gene belongs to the intersection of all four data sets; its official full name is “GTP binding protein 4”, and it is located in Chromosome 10.

Table 5.1: Pairwise intersections between the breast cancer data sets; number of common genes affected by the multiplicative interaction. In order to simplify the notation, consider the first name indicated in the citation of the related reference.

<table>
<thead>
<tr>
<th></th>
<th>Chin</th>
<th>Miller</th>
<th>Sotiriou</th>
<th>Wang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chin</td>
<td>312</td>
<td>26</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>Miller</td>
<td>26</td>
<td>163</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>Sotiriou</td>
<td>24</td>
<td>17</td>
<td>265</td>
<td>20</td>
</tr>
<tr>
<td>Wang</td>
<td>19</td>
<td>24</td>
<td>20</td>
<td>233</td>
</tr>
</tbody>
</table>

Figure 5.7 shows the 3-dimensional surface plot representing the multiplicative effect associated with gene $i$. As can be seen, this type of interaction has a saddle shape. Each point in the surface corresponds to a different sample $j$. In the $x$ and $y$ axes we have $\lambda_1 j$ and $\lambda_2 j$; the $z$ axis represents $\alpha_{i3} \lambda_3 j$. The loading $\alpha_{i3}$ controls how strong is the interaction effect; values close to zero defines flatter surfaces. The sign of $\alpha_{i3}$ determines the orientation of the saddle shape. In each panel, the graph on the left is related to the smallest negative $\alpha_{i3}$, while the graph on the right represents
the largest positive $\alpha_{i3}$.

\[ \begin{array}{cc}
(a) & (b) \\
\end{array} \]

\[ \begin{array}{cc}
(c) & (d) \\
\end{array} \]

**Figure 5.7**: 3-dimensional surface plots representing the multiplicative interaction effect $\alpha_{i3}\lambda_{3j}$ (Assume approach 2 = perfect product). In each panel, the left graph is related to the smallest negative loading; whereas, the graph on the right is related to the largest positive loading. Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005).

### 5.4 Conclusions

In an ordinary factor analysis to describe the expression pattern across samples, the involvement of any gene with the factors is always additive. Biological pathways establishing complex structure of dependencies between genes is a key element suggesting the idea of interaction between factors. As an example, suppose that factor $l \in \{1, 2\}$ is directly associated with some activity involving the group of genes $G_l$, and a third group is influenced by these two groups. It seems reasonable to expect that the interaction between factors 1 and 2 may have significant effects over the
expression of the genes in the third group.

In the current chapter, we have developed the study of multi-factor models including interaction effects between latent factors. The analysis is focused on the particular case of pairwise multiplicative interactions, but the model proposed in Section 5.1 can use any function defining a relationship between a pair of factors. In order to simplify the notation assume, without loss of generality, two target factors and a third one representing the interaction term. Two approaches were considered to define the relationship between \( \lambda_{3j} \) and \( \lambda_{1j}\lambda_{2j} \). Approach 1 assumes a Gaussian prior parameterized with mean \( \lambda_{1j}\lambda_{2j} \) and a small variance \( \nu \). Approach 2 considers the perfect association \( \lambda_{3j} = \lambda_{1j}\lambda_{2j} \). Both approaches provide nice results in the synthetic data analysis developed in Section 5.2. A real data application involving four breast cancer data sets was presented in Section 5.3. That application was motivated by studies in the literature suggesting the presence of Copy Number Alteration in some regions of the genome. The central position of each region is known, but the list of genes involved with the chromosomal duplications was unknown. For this reason, a procedure was defined to select such genes for each region. In order to reduce the computational cost, another selection process was implemented to choose the most interesting genes for the analysis. The model was fitted to a \((3744 \times n)\) matrix \( X \), and we have noted that the parameter \( \nu \) should be extremely small to assure the target interaction effect for \( \lambda_{3j} \) in large data sets. The model assuming approach 2 does not have this problem, and its results indicated 163-312 genes affected by interactions depending on the data. The intersection of any two data sets has at least 17 genes detected with interaction effects. The identification of the model is obtained based on assumptions regarding the gene-factor relationship. In brief, we set prior distributions favoring the association of \( G_l \) with factor \( l \), where \( l \in \{1, 2\} \) is related to a region detected with CNA. The estimate of the factor loadings for \( i \in (G_1 \cup G_2) \) indicated that the CNA interpretation is valid for factors 1 and 2.
Non-linear interactions between factors

The study of non-linear interactions between observed variables have been the focus of many publications in the context of regression problems. In many cases, the proposed model introduce the non-linearity through the specification of Gaussian Process priors. Henao and Winther (2010) consider sparse and identifiable linear latent variable (factor) and linear Bayesian network models for parsimonious analysis of multivariate data. The framework consists of a fully Bayesian hierarchy for sparse models using spike and slab priors, non-Gaussian latent factors and a stochastic search over the ordering of the variables. The authors argue that the model is flexible in the sense that it can be extended by only changing the prior distribution of a set of latent variables to allow for non-linearities between observed variables through Gaussian Process priors. Hoyer et al. (2009) study the relationships between a set of observed variables in the context of Directed Acyclic Graphs (DAG). In brief, each observed variable (node in a DAG) is obtained as a function of its parents plus independent additive noise. An arbitrary function is chosen to define linear/non-linear relationships between the observed values. The paper evaluates whether a DAG is consistent with the data by constructing a non-linear regression of each variable on
its parents, and subsequently testing whether the resulting residuals are mutually independent. Gaussian Process regression and kernelized independence tests are used in the paper. In the spirit of factor analysis, Teh et al. (2005) model the relationships among components of a response vector $y$ using linear (or generalized linear) mixing of underlying latent variables indexed by a covariate vector $x$ (observed values). The authors assume that each latent variable is conditionally independently distributed according to a Gaussian process, with $x$ being the (common) index set. The mean of the response $y$ is then a function of a linear combination of the conditionally independent Gaussian processes.

Most applications of Gaussian Process (GP) models involve learning tasks where both output and input data are assumed to be given at training time. Lawrence (2004) and Lawrence (2005) have proposed a multiple-output GP regression model assuming observed output data, and latent variables as inputs. The approach explores non-linear interactions between the latent factors. The authors introduce a probabilistic interpretation of principal component analysis (PCA) named dual probabilistic PCA (DPPCA). The DPPCA model has the advantage that the linear mappings from the latent space to the data-space can be easily non-linearized through Gaussian processes (DPPCA with a GP introducing non-linearity is then called Gaussian Process Latent Variable Model or GP-LVM). The GP (assumed for latent variables) with an inner product kernel in the covariance function defines a linear association, and it has an interpretation as probabilistic PCA model. GP-LVM can be obtained by replacing this inner product kernel with a non-linear covariance function. The non-linear mappings are designed to address the weaknesses in visualizing data sets that arise when using statistical tools that rely on linear mappings, such as PCA and standard factor models. The analyses are based on optimization via maximum likelihood and sometimes the maximum $a$ posteriori (MAP) estimation; no MCMC algorithm is applied.
Wang et al. (2007) introduces a multi-factor GP model for learning distributions of styles of human motion. Their model can be viewed as a special class of the GP-LVM explored in Lawrence (2005). As in the GP-LVM, they marginalize out the weights (factor loadings), and optimize the latent variables that correspond to the different factors in the model. The GP assumed for each factor is kernelized independently, allowing non-linear mappings from any particular factor to the data. Several researchers have used the GP-LVM to model human poses, for example Grochow et al. (2004) and Urtasun et al. (2005). Wang et al. (2006) extends the GP-LVM by including a dynamical model on the low-dimensional latent space. It thereby models time series data for a single individual, but does not generalize well to multiple styles or activities.

Titsias and Lawrence (2010) studied the GP-LMV model, with focus on the Bayesian framework, to perform non-linear dimensionality reduction. The authors seek to apply the variational inference where a variational distribution is introduced to approximate the true posterior distribution over the latent variables. The main difficulty is that this application of variational Bayes requires to approximately integrate out the latent/input variables that appear non-linearly in the inverse kernel matrix of the GP model, and standard mean field variational methodologies cannot handle this problem. According to the paper, the key ingredient that makes the approach tractable is the application of variational inference in an expanded probability model where the GP prior is augmented to include auxiliary inducing variables.

In GP models, inference is analytically tractable for regression problems, and deterministic approximate inference algorithms are widely used for classification problems. The use of MCMC methods to sample from posterior distributions in a model assuming GP prior have been explored in the literature only for cases with observed input data. As an example, Titsias et al. (2009) describe an MCMC algorithm which constructs proposal distributions by utilizing the GP prior. At each iteration, the
algorithm generates control variables and samples the target function from the conditional GP prior. The control variables are auxiliary points associated with observed input variables defined in the model. An advantage of MCMC over deterministic approximate inference is that the sampling scheme will often not depend on details of the likelihood function, and is therefore very generally applicable. In addition, the development of deterministic approximations is difficult since the likelihood can be highly complex.

The combination of topics “microarray gene expression” and “Gaussian Process for latent variables” have not yet been explored in the literature. In the present section, we are interested in the study of non-linear interactions between factors defined in a factor model. The model is designed for the analysis of expression patterns across samples, and our goal is to identify non-linear relationships between genes included in the data set. We introduce non-linearities through the specification of a GP prior for a set of latent variables, and apply the Bayesian framework using MCMC methods to sample from all posterior distributions defined in the model. Taken together, these aspects regarding the type of data and the modeling process qualify our method as new statistical approach.

Consider the following outline for this chapter. Section 6.1 describes the multifactor model to evaluate non-linear associations between factors. Five different versions of the model are explored, they differ in terms of prior formulations for probability parameters, and the assumption regarding the similarity of the interaction effects affecting distinct genes. Section 6.2 presents a simulated study using three synthetic data sets. Here, our goal is to investigate the performance of the model in term of inference results. In Section 6.3, we develop real data applications related to the Copy Number Alteration problem considered in the previous chapter; different breast cancer data sets are analyzed. In Section 6.4, a second simulated study is developed to compare the results from the factor model with linear and non-linear
structure of interactions; we assume multiplicative interactions in this case. Finally, Section 6.5 indicates the main conclusions of the chapter.

6.1 Factor model with non-linear interactions

Assume the model:

$$X = \alpha \lambda + F + \epsilon,$$  \hspace{1cm} (6.1)

where $\alpha$ is the $(m \times L)$ matrix of factor loadings, $\lambda$ is the $(L \times n)$ matrix of factor scores, $F$ is the $(m \times n)$ matrix of interaction effects, and $\epsilon$ is the $(m \times n)$ noise matrix. Let $\epsilon_{ij} \sim N(0, \sigma^2_i)$ and denote $\sigma^2 = (\sigma^2_1, \sigma^2_2, ..., \sigma^2_m)'$. Note that this model is defined with $L$ factors, $m$ subjects and $n$ samples. We chose to work without the mean expression parameter $\mu$ defined in other chapters. Besides parsimony, this configuration reduces the computational cost to fit large real data sets. In all applications, the rows of matrix $X$ are standardized to define $\mu = 0$.

In this multi-factor model, the interaction effects are defined in the matrix $F$, separated from $\alpha \lambda$. If no constraint is imposed to these elements, the model will experience identifiability issues. As an example, consider the $i$-th row of $\alpha \lambda + F$ and note that $\alpha_i \lambda + F_i = C \alpha_i \lambda + F_i^*$ where $F_i^* = (1 - C)\alpha_i \lambda + F_i$ and $C$ is any real number. As can be seen, we can fit the data in matrix $X$ using infinitely many values for the pair $\alpha \lambda$ and $F$. Another identifiability problem is the fact that the columns of $\alpha$ and the rows of $\lambda$ can switch positions and/or sign, accordingly. Suppose that $G_1, G_2, ..., G_L$ and $G_E$ are disjoint sets of observations partitioning the subjects represented by the rows of $X$. The proposed model is intended for applications where we assume $G_l$ as a group of observations directly associated with factor $l = 1, ..., L$. In this case, $G_E$ is the set of individuals with unknown association with any factor $l$. Our goal is to measure the interaction between factors, and identify the elements in $G_E$ affected by such interactions. We can take advantage of the known subject-
factor relationship involving the elements in \( G_l \), for \( l = 1, \ldots, L \), to impose via prior
distribution a specific configuration for \( \alpha \) and \( F \) in (6.1). In particular, we assume
that most individuals are not affected by interactions; therefore, prior distributions
favoring \( F_i = 0 \) can be applied. According to this assumption, \( \alpha_i \lambda + F_i = \alpha_i \lambda + 0 \)
for most rows \( i \), therefore, \( \alpha \lambda \) and \( F \) can be identified because their rows are not free
to communicate and share values. Further details regarding the identifiability issues
are presented in the next section.

Different versions of the factor model will be explored in our analysis. These
versions differ in terms of prior formulations for \( \alpha_{il} \) and \( F_i \). In all cases, we set the
prior \( \sigma_i^2 \sim IG(a, b) \) for a conjugate analysis, and the specification \( \lambda_j \sim N_L(0, I_L) \).

As usual, a bimodal sparsity promoting prior is chosen for \( \alpha_{il} \). Consider \( \alpha_{il} \sim (1 - h_{il}) \delta_0(\alpha_{il}) + h_{il} N(0, \omega) \) where \( h_{il} \) is a binary indicator variable. We explore two
different forms of expressing our prior uncertainty for the probability that \( h_{il} = 1 \):

1. \( h_{il} \sim \text{Bernoulli}(q_{il}) \) and \( q_{il} \sim \text{Beta}(\gamma_1, \gamma_2) \);

2. \( h_{il} \sim \text{Bernoulli}(q_R), \ R \in \{R_1, R_2, R_3\} \), and \( q_R \sim \text{Beta}(\gamma_{1,R}, \gamma_{2,R}) \). Let \( R = R_1 \)
   if we suspect that gene \( i \) and factor \( l \) are associated, \( R = R_2 \) if no association
   is expected, and \( R = R_3 \) if the relationship is unknown.

According to specification 1, \( q_{il} \) is updated using a single observation \( h_{il} \), and this
strategy can be useful in applications involving large data sets. In specification 2, \( q_R \)
is updated based on the group of \( h_{il} \) such that \((i, l) \in R\). If the group of indices \( R_3 \)
contains a large number of elements and \( \alpha_{il} \neq 0 \) for most \((i, l) \in R_3\), the probability
\( q_{R_3} \) tends to be large which favors \( h_{il} = 1 \). As a result, very few or none \( \alpha_{il} \) related
to \( R_3 \) will be zero, i.e., the level of sparsity is lower than what it should be. In the
next section we show that if \( m \) is small, the model performs well with both prior
specifications for \( h_{il} \).

Assume a mixture prior with two components for the interaction effect vector
One of the components is the degenerated distribution at 0, which allows for the possibility of having $F_i = 0$, i.e., no interaction effect for subject $i$. We will explore two versions of this mixture distribution. The first one assumes that $F_i$ can be different comparing affected subjects; whereas, the second version assumes that $F_i$ is the same for all affected individuals. In the context of gene expression analysis, the version 2 would be less realistic. The expressions for the indicated mixtures are:

1. $(F_i' | \lambda) \sim (1 - z_i)\delta_0(F_i') + z_iN_n(0, K(\lambda))$
2. $(F_i' | F^*) \sim (1 - z_i)\delta_0(F_i') + z_i\delta_{F^*}(F_i')$ and $(F^* | \lambda) \sim N_n(0, K(\lambda))$.

where $z_i$ is an indicator variable, and $K(\lambda)$ is the covariance matrix obtained from the Squared Exponential covariance function depending on $\lambda$.

$$K(\lambda)_{j_1,j_2} = \exp \left\{ -\frac{1}{2l_s^2}||\lambda_{j_1} - \lambda_{j_2}||^2 \right\}, \quad (6.2)$$

where $(j_1, j_2) \in \{1, 2, ..., n\}$, $l_s$ is the characteristic length-scale and $||y||$ represents the Euclidean norm of the vector $y$. The covariance function is a crucial ingredient in the model, as it encodes our assumptions about the function we wish to learn. The Squared Exponential is stationary, isotropic, and probably the most widely-used kernel in the literature. Furthermore, it is infinitely differentiable, which means that a Gaussian Process with this choice has mean square derivatives of all orders, and is thus smooth; see Rasmussen and Williams (2006) for details. Note that if the points $\lambda_{j_1}$ and $\lambda_{j_2}$ are very close in the $\mathbb{R}^L$ space, then the samples $j_1$ and $j_2$ are similar and $K(\lambda)_{j_1,j_2} \approx 1$. On the other hand, the larger the distance between these points, the higher is the dissimilarity between samples $j_1$ and $j_2$, and the closer to 0 is $K(\lambda)_{j_1,j_2}$.

The length-scale $l_s$ is an adjustable parameter that controls how close the points $\lambda_{j_1}$ and $\lambda_{j_2}$ should be in order to be considered associated with each other.

The kernel (6.2) is a non-linear function of elements in $\lambda$. If $z_i = 0 \ \forall i$ in the mixture priors above, the model is said linear because it can be expressed as a linear
combination of $\lambda_{lj} \forall (l, j)$. On the other hand, if $z_i \neq 0$ for some $i$, $X_i$ will depend on $F_i$, which has a non-linear relationship with $\lambda$ through the covariance function; for this reason, we define the model as non-linear.

We also explore different strategies to express our prior knowledge about the indicator $z_i$. Assume the following possibilities:

1. $z_i \sim \text{Bernoulli}(\rho_i)$ and $\rho_i \sim \text{Beta}(\beta_1, \beta_2)$;
2. $z_i \sim \text{Bernoulli}(\rho)$ and $\rho \sim \text{Beta}(\beta_1, \beta_2)$;
3. $z_i \sim \text{Bernoulli}(\rho_R)$, $R \in \{R_1, R_2\}$ and $\rho_R \sim \text{Beta}(\beta_{1,R}, \beta_{2,R})$. Here, $R = R_1$ if we believe that subject $i$ is associated with some factor and is not affected by interactions. Let $R = R_2$ if the association between subject $i$ and any factor is unknown, and an interaction effects may exist.

As will be seen in the next section, all three formulations work well for data sets with few subjects. On the other hand, strategy 1 can be more convenient for applications involving large $m$, because it is less influenced by other observations. The strategy 2 assume a global probability $\rho$ representing the level of subjects affected by interactions. The updating distribution of $\rho$ takes into account all observations $z_i$. We expect very few rows of $F$ indicating non-zero effects; therefore, $\rho$ tends to be very small if $m$ is large. This situation favors $z_i = 0$, and thus the sparsity level in $F$ can be lower than it should be. This same type of problem, related to a large $m$, can occur with $\rho_{R_2}$ in the specification 3. Note that $\rho_{R}$ is updated with $z_i$ for all $i \in R$.

The target joint distribution $p(\alpha, \lambda, F, \sigma^2 | X)$ cannot be evaluated analytically. We use the structure of the Gibbs Sampling algorithm to sample from this distribution; the complete conditional posterior distributions are derived in Appendix F. In particular, the full conditional of $\lambda_{lj}$ depends on which specification we use for $p(F_i | \lambda)$. Appendix F shows that the closed form of the normalizing constant is not
available in this case, and thus an indirect sampling method is required to sample $\lambda_j$. We apply the Metropolis-Hastings algorithm with a random walk proposal distribution for this task. Further details regarding the indicated algorithms can be found in Gamerman and Lopes (2006).

