Applied Bayesian Inference for Diachronic Meaning Change

Schyan Zafar

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Statement of Originality

This thesis is my own work except as specified in the text and the Statement of Authorship forms at the end of Chapters 2, 3 and 5. I declare that the contents of this thesis are original and have not been submitted, in whole or in part, for consideration for any other degree or qualification in this or any other university.

Schyan Zafar

Trinity 2024

In loving memory of my grandfather
Brigadier Syed Mohammad Saleem
who tutored me in my formative years
and imbued me with a lifelong passion for learning

Acknowledgements

This thesis has been a long time coming; and in that time, I have built up a long list of people and entities to acknowledge. I shall try to do justice to all of them.

First off and above everyone else, I owe an enormous debt of gratitude to my supervisor Geoff Nicholls, not just for his role during my DPhil, but for all his support over our long association. Looking back, it was a lucky day for me in December 2005, when I had missed my final undergraduate admission interview earlier, and rushed out of breath into Geoff's office at St Peter's College for the rescheduled one. That first interaction with Geoff turned out to shape my academic journey over the next two decades.

Geoff, thank you for all your help whenever I sought it, whether during my undergraduate studies or during my DPhil, or any time in between. Thank you for offering me this DPhil position, for showing me how to conduct research, and for always being available whenever I needed your time. Thank you also for your understanding, and for not putting me under any pressure, as I went through challenging personal circumstances which delayed my DPhil timeline. You embody everything I could have hoped for in a tutor, mentor or supervisor, and I consider myself extremely privileged to have benefited from working with you.

Next, I thank George Deligiannidis for his supervision of my MSc dissertation, which set me up nicely for the subsequent DPhil; for the brainstorming sessions early on in my research; and for his help with references and advice sporadically throughout my DPhil.

I thank the admin and IT staff at the Department of Statistics for all the work done behind the scenes to keep the research flowing smoothly. In particular, I wish to acknowledge Beverley Lane and Jonathan Whyman for all their help.

I gratefully acknowledge the Engineering and Physical Sciences Research Council (EPSRC) for funding my DPhil for the first 3.5 years. I also gratefully acknowledge Jesus College for the annual research allowance, the various sports grants and the writing-up allowance.

DPhil is about more than just research, and I gained a lot during this time in many other ways. I thank Geoff for the opportunity to conduct undergraduate admission interviews at

St Peter's College. And I thank Robin Evans for the same at Jesus College. I found these to be very enjoyable and useful learning experiences. I thank the Department of Statistics for the opportunity to teach undergraduate classes, especially in Actuarial Science. I also thank Jesus College for the opportunity to teach undergraduate tutorials in Statistics. I found teaching to be a stimulating and satisfying experience.

Further afield, I thank Murray van Zuydam, my line manager at HSBC, for the opportunity to return to actuarial work while suspending my DPhil research. I found the experience refreshing and rewarding (not least in compensation terms!)

Outside of work, I am grateful to the Middle Common Rooms (MCRs) of Jesus College and St Peter's College for providing socially stimulating spaces and events, which greatly enhanced my Oxford experience. I am also grateful to the Oxford University Dancesport Club (OUDC) and the Oxford University Squash Racquets Club (OUSRC) for providing avenues to pursue my two main hobbies and passions besides research, and to all my friends and coaches at these clubs who made my time there enjoyable. To the OUSRC members in particular, whom I had the privilege of serving as Club Secretary and President, thank you for electing me Member of the Year for two years running.

To my colleagues and friends in office 1.19, Anthony, Sam and Rob, thank you for all the intellectual discussions, office banter and many games of bullet chess prior to COVID, and for the four-way chess and Catan sessions during the lockdowns. And Jessie J, thank you for continuing to make me feel part of the office after I became a hermit and switched to working from home post-COVID.

My Oxford experience would be incomplete without all the amazing people I met here. Hamish, thank you for being an incorrigible entertainer with your brazen antics and unfiltered humour. Marta, thank you for your warm friendship and your weekly circle of international home-cooked dinners. Alex G, thank you for the many good times on and off the squash court, and hopefully many more to come. And everyone else that I can't mention here without this turning into a thesis in its own right, thank you for making my time in Oxford so special.

I would be remiss not to acknowledge some very dear friends, with whom my association goes way back, and who have been important in my life over the last few years. Ali and Momina, thank you for making Oxford feel like home as soon as I arrived back here, and for the many games and pizza nights and other good times shared together. Zain, thank you for challenging my hitherto set-in-stone ideas, and for broadening my perspective on

the important things in life. And Uzair, thank you for being a true friend, despite being thousands of miles away, and providing invaluable emotional support during one of the most challenging periods of my life.

Last but one, Anna, thank you for believing in me and my abilities, even at times when I doubted myself. Thank you for your positivity and encouragement, and for lifting me up whenever I was down. I will always cherish the memories of our time together in Oxford.

Finally, I am forever grateful to my family, for their support throughout my life during good and bad times. To my parents especially, thank you for your encouragement which gave me the final push to accept my DPhil offer as I was hesitating at the last minute. And thank you for the upbringing you gave me, which made me the person I am today. You have a part in all my achievements, and I hope that I have made you proud.

Abstract

As a language evolves, the meanings or *senses* of many words in the language change. Examples include "gay", whose predominant sense has changed from bright or cheerful to homosexual; and "mouse", which has acquired a new sense of a computer pointing device in addition to the rodent sense. Modelling words with multiple senses, and learning their diachronic meaning changes from unlabelled text, is a fascinating challenge in statistical inference. One way to approach the problem is through a class of generative Bayesian models derived from the topic modelling literature. In this framework, the sense of a target word is represented as a distribution over context words, and sense prevalence is represented as a distribution over senses, both of which may change with time. This thesis works within this framework to posit new models, model-fitting procedures and inference methods for unsupervised learning of word senses and measurement of diachronic meaning change. Quantifying inferential uncertainty is a particular focus, since this aspect is important for modelling the small and sparse datasets used in our main application. Significant gains are achieved in terms of predictive accuracy, ground-truth recovery, sampling efficiency and scalability. An intuitive method for selecting the learning rate in a generalised Bayes' posterior is also explored. All results are demonstrated on real data from ancient Greek and English, as well as simulated examples.

Contents

Chapter 1

Introduction

"All models are wrong, but some are useful." — George E. P. Box

1 Statistical modelling, inference and sense change

Much has been written on the nature and philosophy of statistical modelling. Yet few statements, in the opinion of this author, capture its essence more succinctly than the famous aphorism of George E. P. Box quoted above. A typical statistical inference pipeline proceeds thus: the statistician observes some data (usually a sample from a larger population), posits a model describing the data generating process (DGP), fits the model to the data, and makes inferences. The model is descriptive, not prescriptive, and as such is inherently *wrong* in the sense that it is a simplified representation of the latent DGP. However, if the model describes the DGP *well enough*, where the definition of 'well enough' is context-dependent, then the model may be *useful*. In that case, inferences may be drawn which, among other things, aid our understanding of the DGP and/or allow us to make predictions or test hypotheses.

Statistical models are stochastic by definition, and inferences drawn using statistical models are therefore subject to uncertainty. Quantifying this uncertainty is important in order to gauge the reliability and variability of inferences. The literature on statistical inference abounds with theories and methods on the subject of quantifying uncertainty, often with no prescribed right or wrong approach, but merely presenting the subjective choices to be made. Amidst all the choices, the classical Bayesian paradigm continues to offer a statistical inference framework that is principled, natural and intuitive for quantifying uncertainty.

The Bayesian inference framework comprises an observation model $p(y|\vartheta)$ for data y given model parameters ϑ, and a prior $\pi(\vartheta)$ which incorporates the modeller's existing knowledge

or beliefs about ϑ. The parameters often have natural or physical interpretations. Having observed y, beliefs about ϑ are updated via Bayes' theorem to give the posterior $\pi(\vartheta|y)$. In this framework, 'modelling' refers to positing both $p(y|\vartheta)$ and $\pi(\vartheta)$, and 'model fitting' refers to computing or sampling from $\pi(\vartheta|y)$. The goal is to obtain posteriors that are well calibrated for the intended task (e.g. prediction or hypothesis testing). This often requires positing good models as well as selecting or developing good model-fitting procedures. Since model parameters are cast as random variables, any inferential uncertainty about them is naturally quantified using posterior probabilities. The principled approach to updating beliefs and quantifying uncertainty about model parameters within a probabilistic framework makes well-specified Bayesian models very useful inferential tools.

Whilst Box's aphorism holds true for modelling virtually all real-world phenomena, it seems particularly appropriate when it comes to natural languages. Computational linguistics, a subfield of artificial intelligence (AI) that uses computational methods to understand and model human language, has advanced exponentially in recent years. However, a language is an incredibly complex system, evolved over hundreds or thousands of years, which makes it challenging to accurately model any aspect of a language. Diachronic meaning change, i.e. the evolution of word meanings or *senses*[1] over time, is no exception. This fundamental linguistic phenomenon is influenced by many factors, including social, cultural and historical, and a full model capturing all these complexities is (at least for now) unachievable.

Historically, diachronic meaning or sense change has been studied by linguists and researchers manually. However, with advancements in technology, computational approaches for studying sense change have become increasingly popular since they allow much larger volumes of textual data to be analysed systematically and efficiently than is possible with any manual study. A common scenario is a researcher seeking to explore the different senses and usages of a given word in a large corpus of unlabelled text. For example, the usage of the word "gay" has changed in recent decades, where it is now very rarely used in its original meanings of bright/showy, bold or merry, and is predominantly used in the meaning of homosexual. Without having to read the entire text, the researcher might wish to

[1] The term 'sense' is used somewhat imprecisely in this thesis. Technically, a distinction exists between polysemy (i.e. multiple meanings for a word) and homography (i.e. words with the same spelling but different meanings). Polysemous words such as 'mouse' (meaning a rodent or a computer pointing device) are listed in a dictionary under the same entry, with the different senses denoted as a numbered list. In contrast, homographic words such as 'bear' (the animal) and 'bear' (to carry) are listed as separate entries. The distinction between polysemes and homographs is in their origin: a polyseme is one word that has changed meaning or acquired new meanings over time, whereas homographs are different words that simply happen to share the same spelling. This distinction is noted here for the record, but is ignored throughout the rest of the thesis. The terms 'meaning' and 'sense' are used interchangeably.

identify and track the different senses of "gay" through time with the aid of a suitable model. Computational linguistics provides many models that allow this kind of analysis.

The usefulness of diachronic sense change models, like any other model, depends on inferential goals. In many cases, the modeller seeks to understand the underlying data structure, so parameter interpretability is a priority. Also, particularly where training data is limited, uncertainty quantification is important. Bayesian inference provides a powerful toolkit for modelling diachronic sense change in such cases. This thesis applies the Bayesian inference framework to develop new and physically interpretable models of diachronic sense change, provide novel methods for fitting these models and, thus, obtain well-calibrated measurements of diachronic sense change with uncertainty quantification. The utility of the models and methods developed herein is demonstrated throughout the thesis with applications on real and synthetic data, showing how they offer significant improvements over the previous best within the class of models considered.

2 Problem setting

This thesis is an application of Bayesian inference to learn word senses and study diachronic sense change from unlabelled text. The research for this thesis started in 2018, and was motivated by the (at the time) recent papers of Frermann and Lapata (2016) and Perrone et al. (2019). In this section, we give an overview of the research problem posed by these papers and how it shaped our research.

2.1 Framework

Frermann and Lapata (2016) introduced a framework for representing word meanings and modelling their temporal changes by adapting from the more familiar dynamic topic modelling literature (Blei and Lafferty, 2006). They gave a generative dynamic Bayesian model of Sense ChANge called SCAN, in which the senses of a target word are represented as probability distributions over context words, and sense prevalence is represented as a probability distribution over senses. Both of these have a time component, which allows the diachronic nature of sense change to be modelled and analysed in a systematic manner.

The model is fitted to a set of text *snippets* centred on the target word. Some example snippets for the word "bug" showing its different senses are given in Table 1. In any snippet, the target word has one of a small number of senses. In applications, we do not know these sense labels and would like to recover them.

Table 1: Example text snippets for target word "bug" taken from Zafar and Nicholls (2024b, Table 1) showing its four different senses. Context words are lemmatised, and stopwords, infrequent words and punctuation are dropped, to get the data used in model fitting.

insect	... insect repellent on a winter trip when there are no bugs around. Your first-aid kit should reflect your personal needs as ...
micro-organism	... These intruders are what cause the fever, for the TB bugs are not virulent enough to cause high temperatures. The effect ...
software glitch	... bug the Quality Assurance people find and $20 for each bug the programmers fix. These are the same programmers who create ...
tapping device	... too much information has been collected through secret informants, wiretaps, bugs, surreptitious mail opening and break-ins, the Church Report had warned ...

In order to infer sense change, our first task is to cluster the snippets so that snippets in the same cluster share the same target-word sense. This is unsupervised Bayesian model-based clustering (Wade, 2023, Section 2a) of snippets: we have a probability distribution over clusterings. In a realisation of this distribution, each snippet at each time has an assigned sense. The evolving sense and sense-prevalence representations are then readily available via the model parameters.

To get the data used in model fitting, context words are lemmatised (i.e. reduced to their root form such as "open" instead of "opening"), and stopwords (i.e. most common words such as "the" and "an"), infrequent words (e.g. with frequency below 10 in the corpus) and punctuation are dropped from the snippets. Each snippet is then regarded as a "bag of words", which is a simplification used frequently within natural language processing (NLP) and means that grammar and syntax are ignored. Referring back to the discussion in Section 1, these simplifications obviously do not represent how language works in practice; however, they make modelling much easier.

Perrone et al. (2019) extended SCAN to include a genre component in the sense prevalence representation, which adds to inferential accuracy in many cases. They called this new model GASC, for genre-aware semantic change, and applied it to model sense change for target words in the Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018). Both SCAN and GASC are generative Bayesian models, and the key attraction of these models is in their physically interpretable parameters which allow the relationship between the data and target-word senses to be described. Being Bayesian, these models are also well suited to quantifying inferential uncertainty. However, this aspect was not discussed by the authors of SCAN and GASC.

2.2 Research aims and direction

Our research set out to develop new models and inference methods for learning target-word senses and measuring diachronic sense change from text snippets in which sense labels are not available. We kept within the same generative modelling framework to retain parameter interpretability, and used the same ancient Greek data for our main application.

Good quality datasets for this kind of analysis are scarce, whereas the ancient Greek corpus was compiled specifically for the purpose of "developing a computational model of semantic change in Ancient Greek" (Vatri and McGillivray, 2018, Abstract). Furthermore, it benefits from expert annotation (Vatri et al., 2019; McGillivray et al., 2019) providing the 'ground truth', which makes it very useful for testing these models. Given that modelling semantic change with this data is of genuine interest to researchers, having this as our main application provides a real use case for the models. Also, this data is hard to model, which makes it a fascinating case study in statistical inference.

Unlike previous authors, our research had an explicit focus on quantifying uncertainty, which is important when working with small[2] datasets. This focus led to the direction of our research being determined quite naturally. Firstly, on attempting to reproduce what was already done by previous authors, it quickly became apparent that fitting the models to this data was no easy task, especially using their choice of sampling method. This led us to seek better model-fitting procedures for the existing models, so that the posteriors could be sampled more accurately and efficiently.

Secondly, even with our best model-fitting procedures, getting convergence on the same posterior in independent runs often proved challenging. This is because the SCAN and GASC posteriors are often multimodal and have highly 'ridged' structures, so any sampler tends to get stuck in one of the modes or explores the ridges very slowly. Aside from the difficulty of model fitting, the multimodality of posteriors makes parameter interpretability problematic. This led us to consider carefully the relationship between context words in snippets and target-word senses, and thus add more structure to the models in order to describe diachronic sense change more accurately.

Thirdly, referring back to the discussion in Section 1 once again, all the models discussed in this thesis, existing or new, are *wrong*; or, using a more helpful description, *misspecified*.

[2] Typically, any dataset with under 40 million tokens is considered 'small', whereas the ancient Greek data has around 10 million tokens.

The bag-of-words assumption in particular assumes an independence that does not actually exist, which potentially causes the posterior variance to be understated. This led us to explore Bayesian approaches for correcting misspecification.

2.3 What this thesis is *not* about

The same year that this research started, 2018, was also the year that two important Large Language Models (LLMs) were first introduced. These were the Generative Pre-trained Transformer (GPT, Radford et al. 2019) and the Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2019), coming on the heel of the seminal Google paper "Attention is all you need" (Vaswani et al., 2017). Since then — and especially since AI has been made accessible to the public through engines like ChatGPT and Google Gemini, in what has been termed by Wikipedia as the "AI boom era" (Wikipedia contributors, 2024) — LLMs have come to dominate the landscape of NLP. In our experience over the last few years, mentioning 'language modelling' tends to make the audience automatically think of LLMs. It is therefore important to clarify that our research is neither based on nor in competition with LLMs in any way.

LLMs are extremely useful tools utilised for a myriad of purposes. However, they are primarily black-box models designed to generate language that resembles human speech. They are not designed for modelling diachronic meaning change with uncertainty quantification for a given target word in a small data sample. In fact, in our experience, LLMs fail to even identify the correct sense for simple English target words in snippets where it is not obvious, let alone obscure target words from ancient Greek. Moreover, they are not designed for interpretability, which is our priority, and uncertainty quantification is well beyond their remit. The architecture behind LLMs, however, *is* used for diachronic sense change modelling, and we review it in this context in Chapter 3.

3 Background

We now show the connection of Frermann and Lapata (2016) and Perrone et al. (2019) to previous research, and give an overview of their models and fitting methods. For convenience, the notation for SCAN/GASC and all acronyms used in this thesis are collated in Appendices B and C respectively.

3.1 Topic models

The SCAN and GASC models are adapted from topic models, as are our new models introduced later, so it helps to understand topic models before going deeper into SCAN and

GASC. We give an overview of the main topic models on which the sense change models discussed in this thesis are based, using a common notational framework to show the links between them.

Probabilistic topic models (e.g. Blei 2012) are generative bag-of-words models that are widely used to infer themes or *topics* from a collection of documents. Suppose we have a set W of D documents. Each document $W_d, d = 1, \ldots, D$, is made up of L_d words taken from a vocabulary $\{1, \ldots, V\}$. The V-sized vocabulary represents the unique tokens after lemmatising and filtering (e.g. to remove stopwords and very low frequency words), and L_d denotes the number of words that remain in document d after filtering. Each document is a bag of words in which word order is unimportant. To emphasise this, we denote the positions occupied by the words as a random permutation $\{i_1, \ldots, i_{L_d}\}$ of size L_d.

3.1.1 Latent Dirichlet Allocation

One of the first[3] and best known topic models is Latent Dirichlet Allocation (LDA) given by Blei et al. (2003). The main premise of LDA is that we assume each document to be a mixture over K topics, where each topic $k \in \{1, \ldots, K\}$ is a probability distribution $\tilde{\psi}^k$ over the vocabulary. The idea is that topic k assigns high probabilities $\tilde{\psi}^k_v$ to words $v \in \{1, \ldots, V\}$ more strongly associated with that topic. Under LDA, each document $d \in \{1, \ldots, D\}$ is generated as follows:

1. draw topic proportions $\tilde{\eta}^d \sim \text{Dir}\,(\eta_1, \ldots, \eta_K)$;
2. for each context position $i \in \{i_1, \ldots, i_{L_d}\}$:
   (a) draw topic assignment $z_{d,i}|\tilde{\eta}^d \sim \text{Mult}\,(\tilde{\eta}^d_1, \ldots, \tilde{\eta}^d_K)$;
   (b) draw word $w_{d,i}|z_{d,i}, \tilde{\psi}^{z_{d,i}} \sim \text{Mult}\left(\tilde{\psi}^{z_{d,i}}_1, \ldots, \tilde{\psi}^{z_{d,i}}_V\right)$.

Dirichlet priors are placed on both topic proportions and topics (hence the model name), that is $\tilde{\eta}^d \sim \text{Dir}\,(\eta_1, \ldots, \eta_K)$ and $\tilde{\psi}^k \sim \text{Dir}\,(\psi_1, \ldots, \psi_V)$, where $\eta \in \mathbb{R}^K_{>0}$ and $\psi \in \mathbb{R}^V_{>0}$ are the respective concentration parameters. The LDA posterior is given by

$$\pi(\tilde{\eta}, \tilde{\psi}, z|W) \propto \pi(\tilde{\psi}) \prod_{d=1}^{D} \pi(\tilde{\eta}^d) \prod_{i=i_1}^{i_{L_d}} \tilde{\eta}^d_{z_{d,i}} \tilde{\psi}^{z_{d,i}}_{w_{d,i}}. \tag{1}$$

[3] Although LDA is widely regarded as one of the first topic models, the same model had in fact been independently developed earlier by Pritchard et al. (2000) for detecting structure in populations. Nevertheless, LDA was amongst the first models to apply this to text data.

3.1.2 Dynamic Topic Model

LDA was extended to the dynamic topic model (DTM) by Blei and Lafferty (2006) to infer the temporal evolution of topics in a document corpus. Suppose each document d has a deterministic mapping $\tau_d \in \{1, \dots, T\}$ to its time label. In DTM, we introduce a new parameter $\phi^t \in \mathbb{R}^K$ to model the topic proportion means over time $t = 1, \dots, T$, and place some prior $\pi(\phi)$ on it describing the temporal structure. η is redefined, so that document topic proportions are now modelled as $\tilde{\eta}^d = \text{softmax}(\eta^d)$, with $\eta^d | \phi^{\tau_d} \sim \mathcal{N}(\phi^{\tau_d}, \text{diag}(\kappa_\eta))$.

We also redefine ψ so that now we have $\psi^{k,t} \in \mathbb{R}^V$ with some temporal structure modelled by the prior $\pi(\psi^{k,\cdot})$ for each topic $k \in \{1, \dots, K\}$. We then set $\tilde{\psi}^{k,t} = \text{softmax}(\psi^{k,t})$. The exact forms of the temporal structures in $\pi(\phi)$ and $\pi(\psi)$ may be ignored for now. Each document $d \in \{1, \dots, D\}$ under DTM is then generated as follows:

1. draw $\eta^d | \phi^{\tau_d} \sim \mathcal{N}(\phi^{\tau_d}, \text{diag}(\kappa_\eta))$ and set $\tilde{\eta}^d = \text{softmax}(\eta^d)$;
2. for each context position $i \in \{i_1, \dots, i_{L_d}\}$:
   (a) draw topic assignment $z_{d,i} | \tilde{\eta}^d \sim \text{Mult}\left(\tilde{\eta}^d_1, \dots, \tilde{\eta}^d_K\right)$;
   (b) draw word $w_{d,i} | z_{d,i}, \tilde{\psi}^{z_{d,i},\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_{d,i},\tau_d}, \dots, \tilde{\psi}_V^{z_{d,i},\tau_d}\right)$.

The DTM posterior is given by

$$\pi(\phi, \eta, \psi, z | W) \propto \pi(\phi)\pi(\psi) \prod_{d=1}^{D} \pi(\eta^d | \phi^{\tau_d}) \prod_{i=i_1}^{i_{L_d}} \tilde{\eta}^d_{z_{d,i}} \tilde{\psi}^{z_{d,i},\tau_d}_{w_{d,i}}. \tag{2}$$

3.1.3 Dynamic Embedded Topic Model

The (static) embedded topic model (ETM) was given by Dieng et al. (2020), which extended LDA to incorporate word embeddings within the model. This is not needed for our exposition. Word embeddings, however, are an important concept which we will expand upon in Chapter 3. For now, a word embedding ρ_v is simply a vector representation of word v in an M-dimensional real space, and ρ denotes the $V \times M$ matrix of word embeddings.

The dynamic embedded topic model (D-ETM) was given by Dieng et al. (2019), incorporating word embeddings into DTM. The embedding matrix ρ is learnt independently of the model. In D-ETM, topic evolution is modelled via a new parameter $\xi^{k,t} \in \mathbb{R}^M$ in the embedding space, with a prior $\pi(\xi^{k,\cdot})$ describing the temporal structure for each topic $k \in \{1, \dots, K\}$. We then redefine $\psi^{k,t} = \rho\xi^{k,t}$. Compared to DTM, only the definition (and hence the prior) for ψ is changed; otherwise, the generative model remains identical to DTM, and is given below for completeness:

1. draw $\eta^d|\phi^{\tau_d} \sim \mathcal{N}\left(\phi^{\tau_d}, \text{diag}(\kappa_\eta)\right)$ and set $\tilde{\eta}^d = \text{softmax}(\eta^d)$;
2. for each context position $i \in \{i_1, \ldots, i_{L_d}\}$:
   (a) draw topic assignment $z_{d,i}|\tilde{\eta}^d \sim \text{Mult}\left(\tilde{\eta}^d_1, \ldots, \tilde{\eta}^d_K\right)$;
   (b) draw word $w_{d,i}|z_{d,i}, \tilde{\psi}^{z_{d,i},\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_{d,i},\tau_d}, \ldots, \tilde{\psi}_V^{z_{d,i},\tau_d}\right)$.

The D-ETM posterior also remains unchanged from DTM at the level of ϕ, η, ψ, z:

$$\pi(\phi, \eta, \psi, z|W) \propto \pi(\phi)\pi(\psi) \prod_{d=1}^{D} \pi(\eta^d|\phi^{\tau_d}) \prod_{i=i_1}^{i_{L_d}} \tilde{\eta}^d_{z_{d,i}} \tilde{\psi}^{z_{d,i},\tau_d}_{w_{d,i}}. \tag{3}$$

3.2 SCAN and GASC models

The SCAN and GASC models will be discussed in greater detail in Chapter 2. Here, we only give a brief overview highlighting their connection to topic models.

SCAN is an adaptation of DTM designed to infer the temporal evolution of the *senses* of a target word, rather than the topics of a document. The model is fitted to a set W of D *snippets*, rather than documents, with each snippet centred on the target word but dropping the target word itself. The snippet window is fixed at L, with $L/2$ context words on either side of the target word, where L is chosen by the modeller. However, after filtering the vocabulary, we are left with L_d words[4] in each snippet $d \in \{1, \ldots, D\}$.

The key difference between SCAN and DTM is that each snippet d in SCAN has only one sense $k \in \{1, \ldots, K\}$, whereas each document d in DTM is a mixture over topics. Therefore, the mixture (topic proportions) parameter $\tilde{\eta}$ is redundant in SCAN and is dropped from the model. Instead, we now define a new sense prevalence distribution $\tilde{\phi}^t = \text{softmax}(\phi^t)$. Snippet sense assignments are then drawn as $z_d|\tilde{\phi}^{\tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\tau_d}, \ldots, \tilde{\phi}_K^{\tau_d}\right), d = 1, \ldots, D$. Compared to DTM, only the function of ϕ (rather than its definition) changes in SCAN.

For each snippet $d \in \{1, \ldots, D\}$, in addition to a deterministic time label $\tau_d \in \{1, \ldots, T\}$, there is also a deterministic genre label $\gamma_d \in \{1, \ldots, G\}$. Arguably, target-word sense prevalence could behave differently depending on the genre. GASC allows this behaviour by including a genre covariate in the model. Hence, we now use $\phi^{g,t} \in \mathbb{R}^K$ to model sense

[4] We take the approach of dropping words from snippets whenever they are filtered from the vocabulary, which leads to a variable snippet length $L_d, d = 1, \ldots, D$. An alternative is to keep the length fixed at L, but fill in the filtered positions by collapsing the remaining (more distant) words together. The former approach with a wider window should give similar results to the latter approach with a narrower window. In our applications, we use $L = 14$ or 20 with the former approach, whereas the authors of SCAN and GASC use $L = 10$ with the latter approach. The choice is subjective and may vary based on the application.

prevalence over time $t = 1, \ldots, T$, with some prior $\pi(\phi^{g,\cdot})$ describing the temporal structure for each genre $g \in \{1, \ldots, G\}$, and we set $\tilde{\phi}^{g,t} = \text{softmax}(\phi^{g,t})$. This is identical to SCAN whenever we have $G = 1$. Also, for both SCAN and GASC, ψ and $\tilde{\psi}$ remain unchanged compared to DTM, except they now model the senses rather than the topics.

Putting this together, under GASC, each snippet $d \in \{1, \ldots, D\}$ is generated as follows:

1. draw sense assignment $z_d | \tilde{\phi}^{\gamma_d, \tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d,\tau_d}\right)$;
2. for each context position $i \in \{i_1, \ldots, i_{L_d}\}$:
   - draw word $w_{d,i} | z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d,\tau_d}, \ldots, \tilde{\psi}_V^{z_d,\tau_d}\right)$.

The GASC posterior is given by

$$\pi(\phi, \psi, z | W) \propto \pi(\phi)\pi(\psi) \prod_{d=1}^{D} \tilde{\phi}_{z_d}^{\gamma_d,\tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{z_d,\tau_d}. \tag{4}$$

A plate diagram for GASC is shown in Appendix A Figure 1.

3.3 Posterior inference

The posteriors (1), (2), (3) and (4) are analytically intractable, and therefore need to be approximated using numerical methods. The approaches for doing so fall into two main classes: variational methods (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017) and Monte Carlo methods (Robert and Casella, 2013).

Variational methods are a class of optimisation techniques. When targeting a posterior density $\pi(\vartheta | y)$, we first posit a family of approximation densities $\mathcal{L}$, and then seek the density $q(\vartheta) \in \mathcal{L}$ minimising the Kullback-Leibler (KL) divergence to $\pi(\vartheta | y)$, that is

$$q^*(\vartheta) = \arg \min_{q(\vartheta) \in \mathcal{L}} \text{KL}\left(q(\vartheta) \parallel \pi(\vartheta | y)\right). \tag{5}$$

This is equivalent to maximising the *evidence lower bound* (ELBO) given by

$$\text{ELBO}(q) = \mathbb{E}_{\vartheta \sim q}\big(\log \pi(\vartheta) p(y | \vartheta) - \log q(\vartheta)\big),$$

which can be done by optimisation techniques such as gradient ascent. A commonly used choice for $\mathcal{L}$ is the mean field variational family, in which we have $q(\vartheta) = \prod_{j=1}^{m} q_j(\vartheta_j)$. This choice makes optimisation efficient, and the $q^*(\vartheta)$ found through mean field variational inference is in many cases a good approximation for $\pi(\vartheta | y)$.

Monte Carlo methods are a class of sampling techniques which, when used appropriately, are guaranteed to converge asymptotically to the target distribution. In practice, within the

wider class of Monte Carlo methods, Markov Chain Monte Carlo (MCMC, Brooks et al. 2011) methods are most frequently used.[5] Classical MCMC methods, such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) are probably already familiar to all readers, and we will cover some more advanced MCMC algorithms in Chapters 2 and 3.

The relative merits of variational inference and MCMC are summarised by Blei et al. (2017, Section 1). Variational inference is computationally fast but approximate, whereas MCMC is expensive but asymptotically exact. Importantly, variational methods tend to underestimate posterior variance. Hence, variational methods may be suitable for large datasets where point estimation is the goal, but MCMC is more appropriate for small datasets where uncertainty quantification is important. Furthermore, variational inference has a mode-seeking behaviour due to the KL objective (5), which can be problematic for posteriors that are ridged or multimodal.

For the models discussed above, the authors of LDA, DTM and D-ETM use variational inference. They marginalise over the discrete z parameter, leaving only continuous parameters, which can be approximated using variational optimisation techniques. When working with topic models, we are usually dealing with large volumes of data, and uncertainty quantification is not the main goal, so variational approaches seem sensible.

However, specifically for LDA, an alternative collapsed Gibbs sampler is given by Griffiths and Steyvers (2004). This works because the Dirichlet priors for $\tilde{\eta}$ and $\tilde{\psi}$ in LDA are conjugate to their respective multinomial sampling distributions, which makes it possible to marginalise $\tilde{\eta}$ and $\tilde{\psi}$ out of the posterior, leaving only the marginal posterior $p(z|W)$. This marginal can be readily sampled with Gibbs sampling. Then, conditioned on the z samples, it is easy to estimate $\tilde{\eta}$ and $\tilde{\psi}$. This collapsed Gibbs sampler only works for LDA, and not for the other models discussed, since the Dirichlet priors are dropped in favour of logistic normals, so the conjugacy is lost.

For SCAN and GASC, the authors use the blocked Gibbs sampling strategy of Mimno et al. (2008). We will examine this more critically in Chapter 2. Briefly, this strategy exploits the conditional independence of ϕ and ψ if we condition on z in (4). The posterior (4) can be

[5] "... most Bayesian inference could be done by MCMC, whereas very little could be done without MCMC." (Geyer, 2011, Section 1.1)
"MCMC sampling has evolved into an indispensable tool to the modern Bayesian statistician." (Blei et al., 2017, Section 1)

factorised as

$$\pi(\phi, \psi, z|W) \propto \pi(\phi)\pi(\psi) \prod_{t=1}^{T} \prod_{k=1}^{K} \left(\prod_{g=1}^{G} (\tilde{\phi}_k^{g,t})^{N_{k,g,t}^z} \right) \left(\prod_{v=1}^{V} (\tilde{\psi}_v^{k,t})^{N_{v,k,t}^{W,z}} \right) \tag{6}$$

where $N_{k,g,t}^z$ is the count of snippets with sense-genre-time equal (k, g, t) under z, and $N_{v,k,t}^{W,z}$ is the count of context word v in snippets W with sense k at time t under assignments z. Conditioning on z in (6) gives, independently,

$$\pi(\phi|z) \propto \pi(\phi) \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{g=1}^{G} (\tilde{\phi}_k^{g,t})^{N_{k,g,t}^z} \quad \text{and} \quad \pi(\psi|W, z) \propto \pi(\psi) \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{v=1}^{V} (\tilde{\psi}_v^{k,t})^{N_{v,k,t}^{W,z}}, \tag{7}$$

which allows ϕ and ψ to be sampled using the auxiliary uniform variable strategy of Mimno et al. (2008). Also, conditioning on ϕ and ψ in (4) gives, independently for each $d \in \{1, \ldots, D\}$,

$$p(z_d|W_d, \phi, \psi) \propto \tilde{\phi}_{z_d}^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{z_d, \tau_d}, \tag{8}$$

which is a multinomial distribution over senses $1, \ldots, K$ and easy to sample. Hence, the blocked Gibbs strategy alternately targets (7) and (8).

4 Thesis outline and contributions

This thesis is in an integrated format consisting of six chapters. Chapter 1 is a thesis introduction; Chapters 2 and 3 are self-contained published papers; Chapter 4 is an unpublished supplement to Chapter 3; Chapter 5 is a self-contained unsubmitted preprint; and Chapter 6 is a thesis conclusion. Each self-contained paper includes an introduction and literature review specific to that paper, so we do not include a separate thesis literature review in order to keep repetition to a minimum. For convenience, all references have been collated into a single Bibliography at the end of the thesis.

A high-level overview of Chapters 2–5 is presented below, outlining their main contributions. These chapters were produced in chronological order, and follow a natural progression of research ideas from one chapter to the next. Some model plate diagrams, notation and acronyms are collated in this chapter's Appendices for easy reference.

Chapter 2: Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference

[Published as Zafar and Nicholls (2022)]

In this chapter, firstly, we give a new model called DiSC (Diachronic Sense Change). DiSC is based on the GASC model, but the key difference is that we impose an additive structure on the sense and time effects by setting $\psi^{k,t} = \chi^k + \theta^t$. (Other differences relate to the time-series priors and hyperparameters.) As well as positing the new model, we also treat prior elicitation explicitly with deductive reasoning, which was not done by previous authors. A plate diagram for DiSC is shown in Appendix A Figure 2.

The new parameters $\chi^k \in \mathbb{R}^V$ and $\theta^t \in \mathbb{R}^V$ model the variations across senses $k \in \{1, \ldots, K\}$ and time periods $t \in \{1, \ldots, T\}$ respectively. The result is a drastic reduction in the dimension of ψ from VKT parameters in GASC to $V(K+T)$ parameters in DiSC, and a significant improvement in predictive accuracy and true-model recovery: on unsupervised estimation and uncertainty quantification in sense prevalence $\tilde{\phi}$, for the ancient Greek word "kosmos", we do nearly as well as an 'oracle' classifier that simply fits a multinomial model conditioned on knowing the true sense labels.

Secondly, we show that the posterior sampling strategy used by the authors of SCAN and GASC is very inefficient, and provide alternatives. In particular, we marginalise over the discrete z parameter in the posterior, and use gradient-based MCMC methods (Roberts and Tweedie, 1996; Roberts and Rosenthal, 2002; Duane et al., 1987; Neal, 2011; Beskos et al., 2013; Hoffman and Gelman, 2014) to target the remaining continuous parameters. This substantially improves sampling efficiency.

Our new DiSC model was posited after careful consideration of the relationship between target-word senses and context words, and describes temporal sense evolution better than its predecessors. Even in synthetic data experiments where we simulate data according to GASC itself, DiSC scores at least as well as GASC on predictive accuracy (as measured by Brier scores). Recall that we are interested in quantifying uncertainty. With DiSC, we are able to do this via posterior credible sets which agree very well with those under expert sense annotation for the ancient Greek word "kosmos".

Chapter 3: An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek

[Published as Zafar and Nicholls (2024a)]

In this chapter, we introduce another new model called EDiSC (Embedded Diachronic Sense Change). EDiSC is an extension of DiSC in which we incorporate word embeddings, analogous to how D-ETM extended DTM. The main novelty in this is to bring together two previously distinct approaches for modelling sense change, i.e. topic-based models and

embedding-based models. This was not straightforward, and again careful statistical modelling and consideration of the model structure were required to make things work.

We posit a new structure $\psi^{k,t} = \rho\xi^{k,t} + \varsigma = \rho(\chi^k + \theta^t) + \varsigma$, where the parameters $\xi^{k,t}, \chi^k, \theta^t \in \mathbb{R}^M$ are all vectors in an M-dimensional embedding space, $\rho \in \mathbb{R}^{V \times M}$ is a matrix of context-word embeddings, and $\varsigma \in \mathbb{R}^V$ is a correction parameter. A plate diagram for EDiSC is shown in Appendix A Figure 3.

The benefits of this new model are twofold. Firstly, EDiSC gives much improved predictive accuracy, ground-truth recovery and uncertainty quantification compared to DiSC. We achieve $\tilde{\phi}$ estimates approaching those given by an 'oracle' classifier conditioned on the truth, for all three ancient Greek target words for which expert sense annotation is available. This is because the embeddings carry extra semantic information about all context words which is lacking in the snippet data. Secondly, the dimension of the parameter space for EDiSC is reduced even further compared to DiSC (since $M < V$), which results in more efficient Monte Carlo sampling and better scalability with increasing data size.

Two important modelling choices need to be made, namely the number of model senses K and the embedding dimension M. Previous authors did not address how to select K for these models, and we were unable to find any principled method for selecting M in the wider embedding literature. In this chapter, we give guidelines on selecting both values based on the widely applicable information criterion (WAIC) (Watanabe, 2010; Vehtari et al., 2017) combined with model-fitting and interpretability considerations.

Finally, even with state-of-the-art MCMC methods, the posteriors for these models are very hard to sample. We show how to overcome this difficulty, to a large extent, with likelihood tempering or annealing (Geman and Geman, 1984; Hajek, 1988).

Chapter 4: Extensions and a Refactorisation of the Embedded Diachronic Sense Change (EDiSC) Model

[Unpublished supplement to Zafar and Nicholls (2024a)]

In this short supplemental chapter, we present two extensions of EDiSC: EDiSC-η and EDiSC-β. Both of these extensions exploit the idea that target-word senses may be correlated for snippets taken from the same document, and both result in greater predictive accuracy compared to EDiSC, thus giving very strong evidence to support this intuition.

The models approach the correlations in different ways. EDiSC-η introduces a new parameter η, similar to topic models, representing the sense proportions in a single document. Separately, EDiSC-β assumes that there is a 'dominant' document sense, and introduces a new parameter β representing the probability of drawing this sense for any snippet from that document. The gains in predictive accuracy from both models are similar.

Separately, this chapter also explores a refactorisation of $\tilde{\psi}$ in the EDiSC model. This was an interesting line of investigation but did not improve model performance.

Chapter 5: Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks

[Preprint available online as Zafar and Nicholls (2024b)]

This exploratory chapter takes a different trajectory to the ones earlier. Whilst all previous chapters explicitly attempt to improve measurements of diachronic sense change and uncertainty quantification, that is not the *goal* for this chapter but rather a by-product.

Generalised Bayesian Inference (GBI, Walker and Hjort 2002; Zhang 2006; Bissiri et al. 2016) attempts to correct model misspecification by introducing a fractional power λ, called the learning rate, on the likelihood in a standard Bayes posterior. We already know that EDiSC is misspecified, so we expect GBI to help, and it does. However, existing approaches for selecting λ (Wu and Martin, 2023) do not work for EDiSC. Therefore, in this chapter, using EDiSC as a case study, we explore a novel approach for selecting the learning rate.

We do this using Posterior Predictive Checks (PPC, Meng 1994; Gelman et al. 1996). PPC is a diagnostic tool used to *detect* model misspecification, not correct for it. We propose a novel method for selecting λ whereby we select the smallest value for which a PPC hypothesis test is not rejected at the 10% level. We show that this leads to optimal or near-optimal λ selection for our three ancient Greek datasets, as well as for ten additional datasets from English. Nevertheless, our approach is only exploratory, and we still need to define the conditions required for it to be applicable more widely.

Appendices

A Model plate diagrams

For reference, plate diagrams for the GASC, DiSC and EDiSC models are shown here for three time periods. Dashed nodes are constant parameters, solid black nodes are latent

variables and solid red nodes are observed variables. The notation is summarised in Appendix B.

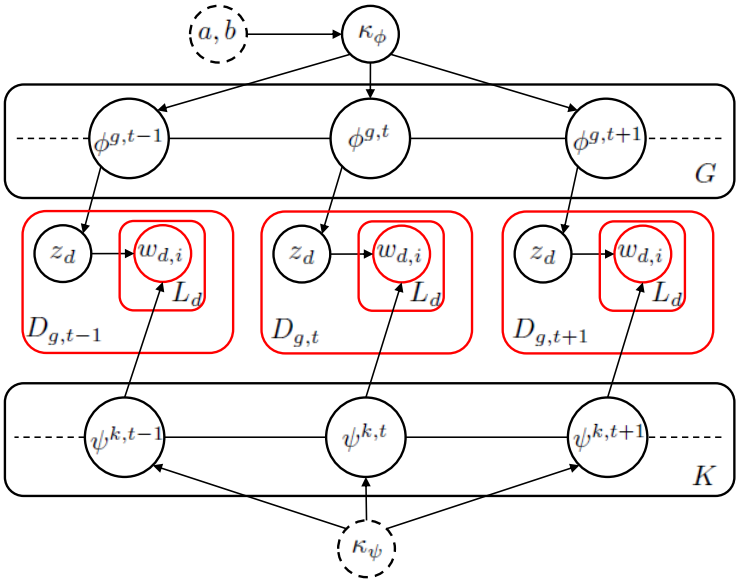


Figure 1: GASC plate diagram

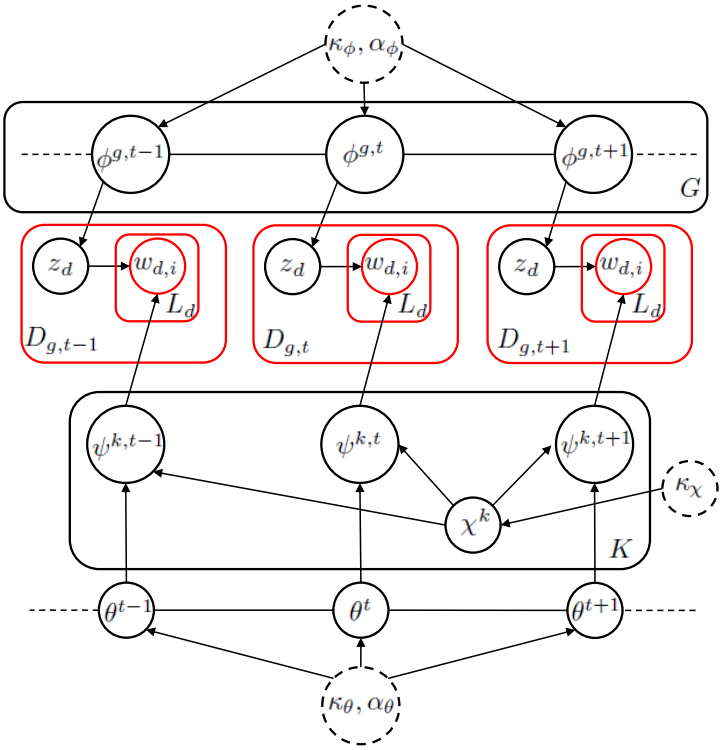


Figure 2: DiSC plate diagram

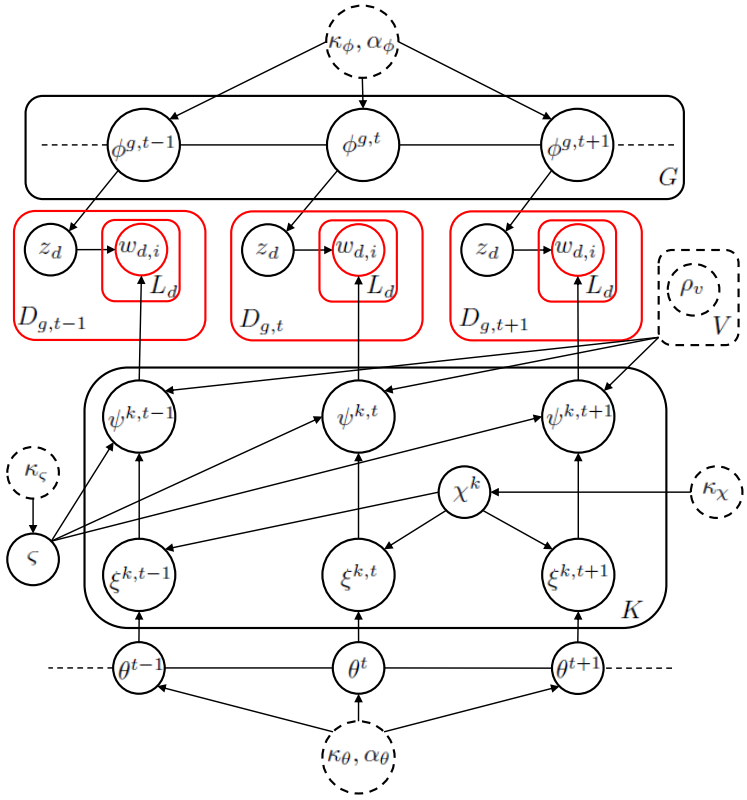


Figure 3: EDiSC plate diagram

B Notation

The notation used most frequently in this thesis is outlined here for easy reference.

Dummy variables

- t — time period.
- g — genre.
- k — sense.
- d — snippet.
- i — context position in a snippet.
- v — a word in the vocabulary.
- w — a word in a snippet.
- n — MCMC iteration.

Constants

- T = number of time periods.
- G = number of genres.

- K = number of target-word senses or meanings.
- D = total number of snippets.
- L = number of context positions in a snippet, not counting the target word itself. There are $L/2$ context positions to each side of the target word.
- V = number of distinct words (or lemmas) in the vocabulary.
- M = dimension of embedding space.
- N = number of MCMC iterations.
- $\rho = V \times M$ matrix of embedding vectors. This is learnt independently of the model and treated as constant.

Deterministic variables

- τ_d = time label for snippet d.
- γ_d = genre label for snippet d.
- o_d = observed true sense label for snippet d. This is not known in general, and is only available for annotated snippets used in testing.
- $D_{g,t}$ = number of snippets for genre(s) g and time(s) t.
- L_d = number of words in snippet d, after filtering stopwords and very low frequency (uninformative) words, treated as deterministic given the data W. The context positions occupied by these words is denoted $\{i_1, \ldots, i_{L_d}\}$, a subset of $\{1, \ldots, L\}$.
- ρ_v = embedding vector in the $\mathbb{R}^M$ space for word v. This is regarded as deterministic since it is learnt independently of the models and treated as fixed for each v.

Random variables

- z_d = sense assignment for snippet d.
- W = set of all snippets.
- W_d = set of words used in snippet d.
- $w_{d,i}$ = word occupying the i^{th} context position in snippet d.

Note that the notation $W, W_d, w_{d,i}$ is used for both random variables and their observed quantities. The context should make it clear which use is intended.

Model parameters

- $\phi = K \times G \times T$ dimensional real array used for modelling sense prevalence.
- $\tilde{\phi} = K \times G \times T$ dimensional array of sense prevalence probabilities.
- $\phi^{g,t} = \left(\phi_1^{g,t}, \ldots, \phi_K^{g,t}\right) = K$-dimensional real vector used for modelling sense prevalence for genre g and time t.
- $\tilde{\phi}^{g,t} = \left(\sum_{k=1}^{K} \exp(\phi_k^{g,t})\right)^{-1} \exp(\phi^{g,t}) = K$-dimensional vector of sense prevalence probabilities for genre g and time t.

- $\psi = V \times K \times T$ dimensional real array used for modelling context-word probabilities.
- $\tilde{\psi} = V \times K \times T$ dimensional array of context-word probabilities (i.e. the senses).
- $\psi^{k,t} = \left(\psi_1^{k,t}, \ldots, \psi_V^{k,t}\right) = V$-dimensional real vector used for modelling context-word probabilities under sense k at time t.
- $\tilde{\psi}^{k,t} = \left(\sum_{v=1}^{V} \exp(\psi_v^{k,t})\right)^{-1} \exp(\psi^{k,t}) = V$-dimensional vector of context-word probabilities under sense k at time t.
- $\chi^k = V$-dimensional real vector (in DiSC) or M-dimensional sense embedding (in EDiSC) used to model variation over senses $k \in \{1, \ldots, K\}$.
- $\theta^t = V$-dimensional real vector (in DiSC) or M-dimensional time embedding (in EDiSC) used to model variation over time $t \in \{1, \ldots, T\}$.
- $\xi^{k,t} = \chi^k + \theta^t = M$-dimensional sense-time embedding in EDiSC.
- $\varsigma = V$-dimensional real vector serving as a 'correction' or 'bias' parameter in EDiSC with $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$.

Hyperparameters

- κ_ϕ = variance hyperparameter for ϕ. This is modelled as an Inv Gamma(a, b) variable in GASC, and as scalar constant in DiSC/EDiSC.
- κ_ψ = scalar variance hyperparameter for ψ in GASC.
- κ_χ = scalar variance hyperparameter for χ in DiSC/EDiSC.
- κ_θ = scalar variance hyperparameter for θ in DiSC/EDiSC.
- κ_ς = scalar variance hyperparameter for ς in EDiSC.
- α_ϕ = AR(1) hyperparameter for $\phi^{g,t}$ in DiSC/EDiSC.
- α_θ = AR(1) hyperparameter for θ^t in DiSC/EDiSC.
- λ = inverse temperature or learning rate.

Indices and counts

- $z = z_{1:D} = (z_1, \ldots, z_D)$ and $\phi^{g,t} = \phi_{1:K}^{g,t} = \left(\phi_1^{g,t}, \ldots, \phi_K^{g,t}\right)$ etc.
- $\{i_1, \ldots, i_{L_d}\}$ = context positions occupied by the L_d words in snippet d.
- $\mathcal{D}(g,t) = \{d : \gamma_d \in g \text{ and } \tau_d \in t\}$ = set of snippet indices for genre(s) g and time(s) t.
- $N_{k,g,t}^z = \sum_{d \in \mathcal{D}(g,t)} \mathbb{I}(z_d = k)$ = count of snippets with sense-genre-time equal (k, g, t) under sense assignments z.
- $N_{v,k,t}^{W,z} = \sum_{d \in \mathcal{D}(\cdot,t)} \mathbb{I}(z_d = k) \sum_{i=i_1}^{i_{L_d}} \mathbb{I}(w_{d,i} = v)$ = count of context word v in snippets W with sense k at time t under assignments z.

C Acronyms

All acronyms used in this thesis are listed alphabetically below for easy reference.

