Bayesian Semi-Modular Inference: Theory and Methods for Inference in Multi-Modular Settings Under Model Misspecification



Chris U. Carmona

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

To my parents: Moyo, Pepe and Yaya.

Acknowledgements

I would like to express my deepest gratitude to Prof. Geoff Nicholls, for his invaluable guidance, dedication, and patience throughout my doctoral journey. More than a supervisor, he has been a trusted mentor enabling my growth as a researcher and individual, a colleague who showed me the joy of real teamwork, and a caring friend who supported me and our group during the difficult times of the Covid-19 pandemic. I feel incredibly privileged to have worked under his kind yet rigorous supervision.

My gratitude extends to other mentors who have greatly impacted my academic and personal growth. Thanks to Prof. Luis Enrique Nieto, who sparked my passion for Bayesian statistics and encouraged me to pursue a DPhil at Oxford. I am tremendously grateful to Dr. Serafín Martinez for being a pillar of unwavering support and motivation throughout these years, especially during the final stretches of writing. His uplifting words and advice kept me persevering.

During my time in Oxford, I was privileged to meet many wonderful and inspiring people who became dear friends. In particular, Fadhel Ayed, Philippe Gagnon, Roman Stasinski, Emilia Pompe, Giuseppe Di Benedetto, Paul Vanetti, Virginia Aglietti, Kamélia Daudel and Claire Huang. Your friendship made my days in Oxford truly memorable and enjoyable. I greatly appreciated all the adventures, conversations, and the caring support you provided. Thanks to all my colleagues, professors and staff from the Statistics Department. Your collective wisdom, encouragement, and community were invaluable.

Most importantly, I wish to thank my partner and my family for their unconditional love and support throughout this journey: Ale Salazar, Moyo, Yaya, Cache, Luke, Ale, Tama, and all my beloved *muégano*.

I gratefully acknowledge the financial support of the institutions that made my research financially viable: Consejo Nacional de Ciencia y Tecnología (CONACYT), Banco de México (Banxico) and the University of Oxford.

Abstract

One of the strengths of Bayesian analysis lies in its ability to incorporate all relevant data into the analysis and address all uncertainties in the model simultaneously. Evidence combination is a core aspect of statistical inference, and the Bayesian approach is particularly well-suited for this task. However, as model structures become larger and more intricate, it becomes increasingly challenging for Bayesians to attempt to model the true data-generating mechanism. Bayesian statistical inference loses predictive optimality when generative models are misspecified.

The increasing complexity of modern applications in statistics has led to an increased interest in multi-modular models. Unfortunately, these models are susceptible to model contamination, as the risk of misspecification in each module accumulates as modules are added. The Cut Posterior and Modularized Inference surged as an approach to address model misspecification by eliminating feedback from suspect modules. Unfortunately, the understanding of such methods and available algorithms for them were limited until recently.

This dissertation provides three major contributions to the growing field of Modular Bayesian Inference. First, we introduce a generalized inferential framework that bridges the gap between standard Bayesian inference and the Cut posterior by proposing an overarching methodology that we call Semi-Modular Inference (SMI). Second, we develop novel algorithms for the computation of the Semi-Modular Posterior, enabling the use of Modularized Inference in large scale applications with thousands of model parameters and multiple cuts. Third, we provide evidence of the benefits of Semi-Modular Inference through several real-world applications, where we outperform both the Cut and Bayes posteriors by efficiently tuning the influence of modules using the Semi-Modular Posterior.

Contents

1 Introduction **1**
1.1 Bayesian Inference in Multi-modular Settings. 1
1.2 Semi-Modular Inference 3
1.3 Computational Methods for Semi-Modular Inference 4
1.4 Applications of Semi-Modular Inference 5

2 Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components **8**
2.1 Introduction 8
2.2 Background methods 10
2.3 Semi-Modular Inference 12
2.4 Analysis with (Semi-)Modular Inference 13
2.5 Examples 20
2.6 Discussion 25
2.7 Supplementary Material 25

3 Scalable Semi-Modular Inference with Variational Meta-Posteriors **47**
3.1 Introduction 47
3.2 Modular Bayesian Inference 51
3.3 Variational Modular Inference 57
3.4 The Variational Meta-Posterior 69
3.5 Experiments 74
3.6 Discussion 81
3.7 Supplementary Material 85

4 Statistical Analysis of the Agroecology of Early 'Cities' in Northern Mesopotamia, the Aegean and Central Europe **106**
4.1 Introduction 106
4.2 Data description 108

4.3 Statistical Models . . . 110
4.4 Single Imputation Analysis . . . 111
4.5 Bayesian Multiple Imputation (BMI) . . . 116
4.6 Semi-Modular Inference . . . 122
4.7 Overall conclusions . . . 126
4.8 Appendix . . . 128

5 Simultaneous Reconstruction of Spatial Frequency Fields and Sample Locations via Bayesian Semi-Modular Inference 132
5.1 Introduction . . . 133
5.2 The Linguistic Atlas of Late Mediaeval English (LALME) Data . . . 136
5.3 Methodology . . . 139
5.4 Model . . . 143
5.5 Semi-Modular Inference for the LALME model . . . 150
5.6 Implementation details . . . 154
5.7 Data Analyses . . . 159
5.8 Conclusions . . . 166
5.9 Supplementary Material . . . 169

6 Conclusions 179
6.1 Summary of contributions . . . 179
6.2 Future work . . . 180

Bibliography 181

Chapter 1

Introduction

1.1 Bayesian Inference in Multi-modular Settings.

The field of statistical inference has undergone significant developments in recent decades, with Bayesian Inference emerging as a central methodology for understanding complex systems and making predictions based on uncertain data. The Bayesian approach provides a coherent framework for modelling and performing inference by incorporating prior knowledge through probability distributions, and updating this knowledge in light of new observations. The versatility and ubiquity of Bayesian Inference is demonstrated by its wide range of applications across disciplines, including natural and social sciences, engineering, computer science, linguistics, archaeology, among many other fields. In particular, when dealing with multiple observation models and data sets, with some model parameters appearing in more than one observation model, a Bayesian multi-modular setting can be employed. This, in principle, would allow for a more comprehensive understanding of the data, by incorporating information from different sources and models simultaneously.

The increasing complexity of modern applications in statistics has led to an increased interest in multi-modular settings in recent years. For example, the Coronavirus Disease 2019 (COVID-19) pandemic has presented new challenges to conventional statistical practice, and the concept of "statistical interoperability" has emerged as an aspiration for the design of effective disease surveillance systems (Nicholson et al., 2022; Richardson, 2022). Modularity is a desirable feature for building inter-operable model structures, as it facilitates breaking down complex models into smaller, self-contained modules that can be developed in parallel and updated as new knowledge becomes available.

However, as model structures become larger and more intricate, it becomes increasingly challenging for Bayesians to attempt to model the true data-generating

mechanism. Multi-modular models are susceptible to model contamination, as the risk of misspecification in each module accumulates as modules are added. If any module is significantly misspecified, it can undermine inference on the joint model. The seminal article by Liu et al. (2009) analyses the dangers of model contamination, and illustrate, with minimal examples, how misspecification in one of the modules can harm inference in the model as a whole.

Model elaboration (Smith, 1986; Gelman et al., 2015) may tackle misspecification in some cases, but in complex setting it could be impractical or at least very challenging. Alternatively, one may consider "inference elaboration" and turn to statistically principled alternatives to Bayesian inference which are robust to some types of misspecification and preserve a coherent update of beliefs (Zhang, 2006; Grünwald and Ommen, 2017; Bissiri et al., 2016; Nicholls et al., 2022).

Modularized Inference (Liu et al., 2009; Plummer, 2015; Jacob et al., 2017) is one such alternative approach which addresses contamination by eliminating feedback from misspecified modules. This approach proceeds by estimating model parameters in multiple stages, sequentially updating beliefs by on a subset of parameters at each stage. Modularized Bayesian Inference and the cut posterior have received increasing attention in recent years, becoming an active area of academic research. There is a plethora of new advances related to the cut posterior, including studies on its asymptotic properties (Pompe and Jacob, 2021; Pompe, 2021) and improved numerical algorithms for sampling from it using MCMC (Plummer, 2015; Jacob et al., 2020; Liu and Goudie, 2022b), ABC (Chakraborty et al., 2023), and Variational Inference (Carmona and Nicholls, 2022; Yu et al., 2021). Different interpretations of the Cut posteriors have also been proposed. An algorithmic interpretation was originally proposed by Spiegelhalter et al. (2014), then Plummer (2015) characterised the *Cut posterior*, Carmona and Nicholls (2020) analysed the cut posterior as a generalised update of beliefs under the framework of Generalised Bayesian Inference Bissiri et al. (2016), while Yu et al. (2021) gave it an interpretation as the solution of a constrained optimisation problem.

Practical applications of modularized Bayesian analysis are numerous. Scientist and practitioners across disciplines seem to have adopted the methods as part of their inferential toolbox. A few examples of these applications in the natural sciences include ecological models of animal abundance (Besbeas et al., 2002; Brooks et al., 2004) and forest dynamics (Ogle et al., 2013), pharmacological PK/PD models (Lunn et al., 2009), epidemiological studies on air pollution (Blangiardo et al., 2011; Szpiro and Paciorek, 2013), Human papillomavirus (HPV) (Maucort-Boulch et al., 2008), and

COVID-19 (Nicholson et al., 2022). In the social sciences we find political models for voting preferences Jessee (2010), two-step econometric models linking unemployment and money growth Murphy and Topel (1985), and anthropological studies on early civilisations (Styring et al., 2017), among many more.

1.2 Semi-Modular Inference

In Chapter 2, we present our first major contribution: Semi-Modular Inference. Conceptually, SMI can be viewed as a smooth interpolation between the standard Bayesian posterior and the Cut posterior, controlled by an influence parameter $\eta \in [0, 1]$. This parameter regulates the feedback from a suspect module into the overall inference process.

The original version of SMI is constructed as a two-stage procedure, which combines principles of Bayesian Multiple Imputation (BMI) and Generalised Bayesian Inference (GBI) in the form of the power-posterior (Bissiri et al., 2016; Holmes and Walker, 2017). In the first stage, the *imputation stage*, we perform Generalised Bayesian Inference (GBI) using the probabilistic model defined for the data by incorporating a scalar $\eta \in [0, 1]$ to reduce the influence of the suspect module on the joint model by raising the corresponding terms to the η power. This stage yields the posterior distribution of link parameters that are shared across modules. Then, in a second *analysis stage*, we relearn the distribution of module-specific parameters conditioning on the "imputed" values from stage 1.

We demonstrate that the Semi-Modular Posterior (SMP) (and the Cut posterior as a particular case) are order-coherent updates of beliefs in the sense of Bissiri et al. (2016) and Nicholls et al. (2022).

We present a simple nested MCMC algorithm to sample from the Semi-Modular Posterior, motivated by the two-stage algorithms for the cut posterior discussed in Plummer (2015). Additionally, we propose choosing of the influence parameter η by maximising an utility function, $U(\eta)$, defined by the analyst according to the inferential goal of the study. We suggest using the Expected Log-pointwise Predictive Density (ELPD) (Vehtari et al., 2017) when predictive performance is the goal.

We illustrate the usage of SMP with two (now classical) examples from the literature of Cut models: the synthetic biased data originally studied by (Liu et al., 2009), and the Human papillomavirus dataset analysed by Plummer (2015). Additionally, we revisit the analysis by Styring et al. (2017) for an archaeological application, and

show that SMI offers improved performance for imputation when compared with the Cut model used in their analysis.

1.3 Computational Methods for Semi-Modular Inference

The Semi-Modular approach for inference poses computational challenges due to the presence of intractable parameter-dependent marginal factors (Plummer, 2015). Previously, algorithms for modularized analysis were limited to MCMC methods. However, these methods, such as those outlined in Plummer (2015); Carmona and Nicholls (2020); Liu and Goudie (2022b) suffer from double asymptotics and require convergence of both the main chain and all conditional sub-chains. Recent developments, such as the unbiased samplers introduced by Jacob et al. (2020) have alleviated some of these challenges, but the scalability issues for MCMC in large-scale applications remain a concern. These computational difficulties may discourage practitioners from taking an exploratory approach to algorithmically "discover" the misspecified modules in settings where multiple cuts are possible.

Our second contribution, in Chapter 3, aims to address these challenges by proposing scalable variational methods for Semi-Modular Inference. We develop principled and scalable algorithms that expand the reach of modularized analysis to large-scale models, both in terms of the number of parameters and the number of candidate cuts. To achieve this objective, we developed a variational framework that allows for end-to-end training of the Variational SMP using modern tools for Automatic Differentiation, as well as recent advances in Stochastic Variational Inference (SVI) and Normalizing Flow (NF).

We define three properties that characterise a useful variational approximation of the SMP. It is shown that the solution by naive minimisation of the Kullback-Leibler Divergence (KL) to the SMI posterior does not satisfy these property as it "leaks" information from the cut module. To circumvent this issue, we define a loss tailored to our goal, and demonstrate that minimisation of such loss can be achieved in two equivalent ways: 1) using a two-stage optimisation procedure, or 2) directly minimising and overall loss using Stochastic Gradient Descent (SGD) in combination with the stop-gradient operator. The second approach has the computational advantage of enabling end-to-end training in a single optimisation loop. In concurrent independent work Yu et al. (2021) also developed variational methods for the Cut posterior. Our approaches match at the Cut posterior, just as SMI interpolates Cut and Bayes, so variational SMI interpolates variational-Cut and variational-Bayes exactly.

Up to this point, we have used numerical algorithms that target the Semi-Modular Posterior work for a single value of the influence parameter η. Approximating the posterior of a different value, say η', would require the analyst to start afresh, without making much use of the samples (or approximation) obtained before. Continuity of the Semi-Modular Posterior on η presents an opportunity that can be exploited for a more efficient exploration.

A contribution on this direction is given by our Variational Meta-Posterior (VMP). The VMP is a family of variational distributions that simultaneously approximate an entire family of posterior distributions by using on a single set of parameters. We present two variants of the VMP: The VMP-map and the VMP-flow. In both cases, the VMP is trained to produce samples from any arbitrary posterior indexed by η, without additional fine tuning. The VMP is not restricted to applications on Semi-Modular Inference, as it could be used to any family of distributions indexed by a hyper-parameter $\eta \in \mathbb{R}^d$.

We illustrate the capabilities of our variational methods with two examples: the Human papillomavirus dataset by Plummer (2015) and an extended version of the synthetic random-effects data studied in (Liu et al., 2009). In both cases, we show that the variational approximation is in good agreement with samples obtained from an MCMC sampler, successfully approximating complex features in the posterior distribution, such as multiple modes. In the second example we show how the VMP can be used to algorithmically identify misspecified modules in a setting with thirty possible cuts.

1.4 Applications of Semi-Modular Inference

In the second part of the thesis, we demonstrate the utility of Semi-Modular Inference and the proposed methods in two substantial real-world applications.

In Chapter 4, we analyse data from archaeological excavations of three settlement sites in northern Mesopotamia, the Aegean and south-west Germany to investigate the relationship between agricultural intensity and population size. The main interest lies in finding evidence for the hypothesis of agricultural extensification, which refers to an expansive production strategy that is land rather than labour limited. Statistically, this translates on testing for a negative effect due to population size in settlements on the manuring used for their crops.

Manuring levels are not measurable from the archaeological remains and must be inferred. A model for imputation is constructed using data from a modern ex-

perimental setting. the imputation model consists of a hierarchical regression model relating manuring level and other covariates such as cereal species and irrigation levels to a response variable measuring the concentration of an nitrogen isotope in the soil. We compare frequentist and Bayesian methods for the analysis, obtaining agreement between the various approaches.

There are two modules in this setting, a Proportional Odds model relating population size with manure level, and a hierarchical regression model for the imputation of manure levels. We use Semi-Modular Inference to control the influence of the imputation module, and select the optimal value of η by maximising the predictive performance measured by the ELPD. The overall conclusion of our analysis supports the hypothesis of agricultural extensification, for the sites in Mesopotamia and the Aegean, but not for the site in Germany.

Finally, in Chapter 4 we present a large-scale application of Semi-Modular Inference in the context of a problem in spatial statistics: The simultaneous reconstruction of spatial frequency fields and sample locations. The general setting studied in this chapter is an extension to the *point-referenced data* setting (Banerjee et al., 2014) expanded to accommodate for joint inference of missing locations. This is a challenging problem for conventional Bayesian Inference if the number of missing locations is large and the generative model for the data is even slightly misspecified. Feedback between estimated fields and the imputed locations may become problematic as incorrect locations may contaminate the fields posterior, which in turn worsen the inference of locations. Semi-Modular Inference offers a solution by controlling the feedback of imputed locations into the latent fields.

The motivating data for this application comes from a large corpus of medieval manuscripts known as The Linguistic Atlas of Late Mediaeval English (LALME) (McIntosh et al., 1986; Benskin et al., 2013). These texts were written in England, parts of Wales, and a small number from southern Scotland, and span a period between 1350 to 1450, approximately. Modern dialectological methods were applied to produce a rich dataset displaying the continuum of linguistic variation in written English during the period. The main inferential goal of the applied problem is to predict the provenance of manuscripts with unknown locations (known as floating linguistic profiles) based on its dialectic characteristics.

We develop a hierarchical model for the motivating data, which broadly consist of a latent multi-layer structure of corregionalised Gaussian Processes and a categorical observational model. Due to the large scale of the data and probabilistic model, variational methods for Semi-Modular Inference are the only viable option we found

to perform inference over the entire dataset. We found it infeasible to use MCMC when the number of floating profiles is at all large, as we could not get convergence in a reasonable time. Nevertheless, by using a reduced subset of the data, we observe good agreement between variational and Monte Carlo methods.

The optimal rate of influence η^* is chosen by minimising the variational-posterior MSE over a set of held-out profiles with known locations. As a final result, we provide estimations for all the floating profiles and a formal quantification of uncertainty. We found general agreement between our predictions and those established by subject-matter experts.

Chapter 2

Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components

Abstract

Bayesian statistical inference loses predictive optimality when generative models are misspecified. Working within an existing coherent loss-based generalisation of Bayesian inference, we show existing Modular/Cut-model inference is coherent, and write down a new family of *Semi-Modular Inference (SMI)* schemes, indexed by an influence parameter, with Bayesian inference and Cut-models as special cases. We give a meta-learning criterion and estimation procedure to choose the inference scheme. This returns Bayesian inference when there is no misspecification. The framework applies naturally to Multi-modular models. Cut-model inference allows directed information flow from well-specified modules to misspecified modules, but not vice versa. An existing alternative power posterior method gives tunable but undirected control of information flow, improving prediction in some settings. In contrast, SMI allows *tunable and directed* information flow between modules. We illustrate our methods on two standard test cases from the literature and a motivating archaeological data set.

2.1 Introduction

Consider statistical inference in a multi-modular setting. The model for the available data is assembled from several component *modules*. Each module describes a probabilistic relation between observable variables (data) and unknown quantities (parameters, latent variables, missing data). Modules may share parameters, missing

data and other latent variables. Figure 2.1 illustrates a simple two-module model with a shared parameter φ (ignore the dashed line explained in section 2.2.1). The first module has data Z and one parameter, φ, while the second module has data Y and two parameters, θ and φ.

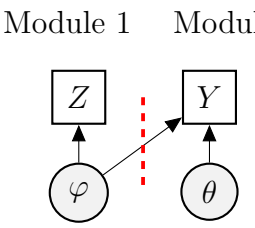


Figure 2.1: Graphical representation of a simple multi-modular model.

In conventional Bayesian inference, parameters are jointly informed by the data and model assumptions shared across the modules. As a consequence, large-scale multi-modular analyses are particularly susceptible to problems arising from model misspecification, as any bad module may distort inference in the model as a whole (Liu et al., 2009).

This drawback has motivated alternative inferential approaches that modify conventional Bayesian learning in order to regulate feedback between modules. *Modular* inference (Liu et al., 2009; Jacob et al., 2017) also known as a *Cut-model* inference (Spiegelhalter et al., 2014; Plummer, 2015) completely eliminates the contribution from some modules to the posterior distribution of parameters in other modules (see section 2.2.1). However, we may do better to moderate, rather than eliminate, the influence of misspecified modules.

We give a new *Semi-Modular Inference (**SMI**)* which *smoothly* regulates the influence of modules on the overall inference. The procedure effectively expands the space of *candidate distributions* [1] in such a way that Bayesian inference and Cut-model inference are particular cases.

When it comes to expanding the inference framework, an "anything goes" approach is clearly suspect. We stay within the class of inference schemes defined by Bissiri et al. (2016). Those authors define and characterise coherent loss-based inference and note that Bayesian inference and the power posterior (Walker and Hjort, 2001) are coherent. We set out existing Cut-model and Power-posterior inference in

[1] following Jacob et al. (2017), we refer to any distribution representing beliefs on parameters θ (or φ, or both) as a *candidate distribution* for the parameters

sections 2.2.1 and 2.2.2, introduce SMI in section 2.3 and then describe the encompassing framework of Bissiri et al. (2016) in section 2.4.1, where we show that SMI and Cut-model inference are coherent.

The SMI posterior interpolates candidate distributions between the conventional Bayesian posterior and the Cut-model posterior. Candidate distributions are indexed by a continuous *degree of influence* parameter η. This controls the contribution of a module to the candidate distribution. When η=0 the candidate distribution is the Cut-model posterior and when η=1 it is the Bayesian posterior. The SMI posterior is not a scheme for model elaboration with an extra parameter. SMI-based inference with any value of η other than $\eta = 1$ is not Bayesian inference.

In section 2.5, we apply SMI to model-based inference for simulated and real-world datasets and evaluate its performance. It is easy to understand why it outperforms Bayesian and Modular inference in these examples. The supplementary material provides proofs and additional numerical experiments. All results are reproducible using the accompanying R package[2].

2.2 Background methods

2.2.1 Modular Inference: cut model

The Cut model is an alternative to Bayesian inference designed to remove unwanted feedback from poorly specified modules. The OpenBUGS manual (Spiegelhalter et al., 2014) describes the cut function with the words "The cut function acts as a kind of valve in the graph: prior information is allowed to flow downwards through the cut, but likelihood information is prevented from flowing upwards". Cut model inference is a form of Bayesian multiple imputation.

Consider again the two-module configuration of fig. 2.1. In standard joint or "full" Bayesian inference, information from the two modules informs every parameter, so in particular the values of Y will in general influence the posterior distribution of φ. The posterior distribution of (φ, θ) in Bayesian inference is

$$p(\varphi, \theta \mid Z, Y) = p(\varphi \mid Z, Y)p(\theta \mid Y, \varphi). \tag{2.2.1}$$

The marginal distribution of φ depends on Y.

Now, suppose that for some reason (usually because we suspect some model misspecification) we want the parameter φ to learn only from module 1, *cutting* the

[2] `https://github.com/chriscarmona/aistats2020smi`

influence from module 2 on φ. The cut is represented in fig. 2.1 by a dashed line on the edge from φ to Y; it denotes an inference structure in which φ influences Y but not vice versa (following Lunn et al. (2009)).

Under this modified scheme, the Cut-model "posterior" for (φ, θ) is

$$p_{cut}(\varphi, \theta \mid Z, Y) = p(\varphi \mid Z)p(\theta \mid Y, \varphi). \tag{2.2.2}$$

Notice that the marginal distribution of φ no longer depends on Y.

Cut-model inference is a form of Bayesian Multiple Imputation (Lunn et al., 2009; Styring et al., 2017) in which we make multiple imputation of φ and then analysis of θ given the imputated distribution of φ. The literature identifies potential advantages: cut models may simplify inference (Cox, 1975); prevent unwanted feedback from suspect models (Lunn et al., 2009); improve MCMC mixing (Plummer, 2015); reduce the MSE in estimates (Liu et al., 2009); increase predictive performance (Jacob et al., 2017); answer the need in some settings to make a sequential analysis in which the data Z is not shared with the analyst carrying out inference for θ.

2.2.2 Power posterior

In the power posterior we raise the likelihood to a power, seeking to improve robustness under model misspecification (Walker and Hjort (2001); Bissiri et al. (2016); Holmes and Walker (2017); Grünwald and Ommen (2017); Miller and Dunson (2018)).

Consider independent data $Y = (Y_1, \ldots, Y_n)$ generated from an unknown true distribution $f^*(Y)$. Assume we have a data model $f(Y|\theta)$ and a prior distribution $p(\theta)$. For a fixed $\eta \in \mathbb{R}$, we define the *η-powered posterior* $p_{pow,\eta}(\theta|x)$ as

$$p_{pow,\eta}(\theta|Y) = \frac{f(Y|\theta)^\eta p(\theta)}{p_\eta(Y)} \tag{2.2.3}$$

with $p_\eta(Y) = \int f(Y|\theta)^\eta p(\theta) d\theta$ the *powered* normalising constant.

The new parameter η is called the *learning rate*, following Grünwald (2012). The learning rate calibrates the influence of the prior relative to that of the data; if $\eta \in [0, 1]$ the prior is given more influence and the data less. When $\eta > 1$ the data is given more prominence, and in the extreme case when η is very large the posterior accumulates around the maximum likelihood estimate for the model. For the misspecification we encounter, we tend to be interested in the case $\eta \in [0, 1]$.

A key point emphasised by Grünwald and Ommen (2017) among others is that this is not simply model elaboration. We do not put a prior on η and learn it in the usual Bayesian way. Roughly speaking the learning rate "corrects" Bayesian inference

and should not be chosen using Bayesian inference but according to other "external" criteria, for example, a predictive loss on test data. Grünwald (2012) and Grünwald and Ommen (2017) propose the *SafeBayes* algorithm to find the optimal learning rate. In that work, the learning rate η is chosen to maximise the "sequentially randomised" Bayesian marginal log-likelihood. This can be interpreted as measure of predictive accuracy. In contrast Holmes and Walker (2017) choose the learning rate by matching the prior expected gain in information between the prior and posterior. This gain in information is quantified by the expected divergence in Fisher information.

2.3 Semi-Modular Inference

In this section, we define ***Semi-Modular Inference (SMI)***, a modification of Bayesian inference in multi-modular settings which allows us to adjust the flow of information between data and parameters in separate modules. Referring to the two-module example in fig. 2.1, we allow the data from module 1 to dominate in inference for φ without entirely discarding the joint structure provided by the full model.

Our approach is motivated by the observation in Plummer (2015) that cut-model inference is Bayesian multiple imputation: we might expect to do better at the second analysis stage of cut-model inference if we can more accurately impute missing values in the first stage. A two stage analysis resembling Cut-model analysis, but using a power posterior in the first stage delivers this. First, we update our beliefs about φ using a power likelihood, with power $\eta \in [0, 1]$ on module 2. The power-posterior improves Bayesian multiple imputation of φ at the expense of θ. In the second stage, we re-learn our beliefs on θ conditional on the learnt distribution of φ.The *degree of influence*, η, controls the contribution of the suspect module in the inference.

2.3.1 SMI distributions

Let $p(Z|\varphi)$ and $p(Y|\varphi, \theta)$ denote the observation models for the two modules. We introduce an auxiliary parameter $\tilde{\theta}$, expanding the model parameters from (φ, θ) to $(\varphi, \theta, \tilde{\theta})$.

We define the η**-smi posterior** as

$$p_{\text{smi},\eta}(\varphi, \theta, \tilde{\theta}|Z, Y) = p_{pow,\eta}(\varphi, \tilde{\theta}|Z, Y)p(\theta|Y, \varphi) \tag{2.3.1}$$

where $p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)$ is the power posterior

$$p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y) \propto p(Z|\varphi)p(Y \mid \varphi, \tilde{\theta})^{\eta}\, p(\varphi, \tilde{\theta}). \tag{2.3.2}$$

Expanding in terms of model elements (details in supplement),

$$p_{\text{smi},\eta}(\varphi,\theta,\tilde{\theta}|Z,Y) \propto p(Z \mid \varphi)\; p(Y \mid \varphi,\tilde{\theta})^{\eta}\; p(Y \mid \varphi,\theta) \\ \times \quad \frac{1}{p(Y \mid \varphi)}\; p(\varphi,\theta,\tilde{\theta}),$$

where $p(Y \mid \varphi) = \frac{1}{p(\varphi)} \int p(Y \mid \varphi,\theta)\; p(\varphi,\theta) d\theta$.

The η-smi posterior of the original parameters is just the marginal,

$$p_{\text{smi},\eta}(\varphi,\theta|Z,Y) = \int p_{\text{smi},\eta}(\varphi,\theta,\tilde{\theta}|Z,Y) d\tilde{\theta}. \tag{2.3.3}$$

The posterior distribution $p_{\text{smi},\eta}(\varphi,\theta|Z,Y)$ interpolates between the Bayesian posterior and the Cut model posterior. When $\eta = 1$ the SMI posterior is the usual Bayesian posterior (eq. (2.2.1)),

$$\begin{aligned} p_{\text{smi},1}(\varphi,\theta|Z,Y) &= p(\varphi|Z,Y) p(\theta|Y,\varphi) \\ &= p(\varphi,\theta|Z,Y), \end{aligned}$$

whereas if $\eta = 0$, the SMI posterior of φ gives back the Cut model (eq. (2.2.2)),

$$\begin{aligned} p_{\text{smi},0}(\varphi,\theta|Z,Y) &= p(\varphi|Z) p(\theta|Y,\varphi) \\ &= p_{cut}(\varphi,\theta|Z,Y). \end{aligned}$$

Semi-modular inference is defined for a fixed degree of influence $\eta \in [0,1]$. Each value of η yields a different *candidate distribution*, $p_{\text{smi},\eta}$, which we call the SMI posterior, representing posterior belief on (φ,θ). Natural questions now are, how to choose in a principled manner the "best" degree of influence, and how and why does SMI help? The answer to the latter question is in a sense straightforward. If a generalised inference scheme achieves a better score, according to some agreed external criterion, we should use it, and not otherwise. This approach is taken in Jacob et al. (2017). We answer the first question in the next section.

2.4 Analysis with (Semi-)Modular Inference

In this section we show that inference with the SMI posterior distribution at fixed η is valid (in the sense of Bissiri et al. (2016)) and give an MCMC algorithm targeting $p_{\text{smi},\eta}(\varphi,\theta|Z,Y)$. We give criteria and estimation procedures for choosing η, and comment on relations with cut-models and the power posterior.

2.4.1 Coherence of (Semi-)Modular Inference

We apply the general framework for updating belief distributions in Bissiri et al. (2016) to show that the SMI posterior - and hence also the cut posterior - are *valid* and *coherent* updates of beliefs. This framework is based on a loss function $l(\theta; y)$ connecting information in the data to the parameters of interest. The log-likelihood is one such loss, but Bissiri et al. (2016) give examples where other losses may be relevant, and Cut-models and SMI prove to be further examples.

Bissiri et al. (2016) characterise a valid belief update. They list a number of axiomatic requirements. We verify that our SMI-update satisfies these axioms in the supplement. The most demanding of these conditions, in our setting, is the coherence condition.

In *Coherent* inference we reach the same posterior distribution, whether we update belief using all data simultaneously or update belief taking the data sequentially in independent blocks. In our two-module setting, this applies in several ways: we can observe responses from different modules one after the other (e.g. first Z, and then Y); we can observe sequential data fragments within the same module (e.g. first Z_1, and then Z_2, with $Z = (Z_1, Z_2)$); any mixture of these.

The generalised update of belief in Bissiri et al. (2016) follows a decision theoretic approach. In single module notation, the generalised posterior distribution p_{l_ρ} arising from a loss $l_\rho(\theta; y)$ in a family of loss functions indexed by ρ, is the probability measure minimising a cumulative loss function $L_\rho(\nu; p_0, Y)$ over choices of probability measure ν,

$$p_{l_\rho}(\theta \mid Y) = \arg\min_{\nu} L_\rho(\nu; p_0, Y).$$

The cumulative loss function, $L_\rho(\nu; p, Y)$ balances the expected loss in the fit to data and the Kullback-Leibler divergence from the posterior to the prior distribution (generically p_0 say), and is defined by

$$L_\rho(\nu; p_0, Y) = \int l_\rho(\theta, Y)\nu(d\theta) + d_{KL}(\nu, p_0).$$

Bissiri et al. (2016) show that the optimal, valid and coherent update of beliefs from prior to posterior is given by

$$p_{l_\rho}(\theta \mid Y) \propto \exp\{-l_\rho(\theta; Y)\}p_0(\theta)$$

The canonical case in the single-module setting is the logarithmic loss function $l(\theta; Y) = -\log f(Y \mid \theta)$, which yields the conventional Bayesian update of beliefs

given by the posterior distribution. The power posterior is obtained by taking the loss function $l_{pow,\rho}(\theta; Y) = -\rho \log f(Y \mid \theta)$.

In the Supplementary material we prove that, for the model in fig. 2.1, the loss function which yields the Cut model posterior is

$$l_{cut}((\varphi, \theta); (Z, Y)) = -\log p(Z \mid \varphi) - \log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi), \tag{2.4.1}$$

and the loss function yielding the SMI posterior is

$$\begin{aligned} l_{\text{smi},\eta}((\varphi, \theta, \tilde{\theta}); (Z, Y)) = &-\log p(Z \mid \varphi) - \eta \log p(Y \mid \varphi, \tilde{\theta}) \\ &- \log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi). \end{aligned} \tag{2.4.2}$$

The $p(Y \mid \varphi)$ terms in each expression are the loss-function expression of the idea of cutting feedback from Y to φ.

We prove that both Cut-model inference and SMI are coherent when we update using the correct associated loss function given above. Detailed proofs are given in the supplementary material.

2.4.2 Targeting the Modular Posterior

Plummer (2015) and Jacob et al. (2017) explain that an MCMC algorithm that correctly targets the cut distribution cannot usually be implemented, due to the presence of the intractable normalising constant $p(Y \mid \varphi)$. A SMI sampler faces the same issue.

In order to sample the SMI-posterior in a single MCMC run we need, refering to eq. (2.3.1), a standard MCMC sampler for $p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)$ and an *exact* sampler for $p(\theta \mid Y, \varphi)$. It is straightforward to check that the transition kernel given by a two-stage update using these two conditional distributions satisfies detailed balance. A proof is given in the supplement.

Exact simulation of $p(\theta \mid Y, \varphi)$ may be impracticable. In practice, there are currently three options, nested MCMC, unbiased MCMC via couplings (Jacob et al., 2020), and tempered transitions (Plummer, 2015).

Unbiased MCMC via couplings simulates samples unbiased in expectation from the Cut-model posterior. The approach uses coupled pairs of Markov Chains sharing a common transition kernel, which almost-surely meet at some finite time $\tau \geq 1$ and stay together thereafter. The same approach applies to SMI, though we have not implemented this.

In our examples we use *nested MCMC*, described in algorithm 1: sample N_1 draws from $p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)$; for each sampled value of φ, run a sub-chain targeting

Algorithm 1 Nested MCMC for SMI posterior eq. (2.3.1)

Input: influence $\eta \in [0, 1]$, Data $(Z, Y) = \{(Z_i, Y_i)\}_{i=1}^n$; observational models $f(Z \mid \varphi)$ and $f(Y \mid \varphi, \theta)$; prior $p(\varphi, \theta)$; run-lengths N_1 and N_2.

Output: Samples $\{(\theta^{(s)}, \varphi^{(s)})\}_{s=1}^{N_1}$ distributed approximately according to the SMI posterior $p_{\text{smi},\eta}(\varphi, \theta \mid Z, Y)$ with influence parameter η.

Stage 1
for $s = 1, \ldots, N_1$ **do**
  Sample $(\tilde{\theta}^{(s)}, \varphi^{(s)}) \sim p_{\eta-pow}(\varphi, \tilde{\theta}|Z, Y)$, using any standard sampler.
end for
Let $\{\varphi^{(s)}\}_{s=1}^{N_1}$ be samples after burn-in and thinning
Stage 2
for $s = 1, \ldots, N_1$ **do**
  for $r = 1, \ldots, N_2$ **do**
    Sample $(\theta^{(s,r)}) \sim p(\theta \mid Y, \varphi^{(s)})$, using any standard sampler.
  end for
  Let $\theta^{(s)} = \theta^{(s,N_2)}$ (final state)
end for
return $\{(\theta^{(s)}, \varphi^{(s)})\}_{s=1}^{N_1}$

$p(\theta \mid \varphi, Z)$ for N_2 steps, where N_2 is large enough to avoid initialisation bias; keep only the last sampled value in this sub-chain. The resulting joint samples $(\varphi, \tilde{\theta}, \theta)$ are approximately distributed according to the SMI posterior. We typically ignore the output for $\tilde{\theta}$ as we target the marginal in eq. (2.3.3).

The validity of this algorithm relies on a double asymptotic regime in N_1 and N_2 (Jacob et al., 2020). Diagnosing convergence of the nested algorithm requires, in principle, checking convergence of the chain in the first stage $\{(\tilde{\theta}^{(s)}, \varphi^{(s)})\}_{s=1}^{N_1}$, and all sub-chains from the second stage $\{\varphi^{(s,r)}\}_{r=1}^{N_2}$; for $s = 1, \ldots, N_1$. This exhaustive diagnostic may be simplified in practice if convergence of the MCMC targeting $p(\theta \mid Y, \varphi)$ is rapid.

In our experiment we evaluated convergence using standard multiple-chain-based diagnostics. These include: visual inspection of the traces, checking $\hat{R} \simeq 1$ using split- and rank-based versions of $\hat{R}$ (Vehtari et al., 2021; Brooks and Gelman, 1998), large effective sample sizes for the bulk and tail (Vehtari, 2021), auto-correlation plots, among others (Robert and Casella, 2004; Gelman et al., 2015; Roy, 2020). For Stage 1 convergence is monitored as with any other generic sampler. For Stage 2, we start by analysing a small subset out of the N_1 available subchains, and validate that the vast majority of subchains are mixing well. Looking at the traces of subchains, we also identify a reasonable value for N_2, the step after which most subchains have reached

the equilibrium distribution.

Another validation for the nested sampler is to compare samples of θ with another independent run using a larger value $N_2' > N_2$. The distributions should coincide. For this comparison, we can use visual inspection of histograms and scatterplots, as well as checking the first few moments of both sets of samples.

Further details of the Nested MCMC are discussed in section 2.7.4.1

2.4.3 Choosing the influence parameter

We work in a $\mathcal{M}$-open setting (Bernardo and Smith, 2000), as model misspecification is our motivation for Semi-Modular Inference. Conventional Bayesian inference is optimal for prediction (for the objective defined below) under an idealised scenario of correct model specification, full availability of data, and no computational restrictions. In the $\mathcal{M}$-open setting the conventional posterior may be outperformed by another candidate distribution (Jacob et al., 2017).

The class of SMI candidate posteriors is indexed by η, so we need to give a procedure to choose a belief update operation from the set of candidate models $\mathcal{M} = \{p_{\text{smi},\eta}; \eta \in [0,1]\}$. Following (Bernardo and Smith, 2000), we should determine η on the basis of expected utility, provided some utility function.

We consider *out-of-sample predictive accuracy* of the model as our utility function. Our criterion is the Expected Log-pointwise Predictive Density (ELPD),

$$\text{elpd}_{z,y}(\eta) = \int\int p^*(z,y) \log p_{\text{smi},\eta}(z,y \mid Z,Y) dz dy, \tag{2.4.3}$$

where p^* is the distribution representing the true data-generating process and

$$p_{\text{smi},\eta}(z,y \mid Z,Y) = \int\int p(z,y \mid \varphi,\theta)\; p_{\text{smi},\eta}(\varphi,\theta \mid Y,Z)\, d\varphi\, d\theta$$

is a candidate posterior predictive distribution, indexed by η. See Vehtari et al. (2017) for further discussion of this measure. We take the value, $\eta = \eta^*$ say, which maximizes the estimated ELPD-function.

We expect this criterion to yield $\eta \simeq 1$ when there is no model misspecification and the data are informative of parameters. The ELPD is a KL-divergence, up to a constant not depending on η. If the posterior distribution of the parameters concentrates on the true values, as it may when there is no model misspecification, then $p_{\text{smi},\eta}(z,y \mid Z,Y)$ coincides with $p^*(z,y)$ when $\eta \simeq 1$ and this choice will minimise the divergence, and maximise the ELPD.

In practice, p^* is unknown, so we use cross-validation and the WAIC estimator (Vehtari et al., 2017; Watanabe, 2009) to approximate the ELPD at a grid of J values of η. This estimator is defined as

$$\widehat{\text{elpd}}_{\text{waic}}(\eta) = \widehat{\text{lpd}} - \widehat{p}_{\text{waic}} \tag{2.4.4}$$

where $\widehat{\text{lpd}}$ is the computed log pointwise predictive density, and is $\widehat{p}_{\text{waic}}$ the estimated effective number of parameters

$$\widehat{\text{lpd}}(\eta) = \sum_{i=1}^{n} \log \mathbb{E}_{\text{smi},\eta}[p(z_i, y_i \mid \varphi, \theta)] \tag{2.4.5}$$

$$\widehat{p}_{\text{waic}}(\eta) = \sum_{i=1}^{n} \text{var}_{\text{smi},\eta}[\log p(z_i, y_i \mid \varphi, \theta)] \tag{2.4.6}$$

We tried both ELPD estimators as a check, and found good agreement. Leave-One-Out Cross-Validation (LOOCV) is natural but expensive. Other estimators are available (Vehtari et al., 2017) and we are exploring these.

One consideration when using a predictive criterion in models with multiple responses is that it is not always obvious whether the criterion should be joint prediction or focus on one of the responses. We may be interested in computing the ELPD to choose η based on prediction of Y for example,

$$\text{elpd}_Y(\eta) = \int p^*(y) \log p_{\text{smi},\eta}(y \mid Z, Y) dz \tag{2.4.7}$$

$$\text{with } p_{\text{smi},\eta}(y \mid Z, Y) = \int \int p(y \mid \varphi, \theta)\; p_{\text{smi},\eta}(\varphi, \theta \mid Y, Z)\, d\varphi\, d\theta$$

The choice of criterion will ultimately depend on the application and the goal of inference. As noted by Jacob et al. (2017), this decision is clear in cases when prediction of a specific response is final goal, but less clear in cases where prediction is not the goal, but rather estimation of parameters in the model. An important challenge in the context of misspecification is that there might be a missmatch in the interpratation of parameters across models. The plan of action described by Jacob et al. (2017) can be adapted to the context of SMI. We start by identifying the module for which the meaning of the parameter of interest is in the intended specification, and maximize the predictive score for that module.

2.4.4 The computational cost of SMI

We compare the computational cost, W_{smi} say, of SMI inference (using nested MCMC at J values of η and determining η^*) to the cost, W_{bm} say, of doing regular Bayes-MCMC on the original problem.

The first stage of algorithm 1 uses the same MCMC updates as Bayesian MCMC and a similar target, so it costs about $W_{bm} = N_1$ units. The second stage uses the same θ-update as Bayes-MCMC so costs no more than $N_2 W_{bm}$. This is not quadratic in N_1 as N_2 is chosen to ensure that the stage two sampler converges but then produces just one draw from its target. The J nested MCMC runs determining η^* parallelise essentially perfectly across cores, and estimation of the WAIC and η^* is in our experience a fast output analysis, so the overall cost is about $W_{\text{smi}} = W_{bm} + N_2 W_{bm}$.

A more careful analysis considering thinning of MCMC chains (run the second half of Algorithm 1 on thinned output from the first half) shows that an overall SMI-cost $W_{\text{smi}} = KW_{bm}$ with $K \simeq 10$ should typically by achievable. This is justified in more detail in the Supplementary Material where the mixing times of the chains are taken into account.

2.4.5 SMI and the Power Likelihood

SMI is a two-stage procedure, fitting a power likelihood for φ and $\tilde{\theta}$, and then recalibrating θ conditional on φ. Does the second stage improve the inference or should we simply stick with $\tilde{\theta}$? The answer depends on the purpose of the inference. If interest lies purely in estimation of φ, we should stay with the power posterior, as the second stage has no effect on the posterior for φ. However, if we are interested in θ or in prediction of Z or Y, the best SMI candidate posterior may have a bigger ELPD and be selected over the power posterior. In Sections 2.5.1 and 2.5.2 we give examples where SMI improves prediction of new data (both sections) and mean square error (Sec. 2.5.1, on synthetic data).

2.4.6 SMI and Bayesian Multiple Imputation

In a Bayesian setting missing observations are unknown quantities inferred jointly with unknown parameters. However, in some circumstances, there is an advantage in adopting different models for imputation and analysis, a situation known as *uncongeniality* (Meng, 1994; Xie and Meng, 2016; Little and Rubin, 2002). This leads to Bayesian Multiple imputation. Since SMI reduces to the Cut-model at $\eta = 0$, we improve on multiple imputation (according to our criterion) if our procedure gives $\eta > 0$. Our analysis in Section 2.5.2 illustrates this.

SMI address a phenomenon known, in the sense of Knuiman et al. (1998)), as "dilution". This is associated with "imputation noise" from uncongenial models. This

noise typically causes the analyst to *shrink* the estimated effect of interest towards zero. Cut-model inference is multiple imputation and suffers from this problem. SMI tends to reduce imputation-noise and dilution. This is picked up in the examples below.

2.5 Examples

Here we present three reproducible examples. In the first two cases, candidate distributions made available by Semi-Modular Inference outperform conventional Bayesian inference and the Cut-model. In the last, the Cut-model, or Bayesian inference are selected. These are special cases of SMI, so the extended inference is doing its job and returning the inference schemes with the best predictive performance.

2.5.1 Simulation study: Biased data

This is a simple synthetic example in which the source of the "misspecification" is a poorly chosen prior. Suppose we have two datasets informing an unknown parameter φ. The first is a "reliable" small sample $Z = (Z_1, \ldots, Z_n)$, $Z_i \sim N(\varphi, \sigma_z^2)$, iid for $i = 1, ..., n$ distribution, with σ_z known; the second is a larger sample $Y = (Y_1, \ldots, Y_m), Y_i \sim N(\varphi + \theta, \sigma_y^2)$ iid for $i = 1, ..., m$, with σ_y known. The "bias" θ is unknown.

This model was used by Liu et al. (2009) and Jacob et al. (2017) as an example where modular/Cut-model approaches improve on Bayesian inference. We show that Semi-Modular Inference outperforms these inference schemes (which are special cases).

We choose true parameter values in such a way that each dataset offers apparent advantages to estimate φ. One dataset is unbiased but has a small sample size, n =25, whereas the second has an unknown bias but more samples, m =50, and smaller variance. Suppose the true generative parameters are φ^* =0, θ^* =1, and we know σ_z =2 and σ_y =1. We assign a constant prior for φ, while θ is subjectively assessed to have a $N(0, \sigma_\theta^2)$ prior. We are over-optimistic about the size of the bias and set σ_θ =0.5.

We calculate the SMI posterior and predictive distributions for each $\eta \in [0, 1]$. The data are synthetic, so we calculate the Squared Errors (SE) $(\hat{\varphi} - \varphi^*)^2$ and $(\hat{\theta} - \theta^*)^2$ between the posterior mean and truth, and $-elpd$ as measures of performance. In the left column of Fig. 2.2 we display these metrics for $\eta \in [0, 1]$, averaged across 1000 synthetic datasets.

As noted in Liu et al. (2009) and Jacob et al. (2017), the cut model posterior outperforms full-Bayes on average elpd (Fig. 2.2: point A<B). SMI offers new candidate distributions which (on average) outperform the cut model and full-Bayes on prediction (point C v.s. A,B), estimating φ^* (F v.s. D,E), and estimating θ (I v.s. G,H). The histogram of optimal η^*-values, for each synthetic dataset, is displayed top-right in Fig. 2.2. In 36% of the synthetic datasets, η^* is not zero (cut model) or one (full-bayes). The two histograms at lower right show the distribution of the SE differences over datasets (the mean is the MSE difference), for estimating φ^*: SMI at η^* against cut model (middle right), and cut model against full-Bayes (bottom right). SMI outperformed the cut model in 46% of datasets (smaller SE), equal in 53.8%, and beaten in 0.2%, while the cut outperformed full-bayes in 61.7% of datasets. See the supplement for further details.

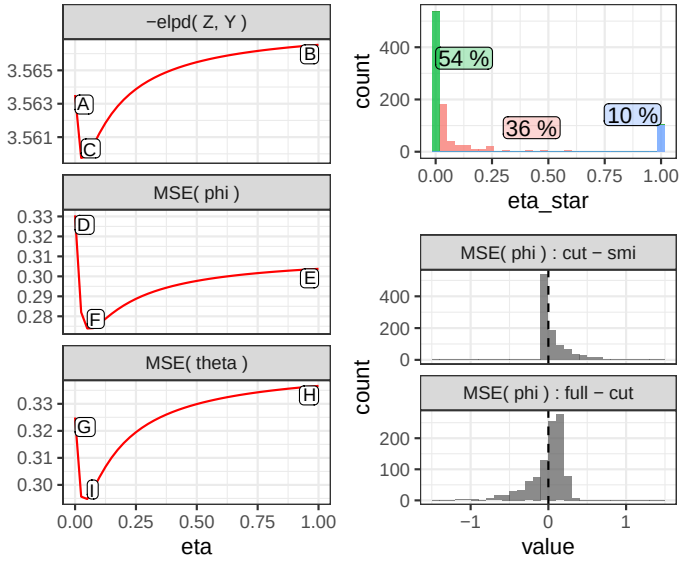


Figure 2.2: Model assessment for biased data example. Left column: average $-elpd$ and MSE for estimates under the SMI posterior with $\eta \in [0, 1]$. Top-right: Histogram of optimal η^* chosen for 1000 synthetic datasets. Bottom-Right: Histograms of differences in MSE between SMI, cut and full-bayes.

2.5.2 Agricultural data

For our second example, we apply SMI to the agricultural dataset introduced by Styring et al. (2017), and analysed using a Cut-model. The authors test for specific agricultural practices in the first urban centres in Mesopotamia. Details of the model

are given in Styring et al. (2017). We give a brief outline here with more detail in the supplement, including a graphical representation.

The observation model has a normal response, Y, and regression parameters, variances and random effects we collect together as a parameter vector ψ. It has a three-level categorical variable "manure-level" M with the same dimension as Y. Manure-level is missing data in roughly half the observations. The generative model for the missing values in M is a proportional odds model with intercept parameters (α_1, α_2) and the effect γ for a higher-level covariate, "site size". Bayesian analysis suggests that the proportional odds module is misspecified. The parameter of scientific interest is γ; the authors test for $\gamma < 0$. In our notation M plays the role of φ above, and γ plays the role of θ.

In Fig. 2.3 we plot density estimates for the posterior distribution of γ at a grid of values of η. The cut posterior is at the bottom ($\eta = 0$). We can see the effect of dilution (see Section 2.4.6) as the mean γ drifts towards zero as η approaches zero. The Bayesian posterior is at the top ($\eta = 1$).

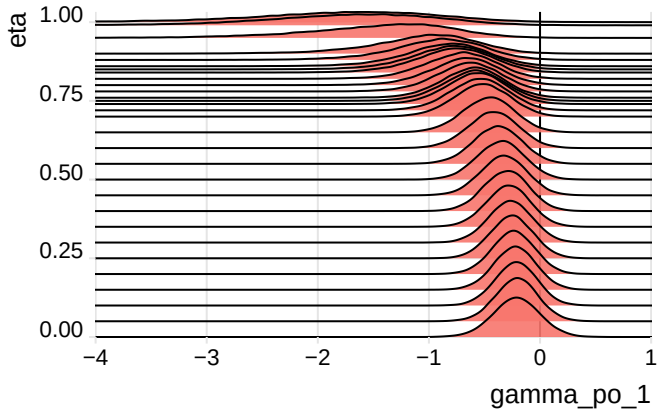


Figure 2.3: Posterior distribution of γ in the *PO* module for different values of $\eta \in [0, 1]$

The estimated ELPD, eq. (2.4.4), is plotted as a function of η in the top panel of Fig. 2.4. The η-value minimising the negative ELPD is 0.82. The evidence for $\gamma < 0$ is much stronger at the candidate posterior as it suffers less *dilution* than the cut posterior. This is quantified in the lower graph in Fig. 2.4 where we plot the posterior odds for $\gamma < 0$ against η. These odds are the Bayes Factor (BF) at $\eta = 1$, because the prior for γ is symmetric about zero. We see the evidence for $\gamma < 0$ is far stronger at the selected η-value (BF=117.11) than it is at the cut-model (BF=5.54).

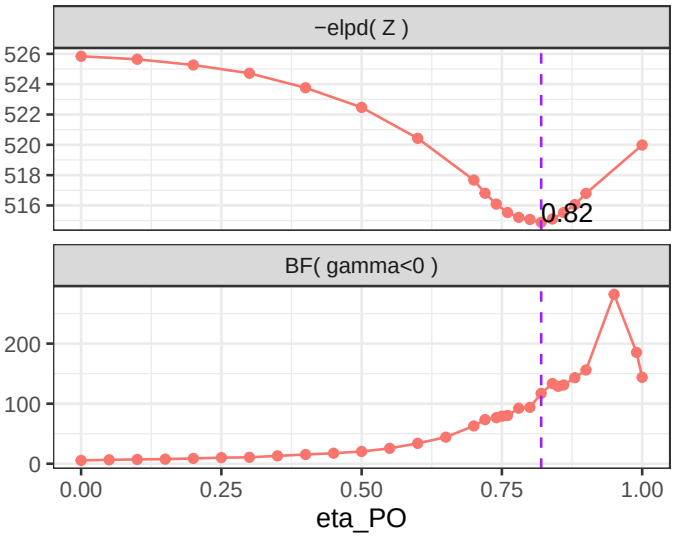


Figure 2.4: Top: Estimated ELPD as predictive criteria for choosing the value of $\eta \in [0, 1]$ The maximum is reached near to $\eta = 0.8$. Bottom: The Bayes Factor for the hypothesis $H_o : \gamma < 0$

2.5.3 Epidemiological data

In our final example, we apply SMI to an epidemiological dataset introduced by Maucort-Boulch et al. (2008), studying the correlation between human papilloma virus (HPV) prevalence and cervical cancer incidence, revisited by Plummer (2015) and Jacob et al. (2017) in the context of cut vs full Bayes models.

The model has two modules: in each population $i = 1, ..., 13$, a Poisson response for the number of cancer cases Y_i in T_i women-years of followup, and a Binomial model for the number Z_i of women infected with HPV in a sample of size N_i from the i'th population,

$$\begin{aligned} Y_i &\sim Poisson(\mu_i) \\ \mu_i &= T_i \exp(\theta_1 + \theta_2 \varphi_i) \\ Z_i &\sim Binomial(N_i, \varphi_i). \end{aligned} \tag{2.5.1}$$

In Fig. 2.5 we show the posterior distribution for the parameters θ_1, θ_2. The left panel shows posterior samples for the cut model posterior ($\eta = 0$) in black and the Bayes posterior ($\eta = 1$) in yellow. The graph agrees with an equivalent plot appearing in Jacob et al. (2017). SMI interpolates between these two distributions. The two panels at the right of Fig. 2.5 show the approximate marginal SMI posteriors for the two parameters.

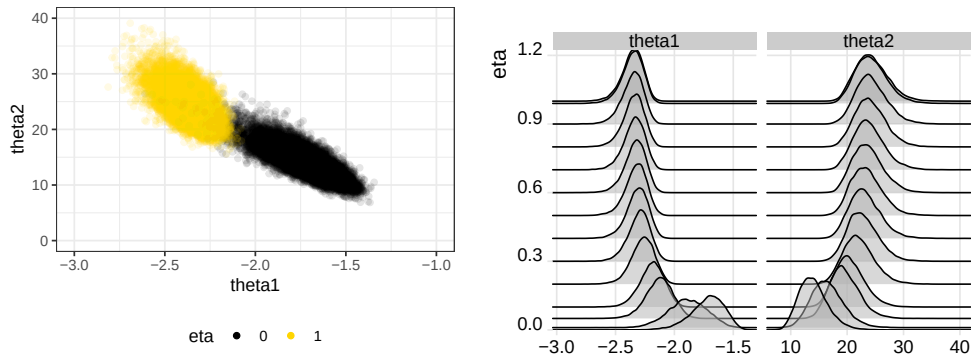


Figure 2.5: Joint SMI posterior for θ_1 and θ_2 for the HPV model using MCMC on the SMI posterior with $\eta \in [0, 1]$

In fig. 2.6 we compare the underlying log-linear relation between the HPV prevalence μ_i and cancer incidence rate φ_i for the 13 groups in the study.

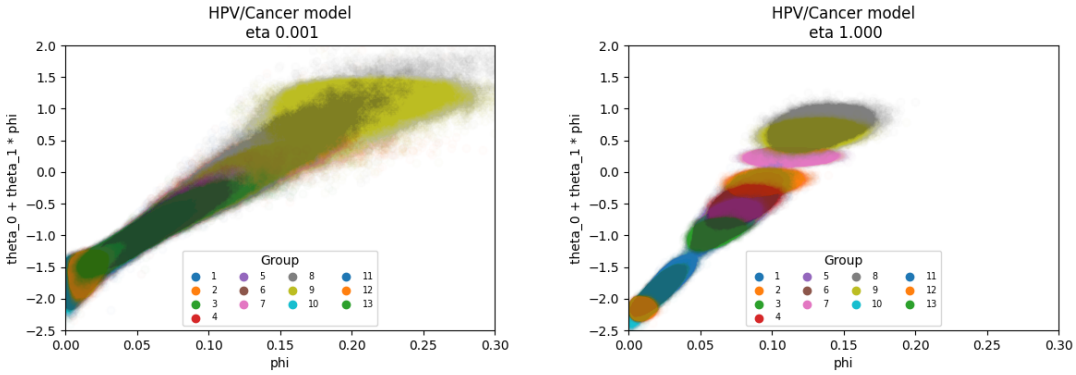


Figure 2.6: Log-linear relation between HPV prevalence and cancer incidence for the 13 groups in the study. Left: Cut Posterior ($\eta = 0$), right: Bayes Posterior ($\eta = 1$)

Lastly, we follow Jacob et al. (2017) and evaluate the predictive performance of the various SMI candidate distributions for $\eta \in [0, 1]$. Our criterion is the ELPD, estimated using the WAIC, and plotted against η in Fig. 2.7. We show the estimated ELPD for prediction of Z and Y separately. Our results expand on and support those reported by Jacob et al. (2017): For the task of predicting the Binomial data (Z), the cut model performs best (lower -elpd at $\eta = 0$ on the left panel of Fig. 2.7). This is expected as the linear relation in the Poisson module is suspected of being misspecified. Eliminating contamination improves prediction of Z. For the task of predicting the Poisson module, full Bayesian analysis performs best (lower -elpd at $\eta = 1$ in the right panel of Fig. 2.7) as the information contained in the Binomial data is reliable and helps correct the misspecified module.

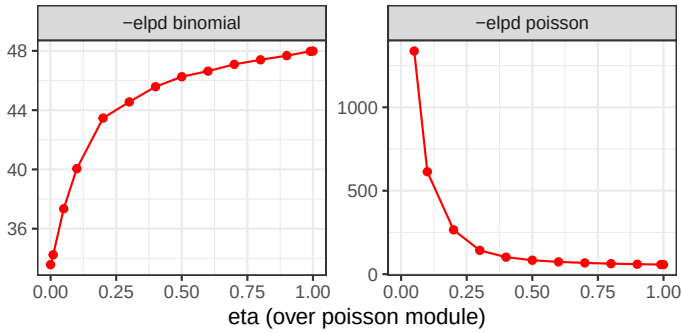


Figure 2.7: Estimated ELPD from eq. (2.4.4) as predictive criteria for selection of $\eta \in [0, 1]$ for the HPV model. The full model ($\eta = 1$) performs better for prediction of the Poisson data, while the cut model ($\eta = 0$) dominates for the Binomial.

2.6 Discussion

We have given an extension of Bayesian inference to a family of inference procedures indexed by an influence parameter. Our inference procedures bring together Bayesian inference and two qualitatively different inference schemes, power-posteriors and Modular inference/Cut-models, used to treat model misspecification. We show the new family is coherent and falls within the larger loss-based inference framework of Bissiri et al. (2016).

We gave a straightforward procedure for choosing the inference scheme according to an external ELPD criterion, which we implemented using LOOCV the WAIC. In different examples this selects Bayesian inference, Cut-model inference and interpolating candidate distributions.

When we encounter model misspecification we may consider model elaboration to improve the fit, but we may alternatively expand the inference framework.

2.7 Supplementary Material

All figures and numerical results presented in the main text and this supplement can be replicated using our R package, `aistats2020smi`, available on Github [3].

```
> # devtools::install_github("chriscarmona/aistats2020smi")
> library(aistats2020smi)
```

[3] `https://www.github.com/chriscarmona/aistats2020smi`

Notation: For the sake of compactness in our derivations, the following expressions are used henceforth,

$$
\begin{aligned}
p(Z) &= \int p(Z \mid \varphi)\, p(\varphi)\, d\varphi \\
p(Z, Y)_\theta &= \int \int p(Z \mid \varphi)\, p(Y \mid \varphi, \theta)\, p(\varphi, \theta)\, d\varphi\, d\theta \\
p(Z, Y_\eta)_{\tilde{\theta}} &= \int \int p(Z \mid \varphi)\, p(Y \mid \varphi, \tilde{\theta})^\eta\, p(\varphi, \tilde{\theta})\, d\varphi\, d\tilde{\theta} \\
p(Y_\eta)_{\tilde{\theta}} &= \int \int p(Y \mid \varphi, \tilde{\theta})^\eta\, p(\varphi, \tilde{\theta})\, d\varphi\, d\tilde{\theta} \\
p(Y, \varphi)_\theta &= \int p(Y \mid \varphi, \theta)\, p(\varphi, \theta)\, d\theta \\
p(Y \mid \varphi)_\theta &= \frac{1}{p(\varphi)}\, p(Y, \varphi)_\theta
\end{aligned}
$$

Here are a few remarks on model misspecification. Classically, misspecification is identified in goodness-of-fit checks as poor posterior predictive performance on held out data. In this paper, a model is relatively more misspecified if it has relatively worse performance in posterior predictive checks. Notice that this may be caused by a misspecified observation model, as in the HPV example in the main text, but unrepresentative prior assumptions may also lead to misspecification. This is illustrated in Section 2.5.1. The example in Section 2.5.2 arguably suffers from both forms of misspecification.

Like the power posterior (Walker and Hjort (2001); Bissiri et al. (2016); Holmes and Walker (2017); Grünwald and Ommen (2017); Miller and Dunson (2018)), the SMI procedure can alternatively be thought of as measuring misspecification. We can measure a model's goodness of fit using the η^*-values of its modules. Using the degree of influence as a measure of misfit has the advantage that it is defined on a standard scale $[0, 1]$ with $\eta^* = 1$ corresponding to no evidence for misfit and $\eta^* = 0$ indicating the module should be removed entirely when we estimate parameters in the other modules, and indicating substantial misspecification.

2.7.1 Explicit formulae for cut posterior

Recall the graphical model analysed in the main text is shown in Figure 3.1.

The **conventional (full) posterior** for this model is:

$$\begin{aligned} p(\varphi, \theta \mid Z, Y) &= p(\varphi \mid Z, Y)\ p(\theta \mid Y, \varphi) \\ &= p(Z, Y, \varphi, \theta)\ \frac{1}{p(Z, Y)_\theta} \\ &= p(Z \mid \varphi)\ p(Y \mid \varphi, \theta)\ p(\varphi, \theta)\ \frac{1}{p(Z, Y)_\theta} \end{aligned} \tag{2.7.1}$$

The **cut posterior** for this model is defined (Plummer, 2015) as:

$$\begin{aligned} p_{cut}(\varphi, \theta \mid Z, Y) &= p(\varphi \mid Z)\ p(\theta \mid Y, \varphi) \\ &= \frac{p(Z \mid \varphi)p(\varphi)}{p(Z)} \frac{p(Y \mid \varphi, \theta)\ p(\varphi, \theta)}{p(Y, \varphi)_\theta} \\ &= p(Z \mid \varphi)\ p(Y \mid \varphi, \theta)\ \frac{1}{p(Z)}\ \frac{1}{p(Y \mid \varphi)_\theta} p(\varphi, \theta) \end{aligned} \tag{2.7.2}$$

Note the relation between the cut posterior and the conventional posterior

$$\begin{aligned} p_{cut}(\varphi, \theta \mid Z, Y) &= p(Z, Y, \varphi, \theta)\ \frac{1}{p(Z)\ p(Y \mid \varphi)_\theta} \\ &= p(\varphi, \theta \mid Z, Y)\ \frac{p(Z, Y)}{p(Z)\ p(Y \mid \varphi)_\theta} \end{aligned}$$

2.7.2 Explicit formulae for SMI posterior

The η**-smi posterior** is defined as

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) &= p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)\ p(\theta \mid Y, \varphi) \\ &= \frac{p(Z \mid \varphi)\ p(Y \mid \varphi, \tilde{\theta})^\eta\ p(\varphi, \tilde{\theta})}{p(Z, Y_\eta)_{\tilde{\theta}}} \frac{p(Y \mid \varphi, \theta)\ p(\varphi, \theta)}{p(Y, \varphi)_\theta} \\ &= p(Z \mid \varphi)\ p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\ \frac{1}{p(Z, Y_\eta)_{\tilde{\theta}}\ p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \\ &\propto p(Z \mid \varphi)\ p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\ \frac{1}{p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \end{aligned} \tag{2.7.3}$$

In the penultimate step, we assume that θ and $\tilde{\theta}$ are conditionally independent given φ in the prior, so $p(\varphi, \theta, \tilde{\theta}) = p(\theta \mid \varphi)p(\tilde{\theta} \mid \varphi)p(\varphi) = p(\varphi, \theta)\ p(\varphi, \tilde{\theta})\frac{1}{p(\varphi)}$.