Table 6.1 provides an identification number for each configuration of prior distributions defining a factor model. As can be seen, we choose to investigate 5 different configurations. In the models 1, 3 and 5, we assume that the interaction effect can differ from row to row in $F$. On the other hand, the same interaction effect is considered for all affected subjects in Models 2 and 4. Note that, Model 5 is the only one using the specifications $h_{il} \sim \text{Bernoulli}(q_R)$ and $z_i \sim \text{Bernoulli}(\rho_R)$. In addition, Models 3 and 4 apply the global Bernoulli probability $\rho$.

Table 6.1: Prior specifications defining different models in the study of non-linear interactions between factors.

<table>
<thead>
<tr>
<th>Model</th>
<th>$h_{il}$</th>
<th>$F_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

In the next section, we develop a simulated study to evaluate the performance of the 5 models indicated in Table 6.1. Three data sets are simulated for the analysis.

6.2 Simulated data analysis

The models defined in the previous section are designed to study a matrix $X$ containing observations for $m$ subjects across $n$ samples. Our target application involves gene expression data and it is defined for a situation where $L = 2$ factors. Suppose that $G_l$ is a group of genes known to be associated with factor $l$, where $l \in \{1, 2\}$. We
assume no association between the subjects in $G_l$ and factor 2 if $l = 1$, or factor 1 if
$l = 2$. Let $G_E$ be a group of extra genes with unknown association gene-factor. The
sets $G_1$, $G_2$ and $G_E$ are disjoint, and only $G_E$ is supposed to be affected by interac-
tion terms involving factors 1 and 2. In our analysis, these assumptions are used to
address the identifiability issues in the factor model. We induce the model identifi-
cation through the choices of Beta prior distributions for the Bernoulli probabilities
related to the indicators $h_{il}$ and $z_i$.

Table 6.2: Prior distributions and initial values specified for the Bernoulli probabil-
ities related to the indicators $h_{il}$ and $z_i$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Param.</th>
<th>Prior</th>
<th>Init. val.</th>
<th>Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3, 4</td>
<td>$q_{il}$</td>
<td>Beta(2,1)</td>
<td>0.9999</td>
<td>(i $\in$ $G_1$, $l = 1$) or (i $\in$ $G_2$, $l = 2$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beta(1,2)</td>
<td>0.0001</td>
<td>(i $\in$ $G_1$, $l = 2$) or (i $\in$ $G_2$, $l = 1$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beta(1,1)</td>
<td>0.5</td>
<td>(i $\in$ $G_E$, $l \in {1, 2}$)</td>
</tr>
<tr>
<td>5</td>
<td>$q_{R_1}$</td>
<td>Beta(2,1)</td>
<td>0.9999</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$q_{R_2}$</td>
<td>Beta(1,2)</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$q_{R_3}$</td>
<td>Beta(1,1)</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1, 2</td>
<td>$\rho_i$</td>
<td>Beta(1,2)</td>
<td>0.0001</td>
<td>i $\in$ ($G_1 \cup G_2$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beta(1,1)</td>
<td>0.5</td>
<td>i $\in$ $G_E$</td>
</tr>
<tr>
<td>3, 4</td>
<td>$\rho$</td>
<td>Beta(1,1)</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\rho_{R_1}$</td>
<td>Beta(1,2)</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\rho_{R_2}$</td>
<td>Beta(1,1)</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2 presents the configuration of priors and initial values for the probability
parameters involved in the proposed factor models. Note that the probability mass is
concentrated around 0 for the Beta(1, 2), around 1 for the Beta(2, 1), and Beta(1, 1)
= $U$(0, 1). These choices work well for the small data sets studied in this section. If $m$
is large, stronger priors should be consider to induce the identification of the model.

In Model 5, $q_{R_1}$ is related to $\{h_{il} : i \in G_1\}$, $q_{R_2}$ is related to $\{h_{il} : i \notin (G_1 \cup G_E)\}$, and
$q_{R_3}$ is associated with $\{h_{il} : i \in G_E\}$. Further, $\rho_{R_1}$ is related to $\{z_i : i \in (G_1 \cup G_2)\}$
and $R_2$ is associated with $\{z_i : i \in G_E\}$. We set $\rho_{R_1} \sim$ Beta(1, 2) to favor $z_i = 0$ (no
interaction effect). Similarly, in Models 1 and 2, the Beta(1, 2) is applied to favor
$F_i = 0$ affecting the subjects in $G_1$ and $G_2$. The configuration of initial values in Table 6.2 is also indicated to induce the identification of the model; for example, in Models 1 and 2 we assume $\rho_i^{(0)} = 0.0001$ for $i \in (G_1 \cup G_2)$ to favor $F_i = 0$.

With respect to the other parameters in the models, we choose $\omega = 10$ in the mixture prior for $\alpha_{il}$, and IG($a = 2.1, b = 1.1$) for $\sigma_i^2$. In addition, we set $l_s = 0.2$ as the characteristic length-scale in (6.2). In terms of initial values, consider $\alpha_{il}^{(0)} = 0, (\sigma_i^2)^{(0)} = 1$ and $\lambda_{ij}^{(0)} \sim N(0,1)$. The initial value of the indicator variables $h_{il}$ and $z_i$ are generated from the Bernoulli distribution assuming, in each model, the corresponding initial probabilities given in Table 6.2. The MCMC algorithm performs 2000 iterations with a burn-in period involving the first 1000 elements of the chains. Fast convergence to the target distribution is observed for all models.

Recall that the elements $\alpha_{il}$ and $\lambda_{l}$ can switch their signs. This issue does not affect the inference involving $q_{il}$, $\rho_i$, and the conditional probabilities $p(h_{il} = 1|\ldots)$ and $p(z_i|\ldots)$. If the posterior estimate of $\alpha_{il}$ and its real value have opposite signs, we simply multiply the estimates of $\alpha_{il}$ and $\lambda_{l}$ by -1 to correct the problem.

In order to simulate the data evaluated in this section, consider the steps below:

1. Set $m = 15$ genes, $n = 100$ samples and $L = 2$ factors;
2. Assume $G_1$ and $G_2$ with 5 genes each ($G_E$ has 5 genes);
3. Set $\alpha_{il} = 0$ for $(i \in G_1, l = 2)$ and $(i \in G_2, l = 1)$,
   Generate $\alpha_{il} \sim N(0,1)$ for $(i \in G_1, l = 1)$ and $(i \in G_2, l = 2)$,
   Generate $\alpha_{il} \sim 0.5\delta_0(\alpha_{il}) + 0.5N(0,0.5)$ for $i \in G_E$;
4. For $i \in G_l$, set $\alpha_{il}$ with the same sign.
   The genes in $G_l$ are expected to show similar correlation between $X_i$ and $\lambda_{l}$;
5. Generate $\lambda_{ij} \sim N(0,1)$;
6. Let $F_i = 0$ for $i \in (G_1 \cup G_2)$, and $F_{ij} = \lambda_{1j}\lambda_{2j}$ for some subjects in $G_E$;
7. Generate $\sigma_i^2$, e.g., from an IG distribution with high probability mass in 0-1;
8. Simulate $\epsilon_{ij} \sim N(0,\sigma_i^2)$, and compute $X$ using expression (6.1).
Figure 6.1: Synthetic matrix $X$. The original data is displayed in the top. The second panel shows $X$ with standardized rows (rows and columns are sorted so that the first principal component is monotone).

Figure 6.2: Results from Model 1: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (a), $\lambda_{lj}$ (b), $\sigma_i^2$ (c) and $F_{ij}$ (d).
Figure 6.1 shows the heat map image of a simulated data set. The rows 1-5 are related to $G_1$, and the rows 11-15 correspond to $G_2$. The matrix $F$ contains interaction effects affecting the subjects 6, 7, 8 and 10. As indicated in the simulating data procedure, we assume the product of factor scores as the true effect for all cases.

Figure 6.2 presents the 95% credible intervals and the posterior mean obtained from Model 1. Note that most true values (circles) are located inside their corresponding interval. Panel (d) suggests a good performance for the identification of subjects affected by interaction terms; most values in matrix $F$ are zero and the model correctly identify them. In this simulated scenario, the true loadings are positive values for both $G_1$ and $G_2$. The dashed line in Panel (a) separates the two factors. We also investigate these types of results for other three simulated data sets with other configurations of $\alpha$; $\lambda_{lj}$, $F_{ij}$, and $\sigma^2_i$ are not altered. The conclusions are essentially the same; see the graphs in Section F.2 of Appendix F.

Figure 6.3 shows in Panel (a) the box plots representing the posterior distribution for the global probability parameter $\rho$ defined in Models 3 and 4. Three different simulated data sets are compared. As can be seen, all distributions concentrate the probability mass below 0.5, which suggests that $F_i = 0$ for most subjects $i$. Panel (b) presents the posterior distributions for the probability parameters $q_R$ and $\rho_R$ in Model 5. Again, three simulated data sets are compared. The factor loadings related to $R_1$ are expected to have non-zero values; the box plot for $q_{R_1}$, located above 0.5, suggests this behavior in all three simulations. Also, we expect $\alpha_{il} = 0$ for the cases related to $R_2$; the box plots for $q_{R_2}$ are located close to zero confirming this assumption. The box plot related to $q_{R_3}$ indicates the level of sparsity in the submatrix containing the loadings for the subjects in $G_E$. In simulation 3, we consider lower sparsity in that submatrix, and this configuration is detected in the corresponding box plot (higher variability and median around 0.5). All three simulations have similar results for $\rho_R$. This probability parameter seems to describe very well the configuration of
matrix $F$. The subjects related to $R_1$ are not affected by interactions; the box plots for $\rho_{R_1}$ indicate low probability of interaction effect. On the other hand, most subjects related to $R_2$ are affected by interactions; the box plots for $\rho_{R_2}$ indicate high probability of interaction effect.

Suppose that $\psi$ is a matrix of size $(d_1 \times d_2)$ containing the parameter $\psi_{ij}$ in the $i$-th row and $j$-th column. Furthermore, assume that $\hat{\psi}_{ij}$ is the estimate for the true value $\psi_{ij}$. Consider the statistic below denominated Average Absolute Distance (AAD).

$$AAD(\psi) = \frac{1}{d_1 d_2} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |\psi_{ij} - \hat{\psi}_{ij}|,$$

(6.3)

where $|y|$ is the absolute value of $y$. We can use this statistic to compare the approximation between posterior estimate and real value in the studies assuming synthetic data. Note that $AAD(\psi) > 0$, and a small value means good approximation.

The results presented in Figure 6.4 are based on the statistic (6.3). Each row of panels is associated with a different parameter, and the columns represent the three simulations. As can be seen, Models 2 and 4 have the lowest $AAD$ in all
comparison of models. This result was expected, because these two models assume the same interaction effect for the affected subjects, and the data is generated with this configuration. In general, the Models 1, 3 and 5 indicate similar results.

Figure 6.4: Average absolute distance between posterior mean and the true value. Comparison involving all models and simulated data sets.

6.3 Real data analysis

Here, we explore the results from real data applications using the factor model with non-linear structure of interactions. Consider again the Copy Number Alteration (CNA) problem investigated in Section 5.3 of the previous chapter. Our study is associated with breast cancer data and we use locations, in the human genome, suggesting the presence of CNA; see Lucas et al. (2010). Our first step is to select a collection of over-expressed genes in the affected region of a chromosome. A range is specified around the central position of the region, and all genes within that range are included. This selection procedure provides disjoint sets of genes associated with different segments with CNA. Three different breast cancer data sets are used: Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005).
Two latent factors are defined in our model for this type of application. In other words, $\lambda$ has two rows of factor scores, and each row describes the expression pattern across samples for the genes associated with a region where the CNA was detected. We will evaluate the model fit assuming three different pairs of chromosome locations. Table 6.3 identifies the position and chromosome number for each region. Denote by $G_1$ the group of genes around the first location in the pair. $G_2$ represents the collection of subjects around the second location. In Section 5.3, we have discussed an issue related to the selection of subjects based on an interval specified around a central location. This strategy can lead to the inclusion of cases unrelated to the CNA in the corresponding region. A cleaning procedure have been proposed to remove the problematic genes from $G_1$ and $G_2$; see Section 5.3 and Appendix E for details. Table 6.3 indicates the number of genes before and after the removal procedure.

Table 6.3: Regions detected with CNA in the human genome. We apply a procedure to remove genes unrelated to the CNA factors. The number of genes before and after this removal is presented.

<table>
<thead>
<tr>
<th>Region</th>
<th>Chr.</th>
<th>Position</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>117,844,879</td>
<td>38</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>35,152,961</td>
<td>50</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>101,400,207</td>
<td>45</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>68,771,985</td>
<td>42</td>
<td>18</td>
</tr>
</tbody>
</table>

The microarrays used in this application have 22283 genes, and each data set contains at least 118 samples available for analysis. In order to reduce the computational cost associated with this large amount of data, we have implemented a gene selection procedure to eliminate the subjects which might not be affected by interactions. This procedure have been applied in Section 5.3, and its full description is shown in Appendix E. In short, the method is based on the data set Chin.
et al. (2006), and we evaluate the pairs of regions (1,4), (2,4) and (3,4); see Table 6.3. Let $G_E$ represent the group of extra genes to be included in the analysis; $G_1$, $G_2$ and $G_E$ are disjoint sets. The selection procedure indicates 3717, 3704 and 3708 elements in $G_E$ for the pairs (1,4), (2,4) and (3,4). For the purpose of comparison, this configuration of $G_E$ is used to study all breast cancer data sets. Our goal is to identify which subjects in $G_E$ are affected by interactions.

The Model 1 in Table 6.1 seems to be more convenient for applications involving large $m$. In this case, we assume a particular Bernoulli probability for each indicator $h_{il}$ and $z_i$, which makes these variables less dependent on other observations. If a large number of indicators $h_{il}$ share the same Bernoulli probability $q_R$, the level of sparsity in $\alpha$ can be incorrectly determined. Note that if most loadings are non-zero values, $q_R$ tend to be large which favors $h_{il} = 1$ for all $(i,l)$ related to $q_R$. Similarly, if a large number of $z_i$ share the same probability $\rho$ (Models 3, 4) or $\rho_R$ (Model 5), and if $F_i = 0$ for most genes, then $\rho$ or $\rho_R$ tend to be small which favors $z_i = 0$ for all involved subjects. Here, the level of sparsity is too high and some interactions effects are neglected. In a real data application, it seems more realistic to assume different interaction effects for different affected genes; for this reason, Model 1 is preferred to Model 2.

Assume the following prior distributions related to Model 1: $\omega = 10$ in the mixture prior for $\alpha_{il}$ and $\sigma_i^2 \sim IG(a = 2.1, b = 1.1)$. Further, we set $l_s = 0.2$ as the characteristic length scale parameter in (6.2). Strong specifications are defined for $q_{il}$ to impose our assumptions regarding the gene-factor relationship for the cases in $G_1$ and $G_2$. Also, this strategy is used to provide the identification of the model. Consider the configuration:

- $p(q_{il} = 1) = 1$ for $(i \in G_1, l = 1)$ or $(i \in G_2, l = 2)$;
- $p(q_{il} = 0) = 1$ for $(i \in G_1, l = 2)$ or $(i \in G_2, l = 1)$;
- Beta(1, 1) for $(i \in G_E, l \in \{1, 2\})$;
Degenerated priors are important to retain the CNA interpretation of factors 1 and 2; otherwise, the gene-factor relationship in \((G_1 \cup G_2)\) can be overwhelmed by the large amount of information in \(G_E\). The \(U(0,1)\) is indicated for all other cases.

We do not expect interaction effects related to the genes in \(G_1\) and \(G_2\). These groups have a strong relationship with one latent factor, and no association with the other. Based on this assumption, we set \(p(\rho_i = 0) = 1\) for \(i \in (G_1 \cup G_2)\). In addition, recall that most rows of \(F\) should be null-vectors to ensure the identification between \(\alpha \lambda\) and \(F\). It is reasonable to expect few genes affected by interactions in the target application; as a result, one might choose a Beta distribution with higher probability mass below 0.5 for \(\rho_i\). We set \(\rho_i \sim \text{Beta}(1,1)\) for \(i \in G_E\), which works well in the applications of this section.

In terms of initial values of the chains, consider the usual choices \(\alpha_{id}^{(0)} = 0\), \((\sigma_i^2)^{(0)} = 1\), and \(\lambda_{ij}^{(0)} \sim N(0,1)\). Let \(F_{ij}^{(0)} = 0\) for all \((i,j)\). The probability \(q_{id}\) is initialized with 1 if \((i \in G_1, l = 1)\) or \((i \in G_2, l = 2)\), 0 if \((i \in G_1, l = 2)\) or \((i \in G_2, l = 1)\) and 0.1 otherwise. The probability \(\rho_i\) has initial value 0 if \(i \in (G_1 \cup G_2)\), and 0.5 if \(i \in G_E\). Finally, consider \(h_{id}^{(0)} \sim \text{Bernoulli}(q_{id}^{(0)})\) and \(z_i^{(0)} \sim \text{Bernoulli}(\rho_i^{(0)})\). The MCMC algorithm is set to perform 600 iterations with the first 300 cases representing the burn-in period; the chains seem to converge in all applications. The Metropolis-Hastings algorithm, used to sample from the full conditional posterior distribution of \(\lambda_{ij}\), has acceptance rate around 31-40%, 15-65% and 67-84% in the applications related to the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005).

The heat map images displaying each data set are presented in Section D.3 of Appendix D; the current section and the applications in Section 5.3 (Chapter 5) examine the same data. The analysis presented in the remainder of this Section is focused on the pair of chromosome locations \((2,4)\). Additional results related to the pairs \((1,4)\) and \((3,4)\) are reported in Section F.3 (Appendix F).
Figure 6.5 shows images indicating the interaction effects in the matrix $F$. The panel on the left represents the full matrix with 3744 rows and 118 columns; the color bar is constrained between (-1,1) for higher contrast. The second panel exhibits the cases $F_i \neq 0$. Note that we can identify 275 genes affected by non-linear interactions involving the two factors related to the expression of genes, detected with CNA, in region 2 and 4. Further, the second image suggests a coherent pattern for groups of subjects; several rows have similar decreasing or increasing estimated effect, as we move from the left to the right, across samples. This result supports the interpretation of $F_i$ as a representation of an interaction effect; on the contrary, a random pattern would be observed for most rows.

![Figure 6.5: Matrix $F$ containing interaction effects: Panel 1 = full matrix (3744 genes), Panel 2 = the cases $F_i \neq 0$. These results are related to the pair of locations (2,4). In Panel 2, rows and columns are sorted so that the 1st principal components are monotone.](image)

Figure 6.6 presents the posterior estimates and the corresponding 95% credible interval for the factor loadings related to genes in $G_1$ and $G_2$. The estimate and interval are computed for the component in the posterior mixture with the highest probability weight. Here, we evaluate the results for three different data sets. As can be seen, most intervals in $G_l$, $l = 1$ or 2, are located together below or above
0 suggesting $\alpha_{il}$ with the same sign. This result supports the association between factors 1-2 and the CNA detected for the genes in $G_1$ and $G_2$. In other words, the estimated interaction effects seems to be a result of the CNA in regions 2 and 4.

