Multilevel Models for the Analysis of Linguistic Data

Craig Alexander

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy

September 2018

Declaration of Authorship

I, CRAIG ALEXANDER, declare that this thesis titled, 'Multilevel Models for the Analysis of Linguistic Data' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

Describing the numerous factors that constrain and promote particular aspects of linguistic behaviour in interaction is very difficult. The recent adoption of more advanced quantitative methods has enhanced this modelling, leading to a greater understanding of linguistic patterns. At the same time, the increase in availability of digital recordings and storage capacity for such recordings is leading to increasingly large corpora of complex linguistic data for such investigations. The Sounds of the City corpus is one such example and is the corpus we model throughout this thesis. The corpus is an electronic real-time corpus of Glaswegian vernacular, which consists of a searchable, multi layered database of 58 hours of recordings from 136 speakers, recorded between 1970 and 2010 with orthographic transcripts and automatically phonemically segmented waveforms, amenable to automatic acoustic analyses of durational and resonance characteristics of speech.

Vowel formant measurements provide a numeric representation of a spoken vowel and are a commonly used metric to measure linguistic variation and change, with each vowel having multiple formant measures, which correspond to the resonances of the vocal tract. The first three vowel formants are important perceptual cues for the successful recognition of vowel qualities. Current quantitative modelling methods consider each formant separately, inferring characteristics on each formant measurement assuming independence between each formant. This assumption for most vowels seems misplaced, as formant measures are often correlated with one another.

In this thesis, we extend upon current modelling techniques applied to sociolinguistic corpora by introducing a Bayesian hierarchical model which models the first three formant measures for each vowel simultaneously, taking

into consideration the correlation present between such measures. We also implement reparameterisation methods to alleviate issues caused by highly correlated samples, which is often observed in MCMC output for models applied to datasets with nested structures, a common feature in sociolinguistic corpora. These models not only account for the complex nested structure of the data and uncover the underlying dynamics of language just like classical mixed effects models, but now additionally account for the correlation between formants, providing a more accurate representation of factors driving linguistic variation and change.

The output from the Bayesian hierarchical model is visualised as a graphical model. Graphical models provide a visual representation of the conditional dependence between variables, making them an attractive inference tool. We combine the hierarchical model and jointly infer the relationship between vowel formant measurements using the precision estimates from the hierarchical model as input to a Bayesian Gaussian graphical model. The resulting graph utilises a chain graph like structure which visually informs the user which factors have a significant effect on vowel variation, corresponding to each formant, and also the relationship present between the first three formants. This novel inference tool helps to aid the understanding of complex model output much like the ones fitted to the Sounds of the City corpus, though can easily be applied to numerous modelling problems.

Acknowledgements

This project was funded by the University of Glasgow's Lord Kelvin Adam Smith scholarship programme. I would like to thank the following people who helped make this thesis possible:

Firstly, I would like to thank each of my supervisors for their patience, guidance and support throughout this project. I would like to thank Dr. Ludger Evers for sharing his extensive knowledge of statistics with me, his novel suggestions for how to approach this problem and always knowing a function in R which avoided numerous 'for' loops! Dr Tereza Neocleous for making sure everything remained on track and being a calming influence throughout. Prof. Jane Stuart-Smith whose enthusiasm for phonetics made researching a new subject area an enjoyable experience and far less taxing than it could have been.

Next I would like to thank the friends I have made over the years at Glasgow and beyond. Without them all, this work would never have been possible. Thanks to Marnie who's been my PhD partner in crime since day one; you've made this journey a lot easier than it could have been. Thanks for reassuring me the 1×10^{100} times I had a meltdown! Vinny, thanks for the numerous pints, expanding our quiz machine knowledge and always being able to answer a stats question. Umberto, thanks for the music, fine dining and the paint skills, they brightened up many a dull day. Lida, thanks for the new song suggestions, the roasts and letting me be excused for a moment. Alan, thanks for the email correspondence and the expertise in the gambling round of the quiz. Fraser, thanks for the travel vlogs and being the other half of "those boys from Glasgow". Irene, thanks for the mini tour of Spain and dragging me to circuits. Paul, thanks for the endless Simpsons jokes, the TV show recommendations

and the annual distillery tour. Sean, thanks for always being free when I needed a hand moving, and dragging me up the climbs on the bike. Lynne, thanks for seshions ™and always being dependable, David, thanks for enduring Rangers with me the past 7 years (I still don't rate Dorrans but) and less thanks for being the worst tipster going! Finally thanks to everyone in the department I have met over the years, particularly my office mates old and new, who made office life all the more entertaining.

Finally, I would like to say a massive thank you to my family throughout my time at university and beyond. Without their input and help over the years, I doubt I would ever have been able to undertake this work. I would especially like to thank my mum, Amanda, for everything over the years. I'd never have written this without your support and input throughout life. Thank you for helping me get here.

This work is dedicated in memory of my grandmother Janet. My first, and best teacher.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sociolinguistics is the study of the intricate relationship between language and society, looking at how culture, society and geography interact with language. Variationist sociolinguistics focusses on a specific branch of this broad subject, taking its focus on the study of language change using quantitative methods. It is the study of linguistic variation and change through observation and interpretation (Tagliamonte, 2012). The core tool of sociolinguistics is the notion of the linguistic variable, which provides a metric to determine if social or linguistic factors are impacting on linguistic variation and change. At its most basic definition, the linguistic variable is two or more ways of saying the same thing. For example, the sound /t/ in 'butter' can be produced as [t] or as a glottal stop. Analysing the variation for the linguistic variable, T-glottalling, has shown that glottal stops are used more by particular social groups (in British English, working-class, male, younger speakers), and in particular linguistic contexts than others (more in word-final position, e.g. 'but' than word-internal position, e.g. 'butter'). It has also shown how T-glottaling has increased in usage over time and space (Smith and Holmes-Elliott, 2017). Variationist sociolinguists have established that linguistic variation is constrained by different social and linguistic factors (Labov, 2001).

Statistical modelling provides a formal assessment of the relationship between the linguistic variable and relevant social and linguistic factors. Traditionally, the most commonly used tool for analysis was logistic regression, first implemented in the variable rule program `Varbrul` (Cedergren and Sankoff, 1974) and then later extended to `Goldvarb`

`2.0` (Rand and Sankoff, 1990). The variable rule program has the structure of a generalised linear model, hence has the capability of performing logistic regression. Unfortunately, `GoldVarb 2.0` lacks the flexibility to perform robust statistical modelling on sociolinguistic corpora, as it fails to capture the additional variability present between Speakers and Word choice within a corpus. Johnson (2009) introduces `Rbrul`, which uses mixed effects modelling at its core, accounting for these additional sources of variation within the corpus.

This recent development of advanced quantitative methods has been a core aspect of analysing and interpreting sociolinguistic patterns, along with the increase in availability of digital recordings allowing the formation of large corpora of complex linguistic data to exist for such investigations. Within such analyses, the linguistic variable of interest may be discrete, e.g. [t] or a glottal stop, or continuous, e.g. formant measures in Hertz for a vowel. As mentioned previously, the common approach to tackling such analyses is to implement mixed effects models (Pinheiro and Bates, 2000), which include random effects which control for experimental variation created by individual speaker or word level variation (Tagliamonte and Baayen, 2012), alongside fixed factors to describe the influence of linguistic and social variables.

When considering formant measures of a vowel as the linguistic variable of interest, for each vowel, we obtain multiple formant measurements, with the first three formants being most commonly modelled (Ladefoged and Johnson, 2014). Until now, it has not been possible for quantitative analysis in sociolinguistics to consider the impact of modelling these multiple formant measurements together, instead fitting models which only consider each formant in turn, thus assuming independence between formants. A speaker's vowel formant measures show a degree of correlation, at least partly as a result of being produced by the same vocal tract. The main question of interest is, to what extent do linguistic and social factors influence the production of vowels, as measured in formants, above and beyond these formant correlations?

We present a Bayesian hierarchical model in Chapter 3 for the analysis of multiple response variables. The functionality is demonstrated through the analysis of the first three formants of the *FLEECE, FACE, TRAP/BATH, LOT, GOAT* and *GOOSE/FOOT* vowels for 31 speakers from the Sounds of the City corpus. The Sounds of the City cor-

pus is an electronic real-time corpus of Glaswegian vernacular (Stuart-Smith et al., 2017), (Stuart-Smith and Lawson, 2017), (Rathcke et al., 2017). This corpus is a searchable, multi layered database of 58 hours of recordings, recorded between 1970 and 2010 with orthographic transcripts and automatically phonemically segmented waveforms, amenable to automatic acoustic analyses of durational (e.g. segment durations) and resonance characteristics of speech, so for vowels, formant measurements in Hertz (Hz). Social factors of interest related to each individual Speaker, for example the Gender, Age and Decade of recording of a Speaker, and linguistic factors such as the Preceding and Following segmental context for the vowel are taken as fixed effects of interest within the model. Their significance is determined within the hierarchical model, which also implements variable selection within the sampler, proposing the addition and removal of coefficients with every iteration of the MCMC sampler to determine the best fitting model.

Sociolinguistic corpora are often nested in design due to the nature of the sampled data. Linguistic variation is produced by speakers who in turn belong to social groupings, nested by e.g. Gender, Age, Decade of birth, and it occurs in words which show differential patterns of use and frequency across speakers. Within an MCMC framework, this often leads to high autocorrelation between parameter samples and thus poor mixing for parameter estimates. This leads to MCMC chains being run for extended periods of time, which from a practical sense in terms of computational time is infeasible. Reparametrisation methods can be implemented to improve MCMC efficiency in nested design problems (Browne et al., 2009). In Chapter 4, we introduce two reparameterisation steps based upon hierarchical centering and parameter expansion, which aim to improve MCMC efficiency greatly in terms of mixing observed between fixed effects coefficients nested within the random effects. We also look at how to improve poor mixing in the precision estimates for effects with a high number of levels.

As we are introducing a model which has now increased in complexity when compared to classical mixed effects models, implemented on one formant at a time, it is imperative that the output of the model is communicated in a clear and concise fashion in order to make this methodology an attractive tool for sociolinguists to use and interpret. In order to do this, we propose the use of graphical models as a visualisation tool of the Bayesian hierarchical model output as discussed in Chapter 6. Graphical models pro-

vide a simple diagrammatic representation of complex probability structures which help to ease understanding of large scale problems. We look to implement a chain graph style structure, where we jointly infer the fixed effects present in the hierarchical model and also the underlying graphical model between the response variables using precision estimates obtained from the hierarchical model. The interpretation is straightforward, with a connection denoted by a direct arrow present between an explanatory variable and a response variable, indicating the explanatory variable is a significant term within the model. This novel inference tool provides a straightforward visual representation of complex model output.

Throughout this thesis, the methods implemented and developed have been constructed to be as generalisable as possible. The motivation for this is to introduce a new inference tool for tackling problems of this nature. The problems tackled within this thesis are not limited to multiple variables characterising a vowel, they can be also used for other multiple variables which characterise other sounds, e.g. stop sounds and sibilant sounds. Nor are they limited to sociolinguistic corpora, there are many practical examples across multiple disciplines which have structured design problems similar to linguistic corpora. Due to the abundance of problems of this nature, we have produced functionality within *R* for the Bayesian hierarchical model which also is capable of producing the chain graph model like structures. The functionality has been created such that for any mixed effects model problem, with univariate or multiple responses or with nested design, it is possible to obtain a hierarchical model and graphical model. An additional aim of this project is to turn this functionality into a package within *R* and also develop a web-based application of the chain graph model using Shiny (Chang et al., 2015). A repository for the code used throughout this thesis can be found at `https://github.com/calex1991/BayesCGModels`.

The work of this thesis has developed and expanded upon the quantitative methods used to tackle questions regarding linguistic variation in variationist sociolinguistics. We have proposed an extension to the mixed effects modelling currently implemented in two main ways: firstly by expanding the model to consider multiple vowel formants simultaneously, taking into account the correlation between these vowel formants, thus obtaining a more accurate representation of the underlying factors influencing vowel change, and

secondly by expressing the model problem in a Bayesian framework, which helps with the creation of the graphical model visualisation. Common issues of poor mixing in MCMC samplers using nested data are also addressed through reparameterisation techniques. Finally, we have constructed a novel inference approach using graphical models to visualise the output of such Bayesian hierarchical models with a view to help simplify interpretation and understanding of the complex model output.

1.1 Thesis Overview

The structure of this thesis is in the following form: Chapter 2 provides some relevant background information on phonetics and sociolinguistics, with further detail on the structure of the Sounds of the City corpus which is the main dataset used throughout this thesis. Chapter 3 details the current methods used to model linguistic variation in the Sounds of the City corpus, then goes on to detail the construction of the Bayesian hierarchical model for multiple vowel formants, with an application to the Sounds of the City corpus. Chapter 4 discusses the reparameterisation methods used to alleviate poor MCMC convergence for nested parameters within the Bayesian hierarchical model, focussing on improvement on mixing of parameter chains. Chapter 6 describes how to construct the chain graph style graphical model for hierarchical model output, describing how we incorporate Bayesian Gaussian graphical model selection within the hierarchical model, using a modification of the PAS algorithm. Chapter 7 discusses the sociolinguistic findings of the resulting graphical models obtained within the chapter. Chapter 8 provides a summary of this thesis, with a discussion on potential future developments from the work undertaken in this thesis.

Derivations of the posterior distributions used to sample the model parameters are detailed in Appendix A.

Chapter 2

Sociolinguistic Background & Data

In this chapter, we provide information about the structure of the Sounds of the City corpus, looking closely at how the work detailed in this thesis can extend upon current findings already obtained from the corpus in terms of how language has changed within Glasgow over the past century. In Section 2.1, we look at how we can model change in language by looking at vowel sounds in detail, and how we can obtain metrics from a vowel utterance that can provide some sense of measure from a particular vowel sound. Section 2.2 looks more closely at the structure of the Sounds of the City corpus, discussing in detail the structure of the corpus, previous findings and results obtained from the corpus in terms of which vowels seem to be providing clearer indicators of vowel change in the Glaswegian vernacular and the structure of the vowel data which form the basis of the analysis presented in this thesis.

2.1 Modelling Vowel Change

The aim of the Sounds of the City project is to study how language has varied over the course of the twentieth century in the city of Glasgow, especially with respect to its pronunciation. Preliminary work on the dataset suggested that a number of aspects of the Glaswegian accent are changing, including some vowels (Stuart-Smith et al., 2017).

Vowel sounds phonetically, in words like. FLEECE and TRAP, are those sounds which are produced without any obstruction to the airflow leaving the vocal tract. Vowels can

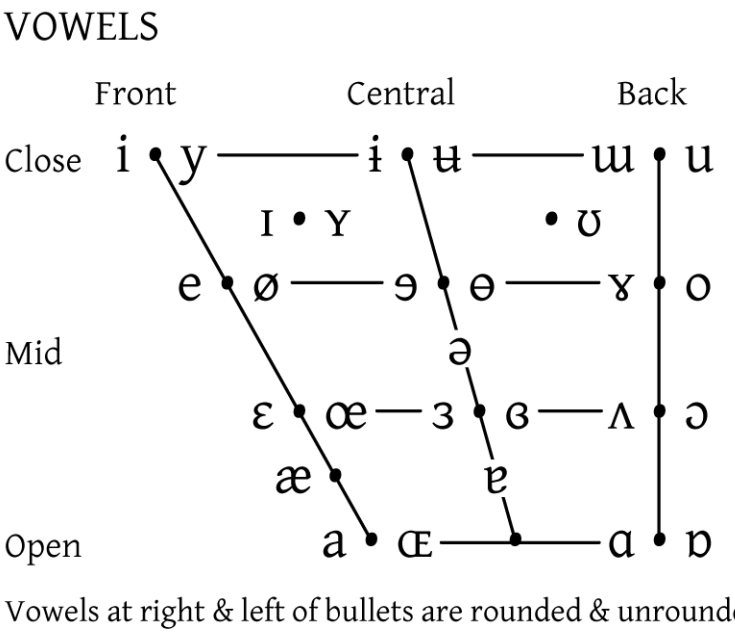

Figure 2.1: Vowel chart of the International Phonetic Alphabet IPA (1999). The dimensions refer to the position of the highest point of the tongue, so 'front' furthest forward as in FLEECE, back furthest back as in CAUGHT.

be described by close listening (auditory analysis) and then transcription using symbols from the International Phonetic Alphabet (IPA, 1999), as detailed in Figure 2.1. Using this system, the vowel in FLEECE is /i/ and in TRAP is /a/. The IPA chart provides a symbolic description of any vowel sound produced in a passage of speech. The IPA chart shown in Figure 2.1 displays the different vowel qualities produced with reference to the position of the highest point of the tongue and the shape of the lips when the vowel sound is produced. The vowel quadrilateral shape is a schema of the vowel space, which is created by the movement of the tongue from front to back, and from closer and further away from the hard palate (roof of the mouth). It was initially based on auditory and quasi articulatory ideas about vowel production (Ladefoged and Johnson, 2014).

When considering evidence for variation and change in vowel sounds, several metrics can be used. In the Sounds of the City corpus, we use acoustic measurements taken from vowels as our metric.

Auditory analyses of vowel sounds using IPA are very common, but they result in discrete variants, and can be subjective. Sociolinguistic analyses usually use acoustic analyses of vowels. In order to differentiate between different acoustic vowel qualities, their differences are studied in terms of spectral frequency and intensity (Johnson, 2011).

This can be done by considering the waveforms and spectrograms obtained from sound recordings.

A spectrogram is a plot of frequency over time which includes a grey scale to highlight the differential patterns of acoustic energy, in terms of amplitude at particular frequencies during the production of a particular sound. White areas indicate areas of minimal noise or silence, while the darkest areas indicate frequencies of high amplitude in comparison to surrounding frequencies which tend to be greyer in appearance. These spectral features result from smoothed fast Fourier transforms applied to the acoustic waveforms resulting from movements of the articulators during speech production. For example, producing a stop sound such as 'p' in 'pin', involves closing the lips, holding them closed, and then releasing the trapped airflow to produce the vowel, for which the tongue body has shifted forward and close to the hard palate, so during a stop acoustically there is much less visible acoustic energy. Vowel sounds are voiced through vocal fold vibration and show resonances at particular frequencies reflecting the shape of the vocal tract configuration for each particular vowel sound. An example of a spectrogram is shown in Figure 2.2 for an utterance of the word *sign* for a female Glaswegian speaker. Here we can see first the acoustic noise corresponding to the turbulent jet of air produced for /s/, then the dark bands of energy reflecting the vowel /ai/, with coloured lines pointing out the first three formants, followed by less visible energy for the nasal /n/ (air escapes through the nose whilst the tongue obstructs the mouth).

Ladefoged (2005) compares the differences in vowel sounds to that of an orchestra. The same note can be played by multiple instruments, which will produce a sound with the same fundamental frequency, which we hear as the pitch of the vowel, but different overtones by instrument. The difference between vowel sounds can be distinguished by such overtones. These resonances of the complex filter formed by the supralaryngeal vocal tract are referred to by phoneticians as formants. A formant is a concentration of acoustic energy around a range of frequencies (i.e. a particular bandwidth) in a speech wave after the wave has been subjected to a spectral analysis (e.g. fast Fourier transform).

As sound waves pass through the oral cavity, they are modified by the differing configurations of the articulators and develop a characteristic pattern of energy along specific frequency ranges that can be interpreted by the brain and the ear as a particular vowel

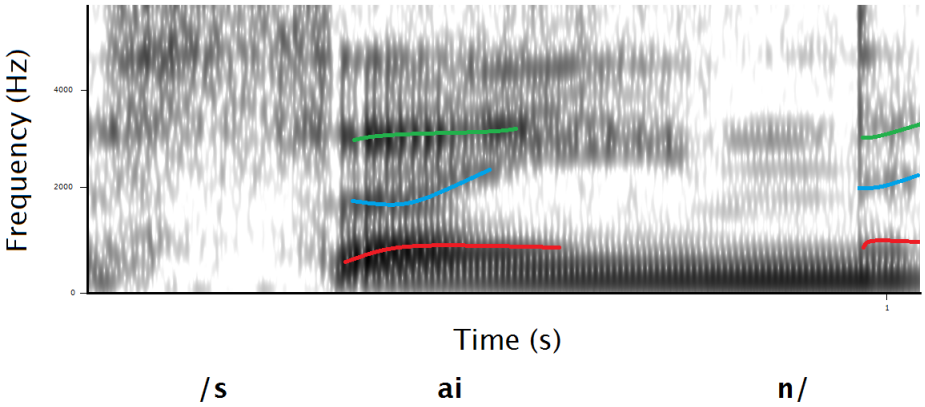

Figure 2.2: Spectrogram of the word *sign* as spoken by a female Glaswegian speaker. The coloured bands correspond to the formants F1 , F2 and F3

.

sound (Johnson, 2011). Formants provide a useful measure to model how the vocal tract acts as a filter in the production of voiced sounds such as vowels, nasals (n.m) and liquids (l, r). The patterns they produce help to define the phonetic quality of sounds and also their place of articulation.

When observing a spectrogram of vowel sounds, a series of thick dark bands can be observed like those in Figure 2.2 which distinguish the individual formants for each vowel sound. Early perception experiments demonstrated the importance of formants for vowel identification and discrimination, specifically the first two formants (Delattre, 1951).

2.2 The Sounds of the City Corpus

The Sounds of the City corpus is a real time corpus of Glaswegian vernacular. Sociolinguistic corpora are of two kinds. 'apparent-time' corpora are where recordings are from the same time point, but from speakers of different ages; speaker age acts as a proxy for time depth. 'real-time' corpora contain recordings made at different points in time (Labov, 1994). The Sounds of the City corpus is both real-time (recordings made at different time points) and apparent-time, from speakers of different ages, at each time point. The corpus consists of recordings of 136 speakers, recorded over 58 hours, comprising some 700,000 words. Recordings were made over four decades from male and female speakers in three age groups, old (67-90), middle-aged (40-55), and young (10-15)

between 1970 and 2010.

In order to gather this real time corpus of data, recordings were collected together from existing sources of different kinds, for example, previous sociolinguistic surveys, footage for broadcast programmes, and oral history interviews. An issue with this approach is that it has lead to an unbalanced design in several respects. Since this is an opportunistic sample, where all possible recordings for Glaswegian dialect were collected from existing sources, there are varying numbers of speakers across groups by generation. Also as the speakers are recorded in different ways, and talk about different things, there are differences in the amount of speech per speaker, and the content, and number of words, used by different speakers.

2.2.1 Vowel Data

This section outlines which vowels were selected, and how the data were extracted and also specifies what the measures and variables are.

Previous research on the SoTC corpus, in conjunction with other research across UK dialects (Stuart-Smith et al., 2017) suggested that the following vowels might be of interest to analyse: *FLEECE, FACE, TRAP, BATH, COT, GOAT, GOOSE* and *FOOT*. Given that the Sounds of the City research is currently analysing these vowels using typical linear mixed effects models (Jose and Stuart-Smith, 2014), these vowels were selected as the basis for this statistical study.

In order to investigate vowel variation and change in this corpus, the speech recordings were first orthographically transcribed producing utterance-level alignment. They were then uploaded to the open-source speech database system, LABB-CAT (Fromont and Hay, 2012), and force-aligned, giving time-aligned segmentation (time stamps) for utterance, word and segments, with corresponding labels. Automated searches were carried out within LABB-CAT to locate and extract all relevant tokens for all vowels. The first 3 formant measures were then taken for each vowel, using the LABB-CAT vowel measurement tool. All of these steps had been carried out prior to this study, as part of the SoTC data processing. Each instance of each vowel, e.g. FLEECE, as uttered in words spoken by the speakers, e.g. beat, bead, sleepy, feet, is thus represented in the form of three formant measures in Hz. These are used as the variables of interest in this

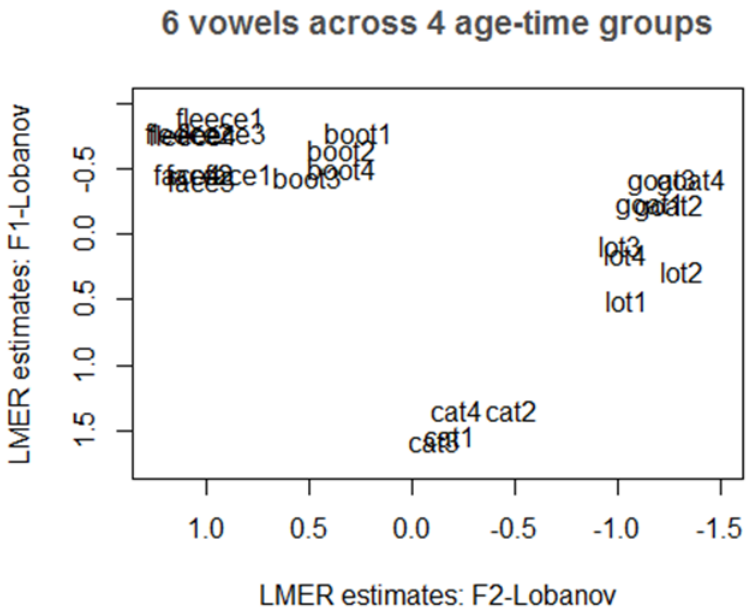

Figure 2.3: Plot of the acoustic normalised F2/F1 vowel space for the *FLEECE, CAT, LOT, GOAT* and *BOOT* vowels where 1= 1970s Old speakers, 2= 2000s Old speakers, 3= 1970s Young speakers, 4= 2000s Young speakers.

analysis.

Figure 2.3 shows the acoustic normalised F2/F1 measures on the vowel space for the *FLEECE, FACE, CAT, LOT, GOAT* and *BOOT* vowels. We observe the vowel layout here is arranged much in the same way as the IPA chart in Figure 2.1, so the *FLEECE* vowel (front and high) is in the top left corner, and the *CAT* vowel (low) is in the bottom. We observe that the different averages for the *FLEECE* and *FACE* vowels basically sit on top of one another across the range of recordings. Conversely, we see that *BOOT* has shifted down in the space (lowering) and *COT* and *GOAT* have shifted up in the space over the real and apparent time represented by the speaker sample. In the *CAT* vowel we see some possible shifting up in the space. Note that in previous work, TRAP and BATH vowels are considered together as CAT, whereas FOOT and GOOSE are considered together as BOOT.

From Figure 2.3, we observe the *BOOT, LOT* and *GOAT* vowels appear to be changing the most over time. Figure 2.4 shows each of these vowels in closer detail, as obtained from Stuart-Smith et al. (2017). We observe a real-time lowering of *BOOT* and a raising of *LOT* and *GOAT* from the 1970s recordings. The apparent-time findings of the elderly

speakers compared with recordings of men born in the same decade (1890s) but recorded during the First World War (marked as X on the plots) suggests that these changes may have started much earlier in the century. This motion of one vowel moving and others following is known as a 'pull-chain' (Labov, 1994). The findings lead researchers to believe that the Glaswegian dialect has some kind of pull chain occurring, which probably started around the mid-19th century, and then took off over the course of the 20th century.

The data analysed here contains quantitative phonetic measures of vowels, consisting of formant measurements taken for the first, second and third formant, denoted F_1, F_2, F_3 respectively. Throughout this analysis, we use the raw mean formant frequency calculated over the duration of each vowel for the first 3 formants; we also analyse normalized F1 and F2. In the data analysed here, Age has two levels: Old and Young speakers

It is common for raw formant measurements to be normalised (Adank et al., 2004). Vowel normalisation techniques have been developed as different speakers have different vocal tract sizes, which in turn causes their formant resonances to differ. This means that, for example the vowel FLEECE produced by a small child and an adult man will show different frequencies relating to their vocal tract size - but listeners will carry out some kind of normalization internally, and will parse both utterances as instances of the vowel /i/. Vowel normalisation is used to compare the vowel realisations by different speakers in meaningful sociolinguistic ways. By eliminating variation caused by physiological differences among speakers, it is easier to determine whether changes and differences in terms of vowel quality are influenced by sociolinguistic factors.

Several normalisation techniques exist to normalise vowel formants. A more detailed description of all these techniques can be found at http://lingtools.uoregon.edu/norm/. One such normalisation method used to model vowel change in the Sounds of the City corpus is the Lobanov normalisation method, which is implemented using Kendall and Thomas (2014), and is defined as

$$F_{s,i}^N = \frac{(F_{s,i}^N - \mu_s^N)}{\sigma_s} \tag{2.1}$$

where $F_{s,i}^N$ is the normalised value of formant $F_{s,i}$, taken on the i^{th} measurement for speaker s. μ_s is the mean formant value for speaker s and σ_s is the standard deviation

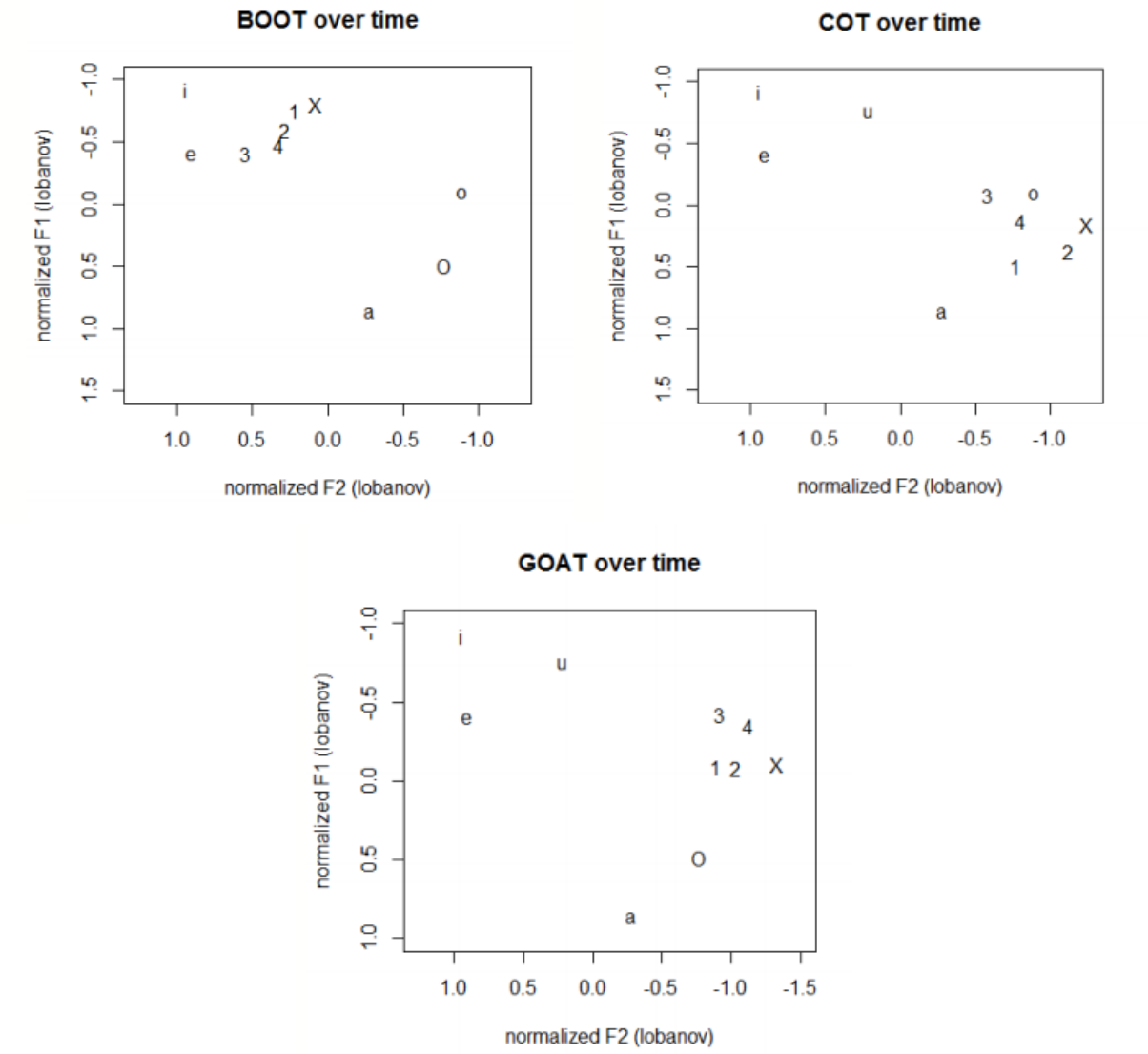

Figure 2.4: Plots of F1 and F2 measures with the relative position for each speaker group for the *BOOT, LOT* (which in this study, is denoted as *COT*) and *GOAT* vowels, where 1= men born in 1890, 2= men born in the 1920s, 3= adolescents born in the 1960s, 4= adolescents born in the 1990s and X = young men born in the 1890s and recorded in 1916/17.

of speaker s corresponding formant values for a given vowel.

The dataset also contains several explanatory variables relating to individual qualities of the speaker. Sociolinguistic theory observes that language variation and change is fundamentally influenced by two kinds of variable, social factors and linguistic factors (Tagliamonte, 2012). Given that the SoTC project is interested in tracking sound change, a key variable is Decade of recording, which in these data has two levels, 1970s and 2000s. Gender and Age of speakers is also of interest to model. There are two sources of variation for these variables. The biological size of the vocal tract is one, as female and younger speakers would be expected to produce slightly higher formant frequency values because they have smaller vocal tracts. The other is social gender and age (Labov, 2001). We are interested in what males and females in a particular community 'do' with their speech to sound like male and female speakers. For example, we would expect men and women to differ in terms of pitch, with males showing lower frequencies than females, but at the same time, communities can acquire pitch norms which may override their biological norms (Eckert and McConnell-Ginet, 2003). An example of this would be Cockney English and French speaking males having higher pitched voices than southern English speaking males. In terms of Age, speakers may also vary because their chronological age reflects a continuation of speech patterns which they acquired much longer ago. For example, a female speaker aged 70 will show speech variation which is typical of both an older woman (physiologically), and will reflect the language system which she acquired as a child, typically around 7-8 years old (Labov, 1994).

Several phonological variables are also recorded, most notably the preceding and following place of articulation of the consonants surrounding the vowel. These linguistic variables are important to include in the analysis as movement of the articulators alters the resonant properties of the vocal tract. For example, if a preceding consonant led to rounding of the lips, this often results in a lowering of the second and third formants because the oral cavity is lengthened, e.g. the sound of /i/ after /s/ (no rounding) and /sh/ (which has lip rounding) in *seep* and *sheep* (Ladefoged and Johnson, 2014).

In Figure 2.5 the vowel measures for old speakers recorded in the 70s and young speakers in the 00s are show. The plot consists of the median formant measures for each speaker for a specific vowel, represented by the large circle. Each individual observation

per speaker is also represented by the fan of points joined to the median circle. This gives a clear indication of the high variability present within each speaker, and the difference in formant measures between speaker. From the plot, we observe a shift in the vowel formant measures from generation, most notably in the *GOAT* vowel. There has been a raising in normalised F2 measures for this vowel. We also observe less variability with young speakers in the 00s. This is mostly due to the improvement in recording techniques between generations.

Figure 2.6 details the effect of preceding and following place of articulation for the *FOOT* vowel. We observe that there is an effect present in terms of the preceding context of consonant in the word, with a clear difference in formant measures for coronal and labial. For following context, there appears to be no clear pattern present within the data.

Looking closer at specific vowels and the effect of generation in Figure 2.7, we observe a generation effect is present when considering the *GOAT* vowel. We observe a raising in the F2 normalised measures for younger speakers in the 00s compared to older speakers in the 70s. When considering the *CAT* vowel, there appears to be no real difference present in formant measures by generation.

It is well known from previous studies that vowel quality, reflected in formant measures here, also varies systematically according to individual speaker, and even according to the words in which they occur. It is therefore common for sociolinguists also to include factors which capture individual speaker and word variation in their models (Drager and Hay, 2012). Accordingly, here in this study, random factors of Speaker and Word are also included in the modelling.

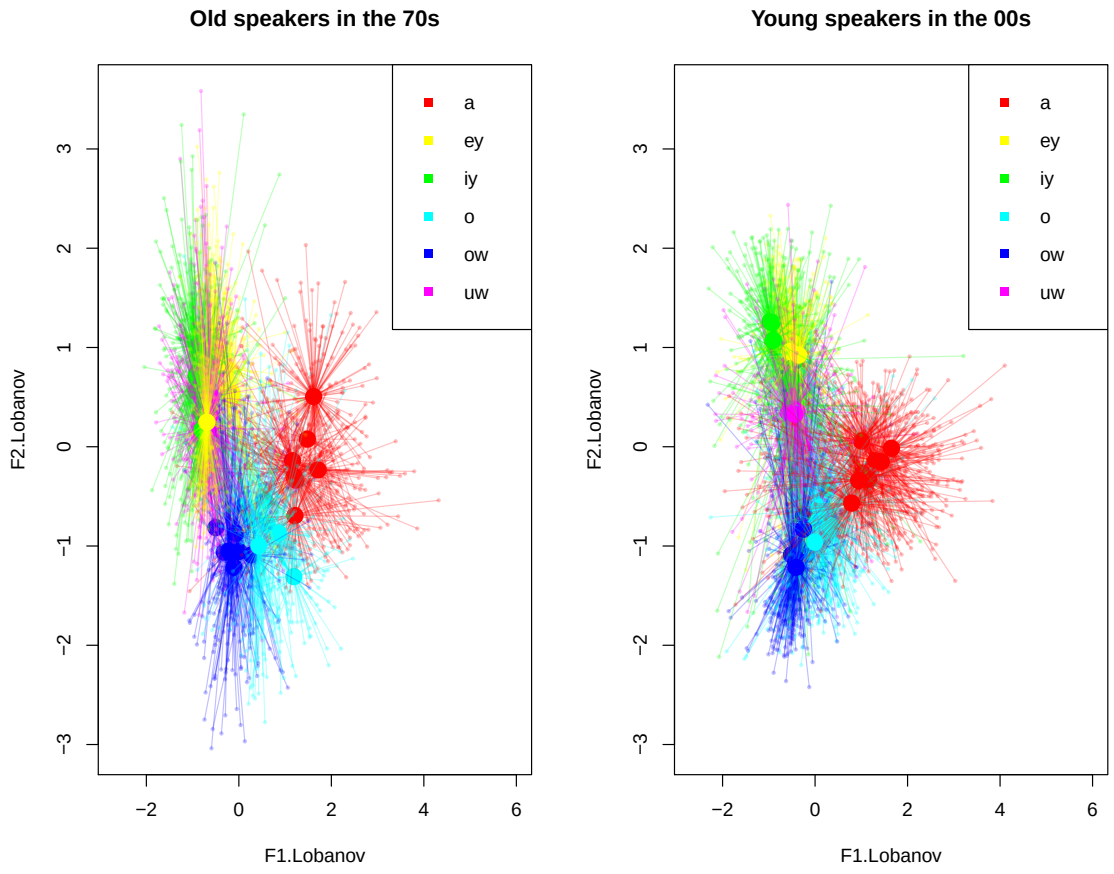

Figure 2.5: Plot of the acoustic normalised F2/F1 vowel space comparing old speakers in the 70s to young speakers in the 00s by vowel. The *GOAT* appears to have shifted more in terms of F1 from generations.

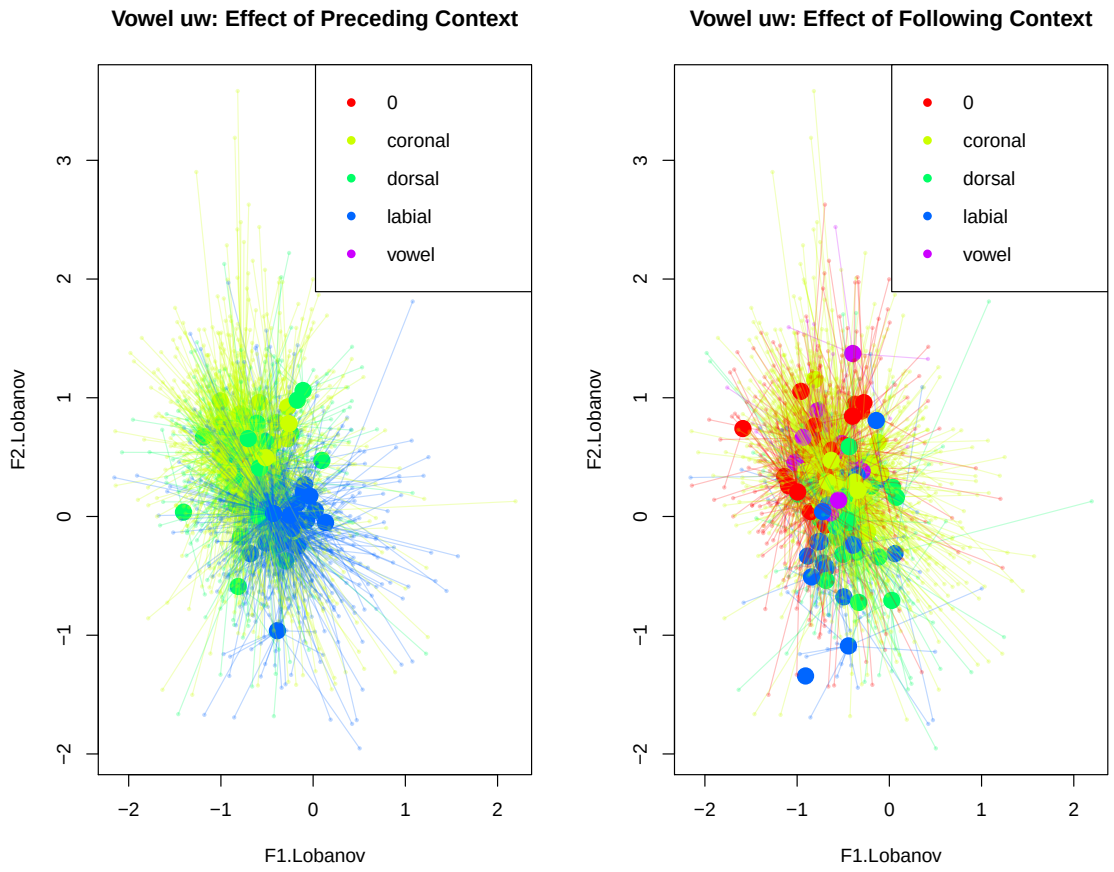

Figure 2.6: Plot of the acoustic normalised F2/F1 vowel space comparing effect of preceding and following context on the *GOOSE* vowel. Preceding context appears to have a clear effect on F1 and F2 measures.

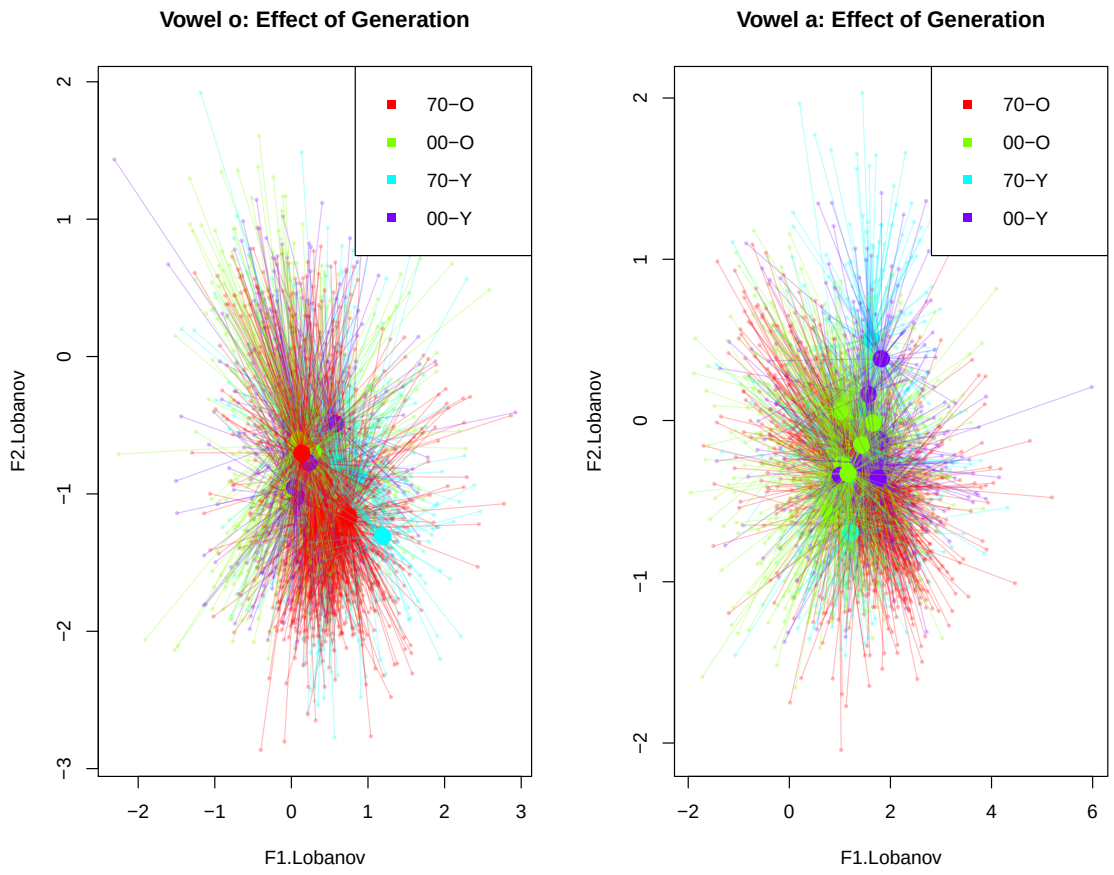

Figure 2.7: Plot of the acoustic normalised F2/F1 vowel space comparing the effect of generation between the *GOAT* and *CAT* vowels. A difference in F1 and F2 measures can be observed for the *GOAT* vowel, but no clear difference observed within the *CAT* vowel.

As we have observed throughout this chapter, it is usual for three formants to be taken from any particular vowel. The current modelling undertaken in Stuart-Smith et al. (2017) looks at the individual first and second formants in turn for each vowel and infers which factors are influencing change for that specific formant. This procedure is standard for sociolinguistic, and indeed phonetic, modelling of vowel variation. The natural progression in a modelling sense is to consider a statistical model which includes all the available formants for a vowel utterance, which in the case of the Sounds of the City corpus would be the first three formants (raw measures) and the first two formants (normalized measures). This thesis will detail the construction of models which can tackle this problem.

2.3 Discussion

In this chapter, we have introduced how we can model vowel variation and change in terms of formant measurements taken from a vowel utterance, providing a continuous metric in terms of frequency that can be used in any potential modelling. We have also discussed the structure of the Sounds of the City corpus, discussing in further detail the current modelling results obtained for vowel variation and change in the city, identifying the *BOOT, COT* and *GOAT* vowels as showing the most change over time within the dialect of the city. The remainder of this thesis will detail how we can extend beyond the current modelling in the Sounds of the City corpus by constructing statistical models which consider all three formants for a vowel in the same model.

Chapter 3

A Bayesian Hierarchical Model for Modelling Linguistic Change in Glaswegian Dialect

In this chapter, we introduce the multiple response Bayesian hierarchical model for modelling linguistic change for vowel formant data. This hierarchical model also carries out model selection within the sampler to select the significant variables present within the model and thus identify the phonetic and social factors which are contributing to linguistic change in the Glaswegian dialect.

The multiple response Bayesian hierarchical model develops on classical mixed effects models used to model linguistic corpora (Johnson (2009), Baayen (2008)) in a twofold manner. Firstly, the model allows for the modelling of multiple response variables within a single model framework, thus taking into consideration possible additional correlation between vowel formants. Secondly, we model in a Bayesian framework which lends itself to the graphical model representation which we discuss in more detail in Chapter 5.

This model aims to determine which social and linguistic factors impact vowel change, with the Decade of Recording (1970s or 2000s) for a specific speaker being the key variable of interest in determining whether there has been a change in vowel formant frequencies, and so vowel quality, over time in the Glaswegian vernacular.

Section 3.1 details the construction of the Bayesian hierarchical model, starting from

the construct of mixed effects models (Johnson, 2009) in Section 3.1.1. Section 3.1.2 extends beyond the classical linear model to the multiple response model, which allows for the modelling of multiple response variables. The Bayesian hierarchical model structure is then explained in Section 3.1.3, with a discussion into the structure of the model and relevant notation detailed. Prior specification is also discussed here and a visualisation of the full model structure is detailed in Figure 3.1. Bayesian posterior inference is then explained further in terms of the structure of the MCMC samplers for parameter inference and also further description on the variable selection undertaken within the model framework.