- ADVI: Automatic Differentiation Variational Inference
- AI: Artificial Intelligence
- AIC: Akaike Information Criterion
- AR: Autoregressive
- BERT: Bidirectional Encoder Representations from Transformers
- BIC: Bayesian Information Criterion
- BS: Brier Score
- CBOW: Continuous Bag of Words
- CCOHA: Clean Corpus of Historical American English
- COHA: Corpus of Historical American English
- D-ETM: Dynamic Embedded Topic Model
- DGP: Data Generating Process
- DiSC: Diachronic Sense Change
- DTM: Dynamic Topic Model
- EDiSC: Embedded Diachronic Sense Change
- ELBO: Evidence Lower Bound
- ELMo: Embeddings from Language Model
- ELPD: Expected Log Pointwise Predictive Density
- ESS: Effective Sample Size
- ETM: Embedded Topic Model
- GASC: Genre-Aware Semantic Change
- GBI: Generalised Bayesian Inference
- GloVe: Global Vectors (for word representation)
- GPC: Generalised Posterior Calibration (algorithm)
- GPT: Generative Pre-trained Transformer
- HMC: Hamiltonian Monte Carlo
- HPD: Highest Posterior Density
- KL: Kullback-Leibler
- LDA: Latent Dirichlet Allocation
- LDMI: Loss Driven Multi-sense Identification
- LLM: Large Language Model
- LOOCV: Leave-One-Out Cross-Validation
- LPD: Log Pointwise Predictive Density

- MALA: Metropolis-Adjusted Langevin Algorithm
- MCMC: Markov Chain Monte Carlo
- NLP: Natural Language Processing
- NUTS: No-U-Turn Sampler
- PPC: Posterior Predictive Checks
- SCAN: (dynamic Bayesian model of) Sense ChANge
- WAIC: Widely Applicable Information Criterion
- WSD: Word Sense Disambiguation

Chapter 2

Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference

Schyan Zafar and **Geoff K. Nicholls**

Abstract

In a bag-of-words model, the *senses* of a word with multiple meanings, e.g. "bank" (used either in a river-bank or an institution sense), are represented as probability distributions over context words, and sense prevalence is represented as a probability distribution over senses. Both of these may change with time. Modelling and measuring this kind of sense change is challenging due to the typically high-dimensional parameter space and sparse datasets. A recently published corpus of ancient Greek texts contains expert-annotated sense labels for selected target words. Automatic sense-annotation for the word "kosmos" (meaning decoration, order or world) has been used as a test case in recent work with related generative models and Monte Carlo methods. We adapt an existing generative sense change model to develop a simpler model for the main effects of sense and time, and give MCMC methods for Bayesian inference on all these models that are more efficient than existing methods. We carry out automatic sense-annotation of snippets containing "kosmos" using our model, and measure the time-evolution of its three senses and their prevalence. As far as we are aware, ours is the first analysis of this data, within the class of generative models we consider, that quantifies uncertainty and returns credible sets for evolving sense prevalence in good agreement with those given by expert annotation.

Keywords: Bayesian inference; diachronic lexical semantics; Markov Chain Monte Carlo methods; natural language processing (NLP); sense change model

1 Introduction

As a natural language evolves, the meanings of words within the language change. The field of diachronic lexical semantics is concerned with how word meanings change over time. Words with multiple meanings or *senses*, and their time-evolution, are of considerable interest within the field. Examples of such words include "mouse" (meaning a rodent or a computer pointing device) and "bank" (meaning a river-bank or a financial institution). Statistical models of diachronic sense change are useful for lexicographic and linguistic research as well as for downstream applications in many natural language processing (NLP) tasks such as information retrieval.

For a word with multiple senses, the intended sense is usually apparent from the context. For example, the different senses of "bank" are obvious in the text snippets "deposited £500 in his bank account" and "plants growing on the bank of the Indus river". We expect certain context words to be used more often than others depending on the intended sense of the target word. In the "bank" example for instance, context words such as "money" or "account" are more likely when "bank" is used in the financial institution sense, whereas context words such as "river" or "stream" are more likely in the river-bank sense. We may also expect changes in the relative frequency of context words over time. E.g. for the financial institution sense of "bank", the context word "specie" (coin) is more likely to be used up to the early 20^{th} century whereas the context word "card" is more likely to be used later on. The prevalence of a sense itself may change over time, e.g. the pointing device sense of "mouse" increased in prevalence over the later half of the 20^{th} century. We are interested in a model that captures all of these features.

In this paper we analyse (as a test case) the sense change for the ancient Greek word "kosmos", meaning decoration, order or world, and quantify the uncertainty in these sense change estimates. Obtaining usefully narrow credible intervals with good coverage is no easy task and, viewed from a statistical perspective, this is our main contribution to the field. We achieve this by careful statistical modelling of the data. We develop a new model of Diachronic Sense Change (DiSC) by adapting the Sense ChANge (SCAN) and Genre-Aware Semantic Change (GASC) models of Frermann and Lapata (2016) and Perrone et al. (2019) respectively. Under this modelling framework, target word senses are represented as probability distributions over context words, sense prevalence is represented as a probability distribution over senses, and both sense and sense prevalence have temporal dependence. Our DiSC model has significantly fewer parameters than SCAN/GASC. However, there is no evidence for any loss of goodness-of-fit, and our model is significantly easier to analyse.

Both aspects are important when the word of interest occurs infrequently in a very large fixed corpus of surrounding text as is the case for "kosmos" in the ancient Greek data. We found that we could only give a reliable and well-calibrated fit for SCAN/GASC when the data were strongly informative of the parameters. We could not find any variant of MCMC that could reliably fit SCAN/GASC to the "kosmos" data even for point estimation of parameters, let alone uncertainty quantification. On the other hand, we were able to fit DiSC to the data, quantify uncertainty, and get a well-calibrated match to the expert sense-annotation. Moreover, in experiments on synthetic data generated according to the SCAN model itself, DiSC scores at least as well as SCAN on sense-labelling, as measured by Brier scores. We attribute this to a favourable bias-variance tradeoff.

These models are related to topic models, though there is no direct correspondence. Variational methods are widely used to fit topic models and can be efficient for inferring posterior means. However, variational methods are less reliable for quantifying uncertainty, and in particular tend to underestimate variance. Markov Chain Monte Carlo (MCMC) methods are particularly challenging for the models under discussion. However, where MCMC methods can be shown to converge, they at least give asymptotically exact posterior summaries. We present a relatively efficient method for fitting these models in such cases, where we marginalise the posterior by summing over the discrete sense labels, and use gradient-based MCMC schemes to target the remaining continuous distributions. We use the occurrences of "bank" in an English text corpus, as a simple illustrative example where all analyses are possible, to compare our sampler against existing MCMC samplers for these models. We compare the models' predictive performance on held-out sense labels for "bank" and for synthetic datasets. We then analyse "kosmos" using our new model and MCMC sampler.

The rest of this paper is structured as follows. In Section 2 we review the existing approaches for modelling diachronic semantic change and parameter inference within the NLP literature. In Section 3 we describe the "bank" and "kosmos" datasets, respectively from the English and ancient Greek corpora. In Section 4 we present our new DiSC model and highlight the differences with the existing SCAN and GASC models. In Section 5 we describe the existing MCMC samplers targeting the posterior, and present our MCMC scheme. Section 6 compares the performance of the MCMC samplers on the SCAN model for "bank". Sections 7–9 look at the application of the models to the English, synthetic and ancient Greek datasets. Section 10 concludes with a discussion of limitations and possible future research. Some further technical details and results are given in the appendices, and R scripts for data extraction and model-fitting are uploaded to GitHub.

2 Related work

The problem of modelling diachronic semantic change has been approached in several different ways within the field of NLP, and detailed overviews of the literature are given by Tahmasebi et al. (2018) and Tang (2018). Broadly, the approaches can be categorised into three groups: topic-based models, graph-based models and embedding models.

The topic model first introduced by Blei et al. (2003) is a generative model called Latent Dirichlet Allocation (LDA). Under LDA, a document of given length is generated by sampling a topic, and then a word given the topic, at each position in the document. LDA is a simple bag-of-words model that captures some basic ideas of meaning via the word-topic associations. The model uses Dirichlet priors for probability distributions over topics and words, and variational inference is commonly used for parameter estimation. Griffiths and Steyvers (2004) give a collapsed Gibbs sampler targeting the marginal posterior integrated over the continuous parameters, which can be used for asymptotically exact inference.

LDA was extended by Blei and Lafferty (2006) to give a dynamic topic model which additionally captures the time-evolution of topics. The dynamic topic model uses logistic normal priors since these are straightforward to model as a time series, as opposed to the parameters of the Dirichlet priors under LDA. Variational inference is used for the parameters, but alternative sampling methods have been proposed in the literature. These include the blocked Gibbs sampler based on auxiliary uniform variables given by Mimno et al. (2008) and a strategy based on auxiliary Polya-Gamma (Polson et al., 2012) variables given by Chen et al. (2013).

The dynamic topic model was adapted by Frermann and Lapata (2016) to explicitly capture the meanings or senses of a given target word (as opposed to topics in documents) and their time-evolution in the SCAN model. The main distinction between the two is that a topic model has an independent topic underlying each context word, whereas in SCAN all context words for a single usage of the target word share the same sense. The logistic normal priors, defined separately in each time period, are connected to their temporal neighbours via an intrinsic Gaussian Markov Random Field (Rue and Held, 2005), which enables modelling the change in adjacent parameters without requiring a global mean. GASC, an extension to SCAN given by Perrone et al. (2019), additionally allows the prevalence of each sense to vary according to the genre of the text in which the target word is used. Both authors use the Gibbs sampler of Mimno et al. (2008) for inferring the model parameters.

A distinct graph-based approach to this problem is given by Mitra et al. (2014, 2015) who use a semantic network model in which words are represented as nodes, and edges between nodes denote word co-occurrence in a sentence. The senses of a word are clustered separately for two different time periods, and then compared across the time periods to identify sense changes as well as sense births, deaths, mergers and splits. A similar approach is used by Tahmasebi and Risse (2017) who cluster senses separately for each time period and track the clusters through time.

Word embeddings are techniques for mapping words onto low-dimensional real vector spaces. In recent years, the neural-network-based Word2vec models developed at Google by Mikolov et al. (2013a) have emerged as the most popular word embedding models, although Stanford's GloVe model developed by Pennington et al. (2014) is a popular alternative that combines the advantages of global co-occurrence matrix factorisation methods with local context window methods. Skip-gram is the more popular of the two Word2vec models (the other being CBOW) and has been used to capture many semantic word relationships (Mikolov et al., 2013b) and linguistic regularities (Mikolov et al., 2013c), but the original model only uses one vector representation for each word and hence does not allow for multiple senses. The original Skip-gram has been extended in several ways to capture multiple senses per word, e.g. the Adaptive Skip-gram model given by Bartunov et al. (2016) and the loss driven multi-sense identification (LDMI) model given by Manchanda and Karypis (2019). A comprehensive review of embedding techniques used for word sense representation is given by Camacho-Collados and Pilehvar (2018).

Models based on word embeddings have been used to track semantic changes over time. These models usually construct the embeddings separately in each time period and then align the vectors across time, e.g. as done by Hamilton et al. (2016) and Kulkarni et al. (2015). Alternative approaches are given by Dubossarsky et al. (2019) who use temporal referencing instead of alignment, and by Rudolph and Blei (2018) who use dynamic embeddings (based on the exponential family embeddings of Rudolph et al. 2016) where word embeddings are set in a probabilistic framework. A different dynamic embedding model (called dynamic Skip-gram) is given by Bamler and Mandt (2017) who use a Kalman filter prior to connect embeddings across time periods.

It is not straightforward to ascertain which of the three approaches, if any, is the best. Whilst word embedding models appear to be the most popular category for semantic modelling (Kutuzov et al., 2018), these models tend to admit either multiple word senses or multiple time periods but not both simultaneously. To the best of our knowledge, there

is currently no embedding model that allows multiple word senses to be modelled consistently across time. The dynamic embedded topic model (Dieng et al., 2019) tracks document topics, but not word senses, across time using word embeddings; so although it is not in substance a model for sense change, it is a step in this direction. Moreover, the embedding and graph-based models are not stochastic-process-based generative models, and this limits their interpretability and Bayesian measures of uncertainty. In contrast, the topic-based SCAN and GASC models are generative, admitting both multiple word senses and multiple time periods, and the model parameters have simple physical interpretations. The main drawback of SCAN and GASC is that they over-parameterise when the interaction between sense and time is weak or weakly evidenced by the data. They are particularly difficult to fit on sparse and noisy data, which are common, and where the parameterisation leads to ridge structures and multi-modality in the posterior. Our model, with an additive effect of sense and time, offers a lower-dimensional alternative.

Quantification of uncertainty in sense change estimates is rare in the field of semantic change detection, and indeed has not been attempted by the authors of SCAN and GASC whose work we build upon. The participating models in the recent SemEval shared-task on semantic change detection (Schlechtweg et al., 2020) give a snapshot of current practice. Few models, if any, attempted to quantify uncertainty. It was not a SemEval assessment criterion, as is typical in this literature.

3 Data

Consider the three text snippets containing the word "bank" in Table 1, where "bank" is used in the sense of a river-bank in the first example, and in the sense of a financial institution in the other two. The first two snippets, written in the time period 1990-2010, are taken from the non-fiction genre, whereas the third snippet, written in the time period 1830-1850, is taken from the magazine genre. The snippets are of equal length with 7 words on either side of "bank". The words in blue are stopwords, i.e. the most common words in the language, which generally do not contribute to the meaning of the target word if we ignore syntax. The words in orange on the other hand appear with a low frequency (called "hapaxes" if they appear exactly once), and are uninformative in the context of the models we consider.

For a given target word, we define the data W as a collection of D snippets with a symmetric context window of $L/2$ words on either side of the target word (ignoring sentence and paragraph boundaries), with the stopwords, uninformative words and punctuation removed,

Table 1: Example text snippets for target word "bank"

" . . . China. The Yellow River had burst its banks, submerging vast areas of farmland, washing away . . . " — "1421: the year China discovered America" (2003) – non-fic – Menzies, Gavin
" . . . to examine whether institutions like the World Bank and the International Monetary Fund needed restructuring . . . " — "The price of loyalty: George W. Bush, the White House, and the education of Paul O'Neill" (2004) – non-fic – Suskind, Ron
" . . . subject of continuing specie payments. Though the Bank of the United States had previously determined . . . " — "Philadelphia Banking" (1839) – mag – US Democratic Review: Nov 1839

and the words lemmatised (i.e. replaced with root words such as "wash" instead of "washing" in the first example). The snippets span multiple discrete and contiguous time periods, and may be taken from any number of text genres. Our model, described in the next section, could be applied to any dataset with these features. We analyse two real datasets in this paper: the context data for target word "bank" extracted from the Corpus of Historical American English (COHA) published by Davies (2010), which is used as an illustrative example, and the context data for target word "kosmos" extracted from the Diorisis Ancient Greek Corpus published by Vatri and McGillivray (2018), which is the focus of this work. We additionally use synthetic data to compare the models' predictive performance on held-out true sense labels.

The "bank" example was used by Frermann and Lapata (2016), and we use it as an illustrative example due to its relative simplicity since the two main senses of "bank" are very distinct. There are c. 74,000 instances of "bank" in COHA covering the years 1810-2009, which we divide into ten 20-year contiguous blocks, and spanning four genres (fiction, non-fiction, news and magazine). We extract snippets of length $L = 14$ words (not counting the target word) around these instances. The snippet length has to strike a balance between including meaningful nearby words and not including noise from distant words, and we found the length $L = 14$ to be sufficient for a human to identify the sense in most cases. For computational ease, we randomly subsample a maximum 100 snippets per genre-time block, giving us 3,685 selected snippets and c. 70,000 non-selected snippets. We manually tag 3,525 of the selected snippets with the correct sense of "bank", grouping together the related meanings of river-bank, edge, tilt or heap, and the related meanings of a financial

Table 2: Frequencies of "bank" in each sense-time block across all genres

Sense	Start of 20-year period									
	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
River-bank	193	95	135	126	175	123	86	89	65	73
Institution	89	189	201	261	199	258	298	275	307	288

Table 3: Frequencies of "kosmos" in each sense-genre-time block

Genre	Sense	Century								
		–7	–6	–5	–4	–3	–2	–1	1	2
Narrative	Decoration	0	0	5	13	10	0	45	123	27
	Order	0	0	10	14	9	0	68	57	17
	World	0	0	0	0	3	0	31	20	2
Non-narrative	Decoration	2	0	9	51	4	37	33	43	2
	Order	1	0	1	29	6	1	2	25	3
	World	0	0	0	52	3	26	15	303	35

institution or a store (e.g. blood bank). The remaining snippets were either ambiguous or used "bank" as a proper noun (e.g. Mr Banks), or a very small number of other senses. We identify and remove stopwords using the R package `Stopwords` (Benoit et al., 2020) as well as part-of-speech tags, marking anything other than nouns, adjectives, verbs and adverbs as stopwords. We further restrict the data to the top 70% most frequently occurring words in the non-selected snippets. This is done for efficiency reasons, and seems to incur little loss of information. We refer to the 30% of words omitted as "uninformative" words: they occur infrequently in the target context. The observation model for context words in the selected snippets is not affected by this registration, as uninformative words are defined by their frequency in non-selected snippets. Using this registration criterion, we have a vocabulary of 973 words.

The "kosmos" example was used by Perrone et al. (2019), and we use it as our test case because, in contrast to "bank", we found it particularly challenging to analyse using existing models and tools. "Kosmos" can be used in one of three senses, i.e. decoration, order or world, and expert sense-annotation is provided by Vatri et al. (2019). We use snippets of length $L = 14$ as before and, following Perrone et al. (2019) for testing purposes, retain only the "collocates", i.e. all snippet instances where an ancient Greek expert was able to identify the sense based on contextual information alone. A "real-use" analysis including non-collocates is given Appendix F. We group the text genres into "narrative" and "non-narrative", and divide time into 9 contiguous centuries from 700 BC to AD 200. We remove

stopwords in the same way as for "bank" plus the additional stopwords identified by Berra (2018). Hapaxes (i.e. words that appear only once in the selected snippets) are removed, approximating the observation model as we do not condition on the event that a context word appears at least twice over all selected snippets. This is in contrast to the registration process for "bank" where we had a large set of non-selected snippets. This leaves us with 1,144 snippets and a vocabulary of 968 words appearing in snippets associated with "kosmos". Tables 2–3 show that the "kosmos" data is a lot more sparse and fragmented than the "bank" data, especially for the early time periods.

4 Prior and observation models

In this section we introduce our new DiSC model and highlight how it improves upon the existing GASC model, noting that GASC is the same as SCAN except that it allows the sense prevalence to vary according to the text genre in addition to time. DiSC, given in Algorithm 1, is a generative model of how the context words in the snippets around a given target word are emitted from a latent stochastic process. We make the simplifying assumption commonly used in NLP that each snippet is a "bag-of-words" (i.e. word order and grammar are ignored) of length L, not counting the target word itself. Stopwords and uninformative words are generated in any snippet d with probabilities q^{SW} and q^{U} respectively, so the number of context words that are neither stopwords nor uninformative, denoted L_d, has a binomial distribution $L_d|L, q^{\mathrm{SW}}, q^{\mathrm{U}} \sim \mathrm{Bin}(L, 1-q^{\mathrm{SW}}-q^{\mathrm{U}})$. The context positions occupied by these words are a random subset $\{i_1, \ldots, i_{L_d}\}$ of size L_d drawn from $\{1, \ldots, L\}$, i.e. the order of words is irrelevant as per the bag-of-words assumption. Once we condition on the L_d values, the likelihood (which we write down in Section 5 below) does not depend on q^{SW} and q^{U}. We can therefore drop q^{SW} and q^{U} from all analysis hereafter.

The snippets span T discrete and contiguous time periods and G text genres, so we have deterministic mappings between each snippet d and its time and genre labels $\tau_d \in \{1, \ldots, T\}$ and $\gamma_d \in \{1, \ldots, G\}$ respectively. The target word can be used in one of K senses in any snippet d, so the single sense assignment z_d for the whole snippet is emitted as a draw from a multinomial distribution over the senses $\{1, \ldots, K\}$ parameterised by the K-dimensional probability vector $\tilde{\phi}^{\gamma_d,\tau_d}$, that is $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \mathrm{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d,\tau_d}\right)$. The vocabulary consists of V words so, given the sense assignment z_d, for each context position $i \in \{i_1, \ldots, i_{L_d}\}$ the context word $w_{d,i}$ is emitted as a draw from a multinomial distribution over the words $\{1, \ldots, V\}$ parameterised by the V-dimensional sense-dependent probability vector $\tilde{\psi}^{z_d,\tau_d}$, that is $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \mathrm{Mult}\left(\tilde{\psi}_1^{z_d,\tau_d}, \ldots, \tilde{\psi}_V^{z_d,\tau_d}\right)$. At this level DiSC is the same as GASC.

In contrast, the dynamic topic model would have $z_{d,i}|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d},\ldots,\tilde{\phi}_K^{\gamma_d,\tau_d}\right)$ independently for each context position i. If we modify this and impose a single topic per document, it becomes precisely the GASC model with $G = 1$; so comparison of DiSC with the dynamic topic model itself is not appropriate.

The probability vectors $\tilde{\phi}^{g,t}$ and $\tilde{\psi}^{k,t}$ are softmax transforms of the real-valued sense prevalence parameter vector $\phi^{g,t}$ and word parameter vector $\psi^{k,t}$ respectively, i.e.

$$\tilde{\phi}^{g,t} = \frac{\exp(\phi^{g,t})}{\sum_{k=1}^{K}\exp(\phi_k^{g,t})} \qquad \text{and} \qquad \tilde{\psi}^{k,t} = \frac{\exp(\psi^{k,t})}{\sum_{v=1}^{V}\exp(\psi_v^{k,t})}. \tag{1}$$

Here $\tilde{\phi}, \phi$ are $K \times G \times T$ dimensional arrays and $\tilde{\psi}, \psi$ are $V \times K \times T$ dimensional arrays. The latent variables ϕ and ψ are not identifiable in this setup, but this is not an issue since these are only used to determine the priors for the probability arrays $\tilde{\phi}$ and $\tilde{\psi}$ which are the variables of interest. The word parameter vector $\psi^{k,t}$, which depends on both the sense k and the time t, is defined as the sum of a word-sense parameter vector χ^k and a word-time parameter vector θ^t. We place independent autoregressive AR(1) priors on the elements of $\phi^{g,t}$ and θ^t, and independent normal priors on the elements of χ^k, as defined in Algorithm 1 lines 2-15. The prior hyperparameters are kept fixed.

The GASC generative model is given in Algorithm 2 in Appendix A. Our DiSC model differs from GASC in three main ways. The fundamental difference is that we assume the effects due to sense and time are additive: the 3-dimensional $V \times K \times T$ array ψ in GASC is replaced by two 2-dimensional arrays in DiSC, i.e. a $V \times K$ array χ and a $V \times T$ array θ, so that we have $\psi^{k,t} = \chi^k + \theta^t$ for $k \in \{1,\ldots,K\}$ and $t \in \{1,\ldots,T\}$. This reduces the dimension of ψ from VKT parameters in GASC to $V(K+T)$ parameters in DiSC. The implications of this for sense change measurements are discussed in Section 4.1.

The second difference is in the modelling of the time series variables $\phi^{g,t}, \theta^t$ in DiSC and $\phi^{g,t}, \psi^{k,t}$ in GASC. Our priors on the time series are autoregressive AR(1) processes with proper stationary distributions, whereas GASC uses improper priors without global means or stationary distributions. For example, the prior distribution of $\phi_k^{g,t}$ in GASC is $\phi_k^{g,t}|\phi_k^{g,t-1}, \kappa_\phi \sim \mathcal{N}\left(\phi_k^{g,t-1}, 2\kappa_\phi\right)$ for $t \in \{2,\ldots,T\}$, and an improper uniform distribution over all real numbers for $t = 1$. It is thus possible for the posterior $\phi_k^{g,t}|W$ to drift off to $\pm\infty$ since there is no global mean to tether the distribution. In contrast, the DiSC priors have a stationary distribution $\phi_k^{g,t}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)$ for all t. We expect this to lead to more homogeneous behaviour at the beginning and end of the time series.

Algorithm 1 DiSC: generative model

PRIOR MODEL

fix hyperparameters $\kappa_\phi, \kappa_\theta, \kappa_\chi, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
initialise at time $t = 1$
for genre $g \in 1 : G$ **do**
    draw sense prevalence parameter $\phi^{g,1} | \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
end for
draw word-time parameter $\theta^1 | \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$
for time $t \in 2 : T$ **do**
    for genre $g \in 1 : G$ **do**
        draw sense prevalence parameter $\phi^{g,t} | \phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \text{diag}(\kappa_\phi)\right)$
    end for
    draw word-time parameter $\theta^t | \theta^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \theta^{t-1}, \text{diag}(\kappa_\theta)\right)$
end for
for sense $k \in 1 : K$ **do**
    draw word-sense parameter $\chi^k | \kappa_\chi \sim \mathcal{N}\left(0, \text{diag}(\kappa_\chi)\right)$
end for
for sense $k \in 1 : K$ and time $t \in 1 : T$ **do**
    set word parameter $\psi^{k,t} = \chi^k + \theta^t$
end for
using softmax (1), transform real arrays ϕ and ψ into probability arrays $\tilde{\phi}$ and $\tilde{\psi}$

OBSERVATION MODEL

fix probabilities of drawing stopwords q^{SW} and uninformative words q^{U}
for snippet $d \in 1 : D$ **do**
    draw number of context words $L_d | L, q^{\text{SW}}, q^{\text{U}} \sim \text{Bin}(L, 1 - q^{\text{SW}} - q^{\text{U}})$
    draw a random subset $\{i_1, \ldots, i_{L_d}\}$ of size L_d from $\{1, \ldots, L\}$
    draw sense assignment $z_d | \tilde{\phi}^{\gamma_d, \tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d, \tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d, \tau_d}\right)$
    for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
        draw context word $w_{d,i} | z_d, \tilde{\psi}^{z_d, \tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d, \tau_d}, \ldots, \tilde{\psi}_V^{z_d, \tau_d}\right)$
    end for
end for

The third difference is that whilst we treat the prior hyperparameter κ_ϕ as fixed, GASC uses a random hyperprior $\kappa_\phi \sim \text{Inv Gamma}(a, b)$ with a, b fixed. However, the data do not inform this parameter at all well: κ_ϕ is conditionally independent of the data W given ϕ, we have a relatively small number of ϕ parameters, and these are in turn conditionally independent of the data W given the unknown sense labels z. In contrast, physical considerations lead to an informative prior for κ_ϕ. The joint posterior (2) below is insensitive to the choice of κ_ϕ for any plausible value of this parameter, so we can fix κ_ϕ without loss using the prior elicitation described in Section 4.2.

4.1 Sense representation and sense change

Following the framework of Frermann and Lapata (2016) for these bag-of-words models, we define sense k of the target word at time t as the distribution $(\tilde{\psi}_1^{k,t}, \ldots, \tilde{\psi}_V^{k,t})$ over words $\{1, \ldots, V\}$ appearing in the context of the target word. "Sense change" may therefore be defined as the evolution of the $V \times K$ matrix $\tilde{\psi}^{\cdot,t}$ over $t \in \{1, \ldots, T\}$. Similarly, "sense prevalence change" may be defined as the evolution of the $K \times G$ matrix $\tilde{\phi}^{\cdot,t}$ over $t \in \{1, \ldots, T\}$. We loosely refer to both evolutions as "sense change" for brevity. This definition of sense change does not correspond to any sort of real meaning change (which we do not define, but we have in mind something like changes in the dictionary definition of the target word), which is a limitation of all bag-of-words models.

We parameterise DiSC to model key structural drivers of sense difference (over senses) and sense change (over time). The additive main effects in DiSC (χ and θ respectively) correspond to these drivers. This allows us to model the main effects, and the data to inform them, in a direct way. In contrast, GASC models the effects and their interaction in a general way without an *explicit* parameterisation or modelling of the main effects. Useful structural information is left out of the GASC prior, as the model imposes no core additive structure. We describe below the main data features that we see as informing sense difference and sense change.

The probability $\tilde{\psi}_v^{k,t}$ of context word $v \in \{1, \ldots, V\}$ being used in a snippet can change when the background usage frequency of word v changes in the text corpus taken as a whole. For example, the background usage frequency of the word "telephone" likely increases over the 20^{th} century. It is then more likely to appear in any context, across all target word senses. This is, in our definition, a form of sense change. We expect this time-effect, captured by θ^t in DiSC, to be a common driver of sense change in these models. The time-effect is not explicitly modelled in GASC.

Similarly, we expect certain words to appear more frequently in the context of particular senses of the target word regardless of their background usage frequency. Examples include "water" and "money" for the two senses of "bank" respectively. This relatively strong sense-effect informs sense difference for the target word, and is basic to the bag-of-words setup. DiSC, in contrast to GASC, models this sense-effect explicitly through χ^k.

On the other hand, we might expect certain words to change in frequency at different rates in the context of different senses of the target word. Such changes might be driven by actual changes in the meaning of the target word. For example, "telephone" might increase

in frequency more rapidly in the financial institution sense of "bank", as the sense changes to reflect more modern ways of banking, than it does in the river-bank sense. This sense-time interaction effect is captured by GASC but not by DiSC. However, GASC does not distinguish between the main and interaction effects, so if we are interested in isolating this behaviour then it is necessary to remove the main effects. We illustrate this in our analysis of the "bank" data in Section 7 below.

Goodness-of-fit is generally retained in DiSC even in the presence of a real interaction. Words contributing to poor fit due to missed interactions are words that *both* appear frequently in the context of multiple target word senses *and* evolve in frequency at different rates, since these have both large percentage change and large value. Large percentage errors in context probability for very small context probabilities are of little consequence. This feature of the data makes DiSC relatively robust to interaction in practice, as demonstrated in the synthetic data experiments in Section 8 below.

The additive structure in DiSC helpfully prevents switching of sense labels between time periods. In MCMC targeting GASC, the senses can settle into different sense-label permutations in different time periods, as ψ-values are only loosely connected temporally via their priors. This is a particular problem when some context words appear frequently across more than one sense. In such cases, the MCMC may get stuck in a local mode where the sense labels are not aligned across time. This is no criticism of GASC, but adds to the challenge of fitting it to data.

One basic and important kind of sense change, i.e. the emergence of a new sense in addition to the existing senses, is captured in both models if K is large enough to accommodate the new sense. The new sense k exists at all times, but its prevalence $\tilde{\phi}_k^{g,t}$ changes from being very small to significant as it emerges with increasing t.

4.2 Hyperparameter settings

We choose the number of senses K equal to the smallest value such that each sense $k \in \{1, \ldots, K\}$ in the model output, as identified by the most probable words under that sense, is distinct and meaningful in the judgement of the expert user. This is practical when DiSC is used as an exploratory tool to discover the lifespan and usage rate of distinct senses without the need for hand-labelling. Alternatively, the expert user may have hand-labelled a small set of snippets, in which case we would have prior knowledge of K. Setting K therefore requires a few trial runs. In our case, $K = 2$ for "bank" and $K = 3$ or 4 for "kosmos" depending on the task (see Section 9 and Appendix F). If the value of K is still

in question, model selection tools may be used, for example classic Bayesian tools such as Bayes factors (Kass and Raftery, 1995) or reversible jump MCMC (Green, 1995), or rather the state-of-the-art on these such as Xing (2021) or Karagiannis and Andrieu (2013) respectively. However, we have not investigated this.

Implementing SCAN/GASC requires setting the hyperparameter κ_ψ in $\psi_v^{k,t}|\psi_v^{k,-t} \sim \mathcal{N}\big(\frac{1}{2}(\psi_v^{k,t-1} + \psi_v^{k,t+1}), \kappa_\psi\big)$ and the hyperparameters a, b in $\kappa_\phi \sim \text{Inv}\,\text{Gamma}(a, b)$. Frermann and Lapata (2016) used the setting $\kappa_\psi = 0.1$, $a = 7$ and $b = 3$ for SCAN whereas Perrone et al. (2019) used $\kappa_\psi = 0.01$, $a = 1$ and $b = 1$ for GASC. We report results for the SCAN choice as these give the best performance for these models.

For the AR(1) processes in DiSC, we use the parameters $\alpha_\phi = \alpha_\theta = 0.9$ — a high value — in order to have weak mean reversion, so that we have a proper process prior without unduly influencing the posteriors.

To set κ_ϕ in DiSC, we elicit a prior by defining what we consider to be an extreme sense prevalence difference. The number of senses K is relatively small (compared to the vocabulary size V). Taking two fixed senses $l, m \in \{1, \ldots, K\}$, we allow a difference as large as $\tilde{\phi}_l^{g,t}/\tilde{\phi}_m^{g,t} \approx 100$ to be possible but extreme. This can easily be adjusted depending on the data-modelling context. On the logit-scale, we therefore assert that $\phi_l^{g,t} - \phi_m^{g,t} > \log 100$ is a 3-sigma event. From the prior stationary distribution in the AR(1) process, $\mathbb{V}(\phi_l^{g,t} - \phi_m^{g,t}) = \frac{2\kappa_\phi}{1-(\alpha_\phi)^2}$; so we express our preference with $3\left(\frac{2\kappa_\phi}{1-(\alpha_\phi)^2}\right)^{\frac{1}{2}} = \log 100$, giving $\kappa_\phi = 0.25$ on rounding. There is no simple comparison with κ_ϕ in SCAN/GASC due to the AR(1) structure in DiSC. If we replace the threshold for a rare event with $\tilde{\phi}_l^{g,t}/\tilde{\phi}_m^{g,t} \approx 10^6$ then we find $\kappa_\phi \approx 2$: the standard deviation changes over a small range from 0.5 to $\sqrt{2}$. Larger values are hard to justify.

The vocabulary size V is typically quite large (c. 1,000 in our examples), so we might expect greater variation in the probabilities for context words in a given sense, perhaps by a factor of 1,000. That is, for any fixed time t, sense k and pair of words $x, y \in \{1, \ldots, V\}$, the ratio of context word probabilities might be as large as $\tilde{\psi}_x^{k,t}/\tilde{\psi}_y^{k,t} \approx 1\,000$. Now $\mathbb{V}(\psi_x^{k,t} - \psi_y^{k,t}) = \mathbb{V}(\chi_x^k - \chi_y^k + \theta_x^t - \theta_y^t) = 2\kappa_\chi + \frac{2\kappa_\theta}{1-(\alpha_\theta)^2}$, so we express our preference with $3\left(2\kappa_\chi + \frac{2\kappa_\theta}{1-(\alpha_\theta)^2}\right)^{\frac{1}{2}} = \log 1\,000$. Attributing the variance in equal parts to χ and θ by setting $\kappa_\chi = \frac{\kappa_\theta}{1-(\alpha_\theta)^2}$, since they are additive effects on the same scale, we have $\kappa_\chi = 1.25$ and $\kappa_\theta = 0.25$ on rounding. As for κ_ϕ, this is robust to the choice of threshold due to the logarithm, so the posterior is relatively insensitive over values plausible *a priori*. Moreover, we have considered two fixed context words, not the most extreme pair, so more extreme variation is allowed.

5 Posterior distribution and MCMC inference

For convenience, we infer posterior distributions for the sense prevalence parameters ϕ, the word-time parameters θ and the word-sense parameters χ, but our interest is in the identifiable probability arrays $\tilde{\phi}$ and $\tilde{\psi}$ which are given as deterministic functions of these latent variables. We expect the sense assignment vector z to be of less interest, except in testing where we have hand-annotated meanings. The joint posterior for ϕ, θ, χ, z given the data W is defined by

$$\pi(\phi, \theta, \chi, z|W) \propto \pi(\phi)\pi(\theta)\pi(\chi)\pi(z|\phi)p(W|z, \theta, \chi) \tag{2}$$

$$= \pi(\phi)\pi(\theta)\pi(\chi) \prod_{d=1}^{D} \tilde{\phi}_{z_d}^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{z_d, \tau_d} \tag{3}$$

$$= \pi(\phi)\pi(\theta)\pi(\chi) \prod_{t=1}^{T} \prod_{k=1}^{K} \left(\prod_{g=1}^{G} (\tilde{\phi}_k^{g,t})^{N_{k,g,t}^z} \right) \left(\prod_{v=1}^{V} (\tilde{\psi}_v^{k,t})^{N_{v,k,t}^{W,z}} \right) \tag{4}$$

where $N_{k,g,t}^z = \sum_{\substack{d:\tau_d=t \\ \text{and } \gamma_d=g}} \mathbb{I}(z_d = k)$ is the number of snippets with sense assignment k and genre g at time t, and $N_{v,k,t}^{W,z} = \sum_{\substack{d:\tau_d=t \\ \text{and } z_d=k}} \sum_{i=i_1}^{i_{L_d}} \mathbb{I}(w_{d,i} = v)$ is the number of occurrences of context word v across all snippets with sense assignment k at time t. The conditional posterior for z is defined, independently for each snippet W_d, by

$$\pi(z_d|W_d, \phi, \psi) \propto \tilde{\phi}_{z_d}^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{z_d, \tau_d}, \tag{5}$$

which is a multinomial distribution over possible senses $\{1, \ldots, K\}$.

The authors of SCAN and GASC both use a blocked Gibbs strategy to alternately sample $z|W, \phi, \psi$ using (5) and $\phi|z$ and $\psi|z, W$. Under the SCAN and GASC models, each column of $\tilde{\phi}$ and $\tilde{\psi}$ has a logistic normal distribution which the authors target with a Gibbs sampler based on the auxiliary uniform variable method of Mimno et al. (2008). An alternative Gibbs sampler is based on auxiliary Polya-Gamma variables (Polson et al., 2012) and an approximate method based on the same is given by Chen et al. (2013). We describe these methods in Appendix B.

Under our DiSC model, since $\tilde{\psi}^{k,t}$ does not have a logistic normal distribution, the latter two methods cannot be used in a straightforward manner, although the auxiliary uniform variable method can be easily adapted. Moreover, both the auxiliary uniform and Polya-Gamma variable methods are very inefficient for these models, whereas the approximate method is obviously not asymptotically exact. We describe our MCMC sampler, which is

asymptotically exact and at least as efficient as the existing samplers, and may be used with any of these models.

We marginalise over the discrete z, removing the need to sample $z|W, \phi, \psi$, and alternate between sampling $\phi|W, \theta, \chi$ and $\theta|W, \phi, \chi$ and $\chi|W, \phi, \theta$ (or, in the case of SCAN/GASC, between $\phi|W, \psi$ and $\psi|W, \phi$). The marginal likelihood for $\phi, \theta, \chi|W$ is

$$p(W|\phi,\theta,\chi) = \prod_{d=1}^{D} p(W_d|\phi,\theta,\chi) \tag{6}$$

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d = k|\phi)p(W_d|z_d = k,\theta,\chi) \tag{7}$$

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d = k|\phi)\prod_{i=i_1}^{i_{L_d}} p(w_{d,i}|z_d = k,\psi) \tag{8}$$

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d} \tag{9}$$

where (6) exploits the conditional independence between the snippets, (7) comes from conditioning on (and summing over) the sense assignment z_d for snippet d, (8) exploits the conditional independence of the context words $w_{d,i}, i \in \{i_1, \ldots, i_{L_d}\}$ in snippet d, and (9) simply picks up the appropriate probabilities from the $\tilde{\phi}$ and $\tilde{\psi}$ arrays. The conditional posteriors for ϕ, θ and χ are therefore

$$\pi(\phi|W,\theta,\chi) \propto \pi(\phi)\prod_{d=1}^{D}\sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d}, \tag{10}$$

$$\pi(\theta|W,\phi,\chi) \propto \pi(\theta)\prod_{d=1}^{D}\sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d} \tag{11}$$

$$\text{and } \pi(\chi|W,\phi,\theta) \propto \pi(\chi)\prod_{d=1}^{D}\sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d} \tag{12}$$

respectively, which we can sample efficiently using gradient-based MCMC methods such as Metropolis-Adjusted Langevin Algorithm (MALA) and Hamiltonian Monte Carlo (HMC). We describe these methods, including the derivation of the gradient vectors and automatic parameter tuning, in Appendix C. At this level, the marginal posterior distribution in GASC is the same as DiSC (the essential difference being the parameterisation of ψ and the priors) and any sampler relevant for DiSC also applies for GASC. This marginal is often tractable for LDA but not used, as the collapsed Gibbs sampler obtained by integrating over the conjugate priors of the continuous parameters is favoured there. That is not possible here as the priors are not conjugate. In addition to these sampling methods, we implemented

a simple random-walk Metropolis sampler, both jointly and marginally over z. This was useful for code-checking but not competitive, and has been omitted.

For both datasets, we find that 20k MCMC iterations for SCAN/GASC with burn-in of 10k, and 10k MCMC iterations for DiSC with burn-in of 5k, are sufficient for convergence, where an iteration is one update over all parameters (except κ_ϕ for SCAN/GASC, which is updated every 50 iterations). Work to date on SCAN and GASC appears to take just 1k MCMC iterations in total, which we found was not nearly enough for convergence on these data. All figures and results reported in the following sections are based on posterior means unless otherwise indicated. The posterior is invariant under sense relabelling in all models, but this behaviour was not observed in any converging run.

6 Experiment 1: Finding the best sampler

We take as our test case for comparing the samplers the problem of fitting the SCAN model (i.e. GASC with one genre) for the target word "bank" in COHA. We are limited in these choices since, as we report in Section 9, no MCMC sampler we tried converged for the SCAN/GASC model on the "kosmos" analysis, and the Poly-Gamma samplers cannot be used with DiSC due to the additive form for ψ. The metric we use for this comparison is the effective sample size (ESS) per hour of CPU time after the burn-in. We implemented the samplers in the R programming language, as efficiently as we could, and using the same functions as far as possible. We checked that they converged to the same posterior distributions for all variables with high precision on small synthetic datasets. All runs are done sequentially on the same Linux PC.

For $\tilde{\phi}$, we are inferring KGT parameters and we take the median ESS per hour over all parameters. For $\tilde{\psi}$, we are inferring VKT parameters but our interest is mainly in the most representative words for each sense; so we take the median ESS per hour across the $\tilde{\psi}$ parameters for the 20 most probable words under each sense marginally over all time periods. The medians, together with the interquartile ranges, are shown in Table 4. Even allowing for uneven coding efficiency, it is clear that the differences are substantial.

The auxiliary uniform variable method (top row) used in the past to fit SCAN and GASC is the least efficient for $\tilde{\phi}$ by an order of magnitude or more, and the auxiliary Polya-Gamma variable method (second row) is by far the least efficient for $\tilde{\psi}$. The approximate Polya-Gamma auxiliary variable method (third row) is efficient, but not asymptotically exact, though fairly accurate in our experiments comparing against asymptotically exact samplers. Our MALA and HMC samplers targeting marginal posteriors are asymptotically

Table 4: Median (interquartile range) ESS per hour of CPU time

Sampling method	ESS for $\tilde{\phi}$	ESS for $\tilde{\psi}$
Aux uniform variable	40 (37 – 44)	419 (255 – 604)
Aux Polya-Gamma	81 (34 – 164)	80 (63 – 102)
Aux Polya-Gamma (approx)	1,557 (723 – 1,694)	671 (546 – 927)
MALA	1,477 (1,236 – 1,683)	382 (270 – 676)
HMC (5 leapfrog steps)	672 (587 – 993)	1,105 (762 – 1,651)
MALA+HMC (variable† steps)	1,613 (864 – 1,816)	1,312 (1,054 – 1,531)

† Randomly chooses 1 or 2 leapfrog steps for ϕ proposals and 1 or 5 leapfrog steps for ψ proposals at each update

exact methods of similar or better efficiency for $\tilde{\phi}$ and $\tilde{\psi}$ respectively, and are therefore preferred. MALA is comparable to a 1-step HMC sampler, so we may combine the strengths of MALA and HMC by randomly choosing either a 1-step or a multiple-step proposal in our HMC sampler at each update. This mixed MALA-HMC sampler (last row) is clearly the most efficient for both $\tilde{\phi}$ and $\tilde{\psi}$.

HMC, whether straight or mixed with MALA, has one additional parameter to tune, i.e. the number of leapfrog steps. Users may therefore prefer MALA for convenience, or mix MALA with the No-U-Turn HMC sampler of Hoffman and Gelman (2014) to avoid tuning the number of steps. Trace plots in Figure 5 in Appendix C illustrate the differences in mixing rates between the samplers for a prevalence parameter (where the difference is most visible) in runs of equal time.

7 Experiment 2: Analysis of sense change for "bank"

We now measure the time-evolution of the prevalence and context word composition of the different senses of a target word in a simple example. Our objective is achieved if the posteriors for $\tilde{\psi}^{k,t}$ for $k \in \{1, \ldots, K\}$ can be interpreted by a human as K unique target word senses, since these senses are automatically identified over time $t \in \{1, \ldots, T\}$ in the snippets. In the case of "bank", both DiSC and SCAN/GASC achieve this since the most probable words under an ordering based on the posterior expectation of time-averaged word probabilities $\frac{1}{T}\sum_{t=1}^{T} \tilde{\psi}^{k,t}$ from both DiSC and SCAN/GASC display the senses of river-bank and institution bank respectively:

```
k=1: river stream   water stand bank    leave  tree    creek  time reach
k=2: bank  national note  money deposit saving reserve credit loan issue
```

Table 5: Brier scores under different genre settings

G	Genre grouping	DiSC	GASC
1	all combined	0.152	0.183
2	fiction vs others	0.183	0.195
2	fic & non-fic vs news & mag	0.153	0.166
4	fic vs non-fic vs news vs mag	0.179	0.182

Table 6: "Bank" confusion matrix statistics

	Sensitivity (true +ve for river-bank sense)	Specificity (true +ve for institution sense)	Accuracy
DiSC	0.935	0.877	0.896
SCAN	0.926	0.874	0.891

The most probable words in each time period can easily be extracted if their evolution is of interest.

In order to assess the performance of the models, we consider their ability to recover the true sense labels $o_d, d \in \{1, \ldots, D\}$ for the $D = 3\,525$ snippets that we manually tagged (cf. Section 3). Let $\hat{p}(z_d = k)$ be the estimated value of $\mathbb{E}_{\phi,\psi|W}\big(p(z_d = k|W_d, \phi, \psi)\big)$ computed on the MCMC output. The Brier score, in our case defined by

$$BS = \frac{1}{D}\sum_{d=1}^{D}\sum_{k=1}^{K}\Big(\hat{p}(z_d = k) - \mathbb{I}(o_d = k)\Big)^2, \tag{13}$$

is a proper scoring rule for multi-category probabilistic predictions $\hat{p}(z_d = k)$, ranging from 0 (best) to 2 (worst), which we use as a criterion for model comparison. If we set $\hat{p}(z_d = k) = \frac{1}{K}$ for all d, k, we get $BS = \big(1 - \frac{1}{K}\big)^2 + (K - 1)\big(\frac{1}{K}\big)^2 = 0.5$ in the case of $K = 2$ senses for "bank", so a model must produce a score lower than this in order to be useful. Using the posterior mean probabilities $\hat{p}(z_d = k)$ obtained from normalising (5), we get the Brier scores shown in Table 5 from running DiSC and GASC under various genre configurations, where we fit both models using MALA. Both models therefore identify the true sense with a good level of accuracy. However, DiSC has slightly better performance in all cases.

We also examine confusion matrices from a simple classification task. We assign snippet d the sense $l = \arg\max_{k \in 1:K} \hat{p}(z_d = k)$. Table 6 shows the key statistics from the confusion matrices produced using this decision rule (treating river-bank as the positive and institution bank as the negative condition) suggesting that the DiSC model is marginally better than SCAN despite having about half as many parameters.

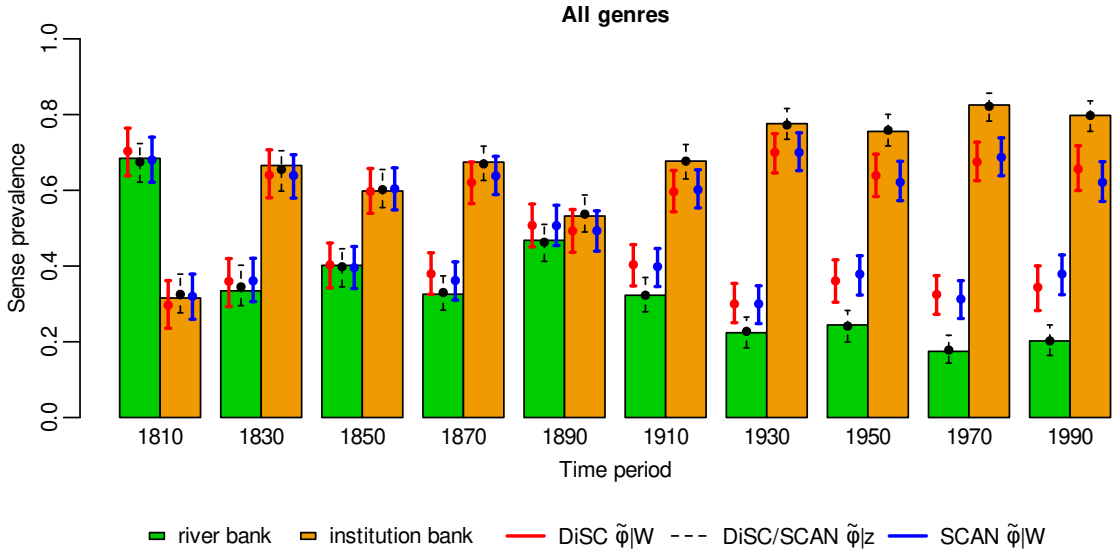


Figure 1: "Bank" expert-annotated empirical sense prevalence (coloured bars with height $N^o_{k,g,t} / \sum_{l=1}^{K} N^o_{l,g,t}$ for each k, g, t) with 95% HPD intervals (error bars) and posterior means (circles) from the model output. Note that there is no perceptible difference between the posteriors $\tilde{\phi}|z$ from DiSC and SCAN.

Marginal Highest Posterior Density (HPD) intervals are a useful visualisation of uncertainty in the model output. In order to form a summary of the evolution of the sense prevalence, we look at the 95% HPD intervals for the marginal posteriors $\tilde{\phi}_k^{g,t}|W$ from the DiSC and SCAN/GASC output. We compare these against independent well-informed estimates. We have the true sense labels $o_d, d \in \{1, \ldots, D\}$ (not available in general) for these data. However, these are only multinomial draws with category-probabilities given by the unknown true prevalence $\tilde{\Phi}$ say. Previous authors have benchmarked against empirical sense probabilities $N^o_{k,g,t} / \sum_{l=1}^{K} N^o_{l,g,t}$ computed using true sense labels. These are MLEs for the components of $\tilde{\Phi}$ in each sense, genre and time period. We expect our HPDs, computed on *unlabelled* data, to match these when the counts are high and the multinomial MLEs have small errors. However, when the labelled data counts are low (as in the "kosmos" data) or vary over a wide range from one time period to another, we should quantify the uncertainty in these independent estimates of the ground-truth prevalence $\tilde{\Phi}$.