**Figure 6.6:** Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ such that $i \in (G_1 \cup G_2)$. The dashed line separates the two factors. Panels (a), (b) and (c) are related to the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005). These results are related to the pair of locations (2,4).

Figure 6.7 shows, in Panels (a) and (c), the 3-dimensional surface plot representing the shape of the estimated interaction effect for two genes. The x and y axes contain the estimated $\lambda_{1j}$ and $\lambda_{2j}$; therefore, each point in the x-y plane is related to a sample (microarray). These shapes are different suggesting distinct interaction effects for those genes. Panels (b) and (d) present the posterior mean used in the z axis of the graph, and the corresponding 95% credible interval indicating our posterior uncertainty related to the estimated surface.
Figure 6.7: 3-dimensional surface plot representing the estimated interaction effect in $F_{1524} \cdot (a)$ and $F_{1945} \cdot (c)$. Panels $(b)$ and $(d)$ contain the posterior mean (x mark), used to create the surfaces, and the corresponding 95% credible interval (bar). This result is related to the data set Chin et al. (2006) and the pair of chromosome locations (2,4).

Table 6.4 compares the list of affected genes related to different breast cancer data sets. The table is divided in three sections representing the pair of regions with CNA. The main diagonal in each section indicates the number of affected genes in the corresponding application. Note that all intersections are non-empty sets, i.e., different data sets indicate the same group of genes as affected by interactions. Given the large number of genes in $G_E$ and the relatively small list of affected cases determined in each application, the identification of elements in the intersections is an important result suggesting a plausible model. The intersections involving all three data sets have 1, 1 and 2 elements considering the pairs (1,4), (2,4) and (3,4).
Table 6.4: Pairwise intersections between the breast cancer data sets; number of common genes affected by interactions. In order to simplify the notation, consider the first name indicated in the citation of the related reference. All target pairs of regions are evaluated.

<table>
<thead>
<tr>
<th></th>
<th>Pair (1,4)</th>
<th></th>
<th>Pair (2,4)</th>
<th></th>
<th>Pair (3,4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chin</td>
<td>Miller</td>
<td>Wang</td>
<td>Chin</td>
<td>Miller</td>
</tr>
<tr>
<td>Chin</td>
<td>139</td>
<td>6</td>
<td>9</td>
<td>275</td>
<td>14</td>
</tr>
<tr>
<td>Miller</td>
<td>6</td>
<td>81</td>
<td>3</td>
<td>14</td>
<td>111</td>
</tr>
<tr>
<td>Wang</td>
<td>9</td>
<td>3</td>
<td>46</td>
<td>19</td>
<td>7</td>
</tr>
</tbody>
</table>

6.4 Comparison between factor models with linear and non-linear structure of interactions

The main aim of this section is to compare the results obtained from the factor models proposed in Chapter 5 and 6. We will use the same data sets simulated for the analysis presented in Section 6.2. Recall that we define $F_{ij} = \lambda_{i1}\lambda_{2j}$ as the true interaction term affecting some subjects $i \in G_E$. Figure 6.8 shows the surface plot representing the saddle shape of the true interaction effect. Since we use the same $\lambda$ in all simulations, this is our target interaction effect for all cases.

![Figure 6.8: 3-dimensional surface plot representing the true interaction effect in all simulated data sets.](image)

In Section 5.1, two approaches have been defined for the factor model with mul-
tiplicative effects. In approach 1, we assume \( \lambda_{3j} \sim N(\lambda_{1j}\lambda_{2j}, \nu) \) introducing the product effect; whereas, approach 2 considers the perfect multiplication \( \lambda_{3j} = \lambda_{1j}\lambda_{2j} \).

Note that the linear factor model defined in the previous chapter is a particular case of (6.1). The interaction effect in the model with non-linear structure is defined in \( F_i \), while this same effect is represented by \( \alpha_{i3}\lambda_3 \) in the model with linear structure. The data set was simulated assuming the framework in Section 6.1. In this case, suppose subject \( i \) is affected by interactions, we can write \( \alpha_{i3} = 1 \) and \( \lambda_3 = F_i \) as the corresponding true values in the linear model from Chapter 5. On the other hand, if subject \( i \) is not affected, we have \( \alpha_{i3} = 0 \). Remember that our analysis is focused on the scenario with \( n_f = 2 \) latent factors.

In terms of prior specifications, initial values and MCMC configuration, consider the same choices defined in the simulated data analysis developed for the involved models in their original chapter. We have already shown that these factor models can provide good results for fitting simulated data. In the current section, we concentrate on the comparison of surface plots to see how well each model can estimate the saddle shape presented in Figure 6.8. It can be anticipated that the linear model assuming multiplicative effects yields the best results. However, the model with non-linear structure is more general and can be used to investigate other types of interaction effects.

Figure 6.9 shows the 3-dimensional surface plots indicating the estimated interaction effect. Note that we can identify the saddle shape in all cases. As one might expect, the linear model from Chapter 5 (Panels c and d) produces a smoother surface than the non-linear model (Panels a and b). The linear model is in advantage, because it assumes the true saddle shape as the target effect. The length scale parameter \( l_s \) defined in (6.2) can be used to control the smoothness of the function related to \( F_i \) in the non-linear models. The current model-fit in Panels (a) and (b) is based on the choice \( l_s = 0.2 \). If this value is increased, the number of neighbors in-
fluencing each point increases; the covariance matrix is then more populated. Figure 6.10 presents the surfaces related to Models 1 and 2 assuming bigger choices of $l_s$. As can be seen, the level of irregularities in the middle of the graph seems reduced with respect to the $l_s = 0.2$ case; this conclusion is more evident for Model 1 with $l_s = 0.5$.

**Figure 6.9**: 3-dimensional surface plot representing the estimated interaction effect. Panel (a) = Model 1, (b) = Model 2, (c) = model with approach 1 (Gaussian prior, Chapter 5), and (d) = model with approach 2 (prefect product, Chapter 5). Models 1 and 2 are fitted assuming $l_s = 0.2$.

The smooth surfaces, for $l_s = 0.5$ in Figure 6.10, seem to be flatter and wider than the other cases. This characteristic can be interpreted as an indication of worse approximation between posterior estimates and true values. The bar plots in Figure 6.11 compares the ADD statistic (6.3) calculated for the parameters in the Models
1-2 with different choices of $l_s$. Note that the approximation is indeed worse when $l_s = 0.5$; the biggest ADD value is observed for $l_s = 0.5$ in all cases.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_10.png}
\caption{3-dimensional surface plot representing the estimated interaction effect. Here, the models are fitted assuming $l_s = 0.3$ or 0.5.}
\end{figure}

The section F.4 in Appendix F contains additional graphs reporting the results for applications involving other data sets (Simulations 2 and 3) and other models (Models 3, 4 and 5). Those graphs provide the same conclusions determined here.

Our final comparison analysis involves the results of the real data applications. Both frameworks from Chapters 5 and 6 have been used to fit the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005); we have studied the pair of regions (2,4), defined in Table 6.3. Each model provides a list of genes affected by interactions; we have found 22 (Chin), 7 (Miller) and 7 (Wang) genes in the
intersection of the lists generated for the same data set. This type of result reinforces the idea that the proposed models can be valid to study interaction effects.

Figure 6.11: Average absolute distance between posterior mean and the true value. Comparison involving Models 1-2 and different choices of the length-scale parameter $l_s$. Only the data set from Simulation 1 is used.

6.5 Conclusions

In this chapter, we have proposed a multi-factor model with a non-linear structure of interactions. Many publications, in the context of regression analysis, have considered Gaussian Process priors to study non-linear interactions between observed variables. This strategy motivated our model proposed in Section 6.1 where the prior distribution for one of the parameters has a Gaussian component with covariance matrix determined via the Square Exponential kernel. Our method evaluates the non-linear structure related to latent factors, and not many publications have considered this topic. The existing studies are based on optimization and deterministic approximations; no MCMC algorithm is applied. In addition, our model can be considered a new approach for its target application, which is the analysis of
coherent expression patterns across samples to identify interaction between factors determining relationship between genes.

In Section 6.1, different prior distributions were considered for the probability parameters defined in the mixture distributions specified for the factor loadings and the interaction term $F_i$. Also, we model the following assumptions: (i) the effect is the same for any pair of affected subjects, and (ii) the effect can be different. Five versions of the model were defined with the combination of the different prior distributions. Assumptions related to the intended type of application were used to specify priors inducing a specific configuration in the factor loadings matrix $\alpha$ and the matrix of interactions $F$. This strategy is important to assure the identification of the model. In Section 6.2, we developed a simulated study to verify the performance of the Models 1-5. Three different synthetic data sets were studied and each one assumes a particular configuration for the factor loadings matrix. All five versions of the model can approximate well the true values of the parameters. In Section 6.3, three breast cancer data sets were used to study interaction effects between factors related to regions, in the human genome, detected with CNA. Only one factor model was applied for this real analysis; the chosen configuration defines individual probabilities parameters in the mixture priors, which is convenient for applications with a large number of subjects. The results indicated some aspects suggesting a plausible model; for example, some affected genes are the same in the analysis of different data sets.
Conclusions and future work

This dissertation have shown different applications of factor models to analyze gene expression data. The data is organized as a matrix $X$ with the rows and columns representing subjects and samples, respectively. We consider two types of technologies to measure the expression values, the first one is the commercial high-density oligonucleotide microarray platform Affymetrix GeneChip®. The second type is a young technology known as high-throughput sequencing data or RNA-Seq. The observations of RNA-Seq are counts, whereas, continuous values represent fluorescent intensities on microarrays. We study the expression pattern across samples to address different problems. The subjects in the rows of $X$ can be: probes or probe sets depending on the problem (for microarrays), and short nucleotide sequences (for RNA-Seq). Each column of $X$ represents a different individual, but all samples are related to the same type of cancer cell. In our real data analyses, different cancer tumors were investigated (Breast, Ovarian, Lung and Brain).

The formulation of the models specifies mixture prior distributions for the factor loadings and, depending on the application, other target parameters. These priors are defined with two components; one of them is a point mass at 0 or a Gaussian
distribution with small variance. This configuration is used to allow for sparsity and then test whether the effect of factors on the subjects is significant or not. Simulated studies have been developed in all chapters to verify the performance of the proposed factor models. The use of synthetic data is convenient for this type of verification given that the true values of the parameters are known. In general, the estimates approximate well the real values which suggests a good behavior of the models.

In Chapter 2 we have considered a one-factor model to generate a detection call (present/absent) for each gene represented on oligonucleotide microarrays. Our method was compared with other two techniques (MAS 5.0 P/A and PANP) proposed in the literature for the same problem. The factor model have shown superior results in all comparison studies based on simulated or real data. However, this good performance depends on the number of microarrays available for analysis, i.e., the factor model is preferred to PANP and MAS 5.0 P/A if the sample size is not too small.

In Chapter 3 we have studied groups of over-expressed genes located altogether in a chromosomal region detected with Copy Number Alteration (CNA) for breast cancer data. Here, each row of $X$ represents a different gene in the group of over-expressed cases. A coherent pattern across samples can be observed in the heat map image of $X$ as a result of the indicated chromosomal abnormality. We have applied a one-factor model to evaluate the expression pattern in $X$, and then determine the presence/absence of CNA. The main idea of this application was to examine copy number changes for the same group of genes in different types of cancer. A probability parameter $q$ was defined in the model to measure our uncertainty about the presence/absence of CNA; the absence of CNA was detected only for a group of genes in Chromosome 1 (Brain cancer).

The high-throughput sequencing data (RNA-Seq) was analyzed in Chapter 4. We have studied the expression pattern across samples to evaluate the frequency
of incorrect sequence mapping to the target gene. We relied on the BLAST search algorithm to identify the possible correct mappings; this algorithm verifies whether an input sequence has similarities with other sequences in a public database. We have observed that coherent patterns are associated with few incorrect mappings, whereas, random patterns indicate cases with many sequences targeting the wrong gene. A factor model, based on the Poisson distribution, was proposed to analyze the count data across samples; it provides a more accurate classification of coherent/random pattern, than the visual inspection of heat map images. This model was implemented using the Gibbs Sampling structure, including the Metropolis-Hastings and the Adaptive Rejection Sampling to generated values from some full conditional distributions.

In Chapter 5 we have proposed a multi-factor model to study interaction effects between latent factors. The analysis was focused on pairwise multiplicative interactions, but any function defining a relationship between a pair of factors can be used. Two approaches were considered to introduce the interaction effect: (1) the product is inserted as the mean of a Gaussian prior, (2) we assume the perfect product between factors in a deterministic setup. In the real data application we have studied four breast cancer data sets. Two factors were defined in the model, and each one is directly associated with the genes located in a particular region, detected with CNA, of the human genome. The main aim was to identify other genes affected by the product interaction of the two factors. A selection process was implemented to choose the most interesting genes for this analysis, nevertheless the matrix $X$ represents a large number of subjects. In this case, approach 1 requires a Gaussian prior with extremely small variance to ensure the multiplicative effect. On the other hand, approach 2 does not suffer from the same problem given its deterministic formulation. Depending on the data set, we have observed 163-312 genes affected by interactions, and the pairwise intersections of these groups have at least 17 elements.
Finally, in Chapter 6 we have developed a multi-factor model with a non-linear structure of interactions. This version is more general and includes the model from Chapter 5 as a particular case. The non-linearities involving the latent factors were introduced through the Squared Exponential kernel, which defines the covariance matrix in the Gaussian component of a mixture prior specified for the parameter representing interaction effects. One version of this prior assumes that the effect can be different comparing affected genes; the less realistic assumption “same effect for any pair of affected subjects” was also studied. In addition, different prior formulations were considered for probability parameters in the mixture prior for interaction effects and factor loadings. As a result, five versions of the model were defined for investigation. Assumptions related to the intended type of application were used to choose the priors and induce a specific configuration in the matrices of factors loadings and interaction effects, which provides the identification of the model. In the real data application, we revisit the two-factor analysis based on regions with CNA. Three breast cancer data sets were explored, and interactions can be identified in all evaluations. The intersections of results from the three data sets are non-empty sets which suggests a plausible model.

7.1 Future work

In this final section, we describe alternatives to improve the analyses and address weaknesses identified in some applications. The models presented in Chapters 4 and 6 are the focus of the discussion below. Three topics are considered: The presence of outliers in the RNA-Seq count data, the smoothness of the function representing interaction effects, and the identification of clusters of genes with the same interaction effect.

In Chapter 4, we have observed outliers in some rows of $X$, i.e., few samples have a much bigger or smaller count than the majority of values in the same row. The
model assuming the ordinary Poisson distribution does not deal very well with these outliers. An alternative to the standard Poisson distribution is known as Generalized Poisson distribution (GPois), which has the probability function below.

\[
p(Y = y \mid \mu, \tau) = \frac{\mu(\mu + y\tau)^{y-1}}{y!} \exp\{-\mu - y\tau\}
\]

where \(\mu > 0\), \(0 \leq \tau \leq 1\), and \(y = 0, 1, 2, 3, \cdots\). When \(\tau = 0\), this distribution reduces to the standard Poisson(\(\mu\)). It can be shown that \(E(Y \mid \mu, \tau) = \mu(1 - \tau)^{-1}\) and \(\text{Var}(Y \mid \mu, \tau) = \mu(1 - \tau)^{-2}\). Consider the index of dispersion represented by the ratio \(\frac{\text{Var}(Y \mid \mu, \tau)}{E(Y \mid \mu, \tau)} = (1 - \tau)^{-2} \geq 1\). It follows that the GPois is suitable for count data observed with a sample variance considerably larger than the sample mean.

In a Poisson outlier context, a sample is thought to have been generated from an ordinary Poisson distribution and interest lies in the ordinary Poisson mean, but it is feared that a few observations from an overdispersed version may have contaminated the data set. Verdinelli and Wasserman (1991) and Scollnik (1995) use a mixture model, involving the GPois and the standard Poisson distributions, to address the outlier issue. Based on their approach, we can rewrite (4.1) in Chapter 4 as follows.

\[
X_{ij} \sim \text{GPois}(\mu_{ij}, w_{ij}\tau) \quad \text{with} \quad w_{ij} \sim \text{Bernoulli}(\xi_{i}) \quad \text{and} \quad 0 < \tau \leq 1.
\]

The parameter \(\xi\) is indexed by \(i\) to indicate that the probability of observing an outlier might differ between rows. We can develop a fully Bayesian analysis for a model assuming the formulation above, which proceeds by means of the Gibbs Sampler. The complete conditional posterior distribution of \(\tau\) is log-concave with respect to its arguments; therefore, the Adaptive Rejection Sampling algorithm can be used to generate samples.

Now, we move to a discussion regarding the smoothness of the function estimated as interaction effect in the factor model with non-linear structure presented.
in Chapter 6. Recall that if we increase the length-scale parameter \( l_s \) in the Squared Exponential (SE) covariance function, we will increase the number of neighbors affecting each point. As a result, the estimated function is flatter and wider, which translates to poor posterior approximations.

The use of a different covariance function can be an alternative to better combine smoothness and good posterior estimation. Of particular interest in this regard is the Matern class of covariance functions.

\[
K(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{r \sqrt{2\nu}}{l_s} \right)^\nu K_\nu \left( \frac{r \sqrt{2\nu}}{l_s} \right),
\]

with positive parameters \( \nu \) and \( l_s \), where \( K_\nu \) is a modified Bessel function (see Abramowitz and Stegun, 1965, sec. 9.6). In the context of Chapter 6, the input \( r \) is the Euclidean length of the vector \( \lambda_{j_1} - \lambda_{j_2} \) with \((j_1, j_2)\) representing distinct columns (samples) of matrix \( X \).

The parameter \( \nu \) is, in fact, a smoothness parameter. The SE covariance function \( \exp\{-r^2/(2l_s^2)\} \) is obtained for \( \nu = \infty \) (see Rasmussen and Williams, 2006, p. 204). Based on the Matern class, the process is \( k \)-times Mean Squared differentiable if and only if \( \nu > k \). The special case obtained by setting \( \nu = 1/2 \) gives the Exponential covariance function \( \exp\{-r/l_s\} \). The corresponding process is Mean Squared continuous but not Mean Squared differentiable, which connects to the idea of Brownian motion.

In summary, we currently control the range of influence between points using the parameter \( l_s \) (small \( l_s \) = local influence, large \( l_s \) = distant neighbors affect each other). In the SE kernel, \( \nu = \infty \) implying that the target function is infinitely differentiable. The Matern class introduces the smoothness parameter \( \nu \), and its choice indicate the number of times the process is Mean Squared differentiable. In order to improve smoothness and retain good posterior approximation, one could try
to balance the choices of \( l_s \) and \( \nu < \infty \).

In our closing discussion, we propose an extension of the factor model in Chapter 6 to cluster subjects with respect to the interaction effects. We have studied two mixture priors for \( F_i \); the first one assumes different interaction effects, and the second version defines the same effect for all affected subjects. These specifications take into account two extreme cases, i.e., the effects are all different or the same. It would be reasonable to consider the intermediate situation, where we identify groups of genes such that the non-linear interaction is the same within each group, but it differs between groups. In order to implement this assumption, we can use the clustering properties of the Dirichlet Process (DP) (Ferguson, 1973, 1974).