From here is easy to see two particular cases of the η-smi posterior: taking marginals over $\tilde{\theta}$ in SMI, the cut model $p_{smi,\eta}(\varphi, \theta \mid Z, Y) = p_{cut}(\varphi, \theta \mid Z, Y)$ when $\eta = 0$; and the conventional posterior $p_{smi,\eta}(\varphi, \theta \mid Z, Y) = p(\varphi, \theta \mid Z, Y)$ when $\eta = 1$.

2.7.3 Coherence of (Semi-)Modular Inference

Coherent update of beliefs

In Bissiri et al. (2016), the conventional update of beliefs provided by Bayes theorem is expanded, providing a generalised framework in which alternative inferential options are justified beyond the conventional posterior.

Let's use single-module notation in this Subsection 2.7.3, so we are aligned with Bissiri et al. (2016). Denote newly observe data as Y and the parameter of interest θ.

The framework established by Bissiri et al. (2016) relies on the idea that an update of beliefs must exist. This update of beliefs is performed, under a decision theory framework, by a function ψ, which turns the prior into posterior beliefs by incorporating new observed data y via a loss function $l(\theta; y)$. that is:

$$p(\theta \mid y) = \psi\{l(\theta; y), p(\theta)\}$$

Such update ψ is **Coherent** if it ensures that we preserve the posterior regardless of the order in which the data was observed. In other words, the posterior is the same whether we update our beliefs by observing all data simultaneously or by observing the data sequentially,

$$\psi\{l(\theta; x_2), \psi\{l(\theta; x_1), p(\theta)\}\} = \psi\{l(\theta; x_1) + l(\theta; x_2), p(\theta)\}$$

The authors show that an optimal, valid and coherent update of beliefs is of the form

$$p(\theta \mid y) = \psi\{l(\theta; y), p(\theta)\} = \frac{\exp\{-l(\theta; y)\}\ p(\theta)}{\int \exp\{-l(\theta; y)\}\ p(\theta)\ d\theta}$$

Coherence of the SMI posterior

Back in our multi-modular setting, the flexibility of the framework allows us to analyse the SMI posterior and the cut model posterior as valid schemes for the update of beliefs, which differ from the traditional fully-Bayesian update.

In the proofs which follow we work with SMI at an arbitrary fixed η. Since this means the results hold at $\eta = 0$, they hold for modular inference/cut models/Bayesian multiple imputation, at least for imputation in the modular setting we consider.

From Eq. 2.7.2 we can see that the loss function underlying the update of beliefs in the cut model is

$$l_{cut}((\varphi, \theta); (Z, Y)) = -\log p(Z \mid \varphi) - \log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi). \tag{2.7.4}$$

Similarly, from Eq. 2.7.3 we can derive the loss function underlying the update of beliefs in SMI the posterior, eq. (2.4.2) in the main text.

Here we prove that the SMI posterior preserves multi-modular coherence. In multi-modular settings, coherence must hold in two ways: 1) by observing responses from different modules one after the other (i.e. first Z, and then Y); and 2) by observing sequential fragments within the same module (e.g. first Z_1, and then Z_2, with $Z = (Z_1, Z_2)$).

The SMI posterior in Eq. 2.7.3 updates the belief distribution by observing the two datasets simultaneously. Under coherent inference, we can also update by observing data only from one module at a time, and still preserve the loss function in eq.2.4.2.

Say our current distribution of beliefs about $(\varphi, \theta, \tilde{\theta})$ is $p(\varphi, \theta, \tilde{\theta})$. Our updated belief by observing *only* Z would be

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) &= \psi\{l((\varphi, \theta, \tilde{\theta}); Z), p(\varphi, \theta, \tilde{\theta})\} \\ &= p(Z \mid \varphi)\ \frac{1}{p(Z)} p(\varphi, \theta, \tilde{\theta}) \\ &= p(Z \mid \varphi) \frac{1}{\int p(Z \mid \varphi)\ p(\varphi)\ d\varphi} p(\varphi, \theta, \tilde{\theta}) \end{aligned} \tag{2.7.5}$$

similarly, if we observe *only* Y our updated beliefs are

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y) &= \psi\{l((\varphi, \theta, \tilde{\theta}); Y), p(\varphi, \theta, \tilde{\theta})\} \\ &= p(Y \mid \varphi, \tilde{\theta})^{\eta} p(Y \mid \varphi, \theta)\ \frac{1}{p(Y_\eta)_{\tilde{\theta}}\ p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \\ &= p(Y \mid \varphi, \tilde{\theta})^{\eta}\ p(Y \mid \varphi, \theta) \frac{1}{\int\int p(Y \mid \varphi, \tilde{\theta})^{\eta}\ p(\varphi, \tilde{\theta})\ d\varphi\ d\tilde{\theta}} \frac{p(\varphi)}{\int p(Y \mid \varphi, \theta)\ p(\varphi, \theta)\ d\theta} p(\varphi \end{aligned} \tag{2.7.6}$$

Coherence when observing data for different modules sequentially.

First, we show that the update from prior to posterior, is equivalent to updating sequentially first Z and then Y. This is (a)=(b)+(c) in the following diagram:

$$p(\varphi, \theta, \tilde{\theta}) \xrightarrow[(b)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) \xrightarrow[(c)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y), \quad p(\varphi, \theta, \tilde{\theta}) \xrightarrow{(a)} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$$

The update (a) is Eq. 2.7.3, updating beliefs by only observing both Z and Y

$$p_{(a)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$$

The update (b) is similar to Eq. 2.7.5, updating beliefs only with data Z.

$$p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z) \tag{2.7.7}$$

$$= p(Z \mid \varphi) \, \frac{1}{p(Z)} p(\varphi, \theta, \tilde{\theta})$$

$$= p(Z \mid \varphi) \, \frac{1}{\int p(Z \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}). \tag{2.7.8}$$

The update (c) is equivalent to Eq. 2.7.6 substituting the current beliefs with $p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z)$, and updating beliefs only with data Y.

$$p_{(b)+(c)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) \propto p(Y \mid \varphi, \tilde{\theta})^{\eta} p(Y \mid \varphi, \theta) \frac{p_{(b)}(\varphi \mid Z)}{\int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta \mid Z) \, d\theta} p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z)$$

$$\propto p(Z \mid \varphi) p(Y \mid \varphi, \tilde{\theta})^{\eta} p(Y \mid \varphi, \theta) \frac{1}{P(Y \mid \varphi)_{\theta}} p(\varphi, \theta, \tilde{\theta}).$$

The equivalence $p_{(b)+(c)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$ is clear by comparing the last formula with smi posterior in Eq. 2.7.3. For the last line, we used the following identity

$$\frac{p_{(b)}(\varphi \mid Z)}{\int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta \mid Z) \, d\theta} = \frac{\int\int p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) d\theta d\tilde{\theta}}{\int\int p(Y \mid \varphi, \theta) \, p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z) \, d\theta d\tilde{\theta}}$$

$$= \frac{\int\int p(Z \mid \varphi) \, \frac{1}{\int p(Z|\varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}) d\theta d\tilde{\theta}}{\int\int p(Y \mid \varphi, \theta) \, p(Z \mid \varphi) \, \frac{1}{\int p(Z|\varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}) \, d\theta d\tilde{\theta}}$$

$$= \frac{p(\varphi)}{\int p(Y \mid \varphi, \theta) \, p(\varphi, \theta) \, d\theta}$$

$$= \frac{1}{p(Y \mid \varphi)_{\theta}}.$$

Now, we want a similar result by first observing Y and then Z, i.e. (a)=(d)+(e) in the following diagram

$$p(\varphi, \theta, \tilde{\theta}) \xrightarrow[(d)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y) \xrightarrow[(e)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) \quad \text{with} \quad p(\varphi, \theta, \tilde{\theta}) \xrightarrow{(a)} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$$

The update (a) is again given by Equation 2.7.3.

The update (d) is Eq. 2.7.6

$$p_{(d)}(\varphi, \theta, \tilde{\theta} \mid Y) \propto p(Y \mid \varphi, \tilde{\theta})^{\eta} p(Y \mid \varphi, \theta) \, \frac{1}{p(Y \mid \varphi)_{\theta}} p(\varphi, \theta, \tilde{\theta}). \tag{2.7.9}$$

The update (d)+(e) is equivalent to Eq. 2.7.5 substituting the current beliefs with $p_{(d)}(\varphi, \theta, \tilde{\theta} \mid Y)$

$$p_{(d)+(e)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) \propto p(Z \mid \varphi) \, p_{(d)}(\varphi, \theta, \tilde{\theta} \mid Y)$$

$$= p(Z \mid \varphi) \, p(Y \mid \varphi, \tilde{\theta})^{\eta} \, p(Y \mid \varphi, \theta) \, \frac{1}{p(Y \mid \varphi)_{\theta}} p(\varphi, \theta, \tilde{\theta}).$$

The equivalence $p_{(d)+(e)}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)$ is direct from comparing the last formula with Eq. 2.7.3.

Coherence when observing data partitioned from the same module.

We now verify that the SMI posterior is coherent when observing a sequential portions of the same module. Define the partitions $Z = (Z_1, Z_2)$ and $Y = (Y_1, Y_2)$.

First, we verify coherence for the partition of data Z. We want to check (b)=(b1)+(b2) in the following diagram. $p(\varphi, \theta, \tilde{\theta}) \xrightarrow[(b1)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z_1) \xrightarrow[(b2)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z)$ with (b) over the composite.

Update (b) is the same as defined above in Equation 2.7.7

Updates (b1) and (b2) are similar to Eq. 2.7.5, substituting the corresponding Z and current state of beliefs

$$
\begin{aligned}
p_{(b1)}(\varphi, \theta, \tilde{\theta} \mid Z_1) &= p(Z_1 \mid \varphi) \, \frac{1}{\int p(Z_1 \mid \varphi) \, p(\varphi) \, d\varphi} p(\varphi, \theta, \tilde{\theta}), \\
p_{(b1)+(b2)}(\varphi, \theta, \tilde{\theta} \mid Z_1, Z_2) &= p(Z_2 \mid \varphi) \, \frac{1}{\int p_{(b1)}(Z_2 \mid \varphi) \, p_{(b1)}(\varphi \mid Z_1) \, d\varphi} p_{(b1)}(\varphi, \theta, \tilde{\theta} \mid Z_1) \\
&\propto p(Z_1 \mid \varphi) \, p(Z_2 \mid \varphi) \, p(\varphi, \theta, \tilde{\theta}) \\
&= p(Z \mid \varphi) \, p(\varphi, \theta, \tilde{\theta}),
\end{aligned}
$$

clearly $p_{(b1)+(b2)}(\varphi, \theta, \tilde{\theta} \mid Z_1, Z_2) = p_{(b)}(\varphi, \theta, \tilde{\theta} \mid Z)$.

Lastly, we verify coherence for the partition of data Y. We want to check (d)=(d1)+(d2) in the following diagram.

$p(\varphi, \theta, \tilde{\theta}) \xrightarrow[(d1)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y_1) \xrightarrow[(d2)]{} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y)$ with (d) over the composite.

Update (d) is the same as defined above in Equation 2.7.9

Updates (d1) and (d2) are similar to Eq. 2.7.6, substituting the corresponding Y

and current state of beliefs

$$
\begin{aligned}
p_{(d1)}(\varphi,\theta,\tilde{\theta} \mid Y_1) =& p(Y_1 \mid \varphi,\tilde{\theta})^{\eta} p(Y_1 \mid \varphi,\theta)\ \frac{1}{p(Y_{1,\eta})_{\tilde{\theta}}\, p(Y_1 \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
\propto& p(Y_1 \mid \varphi,\tilde{\theta})^{\eta} p(Y_1 \mid \varphi,\theta)\ \frac{1}{p(Y_1 \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
=& p(Y_1 \mid \varphi,\tilde{\theta})^{\eta}\ p(Y_1 \mid \varphi,\theta) \frac{p(\varphi)}{\int p(Y_1 \mid \varphi,\theta)\ p(\varphi,\theta)\ d\theta} p(\varphi,\theta,\tilde{\theta}) \\
p_{(d1)+(d2)}(\varphi,\theta,\tilde{\theta} \mid Y_1,Y_2) \propto& p(Y_2 \mid \varphi,\tilde{\theta})^{\eta}\ p(Y_2 \mid \varphi,\theta) \frac{p_{(d1)}(\varphi)}{\int p(Y_2 \mid \varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} p_{(d1)}(\varphi,\theta,\tilde{\theta}) \\
\propto& \Big(p(Y_1 \mid \varphi,\tilde{\theta})\ p(Y_2 \mid \varphi,\tilde{\theta})\Big)^{\eta} \left(p(Y_1 \mid \varphi,\theta)\ p(Y_2 \mid \varphi,\theta)\right) \cdot \\
& \cdot \frac{p_{(d1)}(\varphi)}{\int p(Y_2 \mid \varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} \frac{1}{p(Y_1 \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) \\
\propto& p(Y \mid \varphi,\tilde{\theta})^{\eta} p(Y \mid \varphi,\theta)\ \frac{1}{p(Y \mid \varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}).
\end{aligned}
$$

from here is clear that $p_{(d1)+(d2)}(\varphi,\theta,\tilde{\theta} \mid Y_1,Y_2) = p_{(d)}(\varphi,\theta,\tilde{\theta} \mid Y)$. In the last step, we used the following identity

$$
\begin{aligned}
\frac{p_{(d1)}(\varphi)}{\int p(Y_2 \mid \varphi,\theta)\ p_{(d1)}(\varphi,\theta)\ d\theta} &= \frac{\int\int p_{(d1)}(\varphi,\theta,\tilde{\theta}) d\theta d\tilde{\theta}}{\int\int p(Y_2 \mid \varphi,\theta)\ p_{(d1)}(\varphi,\theta,\tilde{\theta})\ d\theta d\tilde{\theta}} \\
&\propto \frac{\int\int p(Y_1 \mid \varphi,\tilde{\theta})^{\eta}\ p(Y_1 \mid \varphi,\theta) \frac{1}{p(Y_1|\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta}) d\theta d\tilde{\theta}}{\int\int p(Y_2 \mid \varphi,\theta)\ p(Y_1 \mid \varphi,\tilde{\theta})^{\eta}\ p(Y_1 \mid \varphi,\theta) \frac{1}{p(Y_1|\varphi)_{\theta}} p(\varphi,\theta,\tilde{\theta})\ d\theta d\tilde{\theta}} \\
&= \frac{\frac{1}{p(\varphi)} \left(\int p(Y_1 \mid \varphi,\tilde{\theta})^{\eta} p(\varphi,\tilde{\theta}) d\tilde{\theta}\right) \left(\int p(Y_1 \mid \varphi,\theta) p(\varphi,\theta) d\theta\right)}{\frac{1}{p(\varphi)} \left(\int p(Y_1 \mid \varphi,\tilde{\theta})^{\eta} p(\varphi,\tilde{\theta})\ d\tilde{\theta}\right) \left(\int p(Y_1 \mid \varphi,\theta) p(Y_2 \mid \varphi,\theta)\ p(\varphi,\theta)\ d\theta\right)} \\
&= \frac{p(Y_1 \mid \varphi)_{\theta}}{p(Y \mid \varphi)_{\theta}}
\end{aligned}
$$

here again, we assumed that θ and $\tilde{\theta}$ are conditionally independent given φ in the prior, so $p(\varphi,\theta,\tilde{\theta}) = p(\varphi,\theta)\ p(\varphi,\tilde{\theta}) \frac{1}{p(\varphi)}$.

2.7.4 Targeting the modular posterior

Detailed balance of SMI posterior

Here we show that the η-smi posterior (and cut posterior in particular) preserves the detailed balance condition when we use the transition kernel implied by the two-stage MCMC algorithm proposed in the main text.

1. Sample $(\varphi,\tilde{\theta}) \sim p_{pow,\eta}(\varphi,\tilde{\theta} \mid Z,Y) = p(Z \mid \varphi)\ p(Y \mid \varphi,\tilde{\theta})^{\eta} \frac{1}{p(Z,Y_{\eta})_{\tilde{\theta}}} p(\varphi,\tilde{\theta})$

2. Sample $\theta \sim p(\theta \mid Y, \varphi) = p(Y \mid \varphi, \theta)\frac{1}{p(Y|\varphi)_\theta}p(\theta)$

The first step updates $(\varphi, \tilde{\theta})$ using the powered likelihood. It is not difficult to target this posterior using traditional sampling methods.

The second term updates θ exactly from its conditional posterior given data Y from module 2, and a *fixed* value φ.

The transition kernel for one iteration in this scheme is given by

$$K(\varphi', \theta', \tilde{\theta}' \mid \varphi, \theta, \tilde{\theta}) = K(\varphi', \tilde{\theta}' \mid \varphi, \tilde{\theta})p(\theta' \mid Y, \varphi'). \tag{2.7.10}$$

By construction, the first stage of the update (using the powered likelihood) is in detailed balance with the powered likelihood, i.e. satisfies

$$p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)K(\varphi', \tilde{\theta}' \mid \varphi, \tilde{\theta}) = p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')$$

From here we see that the η-smi posterior in Eq. 2.7.3 satisfies detailed balance with the transition kernel in eq. 2.7.10

$$\begin{aligned}
&p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)K(\varphi', \theta', \tilde{\theta}' \mid \varphi, \theta, \tilde{\theta}) \\
&\quad = [p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)p(\theta \mid Y, \varphi)][K(\varphi', \tilde{\theta}' \mid \varphi, \tilde{\theta})p(\theta' \mid Y, \varphi')] \\
&\quad = p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)p(\theta \mid Y, \varphi)\frac{p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')}{p_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)}p(\theta' \mid Y, \varphi') \\
&\quad = p_{pow,\eta}(\varphi', \tilde{\theta}' \mid Z, Y)p(\theta' \mid Y, \varphi')K(\varphi, \tilde{\theta} \mid \varphi', \tilde{\theta}')p(\theta \mid Y, \varphi) \\
&\quad = p_{smi,\eta}(\varphi', \theta', \tilde{\theta}' \mid Z, Y)K(\varphi, \theta, \tilde{\theta} \mid \varphi', \theta', \tilde{\theta}')
\end{aligned}$$

2.7.4.1 Further details about Nested MCMC for SMI posterior.

Our implementation of MCMC targeting the SMI posterior is a two-stage sampler described in Algorithm 1 in the main article. This is arguably the simpler approach that has been discussed in literature about MCMC targeting a modular posterior (e.g. Unbiased MCMC via couplings (Jacob et al., 2020)).

Algorithm 1 does not make specific assumptions about the class of MCMC sampler that is used at each stage. The only requirement is to check convergence in both stages. For the first step, sampling ϕ, we proceed as any traditional implementation of MCMC, running the chains until we have satisfied classical convergence test. The second step, sampling θ, we only need to guarantee that the last sample in every sub-chain are taken after we reached the equilibrium distribution.

Our examples in section 2.5 use standard MCMC at each step. The detailed implementation of the MCMC sampler can be found in the accompanying packages in R `aistats2020smi` and Python `modularbayes`, both available in GitHub [4].

For the epidemiological data, we use Hamiltonian Monte Carlo, specifically the The No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014). In both cases we performed multiple-chain-based diagnostics as described in section 2.4.2. We generated 4 parallel chains. For Stage 1 we generated $N_1 = 10,000$ samples. Traceplots for the cut posterior ($\eta = 0$) of ϕ are shown in fig. 2.8. For Stage 2, we choose $N_2 = 100$ iterations as a conservative length that guaranteed the last iteration was sampled from the equilibrium distribution. We compare these samples with a larger value $N_2' = 500$ and found agreement between the two sets. Traceplots for the cut posterior ($\eta = 0$) of θ are displayed in fig. 2.9, the top panel corresponds to $N_2 = 100$ and the bottom panel to $N_2' = 200$

For the agricultural data in Sec.2.5.2 we use random walk Metropolis-Hastings in both stages.

2.7.4 Computational cost of SMI

Our baseline is standard Bayes-MCMC on the original full model. We suppose for simplicity this was implemented using separate updates for $\theta|\varphi$ and $\varphi|\theta$, though these need not be Gibbs updates. Let $\tau_{\varphi,\theta}$ be the Integrated Autocorrelation Time (IACT) of Bayes-MCMC. If the Effective Sample Size (ESS) of the full Bayes-MCMC output is N then we must have done $T = N\tau_{\varphi,\theta}$ MCMC steps. If *one Bayes-MCMC step updating both θ and φ has unit cost* then the overall cost is $W_{bm} = T$[time]. This doesn't parallelise.

The SMI posterior is

$$p_{\eta-smi}(\varphi, \tilde{\theta}, \theta|Y, Z) = p_{pow,\eta}(\varphi, \tilde{\theta}|Y, Z)p(\theta|Y, \varphi).$$

If we use the same Bayes-MCMC updates to sample $p_{pow,\eta}(\varphi, \tilde{\theta}|Y, Z)$ then the work sampling $(\varphi, \tilde{\theta})$ at one η-value is W_{Bayes}. Thin the φ samples every $\tau_{\varphi,\theta}$ steps to get an ESS about N (this is rough, because the target changes with η, but reasonable if we allow for some tuning of the MCMC with η - we didnt need to tune in our examples).

We use the Bayes-MCMC update for $p(\theta|Y, \varphi)$ in SMI. Let τ_θ be the IACT. Typically, $\tau_\theta < \tau_{\varphi,\theta}$ (the target has lower dimension; illustrative proofs can be given in simple special cases) so take $\tau_\theta = \tau_{\varphi,\theta}$ (conservative, and note that these "side-chains"

[4] R package: `https://github.com/chriscarmona/aistats2020smi`; Python package: `https://github.com/chriscarmona/modularbayes`

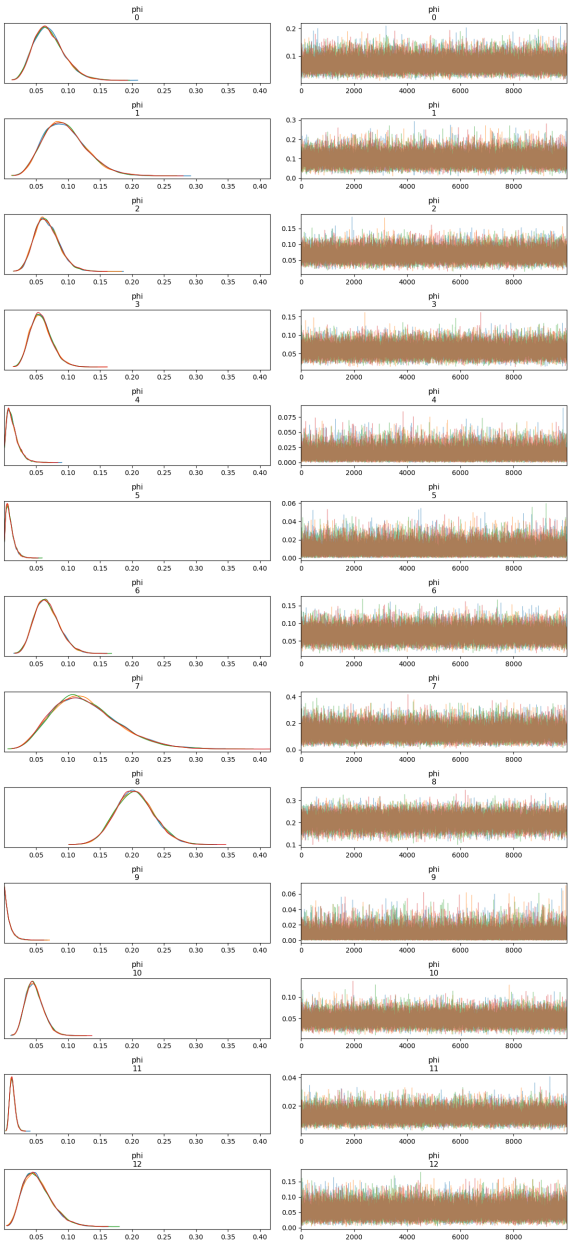
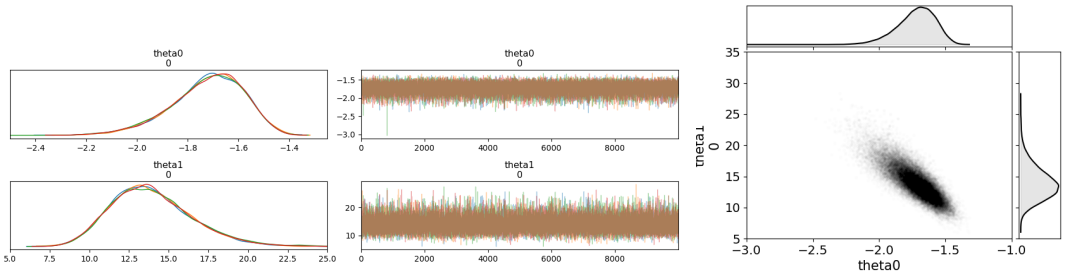


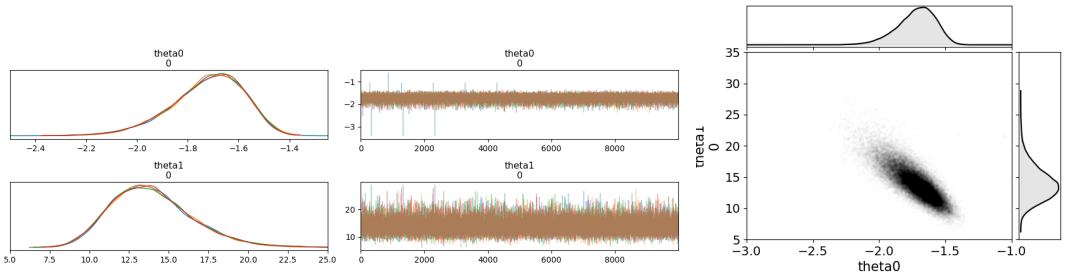
Figure 2.8: Posterior samples of φ for the Cut Posterior ($\eta = 0$) in the Epidemiological model. Obtained in Stage 1 of the nested MCMC algorithm.

targeting $p(\theta|Y, \varphi)$ can be started close to equilibrium using the θ-value output from sampling at the previous step). We run the θ-sampler to equilibrium (initialise $\theta^{(t)}$-run with $\theta^{(t-1)}$). Suppose this takes $K\tau_\theta$ steps ($K \approx 5$ is reasonable).

In the following we assume simulation at different η-values is parallelized over machines, while simulation of $\theta|\phi$ is parallelized over threads on a machine. Suppose we have M_t threads on each of M_p machines. The θ-sampling parallelizes with a small communication overhead if the time to do $K\tau_{\varphi,\theta}$ of the $\theta|\varphi$-updates is significantly



(a) $N_2 = 100$



(b) $N_2' = 500$

Figure 2.9: Posterior samples of θ for the Cut Posterior ($\eta = 0$) in the Epidemiological model. Obtained in Stage 2 of the nested MCMC algorithm. Each sample is the last value of a subchain of lenght N_2. The top panel show samples with $N_2 = 100$, while the bottom panel confirm that the distribution is preserved in an independent run with larger $N_2' = 500$.

larger than the communication time. The cost of the θ-update in the second stage of SMI is no more than the cost of one update in the original Bayes-MCMC where both θ and φ were updated, so a runtime cost for θ-updates equal one unit is conservative. It follows that $(\tilde{\theta}, \varphi, \theta)$-sampling at one η-value costs about

$$W_\eta = W_{bm}(1 + K/M_t).$$

We have to repeat this J times, sampling $p_{\eta-smi}(\varphi, \tilde{\theta}, \theta|Y, Z)$ for each of J different η-values spaced between $\eta = 0$ and $\eta = 1$ ($J \approx 20$ should be enough) using M_p machines. This larger task parallelises essentially perfectly. The total cost is

$$W_{smi} = W_{bm}(1 + K/M_t) \times J/M_p.$$

For eg if we assign resources as $M_p = J$ and $M_t = 1$ (just parallelise over η) the SMI cost is not worse than about 10 times the cost of doing Bayes-MCMC. This reflects our experience.

Finally, we compute and smooth the WAIC across the J runs at different η-values. This part is fast output-analysis. The ESS must be big enough to get stable WAIC

estimates, but WAIC is "nice" to estimate. Very roughly,

$$W_{smi} \simeq 10 W_{bm}$$

should be achievable without a great deal of work on top of the cost of implementing and running standard Bayes-MCMC.

2.7.5 Further Details on Simulation studies: Biased data

Model:

$$Z \mid \varphi \sim N(\varphi, \sigma_z^2)$$
$$Y \mid \varphi, \theta \sim N(\varphi + \theta, \sigma_y^2)$$

with σ_z^2 and σ_y^2 (and other σ's) known.

Priors:

$$\varphi \sim N(0, \sigma_\varphi^2)$$
$$\theta \sim N(0, \sigma_\theta^2)$$
$$\tilde{\theta} \sim N(0, \tilde{\sigma}_\theta^2)$$

In our example on the main text, we know the generative parameters: $(\varphi^*, \theta^*, \tilde{\theta}^*)$. We compare these *true* values with the estimates arising from the η-smi posterior, for different values of $\eta \in [0, 1]$.

SMI posterior

First, derive $p(Y \mid \varphi)_\theta$ as a function of φ

$$\begin{aligned} p(Y \mid \varphi)_\theta &= \frac{1}{p(\varphi)} \int p(Y \mid \varphi, \theta)\, p(\varphi, \theta)\, d\theta \\ &\propto \exp\{-\frac{1}{2}[\varphi^2(\frac{m}{m\sigma_\theta^2 + \sigma_y^2}) - 2\varphi(\bar{Y}\frac{m}{m\sigma_\theta^2 + \sigma_y^2})]\} \end{aligned}$$

Now we can obtain the η-smi posterior

$$\begin{aligned} p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) &= p(Z \mid \varphi)\, p(Y \mid \varphi, \tilde{\theta})^\eta p(Y \mid \varphi, \theta)\, \frac{1}{p(Z, Y_\eta)_{\tilde{\theta}}\, p(Y \mid \varphi)_\theta} p(\varphi, \theta, \tilde{\theta}) \\ &\propto \exp\{-\frac{1}{2}\, [\varphi^2(\frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) + \frac{1}{\sigma_\varphi^2} - \frac{m}{\sigma_y^2 + m\sigma_\theta^2}) \\ &\quad - 2\varphi(\frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \bar{Y}\frac{m}{\sigma_y^2 + m\sigma_\theta^2}) \\ &\quad + \theta^2(\frac{m}{\sigma_y^2} + \frac{1}{\sigma_\theta^2}) - 2\theta(\bar{Y}\frac{m}{\sigma_y^2}) \\ &\quad + \tilde{\theta}^2(\eta\frac{m}{\sigma_y^2} + \frac{1}{\tilde{\sigma}_\theta^2}) - 2\tilde{\theta}(\eta\bar{Y}\frac{m}{\sigma_y^2}) \\ &\quad + 2\varphi\theta(\frac{m}{\sigma_y^2}) + 2\varphi\tilde{\theta}(\eta\frac{m}{\sigma_y^2})]\} \end{aligned}$$

From here we see that the joint posterior distribution for $(\varphi, \theta, \tilde{\theta})$ is a multivariate normal distribution defined as:

$$p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y) = \text{Normal}(\mu, \Sigma), \tag{2.7.11}$$

with

$$\Sigma = \begin{bmatrix} \frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \frac{m}{\sigma_y^2+m\sigma_\theta^2} + \frac{1}{\sigma_\varphi^2} & \frac{m}{\sigma_y^2} & \eta\frac{m}{\sigma_y^2} \\ \frac{m}{\sigma_y^2} & \frac{m}{\sigma_y^2} + \frac{1}{\sigma_\theta^2} & 0 \\ \eta\frac{m}{\sigma_y^2} & 0 & \eta\frac{m}{\sigma_y^2} + \frac{1}{\tilde{\sigma}_\theta^2} \end{bmatrix}^{-1}$$

and

$$\mu = \Sigma \begin{bmatrix} \frac{n}{\sigma_z^2} + \frac{m}{\sigma_y^2}(1+\eta) - \bar{Y}\frac{m}{\sigma_y^2+m\sigma_\theta^2} \\ \bar{Y}\frac{m}{\sigma_y^2} \\ \eta\bar{Y}\frac{m}{\sigma_y^2} \end{bmatrix}.$$

The generative parameters described in the main text are as follows

```
> n=25 # Sample size for Z
> m=50 # Sample size for Y
> phi = 0
> theta = 1 # bias
> sigma_z = 2 # variance for Z
> sigma_y = 1 # variance for Y
```

The true bias is $\theta = 1$. Assume we have an over-optimistic view of the bias, with prior distribution centered in 0 and relatively small prior variance.

In Figure 2.10 we show posterior distributions (mean $\pm$ std. dev.) for a randomly generated dataset ($\bar{Z}$ =-0.0667; $\bar{Y}$ =1.0562) using the generative parameters described in the main text. Note that the conventional bayes ($\eta = 1$) is the worst estimation for the true parameters φ accross all posible candidates $\eta \in [0, 1]$.

```
> # Posterior for conventional bayes eta=1
> posterior = SMI_post_biased_data( Z=Z, Y=Y,
+                                   sigma_z=sigma_z,
+                                   sigma_y=sigma_y,
+                                   sigma_phi=sigma_phi,
+                                   sigma_theta=sigma_theta,
+                                   sigma_theta_tilde=sigma_theta,
+                                   eta=1 )
> posterior = mapply('rownames<-',
+                    posterior, MoreArgs=list(value=param_names))
```

```
> # posterior mean
> posterior$mean

[,1]
phi         0.3511440
theta       0.6528197
theta_tilde 0.6528197
```

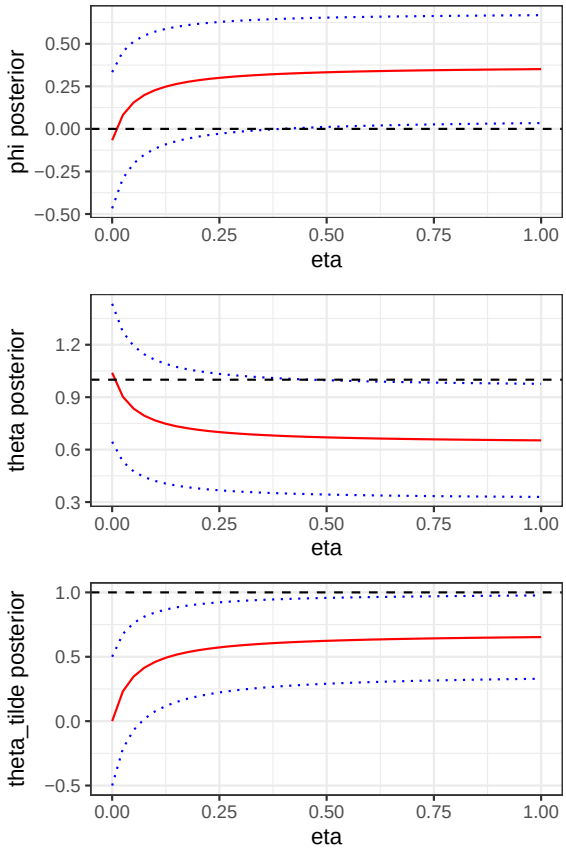


Figure 2.10: Posterior distribution of φ, θ and $\tilde{\theta}$ for a single dataset (Z,Y). A black horizontal line shows the true generative value. The posterior mean is the solid red line and we show intervals with $\pm$ one posterior std. dev. using dotted blue lines.

Mean Square Error (MSE)

From Equation 2.7.3, we can compute the Posterior Squared Error (SE) of estimates arising from the SMI posterior φ, θ and , $\tilde{\theta}$,

$$SE(\varphi) = \Sigma_{[1,1]} + (\mu_{[1]} - \varphi^*)^2$$
$$SE(\theta) = \Sigma_{[2,2]} + (\mu_{[2]} - \theta^*)^2$$
$$SE(\tilde{\theta}) = \Sigma_{[3,3]} + (\mu_{[3]} - \theta^*)^2$$

In the first simulation study of the main text we display the Mean Squared Error (MSE) for φ and θ, which is the result of averaging the posterior SE across datasets. We show that we can reach smaller MSE with values of η other than 0 and 1. To generate these plots, we produced 1000 synthetic datasets, computed MSE using Eq. 2.7.3 on each one, with a grid of values of $\eta \in [0, 1]$. The MSE lines displayed correspond to the *average* MSE across datasets, for each value of η.

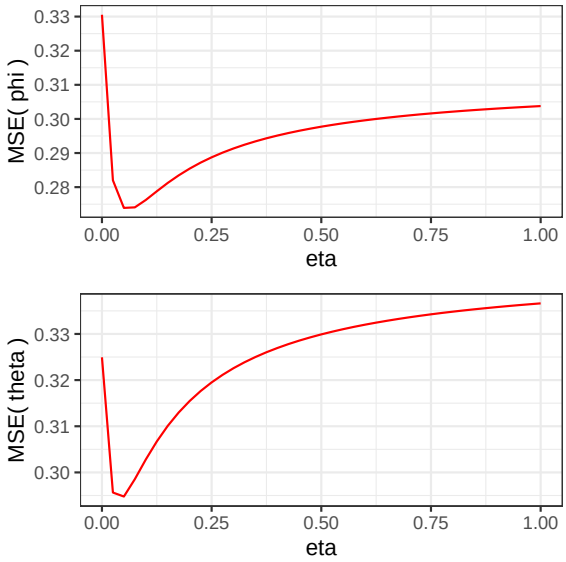


Figure 2.11: Mean Squared Error of the two main parameters under SMI posterior.

Here we also show a comparison between θ and $\tilde{\theta}$. Figure 2.12 shows that $\tilde{\theta}$ is dominated by θ in MSE. The comparison emphasises the convenience of SMI over power likelihood discussed in section 4.4 of the main text.

Expected log pointwise predictive density (elpd)

The elpd is

$$elpd = \int \int p^*(z, y) \log p_{smi,\eta}(z, y \mid Z, Y) dz dy$$

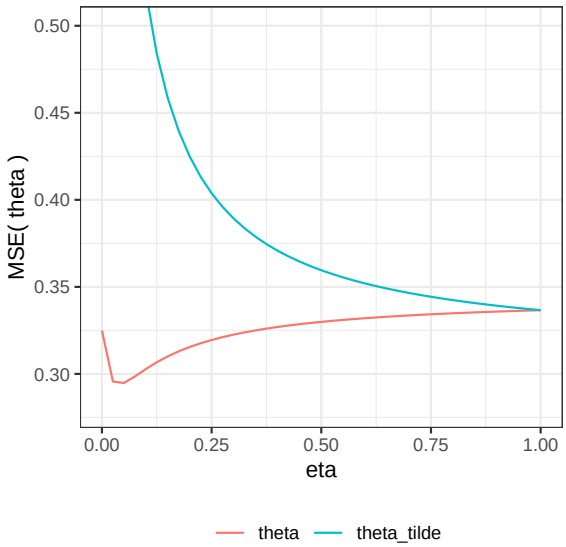


Figure 2.12: Comparison of the bias estimation under SMI (theta) vs powered likelihood (theta tilde)

where p^* is the distribution representing the true data-generating process and

$$p_{smi,\eta}(z, y \mid Z, Y) = \int \int p(z, y \mid \varphi, \theta)\; p_{smi,\eta}(\varphi, \theta \mid Y, Z)\, d\varphi\, d\theta$$

is a candidate posterior predictive distribution, indexed by η.

Let $\begin{bmatrix} a & b \\ b & c \end{bmatrix} = Cov(\varphi, \theta \mid Z, Y)^{-1}$ be the inverse of the posterior covariance matrix of (φ, θ), and $\begin{bmatrix} d \\ e \end{bmatrix} = E(\varphi, \theta \mid Z, Y)$ the posterior means.

Following straightforward Gaussian completion we can show that the joint posterior distribution for φ, θ, and new data z_0 and y_0 is:

$$\begin{aligned} p_{smi,\eta}(z_0, y_0, \varphi, \theta | Z, Y) &\propto p(z_0, y_0 | \varphi, \theta)\; p_{smi,\eta}(\varphi, \theta | Z, Y) \\ &\propto \exp\{-\frac{1}{2}\,[z_0^2(\frac{1}{\sigma_z^2}) + y_0^2(\frac{1}{\sigma_y^2}) + \varphi^2(a + \frac{1}{\sigma_z^2} + \frac{1}{\sigma_y^2}) + \theta^2(c + \frac{1}{\sigma_y^2}) + \\ &\quad - 2z_0\varphi(\frac{1}{\sigma_z^2}) - 2y_0\varphi(\frac{1}{\sigma_y^2}) - 2y_0\theta(\frac{1}{\sigma_y^2}) + 2\varphi\theta(b + \frac{1}{\sigma_y^2}) \\ &\quad - 2\varphi(ad + be) - 2\theta(bd + ce)]\} \end{aligned}$$

So we have

$$p_{smi,\eta}(z_0, y_0, \varphi, \theta \mid Z, Y) = \text{Normal}(\mu, \Sigma), \tag{2.7.12}$$

with

$$\Sigma = \begin{bmatrix} \frac{1}{\sigma_z^2} & 0 & -\frac{1}{\sigma_z^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} & -\frac{1}{\sigma_y^2} & -\frac{1}{\sigma_y^2} \\ -\frac{1}{\sigma_z^2} & -\frac{1}{\sigma_y^2} & a + \frac{1}{\sigma_z^2} + \frac{1}{\sigma_y^2} & b + \frac{1}{\sigma_y^2} \\ 0 & -\frac{1}{\sigma_y^2} & b + \frac{1}{\sigma_y^2} & c + \frac{1}{\sigma_y^2} \end{bmatrix}^{-1}, \text{ and } \mu = \Sigma \begin{bmatrix} 0 \\ 0 \\ ad + be \\ bd + ce \end{bmatrix}.$$

We know the true generative values φ^* and θ^*, so we can compute *elpd* using Monte Carlo samples from the true generative distribution p^* and evaluate this values in the log-density of the bivariate normal $(z_0, y_0|Z, Y)$ from Equation 2.7.12.

In Figure 2.13 we show the Monte Carlo estimation of the *elpd*. We select the optimal η as the value that maximise the *elpd*. To generate this plot, we produced 1000 synthetic datasets, computed *elpd* on each one (using Monte Carlo), with a grid of values of $\eta \in [0, 1]$. The *elpd* line correspond to the *average* elpd across datasets, for each value of η.

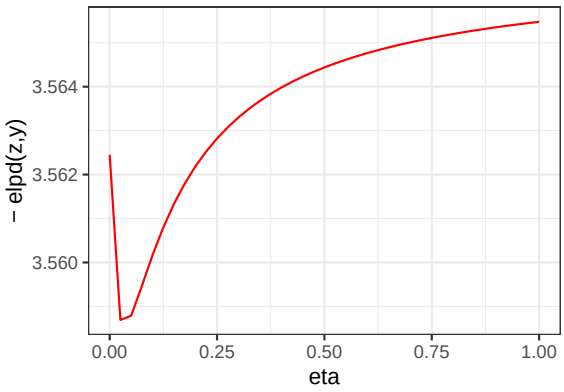


Figure 2.13: ELPD under SMI posterior.

2.7.6 Further Details on Simulation studies: Agricultural data

The aim of the study, described in detail in Styring et al. (2017), is to provide statistical evidence about a specific agricultural practice of the first urban centres in northern Mesopotamia.

The hypothesis is that increased agricultural production to support growing urban populations was achieved by cultivation of larger areas of land, entailing lower manure/midden inputs per unit area. This practice is known as *extensification*.

Our contribution goes into extending the methods used to perform Bayesian analysis in this adverse scenario of model misspecification and copious missing data.

Data

The data consists of measurements of nitrogen and carbon isotopes for a collection of crop remains. There are two datasets: archaeological and modern, which we will denote by $\mathcal{A}$ and $\mathcal{M}$, respectively. First, the *Archaeological* dataset $\mathcal{A}$, consists of data gathered from excavations of antique crop sites in the region of Mesopotamia. Second, the *Modern* dataset, $\mathcal{M}$, was gathered in a controlled experimental setting in recent years. Further characteristics and description of variables on each dataset can be found in the statistical supplement to Styring et al. (2017).

The model

We preserved the model stated in Styring et al. (2017). See Figure 2.14. The main goal of that study is to test the hypothesis of *extensification* in the ancient urban sites. This hypothesis can be condensed in analysing the strength of the effect of the site size S_i on the corresponding manure level M_i for the records in the archaeological data, $\{i \in \mathcal{A}\}$. The more negative the estimated effect, the stronger the evidence supporting the extensification hypothesis.

However, there is no available information about the Manuring levels in the ancient dataset. This is addressed by using a *Data augmentation* perspective, and consider that all the corresponding values of the manure level in the Archaeological data are missing.

The model consists of a two-module model, which integrates archaeological and calibration data so the missing manuring levels can be inferred. The first module is a Proportional Odds model (PO), with the missing Manure Levels in the archaeological data as the ordinal response ($M_i; i \in \mathcal{A}$). The second module consists of a linear Gaussian model (HM), applicable to both datasets, with the Nitrogen level of the crops (Z) as the response, and Manure levels as one of the predictors. The graphical representation of the model is depicted in Figure 2.14

The model can be studied from a multi-modular perspective. Indeed, the supplement of Styring et al. (2017) performs Bayesian Multiple Imputation (BMI) to impute the missing manure levels with parameters learnt from the HM module, and therefore *cutting* the influence from the PO module into this imputation. The parameters of the PO module are then inferred conditional on the imputed values and other information in the archaeological data.

In Figure 2.15 we draw a simplified version of the complete model to clarify how the SMI framework can be applied in this setting. The mapping of variables between

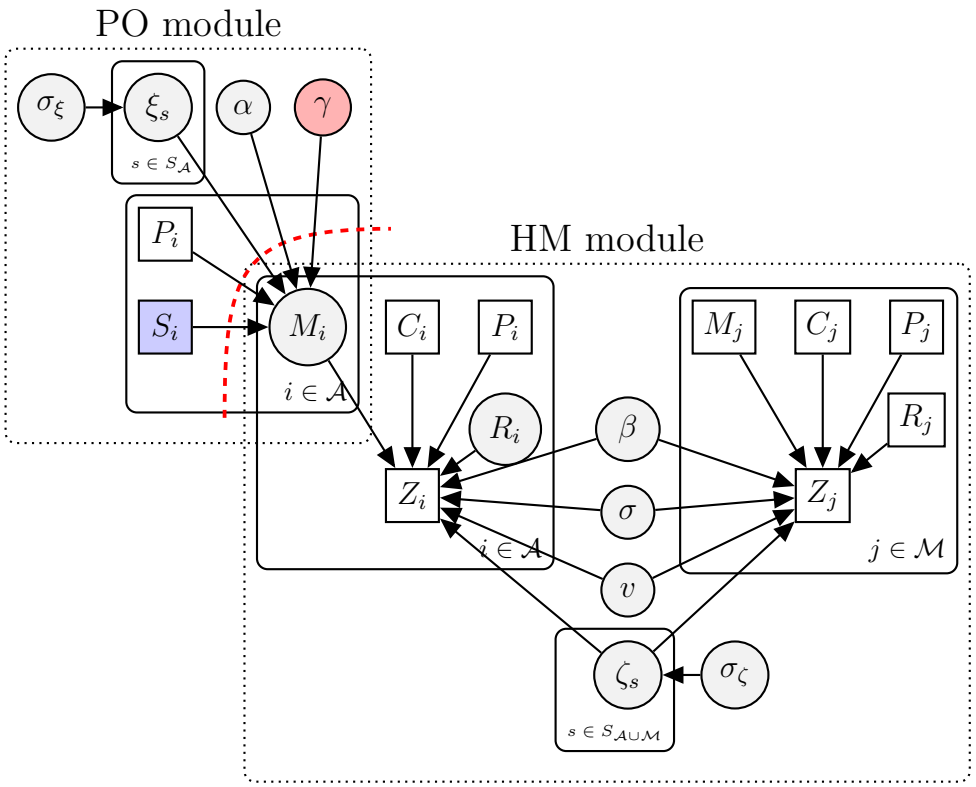


Figure 2.14: Graphical representation of the model for the agricultural data. Squares denote observable variables and circles denote unknown quantities (parameters and missing data).The main interest of the study is on the parameter γ (red circle), effect of size S_i (blue square) on Manure level M_i. The dashed line indicates the cut where SMI is applied for the imputation of missing manure.

the complete and simplified graphs is as follows: the missing manure levels will take the role of ϕ; the observed data in the HM module is Z, the parameters in the PO module will be θ, and the rest of archaeological data in the PO module is Y.

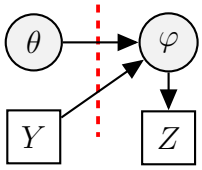


Figure 2.15: Simplified representation of the model for the agricultural data.

The simplified version resembles our graphical model in Figure 3.1. From here it is clear how the Bayesian imputation approach taken by Styring et al. (2017) is equivalent to a cut model: first learn ϕ/manure-level from the HM module, and then

learning θ (ie γ) conditional on ϕ and Y. This scheme yields our known cut posterior from Eq. 2.7.2.

$$P_{cut}(\varphi, \theta \mid Z, Y) = P(\varphi \mid Z)\ P(\theta \mid Y, \varphi)$$

We extend the Bayesian imputation approach in Styring et al. (2017) and apply Semi-Modular Inference to their setting.

$$P_{smi,\eta}(\varphi, \theta \mid Z, Y) = \int P_{pow,\eta}(\varphi, \tilde{\theta} \mid Z, Y)\ P(\theta \mid Y, \varphi) d\tilde{\theta}$$

SMI allows us to expand the space of candidate posteriors in a way that we can control smoothly, rather that eliminating the influence of the PO module in the imputation of the missing Manure Levels.

In Figure 2.3 we display the collection of candidate distributions spanned by SMI by considering a grid of values for $\eta \in [0, 1]$. The distribution at η =0 is comparable to the Bayesian Imputation approach in Styring et al. (2017).

To choose the optimal value, η^* say, we maximise the Expected Log Predictive Density or *ELPD* (Vehtari et al., 2017) to predict a new value of the response in the HM module ($Z_i; i \in \mathcal{A}$ in the diagram). In Figure 2.16 we show the negative ELPD as a function of η. The optimal value is reached at η^* =0.82.

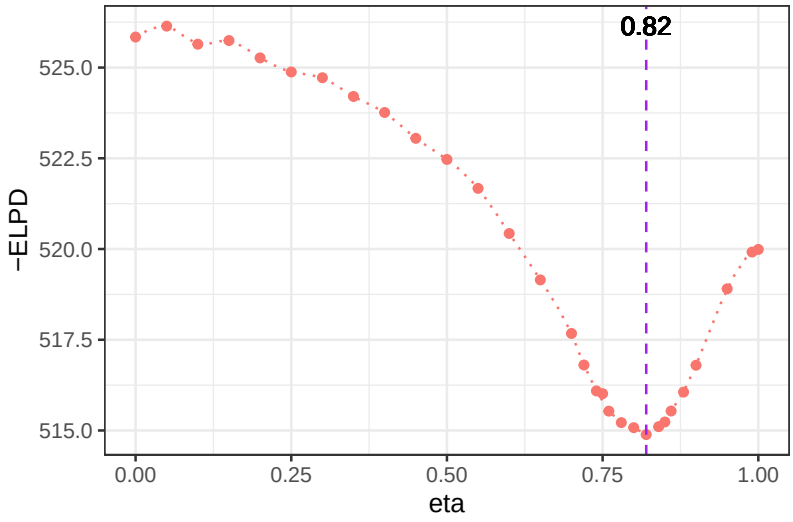


Figure 2.16: Choosing the best SMI posterior candidate by choosing the η-value maximizing the ELPD.

The elpd changes dramatically over the range $0 \leq \eta \leq 1$ and η^* is clearly distinguished from 0 or 1. This impacts downstream inference: the Bayes factor of interest changes from $BF_{smi,0}(\gamma \leq 0)$ =5.54 in the cut model, to $BF_{smi,0.82}(\gamma \leq 0)$ =117.11 in the optimal SMI posterior with η =0.82, a substantial shift in the strength of the

evidence for the *extensification* hypothesis. In Figure 2.17 we show the comparison of Bayes Factors accross values of $\eta \in [0, 1]$ for the hypothesis $\gamma \leq 0$.

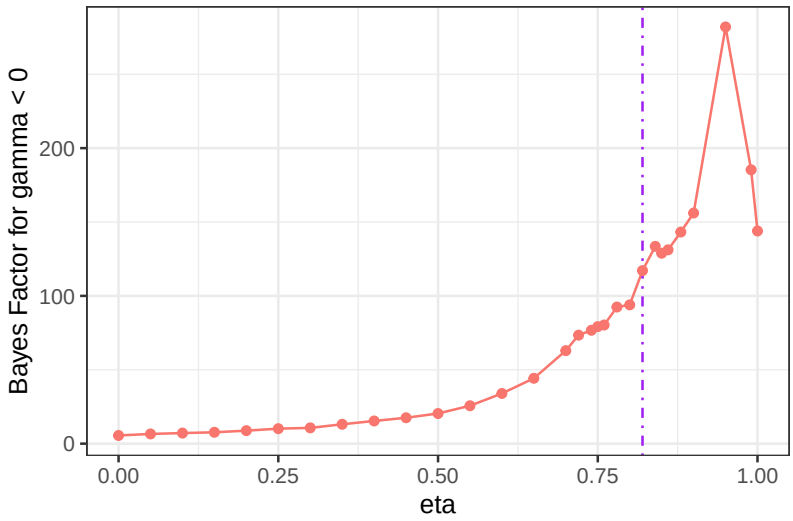


Figure 2.17: Bayes Factor for the hypothesis $\gamma \leq 0$ across all values of η.

Chapter 3

Scalable Semi-Modular Inference with Variational Meta-Posteriors

Abstract

The Cut posterior and related Semi-Modular Inference (SMI) are Generalised Bayes methods for Modular Bayesian evidence combination. Analysis is broken up over modular sub-models of the joint posterior distribution. Model-misspecification in multi-modular models can be hard to fix by model elaboration alone and the Cut posterior and SMI offer a way round this. Information entering the analysis from misspecified modules is controlled by an influence parameter η related to the learning rate. This paper contains two substantial new methods. First, we give variational methods for approximating the Cut and SMI posteriors which are adapted to the inferential goals of evidence combination. We parameterise a family of variational posteriors using a Normalizing Flow for accurate approximation and end-to-end training. Secondly, we show that analysis of models with multiple cuts is feasible using a new Variational Meta-Posterior. This approximates a family of SMI posteriors indexed by η using a single set of variational parameters.

3.1 Introduction

Evidence combination is a fundamental operation of statistical inference. When we have multiple observation models for multiple data sets, with some model parameters appearing in more than one observation model, we have a multi-modular setting in which data sets identify modules. The modules are connected in the graphical model for the joint posterior distribution of the parameters.

Large-scale multi-modular models are susceptible to model contamination, as a hazard for misspecification in each module accumulates as modules are added. If any module is significantly misspecified, it may undermine inference on the joint model (Liu et al., 2009). Model elaboration (Smith, 1986; Gelman et al., 2015) may be impractical or at least very challenging. In this setting we may consider "inference elaboration", and turn to statistically principled alternatives to Bayesian inference such as Generalised Bayes (Zhang, 2006; Grünwald and Ommen, 2017; Bissiri et al., 2016). Some modular inference frameworks allow the analyst to break up the workflow, whilst still implementing a valid belief update (Bissiri et al., 2016; Nicholls et al., 2022). This is discussed in Nicholson et al. (2022) in the broader context of modular "interoperability".

Modular Bayesian Inference (Liu et al., 2009; Plummer, 2015; Jacob et al., 2017; Carmona and Nicholls, 2020; Nicholls et al., 2022) addresses misspecification in a multi-modular setting by controlling feedback from misspecified modules (see section 3.2). Recent applications of multi-modular inference note (eg. Nicholson et al., 2022; Teh et al., 2021) and demonstrate (eg. Carmona and Nicholls, 2020; Yu et al., 2021; Styring et al., 2022) the potential benefits of partially down-weighting the influence of modules, rather than completely removing feedback. Modular Bayesian Inference is characterised by a modified posterior known as the *Cut* posterior (Plummer, 2015). This removes feedback from identified misspecified modules. *Semi-Modular* posteriors (Carmona and Nicholls, 2020) interpolate between Bayes and Cut, controlling feedback using an influence parameter $\eta \in [0, 1]$. This can be identified with the learning rate parameter in a power posterior (Walker and Hjort, 2001; Zhang, 2006; Grünwald and Ommen, 2017). The Cut and SMI posteriors are valid belief updates in the sense of (Bissiri et al., 2016), and part of a larger family of valid inference procedures (Nicholls et al., 2022).

These approaches raise computational challenges due to intractable parameter-dependent marginal factors. In the case of Cut posteriors, nested Monte Carlo samplers are given in Plummer (2015) and Liu and Goudie (2022b) and used in Carmona and Nicholls (2020) to target SMI posteriors. These samplers suffer from double asymptotics, though work well in practice on some target posterior distributions. Jacob et al. (2020) give unbiased Monte Carlo samplers for the Cut posterior and Pompe and Jacob (2021) analyse the asymptotics of the Cut posterior and give two methods to target it: a Laplace approximation and Posterior Bootstrap.

Work to date on Modular Inference focuses on models with a small number of modules and a single pre-specified cut module. We simultaneously adjust the con-

tribution of multiple modules. This allows us to take an exploratory approach and "discover" the misspecified modules. Dealing with multiple cuts and a vector of *influence parameters* is challenging, as each additional "cut" increases the dimension of η and the space of candidate posteriors. One natural approach for selecting a candidate posterior is to take a grid of η-values, sample each distribution (Importance Sampling is not straightforward, as ratios of candidate posteriors are intractable) and evaluate a performance metric on each distribution. This search strategy works for a single cut, but is already inefficient and quickly becomes cumbersome when the number of cuts increases.

In this work, we give a novel variational framework for SMI posteriors which scales to handle multiple cuts. The usual Evidence Lower Bound (ELBO) training utility is intractable, due to the same parameter-dependent marginals that make MCMC sampling difficult. Moreover, the resulting approximation does not meet the original objective of having controlled feedback between modules (see section 3.3). Our solution takes a variational family with a pattern of conditional independence between *shared, extrinsic* and *module-specific, intrinsic* parameters that matches the SMI target, and uses the *stop-gradient* operator to define a modified variational objective. The resulting variational framework gives good approximation, controllable feedback and end-to-end optimisation. In parallel independent work, Yu et al. (2021) give variational methods for Cut-posteriors. Our approaches match at the Cut posterior: just as SMI interpolates Cut and Bayes, so variational SMI interpolates variational-Cut and variational-Bayes exactly.

One of the goals of SMI is to correct for model misspecification, so it is important to get a good variational fit to the SMI-posterior and not make matters worse with a poor approximation. We leverage recent work on relatively expressive variational families using Normalizing Flows (NFs) (Rezende and Mohamed, 2015; Papamakarios et al., 2021), as we get better uncertainty quantification than less expressive mean-field approximations. In particular, we take Flow-based models with universal-approximation capabilities (see Huang et al., 2018; Durkan et al., 2019; Papamakarios et al., 2021) as our default variational families. The conditional independence structure required by the SMI posterior is achieved by defining the *Conditioner* functions of the flow.

We exploit the continuity of the SMI posterior with varying η and introduce the Variational Meta-Posterior (VMP), a variational approximation to the *entire collection* of posteriors indexed by η, using a single set of parameters. We train a function that takes η as input and produces the variational parameters for the corresponding

SMI posterior. We call this function the VMP-map. The Variational Meta-Posterior is key to scalability (as illustrated in our example with 30 potential cuts in section 3.5.2).

The remaining task is to select an SMI posterior (*ie*, η) for downstream analysis. The performance metric deciding the level of influence will depend on the inferential goals. Selection criteria (Wu and Martin, 2023) developed for choosing the learning rate in the power posterior, such as matching information gain (Holmes and Walker, 2017), and predictive performance (Vehtari et al., 2017; Jacob et al., 2017; Wu and Martin, 2021) are relevant. Yu et al. (2021) leverage tractable variational distributions to compute calibrated test statistics (Nott et al., 2021) measuring evidence against Bayes and for Cut. We use the Expected Log-pointwise Predictive Density (ELPD) (Vehtari et al., 2017), which scores predictive performance. Variational methods commonly achieve predictive accuracy comparable with MCMC despite the variational approximation (Wang and Blei, 2019) so this is a happy marriage. We estimate the ELPD using the WAIC (Watanabe, 2012). Fast sampling is available for the variational posterior density and this supports ELPD-estimation for multiple cuts.

In summary, our contributions include:

- a variational framework for approximation of SMI posteriors suitable for modular Bayesian inference;
- approximation of SMI posteriors with Normalizing Flows, underlining the importance of flexible variational families;
- the Variational Meta-Posterior (VMP), a family of variational posteriors indexed by η which approximates a family of SMI posteriors using a single set of parameters;
- end-to-end training algorithms using the *stop-gradient* operator;
- variational methods for identifying misspecified modules and modulating feedback which scale to handle multiple cuts;
- illustrations of the method on real and synthetic data.

We provide code reproducing all results and figures [1].

[1] `https://github.com/chriscarmona/modularbayes`

3.2 Modular Bayesian Inference

In order to fix ideas, we illustrate multi-modular inference using the model structure displayed in fig. 3.1. This structure is already quite rich, as more complex models may sometimes be reduced to this form by grouping together sub-modules into nodes appropriately. Our methods extend straightforwardly to more complex models in a similar fashion to earlier work in this field.

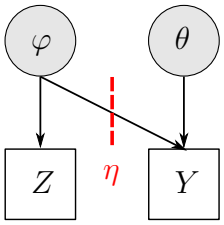


Figure 3.1: Graphical representation of a simple multi-modular model. Grey circles denote unknown quantities to be inferred, and white boxes are fixed quantities. The dashed red line indicates cut feedback from the Y-module into the Z-module. The addition of the η symbol indicates modulated feedback via SMI.

This generic setting has two modules with data $Y = (Y_1, ..., Y_n)$, $Y \in \mathcal{Y}^n$ and $Z = (Z_1, ..., Z_m)$, $Z \in \mathcal{Z}^m$ and continuous parameters $\varphi \in \Omega_\Phi$ and $\theta \in \Omega_\Theta$ of dimension p_φ and p_θ respectively. The generative models for parameters and data are $p(Z \mid \varphi)\, p(\varphi)$ and $p(Y \mid \varphi, \theta)\, p(\varphi, \theta)$. The Bayesian posterior for this model can be written

$$p(\varphi, \theta \mid Z, Y) = p(\varphi \mid Z, Y)\; p(\theta \mid Y, \varphi) \tag{3.2.1}$$

$$\propto p(\varphi, \theta, Z, Y) \tag{3.2.2}$$

where the last line is the natural form for further computation using the joint distribution

$$p(\varphi, \theta, Z, Y) = p(\varphi, \theta)\; p(Z \mid \varphi)\; p(Y \mid \varphi, \theta) \tag{3.2.3}$$

Equation (3.2.1) is given for contrast with the Cut model and SMI below. Note that,

$$p(\varphi \mid Z, Y) \propto p(\varphi)\; p(Z \mid \varphi)\; p(Y \mid \varphi) \tag{3.2.4}$$

and

$$p(\theta \mid Y, \varphi) \propto p(\theta \mid \varphi)\; \frac{p(Y \mid \varphi, \theta)}{p(Y \mid \varphi)}$$

with

$$p(Y \mid \varphi) = \int p(Y \mid \varphi, \theta)\, p(\theta \mid \varphi)\, d\theta.$$

In eq. (3.2.1) the value of φ informs θ. In eq. (3.2.4) the marginal likelihood $p(Y \mid \varphi)$ can be thought of as "feedback" of information from the Y module into the Z-module (Liu et al., 2009; Plummer, 2015; Jacob et al., 2017). Any remaining normalising constants depend only on the data Y, Z.

Cutting Feedback

Several different methods have been proposed to bring the generative models together in a joint distribution for the parameters given data. Besides Bayesian inference itself, these include Markov Melding (Goudie et al., 2019) (which focuses on settings where priors conflict across shared parameters) and Multiple Imputation (Meng, 1994), which discusses inference for "uncongenial" modules, relevant here. Nicholson et al. (2022) discusses the broader concept of "interoperability" of models in multi-modular settings. In this paper we focus on Semi-Modular Inference (SMI) defined in Carmona and Nicholls (2020) and Cut model inference (Plummer, 2015), which is a special case.

Cut-model inference has proven useful in many settings, including complex epidemic models for the Covid pandemic (Teh et al., 2021; Nicholson et al., 2022) and modular models linking isotope analysis and fertiliser use in Archaeological settings (Styring et al., 2017), pharmaco-kinetic and -dynamic models (Lunn et al., 2009) in pharmacological analysis, and health affects and air pollution (Blangiardo et al., 2011).