We therefore treat the true sense labels as a second dataset, and validate our HPD sets estimated on the unlabelled data against HPD intervals for $\tilde{\phi}_k^{g,t}|(z = o)$ computed by taking the true sense labels as data in DiSC. (Whilst this may appear to favour DiSC, in fact, for this part of the model, the conditional distributions of $\tilde{\phi}|z$ are the same in all important respects, differing only by the AR(1) smoothing in DiSC.) A high degree of overlap between the HPD intervals from the two posteriors (based on labelled and unlabelled data) indicates

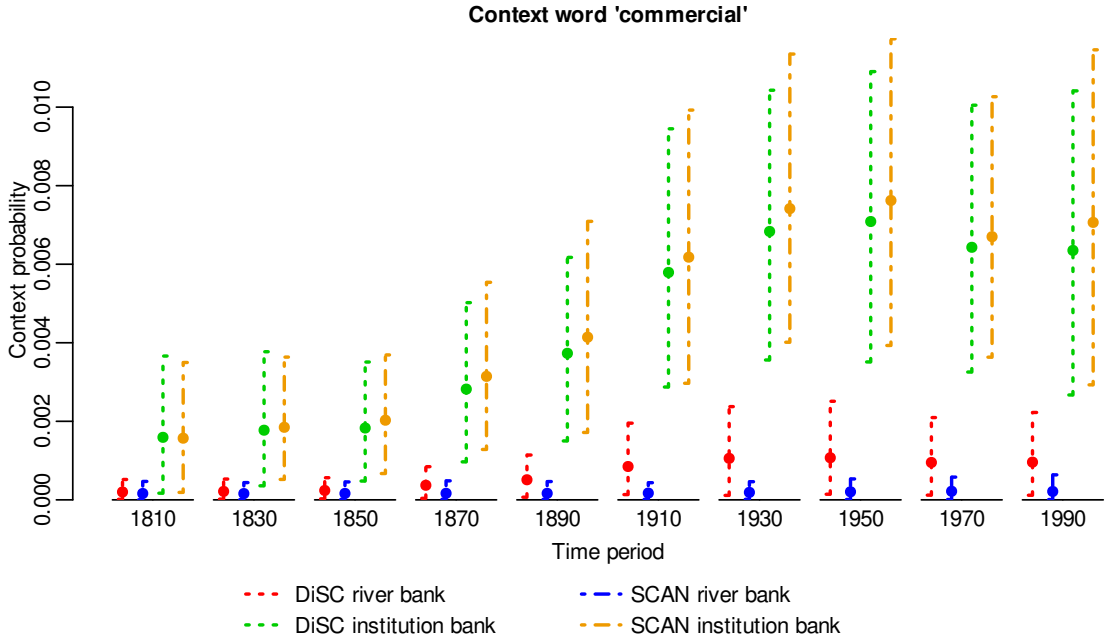


Figure 2: 95% HPD intervals (error bars) and posterior means (circles) for $\tilde{\psi}_v^{k,t}|W$

good model performance. Figure 1 shows that this is the case for both DiSC and SCAN (recall that SCAN is just GASC for $G = 1$) for both senses of "bank" for most time periods, although they start to diverge towards the end, perhaps indicating some over-smoothing across time. The performance of DiSC and SCAN is very similar. The ground-truth prevalence estimates based on MLEs (coloured bars) and labelled-data posteriors (dashed error bars) are very close for these data as the sample sizes are large (cf. Table 2). Equivalent graphs for thinner time intervals are given in Appendix D.

As discussed in Section 4.1, DiSC captures only the main sense and time effects via the additive structure in ψ whereas SCAN includes the sense-time interaction effect. As an example, Figure 2 shows the HPD intervals for $\tilde{\psi}_v^{k,t}|W$ for the context word $v =$ "commercial" under both models. "Commercial" appears predominantly under the institution sense of "bank", and increases in probability over time, which is reflected under both posteriors. However, whilst $\tilde{\psi}_v^{k,t}|W$ under SCAN for the river-bank sense remains relatively flat, in contrast, the DiSC posterior has $\tilde{\psi}_v^{k,t}|W$ increasing for both senses, as the contribution from $\theta_v^t|W$ increases and does not distinguish sense. We know from Brier scores and confusion matrices that DiSC gives better automatic sense-labelling than SCAN or GASC on these criteria. The reason for this is that although SCAN is capturing the interaction here, and DiSC is not, such interactions seem to be rare, and when present involve a context word with a relatively high context frequency $\tilde{\psi}_v^{k,t}|W$ in the evolving sense and low context frequencies

in all other senses (as generic context words have low frequency in any given sense). Our example illustrates this. Small absolute errors in small, uninformative context frequencies are offset by more reliable estimation of larger, more informative context frequencies. This is discussed further in the next section.

8 Experiment 3: Model predictive performance on synthetic data

Since DiSC drops sense-time interaction (cf. Sections 4.1), it is of interest to find examples in which this causes it to fail. We therefore compare the models' predictive performance on held-out sense labels in the presence of a *known* sense-time interaction effect in the data. We try to make this easy for SCAN and hard for DiSC: we calculate Brier scores on synthetic datasets of varying sizes generated using the SCAN model. Using the settings $T = 9$, $K = 3$ and $V \approx 1\,000$ seen in "kosmos", we sample parameters ϕ, ψ, κ_ϕ according to the SCAN prior model, as this generates sense-time interaction. We then sample the true sense labels o and snippets W using the SCAN/DiSC observation model for a fixed number of snippets D/T per time period, using the same stopword probability and context-word registration criterion as for "kosmos" (cf. Section 3). We choose a large D/T so that the data contain lots of sense-time interaction and also strongly inform this interaction. This setup might be expected to favour SCAN and challenge DiSC.

These random synthetic datasets do contain an abundance of sense-time interaction. Consider the following simple measure of the level Λ of interaction in the simulated data. First, we calculate the empirical probabilities $\hat{\tilde{\psi}}_v^{k,t} = \frac{N^{W,o}_{v,k,t}}{N^{W,o}_{\cdot,k,t}}$ for all v, k, t where $N^{W,z}_{\cdot,k,t} = \sum_{v=1}^{V} N^{W,z}_{v,k,t}$. Then, for each context word $v \in \{1, \ldots, V\}$, we fit a linear model

$$\hat{\tilde{\psi}}_v^{k,t} = \alpha_v + \beta_v t + \sum_{l=2}^{K} \delta_v^l \mathbb{I}(l = k) + t \sum_{l=2}^{K} \eta_v^l \mathbb{I}(l = k) + \varepsilon_v \tag{14}$$

where time t is continuous (since discrete t leads to a perfect fit) and ε_v is Gaussian noise. We make an F-test for the interaction effects η_v^l at level 5% and measure the extent of sense-time interaction Λ as the proportion of context words $\{1, \ldots, V\}$ for which the interaction effects in (14) are significant. This is a conservative measure, since it only captures linear sense-time interactions and misses potential situations where context probability increases and then decreases (or vice versa) over time. In the synthetic data with D/T equal 100 and 500, the interaction level is $\Lambda = 0.13$ and $\Lambda = 0.20$ respectively. For comparison, using this measure, we have $\Lambda = 0.144$ and $\Lambda = 0.070$ respectively for the "bank" and "kosmos"

Table 7: Model performance for DiSC and SCAN on synthetic data

	$D/T = 100, \Lambda = 0.131$		$D/T = 500, \Lambda = 0.204$	
	Converges?	Brier score	Converges?	Brier score
DiSC	Yes	0.017	Yes	0.0065
SCAN	No[†]	—	Yes	0.0072

[†] MCMC runs from different starting configurations lead to different equilibrium distributions

datasets, so on this simple measure the extent of interaction in these synthetic data is representative.

Table 7 summarises the results of these experiments. Our MCMC for SCAN on the smaller $D/T = 100$ dataset did not converge, whereas we get good convergence on the larger $D/T = 500$ dataset. DiSC estimates a smaller number of parameters than SCAN, with lower variance but potential bias. This tradeoff seems to be advantageous for sense-labelling: as evidenced by the Brier scores on the larger dataset, DiSC sense-labelling on SCAN-friendly synthetic data is as good as, or better than, SCAN itself.

The Brier score is a proper scoring rule, so we expect SCAN to do better on average over synthetic datasets — on average but not necessarily in probability. Posterior plots similar to Figure 2 (not included in this paper) confirm our intuition that, despite the large proportion of context words with a significant sense-time interaction effect, very few of these words appear with high probability across more than one sense. This may explain the Brier score ordering. It is possible to construct data for which SCAN scores more highly than DiSC by artificially fixing the prior parameters to include a very large proportion of words with *both* high frequency in more than one sense *and* strong interactions (cf. Section 4.1). However, this seems unrepresentative of typical real data. In experiments where we explicitly introduce such strong interactions across multiple senses (discussed in Appendix E), we find that the performance of SCAN is a little better than that of DiSC but the gain is slight. In our experience, modelling sense-time interactions is detrimental for automatic sense-labelling.

9 Experiment 4: Analysis of sense change for "kosmos"

We now come to the principal application. We will make this analysis twice: here, and in Appendix F, depending on whether we exclude (as here) or include non-collocates. All translations of Greek words in this section have been obtained from Wiktionary.

The ancient Greek data for target word "kosmos" (κόσμος) contains considerably fewer

Table 8: "Kosmos" confusion matrix statistics using DiSC

	Positive condition		
	Decoration	Order	World
Sensitivity (true +ve)	0.629	0.547	0.863
Specificity (true −ve)	0.873	0.872	0.815

snippets than the "bank" data (cf. Section 3) whilst using almost the same sized vocabulary, making it a relatively sparse and noisy dataset. The "kosmos" data contains other features making this analysis harder than the "bank" analysis: the three senses of "kosmos" are not as well separated as those of "bank", and a number of context words appear with high probability under more than one sense in the expert annotation $o_d, d \in \{1, \ldots, D\}$. Examples include θεός (divine/god $\{0.26, 0.10, 0.64\}$), ἀνήρ (man $\{0.45, 0.47, 0.08\}$) and πόλις (city $\{0.30, 0.64, 0.06\}$) among others where, given that word v appears as context in a snippet, it appears under sense k with empirical probability $\frac{N^{W,o}_{v,k,\cdot}}{N^{W,o}_{\cdot,k,\cdot}} \Big/ \sum_{l=1}^{K} \frac{N^{W,o}_{v,l,\cdot}}{N^{W,o}_{\cdot,l,\cdot}}$ for $k = \{1, \ldots, K\}$, where $N^{W,z}_{v,k,\cdot} = \sum_{t=1}^{T} N^{W,z}_{v,k,t}$ and $N^{W,z}_{\cdot,k,\cdot} = \sum_{v=1}^{V} N^{W,z}_{v,k,\cdot}$. This makes the task of automatic sense-identification in the "kosmos" data particularly challenging.

Using our DiSC model with two genres, the words most probable *a posteriori* under each sense are as follows:

1	γυνή	χρύσεος	φέρω	κοσμέω	καλός	σῶμα	ὅπλον	ἐσθής	πλῆθος	τάξις
2	πόλις	πολιτεία	πρότερος	ἀνήρ	φέρω	νόμος	μέγας	πάρειμι	ἀρχή	καθίστημι
3	γῆ	οὐρανός	ὅλος	θεός	εἷς	φύσις	καλός	σύμπας	ψυχή	σῶμα

We compare these against ground truth. We identify the senses $\{1, 2, 3\}$ with the labels {decoration, order, world} by mapping our marginal posterior distributions $\frac{1}{T}\sum_{t=1}^{T} \tilde{\psi}^{k,t}$ for $k \in \{1, 2, 3\}$ to their closest empirical distributions $(N^{W,o}_{1,k,\cdot}, \ldots, N^{W,o}_{V,k,\cdot})/N^{W,o}_{\cdot,k,\cdot}$ under the expert-annotated sense labels $o_d, d \in \{1, \ldots, D\}$. We note that certain distinctive words that help identify the sense have been correctly assigned by our model, for example: γυνή (woman) and χρύσεος (golden) for "decoration"; πολιτεία (citizenship) and νόμος (custom/law) for "order"; γῆ (earth) and οὐρανός (sky) for "world". Some words such as τάξις (literally "order" but assigned with "decoration") have been misplaced, although the error is small: if $v =$ τάξις then v is assigned by the model to sense k with probability $p(z_d = k | v \in W_d, \tilde{\psi}) = \sum_{t=1}^{T} \tilde{\psi}^{k,t}_v \Big/ \sum_{l=1}^{K}\sum_{t=1}^{T} \tilde{\psi}^{l,t}_v = \{0.43, 0.42, 0.15\}$ for $k = \{1, 2, 3\}$.

Using the Brier score to assess model performance, the baseline score under uniform random assignment is $BS = 0.67$ for $K = 3$ senses. The realised score of $BS = 0.41$ using DiSC

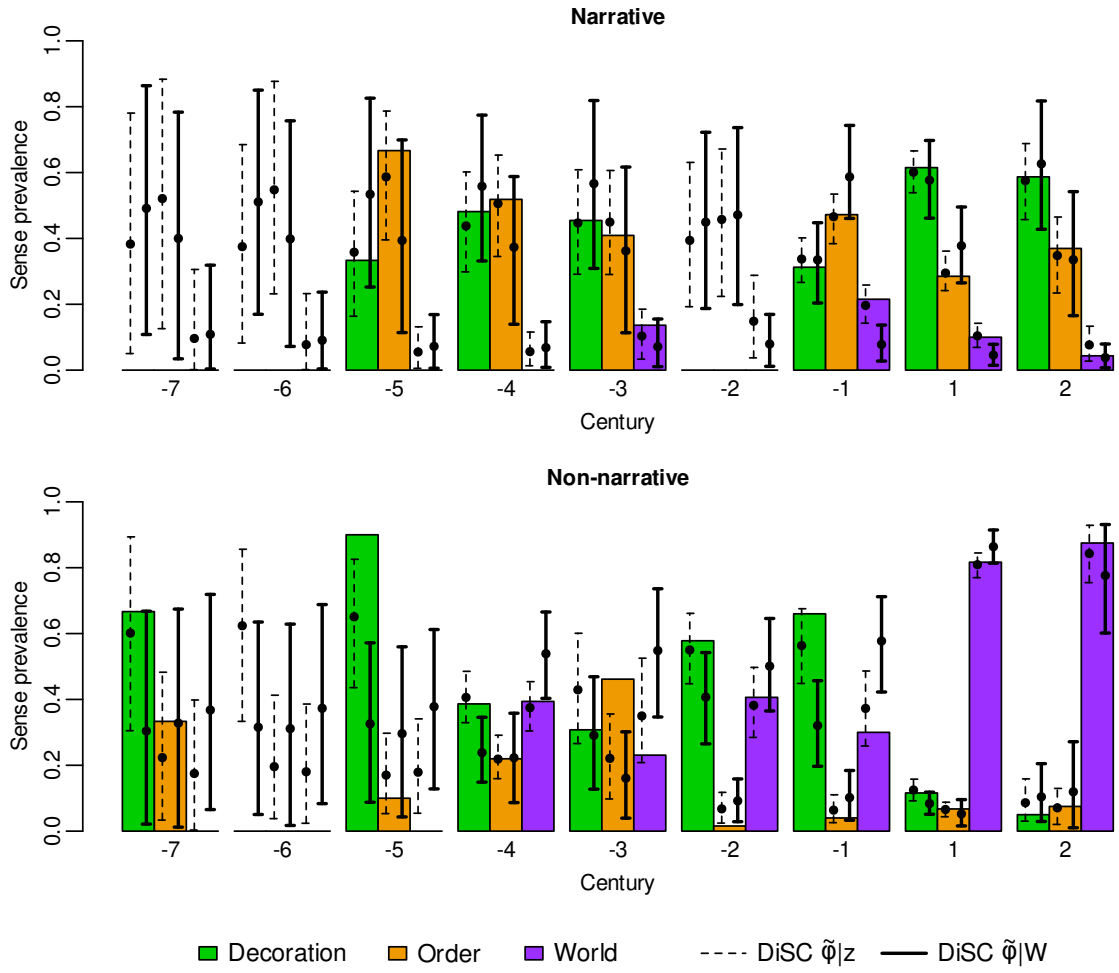


Figure 3: "Kosmos" expert-annotated empirical sense prevalence (coloured bars with height $N^o_{k,g,t} / \sum_{l=1}^{K} N^o_{l,g,t}$ for each k, g, t) with 95% HPD intervals (error bars) and posterior means (circles) from the DiSC output

therefore indicates good model performance. Ignoring genre and setting $G = 1$ gives $BS = 0.47$, i.e. worse than the $G = 2$ case, so the genre-covariate plays a role, in agreement with the views of Perrone et al. (2019). As in the "bank" analysis, the confusion matrix from a similar classification task for the "kosmos" data is summarised in Table 8, indicating generally good model performance. The relatively lower sensitivity for the "decoration" and "order" senses is reflective of the ambiguity discussed above.

In Figure 3 the 95% HPD intervals for the marginal posteriors $\tilde{\phi}^{g,t}_k | W$ from the DiSC output have a high degree of overlap with the posteriors $\tilde{\phi}^{g,t}_k | (z = o)$ based on expert-annotated sense labels for all k, g, t, again indicating good model performance. The DiSC HPD intervals contain the empirical estimates $N^o_{k,g,t} / \sum_{l=1}^{K} N^o_{l,g,t}$ for these data in most cases, with the exceptions being the time periods with little data (cf. Table 3). For these

time periods, ground-truth sense prevalence estimates given by the posteriors $\tilde{\phi}_k^{g,t}|(z = o)$ seem to us a more appropriate basis for comparison than the simple empirical estimates. Adjacent temporal data smooths these posteriors. Also, where there is limited data at both time t and times $t \pm 1$, the increased uncertainty in the ground-truth is quantified in the wider HPD intervals for those times t. The very high degree of overlap between the marginal posteriors $\tilde{\phi}_k^{g,t}|W$ and those for $\tilde{\phi}_k^{g,t}|(z = o)$ indicates we are doing as well, for prevalence estimation with the unlabelled data, as we would do if we actually had the ground truth.

Using GASC, on the other hand, we find that all MCMC fail to converge to the same stationary distribution starting from different random configurations. Exploration showed that the posterior distribution for GASC contains multiple metastable states with very long lifetimes and multiple modes (besides those associated with label switching). We do not see convergence even when using our best samplers with parallel tempering on multiple cores. This seems to be associated with over-parameterisation in the GASC model for this small and noisy dataset: the ψ array has VKT parameters in GASC compared to only $V(K+T)$ parameters in DiSC. This is more than double the number of parameters, and leads to multiple modes of near-equal fit as measured by the log-likelihood.

The senses cannot be reliably identified from the model output as they are not sufficiently distinct. For example, our best run using GASC from six different random starting configurations gave the following most probable words:

1	ἀνήρ	πολιτεία	ἀρχή	γυνή	κύριος	ἀξιόω	τάξις	πρότερος	πάτριος	κόσμος
2	φέρω	γυνή	πόλις	καλός	κοσμέω	θεός	μέγας	τάξις	πολιτεία	σῶμα
3	γῆ	πόλις	ὅλος	θεός	οὐρανός	εἷς	φύσις	σῶμα	πέντε	σύμπας

Since representative words for both "order" and "decoration" (such as πολιτεία and γυνή) appear with high probability under two senses, it is not clear how to assign the sense labels between them in order to make any sort of comparison. The same behaviour is seen further down the sense columns for other less frequent context words. Trying all possible permutations of the sense labels, the best Brier score we get is $BS = 0.73$ — worse than randomly assigning the probability $\hat{p}(z_d = k) = \frac{1}{K}$ for all d, k.

10 Conclusion

We have introduced DiSC, a generative model of diachronic sense change treating sense and time as additive effects, and a gradient-based MCMC method for inferring model parameters.

Whilst we adopt the overall modelling framework of Frermann and Lapata (2016) and Perrone et al. (2019), we found that the specific MCMC samplers used there take around 40 times as long to achieve the same ESS. Our MCMC exploits the fact that it is possible to marginalise by summing over the discrete sense assignments exactly. Gradient-based MCMC on the marginal is, not surprisingly, far more efficient than simple Gibbs sampling on the joint model.

We carried out automatic sense-annotation by identifying word-sense dependence, and discovered sense groupings. We measured time-evolution of sense distributions over context words and of sense prevalence distributions over senses. We showed that, for the well-behaved "bank" data where MCMC targeting DiSC, SCAN and GASC all converge, DiSC has slightly better predictive performance for held-out expert-annotated sense labels despite being the simpler model. We further showed that, for the smaller and noisier "kosmos" data where no well-calibrated fitting procedure is available for SCAN and GASC, DiSC works well in the sense that it gives prevalence estimation intervals close to those we would obtain if we had the true sense labels.

As far as we are aware, our analyses of "kosmos" in the test case Section 9 and "real-use" case Appendix F give the first analysis of sense change for these data that returns good agreement with prevalence from expert annotation. One criticism examined in Section 4.1 is that DiSC does not model the sense-time interaction effect. In our setting, the sense of a target word is defined by the distribution over its context words. DiSC does model temporal change in this distribution, but captures only the main effects associated with changes in the overall usage frequency of the context words across all senses of the target word. It would be interesting to look for further examples of target words where the independent evolution of senses modelled by SCAN/GASC is measurable and impacts sense-labelling.

Our study of synthetic data in Section 8 showed that even when data are simulated under SCAN, with abundant interaction, DiSC gives reliable sense-label predictions and parameter estimates. Furthermore, in many potentially interesting cases where the data are sparse, there is no computational procedure we know of that will actually fit the interaction as parameterised in SCAN/GASC reliably. In future work we would like to find a parameterisation of the sense-time interaction effect that is tractable. DiSC would then be a natural null model to this alternative.

All these models can identify the emergence of new senses, but we have not explored this. It may be possible to include an atom of probability at $\tilde{\phi}_k^{g,t} = 0$ in order to carry out

simultaneous and formal model-selection for the first time t at which $\tilde{\phi}_k^{g,t} > 0$ and there is evidence for a new sense.

DiSC shares some disadvantages with the generative models on which it is based. We target the senses of one word at a time, in contrast to some approaches in the machine learning literature cited in Section 2. However, DiSC does provide well-calibrated uncertainty estimates, as evidenced by our experiments, rather than just point estimates. Another limitation is that the number of senses K needs to be checked in multiple runs. It may be possible to learn K using more formal testing procedures. However, for estimation of sense distributions (over context words), sense prevalence and their evolution from unlabelled text, DiSC works well in a semi-supervised mode in which model selection is based on the requirement that the output be meaningful to the user.

Implementation

R scripts and data files used to produce the results, figures and tables reported in the paper are available from https://github.com/schyanzafar/DiSC.

Acknowledgement

This research is funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/S515541/1.

Appendices

A GASC generative model

The GASC generative model is given in Algorithm 2. Note:

1. Following Frermann and Lapata (2016), the authors of Perrone et al. (2019) give generative prior processes for ϕ and ψ in terms of the full conditional distributions. This is natural as there is an improper initialisation. In lines 3–17 we present the generative process as an equivalent Markov chain (Rue and Held, 2005).
2. In order to make an exact numerical match with GASC, κ_ϕ in Algorithm 2 has a factor of two multiplier; so the κ_ϕ parameters in GASC and DiSC differ by a factor of two even when $\alpha_\phi = 1$ in the AR(1) process.
3. Lines 19 and 21 treat stopwords and uninformative words explicitly. Stopwords may be treated as punctuation and dropped without counting as occupying a context position. Also, Perrone et al. (2019) do not discuss how uninformative words or hapaxes are

Algorithm 2 GASC: generative model

PRIOR MODEL

fix hyperparameters κ_ψ, a, b
draw $\kappa_\phi \sim \text{Inv Gamma}(a, b)$
initialise at time $t = 1$
for genre $g \in 1 : G$ **do**
    draw sense prevalence parameter from improper uniform $\phi^{g,1} \sim \pi(\phi^{g,1}) \propto 1$
end for
for sense $k \in 1 : K$ **do**
    draw word parameter from improper uniform density $\psi^{k,1} \sim \pi(\psi^{k,1}) \propto 1$
end for
for time $t \in 2 : T$ **do**
    for genre $g \in 1 : G$ **do**
        draw sense prevalence parameter $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi \sim \mathcal{N}\left(\phi^{g,t-1}, \text{diag}(2\kappa_\phi)\right)$
    end for
    for sense $k \in 1 : K$ **do**
        draw word parameter $\psi^{k,t}|\psi^{k,t-1}, \kappa_\psi \sim \mathcal{N}\left(\psi^{k,t-1}, \text{diag}(2\kappa_\psi)\right)$
    end for
end for
using softmax (1), transform real arrays ϕ and ψ into probability arrays $\tilde{\phi}$ and $\tilde{\psi}$

OBSERVATION MODEL

fix probabilities of drawing stopwords q^{SW} and uninformative words q^{U}
for snippet $d \in 1 : D$ **do**
    draw number of context words $L_d|L, q^{\text{SW}}, q^{\text{U}} \sim \text{Bin}(L, 1 - q^{\text{SW}} - q^{\text{U}})$
    draw a random subset $\{i_1, \ldots, i_{L_d}\}$ of size L_d from $\{1, \ldots, L\}$
    draw sense assignment $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d,\tau_d}\right)$
    for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
        draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d,\tau_d}, \ldots, \tilde{\psi}_V^{z_d,\tau_d}\right)$
    end for
end for

identified in snippets, so this is a change on their setup. We chose to treat these words in GASC as we do for DiSC as it seemed only an improvement.

B Gibbs samplers

We briefly describe and critique current MCMC methods for sampling the posterior $\phi^{g,t}|z$ in the DiSC and GASC models. Sampling $\psi^{k,t}|z, W$ in the GASC model is very similar. The conditional prior distribution is

$$\phi^{g,t}|\phi^{g,-t} \sim \begin{cases} \mathcal{N}\left(\alpha_\phi \phi^{g,t+1}, \text{diag}(\kappa_\phi)\right) & \text{if } t = 1 \\ \mathcal{N}\left(\frac{\alpha_\phi}{1+(\alpha_\phi)^2}(\phi^{g,t-1} + \phi^{g,t+1}), \text{diag}\left(\frac{\kappa_\phi}{1+(\alpha_\phi)^2}\right)\right) & \text{if } t \in 2 : (T-1) \\ \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \text{diag}(\kappa_\phi)\right) & \text{if } t = T \end{cases} \tag{15}$$

under DiSC. For GASC, we replace κ_ϕ with $2\kappa_\phi$ and substitute $\alpha_\phi = 1$.

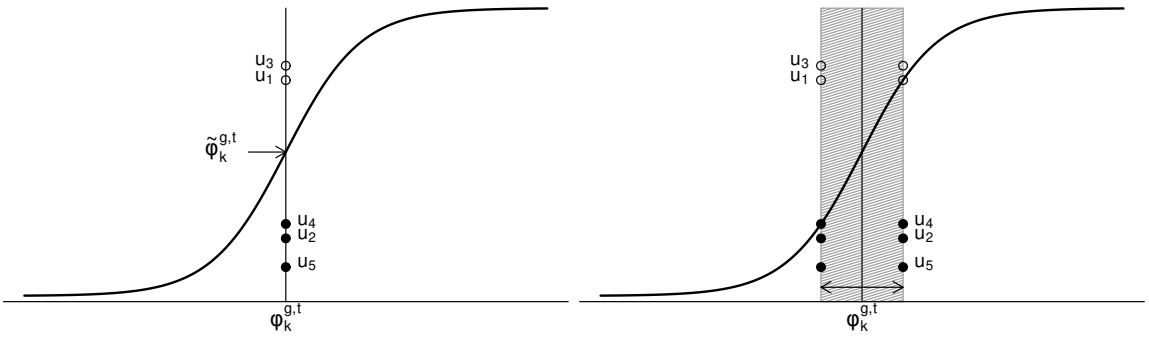


Figure 4: Relationship between $\phi_k^{g,t}$, z, and u in the case of 5 snippets. LEFT: Given $\phi_k^{g,t}$, if $z_d = k$ then u_d (black circle) falls below $\tilde{\phi}_k^{g,t}$; and if $z_d \neq k$ then u_d (white circle) falls above $\tilde{\phi}_k^{g,t}$. RIGHT: With u given, $\phi_k^{g,t}$ can be anywhere within the interval defined by the highest black circle and the lowest white circle. Figure adapted from Mimno et al. (2008).

B.1 Auxiliary uniform variable method

The auxiliary uniform variable method of Mimno et al. (2008), which is based on the method of Groenewald and Mokgatlhe (2005), is used to sample $\phi_k^{g,t}|\phi_{-k}^{g,t}, z$ iteratively over $k \in 1:K$ by moving within a weighted and bounded region, where the weights are determined by the prior distributions and the bounds are determined by $N_{k,g,t}^z$.

Figure 4 shows an example. Given $N_{k,g,t}^z$ snippets assigned sense k and $N_{\cdot,g,t}^z$ snippets in total for time t and genre g, we draw $N_{k,g,t}^z$ uniform variables below the logistic CDF and $N_{\cdot,g,t}^z - N_{k,g,t}^z$ uniform variables above the logistic CDF. That is, for all $d \in \{d' : \tau_{d'} = t \text{ and } \gamma_{d'} = g\}$ we draw

$$u_d \sim \begin{cases} \mathcal{U}(0, \tilde{\phi}_k^{g,t}) & \text{if } z_d = k \\ \mathcal{U}(\tilde{\phi}_k^{g,t}, 1) & \text{if } z_d \neq k, \end{cases} \tag{16}$$

and sample a new $\phi_k^{g,t}$ from the conditional prior (15) truncated to the bounded region

$$\max_{d:z_d=k} \log \frac{Cu_d}{1-u_d} < \phi_k^{g,t} < \min_{d:z_d\neq k} \log \frac{Cu_d}{1-u_d} \tag{17}$$

where

$$C = \sum_{k' \neq k} \exp(\phi_{k'}^{g,t}). \tag{18}$$

The efficiency of this procedure can be improved by sampling two Beta variables instead of $N_{\cdot,g,t}^z$ uniform variables.

The problem with this method is that the bounded region (17) can be very narrow whenever the dimensions $N_{k,g,t}^z$ and $N_{\cdot,g,t}^z$ are large, leading to a very small move in each iteration. This is because sampling a large number of uniform variables below or above the CDF is

likely to result in at least one variable being close to the curve. The resulting convergence is therefore very slow.

B.2 Auxiliary Polya-Gamma variable method

The auxiliary Polya-Gamma variable method of Polson et al. (2012) is used to sample $\phi_k^{g,t}|\phi_{-k}^{g,t}, z$ iteratively over $k \in 1:K$ by first drawing

$$\omega \sim \mathcal{PG}(N_{\cdot,g,t}^z, \eta) \tag{19}$$

from the Polya-Gamma distribution, where $\eta = \phi_k^{g,t} - \log C$ with C as in (18), and then sampling

$$\phi_k^{g,t} \sim \mathcal{N}(m, \Sigma), \tag{20}$$

where $\Sigma = (\sigma^{-2} + \omega)^{-1}$, $m = \Sigma(\mu\sigma^{-2} + N_{k,g,t}^z - \frac{1}{2}N_{\cdot,g,t}^z + \omega \log C)$, and μ and σ^2 are the mean and variance for the relevant conditional prior distribution defined in (15) depending on t.

A draw from a Polya-Gamma distribution is computationally expensive whenever the shape parameter $N_{\cdot,g,t}^z$ is large. This is very often the case with any real dataset, especially for ψ updates. Chen et al. (2013) give an alternative approximate method which cuts the computational cost of each draw from $\mathcal{O}(N_{\cdot,g,t}^z)$ down to $\mathcal{O}(1)$. They note that if $x_i \sim \mathcal{PG}(1, \eta)$ then $\omega = \sum_{i=1}^N x_i \sim \mathcal{PG}(N, \eta)$ by the additive property of the Polya-Gamma distribution. Therefore, by the central limit theorem, ω is approximately Gaussian for large N, and may be obtained by transforming another approximately Gaussian random variable $\lambda \sim \mathcal{PG}(M, \eta)$ viz.

$$\sqrt{\mathbb{V}(\omega)/\mathbb{V}(\lambda)}(\lambda - \mathbb{E}[\lambda]) + \mathbb{E}[\omega] \tag{21}$$

where $\mathbb{E}[\omega] = \frac{N}{2\eta}\tanh\frac{\eta}{2}$ and $\mathbb{V}(\omega)/\mathbb{V}(\lambda) = N/M$. This approximate strategy works well even when $M = 1$. The method is relatively simple. However, parameter inference is not asymptotically exact and, as we saw in Section 6, the method is in any case dominated by a hybrid MALA-HMC sampler which is asymptotically exact.

C Gradient-based MCMC methods

We sample the conditional posteriors (10)–(12) for each variable by proposing an update from a suitable distribution and accepting or rejecting it using the Hastings ratio, iterating over the columns. For example for ϕ, we propose a candidate vector

$$\phi^{*g,t}|\phi^{g,\cdot}, \psi^{\cdot,t}, W_{\mathcal{D}(g,t)} \sim q(\cdot|\phi^{g,\cdot}, \psi^{\cdot,t}, W_{\mathcal{D}(g,t)}) \tag{22}$$

and accept it with probability

$$1 \wedge \frac{\pi(\phi^{*g,t}|\phi^{g,-t})p(W_{\mathcal{D}(g,t)}|\phi^{*g,t},\psi^{\cdot,t})}{\pi(\phi^{g,t}|\phi^{g,-t})p(W_{\mathcal{D}(g,t)}|\phi^{g,t},\psi^{\cdot,t})}\frac{q(\phi^{g,t}|\phi^{*g,\cdot},\psi^{\cdot,t},W_{\mathcal{D}(g,t)})}{q(\phi^{*g,t}|\phi^{g,\cdot},\psi^{\cdot,t},W_{\mathcal{D}(g,t)})} \tag{23}$$

where $p(W_{\mathcal{D}(g,t)}|\phi^{g,t},\psi^{\cdot,t})$ is the likelihood

$$p(W_{\mathcal{D}(g,t)}|\phi^{g,t},\psi^{\cdot,t}) = \prod_{d\in\mathcal{D}(g,t)}\sum_{k=1}^{K}\tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}}\tilde{\psi}_{w_{d,i}}^{k,\tau_d} \tag{24}$$

and $\mathcal{D}(g,t) = \{d : \gamma_d \in g \text{ and } \tau_d \in t\}$ is the set of snippet indices for time(s) t and genre(s) g. Updates for θ, χ for DiSC or ψ for GASC are analogous. The conditional prior density $\pi(\phi^{g,t}|\phi^{g,-t})$ is defined by (15), whereas for θ we have the conditional

$$\theta^t|\theta^{-t} \sim \begin{cases} \mathcal{N}\left(\alpha_\theta\theta^{t+1}, \text{diag}(\kappa_\theta)\right) & \text{if } t = 1 \\ \mathcal{N}\left(\frac{\alpha_\theta}{1+(\alpha_\theta)^2}(\theta^{t-1}+\theta^{t+1}), \text{diag}\left(\frac{\kappa_\theta}{1+(\alpha_\theta)^2}\right)\right) & \text{if } t \in 2:(T-1) \\ \mathcal{N}\left(\alpha_\theta\theta^{t-1}, \text{diag}(\kappa_\theta)\right) & \text{if } t = T, \end{cases} \tag{25}$$

and for χ we have, unconditionally,

$$\chi^k \sim \mathcal{N}\left(0, \text{diag}(\kappa_\chi)\right). \tag{26}$$

In practice, it may be more efficient to update all columns of χ together, which is the approach we take in our R implementation for the applications in this paper.

The proposal density $q(\cdot|\phi^{g,\cdot},\psi^{\cdot,t},W_{\mathcal{D}(g,t)})$ is constructed using gradient-based methods such as MALA (Roberts and Tweedie, 1996) or HMC (Duane et al., 1987), which make proposals in the direction where the posterior density is increasing. MALA, for instance, uses the proposal distribution

$$\phi^{*g,t}|\phi^{g,\cdot},\psi^{\cdot,t},W_{\mathcal{D}(g,t)} \sim \mathcal{N}\left(\phi^{g,t} + \frac{\sigma_\phi^2}{2}\nabla_{\phi^{g,t}}\log\pi(\phi^{g,t}|\phi^{g,-t},\psi^{\cdot,t},W_{\mathcal{D}(g,t)}), \sigma_\phi^2\Sigma_\phi\right) \tag{27}$$

for ϕ updates, and analogous distributions for the other variables, where σ_x^2 and Σ_x are the proposal scale and covariance parameters for variable x respectively. We keep the covariance Σ_x fixed at the appropriate identity matrix, whereas we tune the scale σ_x^2 using the log-adaptive proposals of Shaby and Wells (2010). When using HMC with the leapfrog method (see e.g. Neal 2011), we keep the number of leapfrog steps fixed and use the same technique to tune the leapfrog step size σ_x^2. We experimented with tuning the covariance Σ_x, including varying diagonal elements, without gain.

The tuning is done by calculating the running empirical acceptance rate $\bar{\alpha} = \frac{\#\text{ jumps}}{N}$ based on the last N MCMC iterations, and adjusting the scale parameter up or down via $\log\sigma_x^2 \leftarrow$

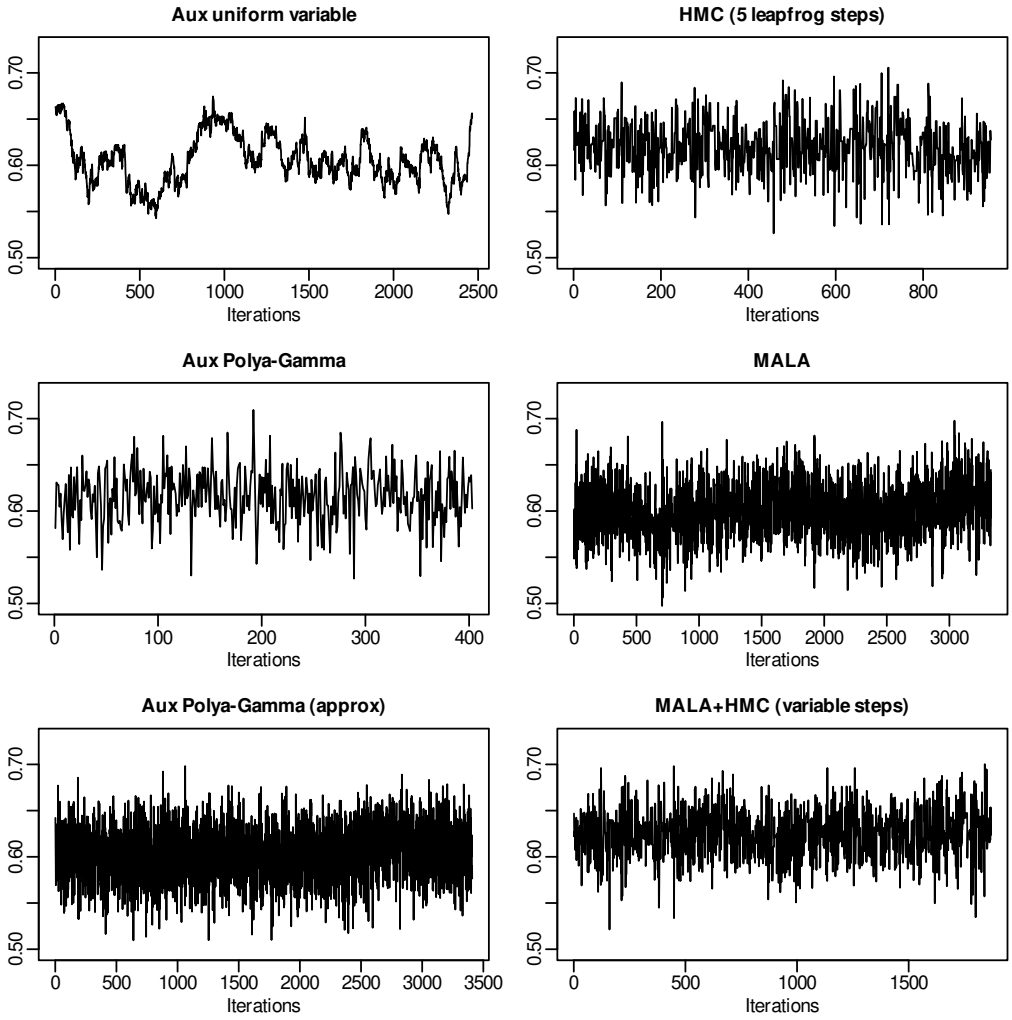


Figure 5: $\tilde{\phi}_k^{g,t}$ trace plots with equal run times (after burn-in) for the institution sense of "bank" at time 1850-70 from the unlabelled SCAN posterior

$\log \sigma_x^2 + C_n(\bar{\alpha} - \alpha^{\text{opt}})$, where C_n is a parameter that decreases with iteration number n (so that σ_x^2 tends to a constant as $n \to \infty$) and α^{opt} is the target optimal acceptance rate. Under certain conditions, optimal asymptotic acceptance rates have been proposed as 0.574 for MALA (Roberts and Rosenthal, 2002) and 0.651 for HMC (Beskos et al., 2013), which are the values we use for α^{opt} in our implementations.

The gradient of the log posterior density in (27) can be broken up into the sum of gradients of the log prior density and the log likelihood. The gradients of the log prior densities $\nabla_{\phi^{g,t}} \log \pi(\phi^{g,t}|\phi^{g,-t})$, $\nabla_{\theta^t} \log \pi(\theta^t|\theta^{-t})$ and $\nabla_{\chi^k} \log \pi(\chi^k)$ are of the form $-V_x^{-1}(x - \mu_x)$ where the mean vector μ_x and covariance matrix V_x for variable x are given by (15), (25) and (26) respectively. The gradients of the log likelihoods are derived in the next subsection.

As we find in Section 6 above, these gradient-based methods give significant efficiency gains over methods like the auxiliary uniform or the (asymptotically exact) auxiliary Polya-Gamma samplers described in Appendix B. This is illustrated in Figure 5 in which the x-axis shows the same elapsed time in CPU seconds across all plots and the y-axis shows the MCMC state for the prevalence parameter in one genre and time period. The gradient-based samplers shown on the right give much more rapid mixing than the other two methods.

C.1 Derivation of $\nabla_{\phi^{g,t}} \log p(W_{\mathcal{D}(g,t)}|\phi^{g,t}, \psi^{\cdot,t})$

From equation (24) we get

$$\log p(W_{\mathcal{D}(g,t)}|\phi^{g,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(g,t)} \log \sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d},$$

and taking derivatives with respect to $\phi_j^{g,t}$ gives

$$\frac{\partial}{\partial \phi_j^{g,t}} \log p(W_{\mathcal{D}(g,t)}|\phi^{g,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(g,t)} \frac{\frac{\partial}{\partial \phi_j^{g,t}} \sum_{k=1}^{K} \tilde{\phi}_k^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{k=1}^{K} \tilde{\phi}_k^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}. \tag{28}$$

Now, $\tilde{\phi}_k^{g,t} = \frac{\exp(\phi_k^{g,t})}{\sum_{k'=1}^{K} \exp(\phi_{k'}^{g,t})}$ so that

$$\frac{\partial \tilde{\phi}_k^{g,t}}{\partial \phi_j^{g,t}} = \frac{\exp(\phi_j^{g,t}) \mathbb{I}(j=k) \sum_{k'=1}^{K} \exp(\phi_{k'}^{g,t}) - \exp(\phi_j^{g,t}) \exp(\phi_k^{g,t})}{\left(\sum_{k'=1}^{K} \exp(\phi_{k'}^{g,t})\right)^2} = \tilde{\phi}_j^{g,t} \left(\mathbb{I}(j=k) - \tilde{\phi}_k^{g,t}\right). \tag{29}$$

Hence we have

$$\frac{\partial}{\partial \phi_j^{g,t}} \sum_{k=1}^{K} \tilde{\phi}_k^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t} = \tilde{\phi}_j^{g,t} \left(\prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{j,t} - \sum_{k=1}^{K} \tilde{\phi}_k^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t} \right),$$

and substituting this into (28) gives

$$\frac{\partial}{\partial \phi_j^{g,t}} \log p(W_{\mathcal{D}(g,t)}|\phi^{g,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(g,t)} \frac{\tilde{\phi}_j^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{j,t}}{\sum_{k=1}^{K} \tilde{\phi}_k^{g,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}} - \tilde{\phi}_j^{g,t} N_{\cdot,g,t}^{z} \tag{30}$$

which are the elements of vector $\nabla_{\phi^{g,t}} \log p(W_{\mathcal{D}(g,t)}|\phi^{g,t}, \psi^{\cdot,t})$ for $j \in \{1, \ldots, K\}$.

C.2 Derivation of $\nabla_{\psi^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$

The gradient vector $\nabla_{\psi^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$ is required for inferring ψ in the GASC model. Differentiating the log likelihood with respect to $\psi_j^{k,t}$ gives

$$\frac{\partial}{\partial \psi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(1:G,t)} \frac{\frac{\partial}{\partial \psi_j^{k,t}} \sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}}. \tag{31}$$

As in (29), we have $\frac{\partial \tilde{\psi}_v^{l,t}}{\partial \psi_j^{k,t}} = \tilde{\psi}_v^{l,t}\left(\mathbb{I}(j=v) - \tilde{\psi}_j^{l,t}\right)$ if $k=l$, and $\frac{\partial \tilde{\psi}_v^{l,t}}{\partial \psi_j^{k,t}} = 0$ otherwise. Hence we have

$$\begin{aligned}
\frac{\partial}{\partial \psi_j^{k,t}} \sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t} &= \frac{\partial}{\partial \psi_j^{k,t}} \tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t} \\
&= \tilde{\phi}_k^{\gamma_d,t} \sum_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t} \left(\mathbb{I}(j=w_{d,i}) - \tilde{\psi}_j^{k,t}\right) \prod_{i' \neq i} \tilde{\psi}_{w_{d,i'}}^{k,t} \\
&= \tilde{\phi}_k^{\gamma_d,t} \sum_{i=i_1}^{i_{L_d}} \left(\mathbb{I}(j=w_{d,i}) - \tilde{\psi}_j^{k,t}\right) \prod_{i'=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i'}}^{k,t} \\
&= \left(\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}\right) \left(\sum_{i=i_1}^{i_{L_d}} \mathbb{I}(j=w_{d,i}) - L_d \tilde{\psi}_j^{k,t}\right)
\end{aligned}$$

and substituting this into (31) gives

$$\frac{\partial}{\partial \psi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(1:G,t)} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \left(\sum_{i=i_1}^{i_{L_d}} \mathbb{I}(j=w_{d,i}) - L_d \tilde{\psi}_j^{k,t}\right) \tag{32}$$

which are the elements of vector $\nabla_{\psi^{k,t}} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t})$ for $j \in \{1, \ldots, V\}$.

C.3 Derivation of $\nabla_{\theta^t} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t})$

The relationship $\psi_j^{k,t} = \chi_j^k + \theta_j^t$ gives $\frac{\partial \psi_j^{k,t}}{\partial \theta_j^t} = 1$ for all $k \in \{1, \ldots, K\}$, so applying the chain rule to (32) we get

$$\begin{aligned}
\frac{\partial}{\partial \theta_j^t} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t}) &= \sum_{k=1}^{K} \frac{\partial}{\partial \psi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t}) \\
&= \sum_{d \in \mathcal{D}(1:G,t)} \left(\sum_{i=i_1}^{i_{L_d}} \mathbb{I}(j=w_{d,i}) - L_d \sum_{k=1}^{K} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \tilde{\psi}_j^{k,t}\right)
\end{aligned} \tag{33}$$

which are the elements of vector $\nabla_{\theta^t} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t})$ for $j \in \{1, \ldots, V\}$.

C.4 Derivation of $\nabla_{\chi^k} \log p(W | \phi, \psi)$

The relationship $\psi_j^{k,t} = \chi_j^k + \theta_j^t$ gives $\frac{\partial \psi_j^{k,t}}{\partial \chi_j^k} = 1$ for all $t \in \{1, \ldots, T\}$, so given the independence between time periods and applying the chain rule to (32) we get

$$\frac{\partial}{\partial \chi_j^k} \log p(W | \phi, \psi) = \sum_{t=1}^{T} \frac{\partial}{\partial \psi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)} | \phi^{\cdot,t}, \psi^{\cdot,t})$$

$$= \sum_{d=1}^{D} \frac{\tilde{\phi}_k^{\gamma_d,\tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,\tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,\tau_d}} \left(\sum_{i=i_1}^{i_{L_d}} \mathbb{I}(j = w_{d,i}) - L_d \tilde{\psi}_j^{k,\tau_d} \right) \tag{34}$$

which are the elements of vector $\nabla_{\chi^k} \log p(W|\phi, \psi)$ for $j \in \{1, \ldots, V\}$.

D "Bank" additional results

We experimented with thinner time intervals on the "bank" data, as shown in Figure 6. As expected, the uncertainty increases due to less data in each interval, but the Brier scores remain comparable to those for the 20-year intervals ($BS = 0.15$ for DiSC). The 5-year graph shows an interesting and rather dramatic change in the empirical sense prevalence around 1925, which is smoothed when we take wider intervals. This sharp change could be caused by changes in the makeup of the source texts in the corpus, e.g. if a number of finance-related texts where introduced at this time. However, we did not identify a change of this sort in the corpus. We should be aware that $\tilde{\phi}$ (and $\tilde{\psi}$) measures prevalence in a language sample (i.e. the corpus) and not prevalence in historical language use itself.

In choosing an appropriate interval, two factors come into play. On the one hand, subject specialists have some understanding of the timescale of variation, and time intervals should be no greater than that scale. On the other hand, it is natural to take them as short as possible, subject to having enough data in each interval to usefully inform parameter estimates. There may be some interplay with κ_ϕ, κ_ψ and κ_θ as these diffusion parameters would naturally be scaled with interval length, so smaller for shorter intervals. When intervals are short, the fit becomes more sensitive to the choice of these hyperparameter priors. Using longer intervals makes the inference more robust to these choices. We encourage user exploration.

E Synthetic data additional results

We constructed some examples with *explicit* sense-time interaction (cf. Section 8) on a small vocabulary of 100 words, so that all words have a relatively high probability of appearing in the context of any sense. 60 context words are designed to have sense-time interaction in examples 1 and 2, the other 40 being noise, whereas all 100 are designed to have sense-time interaction in example 3. For these words, the context probability $\tilde{\psi}_v^{k,t}$ either increases or decreases with time in one sense k, whilst doing the opposite or staying constant in the other senses. It must be emphasised that these examples are highly artificial and not reflective of real-world scenarios. The DiSC and SCAN posteriors $\tilde{\psi}_v|W$ for a single word v in the three examples, together with the true $\tilde{\psi}_v$, are shown in Figure 7. The interaction effect is

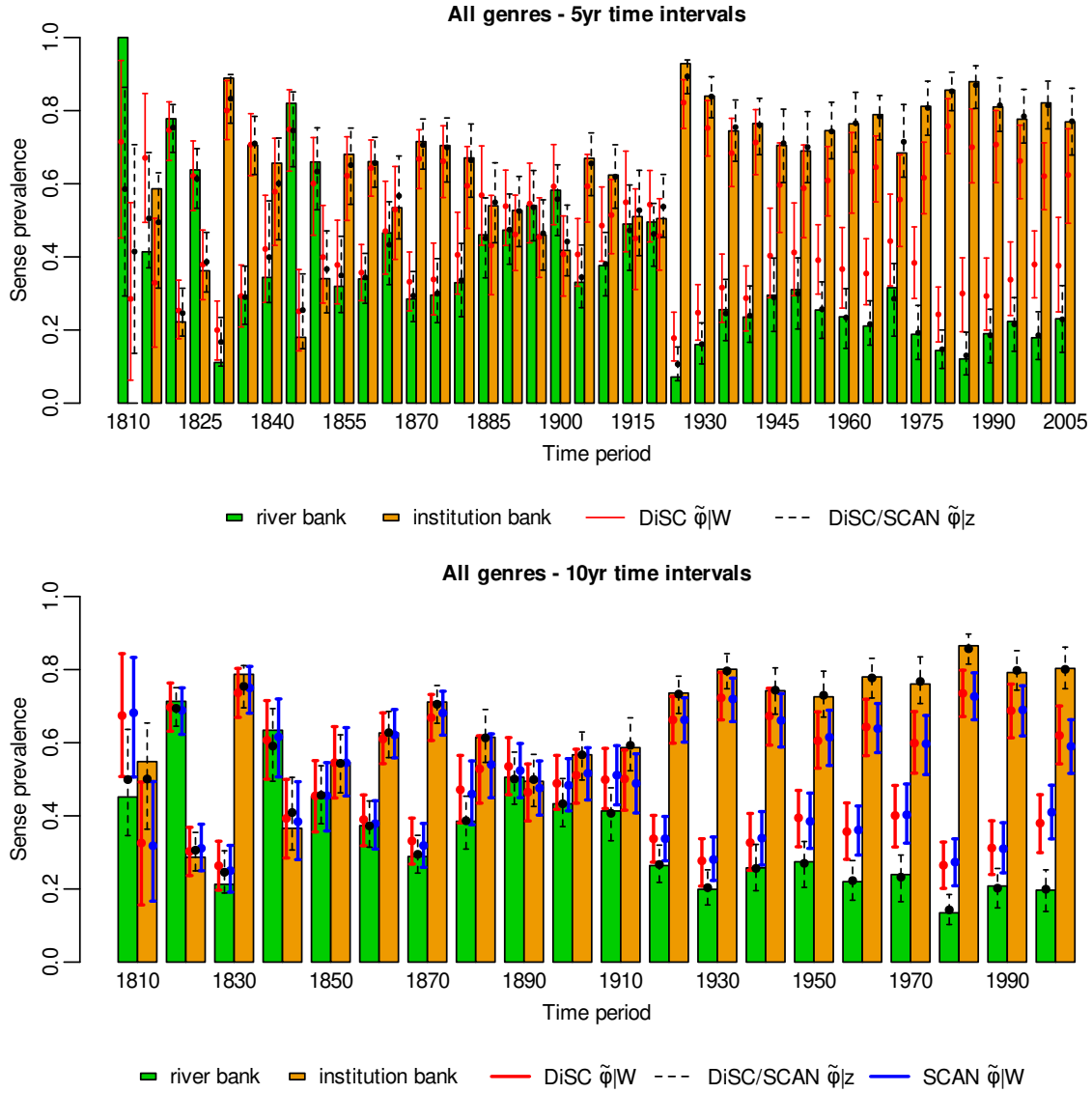


Figure 6: "Bank" expert-annotated empirical sense prevalence (coloured bars with height $N^o_{k,g,t} / \sum_{l=1}^{K} N^o_{l,g,t}$ for each k, g, t) with 95% HPD intervals (error bars) and posterior means (circles) from the model output. The first graph omits SCAN $\tilde{\phi}|W$ error bars since we could not get the MCMC to converge for SCAN.

progressively stronger from example 1 to example 3. It can be seen that the DiSC posterior $\tilde{\psi}_v^{k,t}|W$ is rather flat over time for all senses in example 2, and is thus inaccurate at time 1 for the green sense and at time 9 for the orange sense. Similar inaccuracies can be seen in example 3 for the orange sense at the start and end.