The following result is implied by the Polya urn scheme in Blackwell and MacQueen (1973), and it leads to the so called “Chinese Restaurant Process” (see Aldous, 1985, p. 92).

\[
(\psi_i | \psi_1, \cdots, \psi_{i-1}) \sim \left( \frac{\zeta}{\zeta + i - 1} \right) P_0 + \sum_{j=1}^{i-1} \left( \frac{1}{\zeta + i - 1} \right) \delta_{\psi_j},
\]

where \( \zeta \) is the concentration parameter and \( P_0 \) is the base distribution in the DP. This implies that the \( i \)-th subject is drawn from a new cluster with probability proportional to \( \zeta \) or is allocated to an existing cluster with probability proportional to the number of subjects in that cluster.

As a result, consider the prior specification below for the interaction term affecting subject \( i \).

\[
(F'_i | \lambda) \sim (1 - \rho_i) \delta_0(F_i) + \rho_i \, \text{DP}(\zeta, P_0)
\]

with \( P_0 = N_n[0, K(\lambda)] \), where \( K(\lambda) \) is the covariance matrix depending on \( \lambda \). Let \( c = 1, 2, \cdots, k \) be the cluster index, and assume \( s_c \in \{1, 2, \cdots, m\} \) such that \( F_{s_c} \neq 0 \) and \( s_1 < s_2 < \cdots < s_k \). The following steps are defined in the process.
• $F_{s_1} \sim P_0$;

• $(F_{s_2} \mid F_{s_1}) \sim \left(\frac{\zeta}{\zeta+1}\right) P_0 + \left(\frac{1}{\zeta+1}\right) \delta_{F_{s_1}}$;

• $(F_{s_3} \mid F_{s_1}, F_{s_2}) \sim \left(\frac{\zeta}{\zeta+2}\right) P_0 + \left(\frac{1}{\zeta+2}\right) \delta_{F_{s_1}} + \left(\frac{1}{\zeta+2}\right) \delta_{F_{s_2}}$;

• $(F_{s_k} \mid F_{s_1}, \ldots, F_{s_{k-1}}) \sim \left(\frac{\zeta}{\zeta+k-1}\right) P_0 + \sum_{c=1}^{k-1} \left(\frac{1}{\zeta+k-1}\right) \delta_{F_{s_c}}$;

Note that we do not need to specify the total number of clusters $k$. The data will indicate the most appropriate value through the scheme described above. Because matrix $X$ represents $m$ subjects, we can write $1 \leq k \leq m$. 

114
Appendix A

Posterior computation for Chapter 2

Here we derive the likelihood function and complete conditional posterior distributions associated with the factor model \( X = \mu 1_n + \alpha \lambda + \epsilon \), as indicated in (2.4). Denote \( \sigma^2 = (\sigma^2_1, \sigma^2_2, ..., \sigma^2_m)' \). All prior specifications are defined in Section 2.2, except \( \mu \sim N_m(\mu_0, \Sigma) \). Because we standardize the rows of \( X \), the model (2.1) does not contain \( \mu \). The calculations can be easily adapted to that case by letting \( \mu \) be a null \( m \)-dimensional vector. Define \( X_{\cdot j} \) as the \( m \)-dimensional column vector representing the \( j \)-th column of \( X \). Let \( X_{\cdot j} \sim N_m[\mu + \alpha \lambda_j, D] \) with \( D = \text{diag}(\sigma^2_1, \sigma^2_2, ..., \sigma^2_m) \). We assume conditional independence between samples; therefore,

\[
p(X \mid \mu, \alpha, \lambda, \sigma^2) = \prod_{j=1}^{n} p(X_{\cdot j} \mid \mu, \alpha, \lambda, \sigma^2) \\
= \prod_{j=1}^{n} (2\pi)^{-\frac{m}{2}} |D|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(X_{\cdot j} - \mu - \alpha \lambda_j)' D^{-1}(X_{\cdot j} - \mu - \alpha \lambda_j)] \right\} \\
= (2\pi)^{-\frac{mn}{2}} |D|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} [X_{\cdot j}' D^{-1} X_{\cdot j} - 2\mu' D^{-1} X_{\cdot j} - 2\alpha' \lambda_j D^{-1} X_{\cdot j} + +\mu' D^{-1} \mu + \alpha' \lambda_j^2 D^{-1} \alpha + 2\alpha' \lambda_j D^{-1} \lambda_j] \right\} .
\]
The Bayes theorem using the previous likelihood function and the prior specification $\mu \sim N_m(\mu_0, \Sigma)$ provides the complete conditional posterior distribution $p(\mu | \alpha, \lambda, \sigma^2, X) \propto p(X | \mu, \alpha, \lambda, \sigma^2) p(\mu)$

\[
\propto \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} \left[ -2\mu' D^{-1}(X_j - \alpha\lambda_j) + \mu' D^{-1} \mu \right] \right\} \exp \left\{ -\frac{1}{2} \left( \mu - \mu_0 \right)' \Sigma^{-1} \left( \mu - \mu_0 \right) \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left[ \mu'\left( nD^{-1} + \Sigma^{-1} \right) - 2\mu' \left( \Sigma^{-1} \mu_0 + D^{-1} \left( \sum_{j=1}^{n} [X_j - \alpha\lambda_j] \right) \right) \right] \right\}
\]

The last expression is the kernel of $N_m(\mu_0, \Sigma)$ with $V_\mu = [nD^{-1} + \Sigma^{-1}]^{-1}$ and $M_\mu = V_\mu[\Sigma^{-1}\mu_0 + D^{-1} \left( \sum_{j=1}^{n} [X_j - \alpha\lambda_j] \right)]$.

Suppose $p(\alpha) = (2\pi)^{-\frac{m}{2}} |\Phi_l|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \alpha' \Phi_l^{-1} \alpha \right\}$ with $l \in \{1, 2\}$ indicating a component within the defined mixture prior (2.2). Denote $C_1 = (2\pi)^{-\frac{mn}{2}} |D|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} [X_j' D^{-1} X_j - 2\mu' D^{-1} X_j + \mu' D^{-1} \mu] \right\}$, and $C_{2,l} = (2\pi)^{-\frac{m}{2}} |\Phi_l|^{-\frac{1}{2}}$. Note that

\[
p(X | \mu, \alpha, \lambda, \sigma^2) p(\alpha)
\]

\[
= C_1 C_{2,l} \exp \left\{ -\frac{1}{2} \left[ -2\alpha' \left( \sum_{j=1}^{n} \lambda_j D^{-1}(X_j - \mu) \right) + \alpha' \Phi_l^{-1} + D^{-1} \left( \sum_{j=1}^{n} \lambda_j^2 \right) \alpha \right] \right\}
\]

Assume further $\Phi_l^* = [\Phi_l^{-1} + D^{-1} \alpha \lambda]^{-1}$ and $M_l^* = \Phi_l^* [D^{-1}(X - \mu_1 n) \lambda]$. In addition, $C_{3,l} = \exp \left\{ +\frac{1}{2} \left( M_l^* (\Phi_l^*-1) M_l^* \right) \right\}$ and $C_{4,l} = (2\pi)^{\frac{m}{2}} |\Phi_l^*|^{\frac{1}{2}}$. As a result, we can write

\[
p(X | \mu, \alpha, \lambda, \sigma^2) p(\alpha)
\]

\[
= C_1 C_{2,l} C_{3,l} \exp \left\{ -\frac{1}{2} \left[ -2\alpha' (\Phi_l^*)^{-1} M_l^* + \alpha' (\Phi_l^*)^{-1} \alpha + M_l^* (\Phi_l^*-1) M_l^* \right] \right\}
\]

\[
= C_1 C_{2,l} C_{3,l} C_{4,l} (2\pi)^{-\frac{m}{2}} |\Phi_l^*|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\alpha - M_l^*') (\Phi_l^*)^{-1} (\alpha - M_l^*) \right\}
\]

\[
= C_1 C_{2,l} C_{3,l} C_{4,l} N_m[\alpha | M_l^*, \Phi_l^*].
\]

In summary, the complete conditional posterior distribution $p(\alpha | \mu, \lambda, \sigma^2, X)$ is a mixture of $N(M_l^*, \Phi_l^*)$ where $l \in \{1, 2\}$. Note that $C_{2,l} = N_m[0 | 0, \Phi_l]$ and $C_{3,l}^{-1} C_{4,l}^{-1} = 116$
The kernel of the Inverse Gamma distribution with parameters \(a, b\) posterior probability weight of component \(l = 2\) takes the form below.

\[
q^* = \frac{C_1C_{2,2}C_{3,2}C_{4,2} q}{C_1C_{2,2}C_{3,2}C_{4,2} q + C_1C_{2,1}C_{3,1}C_{4,1} (1 - q)} = \frac{q}{q + \frac{C_{2,1}C_{3,1}C_{4,1}}{C_{2,2}C_{3,2}C_{4,2}} (1 - q)}.
\]

The complete conditional posterior distribution of \(\sigma^2\) is shown below.

\[
p(\sigma^2 | \mu, \alpha, \sigma^2, X) \propto p(X|\mu, \alpha, \lambda, \sigma^2) p(\sigma^2)
\]

The kernel of the Inverse Gamma distribution with parameters \(a^* = a + \frac{n}{2}\) and \(b^* = B + b\) can be identified above.
Consider $X = \mu 1_n + \alpha \lambda + \epsilon$, where $\alpha$ is the $(m \times L)$ matrix of factor loadings and $\lambda$ is the $(L \times n)$ matrix of factor scores. Note that we define a model with $L$ factors. Set $L = 1$ to obtain the single factor specification. The model without $\mu$ is assumed when the rows of $X$ are standardized. In that case, let $\mu$ be a null $m$-dimensional vector in the expressions below. Denote $\sigma^2 = (\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)'$. We can write the likelihood function in two different forms:

- **Likelihood 1:** Define $X\cdot j$ and $\lambda\cdot j$ as the $m$-dimensional and $L$-dimensional column vectors representing the $j$-th column of $X$ and $\lambda$, respectively. Note that $X\cdot j \sim N_m [\mu + \alpha \lambda\cdot j, D]$ with $D = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2)$. We assume conditional independence between samples; therefore, 

$$p(X \mid \mu, \alpha, \lambda, \sigma^2) = \prod_{j=1}^{n} p(X_j \mid \mu, \alpha, \lambda, \sigma^2)$$

$$= \prod_{j=1}^{n} (2\pi)^{-\frac{m}{2}} |D|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(X_j - \mu - \alpha \lambda\cdot j)'D^{-1}(X_j - \mu - \alpha \lambda\cdot j)] \right\}$$
\[
(2\pi)^{-mn/2} |D|^{-n/2} \exp \left\{-\frac{1}{2} \sum_{j=1}^{n} \left[ X_j' D^{-1} X_j - 2\mu' D^{-1} X_j - 2\lambda_j' \alpha' D^{-1} X_j + \mu' D^{-1} \mu + \lambda_j' \alpha' D^{-1} \alpha \lambda_j + 2\mu' D^{-1} \alpha \lambda_j \right] \right\}.
\]

- Likelihood 2: Denote \( X_i \) and \( \alpha_i \) as the \( n \)-dimensional and \( L \)-dimensional row vectors representing the \( i \)-th row of \( X \) and \( \alpha \), respectively. Note that \( X_i' \sim N_n [1_n' \mu_i' + \lambda' \alpha_i', \sigma_i^2 I_n] \) with \( 1_n \) being a \( n \)-dimensional row vector of ones. We assume conditional independence between rows of \( X \); therefore,

\[
p(X \mid \mu, \alpha, \lambda, \sigma^2) = \prod_{i=1}^{m} p(X_i \mid \mu, \alpha, \lambda, \sigma^2) = \prod_{i=1}^{m} \left[ (2\pi)^{-\frac{n}{2}} |\sigma_i^2 I_n|^{-\frac{1}{2}} \right.
\exp \left\{-\frac{1}{2} \left[ (X_i' - 1_n' \mu_i' - \lambda' \alpha_i')' (\sigma_i^2 I_n)^{-1} (X_i' - 1_n' \mu_i' - \lambda' \alpha_i') \right] \right\}
\exp \left\{-\frac{1}{2} \sum_{i=1}^{m} \frac{1}{\sigma_i^2} \left[ X_i X_i' - 2\mu_i 1_n X_i' - 2\alpha_i \lambda X_i' + \mu_i 1_n' \mu_i + \alpha_i \lambda \alpha_i' + 2\mu_i 1_n \lambda \alpha_i' \right] \right\}.
\]

These expressions for the likelihood function of the proposed model are equivalent. The posterior computation for each parameter can be simplified by choosing the appropriate version of the likelihood.

The Bayes theorem using the previous Likelihood function 1 and the prior specification \( \mu \sim N_m (\mu_0, \Sigma) \) provides the following complete conditional posterior distribution for \( \mu \).

\[
p(\mu \mid \alpha, \lambda, \sigma^2, X) \propto p(X \mid \mu, \alpha, \lambda, \sigma^2) \ p(\mu)
\propto \exp \left\{-\frac{1}{2} \sum_{j=1}^{n} \left[ -2\mu' D^{-1} (X_j - \alpha \lambda_j) + \mu' D^{-1} \mu \right] \right\} \exp \left\{-\frac{1}{2} (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0) \right\}
\propto \exp \left\{-\frac{1}{2} \left[ \mu' (nD^{-1} + \Sigma^{-1}) \mu - 2\mu' \left( \Sigma^{-1} \mu_0 + D^{-1} \sum_{j=1}^{n} [X_j - \alpha \lambda_j] \right) \right] \right\}.
\]

119
The last expression is the kernel of \( N_m(M_\mu, V_\mu) \) with \( V_\mu = [nD^{-1} + \Sigma^{-1}]^{-1} \) and \( M_\mu = V_\mu[\Sigma^{-1}\mu_0 + D^{-1}\sum_{j=1}^n(X_j - \alpha \lambda_j)] \).

Consider the following prior specification for components of the factor loadings matrix: \( \alpha_{il} \sim (1 - h_{il})\delta_0(\alpha_{il}) + h_{il}N(0, \omega) \) with \( h_{il} \sim \text{Bernoulli}(q_i) \) and \( q_i \sim \text{Beta}(\gamma_1, \gamma_2) \). The notation \( \delta_0(\alpha_{il}) \) means \( \alpha_{il} = 0 \) with probability 1.

The part of the Likelihood 2 depending on \( \alpha_{il} \) can be written as

\[
p(X \mid \mu, \alpha, \lambda, \sigma^2) \propto \exp \left\{ -\frac{1}{2\sigma_i^2} \left[ \frac{2}{\sigma_i^2} \sum_{j=1}^n \lambda_{ij}^2 - 2\alpha_{il} \sum_{j=1}^n \lambda_{ij}(X_{ij} - \mu_i) + 2\alpha_{il} \sum_{j=1}^n \sum_{l^* \neq l} (\alpha_{il^*} \lambda_{ij} \lambda_{il^*}) \right] \right\}
\]

Denote \( \alpha_{-\{il\}} \) as the set of all loadings in \( \alpha \), except \( \alpha_{il} \). In addition, define \( h \) as the \((m \times L)\) matrix with \( h_{il} \) in the \( i \)-th row and \( l \)-th column. Note that \( (\alpha_{il} \mid h_{il} = 1) \sim N(0, \omega) \). As a result, the corresponding complete conditional posterior distribution is

\[
p(\alpha_{il} \mid \mu, \alpha_{-\{il\}}, \lambda, \sigma^2, h, X) \propto p(X \mid \mu, \alpha, \lambda, \sigma^2) p(\alpha_{il} \mid h_{il} = 1) \propto \exp \left\{ -\frac{1}{2} \left[ \frac{\alpha_{il}^2}{\omega} \left( \frac{1}{\sigma_i^2} \sum_{j=1}^n \lambda_{ij}^2 \right) - 2\alpha_{il} \left( \sum_{j=1}^n \lambda_{ij}(X_{ij} - \mu_i) - \sum_{l^* \neq l} \alpha_{il^*} \lambda_{il^*} \lambda_{ij} \right) \right] \right\}.
\]

The last expression is the kernel of \( N(M_\alpha, V_\alpha) \) with \( V_\alpha = \left[ \frac{1}{\omega} + \frac{1}{\sigma_i^2} \sum_{j=1}^n \lambda_{ij}^2 \right]^{-1} \) and \( M_\alpha = V_\alpha \left[ \frac{1}{\sigma_i^2} \sum_{j=1}^n \lambda_{ij}(X_{ij} - \mu_i) - \sum_{l^* \neq l} \alpha_{il^*} \lambda_{il^*} \right] \). If \( h_{il} = 0 \), the complete conditional posterior distribution will be \( \delta_0(\alpha_{il}) \).

We update \( q_i \) as follows.

\[
p(q_i \mid h) \propto p(h \mid q_i) p(q_i) \propto \left[ \prod_{i=1}^m q_i^{h_{ii}} (1 - q_i)^{1 - h_{ii}} \right] q_i^{\gamma_1 - 1} (1 - q_i)^{\gamma_2 - 1} \propto q_i^{\gamma_1 + \sum_{i=1}^m h_{ii} - 1} (1 - q_i)^{\gamma_2 + m - \sum_{i=1}^m h_{ii} - 1}
\]

The last expression is the kernel of \( \text{Beta}(\gamma_1 + \sum_{i=1}^m h_{ii}, \gamma_2 + m - \sum_{i=1}^m h_{ii}) \).
Define \( q = (q_1, q_2, \ldots, q_L)' \). In order to update \( h_{il} \), compute the conditional probability:

\[
p(h_{il} = 1 \mid \mu, \alpha, \lambda, \sigma^2, q, X) = \frac{q_l}{q_l + (1 - q_l) N[0|M_{\alpha}, V_{\alpha}]}.\]

Notation: \( N[y|M, V] \) is the \( N(M, V) \) probability density function evaluated at \( y \).

Assume the prior specification \( \lambda \cdot j \sim N_{L}(0, I_{L}) \), and consider again the Likelihood version 1. Denote \( \lambda \cdot \{ -j \} \) as the matrix \( \lambda \) without the \( j \)-th column. The complete conditional posterior distribution of \( \lambda \cdot j \) is given by

\[
p(\lambda \cdot j \mid \mu, \alpha, \lambda \cdot \{ -j \}, \sigma^2, X) \propto p(X \mid \mu, \alpha, \lambda, \sigma^2) p(\lambda \cdot j)
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left[ \lambda'_j (\alpha' D^{-1} \alpha + I_L) \lambda_j - 2\lambda'_j \alpha' D^{-1} (X \cdot j - \mu) \right] \right\}
\]

Last expression is the kernel of \( N(M_{\lambda}, V_{\lambda}) \) with \( V_{\lambda} = (\alpha' D^{-1} \alpha + I_L)^{-1} \) and \( M_{\lambda} = V_{\lambda}[\alpha' D^{-1}(X \cdot j - \mu)] \).

Finally, the complete conditional posterior distribution of \( \sigma_i^2 \) is derived below.