Suppose the generative model $p(Y \mid \varphi, \theta)\, p(\varphi, \theta)$ in the Y-module is misspecified via $p(\theta \mid \varphi)$ or $p(Y \mid \varphi, \theta)$. We hope to get a more reliable estimate of φ by "cutting" the feedback from this module into the φ-estimation. This is indicated by the dashed red line in fig. 3.1. Operationally, we drop the factor $p(Y \mid \varphi)$. Following Plummer (2015),

$$\begin{aligned} p_{\text{cut}}(\varphi, \theta \mid Y, Z) &= p(\varphi \mid Z)\, p(\theta \mid Y, \varphi) \\ &\propto \frac{p(\varphi, \theta, Y, Z)}{p(Y \mid \varphi)}. \end{aligned} \tag{3.2.5}$$

Cutting feedback leaves the Cut posterior with the intractable factor $p(Y \mid \varphi)$. Inference with a Cut-posterior is a two-stage operation which can be seen as Bayesian Multiple Imputation. In the first stage we *impute* $\varphi \sim p(\cdot \mid Z)$. This distribution

of imputed φ values is passed to the second *analysis* stage where φ are treated as randomly variable "imputed data" alongside Y, informing $\theta \sim p(\cdot \mid Y, \varphi)$. Looking ahead to SMI, this setup is shown graphically in fig. 3.2, where φ is imputed on the left (appearing in a grey circle as a parameter) and then conditioned on the right (appearing in a white square like Y). In a Cut-posterior $\eta = 0$, and the $\tilde{\theta}, Y$ elements of the graph on the left are absent.

The Cut model posterior is a "belief update", in the sense of Bissiri et al. (2016). It is a rule ψ for updating a prior measure of belief, $p_0(\varphi, \theta)$ say, using a loss $l(\varphi, \theta; Y, Z)$ connecting data and parameter (the -ve log-likelihood is a cannonical loss) to determine a posterior belief measure p_1 say. They write $p_1 = \psi(l, p_0)$. Bissiri et al. (2016) require belief updates $\psi(l, p_0)$ to be coherent: in our notation, if the data are all conditionally independent given the parameters, and $Y = (Y^{(1)}, Y^{(2)})$ and $Z = (Z^{(1)}, Z^{(2)})$ are arbitrary partitions of the data in each module into two sets, then we should arrive at the same posterior $\psi(l(\varphi, \theta; Y, Z), p_0)$ if we take all the data (Y, Z) together or if we update the prior to an intermediate posterior using $(Y^{(1)}, Z^{(1)})$ and then update that intermediate posterior using the rest of the data, $(Y^{(2)}, Z^{(2)})$, that is,

$$\psi(l(\varphi, \theta; Y, Z), p_0) = \psi(l(\varphi, \theta; Y^{(2)}, Z^{(2)}), \psi(l(\varphi, \theta; Y^{(1)}, Z^{(1)}), p_0)). \tag{3.2.6}$$

They show with some generality that a valid belief update must be a Gibbs posterior if it is to be coherent, that is,

$$\psi(l(\varphi, \theta; Y, Z), p_0) \propto \exp(-l(\varphi, \theta; Y, Z))\, p_0(\varphi, \theta).$$

Bayesian inference is coherent because the corresponding loss $l_{\text{bayes}} = -\log(p(Y \mid \varphi, \theta)) - \log(p(Z \mid \varphi))$ is additive for independent data. Carmona and Nicholls (2020) show that the belief update determined by the Cut-model posterior is coherent and Nicholls et al. (2022) show it is valid. This is surprising, as the loss $l_{\text{cut}} = l_{\text{bayes}} + \log(p(Y \mid \varphi))$ is not simply additive. This holds because the "prior" appearing in the marginal $p(Y^{(2)} \mid Y^{(1)}, \varphi)$ in the second belief update is the posterior from the first stage and not $p_0(\theta \mid \varphi)$.

The Cut posterior can also be characterised *via* a constrained optimisation (Yu et al., 2021). Consider the class of all joint densities,

$$\mathcal{F}_{\text{cut}} = \{q(\varphi, \theta) : q(\varphi) = p(\varphi \mid Z)\},$$

for which the φ-marginal $q(\varphi)$ equals $p(\varphi \mid Z)$. Densities in $\mathcal{F}_{\text{cut}}$ are candidate Cut posteriors. Yu et al. (2021) show that, among densities in $\mathcal{F}$, the Cut posterior

in eq. (3.2.5) is the best approximation to the Bayes posterior as measured by KL divergence, that is,

$$p_{\text{cut}}(\varphi, \theta \mid Y, Z) = \arg\min_{q \in \mathcal{F}_{\text{cut}}} D_{KL}(q(\varphi, \theta) \parallel p(\varphi, \theta \mid Y, Z)). \tag{3.2.7}$$

They use this characterisation to motivate a framework for variational approximation of the Cut posterior. Our motivation for variational SMI starts from an equivalent characterisation of SMI.

Statistical inference for the Cut posterior is challenging due to the marginal likelihood factor $p(Y \mid \varphi)$. Several approaches have been suggested. Plummer (2015) gives a nested MCMC scheme: run MCMC targeting $p(\varphi \mid Z)$; for each sampled φ a separate MCMC run targets $p(\theta \mid Y, \varphi)$; this yields $(\varphi, \theta) \sim p_{\text{cut}}$, at least approximately. Nested MCMC for Cut models suffers from double asymptotics but is adequate in some cases (Styring et al., 2017; Teh et al., 2021; Moss and Rousseau, 2022). A recent nested MCMC variant (Liu and Goudie, 2022b) shows efficiency gains for high dimensional targets. An exact unbiased variant of MCMC based on coalescing coupled chains (Jacob et al., 2020) removes the double asymptotics of the nested sampler.

3.2.1 Semi-Modular Inference

Semi-Modular Inference (SMI) (Carmona and Nicholls, 2020) is a modification of Bayesian multi-modular inference which allows the user to adjust the flow of information between data and parameters in separate modules. Cut models stop misspecification in one module from causing bias in others. However, this often leads to variance inflation. Semi-modular posteriors determine a family of candidate posterior distributions indexed by an influence parameter $\eta \in [0, 1]$. They interpolate between the Cut and Bayesian posteriors, expanding the space of candidate distributions and including Bayesian inference and Cut-model inference as special cases.

3.2.1.1 Modulating feedback from data modules

Cut models and SMI are typically presented using models like fig. 3.1 and cutting or modulating feedback from the Y, φ, θ module into the Z, φ module. However, some effective applications of Cut models and SMI cut or modulate feedback from modules that have no data (Jacob et al., 2017; Styring et al., 2017; Carmona and Nicholls, 2020; Yu et al., 2021; Styring et al., 2022). We return to this below and in section 3.7.1.

The **SMI posterior** for the cut in fig. 3.1 is defined as

$$p_{\text{smi},\eta}(\varphi, \theta, \tilde{\theta} \mid Y, Z) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z) p(\theta \mid \varphi, Y) \tag{3.2.8}$$

where $p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Z, Y)$ is the power posterior

$$p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Z, Y) \propto p_{\text{pow},\eta}(\varphi, \tilde{\theta}, Z, Y) \tag{3.2.9}$$

with

$$p_{\text{pow},\eta}(\varphi, \tilde{\theta}, Z, Y) = p(Z \mid \varphi) p(Y \mid \varphi, \tilde{\theta})^{\eta} \, p(\varphi, \tilde{\theta}). \tag{3.2.10}$$

Taking $\eta = 1$ in the η-smi posterior and integrating over $\tilde{\theta}$ gives the conventional posterior in eq. (3.2.1) so that $p_{smi,1}(\varphi, \theta \mid Y, Z) = p(\varphi, \theta \mid Y, Z)$, while $\eta = 0$ gives the Cut posterior in eq. (3.2.5), with $p_{smi,0}(\varphi, \theta \mid Y, Z) = p_{\text{cut}}(\varphi, \theta \mid Y, Z)$.

The SMI-posterior in eq. (3.2.8) is motivated in a similar way to the Cut-posterior. The extra degree of freedom η in the power posterior $p_{\text{pow},\eta}$ down-weights the feedback from the Y-module on φ. It is chosen to give the best possible imputation of φ in the first phase of the inference. The parameters $\tilde{\theta}$ can be thought of as auxiliary parameters introduced for the purpose of imputing φ. This two-stage process is represented in fig. 3.2. As for the Cut-posterior, φ-values from the imputation stage are treated as "imputed data" in the second stage, so they appear as random variables (in a grey circle) on the left, and as conditioned data (in a white square) on the right.

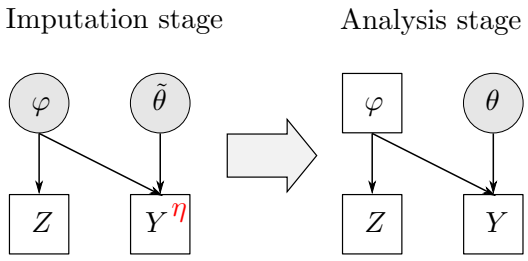


Figure 3.2: Graphical representation of the implicit two-stage inference process in Semi-Modular Inference. Grey circles denote unknown quantities to be infered, and white boxes are fixed quantities.

In sample-based inference for SMI, variants of nested MCMC (Plummer, 2015) which target $p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)$ and then sample $p(\theta \mid Y, \varphi)$ for each sampled φ have the same strengths and weaknesses as they do for the Cut posterior. Efficiency considerations are discussed in Carmona and Nicholls (2020).

SMI can be characterised in the same way as the Cut model in eq. (3.2.7). Consider the class of joint densities,

$$\mathcal{F}_{\text{smi},\eta} = \{q(\varphi, \theta, \tilde{\theta}) : q(\varphi, \tilde{\theta}) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Z, Y)\},$$

in which the marginal $q(\varphi, \tilde{\theta})$ equals the power posterior $p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Z, Y)$. Densities in $\mathcal{F}_{\text{smi},\eta}$ are candidate SMI posteriors. At $\eta = 1$ this is a duplicated Bayes posterior,

$$p_{\text{smi},\eta=1}(\varphi, \theta, \tilde{\theta} \mid Y, Z) = p(\varphi, \tilde{\theta} \mid Y, Z)p(\theta \mid Y, \varphi),$$

in which both $p(\varphi, \tilde{\theta} \mid Y, Z)$ and $\int p_{\text{smi},1}(\varphi, \theta, \tilde{\theta} \mid Y, Z)\, d\tilde{\theta}$ equal $p(\varphi, \theta \mid Y, Z)$ in eq. (3.2.1).

Proposition 1. *The SMI posterior in eq.* (3.2.8) *minimises the following KL-divergence over distributions in* $\mathcal{F}_{smi,\eta}$*,*

$$p_{smi,\eta}(\varphi, \theta, \tilde{\theta} \mid Y, Z) = \arg \min_{q \in \mathcal{F}_{smi,\eta}} D_{KL}(q(\varphi, \theta, \tilde{\theta}) \parallel p(\varphi, \theta, \tilde{\theta} \mid Y, Z)). \tag{3.2.11}$$

Proof. The following is similar to the proof of the corresponding result for the Cut model in Yu et al. (2021). For $q \in \mathcal{F}_{\text{smi},\eta}$, we have

$$q(\varphi, \theta, \tilde{\theta}) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)q(\theta \mid \varphi, \tilde{\theta}),$$

so it is sufficient to show that the KL divergence to the posterior is minimised by $q(\theta \mid \varphi, \tilde{\theta}) = p(\theta \mid Y, \varphi)$ (as that gives $q = p_{\text{smi},\eta}$). We have,

$$\begin{aligned} D_{KL}(q(\varphi, \tilde{\theta})\, q(\theta \mid \varphi, \tilde{\theta}) \parallel p(\varphi, \theta, \tilde{\theta} \mid Y, Z)) &= D_{KL}(q(\varphi, \tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)) \\ &\quad + \quad E_{q(\varphi,\tilde{\theta})}[D_{KL}(q(\theta \mid \varphi, \tilde{\theta}) \parallel p(\theta \mid Y, \varphi))] \\ &= E_{p_{\text{pow},\eta}}[D_{KL}(q(\theta \mid \varphi, \tilde{\theta}) \parallel p(\theta \mid Y, \varphi))], \end{aligned}$$

and the argument of the expectation is non-negative and zero when $q(\theta \mid \varphi, \tilde{\theta}) = p(\theta \mid Y, \varphi), \varphi \in \Omega_\varphi, \theta, \tilde{\theta} \in \Omega_\theta$, so $q = p_{\text{smi},\eta}$ minimises the original target. □

Modulating prior feedback

If a Cut is applied to a *prior density* $p(\theta|\varphi)$, as in Liu et al. (2009); Jacob et al. (2017); Styring et al. (2017) and we simply remove the prior factor at the imputation stage then all that remains in the imputation posterior distribution is the base measure. A detailed example is given in section 3.5.2. The "imputation prior" has been replaced with a constant, and this may be inappropriate in some settings. However, we are free to choose the imputation prior and we should use this freedom, as Moss and

Rousseau (2022) illustrate. Here we outline how this is done in SMI. See section 3.7.1 for detail.

Consider the generative model $\varphi \sim p(\cdot),\ \theta \sim p(\cdot \mid \varphi)$ and $Y_i \sim p(\cdot \mid \varphi, \theta),\ i = 1, ..., n$. This model is shown in the leftmost graph in fig. 3.10 in section 3.7.1. The posterior is

$$p(\varphi, \theta \mid Y) \propto p(Y \mid \varphi, \theta)p(\varphi)p(\theta \mid \varphi).$$

The SMI-posterior is

$$p_{\text{smi},\eta}(\varphi, \theta, \tilde{\theta} \mid Y) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y)p(\theta \mid Y, \varphi), \tag{3.2.12}$$

where now

$$p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y) \propto p(\varphi)p_\eta(\tilde{\theta} \mid \varphi)p(Y \mid \varphi, \tilde{\theta}).$$

The imputation prior $p_\eta(\tilde{\theta} \mid \varphi)$ must satisfy $p_{\eta=1}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)$. Like the Bayes prior $p(\theta \mid \varphi)$, the "Cut prior", $p_{\eta=0}(\tilde{\theta} \mid \varphi) = \tilde{p}(\tilde{\theta})$ say, is a modelling choice. Typically $p(\theta \mid \varphi)$ is a Subjective Bayes prior elicited from physical considerations, but is misspecified, and $\tilde{p}(\tilde{\theta})$ is a non-informative Objective Bayes prior.

This SMI-posterior belief update which cuts feedback in a prior is order coherent in the sense of Bissiri et al. (2016) and Nicholls et al. (2022).

Proposition 3. *The SMI posterior in eq.* (3.2.12) *with cut prior feedback is an order coherent belief update.*

Proof. This follows from the corresponding result for the Cut posterior which is given in proposition 2 in section 3.7.1. □

Taking a *normalised* family $p_\eta(\tilde{\theta} \mid \varphi),\ \eta \in [0, 1]$ of interpolating priors ensures that the marginal prior for $p(\varphi)$ in the imputation doesn't depend on η. An unnormalised family such as $p_\eta(\varphi, \tilde{\theta}) \propto \tilde{p}(\tilde{\theta})^{1-\eta}p(\tilde{\theta} \mid \varphi)^\eta$ has all the desired interpolating properties, but the marginal $p(\varphi)$ in the imputation stage will then depend on η. In some settings (for example when working with normal priors with fixed variance) the two prior parameterisations may be equivalent as η scales the variance.

3.3 Variational Modular Inference

We define a variational approximation for modular posteriors based on the reparametrisation approach. Our strategy has an end-to-end training implementation which avoids two-stage procedures, but converges to the same solution.

3.3.1 Variational Inference and Normalizing Flows

Applications of Variational Inference (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017) were initially focused on Mean Field Variational Inference (MFVI). This class of variational approximations is competitive with MCMC for prediction (Wang and Blei, 2019) but has disadvantages for uncertainty quantification in well specified models, making it less appealing for Bayesian inference for problems with small data sets where MCMC is feasible and well calibrated uncertainty measures are important.

Advances in variational methods have been motivated by its use in generative models in the Machine Learning literature and in particular in the context of Variational Auto-Encoders (VAEs) Kingma and Welling (2014, 2019) and applications in machine vision. Variational families based on Normalizing Flows (NFs) (Rezende and Mohamed, 2015; Papamakarios et al., 2021; Kobyzev et al., 2020) developed in that context offer generative models which are much more expressive than MFVI and give better calibrated measures of uncertainty. Adoption of NFs in applications of statistical modelling and inference, where MCMC and MFVI are the de-facto approaches, has been more limited. Stochastic Variational Inference (SVI) (Hoffman et al., 2013) and Black Box Variational Inference (BBVI) (Ranganath et al., 2014) offer efficient procedures to fit variational families which apply directly to NF parameterisations. Recent advances include new methods for evaluating convergence and adequacy of variational approximation (Yao et al., 2018; Xing et al., 2020; Agrawal et al., 2020; Dhaka et al., 2020).

3.3.2 Variational Bayes in multi-modular models

We begin by giving a standard variational approximation to the Bayes posterior for the multi-modular model. For concreteness, we use the multi-modular model in fig. 3.1. Having established our methods on this class of models, extensions to other dependence structures are straightforward, as we illustrate in section 3.5.2.

We take a parametrisation of the variational posterior in terms of a product

$$q_\beta(\varphi, \theta) = q_{\beta_1}(\varphi) q_{\beta_2}(\theta \mid \varphi), \tag{3.3.1}$$

with each factor using a disjoint subset $\beta_1 \in \Lambda_1$, $\beta_2 \in \Lambda_2$ of a set of variational parameters $\beta = (\beta_1, \beta_2)$, with $\beta \in B$ and $B = \Lambda_1 \times \Lambda_2$. Here $\Lambda_1 = \Re^{L_1}$, $\Lambda_2 = \Re^{L_2}$ and $B = \Re^{L_1+L_2}$ are typically high dimensional real spaces of variational parameters.

Our notation implies a *flow*-based approach but captures a number of other parameterisations. Let $\epsilon = (\epsilon_1, \epsilon_2)$ be a vector of continuous random variables distributed according to a *base* distribution $p(\epsilon) = p(\epsilon_1, \epsilon_2)$, with $\epsilon_1 \in \Re^{p_\varphi}$ and $\epsilon_2 \in \Re^{p_\theta}$. We can for example take $p(\epsilon_1, \epsilon_2)$ to be the φ, θ-prior. Consider a diffeomorphism, $T : \Re^{p_\varphi + p_\theta} \to \Re^{p_\varphi + p_\theta}$ defined by concatenating the two diffeomorphisms expressing φ and θ, so that

$$\begin{aligned} \varphi_{(\beta_1,\epsilon)} &= T_1(\epsilon_1; \beta_1) \\ \theta_{(\beta_2,\epsilon)} &= T_2(\epsilon_2; \beta_2, \epsilon_1) \\ T(\epsilon; \beta) &= (T_1(\epsilon_1; \beta_1), T_2(\epsilon_2; \beta_2, \epsilon_1).)\,, \end{aligned} \tag{3.3.2}$$

For flow-based densities, $T_1 : \Re^{p_\varphi} \to \Re^{p_\varphi}$ and $T_2 : \Re^{p_\theta} \to \Re^{p_\theta}$ have properties listed in section 3.7.3 (see Kobyzev et al., 2020, Sec. 3) which allow us to sample, differentiate and evaluate the densities $q_\beta(\varphi, \theta)$ defined below. However, other familiar variational families such as MFVI can be expressed using eq. (3.3.2). Note that T_2 is a *conditional transformation* that depends on ϵ_1, so it can express correlation between $\varphi_{(\beta_1,\epsilon)}$ and $\theta_{(\beta_2,\epsilon)}$ (see section 3.7.4). In a normalising flow, T_1 and T_2 are compositions of diffeomorphisms, each with their own parameters. This increases the flexibility of the transformation.

The Jacobian matrix, $J_T = \partial T / \partial \epsilon$ is block lower triangular, so its determinant is a product of determinants of $J_{T_1} = \partial T_1 / \partial \epsilon_1$ and $J_{T_2} = \partial T_2 / \partial \epsilon_2$,

$$|J_T| = |J_{T_1}|\,|J_{T_2}|$$

with no cross dependence on β_1, β_2, so that $\nabla_{\beta_2} \log |J_{T_1}| = 0$ and $\nabla_{\beta_1} \log |J_{T_2}| = 0$. The joint variational distribution produced by the flow is then

$$\begin{aligned} q_\beta(\varphi, \theta) &= p(\epsilon_1, \epsilon_2)\,|J_T|^{-1} \\ &= q_{\beta_1}(\varphi) q_{\beta_2}(\theta \mid \varphi), \end{aligned}$$

where

$$q_{\beta_1}(\varphi) = p(\epsilon_1)\,|J_{T_1}|^{-1}\,, \tag{3.3.3}$$

$$q_{\beta_2}(\theta \mid \varphi) = p(\epsilon_2 \mid \epsilon_1)\,|J_{T_2}|^{-1}\,. \tag{3.3.4}$$

We need to be able to evaluate the determinants of the Jacobians J_{T_1} and J_{T_2}. This works for a NF because the matrices are lower trianglular. However, other simpler designs such as MFVI also admit straightforward evaluation.

The optimal variational parameters minimise the KL divergence to the posterior, but will not in general be unique. Let

$$D_B^* = \min_{\beta \in B} D_{KL}(q_\beta(\varphi, \theta) \parallel p(\varphi, \theta \mid Z, Y)), \tag{3.3.5}$$

and

$$B^* = \{\beta \in B : D_{KL}(q_\beta(\varphi, \theta) \parallel p(\varphi, \theta \mid Z, Y)) = D_B^*\}, \tag{3.3.6}$$

and let $\beta^* = (\beta_1^*, \beta_2^*) \in B^*$ be a generic set of parameter values minimising the KL divergence. The definition in eq. (3.3.6) is equivalent to maximising the ELBO,

$$\text{ELBO}_{\text{bayes}} = \mathbb{E}_{(\varphi,\theta)\sim q_\beta(\varphi,\theta)}[\log p(\varphi, \theta, Z, Y) - \log q_\beta(\varphi, \theta)]. \tag{3.3.7}$$

Using the *reparametrisation trick* and expanding the joint distribution

$$\begin{aligned}
\text{ELBO}_{\text{bayes}} &= \mathbb{E}_{\epsilon\sim p(\epsilon)}[\log p(\varphi_{(\beta_1,\epsilon)}, \theta_{(\beta_2,\epsilon)}, Z, Y) - \log q_\beta(\varphi_{(\beta_1,\epsilon)}, \theta_{(\beta_2,\epsilon)})] \\
&= \mathbb{E}_{\epsilon\sim p(\epsilon)}[\log p(Z \mid \varphi_{(\beta_1,\epsilon)}) + \log p(Y \mid \varphi_{(\beta_1,\epsilon)}, \theta_{(\beta_2,\epsilon)}) + \log p(\varphi_{(\beta_1,\epsilon)}, \theta_{(\beta_2,\epsilon)}) \\
&\qquad - \log p(\epsilon) + \log |J_T|],
\end{aligned}$$

and the gradients of the ELBO with respect to the variational parameters (β_1, β_2) are

$$\begin{aligned}
\nabla_{\beta_1} \text{ELBO}_{\text{bayes}} &= \mathbb{E}_{\epsilon\sim p(\epsilon)}[\nabla_\varphi \{\log p(Z \mid \varphi) + \log p(Y \mid \varphi, \theta) + \log p(\varphi)\} \nabla_{\beta_1}\{\varphi\} \\
&\qquad + \nabla_{\beta_1} \log |J_{T_1}|], \qquad (3.3.8) \\
\nabla_{\beta_2} \text{ELBO}_{\text{bayes}} &= \mathbb{E}_{\epsilon\sim p(\epsilon)}[\nabla_\theta \{\log p(Y \mid \varphi, \theta) + \log p(\theta \mid \varphi)\} \nabla_{\beta_2}\{\theta\} \\
&\qquad + \nabla_{\beta_2} \log |J_{T_2}|]. \qquad (3.3.9)
\end{aligned}$$

These gradients are used in Stochastic Variational Inference (Hoffman et al., 2013) to obtain the optimal variational parameters $\beta^* \in B^*$ for approximation of the Bayes posterior.

3.3.3 Variational SMI

In this section we define our variational approximation to the SMI posterior. For this, we expand the variational distribution in eq. (3.3.1) to include the auxiliary parameter $\tilde{\theta}$. Again, we parametrise the variational posterior as a product,

$$q_\lambda(\varphi, \theta, \tilde{\theta}) = q_{\lambda_1}(\varphi) q_{\lambda_2}(\theta \mid \varphi) q_{\lambda_3}(\tilde{\theta} \mid \varphi) \tag{3.3.10}$$

where each factor has its own parameters, $\lambda_1 \in \Lambda_1$ and $\lambda_2, \lambda_3 \in \Lambda_2$ where Λ_1 and Λ_2 are defined above. Let $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ with $\lambda \in \Lambda$ and $\Lambda = \Lambda_1 \times \Lambda_2 \times \Lambda_2$ so that

$\Lambda = \Re^{L_1+2L_2}$. The parameters of $q_{\lambda_2}(\theta \mid \varphi)$ and $q_{\lambda_3}(\tilde{\theta} \mid \varphi)$ both match the variational Bayes parameterisation so we write $(\lambda_1, \lambda_2) \in B$ and $(\lambda_1, \lambda_3) \in B$. Let

$$\mathcal{Q} = \{q_\lambda(\varphi, \theta, \tilde{\theta}); \lambda \in \Lambda\}$$

denote the class of densities in our variational family.

3.3.3.1 The variational-SMI approximation and its properties

The purpose of SMI is to control the flow of information from the Y-module into the posterior distribution for φ. This leads us to define three basic properties that a useful variational approximation of the SMI posterior must possess:

(P1) (expresses Cut) at $\eta = 0$ the optimal variational posterior $q_{\lambda_1^*}(\varphi)$ is completely independent of the generative model for Y, with

$$\lambda_1^* \in \{\lambda_1 \in \Lambda_1 : D_{KL}(q_{\lambda_1}(\varphi) \parallel p(\varphi|Z)) = d_{\text{cut}}^*\}$$

and $d_{\text{cut}}^* = \min_{\lambda_1 \in \Lambda_1} D_{KL}(q_{\lambda_1}(\varphi) \parallel p(\varphi|Z))$.

(P2) (expresses Bayes) at $\eta = 1$ the marginal variational SMI posterior $q_{\lambda_1^*, \lambda_2^*}(\varphi, \theta)$ is equal to the variational Bayes posterior, $q_{\beta_1^*, \beta_2^*}(\varphi, \theta)$ in eq. (3.3.6), so $(\lambda_1^*, \lambda_2^*) \in B^*$;

(P3) (approximates SMI) for $\eta \in [0, 1]$, if $p_{\text{smi},\eta} \in \mathcal{Q}$ then the variational approximation is equal to the target SMI posterior, so $q_{\lambda^*} = p_{\text{smi},\eta}$.

Properties (P1-2) require q_{λ^*} to interpolate a variational approximation to the Cut-posterior (removing all feedback from Y into the variational approximation to the distribution of φ) and our original variational approximation to the Bayes posterior. We will see that a standard variational approximation to the SMI posterior based on the KL divergence between q_λ and $p_{\text{smi},\eta}$ cannot satisfy these properties. We give a variational procedure based on the loss $\mathcal{L}^{(\text{smi},\eta)}(\lambda)$ in eq. (3.3.31), and show in proposition 9 that if we take a variational approximation to $p_{\text{smi},\eta}$ minimising this loss then our variational approximation satisfies properties (P1-3).

3.3.3.2 Defining the variational family

As in section 3.3.2, our notation it set up for flow-based approximation of the SMI posterior, but captures other variational families such as MFVI. We expand the base distribution and diffeomorphism to accommodate the auxiliary $\tilde{\theta}$. The distributions

$p(\epsilon_1)$ and $p(\epsilon_2 \mid \epsilon_1)$ and the transformations T_1 and T_2 are unchanged from section 3.3.2.

Let $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)$ be the vector of random variables for our base distribution, $\epsilon \sim p(\epsilon)$, with ϵ_1 and ϵ_2 as before and $\epsilon_3 \sim \epsilon_2$ so that $\epsilon_3 \in \Re^{p_\theta}$. Consider the extended diffeomorphism T defined by the transformations,

$$\begin{aligned}
\varphi_{(\lambda_1,\epsilon)} &= T_1(\epsilon_1; \lambda_1) \\
\theta_{(\lambda_2,\epsilon)} &= T_2(\epsilon_2; \lambda_2, \epsilon_1) \\
\tilde{\theta}_{(\lambda_3,\epsilon)} &= T_2(\epsilon_3; \lambda_3, \epsilon_1) \\
T(\epsilon; \lambda) &= (T_1(\epsilon_1; \lambda_1),\ T_2(\epsilon_2; \lambda_2, \epsilon_1),\ T_2(\epsilon_3; \lambda_3, \epsilon_1))\,. \qquad (3.3.11)
\end{aligned}$$

See eq. (3.7.7) in section 3.7.4 for further details of these maps in a generic NF setting.

The Jacobian of the transformation, $J_T = \partial T / \partial \epsilon$, is block lower triangular as before, so its determinant factorises $|J_T| = |J_{T_1}(\epsilon_1)|\ |J_{T_2}(\epsilon_1, \epsilon_2)|\ |J_{T_2}(\epsilon_1, \epsilon_3)|$ where we draw attention to the different arguments in the Jacobian factors involving T_2 but omit the λ-dependence. We give more details of the transformation T in section 3.7.3.

Our variational family approximating the SMI posterior $p_{\text{smi},\eta}$ has variational parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ and conditional independence structure

$$\begin{aligned}
q_\lambda(\varphi, \theta, \tilde{\theta}) &= p(\epsilon_1, \epsilon_2, \epsilon_3)\, |J_T|^{-1} \qquad (3.3.12) \\
&= q_{\lambda_1}(\varphi) q_{\lambda_2}(\theta \mid \varphi) q_{\lambda_3}(\tilde{\theta} \mid \varphi),
\end{aligned}$$

where q_{λ_1} and q_{λ_2} are given in eqs. (3.3.3) and (3.3.4) (replacing $\beta \to \lambda$) and

$$q_{\lambda_3}(\tilde{\theta} \mid \varphi) = p(\epsilon_3 \mid \epsilon_1)\, |J_{T_2}|^{-1}\,. \qquad (3.3.13)$$

3.3.3.3 The standard variational loss does not satisfy Properties (P1-2)

A naive application of variational approximation to SMI would minimise the KL divergence to the SMI posterior at $\lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*)$ where

$$\lambda^* = \arg\min_{\lambda \in \Lambda} D_{KL}(q_\lambda(\varphi, \theta, \tilde{\theta}) \parallel p_{\text{smi},\eta}(\varphi, \theta, \tilde{\theta} \mid Z, Y)), \qquad (3.3.14)$$

(ignoring non-uniqueness for brevity). However, this presents two problems, one of principle and one of practice.

The principle of SMI is to control the flow of information from the Y-module into the posterior distribution for φ. This is lost in this setup. The KL-divergence in eq. (3.3.14) is

$$\begin{aligned}
D_{KL}(q_\lambda \parallel p_{\text{smi},\eta}) = {} & D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)) \\
& + \quad E_{\varphi \sim q_{\lambda_1}}[D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi))]. \qquad (3.3.15)
\end{aligned}$$

The first term allows controlled feedback from the Y-module, as the Y-dependence in the power posterior $p_{\text{pow},\eta}$ is controlled by η and vanishes entirely when $\eta = 0$. However, the Y-module influences $q_{\lambda_1}(\varphi)$ when $\eta = 0$ through the second term and so property (P1) cannot hold.

This variational approximation will not in general satisfy property (P2) either. In order for property (P2) to be satisfied at $\eta = 1$ we must have $(\lambda_1^*, \lambda_2^*) = (\beta_1^*, \beta_2^*)$ for some $(\beta_1^*, \beta_2^*) \in B^*$ whenever $(\lambda_1^*, \lambda_2^*, \lambda_3^*)$ satisfy eq. (3.3.14) for some $\lambda_3^* \in \Lambda_2$, that is, the marginal variational SMI distribution for (φ, θ) must coincide with one of the variational Bayes solutions. The optimal λ^* minimise eq. (3.3.15), so they maximise the ELBO,

$$\begin{aligned}\text{ELBO}_{\text{smi naive}} = &\mathbb{E}_{(\varphi,\tilde{\theta})\sim q_{\lambda_1,\lambda_3}(\varphi,\tilde{\theta})}[\log p_{\text{pow},\eta}(\varphi, \tilde{\theta}, Z, Y) - \log q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta})] \\ &+ \mathbb{E}_{(\varphi,\theta)\sim q_{\lambda_1,\lambda_2}(\varphi,\theta)}[\log p(\varphi, \theta, Y) - \log q_{\lambda_2}(\theta \mid \varphi)] \\ &- \mathbb{E}_{\varphi\sim q_{\lambda_1}(\varphi)}[\log p(Y, \varphi)].\end{aligned} \tag{3.3.16}$$

Since the λ^* parameters solve $\nabla_\lambda \text{ELBO}_{\text{smi naive}} = 0$ and the β^* parameters solve $\nabla_\beta \text{ELBO}_{\text{bayes}} = 0$, a necessary condition for (P2) is that $\nabla_\beta \text{ELBO}_{\text{bayes}}|_{\beta=(\lambda_1^*,\lambda_2^*)} = 0$ (ie eqs. (3.3.8) and (3.3.9) at $\beta = (\lambda_1^*, \lambda_2^*)$) at $\eta = 1$. However, using the reparameterisation trick, the $\text{ELBO}_{\text{smi naive}}$-gradients can be written

$$\begin{aligned}\nabla_{\lambda_1} \text{ELBO}_{\text{smi naive}} = \mathbb{E}_{\epsilon\sim p(\epsilon)}[&\nabla_\varphi \left\{\log p(Z \mid \varphi) + \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\varphi)\right\} \nabla_{\lambda_1}\{\varphi\} \\ &+ \nabla_{\lambda_1} \log |J_{T_1}(\epsilon_1)| \\ &+ \nabla_\varphi \left\{\log p(Y \mid \varphi, \theta) + \log p(Y \mid \varphi)\right\} \nabla_{\lambda_1}\{\varphi\}],\end{aligned} \tag{3.3.17}$$

$$\begin{aligned}\nabla_{\lambda_2} \text{ELBO}_{\text{smi naive}} = \mathbb{E}_{\epsilon\sim p(\epsilon)}[&\nabla_\theta \left\{\log p(Y \mid \varphi, \theta) + \log p(\theta \mid \varphi)\right\} \nabla_{\lambda_2}\{\theta\} \\ &+ \nabla_{\lambda_2} \log |J_{T_2}(\epsilon_1, \epsilon_2)|].\end{aligned} \tag{3.3.18}$$

$$\begin{aligned}\nabla_{\lambda_3} \text{ELBO}_{\text{smi naive}} = \mathbb{E}_{\epsilon\sim p(\epsilon)}[&\nabla_\theta \left\{\eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\tilde{\theta} \mid \varphi)\right\} \nabla_{\lambda_3}\{\tilde{\theta}\} \\ &+ \nabla_{\lambda_3} \log |J_{T_2}(\epsilon_1, \epsilon_2)|].\end{aligned} \tag{3.3.19}$$

If these equations and $\nabla_\beta \text{ELBO}_{\text{bayes}}|_{\beta=(\lambda_1^*,\lambda_2^*)} = 0$ all hold at at $\eta = 1$ then

$$\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[\nabla_\varphi \left\{\log p(Y \mid \varphi, \tilde{\theta}) + \log p(Y \mid \varphi)\right\} \nabla_{\lambda_1}\{\varphi\}\right]_{\lambda=\lambda^*} = 0.$$

Our variational framework has to satisfy (P2) for every target $p_{\text{smi},\eta}$ and every variational family $\mathcal{Q}$. However, if we target the loss in eq. (3.3.14) then λ^* would have to satisfy an over-determined system of equations at $\eta = 1$ and this will in general have no solutions.

We learn from this that the loss we seek for the variational SMI approximation is not captured by the KL-divergence in eq. (3.3.14). However, there is a second practical problem with carrying out Stochastic Variational Inference based on this naive variational loss. In practice, in order to minimise eq. (3.3.15), we maximise $\text{ELBO}_{\text{smi,naive}}$ in eq. (3.3.16) using a Monte Carlo estimate of its gradients. The last term in eq. (3.3.16) involves the intractable $E_{\varphi\sim q_{\lambda_1}}[\log(p(Y,\varphi)]$, making the λ_1-variation unrealisable in practice.

3.3.3.4 Objective functions for variational-SMI

One way to characterise variational SMI is by generalising the two-stage optimisation approach given by Yu et al. (2021) for the Cut posterior. We will see that this approach satisfies properties (P1-3), and that the optimal variational parameters are given by minimising a customised variational loss. Let

$$d(q_\lambda, \mathcal{F}_{\text{smi},\eta}) = \min_{\tilde{q}\in\mathcal{F}_{\text{smi},\eta}} D_{KL}(q_\lambda \parallel \tilde{q}). \tag{3.3.20}$$

define the divergence between the density q_λ and the set of densities $\mathcal{F}_{\text{smi},\eta}$.

Proposition 4. *The divergence defined in eq.* (3.3.20) *can be written*

$$d(q_\lambda, \mathcal{F}_{smi,\eta}) = D_{KL}(q_{\lambda_1,\lambda_3}(\varphi,\tilde{\theta}) \parallel p_{pow,\eta}(\varphi,\tilde{\theta} \mid Y,Z)), \tag{3.3.21}$$

and hence does not depend on λ_2.

Proof. See section 3.7.6. □

We now define the optimal variational parameters. These will minimise divergence from distributions in $\mathcal{F}_{\text{smi},\eta}$ and otherwise approximate SMI. First, exploiting proposition 4, (λ_1, λ_3) minimise eq. (3.3.21). Let

$$\begin{aligned} d^*_{\text{smi}} &= \min_{(\lambda_1,\lambda_3)\in B} D_{KL}(q_{\lambda_1,\lambda_3}(\varphi,\tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi,\tilde{\theta} \mid Y,Z)) \\ \Lambda^*_{(1,3)} &= \{(\lambda_1,\lambda_3)\in B : D_{KL}(q_{\lambda_1,\lambda_3}(\varphi,\tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi,\tilde{\theta} \mid Y,Z)) = d^*_{\text{smi}}\}. \end{aligned} \tag{3.3.22}$$

Secondly, λ_2 is chosen for best approximation of $p_{\text{smi},\eta}$ at fixed $(\lambda^*_1, \lambda^*_3) \in \Lambda^*_{(1,3)}$. Let

$$\begin{aligned} D^*_{\text{smi}}(\lambda^*_1) &= \min_{\lambda_2\in\Lambda_2} E_{\varphi\sim q_{\lambda^*_1}}[D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y,\varphi))] \\ \Lambda^*_{(2)}(\lambda^*_1) &= \{\lambda_2\in\Lambda_2 : E_{\varphi\sim q_{\lambda^*_1}}[D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y,\varphi))] = D^*_{\text{smi}}(\lambda^*_1)\} \end{aligned} \tag{3.3.23}$$

The following proposition shows that $\lambda^*_2 \in \Lambda^*_{(2)}(\lambda^*_1)$ targets a good fit to $p_{\text{smi},\eta}$.

Proposition 5. *The set* $\Lambda^*_{(2)}(\lambda^*_1)$ *defined in eq.* (3.3.23) *is equivalently*

$$\begin{aligned}\tilde{D}^*_{smi}(\lambda^*_1) &= \min_{\lambda_2 \in \Lambda_2} D_{KL}(q_{(\lambda^*_1,\lambda_2,\lambda^*_3)} \parallel p_{smi,\eta}) \\ \Lambda^*_{(2)}(\lambda^*_1) &= \{\lambda_2 \in \Lambda_2 : D_{KL}(q_{(\lambda^*_1,\lambda_2,\lambda^*_3)} \parallel p_{smi,\eta}) = \tilde{D}^*_{smi}(\lambda^*_1)\}. \end{aligned} \tag{3.3.24}$$

Proof. Expand the KL divergence in eq. (3.3.24) using eq. (3.3.15) and substitute $(\lambda_1, \lambda_3) = (\lambda^*_1, \lambda^*_3)$. The first term does not depend on λ_2 and the second term gives eq. (3.3.23). □

We now define variational SMI and show it satisfies (P1-3).

Definition 3.3.1. *(Variational SMI)* A variational SMI posterior density is a density $q_{\lambda^*}(\varphi, \theta, \tilde{\theta})$ parameterised in eq. (3.3.10) with $\lambda^* \in \Lambda^*$ where

$$\Lambda^* = \bigcup_{(\lambda^*_1,\lambda^*_3) \in \Lambda^*_{(1,3)}} \left(\bigcup_{\lambda^*_2 \in \Lambda^*_{(2)}(\lambda^*_1)} \{(\lambda^*_1, \lambda^*_2, \lambda^*_3)\} \right), \tag{3.3.25}$$

and $\Lambda^*_{(1,3)}$ and $\Lambda^*_{(2)}(\lambda^*_1)$ are defined in eqs. (3.3.22) and (3.3.23) respectively.

Remark. Our discussion in this section takes η fixed. As we vary η the target $p_{\text{smi},\eta}$ varies, so the set of optimal variational parameters Λ^* depends on η. Below we write $\Lambda^*(\eta)$ when we need to emphasise this dependence.

Remark. The variational SMI parameters $\lambda^* \in \Lambda^*$ are roots of the equations

$$\nabla_{(\lambda_1,\lambda_3)} D_{KL}(q_{\lambda_1,\lambda_3} \parallel p_{\text{pow},\eta}) = 0 \tag{3.3.26}$$

$$\nabla_{\lambda_2} E_{\varphi \sim q_{\lambda_1}} D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi)) = 0 \tag{3.3.27}$$

with positive curvature. We have not substituted $\lambda_1 = \lambda^*_1$ in eq. (3.3.27). As a system, any λ_1 satisfying eq. (3.3.27) is required to be a root (with λ^*_3) of eq. (3.3.26) so the system imposes this condition. This will allow us to solve these equations as a single system using SGD on the loss $\mathcal{L}^{\text{smi},\eta}$ in proposition 9 below, avoiding a two-stage procedure.

Remark. Consider the variational family defined in eq. (3.3.12). Using the reparametrisation trick and expanding terms, eqs. (3.3.26) and (3.3.27) are

$$\begin{aligned} 0 = \mathbb{E}_{\epsilon \sim p(\epsilon)} [&\nabla_\varphi \left\{ \log p(Z \mid \varphi) + \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\varphi) \right\} \nabla_{\lambda_1} \{\varphi\} \\ &+ \nabla_{\lambda_1} \log |J_{T_1}(\epsilon_1)|], \end{aligned} \tag{3.3.28}$$

$$\begin{aligned} 0 = \mathbb{E}_{\epsilon \sim p(\epsilon)} [&\nabla_\theta \left\{ \log p(Y \mid \varphi, \theta) + \log p(\theta \mid \varphi) \right\} \nabla_{\lambda_2} \{\theta\} \\ &+ \nabla_{\lambda_2} \log |J_{T_2}(\epsilon_1, \epsilon_2)|] \end{aligned} \tag{3.3.29}$$

$$\begin{aligned} 0 = \mathbb{E}_{\epsilon \sim p(\epsilon)} [&\nabla_{\tilde{\theta}} \left\{ \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\tilde{\theta} \mid \varphi) \right\} \nabla_{\lambda_3} \{\tilde{\theta}\} \\ &+ \nabla_{\lambda_3} \log |J_{T_2}(\epsilon_1, \epsilon_3)|] \end{aligned} \tag{3.3.30}$$

Notice that the extra terms in eq. (3.3.17) are absent in eq. (3.3.28) so λ^* will not be over-determined when we come to match variational Bayes at $\eta = 1$.

Theorem 1. *Variational SMI in definition 3.3.1 satisfies properties (P1), (P2) and (P3).*

Proof. See section 3.7.6 where P1, P2 and P3 are demonstrated in proposition 6, 7 and 8 respectively. □

3.3.3.5 The overall loss targeted by variational-SMI

We have defined λ^* in two steps, eqs. (3.3.22) and (3.3.23) with two losses, $d(q_\lambda, \mathcal{F}_{\text{smi},\eta})$ and $D_{KL}(q_\lambda \parallel p_{\text{smi},\eta})$. We can bring this together into a single overall loss in two ways. The first (eq. (3.3.31)) is formal and the more useful for computation. The second (eq. (3.3.34)) is useful for mathematical understanding.

For computational purposes we define the loss $\mathcal{L}^{(\text{smi},\eta)}$ targeted by variational SMI using the stop-gradient operator, denoted as $\not\nabla(\cdot)$, acting on $\varphi_{(\epsilon,\lambda_1)}$. Let

$$\mathcal{L}^{(\text{smi},\eta)}(\lambda) = \text{ELBO}_{\text{pow},\eta}(\lambda_1, \lambda_3) + \text{ELBO}_{\text{bayes}\not\nabla(\varphi)}(\lambda_1, \lambda_2) \tag{3.3.31}$$

where

$$\text{ELBO}_{\text{pow},\eta}(\lambda_1, \lambda_3) = \mathbb{E}_{(\varphi,\tilde{\theta})\sim q_{\lambda_1,\lambda_3}}[\log p_{\text{pow},\eta}(\varphi, \tilde{\theta}, Z, Y) - \log q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta})] \tag{3.3.32}$$

$$\text{ELBO}_{\text{bayes}\not\nabla(\varphi)}(\lambda_1, \lambda_2) = \mathbb{E}_{(\varphi,\theta)\sim q_{\lambda_1,\lambda_2}}[\log p(\not\nabla(\varphi), \theta, Z, Y) - \log q_{\lambda_1,\lambda_2}(\not\nabla(\varphi), \theta)]. \tag{3.3.33}$$

with the joint and powered joint distributions given as eqs. (3.2.3) and (3.2.10).

The stop-gradient prevents the contribution of its inputs to be taken into account for computing gradients. Operationally, $\not\nabla(\cdot)$ acts as the identity function in the forward direction, that is, it returns its arguments unchanged. However, it prevents the flow of gradients through that operator in the backward direction in automatic differentiation. It does not prevent backpropagation altogether, but instead prevents an individual tensor from contributing to the gradients that are computed for an expression. (See e.g. Bradbury et al., 2018). More formally, if we consider a differentiable function f, the stop-gradient operator yields a modified function $g = \not\nabla(f)$ with the properties of value preservation, $g(x) = f(x)$, and gradient annihilation, $\nabla g(x) = 0$.

In eq. (3.3.31) we are in effect defining the function and its derivative separately and so this loss is formal and cannot take the place of proposition 10 below in giving meaning to the variation. However it is convenient for implementation, as the `stop_gradient` operator is directly expressed in the automatic differentiation framework we use.

Proposition 9. *The set Λ^* in definition 3.3.1 is the set of solutions of $\nabla_\lambda \mathcal{L}^{(smi,\eta)} = 0$ corresponding to minima.*

Proof. See section 3.7.6. □

An overall loss *function* can be given as follows. Let $v \geq 0$ and

$$\mathcal{L}^{(v)}(\lambda) = d(q_\lambda, \mathcal{F}_{\text{smi},\eta}) + v \cdot D_{KL}(q_\lambda \parallel p_{\text{smi},\eta}) \tag{3.3.34}$$

denote a weighted loss which allows varying levels of priority to be put on proximity to $\mathcal{F}_{\text{smi},\eta}$ and approximation of $p_{\text{smi},\eta}$.

Proposition 10. *Let $\mathcal{L}^*(v) = \min_{\lambda\in\Lambda} \mathcal{L}^{(v)}(\lambda)$ and*

$$\Lambda^*(v) = \{\lambda \in \Lambda : \mathcal{L}^{(v)}(\lambda) = \mathcal{L}^*(v)\}.$$

Under regularity conditions on $\mathcal{F}_{smi,\eta}$ and $p_{smi,\eta}$ given in proposition 11, for every solution $\lambda^* \in \Lambda^*$ in definition 3.3.1 and all sufficiently small $v \geq 0$ there exists a unique continuous function $\lambda^*(v)$ satisfying $\lambda^*(v) \in \Lambda^*(v)$ and

$$\lim_{v\to 0} \lambda^*(v) = \lambda^*.$$

Proof. See section 3.7.6. □

The value of proposition 10 is that it allows us to interpret q_{λ^*}, $\lambda^* \in \Lambda^*$ as minimising a proper loss function $\mathcal{L}^{(v)}$ at small v (approximately). The minimum loss $\mathcal{L}^*(v)$ decreases as we expand the variational family $\mathcal{Q}$ and is zero when $p_{\text{smi},\eta} \in \mathcal{Q}$, in which case $q_{\lambda^*} = p_{\text{smi},\eta}$ for $\lambda^* \in \Lambda^*(v)$. In contrast, although d^*_{smi} decreases as $\mathcal{Q}$ expands, $D^*_{\text{smi}}(\lambda^*_1)$ may increase, though must eventually go to zero when $\mathcal{Q}$ expands to include $p_{\text{smi},\eta}$, by theorem 1. However, $\mathcal{L}^{(v)}$ is not a viable optimisation target at small v because the second term in eq. (3.3.34) is intractable, as we saw in our discussion of eq. (3.3.16).

3.3.3.6 Stochastic gradient descent for variational-SMI

Algorithm 2 gives our Stochastic Gradient Descent method to target the SMI posterior for a fixed value of the influence parameter, η. The algorithm is based on the loss $\mathcal{L}^{(\text{smi},\eta)}$ in eq. (3.3.31) and consists of a single training loop, using the `stop_gradient` operator to avoid two-stage optimisation procedures. This is given for understanding. In section 3.4 and algorithm 3 we will train a "meta-posterior" approximating the whole family of SMI-posteriors as a function of η.

Algorithm 2 Variational Posterior approximation for $p_{\text{smi},\eta}$

Input: $\mathcal{D}$: Data. $p(\varphi, \theta, \mathcal{D})$: Multi-modular probabilistic model. $q_\lambda = (p(\epsilon), T, \lambda)$: variational family. A value of $\eta \in [0, 1]$: Influence parameter(s) for suspect module(s)

Output: Variational approximation $q_{\hat{\lambda}}(\varphi, \theta, \tilde{\theta})$ of the η-SMI posterior.

Initialise variational parameters λ
while SGD not converged **do**
  (Optional) Sample a random minibatch of data $\mathcal{D}^{(b)} \sim \mathcal{D}$.
  for $s = 1, \ldots, S$ **do**
    Sample the base distribution, $\epsilon_s \sim p(\epsilon)$.
    Transform the sampled values $(\varphi_s, \theta_s, \tilde{\theta}_s) \leftarrow T(\epsilon_s; \lambda)$ as in eq. (3.3.11).
  end for
  Compute the Monte Carlo estimate of the loss $\mathcal{L}^{(\text{smi},\eta)}$ in eq. (3.3.31) and its gradients.

$$\hat{\mathcal{L}}^{(\text{smi},\eta)} = \widehat{\text{ELBO}}_{\text{pow},\eta} + \widehat{\text{ELBO}}_{\not\varphi(\varphi)} \tag{3.3.35}$$

  where

$$\widehat{\text{ELBO}}_{\text{pow},\eta} = -\frac{1}{S}\sum_{s=1}^{S}\left[\log p_{\text{pow},\eta}(\varphi_s, \tilde{\theta}_s, \mathcal{D}^{(b)}) - \log q(\varphi_s, \tilde{\theta}_s)\right] \tag{3.3.36}$$

$$\widehat{\text{ELBO}}_{\not\varphi(\varphi)} = -\frac{1}{S}\sum_{s=1}^{S}\left[\log p(\not\varphi(\varphi_s), \theta_s, \mathcal{D}^{(b)}) - \log q(\not\varphi(\varphi_s), \theta_s)\right] \tag{3.3.37}$$

  Update λ using the estimated gradient vector $\nabla_\lambda \hat{\mathcal{L}}^{(\text{smi},\eta)}$
  Check convergence of $q_\lambda(\varphi_s, \theta_s, \tilde{\theta}_s)$
end while
return $\hat{\lambda} = \lambda$

3.3.4 Selecting the SMI posterior

We now give a utility for selection of the influence parameter η. This will depend on the goals of inference. Recall that $Y \in \mathcal{Y}^m$ and $Z \in \mathcal{Z}^n$. In the following we take a predictive loss based on the SMI-predictive distribution for independent new data $(y', z') \in \mathcal{Y} \times \mathcal{Z}$,

$$p_{\text{smi},\eta}(y', z' \mid Y, Z) = \int p(y', z' \mid \varphi, \theta) p_{\text{smi},\eta}(\varphi, \theta \mid Y, Z)\, d\varphi d\theta,$$

and a utility which is equivalent to the negative of the KL divergence to the true generative model for the new data $p^*(y', z')$. This utility is the Expected Log-pointwise Predictive Density (ELPD) (Vehtari et al., 2017),

$$U(\eta) = \int p^*(y', z') \log(p_{\text{smi},\eta}(y', z' \mid Y, Z))\, dy' dz'.$$

In our variational setting, these quantities are replaced by estimates based on our variational approximation q_{λ^*} to $p_{\text{smi},\eta}$. The variational parameters are $\lambda^*(\eta) \in \Lambda^*(\eta)$, per section 3.3.3.4. We define the variational SMI posterior predictive distribution

$$q_\eta(y', z') = \int p(y', z' \mid \varphi, \theta) q_{\lambda^*(\eta)}(\varphi, \theta)\, d\varphi d\theta, \tag{3.3.38}$$

with corresponding utility

$$U(\eta) = \int p^*(y', z') \log(q_\eta(y', z'))\, dy' dz'. \tag{3.3.39}$$

We estimate U in eq. (3.3.39) using the WAIC (Watanabe, 2012), following Vehtari et al. (2017). See section 3.4.3 for further details. When the variational SMI posterior predictive distribution can be calculated in closed form, $U(\eta)$ may be estimated using leave one out cross validation. This is asymptotically equivalent to the WAIC, but will in general be too computationally demanding to compute. In order to complete the inference, we select the optimal influence parameter

$$\eta^* = \arg\max_\eta U(\eta), \tag{3.3.40}$$

and return the final selected variational SMI posterior, $q_{\lambda^*(\eta^*)}(\varphi, \theta)$, for further analysis.

A number of other procedures have been given for selecting the influence in a pure power-posterior setting. Wu and Martin (2023, 2021) introduce a new method and summarise and compare a selection of methods, reflecting different priorities in the inference and corresponding utilities. If our goal is parameter estimation, then a utility that directly targets parameter estimates, rather than predictive distributions, will be preferred. In recent work, Chakraborty et al. (2023) select η as the most Bayes-like (i.e., the largest) value that does not show goodness-of-fit violation with the Cut. Carmona et al. (2023) use a utility tailored to their inference objectives. They have data in which θ is a high dimensional vector, and some of the components of θ are directly measured. They use the posterior mean square error for prediction of known θ-components in a LOOCV framework to select η, linking η-selection to success in parameter estimation.

3.4 The Variational Meta-Posterior

In order to select a variational SMI posterior we estimate the selection criterion $U(\eta)$ in eq. (3.3.39) as a function of η and select an optimal η-value in eq. (3.3.40) and the

variational SMI posterior $q_{\lambda^*(\eta^*)}$. For this we need the fitted parameters $\lambda^*(\eta)$ as functions of η. When $\eta \in [0,1]$ is scalar, we can fit the variational posterior independently at a lattice of η-values, estimate the ELPD at each value, smooth the estimated ELPD values over $\eta \in [0,1]$ and select the η-value maximising this function. However, when we analyse multi-modular models with multiple misspecified modules, the dimension of η grows with the number of bad modules and so independent fitting is not feasible. In this section we give two parameterisations of the Variational Meta-Posterior (VMP), $q_{\lambda(\alpha,\eta)}$ and $q_{\alpha,\eta}$. In the former, based on a "VMP-map", the parameters λ of the NF are expressed as functions of η with parameters α. In the latter, based on a "VMP-flow", η is an additional input to the NF with additional flow parameters μ, and $\alpha = (\lambda, \mu)$.

3.4.1 Motivation and definitions

The SMI-posterior varies continuously with η. Expanding the KL divergence at $\eta+\delta$,

$$D_{KL}(p_{\text{smi},\eta} \parallel p_{\text{smi},\eta+\delta}) = -\delta E_{\text{pow},\eta}\left(\log(p(Y \mid \varphi, \tilde{\theta}))\right) + \log\left(E_{\text{pow},\eta}(p(Y \mid \varphi, \tilde{\theta})^{\delta})\right),$$

and this is continuous and has continuous derivatives in δ if the integrals exist. This motivates parameterisations of $q_{\lambda(\alpha,\eta)}$ and $q_{\alpha,\eta}$ which are continuous in the same sense.

3.4.1.1 The VMP-map

Continuity holds in a stronger sense. Under regularity conditions, a continuous sequence of solutions $\lambda^*(\eta) \in \Lambda^*(\eta)$ passes through any point $\lambda^* \in \Lambda^*(\eta^*)$. Applying the Implicit Function Theorem (as in proposition 11 in section 3.7.6), to the η-dependence of the roots of the functions on the LHS of eqs. (3.3.26) and (3.3.27) we find that, for every $\lambda^* \in \Lambda^*(\eta^*)$, there is a unique continuous function $\lambda^*(\eta)$ satisfying $\lambda^*(\eta) \in \Lambda^*(\eta)$ for η in an open neighborhood of η^* and satisfying $\lambda^*(\eta^*) = \lambda^*$. The regularity conditions require the functions on the LHS of eqs. (3.3.26) and (3.3.27) to be continuously differentiable in λ and η, and the Jacobians of those functions (in (λ_1, λ_3) and λ_2) to be invertible at $\lambda = \lambda^*$.

This motivates a low dimensional reparameterisation of λ which approximates $\lambda^*(\eta)$. Let $\lambda(\alpha, \eta) = f_\alpha(\eta)$, where $\alpha \in A$ is a vector of real parameters and let

$$f_\alpha : H \to \Lambda \tag{3.4.1}$$

be a continuously differentiable mapping parameterised by α. We refer to f_α as the **VMP-map** and define a Variational Meta-Posterior as a family of distributions

$$\mathcal{Q}_{H,\alpha} = \{q_{f_\alpha(\eta)}\ ,\ \eta \in H\}, \quad \alpha \in A. \tag{3.4.2}$$

There is a question of how the α parameters should contribute to the different components of λ. We take $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and

$$f_\alpha(\eta) = (f^{(1)}_{\alpha_1}(\eta), f^{(2)}_{\alpha_2}(\eta), f^{(3)}_{\alpha_3}(\eta))$$

where $f^{(1)}_{\alpha_1} : H \to \Lambda_1$ and $f^{(2)}_{\alpha_2}, f^{(3)}_{\alpha_3} : H \to \Lambda_2$ and set $\lambda_k(\alpha_k, \eta) = f^{(k)}_{\alpha_k}(\eta),\ k = 1, 2, 3$. This gives $f_\alpha : H \to \Lambda$ as before, but breaks up the dependence as

$$(\varphi_{(\lambda_1(\alpha_1,\eta),\epsilon)}, \theta_{(\lambda_2(\alpha_2,\eta),\epsilon)}, \tilde{\theta}_{(\lambda_3(\alpha_3,\eta),\epsilon)}) \sim q_{f_\alpha(\eta)}. \tag{3.4.3}$$

Changing α_1 to improve the fit to the φ distribution at one η does not affect the θ distribution at another η-value, though it will affect the φ distribution there.

A very expressive VMP-map may be undesirable due to a bias-variance trade off in the estimation of λ^*. The estimates $\hat{\lambda}$ of λ^* output by algorithm 2 are estimated independently over η and will not in general lie in $\Lambda^*(\eta)$. Reparameterising with f_α and estimating $\hat{\alpha}$ for best fit across $\eta \in H$ smooths the output $\hat{\lambda}(\eta) = f_{\hat{\alpha}}(\eta),\ \eta \in H$ at the price of some potential bias. Properties (P1-3) hold only approximately on both outputs.

3.4.1.2 The VMP-flow

When we parameterise the VMP with a NF, we model the η-dependence of the variational densities q. We can parameterise the function $\lambda^*(\eta)$ using a VMP-map. Alternatively we can add an η-input to the maps T_1, T_2 (technically an extra conditioner, like ϵ_1). The flow architecture is expanded with extra nodes and weight parameters $\mu = (\mu_1, \mu_2, \mu_3)$ with $\mu \in M$ say. The map T is continuous in its inputs so q will be continuous in η. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ with $\alpha_k = (\lambda_k, \mu_k),\ k = 1, 2, 3$ and $\alpha \in A$ where now $A = \Lambda \times M$. The transformations T_1, T_2 with input η are given in terms of the conditioners and transformers of the NF in eq. (3.7.9) to eq. (3.7.11) in section 3.7.4. Formally,

$$\begin{aligned}
\varphi_{(\alpha_1,\eta,\epsilon)} &= T_1(\epsilon_1; \alpha_1, \eta) \\
\theta_{(\alpha_2,\eta,\epsilon)} &= T_2(\epsilon_2; \alpha_2, (\eta, \epsilon_1)) \\
\tilde{\theta}_{(\alpha_3,\eta,\epsilon)} &= T_2(\epsilon_3; \alpha_3, (\eta, \epsilon_1)) \\
T(\epsilon; \alpha, \eta) &= (T_1(\epsilon_1; \alpha_1, \eta)),\ T_2(\epsilon_2; \alpha_2, (\eta, \epsilon_1)),\ T_2(\epsilon_3; \alpha_3, (\eta, \epsilon_1)))\,.
\end{aligned} \tag{3.4.4}$$

We call this extended flow mapping T with input η a **VMP-flow**. In terms of the new map, the variational densities are

$$q_{\alpha,\eta}(\varphi, \theta, \tilde{\theta}) = q_{\alpha_1,\eta}(\varphi) q_{\alpha_2,\eta}(\theta \mid \varphi) q_{\alpha_3,\eta}(\tilde{\theta} \mid \varphi), \tag{3.4.5}$$

simply replacing $\lambda_k \leftarrow \alpha_k,\ k-1,2,3$ and making the η-dependence explicit as it is a flow input. This gives a second Variational Meta-Posterior as the family of distributions

$$\mathcal{Q}_{H,\alpha} = \{q_{\alpha,\eta}\ ,\ \eta \in H\}, \quad \alpha \in A. \tag{3.4.6}$$

3.4.1.3 The VMP and SMI

Definition 3.4.1. The Variational Meta-Posterior for SMI is given by a pair $(\mathcal{P}_{\text{smi}}, \mathcal{Q}_H)$, where $\mathcal{P}_{\text{smi}} = \{p_{\text{smi},\eta} : \eta \in H\}$ is the family of SMI posteriors indexed by η which we want to approximate and $\mathcal{Q}_H = \cup_{\alpha\in A}\mathcal{Q}_{H,\alpha}$ is the family of all available Variational Meta-Posteriors determined using the VMP-map in eq. (3.4.2) or the VMP-flow in eq. (3.4.6).

3.4.2 Learning the Variational Meta-Posterior for SMI

In this section we give losses for estimation of α^* in the VMP-map and VMP-flow. Let $\eta_{1:R} = (\eta_r)_{r=1,\ldots,R}$ be a given lattice of η-values and let

$$\rho(\eta) = \frac{1}{R}\sum_{r=1}^{R} \delta_{\eta_r}(\eta),\ \eta \in H.$$

The values in $\eta_{1:R}$ would ideally be concentrated around η^*. As this isn't known in advance, concentrating them near the points $\eta_{1:R} \in \{0,1\}^R$ is a useful rule as $\lambda^*(\eta)$ often varies rapidly with η near Cut and Bayes. Adaptive sequential estimation and maximisation of $U(\eta)$ may be of interest in future work.

We take a meta-SMI loss weighted across $\eta \in \eta_{1:R}$. For the VMP-map this is

$$\begin{aligned}\mathcal{L}^{(\text{msmi}-map)}(\alpha) &= \mathbb{E}_{\eta\sim\rho}\left(\mathcal{L}^{(\text{smi},\eta)}(f_\alpha(\eta)\right) \\ &= \frac{1}{R}\sum_{r=1}^{R}\mathcal{L}^{(\text{smi},\eta_r)}(f_\alpha(\eta_r))\end{aligned} \tag{3.4.7}$$

where $\mathcal{L}^{(\text{smi},\eta)}(\lambda)$ is defined in eq. (3.3.31) and we have taken $\lambda = f_\alpha(\eta)$ in order to enforce the parameterisation at each $\eta \in \eta_{1:R}$. For the VMP-flow the loss is

$$\mathcal{L}^{(\text{msmi}-flow)}(\alpha) = \mathbb{E}_{\eta\sim\rho}\left(\mathcal{L}^{(\text{msmi}-flow,\eta)}(\alpha)\right) \tag{3.4.8}$$

where $\mathcal{L}^{(\text{msmi}-flow,\eta)}(\alpha)$ is obtained by substituting $\varphi_{(\alpha_1,\eta,\epsilon)}$ etc into $\mathcal{L}^{(\text{smi},\eta)}$ in eq. (3.3.31) and is defined in detail in eq. (3.7.13) in section 3.7.4.

In the VMP-flow the optimal variational parameters λ^* don't depend on η and this seems to give relatively more rapid and stable convergence in SGD targeting

$\mathcal{L}^{(\text{msmi}-flow)}(\alpha)$ compared to SGD targeting in $\mathcal{L}^{(\text{msmi}-map)}(\alpha)$ (given in algorithm 3). It is a relatively "lightweight" parameterisation, as the dimension of α in the VMP-flow is quite a bit smaller than that of μ in the VMP-map.

In order to estimate $\alpha^* = \arg\min_{\alpha \in A} \mathcal{L}^{(\text{msmi})}(\alpha)$ (dropping the -map and -flow distinction, and ignoring non-uniqueness for brevity) and fit the VMP, we apply SGD, simply replacing $\mathcal{L}^{(\text{smi},\eta)}(\lambda)$ algorithm 2 with $\mathcal{L}^{(\text{msmi})}(\alpha)$ in algorithm 3 in section 3.7.5, and updating α with the gradient $\nabla_\alpha \mathcal{L}^{(\text{msmi})}$ instead of updating λ with the gradient $\nabla_\lambda \mathcal{L}^{(\text{smi},\eta)}$. We implement this using Stochastic Variational Inference (SVI). We take a continuous density ρ and sample a new batch $\eta_{1:R}$ of η-values at each pass of the SGD algorithm. We approximate the family $\mathcal{P}_{\text{smi}}$ in a single end-to-end optimisation, propagating the loss function gradients through the VMP-map or -flow using automatic differentiation. See algorithm 3 in section 3.7.5.

