UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Bandas de predição usando densidade condicional estimada e um modelo LDA com covariáveis.

Gilson Yuuji Shimizu

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Gilson Yuuji Shimizu

Bandas de predição usando densidade condicional estimada e um modelo LDA com covariáveis.

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

USP – São Carlos
Outubro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Shimizu, Gilson Yuuji
S556b Bandas de predição usando densidade condicional estimada e um modelo LDA com covariáveis. / Gilson Yuuji Shimizu; orientador Rafael Izbicki. -- São Carlos, 2021.
87 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Aprendizagem de máquina. 2. Análise de texto. 3. Alocação latente de Dirichlet (LDA). 4. Bandas de predição. 5. Predição conformal. I. Izbicki, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Gilson Yuuji Shimizu

Prediction bands using estimated conditional density and an LDA model with covariates.

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC- USP and to the Department of Statistics – DEs- UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

USP – São Carlos
October 2021

This work is dedicated to my wife Tiemi and my son Jun.

ACKNOWLEDGEMENTS

The author would like to thank the following people:

- Professor Rafael Izbicki for his guidance, support and encouragement in this work and for the teachings in the Statistical Machine Learning course.
- Professor Denis Vale for his guidance in the second part of this work at the University of Florida and for all his support during this period.
- Professor Rafael Stern for the co-orientation in the first part of this work and for the teachings in the Decision Theory course.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

"Trust in the Lord with all your heart and lean not on your own understanding;
in all your ways submit to him, and he will make your paths straight."
Proverbs 3:5-6

ABSTRACT

SHIMIZU, G. **Prediction bands using estimated conditional density and an LDA model with covariates**. 2021. 87p. Thesis (Doctorate in Statistics) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo; Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, 2021.

Machine learning methods are divided into two main groups: supervised and unsupervised methods. In the first part of this work, we develop a method for creating prediction bands that can be applied to supervised problems. Our approach is based on conformal methods, which are very appealing because they create prediction bands that control average coverage assuming solely i.i.d. data. It is also often desirable to control conditional coverage, that is, coverage for every new testing point. However, without strong assumptions, conditional coverage is unachievable. Given this limitation, the literature has focused on methods with asymptotical conditional coverage. In order to obtain this property, these methods require strong conditions on the dependence between the target variable and the features. We introduce two conformal methods based on conditional density estimators that do not depend on this type of assumption to obtain asymptotic conditional coverage: Dist-split and CD-split. While Dist-split asymptotically obtains optimal intervals, which are easier to interpret than general regions, CD-split obtains optimal size regions, which are smaller than intervals. CD-split also obtains local coverage by creating prediction bands locally on a partition of the features space. This partition is data-driven and scales to high-dimensional settings. In a wide variety of simulated scenarios, our methods have a better control of conditional coverage and have smaller length than previously proposed methods.

In the second part, in a context of unsupervised methods, we develop a new version of the Latent Dirichlet Allocation (LDA) model. The LDA model is a popular method for creating mixed-membership clusters. Despite having been originally developed for text analysis, LDA has been used for a wide range of other applications. We propose a new formulation for the LDA model which incorporates covariates. In this model, a negative binomial regression is embedded within LDA, enabling straight-forward interpretation of the regression coefficients and the analysis of the quantity of cluster-specific elements in each sampling units (instead of the analysis being focused on modeling the proportion of each cluster, as in Structural Topic Models). We use slice sampling within a Gibbs sampling algorithm to estimate model parameters. We rely on simulations to show how our algorithm is able to successfully retrieve the true parameter values. The model is illustrated using real data sets from three different areas: text-mining of Coronavirus articles, analysis of grocery shopping baskets, and ecology of tree species on Barro Colorado Island (Panama). This model allows the identification of mixed-membership clusters in discrete data and provides inference on the relationship between covariates and the abundance of these clusters.

Keywords: Machine learning. Text analysis. Latent Dirichlet allocation (LDA). Prediction bands. Conformal prediction.

RESUMO

SHIMIZU, G. **Bandas de predição usando densidade condicional estimada e um modelo LDA com covariáveis**. 2021. 87p. Thesis (Doctorate in Statistics) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo; Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, 2021.

Métodos de machine learning são basicamente divididos em dois grandes grupos: métodos supervisionados e não supervisionados. Na primeira parte deste trabalho nós desenvolvemos um método para criação de bandas de predição que pode ser aplicado em problemas supervisionados. Nossa abordagem é baseada em métodos conformal, que são interessantes porque criam bandas de predição que controlam a cobertura média assumindo somente dados i.i.d.. Geralmente também é desejável controlar a cobertura condicional, ou seja, a cobertura para toda nova amostra de teste. Contudo, sem fortes suposições, a cobertura condicional é inatingível. Dada esta limitação, a literatura tem focado em métodos com cobertura condicional assintótica. A fim de se obter esta propriedade, estes métodos requerem fortes suposições sobre a dependência entre a variável resposta e as características. Nós introduzimos dois métodos conformal baseados em estimadores de densidade condicionais que não dependem deste tipo de suposição para obter cobertura condicional assintótica: Dist-split e CD-split. Enquanto Dist-split obtém intervalos ótimos assintoticamente, que são mais fáceis de interpretar do que regiões de confiança, CD-split obtém regiões de tamanho ótimo, que são menores do que intervalos. CD-split também obtém cobertura local pela criação de bandas de predição localmente numa partição do espaço de características. Esta partição é baseada em dados e permite trabalhar com dados em alta dimensão. Numa grande variedade de cenários simulados, nossos métodos tem melhor controle da cobertura condicional e tem menores comprimentos do que métodos propostos anteriores.