Consider the Likelihood version 2, denote \( \sigma_{-i}^2 = \{ \sigma_1^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \ldots, \sigma_m^2 \} \), and define

\[
B = \frac{1}{2} [X_i X'_i - 2\alpha_i \lambda(X'_i - 1_n \mu'_i) - 2\mu_i 1_n X'_i + \mu_i 1_n 1'n \mu'_i + \alpha_i \lambda \lambda'_i].
\]

\[
p(\sigma_i^2 \mid \mu, \alpha, \lambda, \sigma_{-i}^2, X) \propto p(X \mid \mu, \alpha, \lambda, \sigma^2) p(\sigma_i^2)
\]

\[
\propto (\sigma_i^2)^{-\frac{a}{2}} \exp \left\{ -\frac{B}{\sigma_i^2} \right\} (\sigma_i^2)^{-(a+1)} \exp \left\{ -\frac{b}{\sigma_i^2} \right\} = (\sigma_i^2)^{-(a+\frac{a}{2}+1)} \exp \left\{ -\frac{(B + b)}{\sigma_i^2} \right\}.
\]

This is the kernel of the distribution IG\((a + \frac{a}{2}, B + b)\).
Appendix C

Posterior computation for Chapter 4

This Appendix is divided in two sections. First, we show the details regarding the derivation of the complete conditional posterior distributions for parameters of the Poisson linear factor model. In the second section, we present some aspects related to the application of the Adaptive Rejection Sampling in our study.

C.1 Likelihood function and complete conditional posterior distributions

Suppose $X_{i,j} \sim \text{Poisson}(\mu_{ij})$ and $\log(\mu_{ij}) = \eta_i + \alpha_i \lambda_j$, with $i = 1, \cdots, m$ and $j = 1, \cdots, n$. Notation: $m$ is the total number of nucleotide sequences, $n$ is the number of samples, $L$ is the number of factors, $X$ is the $(m \times n)$ matrix of data, $\alpha$ is the $(m \times L)$ matrix of factor loadings, $\lambda$ is the $(L \times n)$ matrix of factor scores, $\alpha_i = (\alpha_{i1}, \cdots, \alpha_{iL})$, $\lambda_j = (\lambda_{1j}, \cdots, \lambda_{Lj})'$ and $\eta = (\eta_1, \cdots, \eta_m)'$.

Consider the following hierarchical prior specification: $(\eta_i \mid \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2)$, $\theta_i \sim N(\theta_0, \phi)$ and $\sigma_i^2 \sim IG(a, b)$. Denote $\theta = (\theta_1, \cdots, \theta_m)'$ and $\sigma^2 = (\sigma_1^2, \cdots, \sigma_m^2)'$.

In addition, assume the following mixture of Normals as prior distribution for com-
ponents of the factor loadings matrix: $\alpha_{il} \sim (1 - h_{il})N(0, \omega_1) + h_{il}N(0, \omega_2)$ with $h_{il} \sim \text{Bernoulli}(q_l)$ and $q_l \sim \text{Beta}(\gamma_1, \gamma_2)$. Finally, we specify $\lambda_j \sim N_L(0, I_L)$ as the prior distribution for the factor scores in the $j$-th column of $\lambda$, $j = 1, \cdots, n$.

Given $\theta$ and $\sigma^2$, we assume conditional independence and write:

$$p(\eta \mid \theta, \sigma^2) = \prod_{i=1}^m p(\eta_i \mid \theta_i, \sigma^2_i) = (2\pi)^{-m/2} \prod_{i=1}^m (\sigma^2_i)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^m \frac{(\eta_i - \theta_i)^2}{\sigma^2_i} \right\}.$$  

The complete conditional posterior distribution for $\theta_i$ is given by

$$p(\theta_i \mid \eta, \theta_{-i}, \sigma^2) \propto p(\eta_i \mid \theta_i, \sigma^2_i) p(\theta_i)$$

$$\propto \exp \left\{ -\frac{(\eta_i - \theta_i)^2}{2\sigma^2_i} \right\} \exp \left\{ -\frac{(\theta_i - \theta_0)^2}{2\phi} \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \left( \frac{1}{\phi} + \frac{1}{\sigma^2_i} \right) \left[ \theta_i^2 - 2\theta_i \left( \frac{1}{\phi} + \frac{1}{\sigma^2_i} \right) \right] \left( \frac{\eta_i}{\sigma^2_i} + \frac{\theta_0}{\phi} \right) \right\}.$$  

The last expression is the kernel of $N(M_\theta, V_\theta)$, where $V_\theta = \left( \frac{1}{\phi} + \frac{1}{\sigma^2_i} \right)^{-1}$ and $M_\theta = V_\theta \left( \frac{m}{\sigma^2_i} + \frac{\theta_0}{\phi} \right)$.

The complete conditional posterior distribution for $\sigma^2_i$ is shown below.

$$p(\sigma^2_i \mid \eta, \theta, \sigma^2_{-i}) \propto p(\eta_i \mid \theta, \sigma^2) p(\sigma^2_i)$$

$$\propto (\sigma^2_i)^{-a/2} \exp \left\{ -\frac{(\eta_i - \theta_i)^2}{2\sigma^2_i} \right\} \left( \sigma^2_i \right)^{-a+1} \exp \left\{ -\frac{b}{\sigma^2_i} \right\}$$

$$\propto (\sigma^2_i)^{-a+1/2} \exp \left\{ -\frac{1}{\sigma^2_i} \left[ b + \frac{(\eta_i - \theta_i)^2}{2} \right] \right\}.$$  

The last expression is the kernel of $IG\left[ \frac{a}{2}, b + \frac{(\eta_i - \theta_i)^2}{2} \right]$.

Conditional on $\eta$, $\alpha$, $\lambda$ and $\sigma^2$, we assume $X_{ij}$ independent $\forall (i, j)$. As a result, we can write the likelihood function
\[ p(X \mid \eta, \alpha, \lambda, \sigma^2) = \prod_{i=1}^{m} \prod_{j=1}^{n} p(X_{ij} \mid \eta, \alpha, \lambda, \sigma^2) \]

\[ = \prod_{i=1}^{m} \prod_{j=1}^{n} \exp \{ X_{ij}(\eta_i + \alpha_i\lambda_j) - \exp(\eta_i + \alpha_i\lambda_j) \} \frac{1}{X_{ij}!} \]

\[ \exp \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}(\eta_i + \alpha_i\lambda_j) \right\} \exp \left\{ -\sum_{i=1}^{m} \sum_{j=1}^{n} \exp(\eta_i + \alpha_i\lambda_j) \right\} \prod_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{X_{ij}!}. \]

The complete conditional posterior distributions for \( \eta_i \) and \( \lambda_j \) are derived below.

\[ p(\eta_i \mid \eta_{-i}, \alpha, \lambda, \sigma^2, X) \propto p(X \mid \eta, \alpha, \lambda, \sigma^2) p(\eta_i) \]

\[ \propto \exp \left\{ \sum_{j=1}^{n} X_{ij}\eta_i \right\} \exp \left\{ -\sum_{j=1}^{n} \exp(\eta_i + \alpha_i\lambda_j) \right\} \exp \left\{ -\frac{(\eta_i^2 - 2\eta_i\theta_i)}{2\sigma_i^2} \right\}. \]

\[ p(\lambda_j \mid \eta, \alpha, \lambda_{(-j)}, \sigma^2, X) \propto p(X \mid \eta, \alpha, \lambda, \sigma^2) p(\lambda_j) \]

\[ \propto \exp \left\{ \sum_{i=1}^{m} X_{ij}\alpha_i\lambda_j \right\} \exp \left\{ -\sum_{i=1}^{m} \exp(\eta_i + \alpha_i\lambda_j) \right\} \exp \left\{ -\frac{1}{2} \lambda_j^2 \lambda_j \right\}. \]

The closed form of the normalizing constant is not available in the expressions above. An indirect sampling method is required to generate values from the target posterior distributions. We apply the Metropolis-Hastings algorithm with a random walk proposal distribution for these cases.

Denote \( h \) as the set of binary indicators \( h_{il} \) \( \forall(i,l) \), \( h_i = (h_{1,i}, \cdots, h_{m,i})' \) and \( q = (q_1, \cdots, q_L)' \). The complete conditional posterior distribution for \( \alpha_{il} \) is given by

\[ p(\alpha_{il} \mid \eta, \alpha_{-il}, \lambda, \sigma^2, h, q, X) \propto p(X \mid \eta, \alpha, \lambda, \sigma^2, h, q) p(\alpha_{il} \mid h_{il}) \]

\[ \propto \exp \left\{ \sum_{j=1}^{n} X_{ij}\alpha_{il}\lambda_{lj} \right\} \exp \left\{ -\sum_{j=1}^{n} \exp(\eta_i + \alpha_i\lambda_j) \right\} \exp \left\{ -\frac{\alpha_{il}^2}{2\omega_{h_{il}+1}} \right\}. \]

\[ \propto \exp \left\{ -\frac{\alpha_{il}^2}{2\omega_{h_{il}+1}} + \sum_{j=1}^{n} X_{ij}\alpha_{il}\lambda_{lj} - \sum_{j=1}^{n} \exp(\eta_i + \alpha_i\lambda_j) \right\}. \]
Although this univariate conditional density is not of any recognized form, it is log-
concave, and thus samples can be generated efficiently using the Adaptive Rejection
Sampling algorithm. See the next section for further details about the application
of this method.

Given the previous result, we cannot compute the conditional probability of
each component in the posterior mixture. Consider the convenient strategy de-
scribed below for updating the indicator \( h_{il} \). Denote \( P \) as the set of parameters
\( \{\eta, \alpha_{-\{il\}}, \lambda, \sigma^2, h_{-\{il\}}, q\} \).

\[
p(h_{il} = 1 \mid P, \alpha_{il}, X) = \frac{p(X, \alpha_{il} \mid h_{il} = 1, P) \ p(h_{il} = 1 \mid P)}{p(X, \alpha_{il} \mid h_{il} = 1, P) \ p(h_{il} = 1 \mid P) + p(X, \alpha_{il} \mid h_{il} = 0, P) \ p(h_{il} = 0 \mid P)}
\]

\[
= \frac{p(X \mid \alpha_{il}, P) \ p(\alpha_{il} \mid h_{il} = 1, P) \ q_l}{p(X \mid \alpha_{il}, P) \ p(\alpha_{il} \mid h_{il} = 1, P) \ q_l + p(X \mid \alpha_{il}, P) \ p(\alpha_{il} \mid h_{il} = 0, P) \ (1 - q_l)}
\]

\[
= \frac{p(\alpha_{il} \mid h_{il} = 1) \ q_l}{p(\alpha_{il} \mid h_{il} = 1) \ q_l + p(\alpha_{il} \mid h_{il} = 0) \ (1 - q_l)}
\]

\[
= \frac{N[\alpha_{il} \mid 0, \omega_2] \ q_l}{N[\alpha_{il} \mid 0, \omega_2] \ q_l + N[\alpha_{il} \mid 0, \omega_1] \ (1 - q_l)}.
\]

This conditional distribution for the \( h_{il} \) indicators takes exactly the form given by
George and McCulloch (1993) for log-linear models involving a mixture prior distrib-
ution with two Gaussian components. Ntzoufras et al. (2000) is another reference
exploring this Bernoulli probability, which involves only the calculation of the Normal
prior density of the current \( \alpha_{il} \) for both possible values of \( h_{il} \).

Finally, we update \( q_l \) as follows.

\[
p(q_l \mid h) \propto p(h_{il} \mid q_l) \ p(q_l) = \left[ \prod_{i=1}^{m} p(h_{il} \mid q_l) \right] \ p(q_l)
\]

\[
\propto (q_l)^{\gamma_1 + (\sum_{i=1}^{m} h_{il}) - 1} \ (1 - q_l)^{\gamma_2 + m - (\sum_{i=1}^{m} h_{il}) - 1}.
\]

The last expression is the kernel of Beta \( [\gamma_1 + \sum_{i=1}^{m} h_{il}, \gamma_2 + m - \sum_{i=1}^{m} h_{il}] \).
C.2 Derivative free adaptive rejection sampling to update \( \alpha_{il} \)

Denote \( f(\alpha_{il}) = (1/c)g(\alpha_{il}) \) where \((1/c)\) is the unknown normalization constant, and \( g(\alpha_{il}) \) is the available log-concave expression depending on \( \alpha_{il} \). In our study, the domain of the univariate function \( f(\alpha_{il}) \) is unbounded. We apply the derivative free ARS to sample \( \alpha_{il} \) from \( f(\alpha_{il}) \), using the concave function \( \log[g(\alpha_{il})] \). Assume \( g(\alpha_{il}) \) as the expression, shown in the previous section, proportional to the full conditional posterior distribution of \( \alpha_{il} \). The logarithm of that kernel is given below:

\[
\log[g(\alpha_{il})] = -\frac{\alpha_{il}^2}{2\omega_{il}+1} + \alpha_{il} \sum_{j=1}^{n} (X_{ij}\lambda_{lj}) - \sum_{j=1}^{n} \exp(\eta_i + \alpha_i\lambda_{lj}). \quad (C.1)
\]

For the univariate case, \( g(\alpha_{il}) \) is log-concave if \( d\log[g(\alpha_{il})]/d\alpha_{il} \) is a non-increasing function of \( \alpha_{il} \). We have the following derivative result in our application:

\[
\frac{d\log[g(\alpha_{il})]}{d\alpha_{il}} = -\frac{\alpha_{il}}{\omega_{il}+1} + \sum_{j=1}^{n} (X_{ij}\lambda_{lj}) - \sum_{j=1}^{n} \lambda_{lj} \exp(\eta_i + \alpha_i\lambda_{lj}). \quad (C.2)
\]

Suppose \( \alpha_A < \alpha_B \) are two possible values of \( \alpha_{il} \). Consider (C.2), and note that

\[-\frac{\alpha_A}{\omega_{il}+1} + \sum_{j=1}^{n} (X_{ij}\lambda_{lj}) > -\frac{\alpha_B}{\omega_{il}+1} + \sum_{j=1}^{n} (X_{ij}\lambda_{lj}).\]

If \( \lambda_{lj} = 0 \), the last result is enough to show the log-concavity of \( g(\alpha_{il}) \). Otherwise, if \( \lambda_{lj} < 0 \), we can write \( \alpha_A\lambda_{lj} > \alpha_B\lambda_{lj} \) and

\[
\exp(\alpha_A\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}) > \exp(\alpha_B\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}),
\]

\[-\lambda_{lj} \exp(\alpha_A\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}) > -\lambda_{lj} \exp(\alpha_B\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}).\]

If \( \lambda_{lj} > 0 \), then \( \alpha_A\lambda_{lj} < \alpha_B\lambda_{lj} \) and

\[
\exp(\alpha_A\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}) < \exp(\alpha_B\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}),
\]

\[-\lambda_{lj} \exp(\alpha_A\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}) > -\lambda_{lj} \exp(\alpha_B\lambda_{lj} + \eta_i + \alpha_{i(-l)}\lambda_{(-l)j}).\]
In conclusion, the previous results show that $d\log[g(\alpha_{il})]/d\alpha_{il}$ decreases monotonically with increasing $\alpha_{il}$, hence (C.1) is concave as required for the application of the ARS algorithm.

The first step of the ARS algorithm is the specification of a starting grid containing values in the domain of $\log[g(\alpha_{il})]$. Reasonable approximation can be obtained with a number of values in the grid as small as 3, and it improves as the number of points used increases. It is recommended that the grid contains at least one point with positive derivative and one point with negative derivative, if $\log[g(\alpha_{il})]$ allows such choice. We selected a grid with 9 abscissas $\{x_1 < x_2 < \cdots < x_9\}$ for our application.

The next step of the algorithm involves the computation of lower and upper piecewise-linear hulls, i.e., the boundaries for the concave function. Given that the domain of $\log[g(\alpha_{il})]$ is unbounded, the lower hull takes on $-\infty$ to the left of $x_1$ and right of $x_9$. For $\alpha_{il} \in [x_k, x_{k+1}]$ with $k = 1, 2, \cdots, 8$, the lower hull is given by the linear function $L_h(\alpha_{il}) = B_k + \alpha_{il}M_k$ where $B_k = \log[g(x_k)] - M_kx_k$ and $M_k = \frac{\log[g(x_{k+1})]-\log[g(x_k)]}{x_{k+1}-x_k}$. The function $L_h(\alpha_{il})$ represents the secant connecting the points $(x_k, \log[g(x_k)])$ and $(x_{k+1}, \log[g(x_{k+1})])$.

There are two lines forming the upper hulls between the abscissas $x_k$ and $x_{k+1}$, for $k = 2, 3, \cdots, 7$. These line segments are the secants computed as lower hulls for the adjacent intervals $[x_{k-1}, x_k]$ and $[x_{k+1}, x_{k+2}]$. The secants intersect each other at $t_{k-1} = \frac{\log[g(x_k)]-M_{k-1}x_k-B_{k+1}}{M_{k+1}-M_{k-1}}$. As a result, $B_{k-1} + \alpha_{il}M_{k-1}$ is the upper hull for $\alpha_{il} \in [x_k, t_{k-1}]$, and $B_{k+1} + \alpha_{il}M_{k+1}$ is the upper hull for $\alpha_{il} \in [t_{k-1}, x_{k+1}]$. There is no intersection of secants within the intervals $(-\infty, x_1]$, $[x_1, x_2]$, $[x_8, x_9]$ and $[x_9, +\infty)$; therefore, their upper hulls are $B_1 + \alpha_{il}M_1$, $B_2 + \alpha_{il}M_2$, $B_7 + \alpha_{il}M_7$ and $B_8 + \alpha_{il}M_8$ for any $\alpha_{il}$ in those regions, respectively.

The generation of a candidate $\alpha_{il}$ is the next step of the algorithm. First, we
randomly select one of those intervals for which upper hulls were defined. Suppose \( N_L \) and \( N_R \) are the left and right limits of the chosen interval. The candidate value is sampled along the line segment defining the upper hull as indicated below.

1. Generate \( u \) from \( U(0, 1) \);
2. Compute \( \alpha_{il} \) satisfying
\[
\exp(\alpha_{il} M + B) - \exp(N_L M + B) = u \\
\exp(N_R M + B) - \exp(N_R M + B)
\]
where \( M \) and \( B \) are the slope and intercept of the linear function defining the upper hull of the interval.

In particular, if \( N_L = -\infty \), then \( \alpha_{il} = \frac{\log(u)}{M} + N_R \); if \( N_R = +\infty \), then \( \alpha_{il} = \frac{\log(1-u)}{M} + N_L \). The \exp function used in the expression of Step 2 is convenient for the situations \( N_L = -\infty \) with \( M > 0 \), and \( N_R = +\infty \) with \( M < 0 \). The slope \( M \) of the linear segment representing the upper hull for the region to the left of \( x_1 \) is assumed positive; whereas, \( M \) is a negative number for the region to the right of \( x_9 \).

Finally, the acceptance/rejection test is indicated below. Denote \( L_h \) as the lower hull and \( U_h \) as the upper hull for the candidate \( \alpha_{il} \).

1. Generate \( u \sim U(0, 1) \);
2. Accept \( \alpha_{il} \) if \( u \leq \frac{L_h}{U_h} \), i.e., \( u \) is below the lower bound. Otherwise, evaluate the next step;
3. Accept \( \alpha_{il} \) if \( u \leq \frac{\log[g(\alpha_{il})]}{U_h} \), i.e., \( u \) is between the lower bound and the concave function. Otherwise, consider the next step;
4. Reject \( \alpha_{il} \), \( u \) is between \( \log[g(\alpha_{il})] \) and the upper bound.