The Brier scores on these examples are shown in Table 9. The performance of DiSC and SCAN on sense-labelling is comparable in all cases. Even where the interaction effect is very strong in examples 2 and 3, DiSC does not perform much worse than SCAN. Note

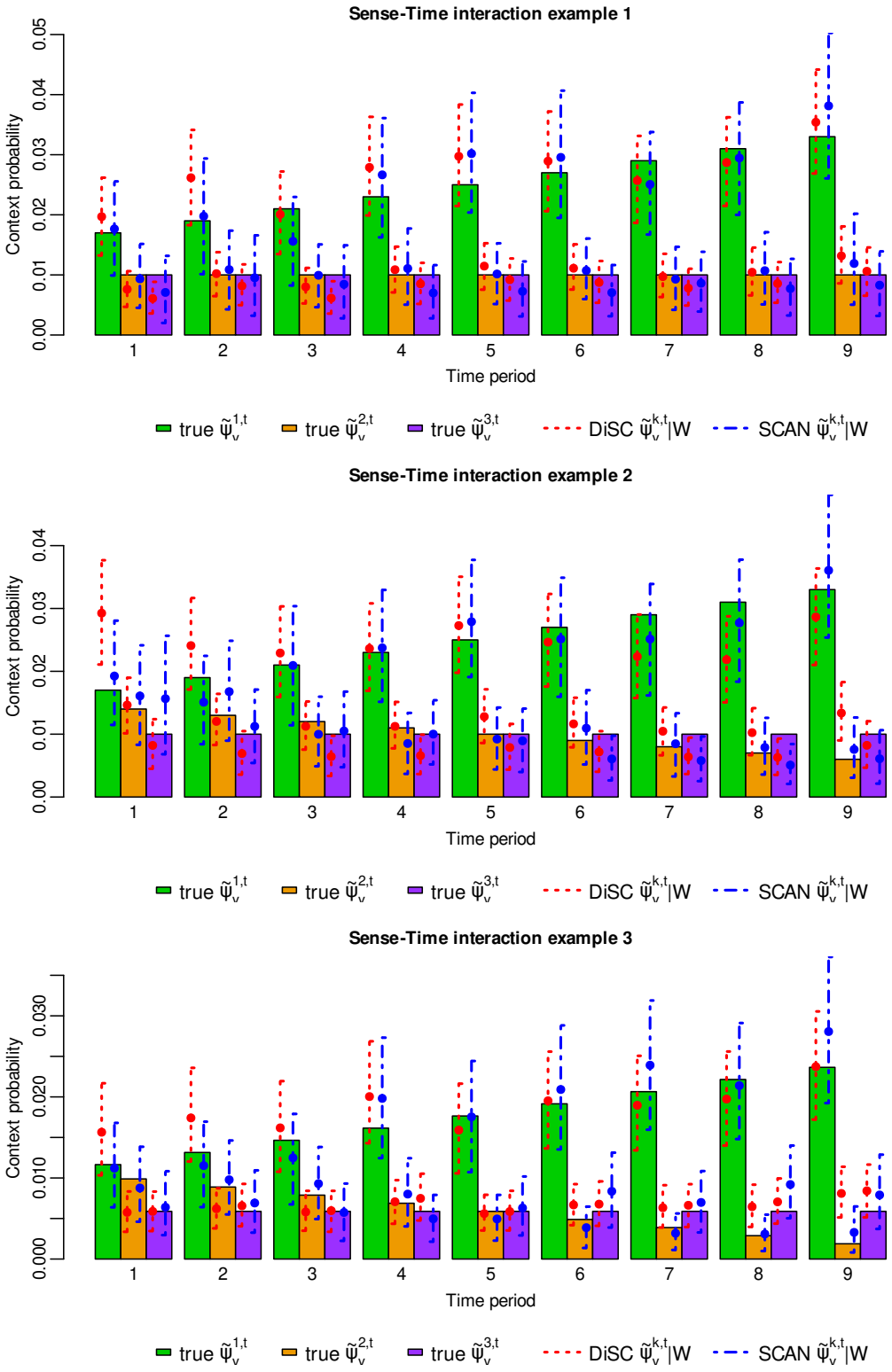


Figure 7: Examples with explicit sense-time interaction, showing the context probabilities for a typical word v under different senses over time. True probabilities are represented by the coloured bars, 95% HPD intervals by the error bars, and posterior means by the circles.

Table 9: Brier scores on synthetic data with explicit sense-time interaction

	Example 1	Example 2	Example 3
DiSC	0.382	0.386	0.252
SCAN	0.387	0.350	0.210

that having more words with explicit interaction in example 3 reduces the proportion of noisy words, and hence makes it easier to identify the sense. We attempted to construct an example where DiSC fails but SCAN performs well, but were unable to do so. (In any example where DiSC failed, so did SCAN.) Moreover, it proved difficult to get SCAN to converge on these examples due to label-switching between time periods and multi-modality. It appears that not only is sense-time interaction rare in any real-world scenario, but even where it does actually exist the gains from omitting it from the model for sense-labelling purposes far outweigh the small loss in accuracy.

F "Kosmos" additional results

It may be argued that retaining only the snippets that an expert was able to annotate from context alone (category "collocates", cf. Section 3) biases the data in the sense that the probabilities $\tilde{\phi}$ and $\tilde{\psi}$ will in general change. A related concern is that we have made the problem easier than the one we face when applying the method to data without knowing which snippets do not admit a human classification. The "kosmos" dataset contains 1,469 snippets, of which 1,144 are of the type "collocates". The remaining 325 snippets therefore represent 22% of the data, which is a very significant level of noise given the small and sparse dataset. In our experiments, we tried to follow as closely as possible what Perrone et al. (2019) did, and therefore removed this noise in order to ensure a fair comparison with GASC.

Recall that, in choosing the number of senses K, we recommend running in a semi-supervised mode: are the most frequently associated words under each sense meaningful to the user? We repeat the analysis with non-collocates included and choose K on this criterion. We find that neither DiSC nor GASC identify recognisable meanings on $K = 3$ senses, and the Brier score is higher than 0.67 (the Brier score under uniform random sense assignment).

When we run DiSC with $K = 4$ senses on all the "kosmos" data, we identify three of these senses with decoration, order or world based on the most probable context words under each sense. The fourth sense has no recognisable meaning and we think of this as a sense

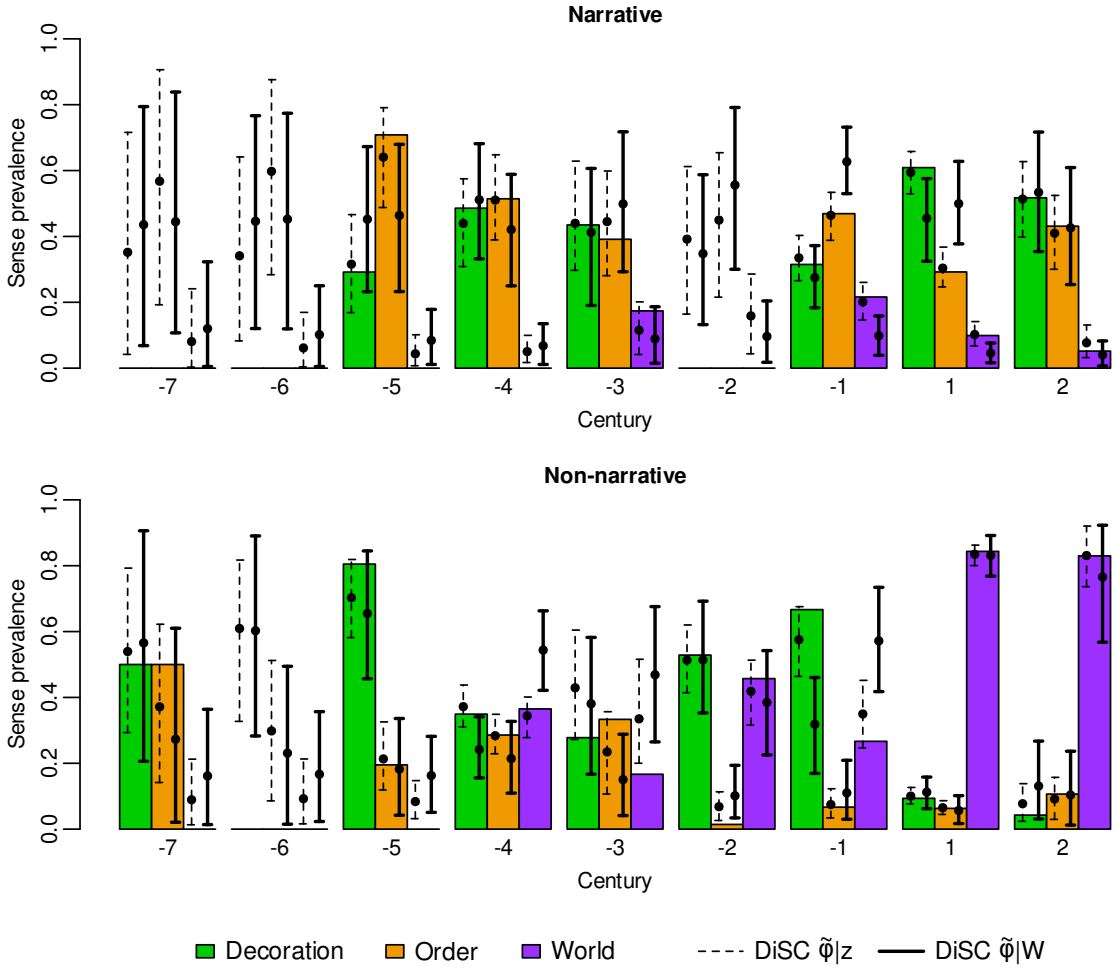


Figure 8: "Kosmos" expert-annotated empirical sense prevalence (coloured bars with height $N^o_{k,g,t} / \sum_{l=1}^{3} N^o_{l,g,t}$ for each k, g, t) with 95% HPD intervals (error bars) and posterior means (circles) from the $K = 4$ DiSC run

the model is using to capture "noise". We were unable to get the GASC model to converge even with this setting — something we found typical for GASC.

We normalise the sense probabilities $\hat{p}(z_d = k)$ (cf. equation (13)) over the three recognised senses and compute the Brier score on type "collocates" data. Thus, in evaluating performance, we exclude snippets a human could not label. This gives a score of 0.38 for $K = 4$ (slightly better than the 0.41 we advertised for $K = 3$ senses on the data with non-collocates removed). The equivalent sense prevalence graph from the DiSC $K = 4$ run is in Figure 8, and shows good agreement with the run conditioned upon the true sense labels. For completeness, if we analyse all the data with $K = 4$ senses, and compute the Brier score on all the data (i.e. not just on collocates), using the sense labels the human expert assigned to non-collocates from broader contextual considerations outside the words

around the target word, we get a score higher than 0.67. Thus, in situations where a human cannot annotate the sense based on context, our model cannot either.

CORRIGENDUM

The last part of Appendix F ("For completeness ... our model cannot either.") in the published version of this paper is incorrect. The conclusion drawn above resulted from a coding error that went undetected prior to publication. In fact, if we analyse all the data with $K = 4$ senses, and compute the Brier score on all the data, we get $BS = 0.40$. This is only slightly worse than the 0.38 score computed on just the collocates. Hence, at least for "kosmos", our model *is* able to annotate the sense based on context remarkably well even where a human cannot.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication, there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (only required where there isn't already a statement of contribution within the paper itself).

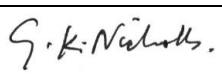
Title of paper	Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference
Publication status	Published
Publication details	Schyan Zafar and Geoff K. Nicholls (2022). Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5):1569–1604. ISSN 0035-9254. DOI: 10.1111/rssc.12591

Student Confirmation

Student name	Schyan Zafar		
Contribution to the paper	I proposed many of the research ideas for this paper; did all the coding and performed the experiments; and drafted the manuscript.		
Signature		Date	7 October 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name and title	Prof. Geoff K. Nicholls		
Supervisor comments			
Signature		Date	7 October 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek

Schyan Zafar and **Geoff K. Nicholls**

Abstract

Word meanings change over time, and word *senses* evolve, emerge or die out in the process. For ancient languages, where the corpora are often small and sparse, modelling such changes accurately proves challenging, and quantifying uncertainty in sense-change estimates consequently becomes important. GASC (Genre-Aware Semantic Change) and DiSC (Diachronic Sense Change) are existing generative models that have been used to analyse sense change for target words from an ancient Greek text corpus, using unsupervised learning without the help of any pre-training. These models represent the senses of a given target word such as "kosmos" (meaning decoration, order or world) as distributions over context words, and sense prevalence as a distribution over senses. The models are fitted using Markov Chain Monte Carlo (MCMC) methods to measure temporal changes in these representations. This paper introduces EDiSC, an Embedded DiSC model, which combines word embeddings with DiSC to provide superior model performance. It is shown empirically that EDiSC offers improved predictive accuracy, ground-truth recovery and uncertainty quantification, as well as better sampling efficiency and scalability properties with MCMC methods. The challenges of fitting these models are also discussed.

Keywords: Bayesian inference; diachronic lexical semantics; MCMC; word embeddings

1 Introduction

Languages evolve with time, and there are many facets to linguistic change that are of considerable interest to researchers across a wide range of academic disciplines. One such

facet is diachronic lexical semantics, which is the study of how the meanings of a word change with time. It is a complex phenomenon that can be influenced by a variety of factors, including social, cultural and technological changes. Computational methods for diachronic semantic change analysis have become increasingly popular in recent years, as they offer a way to study corpora of text data in a systematic and objective manner.

In this paper, the particular problem within diachronic lexical semantics that we are interested in is that of modelling polysemy, i.e. multiple meanings for a word, and homography (a subset of homonymy), i.e. words with the same spelling but different meanings. An example of a polyseme is the word "mouse" (meaning a rodent or a computer pointing device), whereas "bear" (the animal) and "bear" (to carry) are examples of homographs. (We do not distinguish between polysemy and homography in this paper.) We are interested in learning and modelling the different meanings or *senses* of given polysemous/homographic target words over time, and quantifying the uncertainty in these sense-change estimates, from text data that does not have sense labels. The unsupervised nature of this task makes it particularly challenging for small and sparse text datasets.

Perrone et al. (2019), building on the framework of the Bayesian Sense ChANge (SCAN) model of Frermann and Lapata (2016), introduced a model called Genre-Aware Semantic Change (GASC), and applied it to dynamically model the senses of selected target words from an ancient Greek text corpus. In this framework, distinct senses of a target word are represented as distinct distributions over context words, sense prevalence is represented as a distribution over target-word senses, and these distributions are allowed to evolve with time. Zafar and Nicholls (2022) further built on this framework, and modelled sense and time as additive effects in their Diachronic Sense Change (DiSC) model, which offered much improved model performance. They quantified uncertainty in the sense-change estimates using credible sets for the evolving distributions, and showed that these agreed well with those given by expert annotation for their test cases.

Quantification of uncertainty in sense-change estimates is an under-explored area within the field, yet important when working with small datasets. (Typically, any dataset with under 40 million tokens is considered 'small', whereas our ancient Greek data has around 10 million tokens.) It is particularly important for historic corpora where training data are limited and sparse (i.e. with a large proportion of infrequently used words), and inferences drawn as point estimates are therefore unreliable. Fitting models to these data is no easy task, and requires careful statistical modelling to draw accurate and meaningful inferences. The difficulty is compounded since SCAN, GASC and DiSC are all fitted to subsets of text

snippets (i.e. fixed-length windows of context words surrounding the target), thus ignoring the information contained in the wider text corpora outside the snippets, which affects the quality of fit and the inferences drawn.

In this paper, we develop EDiSC, an Embedded Diachronic Sense Change model, which extends the DiSC model by combining it with word embeddings, whereby context words are represented as vectors in an embedding space. This has two main advantages. Firstly, embeddings exploit the wider text corpora to capture useful semantic information about the context words, which is otherwise lost if we focus only on the context of a given target word. This feature of EDiSC leads to improved predictive accuracy, ground-truth recovery and uncertainty quantification. We demonstrate this on challenging test cases from ancient Greek, which is the main focus of this paper, as well as an easier test case from English. Secondly, the dimension of the embedding space is lower than the vocabulary size, and is typically held constant even against an increasing vocabulary size. This results in more efficient Monte Carlo sampling and scalability properties. We demonstrate this via experiments on both real and synthetic data.

The main novelty of this paper is to bring together two previously separate approaches for modelling sense change, namely topic-based models and embedding-based models, to improve model performance. This was not straightforward, as careful statistical modelling and consideration of the model structure are required to make things work. Furthermore, compared to Zafar and Nicholls (2022), we treat model selection explicitly, as well as giving a more thorough treatment of model-fitting methods and convergence issues. We also consider additional ancient Greek test cases, in particular one where accurate sense-change analysis was not achievable using DiSC, but is now possible.

The rest of this paper is organised as follows. In Section 2 we describe the datasets used and our modelling/inference objectives. In Section 3 we discuss related work and where our paper fits within the wider literature. In Section 4 we describe our new model and how it relates to existing models. We also discuss the embeddings used and inference for our model. In Section 5 we show the results of applying the models to our test cases, and assess model performance on predictive accuracy and true-model recovery. We also discuss model selection issues, and sampling efficiency and scalability. Section 6 concludes with a brief discussion. The Appendix contains a discussion on hyperparameter settings, further results not included in the main body, and some other technical details.

Table 1: Example text snippets for target word "bank" taken from Zafar and Nicholls (2022, Table 1). The words are lemmatised, and stopwords in blue and infrequent words in orange are dropped to get the data W.

"... China. The Yellow River had burst its banks, submerging vast areas of farmland, washing away ..."
"... to examine whether institutions like the World Bank and the International Monetary Fund needed restructuring ..."
"... subject of continuing specie payments. Though the Bank of the United States had previously determined ..."

2 Data and problem setting

Whilst the models discussed in this paper could be applied quite widely to gain useful insights on potential target words with sense change, in this paper we are primarily interested in modelling sense change for three target words from the Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018): "kosmos", "mus" and "harmonia". We work with this corpus as it is small and sparse, and therefore challenging for existing methods. We choose these particular test cases since we have the ground truth readily available for these target words, which we can use to accurately assess model performance. Otherwise, we would be limited to qualitative measures of model performance, which are not as convincing. (The ground truth would, of course, not be available for a 'real-use' case.) Fitting the models for these test cases is particularly challenging; so, for demonstration purposes, we also use a simple test case, "bank", from the Corpus of Historical American English (COHA, Davies 2010), where fitting the models is much easier.

COHA consists of 400+ million tokens, from which Zafar and Nicholls (2022) randomly select 3,685 snippets of 14 *lemmatised* context words around the target word "bank", and annotate them with the sense of riverbank or financial institution. (A *lemma* is the root form of the word; so for example "branch" is the lemmatised form of "branching".) The annotation is done to provide ground truth for testing, and is not used in the analysis. Some example snippets are shown in Table 1. Of these sense-labelled snippets, 3,525 snippets of the type *collocates* are used for evaluation, i.e. the snippets where the correct target-word sense could be identified by the reader from context alone. Stopwords and infrequently used words are dropped, leaving a vocabulary of 973 words, and punctuation and sentence boundaries are ignored. Time is divided into ten discrete and contiguous 20-year periods from 1810 to 2010, and a single (combined) genre is used. The choices of snippet length, time discretisation and genre covariate are subjective. These could be informed by domain

knowledge or exploratory analysis, or pre-set based on the quantities of interest. We use the same choices as in previous work since that is not our focus, and they are in any case well informed.

For the ancient Greek data, which is a much smaller corpus with around 10 million tokens, expert sense-annotation is provided by Vatri et al. (2019) for our three target words, with an accompanying explanation by McGillivray et al. (2019, Section 4). Each target word has three true senses: "kosmos" (decoration, order, world), "mus" (mouse, muscle, mussel), "harmonia" (abstract, concrete, musical). Following Perrone et al. (2019), time periods are discrete and contiguous centuries from 800 BC to 400 AD, and the categorical genre covariates are narrative and non-narrative for the "kosmos" data, and technical and non-technical for the "mus" and "harmonia" data. We use snippets of 14 lemmatised context words in line with Zafar and Nicholls (2022). However, in contrast to their setup, we filter the vocabulary based on a minimum count of 10 occurrences in the entire corpus. This filtering is used to reduce noise from rare words. In order to avoid filtering words that are rare in the target-word context, but common outside that context, we base these counts on frequency in the corpus rather than just the snippets. Further, we filter out a smaller list of stopwords than that used by Zafar and Nicholls (2022), preferring to retain some potentially noisy context words rather than lose some context words that may be semantically important. (Previous work identified stopwords using part-of-speech tags, as well as three lists: (a) Berra (2018), and (b) `misc` and (c) `stopwords-iso` from the R package `Stopwords` (Benoit et al., 2020). We continue to use part-of-speech tags as before, but only use the last of the three lists.) Therefore, whilst the Greek datasets used are comparable, they are not identical; and we report all results on the reprocessed data. Also, Zafar and Nicholls (2022) only analysed "kosmos", whereas we now fit the models to all three datasets.

Table 2 summarises the four datasets and some notation that we will use in this paper. Note that the numbers differ from McGillivray et al. (2019, Table 3) because of the slightly different approaches used to extract the snippets. The data W for a given target word consists of D snippets of length L containing context words $w_{d,i}, d \in 1:D, i \in 1:L$, sampled from the vocabulary $1:V$. A subset of D' snippets is of the type collocates. A context position may be empty in the filtered data if a stopword or infrequent word has been dropped; so the bag of words retained in snippet d, denoted W_d, has variable size L_d. The target word itself is excluded. Snippet d has deterministic mappings to genre $\gamma_d \in 1:G$ and time $\tau_d \in 1:T$, which are known from the text that generated the snippet. The target-word sense $z_d \in 1:K$ for snippet d is in general unknown. We use K' to refer to the number of

Table 2: Data summary

Data	(W)	bank	kosmos	mus	harmonia
Snippets	(D)	3,685	1,469	214	653
Collocates	(D')	3,525	1,144	118	451
Vocabulary size	(V)	973	2,904	899	1,607
Snippet length	(L)	14	14	14	14
True senses	(K')	2	3	3	3
Model senses	(K)	2	4	3	4
Text genres	(G)	1	2	2	2
Time periods	(T)	10	9	9	12

true target-word senses, whereas we fit the models using K senses, which may be different to K'. The choice of K is discussed in Section 5.2.

In practice, given a text corpus and target word of interest, the workflow to extract data W might go as follows: identify target-word occurrences in corpus $\rightarrow$ set snippet length $L \rightarrow$ extract snippets and meta-data (time and genre) $\rightarrow$ lemmatise snippets $\rightarrow$ set criteria for stopwords and minimum frequency $\rightarrow$ filter these words from snippets $\rightarrow$ discretise time and set genre covariates to get data W. Given these data, our goal is to dynamically model the target-word senses and sense prevalence where, separately for each target word, we define sense k at time t as the distribution $\tilde{\psi}^{k,t} = \tilde{\psi}^{k,t}_{1:V}$ over context words $1:V$, and we define sense prevalence for genre g at time t as the distribution $\tilde{\phi}^{g,t} = \tilde{\phi}^{g,t}_{1:K}$ over senses $1:K$. We would like to infer the context-word probabilities $\tilde{\psi}^{k,t}_v = p(w_{d,i} = v | z_d = k, \tau_d = t)$ for each (v, k, t) word-sense-time triple, the sense-prevalence probabilities $\tilde{\phi}^{g,t}_k = p(z_d = k | \gamma_d = g, \tau_d = t)$ for each (k, g, t) sense-genre-time triple, and quantify the uncertainty therein. Then, $\tilde{\psi}$ and $\tilde{\phi}$ would together encapsulate the diachronic sense change over time t.

3 Related work

The field of computational lexical semantic change is quite rich and varied. Approaches include topic-based models (e.g. Frermann and Lapata 2016; Perrone et al. 2019; Zafar and Nicholls 2022), on which our current work builds, as well as graph-based models (e.g. Mitra et al. 2014, 2015; Tahmasebi and Risse 2017) and embedding-based models (e.g. Kulkarni et al. 2015; Hamilton et al. 2016; Rudolph and Blei 2018; Dubossarsky et al. 2019; Yüksel et al. 2021). Comprehensive surveys tracing the history of the subject up to 2021 are given by Tang (2018), Kutuzov et al. (2018) and Tahmasebi et al. (2021).

In recent years, embedding-based methods have dominated the landscape of natural lan-

guage processing (NLP). Word embeddings are representations of words in low-dimensional real vector spaces, and embedding models produce these mappings whilst capturing the words' semantic and (sometimes) syntactic relationships. Popular traditional word embedding models include Google's Word2vec (Mikolov et al., 2013a,b,c), Stanford's GloVe (Pennington et al., 2014) and Facebook's FastText (Bojanowski et al., 2017), but there are many others, including variants of these models. Traditional word embeddings are learnt based on patterns of word co-occurrences in text corpora. These are typically context-independent, i.e. the embedding for a word is the same regardless of the context in which it is used. This is a limitation for diachronic sense change modelling, since many words have multiple senses which traditional embeddings fail to capture.

Contextualised word embedding models have been developed in recent years to overcome this limitation by attempting to learn word representations that capture word meanings in their given contexts. Some early contextualised embedding models such as Context2vec (Melamud et al., 2016) and ELMo (Embeddings from Language Model, Peters et al. 2018) are based on recurrent neural network architectures, whereas most modern contextualised embedding models are based on transformer architectures, including the popular GPT (Generative Pre-trained Transformer, Radford et al. 2019) and BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. 2019). Recent surveys discussing modern embedding models, including many variants of the ones mentioned here, are given by Naseem et al. (2021) and Apidianaki (2022).

Word embeddings can be used for a number of NLP tasks, including tasks within computational semantics. Contextualised word embeddings in particular can be used for diachronic sense change modelling, and Montanelli and Periti (2023) survey the recent advances in this area. Most modern methods are data-intensive, and tend to rely on pre-trained models learnt using huge amounts of data. Indeed, virtually all of the approaches for modelling semantic change summarised in Montanelli and Periti (2023, Tables 3-4) rely on some form of pre-training, fine-tuning or domain-adaptation. This restricts the usability of these methods for analysing small, sparse or historic corpora, such as the ancient Greek data used in this paper, for which training data are limited, and pre-trained models are not available.

Word sense disambiguation (WSD) is the related task of identifying the correct sense of a word in a given context. Recent approaches for WSD tend to be either knowledge-based or supervised, utilising sense inventories or sense-annotated corpora, and typically rely on pre-trained language models (Bevilacqua et al., 2021). This again restricts the usage of WSD approaches for our purpose.

Moreover, the goal of WSD is not diachronic sense change modelling, but rather sense induction from context. The usages of a target word may be clustered according to the induced senses, which could then be used for drawing inferences about diachronic sense change, but a clustering approach is not conducive to interpretability or quantification of uncertainty. Graph-based approaches have a similar drawback. On the other hand, generative models designed for the purpose of diachronic sense change modelling have parameters with physical interpretations, and are natural in the context of Bayesian measures of uncertainty.

Probabilistic topic models are generative bag-of-words models that are widely used to infer themes or *topics* from a collection of documents (Blei, 2012). Latent Dirichlet Allocation (LDA, Blei et al. 2003), one of the best-known topic models, is a generative model that represents topics as distributions over words, and documents as mixtures over topics. The dynamic topic model (DTM, Blei and Lafferty 2006) extends LDA to model each topic as a time series, whereas the embedded topic model (ETM, Dieng et al. 2020) extends LDA to incorporate word embeddings within the model. The ETM outperforms LDA in terms of topic quality and predictive performance since it benefits from the extra semantic information captured within the word embeddings that is not present in LDA. Finally, the dynamic embedded topic model (D-ETM, Dieng et al. 2019) brings together the DTM and the ETM by adding a time dimension to the ETM. The D-ETM similarly outperforms the DTM, while requiring less time to fit. New variants of topic models continue to emerge (Churchill and Singh, 2022), showing their continued relevance and importance.

The DTM was adapted by Frermann and Lapata (2016) to capture the evolving *senses* of a given target word in the SCAN model. A snippet under SCAN is the context surrounding an instance of the target word, and may be compared to a document under the DTM. The distinction is that, whilst a document under the DTM is a mixture over multiple topics, each snippet under SCAN has only one sense. The sense prevalence was allowed to vary according to the text genre by Perrone et al. (2019) in the GASC model. Each sense in the SCAN and GASC models has an independent time-evolution, whereas Zafar and Nicholls (2022) imposed an additive structure between the sense-effect and the time-effect in the DiSC model. Therefore, under DiSC, differentiated target-word senses have a common time-evolution, which drastically reduces the dimension of the parameter space. DiSC captures the sense-dependence of snippets better than SCAN/GASC, and results in more accurate sense induction and diachronic sense change modelling.

The relationship between DiSC and the EDiSC model we introduce in this paper is analogous to the relationship between the DTM and the D-ETM. In both cases, the latter extends

the former through incorporating word embeddings within the model. Indeed, EDiSC has been inspired by the D-ETM, and offers similar benefits over DiSC as the D-ETM does over DTM.

4 Model and inference

We first describe DiSC, highlighting its connection to GASC and SCAN, before introducing our new EDiSC model. We then discuss the embeddings used and the inference methods for these models. Prior elicitation with respect to the hyperparameters is discussed in Appendix A.

4.1 Background

DiSC is a generative bag-of-words model for the context words surrounding a given target word, comprising a prior model and an observation model.

Under the DiSC observation model, to generate snippet d, we first sample the sense $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}(\tilde{\phi}^{\gamma_d,\tau_d})$ and then independently sample the words $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}(\tilde{\psi}^{z_d,\tau_d})$ for each context position i in the snippet. The context positions occupied by words in each snippet d are a subset $\{i_1, \ldots, i_{L_d}\}$ of size L_d drawn from the set $\{1, \ldots, L\}$, where L_d is the number of words retained in snippet d after filtering out stopwords and infrequent words. The order of context words is irrelevant due to the bag-of-words assumption. The authors of DiSC treat stopwords and infrequent words explicitly, whereas the authors of GASC and SCAN do not. This is a difference in how context words are registered within the snippets rather than a difference in the models. The observation models are identical for DiSC, GASC and SCAN.

Under the DiSC, GASC and SCAN prior models, $\tilde{\phi}^{g,t}$ and $\tilde{\psi}^{k,t}$ are defined as softmax transforms of real-valued vectors $\phi^{g,t}$ and $\psi^{k,t}$ respectively, that is

$$\tilde{\phi}^{g,t} = \frac{\exp(\phi^{g,t})}{\sum_{k=1}^{K} \exp(\phi_k^{g,t})} \quad \text{and} \quad \tilde{\psi}^{k,t} = \frac{\exp(\psi^{k,t})}{\sum_{v=1}^{V} \exp(\psi_v^{k,t})}. \tag{1}$$

Under the DiSC prior model, for each genre g, the prior on $\phi^{g,t}$ is an AR(1) time series with stationary distribution $\mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$. An additive structure $\psi^{k,t} = \chi^k + \theta^t$ is imposed on ψ. A $\mathcal{N}(0, \text{diag}(\kappa_\chi))$ prior is placed on χ^k, and an AR(1) prior is placed on θ^t with stationary distribution $\mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$. Prior hyperparameters $\kappa_\phi, \kappa_\theta, \kappa_\chi, \alpha_\phi, \alpha_\theta$ are fixed, and $|\alpha_\phi| < 1, |\alpha_\theta| < 1$ to ensure proper priors with mean reversion.

In contrast, GASC and SCAN have no additive structure on $\psi^{k,t}$ and model it as K independent Gaussian time series (without mean-reversion). All time series priors used in GASC

Algorithm 1 EDiSC: generative model

PRIOR MODEL

get word embeddings matrix ρ
fix hyperparameters $\kappa_\phi, \kappa_\theta, \kappa_\chi, \kappa_\varsigma, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
draw bias or correction parameter $\varsigma|\kappa_\varsigma \sim \mathcal{N}(0, \text{diag}(\kappa_\varsigma))$
for genre $g \in 1:G$ **do**
    draw initial sense prevalence parameter $\phi^{g,1}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
    for time $t \in 2:T$ **do**
        draw sense prevalence parameter $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \text{diag}(\kappa_\phi)\right)$
    end for
end for
draw initial time embedding $\theta^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$
for time $t \in 2:T$ **do**
    draw time embedding $\theta^t|\theta^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \theta^{t-1}, \text{diag}(\kappa_\theta)\right)$
end for
for sense $k \in 1:K$ **do**
    draw sense embedding $\chi^k|\kappa_\chi \sim \mathcal{N}(0, \text{diag}(\kappa_\chi))$
    for time $t \in 1:T$ **do**
        set sense-time embedding $\xi^{k,t} = \chi^k + \theta^t$
        set context-word probability parameter $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$
    end for
end for
transform ϕ and ψ into probabilities $\tilde{\phi}$ and $\tilde{\psi}$ using softmax (1)

OBSERVATION MODEL

for snippet $d \in 1:D$ (genre γ_d, time τ_d, length L_d) **do**
    draw sense assignment $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d,\tau_d}\right)$
    for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
        draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d,\tau_d}, \ldots, \tilde{\psi}_V^{z_d,\tau_d}\right)$
    end for
end for

and SCAN are improper without a stationary distribution: $\phi^{g,t}|\phi^{g,-t}, \kappa_\phi \sim \mathcal{N}\left(\frac{1}{2}(\phi^{g,t-1} + \phi^{g,t+1}), \kappa_\phi\right)$ and $\psi^{k,t}|\psi^{k,-t}, \kappa_\psi \sim \mathcal{N}\left(\frac{1}{2}(\psi^{k,t-1} + \psi^{k,t+1}), \kappa_\psi\right)$. Also, $\kappa_\phi \sim \text{Inv Gamma}(a, b)$ in GASC and SCAN. Finally, SCAN has the number of genres $G = 1$ fixed, whereas GASC admits $G \geq 1$; otherwise, the two models are identical.

4.2 Embedded DiSC (EDiSC) model

The EDiSC generative model is given in Algorithm 1 and a plate diagram is shown in Figure 1.

The key idea behind EDiSC is to introduce word embeddings into the model. As such, EDiSC has the same observation model as DiSC, as well as the same prior model for ϕ. However,

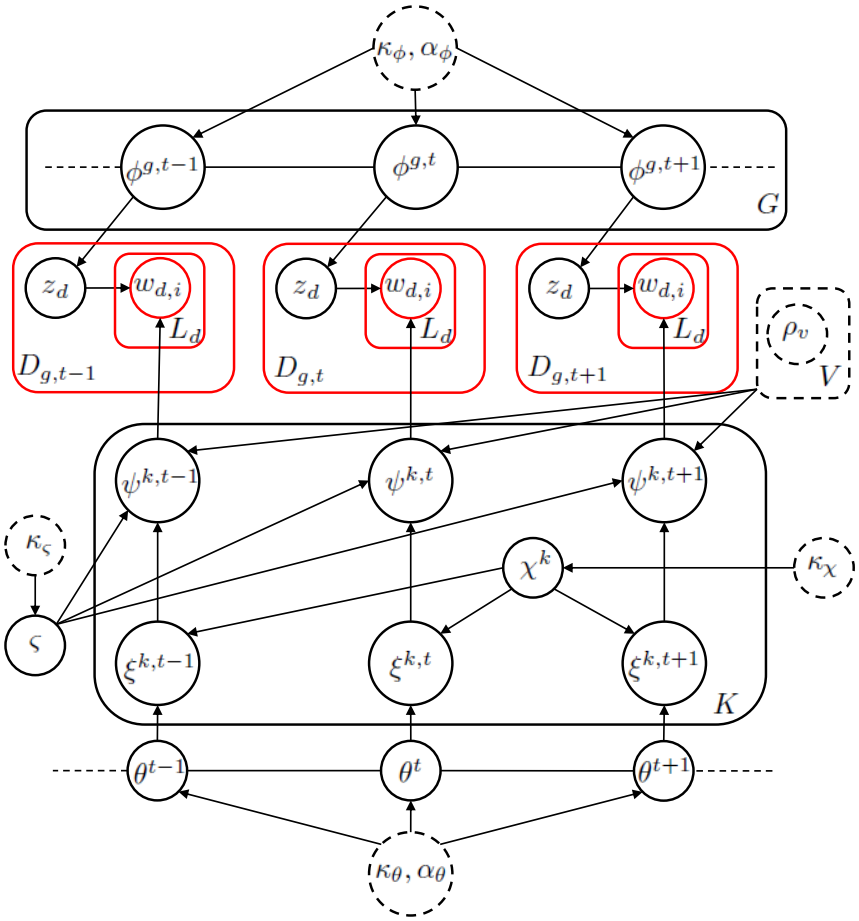


Figure 1: EDiSC plate diagram for three time periods. Dashed nodes are constant parameters, solid black nodes are latent variables and solid red nodes are observed variables. $D_{g,t}$ is the number of snippets for genre g at time t.

we now define $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$, where ρ is a $V \times M$ matrix of word embeddings, $\xi^{k,t}$ is an M-dimensional sense-time embedding vector for sense k at time t, and ς is a V-dimensional bias or correction parameter. The matrix ρ has row vectors $\rho_v = (\rho_{v,1}, \ldots, \rho_{v,M})$, which are M-dimensional word embeddings for words $v \in 1:V$ in the lemmatised vocabulary. The parameter ξ is decomposed as $\xi^{k,t} = \chi^k + \theta^t$, where χ^k is an M-dimensional sense embedding for sense k, and θ^t is an M-dimensional time embedding for time t. We place a $\mathcal{N}(0, \text{diag}(\kappa_\chi))$ prior on χ^k and an AR(1) prior on θ^t with stationary distribution $\mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$. These priors are functionally the same as in DiSC. However, whilst χ^k and θ^t are vectors in a V-dimensional space under DiSC, they are now vectors in the M-dimensional embedding space under EDiSC. Prior hyperparameters are set using quantiles as discussed in Appendix A.

The bias or correction term ς is used because, due to the coupling induced by the embedding

structure, if two words x, y have similar word embeddings ρ_x, ρ_y, then in the absence of a correction term the word parameters $\psi_x^{k,t}, \psi_y^{k,t}$ (and hence the probabilities $\tilde{\psi}_x^{k,t}, \tilde{\psi}_y^{k,t}$) would also be very similar. However, we could have context words x, y appearing with different frequencies in the snippets despite the similarity in embeddings. The correction terms ς_x, ς_y serve to decouple the words and allow them to appear with different probabilities. Put another way, if the smaller M-dimensional $\xi^{k,t}$ vector is able to capture the variation in the larger V-dimensional $\psi^{k,t}$ vector via the product $\rho\xi^{k,t}$, then ς would not be significant; otherwise, ς allows the extra variability required to model $\psi^{k,t}$ accurately.

We briefly mention some other embedded models that we experimented with but discarded. Firstly, we tried $\psi^{k,t} = \rho\chi^k + \theta^t$, using only a sense embedding and not a time embedding, whose predictive performance (as measured by Brier scores) was not as good. Next, we tried $\psi^{k,t} = \rho\xi^{k,t} = \rho(\chi^k + \theta^t)$ without the correction term ς, which generally worked well and gave results comparable to those from our chosen final model. However, an issue common to both these alternatives was that the posteriors under these models were more susceptible to multimodality for the sparse Greek data. Hence, using Markov Chain Monte Carlo (MCMC) to find the right mode was less straightforward. Finally, we tried $\psi^{k,t} = \rho\xi^{k,t}$ without imposing an additive structure $\xi^{k,t} = \chi^k + \theta^t$, so essentially an embedded version of GASC. The multimodality in the posterior for this model was much worse and, even when the MCMC did converge, the performance was much inferior to our chosen model.

4.3 Embeddings

We learn the word embeddings $\rho_v, v \in \{1, \ldots, V\}$, using GloVe. Given the model setup in Section 4.2, a traditional embedding model with one vector representation per word is most suitable for use with EDiSC, as we can learn the embeddings independently and plug them into the model. A contextualised embedding model would not be straightforward to use here, although it would be interesting to investigate how it could work within our framework. However, that is beyond the scope of this paper.

Whilst we deliberately choose a popular traditional embedding model, there is no particular reason for choosing GloVe over Word2vec or FastText, and they should all give similar results in our problem setting. We favour an accessible and universally applicable model over a corpus-specific and/or task-specific one (e.g. Rodda et al. 2019) for consistency between our English and ancient Greek test cases, and because the code to implement GloVe is readily available.

The R code to implement GloVe was adapted from Selivanov (2022). We learnt the embeddings using the settings in Pennington et al. (2014): $x_{\max} = 100$, $\alpha = 3/4$, initial learning rate of 0.05, convergence tolerance of 0.01, context window of 10 words on either side, and adding together the 'in' and 'out' vectors. We pruned the vocabulary based on a minimum count of 10 in the entire corpus in order to reduce noise from rare words. We filtered out stopwords and lemmatised the words before learning the embeddings, thus tailoring the model to our problem setting by sacrificing syntactic information in favour of semantic.

In general, there is no universally optimal method for choosing the embedding dimension M. The choice is usually specific to the corpus and/or task. Even then, it is not an exact science and requires judgement. Clearly, a lower dimension would have benefits in terms of computational cost and memory requirements, whereas a larger dimension may capture more semantic information but also introduce noise via overfitting. Pennington et al. (2014, Section 4.4) suggest that there are "diminishing returns for vectors larger than about 200 dimensions", and Yin and Shen (2018) suggest a method for choosing M based on minimising a loss function. As a rule of thumb, between 50 and 300 is considered an appropriate range (Patel and Bhattacharyya, 2017). We fit our models for M equal 50, 100, 200 and 300, and report the results for these choices in Section 5.1 below. We also include a brief discussion on how we select M out of these choices in Section 5.2.

4.4 Inference

The parameters of interest are the probability arrays $\tilde{\phi}$ and $\tilde{\psi}$, but it is more convenient to target $\phi, \theta, \chi, \varsigma$ given the snippet data W. The posterior for EDiSC is defined by

$$\pi(\phi, \theta, \chi, \varsigma | W) \propto \pi(\phi)\pi(\theta)\pi(\chi)\pi(\varsigma)p(W|\phi, \psi), \tag{2}$$

where the likelihood

$$p(W|\phi, \psi) = \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_d = k|\phi)p(W_d|z_d = k, \psi) = \prod_{d=1}^{D} \sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{w \in W_d} \tilde{\psi}_w^{k, \tau_d} \tag{3}$$

remains unchanged compared to DiSC in terms of ϕ and ψ. The likelihood (3) is obtained by marginalising $p(W, z|\phi, \psi)$ over the unknown sense labels $z = (z_1, \ldots, z_D)$. This leaves us with a posterior defined over continuous variables only, which is convenient for variational inference and gradient-based Monte Carlo. The 'observable' parameters $\tilde{\phi}$ and $\tilde{\psi}$ are identifiable (up to label switching) whereas the logit-scale parameters $\phi, \theta, \chi, \varsigma$ are not. However, non-identifiability at that level is not a concern, since we only care about the interpretable probability arrays, and non-identifiability does not cause any convergence problems in our experiments.

The posterior (2) is quite challenging to sample. This is because of ridge structures and multimodality in the posterior, especially for the sparse ancient Greek data, but also to some extent for the less sparse "bank" data. Before drawing any inferences, it is important to ensure that any method targeting the posterior has converged, that is, different starting configurations and random-number-generator seeds result in the same posterior distribution. Variational methods targeting (2) are highly sensitive to the starting configuration for the optimisation, and therefore fail this test. Using Stan's Automatic Differentiation Variational Inference (ADVI, Kucukelbir et al. 2015) for instance, we are unable to obtain consistent posteriors for most choices of M for all of our test cases. In any case, variational methods typically target local optima: even when they are adequate for predictive inference (targeting the posterior predictive for W), quantification of uncertainty in our target parameters $\tilde{\phi}$ and $\tilde{\psi}$ is poor.

MCMC is the method of choice when sampling from (2), with gradient-based MCMC as described in Zafar and Nicholls (2022, Appendix C) working particularly well, since this provides better coverage of the posterior space and more accurate quantification of uncertainty. We implement all our samplers in the R programming language (R Core Team, 2023) and the scripts are available online. We use Metropolis-Adjusted Langevin Algorithm (MALA, Roberts and Tweedie 1996, Roberts and Rosenthal 2002) and Hamiltonian Monte Carlo (HMC, Duane et al. 1987, Neal 2011, Beskos et al. 2013), as well as the No-U-Turn sampler (NUTS, Hoffman and Gelman 2014) from the Stan software (Stan Development Team, 2023b,a).

In our implementations of MALA and HMC, described in more detail in Appendix D, we use analytically derived gradients; and we target (2) in a Metropolis-within-Gibbs fashion, alternately sampling each variable given the others. Stan, on the other hand, uses automatic numerical differentiation and targets the entire posterior at once. The choice of sampler is not important in any converged MCMC run, since all samplers should converge to the same posterior. However, due to the ridge structures in (2), the samplers may sometimes get stuck in a metastable state and fail to converge. (A metastable state is a region of high density but low total probability mass, separated from the rest of the posterior by regions of low density.) Our HMC sampler generally gives the most consistent performance in this respect. Some of the convergence issues experienced in fitting these models to our test cases are discussed in Appendix C. In Section 5 below, we only report results from converged HMC runs unless otherwise stated.

Table 3: Brier scores for test data using different models with HMC sampling. Best scores are in blue, and scores from the models selected using the criteria set out in Section 5.2 (so independently of the Brier scores) are boxed. 95% confidence intervals for the reported scores are typically within BS $\pm$ 0.005.

	bank	kosmos	mus	harmonia
Uniform predictions	0.500	0.667	0.667	0.667
GASC/SCAN	0.172	*	*	*
DiSC	0.150	0.371	0.203	0.639
EDiSC ($M = 50$)	0.139	0.349	0.135	*
EDiSC ($M = 100$)	0.139	0.329	0.093	*
EDiSC ($M = 200$)	0.133	0.332	0.099	*
EDiSC ($M = 300$)	0.143	0.326	0.101	0.584

* Not converged: MCMC runs from different starting configurations lead to different equilibrium distributions

5 Application and results

We first consider the models' predictive accuracy on held-out true sense labels. We then discuss model selection issues with respect to the choice of K and M. Next, we assess the inferred sense-prevalence evolution of our target words against the ground truth. Finally, we analyse the sampling efficiency and scalability properties of EDiSC vs. DiSC using MCMC methods.

5.1 Predictive accuracy

We use the held-out true sense labels $o_d \in \{1', \ldots, K'\}$, where $d \in \{1', \ldots, D'\}$ are the indices for the type-collocates snippets (cf. Section 2), to assess the models' predictive performance. We quantify this using the Brier score

$$\mathrm{BS} = \frac{1}{D'} \sum_{d=1'}^{D'} \sum_{k=1'}^{K'} \left(\hat{p}(z_d = k) - \mathbb{I}(o_d = k)\right)^2,$$

a proper scoring rule for multi-category probabilistic predictions $\hat{p}(z_d = k)$, ranging from 0 (best) to 2 (worst). Here, $\hat{p}(z_d = k)$ is the estimated value of $\mathbb{E}_{\phi,\psi|W}\big(p(z_d = k|W_d, \phi, \psi)\big)$, computed on the MCMC output by normalising

$$p(z_d = k|W_d, \phi, \psi) \propto \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{w \in W_d} \tilde{\psi}_w^{k, \tau_d} \tag{4}$$

over $k \in \{1, \ldots, K\}$. Recall that we have K' true senses, whereas we run the models using K senses (with $K \geq K'$), so modelled senses may be grouped together to map them onto the true senses. This is discussed further in Section 5.2.

The results for each model and dataset, obtained using converged HMC runs, are summarised in Table 3. Under uniform predictions, if we set $\hat{p}(z_d = k) = \frac{1}{K'}$ for all d, k, we get $\text{BS} = \left(1 - \frac{1}{K'}\right)^2 + (K' - 1)\left(\frac{1}{K'}\right)^2 = 0.5$ in the case of $K' = 2$ for "bank" or 0.667 in the case of $K' = 3$ for the other datasets; so models must produce scores lower than these in order to be useful. In all our test cases, EDiSC with an appropriate dimension M offers a clear improvement over DiSC. Recall that DiSC was already an improvement over GASC/SCAN, so we treat DiSC as the only baseline and do not compare against GASC/SCAN hereinafter.

The extent of improvement provided by EDiSC over DiSC varies depending on the complexity of the target dataset. For the simpler "bank" test case, with a large number of snippets relative to the vocabulary size, the two senses are quite distinct and well-informed by the snippet data, so there is limited scope for improvement. For the more challenging Greek test cases, there are underlying semantic relationships in the wider corpus that cannot be captured through the snippet data alone, but are learnt via the context-word embeddings, so the improvement is more pronounced. The improvement is greatest for "mus", where the small data size limits DiSC performance, but the inclusion of embeddings in EDiSC counteracts by providing added structure via the learnt context-word relationships latent in the corpus.

Note that our goal is to infer $\tilde{\phi}$ and $\tilde{\psi}$ given the snippets, rather than WSD. We obtain the probabilistic predictions (4) as a free byproduct from the converged posteriors, which *could* be used for WSD, but we are not in essence trying to tag instances of our target words with their correct sense. Typically for WSD (and related NLP tasks) with a benchmark dataset, the precision, recall and their harmonic mean (F1 score) are used for assessment. These are appropriate when making 0/1 predictions on sense labels. However, in the case of probabilistic predictions such as ours, a scoring rule is more appropriate. Whilst any proper scoring rule could be used for this purpose, some common alternatives being logarithmic or spherical score, we choose the Brier score for its attractive properties: it is equivalent to the widely used and well-understood mean squared error; the contribution from any one prediction is bounded (in contrast to logarithmic score, which can be unstable); and it has symmetrical penalties for over-confidence in the wrong prediction and under-confidence in the right prediction (in contrast to both logarithmic and spherical scores).

5.2 Model selection

Two important modelling choices to make are the number of model senses K and the embedding dimension M. We first consider the choice of K.

Table 4: "Bank" top 10 context words for each model sense under EDiSC

Sense	Top 10 context words for EDiSC $M = 200$ with $K = 2$ senses									
1	river	stream	water	stand	tree	leave	creek	day	land	reach
2	bank	national	note	money	deposit	reserve	credit	saving	loan	federal
Sense	Top 10 context words for EDiSC $M = 200$ with $K = 3$ senses									
1	river	stream	water	stand	tree	leave	creek	bank	reach	day
2	note	bank	money	deposit	reserve	credit	issue	federal	account	loan
3	national	bank	saving	company	president	city	loan	banking	trust	institution

The models discussed in this paper are useful tools for exploratory analysis of unlabelled snippets and, as such, their success is linked to whether the model output is meaningful to a user. Setting K is like choosing a resolution for how fine we want sense differences to be resolved. This suggests setting K in a semi-supervised mode, where we learn the model parameters unsupervised for a few values of K, and the user assigns meaningful labels to the posterior sense distributions based on the model output. A low value of K that is meaningful to the user would help with interpretability, whereas a higher value may fit the data better, and the user can select K based on this trade-off. We demonstrate what this looks like in practice using the "bank" example.