Section 3.2 applies the hierarchical model to two examples; firstly a simulated example which looks to explore and detail the effectiveness of the variable selection within the model framework, which is discussed in Section 3.2.1 and secondly, an application to the Sounds of the City corpus for raw mean vowel formant measurements and Lobanov normalised formant measurements for all vowels in Section 3.2.2. Within the Sounds of the City analysis, we identify and discuss mixing issues identified due to the nested design of the corpus, which is discussed in further detail in Chapter 4.

3.1 Building the Bayesian Hierarchical Model

In this section, we introduce the classical mixed effects models currently implemented within many sociolinguistic experiments to model linguistic change, and how we extend upon these models to allow for multiple formants at once in a Bayesian framework.

3.1.1 Mixed-Effects Models

Mixed-effects models (West et al., 2007) are the most commonly implemented methods used within the variationist sociolinguistic community to model additional experimental variability present within linguistic corpora (Johnson, 2009). In a mixed-effects model, the response is defined as $\mathbf{y} = (y_1, \ldots, y_N)^\intercal$ and the explanatory variables, commonly referred to as fixed effects, are denoted by $\mathbf{X}$, which is a matrix of $P + 1$ columns and N rows, where the first column is the population intercept. Each of the fixed effects has a corresponding regression coefficient, which is denoted by $\boldsymbol{\beta}$. For example, $\mathbf{X}_j$ has the

corresponding regression coefficient β_j, which defines its level of influence on the response.

Where the mixed effects model differs from a classical linear regression is in the addition of random effects, which are included to control for additional variation present in nested design problems, which in the Sounds of the City corpus is the Speaker and Word choice variation. The random effects design matrix, which is defined by $\mathbf{U}$, is a matrix of indicators with N rows and G columns, where G is the total number of groups within all random effects. The random-effect coefficients are defined as $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_G^\top)^\top$ which consist of a vector of coefficients corresponding to each specific random effect and its respective groups. Each $\mathbf{b}_g$ is assumed to follow a Gaussian distribution with zero mean, and variance respective to group, $\mathbf{b}_g \sim \mathcal{N}(\mathbf{b}_g \mid \mathbf{0}, \sigma^2_{\mathbf{b}_g}\mathbf{I})$. If we consider all groups, the joint distribution is defined as, $\mathbf{b} \sim \mathcal{N}(\mathbf{b} \mid \mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is defined as $\mathbf{G} = \text{blockdiag}(\boldsymbol{\sigma}^2_{\mathbf{b}})$ with $\boldsymbol{\sigma}^2_{\mathbf{b}} = (\sigma^2_{\mathbf{b}_1}\mathbf{I}, \ldots, \sigma^2_{\mathbf{b}_G}\mathbf{I})$.

The mixed effects model is of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma^2_\epsilon \mathbf{I}) \tag{3.1}$$

where the model is assumed to have independent and identically distributed Gaussian errors.

By integrating over $\mathbf{b}$, we can obtain the likelihood, which is defined as:

$$L(\boldsymbol{\beta}, \sigma^2_\epsilon, \mathbf{G} \mid \mathbf{y}, \mathbf{X}, \mathbf{U}) = \mathcal{N}(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \mathbf{U}\mathbf{G}\mathbf{U}^\top + \sigma^2_\epsilon \mathbf{I}). \tag{3.2}$$

3.1.2 Multiple Response Regression

The main drawback to the current modelling techniques implemented within the sociolinguistic community is that mixed effects models, which are implemented do not take into consideration the correlation between dependent variables, namely the first three formants for the same vowel, so they are currently modelled as if they are assumed to be independent. By considering a multiple response linear regression model, we are able to now examine a regression problem where the dependent variable is no longer a single response, but an l length vector of correlated responses, which are defined as $\tilde{\mathbf{y}} = (\mathbf{y}^1, ..., \mathbf{y}^L)^\top$, where each $\mathbf{y}^l$ is an individual response. Like a standard linear regression, there are N observa-

tions, where each observation i consists of P explanatory variables. This can be viewed as a set of l related regression problems for each observation i.

The multiple response regression model is defined as:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{\epsilon} \mid \mathbf{0}, (\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \mathbf{I})\right) \tag{3.3}$$

The regression coefficients are defined as $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^L)^\top$ where each $\boldsymbol{\beta}^l$ is the vector of regression coefficients for the lth response. The corresponding design matrix, defined as $\tilde{\mathbf{X}} = \text{blockdiag}(\mathbf{X}^1, ..., \mathbf{X}^L)$ is constructed in a similar fashion. $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is the m dimensional covariance matrix for the model error.

The likelihood is defined as:

$$L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\epsilon \mid \mathbf{y}, \tilde{\mathbf{X}}) = \mathcal{N}\left(\mathbf{y} \mid \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}, (\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \mathbf{I})\right). \tag{3.4}$$

3.1.3 Bayesian Hierarchical Model

Here, we look to combine the classical mixed effects models with a multiple regression. In addition to this, we also re-express the model in a Bayesian paradigm which helps with the graphical model selection problem detailed in Chapter 6. Within the Bayesian hierarchical model, we also incorporate inter-model selection in order to determine which the most significant social and linguistic factors on impacting vowel formant change.

The likelihood is expressed in a similar fashion to the classical mixed effects model described in Section 3.1.1. The vector of response variables is constructed the same way as in Section 3.1.2, with $\tilde{\mathbf{y}} = (\mathbf{y}^1, ..., \mathbf{y}^L)^\top$, where $l = 1, ..., L$ corresponds to the number of response variables. We denote the current significant social and phonetic factors in the model by $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}$ and their regression coefficients by $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$, where $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}} = \text{blockdiag}(\mathbf{X}^1_{\boldsymbol{\eta}^1}, ..., \mathbf{X}^L_{\boldsymbol{\eta}^L})$ and $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} = \left(\boldsymbol{\beta}^1_{\boldsymbol{\eta}^1}, ..., \boldsymbol{\beta}^L_{\boldsymbol{\eta}^L}\right)^\top$. We also separate the intercepts, $\tilde{\boldsymbol{\beta}}_0$ for each response so they are always included within the model, where $\tilde{\boldsymbol{\beta}}_0 = (\beta^1_0, ..., \beta^L_0)$. The random effects $\tilde{\mathbf{U}}$ and their respective coefficients $\tilde{\mathbf{b}}$ are denoted in a similar fashion to the fixed effects, namely

$\tilde{\mathbf{U}} = \text{blockdiag}\left(\mathbf{U}^1, \ldots, \mathbf{U}^L\right)$ and $\tilde{\mathbf{b}} = \left(\mathbf{b}^1, \ldots, \mathbf{b}^L\right)^\top$. The likelihood is defined as:

$$p\left(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}, \tilde{\mathbf{b}}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}, \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}, \tilde{\mathbf{U}}\right) = \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right). \tag{3.5}$$

The inclusion of the jth column of $\mathbf{X}^l$ is defined by the value of $\eta_j \in \{0, 1\}$, where if the jth variable is included, $\eta_j = 1$, where the corresponding coefficient estimate β_j^l takes some real number value. If the jth variable is removed, $\eta_j = 0$, then the corresponding coefficient estimate is set to $\beta_j^l = 0$. $\mathbf{X}^l_{\boldsymbol{\eta}^l}$ is thus the matrix of relevant explanatory variables, where the number of columns corresponds to the number of non-zero elements in $\boldsymbol{\eta}^l$. $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}$ is constructed by creating a block diagonal matrix which is formed by combining each $\mathbf{X}^l_{\boldsymbol{\eta}^l}$, where $\tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^L)^\top$. The below example provides an illustration of how the notation is used to construct a simple design problem:

$$\mathbf{X}^1 = \begin{pmatrix} x_{1,1}, & x_{2,1}, & x_{3,1} \\ x_{1,2}, & x_{2,2}, & x_{3,2} \\ x_{1,3}, & x_{2,3}, & x_{3,3} \end{pmatrix}; \quad \mathbf{X}^2 = \begin{pmatrix} x_{1,1}, & x_{2,1}, \\ x_{1,2}, & x_{2,2} \\ x_{1,3}, & x_{2,3} \end{pmatrix}; \quad \mathbf{X}^1_{\boldsymbol{\eta}^1} = \begin{pmatrix} x_{1,1}, & x_{3,1} \\ x_{1,2}, & x_{3,2} \\ x_{1,3}, & x_{3,3} \end{pmatrix}; \quad \mathbf{X}^2_{\boldsymbol{\eta}^2} = \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ x_{1,3} \end{pmatrix};$$

$$\boldsymbol{\beta}^1 = \begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix}; \quad \boldsymbol{\beta}^2 = \begin{pmatrix} \beta_1^2 \\ \beta_2^2 \end{pmatrix}; \quad \tilde{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \\ \beta_1^2 \\ \beta_2^2 \end{pmatrix}; \quad \boldsymbol{\beta}^1_{\boldsymbol{\eta}^1} = \begin{pmatrix} \beta_1^1 \\ \beta_3^1 \end{pmatrix}; \quad \boldsymbol{\beta}^2_{\boldsymbol{\eta}^2} = \begin{pmatrix} \beta_1^2 \end{pmatrix}; \quad \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} = \begin{pmatrix} \beta_1^1 \\ \beta_3^1 \\ \beta_1^2 \end{pmatrix} \tag{3.6}$$

$$\boldsymbol{\eta}^1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \boldsymbol{\eta}^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \tilde{\boldsymbol{\eta}} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

In this illustration, we see that $\boldsymbol{\eta}^1$ shows that the second column of the design matrix $\mathbf{X}^1$ is now deemed non-significant, indicating that β_2^1 is now removed from $\boldsymbol{\beta}^1$. Similar is observed for $\boldsymbol{\eta}^2$ where β_2^2 is also removed, leading to the combined $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$

Prior specification

As with the classical mixed effects model in Section 3.1.1, we assume additive i.i.d. Gaussian noise. For the hierarchical model, we assume the errors have precision $\mathbf{\Omega_\epsilon}$. We specify the following conjugate prior on $\mathbf{\Omega_\epsilon}$:

$$\mathbf{\Omega_\epsilon} \sim \mathcal{W}(\nu_\epsilon, \mathbf{S_\epsilon}) \tag{3.7}$$

where ν_ϵ and $\mathbf{S_\epsilon}$ are fixed hyper-parameters that are to be specified.

For the fixed effects coefficients $\tilde{\boldsymbol{\beta}}$, we specify the prior

$$\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\tau} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}^{-1}) \tag{3.8}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)$ and

$$\mathbf{V} = \begin{pmatrix} \tau_1\mathbf{I} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \tau_2\mathbf{I} & \mathbf{0} & \vdots \\ \vdots & & \ddots & \\ \mathbf{0} & \ldots & \mathbf{0} & \tau_L\mathbf{I} \end{pmatrix}$$

The covariance matrix of the fixed effects coefficients is composed of the precision parameters τ_l, which correspond to each response level l, allowing for greater flexibility within the model by permitting specific prior adjustments for each response level. The prior for the precision of the fixed effects coefficients by formant is defined as:

$$\tau_l \sim \mathcal{G}(a_l, b_l) \tag{3.9}$$

with group specific hyperparameters a_l, b_l.

For the random effect parameters $\tilde{\mathbf{b}}_g$, where $\tilde{\mathbf{b}}_g = \left(\mathbf{b}_g^1, \ldots, \mathbf{b}_g^L\right)^\top$, each group follows a Gaussian distributed prior with zero mean and group specific precision matrices $\mathbf{\Omega}_{\tilde{\mathbf{b}}_g}$

$$\tilde{\mathbf{b}}_g \mid \mathbf{\Omega}_{\tilde{\mathbf{b}}_g} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Omega}_{\tilde{\mathbf{b}}_g}^{-1}\right) \tag{3.10}$$

The combined prior for the random effects $\tilde{\mathbf{b}}$ is defined as:

$$\tilde{\mathbf{b}} \mid \boldsymbol{\Omega}_{\tilde{\mathbf{b}}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{-1}\right) \tag{3.11}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\mathbf{b}_1}, \ldots, \boldsymbol{\Omega}_{\mathbf{b}_G}\right)$.

The precision matrices for the random effects have conjugate Wishart prior for each group g:

$$\boldsymbol{\Omega}_{\mathbf{b}_g} \sim \mathcal{W}(\nu_{\mathbf{b}_g}, \mathbf{S}_{\mathbf{b}_g}) \tag{3.12}$$

with group specific hyperparameters $\nu_{\mathbf{b}_g}$ and $\mathbf{S}_{\mathbf{b}_g}$.

A graphical representation of the hierarchical model can be found in Figure 3.1 which details the various levels of input and prior specification of the model.

3.1.4 Bayesian Inference using Markov chain Monte Carlo

In Bayesian inference, the posterior distribution, which is the distribution that contains all the information on the current parameters $\boldsymbol{\theta}$, which for the hierarchical model detailed previously is, $\boldsymbol{\theta} = \left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{b}}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}, \boldsymbol{\Omega}_{\mathbf{b}_g}, \boldsymbol{\tau}\right)^{\top}$, is defined by Bayes theorem. For a given model state and data, $\mathcal{D}$, we define the likelihood to be the probability of $\mathcal{D}$ given the model parameters $\boldsymbol{\theta}$ and model distribution $p(\cdot)$. To obtain the posterior distribution, we multiply the likelihood by the prior distribution on the model parameters, $p(\boldsymbol{\theta})$ and normalise as such:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \propto p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3.13}$$

Markov chain Monte Carlo (MCMC) methods are a group of sampling techniques used to obtain an estimate of a target distribution of interest. They are widely used in Bayesian inference to sample from a posterior distribution of interest for models or to approximate integrals that are extremely difficult or impossible to evaluate. This is performed by sampling values of the parameter of interest, $\boldsymbol{\theta}$, from an approximate distribution and then adjust these draws to better estimate the target posterior, $p(\boldsymbol{\theta} \mid \mathcal{D})$. Each sample is drawn such that the current sample depends only on the previous drawn sample and thus form a Markov chain which, after reaching equilibrium, will effectively

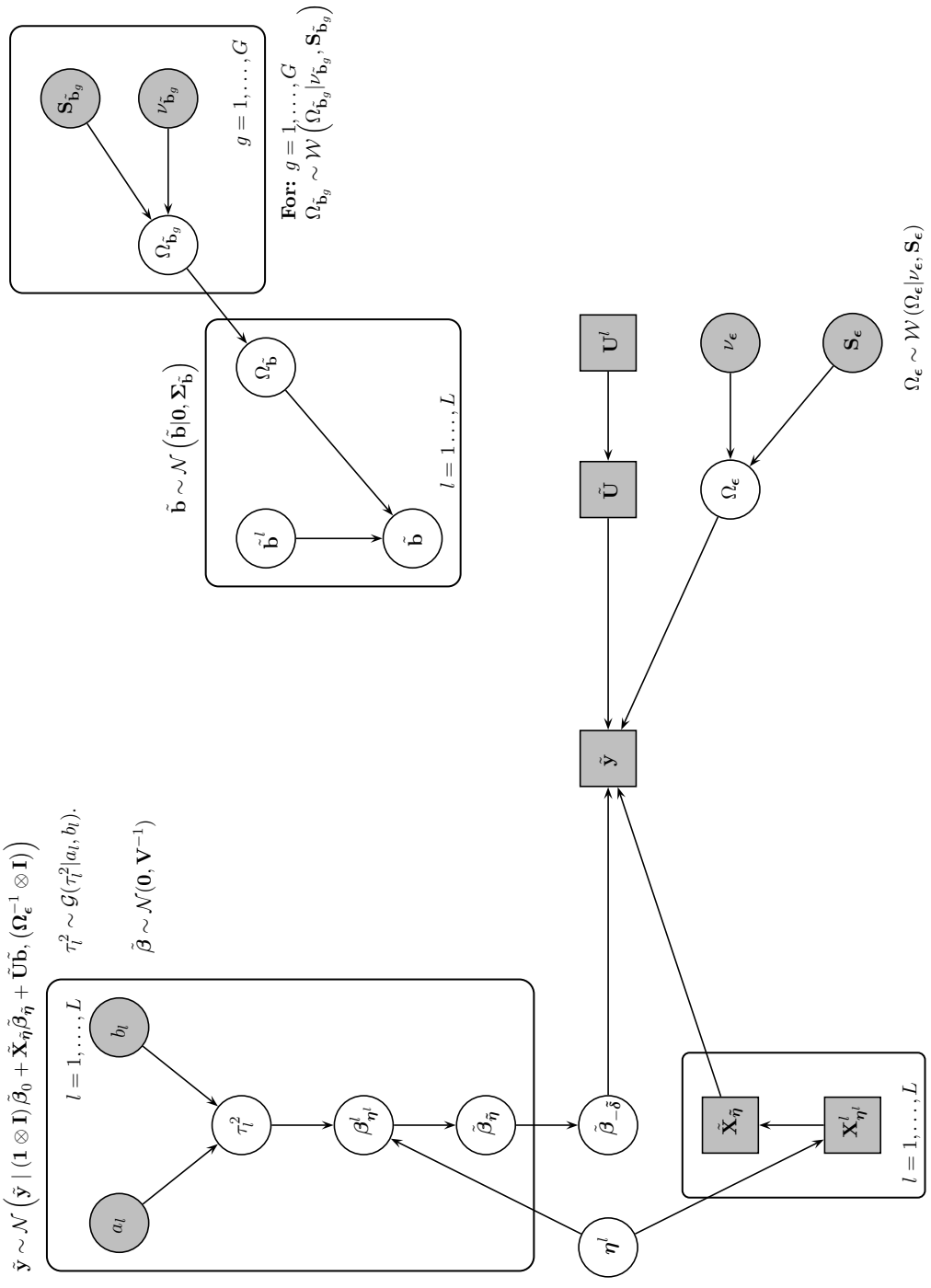

Figure 3.1: Representation of the hierarchical model as a PGM. Nodes which are shaded in grey refer to the fixed hyperparameters and data respectively, whilst nodes in white refer to parameters and hyperparameters that are inferred in the model.

sample from the desired target posterior.

Convergence to the stationary distribution does not occur instantly. As such, we can remove the initial group of samples before convergence has been reached, which is referred to as the burn-in period. Samples can be autocorrelated, leading to poor mixing and approximations of the target distribution. In order to remove this autocorrelation, it is common to take only every ith sample. This is known as thinning.

The two most commonly used MCMC algorithms are the Metropolis-Hastings algorithm and the Gibbs sampler, which is a special case of the Metropolis-Hastings. Within the hierarchical model, we use a combination of both Gibbs and Metropolis-Hastings steps to estimate the model parameters. How the samplers are implemented and constructed with respect to the hierarchical model is discussed in more detail in the following sections.

Gibbs Sampling

Suppose we have a joint distribution $p(\theta_1, \ldots, \theta_k)$ that we wish to sample from. The Gibbs sampler (Geman and Geman, 1984) can be used to sample from this joint distribution, by using the full conditional distributions for each parameter. For a given θ_j, its full conditional is defined as $p(\theta_j \mid \boldsymbol{\theta}_{-j}, \mathcal{D})$. Gibbs sampling requires the conditional distribution to follow a standard distribution. As our parameters within the model follow conjugate priors, all conditional distributions on $\boldsymbol{\theta}$ follow standard distributions which are straightforward to sample from.

For an arbitrary parameter set $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^\intercal$ and data $\mathcal{D}$, the Gibbs sampler works in the following steps:

1. Set initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, ..., \theta_k^{(0)}\right)^\intercal$ to some arbitrary values in the correct parameter space.

2. Generate values of $\boldsymbol{\theta}$ from the respective full conditional distributions for each θ_i as follows:

$$\theta_1^{(1)} \sim p(\theta_1^{(1)} \mid \theta_2^{(0)}, \ldots \theta_k^{(0)})$$
$$\theta_2^{(1)} \sim p(\theta_2^{(1)} \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots \theta_k^{(0)})$$

$$\vdots$$
$$\theta_k^{(1)} \sim p(\theta_k^{(1)} \mid p(\theta_1^{(1)}, \theta_2^{(1)}, \dots \theta_{k-1}^{(1)})$$

3. Repeat step 2 for N iterations of the sampler.

Under reasonable conditions, after a suitable number of samples the algorithm will converge to the target distribution.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (Hastings, 1970), which is a generalisation of the Metropolis algorithm (Metropolis et al., 1953) allows us to make draws from any probability distribution, given the target distribution can be computed at a given value. The acceptance and rejection step is based on the ratio of the posterior and proposal distributions. The parameters are then updated through the MCMC chain.

Given the current sample of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_{(t)}$, we propose a new value, $\boldsymbol{\theta}^*$. This is done using the proposal distribution $q(\boldsymbol{\theta}^* \mid \cdot)$. For continuous data, this distribution will often be centred around the previous value of the chain, $\boldsymbol{\theta}_{(t-1)}$. The distribution of $q(\boldsymbol{\theta}^* \mid \cdot)$ can be of any form, though it must be carefully chosen to improve convergence speed.

The Metropolis-Hastings algorithm works as follows:

1. Begin with initial state $\boldsymbol{\theta}_{(0)}$
2. For t in $1, \dots, T$
   (a) Given the current state $\boldsymbol{\theta}_{(t)}$, sample a new candidate state $\boldsymbol{\theta}^*_{(t)}$ from $q(\boldsymbol{\theta}^*_{(t)} \mid \boldsymbol{\theta}_{(t)})$.
   (b) Calculate the acceptance ratio

$$r = \frac{p\left(\boldsymbol{\theta}^*_{(t)}\right) q\left(\boldsymbol{\theta}_{(t)} \mid \boldsymbol{\theta}^*_{(t)}\right)}{p\left(\boldsymbol{\theta}_{(t)}\right) q\left(\boldsymbol{\theta}^*_{(t)} \mid \boldsymbol{\theta}_{(t)}\right)}$$

   (c) Generate random $u \sim \mathcal{U}(0,1)$, accepting $\boldsymbol{\theta}^*_{(t)}$ if $u < r$. Otherwise, we remain at the current state $\boldsymbol{\theta}_{(t)}$.

3.1.5 Variable Selection

Within our sampler, we implement a model space based approach for variable selection. This approach works by viewing the model space as a whole and placing priors on the number of covariates selected within the model as opposed to placing priors on the individual covariates.

A model space based approach can be implemented by using a Reversible Jump Markov Chain Monte Carlo (RJMCMC). RJMCMC is a technique used for model selection (Green, 1995), which allows the Markov chain to explore model spaces of different dimension. In terms of variable selection, the selected variables are denoted by $\tilde{\boldsymbol{\eta}}$ as shown previously. The model is updated by randomly selecting a variable η_j and then proposing either addition or removal of the selected variable, which translates to either $\eta_i^l = 1$ if the variable is added to the model or $\eta_i^l = 0$ if the variable is removed from the model.

The length of $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ is not fixed, but instead varies throughout the MCMC process, dependent on the current state of $\tilde{\boldsymbol{\eta}}$. The update is performed using a Metropolis-Hastings step, with the acceptance ratio adjusted for the change in dimension.

For a given model state, say $\tilde{\boldsymbol{\eta}}^*$, we can compute the marginal likelihood of the data under this model by the following integral:

$$p(\mathbf{y} \mid \mathbf{X}, \tilde{\boldsymbol{\eta}}^*) = \int p\left(\boldsymbol{y} \mid \tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^*}, \tilde{\mathbf{b}}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}, \tilde{\mathbf{X}}, \tilde{\mathbf{U}}\right) p(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^*} \mid \tilde{\mathbf{X}}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}})\, d\tilde{\beta}_0\, d\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}\, d\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \tag{3.14}$$

Given our newly proposed model state defined by $\tilde{\boldsymbol{\eta}}^1$, the Metropolis-Hastings step computes the ratio of the log marginal likelihoods between the proposed model state $\tilde{\boldsymbol{\eta}}^1$ and the current model state $\tilde{\boldsymbol{\eta}}^0$:

$$\alpha = \frac{p(\mathbf{y} \mid \mathbf{X}, \tilde{\boldsymbol{\eta}}^1)}{p(\mathbf{y} \mid \mathbf{X}, \tilde{\boldsymbol{\eta}}^0)} \tag{3.15}$$

The new model state $\tilde{\boldsymbol{\eta}}^1$, is accepted if $u < \alpha$ where $u \sim \mathcal{U}(0, 1)$ and the relevant coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^1}$ are updated.

3.1.6 Posterior inference

In order to explore the posterior distributions of the hierarchical model, we use an MCMC algorithm, implementing a combination of the methods discussed in Sections 3.1.4 & 3.1.5. Throughout, we have chosen conjugate priors within the model so we can implement Gibbs sampling for the majority of the sampler. The only part of the sampler where we extend beyond using Gibbs is with the model selection, where we implement a Metropolis step for the fixed effects variable selection.

Here, we detail the conditional distributions for all the parameters in the model which are sampled using Gibbs sampling. The step for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^1}$ is split into two steps. Firstly, the current model state is chosen using the RJMCMC step, then secondly we update $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^1}$ using a standard Gibbs step, conditioned on the current model state $\tilde{\boldsymbol{\eta}}^1$.

In order to verify the conditional distributions we obtain, joint distribution tests are implemented, as proposed in Geweke (2004). The motivation behind joint distribution tests is to draw P sets of model parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ from the model's relevant prior distributions. These parameter sets are then used to generate P datasets $\mathcal{D}_1, \ldots, \mathcal{D}_P$. For each combination of parameters and datasets and under the same model and prior specifications, we can run the MCMC sampler to sample from each of the posterior distributions $p(\boldsymbol{\theta}_p \mid \mathcal{D}_p)$ for the P generated datasets. From each of these MCMC chains for each posterior distribution, we can then draw N independent samples of the model parameters $\boldsymbol{\theta}_{p,1}, \ldots, \boldsymbol{\theta}_{p,N}$. In order to determine whether the MCMC samples are sampling from the correct posterior distribution, the next step is to confirm whether the generated samples $\boldsymbol{\theta}_{p,n}$ for $p = 1, \ldots, P$ and $n = 1, \ldots, N$ follow their corresponding prior distribution made to generate the parameter as follows:

$$\frac{1}{P}\sum_{p=1}^{P} p(\boldsymbol{\theta} \mid \mathcal{D}_p) \approx \int p(\boldsymbol{\theta} \mid \boldsymbol{\mathcal{D}})p(\boldsymbol{\mathcal{D}})d\boldsymbol{\mathcal{D}} = \int p(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta})d\boldsymbol{\mathcal{D}} = p(\boldsymbol{\theta}) \tag{3.16}$$

If this follows for a significantly large enough P and N then we can deduce that the MCMC sampler is sampling from the posterior correctly. We have used joint posterior tests to verify all the samplers constructed throughout this research.

We detail the conditional distributions for each parameter within the model. The full derivations of these distributions are detailed further in Appendix A. Using $\boldsymbol{\theta}$ to define those parameters to be conditioned on, where for example $\boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}}$ means 'condition on all parameters excluding $\tilde{\mathbf{b}}$' we obtain:

$$\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} \mid \boldsymbol{\theta}_{\setminus \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} \mid \left[\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}} + \mathbf{V}\right]^{-1} \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{y}}_{\tilde{\boldsymbol{\beta}}}, \left[\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}} + \mathbf{V}\right]^{-1}\right) \tag{3.17}$$

$$\tilde{\mathbf{b}} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}} \propto \mathcal{N}\left(\tilde{\mathbf{b}} \mid \left[\tilde{\mathbf{U}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{U}} + \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}\right]^{-1} \tilde{\mathbf{U}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{y}}_{\tilde{\mathbf{b}}}, \left[\tilde{\mathbf{U}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tilde{\mathbf{U}} + \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}\right]^{-1}\right) \tag{3.18}$$

$$\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}_{\tilde{\mathbf{b}}_g}^{-1} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i} \tilde{\mathbf{b}}_{g_i}^{\top}\right]^{-1}\right) \tag{3.19}$$

$$\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}_{\boldsymbol{\epsilon}}^{-1} + \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\top}\right]^{-1}\right) \tag{3.20}$$

$$\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l^2} \propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}_{\tilde{\boldsymbol{\eta}}_l}^l\|}{2}, b_l + \frac{\sum_{m=1}^{p} (\beta_m^l)^2}{2}\right) \tag{3.21}$$

where we sample $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g respectively, and τ_l for each response level l.

We define $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \otimes \mathbf{I})$, $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_1}, \ldots, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_G}\right)$, $\tilde{\mathbf{y}}_{\tilde{\boldsymbol{\beta}}} = \mathbf{y} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$, $\tilde{\mathbf{y}}_{\tilde{\mathbf{b}}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$ respectively. The distributions can be sampled in any order, where each update uses the most recent version of the conditioned parameters.

The parameters are sampled using the following algorithm:

Algorithm 1: The Bayesian hierarchical model sampler Given initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\mathbf{b}}^{(0)}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(0)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(0)}, \boldsymbol{\tau}^{(0)}\right)$. Then
For $t = 1, \ldots, T$

1. Sample $\tilde{\boldsymbol{\beta}}^{(t)}$ from 3.17.

2. Propose new model state $\tilde{\boldsymbol{\eta}}^{(t)}$. Sample $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^{(t)}}$ from 3.17. Compute 3.15, where $\tilde{\boldsymbol{\eta}}^0$

is the current model state. If $u < \alpha$, where $u \sim \mathcal{U}(0,1)$, set $\tilde{\boldsymbol{\beta}}^{(t)} = \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}^{(t)}}$, else $\tilde{\boldsymbol{\beta}}^{(t)}$ remains the same.

3. Sample $\tilde{\mathbf{b}}^{(t)}$ from 3.18.

4. For $g = 1, \ldots, G$,

   Sample $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{(t)}$ from 3.19.

   Form $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)}$ by $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_1}^{(t)}, \ldots, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_G}^{(t)}\right)$.

5. Sample $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(t)}$ from 3.20.

6. For $l = 1, \ldots, L$,

   Sample $\tau_l^{(t)}$ from 3.21.

   Form $\boldsymbol{\tau}^{(t)} = \left(\tau_1^{(t)}, \ldots, \tau_L^{(t)}\right)$

Looking closer at the posteriors derived for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$, we observe for cases where we have a significantly high number of response variables l, that this step can be somewhat computationally expensive for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$, particularly in terms of the inverses being calculated in the posteriors. When we have a large number of random effects groups g or levels h, we observe similar cost in terms of computation.

In an attempt to bypass such intense calculations, we propose adjustments to the way we sample from the posteriors for both $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$. Instead of sampling all the parameters in one block, we instead sample for $\boldsymbol{\beta}_{\tilde{\boldsymbol{\eta}}}^{l}$ and $\tilde{\mathbf{b}}_{g,h}$, where $h = 1, \ldots, H$ is the number of levels of the corresponding random effect g. Figure 3.2 shows how this sampler modification works visually in terms of the levels of the hierarchical model.

The updated posterior distributions are of the form:

$$\boldsymbol{\beta}_{\boldsymbol{\eta}^l}^{l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}_{\boldsymbol{\eta}^l}^{l}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j}\mathbf{X}_{\boldsymbol{\eta}^l}^{\top}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\mathbf{X}_{\boldsymbol{\eta}^l}^{\top}\mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j}\mathbf{X}_{\boldsymbol{\eta}^l}^{\top}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\right) \tag{3.22}$$

$$\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} \propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right) \tag{3.23}$$

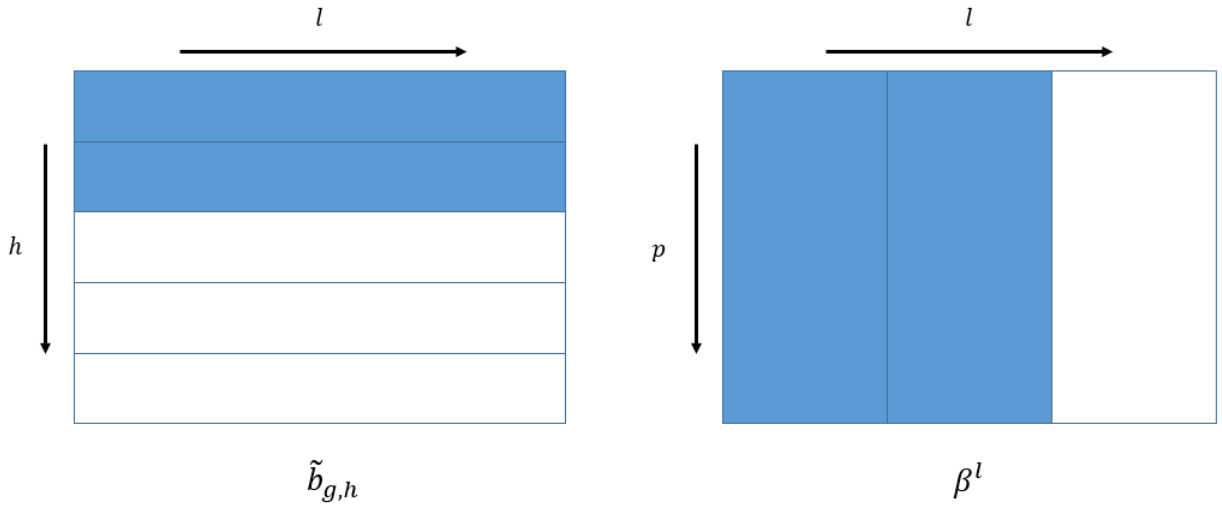

Figure 3.2: Illustration of modified sampler steps. Illustration of modified sampler steps for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$. We observe that we now sample for each $\boldsymbol{\beta}^l$, splitting the sampler up by each response level l. We also now sample by each group $\tilde{\mathbf{b}}_g$, but also sampling for each level of the random effect h, so sampling each $\mathbf{b}_{g,h}$ in turn.

$$\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}_{\tilde{\mathbf{b}}_g}^{-1} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i} \tilde{\mathbf{b}}_{g_i}^{\top}\right]^{-1}\right) \tag{3.24}$$

$$\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}_{\boldsymbol{\epsilon}}^{-1} + \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\top}\right]^{-1}\right) \tag{3.25}$$

$$\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} \propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}^l_{\tilde{\boldsymbol{\eta}}_l}\|}{2}, b_l + \frac{\sum_{m=1}^{p} \left(\beta_m^l\right)^2}{2}\right) \tag{3.26}$$

where $\mathbf{z}_{\boldsymbol{\beta}^l} = \omega_{j,j}\mathbf{y}^l + \sum_{k=1}^{k \neq l} \omega_{j,k}\left(\mathbf{y}^k - \mathbf{X}_{\boldsymbol{\eta}^k}\boldsymbol{\beta}^k\right)$ and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} = \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}_{\tilde{\mathbf{b}}_{-g}}\tilde{\mathbf{b}}_{\tilde{\mathbf{b}}_{-g}}$, where $\tilde{\mathbf{b}}_{-g}$ denotes $\tilde{\mathbf{b}}$ excluding group g and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}$ is the mean value calculated for $\mathbf{y}_{\tilde{\mathbf{b}}_{g,h}}$ for each response level l. Model selection is now performed on each level of $\boldsymbol{\beta}^l$ in turn. By sampling in this fashion, we avoid the computation of large inverses of matrices and improve the computational performance of the sampler.

The parameters are sampled using the following algorithm:

Algorithm 2: The Bayesian hierarchical model sampler - Computationally efficent version Given initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\mathbf{b}}^{(0)}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(0)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(0)}, \boldsymbol{\tau}^{(0)}\right)$. Then

For $t = 1, \ldots, T$

1. For $l = 1, \ldots, L$,

   (a) Sample $\boldsymbol{\beta}^{l,(t)}$ from 3.22.

   (b) Propose new model state $\boldsymbol{\eta}^{l,(t)}$. Sample $\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$ from 3.22. Compute 3.15, where $\boldsymbol{\eta}^{l,0}$ is the current model state. If $u < \alpha$, where $u \sim \mathcal{U}(0,1)$, set $\boldsymbol{\beta}^{l,(t)} = \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$, else $\boldsymbol{\beta}^{l,(t)}$ remains the same.

   Form $\tilde{\boldsymbol{\beta}}^{(t)}_{\tilde{\boldsymbol{\eta}}^{(t)}} = \left(\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{1,(t)}}, \ldots, \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{L,(t)}}\right)$

2. For $g = 1, \ldots, G$

   For $h = 1, \ldots, H$

   (a) Sample $\tilde{\mathbf{b}}^{(t)}_{g,h}$ from 3.23.

   Form $\tilde{\mathbf{b}}^{(t)}_{g} = \left(\tilde{\mathbf{b}}^{(t)}_{g,1}, \ldots, \tilde{\mathbf{b}}^{(t)}_{g,H}\right)^{\top}$

   Form $\tilde{\mathbf{b}} = \left(\tilde{\mathbf{b}}^{(t)}_{1}, \ldots, \tilde{\mathbf{b}}^{(t)}_{G}\right)^{\top}$

3. For $g = 1, \ldots, G$,

   Sample $\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_g}$ from 3.24.

   Form $\boldsymbol{\Sigma}^{(t)}_{\tilde{\mathbf{b}}}$ by $\boldsymbol{\Sigma}^{(t)}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_1}, \ldots, \boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_G}\right)$.

4. Sample $\boldsymbol{\Omega}^{(t)}_{\boldsymbol{\epsilon}}$ from 3.25.

5. For $l = 1, \ldots, L$,

   Sample $\tau^{(t)}_{l}$ from 3.26.

   Form $\boldsymbol{\tau}^{(t)} = \left(\tau^{(t)}_{1}, \ldots, \tau^{(t)}_{L}\right)$

3.2 Analysis using the Bayesian hierarchical model

In this section, we consider two examples to demonstrate the hierarchical model. The first example is a simulated example which aims to test how well variable selection performs in the model framework, followed by an application to the Sounds of the City corpus. All model code has been implemented in the statistical programming language *R* (R Core Team, 2018).

3.2.1 Simulation Study

Here, we consider a simple toy example which looks to demonstrate the effectiveness of the model selection RJMCMC step detailed in Section 3.1.5.

We construct a simple problem with simulated data consisting of 12 fixed effects and two random effects with 1,000 observations. The model is of the form:

$$y_{ijk}^{l} = \mathbf{x}_{ijk}^{\top}\boldsymbol{\beta}^{l} + b_{1,j}^{l} + b_{2,k}^{l} + \epsilon_{ijk}^{l} \tag{3.27}$$

We construct the regression coefficients such that 75% of the coefficients for each response level l were drawn from $\boldsymbol{\beta}_1 \sim \mathcal{N}(10, 1)$ and the remaining 25% drawn from $\boldsymbol{\beta}_2 \sim \mathcal{N}(0, 0.001)$. From this, each response y_i was then generated from the model using the aforementioned regressors, with additive Gaussian noise drawn from $\mathcal{N}(\mathbf{0}, 0.1 \cdot \mathbf{I})$.

For this simulated data, we implement the Bayesian hierarchical model, where we sample 10,000 draws, with a burn-in period of 100 iterations. The hyperparameters are fixed to the following values to give vague prior distributions: $a_l = b_l = 0.001$ and $\nu_{\boldsymbol{\epsilon}} = 4, \mathbf{S}_{\boldsymbol{\epsilon}} = 0.001 \cdot \mathbf{I}$ and $\nu_{\mathbf{b}_g} = 4, \mathbf{S}_{\mathbf{b}_g} = 0.001 \cdot \mathbf{I}$.

Figure 3.3 shows the density plots obtained for the $\boldsymbol{\beta}$ coefficients for the response level $\mathbf{y}^1$. As we have simulated the data we know the true values of the parameter estimates, which are illustrated in each density plot by the red vertical line.

As we can observe from the density plots, the model does well to estimate the parameters in general, and identifies those coefficients that are sampled from $\boldsymbol{\beta}_2$ extremely well, as observed by the sharp spike in their respective densities. We observe multimodality for some of the density plots. The reason for this is the occasional addition of coefficients

that are sampled from $\boldsymbol{\beta}_2$, most notably the estimate for β_6^1, where we observe a mode around 0.2. When this term is included within the model, the remaining coefficients present adjust for the addition of this term, hence the occasional switch in mode due to the correlation between the covariates.

3.2.2 Sounds of the City Corpus

Here, we look at the Sounds of the City corpus as discussed in Section 2.2. The aim of this analysis is to determine what factors may be conditioning variation and change for the *FLEECE, FACE, TRAP,BATH, LOT, GOAT* and *FOOT/GOOSE* vowels within the corpus. Vowel formant measurements on F1, F2 and F3 are taken as the response variables of interest, with models fitted to raw mean vowel formant values for F1, F2 and F3 and Lobanov normalised values for F1 and F2. Random effects are taken for each individual Speaker and choice of Word. Social variables related to the Speaker are taken as fixed effects, with terms relating to Gender of speaker, Decade of recording and Age of speaker taken as predictors within the model alongside word-specific variables relating to the place of articulation of the preceding and following consonant between the vowel utterance within a specific word.

The model equation is of the form

$$y_{ijk}^l = \mathbf{x}_{ijk}^\top \boldsymbol{\beta}^l + \gamma_j^l + \delta_k^l + \epsilon_{ijk}^l \tag{3.28}$$

where we define the formant measures, raw mean or Lobanov normalised, as the response y_{ijk}^l, where y_{ijk}^l is the k^{th} measurement from the j^{th} word of the i^{th} speaker on the l^{th} formant. $\mathbf{x}_{ijk}^\top$ is taken as the vector of explanatory variables containing properties of vowel quality which are attributable to individual speaker and word variation for speaker i and word j and also a level for the population intercept. $\boldsymbol{\beta}^l$ is the corresponding vector of regression coefficients. We define the random effect for speaker by γ_i^l and the word random effect by δ_j^l. Shorthand notations for each of the model parameters are $\boldsymbol{\beta} = \{\boldsymbol{\beta}^l\}_{l=1}^L$; $\boldsymbol{\gamma} = \{\gamma_i^l\}_{i=1}^{I\ \ L}{}_{l=1}$ and $\boldsymbol{\delta} = \{\delta_i^l\}_{j=1}^{J\ \ L}{}_{l=1}$. The precision matrix for the residuals is defined as $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}$ and corresponding precision matrices for the random effects are $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\delta}}$ respectively.

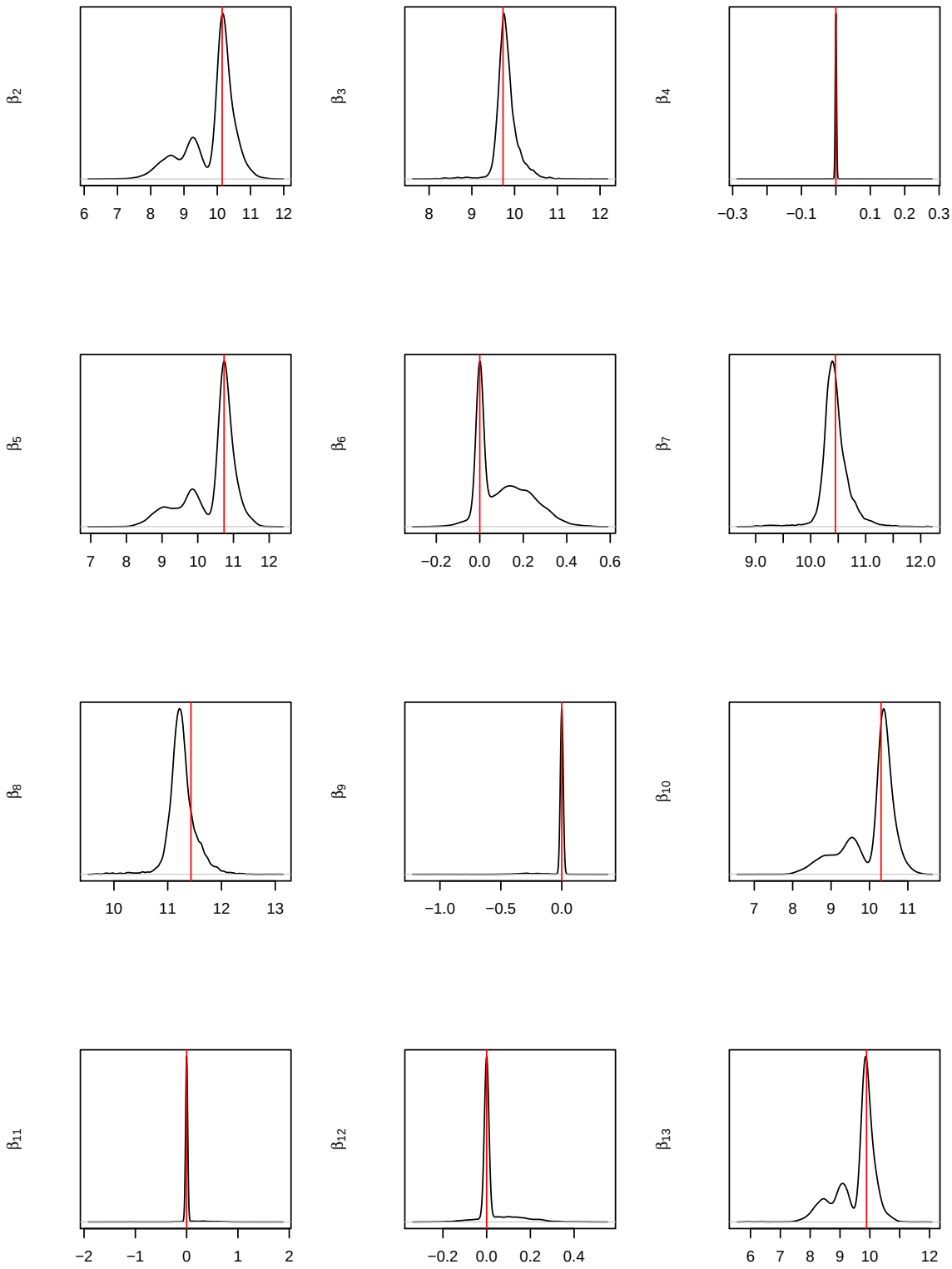

Figure 3.3: Density plots for coefficients from the simulated study Density estimates for the fixed effects coefficients for $\mathbf{y}^1$. We see the coefficients are estimated well from their known values, with the $\boldsymbol{\beta}_2$ coefficients correctly not selected within the model.