Na segunda parte, num contexto de métodos não supervisionados, estudamos uma nova versão do modelo de Alocação Latente Dirichlet (LDA). O modelo LDA é um método popular para criação de mixed-membership clusters. Apesar de ter ficado conhecido na análise de texto, LDA tem sido usado em uma variedade de outras aplicações. Nós propomos uma nova formulação para o modelo LDA que incorpora covariáveis. Neste modelo, uma regressão binomial negativa é embutida dentro do LDA, possibilitando uma interpretação direta dos coeficientes de regressão e análise da quantidade de elementos específicos dos clusters em cada unidade amostral (ao invés da análise ser focada em modelar a proporção de cada cluster, como nos Modelos de Tópicos Estruturados). Nó usamos slice sampling dentro de um algoritmo de Gibbs sampling para estimar os parâmetros. E usamos simulações para mostrar como nosso algoritmo é capaz de estimar com sucesso os

verdadeiros parâmetros do modelo. O modelo é ilustrado usando conjuntos de dados reais de três diferentes áreas: mineração de texto de artigos sobre coronavírus, análise de cestas de supermercados, e análise de espécies de árvores na Ilha de Barro Colorado (Panama). Este modelo permite a identificação de mixed-membership clusters em dados discretos e fornece inferências sobre o relacionamento entre covariáveis e a abundância destes clusters.

Palavras-chave: Aprendizagem de máquina. Análise de texto. Alocação latente de Dirichlet (LDA). Bandas de predição. Predição conformal.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Al	Aluminium
BCI	Barro Colorado Island
CD	Conditional density
CDF	Cumulative distribution function
FCD	Full conditional distributions
FDP	Forest dynamic plot
iid	Independent and identically distributed
LDA	Latent Dirichlet allocation
MCMC	Markov chain Monte Carlo
Mn	Manganese
N	Nitrogen
STM	Structural topic models
Zn	Zinc

CONTENTS

1 THESIS OVERVIEW AND PUBLICATIONS

Machine learning methods (FRIEDMAN *et al.*, 2001; JAMES *et al.*, 2013; IZBICKI; SANTOS, 2018) can be divided into two main areas: supervised methods and unsupervised methods. In supervised settings, we have a response variable that we want to make predictions for given set of covariates. In unsupervised settings, we do not have an observable response variable and we are generally interested in understanding the structure of the data.

This work is divided into two parts that fall into these two areas. In the first part, in a context of supervised methods, we present a method for constructing prediction bands that can be applied to any regression method with the only assumption of i.i.d. data.

In the second part, we present a new formulation of the Latent Dirichlet Allocation (LDA) model. This unsupervised model is often used to find unknown topics in text documents. In this new formulation, covariates are incorporated in order to allow an easy interpretation of the regression coefficients.

The content of this work has appeared previously in the following publications:

- Izbicki, R., Shimizu, G. Y., Stern, R. B. (2020). Flexible distribution-free conditional predictive bands using density estimators. Proceedings of Machine Learning Research (AISTATS Track).
- Valle, D., Shimizu, G., Izbicki, R., Maracahipes, L., Silvério, D., Paolucci, L., Jameel, Y., Brando, P. (2021). The Latent Dirichlet Allocation model with covariates (LDA-cov): a case study on the effect of fire on species composition in Amazonian forests. Ecology and Evolution.
- Shimizu, G., Izbicki, R.,Valle, D. (2021). A new LDA formulation with covariates. Submitted for publication.

Part I

2 INTRODUCTION

Supervised machine learning methods predict a response variable, $Y \in \mathcal{Y}$, based on features, $\mathbf{X} \in \mathcal{X}$, using an i.i.d. sample, $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. While most methods yield point estimates, it is often more informative to present prediction bands, that is, a subset of $\mathcal{Y}$ with plausible values for Y (NETER *et al.*, 1996).

A particular way of constructing prediction bands is through *conformal predictions* (VOVK *et al.*, 2005; VOVK *et al.*, 2009). This methodology is appealing because it controls the *marginal coverage* of the prediction bands assuming solely i.i.d. data. Specifically, given a new instance, $(\mathbf{X}_{n+1}, Y_{n+1})$, a conformal prediction, $C(\mathbf{X}_{n+1})$, satisfies

$$\mathbb{P}\left(Y_{n+1} \in C(\mathbf{X}_{n+1})\right) \geq 1 - \alpha,$$

where $0 < 1 - \alpha < 1$ is a desired coverage level. Besides marginal validity one might also wish for stronger guarantees. For instance, *conditional validity* holds when, for every $\mathbf{x}_{n+1} \in \mathcal{X}$,

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \geq 1 - \alpha.$$

That is, conditional validity guarantees adequate coverage for each new instance and not solely on average across instances.

Unfortunately, conditional validity can be obtained only under strong assumptions about the the distribution of $(\mathbf{X}, Y)$ (VOVK, 2012; LEI; WASSERMAN, 2014; BARBER *et al.*, 2019). Given this result, effort has been focused on obtaining intermediate conditions. For instance, many conformal methods control *local coverage*:

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A) \geq 1 - \alpha,$$

where A is a subset of $\mathcal{X}$ (LEI; WASSERMAN, 2014; BARBER *et al.*, 2019; GUAN, 2019). These methods are based on computing conformal bands using only training instances that fall in A. However, to date, these methods do not scale to high-dimensional settings because it is challenging to create A that is large enough so that many training instances fall in A, and yet small enough so that

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A) \approx \mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}),$$

that is, local validity is close to conditional validity.

Another alternative to conditional validity is *asymptotic* conditional coverage (LEI *et al.*, 2018). Under this property, conditional coverage converges to the specified level

Table 1 – Properties of `Dist-split` and `CD-split`.