A natural way to examine the posterior is to look at the context words v with the highest probabilities $\frac{1}{T}\sum_{t=1}^{T}\tilde{\psi}_v^{k,t}$ under each model sense k, marginally over time. Table 4 shows the top 10 most probable context words for "bank" if we run EDiSC with $K = 2$ and $K = 3$ model senses, using embedding dimension $M = 200$ in both cases. With $K = 2$, the model senses are readily recognisable as riverbank and financial institution respectively. With $K = 3$, sense 1 is recognisable as riverbank, whereas senses 2 and 3 are both recognisable as financial institution. In the latter case, whilst senses 2 and 3 have different distributions, there is some overlap in the most probable words, which is undesirable since we would like the model senses to be as distinct as possible to help with interpretability. We therefore choose $K = 2$ in this case, which is the smallest value giving meaningful model output.

Incidentally, $K = 3$ fits the data better on merging the split financial institution senses of bank, both in terms of predictive accuracy and true-model recovery. However, the true labels are not generally available, and therefore cannot be used for model selection. Ultimately, choosing K is up to the user's judgement.

Another important consideration (for both K and M) is MCMC convergence. As discussed in Section 4.4, the posterior (2) is challenging to sample due to metastability and multimodality. Choosing and carefully tuning a good sampler may help overcome the metastability,

Table 5: "Kosmos" two different modes under EDiSC with $M = 100$ and $K = 3$. Senses 1, 2 and 3 are interpretable as decoration, order and world respectively in the 'correct' mode. The three true senses are not distinguishable in the 'incorrect' mode.

Sense	Top 10 context words for each model sense in the 'correct' mode									
1	ἔχω (to have)	πολύς (many, much)	πᾶς (all)	γυνή (woman)	καλός (beautiful)	μέγας (large)	χρύσεος (golden)	φέρω (to carry)	γίγνομαι (become)	κοσμέω (adorn or arrange)
2	πολιτεία (citizenship)	πᾶς (all)	τάξις (arrangement)	γίγνομαι (become)	ἔρχομαι (to go)	καθίστημι (to set in order)	πόλις (city)	πολύς (many, much	πρότερος (before)	τρέπω (to rotate)
3	πᾶς (all)	οὐρανός (sky)	θεός (divine)	γῆ (earth)	γίγνομαι (become)	κόσμος (kosmos)	λέγω (to say)	ἔχω (to have)	ὅλος (entire)	Ζεύς (Zeus)
Sense	Top 10 context words for each model sense in the 'incorrect' mode									
1	πᾶς (all)	ἔχω (to have)	πολύς (many, much)	γίγνομαι (become)	γυνή (woman)	καλός (beautiful)	φέρω (to carry)	μέγας (large)	πόλις (city)	χρύσεος (golden)
2	θεός (divine)	πατήρ (father)	κύριος (ruling, lord)	κόσμος (kosmos)	οὐρανός (sky)	πᾶς (all)	υἱός (son)	Ἰησοῦς (Jesus)	εἶπον (to speak)	αἰών (lifetime, epoch)
3	πᾶς (all)	οὐρανός (sky)	γῆ (earth)	ἔχω (to have)	γίγνομαι (become)	λέγω (to say)	ὅλος (entire)	κόσμος (kosmos)	φημί (to speak)	φύσις (origin)

Table 6: "Kosmos" top 10 context words for each model sense under EDiSC. Sense 1 corresponds to decoration, sense 2 to order, and senses 3 and 4 to world.

Sense	Top 10 context words for EDiSC $M = 100$ with $K = 4$ senses									
1	ἔχω (to have)	πᾶς (all)	πολύς (many, much)	γυνή (woman)	καλός (beautiful)	μέγας (large)	φέρω (to carry)	χρύσεος (golden)	κοσμέω (adorn or arrange)	γίγνομαι (become)
2	πᾶς (all)	πολιτεία (citizenship)	τάξις (arrangement)	γίγνομαι (become)	ἔρχομαι (to go)	καθίστημι (to set in order)	πόλις (city)	τρέπω (to rotate)	πολύς (many, much)	πολέμιος (belonging to war)
3	πᾶς (all)	γῆ (earth)	οὐρανός (sky)	ἔχω (to have)	γίγνομαι (become)	λέγω (to say)	ὅλος (entire)	κόσμος (kosmos)	φύσις (origin)	ἕκαστος (each, every)
4	θεός (divine)	πατήρ (father)	οὐρανός (sky)	κόσμος (kosmos)	κύριος (ruling, lord)	πᾶς (all)	εἶπον (to speak)	λέγω (to say)	ἔρχομαι (to go)	υἱός (son)

but the posterior may still be multimodal. If we were to condition on the true sense labels, the resulting posterior is unimodal (as it has strongly informative data). We would like to find a model for the unlabelled data that gives a posterior resembling that for the labelled data. We therefore favour models that are more concentrated and unimodal. This can be explored using MCMC. Some configurations of K and M tend to give multimodal posteriors for some datasets. This may indicate model misspecification. Conversely, a model with a unimodal posterior, in which the model senses are interpretable, is indicative of a well-specified model in our setting.

As an example, running EDiSC with $M = 100$ and $K = 3$ on the "kosmos" data, the samplers settle in one of two distinct modes as shown in Table 5. The likelihood is the same

in both cases. In the 'correct' mode, the senses are recognisable due to representative words like γυνή (woman) and χρύσεος (golden) for the "decoration" sense, πολιτεία (citizenship) and τάξις (arrangement) for the "order" sense, and γῆ (earth) and οὐρανός (sky) for the "world" sense. However, in the 'incorrect' mode, the "order" and "world" senses do not appear separated, with some words like οὐρανός (sky) appearing with high probability under both model senses. The Brier scores reflect this: BS = 0.322 in the first case and BS = 0.620 in the second case (the best score obtained through all possible mappings of model senses to true senses). Note that all Greek translations have been obtained from Wiktionary, and we have only included a few representative meanings to help the reader follow.

In general, a model with a multimodal posterior should be avoided since we cannot be sure which (if any) is the 'true' mode. We find that we get unimodal posteriors with $K = 4$ for the "kosmos" data, $K = 3$ for the "mus" data, and $K = 4$ for the "harmonia" data. These are also the lowest K values for which the model senses are interpretable, and so we set K accordingly. With "kosmos", two of the model senses map to the "world" sense, and an example for $M = 100$ is shown in Table 6. Similarly, with "harmonia", two of the model senses map to the "abstract" sense in the converged runs.

In choosing M, the first consideration should be MCMC convergence as discussed above. This is because if the model fails to converge for certain M values, it may be that the embeddings learnt at that dimension do not capture the context-word semantics adequately for our purpose, and so the model is misspecified. After excluding the non-converging settings, model-selection tools may be used. However, the choice of M is ultimately up to the user's judgement, so other factors such as marginal gains or computational costs may be considered. Note that M cannot be decided on the fly: the embeddings must be learnt separately for fixed values of M. Therefore, practically, the choice has to be made out of a handful of predetermined values, which in our case are 50, 100, 200 and 300.

We use the widely applicable information criterion (WAIC) (Watanabe, 2010; Vehtari et al., 2017) to guide our choice as it is a computationally convenient model selection tool for Bayesian inference. It has several slightly different formulations, but the one used in the R packages `LaplacesDemon` (Statisticat and LLC., 2021) and `loo` (Vehtari et al., 2024) is

$$\begin{aligned}\text{WAIC} &= -2(\widehat{\text{LPD}} - \hat{p}_{\text{WAIC}}) \\ &= -2\sum_{d=1}^{D}\Big(\log \mathbb{E}_{\phi,\psi|W}[p(W_d|\phi,\psi)] - \mathbb{V}_{\phi,\psi|W}[\log p(W_d|\phi,\psi)]\Big),\end{aligned} \tag{5}$$

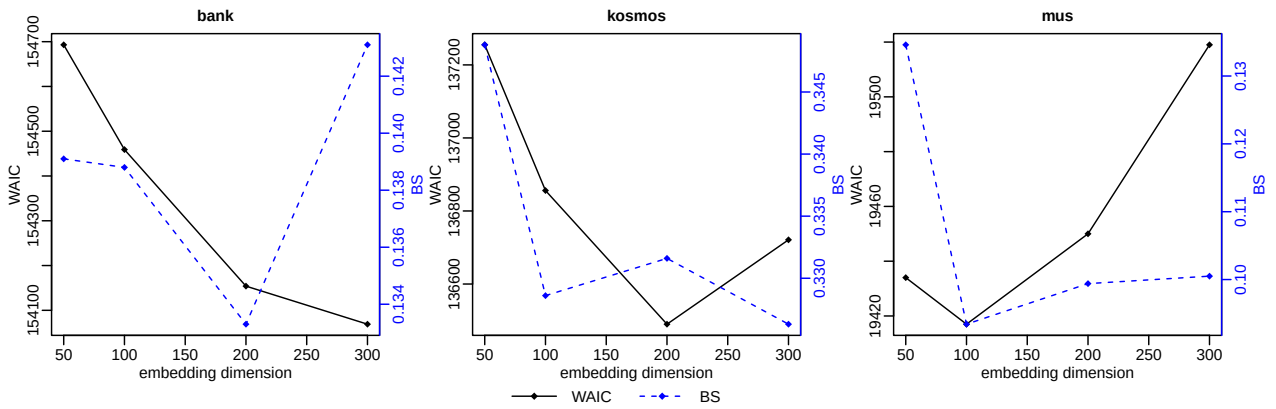


Figure 2: WAIC and Brier scores for different choices of embedding dimension M for the "bank", "kosmos" and "mus" data

where $\widehat{\text{LPD}}$ estimates the log pointwise predictive density (LPD), $\hat{p}_{\text{WAIC}}$ estimates the effective number of parameters, and $(\widehat{\text{LPD}} - \hat{p}_{\text{WAIC}})$ estimates the expected log pointwise predictive density (ELPD). The WAIC is a predictive loss like the Akaike information criterion (AIC), and asymptotically equivalent to selecting the model that maximises the posterior predictive probability density for held-out data in a leave-one-out cross-validation (LOOCV) setup.

Figure 2 shows how the WAIC varies with the choice of M, and also shows the Brier scores for these choices, for the "bank", "kosmos" and "mus" data. There seems to be a loose correlation between the WAIC and BS, which suggests that the WAIC is a sensible tool to use when validation data (with true sense labels) is not available. For the "kosmos" and "mus" data, the WAIC is minimised at $M = 200$ and $M = 100$ respectively, so we go with these choices. For the "bank" data, the WAIC does not seem to have a local minimum within our range. On the other hand, the marginal gains between $M = 200$ and 300 are relatively low, whereas the computational cost is much higher. We therefore select $M = 200$ for "bank", which strikes a good balance. For "harmonia", we only get MCMC convergence for $M = 300$, so that is the only choice.

One may ask why we use the WAIC to guide our choice of M but not of K. This is because, as opposed to M, the WAIC always favours a higher K value in our experiments. This, in turn, is because of how these two parameters interact with the likelihood (3): K directly changes the number of variables used in the likelihood calculation via the dimensions of $\tilde{\phi}$ and $\tilde{\psi}$, whereas M only indirectly affects the likelihood via the relation $\tilde{\psi}^{k,t} = \text{softmax}\left(\rho(\chi^k + \theta^t) + \varsigma\right)$ without changing the dimensions of $\tilde{\psi}$ itself. Vehtari et al. (2017) state that $\hat{p}_{\text{WAIC}}$ in (5) can be severely understated in case of a weak prior, and is

unreliable if any of the terms $\mathbb{V}_{\phi,\psi|W}[\log p(W_d|\phi,\psi)]$ exceed 0.4, which is frequently the case in our experiments. This is not a problem when choosing M, as the effective dimension, estimated by the variance term, does not change much from one M-value to another; so the order of the models is decided mainly by goodness of fit $\widehat{\text{LPD}}$, and the unreliable dimension estimate $\hat{p}_{\text{WAIC}}$ has little impact. However, model selection for K is a trade-off between model fit (LPD) and parsimony (variance); so the poor estimate $\hat{p}_{\text{WAIC}}$ of the effective dimension is an obstacle. In any case, the WAIC should be used as a guide rather than a definitive rule.

5.3 Sense-prevalence estimation

Predictive accuracy and true-model recovery are the two classical goals of statistical inference. However, good performance on one front does not necessarily correlate with good performance on the other. We examined predictive performance using the Brier score. We now assess our fitted models against the 'true' models.

We have the posterior sense-prevalence distributions $\tilde{\phi}|W$ given the *unlabelled* data W. We use the 95% highest posterior density (HPD) intervals as concise visual summaries of the support for the posterior and the uncertainty in the estimates. Following Zafar and Nicholls (2022, Section 7), we would like to compare these HPDs to the unknown true sense prevalence, $\tilde{\Phi}$ say. We do not have $\tilde{\Phi}$: we only have the true sense labels $o_{1:D}$. However, given $o_{1:D}$, estimating $\tilde{\Phi}$ with uncertainty quantification is an easy and classical task: we simply smooth the empirical sense-prevalence probabilities using a time series model for their evolution. This gives us well-calibrated independent estimates $\tilde{\phi}|(z = o)$ given the *labelled* data o, which we use as proxies for $\tilde{\Phi}$. The posteriors $\tilde{\phi}|(z = o)$ would be concentrated on the empirical sense-prevalence probabilities where there are many observations, and would apply some shrinkage and smoothing where snippets are infrequent. If the credible sets from the unlabelled analysis are close to the credible sets from the labelled analysis using the same AR(1) models, this would indicate success: conditioning on the true labels gives the best results achievable with these models.

We show this comparison for both DiSC and EDiSC posteriors on the "kosmos" data in Figure 3, with the models selected as discussed in Section 5.2. We see that both unlabelled posteriors $\tilde{\phi}|W$ (solid error bars) are in surprisingly good agreement with the labelled posterior $\tilde{\phi}|(z = o)$ (dashed error bars). However, EDiSC has generally better range and location: the blue EDiSC error bars generally have higher overlap with the dashed bars, and the circles (posterior means) are closer. The unlabelled analysis is equivalent to many labelled analyses averaged over uncertainty in reconstructed labels. This uncertainty is significant,

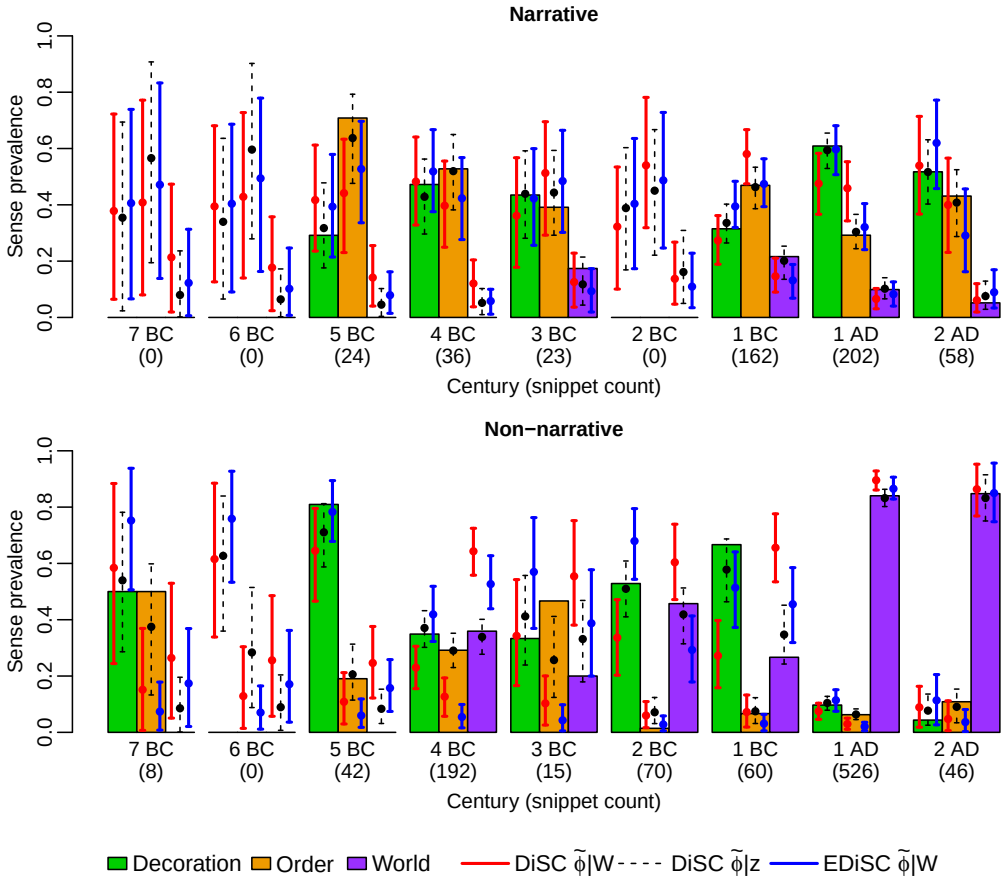


Figure 3: "Kosmos" expert-annotated empirical sense prevalence (coloured bars with height $N^o_{k,g,t} / \sum_{l=1'}^{K'} N^o_{l,g,t}$ for each k, g, t), and 95% HPD intervals (error bars) and posterior means (circles) from the model output. Snippet counts $N^o_{\cdot,g,t}$ are given in brackets below the axes. Note that the labelled posteriors $\tilde{\phi}|z$ from DiSC and EDiSC are identical.

Table 7: "Kosmos" Bayes factors BF_{01} on a $\log_{10}$ scale for nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : DiSC and H_1 : EDiSC. Red indicates incorrect rejection of H_0.

Model	Genre	7 BC	6 BC	5 BC	4 BC	3 BC	2 BC	1 BC	1 AD	2 AD
DiSC	narrative	0.28	0.41	0.47	0.79	0.86	0.62	0.77	0.18	1.20
EDiSC	narrative	0.43	0.63	1.06	1.20	0.92	0.62	1.09	1.63	0.97
DiSC	non-narr	−0.05	0.08	−0.65	$-\infty$	−0.26	0.41	−1.54	0.44	1.10
EDiSC	non-narr	−0.35	0.01	−0.36	$-\infty$	−0.99	0.37	0.77	−0.07	0.92

so it is remarkable how close the unlabelled analysis comes to the labelled analysis. It is also interesting that particular features of the ground truth are reflected well in the posteriors, such as the emergence of the "world" sense of "kosmos" in 4th century BC.

The figure shows the empirical sense-prevalence estimates $N^o_{k,g,t} / N^o_{\cdot,g,t}$ as coloured bars,

where $N^o_{k,g,t}$ is the count of snippets in genre g at time t with sense k under the true sense labels o, and $N^o_{\cdot,g,t} = \sum_{l=1'}^{K'} N^o_{l,g,t}$ is the total snippet count in each g, t (shown in brackets below the axes). The empirical estimates may be of interest, and are shown on the same figure as they are defined on the same space and scale. However, the empirical estimates should not be used to assess true-$\tilde{\Phi}$ recovery if the datasets are small and sparse, as is the case with our ancient Greek data, since they are not smoothed.

To make the comparison more concrete, we quantify both models' performance using Bayes factors to measure agreement between the unlabelled and labelled posteriors. We treat the credible sets from the labelled analysis as the truth and ask, does the unlabelled analysis reject the truth? Let $\mathcal{S}^{g,t}$ be the 95% HPD region for genre g and time t under the labelled posterior $\tilde{\phi}^{g,t}|(z = o)$. For each g and t, we compute the Bayes factor $\text{BF}_{01} = \frac{p(W|H_0)}{p(W|H_1)}$ for the nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : DiSC and H_1 : EDiSC. By the Savage-Dickey density ratio, we have

$$\text{BF}_{01} = \frac{\pi(\tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}|W, H_1)}{\pi(\tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}|H_1)},$$

where the posterior probabilities in the numerator can be computed using our MCMC samples from the unlabelled $\tilde{\phi}^{g,t}|W$ posteriors, and the prior probabilities in the denominator can be computed using softmax-transformed Monte Carlo simulations from the prior $\phi^{g,t}$. We use the same prior simulations for DiSC and EDiSC so that the model comparison is indifferent to simulation error in the denominator. The results are reported in Table 7 on a $\log_{10}$ scale. Positive values indicate evidence in favour of H_0, with higher BF_{01} corresponding to greater overlap between $\tilde{\phi}^{g,t}|W$ and $\tilde{\phi}^{g,t}|(z = o)$, and vice versa for negative values. Using the scale in Kass and Raftery (1995), $\log_{10}(\text{BF}_{01}) < -1$ indicates strong evidence against H_0; so we reject H_0 in favour of H_1 at this threshold. These rejections are highlighted in red in the table. We find that EDiSC gives a higher Bayes factor than DiSC in 61% of cases, and incorrectly rejects H_0 only once versus twice for DiSC. Further, when EDiSC performs better, the improvement can be substantial, for example 1 AD in the narrative genre. The converse is not true: when DiSC performs better, the difference is only slight.

Equivalent figures and tables for the "mus", "harmonia" and "bank" data are given in Appendix B. These similarly show EDiSC outperforming DiSC on true-$\tilde{\Phi}$ recovery in 56, 83 and 90 percent of cases respectively, with fewer incorrect rejections of H_0, less bias, and more accurate and precise credible sets.

Table 8: Median (interquartile range) ESS per hour of CPU time from MALA sampling

Model	ESS for $\tilde{\phi}$		ESS for $\tilde{\psi}$	
DiSC	375	(216 – 531)	391	(213 – 586)
EDiSC ($M = 50$)	1,916	(1,569 – 2,017)	391	(294 – 515)
EDiSC ($M = 100$)	2,192	(1,844 – 2,674)	344	(262 – 459)
EDiSC ($M = 200$)	2,237	(1,810 – 2,424)	303	(215 – 396)
EDiSC ($M = 300$)	1,633	(1,245 – 2,299)	282	(217 – 392)

5.4 Sampling efficiency and scalability

We assess the relative efficiency of sampling the DiSC and EDiSC posteriors using MCMC methods. The form of the model $\psi^{k,t} = \rho(\chi^k + \theta^t) + \varsigma$ in EDiSC, compared to $\psi^{k,t} = \chi^k + \theta^t$ in DiSC, means that additional matrix multiplication is required to calculate the likelihood for EDiSC, which tends to be computationally slow. However, using the effective sample size (ESS) per unit time as a metric for comparing the sampling efficiency, we find that, in fact, the same MCMC applied to EDiSC results in more efficient samples than DiSC. This is because of the lower dimensional model parameters χ, θ for EDiSC.

Table 8 shows the ESS per hour of CPU time from applying MALA to target the $\tilde{\phi}$ and $\tilde{\psi}$ posteriors under DiSC and EDiSC on the "bank" data. We report the medians and interquartile ranges over all the KGT parameters for $\tilde{\phi}$, and over the top 20 most probable words (i.e. over $20KT$ parameters) out of the VKT parameters for $\tilde{\psi}$. We compare implementations not algorithms. However, the comparison is fair as the MALA Monte Carlo is common to both, and the evaluation allows little scope for differential optimisation of implementations. (For example, MALA requires no optimisation of the number of leapfrog steps, as opposed to HMC or NUTS.) All runs were done sequentially on the same PC. We see that whilst the ESS for $\tilde{\psi}$ is of the same order for all models, the ESS for $\tilde{\phi}$ is many times better under EDiSC than under DiSC.

We also analyse how the run times vary with increasing data sizes using synthetic data. Figure 4 summarises the results, where each data point is the mean run-time over three independent runs of 500 MCMC iterations. This comparison favours DiSC, since we use the same number of MCMC iterations for DiSC and EDiSC despite EDiSC being more efficient. We show the results from two different choices for the embedding dimension for EDiSC: $M = 25$ and $M = 200$. Typically, the vocabulary size V would be expected to grow with the data size D (as more snippets tend to bring a larger vocabulary), whereas the embedding dimension would not usually be increased much beyond 200.

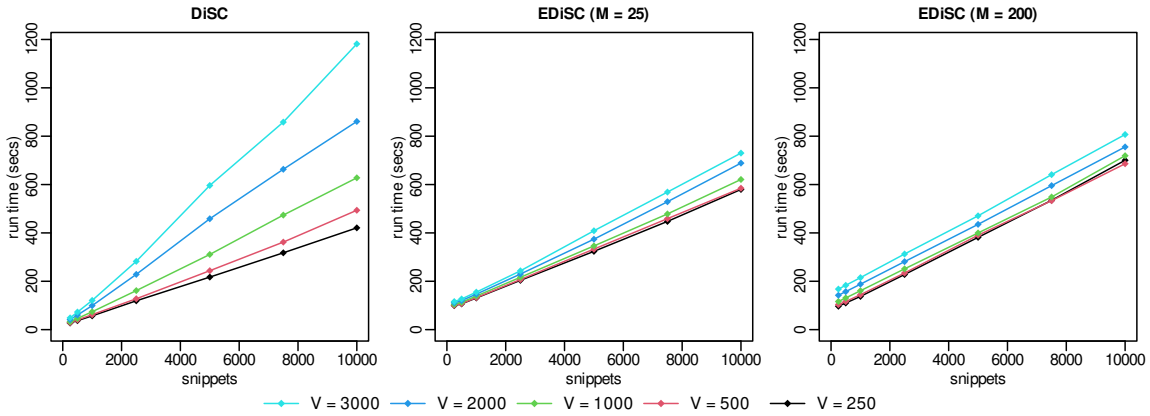


Figure 4: Mean run times in CPU seconds for 500 MCMC iterations on synthetic data using different models, vocabulary sizes (V) and number of snippets (D)

The run times increase linearly with both vocabulary size V and number of snippets D in all cases. However, for $V > 500$ (which is typical) the rate of increase with D is much higher for DiSC than for EDiSC. The interaction effect between V and D on the run time is significant in all cases. However, the interaction effect is much stronger for DiSC than for EDiSC. Moreover, as the embedding dimension M is increased, the interaction effect for EDiSC grows even weaker. As a result, with increasing V and D, run times for DiSC increase much faster than those for EDiSC. The plots show the advantage of using EDiSC over DiSC even in the case of modest data sizes in our synthetic data experiments. Therefore, EDiSC is a lot better suited than DiSC to scaling up for larger data sizes.

This paper focuses on modelling diachronic sense change for our ancient Greek data, where computational issues are a relatively minor concern. However, the models themselves are more widely applicable. Hence, in other situations where efficiency and scalability are more of an issue, these experiments show that there is even more reason to prefer EDiSC over DiSC (or GASC/SCAN for that matter).

6 Discussion

We introduced EDiSC, an embedded version of DiSC, which is a generative model of diachronic sense change that combines word embeddings with DiSC. Experiments on test data show that EDiSC outperforms DiSC, GASC and SCAN, as measured by Brier scores, in terms of accuracy in predicting the unknown sense for the target word in snippets. We estimated the model parameters and quantified the uncertainty in sense-change estimates via HPD intervals, showing that EDiSC outperforms DiSC on recovering the true parameters: estimates obtained using EDiSC on unlabelled data are more closely aligned to those

obtained on labelled data than is the case for DiSC. The good agreement between our HPD intervals computed using unlabelled and labelled data supports our view that it would be very hard to do much better than we have done here, at least in a bag-of-words setting.

We showed that MCMC sampling targeting EDiSC is more efficient than the corresponding sampling for DiSC. Furthermore, EDiSC scales better to large data sizes than DiSC. We considered how fitting these models is challenging due to potential metastability and multimodality in the posteriors, and why variational methods for model fitting fail. We discussed ways of addressing these challenges that work well in our experiments. These include careful model selection with respect to the number of model senses K and the embedding dimension M, as well as MCMC considerations (discussed in the Appendix). More broadly, we gave guidelines for how appropriate values for K and M may be set, such that these meet the user's objectives.

An obvious limitation in the EDiSC model is that it uses traditional word embeddings with only one vector representation per word. Our work is a continuation of existing models (SCAN, GASC, DiSC) using the same general framework, which have previously been used to analyse the ancient Greek data that inspired our research. This framework does not readily admit contextualised word embeddings. It would be interesting to investigate how the framework might be expanded or modified to admit multiple vector representations per word.

Another (necessary) limitation of our work is that we are restricted in what comparisons we can make against other models and methods. As discussed in the literature review, most other methods use some form of pre-training or supervised learning, and in any case have different modelling and inferential goals to ours. However, our models can be generalised and used for wider purposes such as WSD or sense change-point detection if desired; so it would be interesting to investigate how they compare against other methods on shared tasks, or how they might be used in conjunction with other methods from the NLP literature. That is future work for us.

We set out to develop an embedded version of DiSC to model diachronic sense change for our ancient Greek data, drawing parallels from the topic modelling literature, which was an improvement upon the existing model. This objective has been achieved, and prospects for further improvement are good.

Implementation

The code and data used to produce the results reported in this paper are available from https://github.com/schyanzafar/EDiSC.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This research is funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/S515541/1.

Appendices

A Hyperparameter settings

For the AR(1) process hyperparameters $\alpha_\phi, \alpha_\theta$ in EDiSC, a high value admits weak mean reversion without unduly influencing the posteriors. In our model fitting, we therefore experimented with values of 0.9 as in DiSC, and even higher values of 0.99. We found the converged posteriors to be robust to these choices. However, with $\alpha_\theta = 0.99$, convergence problems become more frequent. Moreover, using the WAIC for model selection, values of 0.9 are preferred. Therefore, we continue to use $\alpha_\phi = \alpha_\theta = 0.9$ for EDiSC.

For the κ_ϕ variance hyperparameter, Zafar and Nicholls (2022, Section 4.2) elicit a prior by defining what we consider to be an extreme (i.e. 3-sigma) event, and using this to set quantiles. For any fixed time t, genre g and pair of senses $l, m \in \{1, \ldots, K\}$, a difference in sense prevalence of the order $\tilde{\phi}_l^{g,t} / \tilde{\phi}_m^{g,t} \approx 100$ is considered extreme. Therefore, on the log scale, $\phi_l^{g,t} - \phi_m^{g,t} > \log 100$ is considered a 3-sigma event. Since $\mathbb{V}(\phi_l^{g,t} - \phi_m^{g,t}) = \frac{2\kappa_\phi}{1-(\alpha_\phi)^2}$, we express our preference with $3\left(\frac{2\kappa_\phi}{1-(\alpha_\phi)^2}\right)^{\frac{1}{2}} = \log 100$, giving $\kappa_\phi = \frac{1-(\alpha_\phi)^2}{2}(\frac{1}{3}\log 100)^2 \approx 0.25$ on rounding. The ϕ parameter in EDiSC is identical to that in DiSC, so we continue to use the same value.

For the other variance hyperparameters, still following Zafar and Nicholls (2022, Section 4.2), for any fixed time t, sense k and pair of words $x, y \in \{1, \ldots, V\}$, the ratio of context-word probabilities of the order $\tilde{\psi}_x^{k,t} / \tilde{\psi}_y^{k,t} \approx 1\,000$ is considered extreme. Therefore, on the log scale, $\psi_x^{k,t} - \psi_y^{k,t} > \log 1\,000$ is considered a 3-sigma event.

Now, for EDiSC, we have $\mathbb{V}(\psi_x^{k,t} - \psi_y^{k,t}) = \mathbb{V}(\rho_x^{\mathrm{T}}\xi^{k,t} - \rho_y^{\mathrm{T}}\xi^{k,t} + \varsigma_x - \varsigma_y)$, which simplifies to $\mathbb{V}(\psi_x^{k,t} - \psi_y^{k,t}) = (\rho_x - \rho_y)^{\mathrm{T}}\mathbb{V}(\xi^{k,t})(\rho_x - \rho_y) + \mathbb{V}(\varsigma_x - \varsigma_y)$ since ξ and ς are independent of each other by construction. We have $\mathbb{V}(\varsigma_x - \varsigma_y) = 2\kappa_\varsigma$ whereas $\mathbb{V}(\xi^{k,t}) = \mathbb{V}(\chi^k + \theta^t)$ is an $M \times M$ diagonal matrix with entries $\left(\kappa_\chi + \frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)$. Hence,

$$\mathbb{V}(\psi_x^{k,t} - \psi_y^{k,t}) = (\rho_x - \rho_y)^{\mathrm{T}}(\rho_x - \rho_y)\left(\kappa_\chi + \frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right) + 2\kappa_\varsigma.$$

We approximate $(\rho_x - \rho_y)^{\mathrm{T}}(\rho_x - \rho_y)$ with its median c over all pairs $x, y \in \{1, \ldots, V\}$, and express our preference with $3\left(c(\kappa_\chi + \frac{\kappa_\theta}{1-(\alpha_\theta)^2}) + 2\kappa_\varsigma\right)^{\frac{1}{2}} = \log 1\,000$.

We want the bulk of the variance to be explained by ξ, since ς is only a correction parameter that comes into play when the embeddings for words x, y are similar but the words occur at different frequencies in the snippets. Taking the extreme case $\rho_x = \rho_y$, we would not expect the frequency of x to be too different from that of y; so we assert that $\psi_x^{k,t} - \psi_y^{k,t} > \log 10$ is a 3-sigma event in this extreme case. Also, $\rho_x = \rho_y$ gives $\mathbb{V}(\psi_x^{k,t} - \psi_y^{k,t}) = 2\kappa_\varsigma$, so we express our preference with $3(2\kappa_\varsigma)^{\frac{1}{2}} = \log 10$, giving $\kappa_\varsigma = \frac{1}{2}(\frac{1}{3}\log 10)^2 \approx 0.25$ on rounding.

We have $\xi^{k,t} = \chi^k + \theta^t$, so $\mathbb{V}(\xi^{k,t})$ must be apportioned between $\mathbb{V}(\chi^k)$ and $\mathbb{V}(\theta^t)$. Given our preference $\kappa_\chi + \frac{\kappa_\theta}{1-(\alpha_\theta)^2} = \left((\frac{1}{3}\log 1\,000)^2 - 2\kappa_\varsigma\right)/c$, we set $\kappa_\chi = a_\chi\left((\frac{1}{3}\log 1\,000)^2 - 2\kappa_\varsigma\right)/c$ and $\kappa_\theta = a_\theta(1-(\alpha_\theta)^2)\left((\frac{1}{3}\log 1\,000)^2 - 2\kappa_\varsigma\right)/c$, with $a_\chi + a_\theta = 1$. A plausible choice is $a_\chi = a_\theta = 0.5$ as for DiSC, since χ and θ are additive effects on the same scale. We experimented with this choice, as well as $a_\chi = 0.75, a_\theta = 0.25$ and $a_\chi = 0.25, a_\theta = 0.75$, and found the posteriors to be robust to these choices. Moreover, the WAIC was relatively constant between these choices, so we go with $a_\chi = a_\theta = 0.5$ for simplicity and consistency with DiSC. However, users may adjust the proportions depending on how they want to resolve the variation over senses and time periods for the test case in question. With these choices, and with $\kappa_\varsigma = 0.25$, we get $\kappa_\chi \approx 2.5/c$ and $\kappa_\theta \approx 0.5/c$ on rounding.

To summarise, for EDiSC, we use $\alpha_\phi = \alpha_\theta = 0.9, \kappa_\phi = 0.25, \kappa_\chi = 2.5/c, \kappa_\theta = 0.5/c, \kappa_\varsigma = 0.25$, where c is the median value of $(\rho_x - \rho_y)^{\mathrm{T}}(\rho_x - \rho_y)$ over all pairs $x, y \in \{1, \ldots, V\}$. For comparison, DiSC uses $\alpha_\phi = \alpha_\theta = 0.9, \kappa_\phi = 0.25, \kappa_\chi = 1.25, \kappa_\theta = 0.25$.

Note that, for GASC and SCAN, the hyperparameters to set are κ_ψ and the a, b in $\kappa_\phi \sim \mathrm{Inv\,Gamma}(a, b)$. The authors of GASC preferred $\kappa_\psi = 0.01, a = 1, b = 1$ for the Greek test cases, whereas the authors of SCAN preferred $\kappa_\psi = 0.1, a = 7, b = 3$ for their English test cases. In our comparisons, we try both configurations in our implementation of GASC for the Greek test cases, and stick to the SCAN configuration for our English test case.

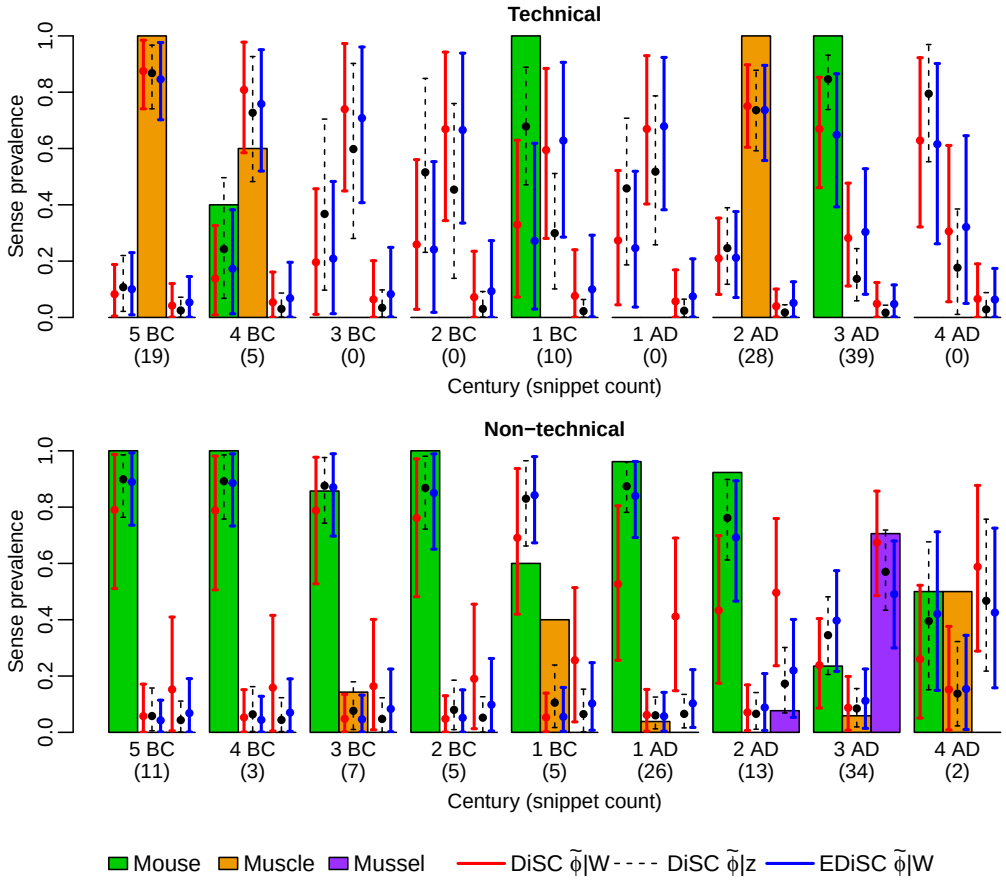


Figure 5: "Mus" expert-annotated empirical sense prevalence (coloured bars), and 95% HPD intervals (error bars) and posterior means (circles) from the model output

Table 9: "Mus" Bayes factors BF_{01} on a $\log_{10}$ scale for nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : DiSC and H_1 : EDiSC

Model	Genre	5 BC	4 BC	3 BC	2 BC	1 BC	1 AD	2 AD	3 AD	4 AD
DiSC	technical	1.37	0.98	0.80	0.61	0.46	1.00	1.56	1.12	0.85
EDiSC	technical	1.30	0.98	0.77	0.51	0.32	0.85	1.44	1.10	0.85
DiSC	non-tech	0.97	0.97	0.85	0.75	0.36	−0.36	0.01	0.83	0.39
EDiSC	non-tech	1.20	1.20	1.09	1.05	0.89	1.24	0.99	0.90	0.48

B Further results

Following on from Section 5.3, we show the sense prevalence graphs for "mus" in Figure 5 and the Bayes factors in Table 9. As in the "kosmos" test case, the EDiSC posterior on the unlabelled data generally matches the ground truth using the labelled data better than DiSC (in 56% of cases), and also gives more precise credible sets for the non-technical genre. Both models pick up on the emergence of the "mussel" sense of "mus" in the non-technical

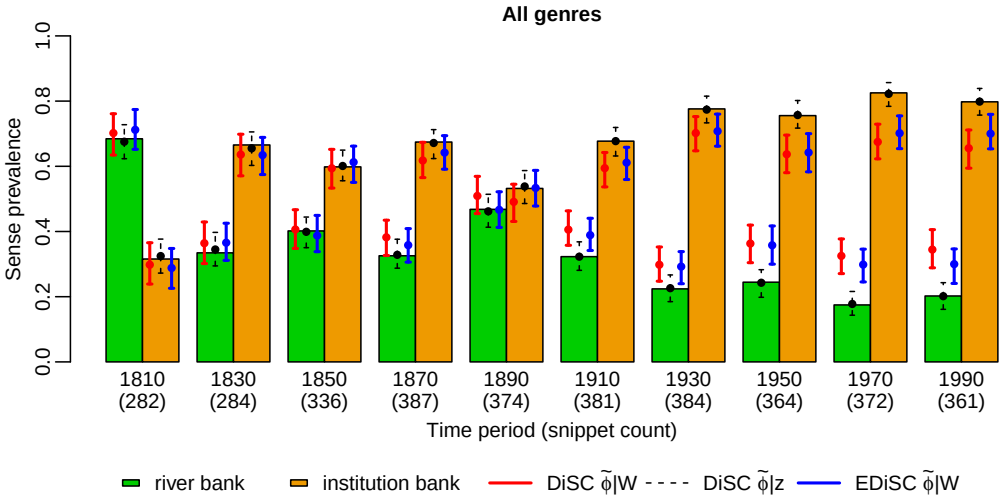


Figure 6: "Bank" manually annotated empirical sense prevalence (coloured bars), and 95% HPD intervals (error bars) and posterior means (circles) from the model output

Table 10: "Bank" Bayes factors BF_{01} on a $\log_{10}$ scale for nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : DiSC and H_1 : EDiSC. Red indicates incorrect rejection of H_0.

Model	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
DiSC	0.85	0.91	0.98	0.67	0.73	−0.05	0.13	−1.60	$-\infty$	−2.30
EDiSC	0.79	0.91	0.98	0.92	0.95	0.42	0.28	−0.91	−1.70	−0.55

genre in 2nd century AD, with EDiSC providing more accurate prevalence estimates. As for "kosmos", when EDiSC outperforms DiSC, it does so much more substantially (e.g. in the 1st and 2nd centuries AD for the non-technical genre) than the converse.

The sense prevalence graphs for "bank" are given in Figure 6 and the Bayes factors in Table 10. This is an easy test case where DiSC was already performing well. The improvement provided by EDiSC is therefore only marginal, but it is nevertheless noticeable for the later time periods. EDiSC outperforms DiSC in 90% of cases, and results in only one incorrect rejection of H_0 as opposed to three for DiSC. We see some divergence between the model posteriors and the ground truth in later time periods, resulting in negative Bayes factors. This is because the modelling choice of $K = 2$ is a little restrictive. In our experiments with higher K values (which we do not show here), the posteriors recover the ground truth much more closely. However, as discussed in Section 5.2, foreknowledge of the truth cannot be used in model selection. We prioritise interpretability, and the model is performing adequately with our modelling choices. The code is available, and we encourage user exploration.

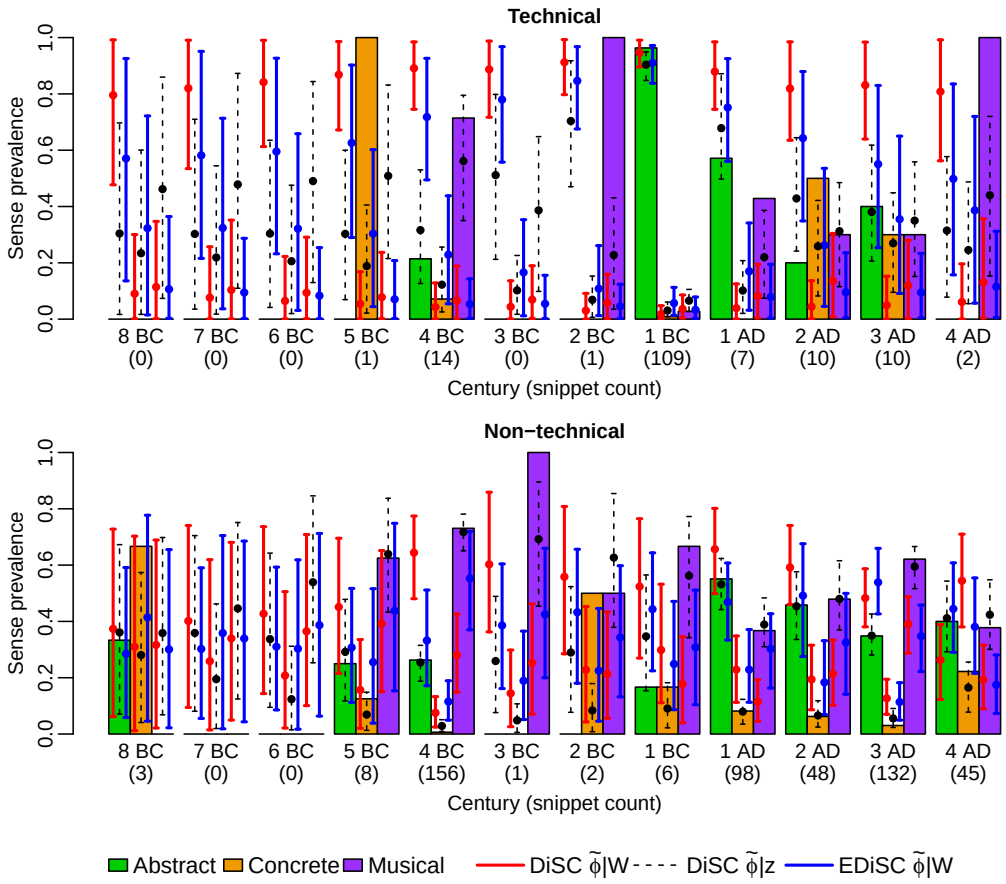


Figure 7: "Harmonia" expert-annotated empirical sense prevalence (coloured bars), and 95% HPD intervals (error bars) and posterior means (circles) from the model output

Table 11: "Harmonia" Bayes factors BF_{01} on a $\log_{10}$ scale for nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : DiSC and H_1 : EDiSC. Red indicates incorrect rejection of H_0.

Model	Genre	8BC	7BC	6BC	5BC	4BC	3BC	2BC	1BC	1AD	2AD	3AD	4AD
DiSC	technical	−0.60	−0.69	−1.06	−1.38	$-\infty$	−0.48	0.43	1.36	0.25	−0.95	−1.16	−1.10
EDiSC	technical	−0.27	−0.38	−0.54	−1.04	−2.24	−0.34	0.43	1.39	0.43	−0.05	−0.25	−0.32
DiSC	non-tech	0.14	0.22	0.32	0.39	$-\infty$	−0.47	−0.46	−0.93	$-\infty$	−0.96	−0.58	$-\infty$
EDiSC	non-tech	0.13	0.16	0.22	0.37	−0.05	0.17	0.11	0.05	0.12	0.39	−0.58	−1.01

Table 12: "Harmonia" counts and top 10 context words for each expert-annotated sense. Repeated words are shown in red.

Sense	Count	Top 10 context words under expert sense-annotation									
abstract	303	γίγνομαι (become)	λόγος (subject matter)	πᾶς (all)	ποιέω (to make)	ἔχω (to have)	ψυχή (spirit, soul)	ἁρμονία (harmonia)	ἀριθμός (number)	εἷς (one)	πολύς (many, much)
concrete	42	ὀστέον (bone, rock)	λίθος (stone)	μέγεθος (height)	ὅσος (how much)	φημί (to speak)	κόσμος (kosmos)	σῶμα (body)	μέγας (large)	οὐκέτι (no more)	πᾶς (all)
musical	308	ῥυθμός (rhythm)	πᾶς (all)	ἔχω (to have)	φημί (to speak)	λόγος (subject matter)	πρότερος (before)	ἁρμονία (harmonia)	ποιέω (to make)	εἷς (one)	καλέω (to call, summon)

The sense prevalence for "harmonia" is shown in Figure 7 and the Bayes factors in Table 11. In contrast to "bank", "harmonia" is a particularly challenging test case, since the data here is particularly sparse. The "concrete" sense of "harmonia" is severely under-represented in the data as shown in Table 12, and there is a high degree of overlap in probable context words under expert-annotation. The under-represented sense has relatively little effect on the likelihood, so the MCMC targeting the posterior struggles to recognise it. The high overlap in probable words means that the true senses are not very distinct, which understandably affects the models' ability to separate them. Some of the overlapping words appear to be function words based on online translations, so perhaps more targeted data filtering (cf. Section 2) aided by expert domain-knowledge would be helpful. However, we have not explored this. Despite these challenges, EDiSC performs remarkably well in recovering the truth, as evidenced by the high degree of overlap between the unlabelled and labelled posteriors. EDiSC outperforms DiSC in 83% of cases, and incorrectly rejects H_0 thrice versus eight times for DiSC. DiSC appears to have a strong bias for the "abstract" sense in the technical genre, but EDiSC fares much better, benefiting from the extra semantic information contained in the embeddings that is lacking in the sparse snippet data. Accurate sense-change analysis for "harmonia" was not possible using DiSC, but is now possible with EDiSC.

C MCMC convergence issues

As discussed in Sections 4.4 and 5.2, MCMC convergence is essential for inference and model selection. However, getting the MCMC to converge for our test cases with these models is not easy. In this section, we discuss some of the issues encountered and possible strategies for overcoming them.

A sampler may sometimes get stuck in a metastable state and fail to converge. Running the sampler for longer often resolves the issue, but this is not guaranteed in practice. As an example, in one of our runs using Stan's NUTS on EDiSC ($M = 300, K = 3$) for "mus", we ran the sampler for 100k iterations, saving every 10th iteration. The sampler was stuck between two different metastable states before eventually finding the correct mode. This can be visualised most easily using the $\tilde{\phi}$ trace plots in Figure 8. The Brier scores in the metastable states are also shown in Figure 9 to demonstrate the issue, but this cannot be used as a diagnostic in practice since we do not have the ground truth. Note that this is not a case of label switching: there is no permutation of model senses $1 : K$ in either of the metastable states that is equivalent to the posterior in the converged state.

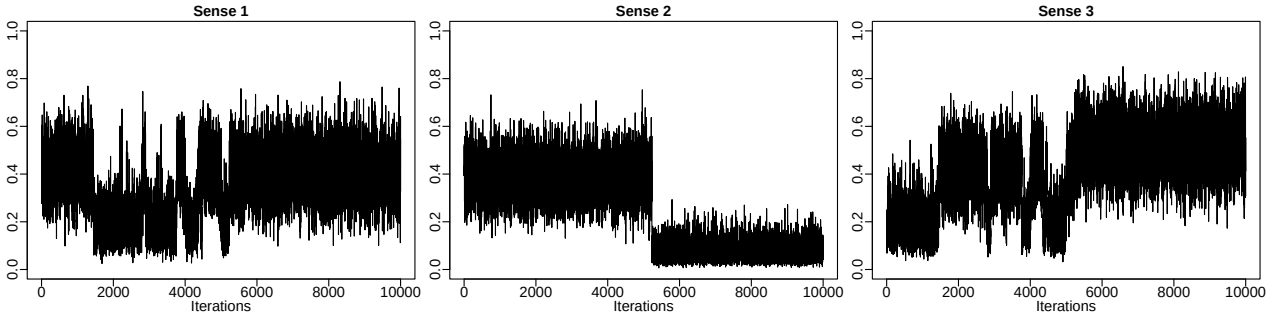


Figure 8: $\tilde{\phi}^{g,t}$ trace plots showing the metastable states before MCMC convergence. Here, we run Stan's NUTS on EDiSC for "mus" with $M = 300, K = 3$, and show the plots for $g = 2, t = 8$.