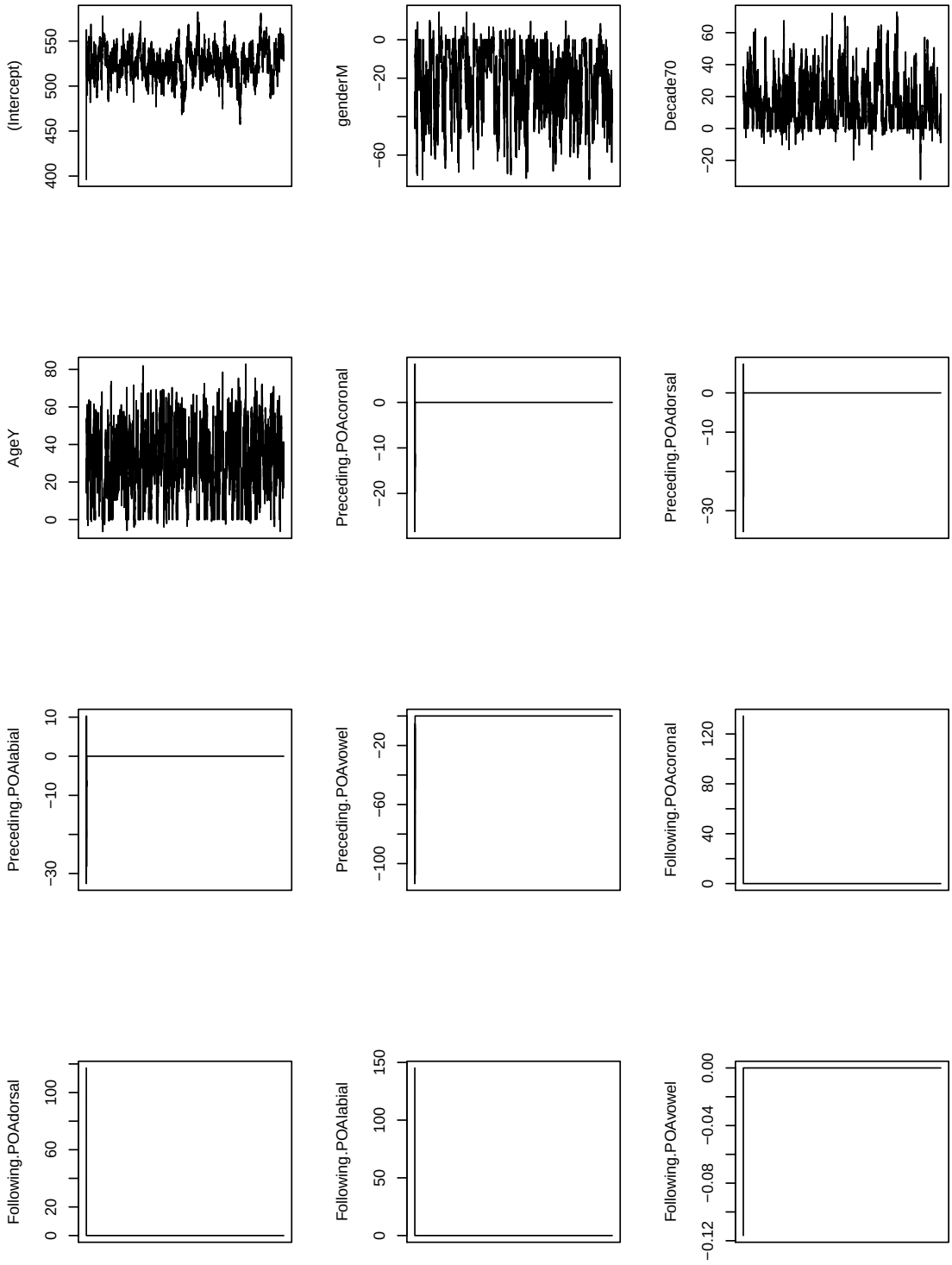

Figure 3.4: Trace plots for the *LOT* vowel model. Trace plots for the fixed effects coefficients obtained from the *LOT* vowel for the F1 raw mean values ran for 10,000 iterations. Poor mixing can be observed for the active terms within the model.

Re-expressing Equation 3.28 in the form of the Bayesian hierarchical model, we denote $\tilde{\mathbf{b}}_g = \left(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\delta}}\right)^\top$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\boldsymbol{\gamma}}, \boldsymbol{\Omega}_{\boldsymbol{\delta}}\right)$.

For each vowel, we sample 10,000 iterations of the MCMC to obtain parameter estimates, on both the raw mean vowel formants and the Lobanov normalised formants. The hyperparameters are fixed in order to give vague prior distributions. Their corresponding values are: $a_l = b_l = 1 \times 10^{-4}$ and $\nu_{\boldsymbol{\epsilon}} = 3, \mathbf{S}_{\boldsymbol{\epsilon}} = 0.001 \cdot \mathbf{I}$ and $\nu_{\mathbf{b}_g} = 3, \mathbf{S}_{\mathbf{b}_g} = 0.001 \cdot \mathbf{I}$.

Focusing on the *LOT* vowel model obtained for raw mean formant measures on F1, F2 and F3, we look closer at the output obtained for this model. Time series plots for the fixed effects parameter estimates for F1 are shown in Figure 3.4. Starting from the saturated model, we observe that the variable selection has within the initial number of iterations removed preceding place of articulation and following place of articulation from the model.

The social factors of Gender, Age and Decade of recording are all conditioning variation and change on F1. F1 has smaller values for 2000s speakers meaning that vowel quality has raised over time (i.e. now sounding less like LOT, and a bit more like GOAT). But there are also two other findings, males in general have smaller values, so they show more raised LOT vowels. Young speakers have higher F1 values, i.e. more open vowels. To properly understand how Decade and Age work, a further analysis with an interaction would be needed here. We also observe on occasion, the removal of each term, when the time series plot 'flattens' on 0, indicating the term has been removed from the active model.

Looking closer at the difference observed between the hierarchical model fit for all formants compared to modelling them all individually, we fit models to the *LOT* vowel for each formant individually, effectively assuming independence between the formants. Table 3.1 details the terms that were selected by the multiple response hierarchical model and also the terms selected when each formant was fitted individually, assuming independence between the formants.

From the results in Table 3.1, we observe that the multiple response model selects the same terms for F1 and F3 as the models for both formants fitted individually. The main difference we observe is in F2, where the preceding place of articulation is included in the single response model, unlike the multiple response case. This is a common occurrence

for results throughout the Sounds of the City corpus for different vowels, where the hierarchical model for multiple responses often produces a model that is more parsimonious than the models where only one formant is modelled at a time.

Table 3.1: Significance of coefficients for fixed effects in the multiple response model and independent model All coefficients selected for the full multiple response model and the individual single response models for raw mean formant measurements for F1, F2 and F3 for the *LOT* vowel. We observe that Preceding place of articulation is included for the F2 model in the univariate case as opposed to the multiple model.

	Multiple			Independent		
	F1	F2	F3	F1	F2	F3
GenderM	✓	✓	✓	✓	✓	✓
Decade70	✓	✗	✓	✓	✗	✓
AgeY	✓	✗	✗	✓	✗	✗
FollowingPOA	✗	✓	✗	✗	✓	✗
PrecedingPOA	✗	✗	✗	✗	✓	✗

If we look closely at the trace plots for the active terms in Figure 3.4, we observe that the variables are mixing quite poorly. This poor mixing indicates that high autocorrelation is present within the samples, meaning that the number of MCMC samples is not actually a good indicator of the amount of observed 'data' we have from the posterior distributions.

In order to obtain a better idea of how many efficient samples we actually draw from the MCMC, we can use the effective sample size (ESS) (Priestley, 1981). The ESS can be interpreted as the number of independent Monte Carlo samples necessary to give the same precision as the MCMC estimator. For example, we could have 1,000 samples from a Markov chain that are the equivalent of 80 independent samples due to the MCMC samples being highly correlated. Conversely we could have 1,000 samples from a different Markov chain are the equivalent of 600 independent samples because although the MCMC samples are dependent, in this sampler they are weakly correlated.

The ESS is defined as:

$$\text{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho(k)} \tag{3.29}$$

where N is the number of MCMC samples and $\rho(k)$ is the correlation at lag k.

If the samples are all independent, the ESS will be the the same as the actual sample size N.

For the *LOT* vowel, we compute the ESS for each of the parameters that are active in the model. These results can be found in Table 3.2.

Table 3.2: Effective sample size (ESS) values for active fixed effects parameters for the *LOT* vowel on raw formant measures for F1, F2 and F3. The model was run for 10,000 iterations. We observe a poor ESS for all of the variables due to the high correlation between samples.

	F1	F2	F3
GenderM	156.09	459.79	93.70
Decade70	134.65	-	79.23
AgeY	214.45	-	-

As we can see from Table 3.2, the ESS for the sampled parameters is very low when compared to the number of iterations, 10,000, with effectively only 5% of the 10,000 iterations being considered effective samples. This suggests the samples are highly correlated and that the MCMC chain should be run for a longer number of iterations to obtain a larger independent sample. From a practical point of view this is computationally expensive. One of the main aims of this work is to encourage the sociolinguistic community to implement the models outlined in this thesis. In order to encourage such use, we want the model to be as computationally efficient as possible. Currently, a run of 100,000 iterations would take over 2 hours to run, which is a significantly longer run time when compared to the run time of *lme4* (Bates et al., 2015).

If we look closer at the design of the Sounds of the City corpus, we observe that the fixed effects are nested within the random effects. For example, the Gender, Age and Decade coefficients are all Speaker-dependent variables and are nested within the random effect. This leads to the highly correlated chains we observe and the resulting poor mixing. The problem is not only limited to the mixing of the fixed effects parameters; we also observe poor mixing in the random effects as seen in Figure 3.5. We also obtain similarly poor ESS estimates which are shown in Table 3.3. Note the improvement for F2 in terms of ESS compared to F1 and F3. This is due in part to only the Gender

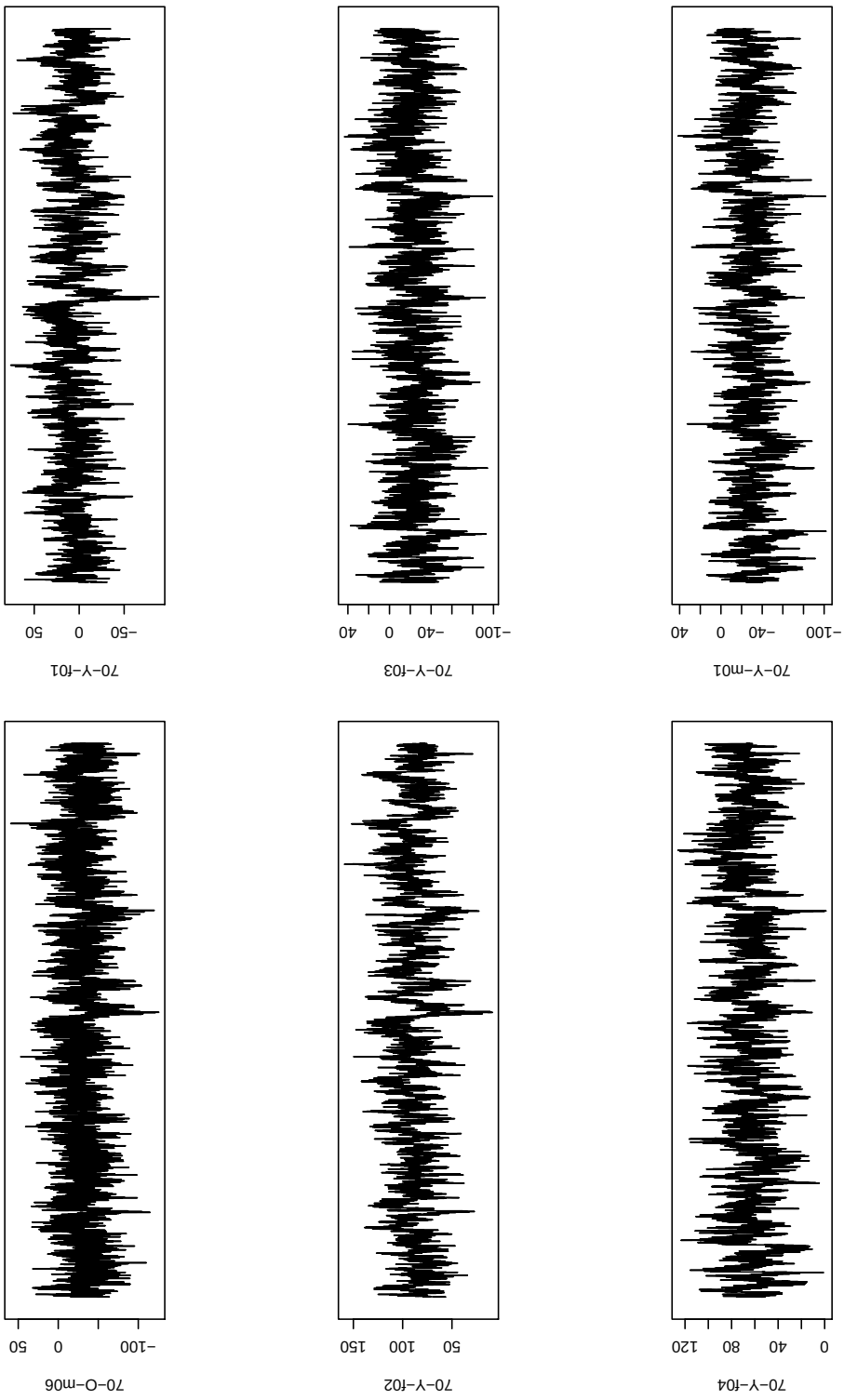

Figure 3.5: Trace plots for the *LOT* vowel model Speaker effect. Trace plots for the Speaker random effect from the *LOT* vowel model for the first six levels for raw mean formant measurements on F1 for 10,000 iterations. We observe similar poor mixing as the fixed effects in Figure 3.4 due to the nested design of the data.

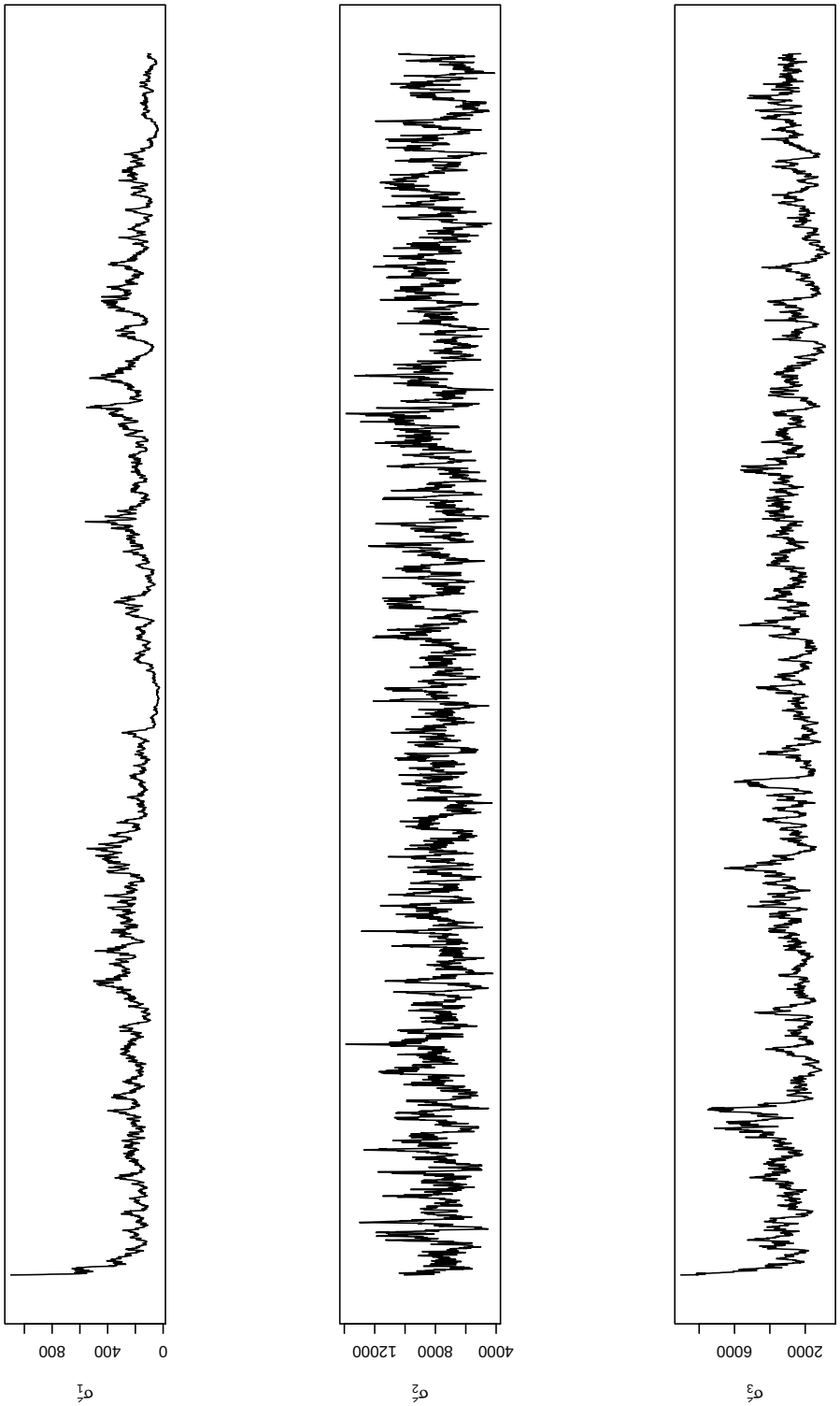

Figure 3.6: Trace plots for the *LOT* vowel model Word effect variance. Trace plots for the Word random effect variance from the *LOT* vowel model for F1, F2 and F3 run for 10,000 iterations. We observe very poor mixing and periods where the sampler becomes stuck at values close to zero.

coefficient being selected for F2, and not also the Decade and Age coefficients, so mixing here is improved due to the reduced number of parameters, thus reducing the correlation between parameters.

Table 3.3: Effective sample size (ESS) values for the Speaker random effect for the first six levels for the *LOT* vowel. Like Table 3.2, we observe a poor ESS for the random effects levels due to the nested design of the data.

	F1	F2	F3
70-O-m06	254.2	2505.3	158.2
70-Y-f01	172.2	1121.9	128.2
70-Y-f02	154.6	797.0	114.1
70-Y-f03	229.2	1692.6	146.3
70-Y-f04	220.3	1168.5	137.2
70-Y-m01	210.7	1197.9	133.2

Another mixing issue can be observed in the precision estimates for the Word random effect. If we look closely at the precision estimates, poor mixing can be observed. Figure 3.6 shows the precision trace plots for F1, F2 and F3. We observe poor mixing for each of the formants and periods where the sampler appears to get stuck at values which are relatively low. This in turn causes poor mixing to the Word random effect coefficients. The trace plots for a sample of Word coefficients is shown in Figure 3.7, where we observe relatively poor mixing, with narrowing and widening of the trace around zero values due to the poor mixing of the variance parameters. Table 3.4 details the ESS for each precision estimate by each formant. We see that the values are extremely low, as would be expected from the trace plots.

Table 3.4: Effective sample size (ESS) values for precision estimates from the Word random effect for F1, F2 and F3 for the *LOT* vowel from the Bayesian hierarchical model ran for 10,000 iterations.The ESS observed is extremely poor for all formant measurements due to the sampler becoming frequently stuck at values close to zero.

	F1	F2	F3
$\omega_{\mathbf{b}_{word}}$	482.2	618.4	395.2

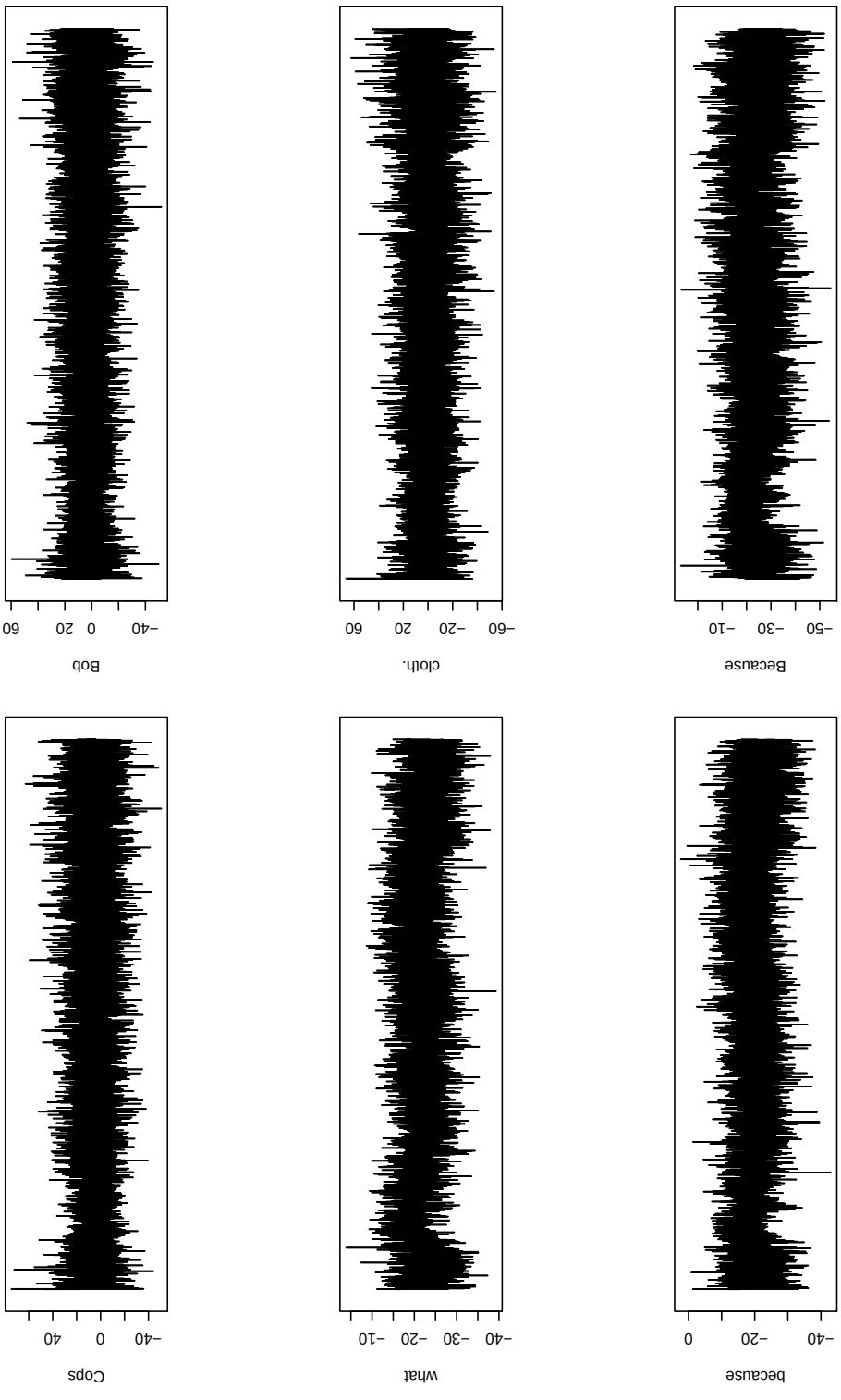

Figure 3.7: Trace plots for the *LOT* vowel model Word effect coefficients. Trace plots for the Word random effect coefficient for a sample of six words for the *LOT* vowel on raw mean formant measurements for F1 for 10,000 iterations. We observe minor narrowing and widening of the chain due to the poor mixing in the precision estimates.

Looking closer at the Word random effect for the LOT vowel, we observe that there are 490 levels over 2,431 observations. Several of the levels are only observed once, which is due to the corpus using spontaneous speech recordings, so it is quite common for one Word to only be uttered once. Due to this, we have a lack of available information on certain levels. This causes the precision to get stuck at values which are very low and causes very high autocorrelation.

In Chapter 4, we will discuss and implement methods which will aim to reduce this autocorrelation found within the MCMC chains, which will lead to improved ESS values resulting in fewer iterations of the MCMC and an improved computational performance time.

3.3 Discussion

In this chapter, we have proposed a Bayesian hierarchical model with the capability to model multiple response variables simultaneously, providing a new modelling approach for the sociolinguistic community to extend upon the current single response mixed effects models that are implemented (Johnson, 2009). The hierarchical model performs well in terms of model selection, as shown by the simulated example in Section 3.2.1, correctly selecting terms of significance. The models obtained can be more parsimonious than modelling the vowel formant measurements individually and provide a more accurate representation due to the extra information sharing between the formants, and their natural correlation.

Several drawbacks have been identified within the hierarchical model when applied to the Sounds of the City corpus in terms of poor mixing of certain parameters. This is due to imbalanced the nested design of the Sounds of the City corpus, which in turn leads to high autocorrelation within the MCMC sampler. An obvious solution is to run the MCMC sampler for a longer number of iterations, so we can obtain a larger sample of independent samples. The drawback to this approach is the significant increase in computational time, which is not practical when compared to *lme4*, which is the standard functionality used.

In order to deal with the high autocorrelation and maintain a reasonable computational time, we look at adapting reparametrisation methods in Chapter 4 which aim to

reduce this autocorrelation and obtain a larger proportion of independent samples from the MCMC chains obtained from the Bayesian hierarchical model.

Chapter 4

Using Reparameterisation Methods to Improve Mixing Within the Hierarchical Model

In this chapter, we aim to resolve the mixing issues highlighted in Chapter 3 which lead to high autocorrelation within MCMC chains. The reason we observe such poor mixing when applying the hierarchical model to the Sounds of the City corpus is due to the nested design of the dataset. The fixed effects of Decade of recording, Gender and Age of speaker are nested within the Speaker random effect. We also have nesting between the Following and Preceding place of articulation and the Word random effect. Also highlighted was the poor mixing of the precision estimates for the Word random effect, with the sampler often becoming stuck at values close to zero.

Reparameterisation schemes can be implemented within MCMC samplers to improve issues with mixing. In this chapter, we introduce two such methods to help alleviate the mixing issues we observe in the Sounds of the City corpus. The poor mixing we observe in Section 3.2.2 for the nested coefficients within the Speaker and Word effect are tackled with using an adaptation of hierarchical centering (Gelfand et al., 1995) and the poor precision mixing for the Word effect using a modification of parameter expansion (Liu et al., 1998).

Implementation of both these schema aims to reduce the time taken to run the MCMC

sampler significantly, which is of key importance to make the model as accessible and efficient as possible for the sociolinguistic community. Within this chapter we aim to show how both methods improve mixing within the model, reducing the number of MC samples required to sample from the target parameter distributions

Section 4.1 looks at how we improve the poor mixing of nested coefficients within the corpus by using an adaptation of hierarchical centering (Gelfand et al., 1995) within the Gibbs sampler. The notion of hierarchical centering is introduced through two motivating examples, a simple univariate response example with a population intercept and one random effect in Section 4.1.1, then a multiple response example in Section 4.1.2 which is similar in design to the Sounds of the City corpus. We then apply the method to the corpus, and comment on the improvements we observe.

Section 4.2 looks at another reparameterisation method we can implement to improve the poor mixing of the precision estimates observed in the Word random effect for the corpus, and it's respective coefficients. The idea of parameter expansion is explained in more detail in Section 4.2.1. We then introduce our modification of parameter expansion where we motivate the problem with a simple univariate example. The problem is then expanded to the multiple response case in Section 4.2.2 with a multiple response example then a direct application to the Sounds of the City corpus.

4.1 Improving nested coefficients mixing using hierarchical centering

When examining the model output for the Sounds of the City corpus in Section 3.2.2, we observed poor mixing within the fixed effects coefficients that are nested within the random effects for Speaker and Word. This nested design induces high correlations within the joint posterior distributions of groups of the parameters. There are several ways to attempt to deal with this correlation. One way is to consider block updating algorithms such as Structured MCMC (SMCMC) (Sargent et al., 2000), which looks to update the parameters in one block. Another approach that can be considered is hierarchical centering, which reparameterises the model in order to remove the correlations we observe.

In this section, we will introduce the notion of hierarchical centering with a simple univariate example which consists of an intercept and a random effect with several levels. We then extend beyond the intercept only case, and introduce using hierarchical centering methods for nested coefficients within multiple random effects for multiple response data, replicating the structure of the Sounds of the City corpus. Finally, we will apply these techniques to the Sounds of the City corpus and observe how they improve upon the poor mixing we observed in Section 3.2.2.

4.1.1 Hierarchical centering

Hierarchical centering (Gelfand et al., 1995) is a method used to improve mixing in MCMC samplers that focuses on the correlation between the fixed effects and the residuals. It can be used to improve mixing in cross classified models but is mainly used for models with nested random effects, like we observe in the Sounds of the City corpus.

To illustrate how hierarchical centering works, we will consider a simple univariate problem

Univariate example

In this example, we consider a model with a single random effect $\boldsymbol{\gamma}$ with four levels. For the first two levels, we assume a population mean β_0 The model is specified as follows:

$$y_{i,1} = \beta_0 + \gamma_1 + \epsilon_{i,1}$$
$$y_{i,2} = \beta_0 + \gamma_2 + \epsilon_{i,2}$$
$$y_{i,3} = \beta_0 + \gamma_3 + \epsilon_{i,3}$$
$$y_{i,4} = \beta_0 + \gamma_4 + \epsilon_{i,4}$$

For the coefficients, we assume conjugate normally distributed priors:

$$\beta_0 \sim \mathcal{N}(0, \sigma_\beta^2), \qquad \gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2)$$

The model error has conjugate inverse gamma prior:

$$\sigma_\epsilon^2 \sim \mathcal{IG}(a_\epsilon, b_\epsilon)$$

as does the random effect variance:

$$\sigma_\gamma^2 \sim \mathcal{IG}(a_\gamma, b_\gamma)$$

The hyperparameters for this model are specified as $\sigma_\beta^2 = 10$, $a_\gamma = b_\gamma = 100$ and $a_\epsilon = b_\epsilon = 100$.

If we look at the model specification, we observe that β_0 is involved in the mean likelihood for each observation, which is as shown the sum of β_0 and all the γ_i's. Hence we observe a strong correlation between our observed β_0 and the random effects. One way we could look to alleviate this correlation is to consider a reparameterisation of the model. We can replace the above model, and re-express in terms of a new variable $\boldsymbol{\delta}$, which for this problem can be constructed as:

$$\delta_1 = \beta_0 + \gamma_1$$
$$\delta_2 = \beta_0 + \gamma_2$$
$$\delta_3 = \beta_0 + \gamma_3$$
$$\delta_4 = \beta_0 + \gamma_4$$

We can now view our model of interest as $y_{i,j} = \delta_j + \epsilon_{i,j}$, where $\delta_j \sim \mathcal{N}(\beta_0, \sigma_\gamma^2)$. To fit this model, we now add an additional step to the Gibbs sampler where we sample β_0 by conditioning on $\boldsymbol{\delta}$. We then obtain the original $\boldsymbol{\gamma}$ values by simply calculating $\gamma_j = \delta_j - \beta_0$.

Figures 4.1 and 4.2 show the traceplots obtained for β_0 and γ_1 for the Gibbs sampler where hierarchical centering has not been implemented and then the sampler with the added centering step respectively. We observe clear differences between both sets of trace plots, with the mixing in Figure 4.2 showing vast improvement over the samples obtained in Figure 4.1. We observe that by centering on β_0, vast improvements are made in terms of mixing and thus obtaining more accurate samples from the MCMC. This is verified by Table 4.1, where we observe that the ESS values improve dramatically.

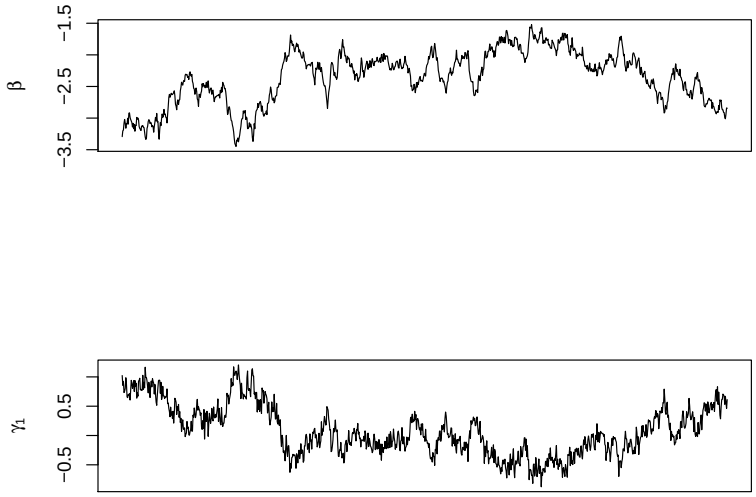

Figure 4.1: Trace plots for the intercept β_0 and random effect level γ_1 with no centering. Trace plots for β_0 and γ_1 for 10,000 iterations from the standard Gibbs sampler. We observe extremely poor mixing in both coefficients, with poor ESS values as shown in Table 4.1.

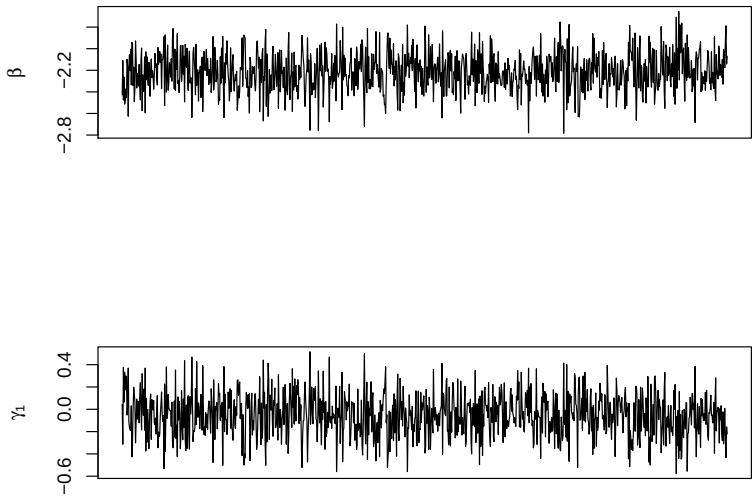

Figure 4.2: Trace plots for the intercept β_0 and random effect level γ_1 with centering. Trace plots for β_0 and γ_1 for 10,000 iterations from the Gibbs sampler with added centering step. The mixing for both coefficients has improved dramatically, with high ESS values as shown in Table 4.1.

Table 4.1: Effective sample size (ESS) values for coefficients from the univariate example for the standard Gibbs sampler and the sampler with added centering step. The ESS improves dramatically when we centre upon the population intercept β_0.

	Standard	Centered
β_0	44.8	9655.2
γ_1	46.5	9522.3
γ_2	49.1	9411.1
γ_3	47.2	9675.1
γ_4	46.1	8995.8

We can also observe the improvements in the sampler in terms of correlation between the parameters in Figures 4.3 and 4.4, where the correlation present between β_0 and γ_1 is shown for both samplers. In Figure 4.3 we observe the high correlation present between the coefficients, which means it will take a significantly large number of samples to obtain even a small effective sample. The density plots also highlight how poor the samples we obtain are, with both containing some multimodality. In Figure 4.4 we observe the improvements that centering can bring to improving mixing. Both terms are now no longer correlated and the sampler is able to efficiently explore the sample space and obtain samples that efficiently approximate the target distributions. Again, this can be seen in the density plots, where the densities now only contain one mode.

4.1.2 Extending Centering to Multiple Nested Coefficients

So far, we have looked at a simple univariate design problem to demonstrate how hierarchical centering works. We now want to extend using centering beyond the population intercept, but also including coefficients which are nested within random effects, much like the Sounds of the City corpus design. We will introduce the nesting notation through a simulated example with a multiple response and one random effect with a nested design within the fixed effects.

To illustrate how we define the nested coefficients, we detail two ways of expressing the following hierarchical model for multiple responses.

$$y_{ij}^l = \boldsymbol{\beta}^l \mathbf{x}_{ij} + \gamma_j^l + \epsilon_{ij}^l \tag{4.1}$$

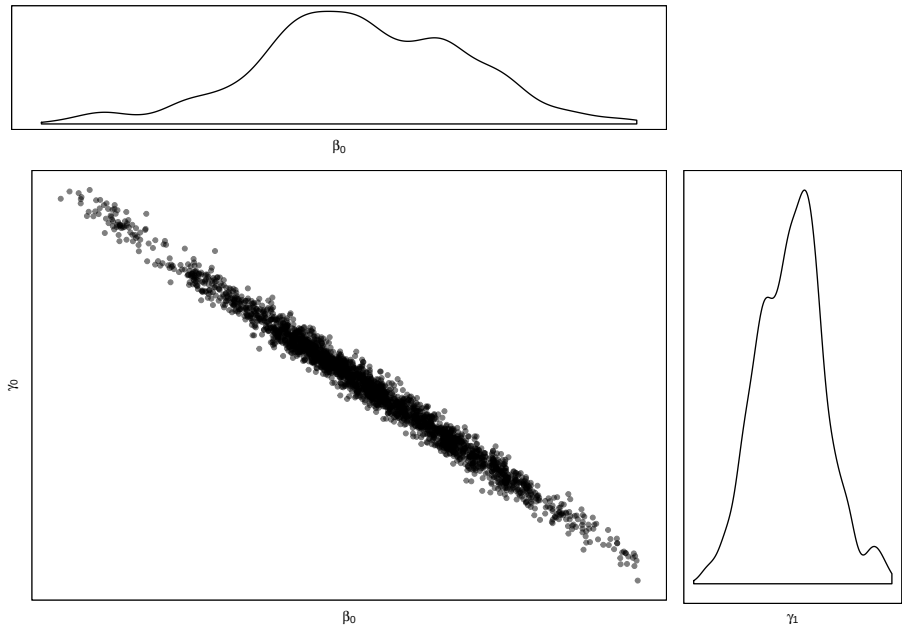

Figure 4.3: Correlation between β_0 and γ_1 coefficients from the standard Gibbs sampler. We observe very strong correlation which causes poor mixing in the sampler and we are unable to explore the full sample space due to the high autocorrelation. This can also be observed by the density plots, which struggle to identify the parameter mode.

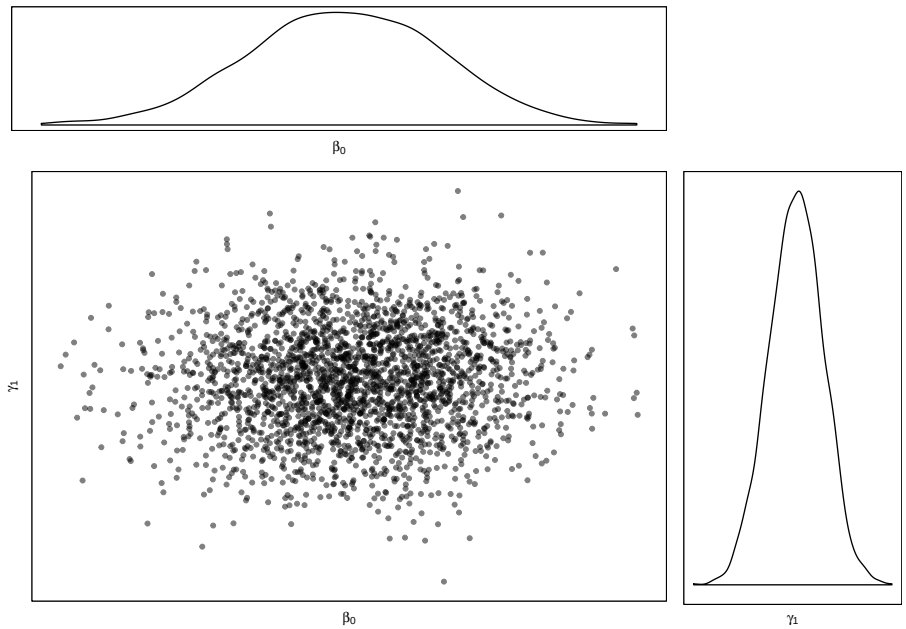

Figure 4.4: Correlation between β_0 and γ_1 coefficients from the centered sampler. We observe almost no correlation between the parameters and are able to fully explore the parameter space freely, leading to improved samples as shown by the density plots.

We can re-express the $\boldsymbol{\beta}^l$ coefficients and split them into two groups, those that are nested within the random effect $\boldsymbol{\gamma}^l$, denoted $\boldsymbol{\beta}^l_{\boldsymbol{\gamma}}$ and the remaining coefficients that are not nested, denoted $\boldsymbol{\beta}^l_{-\boldsymbol{\gamma}}$. By partitioning the fixed effects into these blocks, we can express the model in Equation 4.1 as follows:

$$y^l_{ij} = \boldsymbol{\beta}^l_{\boldsymbol{\gamma}}\mathbf{x}_{ij} + \boldsymbol{\beta}^l_{-\boldsymbol{\gamma}}\mathbf{x}_{ij} + \gamma^l_j + \epsilon_{ij} \tag{4.2}$$

We can sample the model parameters in the same way as the hierarchical model in Section 3.1.3, where the priors on the $\boldsymbol{\beta}^l$ coefficients are defined as follows:

$$\boldsymbol{\beta}^l_{\boldsymbol{\gamma}} \sim \mathcal{N}(\mathbf{0}, \tau_l^2\mathbf{I}) \qquad \boldsymbol{\beta}^l_{-\boldsymbol{\gamma}} \sim \mathcal{N}(\mathbf{0}, \tau_l^2\mathbf{I}). \tag{4.3}$$

We can use the model described in Equation 4.2 to explain how we implement the centering step. The model with a centering step is defined as follows:

$$y^l_{ij} = \delta^l_j + \boldsymbol{\beta}^l_{-\gamma_j}\mathbf{x}_{ij} + \epsilon_{ij}, \qquad \text{where } \delta^l_j = \boldsymbol{\beta}^l_{\gamma_j}\mathbf{x}_{ij} + \gamma^l_j \tag{4.4}$$

We sample the model parameters in the same way as before, but now add in an additional step to sample δ^l_j conditional on $\boldsymbol{\beta}^l_{\gamma_j}$ as follows:

$$\delta^l_j | \boldsymbol{\beta}^l_{\gamma_j} \sim \mathcal{N}(\mathbf{x}_{ij}\boldsymbol{\beta}^l_{\gamma_j}, \sigma^2_{\gamma_j l}) \tag{4.5}$$

Figure 4.5 shows how both Equations 4.2 and 4.4's respective input are constructed.

To illustrate how hierarchical centering can also improve the mixing of nested coefficients, we consider a toy example with three response variables, one random effect and three fixed effects coefficients, one of which is nested within the random effect. The model is constructed in the same fashion as the Bayesian hierarchical model in Section 3.1.3 but now including an additional step in the sampler for centering.

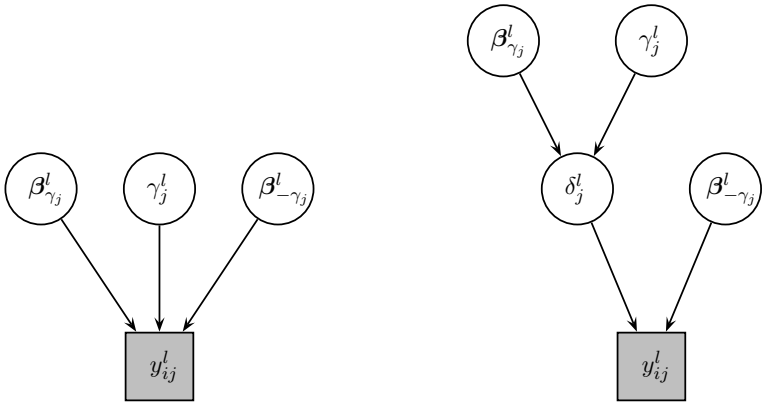

Figure 4.5: Representation of nested coefficients for different samplers Here, we illustrate the notation for the nested coefficients for the standard Gibbs sampler on the left and for the centered sampler on the right. Note the main difference arises from the formation of δ_j.

Posterior inference is now updated from Section 3.1.6 by adding the additional step for hierarchical centering. The full derivation can be found in Appendix A. The posterior distributions are defined as:

$$\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\right) \tag{4.6}$$

$$\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} \propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right) \tag{4.7}$$

$$\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \left[\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\boldsymbol{\delta}}_k, \left[\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\right) \tag{4.8}$$

We sample here for each nested block of coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ for each k, where $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k} = \text{blockdiag}\left(\mathbf{X}^1{}_{\tilde{\boldsymbol{\delta}}_k}, \ldots, \mathbf{X}^l{}_{\tilde{\boldsymbol{\delta}}_k}\right)$, and $\tilde{\boldsymbol{\delta}}_k = \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} + \tilde{\mathbf{U}}_k\tilde{\mathbf{b}}_k$.

$$\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}_{\tilde{\mathbf{b}}_g}^{-1} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}_{g_i}^{\top}\right]^{-1}\right) \tag{4.9}$$

$$\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} \propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}_{\boldsymbol{\epsilon}}^{-1} + \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^{\top}\right]^{-1}\right) \tag{4.10}$$

$$\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} \propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}_{\tilde{\boldsymbol{\eta}}_l}^l\|}{2}, b_l + \frac{\sum_{m=1}^{p}\left(\beta_m^l\right)^2}{2}\right) \tag{4.11}$$

Hyperparameters are fixed to the following values: $a_l = b_l = 0.001$ and $\nu_{\boldsymbol{\epsilon}} = 3, \mathbf{S}_{\boldsymbol{\epsilon}} = 0.001 \cdot \mathbb{I}$ and $\nu_{\mathbf{b}} = 34, \mathbf{S}_{\mathbf{b}} = 0.001 \cdot \mathbb{I}$.

The parameters are sampled using the following algorithm:

Algorithm 3: The Bayesian hierarchical model sampler with centering step
Given initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\mathbf{b}}^{(0)}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(0)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(0)}, \boldsymbol{\tau}^{(0)}\right)$. Then
For $t = 1, \ldots, T$

1. For $l = 1, \ldots, L$,

   (a) Sample $\boldsymbol{\beta}^{l,(t)}$ from 4.6.

   (b) Propose new model state $\boldsymbol{\eta}^{l,(t)}$. Sample $\boldsymbol{\beta}^l_{\boldsymbol{\eta}^{l,(t)}}$ from 4.6. Compute 3.15, where $\boldsymbol{\eta}^{l,(t-1)}$ is the current model state. If $u < \alpha$, where $u \sim \mathcal{U}(0,1)$, set $\boldsymbol{\beta}^{l,(t)} = \boldsymbol{\beta}^l_{\boldsymbol{\eta}^{l,(t)}}$, else $\boldsymbol{\beta}^{l,(t)}$ remains the same.

   Form $\tilde{\boldsymbol{\beta}}^{(t)}_{\tilde{\boldsymbol{\eta}}^{(t)}} = \left(\boldsymbol{\beta}^l_{\boldsymbol{\eta}^{1,(t)}}, \ldots, \boldsymbol{\beta}^l_{\boldsymbol{\eta}^{L,(t)}}\right)$

2. For $g = 1, \ldots, G$

   For $h = 1, \ldots, H$

(a) Sample $\tilde{\mathbf{b}}_{g,h}^{(t)}$ from 4.7.

Form $\tilde{\mathbf{b}}_{g}^{(t)} = \left(\tilde{\mathbf{b}}_{g,1}^{(t)}, \ldots, \tilde{\mathbf{b}}_{g,H}^{(t)}\right)^{\top}$

Form $\tilde{\mathbf{b}} = \left(\tilde{\mathbf{b}}_{1}^{(t)}, \ldots, \tilde{\mathbf{b}}_{G}^{(t)}\right)^{\top}$

3. For $k = 1, \ldots, K$

(a) Sample $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ from 4.8

Form $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}} = \left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_1}, \ldots, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_K}\right)$

4. For $g = 1, \ldots, G$,

Sample $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{(t)}$ from 4.9.

Form $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)}$ by $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_1}^{(t)}, \ldots, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_G}^{(t)}\right)$.

5. Sample $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(t)}$ from 4.10.

6. For $l = 1, \ldots, L$,

Sample $\tau_l^{(t)}$ from 4.11.

Form $\boldsymbol{\tau}^{(t)} = \left(\tau_1^{(t)}, \ldots, \tau_L^{(t)}\right)$

We run the sampler for 10,000 iterations for both the standard Gibbs case and the case with the added centering for comparison. Figure 4.6 shows the improvement we observe in terms of mixing for the population intercept and the nested coefficient for the first response level. This is further verified in Table 4.2 where we see the ESS values improve greatly between the samplers, just as in the univariate case.

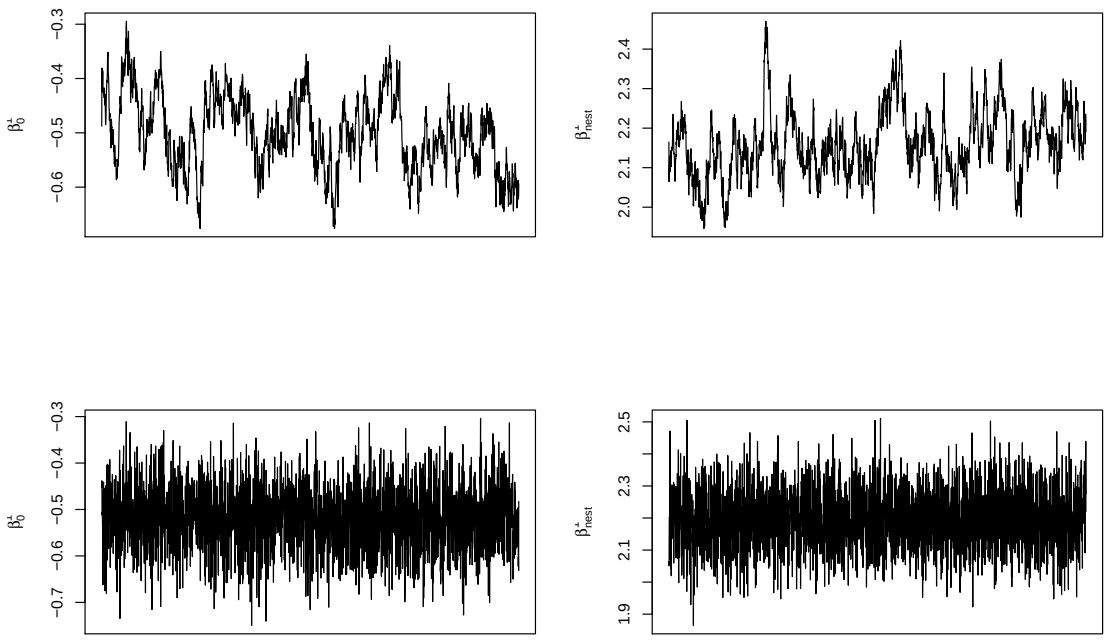

Figure 4.6: Traceplots for the population intercept and nested coefficient for the first response level for the standard Gibbs sampler and the centered sampler for 2,500 iterations. We see a clear improvement in mixing between both samplers for the nested terms and the population intercept.