If the candidate \( \alpha_{il} \) is accepted in Step 3 or rejected in Step 4 above, its value is incorporated in the grid of abscissas for the next iteration of the algorithm. For each iteration of the MCMC method, we run the ARS algorithm until one candidate is accepted.
Appendix D

Posterior computation and additional graphs for Chapter 5

This Appendix has three sections. The first one shows the derivation of the complete conditional posterior distributions for parameters of the linear factor model with multiplicative interactions. The second section presents some additional graphs related to the simulated study developed in Section 5.2. Section D.3 shows additional graphs for the real data analysis presented in Section 5.3.

D.1 Likelihood function and complete conditional posterior distributions

In this Appendix, we present elements of the posterior computation related to the factor model proposed to investigate multiplicative interactions between factors. Consider the model $X = \mu 1_n + \alpha \lambda + \epsilon$, where $X$ is the $(m \times n)$ matrix of data, $\mu$ is an $m$-dimensional column vector of mean expression, $1_n$ is an $n$-dimensional row vector of ones, $\alpha$ is the $(m \times L)$ matrix of factor loadings, $\lambda$ is the $(L \times n)$ matrix of factor scores, $\epsilon$ is the $(m \times n)$ matrix of noise terms, $m$ is the number of genes, and $n$ is
the number of samples, and $L$ is the number of factors plus pairwise multiplicative interaction terms.

We define $\lambda$ with a different structure to introduce multiplicative interactions between factors. Assume $2 \leq n_f < L$ where $n_f$ would be the number of latent factors in a model without interaction effects. Assume further that factor $l = 1, \cdots, n_f$ is represented in the $l$-th row of $\lambda$, and $\lambda_{(n_f+1)j}, \lambda_{(n_f+2)j}, \cdots, \lambda_{Lj}$ are related to $\lambda_{1j}\lambda_{2j}, \lambda_{1j}\lambda_{3j}, \cdots, \lambda_{(n_f-1)j}\lambda_{(n_f)j}$, respectively. Note that $L = n_f + \frac{n_f!}{(n_f-2)!2!}$.

Consider $\epsilon_{ij} \sim N(0, \sigma_i^2)$ and set the conjugate prior distributions $\sigma_i^2 \sim IG(a, b)$ and $\mu \sim N_m(\mu_0, \Sigma)$. In addition, we specify the typical conjugate prior $\lambda_{lj} \sim N(0, 1)$ for $l = 1, \cdots, n_f$. For the other rows, $l = n_f + 1, \cdots, L$, we consider two different approaches:

1. Let $\lambda_{lj} \sim N(\lambda_{l_1j}\lambda_{l_2j}, \nu)$ with $l_1 < l_2 \in \{1, \cdots, n_f\}$ depending on $l \in \{n_f + 1, \cdots, L\}$. The variance $\nu$ must be a small value; otherwise, we would express high uncertainty a priori with respect to the assumption of multiplicative interaction effects.

2. Consider the equality $\lambda_{lj} = \lambda_{l_1j}\lambda_{l_2j}$ with $l_1 < l_2 \in \{1, \cdots, n_f\}$ depending on $l \in \{n_f + 1, \cdots, L\}$. This strategy ensures the multiplicative effect and it is useful for applications involving many genes where $\nu$ is difficult to set.

We use the spike and slab mixture prior for the factor loadings to allow for sparsity. This specification makes it possible to identify which factors and interaction terms have a significant effect. Assume $\alpha_{il} \sim (1 - h_{il})\delta_0(\alpha_{il}) + h_{il}N(0, \omega)$ with $h_{il} \sim \text{Bernoulli}(q_{il})$ and $q_{il} \sim \text{Beta}(\gamma_1, \gamma_2)$.

The structure of the models indicated above is pretty similar to the factor model defined in Appendix B related to Chapter 3. These versions differ only in terms of how $\lambda$ is specified to include interaction terms, and the probability indexing the
Bernoulli distribution for $h_{il}$. As result, the expressions for the complete conditional posterior distributions of $\mu$, $\alpha_{il}$ and $\sigma_{i}^{2}$ are exactly the same as those determined for these parameters in Appendix B. In the remainder of the present Appendix, we derive the full conditionals of $q_{il}$, $h_{il}$ and $\lambda_{lj}$ in the factor model with interactions.

Define $h$ as the $(m \times L)$ matrix with $h_{il}$ in the $i$-th row and $l$-th column. We update $q_{il}$ as follows.

\[
p(q_{il} \mid h) \propto p(h_{il} \mid q_{il}) \propto q_{il}^{h_{il}}(1 - q_{il})^{1 - h_{il}} \quad q_{il}^{\gamma_{1} - 1}(1 - q_{il})^{\gamma_{2} - 1}
\]

The last expression is the kernel of $\text{Beta}(\gamma_{1} + h_{il}, \gamma_{2} + 1 - h_{il})$.

Denote $q$ as the $(m \times L)$ matrix with $q_{il}$ in row $i$ and column $j$. In order to update $h_{il}$, compute the conditional probability:

\[
p(h_{il} = 1 \mid \mu, \alpha, \lambda, \sigma_{i}^{2}, q, X) = \frac{q_{il}}{q_{il} + (1 - q_{il}) \frac{N[0|M_{\lambda}, V_{\alpha}]}{N[0|0, \omega]}}.
\]

where $N[0|M_{\alpha}, V_{\alpha}]$ is the Normal probability density evaluated at 0. The parameters $M_{\alpha}$ and $V_{\alpha}$ can be identified in the Appendix B.

Consider the notation: $\lambda_{-\{lj\}}$ is the set of all elements in $\lambda$, except $\lambda_{lj}$. In addition, let $\{\lambda_{lj}^{*}\}_{l}$ be the set of factor scores $\lambda_{lj}^{*}$, with $l^{*} \in \{n_{f} + 1, \cdots, L\}$, such that the corresponding product term involves $\lambda_{lj}$. For example, $\{\lambda_{lj}^{*}\}_{1} = \{\lambda_{4j}, \lambda_{5j}\}$ if $n_{f} = 3$.

Assume the approach 1 for $\lambda_{lj}$, introducing the product effect through the Gaussian prior. The complete conditional posterior distribution of $\lambda_{lj}$, for $l = 1, \cdots, n_{f}$, is given by the following application of the Bayes theorem.

\[
p(\lambda_{lj} \mid \mu, \alpha, \lambda_{-\{lj\}}, \sigma^{2}, X) \propto p(X, \{\lambda_{lj}^{*}\}_{l} \mid \mu, \alpha, \lambda_{-\{lj\}}, \sigma^{2}) \quad p(\lambda_{lj} \mid \lambda_{-\{lj\}})
\]

In order to simplify the notation consider the particular case $n_{f} = 3$. The adjustment of the notation is straightforward for any case with $n_{f} > 3$. Note that,
\[ \lambda_{ij} \sim N(\lambda_{1j}\lambda_{2j}, \nu), \] 
\[ \lambda_{5j} \sim N(\lambda_{1j}\lambda_{3j}, \nu) \] 
and 
\[ \lambda_{6j} \sim N(\lambda_{2j}\lambda_{3j}, \nu). \]
Without loss of generality let \( l = 1 \). The target full conditional is given by
\[
p(\lambda_{1j} | \mu, \alpha, \lambda_{-\{1j\}}, \sigma^2, X) \propto p(X | \mu, \alpha, \lambda, \sigma^2) p(\lambda_{4j} | \lambda_{1j}, \lambda_{2j}) p(\lambda_{5j} | \lambda_{1j}, \lambda_{3j}) p(\lambda_{1j})
\]
\[
\propto \exp \left\{ -\frac{1}{2} [\lambda_{ij}^2 R_1 - 2\lambda_{ij} R_2] \right\},
\]
where \( R_1 = \left(1 + \sum_{i=1}^m \frac{\alpha_i^2}{\sigma_i^2} + \frac{\lambda_{ij}^2}{\nu} + \frac{\lambda_{ij}^2}{\nu} \right) \)
and
\[
R_2 = \sum_{i=1}^m \frac{(X_{ij} - \mu_i)\alpha_{il}}{\sigma_i^2} - \sum_{i=1}^m \left( \frac{\alpha_i}{\sigma_i^2} \sum_{l \neq 1} \lambda_{ij} \alpha_{il} \right) + \frac{\lambda_{ij}^2}{\nu} + \frac{\lambda_{ij}^2}{\nu}.
\]
The kernel of \( N(M_\lambda, V_\lambda) \) with \( V_\lambda = R_1^{-1} \) and \( M_\lambda = V_\lambda R_2 \) can be identified above.

For \( l = n_f + 1, \cdots, L \) the complete conditional posterior distribution of \( \lambda_{ij} \) is \( N(R_1^{-1} R_2, R_1^{-1}) \) with
\[
R_1 = \left( \frac{1}{\nu} + \sum_{i=1}^m \frac{\alpha_i^2}{\sigma_i^2} \right) \text{ and } R_2 = \frac{\lambda_{1j} \lambda_{2j}}{\nu} + \sum_{i=1}^m \frac{(X_{ij} - \mu_i)\alpha_{il}}{\sigma_i^2} - \sum_{i=1}^m \frac{\alpha_i^2}{\sigma_i^2} \sum_{l \neq l} \lambda_{ij} \alpha_{il} \text{.}
\]
where \( l_1 < l_2 \in \{1, 2, \cdots, n_f\} \) are the indices of the factors scores forming the product interaction term related to \( l = n_f + 1, \cdots, L \).

Consider the approach 2 for \( \lambda_{ij} \), and set \( n_f = 2 \) for simplicity. Here, the prior \( N(0, 1) \) is specified for \( \lambda_{1,j} \) and \( \lambda_{2,j} \). Given these parameters, we set \( \lambda_{3,j} = \lambda_{1,j} \lambda_{2,j} \).
In other words, \( \lambda_{1,j} \) and \( \lambda_{2,j} \) are the target latent factors in the model, and \( \lambda_{3,j} \) is treated as if it were an observed variable. When \( \lambda_{ij} \) or \( \lambda_{2j} \) is generated from its full conditional in any MCMC iteration, the factor score \( \lambda_{3j} \) is corrected to represent the product between latent factors. According to this approach, the complete conditional posterior distribution of \( \lambda_{ij}, l = 1, \cdots, n_f, \) is the \( N(R_1^{-1} R_2, R_1^{-1}) \) as indicated below.
\[
p(\lambda_{ij} | \mu, \alpha, \lambda_{-\{ij\}}, \sigma^2, X) \propto p(X | \mu, \alpha, \lambda, \sigma^2) p(\lambda_{ij}) \propto \exp \left\{ -\frac{1}{2} [\lambda_{ij}^2 R_1 - 2\lambda_{ij} R_2] \right\},
\]
where \( R_1 = \left(1 + \sum_{i=1}^m \frac{\alpha_i^2}{\sigma_i^2} \right) \) and \( R_2 = \sum_{i=1}^m \frac{(X_{ij} - \mu_i)\alpha_{il}}{\sigma_i^2} - \sum_{i=1}^m \frac{\alpha_i^2}{\sigma_i^2} \sum_{l \neq l} \lambda_{ij} \alpha_{il} \text{.}
\]
132
D.2 Additional results for the simulated study

The results presented in Section 5.2 are related to a synthetic data set generated with \( n_f = 2 \) factors. Here, we present additional graphs showing the behavior of the factor model for a data set generated with \( n_f = 3 \).

**Figure D.1**: Synthetic data with multiplicative interaction effects \((n_f = 3)\). The original data is displayed in the top. The second panel shows \( X \) with standardized rows (rows and columns are sorted so that the first principal component is monotone).

**Figure D.2**: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for \( \alpha_{il} \). Dashed lines separate the factors. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).
Figure D.3: Images displaying the posterior mean and the true loading associated with the interaction factor $l = 4$ (row 1), $l = 5$ (row 2) and $l = 6$ (row 3). Box plots representing the distribution of the conditional probability that $h_{il} = 1$ for $l = 4, 5, 6$. Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).
Consider the two versions of the factor model defined in Section 5.1. Panel (a) shows results for the approach 1 (Gaussian prior), and Panel (b) is related to approach 2 (perfect product).

**Figure D.4:** Scatter plots comparing the posterior estimates with the true values.
D.3 Additional results for the a real data analysis

In the next three figures, each panel represents a different breast cancer data set: Panel (a) for Chin et al. (2006), (b) for Miller et al. (2005), (c) for Sotiriou et al. (2006), and (d) for Wang et al. (2005). The credible intervals in Figure D.6 are computed for the component in the posterior mixture with highest probability weight. Assume the factor model with approach 2 (perfect product) in Figures D.6 and D.7.

![Figure D.5: Image of the matrix X containing the data. Rows and columns are sorted so that the first principal component is monotone (rows are standardized).](image)

136
Figure D.6: Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{33}$.

Figure D.7: Posterior mean (x mark) and the 95% credible interval (bar) for $\lambda$. 137
Appendix E

Linear factor model to select genes in the CNA problem

In Chapters 5 and 6, we develop applications related to the Copy Number Alteration (CNA) problem, and involving real data sets with a large number of genes. The MCMC algorithms, implemented for the models proposed in those chapters, can be rather slow to handle the large real data sets. As a solution for this computational difficulty, the linear factor model indicated in this Appendix can be used to select the most interesting genes for those applications. For simplicity, the model is defined without interaction between factors.

This Appendix is divided in three sections. In Section E.1, we describe the factor model, and specify the priors for the involved parameters. The derivation of full conditional posterior distributions is also indicated here. Section E.2 shows some results from a simulated data analysis. Finally, Section E.3 presents a real data application to select genes.
E.1 The two-factor model

Assume the model $X = \alpha \lambda + \epsilon$, where $\alpha$ is the $(m \times L)$ matrix of factor loadings, $\lambda$ is the $(L \times n)$ matrix of factor scores, and $\epsilon$ is the $(m \times n)$ matrix of noise terms. Let $\epsilon_{ij} \sim N(0, \sigma_i^2)$ and denote $\sigma^2 = (\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2)'$. Note that $m$ is the number of genes and $n$ is the sample size. The real data analyses in Chapter 5 and 6 consider two latent factors to investigate interaction effects related to the CNA; for this reason, we set $L = 2$ factors in the current configuration. Note that the mean expression parameter $\mu$ is not included in the model. We intend to fit large data sets; therefore, a parsimonious model would be preferred for this task. If we standardize the rows of $X$, the mean expression can be ignored because it is a value close to zero. Consider this strategy in any application of the proposed model.

In terms of prior specifications assume $\lambda_j \sim N_L(0, I_L)$, $\sigma_i^2 \sim IG(a, b)$ and the spike and slab mixture prior for the factor loadings $\alpha_{il} \sim (1 - h_{il}) \delta_0(\alpha_{il}) + h_{il} N(0, \omega)$. The Bernoulli($q_{il}$) is specified for the indicator variable $h_{il}$, and we set $q_{il} \sim Beta(\gamma_1, \gamma_2)$.

The joint posterior distribution $p(\alpha, \lambda, \sigma^2 | X)$ cannot be analytically evaluated. The Gibbs sampler is implemented to sample from this target probability distribution. The structure of the model above is very similar to the version explored in Appendix B. Let $\mu = 0$ for that case, and consider the likelihood function presented in that section. The derivation of the complete conditional posterior distribution for $\alpha_{il}$, $\lambda_j$ and $\sigma_i^2$ is exactly the same as shown in Appendix B. The mixture prior above specifies $q_{il}$ for each gene $i$ and factor $l$, which is the same assumption defined in Appendix D; therefore, consider $(q_{il}|h) \sim Beta(\gamma_1 + h_{il}, \gamma_2 + 1 - h_{il})$ and the complete conditional probability for $h_{il}$ indicated in Section D.1.

Note that $q_{il}$ is updated based on a single observation $h_{il}$; however, this result is convenient for applications involving large data set. In order to explain this fact, con-
sider $h_{il} \sim \text{Bernoulli}(q_R)$ with $R \in \{R_1, R_2, R_3\}$ and $q_R \sim \text{Beta}(\gamma_{1,R}, \gamma_{2,R})$. Instead of defining a probability $q_{il}$ for each gene $i$ and factor $l$, the suggested version set $R = R_1$ if we suspect that gene $i$ and factor $l$ are associated. Additionally, $R = R_2$ represents no association, and $R = R_3$ is used when the relationship is unknown. In other words, the same $q_R$ is indicated for a group of $h_{il}$, and thus more than one value are used to update $q_R$. This strategy can work pretty well for small data sets, but it can be a problem for cases where $m$ is big. The parameter $q_{R_3}$ is specified for a large number of genes (the cases with unknown association gene-factor), and we expect that most factor loadings are identified as non-zero values. As a result, $q_{R_3}$ is estimated as a large probability, which increases the conditional probability $p(h_{il} = 1 | \cdots)$ for the whole group of $(i, l)$ related to the specification $R_3$. In summary, our goal is to remove from the analysis the genes indicating low probability $p(h_{il} = 1 | \cdots)$, and the model fit using $q_R$ determines high $p(h_{il} = 1 | \cdots)$ for almost all cases. The current specification, $h_{il} \sim \text{Bernoulli}(q_{il})$, is less affected by the behavior of other genes, and it provides useful results for the gene selection. Note that, we can still take into account the groups $(R_1, R_2, R_3)$ to specify Beta distributions for $q_{il}$; this can be used to address the identifiability issues.

E.2 Synthetic data application

In this section, we investigate the performance of the proposed model using simulated data. Assume $\omega = 10$ in the mixture prior for $\alpha_{il}$, and $\sigma_i^2 \sim IG(a = 2,1, b = 1,1)$. In the CNA study, we define $G_l$ as a known list of genes associated with factor $l \in \{1,2\}$. $G_1$ and $G_2$ are disjoint, and factor $l$ is related to some specific region in the genome. According to this information, we can set the following priors for $q_{il}$:

- Beta(2,1) for $(i \in G_1, l = 1)$ and $(i \in G_2, l = 2)$;
- Beta(1,2) for $(i \in G_1, l = 2)$ and $(i \in G_2, l = 1)$;
- Beta(1,1) for $i \in G_E = (G_1 \cup G_2)^C, l \in \{1,2\}$. 

140
In terms of initial values for the chains, consider $\alpha_{il}^{(0)} = 0$, $\lambda_{lj}^{(0)} \sim N(0, 1)$, $\sigma_i^2 = 1$, and $h_{il}^{(0)} \sim \text{Bernoulli}(q_{il}^{(0)})$ where $q_{il}^{(0)}$ is

- 0.9999 for $(i \in G_1, l = 1)$ and $(i \in G_2, l = 2)$;
- 0.0001 for $(i \in G_1, l = 2)$ and $(i \in G_2, l = 1)$;
- 0.5 for $i \in G_E = (G_1 \cup G_2)^C, l \in \{1, 2\}$).

The MCMC algorithm is set to run 2000 iterations and the first 1000 are assumed as the burn-in period. Fast convergence is observed for all chains in any application.