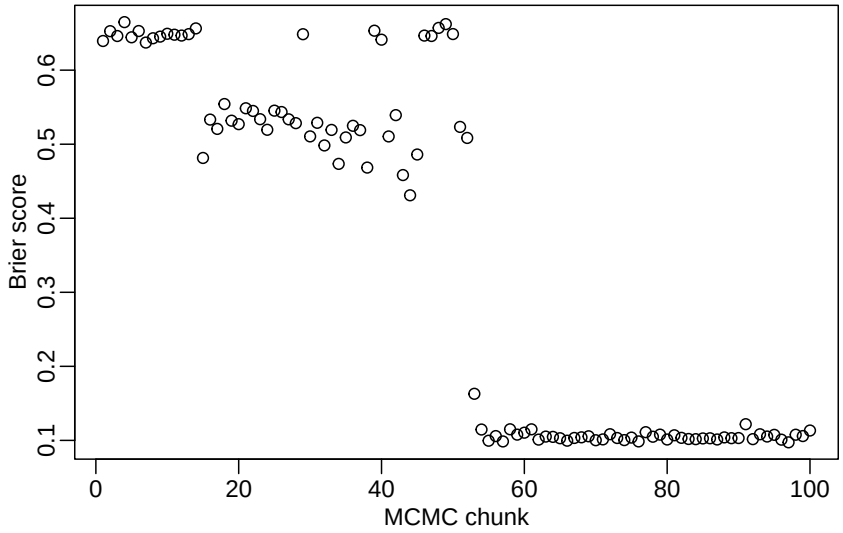


Figure 9: Brier scores computed on sequential MCMC chunks of 100 samples each, showing the two metastable states before convergence

Another technique commonly used in the MCMC literature to overcome metastability is parallel tempering (e.g. the original Geyer et al. 1991 or the state of the art Syed et al. 2021). This uses MCMC on multiple cores to target, in parallel, a *tempered* or *annealed* posterior in which the likelihood is raised to an inverse 'temperature' λ. Samples are obtained from the 'coldest' chain with $\lambda = 1$, which is just the posterior. However, swap moves between chains are proposed, and accepted according to the Hastings ratio, where 'hotter' chains use a schedule of λ values less than 1, thus encouraging mixing by making the target distribution more diffuse. A problem with this approach in our test cases is that, in order to get reasonable acceptance rates for swap moves between chains, the tempering schedules need to have λ values very close to each other. Hence, in order to get mixing at higher temperatures, we need a lot of chains; and this is not achievable with limited computing

resources.

However, the tempering or annealing trick (Geman and Geman, 1984; Hajek, 1988) predates parallel tempering, and can be used on a single core during the burn-in phase of the MCMC (and then switched off, so we target the untempered posterior). This allows more mixing earlier in the chain and ensures that, at the least, the sampler does not get stuck in a metastable state due to the starting configuration. In our HMC implementations, we use a simple tempering schedule

$$\lambda_n = \lambda_{\min} + (1 - \lambda_{\min}) \left(\frac{n}{N_{\text{temp}}} \right)^{\beta}, \tag{6}$$

where n is the MCMC iteration number, N_{temp} is the number of iterations for which to use tempering, $\lambda_{\min}$ is the minimum inverse temperature, and $\beta \leq 1$ determines the rate of change in λ_n (with $\beta = 1$ giving a linear schedule and $\beta < 1$ giving a schedule increasing at a decreasing rate). We find that $\lambda_{\min} = 0.1$, $\beta = 1/3$ and N_{temp} set equal to half the total number of MCMC iterations works adequately in our experiments. Also, we temper the likelihood (3) when targeting ϕ and χ only, since θ and ς are very well-informed by the data as it is.

D Gradient-based MCMC sampling

As mentioned in Section 4.4, for our MALA and HMC implementations of EDiSC, we use analytically derived gradients. We can do MALA or HMC sampling for $\phi, \theta, \chi, \varsigma$ in the same way as discussed in Zafar and Nicholls (2022, Appendix C): alternately sample each variable while conditioning on the others. Inference for ϕ in EDiSC is unchanged compared to DiSC, whereas to get proposals for θ, χ, ς, we need to derive the gradients for the log-likelihood

$$\log p(W|\phi, \psi) = \sum_{d=1}^{D} \log \sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k, \tau_d}$$

with respect to these variables, which we do in this section. Finally, we give the HMC algorithm used in our R implementation.

D.1 Derivation of $\nabla_{\xi^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$

From Zafar and Nicholls (2022) equation (32) we have

$$\frac{\partial}{\partial \psi_v^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) = \sum_{d \in \mathcal{D}(1:G,t)} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \left(\sum_{i=i_1}^{i_{L_d}} \mathbb{I}(v = w_{d,i}) - L_d \tilde{\psi}_v^{k,t} \right) \tag{7}$$

where $\mathcal{D}(g,t) = \{d : \gamma_d \in g \text{ and } \tau_d \in t\}$ is the set of snippet indices for genre(s) g and time(s) t. Also, by the chain rule we have

$$\frac{\partial}{\partial \xi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) = \sum_{v=1}^{V} \frac{\partial}{\partial \psi_v^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) \frac{\partial \psi_v^{k,t}}{\partial \xi_j^{k,t}}.$$

Now $\psi_v^{k,t} = \rho_v^{\mathrm{T}} \xi^{k,t} + \varsigma_v$ gives $\frac{\partial \psi_v^{k,t}}{\partial \xi_j^{k,t}} = \rho_{v,j}$, and so we have

$$\begin{aligned}
\frac{\partial}{\partial \xi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) &= \sum_{v=1}^{V} \rho_{v,j} \frac{\partial}{\partial \psi_v^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) \\
&= \sum_{d \in \mathcal{D}(1:G,t)} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \left(\sum_{i=i_1}^{i_{L_d}} \sum_{v=1}^{V} \rho_{v,j} \mathbb{I}(v = w_{d,i}) - L_d \sum_{v=1}^{V} \rho_{v,j} \tilde{\psi}_v^{k,t} \right) \\
&= \sum_{d \in \mathcal{D}(1:G,t)} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \left(\sum_{i=i_1}^{i_{L_d}} \rho_{w_{d,i},j} - L_d \rho_{\cdot,j}^{\mathrm{T}} \tilde{\psi}^{k,t} \right) \qquad (8)
\end{aligned}$$

which are the elements of vector $\nabla_{\xi^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$ for $j \in \{1, \ldots, M\}$.

D.2 Derivation of $\nabla_{\theta^t} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$

The relationship $\xi_j^{k,t} = \chi_j^k + \theta_j^t$ gives $\frac{\partial \xi_j^{k,t}}{\partial \theta_j^t} = 1$ for all $k \in \{1, \ldots, K\}$, so applying the chain rule to (8) we get

$$\begin{aligned}
\frac{\partial}{\partial \theta_j^t} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) &= \sum_{k=1}^{K} \frac{\partial}{\partial \xi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) \\
&= \sum_{d \in \mathcal{D}(1:G,t)} \left(\sum_{i=i_1}^{i_{L_d}} \rho_{w_{d,i},j} - L_d \sum_{k=1}^{K} \frac{\tilde{\phi}_k^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,t}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,t} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,t}} \rho_{\cdot,j}^{\mathrm{T}} \tilde{\psi}^{k,t} \right)
\end{aligned}$$

which are the elements of vector $\nabla_{\theta^t} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t})$ for $j \in \{1, \ldots, M\}$.

D.3 Derivation of $\nabla_{\chi^k} \log p(W|\phi, \psi)$

The relationship $\xi_j^{k,t} = \chi_j^k + \theta_j^t$ gives $\frac{\partial \xi_j^{k,t}}{\partial \chi_j^k} = 1$ for all $t \in \{1, \ldots, T\}$, so given the independence between time periods and applying the chain rule to (8) we get

$$\begin{aligned}
\frac{\partial}{\partial \chi_j^k} \log p(W|\phi, \psi) &= \sum_{t=1}^{T} \frac{\partial}{\partial \xi_j^{k,t}} \log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t}, \psi^{\cdot,t}) \\
&= \sum_{d=1}^{D} \frac{\tilde{\phi}_k^{\gamma_d,\tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k,\tau_d}}{\sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_d,\tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{l,\tau_d}} \left(\sum_{i=i_1}^{i_{L_d}} \rho_{w_{d,i},j} - L_d \rho_{\cdot,j}^{\mathrm{T}} \tilde{\psi}^{k,\tau_d} \right)
\end{aligned}$$

which are the elements of vector $\nabla_{\chi^k} \log p(W|\phi, \psi)$ for $j \in \{1, \ldots, M\}$.

D.4 Derivation of $\nabla_\varsigma \log p(W|\phi,\psi)$

The relationship $\psi_j^{k,t} = \rho_j^{\mathrm{T}}\xi^{k,t} + \varsigma_j$ gives $\frac{\partial \psi_j^{k,t}}{\partial \varsigma_j} = 1$ for all $k \in \{1,\ldots,K\}$ and for all $t \in \{1,\ldots,T\}$, so given the independence between time periods and applying the chain rule to (7) we get

$$\begin{aligned}
\frac{\partial}{\partial \varsigma_j}\log p(W|\phi,\psi) &= \sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\partial}{\partial \psi_j^{k,t}}\log p(W_{\mathcal{D}(1:G,t)}|\phi^{\cdot,t},\psi^{\cdot,t}) \\
&= \sum_{d=1}^{D}\sum_{k=1}^{K}\frac{\tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}}\tilde{\psi}_{w_{d,i}}^{k,\tau_d}}{\sum_{l=1}^{K}\tilde{\phi}_l^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}}\tilde{\psi}_{w_{d,i}}^{l,\tau_t}}\left(\sum_{i=i_1}^{i_{L_d}}\mathbb{I}(j=w_{d,i}) - L_d\tilde{\psi}_j^{k,\tau_d}\right) \\
&= \sum_{d=1}^{D}\left(\sum_{i=i_1}^{i_{L_d}}\mathbb{I}(j=w_{d,i}) - L_d\sum_{k=1}^{K}\frac{\tilde{\phi}_k^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}}\tilde{\psi}_{w_{d,i}}^{k,\tau_d}}{\sum_{l=1}^{K}\tilde{\phi}_l^{\gamma_d,\tau_d}\prod_{i=i_1}^{i_{L_d}}\tilde{\psi}_{w_{d,i}}^{l,\tau_t}}\tilde{\psi}_j^{k,\tau_d}\right) \qquad (9)
\end{aligned}$$

which are the elements of vector $\nabla_\varsigma \log p(W|\phi,\psi)$ for $j \in \{1,\ldots,V\}$. Note that the term on the left in (9) can be simplified by considering that $\sum_{d=1}^{D}\sum_{i=i_1}^{i_{L_d}}\mathbb{I}(j=w_{d,i}) = N_{j,\cdot,\cdot}^{W,\cdot}$ is the number of times word j appears in snippets W across all senses and time periods.

D.5 HMC sampling scheme

The HMC algorithm used in our implementation is adapted from Neal (2011), and is presented within the context of our EDiSC model and sampling scheme in Algorithm 2. We target one variable at a time, conditioning on all other variables, at the granularity of $\chi, \theta^t, \phi^{g,t}, \varsigma$, in that order. For increased efficiency, we iterate over time $t \in 1:T$ going forward for odd iterations and backward for even iterations. We use 10 leapfrog steps for χ, as this is arguably the most important and challenging variable to sample, and 5 leapfrog steps for all other variables. We run the MCMC for N iterations, which varies considerably between at least 1.5k for "bank" and up to 100k for "mus" in some runs, though typically 3k–10k is sufficient to ensure convergence in our experiments. Gradients for the potential energy $PE(x) = -\log\left(\pi(x)p(W|\phi(x),\psi(x))^{\lambda_n}\right)$ for each variable x take the form $\nabla_x PE(x) = -\nabla_x \log\pi(x) - \lambda_n \nabla_x \log p(W|\phi,\psi)$, where $\nabla_x \log\pi(x)$ are straightforward to compute since the prior densities $\pi(x)$ are normal (see Zafar and Nicholls 2022 Appendix C for explicit expressions), and $\nabla_x \log p(W|\phi,\psi)$ have been derived above.

The tuning scheme used to target an optimal acceptance rate is taken from Shaby and Wells (2010). We target an optimal acceptance rate $\alpha^{\mathrm{opt}} = 0.651$ for HMC as recommended by Beskos et al. (2013). For the initial proposal scales, we slightly adapt the authors' recommendations and use $\sigma_\phi^2 = 2.4^2/(K \times LF_\phi)$, $\sigma_\chi^2 = 2.4^2/((MK)^2 \times LF_\chi)$, $\sigma_\theta^2 = 2.4^2/(M^2 \times LF_\theta)$ and $\sigma_\varsigma^2 = 2.4^2/(V \times LF_\varsigma)$, where LF_x is the number of leapfrog steps used for variable x.

Algorithm 2 Hamiltonian Monte Carlo (HMC) sampling for EDiSC

set number of MCMC iterations N, tempering parameter N_{temp}, tuning parameters $N_{\text{tune}}, N_{\text{stop}}$, and target acceptance rate α^{opt}
for each variable $x \in \{\chi, \theta^t, \phi^{g,t}, \varsigma | g = 1, \ldots, G; t = 1, \ldots, T, \}$ **do**
    set number of leapfrog steps LF_x and initial proposal scale σ_x^2
    set tempering for variable x on or off
end for
initialise ϕ, χ, θ randomly and $\varsigma = \mathbf{0}$
for iteration $n \in 1 : N$ **do**
    for each $x \in \{\chi, \theta^t, \phi^{g,t}, \varsigma | g = 1, \ldots, G; t = 1, \ldots, T, \}$ **do**
        if tempering x and $n \leq N_{\text{temp}}$ **then** compute λ_n using (6); **else** set $\lambda_n = 1$
        draw initial momentum vector $q \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ of same dimension as x
        compute initial potential energy $PE_0 = PE(x) = -\log\left(\pi(x)p(W|\phi(x), \psi(x))^{\lambda_n}\right)$
        compute initial kinetic energy $KE_0 = KE(q) = \frac{1}{2}\sum q^2$
        make a half step for momentum at the beginning $q \leftarrow q - \frac{1}{2}\sigma_x \nabla_x PE(x)$
        save initial $x_0 = x$
        for $l \in 1 : LF_x$ **do**
            make a full step for position $x \leftarrow x + \sigma_x q$
            if $l \neq LF_x$, make a full step for momentum $q \leftarrow q - \sigma_x \nabla_x PE(x)$
        end for
        save final $x_1 = x$
        make a half step for momentum at the end $q \leftarrow q - \frac{1}{2}\sigma_x \nabla_x PE(x)$
        compute final potential and kinetic energies $PE_1 = PE(x)$ and $KE_1 = KE(q)$
        compute Hastings ratio $\alpha = \min\{1, \exp(PE_0 + KE_0 - PE_1 - KE_1)\}$
        with probability α, set $x = x_1$ (accept); else set $x = x_0$ (reject)
        if $n \geq N_{\text{tune}}$ and $n \leq N_{\text{stop}}$ **then**
            compute running acceptance rate $\bar{\alpha} = \frac{\#\text{ accepts}}{N_{\text{tune}}}$ using last N_{tune} iterations
            update proposal scale via $\log \sigma_x^2 \leftarrow \log \sigma_x^2 + C_n(\bar{\alpha} - \alpha^{\text{opt}})$ with $C_n = \left(\frac{n+1}{N_{\text{tune}}}\right)^{-0.8}$
        end if
    end for
end for

These work adequately for our data, but there is no particular reason to stick with these choices. The main consideration is to strike a balance, via trial and error, between setting the initial proposal scales too large (leading to numerical over/underflow) and too small (leading to slower mixing at the start of the chain). The scales are updated via $\log \sigma_x^2 \leftarrow \log \sigma_x^2 + C_n(\bar{\alpha} - \alpha^{\text{opt}})$ at MCMC iteration n, using a running acceptance rate $\bar{\alpha}$ computed on the last $N_{\text{tune}} = 10$ iterations. We typically stop tuning after $N_{\text{stop}} = N/2$ iterations, though it is harmless to continue tuning since $C_n = \left(\frac{n+1}{N_{\text{tune}}}\right)^{-0.8}$ by design converges to zero as $n \to \infty$ and the tuning becomes minuscule.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication, there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (only required where there isn't already a statement of contribution within the paper itself).

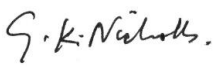
Title of paper	An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek
Publication status	Published
Publication details	Schyan Zafar and Geoff K. Nicholls (2024). An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek. *Computational Statistics & Data Analysis*, 199:108011, ISSN 0167-9473. DOI: 10.1016/j.csda.2024.108011

Student Confirmation

Student name	Schyan Zafar		
Contribution to the paper	I proposed the main research ideas for this paper; did all the coding and performed the experiments; and drafted the manuscript.		
Signature		Date	7 October 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name and title	Prof. Geoff K. Nicholls		
Supervisor comments			
Signature		Date	7 October 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Extensions and a Refactorisation of the Embedded Diachronic Sense Change (EDiSC) Model

Supplement to Chapter 3 "An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek". (This chapter should be read as an extended appendix to Chapter 3 written post-publication.)

Abstract

This exploratory chapter builds on the EDiSC model. In any document where a target word is used multiple times, it might be expected that its sense assignments within the document are correlated in some way. This idea is explored to give two possible extensions to the EDiSC model, both of which improve model performance.

Separately, any context word has a certain probability of being used marginally over the target-word senses, and also a probability of being associated with each of the target-word senses. This idea is used to explore a possible refactorisation of $\tilde{\psi}$ in the EDiSC model, which is interesting but does not improve performance.

1 Correlated sense assignments within a document

Both the DiSC and EDiSC models, like the GASC and SCAN models before them, assume that snippets are generated independently. This may not be the case in practice since, arguably, the snippets drawn from the same document may be correlated in some way. For example, the word "bank" may predominantly be used in the financial institution sense for any snippet drawn from a document discussing finance, or in the riverbank sense if drawn from a document discussing nature. Even without taking into account any further document-

level data, the fact that two snippets, d_1 and d_2 say, come from the same document suggests that their sense assignments z_{d_1} and z_{d_2} are not mutually independent.

In this section, we consider two alternative approaches to extend the EDiSC model to account for the correlation between sense assignments within the same document. The first approach utilises an idea from the topic-modelling literature (e.g. Blei and Lafferty 2006) that a document is a mixture over K topics with associated probabilities $\tilde{\eta}$. Analogously in our case, $\tilde{\eta}$ would denote the mixture probabilities over the K senses (rather than topics) of the target word within a document. We use this idea to give an extension to the model which we will call EDiSC-η. The second approach works on the assumption that there is a dominant document sense, and each snippet takes this sense with probability β. We use this idea to to give an alternative extension to the model which we will call EDiSC-β.

Before introducing the models, we need to define some new notation (contrast with Chapter 3 Section 2). Suppose our D snippets are extracted from a set of E documents. Individual snippets $d \in 1:D$ are extracted from documents $\Delta_d \in 1:E$. The set of snippets in each document $e \in 1:E$ is denoted $S_e := \{d : \Delta_d = e\}$. Since all snippets $d \in S_e$ share the same genre and time labels, we redefine the deterministic mappings so that γ_e and τ_e now refer to the genre and time labels of document e.

1.1 EDiSC-η model

Extending the EDiSC model (Chapter 3 Algorithm 1), we introduce a new parameter η to model the proportions of target-word senses within a document. For each document $e \in 1:E$, we let $\eta^e|\phi, \kappa_\eta \sim \mathcal{N}\left(\phi^{\gamma_e,\tau_e}, \text{diag}(\kappa_\eta)\right)$ and define $\tilde{\eta}^e = \text{softmax}(\eta^e)$ as the multinomial probabilities for the K senses. Then, in the generative model for EDiSC-η, the sense assignments are sampled as $z_d|\tilde{\eta}^{\Delta_d} \sim \text{Mult}\left(\tilde{\eta}_1^{\Delta_d}, \ldots, \tilde{\eta}_K^{\Delta_d}\right)$ for each snippet $d \in 1:D$. Given sense z_d, context words are sampled as before according to $\tilde{\psi}^{z_d,\tau_{\Delta_d}}$. The full EDiSC-$\eta$ model is given in Algorithm 1.

The definitions of ϕ, ψ and $\tilde{\psi}$ remain unchanged from before. However, the EDiSC sense prevalence probabilities $\tilde{\phi}$ no longer play a role in the EDiSC-η generative model. Instead, ϕ is now an AR(1) process representing the means of the new mixture parameter η, and $\tilde{\eta}$ represents the sense prevalence or proportions within documents. The model posterior can be written as

$$\begin{aligned}\pi(\phi, \eta, \psi|W) &\propto \pi(\phi)\pi(\psi)\pi(\eta|\phi)p(W|\psi,\eta) \\ &= \pi(\phi)\pi(\psi)\prod_{e=1}^{E}\pi(\eta^e|\phi)\prod_{d\in S_e}p(W_d|\psi,\eta^e)\end{aligned}$$

Algorithm 1 EDiSC-η: generative model

PRIOR MODEL

get word embeddings matrix ρ
fix hyperparameters $\kappa_\eta, \kappa_\phi, \kappa_\theta, \kappa_\chi, \kappa_\varsigma, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
draw bias or correction parameter $\varsigma|\kappa_\varsigma \sim \mathcal{N}(0, \mathrm{diag}(\kappa_\varsigma))$
for genre $g \in 1:G$ **do**
    draw initial sense prevalence mean $\phi^{g,1}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \mathrm{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
    for time $t \in 2:T$ **do**
        draw sense prevalence mean $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \mathrm{diag}(\kappa_\phi)\right)$
    end for
end for
draw initial time embedding $\theta^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \mathrm{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$
for time $t \in 2:T$ **do**
    draw time embedding $\theta^t|\theta^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \theta^{t-1}, \mathrm{diag}(\kappa_\theta)\right)$
end for
for sense $k \in 1:K$ **do**
    draw sense embedding $\chi^k|\kappa_\chi \sim \mathcal{N}(0, \mathrm{diag}(\kappa_\chi))$
    for time $t \in 1:T$ **do**
        set sense-time embedding $\xi^{k,t} = \chi^k + \theta^t$
        set context-word probability parameter $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$
    end for
end for
transform ψ into probabilities $\tilde{\psi}$ using softmax

OBSERVATION MODEL

for document $e \in 1:E$ **do**
    draw mixture parameter $\eta^e|\phi, \kappa_\eta \sim \mathcal{N}(\phi^{\gamma_e, \tau_e}, \mathrm{diag}(\kappa_\eta))$ and set $\tilde{\eta}^e = \mathrm{softmax}(\eta^e)$
    for snippet $d \in S_e$ **do**
        draw sense assignment $z_d|\tilde{\eta}^e \sim \mathrm{Mult}(\tilde{\eta}^e_1, \ldots, \tilde{\eta}^e_K)$
        for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
            draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_e} \sim \mathrm{Mult}\left(\tilde{\psi}_1^{z_d,\tau_e}, \ldots, \tilde{\psi}_V^{z_d,\tau_e}\right)$
        end for
    end for
end for

$$
\begin{aligned}
&= \pi(\phi)\pi(\psi) \prod_{e=1}^{E} \pi(\eta^e|\phi) \prod_{d \in S_e} \sum_{k=1}^{K} p(W_d, z_d = k|\psi, \eta^e) \\
&= \pi(\phi)\pi(\psi) \prod_{e=1}^{E} \pi(\eta^e|\phi) \prod_{d \in S_e} \sum_{k=1}^{K} p(z_d = k|\eta^e) p(W_d|z_d = k, \psi) \\
&= \pi(\phi)\pi(\psi) \prod_{e=1}^{E} \pi(\eta^e|\phi) \prod_{d \in S_e} \sum_{k=1}^{K} \tilde{\eta}^e_k \prod_{w \in W_d} \tilde{\psi}_w^{k,\tau_e},
\end{aligned}
$$

with the likelihood given by

$$p(W|\psi,\eta) = \prod_{e=1}^{E}\prod_{d\in S_e}\sum_{k=1}^{K}\tilde{\eta}_k^e \prod_{w\in W_d}\tilde{\psi}_w^{k,\tau_e}. \tag{1}$$

Priors $\pi(\phi)$ and $\pi(\psi)$ remain unchanged from before, but we need to elicit a new prior hyper-parameter κ_η for $\pi(\eta|\phi)$. To do this, we follow the same approach as Chapter 3 Appendix A and consider an extreme (i.e. 3-sigma) event.

Given two senses l and m, suppose we consider $\tilde{\eta}_l^e/\tilde{\eta}_m^e = X$ to be extreme. Then, on the log-scale, $\eta_l^e - \eta_m^e > \log X$ is a 3-sigma event. Now,

$$\begin{aligned}\mathbb{V}(\eta_l^e - \eta_m^e) &= \mathbb{E}[\mathbb{V}(\eta_l^e - \eta_m^e|\phi^{\gamma_e,\tau_e})] + \mathbb{V}[\mathbb{E}(\eta_l^e - \eta_m^e|\phi^{\gamma_e,\tau_e})]\\ &= \mathbb{E}(2\kappa_\eta) + \mathbb{V}(\phi_l^{\gamma_e,\tau_e} - \phi_m^{\gamma_e,\tau_e})\\ &= 2\kappa_\eta + \frac{2\kappa_\phi}{1-(\alpha_\phi)^2},\end{aligned}$$

so we express our preference with $3\left(2\kappa_\eta + \frac{2\kappa_\phi}{1-(\alpha_\phi)^2}\right)^{\frac{1}{2}} = \log X$, giving

$$\kappa_\eta = \frac{1}{2}\left(\frac{1}{3}\log X\right)^2 - \frac{\kappa_\phi}{1-(\alpha_\phi)^2}.$$

Arguably, we should allow very extreme variation, since we could plausibly have all the mass within a document concentrated on a single sense, so the ratio X could be infinite. On the other hand, a small prior variance may result in more stable posteriors. Since this is only exploratory analysis, an exact choice is not too important. We have $\kappa_\eta = 5.19$ for $X = 50\,000$ and $\kappa_\eta = 1.34$ for $X = 1\,000$. Rounding these, we experiment with a high value $\kappa_\eta = 5$ and a low value $\kappa_\eta = 1.25$.

The (unnormalised) posterior sense probabilities can be obtained from the expression inside the sum in (1) as

$$p(z_d = k|W_d,\psi,\eta^{\Delta_d}) \propto \tilde{\eta}_k^{\Delta_d}\prod_{w\in W_d}\tilde{\psi}_w^{k,\tau_{\Delta_d}}. \tag{2}$$

These will be required in the computation of the Brier scores later.

1.2 EDiSC-β model

As an alternative to EDiSC-η, we now extend the basic EDiSC by introducing a new parameter β to model the 'sense similarity' within documents. Suppose there is a dominant sense y_e for each document $e \in 1:E$, which is sampled according to the same $\tilde{\phi}^{\gamma_e,\tau_e}$ probabilities.

In the EDiSC-β model, we assume that each snippet $d \in S_e$ takes the sense y_e with probability $\beta_e \in [0,1]$, or the sense z_d is sampled independently as per EDiSC with probability $1-\beta_e$. The definitions of ϕ, ψ and $\tilde{\phi}, \tilde{\psi}$ remain unchanged compared to EDiSC.

We place a Beta(a,b) prior on each β_e, and set the hyperparameters a, b according to what we expect the behaviour to be *a priori*. On the one hand, $a = b = 1$ seems sensible as this corresponds to an uninformative uniform prior. On the other hand, the snippets S_e either have a dominant sense or they do not; so it seems sensible to concentrate the prior mass close to 1 and 0. We explore both intuitions by experimenting with $a = b = 1$ and $a = b = 0.5$. The full EDiSC-β model is given in Algorithm 2.

The model posterior can be written as

$$\begin{aligned}
\pi(\phi,\psi,\beta|W) &\propto \pi(\phi)\pi(\psi)\pi(\beta)p(W|\phi,\psi,\beta) \\
&= \pi(\phi)\pi(\psi)\prod_{e=1}^{E}\pi(\beta_e)\prod_{d\in S_e}p(W_d|\phi,\psi,\beta_e) \\
&= \pi(\phi)\pi(\psi)\prod_{e=1}^{E}\pi(\beta_e)\sum_{l=1}^{K}\prod_{d\in S_e}\sum_{k=1}^{K}p(W_d, z_d=k, y_e=l|\phi,\psi,\beta_e) \\
&= \pi(\phi)\pi(\psi)\prod_{e=1}^{E}\pi(\beta_e)\sum_{l=1}^{K}p(y_e=l|\phi)\prod_{d\in S_e}\sum_{k=1}^{K}p(z_d=k|y_e=l,\phi,\beta_e)p(W_d|z_d=k,\psi) \\
&= \pi(\phi)\pi(\psi)\prod_{e=1}^{E}\pi(\beta_e)\sum_{l=1}^{K}\tilde{\phi}_l^{\gamma_e,\tau_e}\prod_{d\in S_e}\sum_{k=1}^{K}\left(\beta_e\mathbb{I}(k=l)+(1-\beta_e)\tilde{\phi}_k^{\gamma_e,\tau_e}\right)\prod_{w\in W_d}\tilde{\psi}_w^{k,\tau_e},
\end{aligned}$$

with the likelihood given by

$$p(W|\phi,\psi,\beta) = \prod_{e=1}^{E}\sum_{l=1}^{K}\tilde{\phi}_l^{\gamma_e,\tau_e}\prod_{d\in S_e}\sum_{k=1}^{K}\left(\beta_e\mathbb{I}(k=l)+(1-\beta_e)\tilde{\phi}_k^{\gamma_e,\tau_e}\right)\prod_{w\in W_d}\tilde{\psi}_w^{k,\tau_e}. \tag{3}$$

To get the posterior sense probabilities, we consider the expression in (3) for a single document e without marginalising over $z_d = k$, $k \in 1:K$, that is

$$p(W_{S_e}, z_{S_e}|\phi,\psi,\beta_e) = \sum_{l=1}^{K}\tilde{\phi}_l^{\gamma_e,\tau_e}\prod_{d\in S_e}\left(\beta_e\mathbb{I}(z_d=l)+(1-\beta_e)\tilde{\phi}_{z_d}^{\gamma_e,\tau_e}\right)\prod_{w\in W_d}\tilde{\psi}_w^{z_d,\tau_e}. \tag{4}$$

Given a specific snippet $d \in S_e$, let $S_e \backslash d$ denote the indices of snippets in document e excluding snippet d. Now, $p(z_d, z_{S_e\backslash d}|W_{S_e},\phi,\psi,\beta_e) = p(z_{S_e}|W_{S_e},\phi,\psi,\beta_e) \propto p(W_{S_e}, z_{S_e}|\phi,\psi,\beta_e)$, which is given by (4). Therefore, the unnormalised probability $p(z_d|W_{S_e},\phi,\psi,\beta_e)$ can be

Algorithm 2 EDiSC-β: generative model

PRIOR MODEL

get word embeddings matrix ρ
fix hyperparameters $a, b, \kappa_\phi, \kappa_\theta, \kappa_\chi, \kappa_\varsigma, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
draw bias or correction parameter $\varsigma|\kappa_\varsigma \sim \mathcal{N}(0, \text{diag}(\kappa_\varsigma))$
for genre $g \in 1:G$ **do**
    draw initial sense prevalence parameter $\phi^{g,1}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
    for time $t \in 2:T$ **do**
        draw sense prevalence parameter $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \text{diag}(\kappa_\phi)\right)$
    end for
end for
draw initial time embedding $\theta^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$
for time $t \in 2:T$ **do**
    draw time embedding $\theta^t|\theta^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \theta^{t-1}, \text{diag}(\kappa_\theta)\right)$
end for
for sense $k \in 1:K$ **do**
    draw sense embedding $\chi^k|\kappa_\chi \sim \mathcal{N}(0, \text{diag}(\kappa_\chi))$
    for time $t \in 1:T$ **do**
        set sense-time embedding $\xi^{k,t} = \chi^k + \theta^t$
        set context-word probability parameter $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$
    end for
end for
transform ϕ and ψ into probabilities $\tilde{\phi}$ and $\tilde{\psi}$ using softmax

OBSERVATION MODEL

for document $e \in 1:E$ **do**
    draw dominant target-word sense $y_e|\tilde{\phi}^{\gamma_e,\tau_e} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_e,\tau_e}, \ldots, \tilde{\phi}_K^{\gamma_e,\tau_e}\right)$
    draw sense similarity parameter $\beta_e \sim \text{Beta}(a, b)$
    for snippet $d \in S_e$ **do**
        draw $u_d \sim \mathcal{U}(0,1)$
        if $u_d \leq \beta_e$ **then**
            set snippet sense assignment $z_d = y_e$
        else
            draw snippet sense assignment $z_d|\tilde{\phi}^{\gamma_e,\tau_e} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_e,\tau_e}, \ldots, \tilde{\phi}_K^{\gamma_e,\tau_e}\right)$
        end if
        for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
            draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_e} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d,\tau_e}, \ldots, \tilde{\psi}_V^{z_d,\tau_e}\right)$
        end for
    end for
end for

obtained by summing (4) over $z_{d'} = j$, $j \in 1:K$, for $d' \in S_e \backslash d$. Hence we have

$$p(z_d = k|W_{S_e}, \phi, \psi, \beta_e) \propto \sum_{l=1}^{K} \tilde{\phi}_l^{\gamma_e,\tau_e} \left(\left(\beta_e \mathbb{I}(k=l) + (1-\beta_e)\tilde{\phi}_k^{\gamma_e,\tau_e}\right) \prod_{w \in W_d} \tilde{\psi}_w^{k,\tau_e} \times \prod_{d' \in S_e \backslash d} \sum_{j=1}^{K} \left(\beta_e \mathbb{I}(j=l) + (1-\beta_e)\tilde{\phi}_j^{\gamma_e,\tau_e}\right) \prod_{w \in W_{d'}} \tilde{\psi}_w^{j,\tau_e} \right). \quad (5)$$

Table 1: Brier scores on test data using different models

	EDiSC	EDiSC-η		EDiSC-β	
		$\kappa_\eta = 5$	$\kappa_\eta = 1.25$	$a = b = 1$	$a = b = 0.5$
bank	0.133	0.117	0.121	0.119	0.118
kosmos	0.332	0.311	0.311	0.314	0.313
mus	0.093	0.058	0.061	0.062	0.061

Notice that the posterior sense probability (2) for any snippet $d \in S_e$ is conditionally independent of all $S_e \backslash d$ under EDiSC-η, whereas the probability (5) under EDiSC-β is not.

1.3 Results and discussion

We fit both EDiSC-η and EDiSC-β using Stan to avoid deriving and coding the gradients by hand. We compute Brier scores for "bank", "kosmos" and "mus" as in Chapter 3 Section 5.1, estimating $\hat{p}(z_d = k)$ for EDiSC-η and EDiSC-β using (2) and (5) respectively, and using the same modelling choices for M and K as in Chapter 3 Section 5.2. On "harmonia", the converged model senses could not be mapped to the true target-word senses. This could be due to the MCMC sampler used rather than the model (cf. Chapter 3 Appendix C) but we have not verified this.

Assessing the models' performance on 'true-model' recovery as in Chapter 3 Section 5.3 would not be appropriate. The proxy true model used there was just the model conditioned on the true sense labels o. At the level of $\tilde{\phi}$, the DiSC and EDiSC models are identical; so $\tilde{\phi}|(z = o)$ is not biased towards one or the other. However, when comparing EDiSC, EDiSC-η and EDiSC-β, the three models differ in their usage of ϕ; so the equivalent comparison favours whichever one of these models we use to calibrate $\tilde{\phi}|(z = o)$.

The Brier scores are reported in Table 1. We see that both EDiSC-η and EDiSC-β provide an improvement in predictive accuracy over EDiSC of similar magnitude. This supports the hypothesis that target-word senses within a document are correlated to some extent. Given the similar levels of improvement, choosing between the models would be down to the inferential objective, since the interpretation of parameters differs between them.

For EDiSC-η, the higher prior variance $\kappa_\eta = 5$ seems to be marginally better; and for EDiSC-β, the non-uniform prior Beta$(0.5, 0.5)$ seems to be marginally better. However, based on these exploratory analyses, the exact prior hyperparameters do not seem too important, although further experimentation might be useful. Setting the hyperparameters on a scale

varying by the number of snippets per document is another option that might be worth exploring.

As a final thought on this exploratory theme, given that both EDiSC-η and EDiSC-β improve model performance individually, it might be worth experimenting with a combination of both models. For example, we could posit a model that uses document mixture probabilities as in EDiSC-η, as well as a dominant document sense as in EDiSC-β. However, we have not explored this.

2 A refactorisation of $\tilde{\psi}$

[This section may be skipped on a first reading. The model described here did not yield an improvement over EDiSC, but is included here as it may be used in future work.]

In all the models discussed so far, $\tilde{\psi}$ represents the probabilities with which context words are sampled in the generative model. It is answering the question: if the target word is used in a given sense, what is the probability of seeing a particular word in the context? The models do not *directly* answer the opposite question: if a particular word is seen in the context, what is the probability that the target word is used in a given sense? This question can be answered *a posteriori* using Chapter 3 equation (4), or a discriminative model might be posited to address the question directly. Here we explore an alternative factorisation of $\tilde{\psi}$, which attempts to address this discriminative question more directly whilst keeping the generative structure of the EDiSC model.

To motivate the refactorisation, intuitively, certain context words are associated more strongly with particular target-word senses. In many cases, these context words would be assigned a high probability under one sense but not the others, for example the words "river" or "stream" for the riverbank sense of "bank", which allows the model to discriminate the sense. But what if it is a rare word such as "rivage"? Being rare, it would have a very low context probability under any sense; however, it should still have a very strong association with the riverbank sense. The additive structure $\chi^k + \theta^t$ introduced in the DiSC model (Chapter 2 Section 4) addresses this to some extent, as the χ value for the word should be high under one sense only, but the θ value should be low. We now attempt to model this intuition more explicitly.

2.1 Motivation

Consider first the non-embedded model DiSC. The observation model for DiSC draws $w_{d,i}$, the word at position i in snippet d, given the sense assignment z_d and time period τ_d,

according to the probability distribution $\tilde{\psi}^{z_d,\tau_d}$ over context words $1:V$ in the vocabulary. Thus, in this observation model, we have

$$\begin{aligned}\tilde{\psi}_v^{k,t} &= p(w_{d,i}=v|z_d=k,\tau_d=t)\\ &= \frac{p(z_d=k|w_{d,i}=v,\tau_d=t)p(w_{d,i}=v|\tau_d=t)}{\sum_{u=1}^V p(z_d=k|w_{d,i}=u,\tau_d=t)p(w_{d,i}=u|\tau_d=t)}\end{aligned} \tag{6}$$

from an application of the Bayes' theorem. We now define

$$\begin{aligned}\tilde{\chi}_v^k &= p(z_d=k|w_{d,i}=v,\tau_d=t)\\ &= p(z_d=k|w_{d,i}=v)\end{aligned}$$

independently of t as the probability that snippet d has target-word sense k, given that context-word v appears in it. We also define

$$\tilde{\theta}_v^t = p(w_{d,i}=v|\tau_d=t)$$

as the probability of sampling context word v for any position $i \in \{i_1,\ldots,i_{L_d}\}$ in snippet d, given that the snippet appears at time t. Thus, $\tilde{\chi}_v^k$ and $\tilde{\theta}_v^t$ together control $\tilde{\psi}_v^{k,t}$, the probability that context word v appears under sense k at time t, and we have from (6) that

$$\tilde{\psi}_v^{k,t} = \frac{\tilde{\chi}_v^k\tilde{\theta}_v^t}{\sum_{u=1}^V \tilde{\chi}_u^k\tilde{\theta}_u^t}. \tag{7}$$

Notice that for (7) to be a valid probability, it is not necessary for $\tilde{\chi}_v^k$ and $\tilde{\theta}_v^t$ to be probabilities: it is sufficient to have $\tilde{\chi}_v^k, \tilde{\theta}_v^t > 0$. If for example we set $\tilde{\chi}^k = \exp(\chi^k)$ and $\tilde{\theta}^t = \exp(\theta^t)$, then from (7) we recover exactly the DiSC model

$$\tilde{\psi}_v^{k,t} = \frac{\exp(\chi_v^k)\exp(\theta_v^t)}{\sum_{u=1}^V \exp(\chi_u^k)\exp(\theta_u^t)} = \frac{\exp(\chi_v^k+\theta_v^t)}{\sum_{u=1}^V \exp(\chi_u^k+\theta_u^t)} = \frac{\exp(\psi_v^{k,t})}{\sum_{u=1}^V \exp(\psi_u^{k,t})},$$

since we have $\psi^{k,t} = \chi^k + \theta^t$ in DiSC. However, if we set $\tilde{\chi}_v$ to be a softmax over senses, i.e. $\tilde{\chi}_v = \frac{\exp(\chi_v)}{\sum_{l=1}^K \exp(\chi_v^l)}$, and $\tilde{\theta}^t$ to be a softmax over context words, i.e. $\tilde{\theta}^t = \frac{\exp(\theta^t)}{\sum_{w=1}^V \exp(\theta_w^t)}$, then from (7) we get a new factorisation for $\tilde{\psi}$:

$$\tilde{\psi}_v^{k,t} = \frac{\exp(\chi_v^k+\theta_v^t)/\sum_{l=1}^K \exp(\chi_v^l)}{\sum_{u=1}^V \left(\exp(\chi_u^k+\theta_u^t)/\sum_{l=1}^K \exp(\chi_u^l)\right)}. \tag{8}$$

Note that it is enough to set $\tilde{\theta}^t = \exp(\theta^t)$, since the denominator $\sum_{w=1}^V \exp(\theta_w^t)$ cancels out in (8). To keep the notation consistent with DiSC, we redefine

$$\begin{aligned}\psi_v^{k,t} &= \log \tilde{\chi}_v^k\tilde{\theta}_v^t\\ &= \chi_v^k + \theta_v^t - \log\sum_{l=1}^K \exp(\chi_v^l)\end{aligned} \tag{9}$$

so that $\tilde{\psi}_v^{k,t} = \frac{\exp(\psi_v^{k,t})}{\sum_{u=1}^V \exp(\psi_u^{k,t})}$. We call this model DiSC-f, where 'f' denotes the new factorisation.

Algorithm 3 EDiSC-f: generative model

PRIOR MODEL

get word embeddings matrix ρ
fix hyperparameters $\kappa_\phi, \kappa_\theta, \kappa_\chi, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
for genre $g \in 1:G$ **do**
    draw initial sense prevalence parameter $\phi^{g,1}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
    for time $t \in 2:T$ **do**
        draw sense prevalence parameter $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \text{diag}(\kappa_\phi)\right)$
    end for
end for
draw initial time embedding $\dot{\theta}^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$ and set $\theta^1 = \rho\dot{\theta}^1$
for time $t \in 2:T$ **do**
    draw time embedding $\dot{\theta}^t|\dot{\theta}^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \dot{\theta}^{t-1}, \text{diag}(\kappa_\theta)\right)$ and set $\theta^t = \rho\dot{\theta}^t$
end for
for sense $k \in 1:K$ **do**
    draw sense embedding $\dot{\chi}^k|\kappa_\chi \sim \mathcal{N}(0, \text{diag}(\kappa_\chi))$ and set $\chi^k = \rho\dot{\chi}^k$
    for time $t \in 1:T$ **do**
        set context-word probability parameter $\psi^{k,t} = \chi^k + \theta^t - \log\sum_{l=1}^K \exp(\chi^l)$
    end for
end for
transform ϕ and ψ into probabilities $\tilde{\phi}$ and $\tilde{\psi}$ using softmax

OBSERVATION MODEL

for snippet $d \in 1:D$ (genre γ_d, time τ_d, length L_d) **do**
    draw sense assignment $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}\left(\tilde{\phi}_1^{\gamma_d,\tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d,\tau_d}\right)$
    for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
        draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}\left(\tilde{\psi}_1^{z_d,\tau_d}, \ldots, \tilde{\psi}_V^{z_d,\tau_d}\right)$
    end for
end for

2.2 EDiSC-f model

We now define EDiSC-f, the embedded version of this model. We have the $V \times M$ embedding matrix ρ. We introduce new variables $\dot{\chi}^k$ and $\dot{\theta}^t$, which are vectors in the M-dimensional embedding space, and set $\chi^k = \rho\dot{\chi}^k$, $\theta^t = \rho\dot{\theta}^t$. We keep ψ as in (9). If a correction or bias term ς_v is desired, it may be added to (9), but we leave it out since the extra term (compared to EDiSC) $\log\sum_{l=1}^K \exp(\chi_v^l)$ already provides some degree of correction.

We place the same AR(1) priors on $\dot{\theta}$ as in EDiSC. That is, $\dot{\theta}_m^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)$ and $\dot{\theta}_m^t|\dot{\theta}_m^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \dot{\theta}_m^{t-1}, \kappa_\theta\right)$ for $t \in 1:T$. And we place independent Gaussian priors $\dot{\chi}_m^k|\kappa_\chi \sim \mathcal{N}(0, \kappa_\chi)$ on each entry of $\dot{\chi}^k, k \in 1:K$, as in EDiSC. However, we elicit new variance hyperparameters for EDiSC-f as described below. The full EDiSC-f model is given in Algorithm 3.

2.3 Hyperparameters

Following a similar argument to Chapter 3 Appendix A, we elicit priors by defining what we consider to be extreme (i.e. 3-sigma) events, and using these to set quantiles.

$\tilde{\theta}^t$ represents the context-word probabilities at time t across all senses. For any fixed time t and pair of words $x, y \in \{1, \ldots, V\}$, the ratio of context-word probabilities of the order $\tilde{\theta}^t_x / \tilde{\theta}^t_y \approx 1\,000$ may be considered extreme. Therefore, on the log scale, we consider $\theta^t_x - \theta^t_y > \log 1\,000$ to be a 3-sigma event. We have $\mathbb{V}(\theta^t_x - \theta^t_y) = (\rho_x - \rho_y)^{\mathrm{T}} \mathbb{V}(\dot{\theta}^t)(\rho_x - \rho_y)$ where $\mathbb{V}(\dot{\theta}^t)$ is an $M \times M$ diagonal matrix with entries $\frac{\kappa_\theta}{1-(\alpha_\theta)^2}$. Hence $\mathbb{V}(\theta^t_x - \theta^t_y) = (\rho_x - \rho_y)^{\mathrm{T}}(\rho_x - \rho_y)\frac{\kappa_\theta}{1-(\alpha_\theta)^2}$. We approximate $(\rho_x - \rho_y)^{\mathrm{T}}(\rho_x - \rho_y)$ with its median c over all pairs $x, y \in \{1, \ldots, V\}$, and express our preference with $3\left(c\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)^{\frac{1}{2}} = \log 1\,000$. With $\alpha_\theta = 0.9$ as in EDiSC, we get $\kappa_\theta = (1 - (\alpha_\theta)^2)\left(\frac{1}{3}\log 1\,000\right)^2 / c \approx 1/c$ on rounding.

$\tilde{\chi}_v$ represents the snippet-sense probabilities, having observed context word v in the snippet. For a given context word v and pair of target-word senses $k, l \in \{1, \ldots, K\}$, we could have a situation where v only appears under sense k and not under l, and so the ratio $\tilde{\chi^k_v}/\tilde{\chi^l_v}$ is unbounded. On the other hand, we expect very few context words that are highly informative of the sense to behave in this way. For most context words, we would expect the ratio $\tilde{\chi^k_v}/\tilde{\chi^l_v}$ to be closer to 1, since the majority of context words would be used independently of the target-word sense. If we regard $\tilde{\chi^k_v}/\tilde{\chi^l_v} = X$ to be extreme, then on the log scale we consider $\chi^k_v - \chi^l_v > \log X$ to be a 3-sigma event. We have $\mathbb{V}(\chi^k_v - \chi^l_v) = \rho_v^{\mathrm{T}} \mathbb{V}(\dot{\chi}^k - \dot{\chi}^l)\rho_v$ where $\mathbb{V}(\dot{\chi}^k - \dot{\chi}^l)$ is an $M \times M$ diagonal matrix with entries $2\kappa_\chi$. Hence $\mathbb{V}(\chi^k_v - \chi^l_v) = 2\rho_v^{\mathrm{T}}\rho_v\kappa_\chi$. We approximate $\rho_v^{\mathrm{T}}\rho_v$ with its median b over all $v \in \{1, \ldots, V\}$, and express our preference with $3(2b\kappa_\chi)^{\frac{1}{2}} = \log X$. We experiment with two values, $X = 100$ and $X = 10$. This gives $\kappa_\chi = \left(\frac{1}{3}\log X\right)^2 / 2b \approx 1/b$ or $0.25/b$ on rounding respectively.

2.4 Results

The likelihood and inference for EDiSC-f remain unchanged compared to EDiSC in terms of $\tilde{\phi}$ and $\tilde{\psi}$, since all changes are made at the log-scale ψ level. We fit the model using Stan with the same modelling choices M and K as for EDiSC. We report the Brier scores for "bank", "kosmos" and "mus" in Table 2. On "harmonia", again, the converged model senses could not be mapped to the true target-word senses. Disappointingly, we see that the refactorisation leads to a deterioration in predictive accuracy compared to EDiSC.

We also compare the performance of EDiSC-f (for $\kappa_\chi = 0.25/b$) against EDiSC on 'true-model' recovery for "kosmos" as per Chapter 3 Section 5.3. This is the most favourable

Table 2: Brier scores on test data using EDiSC and EDiSC-f

	EDiSC	EDiSC-f	
		$\kappa_\chi = 1/b$	$\kappa_\chi = 0.25/b$
bank	0.133	0.147	0.148
kosmos	0.332	0.338	0.336
mus	0.093	0.118	0.108

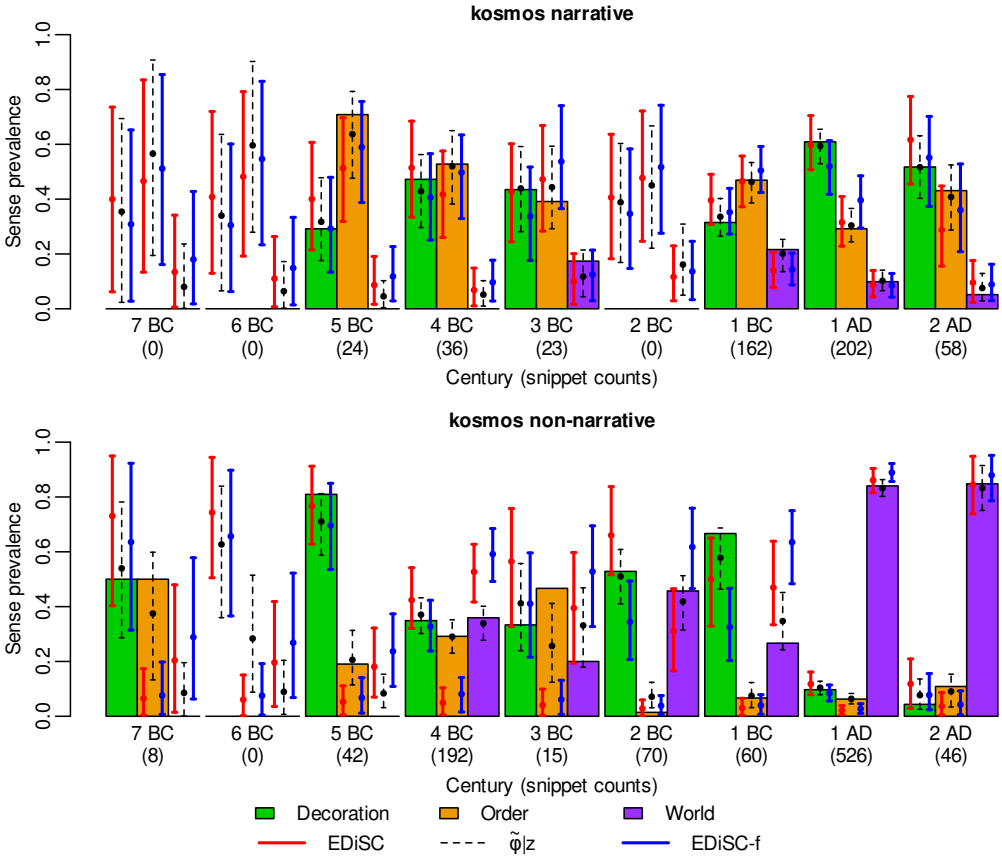


Figure 1: "Kosmos" expert-annotated empirical sense prevalence (coloured bars), and 95% HPD intervals (error bars) and posterior means (circles) from the model output

Table 3: "Kosmos" Bayes factors BF_{01} on a $\log_{10}$ scale for nested 'true' model $H_0 : \tilde{\phi}^{g,t} \in \mathcal{S}^{g,t}$ over each of H_1 : EDiSC and H_1 : EDiSC-f. Red indicates incorrect rejection of H_0.