Table 4.2: ESS values for nested coefficients from the multiple response example for the standard Gibbs sampler and the centered sampler for 2,500 iterations. The ESS improves greatly when we centre on the population intercept and the nested coefficient.

	Standard	Centered
β_0^1	37.3	2472.2
β_{nest}^1	44.1	2481.6
β_0^2	28.9	2500
β_{nest}^2	36.8	2500
β_0^3	36.4	2427.8
β_{nest}^3	50.7	2383.2

4.1.3 Sounds of the City Corpus

Here, we look to verify if the centering improvements we have discussed in this section can help lead to improvement in mixing for the nested coefficients within the Sounds of the City corpus, namely the social factors Gender, Age and Decade of recording nested within the Speaker and linguistic factors of Following and Preceding place of articulation of consonant within the Word choice.

We run the Bayesian hierarchical model to the same specification as in Section 3.2.2, though making two changes. Now, we have included a step for centering of nested terms within the sampler and have disabled the model selection, fitting the model with all fixed effects. The motivation behind this is so we obtain more accurate measurements of ESS, due to the model selection effectively zeroing out coefficients when they are not included within the model. For comparison, we have run the standard Gibbs sampler for the dataset with model selection disabled.

Again, we look at the LOT vowel for the same prior specification in Section 3.2.2. Time series plots for the coefficients fitted to the raw mean formant measurements on F1 are shown in Figure 4.7. We observe a great improvement on the mixing for the nested coefficients within speaker when compared to Figure 3.4. This is further shown in Table 4.3, where we see the ESS for all the coefficients for F1 has improved greatly, with a large improvement for the coefficients nested within Speaker. We do not observe as many problems with nesting for the Word effect, though slight improvement is still shown in terms of ESS when we include the nesting step.

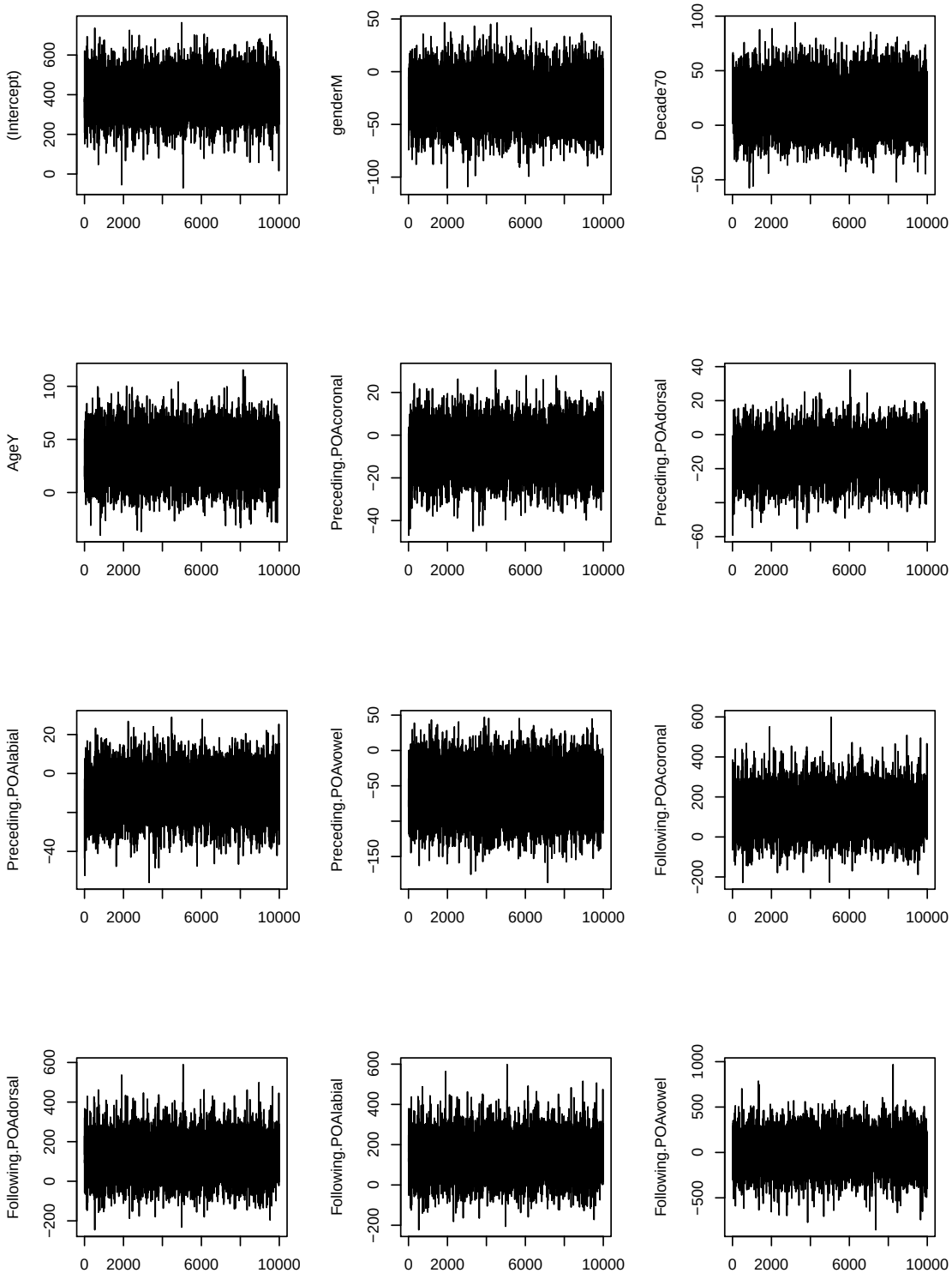

Figure 4.7: Traceplots for the fixed effects for F1 fitted to the LOT vowel for 10,000 iterations with hierarchical centering implemented for nested coefficients. We see a clear improvement in terms of mixing for all the variables comparing to Figure 3.4.

Table 4.3: ESS values obtained for the LOT vowel coefficients for the standard Gibbs sampler and the centered sampler for 10,000 iterations on F1. We observe large improvements in terms of ESS for all parameters, mainly for the terms nested within Speaker.

	Standard	Centered
β^1_0	3224	10000
$\beta^1_{genderM}$	113	6237
$\beta^1_{Decade70}$	147	7532
β^1_{AgeY}	135	6532
$\beta^1_{PrCoronal}$	5142	6947
$\beta^1_{PrDorsal}$	2270	5848
$\beta^1_{PrLabial}$	5037	8009
$\beta^1_{PrVowel}$	8121	9127
$\beta^1_{FoCoronal}$	7232	8217
$\beta^1_{FoDorsal}$	8562	9118
$\beta^1_{FoLabial}$	9946	10000
$\beta^1_{FoVowel}$	10000	10000

4.2 Improving Random Effects Precision Mixing Using Parameter Expansion

One other area we observed poor mixing in the models fitted to the Sounds of the City corpus was within the Word random effect precision estimates. This also leads to an effect on the word effect traceplots for each level, with the word effect appearing to be closely linked to the precision estimate, with the coefficient trace covering more of the posterior space when the precision is not close to zero, and concentrated near zero when the precision trace is stuck around values near zero. As seen in Figure 3.6, the algorithm can get stuck in zero regions for many iterations, leading to poor mixing for the precision estimates and the random effects coefficients.

In this Section, we will introduce the notion of parameter expansion, and propose a simplified case of parameter expansion, applied to some simulated examples. We will then apply this expansion step to a multiple response simulated example, similar in construct to the corpus, then an application to the Sounds of the City corpus to observe how they can improve mixing within the precision step and the word random effect coefficients.

4.2.1 Parameter Expansion Based Mixing Improvements

Parameter expansion was originally proposed by Liu et al. (1998) to speed up the EM algorithm. This was then extended to the Gibbs sampler by Liu and Wu (1999) and then considered for hierarchical models by Gelman et al. (2008). The method is referred to as parameter expansion as our model of interest is expanded by augmenting it with addition parameters to make an expanded model. These additional parameters included within the model framework aren't identifiable within the model, there exists an 'embedded' model that is identifiable and is the original model of interest. This means we can obtain the original parameters of interest from the augmented parameter set.

We propose an idea based on parameter expansion, but simpler in execution. Instead of performing the full expansion step, we effectively update the precision estimate and relevant coefficient parameters by multiplying them by a scalar constant, denoted as some arbitrary value α, and determine whether this modified parameter set yields an improvement on the model by a Metropolis step. We explain this idea in further detail with the use of simulated examples throughout this section.

To illustrate how our adaptation of parameter expansion works, we consider a simple example with a population intercept β_0 and a single random effect $\boldsymbol{\gamma}$ with 80 levels for 100 observations. We propose this structure to emphasise the lack of available information we observe on each level. The model is specified as:

$$y_{ij} = \beta_0 + \gamma_j + \epsilon_{ij} \tag{4.12}$$

For the coefficients, we assume conjugate normally distributed priors:

$$\beta_0 \sim \mathcal{N}(0, \sigma_\beta^2), \qquad \gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2)$$

The model error and random effect variance have conjugate inverse gamma priors:

$$\sigma_\epsilon^2 \sim \mathcal{IG}(a_\epsilon, b_\epsilon) \qquad \sigma_\gamma^2 \sim \mathcal{IG}(a_\gamma, b_\gamma)$$

Hyperparameters are set as $\sigma_\beta^2 = 100$, $a_\gamma = 0.001, b_\gamma = 0.001$ and $a_\epsilon = b_\epsilon = 0.001$.

As highlighted above, we have set the design in such a way that the the effects of each level of the random effect has almost minimal significant effect and the σ^2_γ used to generate the data is set lower than the model error. This should lead to poor mixing within the variance parameter.

In order to deal with this poor mixing, we propose the use of a reparameterisation step, by introducing the additional parameter α, which now changes the model to be defined as:

$$y_{ij} = \beta_0 + \alpha\gamma_j + \epsilon_{ij} \tag{4.13}$$

For α, there are several proposed values we can consider. We could define α by an arbitrary scalar value or to be drawn from a known distribution. We propose sampling α from a Gamma prior as such:

$$\alpha \sim \mathcal{IG}(a_\alpha, b_\alpha) \tag{4.14}$$

The update works using a MH step. Taking the parameters that inhibit poor mixing, which in this example is $\boldsymbol{\gamma}$ and σ^2_γ. Once we have sampled these parameters, we then draw α from the chosen target distribution $q(\alpha)$, which for this example is the Gamma distribution for set hyperparameters a_α and b_α. We then form $\boldsymbol{\gamma}^* = \alpha\boldsymbol{\gamma}$ and $\sigma^2_{\gamma^*} = \alpha^2\sigma^2_\gamma$. From this, we perform the Metropolis step where the parameters $\boldsymbol{\gamma}$ and σ^2_γ will be updated to $\boldsymbol{\gamma}^*$ and $\sigma^2_{\gamma^*}$ respectively if we accept this step. If the step is accepted, we update the parameters such that $\boldsymbol{\gamma}^* = \alpha\boldsymbol{\gamma}$ and $\sigma^2_\gamma = \alpha^2\sigma^2_{\gamma^*}$.

To define the step generally, suppose we have a parameter set $\mathbf{X}$, which has three parameters X_1, X_2 and X_3 and our constant value α and wish to move to the modified parameter set $\mathbf{X}^*$. The variable transformation would be of the form:

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \alpha \end{pmatrix} \longrightarrow \begin{pmatrix} \alpha\mathbf{X}_1 \\ \alpha\mathbf{X}_2 \\ \alpha\mathbf{X}_3 \\ \alpha \end{pmatrix} \tag{4.15}$$

When transforming variables, we must compute the Jacobian of the transformed set of

variables. As we observe in this illustrative example, this would simply be the coefficient α to the power of how many parameters we modify, so in this example the Jacobian is α^3. For any univariate example, the Jacobian will correspond to α to the power of the number of levels of the random effect $\boldsymbol{\gamma} + 2$, with the additional 2 levels coming from the α^2 attached to the variance parameter.

We include this Jacobian result in the acceptance probability, which is a ratio of the densities of the model likelihood, random effect and variance prior distributions for both the standard parameters and the modified parameter sets. We accept the transformed parameters according to the following probability:

$$\phi = \frac{q(\alpha)\mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b}_g, \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}\right)\mathcal{N}\left(\mathbf{b}_g \mid \mathbf{0}, \sigma^2_{\mathbf{b}_g}\mathbf{I}\right)\mathcal{IG}\left(\sigma^2_{\mathbf{b}_g} | a_{\mathbf{b}_g}, b_{\mathbf{b}_g}\right)}{q(1/\alpha)\mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b}^*_g, \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}\right)\mathcal{N}\left(\mathbf{b}^*_g \mid \mathbf{0}, \sigma^2_{\mathbf{b}^*_g}\mathbf{I}\right)\mathcal{IG}\left(\sigma^2_{\mathbf{b}^*_g} | a_{\mathbf{b}^*_g}, b_{\mathbf{b}^*_g}\right)} \left|\mathbf{J}\right| \tag{4.16}$$

where $\left|\mathbf{J}\right| = \alpha^{||\boldsymbol{\gamma}||+2}$ and $||\boldsymbol{\gamma}||$ is the length of $\boldsymbol{\gamma}$.

The motivation as to why this additional step improves mixing as although the sampler can escape values close to 0 for the precision, it quite easily gets stuck again. Multiplying the parameters by α helps to get around this problem, as even a small increase of the variance by α will move both the precision and the parameter estimates together.

Setting $a_\alpha = 20$ and $b_\alpha = 10$, we run the model for the simulated example above. Figure 4.8 shows the traceplots for γ_1 and $\sigma^2_{\boldsymbol{\gamma}}$ for 5,000 iterations of the standard Gibbs sampler and the sampler with the added Metropolis step. We can clearly see the vast improvement on the mixing of $\sigma^2_{\boldsymbol{\gamma}}$, which in the standard Gibbs sampler, is often trapped at 0 for long periods. With the parameter expansion step, we see a greater improvement in mixing, with the sampler exploring the parameter space more freely. Note the mixing for the variance is not perfect, but in comparison to the standard sampler estimate, it has improved dramatically. This in turn also improves the mixing of the coefficient γ_1, which for the standard sampler was often trapped around 0 and unable to explore the full space. With the parameter modification step, the parameter is now mixing extremely well. This is also verified by Table 4.4 where we observe improvement in ESS for the coefficients. The variance parameter also improves, albeit at not quite the same large rate.

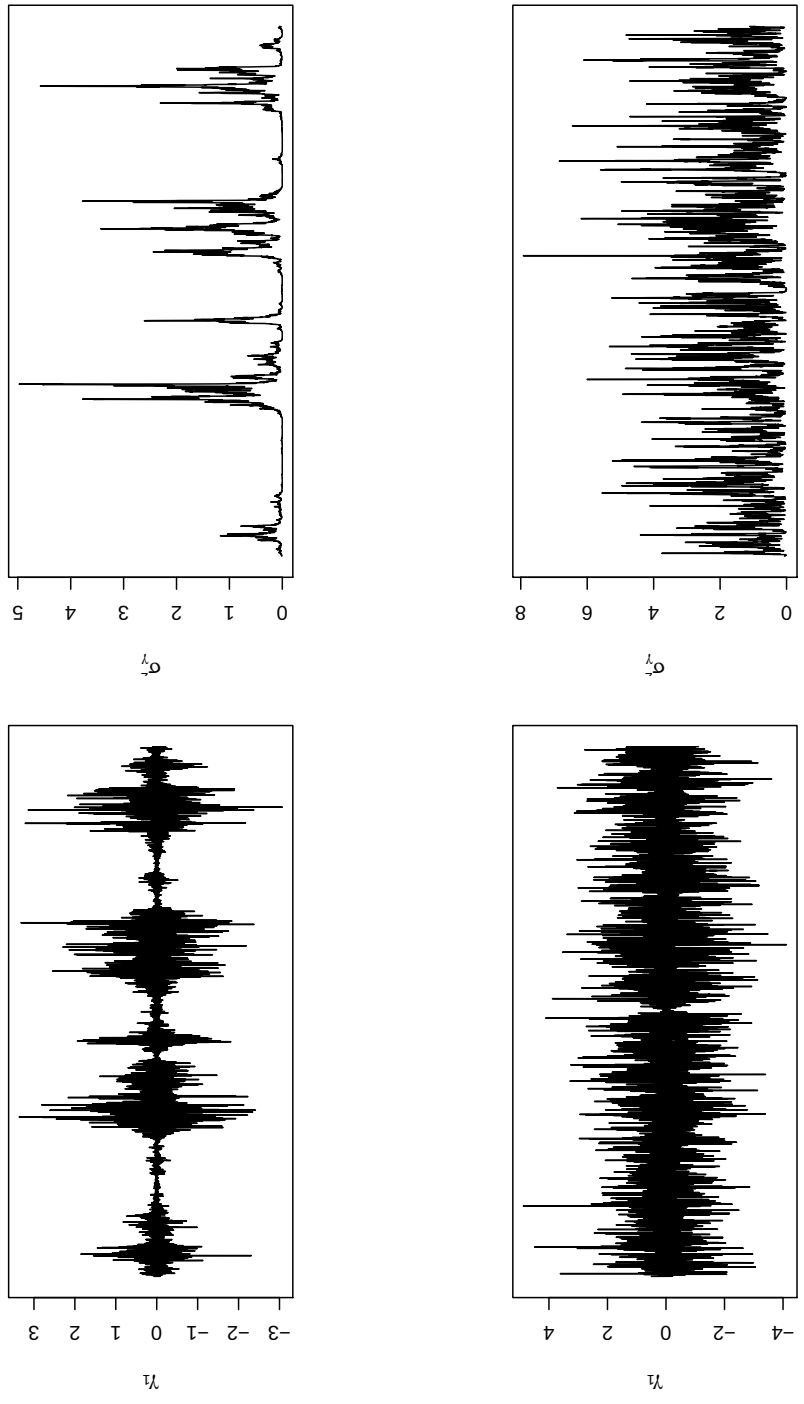

Figure 4.8: Traceplots for the univariate example for the standard Gibbs sampler and one with a parameter expansion step added for 5,000 iterations. We see great improvement in the mixing of γ_1 and σ^2_γ when the parameter expansion step as shown on the second row of plots is included in the sampler.

Table 4.4: ESS values obtained for a selection of γ_j coefficients and σ^2_γ for the standard Gibbs sampler and one with the added parameter expansion step. We see a great improvement in ESS for the coefficients and good improvement for the variance.

	Standard	Expanded
γ_1	185	2174
γ_{11}	283	1741
γ_{21}	105	1899
γ_{31}	703	2708
σ^2_γ	38	507

4.2.2 Multiple Response Expansion - Simulated Example

We now extend beyond the univariate case for the expansion step to the multiple response case. We construct a toy problem in a similar fashion to the multiple response example in Section 4.1.2, though no longer including coefficients that are nested within the random effect. Instead, we construct the random effect in a similar fashion to the univariate example previously, with a population intercept for each response level and a single random effect $\boldsymbol{\gamma}^l$ which has 80 levels for 100 observations, creating a similar structure to the univariate case.

The model is specified as:

$$y^l_{ij} = \beta^l_0 + \gamma^l_j + \epsilon^l_{ij} \tag{4.17}$$

We model this problem using the Bayesian hierarchical model, though this time including a Metropolis step for the parameters that inhibit poor mixing. Now, we extend beyond the univariate example to the multiple response case for the parameter modification step. We do not implement the work discussed in Gelman et al. (2008) which involves the addition of several more sampling steps. Instead, we simply perform a Metropolis step in a similar fashion to the one implemented in Section 4.2.1. We define our parameters that are modified using α as $\tilde{\mathbf{b}}^*_{\mathbf{g}} = \alpha\tilde{\mathbf{b}}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^*_{\tilde{\mathbf{b}}_{\mathbf{g}}} = \alpha\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}$.

The Metropolis step for a given random effect $\tilde{\mathbf{b}}_{\mathbf{g}}$ and its corresponding precision matrix $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}$ for parameter modification by α by the definition of the Bayesian hierarchical

model defined in Section 3.1.3 is:

$$\phi = \frac{q(\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}_g, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}_g|\mathbf{0}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{-1}\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{-1}|\nu_{\tilde{\mathbf{b}}_g}, \mathbf{S}_{\tilde{\mathbf{b}}_g}\right)}{q(1/\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}_g^*, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}_g^*|\mathbf{0}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g^*}^{-1}\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g^*}^{-1}|\nu_{\tilde{\mathbf{b}}_g^*}, \mathbf{S}_{\tilde{\mathbf{b}}_g^*}\right)} \left|\mathbf{J}\right| \tag{4.18}$$

where $\left|\mathbf{J}\right| = \alpha^{\|\tilde{\mathbf{b}}_\mathbf{g}\|+L}$ and $\|\tilde{\mathbf{b}}_\mathbf{g}\|$ is the length of $\tilde{\mathbf{b}}_\mathbf{g}$. The additional expression $L(L+1)$ comes from the number of terms present in the covariance matrix which is of dimension $L \times L$.

The parameter update is accepted if $u < \phi$, where $u \sim \mathcal{U}(0,1)$. This step is performed at the end of the sampler, after $\tilde{\mathbf{b}}_\mathbf{g}$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_\mathbf{g}}$ have been sampled and is performed on each relevant group $\mathbf{g}$ which has poor mixing.

Hyperparameters are fixed to the following values: $a_l = b_l = 0.001$ and $\nu_{\boldsymbol{\epsilon}} = 3, \mathbf{S}_{\boldsymbol{\epsilon}} = 0.001 \cdot \mathbb{I}$ and $\nu_\mathbf{b} = 3, \mathbf{S}_\mathbf{b} = 0.001 \cdot \mathbb{I}$. We also set $a_\alpha = b_\alpha = 10$. We run the model for the simulated example detailed above for 5,000 iterations for the standard Gibbs sampler and also for the sampler with the added parameter expansion step.

Figure 4.9 shows the precision estimates for the random effect for both the standard Gibbs sampler and the one with added reparameterisation step. We observe for the standard sampler that the mixing is relatively poor, getting occasionally trapped near zero values for short periods of time within the sampler. When using the reparameterisation step, we observe a slight improvement in mixing, with the sampler escaping the areas near zero more often than the sampler not implementing the reparameterisation step.

Like in the univariate case, this also leads to improvement in mixing of the corresponding random effects coefficients, as we can observe in Figure 4.10. We observe improved mixing for the first three levels of $\boldsymbol{\gamma}^1$ when implementing the reparameterisation step, in terms of the sampler variability not narrowing near zero values as often and exploring the parameter space more freely. This improvement in mixing is verified in Table 4.5 where we see a very marginal improvement in terms of mixing from both samplers. As the standard sampler was only stuck in values close to zero for marginal periods of time, it is very seldom the sampler accepts a new proposed set of parameters, as demonstrated by the low acceptance rate obtained of 0.12.

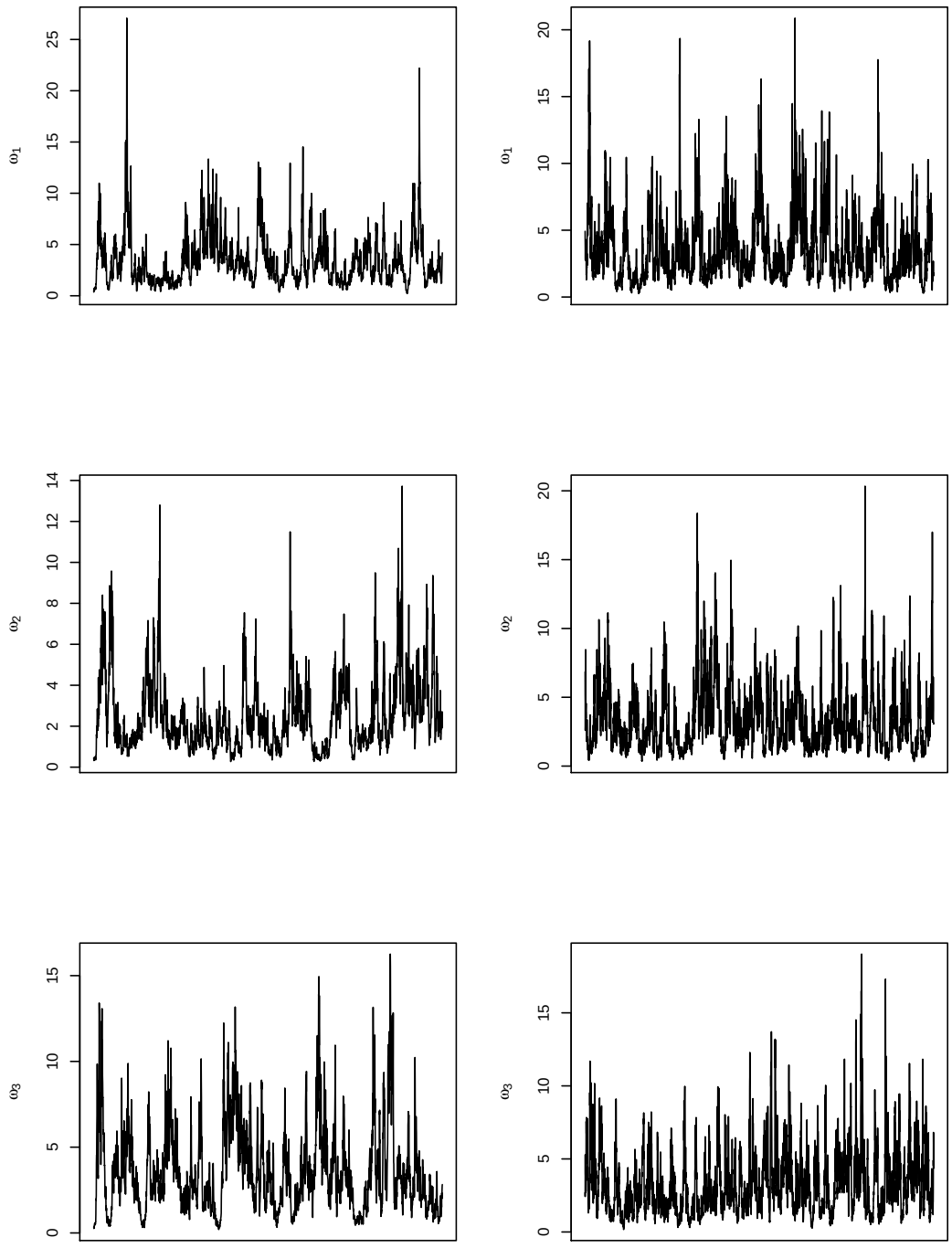

Figure 4.9: Traceplots for the precision estimates for the three response levels for the $\boldsymbol{\gamma}^l$ random effect from the hierarchical model run for 5,000 iterations. The left hand side plots are for the standard model and the right hand plots are with the added parameter expansion step. We observe a small improvement in mixing.

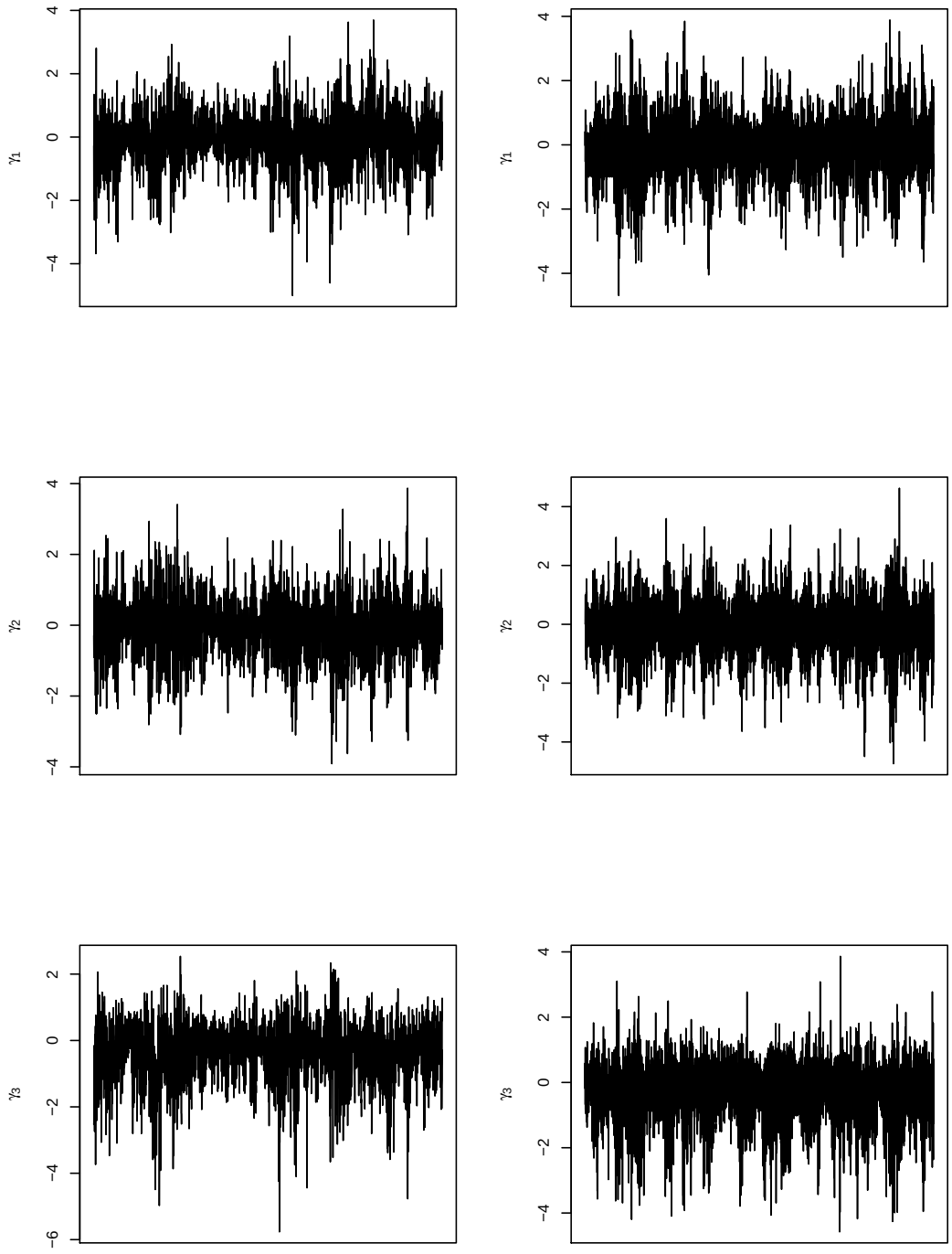

Figure 4.10: Traceplots for the coefficient estimates for γ_1, γ_2 and γ_3 for the standard sampler on the left and the sampler with parameter expansion step on the right. We see a slight improvement in mixing in terms of better variation around zero estimates due to the improvement in precision mixing.

Table 4.5: ESS values obtained for a selection of γ_j^l coefficients and $\sigma^2_{\gamma^l}$ values for the standard Gibbs sampler and the parameter expanded added sampler. We see a great improvement in ESS for the coefficients and good improvement for the variance.

	Standard	Expanded
γ_1^1	1530	2480
γ_{11}^1	478	1006
γ_{21}^1	1245	3255
$\sigma^2_{\gamma^1}$	180	488
$\sigma^2_{\gamma^2}$	220	695
$\sigma^2_{\gamma^3}$	60	401

4.2.3 Sounds of the City Corpus Application

We now look to apply the parameter expansion step to the Sounds of the City corpus, where in Figure 3.6, we observed poor mixing in the precision estimates for the Word random effect, which in turn led to poor mixing in the Word random effects coefficients.

We run the Bayesian hierarchical model to the same specification as in Section 3.2.2, but this time including the hierarchical centering step in Section 4.1.3 and also a reparameterisation step at the end of the sampler for the Word precision estimate and random effects coefficients. We set $a_\alpha = 20 = b_\alpha = 10$. Again, we disable model selection within this test of the model to obtain more accurate measurements for ESS when comparing to the standard sampler results obtained in Section 3.2.2. We also only perform parameter expansion on the first formant, as the other two formants did not exhibit any poor mixing.

The parameters are sampled using the following algorithm:

Algorithm 4: The Bayesian hierarchical model sampler with mixing improvements Given initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\mathbf{b}}^{(0)}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(0)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(0)}, \boldsymbol{\tau}^{(0)}\right)$. Then

For $t = 1, \ldots, T$

1. For $l = 1, \ldots, L$,

   (a) Sample $\boldsymbol{\beta}^{l,(t)}$ from 4.6.

(b) Propose new model state $\boldsymbol{\eta}^{l,(t)}$. Sample $\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$ from 4.6. Compute 3.15, where $\boldsymbol{\eta}^{l,0}$ is the current model state. If $u < \alpha$, where $u \sim \mathcal{U}(0,1)$, set $\boldsymbol{\beta}^{l,(t)} = \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$, else $\boldsymbol{\beta}^{l,(t)}$ remains the same.

Form $\tilde{\boldsymbol{\beta}}^{(t)}_{\tilde{\boldsymbol{\eta}}^{(t)}} = \left(\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{1,(t)}}, \ldots, \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{L,(t)}}\right)$

2. For $g = 1, \ldots, G$

   For $h = 1, \ldots, H$

   (a) Sample $\tilde{\mathbf{b}}^{(t)}_{g,h}$ from 4.7.

   Form $\tilde{\mathbf{b}}^{(t)}_{g} = \left(\tilde{\mathbf{b}}^{(t)}_{g,1}, \ldots, \tilde{\mathbf{b}}^{(t)}_{g,H}\right)^{\top}$

   Form $\tilde{\mathbf{b}} = \left(\tilde{\mathbf{b}}^{(t)}_{1}, \ldots, \tilde{\mathbf{b}}^{(t)}_{G}\right)^{\top}$

3. For $k = 1, \ldots, K$

   (a) Sample $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ from 4.8

   Form $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}} = \left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_1}, \ldots, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_K}\right)$

4. For $g = 1, \ldots, G$,

   Sample $\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_g}$ from 4.9.

   Form $\boldsymbol{\Sigma}^{(t)}_{\tilde{\mathbf{b}}}$ by $\boldsymbol{\Sigma}^{(t)}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_1}, \ldots, \boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_G}\right)$.

5. Sample $\boldsymbol{\Omega}^{(t)}_{\boldsymbol{\epsilon}}$ from 4.10.

6. For $g = 1, \ldots, G$

   (a) Form $\tilde{\mathbf{b}}^{*(t)}_{\mathbf{g}} = \alpha\tilde{\mathbf{b}}^{(t)}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^{*(t)}_{\tilde{\mathbf{b}}_\mathbf{g}} = \alpha\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_\mathbf{g}}$. Compute 4.18. If $u < \phi$, where $u \sim \mathcal{U}(0,1)$, set $\tilde{\mathbf{b}}^{(t)}_{\mathbf{g}} = \alpha\tilde{\mathbf{b}}^{(t)}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_\mathbf{g}} = \alpha\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_\mathbf{g}}$, else $\tilde{\mathbf{b}}^{(t)}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^{(t)}_{\tilde{\mathbf{b}}_\mathbf{g}}$ remain the same.

7. For $l = 1, \ldots, L$,

   Sample $\tau^{(t)}_{l}$ from 4.11.

   Form $\boldsymbol{\tau}^{(t)} = \left(\tau^{(t)}_{1}, \ldots, \tau^{(t)}_{L}\right)$

Table 4.6: Effective sample size (ESS) values for precision estimates from the Word random effect for F1, F2 and F3 for the *LOT* vowel from the Bayesian hierarchical model ran for 10,000 iterations with added parameter expansion step.The ESS observed has significantly improved in comparison to the results shown in Table 3.4.

	F1	F2	F3
$\omega_{\mathbf{b}_{word}}$	592.5	752.8	449.2

Once again, we consider the *LOT* vowel for the same prior specification in Section 3.2.2. Time series plots for the precision estimates for the Word random effect are shown in Figure 4.11. When compared to the plots in Figure 3.6, we see a very slight improvement in terms of mixing, with several values that were trapped near the lower end of the parameter scale mixing more freely. Again, the traceplots are not mixing perfectly, but this slight improvement is still significant and worth implementing due to the minuscule computational cost. This is seen also in Table 4.6, where the effective sample size for each formant has improved over the standard Gibbs sampler.

We also consider the Word effect coefficients, where we observed inconsistent mixing in terms of the trace variablilty as shown in Figure 3.6. Figure 4.12 shows the coefficients after the parameter expansion step. We observe a constant variability throughout the chain for a selection of the coefficients. Table 4.7 shows this improvement in terms of ESS, with all parameters improving with a high number of independent samples recorded.

Table 4.7: ESS values obtained for a sample of Word random effects coefficents for the standard Gibbs sampler and one with the added parameter expansion step for 10,000 iterations for the *LOT* vowel. We see a small improvement in ESS for the coefficients and good improvement for the variance.

	Expanded	Standard
Cops	7980	5240
Bob	6899	4903
what	9649	8171
cloth.	6104	3544
because	8121	7143
Because	8607	5366

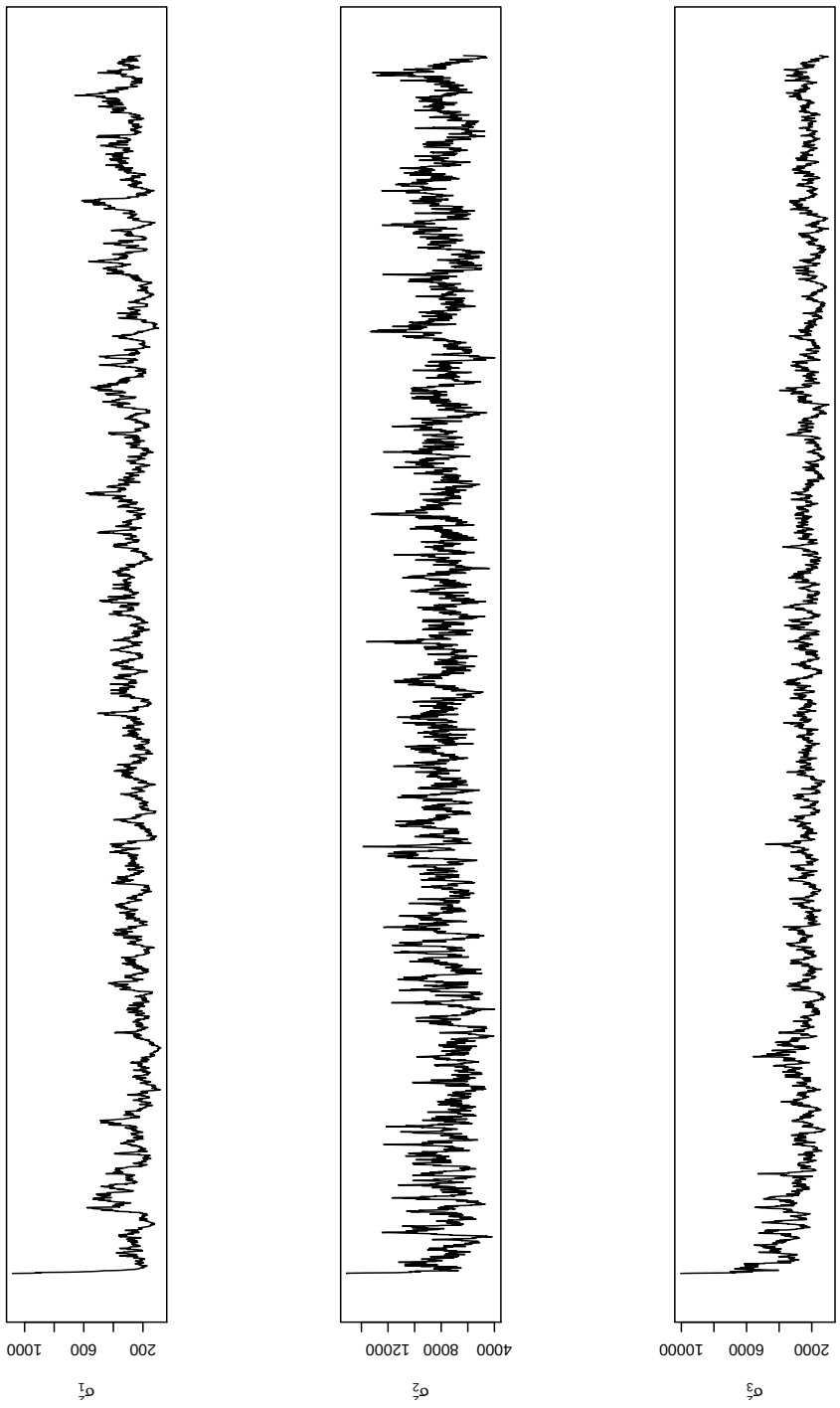

Figure 4.11: Traceplots for the variance estimates for the Word random effect for F1, F2 and F3 for the *LOT* vowel run for 10,000 iterations with the parameter expansion step added. We see a small improvement in mixing when compared to Figure 3.6 for the standard sampler, mainly with the estimates for F1.

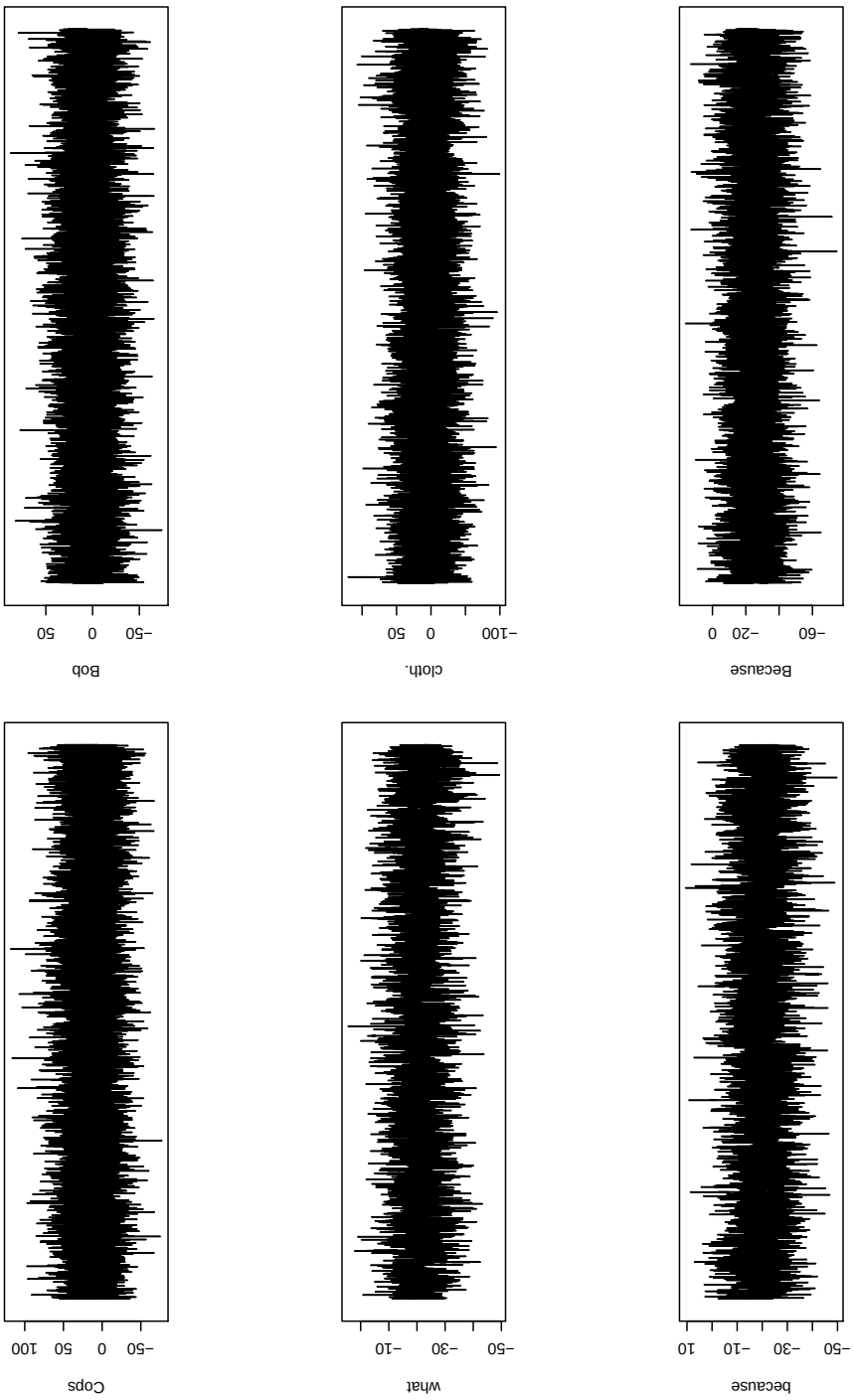

Figure 4.12: Traceplots for a selection of the Word effect coefficients for the *LOT* vowel for F1, for the parameter expanded model ran for 10,000 iterations. We observe an improvement in mixing compared to Figure 3.7, with the trace variance remaining constant due to the improved mixing in the Word effect precision.

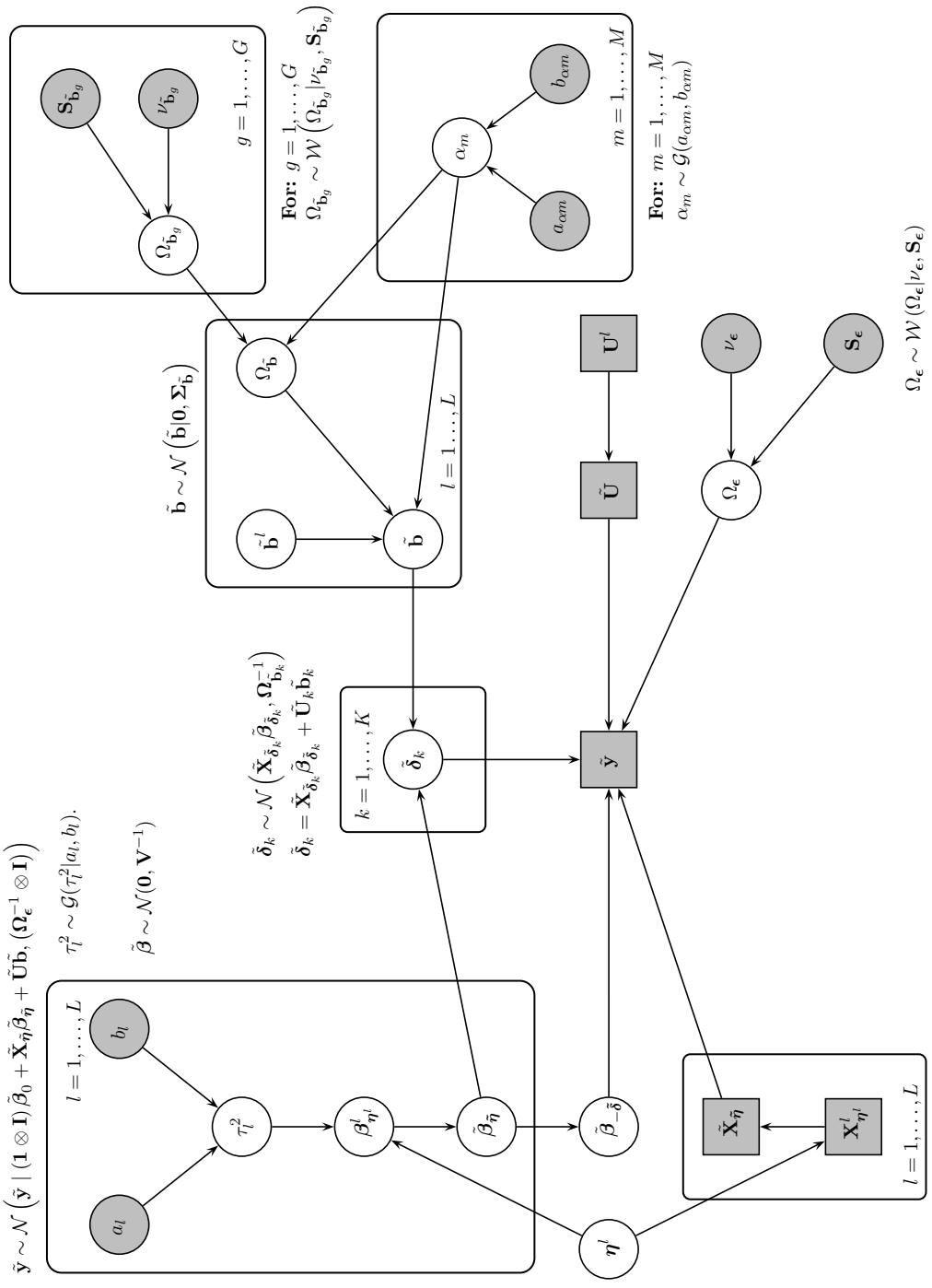

Figure 4.13: Representation of the hierarchical model with added mixing steps as a PGM. The PGM is constructed in a similar style as Figure 3.1, though this time we have included nodes for the hierarchical centering with the $\tilde{\boldsymbol{\delta}}_k$ parameter, and the additional parameter expansion step using α_m

4.3 Discussion

In this chapter, we have introduced two reparameterisation methods to the Bayesian hierarchical model discussed in Chapter 3 to improve mixing issues that were found when applying the model to the Sounds of the City corpus in Section 3.2.2. We have implemented a hierarchical centering (Gelfand et al., 1995) step into the model, which has improved the mixing of the nested fixed effects and respective random effects coefficients greatly. We have also applied an adaptation of parameter expansion (Liu et al., 1998) which has lead to a small improvement in mixing in the precision estimates for the Word random effect and its corresponding coefficients.