When $\dim(H) = C$ (C influence parameters for C cuts) we have vectors $\eta_r = (\eta_{1,r}, ..., \eta_{C,r})$, $r = 1, ..., R$. We defined ρ for resampling purposes as $\eta_{c,r} \sim \text{beta}(a, 1)$, $c = 1, ..., C$ with $a < 1$ independently for each component of η, for example $a = 0.2$. This concentrates sampled η at the Cut boundaries of H as noted above.

3.4.3 Maximising the ELPD using the Variational Meta-Posterior

The VMP allows us to produce posterior samples for any given $\eta \in H$ efficiently. This helps us find the best influence parameter η^* (see section 3.3.4), as we can estimate the utility function $U(\eta)$ accurately in fractions of a second and compare it across the family of SMI posteriors. In settings with a single cut, maximising $U(\eta)$ may be as simple as linear search, but in the case of many potential cuts, with a higher-dimensional H space, we require more elaborate search strategies.

In our case the utility $U(\eta)$ is the ELPD and we estimate this (its negative) using the WAIC estimator given in Vehtari et al. (2017) and minimise the WAIC over $\eta \in H$ with SGD. Denote by $\psi_{\hat{\alpha},\eta,\epsilon} = (\varphi_{\hat{\alpha},\eta,\epsilon}, \theta_{\hat{\alpha},\eta,\epsilon}, \tilde{\theta}_{\hat{\alpha},\eta,\epsilon})$ a full sample parameter vector $\psi_{\hat{\alpha},\eta,\epsilon} \sim q_{\hat{\alpha},\eta}$ from a fitted VMP-flow evaluated at η. Denote by $\psi_{\hat{\alpha},\eta,\epsilon_{1:J}}$ a set of J iid samples from $q_{\alpha,\eta}$ and let $\mathcal{D} = (Y, Z)$ denote the data. The WAIC is a function $\widehat{-\text{elpd}}(\psi_{\hat{\alpha},\eta,\epsilon_{1:J}}, D)$ of the samples and data (Vehtari et al., 2017). In order to implement SGD we need function evaluations and derivatives of $\widehat{-\text{elpd}}(\eta)$ wrt η (keeping only the η dependence). Function evaluations are very fast. In our JAX/TensorFlow setup (Babuschkin et al., 2020; Dillon et al., 2017) we get η-derivatives using automatic differentiation through the functions in the ELPD and all the way into $\psi_{\hat{\alpha},\eta,\epsilon_j}$, $j = 1, ..., J$. This can be seen in operation in our online code.

The main difficulty (in our example in section 3.5.2 where $\eta \in [0,1]^N$) is that the ELPD is clearly non-convex (from our plots), and quite flat when the η components are all close to one. We therefore initialise SGD using an (informed) greedy backward search. This uses backwards selection over cuts starting from Bayes, cutting the module which gives the greatest reduction in $\widehat{-\text{elpd}}$ and stopping when no decrease is possible.

3.5 Experiments

Our experiments illustrate the following points: Variational-SMI with a NF and without (definition 3.3.1) or with (definition 3.4.1) a VMP accurately approximates SMI-posteriors at all η in the examples we consider; the VMP-framework allows us to select an influence parameter *vector* $\eta^* \in H$, where $H = [0,1]^C$ and C is the number of cuts, at values of C which are completely out of reach for one-η-at-a-time MCMC or variational-SMI. MCMC is fine if we want to check the VMP at a handful of η values.

We use two examples which have become default test cases (e.g. Plummer, 2015; Jacob et al., 2017; Carmona and Nicholls, 2020; Liu and Goudie, 2022b; Nicholls et al., 2022). In the first epidemiological example taken from Plummer (2015), we show that VMP agrees with variational-SMI and nested-MCMC (which serves as ground truth) across a range of η-values and in particular at $\eta = 0, 1$, Cut and Bayes. An expressive variational family is needed, so while NFs are effective, MFVI fails. In the second random effects example taken from Jacob et al. (2017), we illustrate variational-SMI with multiple cuts, and compare different methods for estimating the utility $U(\eta)$ which we take as the ELPD throughout. In a companion paper, Carmona et al. (2023), we give an analysis of a spatial model where MCMC at even one η-value is infeasible. This (third) extended example illustrates careful choice of utility for η-selection, as well as being of independent interest in the application domain.

Our variational family has Neural Spline Flow (NSF) transformers (Durkan et al., 2019) with MLP conditioners in eight coupling layers (Dinh et al., 2016), with a MFVI analysis for comparison. See section 3.7.4 for this terminology. We found this arrangement gave an expressive transformation $T = (T_1, T_2, T_2)$ that was easily trained and worked for both examples. Code to replicate all results in this section is available as an open-source repository [2].

[2] `https://github.com/chriscarmona/modularbayes`

Our implementation is based on DeepMind JAX Ecosystem (Babuschkin et al., 2020) and TensorFlow Probability (Dillon et al., 2017). Experiments were carried out using a single Cloud TPU machine type v3-8. Qualitative runtimes to approximate a single SMI posterior for the Random Effects example using our favoured NF were in the range of 10 minutes and sampling 10000 iid samples takes less than a second. This total time is similar to the time to obtain one correlated sample of size $10,000$ at one η-value using nested MCMC. Training the VMP required between 0.5 and 2 hours. However, this training time is compensated by a significant reduction in the search for η^*, as we can generate samples from any $q_{\hat{\alpha},\eta}$ and estimate $U(\eta)$ (the WAIC) in a fraction of a second. Optimisation using greedy initialisation and SGD requires thousands of WAIC-estimates and took about 5 minutes using the VMP, whereas each of these estimates would take 10 minutes using MCMC. Further, we cannot get gradients by automatic differentiation in nested MCMC.

3.5.1 Epidemiological Model

We revisit the well-known epidemiological model for the correlation between Human papillomavirus (HPV) prevalence and cervical cancer incidence (see Maucort-Boulch et al., 2008; Plummer, 2015, for details). In this modular model a small "expensive" prospective trial controlling sample selection from the target population gives straightforward statistical modelling. A second much larger retrospective data set contains information about population parameters, but was gathered with little control over sample selection bias. This sort of data synthesis appears frequently. For example, the simplest Covid prevalence model in Nicholson et al. (2022), which brings together sample survey data and walk-in testing results, belongs to this class.

The data consist of four variables observed from $n = m = 13$ groups of women from n different countries. The model has two modules, a Binomial distribution for the number $Z_i,\ i = 1, ..., n$ of women infected with HPV in a sample of size N_i from the i'th group and a Poisson distribution for the number of cancer cases Y_i during T_i women-years of followup. That is,

$$\begin{aligned} Z_i &\sim Binomial(N_i, \varphi_i). \\ Y_i &\sim Poisson(\mu_i) \\ \mu_i &= T_i \exp(\theta_1 + \theta_2 \varphi_i), \quad i = 1, \ldots, 13. \end{aligned}$$

Following previous authors, the parameter spaces are $\theta \in [0, \infty)^2$ and $\varphi \in [0, \infty)^n$ and the priors are truncated independent normal priors with variance 1000.

Our variational approximation takes an $L = 17$-dimensional independent standard Normal $p(\epsilon) = N(\epsilon; 0, \mathbb{I}_{17})$ as our *base distribution* ($L_1 = 13$ elements for φ, $L_2 = 2$ for θ, and so also $L_2 = 2$ for $\tilde{\theta}$). The NF-conditioner in T_2 (section 3.7.4) takes $\epsilon_{1:13}$ as an input, allowing correlation between φ and θ and between φ and $\tilde{\theta}$, and conditional independence between θ and $\tilde{\theta}$ given φ.

Samples from a VMP $q_{f_{\hat{\alpha}}(\eta)}$ fitted using VMP-NSF-MLP and a VMP-map are shown in fig. 3.3 (at $\eta \in \{0.001, 0.1, 1\}$, corresponding to the Cut, Bayes, and a value of η "halfway" between the two [3]). "Ground truth" SMI distributions obtained using nested-MCMC (Plummer, 2015; Carmona and Nicholls, 2020) are shown in fig. 3.11 for comparison. Samples from the VMP $q_{\hat{\alpha},\eta}$ fitted using a VMP-flow and samples from variational-SMI distributions $q_{\hat{\lambda}(\eta)}$ (estimating $\lambda^*(\eta)$ separately at each η without a VMP) are essentially identical to the variational $q_{f_{\hat{\alpha}}}(\eta)$ posteriors and are omitted. The good agreement here to MCMC shows both that the training losses $\mathcal{L}^{(\text{smi},\eta)}$ and $\mathcal{L}^{(\text{msmi})}$ we wrote down in section 3.3.2 and section 3.4 are doing their job and enforcing a good fit to $p_{\text{smi},\eta}$ over all $\eta \in [0, 1]$, and at the same time interpolating variational approximations with good inferential properties to the Cut (no Y-module feedback) and Bayes (full feedback) posteriors.

In section 3.7.7 we include a comparison with MFVI (see fig. 3.12). This demonstrates its failure, under-dispersed relative to the target $p_{\text{smi},\eta}$, and demonstrates the advantages of using an expressive flow-family. In this example we omit the final stage of an SMI analysis, that is, we do not estimate the ELPD and select η^* and p_{smi,η^*}. As the variational posteriors match nested MCMC, this part of the analysis is the same as that given in Carmona and Nicholls (2020) (though faster, as sampling our flow is *much* faster than nested-MCMC).

3.5.2 Random effects model with Multiple Cuts

The example in this section illustrates the VMP with multiple cuts, demonstrating its convenience in settings with more than two modules. The modules are all potentially misspecified, which in this model gives thirty cuts. The *influence parameter vector* $\eta = (\eta_1, \ldots, \eta_N)$, $\eta \in H$ regulates the influence of each module. We take our random effects model and synthetic data setup from Liu et al. (2009) and Jacob et al. (2017).

Denote by $Y_i = (y_{i,1}, \ldots, y_{i,n_i})$ the data in group $i = 1, ..., N$. We take $N = 30$ and $n_i = 5$ below. The hierarchical Gaussian model with random effects $\beta = (\beta_i)_{i=1,\ldots,N}$

[3] for illustration we take 0.1 instead of 0.5 because posteriors with $\eta \gtrsim 0.2$ are very similar

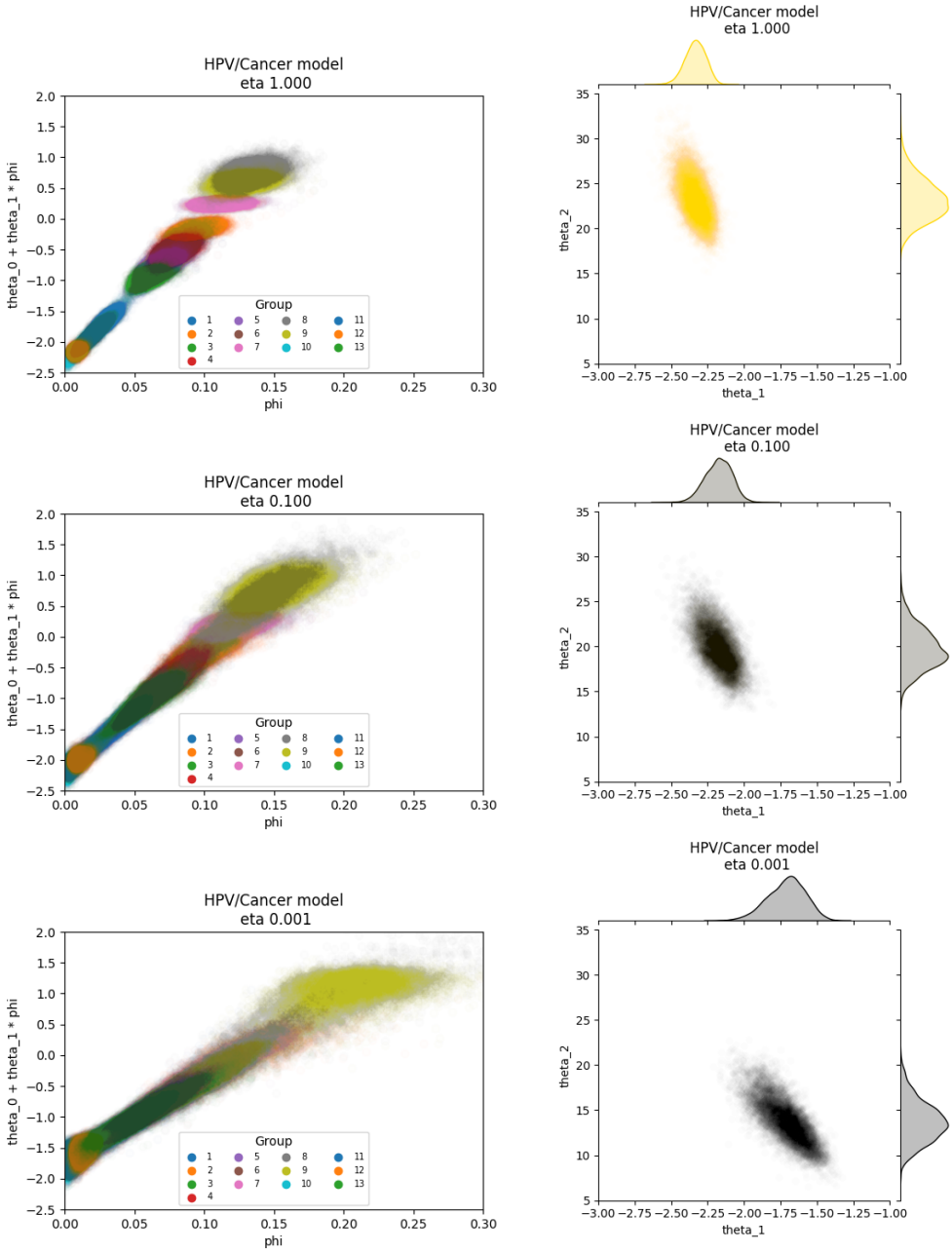


Figure 3.3: Samples from the Variational Meta-Posterior in section 3.5.1, obtained using VMP-NSF-MLP. Rows correspond to influence levels $\eta = (0.001, 0.1, 1)$. The left column shows the relation between HPV prevalence (φ) and cervical cancer incidence $\mu_i = \theta_1 + \theta_2 \varphi_i$ colored by group $i = 1, ..., n$ for the $n = 13$ groups in the data. The right column shows how the joint distribution of slope (θ_1) and intercept (θ_2) varies with η. All samples are produced using a single VMP-map.

and variances $\sigma = (\sigma_i)_{i=1,...,N}$ is specified as follows:

$$Y_{i,j} \sim Normal(\beta_i, \sigma_i^2), \quad i = 1, \ldots, N \; ; \; j = 1, \ldots, n_i$$
$$\beta_i \mid \tau \sim Normal(0, \tau^2),$$

with σ and τ priors

$$p(\sigma_i^2) \propto \sigma_i^{-2}$$
$$p(\tau \mid \sigma) \propto \frac{1}{\tau^2 + \frac{1}{N}\sum_{i=1}^{N} \frac{\sigma_i^2}{n_i}}.$$

A graphical representation of the model and its cuts is displayed in fig. 3.4. The Bayes

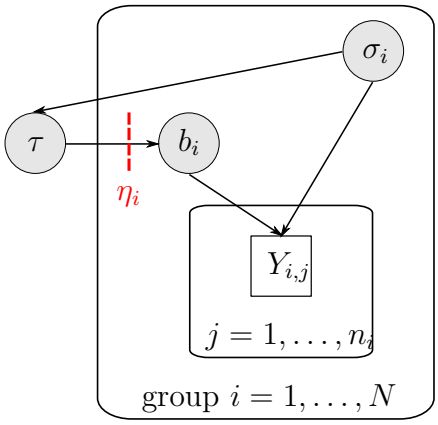


Figure 3.4: Graphical representation of the model for random effects.

posterior is

$$p(\sigma, \beta, \tau \mid Y) \propto p(Y \mid \sigma, \beta)p(\beta \mid \tau)p(\tau \mid \sigma)p(\sigma).$$

Since this is a study on synthetic data, we follow Liu et al. (2009) and take the number of groups N to be large and the number of replications n_i per group to be small as this gives a strong distorting effect under the misspecified model. We simulate data from the true observational model, $p^*(Y) = p(Y \mid \beta^*, \sigma^*)$, with two very large random effects, $\beta_1^* = 10$ and $\beta_2^* = 5$ and zero for the rest, $\beta_i^* = 0$ for $i = 3, \ldots, N$. We take a unit scale for all groups, $\sigma_i^* = 1$ for $i = 1, \ldots, N$. The data are shown in fig. 3.5.

Our choice of how to divide the model into modules depends on how we plan to cut. Liu et al. (2009) have two modules and one cut: Module 1 is the observation module $(Y_i, \beta_i, \sigma_i)_{i=1,\ldots,N}$, while Module 2 is the prior module $(\beta_i, \sigma_i, \tau)_{i=1,\ldots,N}$. Those authors discuss how the Bayes posterior is distorted when the underlying true random effect β_1 in a single group of observations is significantly different from the rest of the groups, so the $\beta|\tau$ prior is misspecified. They take a Cut posterior $p_{\text{cut}}(\sigma, \tilde{\beta}|Y)$, replacing the prior $p(\beta|\tau)$ with an improper imputation prior $\tilde{p}(\tilde{\beta}) \propto 1$. This eliminates *prior feedback* as in section 3.2.1.1 and section 3.7.1 using an implicit imputation prior. This prior is conjugate so the random effects $\tilde{\beta}$ are integrated out to give, at the first stage, $p_{\text{cut}}(\sigma \mid Y) = \prod_i p(\sigma_i|Y_i)$. At the second stage the Cut distribution of $\sigma|Y$ is

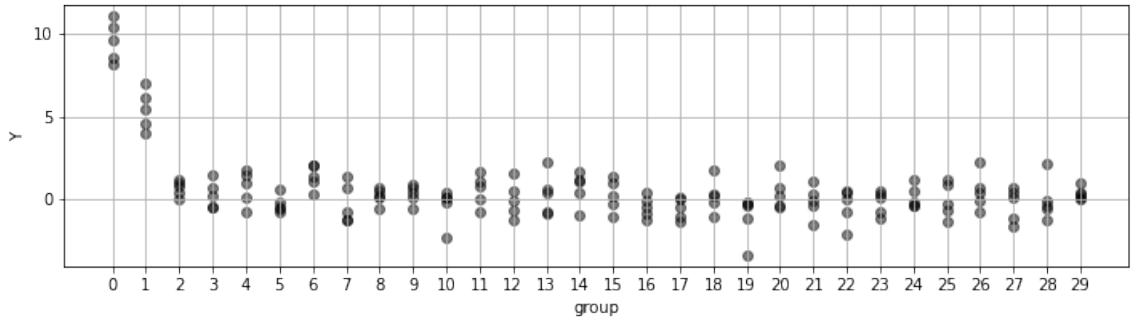


Figure 3.5: Simulated data for the random effects model of section 3.5.2. With $\beta_1^* = 10$, $\beta_2^* = 5$, $\beta_i^* = 0$ for $i = 3, \ldots, N$ and $\sigma_i^* = 1$ for $i = 1, \ldots, N$.

fed into the posterior $p(\beta, \tau | Y, \sigma)$, using the original prior $p(\beta \mid \tau)$ and conditioning on the imputed φ.

In contrast to this all (Bayes) or nothing (Cut) approach, SMI *reduces* the influence of *some* of the groups that may be causing contamination of the posterior. Our modules are for $i = 1, ..., N$ the separate generative models for (Y_i, β_i, σ_i) and $(\beta_i, \sigma_i, \tau)$ so we have in effect $2N = 60$ modules, and we work with the joint distributions involving β. The marginals would still be available, but the higher dimensional joint distributions have interesting shapes and present more of a challenge. Our target SMI posterior is

$$p_{smi,\eta}(\sigma, \beta, \tau, \tilde{\beta}, \tilde{\tau} \mid Y) = p_{pow,\eta}(\sigma, \tilde{\beta}, \tilde{\tau} \mid Y)\, p(\beta, \tau \mid \sigma, Y), \tag{3.5.1}$$

where

$$p_{pow,\eta}(\sigma, \tilde{\beta}, \tilde{\tau} \mid Y) \propto p(Y \mid \tilde{\beta}, \sigma)\, p(\tilde{\tau} \mid \sigma) \prod_{i=1}^{N} \tilde{p}_{\eta_i}(\tilde{\beta}_i \mid \tilde{\tau}), \tag{3.5.2}$$

$$p(\beta, \tau \mid \sigma, Y) \propto p(Y \mid \beta, \sigma)\, p(\tau \mid \sigma) \prod_{i=1}^{N} p(\beta_i \mid \tau) \tag{3.5.3}$$

Performing Semi-Modular Inference using eq. (3.5.1) entails a 30-dimensional influence parameter $\eta \in H$ with $H = [0, 1]^{30}$. The SMI imputation is

$$(\sigma, \tilde{\beta}, \tilde{\tau}) \sim p_{pow,\eta}(\sigma, \tilde{\beta}, \tilde{\tau} \mid Y),$$

and the analysis is

$$(\beta, \tau \mid \sigma) \sim p(\beta, \tau \mid \sigma, Y).$$

The modulated priors $\tilde{p}_{\eta_i}$ used at the imputation stage are chosen to interpolate between the same cut prior $\tilde{p}(\tilde{\theta}) \propto 1$ and analysis priors as before. Following the

discussion in section 3.7.1, we define the modulated imputation prior as a normal density

$$\tilde{p}_{\eta_i}(\tilde{\beta}_i \mid \tilde{\tau}) = \mathcal{N}(\tilde{\beta}_i; 0, \tilde{\tau}/\eta_i),$$

and $\tilde{p}_\eta(\tilde{\beta} \mid \tilde{\tau}) = \prod_i \tilde{p}_{\eta_i}(\tilde{\beta}_i \mid \tilde{\tau})$. This parameterisation gives the Cut prior $\tilde{p}_\eta(\tilde{\beta} \mid \tilde{\tau}) \to \tilde{p}(\tilde{\beta})$ as $\eta \to 0$ and the Bayes prior $\tilde{p}_\eta(\tilde{\beta} \mid \tilde{\tau}) \to p(\tilde{\beta}|\tilde{\tau})$ as $\eta \to 1$.

Our goal is to optimally modulate the feedback from each group into the shared distribution of the β's. Accurate estimates of the utility $U(\eta)$ in eq. (3.3.39) are needed in order to locate the maximum-utility influence-vector η^* in eq. (3.3.40). We could approximate the modular posterior, using either nested-MCMC or variational SMI, at a lattice of η-values in $H = [0, 1]^N$, but this quickly becomes impractical with increasing N. Instead we approximate the candidate SMI posteriors at *all* η in a single function by learning $q_{f_{\hat{\alpha}}(\eta)}$ and $q_{\hat{\alpha},\eta}$, the flow- and map-based Variational Meta-Posteriors. We found we could do this fairly accurately with the essentially same VMP-NSF-MLP meta-variational setup we used in section 3.5.1. The inputs to the flow are $\epsilon_1 \sim N(0, \mathbb{I}_N)$ (these express σ), $\epsilon_2 \sim N(0, \mathbb{I}_{N+1})$ (expressing β, τ) and $\epsilon_3 \sim \epsilon_2$ (expressing $\tilde{\beta}, \tilde{\tau}$). Runtimes for algorithm 3 are manageable as simulation of the VMP,

$$(\sigma, \beta, \tau, \tilde{\beta}, \tilde{\tau}) \sim q_{f_\alpha(\eta)} \tag{3.5.4}$$

at given α, η is fast: set $(\lambda_1, \lambda_2, \lambda_3) = f_\alpha(\eta)$ and then compute $\sigma = T_1(\lambda_1, \epsilon), (\beta, \tau) = T_2(\lambda_2, \epsilon)$ and $(\tilde{\beta}, \tilde{\tau}) = T_2(\lambda_3, \epsilon)$ using the deterministic flow mapping in eq. (3.3.11). Simulation of $q_{\hat{\alpha},\eta}$ is slightly faster.

Our results at $\eta = \mathbf{0}_N, \mathbf{1}_N$ are qualitatively consistent with the Cut and Bayes analyses in Jacob et al. (2017, Sec. 4.4). We took two misspecified effects rather than one, but in other respects the setup is the same. Samples from the exact $p_{\text{smi},\eta}$ posterior produced via MCMC are shown in fig. 3.13 for a selection of variable pairs. Comparing these with the corresponding distributions given by VMP-NSF-MLP in fig. 3.6 using a VMP-map, we see good agreement across each of the three rows/η-configurations, despite the highly irregular contour shapes. We emphasise that *all* samples are produced from a *single* VMP $q_{f_{\hat{\alpha}}(\eta)}$, and we just plug in different η-values to get different rows in fig. 3.6. The VMP-flow density $q_{\hat{\alpha},\eta}$ converged much more rapidly to agreement with the ground truth than did the VMP-map density $q_{f_{\hat{\alpha}}(\eta)}$, but $\mathcal{Q}_{H,\hat{\alpha}}$ approximates $\mathcal{P}_{\text{smi}}$ accurately in both cases. We omit the corresponding VMP-flow plots as they are essentially identical.

Selected components of the VMP-map $f_{\hat{\alpha}}(\eta)$ are shown in fig. 3.7. The surfaces in the top row show a complex structure across the two axes: varying influence

parameters in the misspecified groups has a significant impact on the variational posteriors. However, the surface in the bottom row of plots is almost constant with η_3: cutting one of these "good" groups with labels $3, ..., 30$ doesn't have a strong impact on the variational posterior.

Producing nested-MCMC plots for comparison with the corresponding VMPs at a few η-values is undemanding. However, this is where the contribution from MCMC ends. Accurate estimation of the the utility $U(\eta) = \text{elpd}(\eta)$ over $\eta \in H$ becomes prohibitively expensive using two-stage nested-MCMC methods. The posterior predictive $p(y, z|Y, Z)$ in eq. (3.3.39) must in general be estimated using samples from the VMP. Although the density $q_{\hat{\alpha},\eta}$ is available in closed form and can be sampled independently, it is nevertheless a complicated function. However, as noted above, the simulation in eq. (3.5.4) is fast. In fig. 3.9 we plot (negative) elpd(η)-surfaces using the operational WAIC-estimator. We check this estimate using direct simulation of synthetic data $y' \sim p^*(\cdot)$. In the top row, reducing the feedback from the two misspecified modules improves predictive performance (the ELPD is larger at smaller η_1, η_2-values). In the bottom row, where we vary $(\eta_1, \eta_3) \in [0, 1]$, the rates for one misspecified and one well-specified group, we see that the ELPD surface is relatively flat for η_3, the influence a well-specified group, though trending up with increasing η_3.

Figure 3.8 illustrates the initialisation stage for the minimisation of the WAIC. The second SGD-stage on the WAIC target terminates quickly from this initialisation. The estimated optimal values $\hat{\eta}^*$ (rounded the first decimal place) are

$$\hat{\eta}^* = (0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.9, 1, 1, 0.9, 1, 1, 1, 0.8, 1, 1, 1, 1, 1, 1, 1).$$

This result shows the method is working as information from the first two modules is cut while the rest are Bayes or close to Bayes. This is as expected as we have synthetic data with modules 1 and 2 misspecified. The resulting $\hat{\eta}^*$ gives a $p_{\text{smi},\hat{\eta}^*}$ which is hard to distinguish from the $p_{\text{smi},\eta}$ posterior at $\eta = (0, 0, 1, ..., 1)$ (shown in fig. 3.4) as the SMI-posterior is insensitive to changes to $\eta_3, ..., \eta_N$ close to Bayes values.

3.6 Discussion

We have given variational families and loss functions for approximating Cut- and SMI-posterior distributions. Much of the presentation is agnostic to the details of the variational family. However, we focus on parameterisation based on NFs as this overcomes many well known weaknesses of variational inference. We saw no sign of

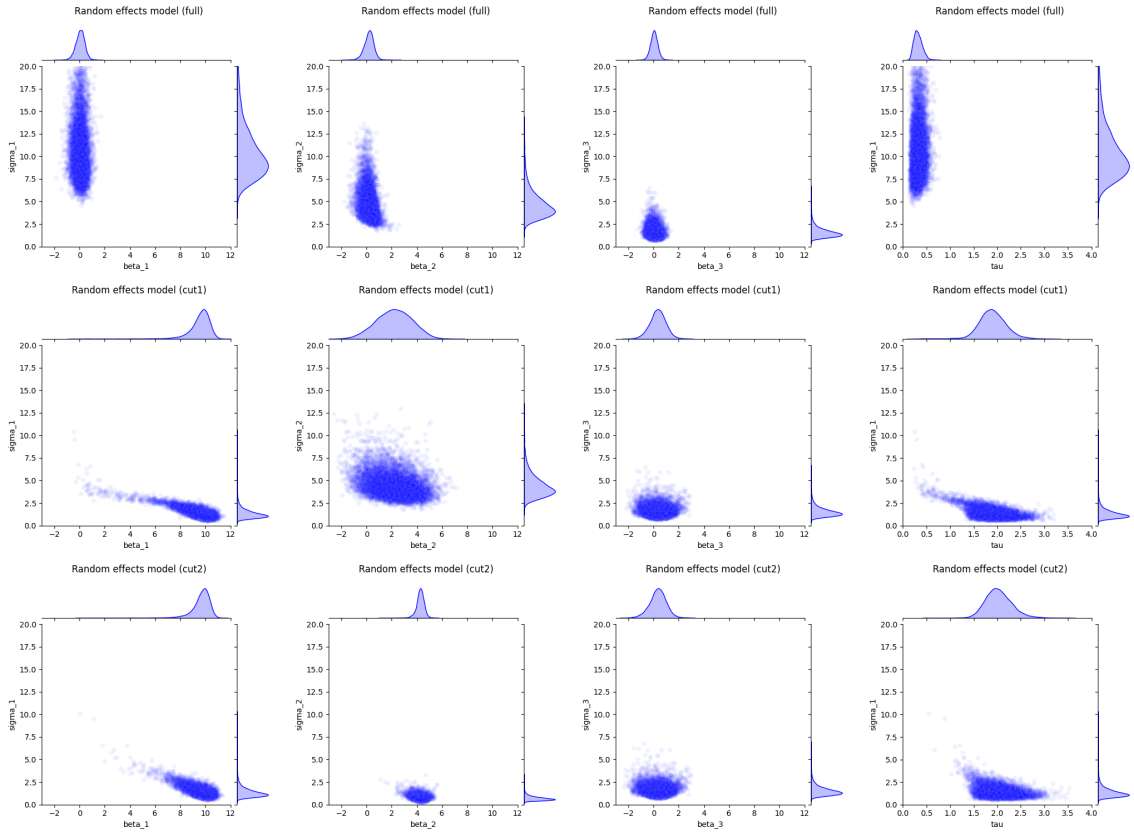


Figure 3.6: Samples from the Variational Meta-Posterior of the Random Effects model, obtained from a VMP-NSF-MLP architecture. Each graph shows the joint distribution of a selected pair of parameters. Rows correspond to three modular *feedback* configurations between groups: (Top row) Bayes, $\eta_1 = ... = \eta_{30} = 1$; (Middle) One Cut module, $\eta_1 = 0$, $\eta_2 = ... = \eta_{30} = 1$; (Bottom) Two Cut Modules, $\eta_1 = \eta_2 = 0$, $\eta_3 = ... = \eta_{30} = 1$. Compare MCMC in fig. 3.13.

underdispersion relative to an MCMC-baseline in our examples. In contrast MFVI approximated the η-dependent mean of the $p_{\text{smi},\eta}$ target well but was significantly underdispersed.

The loss function we use (really two loss functions) in definition 3.3.1 is not the standard KL divergence between the variational density and target, as that loss does not allow control of information flow between modules. Our loss removes dependence on cut modules when we impute parameters in well-specified modules at $\eta = 0$ (the Cut-posterior). Just as $p_{\text{smi},\eta}$ interpolates between the Cut and Bayes posterior distributions, so our variational approximation $q_{\lambda^*(\eta)}$ exactly interpolates between a variational approximation to the Cut due to Yu et al. (2021) and standard variational Bayes. Although the optimized loss need not decrease as we enlarge the variational family, it goes to zero, and we recover the exact target, as the family expands to

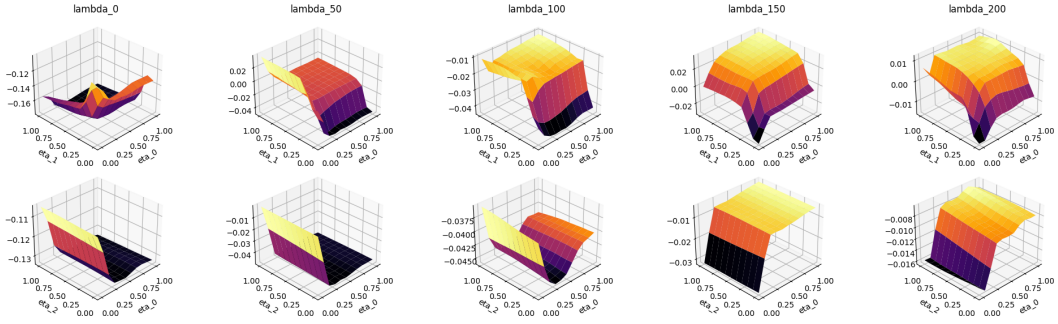


Figure 3.7: Trained VMP-map $\lambda^*(\eta) = f_{\alpha^*}(\eta)$ for selected components of the variational parameter vector $\lambda_i^*(\eta),\ i = 0, 50, 100, 150, 200$ in a Neural Spline Flow. Here $\eta = (\eta_1, \eta_2, \ldots, \eta_{30})$, and we vary two selected η-components at a time, while keeping the rest constant. Top: vary $(\eta_1, \eta_2) \in [0, 1]$ (associated with the misspecified groups) and set $\eta_3 = \ldots = \eta_{30} = 1$. Bottom: vary $(\eta_1, \eta_3) \in [0, 1]$ (one misspecified group, one well-specified) and set $\eta_2 = 0$, $\eta_4 = \ldots = \eta_{30} = 1$.

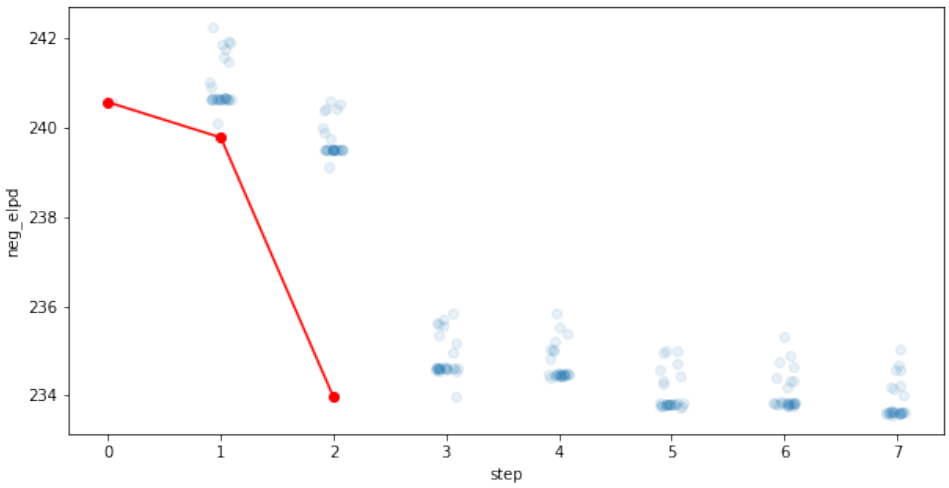


Figure 3.8: Greedy search algorithm to initialize SGD η^*-estimation for the Random Effects model. The red line is the trajectory of the negative ELPD obtained by applying the initialisation algorithm in section 3.4.3. The blue dots represent the $-$ elpd of the "candidate" models at each step for the greedy search, produced by cutting one (additional) module. The search stops after two iterations, cutting the first and second modules, as expected. Each full step takes a fraction of second to compute.

include the target.

In variational SMI our goal is to approximate distributions in the family $\mathcal{P}_{\text{smi}} = \{p_{\text{smi},\eta},\ \eta \in H\}$. We gave a Variational Meta-Posterior $\mathcal{Q}_{H,\alpha} = \{q_{\alpha,\eta},\ \eta \in H\}$ which fits all the distributions in $\mathcal{P}_{\text{smi}}$ at the same time, by taking η as a conditioned quantity in the NF. We called the modified NF the VMP-flow. The SGD in algorithm 3 finds α^* which fits $\mathcal{Q}_{H,\alpha^*}$ to $\mathcal{P}_{\text{smi}}$ in a single joint optimisation for efficient end-to-end

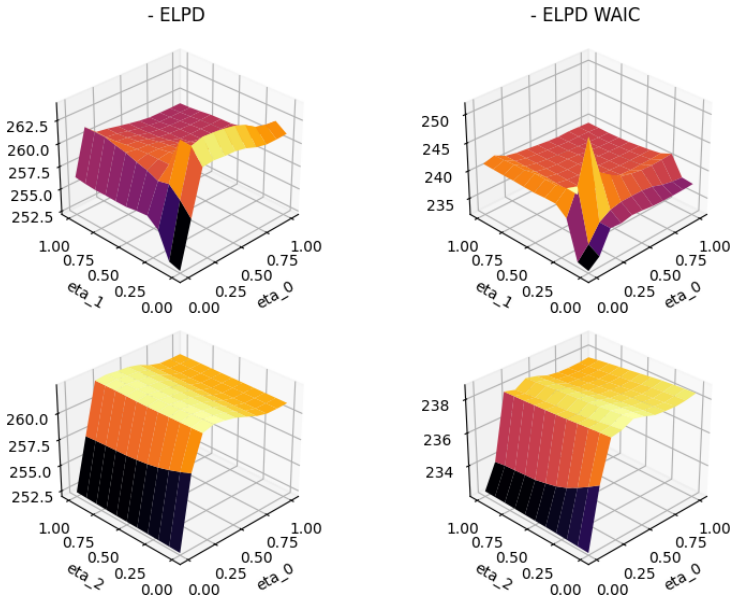


Figure 3.9: Predictive scores comparing performance of the SMI posterior for different values of η in the Random Effects model. We show two approximations to the negative ELPD. (Left column) ELPD estimated using draws from the true generative model. (Right columns) ELPD estimated using the WAIC. These match the left column which acts as a check. (Top row) varying the influence parameters associated with the two misspecified groups, $(\eta_1, \eta_2) \in [0, 1]$, fixing $\eta_2 = \ldots = \eta_{30} = 1$. (Bottom row) varying $(\eta_1, \eta_3) \in [0, 1]$, the rates for one misspecified and one well-specified group, fixing $\eta_2 = 0$ and $\eta_4 = ... = \eta_{30} = 1$.

training. We gave two parameterisations of the VMP. We favor the VMP-flow. Our second parameterisation of the VMP, $\mathcal{Q}_{H,\alpha} = \{q_{\lambda(\alpha,\eta)},\ \eta \in H\}$, trains a VMP-map $f_\alpha(\eta)$ to output the optimal variational parameters $\lambda^*(\eta)$ as functions of η. The two approaches gave similar variational approximations, but training the VMP-flow was faster, as SGD converged more steadily and rapidly with less tuning of optimisation hyper-parameters.

One advantage of using the VMP is that is allows us to modulate feedback from multiple modules at the same time. When we apply a Cut-posterior we have to pre-identify misspecified modules in order to give the locations of cuts. We gave an analysis in which we cut every data-module separately and estimated the associated influence parameters. This allows us to discover rather than pre-identify the cut-modules. Accurate estimation of the ELPD over $\eta \in H$ calls for SMI or variational-SMI posterior samples at C-exponentially many η values for C cuts, and this is prohibitively expensive using one-η-at-a-time methods such as algorithm 2 or nested-MCMC methods.

The main weaknesses of our methods are first, a certain amount of experimentation was needed to find flow architectures that worked well for our targets. However, having found an architecture (with coupling-layer MLP-conditioners and rational spline transformers) that worked well, it worked well for all targets. Further exploration is needed to see if this holds more generally. Tuning of the initialisation and learning rate in algorithm 2 and algorithm 3 were needed also. Another weakness is that the final step of the analysis, after the VMP is trained and we have only to estimate and maximise the ELPD and select the optimal influence parameters η, is not straightforward, at least for high dimensional η. The estimated ELPD is non-convex in η. Working with the VMP makes this step as easy as possible, as sampling the VMP is very fast.

The VMP framework may be useful outside Semi-Modular Inference. Approximating a complete family of models indexed by a set of continuous hyperparameters has potential applications in parameter selection for hyperpriors in standard Variational Bayesian inference. The marginal density $q_{\alpha^*,\eta}(\varphi, \tilde{\theta}),\ \eta \in H$ gives variational approximations to the power posterior distribution $p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)$ at every η at the same time, and this may be all we need if the power posterior rather than SMI is the target.

Acknowledgements

We thank Dennis Prangle, David Nott and Kamélia Daudel for insighful discussions on Variational Methods and Modular Inference.

Research supported with Cloud TPUs from Google's TPU Research Cloud (TRC)

3.7 Supplementary Material

3.7.1 Modulating prior feedback

The material in this section does not address the main point of this paper, which is to present variational inference for SMI. However, the issue we cover here does not seem to have been addressed explicitly in the literature to date, and we need some supporting theory for one of our main examples, in section 3.5.2.

In the standard Cut-model setup given above, a likelihood factor which is present in the analysis stage has been removed in the imputation stage. For example, in eq. (3.2.5), the likelihood $p(Y \mid \varphi, \theta)$ is absent in $p(\varphi \mid Z)$ and present in $p(\theta \mid Y, \varphi)$. In some applications of Cut-posteriors and SMI, going back to Liu et al. (2009) and Jacob

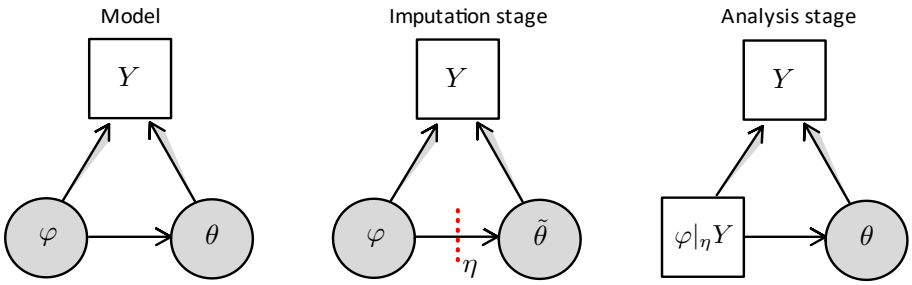


Figure 3.10: Graphical representation of Cut and SMI-posterior inference with a Cut prior. Notation as fig. 3.2. (Left) Graphical model with prior $p(\theta \mid \varphi)$ indicated by arrow $\varphi \to \theta$. (Mid) First (imputation) stage, auxiliary variable $\tilde{\theta}$ introduced for φ imputation. Prior $\varphi|\tilde{\theta}$ modulated as indicated by the red dashed line. (Right) Second (analysis) stage giving inference for θ, now conditioning on the Cut distribution of $\varphi|Y$.

et al. (2017), and including the archaeological applications in Styring et al. (2017), Carmona and Nicholls (2020) and Styring et al. (2022), the misspecified module has no associated data. In these examples the feedback within a prior module is cut. If a Cut is applied to a *prior density* and we simply remove the prior factor at imputation then all that remains in the imputation posterior distribution is the base measure. We have effectively replaced the imputation prior density with a constant, and this may be inappropriate in some settings. However, we are free to choose the imputation prior and we should use this freedom. In this section we discuss Cut-priors as a special case of Cut-posterior inference and point out that the choice of imputation prior must be justified.

In order to see how this works, we take a simply structured example, as we did in fig. 3.1. Consider a generative model of the form

$$\begin{aligned} \varphi &\sim p(\cdot) \\ \theta &\sim p(\cdot \mid \varphi) \\ Y_i &\sim p(\cdot \mid \varphi, \theta), \ i = 1, ..., n. \end{aligned}$$

This is the model represented by the leftmost graph in fig. 3.10. The posterior is

$$p(\varphi, \theta \mid Y) \propto p(Y \mid \varphi, \theta) p(\varphi) p(\theta \mid \varphi). \tag{3.7.1}$$

Suppose we want to cut feedback from θ into φ. In the Cut-posterior for imputation

of φ, the generative model for the data is

$$\begin{aligned}\varphi &\sim p(\cdot)\\ \tilde{\theta} &\sim \tilde{p}(\cdot)\\ Y_i &\sim p(\cdot \mid \varphi, \tilde{\theta}),\ i = 1, ..., n.\end{aligned}$$

This model is represented by the middle graph in fig. 3.10. The new parameter $\tilde{\theta}$ is an auxiliary variable with Cut prior $\tilde{p}(\cdot)$. Some measure for $\tilde{\theta}$ is needed in the imputation of φ as $\tilde{\theta}$ is unknown and present in $p(Y \mid \varphi, \tilde{\theta})$. Work to date takes volume measure in the parameter space of θ without generally remarking that a choice had to be made (for example Equation 17 in Jacob et al. (2017)). Exceptions include Moss and Rousseau (2022), who take a carefully constructed Cut prior $\tilde{p}$, and Styring et al. (2017) and Yu et al. (2021), who remark on the choice. The issue doesn't arise when we remove a likelihood factor, as any parameters enter the likelihood as conditioned variables. In the analysis stage the generative model is

$$\begin{aligned}\theta &\sim p(\cdot \mid \varphi)\\ Y_i &\sim p(\cdot \mid \varphi, \theta),\ i = 1, ..., n.\end{aligned}$$

This model is represented by the rightmost graph in fig. 3.10. The Cut-posterior is

$$p_{\text{cut}}(\varphi, \theta, \tilde{\theta} \mid Y) = p_{\text{cut}}(\varphi, \tilde{\theta} \mid Y) p(\theta \mid Y, \varphi), \tag{3.7.2}$$

where

$$p_{\text{cut}}(\varphi, \tilde{\theta} \mid Y) \propto p(\varphi)\tilde{p}(\tilde{\theta})p(Y \mid \varphi, \tilde{\theta}), \tag{3.7.3}$$

and

$$p(\theta \mid Y, \varphi) = p(\theta \mid \varphi)\frac{p(Y \mid \varphi, \theta)}{p(Y \mid \varphi)}. \tag{3.7.4}$$

The cut prior $\tilde{p}(\tilde{\theta})$ used for imputation in eq. (3.7.3) is a modelling choice, like the Bayes prior $p(\theta \mid \varphi)$ which appears in the posterior. Typically $p(\theta \mid \varphi)$ is a Subjective Bayes prior expressing the relations linking θ and φ available from physical considerations, but is misspecified, and $\tilde{p}(\tilde{\theta})$ will typically be a non-informative Objective Bayes prior.

The loss function $l_{\text{cut}}(Y; \varphi, \theta, \tilde{\theta}, \pi_0)$ for which

$$p_{\text{cut}}(\varphi, \theta, \tilde{\theta} \mid Y) \propto \exp(-l_{\text{cut}}(Y; \varphi, \theta, \tilde{\theta}, \pi_0))\, \pi_0(\varphi, \theta, \tilde{\theta})$$

is a Gibbs posterior is

$$l_{\text{cut}}(Y; \varphi, \theta, \tilde{\theta}, \pi_0) = -\log(p(Y \mid \varphi, \tilde{\theta})) - \log(p(Y \mid \varphi, \theta)) + \log(p(Y \mid \varphi)), \tag{3.7.5}$$

where $\pi_0(\varphi, \theta, \tilde{\theta}) = p(\varphi)\tilde{p}(\tilde{\theta})p(\theta \mid \varphi)$ is the prior which we must specify as part of the loss as it appears in $p(Y \mid \varphi)$ and changes as belief is updated.

This Cut-posterior belief update which cuts feedback in a prior is order coherent in the sense of Bissiri et al. (2016). This is known (Carmona and Nicholls, 2020) for Cut-posteriors with the "standard" setup of section 3.2. However, the Cut-posterior with imputation and analysis priors is qualitatively different.

Proposition 2. *The Cut posterior in eq.* (3.7.2) *with Cut-prior feedback is an order coherent belief update.*

Proof. If we split the data $Y = (Y^{(1)}, Y^{(2)})$ then the imputation prior $\pi_0(\varphi, \tilde{\theta}) = p(\varphi)\tilde{p}(\tilde{\theta})$ is updated to $\pi_1(\varphi, \tilde{\theta}) = p_{\text{cut}}(\varphi, \tilde{\theta} \mid Y^{(1)})$ and the analysis prior $\pi_0(\theta \mid \varphi) = p(\theta \mid \varphi)$ is updated to $\pi_1(\theta \mid \varphi) = p(\theta \mid \varphi, Y^{(1)})$. The update is order coherent if

$$\exp(-l_{\text{cut}}(Y; \varphi, \theta, \tilde{\theta}, \pi_0))\pi_0(\varphi, \theta, \tilde{\theta}) = \exp(-l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1))\pi_1(\varphi, \theta, \tilde{\theta}).$$

Expanding the RHS using the updated prior π_1 and applying eq. (3.7.5),

$$\begin{aligned}
RHS &= \exp(-l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1))p_{\text{cut}}(\varphi, \theta, \tilde{\theta} \mid Y^{(1)}) \\
&\propto \exp(-l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1)) \ \times \ p(\varphi)\tilde{p}(\tilde{\theta})p(Y^{(1)} \mid \varphi, \tilde{\theta}) \ \times \ p(\theta \mid \varphi)\frac{p(Y^{(1)} \mid \varphi, \theta)}{p(Y^{(1)} \mid \varphi)} \\
&= \exp(-l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1) - l_{\text{cut}}(Y^{(1)}; \varphi, \theta, \tilde{\theta}, \pi_0))p(\varphi)\tilde{p}(\tilde{\theta})p(\theta \mid \varphi) \\
&= LHS
\end{aligned}$$

where the last step holds so long as

$$l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1) + l_{\text{cut}}(Y^{(1)}; \varphi, \theta, \tilde{\theta}, \pi_0) = l_{\text{cut}}(Y; \varphi, \theta, \tilde{\theta}, \pi_0).$$

This is the property Nicholls et al. (2022) call prequential additivity. It is easilly verified here. By independence $p(Y \mid \varphi, \theta) = p(Y^{(2)} \mid \varphi, \theta)p(Y^{(1)} \mid \varphi, \theta)$ so,

$$\begin{aligned}
l_{\text{cut}}(Y^{(2)}; \varphi, \theta, \tilde{\theta}, \pi_1) + l_{\text{cut}}(Y^{(1)}; \varphi, \theta, \tilde{\theta}, \pi_0) = &- \log(p(Y \mid \varphi, \tilde{\theta})) - \log(p(Y \mid \varphi, \theta)) \\
&+ \log(p(Y^{(2)} \mid \varphi, Y^{(1)})) + \log(p(Y^{(1)} \mid \varphi))
\end{aligned}$$

with

$$\begin{aligned}
\log(p(Y^{(2)} \mid \varphi, Y^{(1)})) + \log(p(Y^{(1)} \mid \varphi)) &= \log\left(\frac{p(Y^{(2)}, \varphi, Y^{(1)})p(\varphi)}{p(\varphi, Y^{(1)})p(\varphi)}\right) + \log(p(Y^{(1)} \mid \varphi)) \\
&= \log(p(Y \mid \varphi)).
\end{aligned}$$

□

When we do SMI with modulated priors we interpolate between the Bayes posterior in eq. (3.7.1) and Cut posterior in eq. (3.7.2). Let $p_\eta(\tilde{\theta} \mid \varphi)$ be a family of probability densities indexed by η and satisfying $p_{\eta=0}(\tilde{\theta} \mid \varphi) = \tilde{p}(\tilde{\theta})$ (modulated prior equals Cut prior) and $p_{\eta=1}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)$ (modulated prior equals Bayes prior). We take

$$p_{\text{smi},\eta}(\varphi, \theta, \tilde{\theta} \mid Y) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y) p(\theta \mid Y, \varphi) \tag{3.7.6}$$

where in this setting with a Cut prior

$$p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y) \propto p(Y \mid \varphi, \tilde{\theta}) p(\varphi) p_\eta(\tilde{\theta} \mid \varphi),$$

and $p(\theta \mid Y, \varphi)$ is given in eq. (3.7.4), so the second stage analysis is the Cut posterior analysis. This ensures $p_{\text{smi},0} = p_{\text{cut}}$ and $p_{\text{smi},1}$ gives the Bayes posterior. Taking a normalised family $p_\eta(\tilde{\theta} \mid \varphi)$, $\eta \in [0, 1]$ of interpolating priors ensures that the marginal prior for $p(\varphi)$ in the imputation doesn't depend on η. An un-normalised family such as $p_\eta(\varphi, \tilde{\theta}) \propto \tilde{p}(\tilde{\theta})^{1-\eta} p(\tilde{\theta} \mid \varphi)^\eta$ has all the desired interpolating properties, but the marginal $p(\varphi)$ in the imputation stage will then depend on η. In some settings (for example when working with normal priors with fixed variance) the two prior parameterisations may be equivalent as η scales the variance.

SMI with modulated prior feedback is an order coherent belief update. We now give proposition 3, stated in section 3.2.

Proposition 3. *The SMI posterior in eq.* (3.2.12) *with cut prior feedback is an order coherent belief update.*

Proof. Replace $\tilde{p}(\tilde{\theta}) \to p_\eta(\tilde{\theta} \mid \varphi)$ in the proof of proposition 2. □

In contrast to the tempered likelihood in eq. (3.2.10) used in Cut-likelihood SMI in eq. (3.2.8), where the tempered likelihood is not normalised over Y, the modulated prior $p_\eta(\tilde{\theta} \mid \varphi)$ is a normalised distribution over $\tilde{\theta}$. However, SMI in eq. (3.7.6) with a modulated prior is not simply Bayesian inference with a revised prior, as the priors in the imputation and analysis stages are not the same (unless $\eta = 1$).

3.7.2 Variational families

3.7.3 Choice of transformations for SMI

The transformations T_1 and T_2 introduced in eqs. (3.3.2) and (3.3.11) determine our flow-based variational approximation to the SMI-posterior. We make a few standard assumptions about these transformations:

- They are diffemorphisms, meaning that they must be differentiable, invertible and their inverse must be differentiable (needed as we wish to be able to evaluate, differentiate and sample the density q_λ),
- They map the domain of the base distribution onto the domain of the corresponding model parameter (e.g. if φ represents a probability, we need $\varphi = T(\epsilon_1) \in [0, 1]$),
- They support fast sampling from the flow-based density, and fast computation of the density at the sampled values. This allows efficient computation of the forward transformation and the determinant of its Jacobian matrix.

Additionally, there are two attributes of T_1 and T_2 that are important for an accurate approximation, but not strict requirements, namely:

- They are expressive transformations, ideally Normalizing Flows with the property of being a universal approximator,
- T_2 is defined as a conditional transformation, so that correlation in the posterior of φ and θ can be represented.

These requirements are met by Normalizing Flows (NFs). There are now many ways to define expressive flow-based transformations for T_1 and T_2 (see Kobyzev et al., 2020; Papamakarios et al., 2021, for reviews). For our experiments in section 3.5, we defined these transformations by the composition of multiple *coupling* layers based on the Neural Spline Flow (NSF) (Durkan et al., 2019), followed by a last layer that maps to the domain of the model parameters. In order to allow posterior correlation between φ and θ, we defined T_2 as a *conditional* transformation, so that the elementwise transformation from ϵ_2 to θ (and from ϵ_3 to $\tilde{\theta}$) also depends on ϵ_1.

3.7.4 Normalizing Flows.

3.7.4.1 Conditioners and transformers in a general flow

We now define the maps $T_1 : \Re^{p_\phi} \to \Re^{p_\phi}$ and $T_2 : \Re^{p_\theta + p_\phi} \to \Re^{p_\theta}$ in the normalising flows in terms of their transformers and conditioners. The material in the section is based on Papamakarios et al. (2021) and further detail may be found there. Let $x = (x_1, ..., x_p)$ and $v = (v_1, ..., v_q)$ be generic real vectors and let

$$x'_i = \tau(x_i, h_i)$$

transform x_i using parameter vectors $h_i \in \Omega_h,\ i = 1, ..., p$. Here

$$h_i = c_i(x_{<i}; v, w_i)$$

is the output of a conditioner parameterised by a vector w_i with c_i taking as input the conditioning variables $x_{<i} = (x_1, ..., x_{i-1})$ and v. Note that the conditioner argument $x_{<i}$ in c_i changes dimension as we step through the components $i = 1, ..., p$ but all the component updates condition on a common set of shared variables v. In our setting the conditioners $c_i : \Re^{i-1+q} \to \Omega_h$ are MLPs and w_i are the weights in the i'th net. The transformer is a strictly monotonic function. For example, if it is affine then $h_i = (m_i, s_i)$ and $\tau(x_i, h_i) = m_i + s_i x_i$. We experimented with a range of transformers and settled on *rational-quadratic spline transformers* (Durkan et al., 2019).

One pass over the flow composes these maps

$$x_i \leftarrow \tau(x_i, c_i(x_{<i}; v, w_i)) \quad \text{for } i = 1, ..., p$$

to update all components $x \to x'$. Denote by $g : \Re^{p+q} \to \Re^p,\ x' = g(x; v, w)$ a map formed in this way, with w the set of all parameters present in the composition of maps. We make multiple passes over the components, with independent sets of parameters w, permuting the indices in order to get an expressive flow. The autoregressive dependence of x'_i on $x_{<i}$ and the monotone transformer ensure the map is invertible,

$$x = g^{-1}(x'; v, w)$$

given the values of the shared conditioning variables v and parameters w, with a tractable lower triangular Jacobian.

In a variant of this setup we used a composition of *coupling layer conditioners* (Dinh et al., 2016) to define g. These have the advantage that both density evaluation and sampling are fast operations, important for fitting the flow to the target, and then sampling the related flow to estimate the utility. In this case we set $x'_i = x_i,\ i = 1, ..., d$ with $d = \lfloor p/2 \rfloor$ and

$$x'_i = \tau(x_i, h_i), \quad \text{with} \quad h_i = c_i(x_{\leq d}; v, w_i) \qquad \text{for } i = d+1, ..., p.$$

The Jacobian determinant from one application of the map is just $\prod_{i>d} |\partial \tau(x_i, h_i)/\partial x_i|$. Again we compose the maps over different permutations of the indices of x (eight times) so that each entry appears in the conditioner and as output from the transformer to get an overall map $g(x; v, w)$.

3.7.4.2 Normalizing flows for variational SMI

In variational-SMI in section 3.3.3, the NF has the form

$$\begin{aligned}\varphi_{(\lambda_1,\epsilon)} &= T_1(\epsilon_1;\lambda_1)\\ &= g(\epsilon_1;\emptyset,\lambda_1)\end{aligned}$$

as λ_1 parameterises the conditioners of φ with no additional conditioners,

$$\begin{aligned}\theta_{(\lambda_2,\epsilon)} &= T_2(\epsilon_2;\lambda_2,\epsilon_1)\\ &= g(\epsilon_2;\epsilon_1,\lambda_2)\end{aligned}$$

as θ is conditioned on ϵ_1 and hence φ and

$$\begin{aligned}\tilde{\theta}_{(\lambda_3,\epsilon)} &= T_2(\epsilon_3;\lambda_3,\epsilon_1)\\ &= g(\epsilon_3;\epsilon_1,\lambda_3),\end{aligned}$$

similarly $\tilde{\theta}$, and together

$$T(\epsilon;\lambda) = \left(T_1(\epsilon_1;\lambda_1),\ T_2(\epsilon_2;\lambda_2,\epsilon_1),\ T_2(\epsilon_3;\lambda_3,\epsilon_1)\right). \tag{3.7.7}$$

These equations define the maps in eq. (3.3.11). When we want to train this to approximate $p_{\text{smi},\eta}$ we simply retrain at each η where $q_{\lambda^*(\eta)}$ is needed.

3.7.4.3 Parameterising the VMP-map

In the VMP, we consider two ways to parameterise the flow defining the meta-posterior. In the first, in eq. (3.4.3) in section 3.4.2, we define a function (a MLP)

$$f_\alpha(\eta) = (f^{(1)}_{\alpha_1}(\eta), f^{(2)}_{\alpha_2})(\eta), f^{(3)}_{\alpha_3}(\eta))$$

in which the output components correspond to $(\lambda_1(\eta),\lambda_2(\eta),\lambda_3(\eta))$ at any $\eta\in[0,1]^C$. The corresponding mappings in terms of the NFs and their inputs are

$$\begin{aligned}\varphi_{(\lambda_1(\alpha_1,\eta),\epsilon)} &= T_1(\epsilon_1;\lambda_1(\eta))\\ &= g(\epsilon_1;\emptyset,f^{(1)}_{\alpha_1}(\eta))\\ \theta_{(\lambda_2(\alpha_2,\eta),\epsilon)} &= T_2(\epsilon_2;\lambda_2(\eta),\epsilon_1)\\ &= g(\epsilon_2;\epsilon_1,f^{(2)}_{\alpha_2}(\eta))\\ \tilde{\theta}_{(\lambda_3(\alpha_3,\eta),\epsilon)} &= T_2(\epsilon_3;\lambda_3(\eta),\epsilon_1)\\ &= g(\epsilon_3;\epsilon_1,f^{(3)}_{\alpha_3}(\eta)).\\ T(\epsilon;\lambda(\eta)) &= \left(T_1(\epsilon_1;\lambda_1(\eta)),\ T_2(\epsilon_2;\lambda_2(\eta),\epsilon_1),\ T_2(\epsilon_3;\lambda_3(\eta),\epsilon_1)\right).\end{aligned} \tag{3.7.8}$$

This keeps the conditioner-transformer relations unchanged from eq. (3.7.7), and simply injects the right parameters λ into the flow to express $p_{\text{smi},\eta}$ at any particular η.

A perfectly trained universal VMP-map $f_\alpha(\eta)$ would inject an optimal set $\lambda^*(\eta)$ of parameters into the Normalizing Flow at every η. We assume $f_\alpha(\eta)$ is a continuous function of η, motivated by the discussion at the start of section 3.4.1. We require $f_\alpha(\eta)$ to be differentiable almost everywhere in α for the purpose of optimisation. These are minimal assumptions. In order to get properties (P1-3) in this setting, suppose we have $\lambda^*(\eta) \in \Lambda^*$ where $\Lambda^*(\eta)$ is given in definition 3.3.1 (at η) and $f_\alpha(\eta)$ is continuously differentiable in α at each η. Referring to the loss in eq. (3.3.31), let

$$A^* = \{\alpha \in A : \mathcal{L}^{(\text{msmi}-map)}(\alpha) = \min_{a \in A} \mathcal{L}^{(\text{msmi}-map)}(a)\}$$

be the set of optimal α-values. If for some $\alpha^* \in A$, the VMP-map $f_{\alpha^*}(\eta)$ expresses the function $\lambda^*(\eta)$ perfectly at $\eta \in \eta_{1:R}$, that is if $f_{\alpha^*}(\eta_r) \in \Lambda^*(\eta_r),\ r = 1, ..., R$ then

$$A^* = \{\alpha \in A : f_\alpha(\eta_r) \in \Lambda^*(\eta_r),\ r = 1, ..., R\}$$

since these solutions in α and no others minimise $\mathcal{L}^{(\text{smi},\eta)}$ at every η in the sum in eq. (3.4.7). In this case we recover $\lambda^* \in \Lambda^*$ and (P1-3) at each $\eta \in \eta_{1:R}$.

For our examples in section 5, we found three Multi-Layer Perceptrons (MLPs) in parallel, trained with input η and output $f^{(k)}_{\alpha_k}(\eta),\ k = 1, 2, 3$ expressed the non-linear relationships holding between variational parameters, and scaled well with increasing numbers of cuts (dimension of input $\eta \in H$). We made experiments with VMP-maps parameterised with Gaussian Processes (GPs) (Rasmussen and Williams, 2005) and Cubic Splines (Hastie et al., 2001) which we do not report. These may be convenient when the dimension of H is small.

We found in training f_α that α should be initialised to output a constant function $f_{\alpha_0}(\eta) \approx \lambda_0$ independent of η. Here λ_0 are variational parameters obtained in a "pre-training stage" using algorithm 2 for a fixed central value of η. This strategy significantly reduces training time and improves convergence.