Method	Marginal coverage	Asymptotic conditional coverage	Local coverage	Prediction bands are intervals	Can be used for classification?
`Dist-split`	✓	✓	✗	✓	✗
`CD-split`	✓	✓	✓	✗	✓

as the sample size increases. That is, there exist random sets, Λ_n, such that $\mathbb{P}(X_{n+1} \in \Lambda_n | \Lambda_n) = 1 - o_{\mathbb{P}}(1)$ and

$$\sup_{\mathbf{x}_{n+1} \in \Lambda_n} \left| \mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) - (1-\alpha) \right| = o_{\mathbb{P}}(1).$$

In a regression context in which $\mathcal{Y} = \mathbb{R}$, Lei *et al.* (2018) obtains asymptotic conditional coverage under assumptions such as $Y = \mu(\mathbf{X}) + \epsilon$, where ϵ is independent of $\mathbf{X}$ and has density symmetric around 0. Furthermore, the proposed prediction band converges to the interval with the smallest interval among the ones with adequate conditional coverage.

Despite the success of these methods, there exists space for improvement. In many problems the assumption that ϵ is independent of $\mathbf{X}$ and has a density symmetric around 0 is unrealistic. For instance, in heteroscedastic settings (NETER *et al.*, 1996) ϵ depends on $\mathbf{X}$. It is also common for ϵ to have an asymmetric or even multimodal distribution (FREEMAN; IZBICKI; LEE, 2017). Furthermore, in these general settings, the smallest region with adequate conditional coverage might not be an interval, which is the outcome of most current methods.

2.1 Contribution

We propose new methods and show that they obtain asymptotic conditional coverage without assuming a particular type of dependence between the target and the features. Specifically, we propose two methods: `Dist-split` and `CD-split`. While `Dist-split` produces prediction bands that are intervals and easier to interpret, `CD-split` yields arbitrary regions, which are generally smaller and appealing for multimodal data. While `Dist-split` converges to an oracle interval, `CD-split` converges to an oracle region. Furthermore, since `CD-split` is based on a novel data-driven way of partitioning the feature space, it also controls local coverage even in high-dimensional settings. Table 1 summarizes the properties of these methods.

The proposed methods also have desirable computational properties. They are based on fast-to-compute split (inductive)-conformal bands (PAPADOPOULOS, 2008; VOVK, 2012; LEI *et al.*, 2018) and on novel conditional density estimation methods that scale to high-dimensional datasets (LUECKMANN *et al.*, 2017; PAPAMAKARIOS; PAVLAKOU; MURRAY, 2017; IZBICKI; LEE, 2016; IZBICKI; LEE, 2017; DALMASSO *et al.*, 2019; POSPISIL; LEE, 2019) Both methods are easy to compute and scale to large sample sizes as long as the conditional density estimator also does.

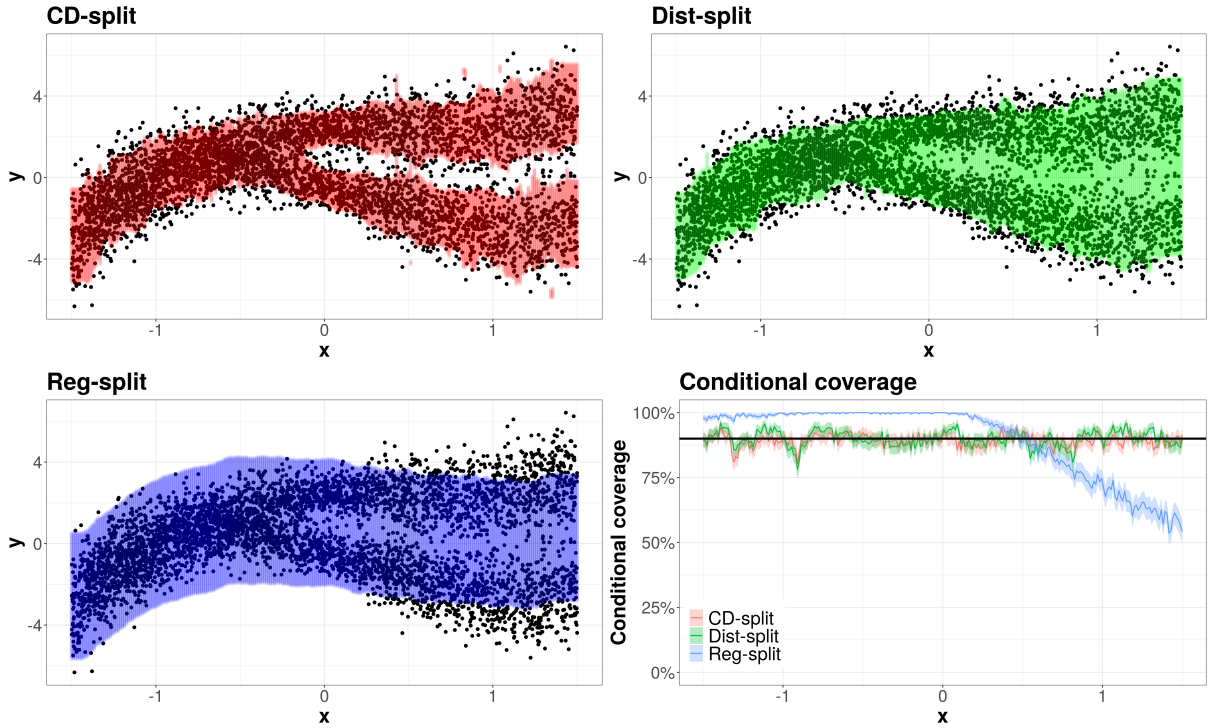


Figure 1 – Comparison between `CD-split`, `Dist-split` and the reg-split method from Lei *et al.* (2018).