Figure 4.13 provides a graphical representation of the hierarchical model, building on Figure 3.1 but now including the updating steps for centering and parameter expansion on the relevant parameters.

With these improvement in mixing, we are able to obtain a greater proportion of independent samples and approximate the target distributions for our parameters of interest within the model. This directly leads to a greater reduction in run time for the model, again improving the usability of the model for the sociolinguistic community.

A drawback to the hierarchical model does still remain. Due to its complex nature, interpretation of the output can be somewhat overwhelming and confusing to users not familiar with its structure. In the next two chapters, we propose and introduce a novel inference tool which uses the output from the hierarchical model and structures the output using graphical models to help aid interpretation and how to fit undirected graphs for multiple precision estimates.

Chapter 5

Using Bayesian Gaussian Graphical Models to Model Response Level Dependency

In this chapter, we look to model the relationship present between the multiple response variables within the Bayesian hierarchical model we have constructed in Chapters 3 and 4 by using graphical models to provide a visualisation of this relationship.

We can infer the relationship between the response variables by using a Bayesian Gaussian graphical model, which uses the model precision to infer the conditional dependencies between responses. Using the precision estimates from the Bayesian hierarchical model as input, we can obtain the best graphical model structure using a modified model selection algorithm.

We introduce the concept of an undirected graph in Section 5.1, extending into the structure of a Gaussian graphical model which we use to infer the conditional dependency present between the response variables. We then extend this to the Bayesian case in Section 5.2, discussing the G-Wishart prior in more detail. Section 5.3 discusses how we can infer the best graphical model for a given precision matrix, extending beyond standard samplers by allowing the input of multiple precision matrices, much like we observe in the Bayesian hierarchical model.

5.1 Graphical Models

In this section we introduce how graphical models are structured, mainly the undirected graphical model. Graphical models provide a visualisation of complex probabilistic models by using graph theory as a framework to represent such structures. They provide a simple to interpret way to visualise the structure of a probabilistic model and show properties such as conditional independence clearly.

A graph $\mathcal{G}$ is composed of two elements: vertices $\mathcal{V}$ and edges $\mathcal{E}$. Vertices represent a random variable, while edges correspond to a conditional dependence between vertices. The graph captures the way in which the joint distribution over all the random variables can be decomposed into a product of factors depending only on a subset of the variables. There are two main categories of graphical model, namely directed graphical models, where the edges of the graph have a particular direction indicated by an arrow, and undirected graphical models, where the edges do not have arrows and thus have no directional influence.

5.1.1 Undirected Graphical Models

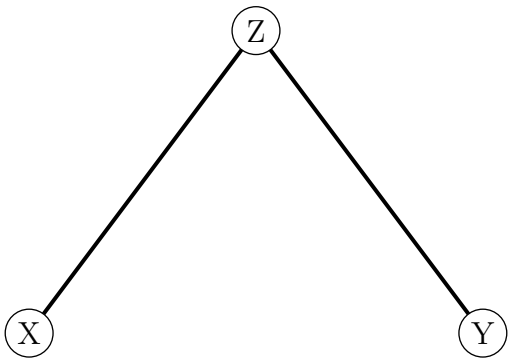

Figure 5.1: Undirected graph example

An undirected graphical model, often referred to as a Markov random field, has a set of vertices and edges just like a directed graph but the difference between these models arises from the construct of the edges; for an undirected graph the edges carry no direction.

Suppose for an undirected graph, we have three vertices, X, Y and Z, for which we

have the following conditional independence statement

$$X \perp\!\!\!\perp Y \mid Z$$

We say Z separates X from Y in the graph G. This relationship can be seen in Figure 5.1.

Gaussian Graphical Models

Let $\boldsymbol{X} = (X_1, ..., X_p) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. From the properties of the multivariate normal distribution, we know the marginal distributions will all follow a normal distribution also. As correlation models have dependence in normal data, the conditional independence structure for a Gaussian graphical model is defined in $\boldsymbol{\Sigma}$. We use the precision matrix to model this structure.

$$\text{Precision of } (X_1, ..., X_p) : \boldsymbol{\Omega} = \begin{pmatrix} \omega_{11} & \dots & \omega_{1p} \\ \vdots & \ddots & \vdots \\ \omega_{1p} & \dots & \omega_{pp} \end{pmatrix} \tag{5.1}$$

Using the precision matrix, we can say that X_k is conditionally independent of X_l given all other X_j if and only if $\omega_{kl} = 0$. To find conditional independence in our variables, we need to observe zeroes in our precision matrix.

For the following precision matrix:

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} & 0 & \omega_{14} \\ \omega_{21} & \omega_{22} & \omega_{23} & 0 \\ 0 & \omega_{32} & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{pmatrix} \tag{5.2}$$

the resulting graphical model is shown in Figure 5.2.

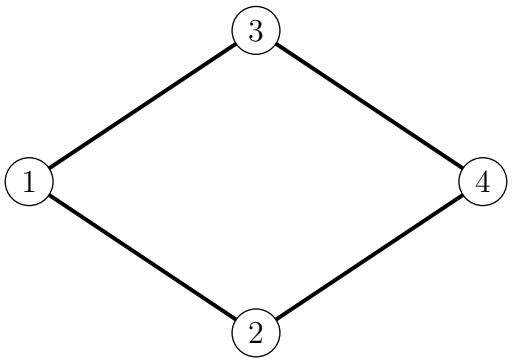

Figure 5.2: Undirected graph for the precision structure in Equation 5.2

5.2 Bayesian Gaussian Graphical Models

Here, we discuss how the Bayesian Gaussian graphical model is constructed which we shall use to model the conditional dependence between the response variables in the hierarchical model.

Let $G = (V, E)$ be an undirected graph, with V denoting the set of vertices and E the set of existing edges (Lauritzen, 2006). Let

$$\mathcal{W} = \{(i, j) \mid i, j \in V, i < j\} \tag{5.3}$$

and $\bar{E} = \mathcal{W} \backslash E$, where $\bar{E}$ denotes the set of non-existing edges. A Gaussian graphical model with respect to G is defined as:

$$\mathcal{M}_G = \{N_p(0, \Sigma) \mid \boldsymbol{\Omega} = \Sigma^{-1}\} \tag{5.4}$$

Here, our $\boldsymbol{\Omega}$ value is obtained from the Bayesian hierarchical model. Later, we will explain how we jointly use the precision estimates for the model error and the random effects present within the model. Let $\boldsymbol{Z} = (Z^{(1)}, \ldots, Z^{(n)})^\intercal$ be an i.i.d. sample of size n from $\mathcal{M}_G$. The likelihood function is defined as

$$P(\boldsymbol{Z} \mid \boldsymbol{\Omega}, G) \propto |\boldsymbol{\Omega}|^{n/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Omega} V)\right\} \tag{5.5}$$

where $V = \boldsymbol{Z}^\intercal \boldsymbol{Z}$.

The prior distribution for the precision matrix now no longer is the Wishart distribution which we used as the conjugate prior in the standard hierarchical model, but is now updated instead to be the G-Wishart distribution, which is the conjugate prior for a Bayesian Gaussian graphical model. The G-Wishart has the following density:

$$P(\mathbf{\Omega} \mid G) = \frac{1}{I_G(\nu, \mathbf{S})} |\mathbf{\Omega}|^{(\nu-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{\Omega S})\right\} \tag{5.6}$$

where $\nu > 2$ is the degree of freedom and $\mathbf{S}$ is a symmetric positive definite matrix corresponding to the relevant precision matrix.

The normalising constant, $I_G(\nu, \mathbf{S})$, is defined as

$$I_G(\nu, \mathbf{S}) = \int_{\mathbb{P}_G} |\mathbf{\Omega}|^{(\nu-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{\Omega S})\right\} dK \tag{5.7}$$

The Wishart density and G-Wishart appear similar at first glance, though there are two main differences, one being the aforementioned normalising constant shown in Equation 5.7. The second is a constraint on the space of matrices obtained through the density, denoted by $\mathbb{P}_G$, which denotes the space of $p \times p$ positive definite matrices with entries (i, j) equal to zero whenever $(i, j) \in \bar{E}$.

For non-decomposable graphs, $I_G(\nu, \mathbf{S})$ has no closed form solution, but it is possible to numerically approximate the integral using a MC method proposed in Atay-Kayis and Massam (2005), which is detailed more in the following section.

Monte Carlo method for computing $I_G(\nu, \mathbf{S})$

Given an arbitrary graph G and given ν and $\mathbf{S}$, to compute the normalising constant, we first compute the Cholesky decomposition $\mathbf{S}^{-1} = T^\top T$. For G, denote by p the number of vertices.

1. Create a $p \times p$ triangular matrix $A = (a_{ij})$ such that $a_{ij} = 0$, if $(i, j) \in \bar{E}$ or if $i = j$, and $a_{ij} = 1$ otherwise.

2. Using A, find h_i, the number of 1's in the ith row of A, and k_i, the number of 1's in the ith column of A. Define $T_{[ij]} = t_{ij}/t_{jj}$. Choose a sample size N and, for $n = 1, ..., N$, go through the following steps:

3. Sample the free variables ψ_{ij}^n, for $(i,j) \in E$ as follows: for $i = 1,...,p$, $\psi_{ii}^n = \sqrt{U_i}$, where $U_i \sim \chi^2_{\nu+h_i}$; then for $i = 1,...,(p-1), j = (i+1),...,p$ and $a_{ij} = 1$, $\psi_{ij}^n = V_{ij}$, where $V_{ij} \sim N(0,1)$.

4. Evaluate ψ_{ij}^n for $(i,j) \in \bar{E}$ as follows, for $i = 1,...,(p-1)$ and for $j = (i+1),...,p$; if $i = 1$ and $a_{ij} = 0$, then $\psi_{ij}^n = -\sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]}$; otherwise, if $i > 1$ and $a_{ij} = 0$ then

$$\psi_{ij}^n = -\sum_{k=i}^{j-1} \psi_{ik} t_{\langle kj]} - \sum_{r=1}^{i-1} \left(\frac{\psi_{ri} + \sum_{l=r}^{i-1} \psi_{rl} t_{\langle li]}}{\psi_{ii}} \right) \left(\psi_{rj} + \sum_{l=r}^{j-1} \psi_{rl} t_{\langle lj]} \right)$$

   The values ψ_{ij}^nm for $(i,j) \in \bar{E}$ are computed line by line and therefore, for a given (i,j), all values ψ_{rs}^n, for $(r,s) < (i,j)$, are available for computing ψ_{ij}^n.

5. Compute $\exp\left\{-\frac{1}{2}\sum_{(i,j)\in\bar{E}} (\psi_{ij}^n)^2\right\}$

6. Compute

$$\hat{J}_{\nu,T}^{MC} = \frac{1}{N} \sum_{k=1}^{N} \left[\exp\left\{ -\frac{1}{2} \sum_{(i,j)\in\bar{E}} (\psi_{ij}^n)^2 \right\} \right] \tag{5.8}$$

   and multiply it by

$$C_{\nu,T} = \prod_{i=1}^{p} (2\pi)^{h_i/2}\, 2^{(\nu+h_i)/2}\, \Gamma\left(\frac{\nu+h_i}{2}\right)\, t_{ii}^{\nu+b_i-1} \tag{5.9}$$

   to obtain $\hat{I}_G(\nu, \mathbf{S})$. Note $b_i = h_i + k_i + 1$

We have implemented this MC algorithm in *R* with the iterative section coded in *C++* (C++, 2017) to improve computational speed with a view to implementation when performing model selection. While testing the algorithm, we discovered a discrepancy in values obtained for $C_{\nu,T}$ shown in Equation 5.9. Using the simple case when G is complete, the G-Wishart distribution reduces to the Wishart distribution, which has a closed form for its normalising constant. The computed normalising constant from the MC algorithm did not yield the same result as the closed form solution.

Using the `BDgraph` package (Mohammadi and Wit, 2016) in *R*, which has an implementation of the MC algorithm, we tested this to compare the output. The results from the `BDgraph` function match that of the closed form solution for the Wishart normalising

constant. Noting this, we have implemented the interpretation of $C_{\nu,T}$ used here, which is defined as

$$\begin{aligned} C_{\delta,T} = &\left(\sum_{i=1}^{i}\sum_{j=1}^{j} A_{ij}/2\right)\ln(\pi) + \left(p + \frac{\nu}{2} + \sum_{i=1}^{i} A_{ii}\right)\ln(2) \\ &+ \sum_{i=1}^{i}\left[\ln\,\Gamma\left(\frac{\nu + h_i}{2}\right)\right] + \sum_{i=1}^{i}(\nu + h_i + k_i)\,\ln|T_{ii}| \end{aligned} \tag{5.10}$$

5.2.1 Sampling from the G-Wishart distribution

There are several sampling methods that can be used to generate from a G-Wishart distribution. See Dobra (2011) for a review of existing methods. Lenkoski (2013) proposes a direct sampling method for the G-Wishart distribution which is detailed below:

Algorithm 5: Direct sampler from precision matrix Given a graph $G = (V, E)$ with precision matrix $\mathbf{\Omega}$, where $\Sigma = \mathbf{\Omega}^{-1}$:

1. Set $\Delta = \Sigma$
2. Repeat for $i = 1, \ldots, p$ until convergence:
   (a) Let $N_i \subset V$ be the set of neighbours of node i in graph G. Form Δ_{N_i} and $\Sigma_{N_i,i}$ and solve
   $$\hat{\beta}_i^* = \Delta_{N_i}^{-1}\Sigma_{N_i,i},$$
   (b) Form $\hat{\beta}_i \in \mathbb{R}^{p-1}$ by copying the elements of $\hat{\beta}_i^*$ to the appropriate locations and zeroes in those locations not connected to i in graph G,
   (c) Replace $\Delta_{i,-i}$ and $\Delta_{-i,i}$ with $\Delta_{-i,-i}\hat{\beta}_i$
3. Return $\mathbf{\Omega} = \Delta^{-1}$

We have provided an implementation of the G-Wishart sampler in our model code, written in *C++* which is implemented within *R* using `RCpp` package (Eddelbuettel and Francois, 2011). The benefits of generating a direct G-Wishart sampler comes in terms of the model selection. Lenkoski (2013) goes on to discuss a model selection algorithm which uses the exchange algorithm (Murray et al., 2006), a popular tool for MCMC

schema when working with models where the likelihood has an intractable normalising constant, much like the G-Wishart distribution. The exchange algorithm is used in the graphical model selection which we discuss in more detail in the next section.

5.3 Bayesian Gaussian graphical model selection

The final part to forming the Bayesian Gaussian graphical model is determining which graph G provides the best description of the relationships between our observations. For a graph with p nodes, there are a total of $2^{p(p-1)/2}$ possible graphs. Even for a moderately sized number of vertices, the problem can explode quickly. Due to this, we need to implement an efficient search algorithm which can explore the space of graphs $\mathcal{G}$ to find the true underlying graph efficiently.

Several model selection algorithms have been proposed, mainly implementing a trans-dimensional MCMC algorithm which explores the model space whilst simultaneously estimating parameters. The most common example of this is the reversible-jump MCMC (Green, 1995). Algorithms of this nature have been implemented in several Gaussian graphical model selection works, such as Dobra (2011). The main drawback to these methods revolves around the calculation of the normalising constant $I_G(\nu, \mathbf{S})$ which requires the use of MC approximation as discussed previously (Atay-Kayis and Massam, 2005).

Lenkoski (2013) and Wang and Li (2012) both propose alternative approaches borrowing ideas from the exchange algorithm and the double Metropolis-Hastings algorithm (Laing, 2010). The work of Wang and Li (2012) does not use a direct G-Wishart sampler, unlike the work of Lenkoski (2013).

One drawback to methods implementing reversible jump steps is that some moves between models may be rejected according to the acceptance probability. This can be inefficient in high-dimensional problems. Wit and Mohammadi (2015) propose an adaptation of the birth-death MCMC (BDMCMC) (Cappe, 2001) where moves between models are always accepted, though the trade-off for this is an increase in computational complexity.

If we look back to the hierarchical model, we note that we have multiple precision

matrices, one for our residual error $\mathbf{\Omega}_{\boldsymbol{\epsilon}}$ and precision matrices for each random effect group $\mathbf{\Omega}_{\tilde{\mathbf{b}}_g}$. Considering all the algorithms we have just outlined, all of them only consider one precision estimate as input for their model selection. Due to this, we shall have to modify any algorithm we consider.

Through the list of algorithms and methods discussed, we have chosen to modify the PAS algorithm discussed in Wang and Li (2012). One of the main changes we implement alongside expanding the algorithm is using a direct G-Wishart sampler, such as the one discussed in Section 5.2.1, which removes the need to use the block Gibbs update and deal with the intractable normalising constants.

Our Bayesian Gaussian graphical model selection problem can be broken down into two parts. The case when our number of responses $V \leq 3$ and when they are $V > 3$. In the case when the number of responses is ≤ 3, we can solve the model selection problem in closed form by exploiting chordality.

A graph G is said to be chordal if every graph cycle of length four or greater has a cycle chord. Put simply, for a given graph, there is no point in the graph where we could cover four or more vertices without encountering a connection between two vertices that includes one we have already covered. Figure 5.3 provides an example of this. As we can see, it is not possible to go to vertices $V = \{1, 2, 3, 4\}$ without hitting either 1 or 4, which contains a connection and is a cycle chord. All graphical models where there are three or less vertices are chordal.

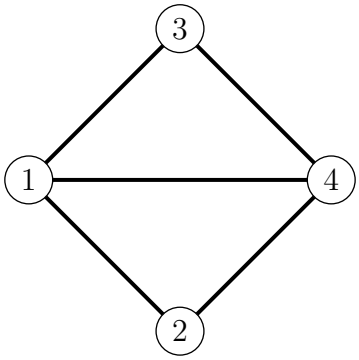

Figure 5.3: Chordal graph example

It is possible to obtain the normalising constant, $I_{\mathrm{G}}(\nu, \mathbf{S})$ for the G-Wishart distri-

bution for chordal graphs via a closed form solution. The normalising constant can be factorised into a product of density functions as shown in Equation 5.11:

$$I_{\mathrm{G}}(\nu, \mathbf{S}) = \frac{\prod_{i=1}^{d} I_{T_i}(\nu, S_{T_i,T_i})}{\prod_{j=1}^{d-1} I_{S_i}(\nu, S_{S_i,S_i})}. \tag{5.11}$$

where T_i are the cliques and S_i are the separators of G.

By exploiting chordality, we can calculate the density straightforwardly. As there are no more than $2^3 = 8$ possible graphs for these cases, this takes insignificant computational time and all possible graphs can be considered at each iteration of the sampler.

For the case when $V > 3$, we use a modification of the PAS algorithm (Wang and Li, 2012), which allows for input from multiple precision matrices like we have in the Bayesian hierarchical model. We detail the modified PAS algorithm below, which due to the independence between the precision estimates, consists of a direct expansion.

Suppose we have two graphs, $G = (V, E)$ and $G' = (V, E')$ which differ by one edge (i, j) and suppose edge $(i, j) \in E$ and $E' = E \backslash (i, j)$ say. The acceptance probability for a move G to G' according to a proposal $q(G'|G)$ is then:

$$\alpha(G \to G') = \min\left[1, \frac{p(G' \mid \mathbf{\Omega}_{\boldsymbol{\epsilon}} \backslash (\omega_{ij}, \omega_{jj}), \hat{\mathbf{y}}) \prod_{k=1}^{K} \left[p(G' \mid \mathbf{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \backslash (\omega_{ij}, \omega_{jj}), \tilde{\mathbf{b}}_{\mathbf{k}})\right] q(G \mid G')}{p(G \mid \mathbf{\Omega}_{\boldsymbol{\epsilon}} \backslash (\omega_{ij}, \omega_{jj}), \hat{\mathbf{y}}) \prod_{k=1}^{K} \left[p(G| \mid \mathbf{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \backslash (\omega_{ij}, \omega_{jj}), \tilde{\mathbf{b}}_{\mathbf{k}})\right] q(G' \mid G)}\right] \tag{5.12}$$

where the conditional posterior odds against the edge (i, j) for a general $\mathbf{\Omega}$ is given by:

$$\frac{p(G' \mid \mathbf{\Omega} \backslash (\omega_{ij}, \omega_{jj}), \mathbf{y})}{p(G \mid \mathbf{\Omega} \backslash (\omega_{ij}, \omega_{jj}), \mathbf{y})} = \frac{p(\mathbf{y}, \mathbf{\Omega} \backslash (\omega_{ij}, \omega_{jj}) \mid G') p(G')}{p(\mathbf{y}, \mathbf{\Omega} \backslash (\omega_{ij}, \omega_{jj}) \mid G) p(G)} \tag{5.13}$$

As shown in Equation (5.6) in Wang and Li (2012), $p(\mathbf{y}, \mathbf{\Omega} \backslash (\omega_{ij}, \omega_{jj}) \mid G')$ has a closed

analytical form. For multiple precision matrices, this can be expressed as:

$$
\begin{aligned}
&p(\mathbf{y}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \backslash (\omega_{ij}, \omega_{jj}) \mid G') \prod_{k=1}^{K} \left[p(\mathbf{y}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \backslash (\omega_{ij}, \omega_{jj}) \mid G') \right] \\
&\quad = (2\pi)^{-\frac{np}{2}} \frac{I\left(b_{\boldsymbol{\epsilon}} + n, S^{\boldsymbol{\epsilon}}_{jj} + D^{\boldsymbol{\epsilon}}_{jj}\right)}{I_{G'}(b_{\boldsymbol{\epsilon}}, D^{\boldsymbol{\epsilon}})} \left| \boldsymbol{\Omega}^{0}_{\boldsymbol{\epsilon} V\backslash j, V\backslash j} \right|^{\frac{n+b_{\boldsymbol{\epsilon}}-2}{2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \left(S^{\boldsymbol{\epsilon}} + D^{\boldsymbol{\epsilon}} \right) \boldsymbol{\Omega}^{0}_{\boldsymbol{\epsilon}} \right\} \right] \\
&\quad \times \prod_{k=1}^{K} \left[(2\pi)^{-\frac{np_{\tilde{\mathbf{b}}_{\mathbf{k}}}}{2}} \frac{I\left(b_{\tilde{\mathbf{b}}_{\mathbf{k}}} + n, S^{\tilde{\mathbf{b}}_{\mathbf{k}}}_{jj} + D^{\tilde{\mathbf{b}}_{\mathbf{k}}}_{jj}\right)}{I_{G'}(b_{\tilde{\mathbf{b}}_{\mathbf{k}}}, D^{\tilde{\mathbf{b}}_{\mathbf{k}}})} \left| \boldsymbol{\Omega}^{0}_{\tilde{\mathbf{b}}_{\mathbf{k}} V\backslash j, V\backslash j} \right|^{\frac{n_{\tilde{\mathbf{b}}_{\mathbf{k}}}+b_{\boldsymbol{\epsilon}}-2}{2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \left(S^{\tilde{\mathbf{b}}_{\mathbf{k}}} + D^{\tilde{\mathbf{b}}_{\mathbf{k}}} \right) \boldsymbol{\Omega}^{0}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \right\} \right] \right]
\end{aligned}
\tag{5.14}
$$

where general $\boldsymbol{\Omega}^0 = \boldsymbol{\Omega}$ except for an entry 0 in the positions (i, j) and (j, i) and an entry c in the position (j, j), where $c = \boldsymbol{\Omega}_{j,V\backslash j} \left(\boldsymbol{\Omega}_{V\backslash j, V\backslash j} \right)^{-1} \boldsymbol{\Omega}_{V\backslash j, j}$. $I(b, D)$ is the normalising constant of a scalar G-Wishart distribution $W_G(b, D)$.

In a similar fashion, a closed form expression is obtained for $p(\mathbf{y}, \boldsymbol{\Omega} \backslash (\omega_{ij}, \omega_{jj}) \mid G)$. For multiple precision matrices, this can be expressed as:

$$
\begin{aligned}
&p(\mathbf{y}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \backslash (\omega_{ij}, \omega_{jj}) \mid G) \prod_{k=1}^{K} \left[p(\mathbf{y}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \backslash (\omega_{ij}, \omega_{jj}) \mid G) \right] \\
&\quad = (2\pi)^{-\frac{np}{2}} \frac{J\left(b_{\boldsymbol{\epsilon}} + n, D^{\boldsymbol{\epsilon}}_{ee} + S^{\boldsymbol{\epsilon}}_{ee}, a_{11}\right)}{I_G(b_{\boldsymbol{\epsilon}}, D_{\boldsymbol{\epsilon}})} \left| \boldsymbol{\Omega}^{1}_{\boldsymbol{\epsilon} V\backslash e, V\backslash e} \right|^{\frac{n+b_{\boldsymbol{\epsilon}}-2}{2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \left(S_{\boldsymbol{\epsilon}} + D_{\boldsymbol{\epsilon}} \right) \boldsymbol{\Omega}^{1}_{\boldsymbol{\epsilon}} \right\} \right] \\
&\quad \times \prod_{k=1}^{K} \left[(2\pi)^{-\frac{n_{\tilde{\mathbf{b}}_{\mathbf{k}}} p}{2}} \frac{J\left(b_{\tilde{\mathbf{b}}_{\mathbf{k}}} + n, D^{\tilde{\mathbf{b}}_{\mathbf{k}}}_{ee} + S^{\tilde{\mathbf{b}}_{\mathbf{k}}}_{ee}, a_{11}\right)}{I_G(b_{\tilde{\mathbf{b}}_{\mathbf{k}}}, D_{\tilde{\mathbf{b}}_{\mathbf{k}}})} \left| \boldsymbol{\Omega}^{1}_{\tilde{\mathbf{b}}_{\mathbf{k}} V\backslash e, V\backslash e} \right|^{\frac{n+b_{\tilde{\mathbf{b}}_{\mathbf{k}}}-2}{2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \left(S_{\tilde{\mathbf{b}}_{\mathbf{k}}} + D_{\tilde{\mathbf{b}}_{\mathbf{k}}} \right) \boldsymbol{\Omega}^{1}_{\tilde{\mathbf{b}}_{\mathbf{k}}} \right\} \right] \right]
\end{aligned}
\tag{5.15}
$$

where

$$
J(h, B, a_{11}) = \left(2\pi B_{22}^{-1} \right)^{\frac{1}{2}} a_{11}^{\frac{h-1}{2}} I(h, B_{22}) \exp\left\{ -\frac{1}{2} \left(B_{11} - B_{22}^{-1} B_{12}^{2} \right) a_{11} \right\} \tag{5.16}
$$

Let general $\boldsymbol{\Omega}^1 = \boldsymbol{\Omega}$ except for entries of $\boldsymbol{\Omega}_{e,V\backslash e} \left(\boldsymbol{\Omega}_{V\backslash e, V\backslash e} \right)^{-1} \boldsymbol{\Omega}_{V\backslash e, e}$ in the positions corresponding to e. We let $A = \boldsymbol{\Omega}_{ee|V\backslash e}$ in Equation 5.16, where $\boldsymbol{\Omega}_{ee|V\backslash e} = \boldsymbol{\Omega}_{ee} - \boldsymbol{\Omega}_{e,V\backslash e} \left(\boldsymbol{\Omega}_{V\backslash e, V\backslash e} \right)^{-1} \boldsymbol{\Omega}_{V\backslash e, e}$ and a_{11} corresponds to the first element of $\boldsymbol{\Omega}_{ee|V\backslash e}$.

We then plug in 5.14 and 5.15 into 5.12 to provide the acceptance rated for a move

from G to G' and obtain:

$$\alpha(G \to G') = \min\left\{1, \frac{p(G')q(G \mid G')I_G(b_\epsilon, D_\epsilon)\prod_{k=1}^K I_G(b_{\tilde{\mathbf{b}}_\mathbf{k}}, D_{\tilde{\mathbf{b}}_\mathbf{k}})}{p(G)q(G' \mid G)I_{G'}(b_\epsilon, D_\epsilon)\prod_{k=1}^K I_{G'}(b_{\tilde{\mathbf{b}}_\mathbf{k}}, D_{\tilde{\mathbf{b}}_\mathbf{k}})} H(e, \mathbf{\Omega}_\epsilon)\prod_{k=1}^K H(e, \mathbf{\Omega}_{\tilde{\mathbf{b}}_\mathbf{k}})\right\} \tag{5.17}$$

where, for a general $\mathbf{\Omega}$,

$$H(e, \mathbf{\Omega}) = \frac{I(b+n, D_{jj}+S_{jj})}{J(b+n, D_{ee}+S_{ee}, a_{11})}\left(\frac{\left|\mathbf{\Omega}^0_{V\setminus j, V\setminus j}\right|}{\left|\mathbf{\Omega}^1_{V\setminus e, V\setminus e}\right|}\right)^{\frac{n+b-2}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{(S+D)\left(\mathbf{\Omega}^0 - \mathbf{\Omega}^1\right)\right\}\right] \tag{5.18}$$

can be analytically evaluated.

Note that the intractable normalising constants still remain within the computation of the acceptance probability. Wang and Li (2012) go on in their work to remove these normalising constants using the exchange algorithm. This involves substituting the normalising constants with an unbiased estimate based on a single sample from the prior, where a new precision matrix $\mathbf{\Omega}'$ is sampled based on the updated graph G'. This gives us an updated acceptance probability of:

$$\begin{aligned}\alpha(G \to G') = \min\Bigg\{1, &\frac{p(G')q(G \mid G')f\left(\mathbf{\Omega}'_\epsilon \setminus (\omega'_{ij}, \omega'_{jj}) \mid G\right)\prod_{k=1}^K f\left(\mathbf{\Omega}'_{\tilde{\mathbf{b}}_\mathbf{k}} \setminus (\omega'_{ij}, \omega'_{jj}) \mid G\right)}{p(G)q(G' \mid G)f\left(\mathbf{\Omega}'_\epsilon \setminus (\omega'_{ij}, \omega'_{jj}) \mid G'\right)\prod_{k=1}^K f\left(\mathbf{\Omega}'_{\tilde{\mathbf{b}}_\mathbf{k}} \setminus (\omega'_{ij}, \omega'_{jj}) \mid G'\right)} \\ &H(e, \mathbf{\Omega}_\epsilon)\prod_{k=1}^K H(e, \mathbf{\Omega}_{\tilde{\mathbf{b}}_\mathbf{k}})\Bigg\}\end{aligned} \tag{5.19}$$

where

$$f(\mathbf{\Omega}' \setminus (\omega'_{ij}, \omega'_{jj}) \mid G') = I(b, D_{jj})\left|\mathbf{\Omega}'_{0, V\setminus j, V\setminus j}\right|^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(D\mathbf{\Omega}'_0)\right\} \tag{5.20}$$

and

$$f(\mathbf{\Omega}' \setminus (\omega'_{ij}, \omega'_{jj}) \mid G) = J(b, D_{ee}, a_{11})\left|\mathbf{\Omega}'_{1, V\setminus e, V\setminus e}\right|^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(D\mathbf{\Omega}'_1)\right\} \tag{5.21}$$

for general $\mathbf{\Omega}$.

We define a modified version of Algorithm 2 in Wang and Li (2012), as we have our own direct sampler for the G-Wishart density detailed below:

Algorithm 6: Modified PAS algorithm for multiple precision matrices Given the current state $\left\{G, \mathbf{\Omega}_{\boldsymbol{\epsilon}}, \mathbf{\Omega}_{\tilde{\mathbf{b}}}, G', \mathbf{\Omega}'_{\boldsymbol{\epsilon}} \backslash (\omega'_{ij}, \omega'_{jj}) \mathbf{\Omega}'_{\tilde{\mathbf{b}}} \backslash (\omega'_{ij}, \omega'_{jj})\right\}$

1. Update $\left\{G', \mathbf{\Omega}'_{\boldsymbol{\epsilon}} \backslash (\omega'_{ij}, \omega'_{jj}) \mathbf{\Omega}'_{\tilde{\mathbf{b}}} \backslash (\omega'_{ij}, \omega'_{jj})\right\}$
   - Propose a new graph G' differing by only one edge from G from the proposal distribution $q(G' \mid G)$.
   - Generate $\mathbf{\Omega}'_{\boldsymbol{\epsilon}}, \mathbf{\Omega}'_{\tilde{\mathbf{b}}}$ using the G-Wishart sampler in Section 5.2.1
2. Update G
   - Exchange G and G'
   - Accept G' with probability α, defined in Equation 5.19.
3. Update $\mathbf{\Omega}_{\boldsymbol{\epsilon}}, \mathbf{\Omega}_{\tilde{\mathbf{b}}}$ conditional on the most recent G using the G-Wishart sampler.

We can implement the modified PAS algorithm when $V > 3$ to perform model selection for the conditional dependence between response variables within our graphical model. In the next chapter, we will detail how we combine the Bayesian Gaussian graphical model within our Bayesian hierarchical model to obtain our full chain graph like graphical structure.

5.4 Discussion

In this chapter, we have discussed how to infer an undirected graphical model to visualise the relationship present between response variables. We have extended beyond the standard model search algorithms for Bayesian Gaussian graphical models, which use one precision matrix as input to infer the graphical model structure, to a multiple precision case, which is the standard output from a Bayesian hierarchical model. By combining all of the precision estimates together, we are able to obtain a more robust measure of the dependency present between the response variables.

In the next chapter, we look to use this model search algorithm to propose a novel inference tool to visualise the output from a Bayesian hierarchical model in the form of a graphical model.

Chapter 6

Visualising Hierarchical Models Using Graphical Models

In this chapter, we look to expand upon the Bayesian hierarchical model we have constructed in Chapters 3 and 4 and implement a novel inference tool which can provide a straightforward representation of which factors are influencing vowel variation and change in the Glaswegian dialect. We use graphical models to aid this visualisation.

The method works by jointly inferring the Bayesian hierarchical model with a Bayesian Gaussian graphical model as discussed in Chapter 5 to model the conditional dependence between responses, using the precision estimates from the hierarchical model as input. From this, we use a chain graph style structure to visualise the combined model output between the fixed effects and the vowel formants.

The motivation behind presenting the hierarchical model in a graphical framework is to ease understanding of the complex model output of the hierarchical model to users less familiar with the construction of multiple response hierarchical models. Through the graphical model visualisation, it is straightforward to infer which factors impact on vowel variation on each formant, and the dependency present between each of the formants.

Section 6.1 introduces some further graphical model concepts which are used within the graph structure. Section 6.2 introduces the chain graph like structure we use to visualise Bayesian hierarchical model output beginning with an explanation of the structure and how it is implemented. We then detail how the sampler for the parameter estimates

is updated to reflect this change. Section 6.3 provides two applications of the chain graph model structure, firstly with a simulated example and then secondly with an application to the Sounds of the City corpus.

6.1 Graphical Models

In this section, we extend to other graphical model structures that are used within the visualisation we implement, discussing the relevant theory behind these different types of models. We will work up to the concept of a chain graph, which is the framework which we build our graphical model upon.

6.1.1 Directed Graphical Models

A directed graphical model, often referred to as a directed acyclic graph (DAG) $\mathcal{D}$ is a graph where all the edges between vertices are directed. An example is shown in Figure 6.1. If we apply the product rule here, we can factorise the joint distribution of this DAG as shown in Equation 6.1. We see from the DAG, that y and z have a dependence on x, and x is independent of the other variables.

$$p(x, y, z) = p(z \mid x)p(y \mid x)p(x) \tag{6.1}$$

The graphical model is constructed by taking each conditional distribution from above and adding a directed link from the vertices corresponding to the variables on which the distribution is conditioned. If we have a link going from vertice x to vertice y then vertice x is called the parent of vertice y and vertice y is the child of vertice x.

We can characterise a DAG by a simple rule for expanding the joint probability in terms of simpler conditional probabilities. Let $X_1, \ldots, X_n$ be a set of random variables represented by corresponding vertices in the graph. Let $\text{pa}[i]$ denote the parents of vertice i and denote $\mathcal{X}_{\text{pa}[i]}$ be the set of variables associated with $\text{pa}[i]$. Then

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathcal{X}_{\text{pa}[i]}) \tag{6.2}$$

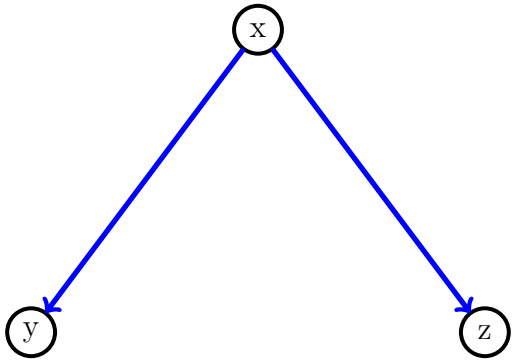

Figure 6.1: Directed acyclic graph example for variables x, y and z

d-separation

Consider a directed graph where X, Y, Z are arbitrary nonintersecting sets of vertices. We wish to ascertain whether a particular conditional independence statement $X \perp\!\!\!\perp Y \mid Z$ is implied. To do this, we consider all the possible paths from any vertice in X to any vertice in Y. Any path is *blocked* if it includes a vertice that either

(a) the arrows on the path meet either head to tail or tail to tail at the vertice, and the vertice is in Z, or

(b) the arrows meet head to head at the vertice, and neither the vertice or any of its descendants is in C.

If all the paths are blocked, then X is said to be d-separated from Y by Z and the joint distribution over all the variables in the graph will satisfy $X \perp\!\!\!\perp Y \mid Z$.

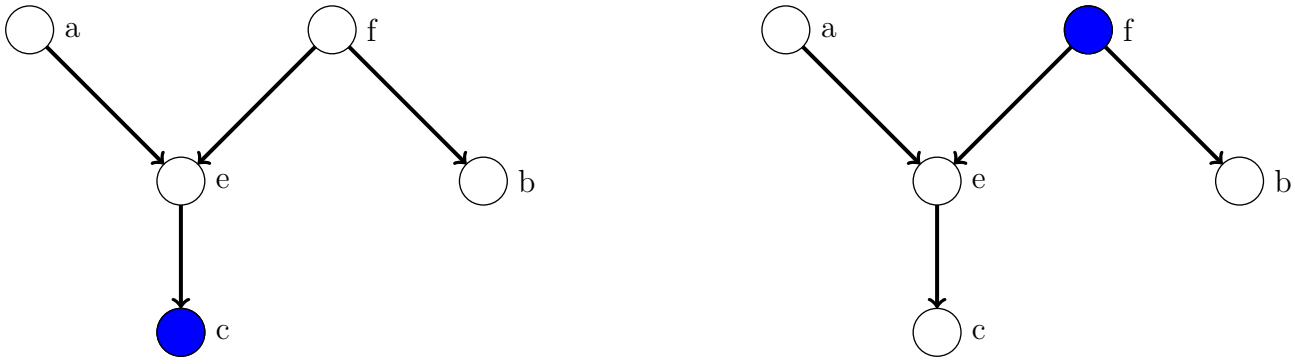

Figure 6.2: d-separation example

In the example in Figure 6.2, starting by considering the left graph, the path from a to b is not blocked by vertice f as it is a tail to tail vertice for this path and is not

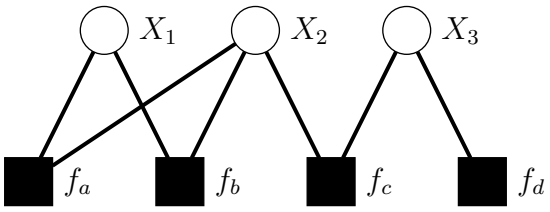

Figure 6.3: Factor graph illustration

observed. It is not blocked by vertice e either, although this is a head to head vertice, it has a descendant c because it is in the conditioning set. So the conditional independence statement $a \perp\!\!\!\perp b \mid c$ does **not** follow from this graph.

With the right hand graph, the path from a to b is blocked by f because this is a tail to tail vertice that is observed, so the conditional independence statement $a \perp\!\!\!\perp b \mid c$ **will** hold for any distribution that factorises according to this graph.

6.1.2 Factor graphs

Both directed and undirected graphs allow a global function of several variables to be expressed as the product of factors over subsets of those variables. Factor graphs make this decomposition explicit by introducing additional vertices for the factors themselves in addition to the vertices representing the variables. They allow us to be more explicit about the details of the factorisation.

The joint distribution over a set of variables can be written as a product of factors like so:

$$p(\boldsymbol{x}) = \prod_s f_s(\boldsymbol{x}_s)$$

where $\boldsymbol{x}_s$ denotes a subset of the variables. Denote the individual variables by x_i.

In a factor graph, there is a vertice for every variable in the distribution. There are also additional factor vertices, which are often depicted by a square, for each factor in the joint distribution. There are also undirected links connecting each factor vertice to all of the variable vertices on which that factor depends.

In the example in Figure 6.3, we can express the factorisation as the following:

$$p(\boldsymbol{x}) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_4)$$

Note here we have two factors f_a and f_b defined over the same set of variables x_1 and x_2. In an undirected graph, the product of two such factors would be joined together in the same clique. The factor graph keeps such factors explicit and thus conveys more detailed information about the underlying factorisation.

Factor graphs are described as *bipartite* because they consist of two distinct kinds of vertices, and all links go between vertices of opposite type.

6.1.3 Chain Graphs

Chain graphs look to combine both directed acyclic graphs and undirected graphs into one graphical form. Vertices are partitioned into blocks, with one common partitioning of blocks being a block of variables of interest and a block of explanatory variables. The edges within blocks are undirected and the edges connecting vertices between blocks are directed.

An important Markov property for chain graphs is the global Markov property (Gottard and Rampichini, 2006), which is based on the definition of the moral graph. Starting from a given chain graph, a moral graph can be obtained by connecting parents of common children and then converting all the arrows into undirected edges. The global Markov property combines the concept of conditional independence to that of separation between vertices in the moral graph. For example, for a given graph, if a set of vertices S separates the vertices in A from the vertices in B so each path from A to B passes by some vertice in S, then $A \perp\!\!\!\perp B \mid S$. These Markov properties induce a factorization of the joint distribution of the variables in a model.

If we look at Figure 6.4, we see the basic structure of a chain graph model for two given blocks of vertices. The dependency structure between the vertices in the left block is modelled using undirected edges, while the dependency structure between blocks is denoted by directed edges.

From Figure 6.4, we can see how it would be possible to model the output of a hierarchical model in such a fashion. If we imagine the black vertices are our explanatory variables within the model, and the white vertices are our response variables, we could visualise the relationships present within the hierarchical model using a structure similar in layout to a chain graph model, albeit not strictly adhering to the global Markov

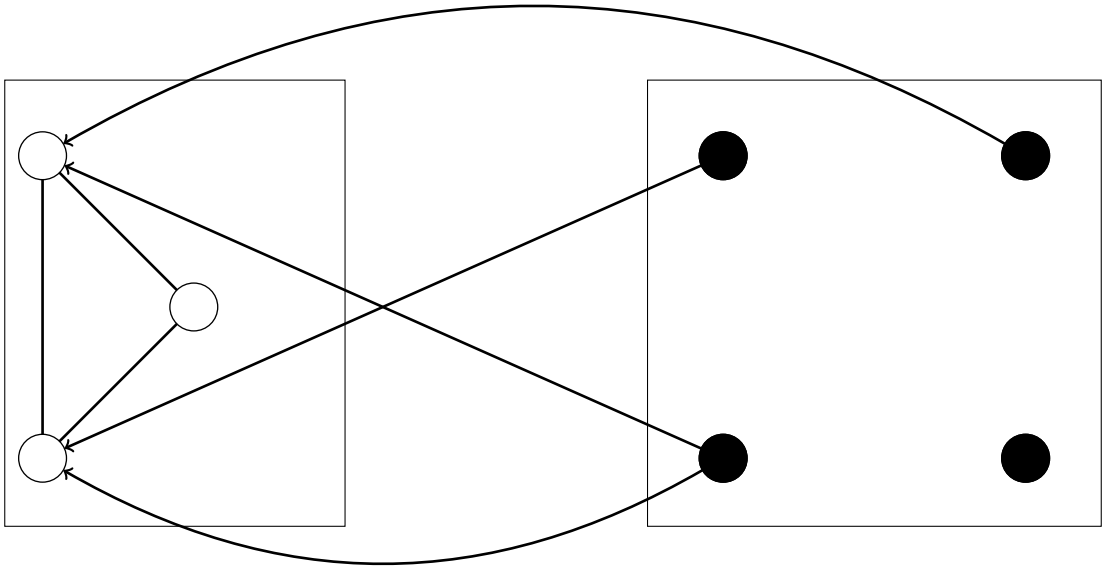

Figure 6.4: Illustration of a chain graph model.

property within the traditional chain graph.

6.2 Using a Chain Graph Style Model for the Hierarchical Model

Here, we look to utilise some of the ideas discussed in Section 6.1 and implement those to allow us to construct a graphical representation of the Bayesian hierarchical model. The visual design of the chain graph model discussed in Section 6.1.3 provides a visualisation that lends itself naturally to a regression design.

Using a chain graph like structure, we can split vertices into partitioned blocks, which could be viewed as a separation between a block of explanatory variables and a block of response variables. The directed edges between vertices in each block corresponds to a predictor variable being a significant predictor of a response variable. The lack of an edge present indicates that the predictor has no significant effect on the response variable, i.e. $\beta_j = 0$. Interaction terms are represented using a factor graph notation, whereby the interaction terms first connect through a relevant factor variable, then an arrow is extended from the factor variable to the response of interest.