3.7.4.4 Parameterising the VMP-flow

The other way we parameterise the meta-posterior is to modify the conditioner. Consider the coupling layer setup. In order to express the η-dependence (in say the coupling layer conditioner), $x'_i = x_i,\ i = 1, ..., d$ and

$$x'_i = \tau(x_i, h_i), \quad h_i = c_i(x_{\leq d}; v, w_i) + c'(\eta; w'_i), \quad \text{for } i = d+1, ..., p.$$

Here η enters via a second additive conditioner c' (which is not indexed by i as it always takes the same conditioning variable η but has parameters w'_i which vary across variables $x_i,\ i > d$). Like v, η is a common conditioner in every transform. When we compose this map to form the overall map, $g(x; (v, \eta), (w, w'))$ say, *each* application of the map has its own sets of parameters w_i and w'_i *for each* $i = d+1, ..., p$. When we parameterise the meta-posterior we write $\alpha_k = (\lambda_k, \mu_k),\ k = 1, 2, 3$ and

$$\begin{aligned}
\varphi_{(\alpha_1,\eta,\epsilon)} &= T_1(\epsilon_1; \alpha_1, \eta) \\
&= g(\epsilon_1; (\emptyset, \eta), (\lambda_1, \mu_1)), && (3.7.9) \\
\theta_{(\alpha_2,\eta,\epsilon)} &= T_2(\epsilon_2; \alpha_2, (\eta, \epsilon_1)) \\
&= g(\epsilon_2; (\epsilon_1, \eta), (\lambda_2, \mu_2)), && (3.7.10) \\
\tilde{\theta}_{(\alpha_3,\eta,\epsilon)} &= T_2(\epsilon_3; \alpha_3, (\eta, \epsilon_1)) \\
&= g(\epsilon_3; (\epsilon_1, \eta), (\lambda_3, \mu_3)), && (3.7.11) \\
T(\epsilon; \alpha, \eta) &= (T_1(\epsilon_1; \alpha_1, \eta),\ T_2(\epsilon_2; \alpha_2, (\eta, \epsilon_1)),\ T_2(\epsilon_3; \alpha_3, (\eta, \epsilon_1)))\,. && (3.7.12)
\end{aligned}$$

so that $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ contain the w-parameters of the "old" conditioner c_i in each of the flows for φ, θ and $\tilde{\theta}$ respectively and $\mu = (\mu_1, \mu_2, \mu_3)$ are the corresponding w'-parameters of the new η-dependent conditioner c'. These are collectively the VMP parameters $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. We take $\mu \in M$ and $\alpha \in A$ with $A = \Lambda \times M$. The influence parameter η enters through the conditioner so it is associated with ϵ_1 as a conditioner for θ and $\tilde{\theta}$ and as a conditioner on its own for φ.

The family of variational densities $q_{\alpha,\eta}(\varphi, \theta, \tilde{\theta})$ defined by the VMP-flow are given in eq. (3.4.4). The optimal parameters $\alpha^* = (\lambda^*, \mu^*)$, $\alpha^* \in A$, minimise a loss $\mathcal{L}^{(\text{msmi}-flow)}(\alpha)$ which is closely related to $\mathcal{L}^{(\text{msmi}-map)}(\alpha)$ in eq. (3.4.7). In more detail,

$$\mathcal{L}^{(\text{msmi}-flow,\eta)}(\alpha) = \text{ELBO}_{\text{pow},\eta}(\alpha_1, \alpha_3) + \text{ELBO}_{\text{bayes}_{\not\nabla(\varphi)}}(\alpha_1, \alpha_2) \tag{3.7.13}$$

where

$$\text{ELBO}_{\text{pow},\eta}(\alpha_1, \alpha_3) = \mathbb{E}_{(\varphi,\tilde{\theta})\sim q_{\alpha_1,\alpha_3,\eta}}[\log p_{\text{pow},\eta}(\varphi, \tilde{\theta}, Z, Y) - \log q_{\alpha_1,\alpha_3,\eta}(\varphi, \tilde{\theta})] \tag{3.7.14}$$

$$\text{ELBO}_{\text{bayes}_{\not\nabla(\varphi)}}(\alpha_1, \alpha_2, \eta) = \mathbb{E}_{(\varphi,\theta)\sim q_{\alpha_1,\alpha_2,\eta}}[\log p(\not\nabla(\varphi), \theta, Z, Y) - \log q_{\alpha_1,\alpha_2,\eta}(\not\nabla(\varphi), \theta)]. \tag{3.7.15}$$

The loss in eq. (3.4.7) given for the map becomes for the flow,

$$\mathcal{L}^{(\text{msmi}-flow)}(\alpha) = \mathbb{E}_{\eta\sim\rho}\left(\mathcal{L}^{(\text{msmi}-flow,\eta)}(\alpha)\right).$$

In order to apply the reparameterisation trick, the expressions for $\varphi_{(\alpha_1,\eta,\epsilon)}, \theta_{(\alpha_2,\eta,\epsilon)}$ and $\tilde{\theta}_{(\alpha_3,\eta,\epsilon)}$ in eq. (3.7.9)-eq. (3.7.11) are substituted into eq. (3.7.14) and eq. (3.7.15) so that gradients in (α_1, α_3) and α_2 can be taken inside the expectation which is now over $\epsilon \sim p(\cdot)$.

3.7.5 Optimisation of the VMP loss

Here we give the algorithm we use to estimate the the parameters α of the VMP-map and -flow. Algorithm 3 gives SGD for the VMP-map loss $\mathcal{L}^{(\text{msmi}-map)}(\alpha)$ (notation $\mathcal{L}^{(\text{msmi})}(\alpha)$ as the algorithm for optimisation of $\mathcal{L}^{(\text{msmi}-flow)}(\alpha)$ is similar).

3.7.6 Proofs for variational SMI properties

Proposition 4. *The divergence defined in eq.* (3.3.20) *can be written*

$$d(q_\lambda, \mathcal{F}_{smi,\eta}) = D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta}) \parallel p_{pow,\eta}(\varphi, \tilde{\theta} \mid Y, Z)),$$

and hence does not depend on λ_2.

Proof. Since $\tilde{q} \in \mathcal{F}_{\text{smi},\eta}$ it can be written $\tilde{q}(\varphi, \theta, \tilde{\theta}) = p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)\tilde{q}(\theta \mid \varphi, \tilde{\theta})$ where $\tilde{q}(\theta \mid \varphi, \tilde{\theta})$ is an arbitrary conditional density. The divergence is

$$\begin{aligned}
\min_{\tilde{q}\in\mathcal{F}_{\text{smi},\eta}} D_{KL}(q_\lambda \parallel \tilde{q}) &= \min_{\tilde{q}(\theta|\varphi,\tilde{\theta})} D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta})q_{\lambda_2}(\theta \mid \varphi) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)\tilde{q}(\theta \mid \varphi, \tilde{\theta})) \\
&= D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)) \\
&\quad + \min_{\tilde{q}(\theta|\varphi,\tilde{\theta})} E_{\varphi\sim q_{\lambda_1}}[D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel \tilde{q}(\theta \mid \varphi, \tilde{\theta}))] \\
&= D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z))
\end{aligned}$$

as the variation over $q(\theta \mid \varphi, \tilde{\theta})$ is over *all* conditional densities with optimum $\tilde{q}(\theta \mid \varphi, \tilde{\theta}) = q_{\lambda_2}(\theta \mid \varphi)$ since the expectation is non-negative and zero in that case. □

Consider property (P1). If $\lambda^* \in \Lambda^*$ with $\lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*)$ are some generic fitted variational parameters, then $q_{\lambda_1^*}(\varphi)$ cannot depend in any way on $p(Y \mid \varphi, \theta)$ at $\eta = 0$, as the power posterior in eq. (3.3.22) is

$$p_{\text{pow},\eta=0}(\varphi, \tilde{\theta} \mid Y, Z) = p(\varphi \mid Z)\, p(\tilde{\theta} \mid \varphi).$$

The Y observation observation model doesn't enter eq. (3.3.22) at $\eta = 0$. Under an additional assumption on the variational family, we can remove any $p(\tilde{\theta} \mid \varphi)$-dependence (so q_{λ^*} is "completely independent of the generative model" at $\eta = 0$).

Proposition 6. *Variational SMI satisfies property (P1) at $\eta = 0$: If the set*

$$\Lambda^*_{(3)} = \{\lambda_3 \in \Lambda_2 : q_{\lambda_3^*}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)\}$$

is non-empty and we set

$$\Lambda^*_{(1)} = \{\lambda_1 \in \Lambda_1 : D_{KL}(q_{\lambda_1}(\varphi) \parallel p(\varphi \mid Z)) = d^*_{cut}\}$$

with d^*_{cut} defined in (P1) then $\Lambda^*_{(1,3)}$ defined in eq. (3.3.22) *satisfies*

$$\Lambda^*_{(1,3)} = \Lambda^*_{(1)} \times \Lambda^*_{(3)},$$

so $q_{\lambda_1^*}$ does not depend in any way on $p(Y \mid \varphi, \theta)$ or $p(\theta \mid \varphi)$ at $\eta = 0$.

Proof. At $\eta = 0$, $p_{\text{pow},\eta=0}(\varphi, \tilde{\theta} \mid Y, Z) = p(\varphi \mid Z)\, p(\tilde{\theta} \mid \varphi)$ in proposition 4. By eq. (3.3.22),

$$\begin{aligned}
d^*_{\text{smi}} &= \min_{(\lambda_1,\lambda_3)\in B} D_{KL}(q_{\lambda_1,\lambda_3} \parallel p(\varphi \mid Z)\, p(\theta \mid \varphi)) \\
&= \min_{(\lambda_1,\lambda_3)\in B} \Big(D_{KL}(q_{\lambda_1}(\varphi) \parallel p(\varphi \mid Z)) + E_{\varphi\sim q_{\lambda_1}}[D_{KL}(q_{\lambda_3}(\tilde{\theta} \mid \varphi) \parallel p(\tilde{\theta} \mid \varphi))] \Big) \\
&= \min_{\lambda_1\in\Lambda_1} D_{KL}(q_{\lambda_1}(\varphi) \parallel p(\varphi \mid Z)) \\
&= d^*_{cut}
\end{aligned}$$

as the expectation is zero (and hence minimised for every argument) when $\lambda_3^* \in \Lambda_2$ satisfies $q_{\lambda_3^*}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)$. Such a λ_3^* exists because $\Lambda^*_{(3)}$ is non-empty by the assumption in the proposition. It follows that λ minimises $d(q_\lambda, \mathcal{F}_{\text{smi},\eta})$ at $\eta = 0$ if and only if $(\lambda_1^*, \lambda_3) \in \Lambda^*_{(1)} \times \Lambda^*_{(3)}$. □

The point here is that the auxiliary variable $\tilde{\theta}$ is present only through its prior in the power posterior at the Cut, $\eta = 0$, but this factor is perfectly expressed by a corresponding factor in the variational approximation, and hence doesnt enter the λ_1-variation. The condition that there is $\lambda_3^* \in \Lambda_2$ such that $q_{\lambda_3^*}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)$ is met by choosing $\epsilon_2, \epsilon_3 \sim p(\cdot \mid \varphi)$, the prior distribution for θ and $\tilde{\theta}$. We can then find λ_3^* to give $T_2(\epsilon_3; \lambda_3^*, \epsilon_1) = (\epsilon_1, \epsilon_3)$ equal to the identity map (possible in a flow-parameterised map, but not in general in MFVI). At this λ_3-value, $\tilde{\theta} = \epsilon_3$ and hence $q_{\lambda_3^*}(\tilde{\theta} \mid \varphi) = p(\tilde{\theta} \mid \varphi)$. If we have a cut prior as in section 3.7.1 then take $\epsilon_2, \epsilon_3 \sim \tilde{p}(\cdot)$, the Cut prior.

The variational approximation to the Cut-posterior defined in proposition 6 is similar to that given in Yu et al. (2021). We focus on flow-based parameterisations of the variational density q_λ, but apart from this our methods coincide at $\eta = 0$.

Consider now property (P2). Taking $\eta = 1$, the power posterior is the Bayes posterior, so eqs. (3.3.6) and (3.3.22) are identical optimisation problems as the λ-dependence is the same. However, this shows that $q_{\lambda_1^*,\lambda_3^*}(\varphi,\tilde{\theta})$ is variational Bayes at $\eta = 1$, and we have to check that $q_{\lambda_1^*,\lambda_2^*}(\varphi,\theta)$ is variational Bayes.

Proposition 7. *Variational SMI satisfies property (P2). Let*

$$\Lambda^*_{(1)} = \bigcup_{(\lambda_1^*,\lambda_3^*)\in\Lambda^*_{(1,3)}} \{\lambda_1^*\}.$$

The set of Bayes and SMI variational posteriors for φ,θ are the same, that is,

$$\bigcup_{\lambda_1^*\in\Lambda^*_{(1)}} \bigcup_{\lambda_2^*\in\Lambda^*_{(2)}(\lambda_1^*)} \{(\lambda_1^*,\lambda_2^*)\} = B^*,$$

when $\eta = 1$.

Proof. When $\eta = 1$ the power posterior for $\varphi,\tilde{\theta}$ is the Bayes posterior,

$$p_{\mathrm{pow},\eta=1}(\varphi,\tilde{\theta} \mid Y,Z) = p(\varphi \mid Y,Z)\, p(\tilde{\theta} \mid Y,\varphi)$$

so eq. (3.3.22) is the same as variational Bayes as determined by eq. (3.3.6). Since $q_{\lambda_3}(\tilde{\theta} \mid \varphi)$ and $q_{\beta_2}(\theta \mid \varphi)$ have the same parameterisation, we have $B^* = \Lambda^*_{(1,3)}$. For any fixed $\lambda_1^* \in \Lambda^*_{(1)}$ let

$$\Lambda^*_{(3)}(\lambda_1^*) = \{\lambda_3 \in \Lambda_2 : D_{KL}(q_{\lambda_1^*,\lambda_3}(\varphi,\tilde{\theta}) \parallel p_{\mathrm{pow},\eta}(\varphi,\tilde{\theta} \mid Y,Z)) = d^*_{\mathrm{smi}}\}$$

so that

$$\Lambda^*_{(1,3)} = \bigcup_{\lambda_1\in\Lambda^*_{(1)}} \bigcup_{\lambda_3\in\Lambda^*_{(3)}(\lambda_1)} \{(\lambda_1,\lambda_3)\} \tag{3.7.19}$$

Fixing $\lambda_1^* \in \Lambda^*_{(1)}$, and taking $\eta = 1$,

$$\begin{aligned} D_{KL}(q_{\lambda_1^*,\lambda_3}(\varphi,\tilde{\theta}) \parallel p_{\mathrm{pow},\eta}(\varphi,\tilde{\theta} \mid Y,Z)) &= D_{KL}(q_{\lambda_1^*} \parallel p(\varphi \mid Y,Z)) \\ &\quad + \quad E_{\varphi\sim q_{\lambda_1^*}}[D_{KL}(q_{\lambda_3}(\tilde{\theta} \mid \varphi) \parallel p(\tilde{\theta} \mid Y,\varphi))] \end{aligned}$$

and the target level is

$$\begin{aligned} d^*_{\mathrm{smi}} &= D_{KL}(q_{\lambda_1^*} \parallel p(\varphi \mid Y,Z)) + \min_{\lambda_3\in\Lambda_2} E_{\varphi\sim q_{\lambda_1^*}}[D_{KL}(q_{\lambda_3}(\tilde{\theta} \mid \varphi) \parallel p(\tilde{\theta} \mid Y,\varphi))] \\ &= D_{KL}(q_{\lambda_1^*} \parallel p(\varphi \mid Y,Z)) + D^*_{\mathrm{smi}}(\lambda_1^*), \end{aligned}$$

so cancelling the common $D_{KL}(q_{\lambda_1^*} \parallel p(\varphi \mid Y,Z))$ term,

$$\begin{aligned} \Lambda^*_{(3)}(\lambda_1^*) &= \{\lambda_3 \in \Lambda_2 : E_{\varphi\sim q_{\lambda_1^*}}[D_{KL}(q_{\lambda_3}(\tilde{\theta} \mid \varphi) \parallel p(\tilde{\theta} \mid Y,\varphi))] = D^*_{\mathrm{smi}}(\lambda_1^*)\} \\ &= \Lambda^*_{(2)}(\lambda_1^*), \end{aligned} \tag{3.7.20}$$

from eq. (3.3.23), so at $\eta = 1$,

$$\bigcup_{\lambda_1 \in \Lambda^*_{(1)}} \bigcup_{\lambda_2 \in \Lambda^*_{(2)}(\lambda_1)} \{(\lambda_1, \lambda_2)\} = \bigcup_{\lambda_1 \in \Lambda^*_{(1)}} \bigcup_{\lambda_3 \in \Lambda^*_{(3)}(\lambda_1)} \{(\lambda_1, \lambda_3)\}$$

by eq. (3.7.20), so the LHS is equal $\Lambda^*_{(1,3)}$ by eq. (3.7.19) and we saw that $\Lambda^*_{(1,3)} = B^*$, giving the set relation claimed in the proposition. We conclude from this that at $\eta = 1$, $q_{\lambda_1^*, \lambda_2^*}(\varphi, \theta)$ is a marginal variational SMI posterior if and only if it is also a variational Bayes posterior. □

Proposition 8. *Variational SMI satisfies property (P3). If $p_{smi,\eta} \in \mathcal{Q}$ then $q_{\lambda^*} = p_{smi,\eta}$ for $\lambda^* \in \Lambda^*$.*

Proof. This is usually immediate for standard variational methods but has to be checked here. If $p_{\text{smi},\eta} \in \mathcal{Q}$ then there exist λ_1, λ_3 such that $q_{\lambda_1,\lambda_3} = p_{\text{pow},\eta}$ and λ_2 such that $q_{\lambda_2}(\theta \mid \varphi) = p(\theta \mid Y, \varphi)$ and since these choices minimise the KL-divergences in propositions 4 and 5 they are the optimal values, so $q_{\lambda^*} = p_{\text{smi},\eta}$. □

Proposition 9. *The set Λ^* in definition 3.3.1 is the set of solutions of $\nabla_\lambda \mathcal{L}^{(smi,\eta)} = 0$ corresponding to minima.*

Proof. Assuming the flow-based construction of the variational family q_λ defined in eq. (3.3.12), the gradients of $\mathcal{L}^{(\text{smi},\eta)}$ are

$$\begin{aligned}
\nabla_{\lambda_1} \mathcal{L}^{(\text{smi},\eta)} &= \mathbb{E}_{\epsilon \sim p(\epsilon)}[\nabla_\varphi \left\{ \log p(Z \mid \varphi) + \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\varphi) \right\} \nabla_{\lambda_1} \{\varphi\} \\
&\qquad + \nabla_{\lambda_1} \log |J_{T_1}(\epsilon_1)|] \qquad (3.7.21) \\
\nabla_{\lambda_2} \mathcal{L}^{(\text{smi},\eta)} &= \mathbb{E}_{\epsilon \sim p(\epsilon)}[\nabla_\theta \left\{ \log p(Y \mid \varphi, \theta) + \log p(\theta \mid \varphi) \right\} \nabla_{\lambda_2} \{\theta\} \\
&\qquad + \nabla_{\lambda_2} \log |J_{T_2}(\epsilon_1, \epsilon_2)|] \qquad (3.7.22) \\
\nabla_{\lambda_3} \mathcal{L}^{(\text{smi},\eta)} &= \mathbb{E}_{\epsilon \sim p(\epsilon)}[\nabla_{\tilde{\theta}} \left\{ \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\tilde{\theta} \mid \varphi) \right\} \nabla_{\lambda_3} \{\tilde{\theta}\} \\
&\qquad + \nabla_{\lambda_3} \log |J_{T_2}(\epsilon_1, \epsilon_3)|]. \qquad (3.7.23)
\end{aligned}$$

Note that the derivation of the gradients incorporate the effect of the stop-gradient operator.

The system of equations determined by $\nabla_\lambda \mathcal{L}^{(\text{smi},\eta)} = 0$ is the same system defined in section 3.3.3.4 for Λ^*, namely, eqs. (3.3.28) to (3.3.30). □

Proposition 10. *Let $\mathcal{L}^*(v) = \min_{\lambda \in \Lambda} \mathcal{L}^{(v)}(\lambda)$ and*

$$\Lambda^*(v) = \{\lambda \in \Lambda : \mathcal{L}^{(v)}(\lambda) = \mathcal{L}^*(v)\}.$$

Under regularity conditions on $\mathcal{F}_{smi,\eta}$ and $p_{smi,\eta}$ given in proposition 11, for every solution $\lambda^* \in \Lambda^*$ in definition 3.3.1 and all sufficiently small $v \geq 0$ there exists a unique continuous function $\lambda^*(v)$ satisfying $\lambda^*(v) \in \Lambda^*(v)$ and

$$\lim_{v \to 0} \lambda^*(v) = \lambda^*.$$

Proof. Take the definition of $\mathcal{L}^{(v)}(\lambda)$ in eq. (3.3.34), use proposition 4 to replace d and eq. (3.3.15) to expand $D_{KL}(q_\lambda \parallel p_{\text{smi},\eta})$. This gives

$$\begin{aligned}\mathcal{L}^{(v)}(\lambda) = (1+v) D_{KL}(q_{\lambda_1,\lambda_3}(\varphi, \tilde{\theta}) \parallel p_{\text{pow},\eta}(\varphi, \tilde{\theta} \mid Y, Z)) \\ + \quad v \cdot E_{\varphi \sim q_{\lambda_1}}[D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi))]\end{aligned}$$

Minima $\lambda^{(v)} \in \Lambda^*(v)$ are stationary points of $\mathcal{L}^{(v)}$ so they solve $\nabla_\lambda \mathcal{L}^{(v)}(\lambda) = 0$ with positive curvature. Using the reparameterisation trick these equations are, for $v > 0$,

$$\begin{aligned}0 = \mathbb{E}_{\epsilon \sim p(\epsilon)}[&(1+v) \nabla_\varphi \left\{ \log p(Z \mid \varphi) + \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\varphi) \right\} \nabla_{\lambda_1}\{\varphi\} \\ &+ v \nabla_\varphi \left\{ \log p(Y \mid \varphi, \theta) - \log p(Y \mid \varphi) \right\} \nabla_{\lambda_1}\{\varphi\} \\ &+ \nabla_{\lambda_1} \log |J_{T_1}(\epsilon_1)|] \qquad (3.7.24) \\ 0 = \mathbb{E}_{\epsilon \sim p(\epsilon)}[&\nabla_\theta \left\{ \log p(Y \mid \varphi, \theta) + \log p(\theta \mid \varphi) \right\} \nabla_{\lambda_2}\{\theta\} \\ &+ \nabla_{\lambda_2} \log |J_{T_2}(\epsilon_1, \epsilon_2)|] \qquad (3.7.25) \\ 0 = \mathbb{E}_{\epsilon \sim p(\epsilon)}[&\nabla_{\tilde{\theta}} \left\{ \eta \log p(Y \mid \varphi, \tilde{\theta}) + \log p(\tilde{\theta} \mid \varphi) \right\} \nabla_{\lambda_3}\{\tilde{\theta}\} \\ &+ \nabla_{\lambda_3} \log |J_{T_2}(\epsilon_1, \epsilon_3)|] \qquad (3.7.26)\end{aligned}$$

The system of equations eqs. (3.7.24) to (3.7.26) converges to the system of equations eqs. (3.3.26) and (3.3.27) as $v \to 0$. Under regularity conditions set out in proposition 11 below, the solutions converge, in the sense that every point in Λ^* is the limit as $v \to 0$ of some continuous sequence of solutions to $\nabla_\lambda \mathcal{L}^{(v)} = 0$. □

We now state the regularity conditions and show solutions converge. Recall that $\Lambda_1 = \Re^{L_1}$ and $\Lambda_2 = \Re^{L_2}$ with $L = L_1 + 2L_2$ so that $\dim(\Lambda) = L$. The system of equations eqs. (3.7.24) to (3.7.26) has the form $F(\lambda, v) = \mathbf{0}_L$ where

$$F(\lambda, v) = f(\lambda) + v \cdot g(\lambda)$$

with $f, g : \Re^L \to \Re^L$ and $F : \Re^{L+1} \to \Re^L$. Equations eqs. (3.3.28) to (3.3.30) are equivalent to $F(\lambda, 0) = \mathbf{0}_L$ which is just $f(\lambda) = \mathbf{0}_L$. Now eqs. (3.3.28) and (3.3.30)

are the reparameterisation of eq. (3.3.26) and eq. (3.3.29) is the reparameterisation of eq. (3.3.27) so we identify $f = (f_1, f_2, f_3)$ with

$$\begin{aligned}
f_1(\lambda_1, \lambda_3) &= \nabla_{\lambda_1} D_{KL}(q_{\lambda_1,\lambda_3} \parallel p_{\text{pow},\eta}) \\
f_2(\lambda_1, \lambda_2) &= \nabla_{\lambda_2} E_{\varphi \sim q_{\lambda_1}} D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi)), \\
f_3(\lambda_1, \lambda_3) &= \nabla_{\lambda_3} D_{KL}(q_{\lambda_1,\lambda_3} \parallel p_{\text{pow},\eta})
\end{aligned}$$

where $f_1 : \Re^{L_1+L_2} \to \Re^{L_1}$, $f_2 : \Re^{L_1+L_2} \to \Re^{L_2}$ and $f_3 : \Re^{L_1+L_2} \to \Re^{L_2}$.

Proposition 11. *Assume Λ^* is not empty and let $\lambda^* \in \Lambda^*$ be given. Assume f and g are continuously differentiable in λ at $\lambda = \lambda^*$ and the Hessians of $D_{KL}(q_{\lambda_1,\lambda_3} \parallel p_{pow,\eta})$ (in λ_1, λ_3) and $E_{\varphi \sim q_{\lambda_1}} D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi))$ (in λ_2) are invertible at $\lambda = \lambda^*$. For every $\lambda^* \in \Lambda^*$ there is $\delta > 0$ and a unique continuous function $\lambda^*(v)$ satisfying $F(\lambda^*(v), v) = 0$ for $0 \leq |v| \leq \delta$ and $\lambda^*(0) = \lambda^*$ at $v = 0$.*

Proof. If the Jacobian $\partial F/\partial\lambda$ is invertible at $\lambda = \lambda^*$ then by assumptions of the proposition and the Implicit Function Theorem, there exists $\delta > 0$ and a unique continuous function $\lambda^* : \Re \to \Re^L$ satisfying $F(\lambda^*(v), v) = 0$ for $0 \leq |v| \leq \delta$. It follows that every point in Λ^* is the limit as $v \to 0$ of some continuous sequence of solutions to $\nabla_\lambda \mathcal{L}^{(v)} = 0$.

The Jacobian $\partial F/\partial\lambda = \partial f/\partial\lambda$ at $v = 0$ since the additive term in g does not contribute at $v = 0$. It follows that the Jacobian $\partial F/\partial\lambda$ is invertible at $\lambda = \lambda^*$ if $|\partial f/\partial\lambda|_{\lambda=\lambda^*} \neq 0$. The Jacobian matrix $\partial f/\partial\lambda$ has a block structure,

$$\begin{aligned}
\frac{\partial f}{\partial \lambda} &= \begin{bmatrix} \partial f_1/\partial\lambda_1 & \mathbf{0}_{L_1 \times L_2} & \partial f_1/\partial\lambda_3 \\ \partial f_2/\partial\lambda_1 & \partial f_2/\partial\lambda_2 & \mathbf{0}_{L_2 \times L_2} \\ \partial f_3/\partial\lambda_1 & \mathbf{0}_{L_1 \times L_2} & \partial f_3/\partial\lambda_3 \end{bmatrix} \\
&= \begin{bmatrix} L_1 \times L_1 & L_1 \times L_2 & L_1 \times L_2 \\ L_2 \times L_1 & L_2 \times L_2 & L_2 \times L_2 \\ L_2 \times L_1 & L_2 \times L_2 & L_2 \times L_2 \end{bmatrix}
\end{aligned}$$

with block dimensions in the second line. It follows that the determinant is

$$\left|\frac{\partial f}{\partial \lambda}\right| = \begin{vmatrix} \partial f_1/\partial\lambda_1 & \partial f_1/\partial\lambda_3 \\ \partial f_3/\partial\lambda_1 & \partial f_3/\partial\lambda_3 \end{vmatrix} \times \left|\partial f_2/\partial\lambda_2\right|.$$

This is just the product of the determinants of the Hessians of $D_{KL}(q_{\lambda_1,\lambda_3} \parallel p_{\text{pow},\eta})$ (in λ_1, λ_3) and $E_{\varphi \sim q_{\lambda_1}} D_{KL}(q_{\lambda_2}(\theta \mid \varphi) \parallel p(\theta \mid Y, \varphi))$ (in λ_2), so $\partial F/\partial\lambda$ is invertible if these Hessians are invertible at $\lambda = \lambda^*$. □

3.7.7 Further details of Experiments.

3.7.8 Epidemiological Data

MCMC samples from the posterior distribution of the epidemiological model in section 3.5.1 are shown in fig. 3.11. Samples from the MFVI approximation fitted using variational SMI separately at each η are shown fig. 3.12.

3.7.9 Random effects model

Samples from the posterior distribution of the Random Effects model, obtained via MCMC are shown in fig. 3.13.

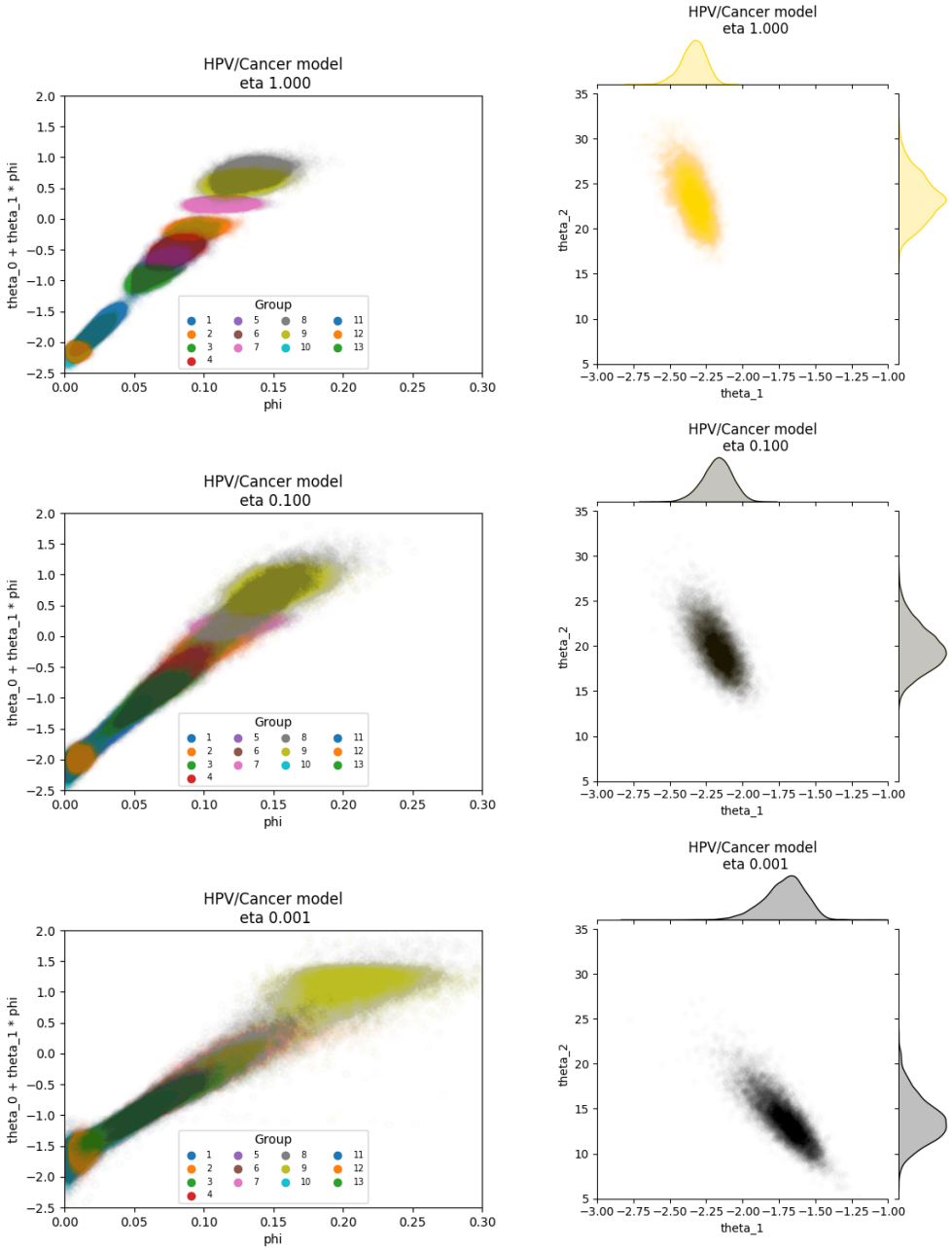


Figure 3.11: Samples from the posterior distribution of the epidemiological model, obtained via MCMC. Rows correspond to three rates of *feedback* from the Poisson module, $\eta = (0.001, 0.1, 1)$. In the left column, we plot the relation between HPV prevalence (ϕ) and cervical cancer incidence (μ) for the 13 groups in the data. In the right column, the joint distribution of slope (θ_1) and intercept (θ_2) of such relation.

Algorithm 3 Variational Meta-Posterior approximation for $\mathcal{P}_{\text{smi}} = \{p_{\text{smi},\eta},\ \eta \in H\}$

Input: $\mathcal{D}$: Data. $p(\varphi, \theta, \mathcal{D})$: Multi-modular probabilistic model. $q_\lambda = (p(\epsilon), T, \lambda)$: Variational family. $f_\alpha(\eta)$: VMP-map. ρ: η-weighting distribution over H.

Output: VMP-map $f_{\hat{\alpha}}(\eta)$ giving the Variational approximation for the $\mathcal{P}_{\text{smi}}$ family.

Initialise mapping parameters α
while SGD not converged **do**
  Sample $\mathcal{D}^{(b)} \sim \mathcal{D}$ (random minibatch of data).
  Sample R values of $\eta_r \sim \rho,\ r = 1, ..., R$.
  for $r = 1, \ldots, R$ **do**
    Obtain variational parameters $\lambda_r = f_\alpha(\eta_r)$.
    for $s = 1, \ldots, S$ **do**
      Sample the base distribution, $\epsilon_{r,s} \sim p(\cdot)$.
      Transform the sampled values $(\varphi_{r,s}, \theta_{r,s}, \tilde{\theta}_{r,s}) \leftarrow T_{\lambda_r}(\epsilon_{r,s})$ as in eq. (3.3.11).
    end for
  end for
  Compute the Monte Carlo estimate of the loss $\mathcal{L}^{(\text{msmi})}$ in eq. (3.4.7) and its gradients.

$$\widehat{\mathcal{L}}^{(\text{msmi})} = \widehat{\text{ELBO}}^{(\text{msmi})}_{\text{pow}} + \widehat{\text{ELBO}}^{(\text{msmi})}_{\not\varphi(\varphi)} \tag{3.7.16}$$

  where

$$\widehat{\text{ELBO}}^{(\text{msmi})}_{\text{pow}} = -\frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S}\left[\log p_{\text{pow},\eta_r}(\varphi_{r,s}, \tilde{\theta}_{r,s}, \mathcal{D}^{(b)}) - \log q(\varphi_{r,s}, \tilde{\theta}_{r,s})\right] \tag{3.7.17}$$

$$\widehat{\text{ELBO}}^{(\text{msmi})}_{\not\varphi(\varphi)} = -\frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S}\left[\log p(\not\varphi(\varphi_{r,s}), \theta_{r,s}, \mathcal{D}^{(b)}) - \log q(\not\varphi(\varphi_{r,s}), \theta_{r,s})\right] \tag{3.7.18}$$

  Update α using the estimated gradient vector $\nabla_\alpha \widehat{\mathcal{L}}^{(\text{msmi})}$
  Check convergence of $q_{f_\alpha(\eta)}(\varphi, \theta)$ for multiple $\eta \in H$
end while
return $\hat{\alpha} = \alpha$

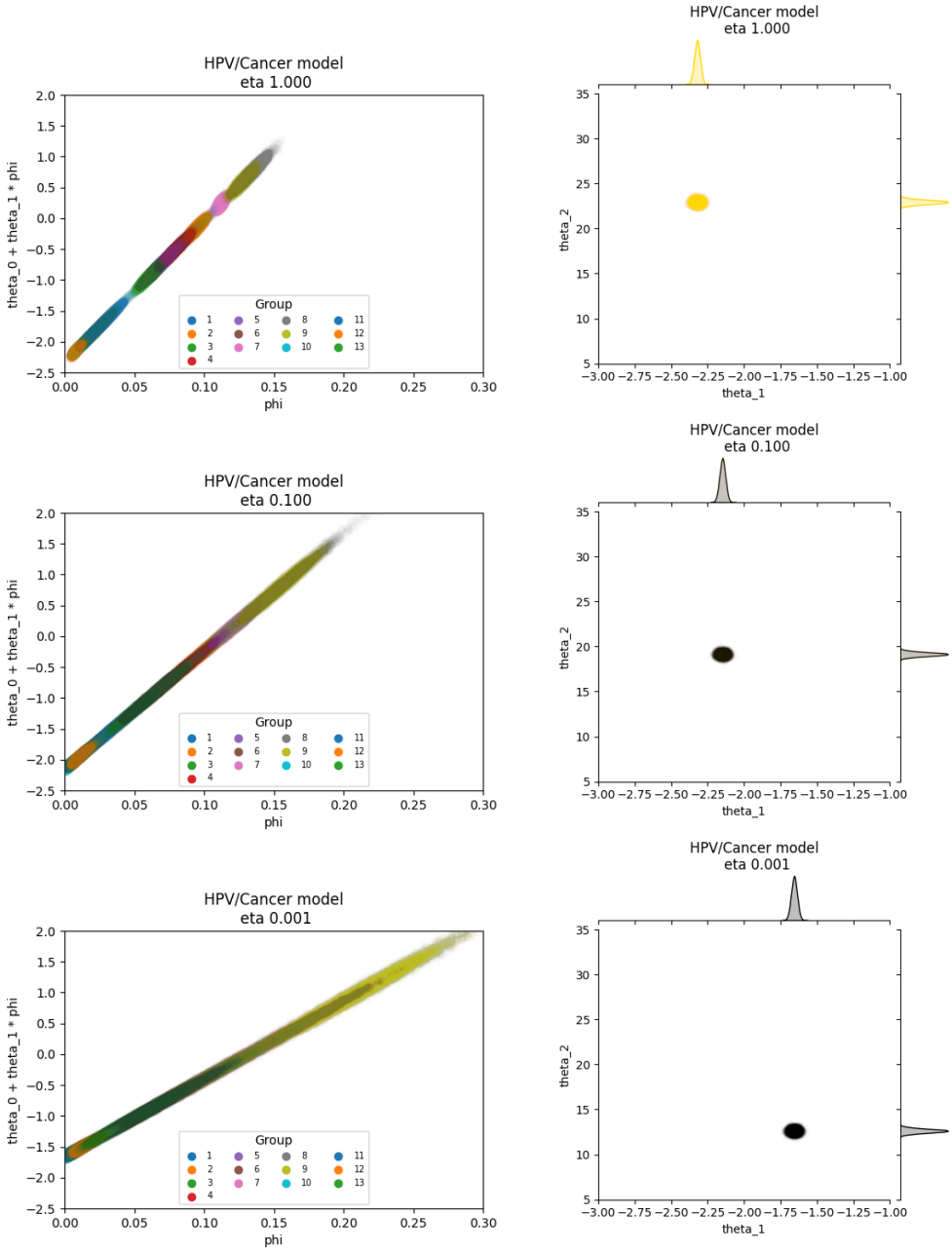


Figure 3.12: Samples from the variational posterior distribution of the epidemiological model, obtained using a Mean-Field approximation. Plots are interpreted as in fig. 3.11. We train one variational posterior for each rate $\eta = (0.001, 0.1, 1)$ separately. The approximations clearly underestimate posterior variance.

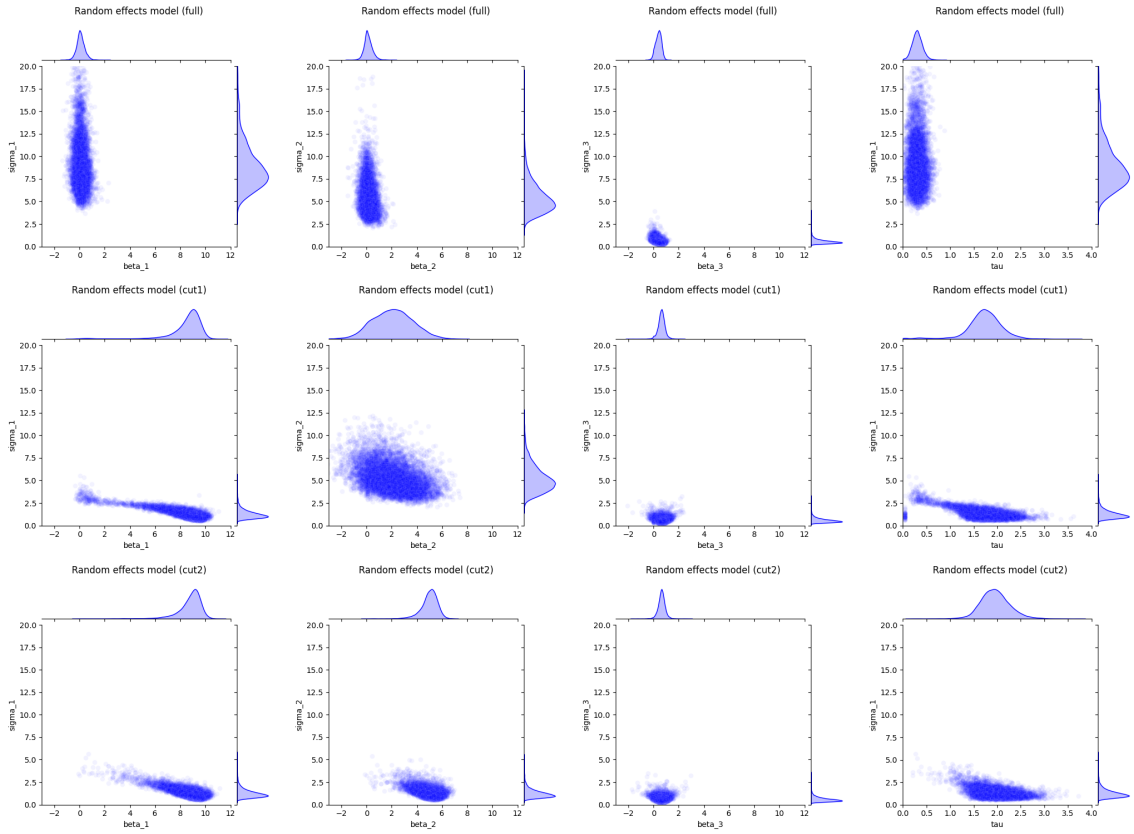


Figure 3.13: Samples from the posterior distribution of the Random Effects model, obtained via MCMC. Each graph shows the joint distribution of a selected pair of parameters. Rows correspond to three modular *feedback* configurations between groups: (Top row) Bayes, $\eta_1 = ... = \eta_{30} = 1$; (Middle) One Cut module, $\eta_1 = 0$, $\eta_2 = ... = \eta_{30} = 1$; (Bottom) Two Cut Modules, $\eta_1 = \eta_2 = 0$, $\eta_3 = ... = \eta_{30} = 1$.

Chapter 4

Statistical Analysis of the Agroecology of Early 'Cities' in Northern Mesopotamia, the Aegean and Central Europe

Abstract

Agricultural extensification refers to an expansive, lowinput production strategy that is land rather than labour limited. Here, we present a robust method, using the archaeological proxies of cereal grain nitrogen isotope values and settlement size, to investigate the relationship between agricultural intensity and population size at Neolithic to Bronze/Iron Age settlement sites in northern Mesopotamia, the Aegean and south-west Germany. We conclude that urban form—in particular, density of occupation—as well as scale shaped the agroecological trajectories of early cities. Whereas high-density urbanism in northern Mesopotamia and the Aegean entailed radical agricultural extensification, lower density urbanism in south-west Germany afforded more intensive management of arable land. We relate these differing agricultural trajectories to long-term urban growth/collapse cycles in northern Mesopotamia and the Aegean, on the one hand, and to the volatility of early Iron Age elite power structures and urban centralization in south-west Germany, on the other.

4.1 Introduction

This document is the statistical supplement to "Urban form and scale shaped the agroecology of early cities in northern Mesopotamia, the Aegean and central Europe". We test for an effect due to site size on manure level in the archaeological sites

presented in the main paper. We consider new datasets from two archaeological contexts for this test: one from Early Neolithic-Late Bronze Age Greece (the `aegean` data), and a second from Early Neolithic-Early Iron Age south-west Germany (the `swgermany` data). The results given in section 4.6.3 are reported in the main paper. Our analysis follows the supplement to Styring et al. (2017) in substance. Those authors test for the same effect in similarly structured archaeological data from urban centres in northern Mesopotamia. We use methods from that source throughout this document. However, we improve on the statistical methodology of Styring et al. (2017) in the section 4.6.

The document is organised as follows. In section 4.2 we summarise the two new archaeological datasets and the modern dataset we use for calibration. In section 4.4 we present a *single imputation* analysis, which introduces the statistical models in a simple inferential framework. In section 4.5 we repeat the test using *Bayesian Multiple Imputation* (BMI), which accounts for uncertainty in the imputation. Up to this point we are using the same methods as Styring et al. (2017) [1]. One weakness of multiple imputation, noted by Styring et al. (2017), is that it usually gives conservative estimates for the strength of any effect. The effect is "diluted" in the sense of Knuiman et al. (1998) due to "imputation noise". This could be ignored by Styring et al. (2017) as the effect-size of interest remained significant despite dilution. In section 4.6 we repeat the analysis using *Semi-Modular Inference* (SMI) (Carmona and Nicholls, 2020). This method is closely related to BMI. It measures model misspecification and down-weights but does not remove misspecified model components at the imputation stage. Down-weighting reduces dilution and gives a more accurate measure of evidence than BMI.

Our analyses show a clear (negative) effect due to site size on manure-level in the `nmeso` and `aegean` data, and a potential positive effect in the `swgermany` data. Along the way we reproduce the results of Styring et al. (2017) using a completely independent code base and get good agreement. The results in this supplement can be reproduced using the accompanying scripts in `R`. Data and functions are bundled in the `agricurbayes` package.

[1] Our model for Bayesian Multiple Imputation differs slightly from Styring et al. (2017), as we incorporate the effect due to date in the `PO` module (See section 4.5). The conclusions and evidence for extensification remain unchanged.

4.2 Data description

4.2.1 Modern calibration data

The modern data consist of agricultural records gathered in experimental settings in which manure level is an independent variable and nitrogen isotope ratio is the measured outcome. This modern calibration dataset is the same as the one used by Styring et al. (2017). The dataset has observations on 268 grain samples. Six variables are of interest here.

normd15n : Measured cereal grain-sample nitrogen isotope ratio.

manure_level : The level of manuring treatment the cereal grain-sample received. A 3 level ordinal variable with levels "low", "medium" and "high".

site : The name of the site where the grain-sample was gathered. A 19-level categorical variable.

year : The year in which the grain-sample was gathered.

category : The grain-sample cereal species. A 2-level variable with levels "barley" and "wheat".

rainfall : annual rainfall estimate for the cereal grain sample site, obtained by interpolation of average monthly local climate data for 1960 - 1990, taken from the WorldClim database.

The data were gathered in experiments conducted under different conditions at different times at different sites. Importantly, the manuring treatments vary qualitatively and quantitatively from site to site. These manure-level records were mapped to a 3-level ordinal scale with levels low, medium and high. In fig. 4.10 of section 4.8.1 we plot pairwise relations between key variables in the modern data. In particular, note the positive association between Nitrogen Isotope (`normd15n`) and Manure levels (`manure_level`), and the negative relation between `normd15n` and rainfall levels. These are further illustrated in fig. 4.1.

```
> mod.dat <- agricurb_data %>%
+   dplyr::filter( dataset=="modern" )
```

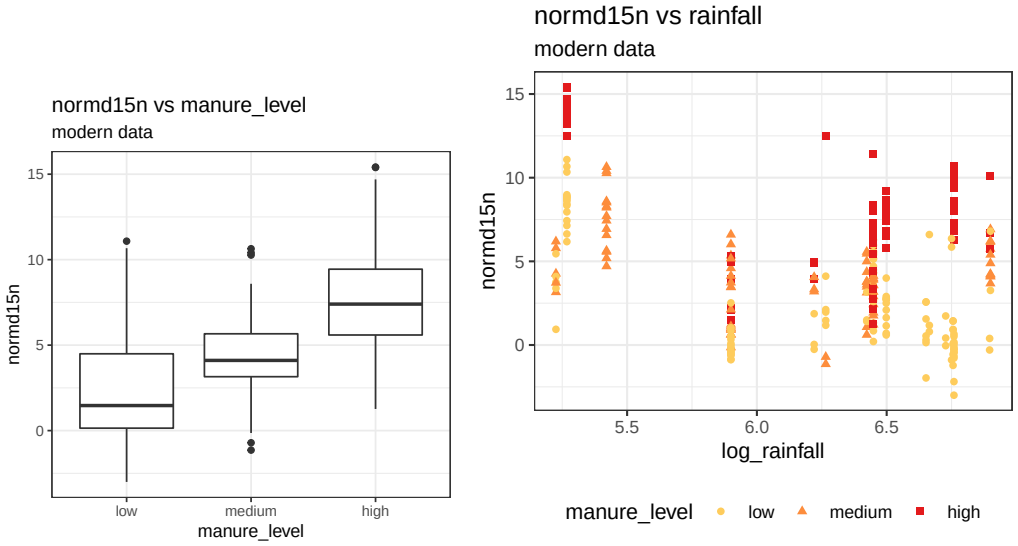


Figure 4.1: Key relations in the modern Data. Left: Positive relation between Nitrogen Isotope (`normd15n`, y-axis) and `manure_level` (x-axis). Right: Negative relation between `normd15n` (y-axis) and `log(rainfall)` (x-axis), with point shapes and colours distinguishing different manure levels.

4.2.2 The Archaeological data

The `aegean` and `swgermany` datasets contain observations on 181 and 435 grain-samples respectively. The variables are the same as those used in Styring et al. (2017) in the Mesopotamian context. We include the Northern Mesopotamia dataset (`nmeso`, 269 observations) studied in Styring et al. (2017) for comparison in all analyses.

normd15n : measured grain-sample nitrogen isotope ratio.

site : site name for sample-context. A categorical variable with a level for each site ID, (5 levels for `nmeso`, 6 for `aegean`, and 9 for `swgermany`).

phase : phase name for grain-sample context. A categorical variable with a level for each phase ID, giving the temporal context of the sample (14 levels for `nmeso`, 31 for `aegean`, and 5 for `swgermany`).

date : date of grain-sample context.

size : size of context-site in hectares at sample-context date.

category : grain-sample species. A 2-level categorical variable with levels "barley" and "wheat".

min_rainfall, max_rainfall : upper and lower bounds on rainfall at grain-sample context at context date.

rainfall : estimated annual rainfall at sample-context at context-date.

Note the negative relation between `normd15n` and `log(rainfall)` in all datasets in fig. 4.2. In contrast the relation between `normd15n` and site `size` shown in fig. 4.3 varies across datasets. For the `nmeso` and `aegean` datasets this relation is negative, whereas for the `swgermany` dataset it may be slightly positive, if there is any relation at all. In figs. 4.12 and 4.13 in section 4.8.1 we plot pairwise relations between all key variables in the `aegean` and `swgermany` datasets.

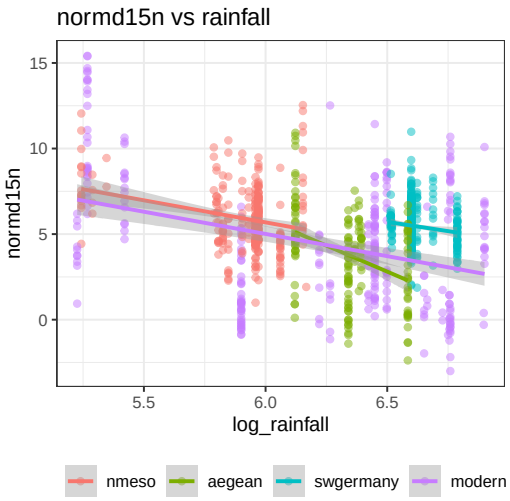


Figure 4.2: Negative relation between levels of Nitrogen isotope and (log) rainfall. Color indicates a dataset. Solid lines regress these two variables, separately for each dataset.

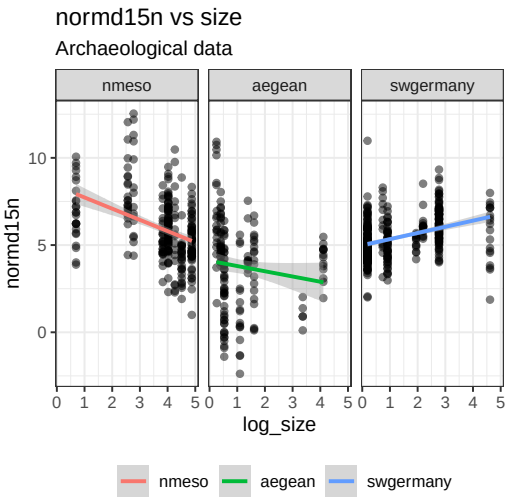


Figure 4.3: Relation between levels of Nitrogen isotope and site size for the three archaeological datasets. Solid lines regress these two variables.

4.3 Statistical Models

Our model definition follows the statistical supplement to Styring et al. (2017).

The model is composed by two modules. The first module takes the nitrogen levels `normd15n` as the continuous response variable and uses four predictors: `manure_level`, grain category, rainfall, and the site of the crop. We use a Hierarchical Gaussian Model for this module, and denote it as `HM` hereafter. The second module takes `manure_level` as a 3-level ordinal response variable, and uses two predictors:

size of the field, and site of the crop. We use a Proportional Odds model (McCullagh, 1980) for this module, and denote it as `PO` hereafter.

The `HM` module allow us to capture the relation between `manure_level` and `normd15n` using the modern data, and to impute the missing Manure Levels in the archaelogical data using this relationship. The `PO` module aims at discovering the relation between the imputed archaelogical `manure_level` and the size of fields, which is the main parameter of interest.

We use this multi-modular model across the following sections, performing inference in three different ways: Single Imputation (section 4.4), Bayesian Multiple Imputation (section 4.5), and Semi-Modular Inference (section 4.6).

An alternative model formulation could incorporate a continuos latent variable for the `manure_level`. A common approach to model ordinal data with a latent continuos variable is by introducing *thresholds* of *cut-off* points (e.g. Carmona et al., 2019; Kottas et al., 2005). For each ordinal manure level m_i, we define a latent variable z_i satisfying $m_i = k$ iff $\gamma_{k-1} < z_i \leq \gamma_k$,for $k = 1, 2, 3$, where $\gamma_0, \ldots, \gamma_3$ are fixed thresholds with $\gamma_0 = -\infty$ and $\gamma_3 = \infty$. Under this formultaion, we would use the `HM` to impute continuous latent manure levels, and the `PO` module would be modified to accomodate for a continuous response, possibly a Gaussian linear model. One disadvantge of this approach is that the estimation of parameters in the `HM` require imputation of all latent manure levels, including those in the modern data.

4.4 Single Imputation Analysis

In this section we use single imputation to test for an effect due to site size on manuring levels. The aim is to introduce the reader to the models and reasoning in a simple inferential setting. The results here give a sanity check for the more sophisticated analysis in the sections to follow.

Our analysis in this section follows Section 3 of the statistical supplement to Styring et al. (2017).

Our plan is (Step 1) use the modern dataset to capture the relation between `manure_level` and `normd15n`. Then (Step 2), invert that relation to estimate the missing `manure_level`-values in the archaeological data from their observed `normd15n` values. This determines `manure_level.imputed` values for the archaeological data. Finally (Step 3), we model the relation between the `manure_level.imputed`-values and `size`, the size of the site associated with the grain-sample context.

4.4.1 (Step 1) Select the model linking normd15n and manure_level

We first fit a model linking normd15n and manure_level. We have the same modern data as Styring et al. (2017) and the same model, a hierarchical normal linear model, denoted HM, regressing normd15n on log(rainfall), with a random effect on the intercept due to site, a random effect due to manure_level within site, and a variance offset for wheat over barley. In R formula notation, (using the nlme package (Pinheiro et al., 2019)):

```
> HMmodel <- nlme::lme( normd15n ~ log_rainfall + manure_level,
+                 random=~1|site/manure_level,
+                 weights=varIdent(form=~1|category),
+                 data=mod.dat, method='REML' )
```

The variable-dependencies in this model were chosen using physical considerations, deviance tests, goodness-of-fit plots and predictive testing as described in Section 3.2.1 and Appendix B of the statistical supplement of Styring et al. (2017). We include our own calculations leading to the same model selected by Styring et al. (2017) in our accompanying scripts. We fit this model to the modern data and estimate model parameters.

4.4.2 (Step 2) Impute the missing manure_level values

Single imputation is carried out as described in Section 3.2.2 of the statistical supplement to Styring et al. (2017). When we fix a value for an unknown manure level we fix the fitted normd15n-value for that observation in the HM module. Imputed manuring level values manure_level.imputed are chosen to minimise the sum of squared differences between fitted and observed values for normd15n.

4.4.3 (Step 3) Test for an association between the imputed manure level and site size

We can now estimate the effect of site size on manure_level.imputed in the archaeological data. We use a *Proportional Odds model* (denoted as PO hereafter) (McCullagh, 1980) with 3-level ordinal response manure_level.imputed and size as a potential explanatory variable. Proportional Odds models are GLM-like models widely used for modelling the dependence of an ordinal response on categorical and continuous covariates. See e.g. Agresti (2002) for more details of the model. HM and PO are submodels or *modules* of the overall model.

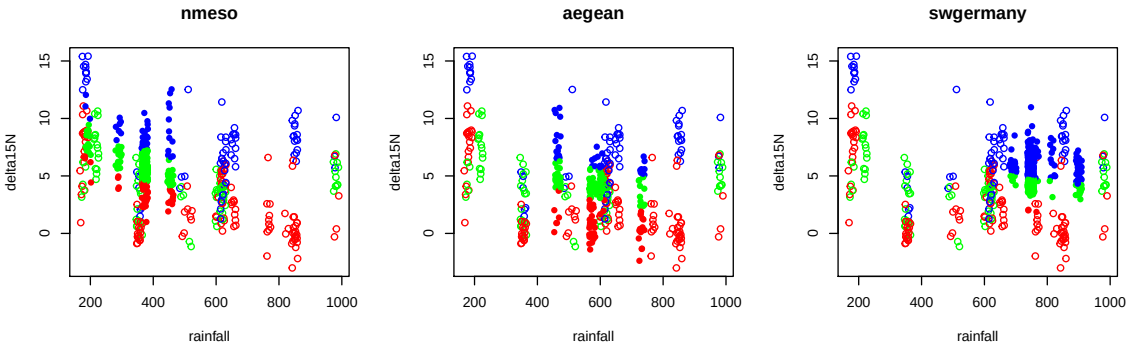


Figure 4.4: Observed `normd15n`-values colored by imputed manure level for the three archaeological datasets. Each panel shows the relation between `normd15n` and log(rainfall); each point corresponds to a single cereal grain-sample. modern(Archaeological) data plotted with open(filled) circles. Colors correspond to manuring level (red=low, green=medium, blue=high). For the modern data the manuring levels are known, for the Archaeological data, these are the `manure_level.imputed`-values.

We carry out model selection on the covariates of the archaelogical data in order to choose the model for `PO`. Physical considerations must play a role as there are many models, and we need to avoid data-dredging. We compare models using information criteria (BIC and AIC) and deviance tests. In (Styring et al., 2017) model selection was carried out on the `nmeso` dataset. Effects due to `date` and random effects due to `phase` must be investigated on physical grounds. However these effects were not supported by the data in the `nmeso` case. The final selected model has a fixed effect due to `size`, a random effect on the `site` id, and a logit link function. In R formula notation (using the ordinal package (Christensen, 2010)),

```
> PO_nmeso = ordinal::clmm( manure_level.imputed ~ scale(size) + (1|site
+                           link = "logit",
+                           data = agricurb_data %>%
+                             dplyr::filter( dataset=="nmeso" ) %>%
+                             dplyr::mutate( site=as.factor(site) ) )
> coefficients(summary(PO_nmeso))

Estimate Std. Error    z value      Pr(>|z|)
low|medium   -1.5326947  0.4107875 -3.731113 0.00019063570
medium|high   1.6673236  0.4150418  4.017242 0.00005888314
scale(size) -0.4636204  0.1715456 -2.702607 0.00687980288
```

The first two rows in the output table give the `PO` module intercepts separating the levels of `manure_level.imputed`, and fit the simple overall proportion of `manure_level.imputed`-values in each level. The estimated coefficient for `size`, -0.4636, is negative and significant (p-value=0.0069). This means that the probability for a higher manure level tends to decrease as site-size increases, in agreement with Styring et al. (2017), and supporting their main finding of agricultural extensification in the `nmeso` dataset.

We applied the same reasoning to our two new datasets. For the `aegean` dataset, we arrive at a similar model, with the difference that `size` is included on a logarithmic scale for best fit.

```
> PO_aegean = ordinal::clmm( manure_level.imputed ~ scale(log_size) + (1
+                            link = "logit",
+                            data = agricurb_data %>%
+                              dplyr::filter( dataset=="aegean" ) %>%
+                              dplyr::mutate( site=as.factor(site) ) %>%
+                              dplyr::mutate( phase=as.factor(phase) ) )
> coefficients(summary(PO_aegean))

                Estimate Std. Error   z value   Pr(>|z|)
low|medium      -0.9288111  0.5528567 -1.680021 0.09295315
medium|high      1.6158724  0.5673630  2.848040 0.00439894
scale(log_size) -0.5076650  0.2335632 -2.173566 0.02973775
```

The estimated coefficient for `size`, -0.5077, is negative and significant (p-value=0.0297), providing evidence that manure level outcome decreases with increasing site size at this site as well.

We found no effect due to `date` in the `aegean` data. However, a model with random effects on `site` and `phase` but not `size` performs as well as `PO_aegean` given above. We have then to choose between `size` and `phase`. The experimental design in the `aegean` data makes it hard to disentangle the two variables. This is due to colinearity between `size` and `site:phase`: there are few observations at each level defined by `site:phase` and they share the same value of `size` within each level. Any change in the linear predictor associated with variation of the variable `size` can be reproduced using a linear combination of the dummy variables associated with `site:phase`. We kept `size` and dropped `phase` based on two considerations: 1) the results from the `nmeso` data, where the experimental design allows us to separate the effects of the

two factors, and the high-level similarities between agricultural practices in the two cultures; and 2) the higher dimension of the random effects for the categorical variable `phase` make it possible for any effect attributable to the one dimensional continuous variable `size` to be represented as an effect due to `phase`, but not in general vice-versa. We accept the more parsimonious explanation.

In the `swgermany` data, we will see that `size` is not a significant explanatory variable. If it is included, the associated coefficient is positive, but not significant. The `swgermany` data also show qualitatively different covariate dependence to the other datasets. The model selected using information criteria and deviance tests includes random effects on `site` ID, but does not include `size`. We preserve log-size in the model in later analysis as this effect is of interest. However we can expect the associated parameter to be close to zero.

```
> PO_swgermany = ordinal::clmm( manure_level.imputed ~ scale(log_size) +
+                               link = "logit",
+                               data = agricurb_data %>%
+                                 dplyr::filter( dataset=="swgermany" ) %>%
+                                 dplyr::mutate( site=as.factor(site) ) )
> coefficients(summary(PO_swgermany))

                 Estimate Std. Error    z value               Pr(>|z|)
low|medium     -5.4747248  0.7170897 -7.6346445 0.00000000000002264445
medium|high    -1.1441110  0.4281816 -2.6720230 0.00753954827467872256
scale(log_size)  0.2503616  0.3679126  0.6804919 0.49619307069162338353
```

The estimated coefficient associated with `size` in the `swgermany` data, 0.2504, is positive but not significant (p-value=0.4962).

4.4.4 Conclusions from Single Imputation

The effect of site `size` on `manure_level.imputed` in the `nmeso` and `aegean` datasets is significant at level 0.05 (respectively, p=0.0069 and p=0.0297). This agrees with previous work on the `nmeso` data and supports the hypothesis of extensification in the new `aegean` data. The effect due to `size` in the `aegean` data is hard to separate from the effect due to `phase`. We report the test for the more parsimonious model. The effect in the `swgermany` data is not significant (p=0.5), so we find no support for extensification at the `swgermany` sites.

Single imputation conditions on a single set of `manure_level.imputed`-values, so it does not allow for uncertainty in imputed values. Our analysis above also ignored uncertainty in rainfall values for Archaeological contexts. However, we will see that the results from our most careful analysis (SMI, summarised in section 4.6.3) are in good agreement, on all three datasets, with the conclusions of the simple analysis given in this section.

4.5 Bayesian Multiple Imputation (BMI)

In this section we give a Bayesian analysis of the effect of site size on manuring levels in the archaeological data. We account for essentially all sources of uncertainty in this analysis. In section 4.5.2, we report that model misspecification undermines a conventional *full* Bayesian approach to this particular problem. In section 4.5.3 we use Bayesian Multiple Imputation (BMI), following the Styring et al. (2017) analysis of the `nmeso` data. We expand the model used by those authors by including a possible effect due to `date` in the `PO` module. The effect due to `date` was not found significant in the single imputation analysis. We nevertheless include it in the Bayesian analysis in order to make that analysis self-contained, and because it is has been raised as a possible confounder. BMI is relatively robust to misspecification (see Lunn et al. (2009); Plummer (2015) where it is referred to as a "cut model"). Thousands of sets of `manure_level.imputed` values are sampled according to their posterior distribution in the `HM` module. Each set of `manure_level.imputed` values determines a posterior distribution for the size-effect. These distributions are averaged to get a posterior for the effect of interest which allows for uncertainty in the unknown manure levels (and `rainfall`).

Our full R-implementation of the methods used in this document is available via the R package `agricurbayes`. Supplementary scripts replicate the results and figures in this section using functions available in that package.