In a wide variety of simulation studies, we show that our proposed methods obtain better conditional coverage and smaller band length than alternatives in the literature. For example, Figure 1 illustrates `CD-split`, `Dist-split` and the reg-split method from Lei *et al.* (2018) on the toy example from Lei and Wasserman (2014). The bottom right plot shows that both `CD-split` and `Dist-split` get close to controlling conditional coverage. Since `Dist-split` can yield only intervals, `CD-split` obtains smaller bands in the region in which $\mathbf{Y}$ is bimodal. In this region `CD-split` yields a collection of intervals around each of the modes.

This first part of the thesis is organized as follows: Chapter 3 and Chapter 4 introduce, respectively, `Dist-split`and `CD-split`. Experiments are shown in Chapter 5. All proofs can be found in the Appendix.

Notation. Unless stated otherwise, we study a univariate regression setting such that $\mathcal{Y} = \mathbb{R}$. Data from an i.i.d. sequence is split into two parts, $\mathbb{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ and $\mathbb{D}' = \{(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)\}$. Both datasets have the same size solely to simplify notation. Also, the new instance, $(\mathbf{X}_{n+1}, Y_{n+1})$, has the same distribution as the other sample units. Finally, $q(\alpha; \{t_1, \ldots, t_m\})$ is the α empirical quantile of $\{t_1, \ldots, t_m\}$.

3 DIST-SPLIT

The `Dist-split` method is based on the fact that, if $F(y|\mathbf{x})$ is the conditional distribution of Y_{n+1} given $\mathbf{X}_{n+1}$, then $F(Y_{n+1}|\mathbf{X}_{n+1})$ has uniform distribution. Therefore, if $\widehat{F}$ is close to F, then $\widehat{F}(Y_{n+1}|\mathbf{X}_{n+1})$ approximately uniform, and does not depend on $\mathbf{X}_{n+1}$. That is, obtaining marginal coverage for $\widehat{F}(Y_{n+1}|\mathbf{X}_{n+1})$ is close to obtaining conditional coverage.

Definition 3.1 (`Dist-split` prediction band)**.** Let $\widehat{F}(y|\mathbf{x}_{n+1})$ be an estimate based on $\mathbb{D}'$ of the conditional distribution of Y_{n+1} given $\mathbf{x}_{n+1}$. The `Dist-split` prediction band, $C(\mathbf{x}_{n+1})$, is

$$\begin{aligned} C(\mathbf{x}_{n+1}) &:= \left\{y : q(.5\alpha; \mathcal{T}(\mathbb{D})) \leq \widehat{F}(y|\mathbf{x}_{n+1}) \leq q(1-.5\alpha; \mathcal{T}(\mathbb{D}))\right\} \\ &= \left[\widehat{F}^{-1}(q(.5\alpha; \mathcal{T}(\mathbb{D}))|\mathbf{x}_{n+1}); \widehat{F}^{-1}(q(1-.5\alpha; \mathcal{T}(\mathbb{D}))|\mathbf{x}_{n+1})\right] \end{aligned}$$

where $\mathcal{T}(\mathbb{D}) = \left\{\widehat{F}(Y_i|\mathbf{X}_i), i = 1, \ldots, n\right\}$ and $\widehat{F}^{-1}$ is the generalized inverse of a cdf.

Algorithm 1 shows an implementation of `Dist-split`.

Algorithm 1 `Dist-split`

Input: Data $(\mathbf{X}_i, Y_i)$, $i = 1, ..., n$, miscoverage level $\alpha \in (0, 1)$, algorithm $\mathcal{B}$ for fitting conditional cumulative distribution function

Output: Prediction band for $\mathbf{x}_{n+1} \in \mathbb{R}^d$

1: Randomly split $\{1, 2, ..., n\}$ into two subsets $\mathbb{D}$ e $\mathbb{D}'$
2: Fit $\hat{F} = \mathcal{B}(\{(\mathbf{X}_i, Y_i) : i \in \mathbb{D}'\})$ **// Estimate cumulative distribution function**
3: Let $\mathcal{T}(\mathbb{D}) = \{\hat{F}(y_i|\mathbf{x}_i), i \in \mathbb{D}\}$
4: Let $t_1 = q(\alpha/2; \mathcal{T}(\mathbb{D}))$ and $t_2 = q(1-\alpha/2; \mathcal{T}(\mathbb{D}))$ **// Compute the quantiles of the set $\mathcal{T}(\mathbb{D})$**
5: **return** $\left\{y : t_2 \geq \widehat{F}(y|\mathbf{x}_{n+1}) \geq t_1\right\}$

`Dist-split` adequately controls the marginal coverage. Furthermore, it exceeds the specified $1-\alpha$ coverage by at most $(n+1)^{-1}$. These results are presented in Theorem 3.2.

Theorem 3.2 (Marginal coverage)**.** *Let $C(\boldsymbol{X}_{n+1})$ be such as in Definition 3.1. If both $F(y|\mathbf{x})$ and $\widehat{F}(y|\mathbf{x})$ are continuous for every $\mathbf{x} \in \mathcal{X}$, then*

$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in C(\boldsymbol{X}_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}.$$

Under additional assumptions `Dist-split` also obtains asymptotic conditional coverage and converges to an optimal oracle band. Two types of assumptions are required. First, that the conditional density estimator, $\widehat{F}$ is consistent. This assumption is an

adaptation to density estimators of the consistency assumption for regression estimators in Lei *et al.* (2018). Also, we require that $F(y|\mathbf{x})$ is differentiable and $F^{-1}(\alpha^*|\mathbf{x})$ is uniformly smooth in a neighborhood of $.5\alpha$ and $1-.5\alpha$. These assumptions are formalized below.