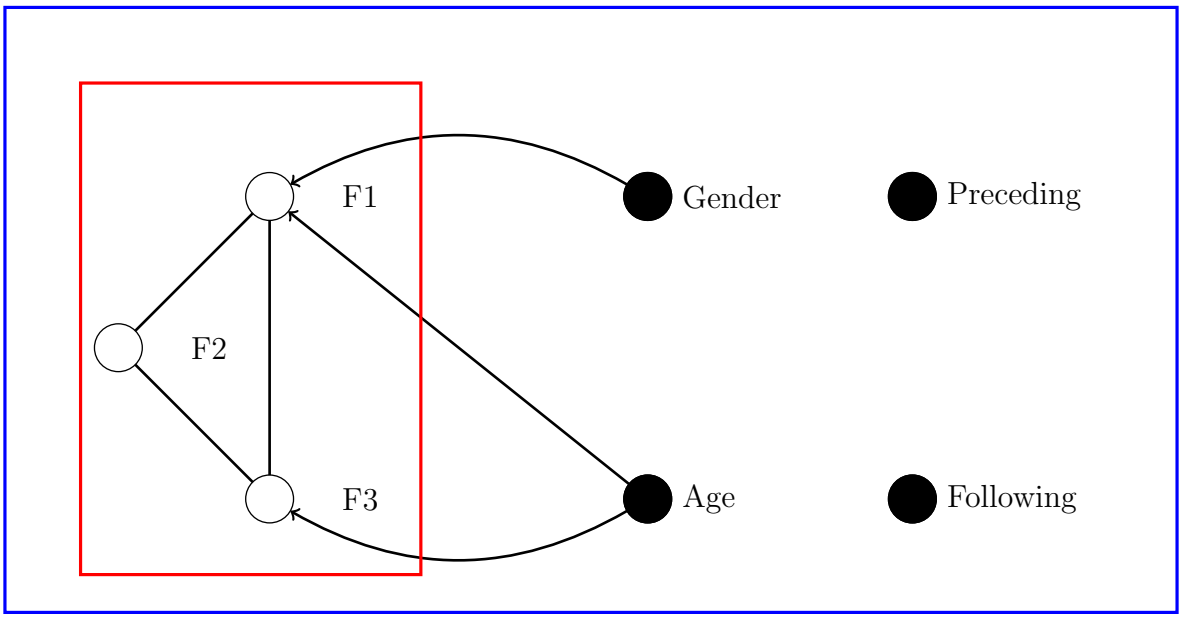

Figure 6.5: Illustration of the Chain graph style model output This graph is stylised to the Sounds of the City corpus. The directed edges are modelled by the Bayesian hierarchical model, the undirected graph for the response variables modelled using the Bayesian graphical model.

The relationship between variables in each block can be modelled using an undirected graph structure. The relationship between the response variables is of more interest than the relationship between the explanatory variables, so our model will focus on the relationship between the response variables only. This model could be extended to include the graphical relationship between the explanatory variables, where the relationship could be modelled either by a Bayesian Gaussian graphical model or a log-linear model if our explanatory variables are discrete, though we do not discuss this in detail here. The undirected graph for the response variables can be modelled using a Bayesian Gaussian graphical model, where the precision estimates from the hierarchical model are used as input.

Figure 6.5 illustrates how we can construct the graphical model visualisation, with the corresponding modelling techniques used to construct each part of the graph highlighted.The DAG is constructed using the Bayesian hierarchical model output. The relationship between the response variables is modelled using the Bayesian Gaussian graphical model. The example in Figure 6.5 has been stylised to the Sounds of the City corpus.

6.2.1 Updating the Hierarchical Model

Now that we have samplers for the Bayesian Gaussian graphical model, we need to update the Bayesian hierarchical model to incorporate graphical model selection. This is done by updating our priors on the precision estimates. We now change from the Wishart prior to the G-Wishart prior. This leads to the following new priors for the model precisions:

$$\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \sim \mathcal{W}_G(\nu_{\boldsymbol{\epsilon}}, \mathbf{S}_{\boldsymbol{\epsilon}}) \qquad \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}} \sim \mathcal{W}_G(\nu_{\tilde{\mathbf{b}}_{\mathbf{g}}}, \mathbf{S}_{\tilde{\mathbf{b}}_{\mathbf{g}}}) \tag{6.3}$$

Using our new priors and including the step for hierarchical centering and parameter expansion, our updated Gibbs sampler is of the following form:

$$\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\right) \tag{6.4}$$

$$\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} \propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right) \tag{6.5}$$

$$\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} \propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid\mid \left[\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\boldsymbol{\delta}}_k, \left[\tilde{\mathbf{X}}^\top_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\right) \tag{6.6}$$

$$\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} \propto \mathcal{W}_G\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}^{-1}_{\tilde{\mathbf{b}}_g} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}^\top_{g_i}\right]^{-1}\right) \tag{6.7}$$

$$\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} \propto \mathcal{W}_G\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}^{-1}_{\boldsymbol{\epsilon}} + \sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^\top\right]^{-1}\right) \tag{6.8}$$

$$\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} \propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}^l_{\tilde{\boldsymbol{\eta}}^l}\|}{2}, b_l + \frac{\sum_{m=1}^{p}\left(\beta^l_m\right)^2}{2}\right) \tag{6.9}$$

where we sample $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g respectively, every $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_{\mathbf{k}}}$ is sampled for every group of random effects which has nested coefficients k and τ_l^2 for each response level l.

We define $\mathbf{z}_{\boldsymbol{\beta}^l} = \omega_{j,j}\mathbf{y}^l + \sum_{k=1}^{k\neq l}\omega_{j,k}\left(\mathbf{y}^k - \mathbf{X}_{\boldsymbol{\eta}^k}\boldsymbol{\beta}^k\right)$ and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} = \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}_{\tilde{\mathbf{b}}_{-g}}\tilde{\mathbf{b}}_{\tilde{\mathbf{b}}_{-g}}$, where

$\tilde{\mathbf{b}}_{-g}$ denotes $\tilde{\mathbf{b}}$ excluding group g and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}$ is the mean value calculated for $\mathbf{y}_{\tilde{\mathbf{b}}_{g,h}}$ for each response level l. for each response level l. and $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$ respectively.

The nesting step in Equation 6.6 has parameters defined as $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k} = \text{blockdiag}\left(\mathbf{X}^1{}_{\tilde{\boldsymbol{\delta}}_k}, \ldots, \mathbf{X}^l{}_{\tilde{\boldsymbol{\delta}}_k}\right)$, where $\tilde{\boldsymbol{\delta}}_k = \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} + \tilde{\mathbf{U}}_k\tilde{\mathbf{b}}_k$ for each block of nested coefficients k.

The Bayesian Gaussian graphical model selection step occurs prior to the draws for the precision matrices. Once the process has determined our current G, we draw $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for $g = 1, \ldots G$ from Equations 6.8 and 6.7 respectively.

Finally, we perform parameter expansion at the end of the sampler for the relevant random effects coefficients and precisions by performing a Metropolis-Hastings step as detailed in 4.2.2 , accepting with probability ϕ, shown in Equation 4.18.

Figure 6.6 provides a graphical representation of the full hierarchical model with graphical model selection. The main difference with this model from the representation in Figure 4.13 is the change in prior for the precision estimates from the Wishart to the G-Wishart distribution.

The model is constructed by implementing the following algorithm:

Algorithm 6: The Bayesian hierarchical model sampler with mixing improvements Given initial parameter estimates $\boldsymbol{\theta}^{(0)} = \left(\tilde{\boldsymbol{\beta}}^{(0)}, \tilde{\boldsymbol{\eta}}^{(0)}, \tilde{\mathbf{b}}^{(0)}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(0)}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(0)}, \boldsymbol{\tau}^{(0)}\right)$.
Then
For $t = 1, \ldots, T$

1. For $l = 1, \ldots, L$,

   (a) Sample $\boldsymbol{\beta}^{l,(t)}$ from 6.4.

   (b) Propose new model state $\boldsymbol{\eta}^{l,(t)}$. Sample $\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$ from 6.4. Compute 3.15, where $\boldsymbol{\eta}^{l,0}$ is the current model state. If $u < \alpha$, where $u \sim \mathcal{U}(0,1)$, set $\boldsymbol{\beta}^{l,(t)} = \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{l,(t)}}$, else $\boldsymbol{\beta}^{l,(t)}$ remains the same.

   Form $\tilde{\boldsymbol{\beta}}^{(t)}_{\tilde{\boldsymbol{\eta}}^{(t)}} = \left(\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{1,(t)}}, \ldots, \boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^{L,(t)}}\right)$

2. For $g = 1, \ldots, G$

   For $h = 1, \ldots, H$

   (a) Sample $\tilde{\mathbf{b}}^{(t)}_{g,h}$ from 6.5.

Form $\tilde{\mathbf{b}}_g^{(t)} = \left(\tilde{\mathbf{b}}_{g,1}^{(t)}, \ldots, \tilde{\mathbf{b}}_{g,H}^{(t)}\right)^\top$

Form $\tilde{\mathbf{b}} = \left(\tilde{\mathbf{b}}_1^{(t)}, \ldots, \tilde{\mathbf{b}}_G^{(t)}\right)^\top$

3. For $k = 1, \ldots, K$

   (a) Sample $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ from 6.6

   Form $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}} = \left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_1}, \ldots, \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_K}\right)$

4. Using precision estimates $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t-1)}$ and $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(t-1)}$, obtain the current graph G using 5.11 if $V \leq 3$, else, use Algorithm 5.

5. For $g = 1, \ldots, G$,

   Sample $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{(t)}$ from 6.7.

   Form $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)}$ by $\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}^{(t)} = \text{blockdiag}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_1}^{(t)}, \ldots, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_G}^{(t)}\right)$.

6. Sample $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{(t)}$ from 6.8.

7. For $g = 1, \ldots, G$

   (a) Form $\tilde{\mathbf{b}}_{\mathbf{g}}^{*(t)} = \alpha \tilde{\mathbf{b}}_{\mathbf{g}}^{(t)}$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}^{*(t)} = \alpha \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}^{(t)}$. Compute 4.18. If $u < \phi$, where $u \sim \mathcal{U}(0,1)$, set $\tilde{\mathbf{b}}_{\mathbf{g}}^{(t)} = \alpha \tilde{\mathbf{b}}_{\mathbf{g}}^{(t)}$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}^{(t)} = \alpha \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}^{(t)}$, else $\tilde{\mathbf{b}}_{\mathbf{g}}^{(t)}$ and $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}^{(t)}$ remain the same.

8. For $l = 1, \ldots, L$,

   Sample $\tau_l^{(t)}$ from 6.9.

   Form $\boldsymbol{\tau}^{(t)} = \left(\tau_1^{(t)}, \ldots, \tau_L^{(t)}\right)$

Performing all these steps, we can obtain a chain graph model like structure as shown in Figure 6.5.

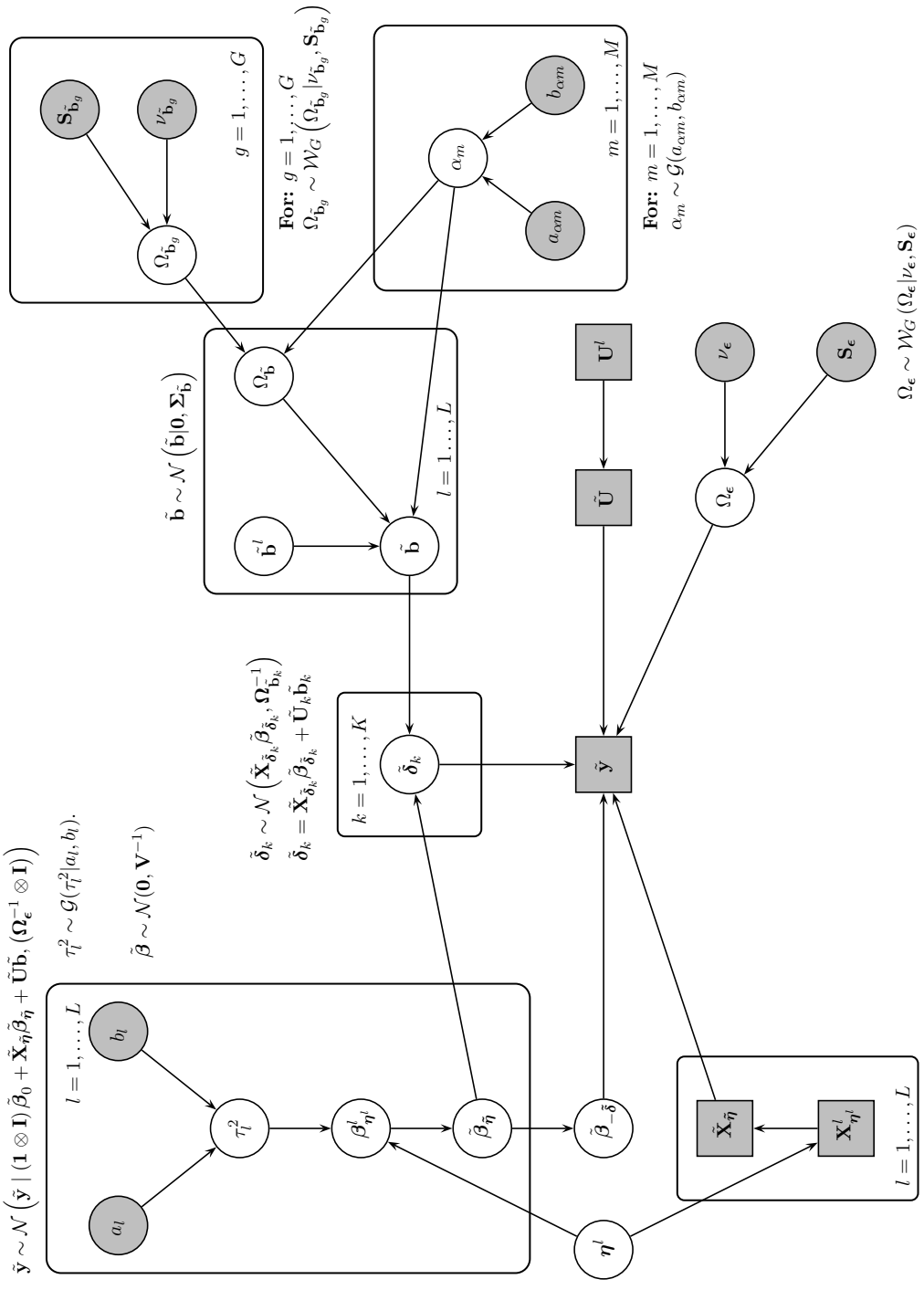

Figure 6.6: Representation of the full hierarchical model with graph selection as a PGM. The PGM is constructed in a similar style as Figure 4.13, though this time we have updated the priors on the precision parameters to be adjusted for a G-Wishart distribution.

6.3 Application of Graphical Models

In this section, we will provide an implementation of the graphical models we have discussed in this chapter in the way of a simulated example constructed in a similar fashion to the multiple response nesting problem in Section 4.1.2 and also with an application to the Sounds of the City corpus.

6.3.1 Simulated Example

To illustrate the chain graph model structure, we consider a simulated example which is constructed in a similar way to the problem in Section 4.1.2, though now we consider four response variables instead of three, which are independent of one another. We consider a simple design problem with four fixed effects, where one is nested within the one random effect denoted by the *Gender* variable, also within the model, with some coefficients being randomly assigned zero coefficient values.

We run the sampler for 10,000 iterations with the nesting step added for the nested coefficient. Hyperparameters are set at $a_l = b_l = 1 \times 10^{-3}$ and $\nu_{\boldsymbol{\epsilon}} = 3, \mathbf{S}_{\boldsymbol{\epsilon}} = 0.001 \cdot \mathbf{I}_3$ and $\nu_{\mathbf{b}_g} = 3, \mathbf{S}_{\mathbf{b}_g} = 0.001, \cdot \mathbf{I}_3$.

We look to produce the four "best" graphs, determined by their posterior probability, which corresponds to the number of times a particular graph is selected. Figure 6.7 highlights the four top graphs selected by their model posterior probability. We observe the top two graphs differ only by the significance of one term, the Gender coefficient on Y_3, which for model 1 is not present, and for model 2 is present. We also notice a similar trend between models 1 and 2 and models 3 and 4, where the gender coefficient is not present in Y_2 for our top 2 models, but is selected in models 3 and 4.

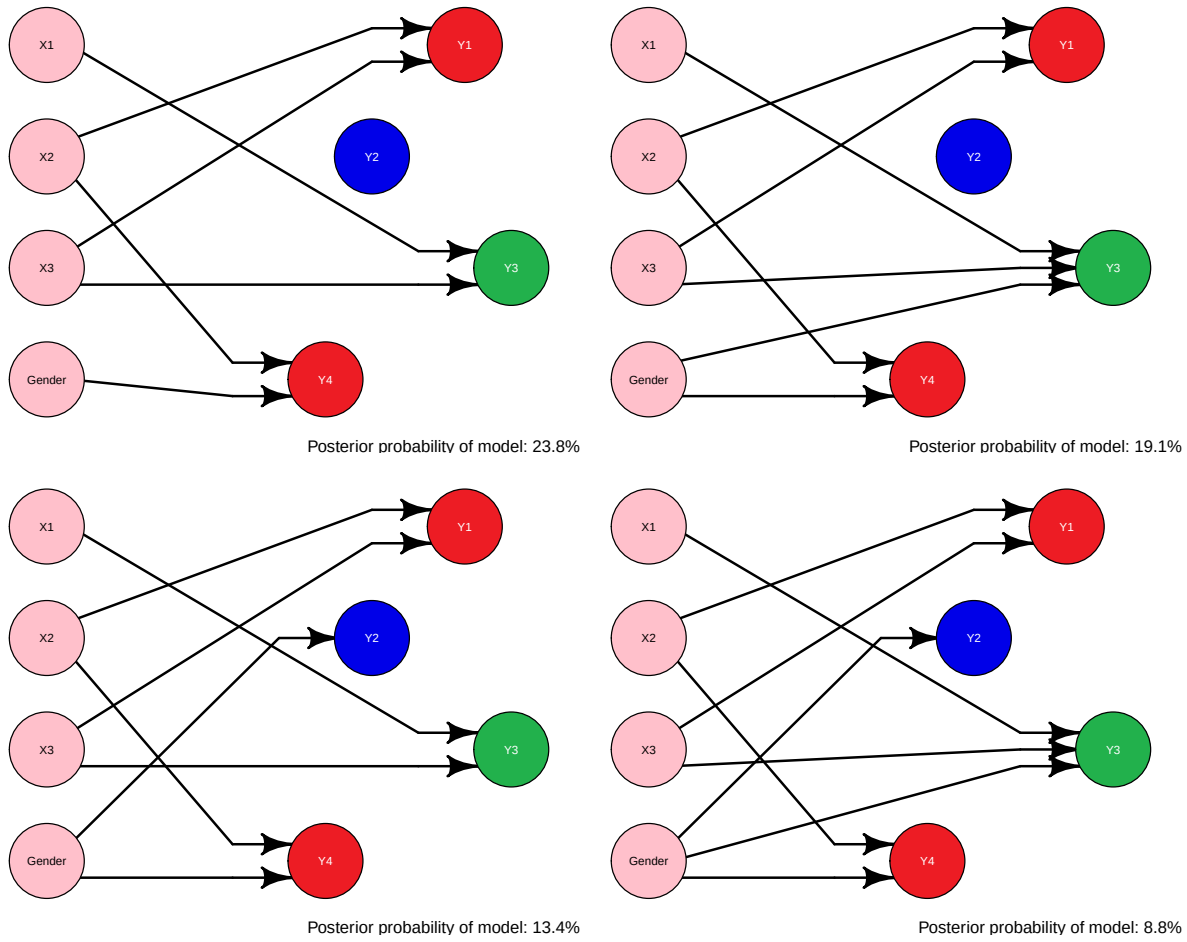

Figure 6.7: Graphical models obtained for the simulated example. The best four graphs, determined by posterior probability for the simulated example, run for 10,000 iterations. The top two graphs are selected for similar times, differing only by the significance of the Gender coefficient on the Y_3 response.

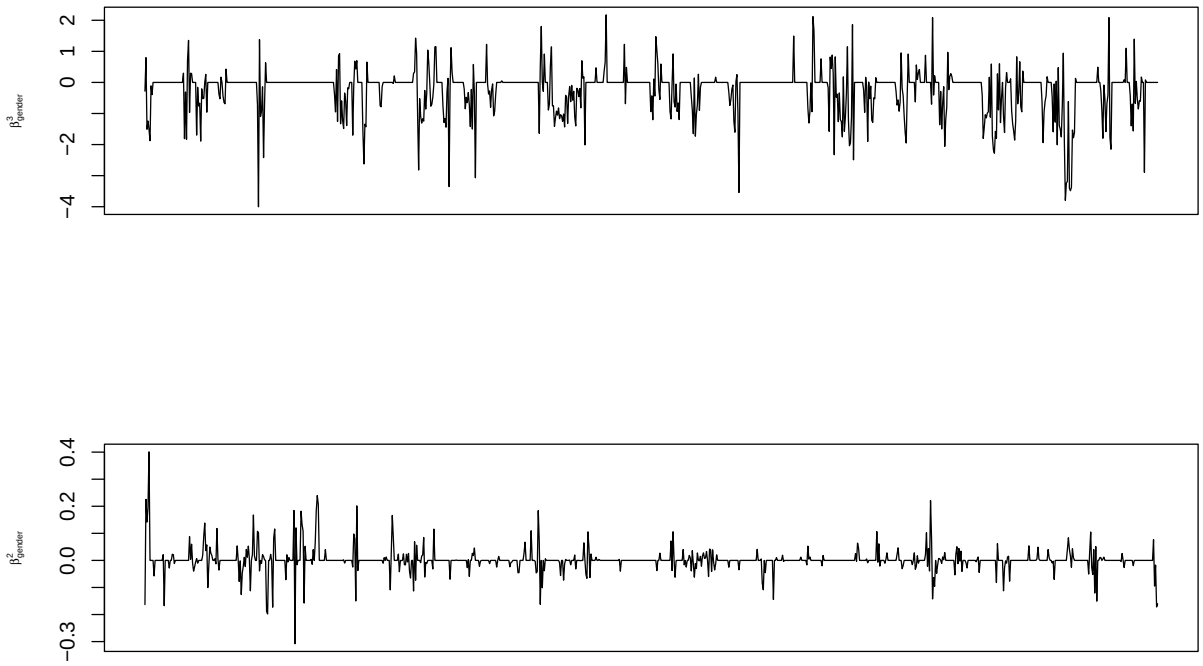

Figure 6.8: Trace plots for Gender coefficient on Y_2 and Y_3. Traceplots for the Gender coefficent on Y_2 and Y_3 for 10,000 iterations. We observe periods in the sampler where the terms are not selected (at zero) and smaller periods where the term is added to the model.

We can see closer at how often both terms are selected by their traceplots in Figure 6.8. We see from these traceplots that there are periods for both coefficients where both terms are not included in the model. The effect of gender on Y_3 does appear to be larger than on Y_2, but a model including both could be considered.An adjustment on the hyperparameters for $\tilde{\boldsymbol{\beta}}$ could lead to a more parsimonious model, leading to a reduction in the number of times these terms are selected. It is worth noting that for Y_2, the coefficient is not significant, but for Y_3, it is significant, albeit with a small coefficient value.

6.3.2 Sounds of the City Corpus

We will now apply the chain graph model structure to the Sounds of the City corpus data. We now implement the combined sampler in Section 6.2.1, implementing the centering step for the nested coefficients within Speaker and Word and also the parameter expansion step for the Word effect.

We focus on the *GOAT* vowel for raw mean formant measurements on F1, F2 and F3, with all 2-way interactions across the fixed effects. We run the sampler for 10,000 iterations with fixed hyperparameters $a_l = b_l = 1 \times 10^{-3}$ and $\nu_{\boldsymbol{\epsilon}} = 3, \mathbf{S}_{\boldsymbol{\epsilon}} = 1 \times 10^{-3} \cdot \mathbb{I}_3$, $\nu_{\mathbf{b}_g} = 3, \mathbf{S}_{\mathbf{b}_g} = 1 \times 10^{-3} \cdot \mathbb{I}_3$ and $a_\alpha = 500, b_\alpha = 510$. For graphical model selection, we are able to exploit chordality as we have only three nodes, so every possible graph is considered at each stage and the best fitting graph to the precision estimates selected. Model selection is also enabled within the sampler for model fitting.

We obtain the best four graphs by posterior probability as shown in Figure 6.9. For the best graph by posterior probability, we observe that the raw mean formant measurements share a conditional dependence, as shown by their fully connected graph. F1 is influenced by Age, with the vowel lowering in younger speakers and Gender influences F3, with females showing higher frequency values than males. The model also finds effects for following place of articulation in F2, with frequency values falling in general, though with greater levels of magnitude for the dorsal and labial factors, this would indicate vowel quality retraction in these contexts. Most importantly, we observe a change in F3 for Decade, indicating that for recordings measured in the 2000s, frequency values are increasing compared to recordings taken in the 1970s, indicating a shortening of the front cavity; this could relate to less lip rounding for this vowel over time (this is a new finding, since previous work has not analysed F3).

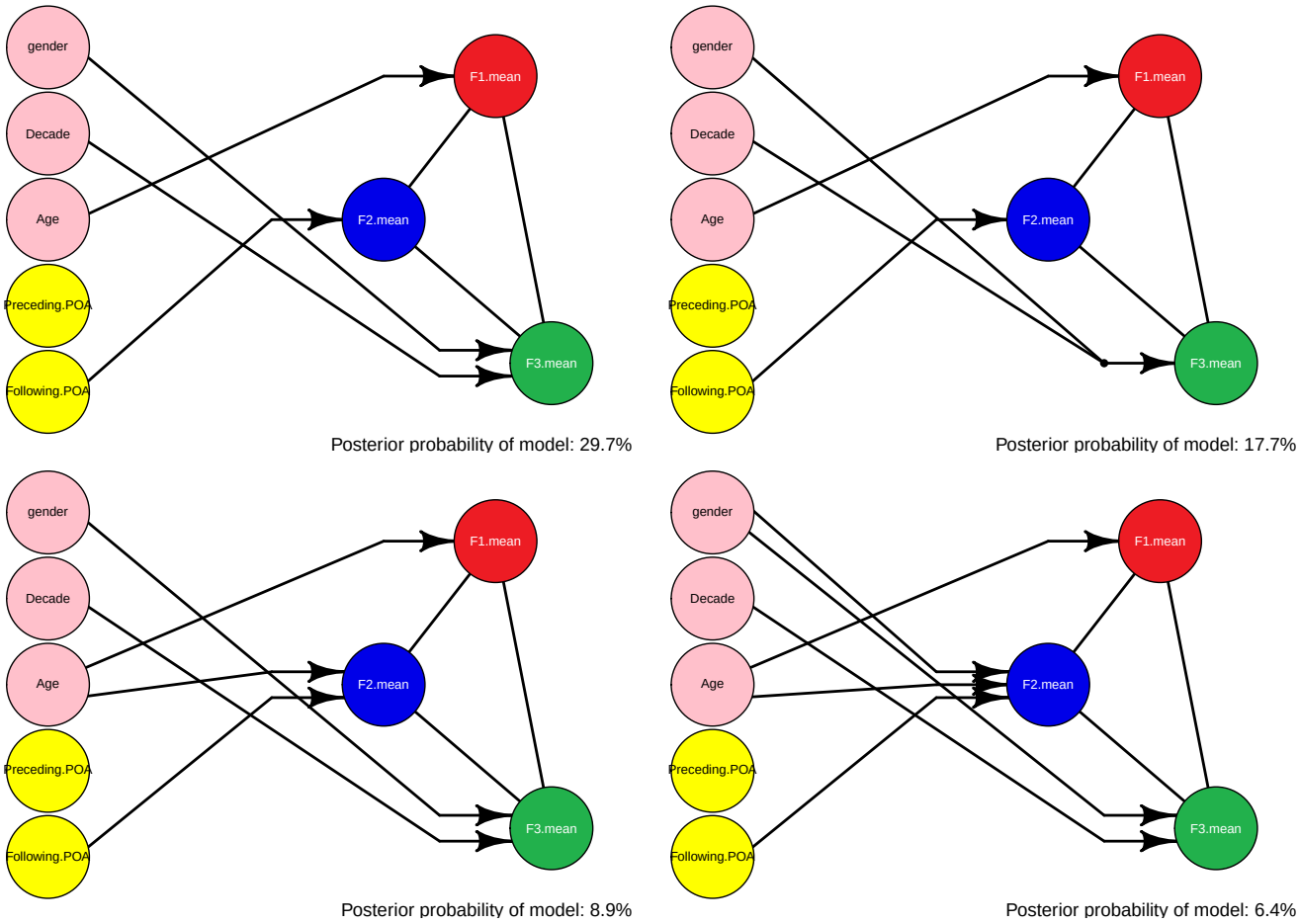

Figure 6.9: Graphical models obtained for *GOAT* vowel The best four graphical models by posterior probability obtained for the *GOAT* vowel. We observe a prominent Gender and Decade effect on F3 across all models.

If we consider the second most often selected model, the main difference comes with the addition of an interaction between Gender and Decade acting upon F3, instead of the factors being independent of one another. This is quite similar to the main model, with the difference being in the additional interpretation of the interaction, where we see that over time, the vowel frequency is smaller in F3 for females. In general there is a shift in F3 by Decade, such that the F3 values are increasing (less lip rounding), but less so for females, who either show smaller F3 values, or not so much increased F3 values. Either way, that would suggest the change towards less lip rounding for GOAT is not progressing as much in females.

As both the most selected models differ by only one term, the interaction between Gender and Decade, we take a closer look at the traceplots for Gender, Decade and their interaction in Figure 6.10. We observe from the traceplots that Gender is always included in the model, and Decade is in the model almost always, with only small periods of the sampler do we observe the coefficient is zero. The interaction between Gender and Decade on F3 on the other hand is not often present in the active model. If we were to consider a more parsimonious model by updating the hyperparameters on $\boldsymbol{\tau}$, it is likely we would see this interaction removed from the active model.

To highlight how considering the hierarchical model with all formants is important, we fit the *GOAT* vowel using each formant individually, assuming independence between the formants. Figure 6.11 shows the three graphs obtained for F1, F2 and F3 independently. We observe that F1 matches well with the best graphs in Figure 6.9, selecting only Age. For F2, we observe Gender has now been included, and is selected 90% of the time. This does not match with the full graphical models, with Gender only appearing in the fourth best model, of posterior probability 6.4%. For F3, we observe that Gender and Decade are selected, matching well with the best full graphs. We note that the posterior probability for this graph is 60%, which is low compared to the graphs for F1 and F2. This is due to the interaction between Gender and Decade being selected at other points in time within the sampler.

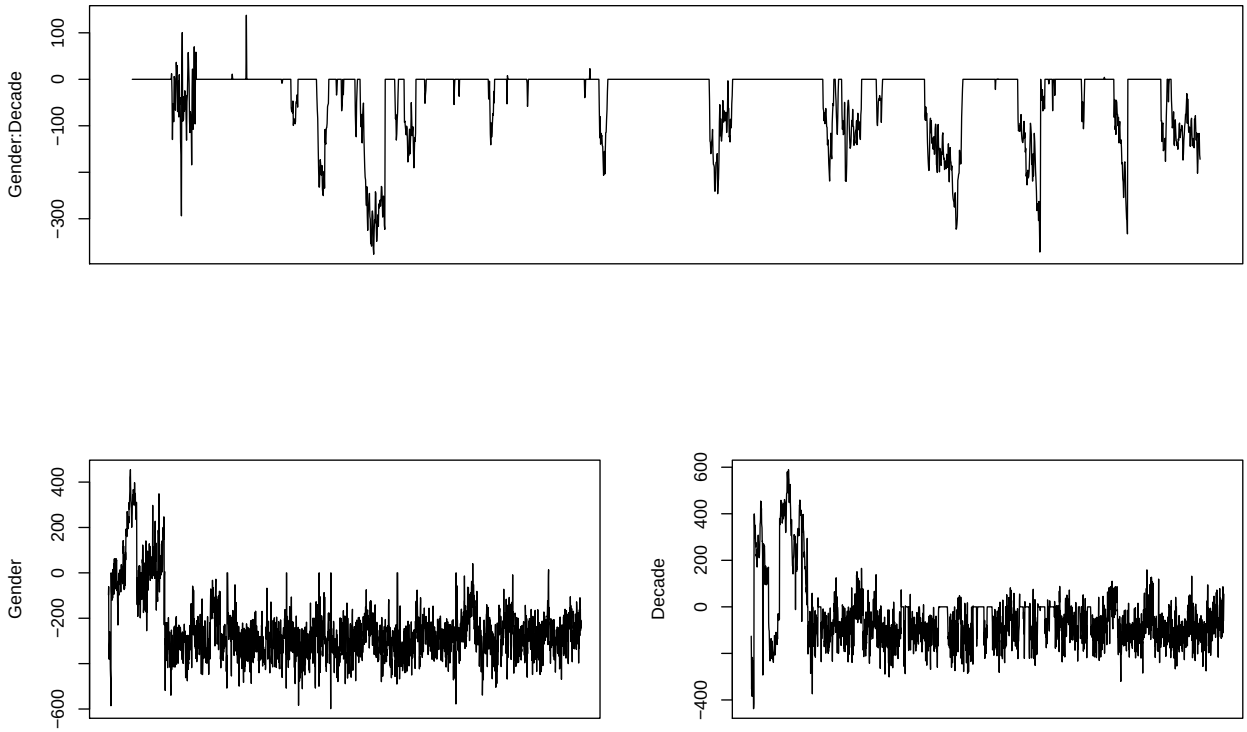

Figure 6.10: Traceplots for Gender, Decade and Gender:Decade interaction Traceplots for the Gender, Decade and Gender:Decade coefficents on $F3$ for the *GOAT* vowel for 10,000 iterations. We observe that Gender is selected always within the model, with Decade also selected frequently. The interaction between both is selected for inconsistent periods in the sampler.

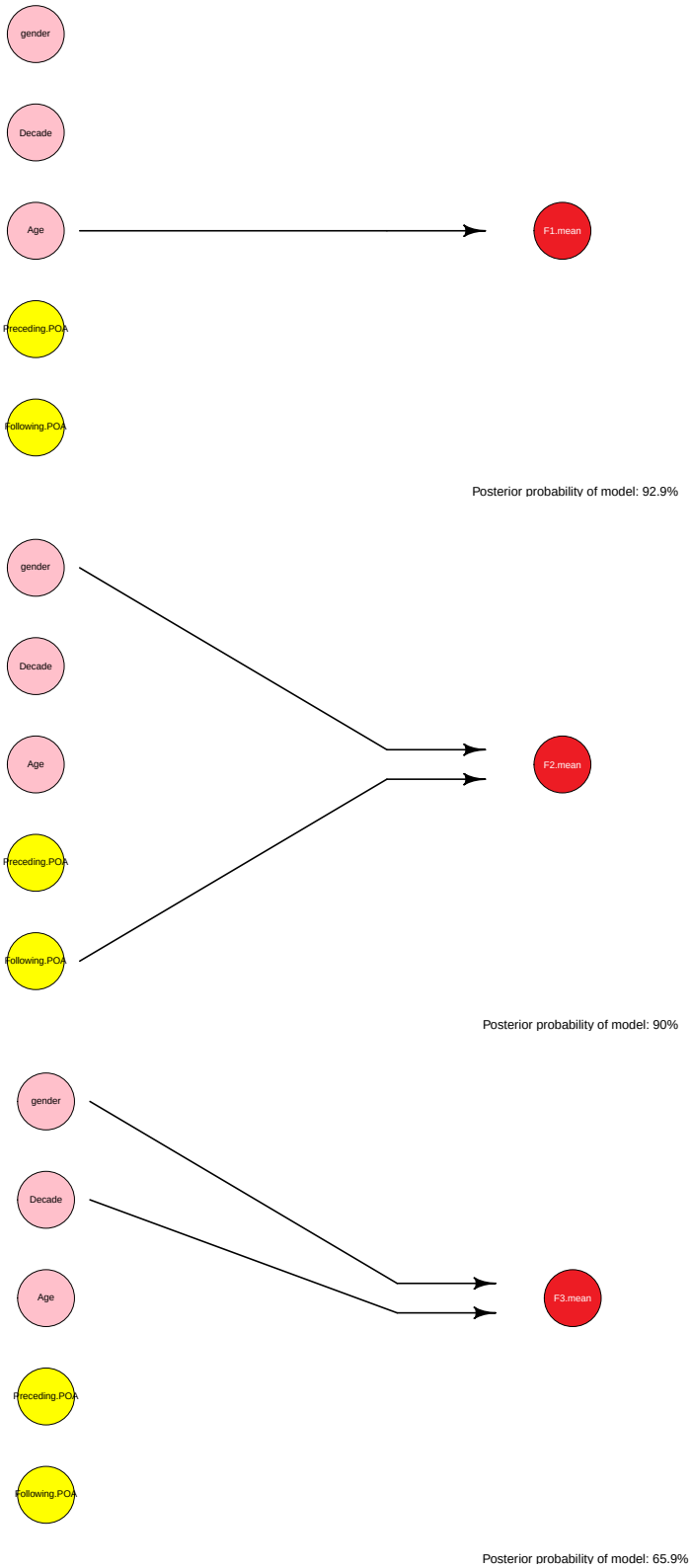

Figure 6.11: Graphs obtained for *GOAT* vowel for F1, F2 and F3 Graphs obtained for the *GOAT* vowel for 10,000 iterations fitting to each formant independently. We observe that Gender is now a significant term for F2, when it is not selected by the top models in Figure 6.9. The much lower posterior probability for the F3 model is due to the interaction between Decade and Gender at times being selected.

6.4 Discussion

In this chapter, we have introduced a novel inference tool which looks to combine the output of a Bayesian hierarchical model and present it as a graphical model, through a chain graph model structure. By implementing the hierarchical model in Chapter 3 and using the mixing modifications discussed in Chapter 4, we have been able to produce graphs for the Sounds of the City corpus for the raw formant measurements on F1, F2 and F3 and the Lobanov normalised measurements on F1 and F2, as shown in Appendix **??**. Through the graphical model, it is straightforward to see instantly which variables have a direct influence on vowel variation and change for this corpus.

Although we have applied this model specifically to the Sounds of the City corpus, it is imperative to highlight that this approach can be easily applied to any linguistic corpora, indeed whether it be on vowel change or any other linguistic phenomenon with multiple response variables. Good examples are fricatives like /s/, which are often characterised through 3 or 4 dependent variables, or stop sounds like /p/, which again are viewed using more than one acoustic variable. This modelling approach can easily be applied to data problems of a similar construct to the Sounds of the City corpus and provides a simple to understand representation of what can be in some cases vast and complex levels of output.

Chapter 7

Graphical Model Output - Sociolinguistic Discussion of Results

In this chapter, we discuss some of the results obtained from the models obtained from the corpus and fitted using the Bayesian hierarchical model. We look at how each of the vowel sounds have changed within the Sounds of the City corpus, and how these compare to the results obtained in Stuart-Smith et al. (2017), and comparing results between the raw formant data and the Lobanov normalised data. We also identify new questions of interest that have been identified through this modelling for sociolinguists to research further.

7.1 Sounds of the City Corpus - Results

Looking at the results shown within this chapter, where for each vowel we fit the Bayesian hierarchical model to the raw mean formant values for F1, F2 and F3 and the Lobanov normalised formant values for F1 and F2. Each model was run for 10,000 iterations and included the possible selection of all three-way interactions for each of the five predictor variables. Each model was fit with the same prior specification as used in Section 2.2. The four most selected models by posterior probability are shown for each vowel.

In Stuart-Smith et al. (2017), the vowels that were identified as changing within Glasgow over time were the *BOOT, COT* and *GOAT* vowels. The equivalent vowels we

observe in our results are the *FOOT, GOOSE, LOT* and *GOAT* vowels, where *FOOT* and *GOOSE* correspond to *BOOT* for Glaswegian speakers. In the new analyses presented here, raw F1, F2, F3 were modelled together as response variables. The results obtained were based on models run for F1 and F2 individually, assuming independence between the formants. From these results, we note several interesting new findings. For the *FOOT* vowel, Decade is a significant variable for most of the models, but it is always significant in F3. This differs from the results in Stuart-Smith et al. (2017), mainly as no modelling has been considered on F3, with only F1 and F2 formant measures considered, but also because the new modelling considers all three formants together. An overall new finding is a shift from the role of F1 (relating to lowering and raising of the vowels) to F3 (relating to changes in lip position). This is different for the *GOOSE* vowel, where Decade is again significant, but in F2 this time.

The findings for the *FOOT* vowel highlight the possible increase in complexity when discussing interpretation of the results in a sociolinguistic sense. Changes in F3 are linked to rounding of the lips, with more rounding present resulting in a lowering of formant measurements. For the *FOOT* vowel, we observe an increase in frequency for speakers obtained from 2000s recordings. This would indicate that less lip rounding is taking place now as opposed to previous decades, showing linguistic variation over time. This finding is entirely new: all previous work to date on the SoTC corpus has failed to capture aspects of vowel variation and change related to lip rounding. For *GOOSE*, the change in Decade is taking place within F2, which that the front/back position of the tongue during the vowel, with an increase in F2 indicating more fronting of the tongue. We observe an increase in F2 for 2000s recorded speakers, indicating that this vowel is fronting for these speakers. Again, this is a new finding with respect to existing SoTC analysis.

For the *GOAT* vowel, we again observe a prominent inclusion of Decade for F3, as opposed to F1 or F2. We see an increase in frequency for F3 for speakers from 2000s recordings, indicating less lip rounding than was found in previous decades. For the *LOT* vowel, we again observe a similar pattern, with the Decade term being significant in F3, with speakers from the 2000s showing increases in formant frequency. This means that for all three vowels shown by previous modelling to be changing with respect to vowel height (F1), we now find once correlations between the three formants are controlled for,

that changes with respect to lip position (F3) may be even more important. Changing to multiple response modelling fundamentally shifts the possible perspective on variation and change in these vowels.

General features that can be noted for all of the vowel formants, raw and normalized measures, is the significance of the Gender coefficient, which is a significant term for every vowel, indicating that Gender is one of the important impactors on vowel variation and change in this corpus. Note that the presence of Gender as a significant factor for both raw and normalised formant measures shows that Gender here has to be acting as a factor above and beyond physiological differences influencing vowel formants. Considering the formants also, most raw formant measure models show a fully connected graph present between the formants, indicating that there is indeed a correlation present between them all for each vowel, again highlighting the need to consider a model which can consider multiple response variables.

Another interesting observation is now the inclusion of Decade for several other vowels which is of interest because these vowels (*FLEECE, FACE* and *TRAP/BATH (CAT)*) were not thought to be changing. This could possibly be due to the more complex modelling we implement, finding relationships that could not be observed before. It is also possible that due to the models being fitted having to consider all three-way interactions, we have not fully explored the model space efficiently in the number of iterations, and a longer chain may have to be run in order to explore the full model space more thoroughly. This is reflected in the small posterior probabilities often observed for the most selected models, with one model seldom selected more often than the other options.

Looking closer at the Lobanov normalised vowel results in comparison to the raw formant measures, we observe several differences. Firstly, the models here only consider F1 and F2, so the results are not directly comparable with the raw formant results. For the *GOOSE* vowel, we observe a prominent significance for Decade across all models, which ties in with the result shown for the *BOOT* vowel shown in Chapter 2. For *FOOT*, we do not observe any notable Decade effect in the first three models obtained, with Following and Preceding Place of Articulation being more prominent indicators of vowel quality change, as with Age. *LOT* and *GOAT* also do not show any prominent Decade effect, though the *LOT* vowel models all have relatively low posterior probability.

GOAT appears to be more influenced by the social factors Gender and Age and also both linguistic factors. For the remaining vowels, Decade does appear as a significant term occasionally, though once again the models have low posterior probability. There is also another very clear difference: the BATH and GOAT models do not show correlation between F1 and F2. The reasons for this, with respect to the raw F1, F2, and F3 models, are not immediately clear. Comparison with models of just raw F1 and F2 formants may be informative.

Due to the design of the Sounds of the City corpus, and recordings being made on spontaneous speech, it is difficult to obtain an equally balanced sample size of vowels. For example, the *BATH* vowel contains only 327 observations, which is significantly smaller than the remaining vowels and could explain the peculiar difference in models observed for this vowel. Another point to note is that due to the design of the study, we obtain many observations on an individual speaker, which gives us a good indicator of the individual's speech characteristics, but have a relatively small sample of different speakers across different groups. This lack of available information within the data could also explain some of the varying results we observe.

7.2 Raw Mean Formant Results

Here, we provide the graphical models obtained for vowels based on their raw mean formant values for F1, F2 and F3. The models obtained were ran for 5,000 iterations of the sampler and included all three-way interactions for each of the five predictor variables. The prior specifications for all hyperparameters are set to the same specification as the models fitted in Section 3.2.2. The best four models by posterior probability are shown for each vowel.

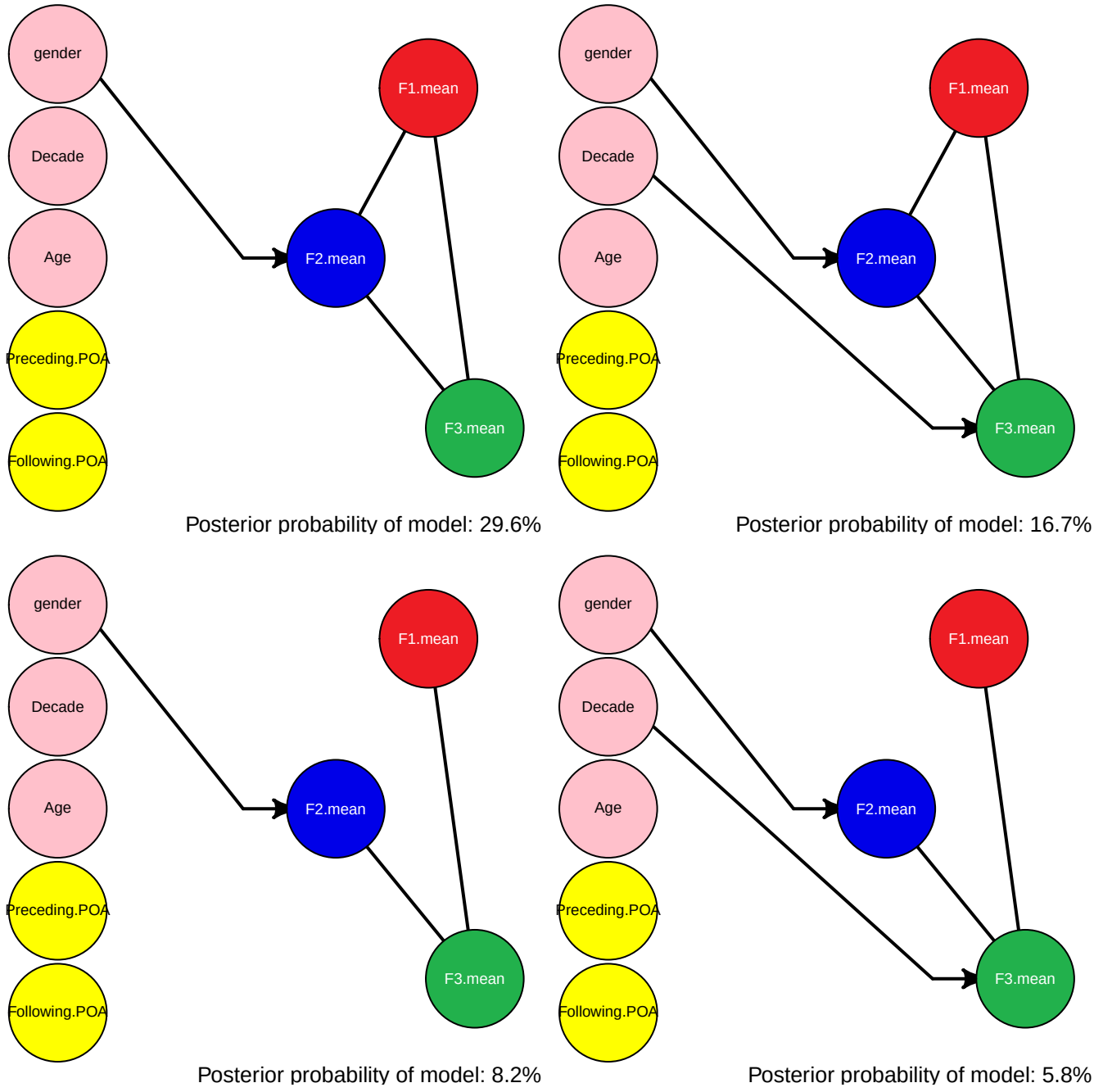

Figure 7.1: *BATH* vowel for raw mean formants.