Model	Genre	7 BC	6 BC	5 BC	4 BC	3 BC	2 BC	1 BC	1 AD	2 AD
EDiSC	narrative	0.43	0.63	1.06	1.20	0.92	0.62	1.09	1.63	0.97
EDiSC-f	narrative	0.36	0.55	0.90	1.13	0.84	0.64	1.27	1.14	1.19
EDiSC	non-narr	−0.35	0.01	−0.36	$-\infty$	−0.99	0.37	0.77	−0.07	0.92
EDiSC-f	non-narr	−0.91	−0.37	−0.77	$-\infty$	−0.68	0.27	−1.07	0.34	1.06

comparison for EDiSC-f, as its performance is worse in the other experiments. The sense prevalence graphs are shown in Figure 1, and the corresponding Bayes factors in Table 3. The performance of EDiSC-f is somewhat comparable to EDiSC, but on the whole slightly worse even in this best case. Hence, even though this line of exploration seemed interesting, we discard it based on the results.

Whilst the EDiSC-f model did not yield an improvement over EDiSC, it does show that there is more than one logical way of modelling the sense and time effects (i.e. other than as additive). The alternative modelling may prove useful to generate ideas for future model development.

Chapter 5

Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks

Schyan Zafar and **Geoff K. Nicholls**

Abstract

Generalised Bayesian Inference (GBI) attempts to address model misspecification in a standard Bayesian setup by tempering the likelihood. The likelihood is raised to a fractional power, called the learning rate, which reduces its importance in the posterior and has been established as a method to address certain kinds of model misspecification. Posterior Predictive Checks (PPC) attempt to detect model misspecification by locating a diagnostic, computed on the observed data, within the posterior predictive distribution of the diagnostic. This can be used to construct a hypothesis test where a small p-value indicates potential misfit. The recent Embedded Diachronic Sense Change (EDiSC) model suffers from misspecification and benefits from likelihood tempering. Using EDiSC as a case study, this exploratory work examines whether PPC could be used in a novel way to set the learning rate in a GBI setup. Specifically, the learning rate selected is the lowest value for which a hypothesis test using the log likelihood diagnostic is not rejected at the 10% level. The experimental results are promising, though not definitive, and indicate the need for further research along the lines suggested here.

Keywords: likelihood tempering; model misspecification; power likelihood; power posterior

1 Introduction

Bayesian inference has long been a cornerstone of statistical analysis. It provides a coherent framework for updating prior beliefs $\pi(\cdot)$ about model parameter(s) ϑ in light of observed data y. However, traditional Bayesian methods assume that the observation model $y \sim p(\cdot|\vartheta)$

is correctly specified, an assumption that is often violated in practice. Generalised Bayesian Inference (GBI) extends traditional Bayesian methods to address model misspecification, offering a more robust approach to inference when the likelihood function does not perfectly represent the true data generating process (DGP). In its simplest form, GBI works by raising the likelihood in the standard Bayes' posterior to a fractional power $\lambda \in [0, 1]$ called the learning rate — a trick known as *tempering* in some contexts — giving the generalised Bayes' posterior

$$\pi_\lambda(\vartheta|y) \propto \pi(\vartheta)p(y|\vartheta)^\lambda. \tag{1}$$

It has been shown (Bissiri et al., 2016) that GBI with a tempered likelihood is a valid statistical procedure well supported by decision theory, and that (1) with some $\lambda < 1$ leads to more robust posterior belief updates compared to standard Bayes when the model is misspecified (Grünwald and van Ommen, 2017). The key question then is how to select an appropriate learning rate. This is a subject of ongoing research, and several methods for selecting λ have been proposed in the literature. Some recent papers on this subject include Grünwald and van Ommen (2017), Holmes and Walker (2017), Lyddon et al. (2019) and Syring and Martin (2018), and are reviewed in Wu and Martin (2023). However, these methods can only be implemented efficiently for models where the posterior summaries, integrals or other quantities required by the methods can be computed quickly. For high-dimensional models where the posteriors necessitate computationally expensive Markov Chain Monte Carlo (MCMC) sampling, the methods can be inefficient or even prohibitive. There is therefore a need for new methods for selecting λ that can be implemented efficiently in such cases.

Posterior Predictive Checks or Checking (PPC) is a classic tool used in Bayesian analysis to assess whether the posited model does a good job of capturing aspects of the distribution of the data that are important to the modeller. PPC uses the posterior distribution of the parameters to simulate replicated datasets, and hence predict the distribution of a suitable diagnostic function of the data (or a joint function of the data and parameters). This is compared to the same diagnostic function computed on the observed data to validate the model's assumptions and fit. For a well-specified model, we should not be able to tell the observed diagnostic apart from its posterior predictions. Conversely, for a misspecified model, the observed diagnostic should be an outlier within the posterior predictive distribution. This property may be used to construct a hypothesis test (notwithstanding any miscalibration) where a 1-sided p-value too close to 0 or 1 would indicate potential misfit.

A diagnostic commonly used in PPC is the log likelihood, which can be interpreted as a measure of agreement between the model parameters and data. Starting from a place of known model misspecification, specifically a model that follows the data too closely (i.e. overfits), the observed log likelihood should be large compared to the posterior predictive distribution; that is, it should fall in the right tail of the distribution, as the agreement between the parameters and data should be atypically good due to overfitting. If we then temper the likelihood with a learning rate $\lambda < 1$, the observed log likelihood should shift towards the left as we reduce its importance in the posterior, ultimately becoming an outlier in the left tail as $\lambda \to 0$. Intuitively, between these extremes, there should be a region where the observed diagnostic is typical of a well-specified model. The question we explore in this paper is whether we can use this intuition to select an appropriate λ; that is, pick a learning rate for which a hypothesis test using the log likelihood diagnostic is not rejected at a suitable significance level. We find that a 10% threshold works well in all our real-data examples. Our approach resembles Chakraborty et al. (2023). However, to the best of our knowledge, PPC has not previously been used to guide the choice of learning rate in quite this way.

This is exploratory work in its initial stage. We do not give any theoretical guarantees, nor make any claims that the approach studied in this paper could be generalised. In fact, it is easy to construct examples where the approach does *not* work, so it is certainly not universally applicable. We elaborate further on this in the discussion at the end, where we give some intuition for when the approach is valid.

This work has been motivated by the recent Embedded Diachronic Sense Change (EDiSC) model (Zafar and Nicholls, 2024a), which is exceptionally hard to fit to the data in question, and poses a unique set of challenges. It was found experimentally that EDiSC benefitted from likelihood tempering, making it an interesting case study for GBI. However, the existing state-of-the-art approaches for tuning λ discussed in Wu and Martin (2023) proved particularly challenging to implement, which motivated the search for a novel and readily implementable solution. Our approach for tuning λ is intuitive and gives promising results in this case study, though we acknowledge its limited scope. Crucially, the learning rate selected using our method gives an improvement in the model's predictive performance, in all the examples studied, compared to standard Bayes. The next stage of this work would be to investigate what other models and settings, if any, could this approach be generalised to.

The rest of this paper is organised as follows. In Section 2 we give some background on model

misspecification, GBI and PPC. In Section 3 we introduce our novel approach for selecting the learning rate using PPC. In Section 4 we describe the EDiSC model and inferential problem that motivated this work. In Section 5 we describe the data used to develop and test our method, and our performance measurement criteria. In Section 6 we present our experimental results on development and test data. Finally, in Section 7 we conclude with a discussion of possible conditions required for our method to work. Some further insights are given in the Appendix.

2 Background

In this section, we set out the notation and describe the concepts that will be required to develop our proposed method for setting the learning rate λ using PPC.

2.1 Model misspecification

Suppose we have a parametric model $y \sim p(\cdot|\vartheta)$ for data $y = (y_1, \ldots, y_D)$ depending on some (possibly multi-dimensional) parameter $\vartheta \in \Theta$ via likelihood $p(y|\vartheta)$. The data may come in pairs (x_d, y_d), with covariates $x_d = (x_{d,1}, \ldots, x_{d,p})$ for each $d = 1, \ldots, D$. Conditioning on the covariates via likelihood $p(y_d|x_d, \vartheta)$ is implied in that case, but we do not explicitly show it in this paper since $x = (x_1, \ldots, x_D)$ is not a random variable. In standard Bayesian inference, we place a prior $\pi(\vartheta)$ on the parameter and update our posterior beliefs about ϑ via Bayes' theorem:

$$\pi(\vartheta|y) \propto \pi(\vartheta)p(y|\vartheta). \tag{2}$$

In a correctly specified model, there exists a $\vartheta^* \in \Theta$ such that $y \sim p(\cdot|\vartheta^*)$ is the true DGP. Furthermore, under regularity conditions, the Bayes' posterior (2) converges to a normal distribution concentrated on the true ϑ^* in the limit of infinite data $D \to \infty$ by the Bernstein-von Mises theorem. This ϑ^* also coincides with the asymptotic value of the frequentist maximum likelihood estimate.

In practice, however, a statistical model is rarely well specified, since a model is by definition a simplified representation of some real-world phenomenon. In a misspecified model, say where the true DGP is $y \sim h(\cdot)$ for some $h(\cdot) \neq p(\cdot|\vartheta)$, a 'true' ϑ^* does not exist. Instead, as $D \to \infty$, under regularity conditions, both the Bayes' posterior (2) and the maximum likelihood estimator concentrate on the *pseudo-true* parameter

$$\vartheta^\dagger = \arg\min_{\vartheta \in \Theta} \mathrm{KL}\left(h(\cdot) \,\middle\|\, p(\cdot|\vartheta)\right), \tag{3}$$

(Kleijn and van der Vaart, 2012) where $\mathrm{KL}\left(h(\cdot) \parallel p(\cdot|\vartheta)\right) = \int h(y) \log\left(h(y)/p(y|\vartheta)\right) \mathrm{d}y$ is the Kullback-Leibler divergence. When ϑ has some physical meaning, this $\vartheta^\dagger$ is still interpretable and useful, and tells us something about the phenomenon being studied. However, since KL divergence places a lot of importance on the tail behaviour of the distributions in question, $\vartheta^\dagger$ learnt according to (3) can be sensitive to tail misspecifications, particularly if $h(\cdot)$ has heavier tails than $p(\cdot|\vartheta)$ (Jewson et al., 2018). In other words, the Bayes' posterior (2) lacks robustness and may overfit the model. Model misspecification can, of course, be more general, including where $\vartheta^\dagger$ has no meaningful interpretation.

2.2 Generalised Bayesian Inference

GBI, in its simplest form dating back to Walker and Hjort (2002) and Zhang (2006), attempts to address some kinds of model misspecification by reducing the importance of the likelihood in the standard Bayes' posterior (2) via tempering. The likelihood is raised to a fractional power $\lambda \in [0, 1]$, called the learning rate, where $\lambda = 1$ returns standard Bayes and $\lambda = 0$ gives just the prior. For any other λ, we get the generalised Bayes' posterior

$$\pi_\lambda(\vartheta|y) \propto \pi(\vartheta)p(y|\vartheta)^\lambda. \tag{4}$$

The smaller the value of λ, the slower the rate at which posterior beliefs about ϑ are updated with an increasing number of samples D, and vice versa. Importantly, the interpretation of the model and parameters remains unchanged with this form of GBI: only the rate of concentration of (4) around (3) is affected, and the Bernstein-von Mises theorem (under regularity conditions) still applies (Miller, 2021). Other forms of GBI extend Bayesian inference in more general and flexible ways, including loss-function-based (Bissiri et al., 2016) and scoring-rule-based (Pacchiardi et al., 2021) inference. See Pacchiardi (2022, Section 1.2) for a nice summary of how the generalisations fit together. These are beyond the scope of this paper.

Intuitively (Walker et al., 2005), a $\lambda < 1$ makes the generalised Bayes' posterior (4) more robust to inconsistencies arising from overfitting compared to (2) in any finite data sample, since the information coming from the data is reduced. However, too small a value can result in not learning enough about the DGP if the signal from the data gets too weak. Hence, we need to set the learning rate λ to a value suitable for the inferential task. Common inferential tasks include prediction and 'true model' recovery (in the sense of (3) for example), and it is well known that an optimal model for one task is not necessarily optimal for the other.

Four recent methods for selecting the learning rate have been reviewed in Wu and Martin (2023). The first of these is the SafeBayes algorithm of Grünwald and van Ommen (2017),

which involves a grid search over potential λ values, and selects the learning rate minimising a loss function. The loss function is computed by fitting the model iteratively, adding one data point in each iteration, and calculating the cumulative "posterior-expected posterior-randomised (log) loss" of predicting the next outcome in each fit. This method becomes computationally infeasible when we have hundreds or thousands of data points, and model fitting is done using MCMC, with each fit potentially taking hours to run.

Next, the (distinct) information-matching strategies of Holmes and Walker (2017) and Lyddon et al. (2019) give an explicit "oracle" learning rate as a formula. The computation of both oracles requires evaluating expectations with respect to the true DGP and using the pseudo-true $\vartheta^\dagger$. Since both of these are unknown, they are estimated by the empirical distribution of observed data and the maximum likelihood parameter estimates respectively. In our setting, with sparse high-dimensional data and high-dimensional parameters, these quantities are very hard to estimate with the accuracy needed to make these methods work.

Finally, the bootstrap-motivated generalised posterior calibration (GPC) algorithm of Syring and Martin (2018) attempts to tune λ iteratively over model fittings so that the generalised Bayesian posterior credible sets achieve the nominal frequentist coverage probability. This method is computationally expensive, even in simple applications where posterior credible sets are readily available, since it requires multiple model fittings using bootstrapped data for each candidate learning rate. Additionally, in high-dimensional applications where the posteriors are sampled using MCMC, a prohibitively large number of samples per model fitting is required to compute accurate coverage probabilities. A GPC variant is given by Winter et al. (2023), and similar remarks apply.

Separately, Carmona and Nicholls (2020) use the widely applicable information criterion (WAIC) (Watanabe, 2010; Vehtari et al., 2017), a predictive loss, to select the learning rate. This works well when the WAIC is a reliable estimator for the expected log pointwise predictive density (ELPD), which is not the case in our setting (Zafar and Nicholls, 2024a, Section 5.2).

Setting the learning rate is a subject of ongoing research, and our literature review did not find any other methods differing substantially from those discussed in the papers cited above.

2.3 Posterior Predictive Checks

PPC is a classic (Guttman, 1967; Rubin, 1984) diagnostic tool used frequently in Bayesian analysis to assess a model's ability to describe the data. Given observed data $y^{\text{obs}} = (y_1^{\text{obs}}, \ldots, y_D^{\text{obs}})$, PPC works by simulating $y^{\text{rep}} \sim p(\cdot|y^{\text{obs}})$ from the posterior predictive distribution

$$p(\cdot|y^{\text{obs}}) = \int_\Theta p(\cdot|\vartheta)\pi(\vartheta|y^{\text{obs}})\,\mathrm{d}\vartheta \tag{5}$$

to give replicated data $y^{\text{rep},n}, n = 1, \ldots, N$, where each replicate $y^{\text{rep},n} = (y_1^{\text{rep},n}, \ldots, y_D^{\text{rep},n})$ is a dataset of the same size as y^{obs}, using the same covariates x (if applicable). In a basic PPC, a diagnostic function $s(y)$ of the data is chosen, which may be some quantity of interest to the modeller such as the sample mean $\bar{y} = \frac{1}{D}\sum_{d=1}^D y_d$ or variance $\frac{1}{D-1}\sum_{d=1}^D (y_d - \bar{y})^2$, to investigate the compatibility between the data y and the model. The diagnostic is computed on both observed and replicated data to give, respectively, the observed statistic $s(y^{\text{obs}})$ and samples $s(y^{\text{rep},1}), \ldots, s(y^{\text{rep},N})$ from the *reference distribution* of $s(y^{\text{rep}})$. The idea is that, if the model is true, $s(y^{\text{rep}})$ has a mixture distribution over $\vartheta \in \Theta$ with a component $\vartheta = \vartheta^*$ that matches the generative model for $s(y^{\text{obs}})$. Conversely, if $s(y^{\text{obs}})$ is an outlier within the distribution of $s(y^{\text{rep}})$, it is indicative of a misspecified model.

Meng (1994) and Gelman et al. (1996) extend the basic PPC to use a "realised discrepancy" diagnostic function $s(y, \vartheta)$, which depends on both the data and the model parameters, and quantifies the discrepancy between them. This could, for instance, be the negative log likelihood $-\sum_{d=1}^D \log p(y_d|\vartheta)$ or the chi-squared diagnostic $\frac{1}{D}\sum_{d=1}^D \frac{(y_d - \mathbb{E}(y_d|\vartheta))^2}{\mathbb{V}(y_d|\vartheta)}$. Quoting Gelman et al. (1996, Section 2.2), "the focus here is to measure discrepancies between a model and the data, not to test whether a model is true". The idea is conceptually the same as for $s(y)$: we locate $s(y^{\text{obs}}, \vartheta)$ within the reference distribution of $s(y^{\text{rep}}, \vartheta)$ and check whether it is an outlier. Note that, in this paper, we take the diagnostic to be the (positive) log likelihood, so $s(y, \vartheta)$ is interpreted as the *agreement*, rather than the discrepancy, between y and ϑ.

A common way to check whether $s(y^{\text{obs}}, \vartheta)$ is an outlier within the distribution of $s(y^{\text{rep}}, \vartheta)$ is to use a hypothesis test with a 1-sided 'p-value'

$$p_{\text{ppc}} = \mathbb{P}\left(s(y^{\text{rep}}, \vartheta) < s(y^{\text{obs}}, \vartheta) \mid y^{\text{obs}}\right), \tag{6}$$

but other measures of surprise (Bayarri and Morales, 2003) or visual approaches (Gelman et al., 1996; Gelman, 2004) may also be used. The direction of the inequality in (6) may be reversed depending on the specific diagnostic $s(y, \vartheta)$ chosen in practice, or indeed a 2-sided 'p-value' may be more appropriate. We write 'p-value' within quotes since it is well known

(Meng, 1994; Bayarri and Berger, 2000; Robins et al., 2000) that this hypothesis test is in general not calibrated; that is, under the null hypothesis of a well-specified model, the distribution of p_{ppc} is not uniform over $[0, 1]$. The miscalibration results from the double use of the data: we use y^{obs} to generate replicated data in (5), and the same data to compute p_{ppc} in (6).

A number of alternatives and solutions have been proposed in the literature to address the miscalibration. These include post-hoc corrections (Robins et al., 2000; Hjort et al., 2006), partial posterior predictive and conditional predictive p-values (Bayarri and Berger, 2000), split predictive checks (Li and Huggins, 2022), and holdout predictive checks (Moran et al., 2023) among others. Moran et al. (2023, Sections 1.1 & 4) give a brief summary and criticism of all these methods. Nevertheless, despite being miscalibrated, it is still the case that a value of p_{ppc} in (6) too close to 0 or 1 would potentially indicate misfit. The only consequence of the miscalibration is that, if we reject the null hypothesis of a well-specified model at significance level α, the probability of a type I error will usually be smaller than α (Meng, 1994).

It is, of course, not possible to compute (6) directly, since the parameter ϑ is unknown. However, it is easy to estimate (6) using Monte Carlo methods — a property that makes PPC a natural and computationally efficient diagnostic tool to use when the posterior is sampled with MCMC. Given posterior samples $\vartheta^n \sim \pi(\cdot|y^{\text{obs}}), n = 1, \ldots, N$, we first simulate a replicated dataset $y^{\text{rep},n} \sim p(\cdot|\vartheta^n)$ for each n using the posited model. Then, for each pair $(y^{\text{rep},n}, \vartheta^n)$, we compute $s(y^{\text{rep},n}, \vartheta^n)$ to get the reference empirical distribution of the diagnostic.

For the observed diagnostic, there are a few options. One option is to use the same posterior samples ϑ^n to compute $s(y^{\text{obs}}, \vartheta^n)$ for each n. This gives us a diagnostic pair (replicated and observed) for each posterior sample, and we can compute

$$p_{\text{ppc}} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(s(y^{\text{rep},n}, \vartheta^n) < s(y^{\text{obs}}, \vartheta^n)\right). \tag{7}$$

This option is favoured by Gelman et al. (1996), who also recommend making a scatter plot of $s(y^{\text{obs}}, \vartheta^n)$ against $s(y^{\text{rep},n}, \vartheta^n)$ for visual inspection. A second option is to use an average $\bar{s}(y^{\text{obs}})$, such as the mean $\frac{1}{N}\sum_{n=1}^{N} s(y^{\text{obs}}, \vartheta^n)$ or $\text{median}_n s(y^{\text{obs}}, \vartheta^n)$ over $n \in \{1, \ldots, N\}$, to approximate $s(y^{\text{obs}}, \vartheta)$, and then compute

$$p_{\text{ppc}} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(s(y^{\text{rep},n}, \vartheta^n) < \bar{s}(y^{\text{obs}})\right). \tag{8}$$

A third option is to use a summary statistic such as the posterior mean $\bar{\vartheta} = \frac{1}{N}\sum_{n=1}^{N}\vartheta^n$ to approximate ϑ, and use this to compute

$$p_{\text{ppc}} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(s(y^{\text{rep},n},\vartheta^n) < s(y^{\text{obs}},\bar{\vartheta})\right). \tag{9}$$

We will use the third option in our experiments for the reasons given below.

If the posterior distribution is concentrated on the pseudo-true (3) then the three p-values above will be similar. However, in practice, the regularity conditions for the Bernstein-von Mises theorem do not always hold, and in any case the convergence to (3) may be extremely slow. Thus, with finite data, the posterior samples $\vartheta^1,\ldots,\vartheta^N$ could be far from the pseudo-true $\vartheta^\dagger$, especially in high dimensions. This is indeed the case for all real-data examples studied in this paper. Even if the model is well specified, the diffuseness of the posterior results in the agreement $s(y^{\text{obs}},\vartheta^n)$ being deflated by the parameter variance, and typically much lower than $s(y^{\text{obs}},\vartheta^*)$. We wish to remove this parameter variance.

If the model is correct then the agreement between the observed data and the true parameter should be about the same as the agreement between simulated data and the parameter at which it was simulated, so $s(y^{\text{obs}},\vartheta^*)$ and $s(y^{\text{rep},n},\vartheta^n)$ should have similar magnitude. We assume that the posterior mean $\bar{\vartheta}$ is a reasonable estimate for the true parameter ϑ^*, and test a null hypothesis of a well-specified model using the p-value in (9) rather than (7) or (8). This is quite a strong assumption in our setting. After all, we are interested in correcting for misspecification, and this may bias the posterior mean. However, as we shall see in Section 6, our approach for correcting misspecification (set out in Section 3 below) seems quite effective, which suggests that misspecification is causing under-dispersion rather than biasing the mean.

3 Setting λ using PPC

We now describe our proposed method for setting the learning rate λ using PPC, which bears strong resemblance to the approach taken by Chakraborty et al. (2023, Section 4.2) to set the influence parameter in semi-modular inference (Carmona and Nicholls 2020, a variant of generalised Bayes). We start with observed data $y^{\text{obs}} = (y_1^{\text{obs}},\ldots,y_D^{\text{obs}})$, parametric model $y^{\text{obs}} \sim p(\cdot|\vartheta)$, prior $\pi(\vartheta)$, and candidate learning rates $\lambda_r, r = 1,\ldots,R$, including $\lambda_1 = 1$. We also choose a diagnostic function $s(y,\vartheta)$ and a significance level α appropriate for the application. For each $\lambda \in \{\lambda_1,\ldots,\lambda_R\}$, we target the generalised Bayes' posterior

$$\pi_\lambda(\vartheta|y^{\text{obs}}) \propto \pi(\vartheta)p(y^{\text{obs}}|\vartheta)^\lambda \tag{10}$$

Algorithm 1 Setting the generalised Bayes' learning rate using Posterior Predictive Checks

for observed data $y^{\text{obs}} = (y_1^{\text{obs}}, \dots, y_D^{\text{obs}})$, specify model $y^{\text{obs}} \sim p(\cdot|\vartheta)$ and prior $\pi(\vartheta)$
specify candidate learning rates $\lambda_1, \dots, \lambda_R$, with $\lambda_1 = 1$
choose diagnostic function $s(y, \vartheta)$ and significance level α
for learning rate $\lambda \in \{\lambda_1, \dots, \lambda_R\}$ **do**
    run MCMC targeting $\pi_\lambda(\vartheta|y^{\text{obs}}) \propto \pi(\vartheta)p(y^{\text{obs}}|\vartheta)^\lambda$
    obtain posterior samples $\vartheta^{\lambda,n} \sim \pi_\lambda(\cdot|y^{\text{obs}})$, $n = 1, \dots, N$
    for $n \in \{1, \dots, N\}$ **do**
        simulate replicated dataset $y^{\text{rep},\lambda,n} \sim p(\cdot|\vartheta^{\lambda,n})$ and evaluate diagnostic $s(y^{\text{rep},\lambda,n}, \vartheta^{\lambda,n})$
    end for
    compute posterior mean $\bar{\vartheta}^\lambda = \frac{1}{N}\sum_{n=1}^N \vartheta^{\lambda,n}$ and evaluate observed diagnostic $s(y^{\text{obs}}, \bar{\vartheta}^\lambda)$
    compute p-value $p^\lambda_{\text{ppc}} = \frac{1}{N}\sum_{n=1}^N \mathbb{I}\left(s(y^{\text{rep},\lambda,n}, \vartheta^{\lambda,n}) < s(y^{\text{obs}}, \bar{\vartheta}^\lambda)\right)$
end for
set learning rate $\lambda^\dagger = \arg\min_{\lambda\in\{\lambda_1,\dots,\lambda_R\}}\{p^\lambda_{\text{ppc}} : p^\lambda_{\text{ppc}} > \alpha\}$

to obtain posterior samples $\vartheta^{\lambda,n} \sim \pi_\lambda(\cdot|y^{\text{obs}})$, $n = 1, \dots, N$. Using these, we generate replicated data $y^{\text{rep},\lambda,n} \sim p(\cdot|\vartheta^{\lambda,n})$ to perform a PPC and compute the p-value as in (9), denoted

$$p^\lambda_{\text{ppc}} = \frac{1}{N}\sum_{n=1}^N \mathbb{I}\left(s(y^{\text{rep},\lambda,n}, \vartheta^{\lambda,n}) < s(y^{\text{obs}}, \bar{\vartheta}^\lambda)\right), \tag{11}$$

where $\bar{\vartheta}^\lambda = \frac{1}{N}\sum_{n=1}^N \vartheta^{\lambda,n}$. Then, we select the rate minimising (11) subject to $p^\lambda_{\text{ppc}} > \alpha$, that is

$$\lambda^\dagger = \arg \min_{\lambda\in\{\lambda_1,\dots,\lambda_R\}}\{p^\lambda_{\text{ppc}} : p^\lambda_{\text{ppc}} > \alpha\}. \tag{12}$$

The method is summarised in Algorithm 1.

The direction of the inequality in (11) assumes that we have chosen a diagnostic $s(y, \vartheta)$ where a large value indicates greater agreement between the data and parameters, such as the log likelihood $\log p(y|\vartheta)$. If the opposite is true, we could simply take the negative function as our diagnostic. This convention helps us to think in log-likelihood terms, which is the chosen diagnostic in our application, rather than discrepancy terms (as is usual with PPC).

Intuitively, our method works on the premise that, at $\lambda = 1$, we are overfitting the model; so the agreement is high and $s(y^{\text{obs}}, \bar{\vartheta}^\lambda)$ must lie in the right tail of the reference distribution of $s(y^{\text{rep},\lambda}, \vartheta^\lambda)$. Then, as we reduce λ, the generalised Bayes' posterior (10) tracks the data less closely, which reduces the agreement and moves $s(y^{\text{obs}}, \bar{\vartheta}^\lambda)$ to the left relative to the reference distribution. Reducing λ reduces the amount of information coming from the data as it shifts the weight in (10) from $p(y^{\text{obs}}|\vartheta)$ to $\pi(\vartheta)$. This makes sense when the data is actually

correlated but the fitted model assumes independence, since then the effective sample size of the data is a lot smaller than its nominal sample size, and removing information from the data is a good thing. However, this makes the fit worse. Our method implicitly assumes that this tradeoff is desirable up to the point that the agreement $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ is representative of a well-specified model. We therefore stop at the point beyond which the PPC would reject the model as misspecified.

The significance level α in a hypothesis test is always subjective, but 0.05 is usually a popular choice. However, as noted in Section 2.3, the p^{λ}_{ppc} in (11) is not a calibrated p-value. It is, in fact, biased towards 0.5 (Robins et al., 2000), and the true probability of a Type I error is typically less than α (Meng, 1994). Therefore, as a rough guide, if we wish to target a Type I error probability of 0.05, we should set α to a value higher than 0.05. In our exploratory analysis, we find that $\alpha = 0.1$ serves us well.

For the candidate values $\lambda_r, r = 1, \ldots, R$, we could set these to $\lambda_r = 2^{-(r-1)\kappa}$ for some κ as in Grünwald and van Ommen (2017). Alternatively, we could set these at suitably small fixed-size decrements from 1. We could also proceed heuristically, and tune λ to zero-in on $p^{\lambda}_{\text{ppc}} = \alpha$, rather than doing a grid search. The exact choice is not that important, and depends on the application as well as computational resources available. In our exploratory analysis, we use decrements of 0.1, which is sufficient for our purpose.

4 Model and inference

In this section, we describe the model and inferential problem that motivated this work. The recent EDiSC model (Zafar and Nicholls, 2024a), like its predecessor models DiSC (Zafar and Nicholls, 2022), GASC (Perrone et al., 2019) and SCAN (Frermann and Lapata, 2016), is used to infer and analyse the evolving meanings or senses of a given target word in an unsupervised setting. It is a bag-of-words model in which grammar and syntax are ignored, and is closely related to topic models (Blei and Lafferty, 2006; Dieng et al., 2019, 2020). The model is fitted to a set of text snippets centred on the target word. It exploits the idea that context words inform the sense of the target word in any given snippet. An example is the word "bank" with two distinct senses of riverbank and financial institution. Context words like "water" or "stream" would indicate the former sense, whereas context words like "money" or "finance" would indicate the latter. Another example for the word "bug" is given in Table 1. We can thus posit a generative model where context words are sampled conditional on the target-word sense. If the data spans multiple time periods, a time dimension in the model captures the diachronic sense change.

Table 1: Example text snippets for target word "bug" showing its four different senses. Context words are lemmatised, and stopwords, infrequent words and punctuation are dropped, to get the data used in model fitting.

insect	... insect repellent on a winter trip when there are no bugs around. Your first-aid kit should reflect your personal needs as ...
micro-organism	... These intruders are what cause the fever, for the TB bugs are not virulent enough to cause high temperatures. The effect ...
software glitch	... bug the Quality Assurance people find and $20 for each bug the programmers fix. These are the same programmers who create ...
tapping device	... too much information has been collected through secret informants, wiretaps, bugs, surreptitious mail opening and break-ins, the Church Report had warned ...

For a detailed exposition of the model and notation, we refer the reader to Zafar and Nicholls (2024a, Sections 2 & 4). Here, we summarise only the elements required for the current paper. The data W for a target word consists of D snippets, and spans T discrete contiguous time periods and G genres. Snippets are a fixed window of L lemmatised context words around the target word; but stopwords and very low frequency words are filtered out, leaving some context positions empty. Each snippet $W_d, d = 1, \ldots, D$, belongs to time period τ_d and genre γ_d. In the generative model, for each W_d, we first sample its sense z_d, and then sample context words given the sense. Sense z_d is sampled from a multinomial sense-prevalence distribution $\tilde{\phi}^{\gamma_d,\tau_d}$ over K senses (indexed by genre and time) so that $p(z_d = k|\tilde{\phi}^{\gamma_d,\tau_d}) = \tilde{\phi}_k^{\gamma_d,\tau_d}$ for all senses $k \in \{1, \ldots, K\}$. Given z_d, at each position i in W_d, context words $w_{d,i}$ are sampled independently from a multinomial distribution $\tilde{\psi}^{z_d,\tau_d}$ over the V-sized lemmatised vocabulary (indexed by sense and time) so that $p(w_{d,i} = v|z_d, \tilde{\psi}^{z_d,\tau_d}) = \tilde{\psi}_v^{z_d,\tau_d}$ for all words $v \in \{1, \ldots, V\}$. The inferential task is to learn $\tilde{\phi}$ and $\tilde{\psi}$ given W.

We use the notation (z_d, γ_d, τ_d) for snippet-specific sense-genre-time triples, and (k, g, t) for their generic equivalents. We define $\tilde{\phi}^{g,t} = \text{softmax}(\phi^{g,t})$ and $\tilde{\psi}^{k,t} = \text{softmax}(\psi^{k,t})$, and place priors on the real arrays ϕ and ψ as described in Zafar and Nicholls (2024a, Sections 4). In the notation of Section 2.1, we thus have $\vartheta = (\phi, \psi)$ and $y = (W, z)$. Each data 'point' $y_d = (W_d, z_d)$ is a composite made up of an independent sample z_d and a set of conditionally independent samples $w \in W_d$, and has deterministic covariates $x_d = (\gamma_d, \tau_d)$. Snippets W are observed data, whereas sense assignments $z = (z_1, \ldots, z_D)$ are missing data. Taking a Bayesian approach, z is treated as a latent variable and marginalised out analytically to

give the likelihood

$$p(W|\phi,\psi) = \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d = k|\phi)p(W_d|z_d = k,\psi) = \prod_{d=1}^{D}\sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d,\tau_d} \prod_{w\in W_d} \tilde{\psi}_w^{k,\tau_d}. \tag{13}$$

It is easy to see why this model is inherently misspecified. In any natural language, syntax and grammar are important, and context words are neither independent nor random. Target-word sense depends on more than just the time and genre in reality. Snippets are not actually generated in isolation, but are found within a larger document or corpus. Lemmatising the vocabulary and filtering out stopwords and very low frequency words is a gross simplification of actual usage. Furthermore, each word in the vocabulary is represented by an M-dimensional embedding vector (learnt independently of the snippet data), and there would inevitably be further misspecifications in the embedding model.

Notwithstanding the above, in most real cases, there is at least some dependence between any target-word sense and the context words used for that sense. Also, 'physical' parameters retain their meaning under misspecification: there must always be some actual sense-prevalence frequency. It follows that the inferred $\tilde{\psi}$ and $\tilde{\phi}$ are well defined despite the misspecification. Our goal, then, is to learn the parameters for which the posited model (13) is closest to the latent 'true' DGP in the sense of (3).

There are, in the fitted model, many assumptions of independence, whereas in fact there must be very complex correlations. Fitting a model that assumes independence to correlated data often gives credible sets that are too small. However, a realistic parametric model for these complex correlations seems infeasible, and non-parametric approaches are unlikely to help due to the sparsity of data. Faced with this, we turn to GBI: instead of elaborating the model, we modify the inference framework and target the Generalised Bayes' posterior

$$\pi_\lambda(\phi,\psi|W) \propto \pi(\phi,\psi)p(W|\phi,\psi)^\lambda. \tag{14}$$

The posterior (14) suffers from ridge structures and multimodality, making simple variational methods unreliable (Zafar and Nicholls, 2024a, Section 4.4). The dimension is too high for quadrature, so we are restricted to MCMC methods. For efficiency reasons, we use the No-U-Turn sampler (NUTS, Hoffman and Gelman 2014) from the Stan software (Stan Development Team, 2023b,a) for all MCMC used in this paper.

5 Data and evaluation framework

Here we describe the data used to develop and test our method, as well as the criteria used for measuring performance.

5.1 Development and test data

Zafar and Nicholls (2024a, Section 2) describe the four datasets used in their experiments. Three of these — the main focus of their paper — come from the Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018), with expert sense-annotation provided by Vatri et al. (2019). These are for the target words "kosmos" (with decoration, order and world senses), "mus" (with mouse, muscle and mussel senses) and "harmonia" (with abstract, concrete and musical senses). We use these three datasets for exploratory analysis and to develop our method for setting the learning rate λ given in Section 3. Since we explored a number of λ-selection methods (which we do not report) on these data, it would be misleading to evaluate the performance of our chosen method on just these data. We therefore test our method on the additional datasets described below.

The fourth dataset used by Zafar and Nicholls (2024a) is for the target word "bank" from the Corpus of Historical American English (COHA, Davies 2010, 2012). The authors manually annotated 3,525 snippets for "bank" with the sense of riverbank or financial institution. We split this data into five roughly equal-sized subsets to increase the datasets available for testing. For the "bank" and ancient Greek data, we use the same modelling choices (number of senses K and embedding dimension M) as Zafar and Nicholls (2024a, Section 5.2).

In addition to these datasets, we extract five new datasets from the Clean Corpus of Historical American English (CCOHA, Alatrash et al. 2020), an updated version of COHA with fewer typographical errors, and manually annotate the snippets with the correct sense. These are for the target words "chair" (chairman or furniture), "apple" (Apple Inc. company or fruit), "gay" (bright/showy, forward/bold, merry/lively/social or homosexual), "mouse" (computer pointing device or rodent) and "bug" (insect, microorganism, software glitch or tapping device). The "chair", "gay" and "bug" data are obtained from the corpus using stratified random sampling over genres and time periods, whereas the "apple" and "mouse" samples are selected to ensure both senses are represented adequately. Embeddings of dimension $M = 200$ are learnt for all context words in the corpus with minimum frequency 10 using GloVe (Pennington et al., 2014). Compared to the earlier "bank" data, we retain a larger vocabulary in the filtering process so as not to lose important semantic information, but consequently also retain more noise.

The datasets are summarised in Table 2. Note that the manual sense annotations are used only in the evaluation and not in model fitting. The notation K' refers to the number of true target-word senses, whereas the models are fitted using K senses, which may be different to K'. Model selection with respect to the choice of K is done as per Zafar and

Table 2: Data summary

Corpus	Target word	Snippets (D)	Vocab (V)	Length (L)	True senses (K')	Model senses (K)	Genres (G)	detail	Time periods (T)	detail
Diorisis	kosmos	1,469	2,904	14	3	4	2	narrative, non-narr	9	700 BC to 200 AD, centuries
	mus	214	899	14	3	3	2	technical and	9	500 BC to 400 AD, centuries
	harmonia	653	1,607	14	3	4	2	non-technical	12	800 BC to 400 AD, centuries
COHA	bank split 1	704	736	14	2	2	1		10	1810–2010, 20yr intervals
	bank split 2	708	717	14	2	2	1	merged news,	10	1810–2010, 20yr intervals
	bank split 3	703	728	14	2	2	1	magazine	10	1810–2010, 20yr intervals
	bank split 4	704	742	14	2	2	1	fiction, non-fic	10	1810–2010, 20yr intervals
	bank split 5	706	735	14	2	2	1		10	1810–2010, 20yr intervals
CCOHA	chair	745	3,180	20	2	2	4	news, magazine,	10	1820–2020, 20yr intervals
	apple	1,154	3,737	20	2	2	4	fic, non-fic/acad	6	1960–2020, 10yr intervals
	gay	650	3,071	20	2	4	1	merged news,	5	1920–2020, 20yr intervals
	mouse	584	2,439	20	2	3	1	magazine,	4	1940–2020, 20yr intervals
	bug	522	2,475	20	4	4	1	non-fic/acad	8	1980–2020, 5yr intervals

Nicholls (2024a, Section 5.2). For target word "gay", during the manual annotation, it was often quite difficult for us to identify the correct sense out of the three non-homosexual senses owing to the very subtle differences between them; hence, for evaluation purposes, we combined the three senses into a single non-homosexual sense.

5.2 Assessing model performance

Zafar and Nicholls (2024a, Section 5) assess model performance on two fronts: predictive accuracy and true-model recovery. The true model is obviously unknown, but a proxy 'true' model is used instead. Given true sense assignments $o = (o_1, \ldots, o_D)$, the standard Bayes' posterior conditioned on the truth, $\tilde{\phi}|(z = o)$, is used as a well-calibrated independent estimate for the unknown true sense prevalence $\tilde{\Phi}$; and the model posterior $\tilde{\phi}|W$ is compared against it. Then, the overlap between $\tilde{\phi}|(z = o)$ and $\tilde{\phi}|W$ quantifies performance on true-model recovery. However, using this comparison in a GBI setup would not be appropriate since, arguably, we could have a generalised Bayes' posterior $\tilde{\phi}^{\lambda}|(z = o)$, for some $\lambda < 1$, that is in fact closer to $\tilde{\Phi}$ than $\tilde{\phi}|(z = o)$; and finding a suitable λ value is the very problem we are addressing. On the other hand, assessing predictive accuracy only requires the objective truth $o = (o_1, \ldots, o_D)$, and is therefore a more suitable measure of performance.

Following Zafar and Nicholls (2024a, Section 5.1), we assess predictive accuracy under each learning rate λ using the Brier score

$$\mathrm{BS}_{\lambda} = \frac{1}{D} \sum_{d=1}^{D} \sum_{k=1'}^{K'} \left(\hat{p}_{\lambda}(z_d = k) - \mathbb{I}(o_d = k)\right)^2, \tag{15}$$

a proper scoring rule for multi-category probabilistic predictions $\hat{p}_{\lambda}(z_d = k)$, ranging from 0 (best) to 2 (worst). Here, $\hat{p}_{\lambda}(z_d = k)$ is the estimated value of $\mathbb{E}_{(\phi,\psi)^{\lambda}|W}\big(p(z_d = k|W_d, \phi, \psi)\big)$, computed on the MCMC output $(\phi, \psi)^{\lambda}|W$ targeting the generalised Bayes' posterior (14)

with learning rate λ. The estimate is obtained by normalising

$$p(z_d = k|W_d, \phi, \psi) \propto \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{w \in W_d} \tilde{\psi}_w^{k, \tau_d} \tag{16}$$

over senses $k \in \{1, \ldots, K\}$ for each posterior sample, and then averaging the normalised (16) across samples. Recall that we have K' true senses, whereas we run the models using K senses (with $K \geq K'$), so modelled senses may be grouped together to map them onto the true senses.

Within the context of this exploratory work, we would ideally like to select the 'optimal' learning rate $\lambda^* = \arg\min_\lambda \mathrm{BS}_\lambda$. However, unless we use the true sense assignments in some way, this can only be coincidental, since GBI attempts only to correct model misspecification in the sense implicit in Section 3 rather than specifically find the best predictive model. We expect it to do reasonably well, as our criterion for selecting λ should improve posterior calibration, and the Brier score is sensitive to calibration (among other factors). Therefore, as long as the $\lambda^\dagger$ selected as per (12) gives a Brier score $\mathrm{BS}_{\lambda^\dagger} \leq \mathrm{BS}_1$ and reasonably close to the optimal BS_{λ^*}, that would indicate success.

6 Experiments

We describe our experiments on development data, followed by the results on test data.

6.1 Method development

Here we give the results of experiments performed on the three ancient Greek datasets to develop our λ-selection method. We start by exploring whether the model is misspecified for these data using PPC. We fit the model using MCMC, and obtain posterior samples $(\phi, \psi)^{\lambda,n}, n = 1, \ldots, N$, given observed data W^{obs}. To generate replicated data $W^{\mathrm{rep},\lambda,n}$, for each snippet $d \in \{1, \ldots, D\}$, we first simulate the sense assignment $z_d^{\mathrm{rep},\lambda,n} \sim p\left(\cdot|W_d^{\mathrm{obs}}, \phi^{\lambda,n}, \psi^{\lambda,n}\right)$ with $p(\cdot|W_d, \phi, \psi)$ as in (16). Then, we simulate context words $w_{d,i}^{\mathrm{rep},\lambda,n}|(z_d^{\mathrm{rep},\lambda,n} = k) \sim \tilde{\psi}^{\lambda,n}_{k,\tau_d}$ independently for each position i in $W_d^{\mathrm{rep},\lambda,n}$. For the diagnostic function $s(W, \phi, \psi)$, we use the log likelihood $\log p(W|\phi, \psi)$, with $p(W|\phi, \psi)$ as in (13), which gives the best results out of all the options explored. (The other options included the chi-squared statistic based on observed and expected counts, and various divergences between the parameters and the empirical distribution.)

Figure 1 shows the PPCs in the $\lambda = 1$ case. The solid black line shows the distribution of $\log p\left(W^{\mathrm{obs}}|(\phi, \psi)^\lambda\right)$ computed on observed data at posterior samples; the red line shows the reference distribution of $\log p\left(W^{\mathrm{rep},\lambda}|(\phi, \psi)^\lambda\right)$ computed on replicated data at posterior

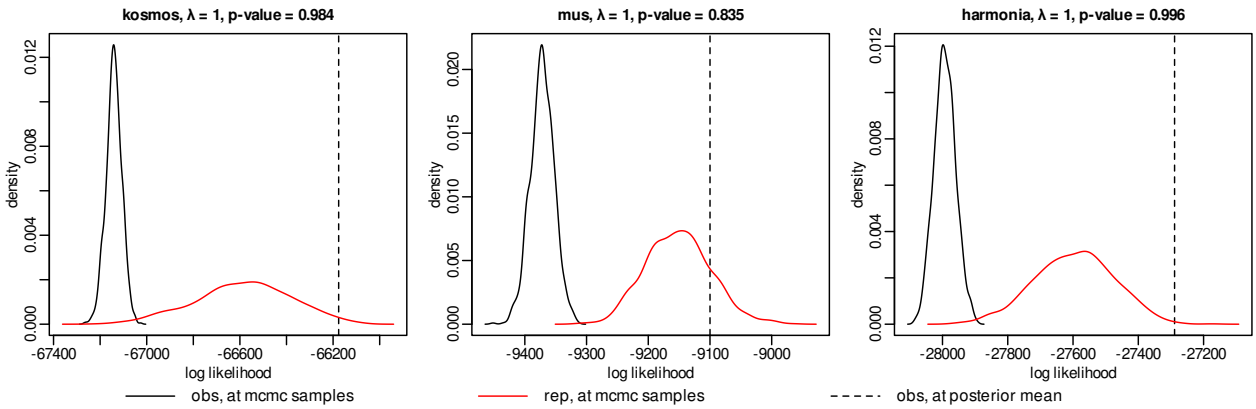


Figure 1: Posterior predictive checks on development data for $\lambda = 1$, showing the distribution of the log-likelihood diagnostic

samples; and the dashed black line shows the point-value $\log p\left(W^{\text{obs}}|(\bar{\phi},\bar{\psi})^{\lambda}\right)$ computed on observed data at the posterior means $(\bar{\phi},\bar{\psi})^{\lambda}$. (Note that we actually take posterior means for the softmax-transformed probabilities $(\tilde{\phi},\tilde{\psi})^{\lambda}$, but write $(\bar{\phi},\bar{\psi})^{\lambda}$ to keep the notation tidy.) For all three datasets, the observed $\log p\left(W^{\text{obs}}|(\phi,\psi)^{\lambda}\right)$ lies well to the left relative to the reference distribution, suggesting that the individual posterior samples do not describe the observed data very well. On the other hand, $\log p\left(W^{\text{obs}}|(\bar{\phi},\bar{\psi})^{\lambda}\right)$ lies to the right in all cases, suggesting that the posterior means tend to overfit the observed data, at least for "kosmos" and "harmonia"; so there is potential misspecification, possibly due to the unmodelled latent correlations mentioned in Section 4. As discussed at the end of Section 2.3, we use the test statistic $\log p\left(W^{\text{obs}}|(\bar{\phi},\bar{\psi})^{\lambda}\right)$ based on posterior means, and do not refer to $\log p\left(W^{\text{obs}}|(\phi,\psi)^{\lambda}\right)$ henceforth in this section.

We now examine the behaviour of p-values p^{λ}_{ppc} and Brier scores BS_{λ} as we reduce the learning rate λ. This is shown in Figure 2 for $\lambda = 1, 0.9, \ldots, 0.1$. Firstly, we note that p^{λ}_{ppc} (solid red line) always decreases with λ until it hits zero. This corresponds to the observed diagnostic $\log p\left(W^{\text{obs}}|(\bar{\phi},\bar{\psi})^{\lambda}\right)$ always shifting to the left relative to the reference distribution of $\log p\left(W^{\text{rep},\lambda}|(\phi,\psi)^{\lambda}\right)$, as is shown explicitly in Figure 3 for two different λ values for each dataset. This is as we would expect, since decreasing λ corresponds to a smaller weight being given to the likelihood in the generalised Bayes' posterior (14).

Secondly, we note that BS_{λ} (solid black line in Figure 2) decreases (i.e. improves) with λ up to the point that it hits the optimal BS_{λ^*} (blue circle). If we keep reducing λ beyond this point, either BS_{λ} starts to increase again (as for "mus"), or the inferred parameters stop being meaningful (which we will refer to as 'collapse' for the reasons given in the Appendix). The latter is guaranteed to happen in any case for a sufficiently small λ when

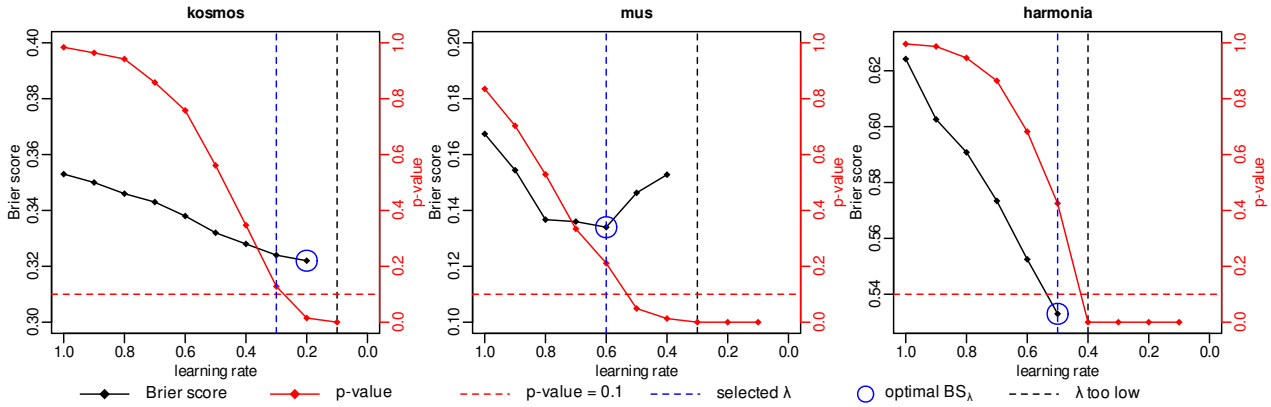


Figure 2: Brier scores and p-values for varying learning rates on development data

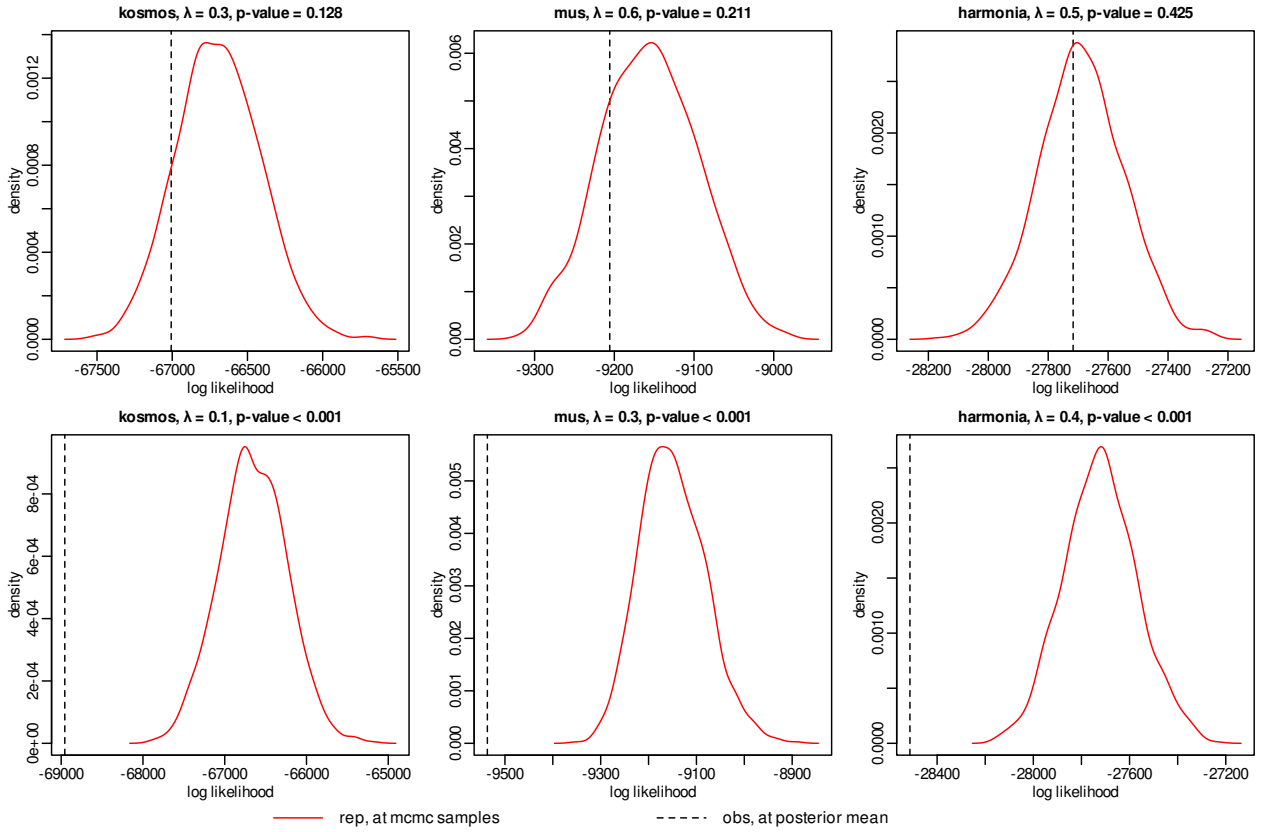


Figure 3: Posterior predictive checks on development data for $\lambda^{\dagger} = \arg\min_{\lambda}\{p^{\lambda}_{\text{ppc}} : p^{\lambda}_{\text{ppc}} > 0.1\}$ (top row) and the λ value at the point of collapse (bottom row)

the prior dominates the tempered likelihood in (14). Brier score cannot be computed after collapse, since there is no mapping between modelled senses and true senses, so the BS_{λ} line ends just before this point.