Assumption 3.3 (Consistency of density estimator)**.** There exist $\eta_n = o(1)$ and $\rho_n = o(1)$ such that

$$\mathbb{P}\left(\mathbb{E}\left[\sup_{y\in\mathcal{Y}}\left(\widehat{F}(y|\mathbf{X}) - F(y|\mathbf{X})\right)^2 \Big|\widehat{F}\right] \geq \eta_n\right) \leq \rho_n.$$

Assumption 3.4. For every $\mathbf{x} \in \mathcal{X}$, $F(y|\mathbf{x})$ is differentiable. Also, if $q_\alpha = F^{-1}(\alpha)$, then there exists $M^{-1} > 0$ such that $\inf_{\mathbf{x}} \frac{dF(y|\mathbf{x})}{dy} \geq M^{-1}$ in a neighborhood of $q_{0.5\alpha}$ and of $q_{1-0.5\alpha}$.

Given the above assumptions, `Dist-split` satisfies desirable theoretical properties. First, it obtains asymptotic conditional coverage. Also, `Dist-split` converges to the optimal *interval* according to the commonly used (PARMIGIANI; INOUE, 2009) loss function

$$L((a,b), Y_{n+1}) = \alpha(b-a) + (a - Y_{n+1})_+ + (Y_{n+1} - b)_+,$$

that is, `Dist-split` satisfies

$$C(\mathbf{X}_{n+1}) \approx \left[F^{-1}(.5\alpha|\mathbf{X}_{n+1}); F^{-1}(1-.5\alpha|\mathbf{X}_{n+1})\right]$$

These results are formalized in Theorem 3.5.

Theorem 3.5. *Let $C_n(\mathbf{X}_{n+1})$ be the prediction band in Definition 3.1 and $C^*(\mathbf{X}_{n+1})$ be the optimal prediction interval according to*

$$L((a,b), Y_{n+1}) = \alpha(b-a) + (a - Y_{n+1})_+ + (Y_{n+1} - b)_+.$$

Under Assumptions 3.3 and 3.4,

$$\lambda(C_n(\mathbf{X}_{n+1}) \Delta C^*(\mathbf{X}_{n+1})) = o_{\mathbb{P}}(1),$$

where λ is the Lebesgue measure.

Corollary 3.6. *`Dist-split` achieves asymptotic conditional coverage under Assumptions 3.3 and 3.4.*

`Dist-split` converges to the same oracle as recently proposed conformal quantile regression methods (ROMANO; PATTERSON; CANDÈS, 2019; SESIA; CANDÈS, 2019). However, the experiments in Chapter 5 show that `Dist-split` usually outperforms these methods.

If the distribution of $Y|\mathbf{x}$ is not symmetric and unimodal, `Dist-split` may obtain larger regions than necessary. For example, a union of two intervals better represents a bimodal distribution than a single interval. The next section introduces `CD-split` which obtains prediction bands that are more general than intervals.

4 CD-SPLIT

The intervals output by `Dist-split` are wider than necessary when the target distribution is multimodal, such as in Figure 1. In order to overcome this issue, `CD-split` yields prediction bands that approximate $\{y : f(y|\mathbf{x}_{n+1}) > t\}$, the highest posterior region.

A possible candidate for this approximation is $\left\{y : \widehat{f}(y|\mathbf{x}_{n+1}) > t\right\}$, where $\widehat{f}$ is a conditional density estimator. However, the value of t that guarantees conditional coverage varies according to $\mathbf{x}$. Thus, in order to obtain conditional validity, it is necessary to choose t adaptively. This adaptive choice for t is obtained by making $C(\mathbf{x}_{n+1})$ depend only on samples close to $\mathbf{x}_{n+1}$, similarly as in Lei and Wasserman (2014), Barber *et al.* (2019), Guan (2019).

Definition 4.1 (`CD-split` prediction band)**.** Let $\widehat{f}(y|\mathbf{x}_{n+1})$ be a conditional density estimate obtained from data $\mathbb{D}'$ and $0 < 1 - \alpha < 1$ be a coverage level. Let d be a distance on the feature space and $\mathbf{x}_1^c, \ldots, \mathbf{x}_J^c \in \mathcal{X}$ be centroids chosen so that $d(\mathbf{x}_i^c, \mathbf{x}_j^c) > 0$. Consider the partition of the feature space that associates each $\mathbf{x} \in \mathcal{X}$ to the closest $\mathbf{x}_j^c$, i.e., $\mathcal{A} = \{A_j : j = 1, \ldots, J\}$, where $A_j = \left\{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \mathbf{x}_j^c) < d(\mathbf{x}, \mathbf{x}_k^c) \text{ for every } k \neq j\right\}$. The `CD-split` prediction band for Y_{n+1} is:

$$C(\mathbf{x}_{n+1}) = \left\{y : \widehat{f}(y|\mathbf{x}_{n+1}) \geq q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))\right\},$$

where $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \{\widehat{f}(y_i|\mathbf{x}_i) : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\}$, and $A(\mathbf{x}_{n+1})$ is the element of $\mathcal{A}$ to which $\mathbf{x}_{n+1}$ belongs to.

Remark 1 (Multivariate responses)**.** *Although we focus on univariate targets,* ***`CD-split`*** *can be extended to the case in which $\mathbf{Y} \in \mathbb{R}^p$. As long as an estimate of $f(\mathbf{y}|\mathbf{x})$ is available, the same construction can be applied.*

The bands given by `CD-split` control local coverage in the sense proposed by Lei and Wasserman (2014).