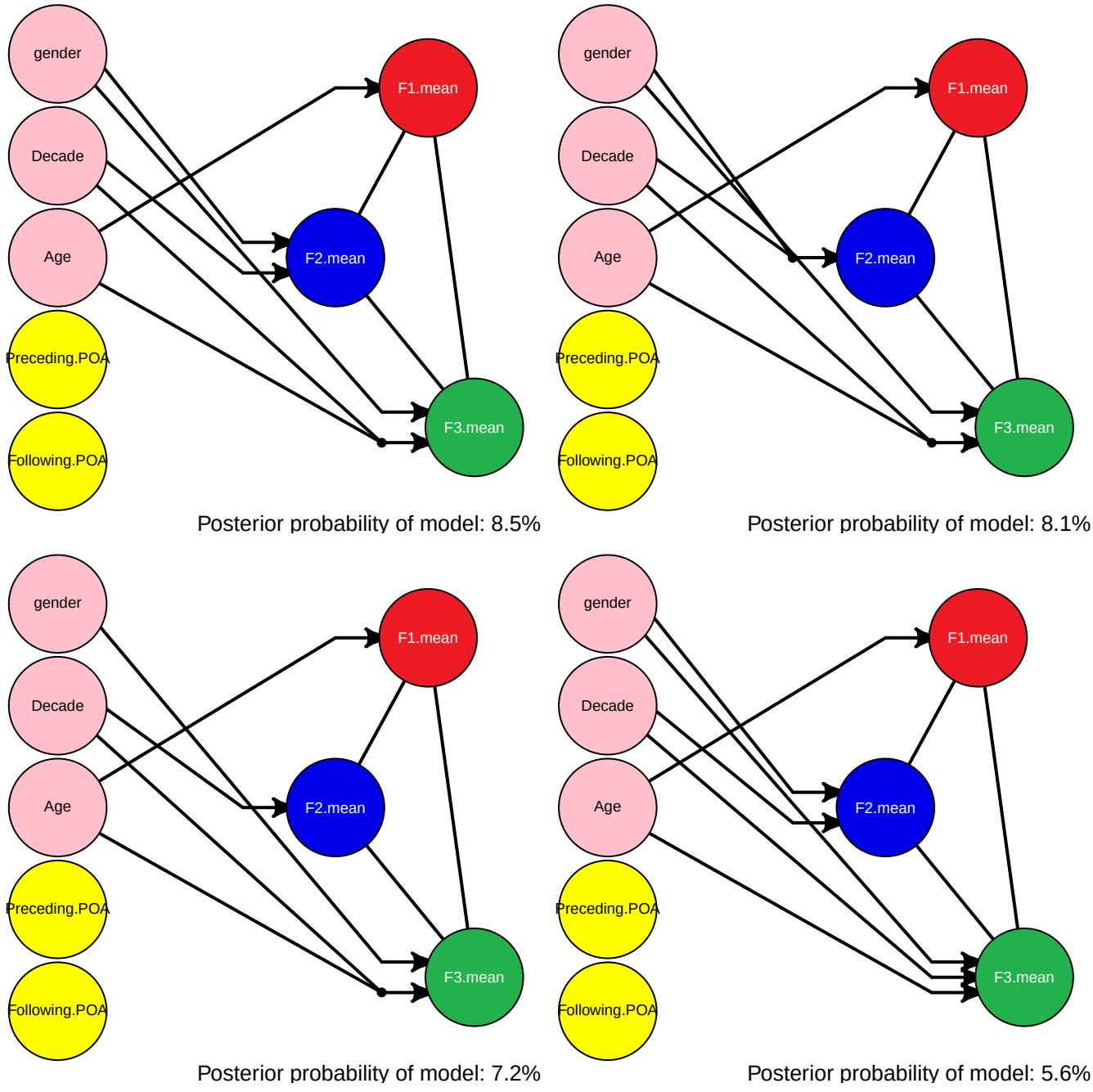

Figure 7.2: *FACE* vowel for raw mean formants.

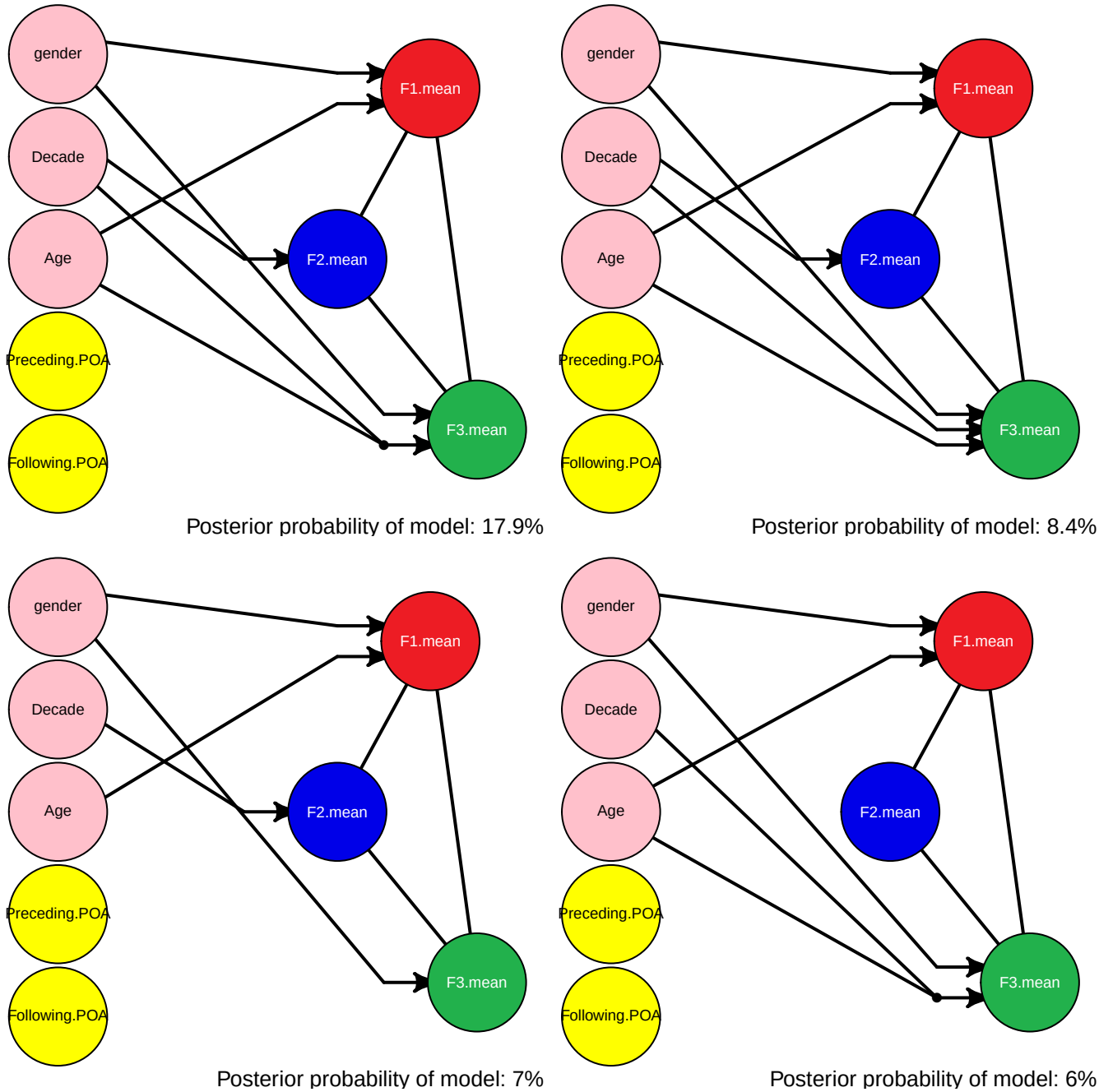

Figure 7.3: *FLEECE* vowel for raw mean formants.

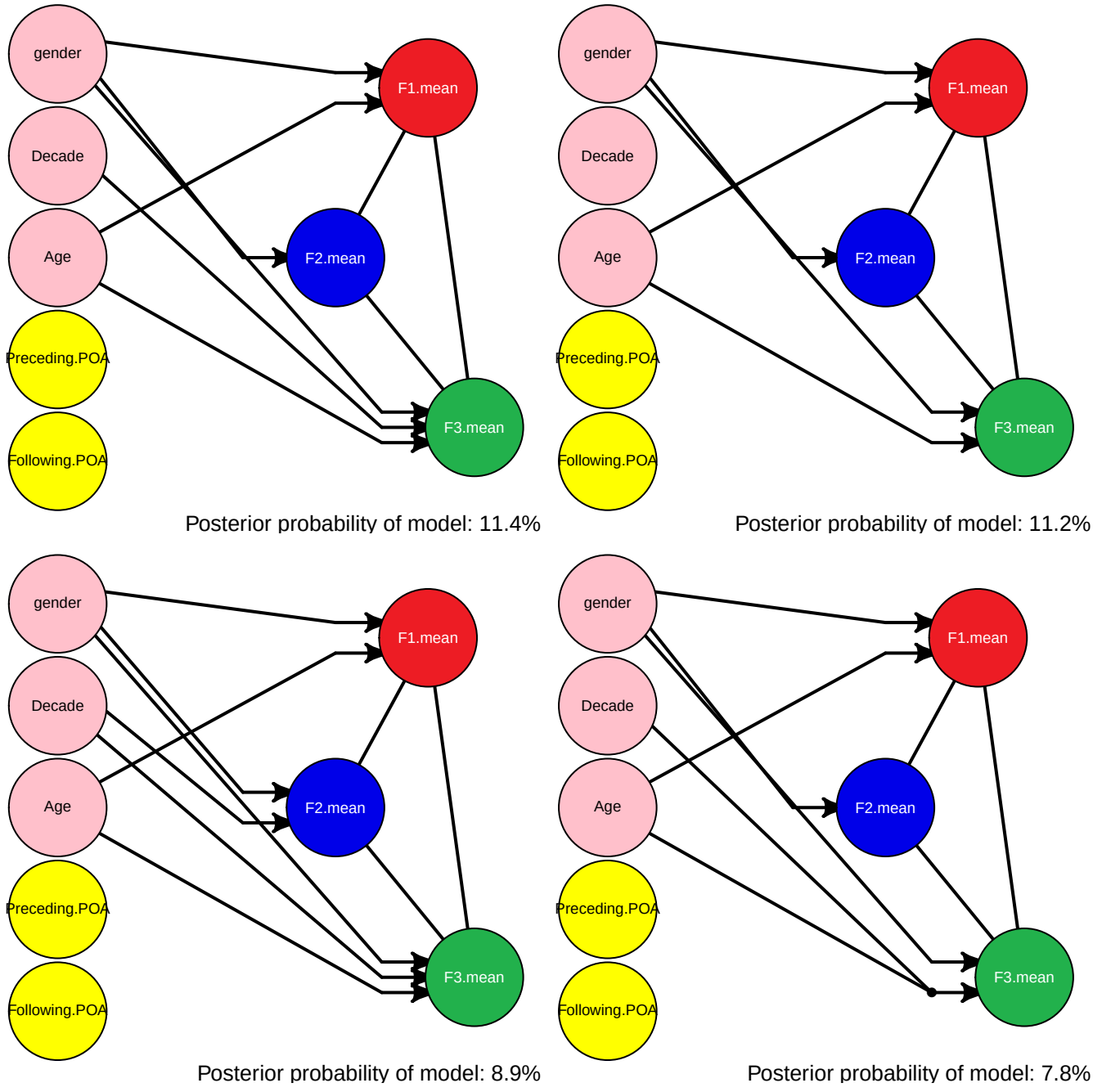

Figure 7.4: *FOOT* vowel for raw mean formants.

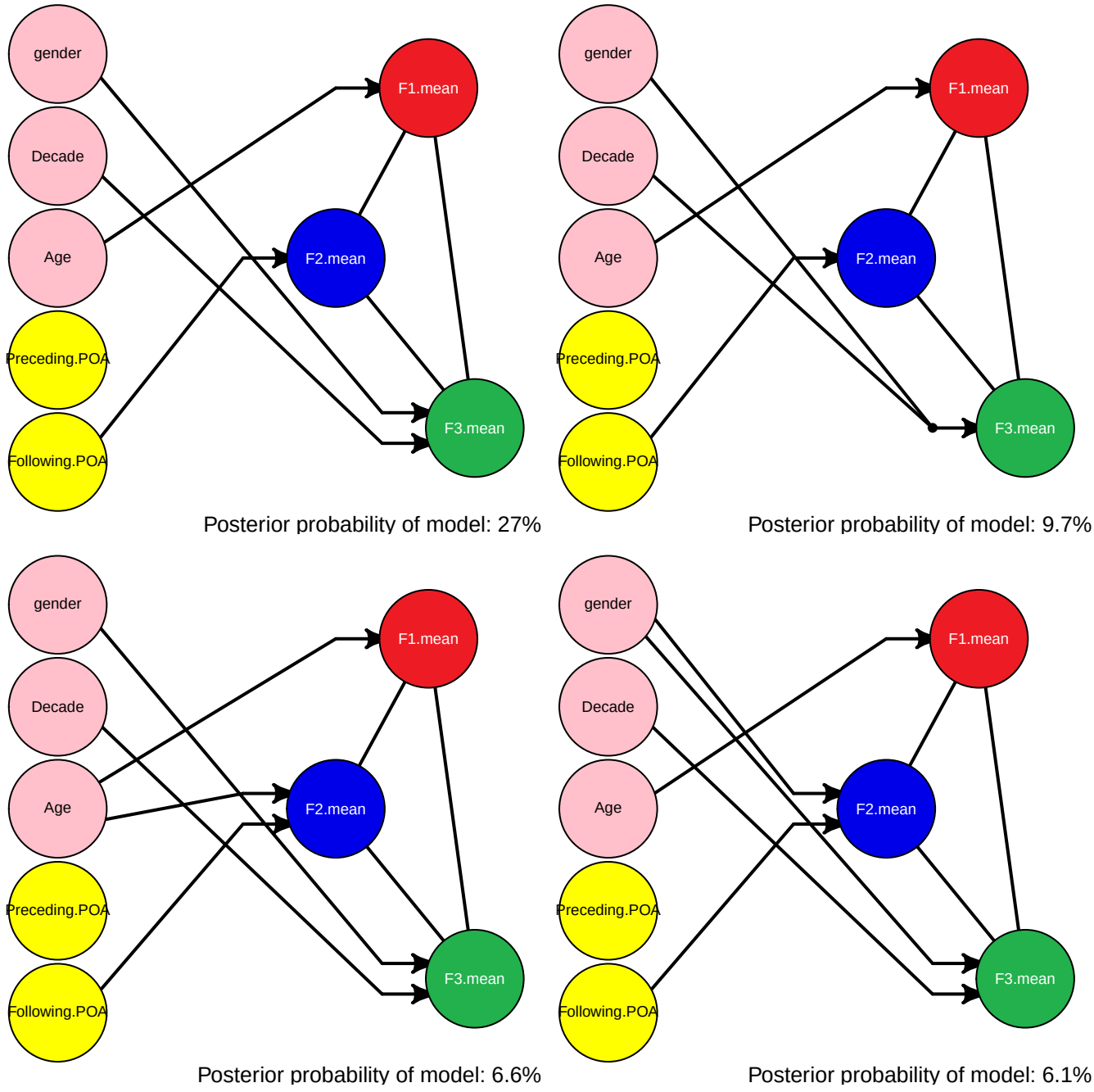

Figure 7.5: *GOAT* vowel for raw mean formants.

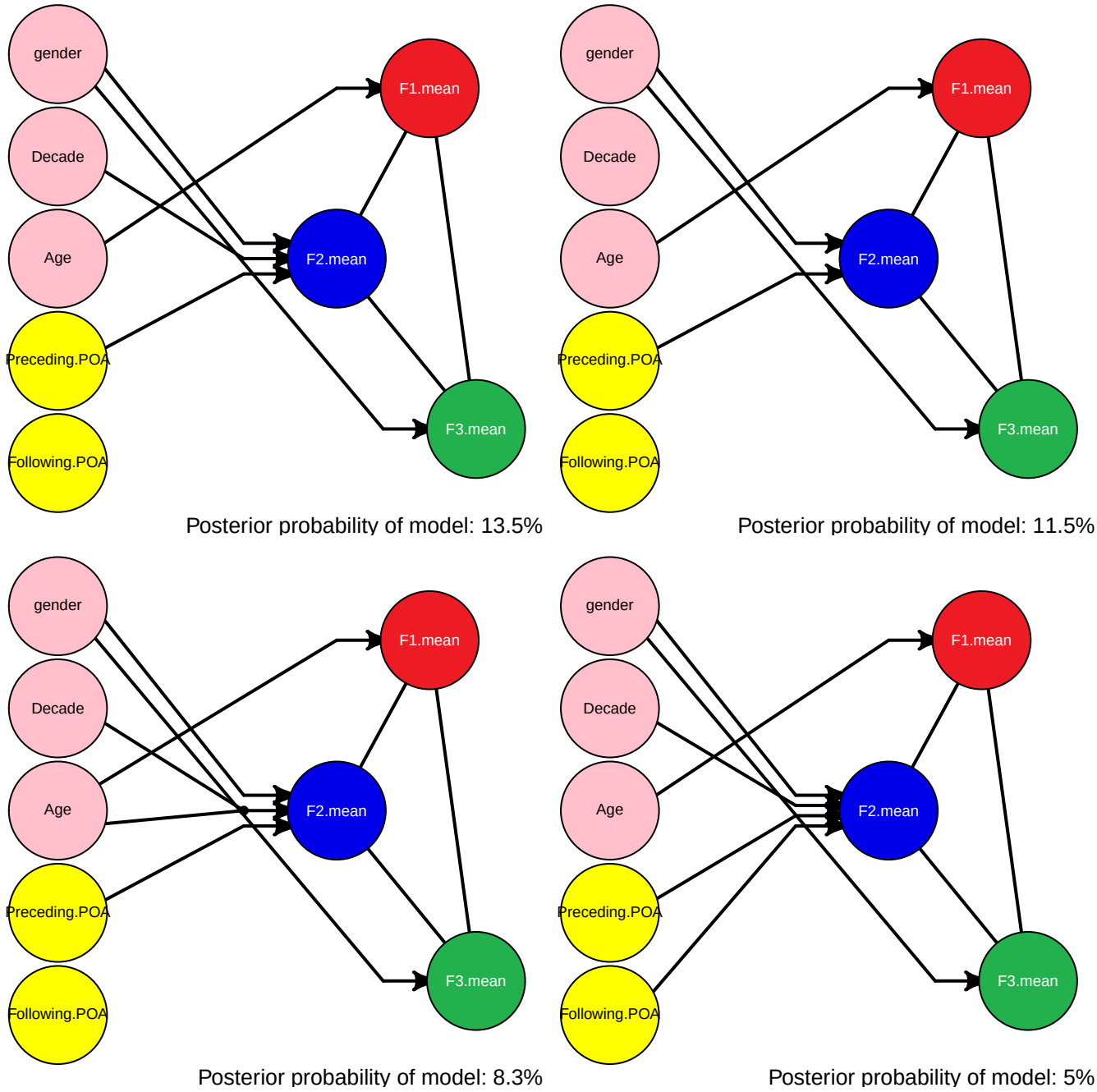

Figure 7.6: *GOOSE* vowel for raw mean formants.

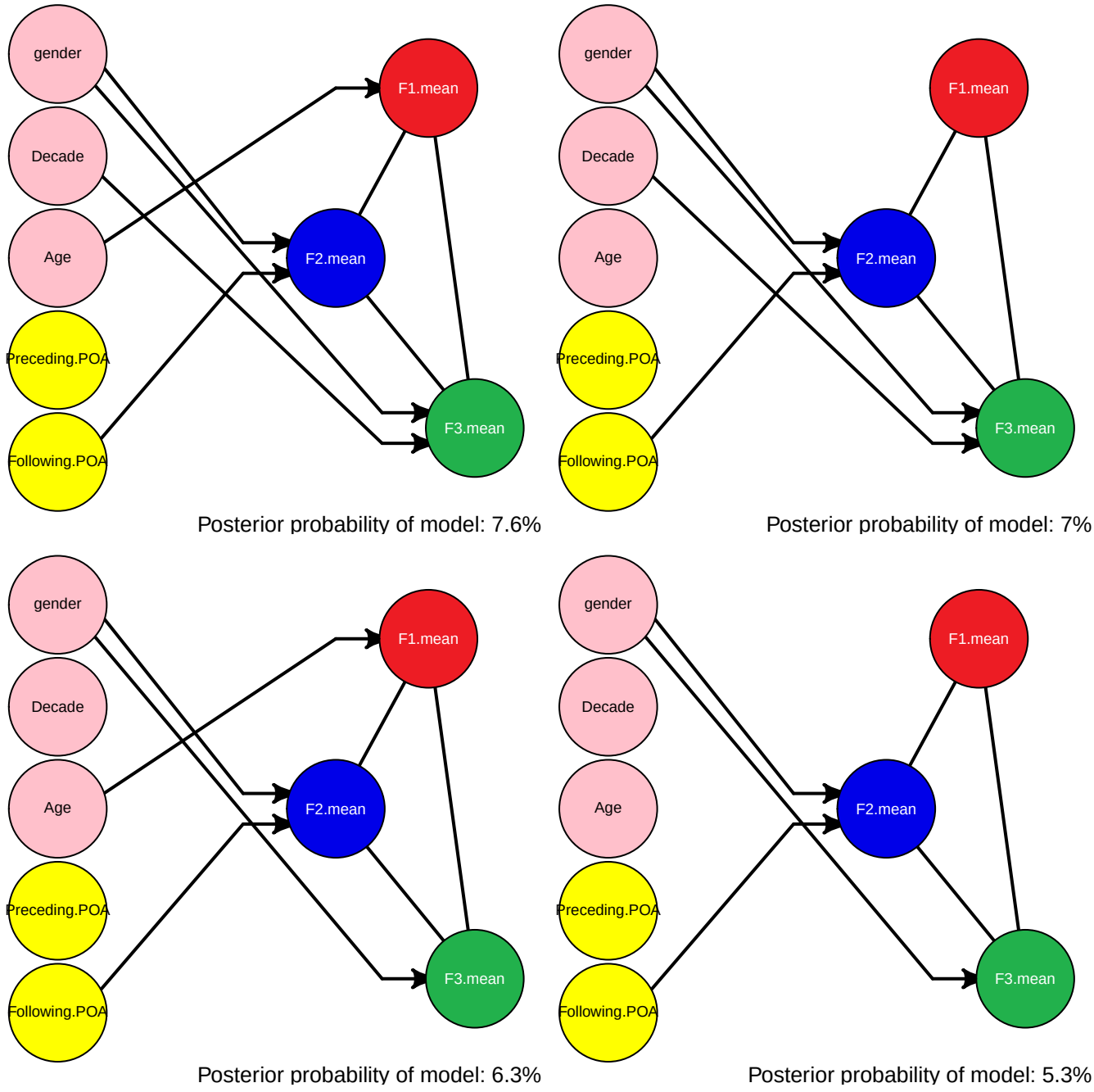

Figure 7.7: *LOT* vowel for raw mean formants.

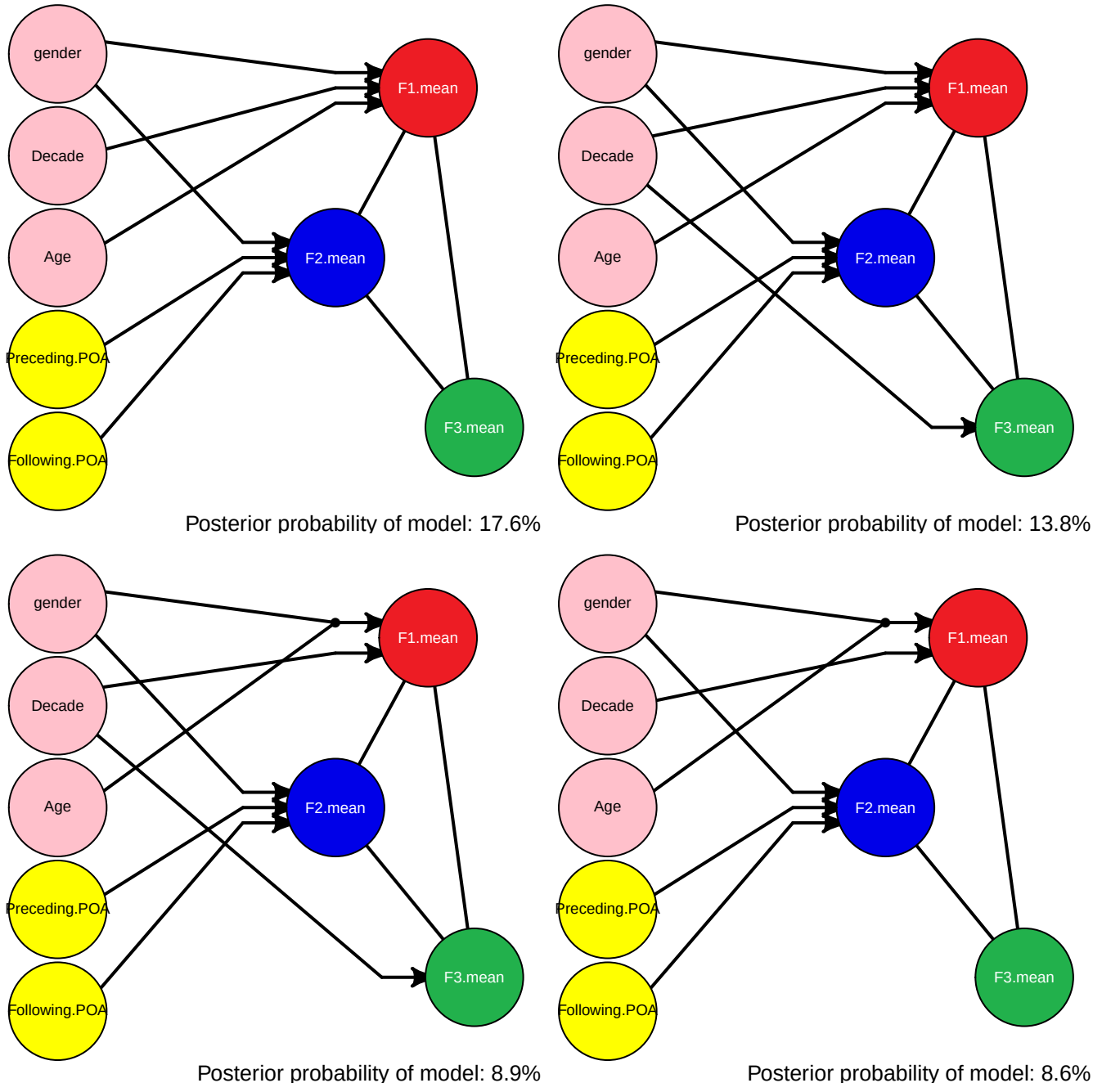

Figure 7.8: *TRAP* vowel for raw mean formants.

7.3 Lobanov Normalised Formant Results

Here, we provide the graphical models obtained for vowels based on their Lobanov normalised formant values for F1 and F2. The models obtained were ran for 5,000 iterations of the sampler and included all three-way interactions for each of the five predictor variables. The prior specifications for all hyperparameters are set similarly to the specification in Section 3.2.2, though we specify different values for a_l and b_l, namely $a_l = b_l = 1 \times 10^{-2}$. The best four models by posterior probability are shown for each vowel.

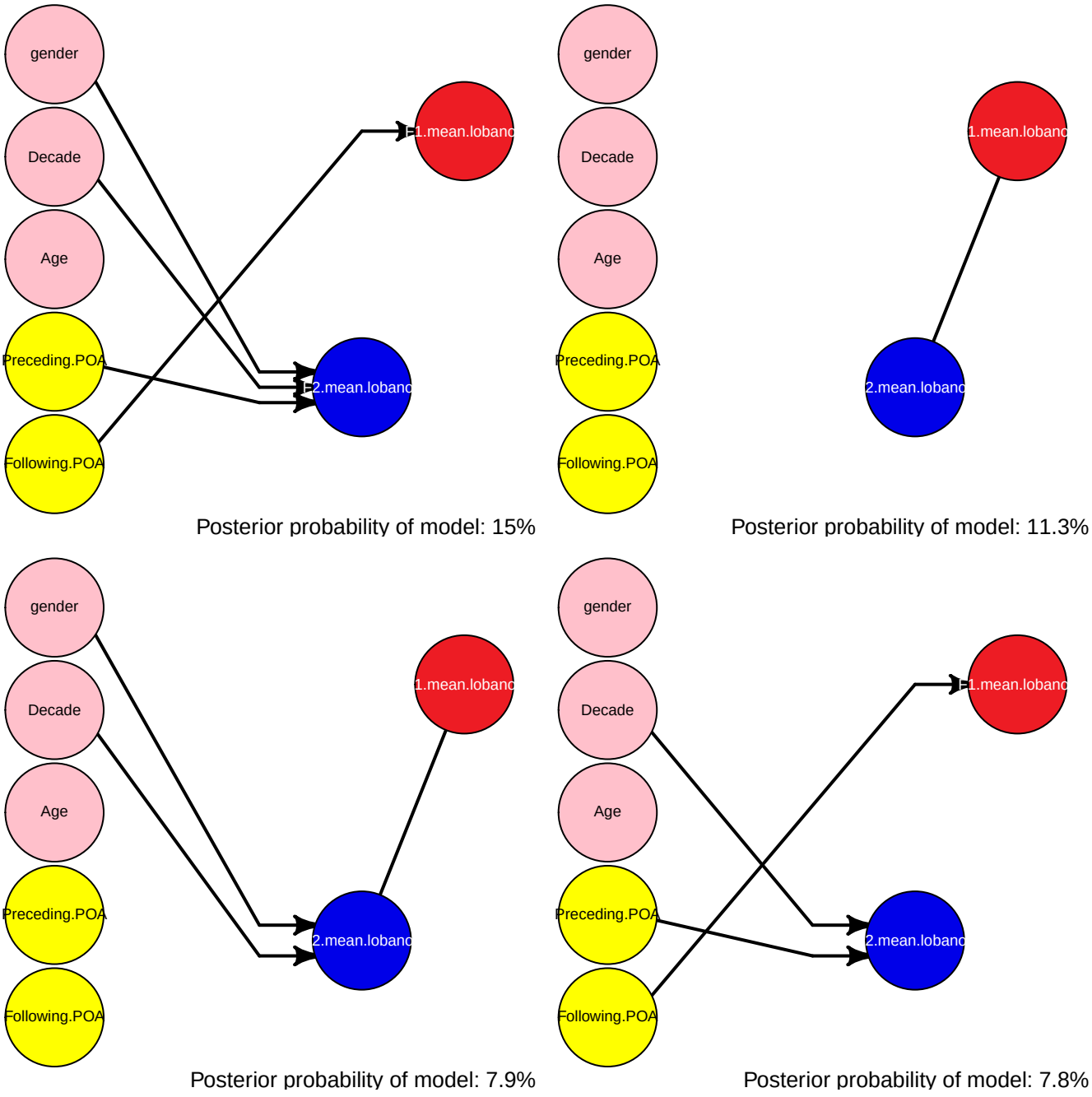

Figure 7.9: *BATH* vowel for Lobanov normalised formants.

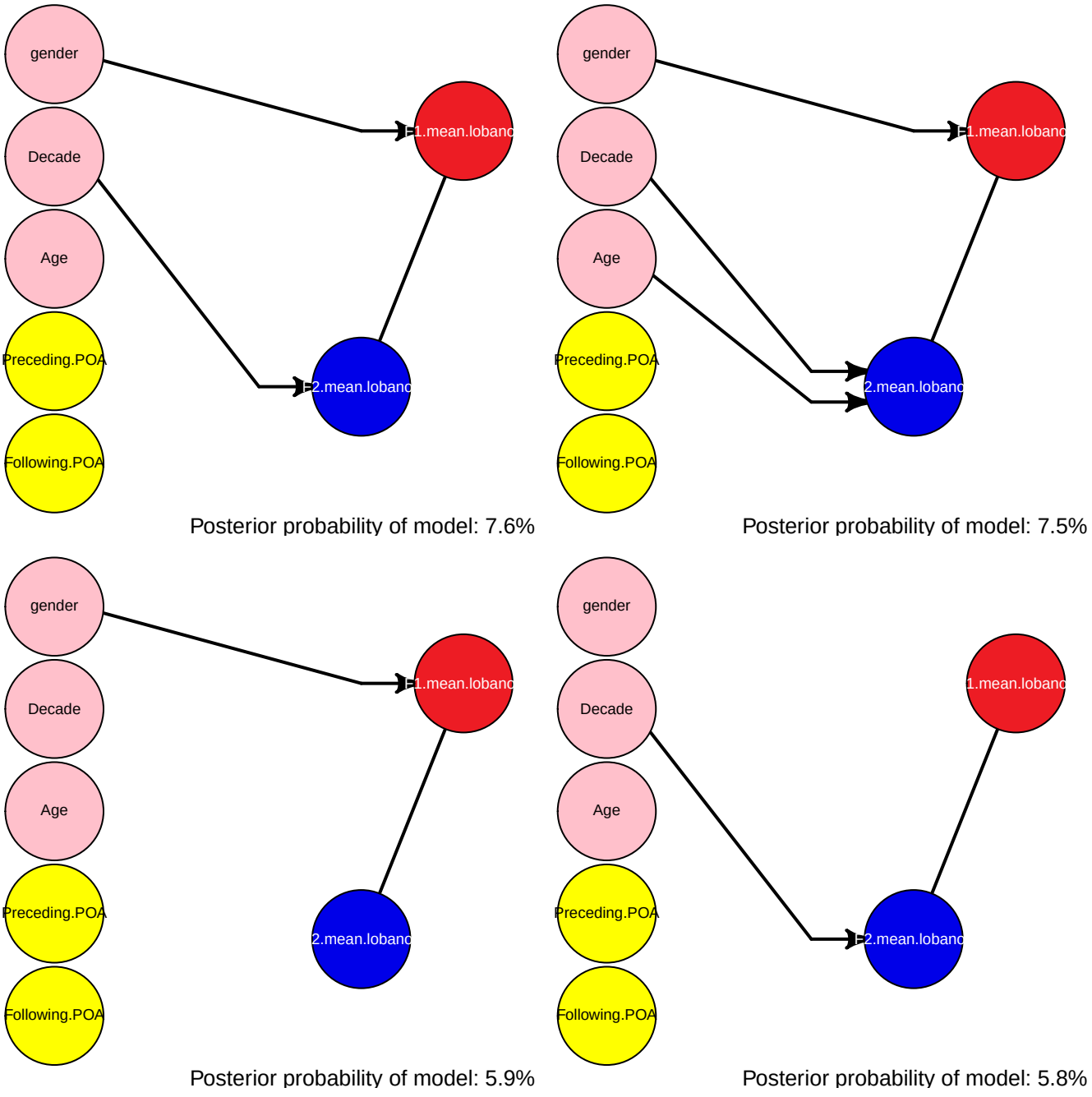

Figure 7.10: *FACE* vowel for Lobanov normalised formants.

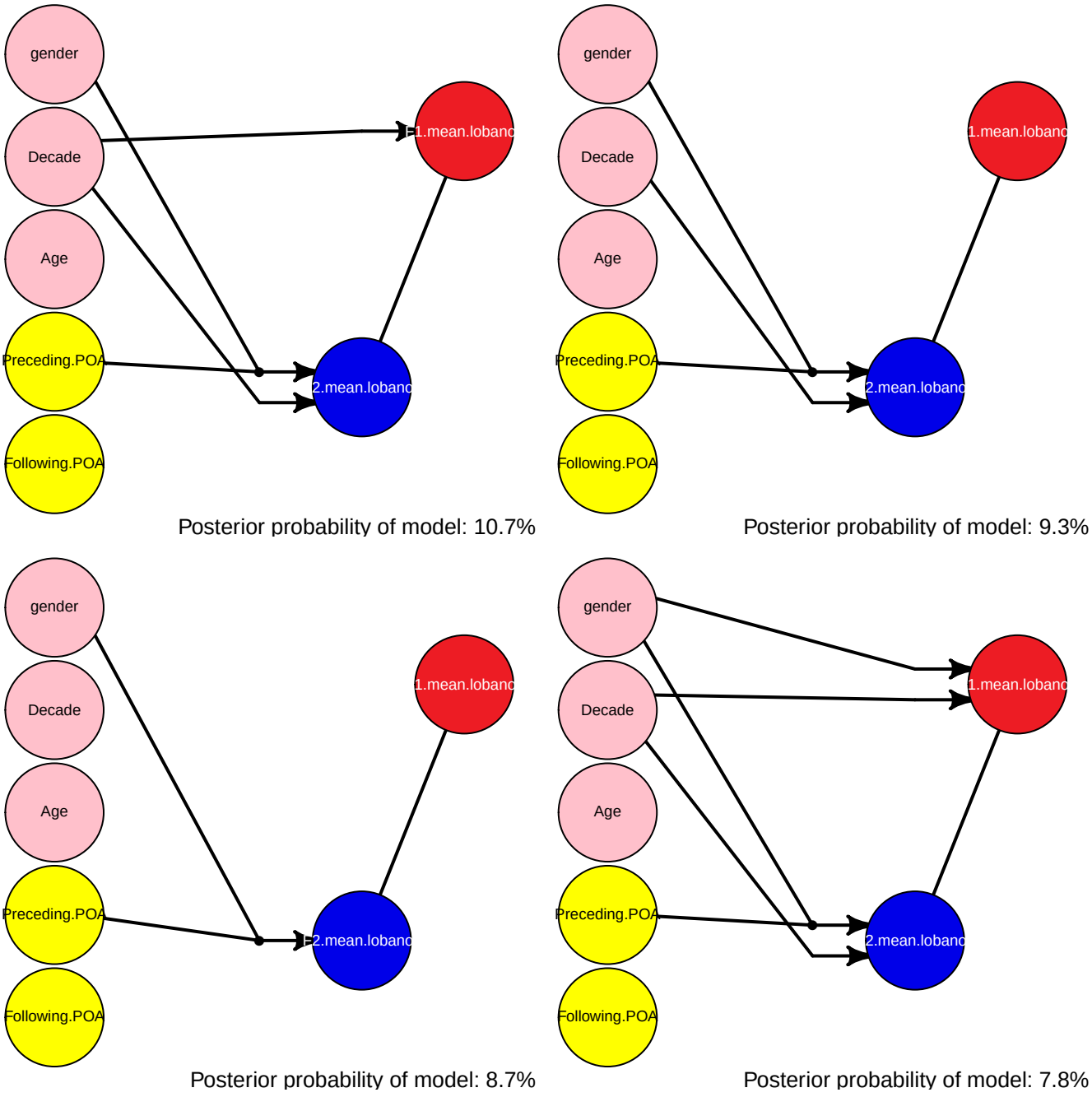

Figure 7.11: *FLEECE* vowel for Lobanov normalised formants.

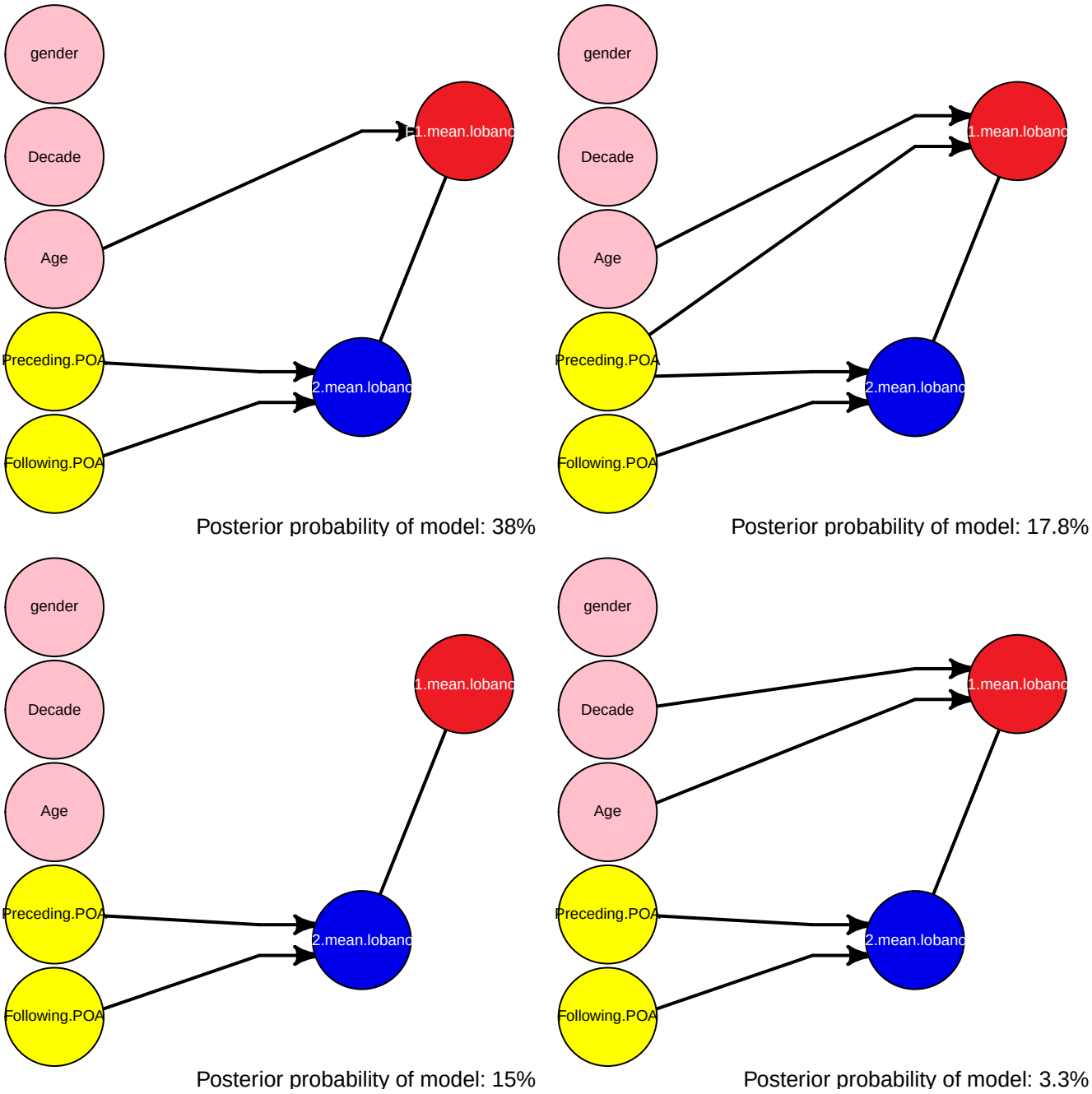

Figure 7.12: *FOOT* vowel for Lobanov normalised formants.

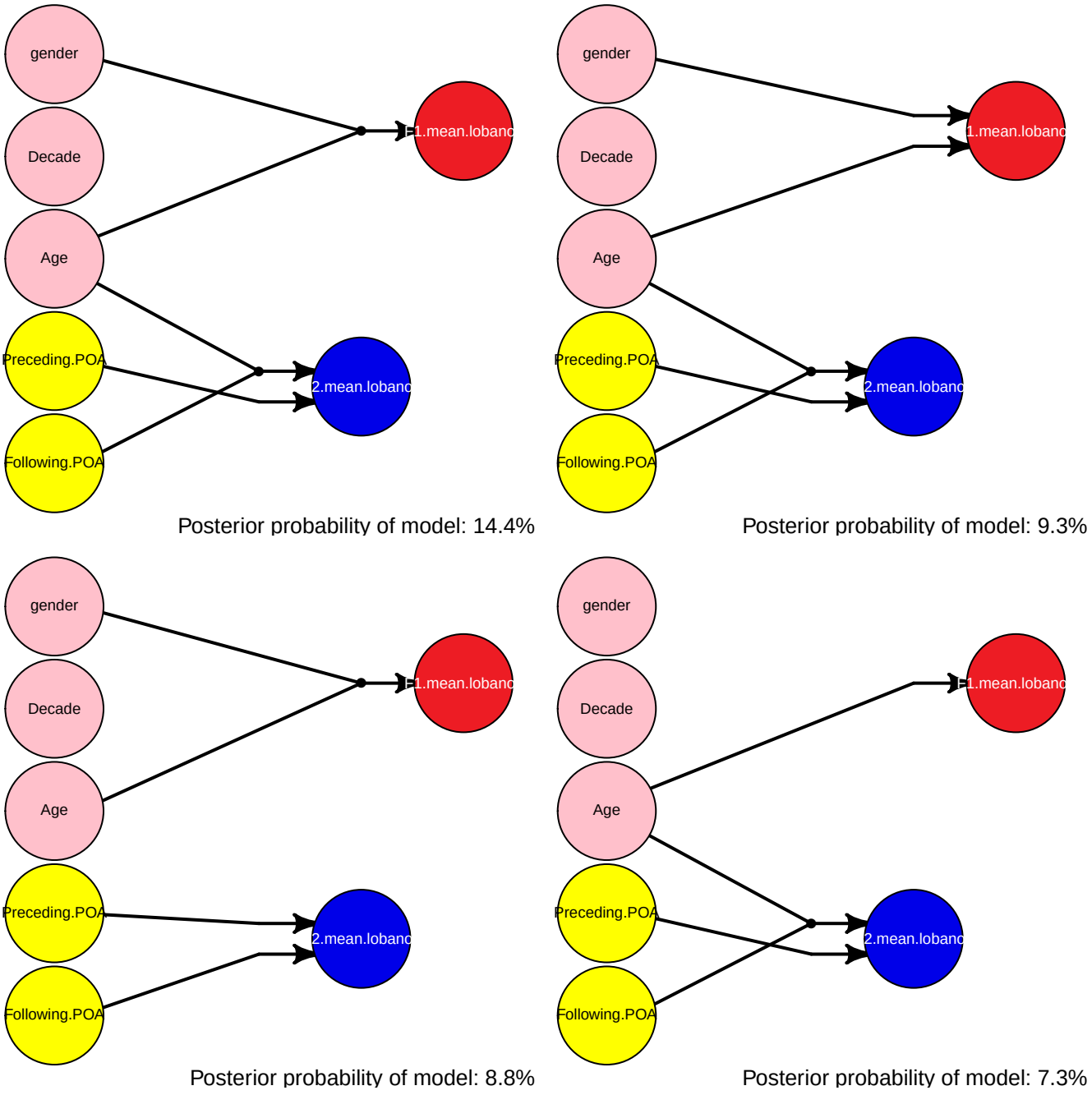

Figure 7.13: *GOAT* vowel for Lobanov normalised formants.