The matrix $X$ will be generated based on a real data set. We choose two groups of genes: $G_1$ (15 genes) is located around the position 112,006,088 in the chromosome 6, and $G_2$ (35 genes) is located around the position 204,828,862 in the chromosome 1. All the genes in each group are located within the distance 2,000,000 to the central position. No other gene is included in $X$, i.e., $m = 50$ and $G_E$ is an empty set. We use the breast cancer data set indicated in Miller et al. (2005) which contains $n = 251$ samples. The observation $X_{ij}$ is the RMA output for gene $i$ and sample $j$; in addition, the rows of $X$ are standardized to remove the mean expression. The two-factor model is fitted to this data using the configuration above for priors and initial values. The posterior mean of $\alpha$, $\lambda$ and $\sigma^2$ are assumed as real values. In the final step, we generate $\epsilon_{ij} \sim N(0, \sigma_i^2)$ and compute $X = \alpha \lambda + \epsilon$ to obtain the synthetic observations.

The selection procedure, defining the interval $[2,000,000 - P, 2,000,000 + P]$ around the position $P$ detected with CNA, can be inefficient by including unrelated genes. This issue is discussed with more details in the next section. The estimated $\alpha$ matrix is expected to display the configuration: all elements in each set $\{\alpha_{i1} : i \in G_1\}$ and $\{\alpha_{i2} : i \in G_2\}$ should have the same, and the remaining cases should be zero. In the model fit described above to generate $X$, we observe some genes violating this configuration. These might be the cases unrelated to the CNA in the studied region.
In order to correct the problem, we remove these “bad” genes from the \((G_1 \cup G_2)\) list and assume them as members of \(G_E\) in the simulated data analysis. As a result, \(G_1\) now has 5 genes, \(G_2\) has 20, and \(G_E\) contains 25 elements.

\[
(G_1 \cup G_2)
\]

\[
G_E
\]

\[
G_1 \text{ now has 5 genes, } G_2 \text{ has 20, and } G_E \text{ contains 25 elements.}
\]

\((a)\) \((b)\) \((c)\) \((d)\)

**Figure E.1:** Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar). Panel (a) shows only the factor loadings for \(i \in (G_1 \cup G_2)\). Panel (b) shows all \(\alpha_{il}\). Panels (c) and (d) are related to \(\lambda\) and \(\sigma^2\), respectively. Intervals for \(\alpha_{il}\) are computed for the component with highest probability weight \textit{a posteriori}.

Figure E.1 presents the 95% credible intervals and compares the posterior estimates with the corresponding real values. As can be seen, most true values are located inside the 95% intervals suggesting a good performance of the model to fit the data. Panel (a) shows the factor loadings related to the genes in \(G_1\) and \(G_2\); the dashed line separates the groups. Note that the model fit satisfies our assumption regarding the relationship between these genes and the latent factors. Panel (b) presents results for all \(\alpha_{il}\).
E.3 Real application: gene selection

In this section, we apply the two-factor model, proposed in Section E.1, to select
genes for the applications involving the models with interaction effects. As indicated
in the previous section, the gene lists $G_1$ and $G_2$ may contain some genes unrelated
to the CNA detected in their corresponding genome region. The first step in the
selection procedure involves the identification and removal of the problematic genes.
In brief, we fit the two-factor model to the matrix $X$ containing only the genes in
$(G_1 \cup G_2)$. In terms of prior specifications and initial values, consider the choices
presented in Section E.2. The estimated factor loadings matrix is expected to show
$\{\alpha_{i1} : i \in G_1\}$ with the same sign, $\{\alpha_{i2} : i \in G_2\}$ with the same sign, and $\alpha_{il} = 0$
for all other cases. In order to verify the sign requirement for $\{\alpha_{i1} : i \in G_1\}$ or
$\{\alpha_{i2} : i \in G_2\}$, we assume the most common sign in that set as the true one, and
any gene showing the opposite sign is considered problematic. In conclusion, the
final result is the updated $G_1$ and $G_2$ lists containing only the genes satisfying our
assumptions for $\alpha$.

Table E.1: Regions detected with CNA in the human genome. Number of genes
before and after the cleaning procedure.

<table>
<thead>
<tr>
<th>Region</th>
<th>Chr.</th>
<th>Position</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Before</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>117,844,879</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>35,152,961</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>101,400,207</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>68,771,985</td>
<td>42</td>
</tr>
</tbody>
</table>

Section E.2 shows an application of the “cleaning” procedure for two groups of
genes selected in regions of chromosome 1 and 6. In the next application, we will
consider other three pairs of regions. Each pair combines two of the positions defined
in Table E.1. In addition, we will use the breast cancer data set ($n = 118$ samples)
indicated in Chin et al. (2006). Table E.1 presents the number of genes before and after the application of the cleaning procedure. As can be seen, the smallest list contains 13 genes.

If we apply the cleaning procedure to evaluate pairs of regions in other breast cancer data sets (Miller et al. 2005 and Wang et al. 2005), the intersection of \((G_1 \cup G_2)\) from two data sets has at least 5 genes in \(G_1\) or \(G_2\) for the pairs \((1,4), (2,4)\) and \((3,4)\). For this reason, we choose to investigate the results for these combinations of regions. In any application, assume \(G_1\) and \(G_2\) as the gene lists after the cleaning procedure. The next step of the analysis is the selection of genes via the two-factor model proposed in this Appendix. The model is fitted to the \((22283 \times 118)\) matrix \(X\) representing the expression across samples for all available genes. This is a real data application, involving a large data set; therefore, we will assume stronger prior specifications for \(q_{il}\) to assure the identifiability. Consider:

- Beta(10,1) for \((i \in G_1, l = 1)\) and \((i \in G_2, l = 2)\);
- Beta(1,10) for \((i \in G_1, l = 2)\) and \((i \in G_2, l = 1)\);
- Beta(1,5) for \(i \in G_E = (G_1 \cup G_2)^C, l \in \{1,2\}\).

In addition, consider the same initial values suggested in Section E.2 for \(q_{il}\). The only difference is the choice 0.1 for \(i \in G_E = (G_1 \cup G_2)^C\) and \(l \in \{1,2\}\), i.e., low initial probability to favor sparsity for the cases with unknown association gene-factor. The priors for other parameters are exactly the same as in Section E.2.

Figure E.2 shows several bars containing percentiles (25th, 50th, 75th), the minimum and maximum values related to the posterior distribution of the conditional probability that \(h_{il} = 1\). Each bar is associated with a pair \((i, l)\) in the factor loadings matrix, and they are ordered with respect to the median. We have strong evidence to support \(\alpha_{il} \neq 0\), if the corresponding bar indicates low variability and is located close to 1. In this case, we can say that the gene \(i\) is affected by factor \(l\). In order to
select interesting genes for the study of interaction effects, we impose two requirements. Our posterior estimate is $\alpha_{il} \neq 0$ if the mean of $p(h_{il} = 1|\cdots)$ is larger than 0.5. In the gene selection, we will assume a stronger requirement to accept $\alpha_{il} \neq 0$; the 25th percentile must be above 0.5. The second requirement is that the gene $i$ must be affected by both factors. It seems reasonable to assume that genes affected by interactions tend to be associated with both factors. The genes in $G_1$ and $G_2$ are associated with only one factor; therefore, they are not selected. As a result of the selection process, the group $G_E$ includes 3717 genes (factors 1 & 4), 3704 genes (factors 2 & 4) and 3708 genes (factors 3 & 4).

![Figure E.2: Distribution of the conditional probability that $h_{il} = 1$. Each bar contains the 25th, 50th and 75th percentiles. Panels (a), (b) and (c) represent the pairs (1,4), (2,4) and (3,4), respectively.](image)
Appendix F

Posterior computation and additional graphs for Chapter 6

This Appendix is divided in four sections. Section F.1 shows the likelihood function and the derivation of complete conditional posterior distributions for the parameters in the factor model with non-linear structure of interactions. Section F.2 presents additional graphs for the simulated study developed in Section 6.2. The extra results in Section F.3 are related to the real data analysis developed in Section 6.3. Finally, Section F.4 presents extra graphs to complement the comparison analysis in Section 6.4.

F.1 Likelihood function and complete conditional posterior distributions

Assume the model \( X = \alpha \lambda + F + \epsilon \), where \( \alpha \) is the \((m \times L)\) matrix of factor loadings, \( \lambda \) is the \((L \times n)\) matrix of factor scores, \( F \) is the \((m \times n)\) matrix of interaction effects, and \( \epsilon \) is the \((m \times n)\) matrix of idiosyncratic noise terms. Note that we define a model with \( L \) factors, \( m \) genes and \( n \) samples. Let \( \epsilon_{ij} \sim N(0, \sigma_i^2) \) and denote
\[ \sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)' . \]

We assume that the observations \( X_{ij} \) are conditionally independent given the values of parameters specified in the model. The likelihood function can be written in two different forms:

- **Likelihood 1:** Define \( X_j \) and \( F_j \) as the \( m \)-dimensional column vectors representing the \( j \)-th column of \( X \) and \( F \). In addition, \( \lambda_j \) (\( L \)-dimensional) is the \( j \)-th column of \( \lambda \). Note that \( X_j \sim N_m [\alpha \lambda_j + F_j, D] \) with \( D = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2) \).

As a result, we can express
\[
p(X | \alpha, \lambda, F, \sigma^2) = \prod_{j=1}^{n} p(X_j | \alpha, \lambda, F, \sigma^2)
\]
\[
= \prod_{j=1}^{n} (2\pi)^{-\frac{m}{2}} |D|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left[ (X_j - \alpha \lambda_j - F_j)'D^{-1}(X_j - \alpha \lambda_j - F_j) \right] \right\}
\]
\[
= (2\pi)^{-\frac{mn}{2}} |D|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} \left[ X_j'D^{-1}X_j - 2F_j'D^{-1}X_j - 2\lambda_j\alpha'D^{-1}X_j + F_j'D^{-1}F_j + \lambda_j\alpha'D^{-1}\lambda_j + 2F_j'D^{-1}\lambda_j \right] \right\} .
\]

- **Likelihood 2:** Denote \( X_i \) and \( F_i \) as the \( n \)-dimensional row vectors representing the \( i \)-th row of \( X \) and \( F \). Also, let \( \alpha_i \) (\( L \)-dimensional) be the \( i \)-th row of \( \alpha \). Note that \( X_i' \sim N_n [\lambda\alpha_i' + F_i', \sigma_i^2 I_n] \). In this case, we can write
\[
p(X | \alpha, \lambda, F, \sigma^2) = \prod_{i=1}^{m} p(X_i | \alpha, \lambda, F, \sigma^2) = \prod_{i=1}^{m} \left[ (2\pi)^{-\frac{n}{2}} |\sigma_i^2 I_n|^{-\frac{1}{2}} \right]
\]
\[
\exp \left\{ -\frac{1}{2} \left[ (X_i' - \lambda\alpha_i' - F_i')'(\sigma_i^2 I_n)^{-1}(X_i' - \lambda\alpha_i' - F_i') \right] \right\}
\]
\[
= (2\pi)^{-\frac{mn}{2}} \left( \prod_{i=1}^{m} \sigma_i^2 \right)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{1}{\sigma_i^2} \left[ X_iX_i' - 2F_iX_i' - 2\alpha_i\lambda X_i' + F_iF_i' + \alpha_i\lambda\alpha_i' + 2F_i\lambda\alpha_i' \right] \right\} .
\]
These expressions for the likelihood function are equivalent. For each parameter, one of them simplifies the derivation of the corresponding full conditional posterior distribution.

Consider the notation \( \sigma_{i-1}^2 = \{\sigma_1^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \ldots, \sigma_m^2\} \), and set \( \text{IG}(a,b) \) as the prior distribution for \( \sigma_i^2 \). The Bayes theorem using the Likelihood version 2 provides the conditional distribution indicated below.

\[
p(\sigma_i^2 | \alpha, \lambda, F, \sigma_{-i}^2, X) \propto p(X|\alpha, \lambda, F, \sigma_i^2) p(\sigma_i^2)
\]

\[
\propto (\sigma_i^2)^{-\frac{n}{2}} \exp\left\{ -\frac{B}{\sigma_i^2} \right\} (\sigma_i^2)^{-(a+1)} \exp\left\{ -\frac{b}{\sigma_i^2} \right\} = (\sigma_i^2)^{-(a+\frac{n}{2}+1)} \exp\left\{ -\frac{(B+b)}{\sigma_i^2} \right\}.
\]

where \( B = \frac{1}{2} [X_i'X_i - 2\alpha_i\lambda(X_i' - F_i') - 2F_iX_i' + F_iF_i' + \alpha_i\lambda\lambda'] \). The \( \text{IG}(a + \frac{n}{2}, B + b) \) can be identified as the posterior distribution in this case.

We propose different versions of the factor model used to evaluate non-linear interactions between factors. These versions differ in terms of prior specifications in the hierarchical structure. The remainder of this Section is divided in three subsections presenting the prior specifications and the derivation of full conditionals for a set of parameters. Subsections F.1.1, F.1.2 and F.1.3 show the results related to \( \alpha_{il}, F \) and \( \lambda \), respectively.

**F.1.1 Prior and complete conditional for \( \alpha_{il} \) and its associated parameters**

Consider the mixture prior: \( \alpha_{il} \sim (1-h_{il})\delta_0(\alpha_{il}) + h_{il}N(0,\omega) \) where \( h_{il} \) is an indicator variable. In this section, we explore two different forms of expressing our prior uncertainty for the probability that \( h_{il} = 1 \). The specifications are:

1. \( h_{il} \sim \text{Bernoulli}(q_{il}) \) and \( q_{il} \sim \text{Beta}(\gamma_1, \gamma_2) \);
2. \( h_{il} \sim \text{Bernoulli}(q_R), R \in \{R_1, R_2, R_3\}, \) and \( q_R \sim \text{Beta}(\gamma_{1,R}, \gamma_{2,R}) \). We set \( R = R_1 \) if we suspect that gene \( i \) and factor \( l \) are associated. Additionally, \( R = R_2 \) represents no association, and \( R = R_3 \) is used for unknown relationships.
The part of Likelihood 2 depending on $\alpha_{il}$ can be written as:

$$p(X \mid \alpha, \lambda, F, \sigma^2) \propto$$

$$\propto \exp \left\{ -\frac{1}{2\sigma_i^2} \left[ \alpha_{il}^2 \sum_{j=1}^{n} \lambda_{lj}^2 - 2\alpha_{il} \sum_{j=1}^{n} \lambda_{lj} (X_{ij} - F_{ij}) + 2\alpha_{il} \sum_{j=1}^{n} \sum_{l^* \neq l} (\alpha_{il^*} \lambda_{lj} \lambda_{l^*j}) \right] \right\}$$

Denote $\alpha_{-\{il\}}$ as the set of all loadings in $\alpha$, except $\alpha_{il}$. In addition, define $h$ as the $(m \times L)$ matrix with $h_{il}$ in the $i$-th row and $l$-th column. Note that $(\alpha_{il} | h_{il} = 1) \sim N(0, \omega)$. As a result, the complete conditional posterior distribution is given by

$$p(\alpha_{il} \mid \alpha_{-\{il\}}, \lambda, F, \sigma^2, h, X) \propto p(X \mid \alpha, \lambda, F, \sigma^2) \ p(\alpha_{il} \mid h_{il} = 1)$$

$$\propto \exp \left\{ -\frac{1}{2} \left[ \alpha_{il}^2 \left( \frac{1}{\omega} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n} \lambda_{lj}^2 \right) - 2\frac{\alpha_{il}}{\sigma_i^2} \left( \sum_{j=1}^{n} \lambda_{lj} [X_{ij} - F_{ij} - \sum_{l^* \neq l} \alpha_{il^*} \lambda_{l^*j}] \right) \right] \right\}.$$  

The last expression is the kernel of $N(M_\alpha, V_\alpha)$ with $V_\alpha = \left[ \frac{1}{\omega} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n} \lambda_{lj}^2 \right]^{-1}$ and $M_\alpha = V_\alpha \left[ \frac{1}{\sigma_i^2} \sum_{j=1}^{n} \lambda_{lj} (X_{ij} - F_{ij} - \sum_{l^* \neq l} \alpha_{il^*} \lambda_{l^*j}) \right]$. If $h_{il} = 0$, the complete conditional posterior distribution will be $\delta_0(\alpha_{il})$.

Let $q$ be the $(m \times L)$ matrix of probabilities $q_{il}$. If we assume the prior specification 1 for $h_{il}$, we will update this indicator using the conditional probability

$$p(h_{il} = 1 \mid \alpha, \lambda, F, \sigma^2, q, X) = \frac{q_{il}}{q_{il} + (1 - q_{il}) \frac{N[y|M_\alpha, V_\alpha]}{N[0, \omega]}}.$$  

where $N[y|M, V]$ is the $N(M, V)$ density function evaluated at $y$.

Also, under specification 1, $q_{il}$ is updated using the Beta$(\gamma_1 + h_{il}, \gamma_2 + 1 - h_{il})$ according to the result: $p(q_{il} \mid h_{il}) \propto p(h_{il} \mid q_{il})p(q_{il}) \propto q_{il}^{h_{il}}(1 - q_{il})^{1 - h_{il}} q_{il}^{-\gamma_1 - 1}(1 - q_{il})^{\gamma_2 - 1} \propto q_{il}^{\gamma_1 + h_{il} - 1}(1 - q_{il})^{\gamma_2 + 1 - h_{il} - 1}$.

The prior specification 2 provides the following conditional probability to update the indicator $h_{il}$.

$$p(h_{il} = 1 \mid \alpha, \lambda, F, \sigma^2, q_R, X) = \frac{q_{R}}{q_{R} + (1 - q_{R}) \frac{N[y|M_\alpha, V_\alpha]}{N[0, \omega]}}.$$  

149
Recall that $R = \{R_1, R_2, R_3\}$ depending on our prior knowledge about the relationship between gene $i$ and factor $l$. Denote $\{R\}$ as the set of pairs $(i,j)$ associated with $q_R$, and $|\{R\}|$ is the cardinality of $\{R\}$. The probability $q_R$ is updated using the Beta($\gamma_1 + \sum_{(i,l) \in \{R\}} h_{il}, \gamma_2 + |\{R\}| - \sum_{(i,l) \in \{R\}} h_{il}$) as shown below.

$$p(q_R \mid h) \propto p(h_{\{R\}} \mid q_R) p(q_R)$$

$$\propto q_R^{\gamma_1 + \sum_{(i,l) \in \{R\}} h_{il}}(1 - q_R)^{\sum_{(i,l) \in \{R\}} (1 - h_{il})} d_R (\gamma_2, R - 1) - (\sum_{(i,l) \in \{R\}} h_{il}) - 1$$

\[ F.1.2 \text{ Prior and complete conditional for } F \text{ and its associated parameters} \]

A mixture prior, with a degenerated distribution at 0, is defined for the interaction effect vector $F_i$. Two versions are explored, the first one assumes that $F_i$ can be different comparing affected genes. The second version is less realistic, and assumes that $F_i$ is the same for all gene $i$ affected by interactions.