Definition 4.2 (Local validity; Definition 1 of Lei and Wasserman (2014))**.** Let $\mathcal{A} = \{A_j : j \geq 1\}$ be a partition of $\mathcal{X}$. A prediction band C is locally valid with respect to $\mathcal{A}$ if, for every j and $\mathbb{P}$,

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A_j) \geq 1 - \alpha$$

Theorem 4.3 (Local and marginal validity)**.** *The* ***`CD-split`*** *band is locally valid with respect to $\mathcal{A}$. It follows from Lei and Wasserman (2014) that the* ***`CD-split`*** *band is also marginally valid.*

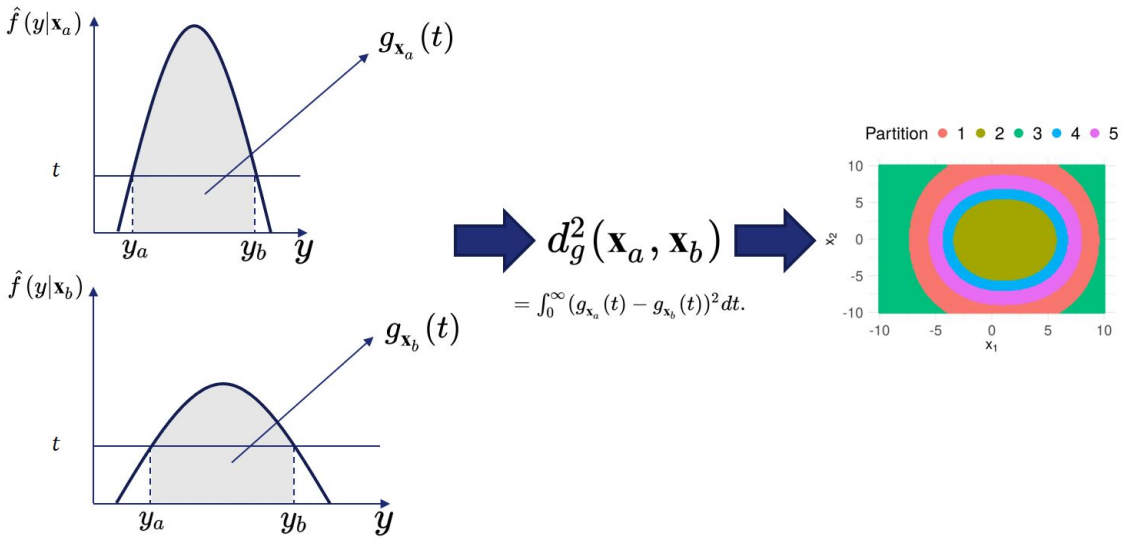


Figure 2 – Illustration of the profile distance, which is used in `CD-split` for partitioning the feature space.

Although `CD-split` controls local coverage, its performance drastically depends on the chosen partition of the feature space. If the partition is not chosen well, local coverage may be far from conditional coverage. For instance, if the partition is defined according to the Euclidean distance (LEI; WASSERMAN, 2014; BARBER *et al.*, 2019), then the method will not scale to high-dimensional feature spaces. In these settings small Euclidean neighborhoods have few data points and, therefore, large neighborhoods must be taken. As a result, local coverage is far from conditional coverage. We overcome this drawback by using a specific data-driven partition. In order to build this metric, we start by defining the profile of a density, which is illustrated in Figure 2.

Definition 4.4 (Profile of a density)**.** For every $\mathbf{x} \in \mathbb{R}^d$ and $t \geq 0$, the profile of $\widehat{f}(y|\mathbf{x})$, $g_{\mathbf{x}}(t)$, is

$$g_{\mathbf{x}}(t) := \int_{\{y:\widehat{f}(y|\mathbf{x}) \geq t\}} \widehat{f}(y|\mathbf{x}) dy.$$

The profile of a density is the cumulative distribution function associated to its level sets. It is used to define the profile distance in the feature space.

Definition 4.5 (Profile distance)**.** The profile distance[1] between $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{X}$ is

$$d_g^2(\mathbf{x}_a, \mathbf{x}_b) := \int_0^\infty \left(g_{\mathbf{x}_a}(t) - g_{\mathbf{x}_b}(t)\right)^2 dt,$$

Contrary to the Euclidean distance, the profile distance is appropriate even for high-dimensional data. For instance, two points might be far in Euclidean distance and

[1] The profile distance is a metric on the quotient space $\mathcal{X}/\sim$, where $\sim$ is the equivalence relation $\mathbf{x}_a \sim \mathbf{x}_b \iff g_{\mathbf{x}_a} = g_{\mathbf{x}_b}$ a.e.

still have similar conditional densities. In this case one would like these points to be on the same partition element. The profile obtains this result by measuring the distance between instances based on the distance between their conditional densities. By grouping points with similar conditional densities, the profile distance allows partition elements to be larger without compromising too much the approximation of local validity to conditional validity. This property is illustrated in the following examples.

Example 4.6. [**Location family**] Let $h(y)$ be a density, $\mu(\mathbf{x})$ a function, and $Y|\mathbf{x} \sim h(y-\mu(\mathbf{x}))$. In this case, $d_g(\mathbf{x}_a, \mathbf{x}_b) = 0$, for every $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^d$. For instance, if $Y|\mathbf{x} \sim N(\beta^t\mathbf{x}, \sigma^2)$, then all instances have the same profile. Indeed, in this special scenario, if `CD-split` uses a unitary partition, then conditional validity is obtained.

Example 4.7. [**Irrelevant features**] If $\mathbf{x}_S$ is a subset of the features such that $f(y|\mathbf{x}) = f(y|\mathbf{x}_S)$, then $d_g(\mathbf{x}_a, \mathbf{x}_b)$ does not depend on the irrelevant features, S^c. While irrelevant features do not affect the profile distance, they can have a large impact in the Euclidean distance in high-dimensional settings.

Also, if all samples that fall into the same partition as $\mathbf{x}_{n+1}$ have the same profile as $\mathbf{x}_{n+1}$ according to f, then the statistics used in `CD-split` , $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D})$ are i.i.d. data. Thus, the quantile used in `CD-split` will be the α quantile of $f(Y_{n+1}|\mathbf{x}_{n+1})$. This in turn makes $C(\mathbf{x}_{n+1})$ the smallest prediction band with conditional validity of $1-\alpha$. Theorem 4.8, below, formalizes this statement.