Finally, we note that as soon as the parameters collapse (dashed black line in Figure 2), p^{λ}_{ppc}

immediately drops to zero. PPCs at the point of collapse are shown in the bottom row of Figure 3, clearly showing the observed diagnostic $\log p\left(W^{\text{obs}}|(\bar{\phi},\bar{\psi})^{\lambda}\right)$ at these λ values to be an outlier within the reference distribution of $\log p\left(W^{\text{rep},\lambda}|(\phi,\psi)^{\lambda}\right)$. The converse is not true: as $p^{\lambda}_{\text{ppc}} \to 0$, collapse may not yet have happened (as for "kosmos" and "mus"), and BS_{λ} may still be improving (as for "kosmos"). However, it appears that a very low value of p^{λ}_{ppc} does correlate, albeit loosely, with an under-optimised posterior in some sense. This suggests that a cutoff point (i.e. significance level) α for p^{λ}_{ppc} could be a suitable criterion for λ selection. As discussed in Section 3, a conservative approach would be appropriate due to the miscalibrated p-values, so we set the cutoff point reasonably high at $\alpha = 0.1$ (dashed red line). Out of the candidate λ values, we then select the learning rate $\lambda^{\dagger}$ as per (12) (dashed blue line). For our development data, this $\lambda^{\dagger}$ gives the optimal BS_{λ^*} for "mus" and "harmonia", and very close to the optimal BS_{λ^*} for "kosmos". By design, the misspecification at these $\lambda^{\dagger}$ values is not significant under PPC, as shown in the top row of Figure 3.

6.2 Results on test data

We now give the results of applying our λ-selection method on the 10 test datasets in Table 3. For the five split "bank" datasets, our method consistently selects $\lambda^{\dagger} = 0.4$. The $\text{BS}_{\lambda^{\dagger}}$ for the selected rate is optimal, or almost optimal, for splits 2–4. For splits 1 and 5, $\text{BS}_{\lambda^{\dagger}}$ is still closer to the optimal BS_{λ^*} than BS_1, which is a successful outcome as per our criteria set out at the end of Section 5.2. Furthermore, the fact that our method selects the same $\lambda^{\dagger}$ for all five splits makes sense: the split datasets are a random partition of a larger dataset, and we expect misspecification to be similar in nature and degree across partitions. The learning-rate estimates are robust to other details of the data.

Out of the other five test datasets, for all except the "chair" data, our method again selects a rate $\lambda^{\dagger}$ that is either optimal or almost optimal as measured by Brier score. Interestingly, p^{λ}_{ppc} for the "apple" and "mouse" data drops more slowly with decreasing λ compared to the other datasets, and our method selects relatively lower $\lambda^{\dagger}$ values for these data. As mentioned in Section 5.1, these data samples are subsets of snippets selected from the corpus to ensure adequate representation of both target-word senses. The selection mechanism tends to select snippets with strong and clearly evidenced meanings, and as a consequence the model senses in the posterior are more sharply separated. Hence, λ needs to be reduced more to have enough of a 'softening' effect on the sense separation.

On the "chair" data, the $\text{BS}_{\lambda^{\dagger}}$ returned is an improvement on BS_1, though it is not as close to the optimal BS_{λ^*} as in the other examples. Recall (from Section 5.2) that we are using

Table 3: Brier scores BS_λ and p-values p^λ_{ppc} for candidate learning rates λ on test data. Optimal scores BS_{λ^*} are in blue. Scores $\mathrm{BS}_{\lambda^\dagger}$ for the learning rates $\lambda^\dagger$ selected using our method are boxed. Missing values for BS_λ indicate collapse. All values are accurate to 3 s.f.

	bank split 1		bank split 2		bank split 3		bank split 4		bank split 5	
λ	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}
1	0.135	0.863	0.115	0.832	0.106	0.833	0.144	0.882	0.189	0.833
0.6	0.128	0.428	0.111	0.393	0.103	0.412	0.139	0.459	0.178	0.374
0.5	0.125	0.256	0.110	0.233	0.103	0.258	0.138	0.266	0.173	0.213
0.4	0.124	0.114	0.109	0.119	0.103	0.126	0.138	0.138	0.168	0.105
0.3	0.121	0.045	0.108	0.030	0.104	0.036	0.137	0.048	0.162	0.030
0.2	0.118	0.009	0.109	0.009	0.106	0.007	0.137	0.009	0.158	0.007
0.1	0.122	<0.001	0.113	<0.001	0.113	0.001	0.141	<0.001	0.152	<0.001

	chair		apple		gay		mouse		bug	
λ	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}	BS_λ	p^λ_{ppc}
1	0.140	0.656	0.0510	0.981	0.343	0.776	0.0474	0.968	0.300	0.974
0.6	0.133	0.134	0.0496	0.899	0.299	0.217	0.0382	0.796	0.268	0.567
0.5	0.129	0.037	0.0488	0.794	0.291	0.111	0.0359	0.692	0.267	0.222
0.4	0.126	0.009	0.0484	0.652	0.282	0.021	0.0348	0.474	0.265	0.066
0.3	0.124	<0.001	0.0479	0.403	0.295	<0.001	0.0335	0.245		<0.001
0.2	0.120	<0.001	0.0467	0.208		<0.001	0.0337	0.077		<0.001
0.1	0.125	<0.001	0.0461	0.029		<0.001	0.0393	0.004		<0.001

predictive accuracy as a criterion for performance measurement only for the sake of objectivity, since we do not know the true DGP. However, the goals of predictive accuracy and true-model recovery are not always in sync (witness the contrasting consistency behaviour of AIC and BIC). GBI attempts to correct model misspecification, and is therefore more aligned with the latter than the former. It seems that the two goals are less synchronised for the "chair" data than the other datasets. In any case, as set out in the introduction, our method is not guaranteed to find the optimal model — we promote it as an intuitively well-founded and computationally efficient way to select λ, and one that works well in all of our test cases.

7 Discussion

In this exploratory work, we considered the problem of selecting an appropriate learning rate λ for GBI, within the specific context of the recent EDiSC model that is used to quantify changes in target-word senses over time. We argued that it makes intuitive sense to set the learning rate using PPC based on the log likelihood diagnostic, such that we select the lowest rate where a PPC is not rejected at the (nominal) 10% significance level. This approach is computationally efficient and can be readily implemented with MCMC sampling. We developed this approach using experiments on three datasets, and tested it

on 10 new datasets. We found the approach to work very well in all cases as measured by predictive accuracy quantified using Brier scores: the accuracy attained with a learning rate selected using our method is very close to, or exactly at, the optimal level. Some further insights into the mechanism by which tempering helps are given in the Appendix.

Our λ-selection method is adapted to our model and training data; and although it is effective in our setting, it is not universally applicable. The method seems well suited to cases where the posterior mean overfits the data, perhaps due to unmodelled latent correlations in the data. In these cases, for $\lambda = 1$ (i.e. standard Bayes), the agreement score $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ for observed data at the posterior mean is typically higher than the agreement scores $s(y^{\text{rep},\lambda,n}, \vartheta^{\lambda,n}), n = 1, \ldots, N$, for replicated data at posterior samples. As λ decreases, we remove information entering the analysis from the data, and the fitted agreement score $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ reduces towards the replicated agreement scores. This reflects the fact that the model is over-valuing the information in the data: the data contains latent unmodelled correlations, so its effective sample size is lower than its nominal sample size.

We ran additional tests of the method for simple Bayesian linear models, using simulated small datasets and/or low-dimensional parameters, and found that the method was not useful for selecting λ. In these tests, the property mentioned above was not present by construction. Therefore, the fitted agreement score $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ at $\lambda = 1$ was in a random location relative to the replicated agreement scores $s(y^{\text{rep},\lambda,n}, \vartheta^{\lambda,n}), n = 1, \ldots, N$. Taking $\lambda < 1$ only decreases the fitted agreement, potentially moving it further away from the replicated agreement scores. In future work, we would like to experiment further, using simulated data designed more accurately to mimic the misspecification seen in our setting, in order to provide other use cases.

The reasons for the behaviour described above are not obvious. However, it is clear that a reasonably large dataset is required to see this behaviour, and a high-dimensional parameter space seems warranted (which is a natural consequence of the data in our setting). The posterior should display a complex underlying correlation structure, so that individual posterior samples $\vartheta^{\lambda,n}, n = 1, \ldots, N$, do not describe the data well but the posterior *mean* $\bar{\vartheta}^{\lambda}$ does so. This corresponds, at $\lambda = 1$, to the agreement score $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ at the posterior mean being higher than the equivalent agreement scores $s(y^{\text{obs}}, \vartheta^{\lambda,n})$ at the posterior samples. As seen in Figure 1, this is the case in our real-data setting. However, in the additional tests mentioned above where our λ-selection method was not useful, $s(y^{\text{obs}}, \bar{\vartheta}^{\lambda})$ was generally representative of $s(y^{\text{obs}}, \vartheta^{\lambda,n})$ at the posterior samples.

Secondly, the likelihood equation (13) plays a role. The actual data, as discussed in Section 4, exists in pairs with a snippet W_d and a sense z_d for all $d = 1, \ldots, D$. However, the latent sense assignments z are not observed, so (13) is computed marginally over z. In our experience, GBI methods have most to offer when the model has a high-dimensional latent parameter that is misspecified in the model, which is certainly the case here. Each snippet brings with it one of these missing z_d, so they are numerous. The model plays an important role in weighting sense assignments, as any W_d itself may contain relatively little information about z_d. Hence, if the model is misspecified, methods treating misspecification have a role as well. On the other hand, if the data were strongly informative of the sense, this would not be the case.

Thirdly, the power λ in (14) goes outside the sum over possible sense assignments in (13), so the power is on the partially marginalised likelihood. In this sum, if the fit is good, the term for the correct sense assignment would be largest. Taking the power *outside* the sum has the effect of 'protecting' the largest term in the sum when reducing λ.

Whilst the above explanations give some intuition as to when our λ-selection approach might work, we have not yet defined the exact conditions required. However, the systematic nature of results seen in all the real-data examples studied in this paper, and the success of our method in all cases, suggests that this line of investigation is worth exploring further.

Implementation

The code and data used to produce the results reported in this paper are available from https://github.com/schyanzafar/GBI.

Appendix

Analysis of generalised Bayes' posterior

To gain some insight into the mechanism by which tempering the likelihood improves model performance in our setting, it helps to analyse what happens to the inferred parameters $(\tilde{\phi}, \tilde{\psi})^\lambda$ as we reduce λ. We show these analyses for "mus" only, but the results discussed here are also typical for "kosmos" and "harmonia".

As discussed in Zafar and Nicholls (2024a, Section 5.2), a natural way to examine the model output is to look at the context words with the highest probabilities under each model sense, marginally over time, using the posterior mean $\bar{\psi}^\lambda$. We show the top 10 context words under

Table 4: Top 10 context words under each model sense of "mus" for different learning rates. Words repeated across multiple senses are indicated in red.

Sense	Top 10 context words $\lambda = 1$									
1	λέγω	φημί	γῆ	γίγνομαι	πᾶς	εἶτα	νῦν	μῦς	γαλέη	πλῆθος
2	νεῦρον	ὀστέον	φλέψ	βραχίων	ἔχω	τένων	ἄρθρον	μῦς	πυρετός	μυελός
3	σωλήν	κτείς	ὄστρεον	χρύσεος	λεπάς	χήμη	κόγχη	ὄστρειον	πίννα	νεῦρον
Sense	Top 10 context words $\lambda = 0.6$									
1	λέγω	φημί	γῆ	γίγνομαι	πᾶς	πολύς	ἔχω	πλῆθος	εἶτα	νῦν
2	νεῦρον	ὀστέον	φλέψ	βραχίων	ἔχω	τένων	ἄρθρον	μῦς	πυρετός	κίνησις
3	σωλήν	κτείς	λεπάς	χήμη	ὄστρεον	πίννα	χρύσεος	ὄστρειον	κόγχη	μῦς
Sense	Top 10 context words $\lambda = 0.3$									
1	νεῦρον	ὀστέον	φλέψ	λέγω	φημί	μῦς	ἔχω	γίγνομαι	γῆ	πολύς
2	νεῦρον	ὀστέον	φλέψ	μῦς	βραχίων	ἔχω	λέγω	φημί	ἄρθρον	ὦμος
3	νεῦρον	μῦς	σωλήν	κτείς	φημί	λέγω	λεπάς	ὀστέον	βραχίων	πίννα

the model fits for $\lambda = 1$ (standard Bayes), $\lambda = 0.6$ (optimal) and $\lambda = 0.3$ (collapsed) in Table 4. With some knowledge of ancient Greek, or with the help of a dictionary (e.g. Wiktionary), or by comparing against expert annotation, the three model senses in the first two cases can be identified as mouse, muscle and mussel respectively. The output for $\lambda = 1$ and $\lambda = 0.6$ is quite similar, and the model senses retain their separate identities in both cases. However, for $\lambda = 0.3$, the model senses are no longer distinguishable from each other, as they are displaying very similar sets of top words; or, in other words, they have 'collapsed'. They do, nevertheless, still reflect the overall context-word probabilities (across all senses) learnt from the data. If we keep reducing λ, eventually all model senses look the same, and all context words take uniform probabilities as per the prior.

Sense prevalence $\tilde{\phi}$ is a lower-dimensional object reflecting the behaviour of higher-dimensional senses $\tilde{\psi}$ (where each sense k for time t is a distribution $\tilde{\psi}^{k,t}$ over context words $1, \ldots, V$) and is therefore easier to visualise. We argued in Section 5.2 that a comparison of model posterior $\tilde{\phi}^{\lambda}|W$ against the proxy 'true' model posterior $\tilde{\phi}|(z = o)$ should not be used to measure performance. However, this comparison does help us analyse the posterior behaviour, and is shown in Figure 4 for several learning rates. We see that, as we reduce λ from 1 to 0.6, the effect is a slight increase in variance without much change in location. This is the same effect that makes likelihood tempering useful in an MCMC convergence context (cf. Zafar and Nicholls 2024a, Appendix C). In this case, if $\tilde{\phi}^{1}|W$ is slightly misspecified in the sense of not achieving enough overlap with the proxy-true $\tilde{\phi}|(z = o)$, the increased variance allows $\tilde{\phi}^{0.6}|W$ to correct the misspecification by covering more of the posterior space and increasing the overlap. As a consequence, this correction improves overall predictive accuracy on average.

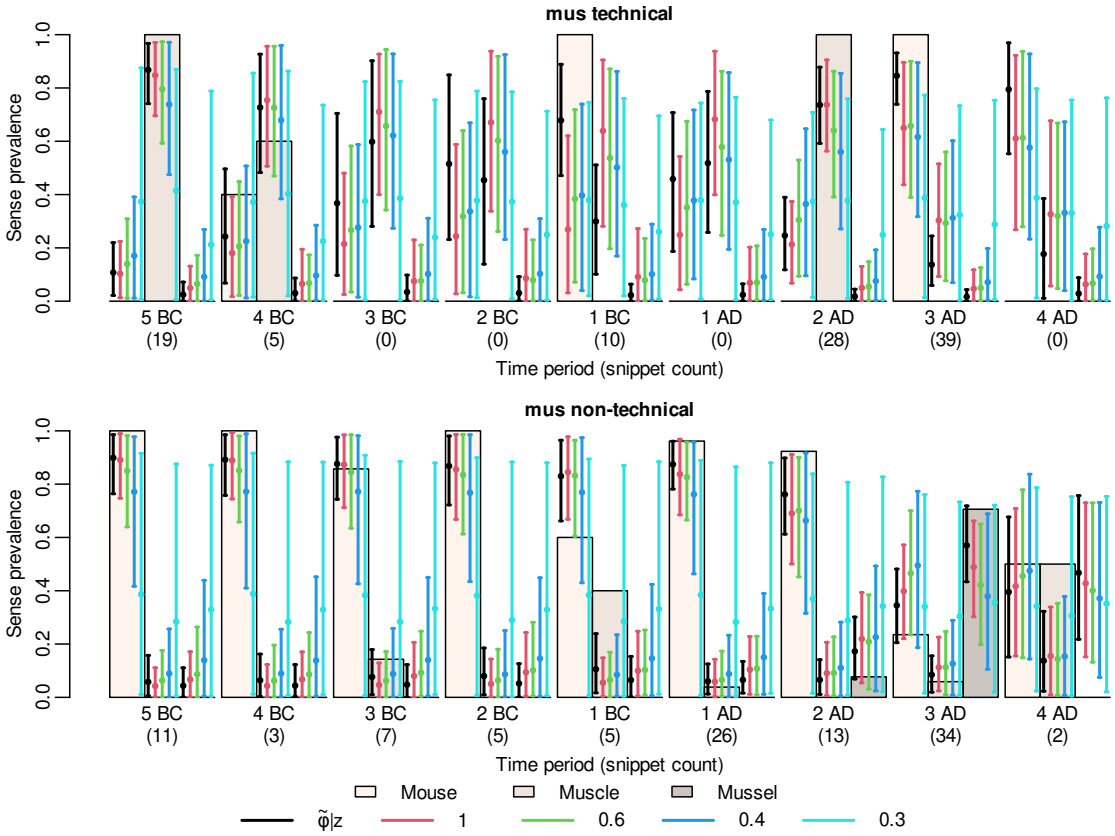


Figure 4: 95% HPD intervals (error bars) and posterior means (circles) for the model output on "mus" data. Results are shown for the proxy truth $\tilde{\phi}|z$ and the posteriors $\tilde{\phi}^{\lambda}|W$ for $\lambda \in \{1, 0.6, 0.4, 0.3\}$. Expert-annotated empirical sense prevalence (solid bars) for the three true senses is also shown.

When we reduce λ further to 0.4, the posterior location starts to gets less accurate due to lower influence from the data, and the variance increases to cover regions of the posterior space that do not overlap with the proxy truth. Thus, whilst parts of the posterior still return good predictive accuracy, the overall accuracy reduces on average. However, up to this point, the model may still be described as somewhat 'well specified' in the sense of achieving reasonable overlap with the proxy-true $\tilde{\phi}|(z = o)$. Reducing λ further to 0.3, we see that the posterior $\tilde{\phi}^{\lambda}|W$ now becomes too diffuse, and too uniform across senses, to be described as well specified in any way. This explains why a reduction in λ helps up to a point, and why the degree of misspecification (which we measure using PPC) is an intuitive choice to demarcate this point.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication, there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (only required where there isn't already a statement of contribution within the paper itself).

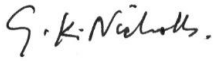
Title of paper	Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks
Publication status	Unpublished and unsubmitted work written in a manuscript style
Publication details	Schyan Zafar and Geoff K. Nicholls (2024). Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks. *arXiv e-prints*, art. arXiv:2410.01475. DOI: 10.48550/arXiv.2410.01475

Student Confirmation

Student name	Schyan Zafar		
Contribution to the paper	I proposed many of the research ideas for this paper; did all the coding and performed the experiments; and drafted the manuscript.		
Signature		Date	7 October 2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name and title	Prof. Geoff K. Nicholls		
Supervisor comments			
Signature		Date	7 October 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

Conclusion

"A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions." — Michael J. Moroney

The above quote is perhaps not as famous as the aphorism with which this thesis opened, but it encapsulates well the nature of what this thesis has accomplished. A good statistical analysis is like a surgical procedure, where a methodical approach, derived through careful consideration and executed with precision, is used to extract meaningful insights from often complex data with inherent uncertainty. This thesis is an application of one such analysis. We recap its main contributions below.

1 Summary and contributions

In Chapter 2, the contributions were twofold: a new DiSC model and a novel MCMC sampling strategy for this class of generative models. The goal of this chapter was to develop new models and methods for learning target-word senses from unlabelled snippets, and obtaining well-calibrated measurements of diachronic sense change with uncertainty quantification. Taken together, our innovations helped us achieve this goal for the applications considered.

The main difference in DiSC compared to SCAN/GASC from which it derived is in modelling the sense and time effects as additive via $\psi^{k,t} = \chi^k + \theta^t$. This seemingly simple change was the result of a long process of critical examination of SCAN/GASC, and was obtained after several discarded approaches. The problem with SCAN/GASC is that each sense $\tilde{\psi}^{k,\cdot}, k \in \{1, \ldots, K\}$, is allowed to evolve independently over time: the behaviour in any one sense has no effect on the others. This is of little consequence if we have large volumes

of data informing all model senses reasonably well, but results in over-parameterisation if the dataset is small and sparse. More structure is therefore needed in the model to relate the senses to each other, but it is not obvious how this may be done. We settled on the additive structure in DiSC by thinking about how parameter dimension might be reduced, and reasoning by analogy with classical regression: is an additive structure enough or do we need interaction? Physical considerations played a part, as we considered how context-word usage frequencies behave in practice. The gains are clear, to the extent that analysis for some data is now possible when previously it was not.

The novel MCMC sampling strategy is likewise straightforward, but the gains in efficiency are significant. This too was the result of critically examining the existing strategy used by previous authors, which was slow because of the conditioning on the discrete z, and the resultant restricted region of the posterior space explored in each iteration. It therefore made sense to marginalise out the discrete parameter and target the remaining continuous parameters with state-of-the-art (gradient-based) MCMC methods.

In Chapter 3, firstly, the main contribution was to introduce the new EDiSC model, which brings together two previously distinct approaches for modelling sense change, i.e. topic-based models and embedding-based models. This, compared to previous models, led to significant gains in predictive accuracy and true-model recovery as well as sampling efficiency and scalability.

It was natural to want to incorporate word embeddings into the model similar to how it was done for topic models; however, once again, the modelling was not straightforward, and a few different alternatives were considered as discussed. In particular, the correction parameter ς in $\psi^{k,t} = \rho(\chi^k + \theta^t) + \varsigma$ was the result of careful deliberation, and led to much better behaved posteriors than if we simply took $\psi^{k,t} = \rho(\chi^k + \theta^t)$ or perhaps $\psi^{k,t} = \rho\chi^k + \theta^t$.

Secondly, we showed how model selection with respect to the number of senses K and the embedding dimension M may be done in this setting. These are obviously important modelling decisions, and existing literature did not address the issue. These models are primarily exploratory tools used to discover latent structure in text data, so interpretability is a major consideration. Hence, we proposed setting K in a semi-supervised mode where the output was most meaningful to the user. For M, we proposed using the WAIC along with computational considerations to guide the choice. We gave guidelines rather than prescribe a method, but the guidelines are principled and lead to good predictive accuracy using the

selected models. Thirdly, we showed how to overcome some MCMC convergence issues using likelihood tempering.

In Chapter 4, the main contribution was to show two different ways of correlating target-word senses within the same document, both of which improve predictive accuracy. Secondly, not a contribution as such, but the refactorisation of $\tilde{\psi}$ shows the difficulty and importance of careful statistical modelling in this context. As discussed above, relating the model senses to each other is important, but how to do so is not obvious. This refactorisation showed an alternative approach in which the sense and time effects were multiplicative rather than additive. It did not lead to an improvement, but certainly gave food for thought.

In Chapter 5, the main contribution was to show a novel application of PPC in a GBI context to select the learning rate. PPC is a diagnostic tool used to *detect* model misspecification, not correct for it. We explored a new use of an established concept, and showed that systematic results are obtained in our setting, allowing us to select the learning rate such that the null hypothesis of a well-specified model is not rejected under a PPC at the 10% level. As with most of the other thesis contributions discussed above, this approach was the result of many months of work, during which quite a few other ideas were considered (and discarded). To avoid a multiple-testing hazard, we tested our method on ten new datasets not used in the method development, and showed that we obtain optimal or near-optimal learning rates for all of them. Whilst this chapter is only exploratory, it opens up a very interesting avenue for further investigation to generalise our method more widely.

As a by-product of the above, we have also created new annotated datasets that could be used for testing other models and methods for (diachronic or synchronic) sense change analysis.

2 Possible extensions

Notwithstanding all the contributions summarised above, further improvements to the models and methods developed in this thesis can obviously be made. We consider some possible extensions and directions for future research.

The models discussed in this thesis make use of some metadata extracted from the corpus in addition to the snippet data. The time and genre labels are examples of such metadata, and we utilise further metadata (i.e. the document information) in the EDiSC extensions discussed in Chapter 4. The models could be extended to use additional metadata, such as the original target-word form (rather than the lemmatised form) and/or part-of-speech

tags, since these could be very informative of the target-word sense. For example, if the word "banking" is used, it would almost inevitably be used in the financial institution sense; or if "bank" is used as a verb, chances are again very high that it is used in the institution sense. There is probably no single correct way of incorporating these metadata, and careful consideration of the model structure would again be required. One possibility is to incorporate these as additional layers in the hierarchical generative model. For example, the word form, x_d say, for snippet d could be drawn before the sense z_d, and z_d could be drawn conditional on x_d. Another possibility is to keep z_d unchanged, but to draw the data in pairs (W_d, x_d). Part-of-speech tags could be incorporated in a similar way. Some trial and error would be required to make things work, but the gains could potentially be quite significant, especially for target words where there is a strong correlation between the sense and word form and/or part of speech.

As discussed in Chapter 2, the gains made by DiSC through modelling sense and time as additive effects come at the cost of losing the sense-time interaction effect. This is fine in most real-use cases, since such interactions are rare, as evidenced by the fact that we had to work really hard to contrive an example where SCAN performed better than DiSC. However, where an interaction effect does genuinely exist, it would potentially be beneficial to incorporate that in the model. Therefore, some thought should perhaps be given to how such an interaction effect may be included without compromising the performance gains made. Consideration of the DiSC-f/EDiSC-f models introduced in Chapter 4 might generate some ideas, since this is effectively a different way of modelling the sense and time effects. Separately, we made a suggestion in Chapter 4 that, since EDiSC-η and EDiSC-β are both individually helpful, a combination of both models might be even more useful. This intuition might also be worth testing.

A separate line of research is to use some of the ideas developed in this thesis, in particular the additive sense and time effects, in the context of topic models. After all, these sense change models are closely related to topic models; so if a method is useful here then it may very well be useful there. Conversely, it is worth exploring other ideas from the topic modelling literature, and incorporating some of these into sense change models. One important respect in which this may be done is to relax the bag-of-words assumption, which is a major limitation of the models discussed. Some work has been done in this direction in topic modelling (Griffiths et al., 2004; Wallach, 2006). Another interesting respect is the use of Bayesian non-parametrics to tune K on the fly (Teh et al., 2004; Blei et al., 2010), which might prove useful in our context. Admittedly, these papers are bit dated, but some useful ideas may still be found in there. Relaxing the bag-of-words assumption, in particular, could

be cast within the EDiSC framework and considered together with contextualised (rather than static) word embeddings for a refreshing upgrade. However, we have not looked further into this. Churchill and Singh (2022) provide a concise and up-to-date summary of topic modelling, so that might be a starting point for this kind of investigation.

Finally, for the exploratory analysis in Chapter 5, the direction of future research is quite obvious: we need to determine the conditions required for our λ-selection method to work. A starting point for this is to test our intuition, using a wider range of synthetic data experiments, that the nature of misspecification which makes our method feasible is indeed posterior under-dispersion resulting from unmodelled latent correlations. Going further, we would need to examine the theoretical framework underpinning GBI and PPC more closely, so that theoretical guarantees may be given for the method to work under the right conditions.

3 Final thoughts

Research is a never-ending process, so there is always something more that could be done. However, it is important to bear in mind our starting point, where no well-calibrated procedure existed for measuring diachronic meaning change for the ancient Greek data using the existing SCAN and GASC models. Considering this, we have advanced this class of generative Bayesian models and inference methods very far indeed.

A limitation of this thesis is that we restricted all our analyses to manually annotated datasets where ground truth was available for testing. To verify the validity of our posited models and methods, this was essential. However, a real use of this thesis would be to apply these models and methods 'in the dark', i.e. on data that is genuinely unlabelled where a researcher is interested in finding structure within it. We believe that linguistic researchers will find these models and methods useful for such exploratory analyses.

The application need not be restricted to polysemous/homographic words: EDiSC could very well be fitted to any kind of text data to find sensible word groupings via the $\tilde{\phi}$ and $\tilde{\psi}$ parameters. An interesting application[1], for instance, might be to apply EDiSC to a set of political speeches, for some target words that are not necessarily polysemous but just of general interest. 'Sense' in this setting is simply a distribution over context words, and the inferred 'senses' could be very revealing, especially during times of historical change (e.g. a revolution) when the 'sense' could change dramatically. If that is the case, the model could

fit two different senses and observe a quick transition from one sense to the other.

The chief utility of the models and methods developed in this thesis is to allow target-word senses and diachronic meaning change to be inferred from unlabelled snippets with well-calibrated uncertainty quantification, within the familiar Bayesian framework, via interpretable parameters. On estimating and quantifying uncertainty in sense prevalence $\tilde{\phi}$ for the ancient Greek target words, without having sense labels, we do nearly as well as an 'oracle' classifier that simply fits a multinomial model conditioned on knowing the true sense labels. This shows that, though improvements are no doubt possible, the scope for improvement in $\tilde{\phi}$-estimation from snippet data (even with grammar and human-like knowledge of context) must be limited. Referring to Box's aphorism once again, the models developed in this thesis are, naturally, wrong. *All* models are wrong! However, with experiments throughout the thesis, particularly on uncertainty quantification, we have demonstrated that our models are indeed very useful.

[1] This suggestion was offered by Janet B. Pierrehumbert, Professor of Language Modelling at the Oxford e-Research Centre. Due to limitations of time, we were unable to implement it.

Bibliography

Alatrash, R., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2020). CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.859. 148

Apidianaki, M. (2022). From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, **49**(2), 465–523. ISSN 0891-2017. doi:10.1162/coli_a_00474. URL https://doi.org/10.1162/coli_a_00474. 87

Bamler, R. and Mandt, S. (2017). Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. PMLR. URL https://proceedings.mlr.press/v70/bamler17a.html. 41

Bartunov, S., Kondrashkin, D., Osokin, A., and Vetrov, D. (2016). Breaking Sticks and Ambiguities with Adaptive Skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 130–138, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/bartunov16.html. 41

Bayarri, M. J. and Berger, J. O. (2000). P Values for Composite Null Models. *Journal of the American Statistical Association*, **95**(452), 1127–1142. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2669749. 142

Bayarri, M. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference*, **111**(1), 3–22. ISSN 0378-3758. doi:https://doi.org/10.1016/S0378-3758(02)00282-3. URL https://

www.sciencedirect.com/science/article/pii/S0378375802002823. Special issue I: Model Selection, Model Diagnostics, Empirical Bayes and Hierarchical Bayes. 141

Benoit, K., Muhr, D., and Watanabe, K. (2020). *Stopwords: Multilingual Stopword Lists*. URL https://CRAN.R-project.org/package=stopwords. R package version 2.1. 44, 85

Berra, A. (2018). *Ancient Greek and Latin Stopwords for Textual Analysis*. URL https://github.com/aurelberra/stopwords. Greek v2.7 as of 30 Oct 2018. 45, 85

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, **19**(5A), 1501 – 1534. doi:10.3150/12-BEJ414. URL https://doi.org/10.3150/12-BEJ414. 27, 69, 94, 116

Bevilacqua, M., Pasini, T., Raganato, A., and Navigli, R. (2021). Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2021/593. URL https://doi.org/10.24963/ijcai.2021/593. Survey Track. 87

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. doi:https://doi.org/10.1111/rssb.12158. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12158. 29, 136, 139

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, **55**(4), 77–84. ISSN 0001-0782. doi:10.1145/2133806.2133826. URL https://doi.org/10.1145/2133806.2133826. 21, 88

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi:10.1145/1143844.1143859. URL https://doi.org/10.1145/1143844.1143859. 17, 22, 40, 88, 122, 145

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, **3**(null), 993–1022. ISSN 1532-4435. URL https://dl.acm.org/doi/abs/10.5555/944919.944937. 21, 40, 88

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *J. ACM*, **57**

(2). ISSN 0004-5411. doi:10.1145/1667053.1667056. URL https://doi.org/10.1145/1667053.1667056. 164

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. doi:10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773. 24, 25

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. ISSN 2307-387X. doi:10.1162/tacl_a_00051. URL https://doi.org/10.1162/tacl_a_00051. 87

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC., New York, 1st edition. ISBN 9780429138508. doi:https://doi.org/10.1201/b10905. URL https://www.mcmchandbook.net/HandbookTableofContents.html. 25

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, **63**, 743–788. doi:https://doi.org/10.1613/jair.1.11259. URL https://www.proquest.com/scholarly-journals/word-sense-embeddings-survey-on-vector/docview/2554077036/se-2?accountid=13042. 41

Carmona, C. and Nicholls, G. (2020). Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4226–4235. PMLR. URL https://proceedings.mlr.press/v108/carmona20a.html. 140, 143

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**(3), 33. doi:10.1007/s11222-023-10207-5. URL https://doi.org/10.1007/s11222-023-10207-5. 137, 143

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). Scalable inference for logistic-normal topic models. In *NIPS*, volume 26, pages 2445–2453. Citeseer. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/285f89b802bcb2651801455c86d78f2a-Paper.pdf. 40, 51, 67

Churchill, R. and Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, **54**(10s), 1–35. ISSN 0360-0300. doi:10.1145/3507900. URL https://doi.org/10.1145/3507900. 88, 165

Davies, M. (2010). The Corpus of Historical American English: 400 million words, 1810-2009. URL http://corpus.byu.edu/coha/. 43, 84, 148

Davies, M. (2012). The 400 million word Corpus of Historical American English (1810–2009). In *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010*, volume 325, pages 231–262. John Benjamins Publishing. URL https://www.jbe-platform.com/content/books/9789027273192-cilt.325.11dav. 148

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. 20, 87

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019). The Dynamic Embedded Topic Model. *arXiv e-prints*, art. arXiv:1907.05545. doi:10.48550/arXiv.1907.05545. 22, 42, 88, 145

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, **8**, 439–453. ISSN 2307-387X. doi:10.1162/tacl_a_00325. URL https://doi.org/10.1162/tacl_a_00325. 22, 88, 145

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, **195**(2), 216–222. ISSN 0370-2693. doi:10.1016/0370-2693(87)91197-X. URL https://www.sciencedirect.com/science/article/pii/037026938791197X. 27, 68, 94

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. *arXiv e-prints*, art. arXiv:1906.01688. doi:10.48550/arXiv.1906.01688. 41, 86

Frermann, L. and Lapata, M. (2016). A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, **4**, 31–45. ISSN 2307-387X. doi:10.1162/tacl_a_00081. URL https://doi.org/10.1162/tacl_a_00081. 17, 20, 38, 40, 43, 48, 50, 63, 64, 82, 86, 88, 145

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**(410), 398–409. doi:10.1080/01621459.1990.10476213. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476213. 25

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, **13**(4), 755–779. doi:10.1198/106186004X11435. URL https://doi.org/10.1198/106186004X11435. 141

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**(4), 733–760. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306036. 29, 141, 142

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, **PAMI-6**(6), 721–741. doi:10.1109/TPAMI.1984.4767596. 25, 28, 114

Geyer, C. (2011). Introduction to MCMC. In *Handbook of Markov Chain Monte Carlo*, pages 3–48. Chapman & Hall/CRC. doi:10.1201/b10905. URL https://www.mcmchandbook.net/HandbookChapter1.pdf. 25

Geyer, C. J., Keramidas, E., and Kaufman, S. (1991). Markov Chain Monte Carlo Maximum Likelihood. Interface Foundation of North America. URL https://hdl.handle.net/11299/58440. 113

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732. ISSN 0006-3444. doi:10.1093/biomet/82.4.711. URL https://doi.org/10.1093/biomet/82.4.711. 50

Griffiths, T., Steyvers, M., Blei, D., and Tenenbaum, J. (2004). Integrating topics and syntax. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf. 164

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, **101**(suppl_1), 5228–5235. doi:10.1073/pnas.0307752101. URL https://www.pnas.org/doi/abs/10.1073/pnas.0307752101. 25, 40

Groenewald, P. C. and Mokgatlhe, L. (2005). Bayesian computation for logistic regression. *Computational Statistics & Data Analysis*, **48**(4), 857 – 868. ISSN 0167-9473.

doi:https://doi.org/10.1016/j.csda.2004.04.009. URL https://www.sciencedirect.com/science/article/pii/S0167947304001148. 66

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, **12**(4), 1069 – 1103. doi:10.1214/17-BA1085. URL https://doi.org/10.1214/17-BA1085. 136, 139, 145

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, **29** (1), 83–100. doi:https://doi.org/10.1111/j.2517-6161.1967.tb00676.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1967.tb00676.x. 141

Hajek, B. (1988). Cooling schedules for optimal annealing. *Mathematics of operations research*, **13**(2), 311–329. doi:10.1287/moor.13.2.311. URL https://doi.org/10.1287/moor.13.2.311. 28, 114

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv e-prints*, art. arXiv:1605.09096. doi:10.48550/arXiv.1605.09096. 41, 86

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109. ISSN 0006-3444. doi:10.1093/biomet/57.1.97. URL https://doi.org/10.1093/biomet/57.1.97. 25

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, **101** (475), 1157–1174. doi:10.1198/016214505000001393. URL https://doi.org/10.1198/016214505000001393. 142

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. URL http://jmlr.org/papers/v15/hoffman14a.html. 27, 54, 94, 147

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. ISSN 0006-3444. doi:10.1093/biomet/asx010. URL https://doi.org/10.1093/biomet/asx010. 136, 140

Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principles of Bayesian Inference Using General Divergence Criteria. *Entropy*, **20**(6). ISSN 1099-4300. doi:10.3390/e20060442. URL https://www.mdpi.com/1099-4300/20/6/442. 139

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**(2), 183–233. ISSN 1573-0565. doi:10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178. 24

Karagiannis, G. and Andrieu, C. (2013). Annealed Importance Sampling Reversible Jump MCMC Algorithms. *Journal of Computational and Graphical Statistics*, **22** (3), 623–648. doi:10.1080/10618600.2013.805651. URL https://doi.org/10.1080/10618600.2013.805651. 50

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**(430), 773–795. doi:10.1080/01621459.1995.10476572. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572. 50, 103

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, **6**(none), 354 – 381. doi:10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675. 139

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic Variational Inference in Stan. *arXiv e-prints*, art. arXiv:1506.03431. doi:10.48550/arXiv.1506.03431. 94

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi:10.1145/2736277.2741627. URL https://doi.org/10.1145/2736277.2741627. 41, 86

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1117. 86

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. *arXiv e-prints*, art. arXiv:1806.03537. doi:10.48550/arXiv.1806.03537. 41

Li, J. and Huggins, J. H. (2022). Calibrated Model Criticism Using Split Predictive Checks. *arXiv e-prints*, art. arXiv:2203.15897. doi:10.48550/arXiv.2203.15897. 142

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, **106**(2), 465–478. ISSN 0006-3444. doi:10.1093/biomet/asz006. URL https://doi.org/10.1093/biomet/asz006. 136, 140

Manchanda, S. and Karypis, G. (2019). Distributed representation of multi-sense words: A loss-driven approach. *arXiv e-prints*, art. arXiv:1904.06725. doi:10.48550/arXiv.1904.06725. 41

McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., and Vatri, A. (2019). A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, **34**(4), 893–907. ISSN 2055-7671. doi:10.1093/llc/fqz036. URL https://doi.org/10.1093/llc/fqz036. 19, 85

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany, August 2016. Association for Computational Linguistics. doi:10.18653/v1/K16-1006. URL https://aclanthology.org/K16-1006. 87

Meng, X.-L. (1994). Posterior Predictive p-Values. *The Annals of Statistics*, **22**(3), 1142 – 1160. doi:10.1214/aos/1176325622. URL https://doi.org/10.1214/aos/1176325622. 29, 141, 142, 145

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092. ISSN 0021-9606. doi:10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114. 25

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, art. arXiv:1301.3781. doi:10.48550/arXiv.1301.3781. 41, 87

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, volume 26, pages 3111–3119. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. 41, 87

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746–751, Atlanta, Georgia, June 2013c. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090. 41, 87

Miller, J. W. (2021). Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors. *Journal of Machine Learning Research*, **22**(168), 1–53. URL http://jmlr.org/papers/v22/20-469.html. 139

Mimno, D., Wallach, H., and McCallum, A. (2008). Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, volume 61. URL https://api.semanticscholar.org/CorpusID:15401050. 25, 26, 40, 51, 66

Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., and Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. *arXiv e-prints*, art. arXiv:1405.4392. doi:10.48550/arXiv.1405.4392. 41, 86

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., and Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, **21**(5), 773–798. doi:10.1017/S135132491500011X. 41, 86

Montanelli, S. and Periti, F. (2023). A Survey on Contextualised Semantic Shift Detection. *arXiv e-prints*, art. arXiv:2304.01666. doi:10.48550/arXiv.2304.01666. 87

Moran, G. E., Blei, D. M., and Ranganath, R. (2023). Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **86**(1), 194–214. ISSN 1369-7412. doi:10.1093/jrsssb/qkad105. URL https://doi.org/10.1093/jrsssb/qkad105. 142

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **20**(5). ISSN 2375-4699. doi:10.1145/3434237. URL https://doi.org/10.1145/3434237. 87

Neal, R. (2011). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall/CRC. doi:10.1201/b10905. URL https://www.mcmchandbook.net/HandbookChapter5.pdf. 27, 68, 94, 116

Pacchiardi, L. (2022). *Statistical inference in generative models using scoring rules*. PhD thesis, University of Oxford. URL https://ora.ox.ac.uk/objects/uuid:67db26b1-3dbf-4087-bbd1-7297769ff37f. 139

Pacchiardi, L., Khoo, S., and Dutta, R. (2021). Generalized Bayesian Likelihood-Free Inference. *arXiv e-prints*, art. arXiv:2104.03889. doi:10.48550/arXiv.2104.03889. 139

Patel, K. and Bhattacharyya, P. (2017). Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 31–36, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-2006. 93

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162. 41, 87, 93, 148

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., and McGillivray, B. (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4707. URL https://aclanthology.org/W19-4707. 17, 18, 20, 38, 40, 44, 50, 61, 63, 64, 75, 82, 85, 86, 88, 145

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202. 87

Polson, N. G., Scott, J. G., and Windle, J. (2012). Bayesian inference for logistic models using Polya-Gamma latent variables. *arXiv e-prints*, art. arXiv:1205.0310. doi:10.48550/arXiv.1205.0310. 40, 51, 67

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**(2), 945–959. ISSN 1943-2631. doi:10.1093/genetics/155.2.945. URL https://doi.org/10.1093/genetics/155.2.945. 21

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 94

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9. URL https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf. 20, 87

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York. ISBN 9781475741452. doi:https://doi.org/10.1007/978-1-4757-4145-2. 24

Roberts, G. O. and Rosenthal, J. S. (2002). Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **60**(1), 255–268. ISSN 1369-7412. doi:10.1111/1467-9868.00123. URL https://doi.org/10.1111/1467-9868.00123. 27, 69, 94

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**(4), 341–363. ISSN 13507265. URL https://projecteuclid.org/journals/bernoulli/volume-2/issue-4/Exponential-convergence-of-Langevin-distributions-and-their-discrete-approximations/bj/1178291835.full. 27, 68, 94

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association*, **95**(452), 1143–1156. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2669750. 142, 145

Rodda, M. A., Probert, P., and McGillivray, B. (2019). Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique des Langues*, **60**(3), 63–87. URL https://aclanthology.org/2019.tal-3.4. 92

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, **12**(4), 1151 – 1172. doi:10.1214/aos/1176346785. URL https://doi.org/10.1214/aos/1176346785. 141

Rudolph, M. and Blei, D. (2018). Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi:10.1145/3178876.3185999. URL https://doi.org/10.1145/3178876.3185999. 41, 86

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. (2016). Exponential Family Embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 478–486. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf. 41

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, New York, 1st edition. ISBN 9780429208829. doi:https://doi.org/10.1201/9780203492024. 40, 64

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi:10.18653/v1/2020.semeval-1.1. URL https://aclanthology.org/2020.semeval-1.1. 42

Selivanov, D. (2022). Glove word embeddings. URL https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html. Accessed 2022-06-01. 93

Shaby, B. and Wells, M. T. (2010). Exploring an adaptive Metropolis algorithm. *Technical report*. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a4aa10db04b9094cc9a913883e7a79e4dc8e915a. 68, 116

Stan Development Team (2023a). RStan: the R interface to Stan. URL https://mc-stan.org/. R package version 2.32.3. 94, 147

Stan Development Team (2023b). Stan modeling language user's guide and reference manual. URL https://mc-stan.org/docs/2_26/reference-manual/index.html. Stan version 2.26.1. 94, 147

Statisticat and LLC. (2021). *LaplacesDemon: Complete Environment for Bayesian Inference*. URL https://web.archive.org/web/20150206004624/http://www.Bayesian-inference.com/software. R package version 16.1.6. 99

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2021). Non-Reversible Parallel Tempering: A Scalable Highly Parallel MCMC Scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**(2), 321–350. ISSN 1369-7412. doi:10.1111/rssb.12464. URL https://doi.org/10.1111/rssb.12464. 113

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. *Biometrika*, **106**(2), 479–486. ISSN 0006-3444. doi:10.1093/biomet/asy054. URL https://doi.org/10.1093/biomet/asy054. 136, 140

Tahmasebi, N. and Risse, T. (2017). Finding Individual Word Sense Changes and their Delay in Appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria, September 2017. INCOMA Ltd. doi:10.26615/978-954-452-049-6_095. URL https://doi.org/10.26615/978-954-452-049-6_095. 41, 86

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of Computational Approaches to Lexical Semantic Change. *arXiv e-prints*, art. arXiv:1811.06278. doi:10.48550/arXiv.1811.06278. 40

Tahmasebi, N., Borin, L., and Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. URL https://doi.org/10.5281/zenodo.5040302. 86

Tang, X. (2018). A State-of-the-Art of Semantic Change Computation. *arXiv e-prints*, art. arXiv:1801.09872. doi:10.48550/arXiv.1801.09872. 40

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, **24**(5), 649–676. doi:10.1017/S1351324918000220. 86

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/fb4ab556bc42d6f0ee0f9e24ec4d1af0-Paper.pdf. 164

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 20

Vatri, A. and McGillivray, B. (2018). The Diorisis Ancient Greek corpus. *Research Data Journal for the Humanities and Social Sciences*, **3**(1), 55 – 65. doi:10.1163/24523666-01000013. URL https://brill.com/view/journals/rdj/3/1/article-p55_55.xml. 18, 19, 43, 84, 148

Vatri, A., Lähteenoja, V., and McGillivray, B. (2019). Ancient Greek semantic change - annotated datasets and code. *figshare*. doi:10.6084/m9.figshare.c.4445420. URL https://figshare.com/collections/Ancient_Greek_semantic_change_-_annotated_datasets_and_code/4445420/1. 19, 44, 85, 148

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. ISSN 1573-1375. doi:10.1007/s11222-016-9696-4. URL https://doi.org/10.1007/s11222-016-9696-4. 28, 99, 100, 140

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2024). loo: Efficient leave-one-out cross-validation and waic for Bayesian models. URL https://mc-stan.org/loo/. R package version 2.7.0. 99

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **381**. doi:10.1098/rsta.2022.0149. 18

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. ISSN 1935-8237. doi:10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001. 24

Walker, S. and Hjort, N. L. (2002). On Bayesian Consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **63**(4), 811–821. ISSN 1369-7412. doi:10.1111/1467-9868.00314. URL https://doi.org/10.1111/1467-9868.00314. 29, 139

Walker, S. G., Lijoi, A., and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika*, **92**(4), 765–778. ISSN 0006-3444. doi:10.1093/biomet/92.4.765. URL https://doi.org/10.1093/biomet/92.4.765. 139

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 977–984, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi:10.1145/1143844.1143967. URL https://doi.org/10.1145/1143844.1143967. 164

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of machine learning research*, **11**(116), 3571–3594. ISSN 1532-4435. URL http://jmlr.org/papers/v11/watanabe10a.html. 28, 99, 140

Wikipedia contributors (2024). Attention is all you need — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Attention_Is_All_You_Need&oldid=1248164812. [Online; accessed 8-October-2024]. 20

Winter, S., Melikechi, O., and Dunson, D. B. (2023). Sequential Gibbs Posteriors with Applications to Principal Component Analysis. *arXiv e-prints*, art. arXiv:2310.12882. doi:10.48550/arXiv.2310.12882. 140

Wu, P.-S. and Martin, R. (2023). A Comparison of Learning Rate Selection Methods in Generalized Bayesian Inference. *Bayesian Analysis*, **18**(1), 105 – 132. doi:10.1214/21-BA1302. URL https://doi.org/10.1214/21-BA1302. 29, 136, 137, 139

Xing, H. (2021). Improving Bridge estimators via f-GAN. *arXiv e-prints*, art. arXiv:2106.07462. doi:10.48550/arXiv.2106.07462. 50

Yin, Z. and Shen, Y. (2018). On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/b534ba68236ba543ae44b22bd110a1d6-Paper.pdf. 93

Yüksel, A., Uğurlu, B., and Koç, A. (2021). Semantic change detection with gaussian word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3349–3361. doi:10.1109/TASLP.2021.3120645. 86

Zafar, S. and Nicholls, G. K. (2022). Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5), 1569–1604. ISSN 0035-9254. doi:10.1111/rssc.12591. URL https://doi.org/10.1111/rssc.12591. 26, 82, 83, 84, 85, 86, 88, 94, 101, 107, 114, 116, 145

Zafar, S. and Nicholls, G. K. (2024a). An Embedded Diachronic Sense Change Model with a Case Study from Ancient Greek. *Computational Statistics & Data Analysis*, **199**, 108011. ISSN 0167-9473. doi:10.1016/j.csda.2024.108011. URL https://www.sciencedirect.com/science/article/pii/S0167947324000951. 27, 28, 137, 140, 145, 146, 147, 148, 149, 156, 157

Zafar, S. and Nicholls, G. K. (2024b). Exploring Learning Rate Selection in Generalised Bayesian Inference using Posterior Predictive Checks. *arXiv e-prints*, art. arXiv:2410.01475. doi:10.48550/arXiv.2410.01475. 18, 29

Zhang, T. (2006). From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, **34**(5), 2180 – 2210. doi:10.1214/009053606000000704. URL https://doi.org/10.1214/009053606000000704. 29, 139