We examine the version 1 first. Consider the prior distribution: $(F_i' \mid \lambda) \sim (1 - z_i)\delta_0(F_i') + z_i N_n[0, K(\lambda)]$ where $z_i$ is an indicator variable, and $K(\lambda)$ is the covariance matrix based on the Squared Exponential covariance function depending on $\lambda$.

$$K(\lambda)_{j_1,j_2} = \exp \left\{ -\frac{1}{2l_s^2} ||\lambda_{j_1} - \lambda_{j_2}||^2 \right\},$$

where $(j_1, j_2) \in \{1,2,\ldots,n\}$, $l_s$ is the characteristic length-scale and $||y||$ represents the Euclidean norm of the vector $y$.

In the same way as the previous section, we consider different strategies to express our prior knowledge about the indicator $z_i$. Assume the following possibilities:

1. $z_i \sim \text{Bernoulli}(\rho_i)$ and $\rho_i \sim \text{Beta}(\beta_1, \beta_2)$;
2. $z_i \sim \text{Bernoulli}(\rho)$ and $\rho \sim \text{Beta}(\beta_1, \beta_2)$;
3. $z_i \sim \text{Bernoulli}(\rho_R)$, $R \in \{R_1, R_2\}$ and $\rho_R \sim \text{Beta}(\beta_{1,R}, \beta_{2,R})$. In this case, $R = R_1$ if we believe that gene $i$ is associated with some factor and interactions.
are not affecting this gene. On the other hand, set \( R = R_2 \) if the association between gene \( i \) and any factor is unknown; interaction effects may exist here.

The part of Likelihood 2 depending on \( F_i \) is presented below:

\[
p(X \mid \alpha, \lambda, F, \sigma^2) \propto \exp \left\{ -\frac{1}{2\sigma_i^2} \left[ F_i F_i' - 2F_i (X_i' - \lambda' \alpha_i') \right] \right\}.
\]

Note that \((F_i \mid \lambda, z_i = 1) \sim N(0, K(\lambda))\); as a result, the complete conditional posterior distribution is given by

\[
p(F_i \mid \alpha, \lambda, \sigma^2, z_i = 1, X) \propto p(X \mid \alpha, \lambda, F_i, \sigma^2) p(F_i \mid \lambda, z_i = 1)
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left[ F_i \left( \frac{1}{\sigma_i^2} I_n + K(\lambda)^{-1} \right) F_i' - 2F_i \frac{1}{\sigma_i^2} (X_i' - \lambda' \alpha_i') \right] \right\}.
\]

The last expression can be recognized as the kernel of \( N(M_F, V_F) \) with parameters 
\( V_F = \left[ \frac{1}{\sigma_i^2} I_n + K(\lambda)^{-1} \right]^{-1} \) and 
\( M_F = V_F \left[ \frac{1}{\sigma_i^2} (X_i' - \lambda' \alpha_i') \right] \). If \( z_i = 0 \), the complete conditional posterior distribution will be \( \delta_0(F_i) \).

The entries of the covariance matrix \( K(\lambda) \) are non-linear functions of elements in \( \lambda \). Note that \( V_F \) depends on \( \lambda \) via \( K(\lambda) \), and the expression of the posterior mean \( M_F \) depends on \( V_F \). In summary, \( F_i \) has a non-linear relationship with \( \lambda \); for this reason we designate the structure of the model as non-linear.

The expression of the conditional probability \( p(z_i = 1 \mid \cdots) \) depends on the prior specifications enumerated above. The case number 1 provides the following result

\[
p(z_i = 1 \mid \alpha, \lambda, F, \sigma^2, \rho_i, X) = \frac{\rho_i}{\rho_i + (1 - \rho_i) \frac{\text{N}(0, M_F, V_F)}{\text{N}(0, 0, K(\lambda))}}.
\]

In addition, \( \rho_i \) is updated via Beta\((\beta_1 + z_i, \beta_2 + 1 - z_i)\) as shows the result: \( p(\rho_i \mid z_i) \propto p(z_i \mid \rho_i) p(\rho_i) \propto \rho_i^{z_i} (1 - \rho_i)^{1-z_i} \rho_i^{\beta_1-1} (1 - \rho_i)^{\beta_2-1} \propto \rho_i^{\beta_1+z_i-1} (1 - \rho_i)^{\beta_2+1-z_i-1} \).

If we assume the prior specification 2, the conditional probability to update \( z_i \) is

\[
p(z_i = 1 \mid \alpha, \lambda, F, \sigma^2, \rho, X) = \frac{\rho}{\rho + (1 - \rho) \frac{\text{N}(0, M_F, V_F)}{\text{N}(0, 0, K(\lambda))}}.
\]
Let \( z = \{z_1, ..., z_m\} \), the probability \( \rho \) is updated via \( \text{Beta}(\beta_1 + \sum_{i=1}^{m} z_i, \beta_2 + m - \sum_{i=1}^{m} z_i) \) as indicated below.

\[
p(\rho \mid z) \propto p(z \mid \rho) p(\rho) \propto \left[ \prod_{i=1}^{m} \rho^{z_i}(1-\rho)^{1-z_i} \right] \rho^{\beta_1-1}(1-\rho)^{\beta_2-1}
\]

\[
\propto \rho^{\beta_1 + \sum_{i=1}^{m} z_i - 1} (1-\rho)^{\beta_2 + m - \sum_{i=1}^{m} z_i - 1}.
\]

Finally, consider the prior specification 3 and denote: \( \{R\} \) is the set of indices \( i \) representing genes for which we assume the probability \( \rho_R \), and \(|\{R\}|\) is the cardinality of \( \{R\} \). Recall that we can have \( R = R_1 \) or \( R_2 \). The conditional probability to update \( z_i \) takes the form

\[
p(z_i = 1 \mid \alpha, \lambda, F, \sigma^2, \rho_R, X) = \frac{\rho_R}{\rho_R + (1-\rho_R) \frac{N[0|M_F, V_F]}{N[0, K(\lambda)]}}.
\]

The complete conditional distribution for \( \rho_R \) is derived below.

\[
p(\rho_R \mid z) \propto p(z \mid \rho_R) p(\rho_R) \propto \rho_R^{\sum_{i \in \{R\}} z_i (1-\rho_R)} \rho_R^{\beta_1 + \sum_{i \in \{R\}} (1-\rho_R)} \rho_R^{\beta_2, R - 1} (1-\rho_R)^{\beta_2, R - 1}
\]

\[
\propto \rho_R^{\beta_1, R + \sum_{i \in \{R\}} z_i - 1} (1-\rho_R)^{\beta_2, R + |\{R\}| - \sum_{i \in \{R\}} z_i - 1}.
\]

This expression is the kernel of \( \text{Beta}(\beta_1, R + \sum_{i \in \{R\}} z_i, \beta_2, 2 + |\{R\}| - \sum_{i \in \{R\}} z_i) \).

Now, we derive the full conditional distributions for the model assuming the second version of the mixture prior for \( F_i \). In this case, we assume that the interaction effect is the same for all affected genes. Consider the prior: \( (F'_i \mid F^*) \sim (1-z_i)\delta_0(F'_i) + z_i\delta_{F^*}(F'_i) \) and \( (F^* \mid \lambda) \sim N_n[0, K(\lambda)] \). Once again, \( z_i \) is an indicator variable, and we use different strategies to express our prior uncertainty about the probability that \( z_i = 1 \). In this case, we consider the strategies enumerated as 1 and 2 in the list indicated for the previous analysis.

In order to update \( F^* \), consider the following expression proportional to the
likelihood function 2.

\[ p(X \mid \alpha, \lambda, F, \sigma^2) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{z_i}{\sigma_i^2} [F^{*'}F^{*} - 2F^{*'}(X'_i - \lambda'\alpha'_i)] \right\}. \]

The complete conditional distribution for \( F^* \) is given by

\[ p(F^* \mid \alpha, \lambda, \sigma^2, z, X) \propto p(X \mid \alpha, \lambda, F^*, \sigma^2, z)\ p(F^* \mid \lambda) \]

\[ \propto \exp \left\{ -\frac{1}{2} \left[ F^{*'} \left( \sum_{i=1}^{m} \frac{z_i}{\sigma_i^2} I_n + K(\lambda)^{-1} \right) F^* - 2F^{*'} \left( \sum_{i=1}^{m} \frac{z_i}{\sigma_i^2} (X'_i - \lambda'\alpha'_i) \right) \right] \right\}. \]

Let \( V_{F^*} = \left[ (\sum_{i=1}^{m} \frac{z_i}{\sigma_i^2})I_n + K(\lambda)^{-1} \right]^{-1} \) and \( M_{F^*} = V_{F^*} \left[ \sum_{i=1}^{m} \frac{z_i}{\sigma_i^2} (X'_i - \lambda'\alpha'_i) \right] \); the kernel of \( N(M_{F^*}, V_{F^*}) \) can be recognized above.

Assuming the prior specification 1 for the indicator \( z_i \), we can express \( (F'_i \mid F^*) \sim (1 - \rho_i)\delta_{0}(F'_i) + \rho_i\delta_{F^*}(F'_i) \). This result allows us to derive the conditional probability to update \( z_i \) as indicated below. Denote \( F_{i(-i)} \) as the set of all rows in matrix \( F \), except \( F_i \).

\[ p(F'_i = 0 \mid \alpha, \lambda, F_{i(-i)}, \sigma^2, \rho_i, X) \propto p(X \mid \alpha, \lambda, F'_i = 0, F_{i(-i)}, \sigma^2, \rho_i)\ p(F'_i = 0 \mid \rho_i) \]

\[ \propto \exp \left\{ -\frac{1}{2\sigma_i^2} [\mathbf{0}'\mathbf{0} - 2\mathbf{0}'(X'_i - \lambda'\alpha'_i)] \right\} (1 - \rho_i) \propto (1 - \rho_i). \]

\[ p(F'_i = F^* \mid \alpha, \lambda, F_{i(-i)}, \sigma^2, \rho_i, X) \propto p(X \mid \alpha, \lambda, F'_i = F^*, F_{i(-i)}, \sigma^2, \rho_i)\ p(F'_i = F^* \mid \rho_i) \]

\[ \propto \exp \left\{ -\frac{1}{2\sigma_i^2} [F^{*'}F^* - 2F^{*'}(X'_i - \lambda'\alpha'_i)] \right\} \rho_i. \]

We normalize these results to obtain a value between 0 and 1; therefore, \( p(z_i = 1 \mid \alpha, \lambda, F_{i(-i)}, \sigma^2, \rho, X) \) takes the form

\[ \frac{\rho_i \exp \left\{ -\frac{1}{2\sigma_i^2} [F^{*'}F^* - 2F^{*'}(X'_i - \lambda'\alpha'_i)] \right\}}{(1 - \rho_i) + \rho_i \exp \left\{ -\frac{1}{2\sigma_i^2} [F^{*'}F^* - 2F^{*'}(X'_i - \lambda'\alpha'_i)] \right\}}. \]

The prior specification number 2 for \( z_i \) provides almost the same result, i.e., we update the indicator variable using the expression above with \( \rho_i \) replaced by \( \rho \). Note
that the full conditionals for $\rho_i$ and $\rho$ depends only on the indicators $z_i$; therefore, consider the same Beta distributions derived above for these elements.

### F.1.3 Prior and complete conditional for $\lambda$

We can write the following expression, depending on $\lambda_j$, proportional to the Likelihood version 1.

$$
 p(X \mid \alpha, \lambda, F, \sigma^2) \propto \exp \left\{-\frac{1}{2} \left[ \lambda_j' \alpha' D^{-1} \alpha \lambda_j - 2 \lambda_j' \alpha' D^{-1} (X_j - F_j) \right] \right\}.
$$

Assume the prior specifications:

$$
 \lambda_j \sim N_L(0, I_L) \quad \text{and} \quad (F'_i \mid \lambda) \sim (1 - z_i) \delta_0 (F'_i) + z_i N_n[0, K(\lambda)].
$$

Note that

$$
 p(F \mid \lambda, z) = \prod_{i=1}^m p(F'_i \mid \lambda, z_i) = \prod_{i=1}^m [(1 - z_i) + z_i N_n[F'_i, 0, K(\lambda)]].
$$

Denote $\lambda_{(-j)}$ as the matrix $\lambda$ without the $j$-th column. The complete conditional posterior distribution of $\lambda_j$ is given by

$$
 p(\lambda_j \mid \alpha, \lambda_{(-j)}, F, \sigma^2, X) \propto p(X, F \mid \alpha, \lambda, \sigma^2, z) \cdot p(\lambda_j)
$$

$$
 \propto p(X \mid \alpha, \lambda, F, \sigma^2) \cdot p(F \mid \lambda, z) \cdot p(\lambda_j)
$$

$$
 \propto \exp \left\{-\frac{1}{2} \left[ \lambda_j' \left( \alpha' D^{-1} \alpha + I_L \right) \lambda_j - 2 \lambda_j' \alpha' D^{-1} (X_j - F_j) \right] \right\}
$$

$$
 |K(\lambda)|^{-\frac{1}{2}} \sum_{i=1}^m z_i \exp \left\{-\frac{1}{2} \sum_{i=1}^m F_i K(\lambda)^{-1} F_i' \right\}
$$

$$
 \propto N_L[\lambda_j \mid M_\lambda, V_\lambda] \cdot |K(\lambda)|^{-\frac{1}{2}} \sum_{i=1}^m z_i \exp \left\{-\frac{1}{2} \sum_{i=1}^m F_i K(\lambda)^{-1} F_i' \right\}
$$

where $V_\lambda = [\alpha' D^{-1} \alpha + I_L]^{-1}$ and $M_\lambda = V_\lambda \alpha' D^{-1} (X_j - F_j)$. The closed form of the normalizing constant is not available in this expression. An indirect method is required to generate samples from the target distribution. We will consider the Metropolis-Hastings with a random walk as the proposal distribution.
Now, assume that the interaction effect is $F^*$ for any affected gene, i.e., $(F_i' \mid F^*) \sim (1 - z_i)\delta_0(F_i') + z_i\delta_{F^*}(F_i')$ with $(F^* \mid \lambda) \sim N_n[0, K(\lambda)]$. The complete conditional posterior distribution is given by

\[
p(\lambda, j \mid \alpha, \lambda_{(-j)}, F^*, \sigma^2, z, X) \propto p(X, F^* \mid \alpha, \lambda, \sigma^2, z) \, p(\lambda, j)
\]

\[
\propto p(X \mid \alpha, \lambda, F, \sigma^2) \, p(F^* \mid \lambda) \, p(\lambda, j)
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left[ \lambda_j' \left( \alpha' D^{-1} \alpha + I_L \right) \lambda_j - 2\lambda_j' \alpha' D^{-1} (X_j - F_j) \right] \right\}
\]

\[
\times |K(\lambda)|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} F^* K(\lambda)^{-1} F^* \right\}
\]

\[
\propto N_L[\lambda, j \mid M_\lambda, V_\lambda] \mid K(\lambda)|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} F^* K(\lambda)^{-1} F^* \right\}
\]

where $M_\lambda$ and $V_\lambda$ have the same expressions identified in the previous case. Once again, the closed form of the normalizing constant is not available in the kernel above. An indirect method is required to sample from the target distribution; we will consider the Metropolis-Hastings with a random walk as the proposal distribution.
F.2 Additional results for the simulated study

Section 6.2 presents the main results related to the simulated scenario 1 (Simulation 1). Here, we show graphs indicating the performance of the Models 1, 2, 3, 4 and 5 to fit other two synthetic data sets (Simulations 2 and 3).

**Simulation 2**

**Simulation 3**

**Figure F.1**: Synthetic matrix \(X\). The original data is displayed in the top of each panel. The image in the bottom shows \(X\) with standardized rows (rows and columns are sorted so that the first principal component is monotone).
Figure F.2: Results from Model 1: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$, $\lambda_{lj}$, $\sigma^2_i$ and $F_{ij}$. Panels in rows 1-2 = Simulation 2. Panels in rows 3-4 = Simulation 3.
Figure F.3: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 1.
Figure F.4: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 2.
Figure F.5: Real value (circle), Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ (Column 1) and $F_{ij}$ (Column 2). Rows 1-4 correspond to the Models 2, 3, 4 and 5, respectively. Results related to Simulation 3.
F.3 Additional results for the real data analysis

Section 6.3 presents the main results for a model assuming factors related to the pair of chromosome regions (2,4); see Table 6.3. In the current Appendix section, we show graphs exploring the results related to the pairs of locations (1,4) and (3,4).

Figure F.6: Matrix \( F \) containing the interaction effects. Column 1: full matrix (3744 genes); Column 2: matrix with the cases \( F_i \neq 0 \) (rows and columns are sorted so that the 1st principal components are monotone); Row 1: Pair of chromosome locations (1,4); Row 2: Pair of locations (3,4).
Figure F.7: Posterior mean (x mark) and the 95% credible interval (bar) for $\alpha_{il}$ $(i \in G_1 \cup G_2)$. The dashed line separates the two factors. Rows 1-3 are related to the data sets Chin et al. (2006), Miller et al. (2005) and Wang et al. (2005). Columns 1-2 correspond to the pairs of chromosome locations (1,4) and (3,4).
Figure F.8: Column 1: 3-dimensional surface plot representing the estimated interaction effect in $F_{1859}$ (row 1), $F_{3155}$ (row 2), $F_{2006}$ (row 3) and $F_{2719}$ (row 4). Column 2: posterior mean (x mark), used to create the surfaces, and the corresponding 95% credible interval. Consider the data set Chin et al. (2006); Rows 1-2 = pair of chromosome locations (1,4), Rows 3-4 = pair (3,4).
F.4 Additional results for the analysis comparing the models with linear and non-linear structure

Section 6.4 compares the frameworks from Chapters 5 and 6 in a study of synthetic data representing a scenario 1 (Simulation 1). In addition, only the Models 1 and 2 (non-linear structure) are investigated in that section. Here, we explore results for other two simulations, as well as the versions 3, 4 and 5 of the non-linear factor model.

(Figures F.9) 3-dimensional surface plot representing the estimated interaction effect. According to Chapter 5: Approach 1 = model assuming $\lambda_{3j} \sim N(\lambda_{1j}\lambda_{2j}, \nu)$, and Approach 2 = model assuming the equality $\lambda_{3j} = \lambda_{1j}\lambda_{2j}$. 

164
Figure F.10: 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 1.
Figure F.11: 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 2. The models are fitted assuming $l_s = 0.2$. 
Figure F.12: 3-dimensional surface plot representing the estimated interaction effect. These results are related to the Simulation 3. The models are fitted assuming $l_s = 0.2$. 

167
Bibliography


Biography

Vinicius Diniz Mayrink was born on November 4, 1981 in Joao Monlevade, Brazil. In 2004, he received his Bachelor degree in Statistics from the Federal University of Minas Gerais (UFMG, Brazil) working with Dr. Rosangela Loschi. In 2006, he received his first Master of Statistics degree from the Federal University of Rio de Janeiro (UFRJ, Brazil) under the supervision of Prof. Dani Gamerman. In the first semester of 2007, Vinicius worked as an instructor teaching statistical courses at the Federal University of Ouro Preto (UFOP, Brazil). He moved to the United States to start his Ph.D. studies in August, 2007. In 2009, he completed the requirements for his Master of Statistics degree at Duke University. In 2011, Vinicius received the degree of Doctor of Philosophy from the Department of Statistical Science in the Graduate School of Duke University. Dr. Joseph Lucas worked as his advisor from 2008 to 2011.