Theorem 4.8. *Assume that all samples that fall into the same partition as $\mathbf{x}_{n+1}$, say $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_m, Y_m)$, are such that $g_{\mathbf{x}_i} = g_{\mathbf{x}_{n+1}}$, and that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ is continuous as a function of y for every $\mathbf{x} \in \mathcal{X}$. Let $T_m := q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))$ be the cutoff used in `CD-split`. For every $\alpha \in (0,1)$*

$$T_m \xrightarrow[a.s.]{m \longrightarrow \infty} t^*$$

where $t^* = t^*(\mathbf{x}_{n+1}, \alpha)$ is the cutoff associated to the oracle band, the smallest predictive region with coverage $1-\alpha$.

Given the above reasons, the profile density captures what is needed of a meaningful neighborhood that contains many samples even in high dimensions. Indeed, consider a partition of the feature space, $\mathcal{A}$, that has the property that all samples that belong to the same element of $\mathcal{A}$ have the same oracle cutoff t^*. Theorem 4.9 shows that the coarsest partition that has this property is the one induced by the profile distance.

Theorem 4.9. *Assume that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ is continuous as a function of y for every $\mathbf{x} \in \mathcal{X}$. For each sample $\mathbf{x} \in \mathcal{X}$ and miscoverage level $\alpha \in (0,1)$, let $t^*(\mathbf{x}, \alpha)$ be the cutoff of the oracle band for $f(y|\mathbf{x})$ with coverage $1-\alpha$. Consider the equivalence relation $\mathbf{x}_a \sim \mathbf{x}_b \iff d_g(\mathbf{x}_a, \mathbf{x}_b) = 0$.*

(i) If $\mathbf{x}_a \sim \mathbf{x}_b$*, then* $t^*(\mathbf{x}_a, \alpha) = t^*(\mathbf{x}_b, \alpha)$ *for every* $\alpha \in (0, 1)$

(ii) If $\sim'$ *is any other equivalence relation such that* $\mathbf{x}_a \sim' \mathbf{x}_b$ *implies that* $t^*(\mathbf{x}_a, \alpha) = t^*(\mathbf{x}_b, \alpha)$ *for every* $\alpha \in (0, 1)$*, then* $\mathbf{x}_a \sim' \mathbf{x}_b \Rightarrow \mathbf{x}_a \sim \mathbf{x}_b$.

To conclude we observe that if a conformal method converges to the highest predictive density set, then it satisfies asymptotic conditional coverage (IZBICKI; SHIMIZU; STERN, 2021).

Based on the above motivation, we use `CD-split` with the profile distance. In order to compute the prediction bands, we need to define the centroids $\mathbf{x}_i^c$. Ideally, the partitions should be such that: (i) all sample points inside a given element of the partition have similar profile, and (ii) sample points that belong to different elements of the partition have profiles that are very different from each other. We accomplish this by choosing the partitions by applying a k-means++ clustering algorithm (ARTHUR; VASSILVITSKII, 2007) using the profile distance. This is done by applying the standard (Euclidean) k-means++ algorithm to the data points $\mathbf{w}_i := \tilde{g}(\mathbf{x}_i)$, where $\tilde{g}(\mathbf{x}_i)$ is a discretization of the function $g(\mathbf{x}_i)$, obtained by evaluating $g(\mathbf{x}_i)$ on a grid of values. The points, $\mathbf{w}_1^c, \ldots, \mathbf{w}_J^c$, are the centroids of such clusters. Figure 3 illustrates the partitions that are obtained in one dataset. The profile distance allows samples that are far from each other in the Euclidean sense to fall into the same element of the partition. This is the key reason why our method scales to high-dimensional datasets. Algorithm 2 shows pseudo-code for implementing `CD-split`.

Algorithm 2 `CD-split`

Input: Data $(\mathbf{x}_i, Y_i)$, $i = 1, ..., n$, miscoverage level $\alpha \in (0, 1)$, algorithm $\mathcal{B}$ for fitting conditional density function, number of elements of the partition J.
Output: Prediction band for $\mathbf{x}_{n+1} \in \mathbb{R}^d$

1: Randomly split $\{1, 2, ..., n\}$ into two subsets $\mathbb{D}$ and $\mathbb{D}'$
2: Fit $\hat{f} = \mathcal{B}(\{(\mathbf{x}_i, Y_i) : i \in \mathbb{D}'\})$ **// Estimate density function**
3: Compute $\mathcal{A}$, the partition of $\mathcal{X}$, by applying k-means++ on the profiles of the samples in $\mathbb{D}$'
4: Compute $g_{\mathbf{x}_{n+1}}(t) = \int_{\{y:\hat{f}(y|\mathbf{x})\geq t\}} \hat{f}(y|\mathbf{x})dy$, for all $t > 0$ **// Profile of the density (Definition 4.4)**
5: Find $A(\mathbf{x}_{n+1}) \in \mathcal{A}$, the element of $\mathcal{A}$ such that $\mathbf{x}_{n+1} \in \mathcal{A}$
6: Compute $g_{\mathbf{x}_i}(t) = \int_{\{y:\hat{f}(y|\mathbf{x})\geq t\}} \hat{f}(y|\mathbf{x})dy$, for all $t > 0$ and $i \in \mathbb{D}$ **// Profile of the densities (Definition 4.4)**
7: Let $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \{\hat{f}(y_i|\mathbf{x}_i), i \in \mathbb{D} : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\}$
8: Let $t = q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))$ **// Compute the α- quantile of the set $\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D})$**
9: **return** $\left\{y : \hat{f}(y|\mathbf{x}^*) \geq t\right\}$