4.5.1 Common notation

We begin with the `HM` module introduced in section 4.4. Let Y_{arc}, Y_{mod} be the `normd15n`-values in the archaeological and modern data, respectively. Let M_{arc} and R_{arc} respectively denote the unknown true manure and rainfall levels for the archaeological data. Here M_{arc} plays the same role as `manure_level.imputed` in the single imputation above. Let $\beta = (\beta_1, \ldots, \beta_4)$ be the four fixed effects in the `HM` module associated with the intercept, R_{arc} and M_{arc} (three levels, so two offsets for medium

and high levels relative to baseline which is low-level manuring). The sets of site-level ID's are

$$S_M = \{s; s \text{ is an ID for a site in the modern dataset}\},$$

and

$$S_A = \{s; s \text{ is an ID for a site in the archaeological dataset under analysis}\},$$

and let $\zeta = (\zeta_s)_{s\in\{S_A\cup S_M\}}$ be the mean-zero normal random effects for `site`-levels for the modern and archaeological site, with variance σ^2_ζ. We make a separate analysis for each archaeological dataset, so the site ID's in S_A, the dimension of ζ, and σ^2_ζ vary from one analysis to another. Finally, let σ^2 be the variance of the `normd15n` response given the value of its linear predictor. Denote by $P(Y_{mod} \mid \beta, \zeta)$ the observation model density for the modern data in the `HM` module, where manuring levels are known, and by $P(Y_{arc} \mid \beta, \zeta, M_{arc})$ the same for the archaeological data, where they are unknown.

For the Proportional Odds module `PO`, let γ denote the effect due to `size` on manure level M_{arc} in the archaeological data under analysis. Let τ denote the corresponding effect due to `date`. Let $\alpha = (\alpha_1, \alpha_2)$ be the intercepts for our two-level ordinal `PO`-response M_{arc}. Let $\xi = (\xi_s)_{s\in S_A}$ be mean-zero normal random effects due to `site` on manure-level (only relevant for archaeological sites), with variance σ^2_ξ. Let $P(M_{arc} \mid \alpha, \gamma, \tau, \xi)$ be the (proportional-odds) probability mass function for the missing manure-levels in the archaeological data.

Denote by $\phi = (\alpha, \tau, \xi, \sigma_\xi)$ all the variables of the `PO` module except γ. Denote by $\theta = (\beta, \zeta, \sigma, \sigma_\zeta, R_{arc})$ all the variables of the `HM` module except M_{arc}.

Our parameter of interest is γ, the effect due to size on manure level, which is linked to the hypothesis of extensification.

4.5.2 Remarks on full Bayesian Inference and priors

In the following we take flat prior distributions, $P(\beta) \propto 1$, $P(\alpha, \gamma, \tau) \propto 1$, subject to conservative upper and lower bounds $(\alpha, \gamma, \tau) \in [-5, 5]^3$ on the `PO` module parameters. These are justified below.

The prior for the variance of the random effects in both models is an inverse gamma, $\sigma_\zeta, \sigma_\xi \sim \text{InvGamma}(\alpha = 2, \beta = 1)$. This prior concentrates most of its probability (around 70%) in values smaller than 1 and gives enough room for larger values.

The prior for rainfall $P(R_{arc})$ is $R_{arc} \sim U(\texttt{min_rainfall}, \texttt{max_rainfall})$ as these bounds are fairly tight. The joint posterior distribution conditional on all observed

data is

$$P(\alpha,\gamma,\tau,\xi,M_{arc},\beta,\zeta,M_{arc},R_{arc}\mid Y_{mod},Y_{arc}) \propto P(Y_{mod}\mid\beta,\zeta)\ P(Y_{arc}\mid\beta,\zeta,M_{arc},R_{arc})\cdot \\ P(R_{arc})\ P(\zeta\mid\sigma_\zeta)\ P(\beta,\sigma_\zeta)\cdot \\ P(M_{arc}\mid\alpha,\gamma,\tau,\xi)P(\xi\mid\sigma_\xi)\ P(\alpha,\gamma,\tau,\sigma_\xi). \tag{4.5.1}$$

Explicit expressions for these densities and probability mass functions are given in Styring et al. (2017).

The parameters of the `HM` module are well-informed by the data, and any reasonable prior leads to the same conclusions, so we make no careful prior elicitation for that module. However, (Styring et al., 2017) point out that, without a proper prior for the parameters of the `PO` module, the posterior is improper (the normalising constant is infinite).

This is related to the problem of linear separability in multinomial logistic regression (Albert and Anderson, 1984), where a maximum likelihood estimate need not exist, and is compounded by our use of data-augmentation in the `PO` module, where M_{arc} is missing data. Our conservative bounds make the prior proper and remove this problem.

We have then to justify our choice of bounds. Firstly, values of α and γ outside $[-5,5]$ give models in which very little variation in manure-level is allowed at a site. It may be argued that this is not completely un-physical. However, values as large as $\pm 4, \pm 5$ are already strongly informative. Also, the outcome of our test is not sensitive to the choice of bounds: we measure evidence for the hypothesis that $\gamma < 0$; the absolute value of γ is not of interest. We have experimented with more extreme bounds and find that the evidence for $\gamma < 0$ in the `nmeso` and `aegean` data only increases. However, although the dependence of the evidence on widening bounds is understood, accurate MCMC estimation of the evidence becomes impractical for significantly wider bounds. Further exploration was in any case uninteresting for the reasons given above.

However, there is a further problem. There is model misspecification in the `PO` module, at least for the `nmeso` and `aegean` data. This is discussed in Styring et al. (2017) and motivated their use of misspecification-robust BMI. The most straightforward evidence for misspecification is found in the SMI analysis below, where the control parameter η is a summative measure of misspecification. Our estimated values show that misspecification is present. The model elaborations we explored to identify and correct misspecification did not remove the problem.

If we naively apply full Bayesian inference then we find the posterior for γ puts all its mass close to the lower bound $\gamma \simeq -5$ and follows the lower bound as it is moved to lower values. This cannot be accepted as evidence that γ is negative as the model is misspecified. Repeating the analysis on synthetic data with the same covariates as the real data, where there is no model misspecification, gives posteriors for γ which cover the fixed synthetic true value, with tails which go rapidly to zero, so that conservative upper and lower limits on γ play no practical role in the analysis. This qualitatively different behaviour for synthetic data is further evidence for misspecification.

Continuing the reasoning of (Styring et al., 2017), we could impose strongly informative prior distributions for parameters of the `PO` module or M_{arc} itself.

However, we have no concrete prior information justifying informative priors on the PO-module parameters, or restrictions on M_{arc}. Styring et al. (2017) refer to the literature on model misspecification in evidence combination. This treats misspecification by replacing full Bayesian inference with BMI, also known as "cut models". This is the approach we now apply.

4.5.3 Bayesian Multiple Imputation analysis

The analysis presented here follows the BMI-analysis for the `nmeso` data in Styring et al. (2017). We make some slight revisions and extend the analysis to the `aegean` and `swgermany` data. The analysis is applied to one dataset at a time (with the same modern calibration data used throughout).

(Step 1) This is the imputation step. We sample the posterior for M_{arc} in the `HM` module. In the full Bayesian analysis the `PO` module acts as a prior on M_{arc}. Here we cut that feedback and take a flat prior, $P(M_{arc})$. Parameter priors, collectively $P(\theta)$, are otherwise unchanged. The imputation posterior is

$$P(M_{arc}, \theta \mid Y_{mod}, Y_{arc}) \propto P(Y_{mod}, Y_{arc} \mid \theta, M_{arc})\ P(\theta) P(M_{arc}). \tag{4.5.2}$$

In terms of the low-level variables,

$$P(Y_{mod}, Y_{arc} \mid \theta, M_{arc}) = P(Y_{mod} \mid \beta, \zeta)\ P(Y_{arc} \mid \beta, \zeta, M_{arc}, R_{arc}),$$

as before (first line of eq. (4.5.1)).

(Step 2) We now treat M_{arc} as data in the `PO` module. All `PO` priors, collectively $P(\phi)$, are unchanged. The `PO` posterior conditioned on M_{arc} is

$$P(\gamma, \phi \mid M_{arc}) = \frac{P(M_{arc} \mid \phi, \gamma)\ P(\phi)}{D(M_{arc})}, \tag{4.5.3}$$

where $D(M_{arc})$ is an intractable normalising constant. In terms of the low-level variables,

$$P(M_{arc} \mid \phi, \gamma) = P(M_{arc} \mid \alpha, \gamma, \tau, \xi),$$

as in eq. (4.5.1).

Many of the model elements are unchanged from the deprecated full Bayes analysis. However, we now carry out MCMC targeting

$$P_{BMI}(\gamma, \phi, M_{arc}, \theta \mid Y_{mod}, Y_{arc}) \equiv P(\gamma, \phi \mid M_{arc}) P(M_{arc}, \theta \mid Y_{mod}, Y_{arc}), \qquad (4.5.4)$$

the revised BMI-posterior for γ.

The dependence on $D(M_{arc})$ may look awkward, but is straightforward to handle using Monte Carlo methods. The M_{arc} distribution is determined by the posterior in eq. (4.5.2), and then for any fixed M_{arc}, we can sample from $P(\gamma, \phi \mid M_{arc})$ in eq. (4.5.3) using standard MCMC methods. This sampling scheme follows Plummer (2015). The MCMC algorithm is outlined in Styring et al. (2017). The `agricurbayes` package contains our independent implementation set out in detail in the supplement to Carmona and Nicholls (2020). fig. 4.5 shows the posterior distribution of γ under P_{BMI}.

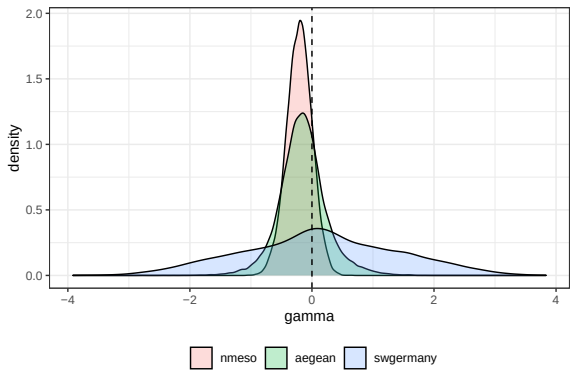


Figure 4.5: Posterior distribution of γ, the effect due to size on manuring levels in the archaeological data.

We measure evidence in favour of extensification (manure level outcome decreases with increasing site size) using the posterior probability that the effect of size, γ, is negative.

$$P(\gamma < 0 \mid Y_{mod}, Y_{arc}) = \int_{-\inf}^{0} P_{BMI}(\gamma \mid Y_{mod}, Y_{arc})\, d\gamma \qquad (4.5.5)$$

This probability is estimated using output from our MCMC algorithm targeting P_{BMI} in eq. (4.5.4).

```
> # Posterior probability of negative gamma #
> p_gamma_leq_0 <- apply( mcmc_gamma < 0, 2, mean )
> p_gamma_leq_0

nmeso    aegean swgermany
0.8421000 0.6926333 0.4554000
```

The posterior probability for $\gamma < 0$ in the `nmeso` data is 0.8421. The corresponding posterior probabilities for the other datasets may be read off from the output table. Our prior for γ puts equal weight on the two hypotheses $\gamma < 0$ and $\gamma \geq 0$ so the Bayes Factor (Kass and Raftery, 1995) $B_{\gamma<0}$ comparing these two hypotheses is

$$B_{\gamma<0}^{(BMI)} = \frac{P_{BMI}(\gamma < 0 \mid Y_{mod}, Y_{arc})}{P_{BMI}(\gamma \geq 0 \mid Y_{mod}, Y_{arc})}. \tag{4.5.6}$$

```
> # Bayes factor for negative gamma #
> BF_gamma_leq_0 = p_gamma_leq_0 / (1-p_gamma_leq_0)
> BF_gamma_leq_0

nmeso    aegean swgermany
5.3331222 2.2534432 0.8362101
```

The `nmeso` and `aegean` datasets give Bayes factors equal to 5.3331 and 2.2534 respectively, evidence for $\gamma < 0$. This is in line with the findings in Styring et al. (2017) and section 4.4. These Bayes factors are likely to underestimate the strength of the evidence due to dilution from noise in the imputation. The `swgermany` data gave a Bayes factor equal to 0.8362, close to 1, and indicating no evidence for a negative γ. The posterior distribution is centred around zero so the effect is not significant here. This is all in line with our findings in section 4.4.

4.5.4 Conclusions from Bayesian Multiple Imputation

We find moderate evidence for a negative effect of site `size` on manure level in the `nmeso` and `aegean` datasets (respectively, $B_{\gamma<0}^{(BMI)}$ =5.33 and $B_{\gamma<0}^{(BMI)}$ =2.25). This agrees with previous work on the `nmeso` data and does not contradict the hypothesis of extensification in the new `aegean` data. We find no evidence for an effect in the `swgermany` data ($B_{\gamma<0}^{(BMI)}$ =0.8362) so no support for extensification at the `swgermany` sites. These measures allow for all major sources of uncertainty (including now any possible confounding due to `date`). They are conservative due to the effect of dilution in BMI.

4.6 Semi-Modular Inference

The strength of the evidence for $\gamma < 0$, the effect of `size` on M_{arc}, measured using BMI in the `nmeso` and `aegean` datasets in section 4.5 is real but not overwhelming. Recent work on evidence combination, in particular *Semi-Modular Inference (SMI)* (Carmona and Nicholls, 2020), adjusts for misspecification without dilution.

4.6.1 Outline of SMI

Semi-Modular Inference (SMI) combines BMI (Liu et al., 2009; Lunn et al., 2009; Plummer, 2015) and the power-posterior (Bissiri et al., 2016; Holmes and Walker, 2017; Grünwald and Ommen, 2017), two inferential procedures treating model misspecification in Bayesian inference. The contribution each module (here `HM` and `PO`) makes to the inference is controlled by a parameter η so that misspecified modules can be down-weighted. In a misspecified setting this may give more accurate imputation and hence better downstream inference than either of fully Bayesian inference or BMI. It will not in general do worse. This is because SMI defines a continuum of inference procedures, indexed by η, interpolating between BMI (at $\eta = 0$) and fully Bayesian inference (at $\eta = 1$). It includes these methods as special cases, and selects the procedure with best predictive performance.

In our context, we impute M_{arc} using an η-*tempered* version of the full model, which reduces (but does not eliminate) the influence of the `PO` module by raising its likelihood to a power $\eta \in [0, 1]$. This setup is used in the imputation of M_{arc}. The second stage of the analysis is the same as step 2 of BMI. In the notation of section 4.5 the SMI posterior for our model is

$$P_\eta(\gamma, \phi, \tilde{\gamma}, \tilde{\phi}, M_{arc}, \theta \mid Y_{mod}, Y_{arc}) = P(\gamma, \phi \mid M_{arc})\ P_\eta(\tilde{\gamma}, \tilde{\phi}, M_{arc}, \theta \mid Y_{mod}, Y_{arc}). \tag{4.6.1}$$

Here $\tilde{\gamma}$ and $\tilde{\phi}$ are parameters used only at the imputation stage, $P(\gamma, \phi \mid M_{arc})$ is unchanged from eq. (4.5.3) and

$$\begin{aligned} P_\eta(\tilde{\gamma}, \tilde{\phi}, M_{arc}, \theta \mid Y_{mod}, Y_{arc}) \propto & P(Y_{mod} \mid \beta, \zeta)\ P(Y_{arc} \mid \beta, \zeta, M_{arc}, R_{arc}) \cdot \\ & P(R_{arc})\ P(\zeta \mid \sigma_\zeta)\ P(\beta, \sigma_\zeta) \cdot \\ & P(M_{arc} \mid \alpha, \gamma, \tau, \xi)^\eta P(\xi \mid \sigma_\xi)\ P(\alpha, \gamma, \tau, \sigma_\xi). \end{aligned} \tag{4.6.2}$$

Note the similarity between eq. (4.5.4) (BMI-posterior) and eq. (4.6.1) (SMI posterior). The two-stage inference set out in Plummer (2015) remains, but the BMI imputation posterior in eq. (4.5.1) is replaced in the SMI imputation posterior (eq. (4.6.2))

by something very like the full posterior in eq. (4.5.1). The change is that the `PO` module likelihood in eq. (4.6.2) is raised to a power η to reduce its influence. The SMI-imputation step uses a power posterior.

We have finally to choose a value for η. Following Carmona and Nicholls (2020), we define the optimal η as the value $\eta = \eta^*$ say, which maximises the Expected Log Predictive Density or *ELPD* (Vehtari et al., 2017) for the `normd15n`-data. This is a natural performance measure, accounting for uncertainty in unknown parameters. It is equivalent to carrying out Leave-One-Out Cross-Validation (LOOCV) on held out `normd15n` responses and selecting the η-value that minimises the mean square error for prediction. It can be computed from the MCMC output. Further details of SMI are given in Carmona and Nicholls (2020).

We use the priors given in section 4.5.2 and the observations models `HM` and `PO` given in section 4.4.

4.6.2 Results from SMI

We use MCMC to sample P_η in a two-step procedure similar to that outlined for P_{BMI}.

In fig. 4.6 we show the estimated *ELPD* as a function of $\eta \in [0, 1]$.

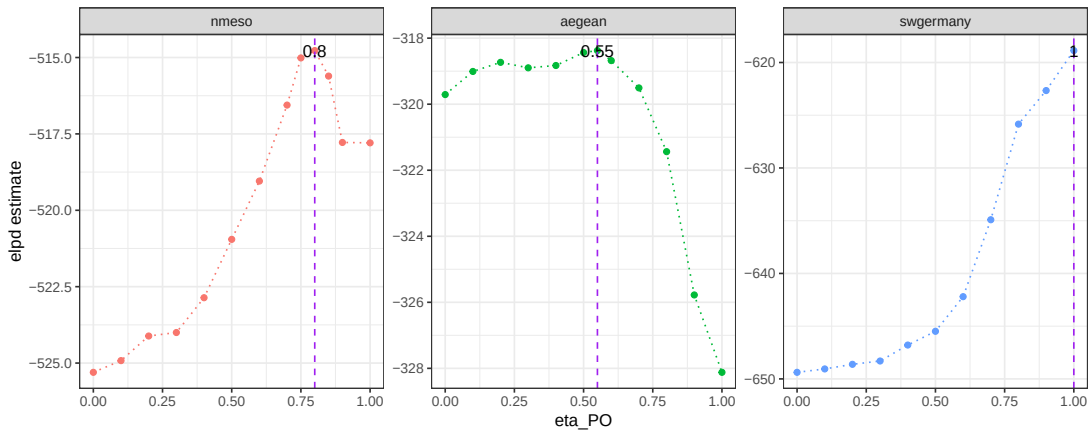


Figure 4.6: Estimated predictive performance measured by ELPD (y-axis) as a function of the control parameter η (x-axis) under SMI for each of the three datasets (see panel titles). Vertical lines give the optimal values $\eta = \eta^*$ of the control parameter for each dataset.

The optimal values for η, denoted η^*, are approximately 0.8 for `nmeso`, 0.55 for `aegean`, and 1 for `swgermany`. The fact that $\eta^* < 1$ in the `nmeso` and `aegean` data is evidence for misspecification. There is no evidence for misspecification in the `swgermany` data where the optimal imputation scheme is full Bayes itself.

We plot in fig. 4.7 credible intervals for the SMI-posterior of γ, the effect of `size` on manure level M_{arc}, as a function of the control parameter η.

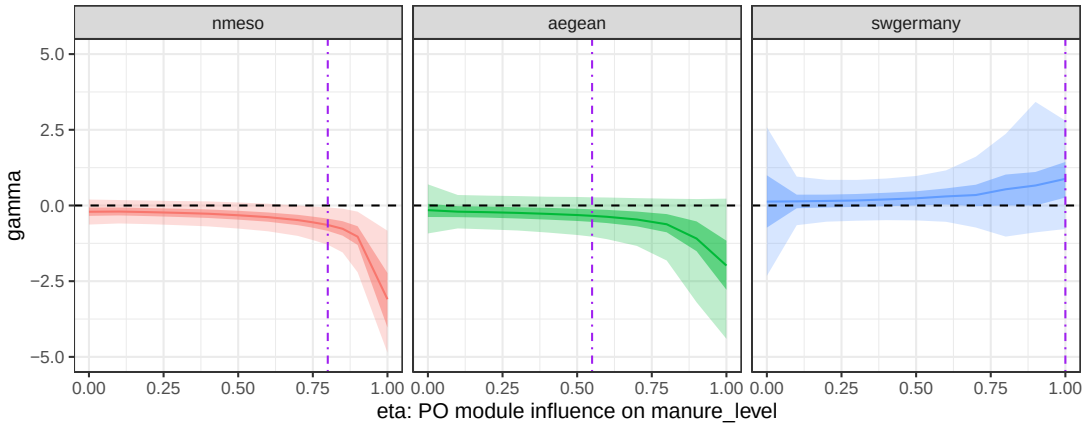


Figure 4.7: Posterior mean and credible intervals of γ under Semi-Modular Inference. The y-axis correspond to the value of γ, while the x-axis corresponds to values of the degree of influence $\eta \in [0, 1]$.

The plots show the posterior mean (solid line) and the 95% and 50% credible intervals (shades) at each η-value (ie each vertical slice). Panels correspond to datasets `nmeso` (left), `aegean` (centre), and `swgermany` (right). The vertical lines give the optimal value of η in each dataset. Our selected SMI posterior is the posterior given by the credible intervals at this line and is shown in fig. 4.8 below. Referring to fig. 4.7, the `nmeso` and `aegean` datasets (left and centre) support similar negative values for the size effect γ at all values of the control parameter η, whereas the `swgermany`-posterior is close-to or slightly above zero.

The posterior credible interval for γ determined by BMI in section 4.5.3 (fig. 4.5), and Styring et al. (2017) supplement Figure 7. (left) page 16, can be read off at $\eta = 0$ in the leftmost `nmeso` plot. Credible intervals shrink towards zero as $\eta \to 0$ as imputation noise rises. The posterior diverges as η approaches $\eta = 1$ and we move towards fully Bayesian inference. This is a consequence of the combined effects of misspecification, linear separability and the conservative bounds used in our non-informative priors, as discussed in section 4.5.2.

fig. 4.8 shows the posterior distribution of γ at $\eta = \eta^*$ for each dataset. For the `nmeso` and `aegean` datasets, the posterior distribution puts little weight on $\gamma > 0$, evidence of a negative effect of `size` on estimated M_{arc}-values.

The evidence in favor of the agricultural extensification hypothesis is measured by the posterior probability that the effect γ is negative, $P_{\eta^*}(\gamma < 0 \mid Y_{mod}, Y_{arc})$.

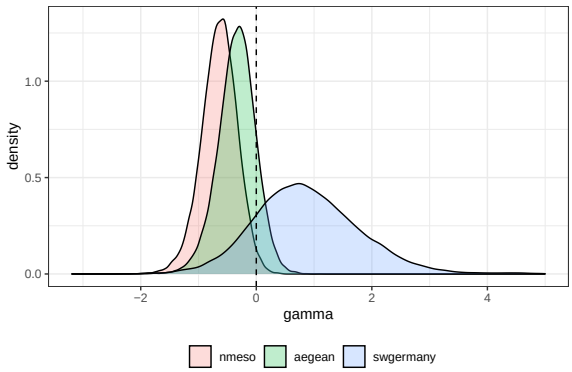


Figure 4.8: Posterior distribution of γ under SMI for the optimal degree of influence, η^*.

```
> # Posterior probability of negative gamma #
> p_gamma_leq_0_smi <- apply( mcmc_best_smi_gamma < 0, 2, mean )
> p_gamma_leq_0_smi

nmeso    aegean swgermany
0.9849333 0.8615000 0.1554333
```

The posterior probabilities for $\gamma < 0$ is higher in `nmeso` and `agean` than we saw in BMI, section 4.5 as dilution is reduced. Some evidence for $\gamma > 0$ in `swgermany` now appears.

The corresponding Bayes Factors for the hypothesis $\gamma < 0$ are given by

$$B_{\gamma<0}^{(\eta^*)} = \frac{P_{\eta^*}(\gamma < 0 \mid Y_{mod}, Y_{arc})}{1 - P_{\eta^*}(\gamma < 0 \mid Y_{mod}, Y_{arc})}$$

```
> # Bayes factor for negative gamma #
> BF_gamma_leq_0_smi = p_gamma_leq_0_smi / (1-p_gamma_leq_0_smi)
> BF_gamma_leq_0_smi

nmeso     aegean  swgermany
65.3716814  6.2202166  0.1840392
```

These Bayes factors (65 and 6) give strong support for $\gamma < 0$ in the `nmeso` and substantial support in the `aegean` data. The Bayes factor for $\gamma < 0$ in the `swgermany` data (0.184) gives substantial evidence for $\gamma > 0$.

4.6.3 Conclusions from SMI

We find evidence for a negative effect of site `size` on imputed manure level M_{arc} in the `nmeso` and `aegean` datasets (respectively, $B^{(\eta^*)}_{\gamma<0}$ =65.3717 and $B^{(\eta^*)}_{\gamma<0}$ =6.2202). This improves on the conservative BMI estimate for the `nmeso` data (taking into account any possible effect due to `date`) and provides clear evidence for the hypothesis of extensification in the new `aegean` data.

There is some evidence for a positive effect in the `swgermany` data ($B^{(\eta^*)}_{\gamma<0}$ =0.184, or in other words a Bayes factor of about 5.4 for the opposite hypothesis, that γ is positive). The conclusion is less clear in this case, and not only for the reason that the evidence is weaker. We entered the analysis aiming to measure evidence for $\gamma < 0$ and so there is no question, for the `nmeso` and `aegean` analyses, of our changing the hypothesis on seeing the data. In claiming evidence for $\gamma > 0$ in the `swgermany` analysis we would be changing our hypothesis after seeing the data. This undermines the analysis (see discussion in Cox (2006)) and, taken with the real but not overwhelming strength of the evidence (Bayes Factor 5.4), we would be inclined to leave this as undecided.

Finally, we visualise the size effect. In fig. 4.9 we show the relation between size and manuring level estimated at $\eta = \eta^*$. The pair of plots in each row share a dataset, `nmeso` (top), `aegean` (middle), and `swgermany` (bottom). Plots in the left column show, for each missing manure level in M_{arc}, the posterior probability the missing level takes the value *low* as a function of `size`. The right column shows the posterior probability for the missing manure level to take either of the values *low* or *medium*. Tighter credible intervals (shaded area) indicate a clearer relation between estimated M_{arc} and `size`. The negative effect of γ is evident for `nmeso` and `aegean`: the posterior probability for lower manure levels grows as `size` increases. For `swgermany` the effect is positive.

4.7 Overall conclusions

We reported results from three analyses of our data. Single imputation gives a simple introduction to the models and inference problem and gave an easy-to-understand check on the more sophisticated analyses which followed. Bayesian Multiple Imputation is a familiar method, allowed direct comparison with Styring et al. (2017), and has much in common with Semi-Modular Inference.

All three analyses, single imputation (section 4.4), BMI (section 4.5) and SMI (section 4.6) give consistent evidence for the hypothesis of extensification in the `nmeso`

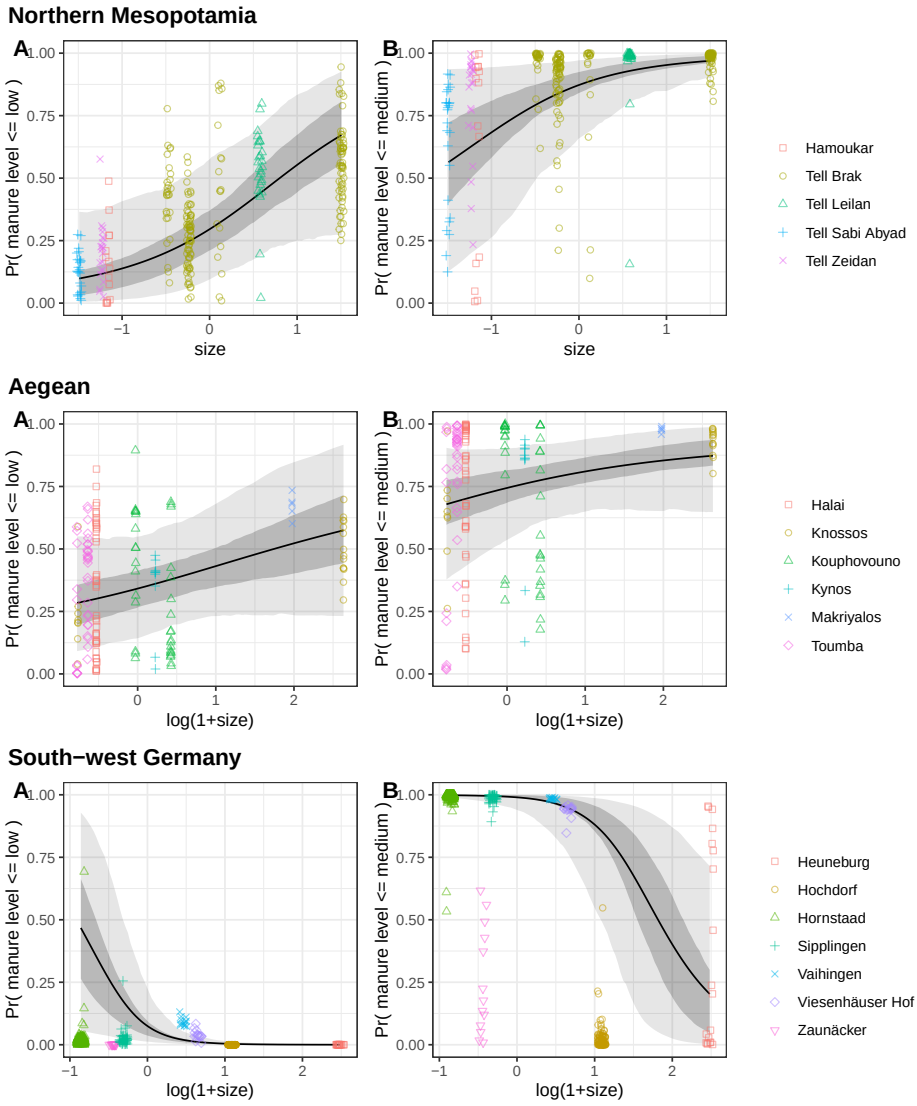


Figure 4.9: The posterior probability (y-axis) that a given archaeological cereal grain sample has a manuring level equal m or lower is plotted against site size (x-axis). (a) Manuring level m=low. (b), Manuring level m=medium. Shaded areas give 50% and 90% credible intervals for the probability. Each row corresponds to a dataset: `nmeso` (top), `aegean` (middle), and `swgermany` (bottom). Colors distinguish different sites.

and `aegean` data. In the main paper we report the SMI measures reported in section 4.6.3,as this is our favored analysis. It improves on single imputation as it allows for all major sources of uncertainty and improves on BMI as it greatly reduces the dilution effect we saw in BMI (see in particular our comments on fig. 4.7).

4.8 Appendix

4.8.1 Pairwise relation of variables on modern and archaeological data

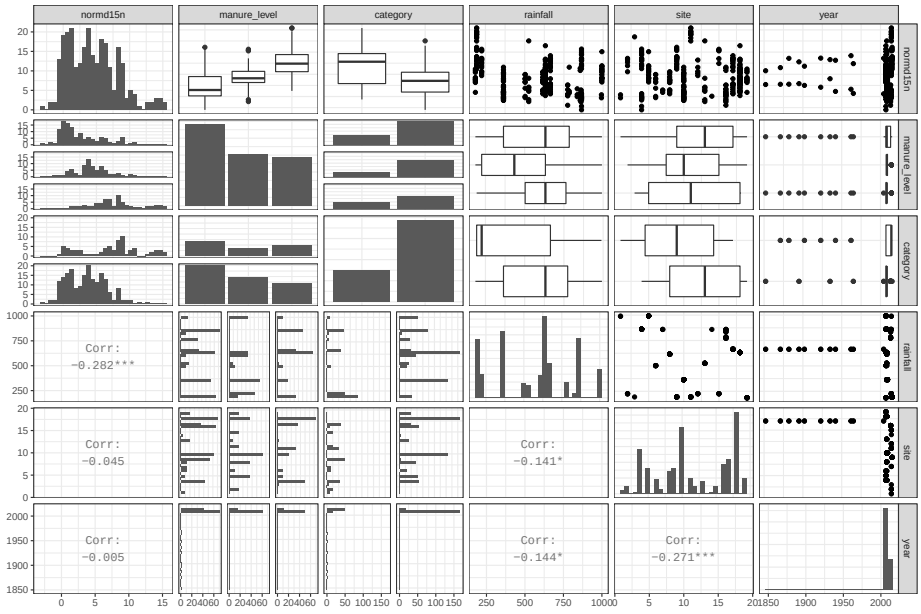


Figure 4.10: Pairwise relation between key variables in the modern data

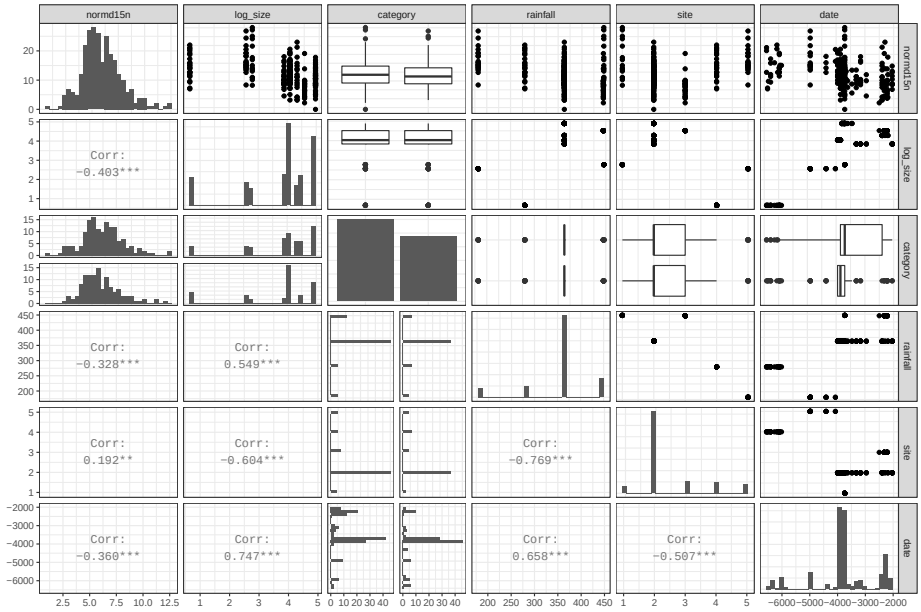


Figure 4.11: Pairwise relation between key variables in the archaeological data from Northern Mesopotamia (`nmeso`)

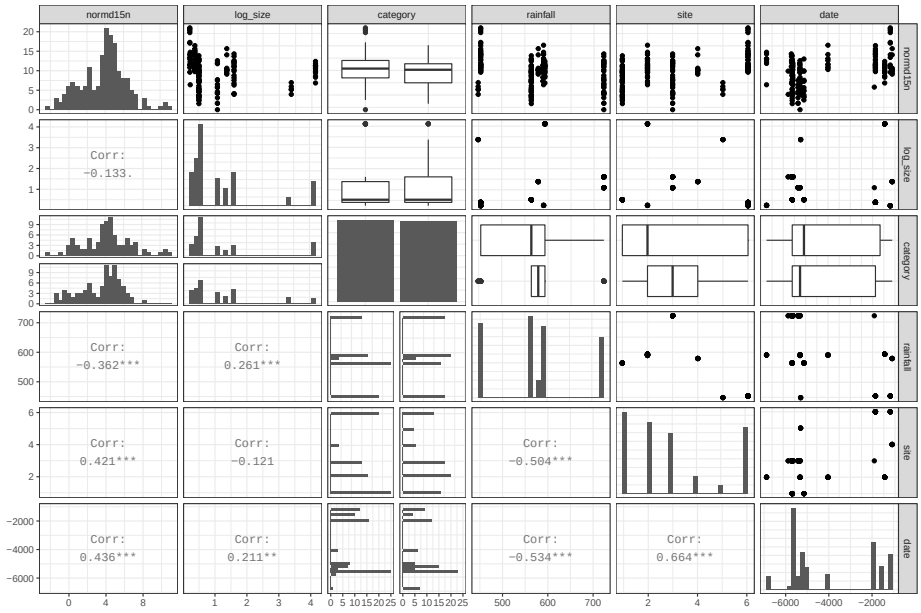


Figure 4.12: Pairwise relation between key variables in the archaeological data from the Greece (`aegean`)

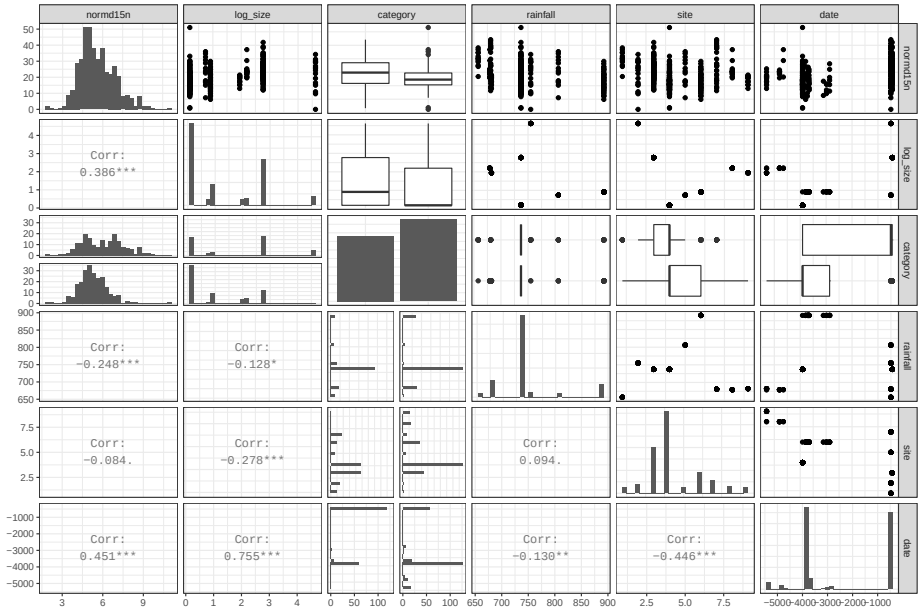


Figure 4.13: Pairwise relation between key variables in the archaeological data from the Germany (`swgermany`)

4.8.2 MCMC analysis for SMI

Here we show simple convergence plots for the MCMC of the SMI posterior under the best η. The MCMC was ran for a total of 310,000 iterations, with a warm-up period of 10,000 and a thinning of 10. fig. 4.14 show the traces for the main parameters in the model, showing no sign of trending. Parameter correlations are displayed as a heatmap in fig. 4.15.

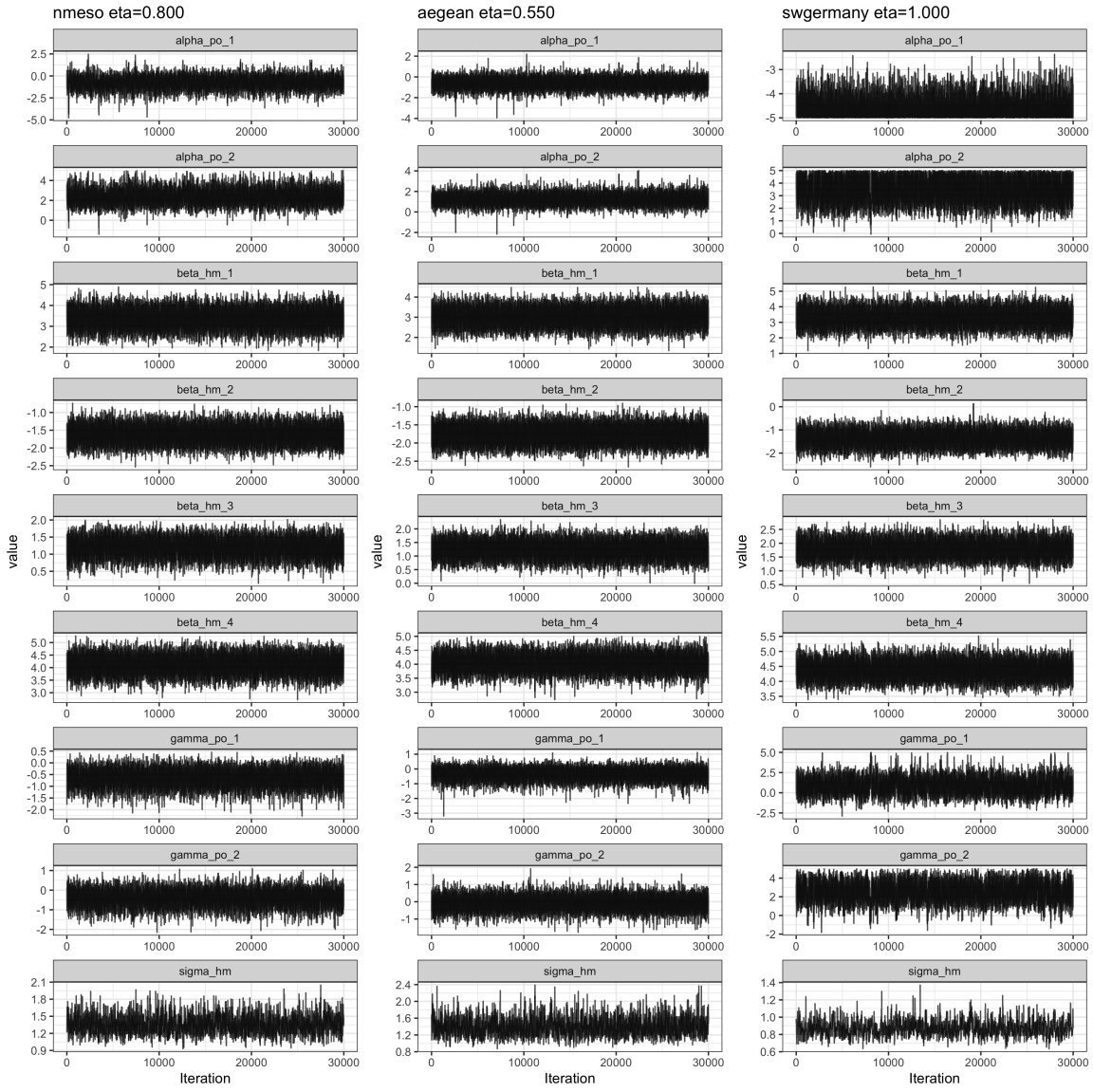


Figure 4.14: Trace plot of the Monte Carlo samples for the main parameters in the model.

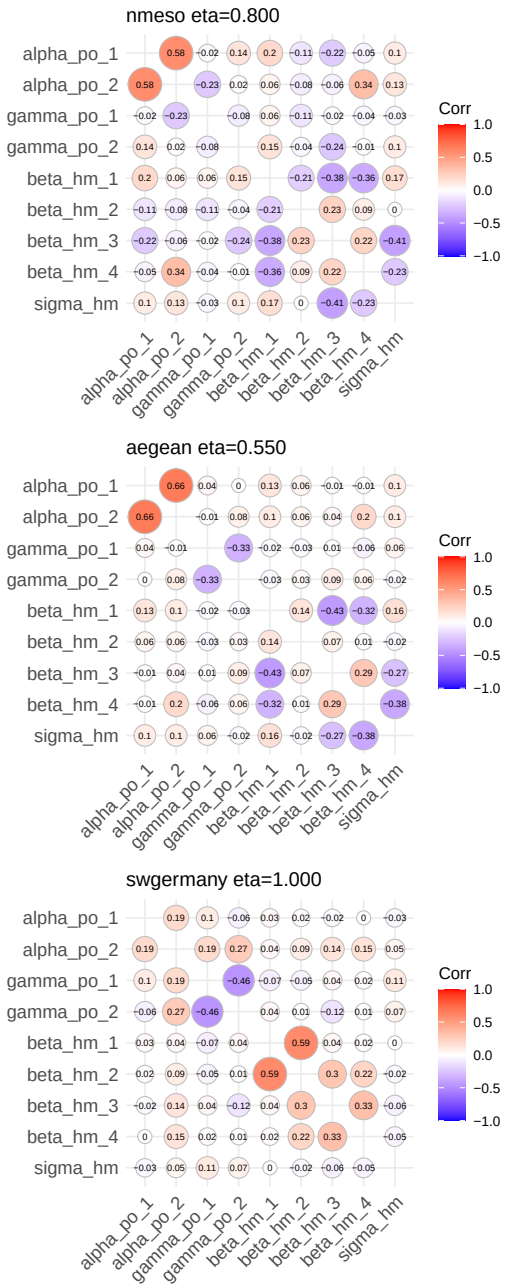


Figure 4.15: Heatmap with correlations of the principal parameters in the model (`gamma_po_1` corresponds to γ, the effect due to `size` in `PO`, whereas `gamma_po_2` is the effect due to `date`, denoted by τ in the text.)

Chapter 5

Simultaneous Reconstruction of Spatial Frequency Fields and Sample Locations via Bayesian Semi-Modular Inference

Abstract

Traditional methods for spatial inference estimate smooth interpolating fields based on features measured at well-located points. When the spatial locations of some observations are missing, joint inference of the fields and locations is possible as the fields inform the locations and *vis versa*. If the number of missing locations is large, conventional Bayesian Inference fails if the generative model for the data is even slightly misspecified, due to feedback between estimated fields and the imputed locations. Semi-Modular Inference (SMI) offers a solution by controlling the feedback between different modular components of the joint model using a hyper-parameter called the influence parameter. Our work is motivated by linguistic studies on a large corpus of late-medieval English textual-dialects. We simultaneously learn dialect fields using dialect features observed in "anchor texts" with known location *and* estimate the location of origin for "floating texts" of unknown origin. The optimal influence parameter minimises a loss measuring the accuracy of held-out anchor data. We compute a (flow-based) variational approximation to the SMI posterior for our model. This allows efficient computation of the optimal influence. MCMC-based approaches, feasible on small subsets of the data, are used to check the variational approximation.

5.1 Introduction

In the conventional setting of *point-referenced data* (Banerjee et al., 2014), data are vectors $y_p \in \mathcal{Y},\ p \in A,\ A = \{1, 2, ..., N\}$ measured at fixed locations $x_p \in \mathcal{X}$ in a region $\mathcal{X} \subseteq \mathbb{R}^2$, collectively $X_A = (x_p)_{p\in A}$ and $Y_A = (y_p)_{p\in A}$. Statistical methods commonly assume that the locations are known and specify a hierarchical generative model, assuming that data at $x \in \mathcal{X}$ are generated conditional on the local value $\Phi(x)$ of a multi-dimensional latent spatial field $\Phi : \mathcal{X} \to \mathbb{R}^{d_\phi}$. Allowing other global parameters θ, the observation model has the form

$$y_p \sim P(\cdot|\Phi(x_p), \theta)\ ,\ p \in A.$$

Let $\Phi(X_A) = (\Phi(x_p))_{p\in A}$ give the field values at the observation locations. A Bayesian approach assigns prior distributions to the latent fields and global parameters, and targets the posterior distribution conditional on these "anchor" data,

$$P(\Phi(X_A), \theta|Y_A) \propto P(Y_A|\Phi(X_A), \theta)P(\Phi(X_A), \theta).$$

There is a large literature for this setting with known locations, covering different specifications for the response vector and prior. However, few methods have been developed to handle the situation in which some sample locations are missing and estimation of these locations is the goal.

Suppose that we are given an additional set of M measurements $y_p,\ p \in \bar{A},\ \bar{A} = \{N + 1, ..., N + M\}$ with unknown locations. We call these data "floating" data $Y_{\bar{A}} = (y_p)_{p\in\bar{A}}$ and the data $y_p,\ p \in A$ "anchor data" and assume the floating and anchor data are realisations of a common generating process. The observed data is then $X_A, Y_A, Y_{\bar{A}}$. The missing locations of floating data, $X_{\bar{A}} = (x_p)_{p\in\bar{A}}$, enter as unknown quantities in the model. Let $\mathcal{P} = A \cup \bar{A}$, let $X_\mathcal{P} = (X_A, X_{\bar{A}})$ be the set of all locations for observations, known and unknown and let $Y_\mathcal{P} = (Y_A, Y_{\bar{A}})$. The posterior distribution is now

$$P(\Phi(X_\mathcal{P}), \theta, X_{\bar{A}}|Y_\mathcal{P}) \propto P(Y_A|\Phi(X_A), \theta)P(Y_{\bar{A}}|\Phi(X_{\bar{A}}), \theta)P(\Phi(X_\mathcal{P}), \theta, X_{\bar{A}}).$$

where $\Phi(X_\mathcal{P}) = (\phi(x_p))_{p\in\mathcal{P}}$ are the values of the multidimensional latent field at all locations.

The first contribution of this work is to provide a statistical framework to estimate the missing locations $x_p \in \bar{A}$ jointly with estimation of the fields Φ. Bayesian inference for missing locations is explored in Giorgi and Diggle (2015). They develop an

MCMC procedure to target missing locations in a rectangular region in $\mathbb{R}^2$, assuming normally distributed observations with a single spatial field parameterising the mean. They focus on cases where the number of floating observations is at most two. One important challenge in our setting is the sparsity of the data vectors $y_p,\ p \in \mathcal{P}$ and the high dimension of Φ. In the model below, the fields Φ are represented using Gaussian Processes. Using a softmax transformation, each component field in Φ parameterises the probability to select one level of one categorical variable. This is the GP-Generalized Linear Model (GLM) modelling framework of Diggle et al. (1998); Savitsky et al. (2011). We have observations of 71 categorical variables at each location in our main analysis, and each categorical variable has many levels, so we have hundreds of spatial fields.

This data-sparsity is a distinctive feature of the model: if we take measurements of an additional categorical variable at each location, we add a set of fields for each level of the new variable. If we add data in this way, we increase the dimension of the parameter space significantly. However, if we add new anchor measurements x_p, y_p to A, keeping the number of variables observed at each location fixed (so fixed dimension for Φ), then under regularity conditions and in the well-specified setting we expect the fields to concentrate on their true values. Consistency results for the unknown locations X_A of the floating observations are outside the scope of our work; we plot marginal posterior distributions for locations based on the data we actually have.

When the fields are informed by sparse and indirect measurements, and just a subset of those observations are anchors, the analysis is exposed to misspecification, and this motivates our use of specialised methods for handling model misspecification; extending these methods to work in our setting is a second contribution of this paper. In a straightforward Bayesian analysis the posterior distribution for the unknown locations $X_{\bar{A}}$ and latent fields $\Phi(X_{\mathcal{P}})$ appear in a single joint distribution conditioned on the known locations X_A and the response data $Y_{\mathcal{P}}$. In this setting response data $Y_{\bar{A}}$ measured at unknown locations also inform, or "feed-back", into the posterior distribution for latent fields. If the number M of observations with missing locations is large compared to the number N of observations with known locations, and the model is misspecified, then this feedback can pose real problems: incorrect locations may distort the field posterior, which in turn biases location estimates.

In an analysis of the same data, Haines (2016) gets around this problem by restricting analysis to a subset of the data and reducing the number of float locations, but this is unsatisfactory. Model elaboration (Gelman et al., 2015) helped, and we

present here our best post-elaboration model, a significant improvement over that given in Haines (2016). However goodness of fit problems remain as we show below.

Bayesian inference in the $\mathcal{M}$-open setting, where the true generative model is not in the class of models expressed by the prior and likelihood, has progressed a great deal in the last decade (Zhang, 2006; Liu et al., 2009; Plummer, 2015; Bissiri et al., 2016; Grünwald and Ommen, 2017; Carmona and Nicholls, 2020). The approach we take, based on Bayesian Multiple Imputation (BMI), draws on this literature. BMI can be understood as a two-stage inferential process. First, in an *imputation stage*, the posterior distribution for the unknown spatial fields and background parameters is formed using the anchor data alone. Then in a second *analysis stage*, the posterior for float locations is conditioned on the measurements at those locations and the imputed field distributions. The imputed fields are random variables passed from the imputation to the analysis, hence the BMI designation. This two-stage approach to estimation of missing observation-locations was taken in Wasser et al. (2004, 2007). Although this approach has many Bayesian elements, it does not define a standard Bayesian posterior.

In the Bayesian Multiple Imputation setup, feedback from float locations into field estimates is "cut", so we call the resulting joint "posterior distribution" a *Cut posterior* (Liu et al., 2009; Plummer, 2015). However, in mildly misspecified settings, completely cutting feedback may be unnecessarily conservative and some scheme allowing some controlled feedback from floating locations into field estimates may do better. In this paper we "select the inference" (Jacob et al., 2017) in a single overarching framework called Semi-Modular Inference (SMI) (Carmona and Nicholls, 2020, 2022; Nicholls et al., 2022). This includes Cut-model inference and Bayesian inference as special cases. Sampling an SMI- or Cut-posterior using Monte Carlo methods is more challenging than sampling the equivalent standard Bayesian posterior distribution (Plummer, 2015; Liu and Goudie, 2022b), and may require two-stage MCMC methods (Jacob et al., 2020) or variational approximations (Yu et al., 2021; Carmona and Nicholls, 2022).

The generative model we write down for for the float locations, background parameters, fields and observations is developed from the model given in Haines (2016). That work introduces the data and develops models in more detail than we do here. It gives comparisons with alternative methods including k-Nearest Neighbours. These are not competitive with Cut posterior inference, though Nearest Neighbours shows promise. We refer the reader to Haines (2016) for these experiments.

5.2 The Linguistic Atlas of Late Mediaeval English (LALME) Data

We now describe the data in detail. See section 5.9 in the Supplementary Material for further background from the perspective of historical linguistics and Haines (2016) for exploration of these data from a statistical perspective.

The Linguistic Atlas of Late Mediaeval English (LALME), originally published by McIntosh et al. (1986) and later revised and supplemented by Benskin et al. (2013), provides a corpus of lexico-grammatical data across a range of late mediaeval texts. The Atlas is founded on inventories of spellings for 1511 texts written in England, parts of Wales, and a small number from southern Scotland, and spans approximately 1350CE to 1450CE. At the time spellings were not standardised, and people tended to "spell as they spoke", so spatial variation in spelling reflects dialectical variation. In order to map this variation, the Atlas contains data on the usage of differing forms of words and grammatical inflexions in different texts.

The Atlas differentiates between different *forms* (loosely, different spellings, indicated below by $\langle\rangle$) of a corresponding *item* (loosely, different words, indicated by $\{\}$). We expand on these definitions in section 5.9.3. For example, the item {AGAINST} has 165 varying forms, including ⟨AGEINST⟩, ⟨AYEINST⟩, ⟨AGAINz⟩ and ⟨TOYENST⟩. The variation between capitalised and non-capitalised letters represents the use of manuscript symbols, such as extra symbols and accented letters, rather than the presence of capitalisation itself. A single item can have many forms, some as many as several hundred. The Atlas authors start with a list of items. They have P texts and take each text and go through each item in the list recording the forms of the item which appear in the text. Each text document is then summarised by a record called a Linguistic Profile (LP) recording for each item the forms it contains.

Denote by $\mathcal{I} = \{1, 2, ..., I\}$ the set of all item labels (distinct words and grammatical concepts) and for each item $i \in \mathcal{I}$ let $\mathcal{F}_i = \{1, 2, ..., F_i\}$ give the set of forms it takes. For example if $i \in \mathcal{I}$ is the item label for {AGAINST} then each of the forms listed above will have a corresponding label $f \in \mathcal{F}_i$. As explained below, we focus on $I = 71$ selected forms which we subject to re-registration in a linguistically informed coarsening process detailed below and in section 5.9.4. There are in all $d_\phi = \sum_{i=1}^{I} F_i$ forms in these data. Let $\mathcal{P} = \{1, 2, ..., P\}$ give the set of profile labels, one for each text. An observation $y_{p,i,f} = 1$ indicates that form f of item i is present in the text summarised by profile p, and $y_{p,i,f} = 0$ is observed otherwise. Let $y_{p,i} = (y_{p,i,f})_{f \in \mathcal{F}_i}$

and $y_p = (y_{p,i})_{i \in \mathcal{I}}$, so that data $y_p \in \{0,1\}^{d_\phi}$ records the presence and absence of forms for the LP for the text with label $p \in \mathcal{P}$.

The 1511 LPs are divided into two categories: 300 anchor profiles and 1211 floating profiles. Anchor profiles are localised on the basis of non-linguistic evidence. In rare cases, a scribe with known biography can be identified, but most anchors are legal or administrative documents whose place of origin is explicitly stated. In this case it is assumed (if there is no evidence to the contrary) to represent the local forms of language. McIntosh et al. (1986) deduce the location of floating profiles using a sequential method they call the 'fit-technique'. This is described in more detail in section 5.9.2, what follows is an outline. For each form that appears in a floating profile, anchor texts and previously located floating profiles are used to delimit the region in which this form was used. The unknown profile location should lay in the intersection of these regions across all the forms it displays. Further linguistic considerations were used to resolve conflicts and select a location. The located float is then treated as an anchor and the process continued until all the floats are positioned. If someone wishes to locate a new text they treat it as the $M + 1$'st float and apply the same method.

Statistical modelling of these data is challenging, as a given item can have hundreds of forms, and many of these may be rare. In order to treat this a coarsened version of the data was made. This coarsening is described in section 5.9.4. Forms with very similar spellings were merged according to linguistic criteria developed by the authors of the Atlas. The coarsened data contains, for the purpose of location estimation, most of the information in the original data, but is significantly less sparse. In this paper we analyse the coarsened data.

We focus on a rectangular region, displayed in fig. 5.1, spanning 200 km east-west and 180 km north-south. The region contains $N = 120$ anchor profiles with labels $A = \{1, ..., N\}$. A further $M = 247$ floating profiles with labels $\bar{A} = \{N + 1, ..., N_M\}$ were assigned to this region using the fit-technique (so that $P = 367$ above). We are assuming the Atlas float-locations are sufficiently accurate for this purpose. Our analysis below supports this, so the selection rule and analysis results are at least consistent. The data $X_A, Y_A, Y_{\bar{A}}$ and missing data $X_{\bar{A}}$ are now as defined in the introduction. There are in all $d_\phi = 741$ forms in the coarsened data displayed in the region of interest. As an example of dialectal variation, in fig. 5.2 we display the locations corresponding to usage of coarsened forms of the item {SUCH}.

The third contribution of our work is our analysis of the Atlas data itself. We estimate and quantify uncertainty in the location of floating profiles given the anchor

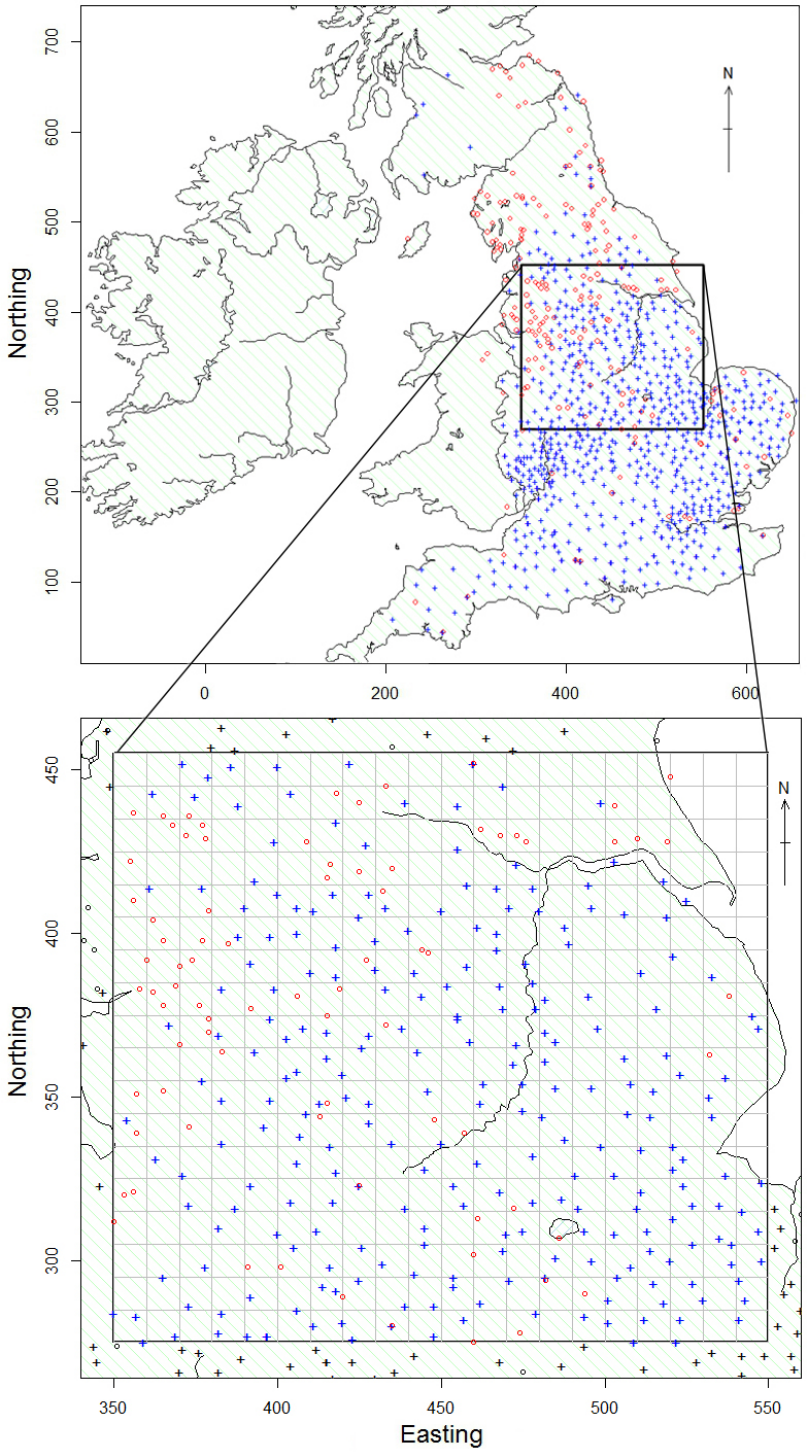


Figure 5.1: The locations of the anchor profiles (red) and the locations of the floating profiles estimated using the fit-technique (blue). We focus on a 200 km × 180 km region, which is expanded in the bottom panel.

profiles. This is the same problem the Atlas authors faced. We get good agreement

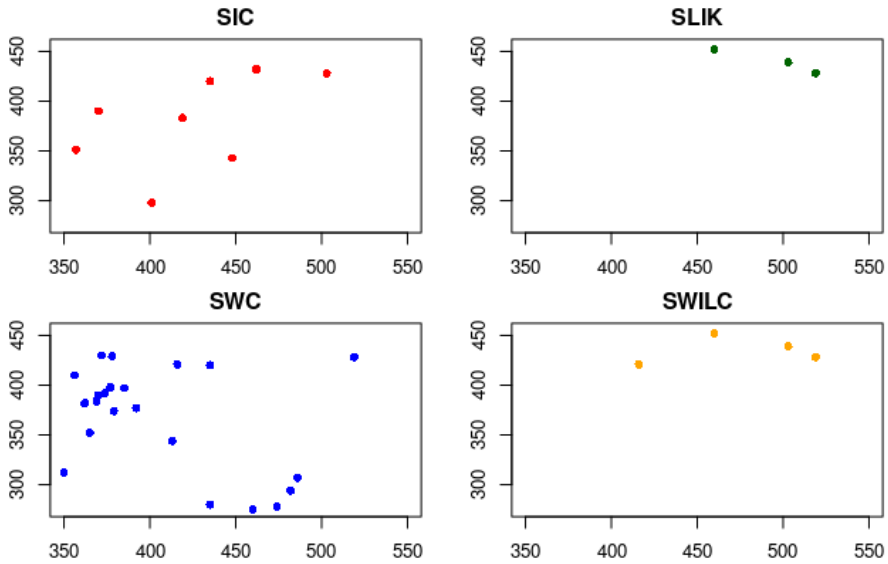


Figure 5.2: The anchor locations at which ⟨SIC⟩, ⟨SLIK⟩, ⟨SWC⟩ and ⟨SWILC⟩, the four most common forms of the item {SUCH} are recorded.

between the credible regions we estimate for floating profiles and the Atlas locations estimated using the fit-technique. Our approach is automated, and provides well-calibrated uncertainty measures for location. The level of location precision achievable with these data has been in question and this is not easily quantified using the non-statistical fit-technique. Our analysis shows that location accuracy of order thirty kilometres is possible, underpinning the value of the Atlas. One application of the Atlas is to locate new floating profiles. We validate our method by treating anchors as if they were floating profiles and predicting their locations. This demonstrates the feasibility of locating new texts.

5.3 Methodology

5.3.1 Gaussian Processes for Latent Fields

Gaussian Processes (GPs) (Rasmussen and Williams, 2005) form a flexible prior over functions and have been used to model latent spatial fields (Diggle et al., 2013; Lloyd et al., 2015; Aglietti et al., 2019). Bayesian inference for indirectly observed latent spatial fields using GPs gives well calibrated measures of uncertainty when the prior and observation models are well specified.

We have a large number of spatial fields in our motivating problem, reaching many hundreds. Moreover, fields representing different forms of a given item (some particular word) are anti-correlated at each location: if one spelling of a word is

common at some location then other spellings are typically uncommon. However, frequencies of spellings of different words are expected to be correlated by shared dialect: similar patterns of spelling preferences across different words recur at different locations.

In order to capture correlation between intensity fields and improve scalability on the dictionary dimension, we follow the Linear Model of Coregionalization (LMC) (Journel and Huijbregts, 1978; Álvarez et al., 2012) in which a large number of fields are given as different linear combinations of a small number of basis fields.