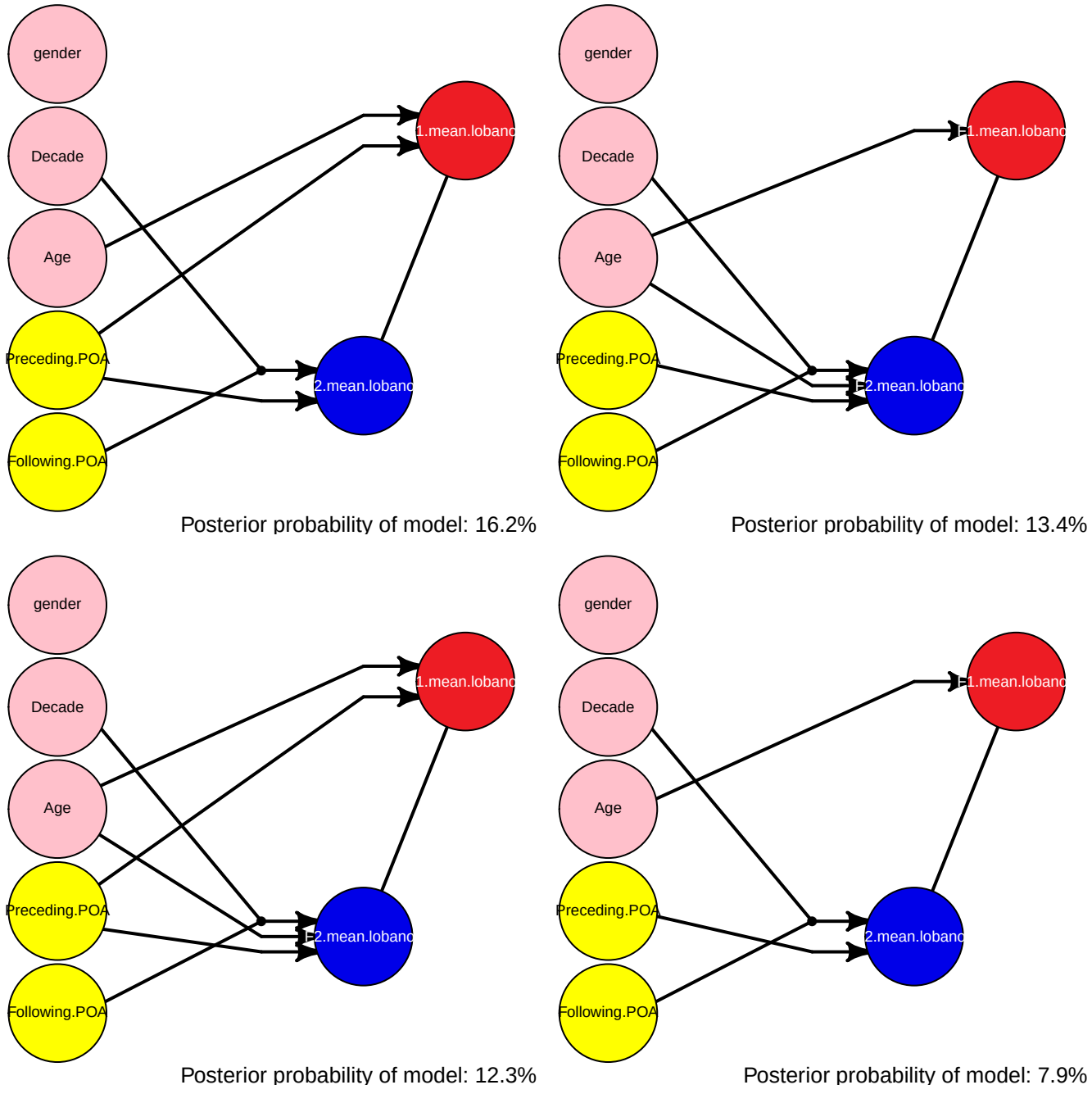

Figure 7.14: *GOOSE* vowel for Lobanov normalised formants.

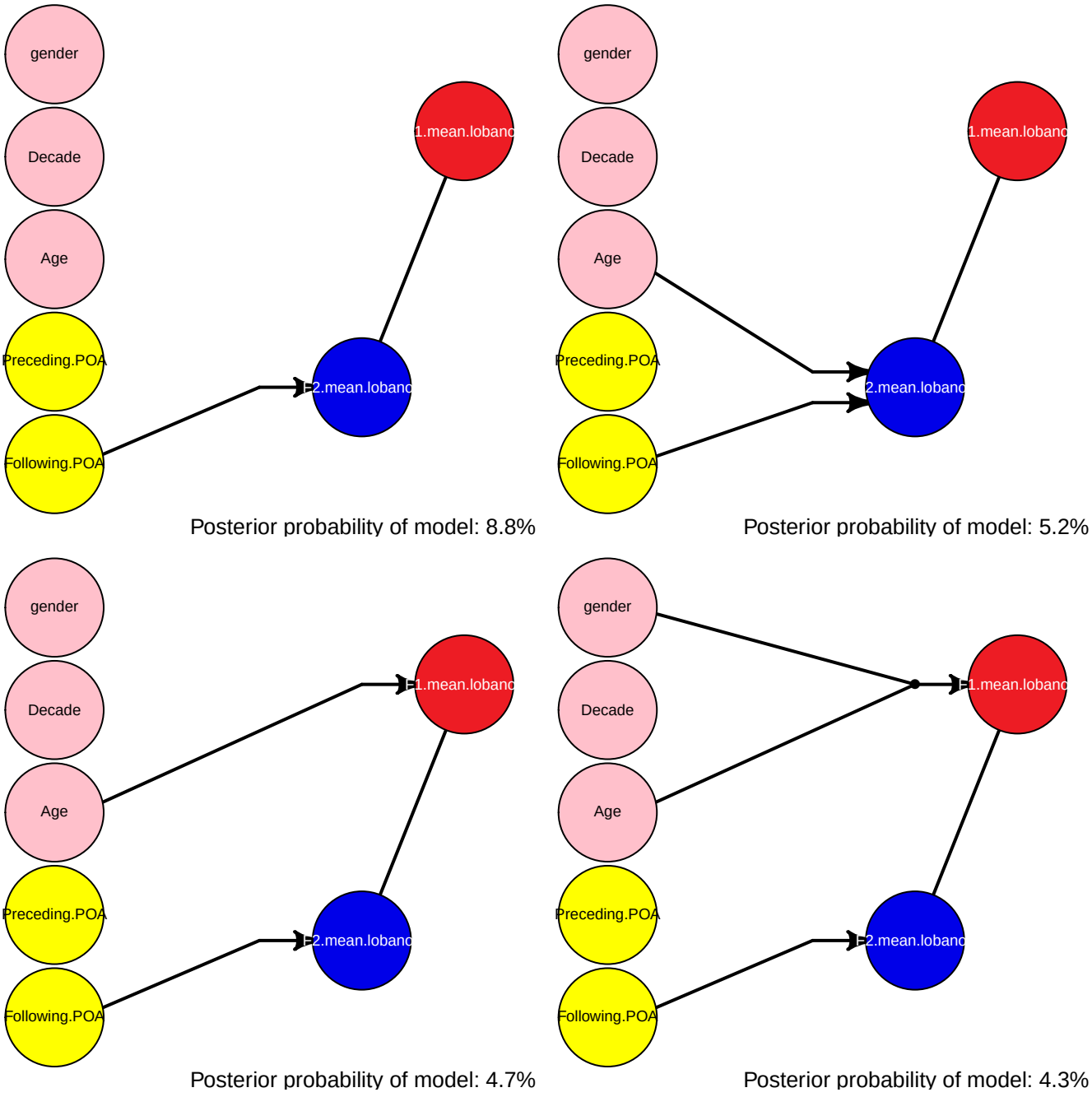

Figure 7.15: *LOT* vowel for Lobanov normalised formants.

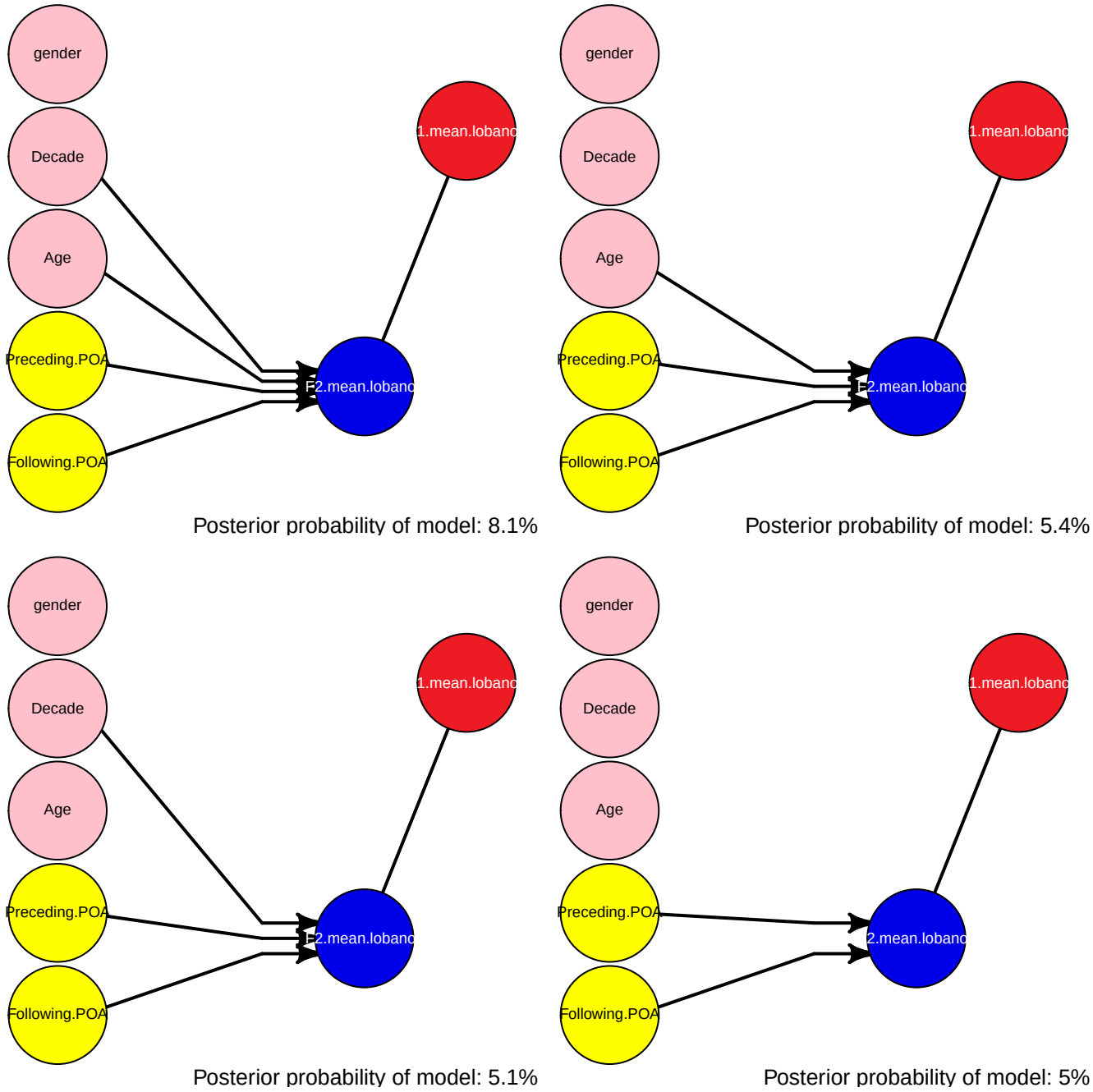

Figure 7.16: *TRAP* vowel for Lobanov normalised formants.

7.4 Discussion

In this chapter, we have provided a brief discussion of the results obtained fitting the Bayesian hierarchical model to the Sounds of the City corpus, and how these results differ from the findings in Stuart-Smith et al. (2017). We have identified several key points of interest, most notably the presence of significant effects in F3 for raw mean formant data, which have not been considered in this corpus study and could lead to potential future studies of interest.

Chapter 8

Conclusions and Further Work

The aim of this thesis has been to extend upon quantitative methods used within variationist sociolinguistic experiments to allow for more complete modelling by permitting the use of several linguistic variables on one observation. Based on this objective, we have created a Bayesian hierarchical model in Section 3.1.3 which allows for multiple response variables, which are here applied to multiple formant measurements from a particular vowel. Problems caused by nested designs, a common feature in linguistic corpora, in terms of poor MCMC mixing are alleviated with a modified sampler, shown in Figure 4.13, which includes additional steps for hierarchical centering and parameter expansion. The output from the Bayesian hierarchical model is then presented using a chain graph model like structure, constructed using a novel inference method combining hierarchical model output and Bayesian Gaussian graphical models. By implementing this method, we improve the readability of the hierarchical model output, which in turn increases the attractiveness of this new method to be implemented by sociolinguists. The following sections summarise the work that has taken place in this thesis and additional proposals for further work in this area.

8.1 Methodological Advances

The methodological advances from this work can be broadly split into two parts; the Bayesian hierarchical model and its mixing improvements as discussed in Chapters 3 and 4. The chain graph model structure visualisation and Bayesian Gaussian graphical model selection for multiple precision matrices are discussed in Chapters 5 & 6.

8.1.1 Bayesian hierarchical model with mixing improvements

In Section 3.1.3, we introduce the general structure of the Bayesian hierarchical model, which allows for multiple response variables, with a visual representation of the hierarchical model in the form of a probabilistic graphical model which can be found in in Figure 3.1. The Bayesian hierarchical model is similar in structure to the classical mixed effects models used in sociolinguistic studies (Johnson, 2009), but has two extensions: firstly, the expansion to allow for multiple response variables and secondly, we have now expressed the model in a Bayesian paradigm. Both extensions are extremely beneficial, with the multiple response modelling now allowing for several formants to be modelled simultaneously, taking into consideration the correlation present between the formant measurements and thus obtaining a more accurate representation of which underlying factors are contributing to vowel change. The move into a Bayesian framework is beneficial mainly for the chain graph model extension, allowing us to implement the G-Wishart prior for the precision estimates to obtain the Bayesian Gaussian graphical model, modelling the relationship between response variables.

In the remainder of Chapter 3, we applied the hierarchical model to the Sounds of the City corpus where we observed several issues with the mixing of parameters in the MCMC output. Two mixing issues were observed: firstly, due to the nested design of the corpus, fixed effects were nested within random effects, namely the social factors gender, age and decade of recording which are nested within the speaker and also the linguistic factors of following and preceding place of articulation of consonants within the word choice. This leads to extremely poor mixing between the fixed effect and random effect coefficients. Another issue was identified in the mixing of the precision estimates for word choice, where the sampler was often found to be getting stuck around smaller values. This also impacted the relevant random effects coefficients, where the trace variance for the coefficients would 'shrink' when the precision sampler was trapped at small values.

Chapter 4 focussed on addressing these mixing issues through the implementation of reparameterisation methods within the MCMC. Hierarchical centering (Gelfand et al., 1995) was used to address the poor mixing observed through the nested coefficients. This step involves forming $\tilde{\boldsymbol{\delta}}_k = \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} + \tilde{\mathbf{U}}_k\tilde{\mathbf{b}}_k$, where $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ is the set of coefficients which are

nested within the random effect $\tilde{\mathbf{b}}_k$. From this, we can sample $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ conditional on $\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}$. Performing this step reduces the correlation present within the sampler, and allows us to explore the parameter space more freely and converge to the target distribution in fewer samples. The improvements on the corpus data by implementation of this step are shown in Section 4.1.3. Secondly, we implement a step influenced by parameter expansion to address the poor mixing observed in the precision estimate for the word choice effect, where the sampler was often stuck at zero. We implemented an adaptation of the method proposed in Gelman et al. (2008) which works well in a practical MCMC setting. We sample $\alpha_m \sim \mathcal{G}(a_{\alpha m}, b_{\alpha m})$ for each random effect m with poor precision mixing. We then define $\tilde{\mathbf{b}}^*_{\mathbf{g}} = \alpha\tilde{\mathbf{b}}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^*_{\tilde{\mathbf{b}}_{\mathbf{g}}} = \alpha\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}$. A Metropolis-Hastings step is then performed whether to accept or reject the modified parameters over the current model parameters. Section 4.2.2 shows this applied to the Sounds of the City corpus, where we again observe a slight improvement in the mixing of the precision and random effects coefficient mixing.

By implementing the mixing improvements, the Bayesian hierarchical model is now able to perform MCMC more efficiently, which leads to a significant reduction in computational time, which is a key point in order to promote this approach as a usable tool for sociolinguistic analysis.

8.1.2 Chain graph model visualisation

The second development we have implemented was introduced in Chapters 5 & 6, where we introduce a new inference tool that combines a Bayesian hierarchical model and visualises the output as a chain graph model like structure, jointly inferring a Bayesian Gaussian graphical model to model the relationship between the response variables. The main point of development within this chapter was Bayesian Gaussian graphical model selection for the response variable graph in Chapter 5. Many algorithms exist for Gaussian graphical model selection, and are discussed in depth in Section 5.3. In order to jointly infer the relationship using output from the hierarchical model, we must use the precision estimates from the model, which provide the information to best model the undirected graph. The initial drawback we had for this was that all selection algorithms consider only one precision input at a time. Due to this, we formed the modified PAS

algorithm (5.3), based on the PAS algorithm of Wang and Li (2012). Using this algorithm and also implementing factorisation properties for chordal graphs, we have constructed a Gaussian graphical model selection algorithm that can be implemented for multiple precision inputs.

The final step involved updating the Bayesian hierarchical model to allow for Gaussian graphical model selection. The main change here comes from the update in prior distributions for the precisions, changing from the Wishart distribution to the G-Wishart distribution. The updated posteriors are shown in Section 6.2.1 and also visualised in Figure 6.6. From this update, we have obtained graphs for the Sounds of the City corpus which are shown in Chapter 7, where the best four graphs by posterior probability are shown for each vowel for both the raw mean formant measurements and the Lobanov normalised measurements.

8.2 Sociolinguistic Advances

In terms of direct advances for variationist sociolinguistic studies, we have introduced a new statistical tool that can be easily applied to a variety of corpora. Previous studies on the Sounds of the City corpus (Stuart-Smith et al., 2017) considered models only capturing one formant measurement per time on vowel sounds. As we have observed in Section 6.3.2, without modelling all formant measures together, it is possible we can observe relationships between fixed effects and formants that could be weaker than if we considered all formants in the same model. The inclusion of inter-model selection within the hierarchical model also removes the need to fit multiple models to the corpus, which increases time in terms of model fitting and additional interpretation.

The chain graph visualisation provides a clear picture of the underlying model, which helps provide an instantaneous image of what factors are influencing vowel variation and change. The chain graph output also acts as an incentive for sociolinguists to implement the models we have discussed, as the visualisation helps to give a clear initial impression of the model output. Without the addition of the graph, the increased complexity in modelling could possibly have discouraged users from implementing the model.

8.3 Further Work

The Bayesian hierarchical model developed provides a clear improvement on the single response mixed effects models used in many sociolinguistic corpus studies, with the chain graph model visualisation providing a useful tool for an initial impression of the model output. However, we can improve on the hierarchical model specification and the graphical model visualisation even further.

One of the key points to highlight from the development of the hierarchical model and resulting chain graph style visualisation is how this model can be applied to any mixed effects problem, not just to sociolinguistic data. In order to develop further on this point, a natural step would be to produce a *R* package from the constructed model, using the model code generated (which can be found here `https://github.com/calex1991/BayesCGModels`). Another extension could be to provide an interactive web based application using Shiny, which could take the hierarchical model output and produce an interactive graph with extensions beyond the current graphs, namely inclusion of summary statistics of model parameters and a clearer visualisation of the inclusion of a term within the model.

Another extension which could be considered is further development on variable selection within the hierarchical model. This could be expanded twofold, firstly, with variable selection considered for random effects. This is perhaps not an issue for linguistic corpora, where speaker and word choice almost always have a significant effect within models, but for different data problems, this could be an issue. Different variable selection methods could also be considered in general, compared to the current implementation in Section 3.1.5.

The Bayesian Gaussian graphical model selection can also be studied in further detail. With a wealth of algorithms being developed in the past few years, improvements are being made in the field constantly. The work of Wit and Mohammadi (2015) uses Birth-Death MCMC (BDMCMC) to select the best fitting graph to observed data. Results produced using this method suggest that it outperforms many other model search algorithms, including the PAS algorithm which we have adapted from for the modified PAS algorithm implemented currently. Further research into these methods and feasibility

of expansion to the multiple precision case could lead to improvement in computational times and model fitting.

Appendix A

Posterior Distributions

In this appendix, we shall derive the conditional distributions shown in Sections 3.1.6 4.1.2 and 6.2.1 which are used to sample from the Bayesian hierarchical model, the hierarchical model with added hierarchical centering and parameter expansion based mixing improvements and the hierarchical model embedded in the chain graph structure respectively.

A.1 Derivation of Posteriors

Here, we show how the conditional distributions for each of the samplers are obtained by using standard results for Gaussian distributions as discussed in Bishop (2006).

If we have a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ which is in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda}^{-1}) \tag{A.1}$$

and

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{A.2}$$

then the marginal distribution of $\mathbf{y}$, and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$, are

given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\top) \tag{A.3}$$

and

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{\Sigma}\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\mu\}, \mathbf{\Sigma}) \tag{A.4}$$

where

$$\mathbf{\Sigma} = (\mathbf{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A}) \tag{A.5}$$

A.2 Standard Bayesian hierarchical model

The conditional distributions derived here are laid out in a similar fashion to Section 3.1.6 which details the sampler for the standard Bayesian hierarchical model which is presented in a general form for multiple random effects and multiple response variables. Using the standard results for conditional Gaussian distributions discussed in Section A.1 and from Figure 3.1, we compute the conditional distributions for $\tilde{\boldsymbol{\beta}}_{\tilde{\eta}}$ and $\tilde{\mathbf{b}}$,, where $\boldsymbol{\theta}$ is defined as a vector of all the model parameters and hyperparameters:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} \,|\, \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\eta}}} &\propto \mathcal{N}\left(\boldsymbol{y} \,|\, (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\eta}}\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\mathbf{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} \,|\, \mathbf{0}, \mathbf{V}^{-1}\right) \\
\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} \,|\, \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\eta}}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} \,|\, \left[\tilde{\mathbf{X}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{X}} + \mathbf{V}\right]^{-1}\tilde{\mathbf{X}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{y}}_{\tilde{\boldsymbol{\beta}}}, \left[\tilde{\mathbf{X}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{X}} + \mathbf{V}\right]^{-1}\right)
\end{aligned} \tag{A.6}$$

where we define $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = (\mathbf{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I})$, $\tilde{\mathbf{X}}_{\tilde{\eta}} = \text{blockdiag}(\mathbf{X}^1_{\boldsymbol{\eta}^1}, ..., \mathbf{X}^L_{\boldsymbol{\eta}^L})$, and $\tilde{\mathbf{y}}_{\tilde{\boldsymbol{\beta}}} = \mathbf{y} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$.

$$\begin{aligned}
\tilde{\mathbf{b}} \,|\, \boldsymbol{\theta}_{\setminus\tilde{\mathbf{b}}} &\propto \mathcal{N}\left(\boldsymbol{y} \,|\, (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\eta}}\tilde{\boldsymbol{\beta}}_{\tilde{\eta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\mathbf{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{N}\left(\tilde{\mathbf{b}} \,|\, \mathbf{0}, \mathbf{\Sigma}_{\tilde{\mathbf{b}}}\right) \\
\tilde{\mathbf{b}} \,|\, \boldsymbol{\theta}_{\setminus\tilde{\mathbf{b}}} &\propto \mathcal{N}\left(\tilde{\mathbf{b}} \,|\, \left[\tilde{\mathbf{U}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{U}} + \mathbf{\Sigma}_{\tilde{\mathbf{b}}}\right]^{-1}\tilde{\mathbf{U}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{y}}_{\tilde{\mathbf{b}}}, \left[\tilde{\mathbf{U}}^\top\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\tilde{\mathbf{U}} + \mathbf{\Sigma}_{\tilde{\mathbf{b}}}\right]^{-1}\right)
\end{aligned} \tag{A.7}$$

where we define $\tilde{\mathbf{U}} = \text{blockdiag}\left(\mathbf{U}^1, \ldots, \mathbf{U}^L\right)$, $\mathbf{\Sigma}_{\tilde{\mathbf{b}}} = \text{blockdiag}\left(\mathbf{\Omega}_{\tilde{\mathbf{b}}_1}, \ldots, \mathbf{\Omega}_{\tilde{\mathbf{b}}_G}\right)$, and $\tilde{\mathbf{y}}_{\tilde{\mathbf{b}}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$.

We can then obtain the conditional distributions for the precision parameters:

$$\begin{aligned}
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \nu_{\mathbf{b}_g}, \mathbf{S}_{\mathbf{b}_g}\right) \\
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}_{\tilde{\mathbf{b}}_g}^{-1} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}_{g_i}^{\top}\right]^{-1}\right)
\end{aligned} \tag{A.8}$$

where we sample each $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g.

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \nu_{\boldsymbol{\epsilon}}, \mathbf{S}_{\boldsymbol{\epsilon}}\right) \\
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}_{\boldsymbol{\epsilon}}^{-1} + \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^{\top}\right]^{-1}\right)
\end{aligned} \tag{A.9}$$

where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$.

We then sample the prior variance parameter $\boldsymbol{\tau}$ for the fixed effects coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$

$$\begin{aligned}
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}_l} \mid \mathbf{0}, \tau_l \mathbf{I}\right) \mathcal{G}(\tau_l \mid a_l, b_l) \\
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}_{\tilde{\boldsymbol{\eta}}_l}^l\|}{2}, b_l + \frac{\sum_{m=1}^{p} \left(\beta_m^l\right)^2}{2}\right)
\end{aligned} \tag{A.10}$$

where we sample for each l separately and form $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_l)$.

A.3 Bayesian hierarchical model with efficient sampling of $\tilde{\beta}$ and $\tilde{\mathbf{b}}$

The conditional distributions derived here are similar to those in 3.1.6 which details the sampler for the standard Bayesian hierarchical model, but here we propose different samplers for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$. Instead of sampling all the parameters in one block, we instead sample for $\boldsymbol{\beta}_{\tilde{\boldsymbol{\eta}}}^l$ and $\tilde{\mathbf{b}}_{g,h}$, where $h = 1, \ldots, H$ is the level of the corresponding random effect g. Using the standard results for conditional Gaussian distributions discussed in Section

A.1 we compute the conditional distributions for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$, where $\boldsymbol{\theta}$ is defined as a vector of all the model parameters and hyperparameters:

$$\begin{aligned}
\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} \mid \mathbf{0}, \mathbf{V}^{-1}\right) \\
\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1} \mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j}\mathbf{X}^\top_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\right)
\end{aligned} \tag{A.11}$$

where we define $\mathbf{z}_{\boldsymbol{\beta}^l} = \omega_{j,j}\mathbf{y}^l + \sum_{k=1}^{k\neq l}\omega_{j,k}\left(\mathbf{y}^k - \mathbf{X}_{\boldsymbol{\eta}^k}\boldsymbol{\beta}^k\right)$. Model selection is now performed on each level of $\boldsymbol{\beta}^l$ in turn.

$$\begin{aligned}
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{N}\left(\tilde{\mathbf{b}} \mid \mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}\right) \\
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right)
\end{aligned} \tag{A.12}$$

where we define $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} = \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}_{\tilde{\mathbf{b}}_{-g}}\tilde{\mathbf{b}}_{\tilde{\mathbf{b}}_{-g}}$, where $\tilde{\mathbf{b}}_{-g}$ denotes $\tilde{\mathbf{b}}$ excluding group g and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}$ is the mean value calculated for $\mathbf{y}_{\tilde{\mathbf{b}}_{g,h}}$ for each response level l.

We can then obtain the conditional distributions for the precision parameters:

$$\begin{aligned}
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \nu_{\mathbf{b}_g}, \mathbf{S}_{\mathbf{b}_g}\right) \\
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}^{-1}_{\tilde{\mathbf{b}}_g} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}} \tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}^\top_{g_i}\right]^{-1}\right)
\end{aligned} \tag{A.13}$$

where we sample each $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g.

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \nu_{\boldsymbol{\epsilon}}, \mathbf{S}_{\boldsymbol{\epsilon}}\right) \\
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}^{-1}_{\boldsymbol{\epsilon}} + \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^\top\right]^{-1}\right)
\end{aligned} \tag{A.14}$$

where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$.

We then sample the prior variance parameter $\boldsymbol{\tau}$ for the fixed effects coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$

$$\begin{aligned}
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}_l} \mid \mathbf{0}, \tau_l \mathbf{I}\right) \mathcal{G}(\tau_l \mid a_l, b_l) \\
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}^l_{\tilde{\boldsymbol{\eta}}_l}\|}{2}, b_l + \frac{\sum_{m=1}^{p} \left(\beta^l_m\right)^2}{2}\right)
\end{aligned} \tag{A.15}$$

where we sample for each l separately and form $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_l)$.

A.4 Bayesian hierarchical model with hierarchical centering and parameter expansion

The conditional distributions derived here are laid out in a similar fashion to Section 4.1.2 which details the sampler for the standard Bayesian hierarchical model with modifications for the hierarchical centering step and the simplification of the parameter expansion step.. Using the standard results for conditional Gaussian distributions discussed in Section A.1 and from Figure 4.5, we compute the conditional distributions for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$, where $\boldsymbol{\theta}$ is defined as a vector of all the model parameters and hyperparameters:

$$\begin{aligned}
\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\, \tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}} \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}} \tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} \mid \mathbf{0}, \mathbf{V}^{-1}\right) \\
\boldsymbol{\beta}^l_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus \boldsymbol{\beta}^l_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j} \mathbf{X}^\top_{\boldsymbol{\eta}^l} \mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2} \mathbf{I}\right]^{-1} \mathbf{X}^\top_{\boldsymbol{\eta}^l} \mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j} \mathbf{X}^\top_{\boldsymbol{\eta}^l} \mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2} \mathbf{I}\right]^{-1}\right)
\end{aligned} \tag{A.16}$$

where we define $\mathbf{z}_{\boldsymbol{\beta}^l} = \omega_{j,j} \mathbf{y}^l + \sum_{k=1}^{k \neq l} \omega_{j,k} \left(\mathbf{y}^k - \mathbf{X}_{\boldsymbol{\eta}^k} \boldsymbol{\beta}^k\right)$. Model selection is now performed on each level of $\boldsymbol{\beta}^l$ in turn.

$$\begin{aligned}
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1} \otimes \mathbf{I})\, \tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}} \tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}} \tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right) \mathcal{N}\left(\tilde{\mathbf{b}} \mid \mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}\right) \\
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus \tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}} \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}} \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}} \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right)
\end{aligned} \tag{A.17}$$

where we define $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} = \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}_{\tilde{\mathbf{b}}_{-g}}\tilde{\mathbf{b}}_{\tilde{\mathbf{b}}_{-g}}$, where $\tilde{\mathbf{b}}_{-g}$ denotes $\tilde{\mathbf{b}}$ excluding group g and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}$ is the mean value calculated for $\mathbf{y}_{\tilde{\mathbf{b}}_{g,h}}$ for each response level l.

We now define the conditional distribution for fixed effects coefficients which are nested within specific random effects:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\delta}}_k \mid \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}, \boldsymbol{\Omega}^{-1}_{\tilde{\mathbf{b}}_k}\right)\mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \mathbf{0}, \mathbf{V}^{-1}\right) \\
\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \left[\tilde{\mathbf{X}}^{\top}_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\tilde{\mathbf{X}}^{\top}_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\boldsymbol{\delta}}_k\left[\tilde{\mathbf{X}}^{\top}_{\tilde{\boldsymbol{\delta}}_k}\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\right)
\end{aligned} \tag{A.18}$$

We sample here for each nested block of coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ for each k, where

$\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k} = \text{blockdiag}\left(\mathbf{X}^1{}_{\tilde{\boldsymbol{\delta}}_k}, \ldots, \mathbf{X}^l{}_{\tilde{\boldsymbol{\delta}}_k}\right)$, and $\tilde{\boldsymbol{\delta}}_k = \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} + \tilde{\mathbf{U}}_k\tilde{\mathbf{b}}_k$.

We can then obtain the conditional distributions for the precision parameters:

$$\begin{aligned}
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}^{-1}_{\boldsymbol{\epsilon}}\otimes\mathbf{I}\right)\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \nu_{\mathbf{b}_g}, \mathbf{S}_{\mathbf{b}_g}\right) \\
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\boldsymbol{g}}}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}^{-1}_{\tilde{\mathbf{b}}_g} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}}\tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}^{\top}_{g_i}\right]^{-1}\right)
\end{aligned} \tag{A.19}$$

where we sample each $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g.

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}^{-1}_{\boldsymbol{\epsilon}}\otimes\mathbf{I}\right)\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \nu_{\boldsymbol{\epsilon}}, \mathbf{S}_{\boldsymbol{\epsilon}}\right) \\
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{W}\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}^{-1}_{\boldsymbol{\epsilon}} + \sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}^{\top}_i\right]^{-1}\right)
\end{aligned} \tag{A.20}$$

where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$.

We then sample the prior variance parameter $\boldsymbol{\tau}$ for the fixed effects coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$

$$\begin{aligned}
\tau_l \mid \boldsymbol{\theta}_{\setminus\tau_l} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}_l} \mid \mathbf{0}, \tau_l\mathbf{I}\right)\mathcal{G}(\tau_l \mid a_l, b_l) \\
\tau_l \mid \boldsymbol{\theta}_{\setminus\tau_l} &\propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}^l_{\tilde{\boldsymbol{\eta}}_l}\|}{2}, b_l + \frac{\sum_{m=1}^{p}\left(\beta^l_m\right)^2}{2}\right)
\end{aligned} \tag{A.21}$$

where we sample for each l separately and form $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_l)$.

For the parameter expansion step, we perform a Metropolis-Hastings step on the random effect precision $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_m}$ and coefficients $\tilde{\mathbf{b}}_m$ by sampling α_m from the following distribution:

$$\alpha_m \propto \mathcal{G}\left(a_\alpha, b_\alpha\right) \tag{A.22}$$

We define $\tilde{\mathbf{b}}_{\mathbf{g}}^{*} = \alpha\tilde{\mathbf{b}}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^{*}_{\tilde{\mathbf{b}}_{\mathbf{g}}} = \alpha\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}$ and accept the move with probability

$$\phi = \frac{q(\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}_g, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}_g|\mathbf{0}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{-1}\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}^{-1}|\nu_{\tilde{\mathbf{b}}_g}, \mathbf{S}_{\tilde{\mathbf{b}}_g}\right)}{q(1/\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}_g^{*}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}_g^{*}|\mathbf{0}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g^{*}}^{-1}\right)\mathcal{W}\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g^{*}}^{-1}|\nu_{\tilde{\mathbf{b}}_g^{*}}, \mathbf{S}_{\tilde{\mathbf{b}}_g^{*}}\right)} \left|\mathbf{J}\right| \tag{A.23}$$

where $\left|\mathbf{J}\right| = \alpha^{\|\tilde{\mathbf{b}}_{\mathbf{g}}\|+(L(L+1))}$ and $\|\tilde{\mathbf{b}}_{\mathbf{g}}\|$ is the length of $\tilde{\mathbf{b}}_{\mathbf{g}}$. The additional expression $L(L+1)$ comes from the number of terms present in the covariance matrix which is of dimension $L \times L$.

The parameter update is accepted if $u < \phi$, where $u \sim \mathcal{U}(0,1)$.

A.5 Bayesian chain graph hierarchical model

The conditional distributions derived here are laid out in a similar fashion to Section 6.2.1 which adapts upon the mixing sampler by allowing for Bayesian Gaussian graphical model selection, with updates on the precision estimates, with adjustments for the G-Wishart distribution. Using the standard results for conditional Gaussian distributions discussed in Section A.1 and from Figure 6.6, we compute the conditional distributions for $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$ and $\tilde{\mathbf{b}}$, where $\boldsymbol{\theta}$ is defined as a vector of all the model parameters and hyperparameters:

$$\begin{aligned}
\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} \mid \mathbf{0}, \mathbf{V}^{-1}\right) \\
\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^l} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\beta}^{l}_{\boldsymbol{\eta}^l}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\eta}^l} \mid \left[\omega_{j,j}\mathbf{X}^{\top}_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\mathbf{X}^{\top}_{\boldsymbol{\eta}^l}\mathbf{z}_{\boldsymbol{\beta}^l}, \left[\omega_{j,j}\mathbf{X}^{\top}_{\boldsymbol{\eta}^l}\mathbf{X}_{\boldsymbol{\eta}^l} + \frac{1}{\tau_l^2}\mathbf{I}\right]^{-1}\right)
\end{aligned} \tag{A.24}$$

where we define $\mathbf{z}_{\boldsymbol{\beta}^l} = \omega_{j,j}\mathbf{y}^l + \sum_{k=1}^{k\neq l}\omega_{j,k}\left(\mathbf{y}^k - \mathbf{X}_{\boldsymbol{\eta}^k}\boldsymbol{\beta}^k\right)$. Model selection is now performed on each level of $\boldsymbol{\beta}^l$ in turn.

$$\begin{aligned}
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus\tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{N}\left(\tilde{\mathbf{b}} \mid \mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\mathbf{b}}}\right) \\
\tilde{\mathbf{b}}_{g,h} \mid \boldsymbol{\theta}_{\setminus\tilde{\mathbf{b}}_{g,h}} &\propto \mathcal{N}\left(\tilde{\mathbf{b}}_{g,h} \mid \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1} n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}, \left[\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} + n_{\tilde{\mathbf{b}}_{g,h}}\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}\right]^{-1}\right)
\end{aligned} \tag{A.25}$$

where we define $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} = \bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}_{\tilde{\mathbf{b}}_{-g}}\tilde{\mathbf{b}}_{\tilde{\mathbf{b}}_{-g}}$, where $\tilde{\mathbf{b}}_{-g}$ denotes $\tilde{\mathbf{b}}$ excluding group g and $\bar{\mathbf{y}}_{\tilde{\mathbf{b}}_{g,h}}$ is the mean value calculated for $\mathbf{y}_{\tilde{\mathbf{b}}_{g,h}}$ for each response level l.

We now define the conditional distribution for fixed effects coefficients which are nested within specific random effects:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\delta}}_k \mid \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}, \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}^{-1}\right)\mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \mathbf{0}, \mathbf{V}^{-1}\right) \\
\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \boldsymbol{\theta}_{\setminus\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} \mid \left[\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}^{\top}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}^{\top}\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_k}\tilde{\boldsymbol{\delta}}_k, \left[\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}^{\top}\boldsymbol{\Sigma}_{\tilde{\mathbf{b}}_k}\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\right]^{-1}\right)
\end{aligned} \tag{A.26}$$

We sample here for each nested block of coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k}$ for each k, where

$\tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k} = \text{blockdiag}\left(\mathbf{X}^1{}_{\tilde{\boldsymbol{\delta}}_k}, \ldots, \mathbf{X}^l{}_{\tilde{\boldsymbol{\delta}}_k}\right)$, and $\tilde{\boldsymbol{\delta}}_k = \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\delta}}_k}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\delta}}_k} + \tilde{\mathbf{U}}_k\tilde{\mathbf{b}}_k$.

We can then obtain the conditional distributions for the precision parameters:

$$\begin{aligned}
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{W}_G\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \nu_{\mathbf{b}_g}, \mathbf{S}_{\mathbf{b}_g}\right) \\
\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}} &\propto \mathcal{W}_G\left(\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g} \mid n_{\tilde{\mathbf{b}}_g} + \nu_{\tilde{\mathbf{b}}_g}, \left[\mathbf{S}_{\tilde{\mathbf{b}}_g}^{-1} + \sum_{i=1}^{n_{\tilde{\mathbf{b}}_g}}\tilde{\mathbf{b}}_{g_i}\tilde{\mathbf{b}}_{g_i}^{\top}\right]^{-1}\right)
\end{aligned} \tag{A.27}$$

where we sample each $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_g}$ for each group g.

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{N}\left(\boldsymbol{y} \mid (\mathbf{1}\otimes\mathbf{I})\,\tilde{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}_{\tilde{\boldsymbol{\eta}}}\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\otimes\mathbf{I}\right)\right)\mathcal{W}_G\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \nu_{\boldsymbol{\epsilon}}, \mathbf{S}_{\boldsymbol{\epsilon}}\right) \\
\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}} &\propto \mathcal{W}_G\left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mid n + \nu_{\boldsymbol{\epsilon}}, \left[\mathbf{S}_{\boldsymbol{\epsilon}}^{-1} + \sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^{\top}\right]^{-1}\right)
\end{aligned} \tag{A.28}$$

where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{U}}\tilde{\mathbf{b}}$.

We then sample the prior variance parameter $\boldsymbol{\tau}$ for the fixed effects coefficients $\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}}$

$$\begin{aligned}
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\eta}}_l} \mid \mathbf{0}, \tau_l \mathbf{I}\right) \mathcal{G}(\tau_l \mid a_l, b_l) \\
\tau_l \mid \boldsymbol{\theta}_{\setminus \tau_l} &\propto \mathcal{G}\left(\tau_l \mid a_l + \frac{\|\boldsymbol{\beta}^l_{\tilde{\boldsymbol{\eta}}_l}\|}{2}, b_l + \frac{\sum_{m=1}^{p} \left(\beta^l_m\right)^2}{2}\right)
\end{aligned} \tag{A.29}$$

where we sample for each l separately and form $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_l)$.

For the parameter expansion step, we perform a Metropolis-Hastings step on the random effect precision $\boldsymbol{\Omega}_{\tilde{\mathbf{b}}_m}$ and coefficients $\tilde{\mathbf{b}}_m$ by sampling α_m from the following distribution:

$$\alpha_m \propto \mathcal{G}\left(a_\alpha, b_\alpha\right) \tag{A.30}$$

We define $\tilde{\mathbf{b}}^*_{\mathbf{g}} = \alpha \tilde{\mathbf{b}}_{\mathbf{g}}$ and $\boldsymbol{\Omega}^*_{\tilde{\mathbf{b}}_{\mathbf{g}}} = \alpha \boldsymbol{\Omega}_{\tilde{\mathbf{b}}_{\mathbf{g}}}$ and accept the move with probability

$$\phi = \frac{q(\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}_g, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}_g|\mathbf{0}, \boldsymbol{\Omega}^{-1}_{\tilde{\mathbf{b}}_g}\right)\mathcal{W}\left(\boldsymbol{\Omega}^{-1}_{\tilde{\mathbf{b}}_g}|\nu_{\tilde{\mathbf{b}}_g}, \mathbf{S}_{\tilde{\mathbf{b}}_g}\right)}{q(1/\alpha)\mathcal{N}\left(\mathbf{y}|\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{U}}\tilde{\mathbf{b}}^*_g, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1}\right)\mathcal{N}\left(\tilde{\mathbf{b}}^*_g|\mathbf{0}, \boldsymbol{\Omega}^{-1}_{\tilde{\mathbf{b}}^*_g}\right)\mathcal{W}\left(\boldsymbol{\Omega}^{-1}_{\tilde{\mathbf{b}}^*_g}|\nu_{\tilde{\mathbf{b}}^*_g}, \mathbf{S}_{\tilde{\mathbf{b}}^*_g}\right)} \left|\mathbf{J}\right| \tag{A.31}$$

where $\left|\mathbf{J}\right| = \alpha^{\|\tilde{\mathbf{b}}_{\mathbf{g}}\|+(L(L+1))}$ and $\|\tilde{\mathbf{b}}_{\mathbf{g}}\|$ is the length of $\tilde{\mathbf{b}}_{\mathbf{g}}$. The additional expression $L(L+1)$ comes from the number of terms present in the covariance matrix which is of dimension $L \times L$.

The parameter update is accepted if $u < \phi$, where $u \sim \mathcal{U}(0,1)$.

-

References

(2017). *ISO International Standard ISO/IEC 14882:2014(E) Programming Language C++*. International Organization for Standardization (ISO), Geneva, Switzerland. 84

Adank, P., Smits, R., and Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5):3099–3107. 12

Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92:317–335. 83, 86

Baayen, R. H. (2008). *Analyzing linguistic data*. Cambridge University Press, Cambridge. 20

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. 42

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 141

Browne, W., Steele, F., Golalizadeh, M., and Green, M. (2009). The use of simple reparameterizations to improve the efficiency of markov chain monte carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society: Series A*, 172:579–598. 3

Cappe, O., e. a. (2001). Reversible jump birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society, Series B*, 65:679–700. 86

Cedergren, H. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, (50):333–355. 1

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2015). *shiny: Web Application Framework for R*. R package version 0.11.1. 4

Delattre, P. e. a. (1951). Voyelles synthetiques a deux formantes et voyelles cardinales. *Le Maitre Phonetique*. 9

Dobra, A. e. a. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106. 85, 86

Drager, K. and Hay, J. (2012). Exploiting random intercepts : two case studies in sociophonetics. *Language Variation and Change*, 24:59–78. 15

Eckert, P. and McConnell-Ginet, S. (2003). *Language and Gender*. Cambridge University Press, Cambridge. 14

Eddelbuettel, D. and Francois, R. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40:1–18. 85

Fromont, R. and Hay, J. (2012). Labb-cat: an annotation store. *Proceedings of Australasian Language Technology Association Workshop*, pages 113–117. 10

Gelfand, A., Sahu, S., and Carlin, B. (1995). Efficient parameterizations for normal linear mixed models. *Biometrika*, 82:479–488. 49, 50, 51, 78, 136

Gelman, A., van Dyk, D., Huang, Z., and Boscardin, W. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17:95–122. 64, 68, 137

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 28

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99:799–804. 31

Gottard, A. and Rampichini, C. (2006). Ghain graphs for multilevel models. *Statistics and Probability Letters*, 77:312–318. 97

Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732. 30, 86

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*. 29

IPA (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press. xi, 7

Johnson, D. (2009). Getting off the goldvarb standard: introducing rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, pages 359–383. 2, 20, 21, 47, 136

Johnson, K. (2011). *Acoustic and Auditory Phonetics*. Oxford: Blackwell. 7, 9

Jose, B. and Stuart-Smith, J. (2014). Methodological issues in a real-time study of glaswegian vowels: Automation and comparability. 10

Kendall, T. and Thomas, E. R. (2014). *vowels: Vowel Manipulation, Normalization, and Plotting*. R package version 1.2-1. 12

Labov, W. (1994). *Principles of Lingustic Change, Vol 1: Internal Factors*. Blackwell, Oxford. 9, 12, 14

Labov, W. (2001). *Principles of Linguistic Change*. Oxford:Blackwell. 1, 14

Ladefoged, P. (2005). *Vowels and Consonants: An Introduction to the Study of Languages*. Oxford: Blackwell. 8

Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Cengage learning, Stamford, CT, 7 edition. 2, 7, 14

Laing, F. (2010). A double metropolis-hastings sampler for spatial models with intractiable normalizing constants. *Jounrnal of Statistical Computing and Simulation*, 80:1007–1022. 86

Lauritzen, S. L. (2006). *Elements of Graphical Models*. Springer. 82

Lenkoski, A. (2013). A direct sampler for g-wishart variates. *Stat*, 2:119–128. 85, 86

Liu, C., Rubin, D., and Wu, Y. (1998). Parameter expansion to accelerate em: The px-em algorithm. *Biometrika*, 85:755–770. 49, 64, 78

Liu, C. and Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274. 64

Metropolis, N., Rosenbluth, A., Rosenbluth, M., and Teller, A. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092. 29

Mohammadi, A. and Wit, E. (2016). *BDgraph: Bayesian Graph Selection Based on Birth-Death MCMC Approach*. R package version 2.27. 84

Murray, I., Ghahramani, Z., and MacKay, D. (2006). Mcmc for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*. 85

Pinheiro, J. and Bates, D. (2000). *Mixed effects models in S and S-plus*. Springer. 2

Priestley, M. (1981). *Spectral Analysis and Time Series 1*. Academic Press. 41

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 36

Rand, D. and Sankoff, D. (1990). *GoldVarb: A variable rule application for the Macintosh*. Centre de recherches mathematiques, Universite de Montreal. 2

Rathcke, T., Stuart-Smith, J., Torsney, B., and Harrington, J. (2017). The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies. *Speech Communication*, 86:24–41. 3

Sargent, D., Hodges, J., and Carlin, B. (2000). Structured markov chain monte carlo. *Jounal of Computational and Graphical Statistics*, 9:217–234. 50

Smith, J. and Holmes-Elliott, S. (2017). The unstoppable glottal: tracking rapid change in an iconic British variable 1. *English Language & Linguistics*, pages 1–33. 1

Stuart-Smith, J., José, B., Rathcke, T., Macdonald, R., and Lawson, E. (2017). Changing Sounds in a Changing City: An Acoustic Phonetic Investigation of Real-Time Change over a Century of Glaswegian. In Montgomery, C. and Moore, E., editors, *Language and a Sense of Place: Studies in Language and Region*, pages 38–65. Cambridge University Press, Cambridge. 3, 6, 10, 11, 15, 113, 114, 134, 138

Stuart-Smith, J. and Lawson, E. (2017). Scotland: Glasgow/the central belt. In Hickey, R., editor, *Listening to the Past: Audio Records of Accents of English*, pages 171–98. Cambridge University Press, Cambridge. 3

Tagliamonte, S. (2012). *Variationist Sociolinguistics*. Wiley-Blackwell. 1, 14

Tagliamonte, S. and Baayen, R. (2012). Models, forests and trees of york english;was/were variation as a case study for statistical practice. *Language Variation and Change*, 24:135–178. 2

Wang, H. and Li, S. (2012). Efficient gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 4:1470–1475. 86, 87, 88, 90, 91, 138

West, B., Welch, K., and Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall. 21

Wit, E. C. and Mohammadi, A. (2015). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10:109–138. 86, 139