In order to improve scalability with increasing numbers of profiles and to make variational inference more straightforward, we use a sparse approximation to the fields via inducing points (Titsias, 2009; Bonilla et al., 2019; Quiñonero-Candela and Rasmussen, 2005). Inducing values augment the prior specification and serve as *global parameters* that are convenient for approximating the posterior distribution of the entire field via Stochastic Variational Inference (SVI) (Hensman et al., 2013, 2015). We treat the field values at inducing points as random variables in the posterior and marginalise (approximately) over the field values at floating profiles. Targeting a posterior on the joint distribution of the floating profile locations and the field values at those locations would be challenging for variational inference due to complex co-dependence. By contrast, the inducing point locations are fixed.

5.3.2 Semi-Modular Inference

Cut models (Liu et al., 2009; Plummer, 2015) and Semi-Modular Inference (SMI) (Carmona and Nicholls, 2020, 2022) are forms of Bayesian Multiple Imputation that are useful in settings where some of the priors or likelihoods of the generative model for the data are mispecified. The idea is to break the model up into "good" modules (thought to be well-specified) and "bad" (misspecified) modules. These methods are natural when there are several datasets in the analysis, each with their own observation model. In that setting one module may be associated with each dataset. Each dataset defines a posterior distribution for the variables appearing in the observation model for that dataset. This definition of a "module" has been formalised in Liu and Goudie (2022a).

Consider for example a model with dependence structure illustrated in fig. 5.3. The modules are $\{Y, \psi, \theta\}$ and $\{Z, \psi\}$. The two posteriors are respectively $p(\psi|Z)$ and $p(\theta, \psi|Y)$. Suppose the Z-module is trusted and the Y-module is not.

5.3.2.1 Cut-Model inference

In Cut model inference, parameters shared between good and bad modules are "imputed" in a first stage analysis using only the information in the good module. In our example the Cut-posterior for ψ is just $p(\psi|Z)$. Parameters in the bad modules are then "analysed" in a second analysis conditioned on the imputed parameters from the first stage. The idea is to stop mispecification in the bad module biasing the posterior for parameters it shares with the good module. In the example above the joint Cut posterior for ψ, θ is

$$\begin{aligned} p_{cut}(\psi, \theta|Z, Y) &= p(\psi|Z)\ p(\theta|\psi, Y) \\ &\propto p(Z|\psi)p(\psi)\ \times\ p(Y|\psi, \theta)\frac{p(\theta|\psi)}{p(Y|\psi)}. \end{aligned} \tag{5.3.1}$$

This is "multiple imputation" because the imputed parameters ψ have a posterior distribution from the first stage, and the second stage conditions on these random variables. Yu et al. (2021) show that the Cut-posterior is the joint distribution of ψ, θ minimising the KL-divergence to the Bayes posterior

$$\begin{aligned} p(\psi, \theta|Z, Y) &= p(\psi|Z, Y)\ p(\theta|\psi, Y) \\ &\propto p(Z|\psi)\ p(\psi)\ \times\ p(Y|\psi, \theta)\ p(\theta|\psi) \end{aligned}$$

over all distributions $q(\psi, \theta)$ with ψ-marginal $q(\psi) = p(\psi|Z)$. The Cut-posterior is not a Bayesian belief update which uses the likelihood as a utility and does not have frequentist coverage of the pseudo-true parameter asymptotically (Pompe and Jacob, 2021). However, Carmona and Nicholls (2020) show that the Cut-model belief update uses a different utility and is coherent in the sense of Bissiri et al. (2016), and Nicholls et al. (2022) shows that the posterior is valid for this utility.

5.3.2.2 Semi-Modular Inference

The setup in SMI is similar. However, in SMI some information from the mispecified module is used in the imputation stage. The idea is to allow the misspecified module to bias the posterior for shared parameters in order to reduce their variance. This trade off is modulated by an *influence parameter* η. The posterior used for imputation of the shared parameter ψ is the "power posterior" (Friel and Pettitt, 2008),

$$p_{\text{pow},\eta}(\psi, \tilde{\theta}|Z, Y) \propto p(Z|\psi)\ p(Y|\psi, \tilde{\theta})^{\eta}\, p(\tilde{\theta}, \psi). \tag{5.3.2}$$

An auxiliary "copy" of the θ parameter is needed in the imputation stage in order to access information from the likelihood $p(Y|\psi, \tilde{\theta})$. In the analysis stage we condition

on the marginal random variable $\psi \sim p_{\text{pow},\eta}(\psi|Z,Y)$, so the full joint Semi-Modular Posterior (SMP) can be taken to be

$$p_{\text{smi},\eta}(\psi,\theta,\tilde{\theta} \mid Z,Y) = p_{\text{pow},\eta}(\psi,\tilde{\theta} \mid Z,Y)\; p(\theta|\psi,Y). \tag{5.3.3}$$

This reduces to p_{cut} when $\eta = 0$ and gives standard Bayesian inference when $\eta = 1$ so the SMP uses the power posterior to give a sequence of distributions interpolating Bayes and Cut (Carmona and Nicholls, 2020). Recent variants of SMI suggest alternative interpolating sequences of distributions (Liu and Goudie, 2021; Chakraborty et al., 2023; Nicholls et al., 2022) within the same framework.

5.3.2.3 Targeting the Semi-Modular Posterior

Targeting the SMP with MCMC is not straightforward due to the marginal factor $p(Y|\psi)$ in eq. (5.3.1) and eq. (5.3.3). A form of nested MCMC (Plummer, 2015) is straightforward and adequate for many simple problems (Carmona and Nicholls, 2020). For any given value of the influence parameter η, we first obtain T_1 samples $(\psi^{(t)},\tilde{\theta}^{(t)})_{t=1}^{T_1}$ from the power posterior $p_{\text{pow},\eta}(\psi,\tilde{\theta} \mid Z,Y)$ using MCMC. In the second stage, for each sample $\psi^{(t)}$, we run a sub-chain $(\theta^{(t,\tau)})_{\tau=1}^{T_2}$ targeting $p(\theta|\psi^{(t)},Y)$ to equilibrium, preserving only the last sample $\theta^{(t)} = \theta^{(t,T_2)}$ of each sub-chain. The samples returned by the two-stage sampler are $(\psi^{(t)},\theta^{(t)},\tilde{\theta}^{(t)})_{t=1}^{T_1}$. This can be done using standard MCMC algorithms. Asymptotically in T_1 and T_2, the triples $(\psi^{(t)},\tilde{\theta}^{(t)},\theta^{(t)}) \sim p_{\text{smi},\eta}$ have the desired distribution. We found this worked well on our problem if the number of items and floats in the data was small (of order ten items and ten floats), but too slow on the full dataset. This motivates our use of Variational SMI. Besides this nested MCMC, other Monte-Carlo methods (Jacob et al., 2020; Liu and Goudie, 2022b) have been given for the Cut-model. These could be used for SMI. However, although they remove the double asymptotics and provide better guarantees on sample quality, they are in general slower than Nested MCMC and we have not explored them here. Some further implementation detail is given in section 5.6.2.

In Variational SMI, the aim is to find a distribution $q_\lambda(\psi,\theta,\tilde{\theta})$ that approximates the SMP in eq. (5.3.3). Straightforward minimisation of the KL-divergence loss to the SMP over a variational family q_λ is neither feasible (on account of the intractible marginal $p(Y|\psi)$ factor) nor in fact desirable (as we would loose control of the influence of bad modules on parameters shared with good nodules). Recently, Yu et al. (2021) and Carmona and Nicholls (2022) gave a variational loss for approximation of Cut-models and the Semi-Modular Posterior that preserves control of information flow in

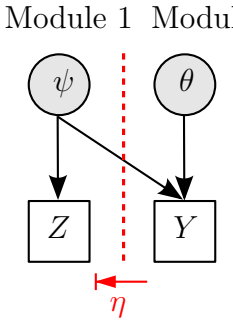


Figure 5.3: A Cut-model with two modules. The first module has parameter ψ and data Z, while the second is defined conditionally upon ψ, with parameter θ and data Y.

the approximating posterior and is feasible to fit using Stochastic Gradient Descent (SGD). We explain how this applies in our setting in section 5.6.3.

5.4 Model

In this section we give our statistical model for the LALME data described in section 5.2.

5.4.1 Observation model

Imagine a medieval scribe writing text $p \in \mathcal{P}$. They use item $i \in \mathcal{I}$ a certain number of times, $z_{p,i}$. Each time they use item i they choose a form $f \in \mathcal{F}_i$ for the item to take (informally, each time they write a word they choose a spelling). The probability $\phi_{i,f}(x_p)$ for them to choose any particular form depends on their location x_p (recall, location here is the source of the scribe's dialect, for example, their place of schooling; for anchor texts this is taken to be the place at which the text was created, in the absence of evidence to the contrary). For $x \in \mathcal{X}$ let

$$\phi_i(x) = (\phi_{i,1}(x), \ldots, \phi_{i,F_i}(x))$$

give the vector of form-selection probabilities for item i at location x, with $\sum_{f=1}^{F_i} \phi_{i,f}(x) = 1$. Suppose that, by the end of the text they use form $f \in \mathcal{F}_i$ in all $z_{p,i,f}$ times so that $z_{p,i} = \sum_{f \in \mathcal{F}_i} z_{p,i,f}$.

Some items represent words which express ideas which come up more often in texts. For $i \in \mathcal{I}$ let μ_i give the mean number of times item i appears in any given text, let $\mu = (\mu_i)_{i \in \mathcal{I}}$ and

$$(z_{p,i} \mid \mu_i) \sim \text{Poisson}(\mu_i) \quad \text{iid for } p \in \mathcal{P}.$$

This does not depend on location x_p as the selection of items (which depends on the topic of the text) is assumed to be location-independent. Let $\mathcal{Z}_{p,i} = (z_{p,i,1}, \ldots, z_{p,i,F_i})$. If forms are chosen independently then

$$(\mathcal{Z}_{p,i} \mid z_{p,i}, \phi_i(x_p)) \sim \text{Multinomial}(\phi_i(x_p), z_{p,i}) \quad \text{iid for } p \in \mathcal{P} \text{ and } i \in \mathcal{I}$$

so the joint distribution for the count data $Z_{\mathcal{P}} = (\mathcal{Z}_{p,i})_{p\in\mathcal{P}}^{i\in\mathcal{I}}$ is

$$p(Z_{\mathcal{P}}|\mu, \Phi(X_{\mathcal{P}})) = \prod_{p\in\mathcal{P}} \prod_{i\in\mathcal{I}} \prod_{f\in\mathcal{F}_i} \frac{e^{-\mu_i\phi_{i,f}(x_p)}(\mu_i\phi_{i,f}(x_p))^{z_{p,i,f}}}{z_{p,i,f}!}$$

where $\Phi(X_{\mathcal{P}}) = (\phi_i(x_p))_{p\in\mathcal{P}}^{i\in\mathcal{I}}$ collects the form frequencies for all items at all profile locations. Here $\Phi(X_{\mathcal{P}}) \in [0,1]^{d_\Phi}$ is a $d_\Phi = \sum_{i\in\mathcal{I}} F_i$ dimensional vector.

We now give the observation model for the observed presence/absence data $y_{p,i,f} \in \{0,1\},\ p \in \mathcal{P},\ I \in \mathcal{I},\ f \in \mathcal{F}_i$. We allow zero inflation in these records in order to account for missing items, so we do not simply set $y_{p,i,f} = \mathbb{I}(z_{p,i,f} > 0)$. Also, texts have topics, and this may preclude a given item appearing. Some "specialist" items may appear frequently when on-topic and otherwise be absent. Let $\zeta = (\zeta_i)_{i\in\mathcal{I}},\ \zeta_i \in [0,1]$ give the zeroing-probability parameters which vary from item to item. With these

$$p(y_{p,i,f} = 0 \mid \mu_i, \zeta_i, \phi_{i,f}(x_p)) = \zeta_i + (1-\zeta_i)\, e^{-\mu_i\phi_{i,f}(x_p)},$$

and so the joint likelihood for the binary data is

$$\begin{aligned} p(Y_{\mathcal{P}} \mid \mu, \zeta, \Phi(X_{\mathcal{P}})) &= \prod_{p\in\mathcal{P}} \prod_{i\in\mathcal{I}} \prod_{f\in\mathcal{F}_i} p(y_{p,i,f} \mid \mu_i, \zeta_i, \phi_i(x_p)) \\ &= \prod_{p\in\mathcal{P}} \prod_{i\in\mathcal{I}} \prod_{f\in\mathcal{F}_i} \zeta_i(1-y_{p,i,f}) + (1-\zeta_i)\beta_{i,f}(x_p) \end{aligned} \tag{5.4.1}$$

where $\beta_{i,f}(x_p) = (1-e^{-\mu_i\phi_{i,f}(x_p)})^{y_{p,i,f}}(e^{-\mu_i\phi_{i,f}(x_p)})^{1-y_{p,i,f}}$. Haines (2016) includes an additional outlier model which identifies some items within some profiles as unreliable.

5.4.2 Prior distributions

The parameters in the observational model in eq. (5.4.1) are the item intensities $\mu \in [0,\infty)^I$, the zero-inflation parameters $\zeta \in [0,1]^I$ and the form-probability fields $\Phi(X_{\mathcal{P}})$ at all locations $X_{\mathcal{P}} = (X_A, X_{\bar{A}})$. The unknown locations of floating profiles, $X_{\bar{A}} = (X_i)_{i\in\bar{A}}$, are also parameters of the model. In the following we assume an independent prior for each of these quantities.

We define the prior for the item intensity parameters μ following prior elicitation from an Atlas author. We use Gamma distributions with common shape and rate parameters $a = b = 3$ so

$$p(\mu) = \prod_{i=1}^{I} \text{Gamma}(\mu_i\,;\, a, b).$$

For the zero-inflation probabilities we use uniform priors so that

$$p(\zeta) = \prod_{i \in \mathcal{I}} \mathbb{I}(0 \leq \zeta_i \leq 1).$$

This reflects the fact that we expect zero inflation events to be rather common. For the missing locations $p(X_{\bar{A}})$, we assume independent uniform priors over the rectangular region of interest, rescaling easting $\times$ northing into $\mathcal{X} = [0, 1] \times [0, 0.9]$ and yielding

$$p(X_{\bar{A}}) \propto \prod_{p=N+1}^{N+M} \mathbb{I}(x_p \in \mathcal{X}).$$

It may be possible to elicit more informative priors for $x_p,\ p \in \bar{A}$, for example, proportional to population density if historical population density maps were available.

The latent linguistic fields giving form-usage probabilities have one field for each form of each item, in total $d_\phi = \sum_{i=1}^{I} F_i$ fields. Denote the collection of all *probability fields* as $\Phi = \{\phi_{i,f}(x)\}_{i \in \mathcal{I}, f \in \mathcal{F}_i, x \in \mathcal{X}}$. In building a model for these fields, three considerations apply: first, at any given location $x \in \mathcal{X}$, the fields associated with item i must sum to one, $\sum_{f \in \mathcal{F}_i} \phi_{i,f}(x) = 1$; second, field values at nearby locations should be correlated; and third, fields of different items should be correlated by substructure in the population of speakers.

In order to satisfy the location-additivity requirement, we parameterise the probability fields $\phi_{i,f}(x),\ i \in \mathcal{I}, f \in \mathcal{F}_i, x \in \mathcal{X}$ using a *softmax* transformation of *intensity fields* $\gamma_{i,f}(x) \in \mathbb{R}$ such that

$$\phi_{i,f}(x) = \frac{\exp(\gamma_{i,f}(x))}{\sum_{f' \in \mathcal{F}_i} \exp(\gamma_{i,f'}(x))}.$$

The second and third requirements are satisfied by using Gaussian Process (GP) priors (Rasmussen and Williams, 2005) for the form intensity fields. Specifically, we use a Linear Model of Coregionalization (LMC) (Journel and Huijbregts, 1978; Álvarez et al., 2012) to induce correlation across locations and between items. Let $\mathcal{B} = \{1, ..., B\}$ and $\Gamma = \{\gamma_b(x)\}_{b \in \mathcal{B},\, x \in \mathcal{X}}$ denote a collection of B independent GP *basis fields* with common kernel,

$$\gamma_b(x) \sim GP(0, k(\cdot, \cdot))\ ,\ b \in \mathcal{B},$$

and define the intensity fields as a linear combination these basis GPs by using a set of mixing weights and offsets,

$$\gamma_{i,f}(x) = a_{i,f} + \sum_{b=1}^{B} \gamma_b(x)\ w_{i,f,b}. \tag{5.4.2}$$

The use of LMC in a GP setting seems to be the key to the significant performance-improvement we see over Haines (2016), who used an independent Markov Random Field with a local L1 smoothness penalty for each field $\phi_{i,f}(x)$. Haines (2016) saw a complete collapse in reconstruction accuracy as items and floating profiles are added to a Bayesian analysis. The present model, in which B need not grow indefinitely with increasing I, avoids this collapse.

Let $a = (a_{i,f})_{i\in\mathcal{I},f\in\mathcal{F}_i}$ and $W = (w_{i,f,b})_{i\in\mathcal{I},f\in\mathcal{F}_i,b\in\mathcal{B}}$. We set independent Laplace priors for the weights to induce sparsity, and Gaussian priors for the offsets

$$p(a) = \prod_{i=1}^{I}\prod_{f=2}^{F_i} p(a_{i,f}) \qquad \text{with } a_{i,f} \sim \text{Normal}(0, \sigma_a)$$

$$p(W) = \prod_{b=1}^{B}\prod_{i=1}^{I}\prod_{f=1}^{F_i} p(w_{i,f,b}) \qquad \text{with } w_{i,f,b} \sim \text{Laplace}(0, \sigma_w)$$

In order to avoid identifiability issues, we set the offset for the first form of each item $a_{i,1}$ to 0. In our experiments, we use a Squared Exponential Covariance for the Kernel of the basis GPs. The kernel hyper-parameters were set via Hyper-Parameter Optimisation (HPO) (described in section 5.6.4).

The Linear Model of Coregionalization not only induces prior correlation between items, but also reduces the model dimension significantly. We only need B GPs priors, instead of the $\sum_{i=1}^{I} F_i - I$ priors required if we set independent priors for forms within an item. In our experiments, we find that a small number of basis fields (around $B \simeq 10$) are enough to capture the geographical structure of the linguistic fields (see section 5.7). This is orders of magnitude smaller than directly setting a prior on the hundreds of intensity fields. However, it is also desirable from a prior elicitation perspective, as a basis field which contributes significantly may be interpreted as a spelling "dialect".

Let $\Gamma_b(X_\mathcal{P}) = (\gamma_b(x_p))_{p\in\mathcal{P}}$ be the vector of values taken by one basis field $b \in \mathcal{B}$ at all profile locations, so that $\Gamma_b(X_\mathcal{P}) \in \mathbb{R}^P$ and let $\Gamma(X_\mathcal{P}) = (\Gamma_b(X_\mathcal{P}))_{b\in\mathcal{B}}$ be the vector of all basis-field values at all profile locations, with $\Gamma(X_\mathcal{P}) \in \mathbb{R}^{B\times P}$. In this construction, the prior for the probability fields $\Phi(X_\mathcal{P})$ at profile locations is determined by the GPs prior of the basis fields and the mixing coefficients, $p(\Gamma(X_\mathcal{P}), a, W)$.

5.4.3 Exact Joint Probabilistic Model

In this section we set out the joint model for the parameters and data, collecting all the elements described above into a joint probabilistic model. In the next section we will give an approximation to this distribution which is needed for tractability.

Given the deterministic relation between the probability fields Φ and their constituents (Γ, a, W) under the LMC, we rewrite the observational model from eq. (5.4.1) as

$$p(Y_{\mathcal{P}} \mid \mu, \zeta, \Phi(X_{\mathcal{P}})) \equiv p(Y_{\mathcal{P}} | \mu, \zeta, a, W, \Gamma(X_{\mathcal{P}})).$$

Let $\Gamma(X_A) = (\Gamma_b(X_A))_{b\in\mathcal{B}}$ and $\Gamma(X_{\bar{A}}) = (\Gamma_b(X_A))_{b\in\mathcal{B}}$ give the vectors of basis fields at anchors and floats respectively. The exact joint probabilistic model of data and parameters is

$$\begin{aligned} p(Y_{\mathcal{P}}, \mu, \zeta, a, W, \Gamma(X_{\mathcal{P}}), X_{\bar{A}}) =& p(Y_A|\mu, \zeta, a, W, \Gamma(X_A))\; p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\bar{A}})) \\ & \times p(\Gamma(X_A), \Gamma(X_{\bar{A}}) \mid X_{\bar{A}}) \\ & \times p(X_{\bar{A}})\; p(\mu, \zeta, a, W) \end{aligned} \tag{5.4.3}$$

Exact inference on this probabilistic model becomes computationally infeasible with a large number of profiles in the data. The joint distribution of the values $\Gamma_b(X_{\mathcal{P}})$ taken by one basis field $b \in \mathcal{B}$ at all profile locations, is a multivariate Gaussian of dimension $P = N + M$, with zero mean and covariance matrix $K_P = (k(x_i, x_j))_{i,j\in\mathcal{P}}$,

$$\Gamma_b(X_{\mathcal{P}}) \sim \text{Normal}(0, K_P),$$

and we need to compute the inverse of K_P many times over when targeting the posterior. This wont scale to the full LALME dataset, which contains thousands of profiles P.

5.4.4 Sparsity in Latent Fields via Inducing Points.

In order to overcome the scalability challenge on the number of profiles, we introduce a sparse approximation to the joint distribution in eq. (5.4.3) via inducing points (Titsias, 2009; Bonilla et al., 2019; Quiñonero-Candela and Rasmussen, 2005). Inducing values augment the prior specification and serve as *global parameters* that are particularly convenient in SVI (Hensman et al., 2013, 2015). There are two steps to the approximation: we indicate the exact joint model by $p()$ in eq. (5.4.3) and the first and second stage of the approximation by $p^{(1)}$ and $p^{(2)}$ in eq. (5.4.6) and eq. (5.4.9) below.

Consider a set of U *inducing points*, $X_{\mathcal{U}} = (x_u)_{u\in\mathcal{U}}$, $\mathcal{U} = \{1, \ldots, U\}$ and for $b \in \mathcal{B}$ let $\Gamma_b(X_{\mathcal{U}}) = (\gamma_b(x_u))_{u\in\mathcal{U}}$. Here $\Gamma_b(X_{\mathcal{U}})$ is a multivariate Gaussian random variable of dimension U, with zero mean and covariance matrix $K_U = (k(x_i, x_j))_{i,j\in\mathcal{U}}$,

$$\Gamma_b(X_{\mathcal{U}}) \sim \text{Normal}(0, K_U).$$

We call $\Gamma_b(X_{\mathcal{U}})$ the *inducing values* of basis field b. Let $\Gamma(X_{\mathcal{U}}) = (\Gamma_b(X_{\mathcal{U}}))_{b\in\mathcal{B}}$ be the vector of all inducing values for all basis fields.

In the sparse approximation the prior distribution for the basis GPs at profiles, $p(\Gamma(X_{\mathcal{P}}) = (\Gamma(X_A), \Gamma(X_{\bar{A}}))$, and inducing locations, $\Gamma(X_{\mathcal{U}})$, is approximated by assuming conditional independence between fields values, $\Gamma(X_A)$ and $\Gamma(X_{\bar{A}})$, at anchor and floating profiles given the inducing values, that is

$$\begin{aligned} p(\Gamma(X_{\mathcal{P}}), \Gamma(X_{\mathcal{U}}) \mid X_{\bar{A}}) &= p(\Gamma(X_A)), \Gamma(X_{\bar{A}}) \mid \Gamma(X_{\mathcal{U}}), X_{\bar{A}})\; p(\Gamma(X_{\mathcal{U}}) \\ &\approx p(\Gamma(X_A) \mid \Gamma(X_{\mathcal{U}}))\; p(\Gamma(X_{\bar{A}}) \mid \Gamma(X_{\mathcal{U}}), X_{\bar{A}})\; p(\Gamma(X_{\mathcal{U}})). \end{aligned} \tag{5.4.4}$$

This is likely to be a better approximation than is often seen in inducing point approximations, as it retains correlation within the anchor group and within the group of floats. One popular setup (Hensman et al., 2015) would take all fields at all profile locations to be jointly independent given the inducing values.

In our experiments, we define a homogeneous square grid of inducing points along the region of interest by dividing in 10 sections vertically and horizontally, for a total of 121 inducing points. The space between inducing points corresponds to approximately 20 km in the map which is somewhat shorter than the scale of variation we recover (see fig. 5.15). Our approximation will be good if the inter-point spacing of inducing points is smaller than the length scale for variation in the basis fields.

We now specify the factors appearing in eq. (5.4.4). In fig. 5.4 we display a graphical representation of the model we are describing. The joint distribution of the values of basis field b at anchor locations conditioned on the values taken by field b at inducing points is

$$(\Gamma_b(X_A) \mid \Gamma_b(X_{\mathcal{U}})) \sim \text{Normal}(K_{A,U}K_U^{-1}\Gamma_b(X_{\mathcal{U}}),\; K_A - K_{A,U}K_U^{-1}K_{A,U}^t), \tag{5.4.5}$$

where $K_A = (k(x_p, x_q))_{p,q\in A}$ is the covariance matrix between anchor locations, and $K_{A,U} = (k(x_i, x_j))_{i\in A\,,\, j\in U}$ is the cross-covariance matrix between anchor profiles and inducing locations. The conditional distribution $(\Gamma_b(X_{\bar{A}}) \mid \Gamma_b(X_{\mathcal{U}}), X_{\bar{A}})$ of the basis field values at floating profiles locations $X_{\bar{A}}$ is given by eq. (5.4.5) with $A \to \bar{A}$.

By the prior independence of basis fields we have $p(\Gamma(X_{\mathcal{U}})) = \prod_{b=1}^{B} p(\Gamma_b(X_{\mathcal{U}}))$, a product of multivariate Gaussians of dimension U. Similarly,

$$p(\Gamma(X_A) \mid \Gamma(X_{\mathcal{U}})) = \prod_{b=1}^{B} p(\Gamma_b(X_A) \mid \Gamma_b(X_{\mathcal{U}})),$$

a product of multivariate Gaussians of dimension N. For the floating profiles,

$$p(\Gamma(X_{\bar{A}}) \mid \Gamma(X_{\mathcal{U}}), X_{\bar{A}}) = \prod_{b=1}^{B} p(\Gamma_b(X_{\bar{A}}) \mid \Gamma_b(X_{\mathcal{U}}), X_{\bar{A}}),$$

the factors each have dimension M. The approximate joint density after this first factorising approximation is

$$\begin{aligned} p^{(1)}(Y_{\mathcal{P}}, \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), \Gamma(X_{\mathcal{P}}), X_{\bar{A}}) =& p(Y_A|\mu, \zeta, a, W, \Gamma(X_A))\; p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\bar{A}})) \\ &\times p(\Gamma(X_A) \mid \Gamma(X_{\mathcal{U}}))\; p(\Gamma(X_{\bar{A}}) \mid \Gamma(X_{\mathcal{U}}), X_{\bar{A}}) \\ &\times p(X_{\bar{A}})\; p(\Gamma(X_{\mathcal{U}}))\; p(\mu, \zeta, a, W) \end{aligned} \tag{5.4.6}$$

We make an additional approximation in order to compute the joint log-probability: we make an approximate integration over the fields at profile locations, following Hensman et al. (2015),

$$\begin{aligned} \log p(Y_A \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})) &= \log \left[\mathbb{E}_{p(\Gamma(X_A)|\Gamma(X_{\mathcal{U}}))}\, p(Y_A|\mu, \zeta, a, W, \Gamma(X_A))\right] \\ &\geq \mathbb{E}_{p(\Gamma(X_A)|\Gamma(X_{\mathcal{U}}))} \left[\log p(Y_A|\mu, \zeta, a, W, \Gamma(X_A))\right] \\ &\equiv \log \tilde{p}(Y_A \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})) \qquad (5.4.7) \\ \log p(Y_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}}) &\geq \mathbb{E}_{p(\Gamma(X_{\bar{A}})|\Gamma(X_{\mathcal{U}}), X_{\bar{A}})} \left[\log p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\bar{A}}), X_{\bar{A}})\right] \\ &\equiv \log \tilde{p}(Y_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}}). \qquad (5.4.8) \end{aligned}$$

We approximate the likelihoods on the left with the lower bounds on the right, so that the final approximated joint probability distribution is

$$\begin{aligned} p^{(2)}(Y_{\mathcal{P}}, \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}}) = &\tilde{p}(Y_A \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})) \\ &\times \tilde{p}(Y_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}}) \\ &\times p(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})) + \log p(X_{\bar{A}}). \end{aligned} \tag{5.4.9}$$

Whenever we need to evaluate eq. (5.4.7) or eq. (5.4.8) we estimate their values by drawing joint samples $\Gamma(X_A) \sim p(\cdot \mid \Gamma(X_{\mathcal{U}}))$ and $\Gamma(X_{\bar{A}}) \sim p(\cdot \mid \Gamma(X_{\mathcal{U}}), X_{\bar{A}})$ and averaging. We found by experiment that 100 samples was sufficient for stable downstream inference.

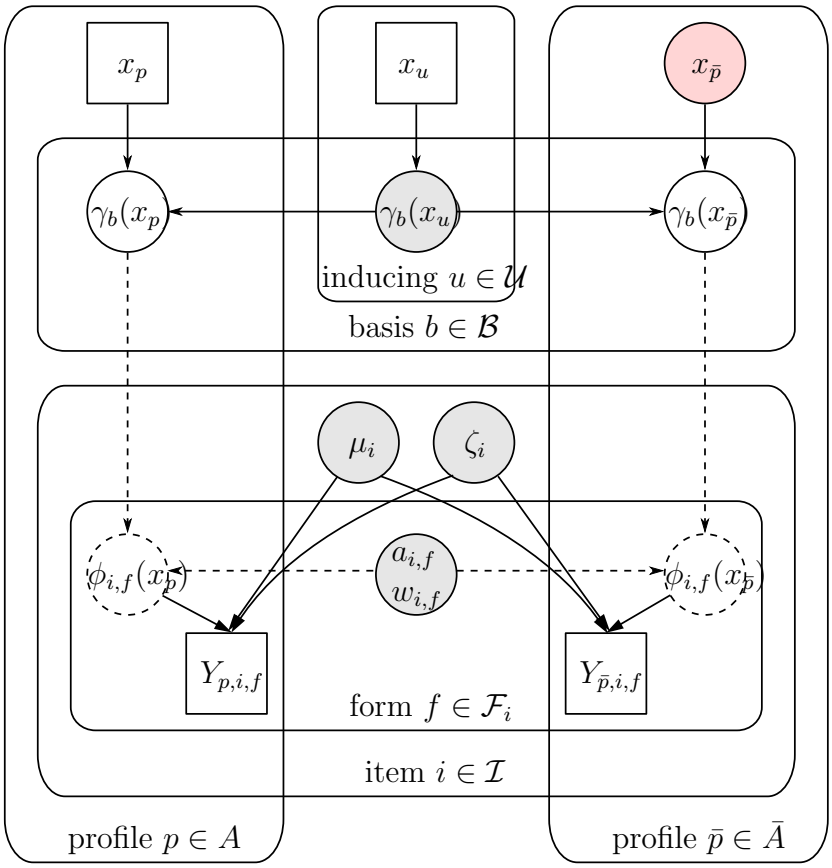


Figure 5.4: Probabilistic Graphical Model for the Linguistic Data. Squares represent observed data, circles represent model parameters. Global parameters are at the core of the model, colored in grey, and the missing locations colored in red. Probabilistic relationships are represented by solid lines, while dashed lines represent deterministic relations. Plates denote multiplicity along sets.

5.5 Semi-Modular Inference for the LALME model

In the context of a different model for the spatial fields, Haines (2016) observes that jointly inferring the spatial fields and the locations of all floating profiles yields poor results, with floating profiles often clustering together at the borders of the region. Model error in the field-prior is amplified by feedback from floating profiles and can dominate information from anchors. This suggested a *Cut Posterior* setup (Haines, 2016, p. 125) in which the posterior distribution of the fields was estimated using the anchor profiles alone, and then the posterior distribution of the floating-profile locations was obtained by averaging over the field posteriors. This two-stage approach, which eliminates information flow from the floating profiles to the spatial fields, out-

performed Bayesian inference.

Reducing feedback between *modules* is an effective way to alleviate model contamination (Liu et al., 2009; Jacob et al., 2017). A graphical representation of this Cut in the LALME model is displayed in fig. 5.5. If we group together elements of the probabilistic model in fig. 5.5 we get exactly the two-module model in fig. 5.3,

$$\begin{aligned} Z &\equiv Y_A, \qquad Y \equiv Y_{\bar{A}}, \\ \theta &\equiv X_{\bar{A}}, \qquad \psi \equiv (\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})). \end{aligned} \tag{5.5.1}$$

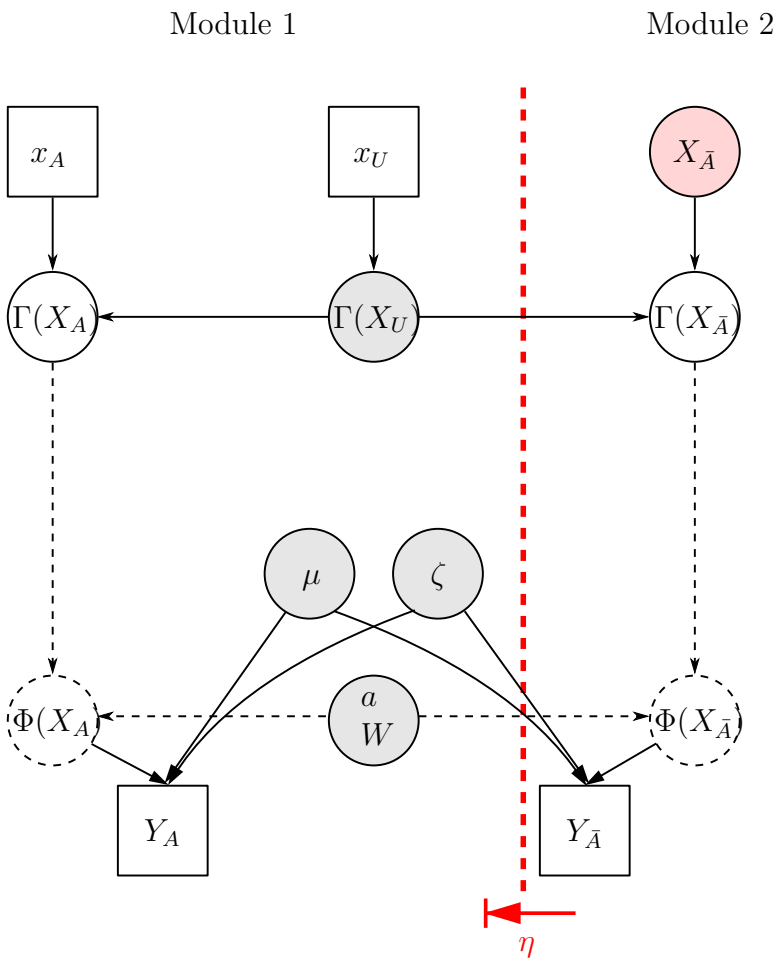


Figure 5.5: Probabilistic Graphical Model for the Linguistic Data. The plates are omitted for simplicity. Notation is otherwise as fig. 5.4. The red dashed line indicates the cut. In the cut-posterior, parameters on the left side of the cut are inferred independently of any information from the right side. In SMI, information from the right is downweighted.

We begin by giving the cut posterior and then extend the analysis to give the SMP. Using our model specification, the cut posterior eliminating the feedback from

floating profiles into linguistic fields is given by

$$p_{cut}(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}} \mid Y_{\mathcal{P}}) = p(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}) \mid Y_A) \times \quad p(X_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), Y_{\bar{A}}), \tag{5.5.2}$$

which, when written in the notation of eq. (5.5.1), is identical to eq. (5.3.1). The approximations given in eq. (5.4.7) and eq. (5.4.8) are needed in both stages, so that

$$p(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}) \mid Y_A) \propto \tilde{p}(Y_A \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))\; p(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})),$$
$$p(X_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), Y_{\bar{A}}) \propto \tilde{p}(Y_{\bar{A}} | \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}})\; p(X_{\bar{A}}),$$

where we omitted a factor $p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))$ in the second denominator. This is constant for any fixed set of conditioning variables, and corresponds to $p(Y|\psi)$ in eq. (5.3.1). This wont play a role in our MCMC or SVI due to the two-stage structure they share.

We now extend the modularized analysis of Haines (2016) to Semi-Modular Inference (SMI) (Carmona and Nicholls, 2020). In order to define the Semi-Modular Posterior in our setting, we simply drop the objects defined in eq. (5.5.1) into the power posterior eq. (5.3.2) and SMI posterior eq. (5.3.3). We define a scalar influence parameter $\eta \in [0, 1]$ which enters as a power on the likelihood factor for floats and define the *powered* joint distribution

$$p_{\text{pow},\eta}(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}} \mid Y_{\mathcal{P}}) \propto \tilde{p}(Y_A|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})) \times \tilde{p}(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}})^{\eta} \times p(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))\; p(X_{\bar{A}}). \tag{5.5.3}$$

We have replaced exact likelihoods $p(Y_A|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))$ and $p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})$ with the lower bound approximations given in eq. (5.4.7) and eq. (5.4.8). The notation of eq. (5.5.1) identifies eq. (5.5.3) with eq. (5.3.2).

The Semi-Modular Posterior is obtained by using this power-posterior and creating auxiliary copies of selected model parameters, namely $\tilde{\theta} \equiv \tilde{X}_{\bar{A}}$. The Semi-Modular Posterior for the LALME model is then

$$p_{\text{smi},\eta}(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), X_{\bar{A}}, \tilde{X}_{\bar{A}} \mid Y_{\mathcal{P}}) \propto p_{\text{pow},\eta}(\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), \tilde{X}_{\bar{A}} \mid Y_{\mathcal{P}}) \times p(X_{\bar{A}} \mid \mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}), Y_{\bar{A}}). \tag{5.5.4}$$

In the notation of eq. (5.5.1), eq. (5.5.4) is

$$p_{\text{smi},\eta}(\psi, \theta, \tilde{\theta} \mid Y_A, Y_{\bar{A}}) = p_{\text{pow},\eta}(\psi, \tilde{\theta}|Y_A, Y_{\bar{A}})\; p(\theta|\psi, Y_{\bar{A}}) \tag{5.5.5}$$

and so we can identify the Semi-Modular Posterior in eq. (5.5.4) for the model in fig. 5.5 with the Semi-Modular Posterior in eq. (5.3.3) for the model in fig. 5.3. The Semi-Modular Posterior in eq. (5.5.4) interpolates between the cut posterior in eq. (5.5.2) and the (approximate) Bayes posterior in eq. (5.4.9).

5.5.1 Choosing the rate of influence

The selection of the influence parameter η will depend on the goals of inference. We define a *loss function* $V : [0, 1] \to \mathbb{R}$ that measures success in the inference and chose the influence parameter that minimises this loss,

$$\eta^* = \arg\min_{\eta} V(\eta).$$

The SMP distribution corresponding to η^* is used for downstream analyses.

In the context of the present problem, the loss function should measure the accuracy of our $X_{\bar{A}}$-estimation as a function of the influence η of the float data $Y_{\bar{A}}$. We do not of course know the true float locations, $X^*_{\bar{A}}$ say, but we do have an obvious proxy in the anchor locations X_A. We therefore take the Posterior Mean Squared Error for predicting anchor locations as our loss. We treat $K = 20$ anchors as floats and adjust η so that the posterior for their locations are concentrated on their true locations using the posterior MSE as a loss encapsulating a penalty for the bias and variance of location estimates.

Let $\mathcal{K} = \{i_1, ..., i_K\}$ denote the indices of the "held out" anchors, so that $\mathcal{K} \subset \mathcal{I}$, and $A \setminus \mathcal{K}$ are the indices of the remaining anchors. Relabel the true locations of the held out anchors $X^*_{\mathcal{K}}$ and let $X_{\bar{A}\cup\mathcal{K}} \in \mathcal{X}^{M+K}$ denote all the locations we wish to estimate in this cross-validation analysis and let $\tilde{X}_{\bar{A}\cup\mathcal{K}}$ be a second copy of the same parameters. In this analysis the SMP in eq. (5.5.5) is

$$p^{(val)}_{\text{smi},\eta}(\psi, X_{\bar{A}\cup\mathcal{K}}, \tilde{X}_{\bar{A}\cup\mathcal{K}} \mid Y_{\mathcal{P}}) = \; p_{\text{pow},\eta}(\psi, \tilde{X}_{\bar{A}\cup\mathcal{K}} | Y_{A\setminus\mathcal{K}}, Y_{\bar{A}\cup\mathcal{K}})\; p(X_{\bar{A}\cup\mathcal{K}} | \psi, Y_{\bar{A}\cup\mathcal{K}}), \quad (5.5.6)$$

and we have the standard SMI setup of eq. (5.3.3) with

$$\begin{aligned} Z &\equiv Y_{A\setminus\mathcal{K}}, \qquad & Y &\equiv Y_{\bar{A}\cup\mathcal{K}}, \\ \theta &\equiv X_{\bar{A}\cup\mathcal{K}}, \qquad & \psi &\equiv (\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}})). \end{aligned} \quad (5.5.7)$$

Our loss function is then,

$$V(\eta) = K^{-1} \sum_{p\in\mathcal{K}} \mathbb{E}_{p^{(val)}_{\text{smi},\eta}} \left[\|x_p - X^*_p\|^2 \right], \quad (5.5.8)$$

where $\|x_p - X_p^*\|$ is Euclidean norm and for $p \in \mathcal{K}$, x_p and X_p^* are the posterior and known true profile locations for the $K = 20$ held out anchors.

A design like this is discussed in Mao et al. (2022). If, for example, the pairs $(x_p, y_p)_{p\in\mathcal{P}}$ were exchangeable (so $x_p,\ p \in \bar{A}$ is unknown, but the missingness is random) then the distribution of the "held out" anchor loss $\|x_p - X_p^*\|^2,\ p \in \mathcal{K}$ at any given η would be the same as the "ideal" loss $\|x_p - X_p^*\|^2$ for $p \in \bar{A}$ a float, and hence a good proxy for the ideal target loss. Exchangeability does not hold in our spatial setting so this idealised. A form of local-exchangeability developed by Mao et al. (2022) may hold. However, we have not explored this.

5.6 Implementation details

5.6.1 Data settings

Recall our description of the LALME data from section 5.2. The observations come from a rectangular region spanning 200 km east-west and 180 km north-south. The region contains $N = 120$ anchor profiles and $M = 247$ floating profiles. In these profiles we have in all $I = 71$ items producing $\sum_{i=1}^{I} f_i = 741$ distinct forms. We consider four datasets in our experiments: the *full dataset* containing all data in the region, a *validation dataset* obtained from the full dataset by treating twenty anchors as floats, a *medium dataset* with all the floating profiles but just a subset of the items and a *small dataset* with just a subset of floating profiles and items,

Full dataset: 367 profiles ($N = 120, M = 247$), $I = 71$ items, $\sum_{i=1}^{I} f_i = 741$ forms,

Validation dataset: 367 profiles ($N = 100, M = 267$), $I = 71$ items, $\sum_{i=1}^{I} f_i = 741$ forms,

Medium dataset: 367 profiles ($N = 120, M = 247$), $I = 8$ items, $\sum_{i=1}^{I} f_i = 79$ forms,

Small dataset: 120 profiles ($N = 120, M = 10$), $I = 8$ items, $\sum_{i=1}^{I} f_i = 79$ forms.

In table 5.1 we provide a count of the number of parameters required by the LALME model for the small and full datasets. The small dataset is motivated as a baseline to compare the estimation between MCMC and variational methods. We found it infeasible to use MCMC when the number of floating profiles is at all large, as we could not get convergence in a reasonable time. This is demonstrated using the medium dataset.

We set the number of basis fields in the LMC model in eq. (5.4.2) to $B = 10$ in all analyses. The number of fields has to be large enough to allow the LMC

model to express complex spatial structures over dialect fields. The B values selected are "large enough", in the sense that not all the basis fields are needed: the posterior distribution for some basis fields concentrates on zero; the model has captured enough geographical structure with the remaining fields to model the data and any increase in B would just add redundant basis fields. In fig. 5.15 we show the $B = 10$ basis fields for the full dataset and observe that some of the fields are practically flat over the entire region.

The number of inducing points is $U = 121$ in all analyses. These correspond to the vertices of a homogeneous rectangular grid dividing the region vertically and horizontally into 10^2 sections of roughly 20 km^2. This reflects a prior belief that 20 km is at or below the length scale over which dialectic variation occurs. With these inducing points we could reconstruct variation on shorter scales if it were there, though we would not accurately quantify uncertainty in such structures due to the nature of the approximations we make. The variational setup would in principle allow us to learn the inducing locations as one more parameter in the model. However, this would be computationally demanding as it requires re-computation of covariance matrices at each iteration.

Parameter	Dimension	Full dataset	Small dataset
μ	I	71	8
ζ	I	71	8
a	$\sum f_i$	741	72
W	$B\sum f_i$	7,410	720
$\Gamma(X_{\mathcal{U}})$	BU	1,210	1,210
Subtotal (global params)		9503	2018
$X_{\bar{A}}$	$2M$	494	20
Subtotal (with floating locations)		9997	2038
$\Gamma(X_A)$	BN	1,200	1,200
$\Gamma(X_{\bar{A}})$	BM	2,470	100
Total		13,667	3,338

Table 5.1: Number of parameters in the model. The parameters $\Gamma(X_A) and \Gamma(X_{\bar{A}})$ are integrated out of the final posteriors in eq. (5.5.4) and eq. (5.5.6) using Monte Carlo and the approximation in eq. (5.4.9) (these are not output by the x trained Normalizing Flow).

5.6.2 Inference via MCMC

We use Markov chain Monte Carlo (MCMC) methods to perform inference for the LALME model using the small dataset.

Our implementation follows the two-stage MCMC procedure to target the Semi-Modular Posterior (Carmona and Nicholls, 2020) given in section 5.3.2.3. The two stage approach is needed to deal with the intractable ψ-dependent normalising constant in $p(\theta|\psi, Y) = p(Y_{\bar{A}}|\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))$. In order to compare SMPs across different values of the influence parameter η, we repeat this over a grid of values of $\eta \in [0, 1]$.

In both stages, we use a standard No U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) implemented in Blackjax (Lao and Louf, 2020; Lao et al., 2020). This software is convenient for the implementation of the nested MCMC algorithm, as it allows easily to parallelize the sub-chains in the second stage of sampling thanks to the vectorization capabilities of Jax (Bradbury et al., 2018). Moreover, we can make a reasonable comparison between the cost of inference using MCMC and variational methods by using a shared implementation of the joint log-probability function.

We provide software in our supplementary material.

5.6.3 Inference via Variational Methods

We obtain an approximation to the Semi-Modular Posterior in eq. (5.5.4) at each $\eta \in [0, 1]$, using the variational methods given in Carmona and Nicholls (2022). In this setup we learn a Variational Meta-Posterior (VMP) which parameterises an approximation to the entire family of SMI posteriors indexed by the influence parameter η. This is similar in spirit to Giordano et al. (2018, 2022) who consider perturbations of prior hyper-parameters in variational Bayes posteriors. In contrast we learn the entire family of Variational Meta-Posteriors and vary the influence parameter η indexing the inference scheme. We use a VMP-flow (Carmona and Nicholls, 2022) constructed by composing several 8 conditional coupling layers of Neural Spline Flows (NSFs) (Durkan et al., 2019; Dinh et al., 2016) followed by a layer to map the flow into the domain of the model parameters.

In order to construct a variational family for approximation of the SMP in eq. (5.5.5), take $\psi = (\mu, \zeta, a, W, \Gamma(X_{\mathcal{U}}))$ as above, and write the joint variational posterior as

$$q_{\lambda_{1:3},\eta}(\psi, X_{\bar{A}}, \tilde{X}_{\bar{A}}) = q_{\lambda_1,\eta}(\psi)\; q_{\lambda_2,\eta}(X_{\bar{A}} \mid \psi)\; q_{\lambda_3,\eta}(\tilde{X}_{\bar{A}} \mid \psi) \tag{5.6.1}$$

where $\lambda_{1:3} = \{\lambda_1, \lambda_2, \lambda_3\}$ are variational parameters which do not depend on η and we use NSFs for each factor. The variational family in eq. (5.6.1) depends on η using the parameterisation given in Carmona and Nicholls (2022) which adds η to the conditioner in our coupling-layer conditioner. The trained family of variational densities $q_{\lambda_{1:3},\eta},\ \eta \in [0,1]$ is called the Variational Meta-Posterior (VMP).

The optimal λ-parameters $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ maximise a variant of the Evidence Lower Bound (ELBO) as explained in Carmona and Nicholls (2022). A similar revised target utility is given in Yu et al. (2021). This modification of the usual variational utility is needed in order to stop the variational approximation introducing uncontrolled information flow between modules. The utilities connecting the Semi-Modular Posterior in eq. (5.5.5) to the variational family eq. (5.6.1) are

$$\begin{aligned}
\text{ELBO}_{\text{pow},\eta}(\lambda_1, \lambda_3) &= \mathbb{E}_{(\psi,\tilde{X}_{\bar{A}})\sim q_{\lambda_1,\lambda_3}}[\log p_{\text{pow},\eta}(\psi, \tilde{X}_{\bar{A}}, Y_A, Y_{\bar{A}}) - \log q_{\lambda_1,\lambda_3}(\psi, \tilde{X}_{\bar{A}})] \\
\text{ELBO}_{\text{bayes}}(\lambda_1, \lambda_2) &= \mathbb{E}_{(\psi,X_{\bar{A}})\sim q_{\lambda_1,\lambda_2}}[\log p(\psi, X_{\bar{A}}, Y_{\bar{A}}) - \log q_{\lambda_1,\lambda_2}(\psi, X_{\bar{A}})]
\end{aligned}$$

and the simplest two-stage variational scheme at a fixed η value would take

$$\begin{aligned}
\hat{\lambda}_1^{(\eta)}, \hat{\lambda}_3^{(\eta)} &= \arg\max_{\lambda_1,\lambda_3} \text{ELBO}_{\text{pow},\eta}(\lambda_1, \lambda_3), \\
\hat{\lambda}_2^{(\eta)} &= \arg\max_{\lambda_2} \text{ELBO}_{\text{bayes}}(\hat{\lambda}_1^{(\eta)}, \lambda_2).
\end{aligned}$$

These $\hat{\lambda}_{1:3}^{(\eta)}$ depend on η so this is not yet the VMP. As explained in Carmona and Nicholls (2022), the two stage scheme can be implemented in a single end-to-end scheme using the stop-gradient operator $\not\nabla(\psi)$. This "protects" its argument from any derivative in any term in which it appears. The ELBO for the second stage becomes

$$\text{ELBO}_{\text{bayes}\not\nabla(\psi)}(\lambda_1, \lambda_2) = \mathbb{E}_{(\psi,X_{\bar{A}})\sim q_{\lambda_1,\lambda_2}}[\log p(\not\nabla(\psi), X_{\bar{A}}, Y_{\bar{A}}) - \log q_{\lambda_1,\lambda_2}(\not\nabla(\psi), X_{\bar{A}})]$$

and the overall utility is

$$\mathcal{L}_{(\text{smi},\eta)}(\lambda_{1:3}) = \text{ELBO}_{\text{pow},\eta}(\lambda_1, \lambda_3) + \text{ELBO}_{\text{bayes}\not\nabla(\psi)}(\lambda_1, \lambda_2). \tag{5.6.2}$$

In a VMP we learn the variational approximation at each η simultaneously. Following Carmona and Nicholls (2022) we take an auxiliary distribution $\rho(\eta)$ over $[0,1]$ and define

$$\mathcal{L}_{(\text{msmi}-\text{flow})}(\lambda_{1:3}) = E_{\eta\sim\rho(\cdot)}(\mathcal{L}_{(\text{smi},\eta)}(\lambda_{1:3})). \tag{5.6.3}$$

The optimal variational parameters are then

$$\hat{\lambda}_{1:3} = \arg\max_{\lambda_{1:3}} \mathcal{L}_{(\text{msmi}-\text{flow})}(\lambda_{1:3}). \tag{5.6.4}$$

These $\hat{\lambda}_{1:3}$-values do not depend on η and define $q_{\hat{\lambda}_{1:3},\eta}$, the VMP. The optimisations are carried out using the reparameterisation trick and SGD.

We take $\rho(\cdot)$ to be Beta$(1/2, 1/2)$, reflecting our experience that accurate estimation of the VMP near Bayes and Cut ($\eta \simeq 0, 1$) is important, because the utility $V(\eta)$ is often rapidly varying in these regions. We will estimate $\mathcal{L}_{(\text{msmi}-\text{flow})}$ using samples $\eta \sim \rho(\cdot)$, so taking a continuous density with support on all $[0, 1]$ but more samples at the extremes is helpful for convergence of SGD over $\lambda_{1:3}$. However, NSFs are relatively expressive, and if there exists $\hat{\lambda}_{1:3}$ maximising $\mathcal{L}_{(\text{smi},\eta)}(\lambda_{1:3})$ at every η then $\hat{\lambda}_{1:3}$ also maximises $\mathcal{L}_{(\text{msmi}-\text{flow})}(\lambda_{1:3})$ for *any* distribution ρ over $[0, 1]$, so while ρ defines the absolute utility of $\lambda_{1:3}$ it has less impact on the final fitted VMP.

In order to find the value of η maximising the utility $V(\eta)$ in eq. (5.5.8), we target the validation posterior in eq. (5.5.6). The variational family targeting the validation-SMP in eq. (5.5.6) has the form

$$q^{(val)}_{\lambda_{1:3},\eta}(\psi, X_{\bar{A}\cup\mathcal{K}}, \tilde{X}_{\bar{A}\cup\mathcal{K}}) = q_{\lambda_1,\eta}(\psi)\; q_{\lambda_2,\eta}(X_{\bar{A}\cup\mathcal{K}} \mid \psi)\; q_{\lambda_3,\eta}(\tilde{X}_{\bar{A}\cup\mathcal{K}} \mid \psi).$$

Having estimated $q^{(val)}_{\hat{\lambda}_{1:3},\eta}$ as a function of η using the same methods as for $q_{\hat{\lambda}_{1:3},\eta}$, we choose η to minimise the loss in eq. (5.5.8). This loss function is approximated by taking the variational-posterior Mean Squared Error of the reconstructed anchor locations,

$$V^{(val)}(\eta) = K^{-1} \sum_{p\in\mathcal{K}} \mathbb{E}_{q^{(val)}_{\hat{\lambda}_{1:3},\eta}} \left[\|x_p - X^*_p\|^2\right] \tag{5.6.5}$$

and then η^* is estimated by simply evaluating this at a selection of η-values.

5.6.4 Hyper-Parameters Optimisation (HPO)

We used Hyper-Parameter Optimisation to select a few parameters used for training the model. In particular, we selected four quantities: the decay factor and peak value for the AdaBelief learning rate schedule (Zhuang et al., 2020), as well as the amplitude and length scale for training the Exponential Quadratic kernel for the GP prior of basis fields. The target metric to minimise was the Mean Squared Error defined in eq. (5.6.5) using the validation data setting described above. Search was performed via Bayesian Optimisation (BP) (Snoek et al., 2012) implemented in the Syne Tune library (Salinas et al., 2022), and executed in 40 distributed training jobs using Amazon SageMaker (Perrone et al., 2020). A pleasing observation during the optimisation process was that the metric values stayed in a stable narrow range after a few (≈ 10) iterations.

5.7 Data Analyses

In this section we present our experimental results for the LALME model in section 5.4.

Our discussion will proceed in an incremental complexity. We start with a conventional Bayesian analysis which estimates the Bayes posterior on floating profiles using the small dataset via MCMC. This is useful for comparison with our later analysis on the full dataset. We then give the VMP and MCMC approximations to the SMI-posterior in eq. (5.5.4) for the predicted locations of floating profiles using the small dataset. We trust the MCMC to give a reasonable summary of this SMI-posterior at any fixed η so this gives a check on the VMP: is our NSF sufficiently expressive, and our ELBO-maximisation via SGD sifficiently accurate, to reproduce the ground truth distributions given by MCMC on a dataset small enough to allow MCMC simulation in reasonable time. We illustrate the difficulties with MCMC when the number of floating profiles is too large using the medium dataset.

Having established our methods on the small dataset we then apply them to the full dataset where MCMC is not feasible. We select the influence parameter η using the validation dataset. The same analysis is used to measure the accuracy of our location estimates for the held-out anchors. We then present the results of variational inference on the full dataset with the Semi-Modular Posterior at the optimal η-value.

5.7.1 Conventional Bayesian Posterior on the Small dataset.

We start off with a small baseline to assess our model in a manageable setting. The small dataset is obtained by taking 8 items, all 120 anchors and 10 floating profiles from the full dataset. With fewer items we require fewer dialect basis fields, and this gives a helpful reduction in parameter dimension. With a small number of floating profiles, the fields are mainly informed by the anchor locations, reducing any damage due to feedback between imputed locations for floating profiles and latent fields.

In this setting, MCMC on the conventional Bayesian posterior in eq. (5.4.9) (conditioned on $Y_{\mathcal{P}}$) is satisfactory overall. We observe good mixing for the global parameters μ and ζ in fig. 5.6, comparing samples for four chains of the NUTS MCMC sampler. The zeroing probabilities $\zeta_i,\ i \in \mathcal{I}$ are not small, ranging from around 0.3 to 0.7. The mean useage frequencies μ vary from around 5 to 15. Marginal posterior location-distributions for a subset of 10 representative floating profiles are displayed (in blue) in fig. 5.7, and their corresponding traces for northing and easting coordinates in fig. 5.8. We select 10 floating profiles which "cover" the window $\mathcal{X}$ (using

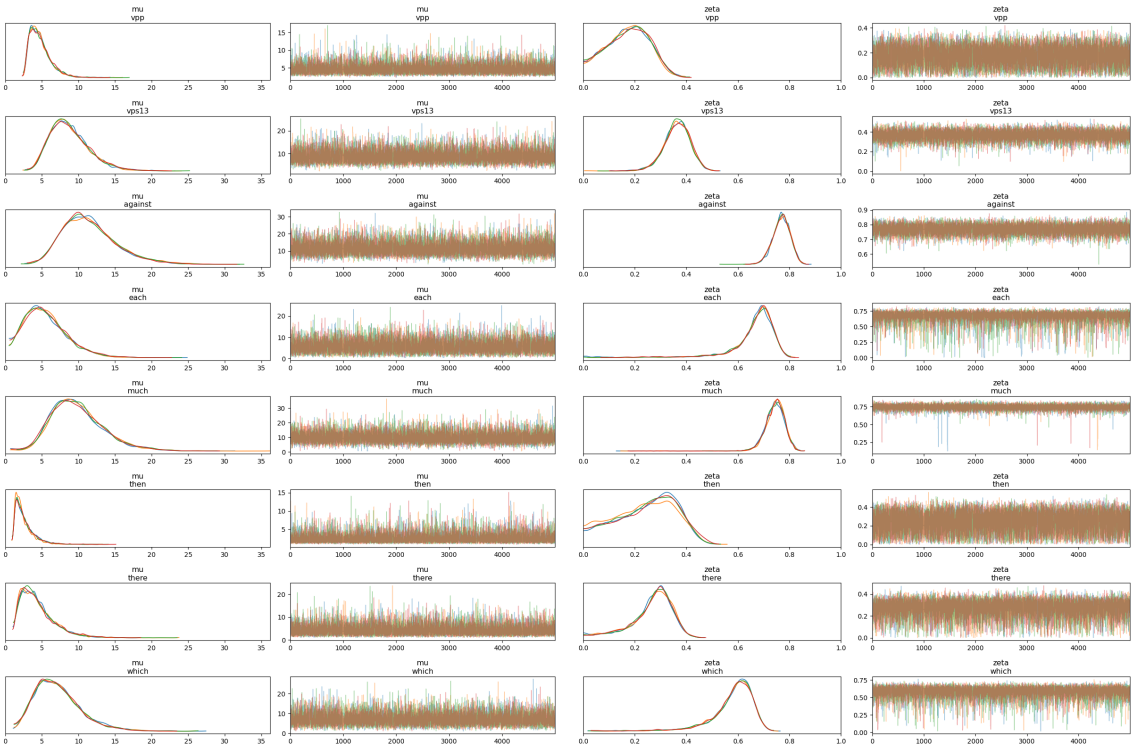


Figure 5.6: Distributions and trace plots for μ_i (left) and ζ_i (right), $i = 1, \ldots, I$. We compare four MCMC chains targeting the conventional Bayesian posterior (eq. (5.4.9) or equivalently $\eta = 1$ in eq. (5.5.4)) on the small dataset.

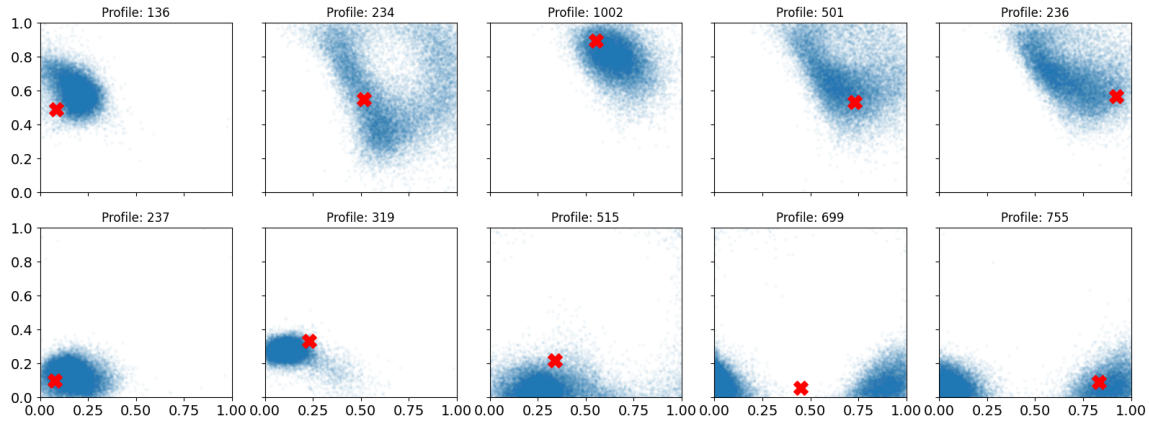


Figure 5.7: Posterior samples (in blue) for the 2D locations of a subset of floating profiles. The samples correspond to the Bayesian posterior on the small dataset. The red cross indicates the Atlas location estimated using the fit-technique. Samples are output from MCMC targeting eq. (5.4.9) or equivalently $\eta = 1$ in eq. (5.5.4).

the Atlas location estimates to guide our choice) as we expect different behaviour in different sectors due to the uneven distribution of anchors (see fig. 5.1).

We observe some issues with convergence and mixing of some chains for the floating locations even in this small-data setting. For example, chains for coordinates of profiles 699 and 755 (last two rows from top in fig. 5.8 and bottom right corner of the grid in fig. 5.7) move very slowly between different modes. The problem gets worse if we increase the number of floating profiles or the number of items. We ran an experiment using the medium dataset, including all floating profiles but still only 8 items. We display MCMC traces in fig. 5.16. These mixing problems with the floating locations are related to the issues reported by Haines (2016), which may be caused by misspecified feedback between floating profile locations $X_{\bar{A}}$ and linguistic fields Φ. Although the MCMC might of course be improved, the VMP offers other advantages besides stable estimation. Separate runs at different η values are not required as the VMP parameterises a smooth sequence of SMI-posteriors over $\eta \in [0, 1]$ in a single end-to-end training scheme.

5.7.2 Variational SMI on the Small dataset.

We now expand the analysis to consider the Variational Meta-Posterior on the small dataset. The purpose of this section is show that the Variational Meta-Posterior we fit shows reasonable agreement to the MCMC ground truth for the SMI-posterior at each η-value.

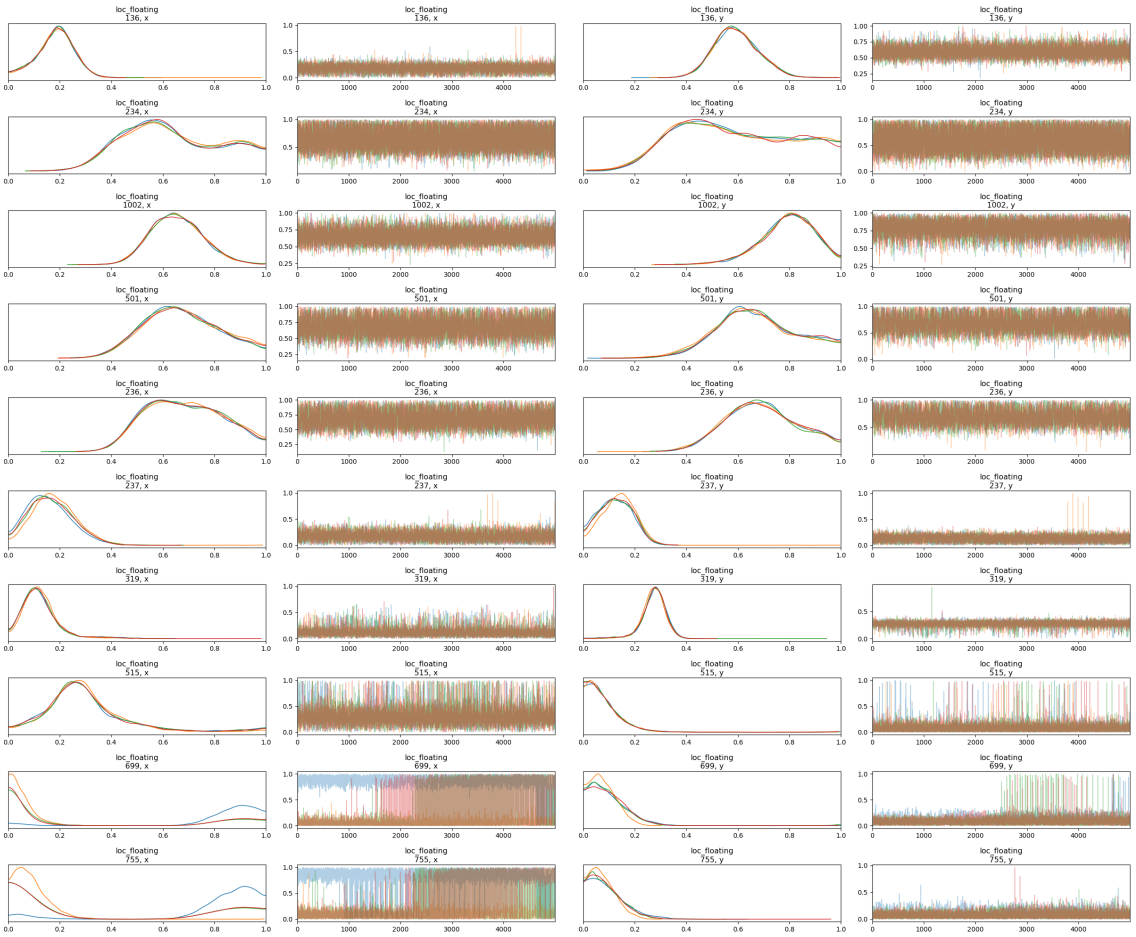


Figure 5.8: Distributions and trace plots for a subset of floating locations. We compare four MCMC chains for the conventional Bayesian posterior ($\eta = 1$) on the small dataset.

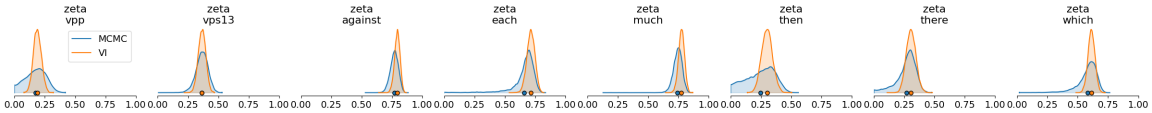
We observe good agreement for global model parameters. Comparisons of the SMP for ζ and μ determined by the VMP at η-values $0, 0.25, 0.75$ and 1 are shown in fig. 5.9 and fig. 5.10 respectively. The VMP concentrates on similar ranges for the two methods, across items and along degrees of influence η. The variational SMP is narrower, as one would expect from variational approximations. Our approximation based on Neural Spline Flow alleviates part of the underestimation of variance that one would observe with less expressive families such as a Mean Field parametrisation. The distributions of these global parameters are insensitive to η, showing that they get relatively little information from the floating profiles.

The approximation to the posterior locations of floating profiles is also good overall. In fig. 5.11 we overlay Highest Posterior Density (HPD) credible regions (with levels $0.05, 0.5, 0.95$) of the variational posterior for the locations of ten representative floating profiles at η-values $1, 0.75, 0.25$ and 0. The top grid of plots in fig. 5.11(a) corresponds to the Bayes posterior ($\eta = 1$) and are identical to those in fig. 5.7. The bottom grid of plots fig. 5.11(d) corresponds to the Cut Posterior ($\eta = 0$), which completely eliminates feedback from the floating profiles. Other values of η interpolate between the Cut and Bayes posterior, allowing partial feedback from the floating profiles. The main posterior modes are recovered, with no evidence for under-dispersion, and we find agreement even in complex multi-modal structures (for example profile 755, far right in the bottom row of grids a-d). However, there are also regions of disagreement, (for example profile 699, second from right in the bottom row of grids a-d, where a mode is missed) including regions frequently visited by the MCMC chain where the variational posterior do not assign high probability.

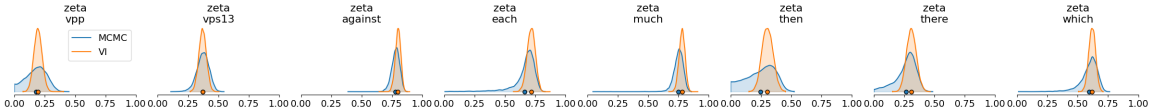
5.7.2.1 Variational Inference on the full dataset

We now consider the analysis of the full data. This is done in two stages. First we fit the validation dataset, visualise the results (in fig. 5.12) for prediction of held-out anchors, and then select the optimal influence parameter η^* to minimise the variational-posterior MSE (in fig. 5.13). We then refit the VMP to the full dataset using the utility in eq. (5.6.3)) and report selected posterior distributions for the locations of floats, predicted now using all the anchors and the optimal level of influence (in fig. 5.14).

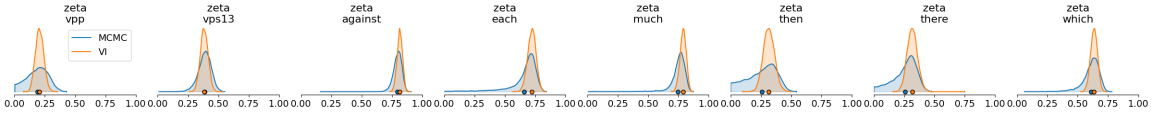
The comparison of our variational approximations in section 5.7.2 with MCMC "ground truth" gave good results in general for the small dataset. The SMP distributions for global parameters and many profile locations were well represented, and the distributions of most profiles were accurately captured. It is important to bear in



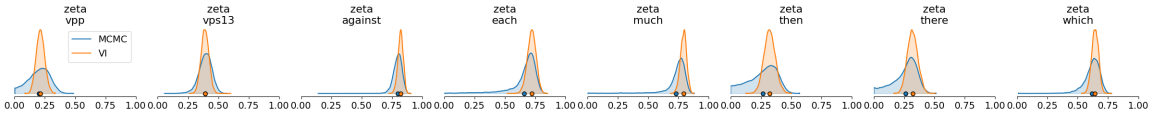
(a) $\eta = 1.0$



(b) $\eta = 0.75$

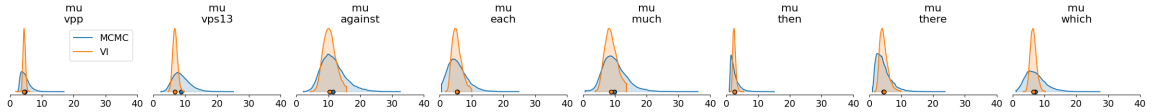


(c) $\eta = 0.25$

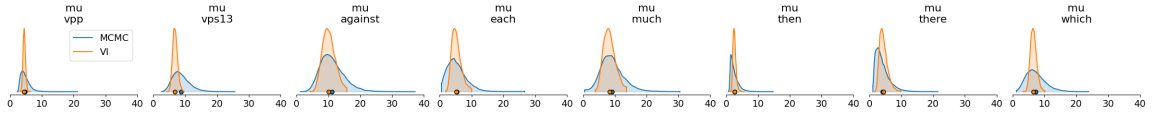


(d) $\eta = 0.0$

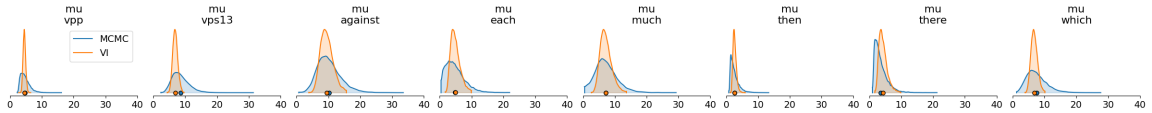
Figure 5.9: Comparison of the SMP distributions for ζ on the small dataset across different values of η. We compare the MCMC samples (blue) with the Variational posterior approximation (orange).



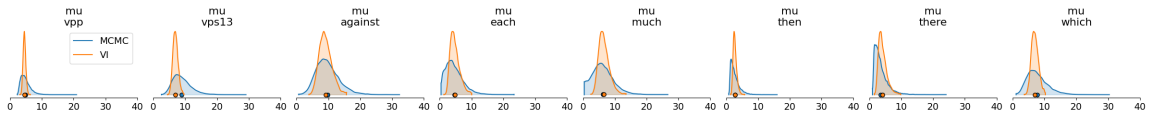
(a) $\eta = 1.0$



(b) $\eta = 0.75$



(c) $\eta = 0.25$



(d) $\eta = 0.0$

Figure 5.10: Comparison of the SMP distributions for μ on the small dataset across different values of η. We compare the MCMC samples (blue) with the Variational posterior approximation (orange).

mind that in the setting with a large number of floating profiles, multiple disconnected modes may be present in the posterior of imputed locations, which complicates inference for both MCMC and VI. However, the VMP location distributions are in pleasing agreement with the Atlas float-locations assigned by the fit-technique (association of red crosses and contours in fig. 5.11) and this suggests the variational approximation may give useful location predictions. We cant use Atlas locations for floats for inference purposes. Fortunately, as discussed in section 5.5.1, we can use held-out anchor profiles to quantify the accuracy of the SMP location distributions estimated using the VMP.

In fig. 5.12 we plot (in the same format as fig. 5.11) the marginal distributions $q^{(val)}_{\hat{\lambda}_{1:3},\eta}(x_p),\ p \in \mathcal{K}$ (see definition the end of section 5.6.3) of fifteen of the $K = 20$ held-out anchors for η-values $1, 0.75, 0.25$ and 0. These η-values are chosen for illustration as the estimated loss $\hat{V}^{(val)}(\eta)$ has a minimum at around $\eta = 0.7$. We draw attention to three features of this figure. First, the locations of the held-out anchors are clearly well-predicted by the marginal posteriors at each η. Second, the distributions tend to become more diffuse as we move down through the grids from (a) to (d) becoming really quite diffuse in the cut model in (d). This shows that the information in the floating profiles is informing the fields Φ at larger η values where some modulated feedback is allowed. Third, the distributions in fig. 5.12 are more uni-modal in appearance than those those in fig. 5.11. We believe this shows the usual concentration of the posterior with increasing data (here both I, the number of items and M the number of floating profiles have increased). This means it is likely that the variational approximations are more accurate than those in section 5.7.2 as the target is more easily matched.

We now present results for the choice of η. The loss $V^{(val)}$ is estimated by sampling $q^{(val)}_{\hat{\lambda}_{1:3},\eta}(x_p),\ p \in \mathcal{K}$ and averaging. The η-dependence of $\hat{V}^{(val)}(\eta)$ is plotted in fig. 5.13 at left. It has a minimum at around $\eta \approx 0.7$. Given the generally uni-modal shape of the marginal posterior location distributions in fig. 5.12, the posterior mean is a reasonable estimator. At right in fig. 5.13 we plot

$$K^{-1} \sum_{p \in \mathcal{K}} \| \bar{x}_p - X^*_p \| \tag{5.7.1}$$

with $\bar{x}_p = \hat{\mathbb{E}}_{q^{(val)}_{\hat{\lambda}_{1:3},\eta}}[x_p]$. This measures the distance between point prediction and truth, as it removes the effect of posterior dispersion present in $V^{(val)}$ but remains sensitive to all the biasing effects of model misspecification. The resulting estimated average misplacement distance for location prediction of the held out anchors is around 28

kilometres and we take this as our estimate of the scale of the error for our prediction of locations for floating profiles in fig. 5.14 below.

In the last stage of this analysis, we refit the VMP to the posterior in eq. (5.5.4) for the full dataset and plot marginal posterior distributions for the locations of selected profiles along with the atlas locations in fig. 5.14. This is the main "output" result of this paper. The optimal influence η^* lays somewhere between grids (b) and (c). The distributions (in blue) are somewhat more concentrated than those for the held-out anchors in fig. 5.12. This may reflect differences in the text-type. Anchors are generally legal documents with locations indicated in the text. Unlocated texts corresponding to floating profiles are more hetrogeneous. Also, the disagreement between posterior distributions for float locations and Atlas fit-technique locations is somewhat larger than that seen for held out anchors. However, the fit-technique is itself subject to uncertainty so this is to be expected.

The VMP-means for the basis GP fields Γ offer some insight on the real underlying structure of the form-probability fields Φ. These are plotted in fig. 5.15 at the same selection of η-values. The scales given by the shading-bar to the right of each graph are important here. For example, in the bottom row of plots, corresponding to the cut at $\eta = 0$, the scales for graphs 1,3,4,5,9 (reading from left) show that the fields show small or negligible variation across the region $\mathcal{X}$. These basis fields could be dropped from the analysis with relatively little loss in the quality of the fit. As we move down a column we see the basis field evolves, though retains its basic form. The basis fields have a label switching symmetry in the posterior, but VMP is helpfully breaking that symmetry and identifying fields across η-values: the same set of $\hat{\lambda}_{1:3}$-values are used in each posterior. Fields six and ten capture variations which are aligned roughly north-south and east-west respectively. We conjecture that this reflects the imprint of the population history of speakers in this region on the spatial patterning of dialect.

5.8 Conclusions

5.8.1 Main results

We have shown how to simultaneously estimate spatial fields and measurement locations using a set of anchored measurements. We selected an inference framework (from Cut models and SMPs and the Bayesian posterior) which makes some allowance for misspecification. This methodology is needed due to the high-dimension of the missing data (the fields) and the difficulty of forming an accurate generative model

for the data. The selected value of the influence parameter is less than one. Referring to fig. 5.12, this makes sense: the posterior distributions clearly cover the true held-out anchor locations more reliably at η values smaller than one. This confirms that some down-weighting of the information from floating profiles is desirable.

Was it worth it? Does information from the floating profiles actually help? One could simply infer the posterior distribution of the spatial fields from the anchors and use this to predict float locations. This would be cut-model inference and corresponds to the bottom grid in fig. 5.12 with $\eta = 0$. Whilst these marginal posterior distributions do cover the true values well, they are clearly much more diffuse than the distributions in the grid at larger η, as they throw out all information from the floats. This shows that allowing some feedback is useful.

One nice feature of the inference framework here is the possibility of using the location error itself as a loss for selection of the influence parameter η. Previous work used the elpd as a guide. This reflects that fact that the location parameters $x_p,\ p \in \bar{A}$ are "physical" quantities, with a clear meaning independent of the model, misspecified or otherwise. In fact a loss of this sort will be available in many problems that have the form of supervised classification: if x_p is the "class label" for observation p and y_p is a feature vector then x_p are known in the training data $p \in A$ and unknown in the unlabelled data $p \in \bar{A}$ intended for classification. In our case the class labels are continuous, but the physical meaning for the labels, and the possibility for constructing a loss minimising the predicted classification error, remains.

Our final results in fig. 5.14 show our predictions for the locations of floating profiles. These profiles were selected to cover the region, and are representative of typical performance. There is good agreement between our predicted locations and those established using the fit technique. We provide a formal quantification of uncertainty missing in that non-statistical method. The application we have given was selected to demonstrate the effectiveness of our method and agreement with the Atlas. Our methods could be used to predict the location of new profiles. From the perspective of users this would be the main application of interest. In principle this should be done by adding the new profile to the set of floats and repeating the simulation in section 5.7.2.1. However, a single profile is unlikely to change the distribution to any great degree, so treating the existing analysis as the first stage of a cut model for Bayesian multiple imputation of ψ, and considering the posterior distribution of the new profile location conditioned on its y_p-values and the imputed (random) fields, should lead to efficiency gains with little loss of accuracy.

What really was the misspecification? The generative model given in this paper is a significant improvement on that in Haines (2016), we believe because of the reduction in field dimension due to working with basis-fields, and the natural way this introduces correlation between fields. However, we ignored correlation between the forms of an item selected by a scribe. They are unlikely to use multinomial sampling of forms. Also, where a scribe is copying an extant text, they may sometimes copy the spelling in that text. This may bring forms from far afield into the profile. Finally, the locations will not be missing at random from the LPs as the anchor texts tend to be legal documents while the floating texts are more heterogeneous. There are no doubt other processes which we do not understand. The use of an inference framework that accommodates these errors, rather than further model-elaboration, allowed us to make progress without ultimately resolving these issues.

5.8.2 Future work

Our Semi-Modular analysis reduces the influence of all floating profiles homogeneously by using a single commom influence parameter. Nevertheless, there are other possible ways to *modularize* our probabilistic model for the LALME data. Two extensions may be of interest for future exploration. First, we could reduce the influence of individual profiles differently, replacing $\eta \to \eta_p,\ p \in \bar{A}$. This would be equivalent to introducing *multiple cuts* in the model, and in practice could be achieved by imposing different powers η_p on the likelihood of each profile. This would require an elpd-style training loss for η selection. Second, we could reduce the influence of some items. The geographical structure may be more marked for some items, they may be more informative and associated observations more easily modelled. Some words are clearly differentiated by region, while others may be used equally across regions. We may find we can improve our location accuracy by reducing the influence of unstructured or misspecified items. This case would also entail multiple cuts and would be implemented by introducing a set of influence-parameters $\eta_i,\ i \in \mathcal{I}$ to be applied as per-factor powers in the likelihood product over items. Carmona and Nicholls (2022) shows how a loss on mutiple cuts may be optimised in a VMP setting using SGD.

5.9 Supplementary Material

Further background on the data

5.9.1 LALME

The Linguistic Atlas of Late Mediaeval English (McIntosh et al., 1986; Benskin et al., 2013) is a survey of local variation in the written language for the period c. 1350 to c. 1450. Its results are presented as separate maps for each of 62 selected items. Here an "item" is a word or other linguistic category, such as a grammatical inflexion or an etymological segment, here marked by {}; the variant realisations of an item are its "forms", marked by ⟨⟩. An item map displays all of the variant forms for that item as recorded at the irregular grid of survey points shown as red and blue dots in fig. 5.1. Additionally, and for the same survey points, twelve hundred dot maps report the distributions of select variants or sets of variants for another 218 selected items, as well as analyses of such distributions within the item maps.

The survey includes most of England and such few parts of Wales from which English writing is known, and provides token coverage of southern Scotland. Scotland is now treated comprehensively by Williamson (2008). Ideally the period of survey would be much less, to guard against interference from variation over time with variation over space, but the accidents of manuscript production and survival, and the displacement of local usage by what was becoming a national written standard, would leave large areas unrepresented if this time-span were much reduced (McIntosh, 1963; McIntosh et al., 1986). Even so, and in spite of the 1511 sources, coverage for England remains geographically uneven; for parts of the north and the south-west it is very sparse, whereas for East Anglia, the London area, and parts of the West Midlands, it leaves little room for additional material.

5.9.2 Survey points and the fit-technique

The Leeds Survey of Modern English Dialects (Orton and Dieth, 1962) could choose its three hundred survey points as more or less evenly spaced across the land, but a survey of the mediaeval language has to make the best of what it can find. The Atlas's 'informants' were its sample manuscripts, represented by 1511 linguistic profiles. An additional challenge is that the local origins for the bulk of these informants were at the outset unknown: the Atlas had to be constructed as an ever tightening taxonomic matrix within which unlocalised informants could be placed. Since regional dialect changes for the most part in an orderly way over space (that is, the regional

distributions of dialectal features are for the most part cohesive), it follows that a taxonomic placing within a geographically rooted irregular grid is an approximation to a local origin, and the denser the grid, the more closely will that origin be defined; informants so placed can be incorporated as additional survey points, and contribute in their turn to the placing of others. This procedure – McIntosh's 'fit'-technique (Benskin, 1991; McIntosh et al., 1986; Benskin et al., 2013) – defines relative placings within a grid variously anchored by informants for which there is good non-linguistic evidence of local origin. In rare cases, a scribe can be identified and his biography is known, but most anchors are legal or administrative documents whose place of origin is explicitly stated, and which (though it is not a foregone conclusion) can be trusted at least provisionally to represent the local forms of language (Benskin, 2004). In one or two cases, anchors found after the Atlas was published have confirmed placings previously determined by 'fit'.

The fit-technique is mechanical and cartographic, proceeding from a set of maps each of which displays the geographical distributions of the variant forms for a single item. A local dialect can be regarded as an assemblage of such forms, some few of which may be confined to that assemblage, but most if not all of which are shared with one or another of the neighbouring dialects; a local dialect is thus far an intersection of the distributions of its constituent forms. The local usage in any of the Atlas inhabited blank spaces is hence assumed to have been a sub-set of the aggregate of the forms in the survey points that delimit that space, with few of its forms (if any) that are not predictable from those delimiters. Conversely, when such a combination is identified, it is assumed, in default of evidence to the contrary, to represent the local usage proper to that space (McIntosh et al., 1986). The fit technique treats dialectal distributions in terms of areas and overlaps, and proceeds by the cumulative elimination, map by map, of all the areas which are documented but in which the constituent features of the unlocalised dialect are not recorded. What remains is the area which accommodates all its features as an assemblage; in the first instance it is a taxonomic sector, and by implication the dialect's area of origin. The present paper considers a statistical alternative to this procedure.

5.9.3 Linguistic profiles

The core of the Atlas comprises some 1511 taxonomic inventories, whose contents are the spellings elicited by a fixed-format survey questionnaire; a completed survey questionnaire is a linguistic profile ('LP') for an individual scribe. In all the forms of 623 distinct items were recorded. The items were unevenly surveyed over space. Details of

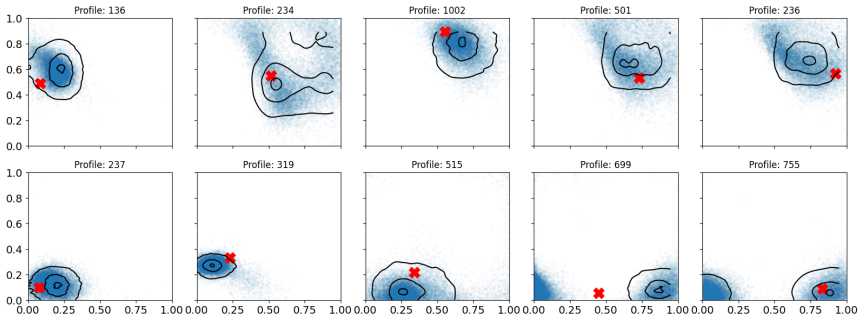
how the list of recorded items developed are recorded in the Atlas. It should perhaps be emphasised that the survey is not concerned with peculiar local vocabulary. The choice of an item was governed by two considerations: (a) it must be represented by at least two co-variant forms of philological interest, and (b) it must be in common use, with a good chance of appearing in whatever text might be available. Most items are hence commonplace vocabulary, like {THESE}, {SHE}, {WHICH}, {SHALL}, {IF}, {TOGETHER}, and {TWO}. Grammatical inflexions are almost by definition of common occurrence; among those used in Atlas are the suffixes of the {present participle} (so ⟨∼ING⟩, ⟨∼YnG⟩, ⟨∼AND⟩, ⟨∼END⟩, etc.) and the {noun plural} (so ⟨∼ES⟩, ⟨∼IS⟩, ⟨∼US⟩, etc.). The Atlas is in the first instance a survey of written language: much of the variation it records corresponds to variation in the spoken language, but some of it, like that between ⟨∼YnG⟩and ⟨∼ING⟩, demonstrably does not. The idea that any given word in English should have one and only one spelling, even if its pronunciation was a constant, was alien to perhaps the majority of scribes; spelling systems had their own generative logic, which could be developed in different ways. Although some scribes were very consistent, it is not unusual to find in the course of a long text that many words, perhaps even a majority, have several forms in either free or conditioned variation. Since the Atlas was intended as a research tool, all variants elicited by the questionnaire, regardless of their presumed significance for the spoken language, are separately recorded; transcribed from manuscript, they are represented in raw state so far as conversion into ASCII characters allows. The contents of the completed survey questionnaires are correspondingly prolific and diverse.

5.9.4 Coarsening

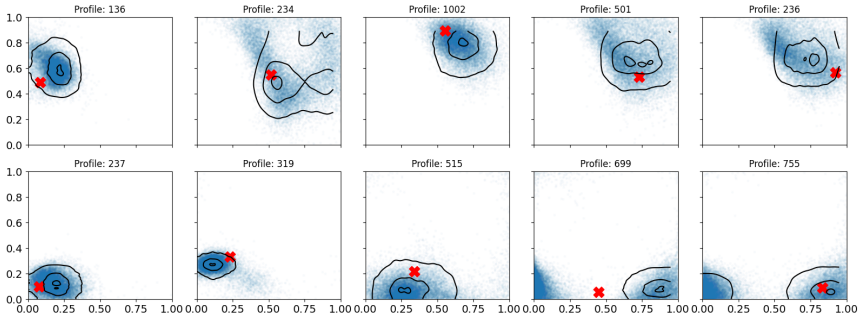
The number of forms corresponding to a single item can be very large, commonly in the range of 100–200; notorious is {THROUGH}, for which the Atlas records over four hundred variants. After analysing this primary data, Haines (2016) adopted a coarsened version, in which the forms of an item under consideration were classified and treated as subsets. For experiments with multi-dimensional scaling as an alternative to the 'fit'-technique (Benskin, 1988), a segmentation dictionary had been developed in course of making the Atlas. For each item, a template consisting of vowel and consonant segments was devised from which all of its variants could be generated. For example, the template for {THEM} is C1+V1+C2+V2+C3. These can be seen as slots into which particles can be dropped. In {THEM}, C1 can be null (in the forms ⟨AM⟩ and ⟨EM⟩) or H (so ⟨HAM⟩, ⟨HEM⟩, etc.) or TH (so ⟨THEM⟩, ⟨THAIM⟩,

etc.) or the mediaeval 'thorn' (so ⟨yEM⟩, ⟨yEIM⟩, etc.) or mediaeval 'yogh' (for which ⟨zAM⟩, ⟨zEOM⟩, etc.); V1 can be A, E, EE, EO, I, Y, O, OE, U, UE, EI, EY, AI, AIE, AY; C2 is usually M, but often abbreviated (so ⟨HEm⟩, ⟨THAm⟩, etc.), rarely MM (for example ⟨HEMME⟩) or mM (for example ⟨yAmME⟩), but can be S (seen in ⟨AS⟩, ⟨ES⟩, ⟨HIS⟩, etc.) or N (as ⟨HAN⟩) or NN (in ⟨yENN⟩); V2 is usually null, but may be E (as in ⟨THAIME⟩); C3 is usually null, but may be M (in ⟨HEMEM⟩) or N (in ⟨HYMEN⟩). These do not exhaust the possibilities for the 116 different forms (letters may be written superscript, as for example in ⟨HA$^{\mathrm{A}}$M⟩and ⟨Y$^{\mathrm{EM}}$⟩), but are enough to demonstrate the principle. Via such a template, comparison of LPs in respect of them can be confined to a single segment, and the number of variants reduced to easily managed proportions by identifying different particles within a slot. Thus for C1, a binary contrast of H versus TH and its mediaeval equivalent 'thorn' includes most by far of the 116 variants and so removing the distinction between H and TH in C1 of {THEM} is a coarsening; the contrast is implicitly phonological, as is that between M (including abbreviated m) and S for C2. Philological judgement, by no means infallible, is called for at every stage, but such restrictions relieve the input data of what is likely to be a great deal of noise.

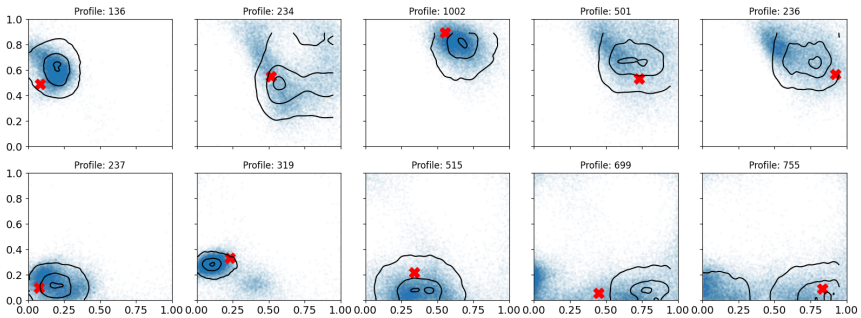
For present purposes, the area investigated was also confined. We focus on a rectangular area spanning 200km east-west and 180km north-south, as displayed in fig. 5.1. This contains 120 anchor Linguistic Profiles (LPs) and a further 247 floating LPs that have been assigned locations via the fit-technique.



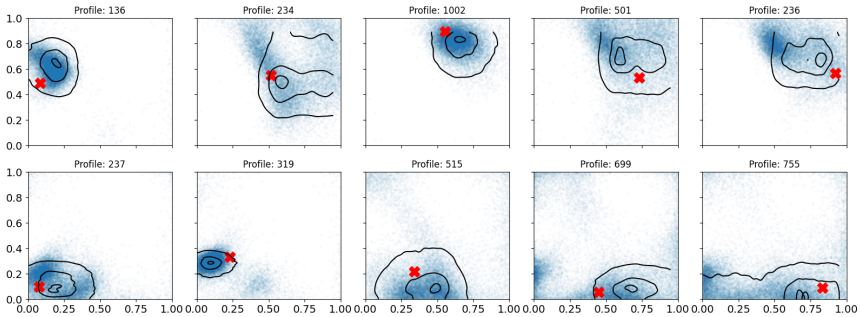
(a) $\eta = 1.0$



(b) $\eta = 0.75$

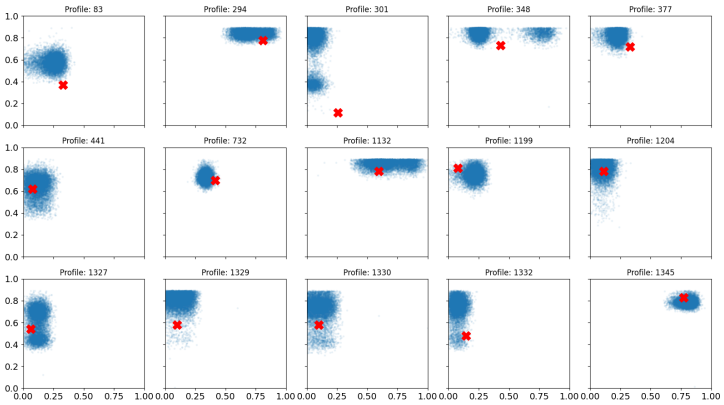


(c) $\eta = 0.25$

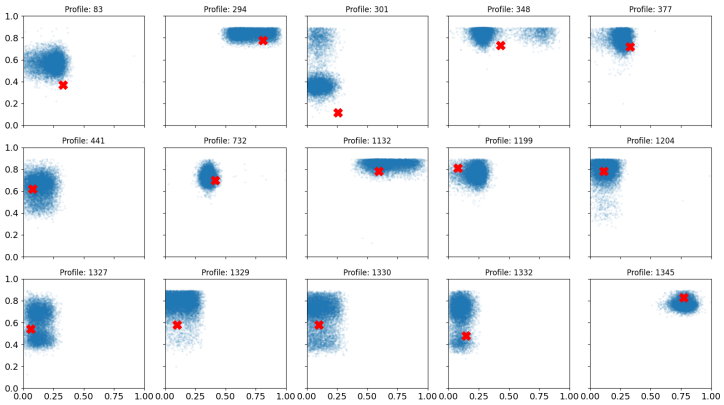


(d) $\eta = 0.0$

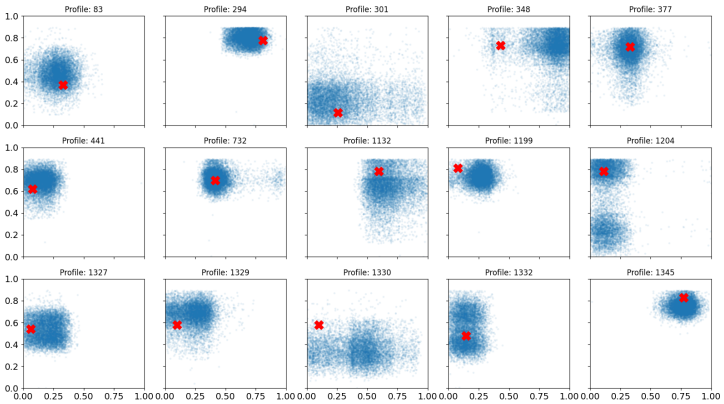
Figure 5.11: Posterior distributions of ten floating locations obtained using SMI with varying values of η. Small Dataset. Red cross denotes the floating-profile location assigned by the "fit-technique". We compare the MCMC samples (blue scatter plot) with the Variational Meta-Posterior approximation (level sets at levels $0.05, 0.5, 0.95$).



(a) $\eta = 1.0$



(b) $\eta = 0.75$



(c) $\eta = 0.0$

Figure 5.12: Posterior distributions of imputed locations for fifteen (random) held-out anchor profiles obtained using the Variational Meta-Posterior on the entire LALME dataset. Darker regions denote higher posterior probabilities, and red crosses mark true locations of profiles.

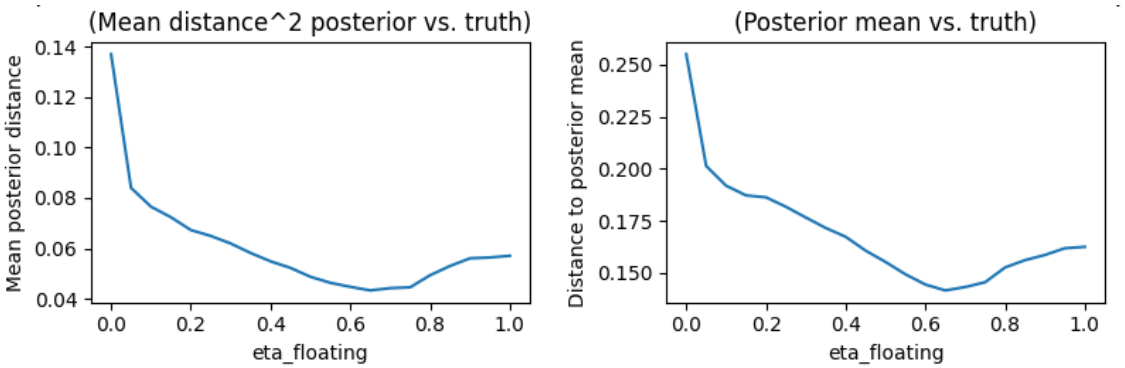
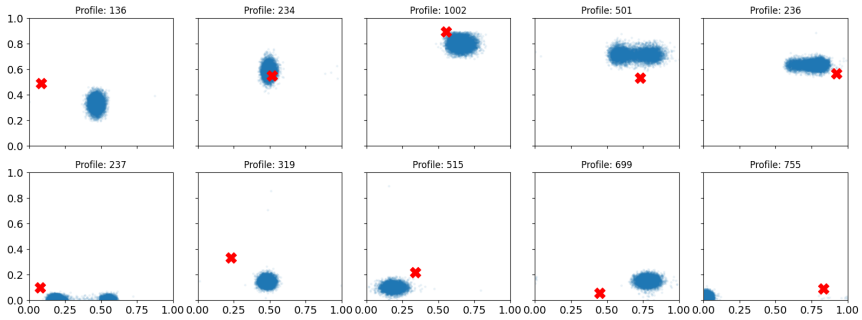
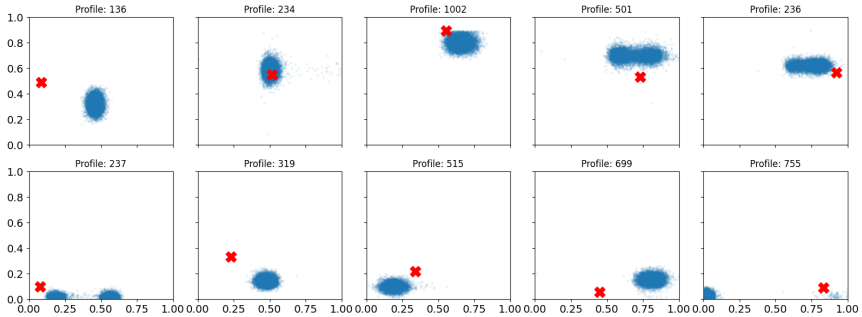


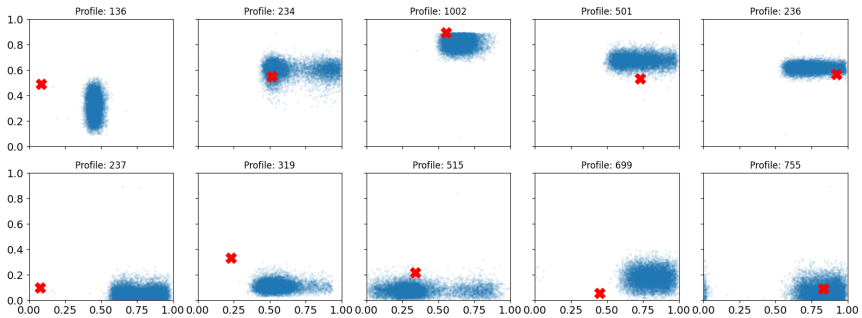
Figure 5.13: Error in location reconstruction for held-out profiles as a function of the influence parameter. Left: Variational-posterior MSE to true location, in eq. (5.6.5). Right: Euclidean distance between the true location and the posterior location mean, in eq. (5.7.1). In both cases, we average over twenty random held-out profiles, including those shown in fig. 5.12. The best Semi-Modular Posterior is obtained with $\eta \approx 0.75$ producing an average held-out error of around 30km on point estimates.



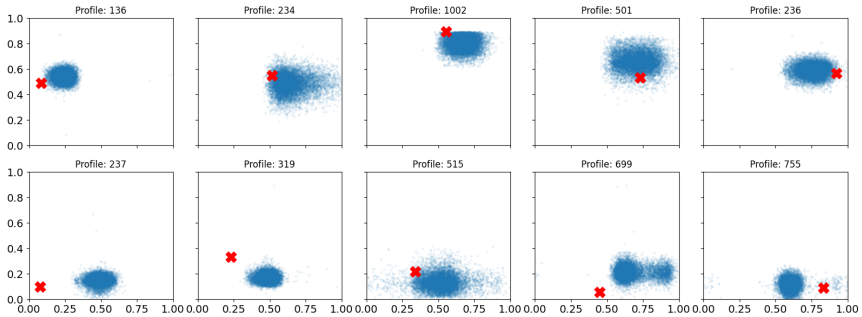
(a) $\eta = 1.0$



(b) $\eta = 0.75$

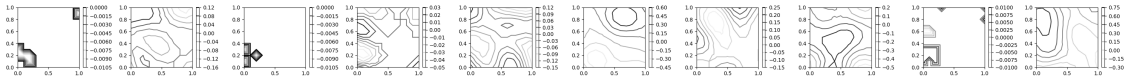


(c) $\eta = 0.25$

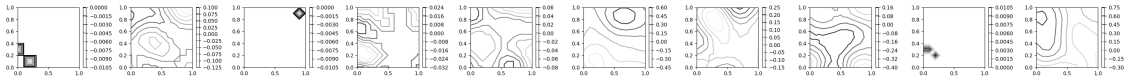


(d) $\eta = 0.0$

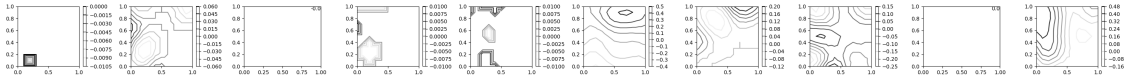
Figure 5.14: Posterior distributions of locations of ten randomly chosen floating LPs, obtained using Variational Meta-Posterior with varying values of η. Darker regions denote higher probabilities on each plot, and red crosses mark location assigned by the "fit-technique".



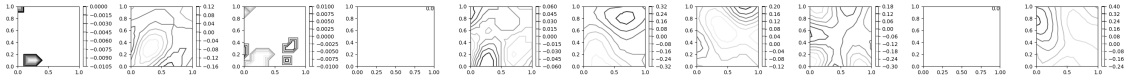
(a) $\eta = 1.0$



(b) $\eta = 0.75$



(c) $\eta = 0.25$



(d) $\eta = 0.0$

Figure 5.15: Visualization of the posterior basis GP fields Γ, represented by the level plots of the Semi-Modular Posterior mean. Darker lines denote higher levels within each plot. We set the number of basis fields equal to 10, but observe from the scale bars that the posterior concentrates on a smaller number of *effective* basis fields, leaving a few flat fields (eg. first field in every row).

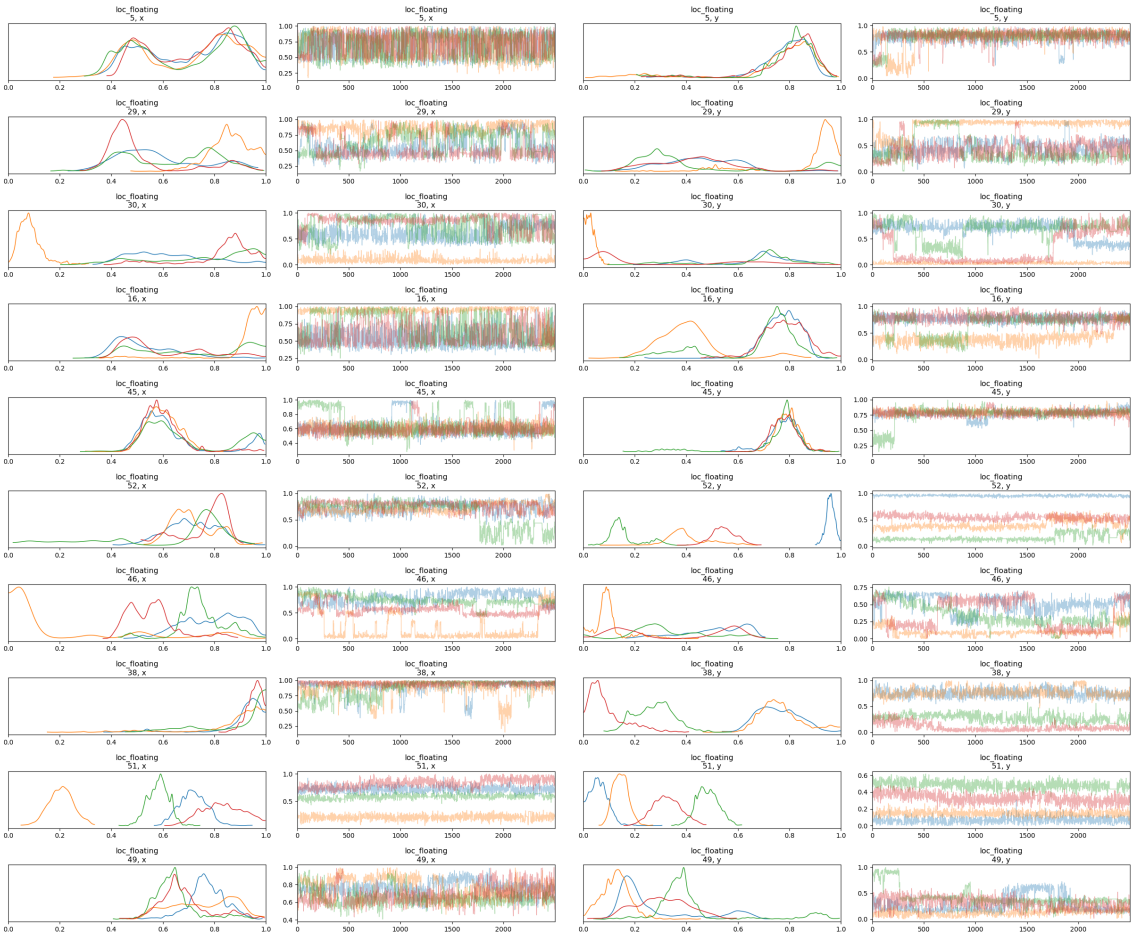


Figure 5.16: Distributions and trace plots for a subset of floating locations. We compare four MCMC chains for the conventional Bayesian posterior ($\eta = 1$) on the medium dataset including all (247) floating profiles and 8 items.

Chapter 6

Conclusions

In this thesis we presented multiple contributions to the field of Generalized Bayesian inference, with a particular focus on evidence combination and the problem of model misspecification.

The proposed framework known as Semi-Modular Inference has been well received by the statistical community and have opened new frontiers for theory and practice involving data integration. A number of extensions and related studies have been published since the original article (Carmona and Nicholls, 2020). For example, alternative versions of the Semi-Modular Posterior are studied in Nicholls et al. (2022); Frazier and Nott (2022), novel sampling algorithms for SMI known as Posterior Bootstrap for Modular Inference (PBMI) are proposed by Pompe (2021), procedures for likelihood free inference using SMI are given in Chakraborty et al. (2023), and an interpretation of SMI as a Geographically Weighted Regression (GWR) is discussed in Liu and Goudie (2021).

We close this dissertation by summarising the main contributions discussing potential future directions.

6.1 Summary of contributions

In Chapter 2, we introduced the Semi-Modular Inference framework, a generalisation of Bayesian inference that brings together Bayesian Inference and two qualitatively different inference schemes, power-posteriors and Cut models, used to treat model misspecification. We show that the proposed scheme is an order-coherent update of beliefs in the sense of Bissiri et al. (2016) and Nicholls et al. (2022). We give an algorithm to sample from the Semi-Modular Posterior and suggest criteria for choosing the influence parameter η according to an external utility function. Open-source code and data are provided to reproduce our experimental results.

In Chapter 3, we gave variational families and algorithms for approximating the Cut- and SMI-posterior distributions. The proposed methods can be adapted to a wide range of variational families, including Normalizing Flow to overcome known limitations of variational inference such as under dispersion of the approximated posterior. We showed that minimisation of the standard Kullback-Leibler Divergence does not allow control of information flow between modules and propose a loss function tailored to modularized inference. We introduced Variational Meta-Posteriors which is a new object capable of approximation an entire family of posterior distributions using a single set of trainable parameters. Advantages of the VMP for the efficient exploration of modularized distributions were discussed and illustrated with two examples.

In Chapter 4, we applied Semi-Modular Inference on an archaeological application to unveil agricultural practices in ancient civilisations. We provided three alternative analyses of the multi-modular dataset, obtaining consistent evidence for the hypothesis of extensification in urban settlements of Mesopotamia and the Aegean. The analyses extended work by Styring et al. (2017) and provide improved estimations with SMI, in particular, reducing the dilution effect observed for the Cut posterior.

In Chapter 5, we developed a new statistical framework for point-referenced data that enables simultaneous reconstruction of latent spatial fields and imputation of missing sample locations. We presented a hierarchical model tailored to the motivating linguistic dataset, but applicable to other applications where geographical variations are assumed. Point estimates and uncertainty quantification for the missing documents were provided, matching estimations produced by subject-matter experts. This application provided a substantial example where variational methods for Semi-Modular Inference solves an unreachable problem for standard MCMC methods.

For all our methodological contributions, we provided open-source implementations, along with public data to facilitate reproducibility of our experiments.

6.2 Future work

The field of Modularized Bayesian Inference is an active area of research, with potential for further developments and extensions to be explored in future studies. In this context, we propose several research directions and open questions that warrant further investigation.

- Facilitate adoption of Semi-Modular Inference for practitioners by implementing friendly interfaces based on widely-used frameworks such as Stan, Pyro,

TensorFlow Probability or BlackJax. Perhaps resembling the original interface of the `cut` function from OpenBUGS.

- Further studies on alternative versions of the Semi-Modular Posterior and their properties. For example, those discussed by Nicholls et al. (2022) and Chakraborty et al. (2023).

- Further analysis of multiple cuts. In section 3.5.2 we provide one example where SMI accommodates for multiple cuts, but further research is needed to extend the framework to other scenarios, as it is discussed in Liu (2022).

- Online learning for the Semi-Modular Posterior and dynamic choice of the rate of influence.

- Use of the Variational Meta-Posterior beyond Semi-Modular Inference. For example, one could perform sensitivity analyses to the choice of prior hyper-parameters.

- Better understanding of the training process for the Variational Meta-Posterior. For example, is there a "best choice" for the lattice $\rho(\eta)$ mentioned in section 3.4.2? Instead of a fixed lattice, perhaps an evolving algorithm that trading off exploration and exploitation during training time.

- Methods for assessing convergence of the Variational Meta-Posterior.

- Further studies of the LALME dataset. For example, incorporating data outside the rectangular region used for our analysis.

Bibliography

Aglietti, V., Bonilla, E. V., Damoulas, T., and Cripps, S. (2019). Structured Variational Inference in Continuous Cox Process Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Agrawal, A., Sheldon, D., and Domke, J. (2020). Advances in black-box VI: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 2020-Decem. arXiv: 2006.10343 ISBN: 2006.10343v1.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, USA. Series Title: Wiley Series in Probability and Statistics ISSN: 0040-1706.

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1):1.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic, H., Mikulik, V., Norman, T., Quan, J., Papamakarios, G., Ring, R., Ruiz, F., Sanchez, A., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Wang, L., Stokowiec, W., and Viola, F. (2020). The DeepMind JAX Ecosystem.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, New York, 2 edition.

Benskin, M. (1988). The numerical classification of languages, and dialect maps for the past. In Reenen, P. v. and Reenen-Stein, K. v., editors, *Distributions spatiales et temporelles, constellations des manuscrits/Spatial and Temporal Distributions, Manuscript Constellations: Etudes de variation linguistiques offertes à Anthonij Dees à l'occasion de son 60ème anniversaire / Studies in language variation offered to*

Anthonij Dees on the occasion of his 60th birthday, page 13. John Benjamins Publishing Company. Publication Title: Distributions spatiales et temporelles, constellations des manuscrits/Spatial and Temporal Distributions, Manuscript Constellations: Etudes de variation linguistiques offertes à Anthonij Dees à l'occasion de son 60ème anniversaire / Studies in language variation offered to Anthonij Dees on the occasion of his 60th birthday.

Benskin, M. (1991). The "fit" technique explained. In Riddy, F., editor, *Regionalism in Late Medieval Manuscripts and Texts: Essays Celebrating the Publication of a Linguistic Atlas of Late Mediaeval English*, pages 9– 26. Cambridge : D.S. Brewer.

Benskin, M. (2004). Chancery Standard. In Kay, C., Hough, C., and Wotherspoon, I., editors, *New Perspectives on English Historical Linguistics Volume II: Lexis and Transmission*, Current Issues in Linguistic Theory, pages 21–26. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Benskin, M., Laing, M., Karaiskos, V., and Williamson, K. (2013). An Electronic Version of A Linguistic Atlas of Late Mediaeval. url: https://www.lel.ed.ac.uk/ihd/elalme/elalme.html.

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA. Series Title: Wiley Series in Probability and Statistics ISSN: 0957-0233.

Besbeas, P., Freeman, S. N., Morgan, B. J. T., and Catchpole, E. A. (2002). Integrating Mark–Recapture–Recovery and Census Data to Estimate Animal Abundance and Demographic Parameters. *Biometrics*, 58(3):540–547.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Blangiardo, M., Hansell, A., and Richardson, S. (2011). A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. *Atmospheric Environment*, 45(2):379–386.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877. Publisher: American Statistical Association.

Bonilla, E. V., Krauth, K., and Dezfouli, A. (2019). Generic inference in latent Gaussian process models. *Journal of Machine Learning Research*, 20. arXiv: 1609.00577.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. (2018). JAX: composable transformations of Python+NumPy programs.

Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/10618600.1998.10474787.

Brooks, S. P., King, R., and Morgan, B. J. (2004). A Bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation*, 27(1):515–529.

Carmona, C., Nieto-Barajas, L., and Canale, A. (2019). Model-based approach for household clustering with mixed scale variables. *Advances in Data Analysis and Classification*, 13(2):559–583. arXiv: 1612.00083.

Carmona, C. U., Haines, R. A., Anderson Loake, M., Benskin, M., and Nicholls, G. K. (2023). Simultaneous Reconstruction of Spatial Frequency Fields and Sample Locations via Bayesian Semi-Modular Inference.

Carmona, C. U. and Nicholls, G. K. (2020). Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components. In Silvia, C. and Calandra, R., editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, pages 4226–4235. PMLR. arXiv: 2003.06804.

Carmona, C. U. and Nicholls, G. K. (2022). Scalable Semi-Modular Inference with Variational Meta-Posteriors. arXiv: 2204.00296.

Chakraborty, A., Nott, D. J., Drovandi, C. C., Frazier, D. T., and Sisson, S. A. (2023). Modularized Bayesian analyses and cutting feedback in likelihood-free inference. *Statistics and Computing*, 33(1):33. arXiv:2203.09782 [stat].

Christensen, R. H. B. (2010). Ordinal—regression models for ordinal data.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Dhaka, A. K., Catalina, A., Andersen, M. R., Magnusson, M., Huggins, J. H., and Vehtari, A. (2020). Robust, Accurate Stochastic Optimization for Variational Inference. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Proceedings of the 34th Conference on Neural Information Processing Systems, NeurIPS 2020*, pages 10961–10973. Curran Associates, Inc. arXiv: 2009.00666.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. *Statistical Science*, 28(4). arXiv:1312.6536 [stat].

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. (2017). TensorFlow Distributions. arXiv:1711.10604 [cs, stat].

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using Real NVP. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*. arXiv: 1605.08803.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural Spline Flows. In *Proceedings of the 33rd Conference on Neural Information Processing Systems, NeurIPS 2019*. arXiv: 1906.04032.

Frazier, D. T. and Nott, D. J. (2022). Cutting feedback and modularized analyses in generalized Bayesian inference. arXiv: 2202.09968.

Friel, N. and Pettitt, A. N. (2008). Marginal Likelihood Estimation via Power Posteriors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(3):589–607. Publisher: [Royal Statistical Society, Wiley].

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2015). *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 3 edition.

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19:1–49. arXiv: 1709.02536.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2022). Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. *Bayesian Analysis*, -1(-1):1–67. Publisher: International Society for Bayesian Analysis.

Giorgi, E. and Diggle, P. J. (2015). On the inverse geostatistical problem of inference on missing locations. *Spatial Statistics*, 11:35–44.

Goudie, R. J. B., Presanis, A. M., Lunn, D., De Angelis, D., and Wernisch, L. (2019). Joining and splitting models with Markov melding. *Bayesian Analysis*, 14(1). arXiv:1607.06779 [stat].

Grünwald, P. (2012). The Safe Bayesian. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, volume 7568 LNAI, pages 169–183. Springer Berlin Heidelberg. ISSN: 03029743.

Grünwald, P. and Ommen, T. v. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069–1103. Publisher: International Society for Bayesian Analysis.

Haines, R. A. (2016). *Simultaneous Reconstruction of Spatial Frequency Fields and Field Sample Locations*. PhD thesis, University of Oxford.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer New York, New York, NY. Series Title: Springer Series in Statistics.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian Processes for Big Data. In *Proceedings of the 29th Conference, of Uncertainty in Artificial Intelligence, UAI 2013*, pages 282–290. arXiv: 1309.6835.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015*. arXiv: 1411.2005.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Huang, C. W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 5, pages 3309–3324. arXiv: 1804.00779.

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. arXiv: 1708.08719 Pages: 1-31.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600. arXiv: 1708.03625 Publisher: Wiley.

Jessee, S. A. (2010). Partisan Bias, Political Information and Spatial Voting in the 2008 Presidential Election. *The Journal of Politics*, 72(2):327–340.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. arXiv: 1312.6114.

Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392. arXiv: 1906.02691.

Knuiman, M. W., Divitini, M. L., Buzas, J. S., and Fitzgerald, P. E. (1998). Adjustment for Regression Dilution in Epidemiological Regression Analyses. *Annals of Epidemiology*, 8(1):56–63. ISBN: 1047-2797 (Print).

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1. arXiv: 1908.09257.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian Modeling for Multivariate Ordinal Data. *Journal of Computational and Graphical Statistics*, 14(3):610–625. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].

Lao, J. and Louf, R. (2020). Blackjax: A sampling library for JAX.

Lao, J., Suter, C., Langmore, I., Chimisov, C., Saxena, A., Sountsov, P., Moore, D., Saurous, R. A., Hoffman, M. D., and Dillon, J. V. (2020). tfp.mcmc: Modern Markov Chain Monte Carlo Tools Built for Modern Hardware. arXiv:2002.01184 [cs, stat].

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, NJ, USA. ISSN: 0004-3591.

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150. ISBN: 1936-0975.

Liu, Y. (2022). *Modularized Bayesian Inference: Methodology, Algorithm, Theory And Application.* Thesis, University of Cambridge. Accepted: 2022-11-08T14:10:29Z.

Liu, Y. and Goudie, R. J. B. (2021). Generalized Geographically Weighted Regression Model within a Modularized Bayesian Framework. arXiv: 2106.00996.

Liu, Y. and Goudie, R. J. B. (2022a). A General Framework for Cutting Feedback within Modularized Bayesian Inference. arXiv:2211.03274 [math, stat].

Liu, Y. and Goudie, R. J. B. (2022b). Stochastic approximation cut algorithm for inference in modularized Bayesian models. *Statistics and Computing*, 32(1):7. arXiv: 2006.01584.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational Inference for Gaussian Process Modulated Poisson Processes. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1814–1822. PMLR. ISSN: 1938-7228.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(1):19–38. ISBN: 1092800891.

Mao, H., Martin, R., and Reich, B. (2022). Valid model-free spatial prediction. arXiv:2006.15640 [math, stat].

Maucort-Boulch, D., Franceschi, S., and Plummer, M. (2008). International Correlation between Human Papillomavirus Prevalence and Cervical Cancer Incidence. *Cancer Epidemiology Biomarkers & Prevention*, 17(3):717–720. ISBN: 1055-9965 (Print).

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142.

McIntosh, A. (1963). A new approach to Middle English dialectology. *English Studies*, 44(1-6):1–11. Publisher: Routledge.

McIntosh, A., Samuels, M. L., and Benskin, M. (1986). *A Linguistic Atlas of Late Mediaeval English*. Aberdeen University Press, Aberdeen.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–558.

Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association*, pages 1–13. Publisher: American Statistical Association.

Moss, D. and Rousseau, J. (2022). Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models. arXiv:2203.06081 [math, stat].

Murphy, K. M. and Topel, R. H. (1985). Estimation and Inference in Two-Step Econometric Models. *Journal of Business & Economic Statistics*, 3(4):370–379. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Nicholls, G. K., Lee, J. E., Wu, C.-H., and Carmona, C. U. (2022). Valid belief updates for prequentially additive loss functions arising in Semi-Modular Inference. arXiv:2201.09706 [stat].

Nicholson, G., Blangiardo, M., Briers, M., Diggle, P. J., Fjelde, T. E., Ge, H., Goudie, R. J. B., Jersakova, R., King, R. E., Lehmann, B. C. L., Mallon, A.-M., Padellini, T., Teh, Y. W., Holmes, C., and Richardson, S. (2022). Interoperability of Statistical Models in Pandemic Preparedness: Principles and Reality. *Statistical Science*, 37(2):183–206. Publisher: Institute of Mathematical Statistics.

Nott, D. J., Seah, M., Al-Labadi, L., Evans, M., Ng, H. K., and Englert, B.-G. (2021). Using Prior Expansions for Prior-Data Conflict Checking. *Bayesian Analysis*, 16(1):203–231. Publisher: International Society for Bayesian Analysis.

Ogle, K., Barber, J., and Sartor, K. (2013). Feedback and Modularization in a Bayesian Meta–analysis of Tree Traits Affecting Forest Dynamics. *Bayesian Analysis*, 8(1):133–168. ISBN: 1931-6690.

Orton, H. and Dieth, E. (1962). *Survey of English Dialects*. University of Leeds. Google-Books-ID: FqrJzgEACAAJ.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22. arXiv: 1912.02762.

Perrone, V., Shen, H., Zolic, A., Shcherbatyi, I., Ahmed, A., Bansal, T., Donini, M., Winkelmolen, F., Jenatton, R., Faddoul, J. B., Pogorzelska, B., Miladinovic, M., Kenthapadi, K., Seeger, M., and Archambeau, C. (2020). Amazon SageMaker Automatic Model Tuning: Scalable Black-box Optimization. Technical report, Amazon. arXiv: 2012.08489.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2019). nlme: Linear and Nonlinear Mixed Effects Models.

Plummer, M. (2015). Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43.

Pompe, E. (2021). *Monte Carlo algorithms for Bayesian inference*. http://purl.org/dc/dcmitype/Text, University of Oxford.

Pompe, E. and Jacob, P. E. (2021). Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap. arXiv: 2110.11149.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In Kaski, S. and Corander, J., editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, AISTATS 2014*, volume 33, pages 814–822, Reykjavik, Iceland. PMLR. arXiv: 1401.0118 Series Title: Proceedings of Machine Learning Research ISSN: 15337928.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

Rezende, D. and Mohamed, S. (2015). Variational Inference with Normalizing Flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538, Lille, France. PMLR. Series Title: Proceedings of Machine Learning Research.

Richardson, S. (2022). Statistics in times of increasing uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4):1471–1496.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer New York, New York, NY. arXiv: 1011.1669v3 Series Title: Springer Texts in Statistics ISSN: 0162-1459.

Roy, V. (2020). Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412. _eprint: https://doi.org/10.1146/annurev-statistics-031219-041300.

Salinas, D., Seeger, M., Klein, A., Perrone, V., Wistuba, M., and Archambeau, C. (2022). Syne Tune: A Library for Large Scale Hyperparameter Tuning and Reproducible Research. In *Proceedings of the First International Conference on Automated Machine Learning*, pages 16/1–23. PMLR. ISSN: 2640-3498.

Savitsky, T., Vannucci, M., and Sha, N. (2011). Variable Selection for Nonparametric Gaussian Process Priors: Models and Computational Strategies. *Statistical Science*, 26(1).

Smith, A. F. M. (1986). Some Bayesian Thoughts on Modelling and Model Choice. *The Statistician*, 35(2):97. Publisher: JSTOR.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2014). OpenBUGS User Manual. url: https://www.mrc-bsu.cam.ac.uk/software/bugs/openbugs/.

Styring, A. K., Carmona, C. U., Isaakidou, V., Karathanou, A., Nicholls, G. K., Sarpaki, A., and Bogaard, A. (2022). Urban form and scale shaped the agroecology of early 'cities' in northern Mesopotamia, the Aegean and Central Europe. *Journal of Agrarian Change*.

Styring, A. K., Charles, M., Fantone, F., Hald, M. M., McMahon, A., Meadow, R. H., Nicholls, G. K., Patel, A. K., Pitre, M. C., Smith, A., So?tysiak, A., Stein, G., Weber, J. A., Weiss, H., and Bogaard, A. (2017). Isotope evidence for agricultural extensification reveals how the world's first cities were fed. *Nature Plants*, 3(6).

Szpiro, A. A. and Paciorek, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, 24(8):501–517. arXiv: 1210.7351.

Teh, Y. W., Bhoopchand, A., Diggle, P., Elesedy, B., He, B., Hutchinson, M., Paquet, U., Read, J., Tomasev, N., and Zaidi, S. (2021). Efficient Bayesian Inference of Instantaneous Reproduction Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the Covid-19 Epidemic in British Local Authorities.

Titsias, M. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, AISTATS 2009*, volume 5, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR. Series Title: Proceedings of Machine Learning Research.

Vehtari, A. (2021). Comparison of MCMC effective sample size estimators.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432. arXiv: 1507.04544 Publisher: Springer US ISBN: 1507.04544.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing

Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718. Publisher: International Society for Bayesian Analysis.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305. ISBN: 1601981848.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821.

Wang, Y. and Blei, D. M. (2019). Variational Bayes under Model Misspecification. In *Proceedings of the 33rd Conference on Neural Information Processing Systems, NeurIPS 2019*, pages 577–584. arXiv: 1905.10859.

Wasser, S. K., Mailand, C., Booth, R., Mutayoba, B., Kisamo, E., Clark, B., and Stephens, M. (2007). Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences*, 104(10):4228–4233. Publisher: Proceedings of the National Academy of Sciences.

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B., and Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences*, 101(41):14847–14852. Publisher: Proceedings of the National Academy of Sciences.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press. Series Title: Cambridge Monographs on Applied and Computational Mathematics.

Watanabe, S. (2012). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14:867–897. arXiv: 1208.6338 ISBN: 1532-4435.

Williamson, K. (2008). A Linguistic Atlas of Older Scots, Phase 1: 1380-1500. http://www.lel.ed.ac.uk/ihd/laos1/laos1.html.

Wu, P.-S. and Martin, R. (2021). Calibrating generalized predictive distributions. arXiv: 2107.01688.

Wu, P.-S. and Martin, R. (2023). A Comparison of Learning Rate Selection Methods in Generalized Bayesian Inference. *Bayesian Analysis*, 18(1):105–132. Publisher: International Society for Bayesian Analysis.

Xie, X. and Meng, X.-L. (2016). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, 27:1485–1594.

Xing, H., Nicholls, G., and Lee, J. K. (2020). Distortion estimates for approximate Bayesian inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1208–1217. PMLR. ISSN: 2640-3498.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but Did It Work?: Evaluating Variational Inference. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 12:8887–8895. arXiv: 1802.02538 Publisher: PMLR Place: Stockholmsmässan, Stockholm Sweden ISBN: 9781510867963.

Yu, X., Nott, D. J., and Smith, M. S. (2021). Variational inference for cutting feedback in misspecified models. arXiv: 2108.11066.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321. Conference Name: IEEE Transactions on Information Theory.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S., Dvornek, N., Papademetris, X., and Duncan, J. S. (2020). AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients. arXiv:2010.07468 [cs, stat].

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for Vector-Valued Functions: A Review. *Foundations and Trends in Machine Learning*, 4(3):195–266. arXiv: 1106.6251.