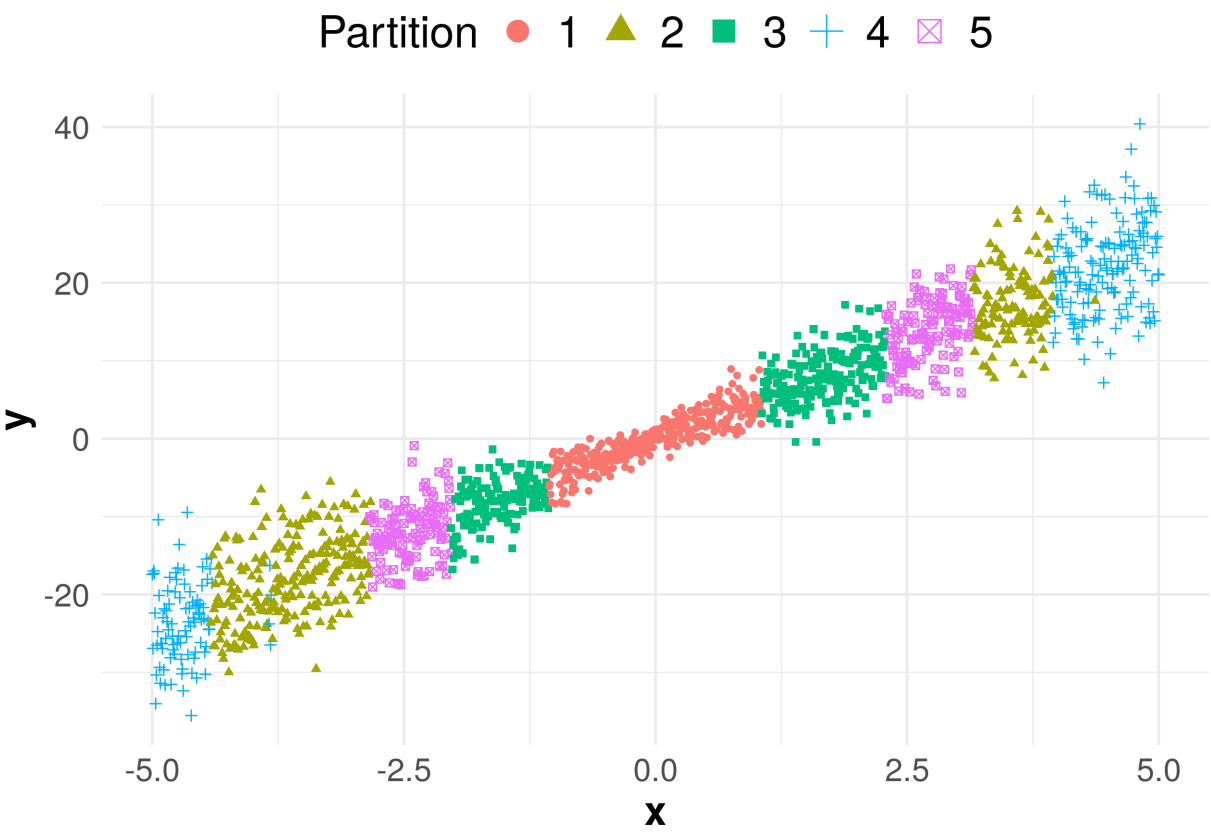


Figure 3 – Scatter plot of data generated according to $Y|x \sim N(5x, 1+|x|)$. Colors indicate partitions that were obtained using the profile of the estimated densities. Note that points that are far from each other on the x-axis can have similar densities and belong to the same element of the partition. This allows larger partition elements while preserving the optimal cutoff (Theorem 4.9).

4.1 Multiclass classification

If the sample space $\mathcal{Y}$ is discrete, we use a similar construction to that of Definition 4.1. More precisely, the `CD-split` prediction band is given by

$$C(\mathbf{x}_{n+1}) = \left\{y : \widehat{\mathbb{P}}(Y = y|\mathbf{x}_{n+1}) \geq q(\alpha; \mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}))\right\}, \text{ where}$$
$$\mathcal{T}(\mathbf{x}_{n+1}, \mathbb{D}) = \left\{\widehat{\mathbb{P}}(Y_i = y_i|\mathbf{x}_i), i = 1, \ldots, n : \mathbf{x}_i \in A(\mathbf{x}_{n+1})\right\},$$

$A(\mathbf{x}_{n+1})$ is the element of $\mathcal{A}$ to which $\mathbf{x}_{n+1}$ belongs to, and

$$d_g^2(\mathbf{x}_a, \mathbf{x}_b) = \sum_{y \in \mathcal{Y}} \left(\widehat{\mathbb{P}}(Y = y|\mathbf{x}_a) - \widehat{\mathbb{P}}(Y = y|\mathbf{x}_b)\right)^2.$$

Theorems analogous to those presented in the last section hold in the classification setting as well.

Remark 2. *While* `CD-split` *controls the coverage of* C *conditional on the value* $\mathbf{x}_{n+1}$*, in a classification setting some methods control* class-specific coverage *(SADINLE; LEI;*

Figure 4 – Prediction bands for some instances of the Fashion-MNIST dataset (XIAO; RASUL; VOLLGRAF, 2017) with $\alpha = 0.01$.

WASSERMAN, 2019), defined as

$$\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | Y_{n+1} = y) \geq 1 - \alpha_y.$$

The Figure 4 shows an example of the CD-split method in a classification setting applied to the traditional Fashion-MNIST dataset (XIAO; RASUL; VOLLGRAF, 2017).

5 EXPERIMENTS

We consider the following settings with $d = 20$ covariates:

- [**Asymmetric**] $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} = 5x_1 + \epsilon$, where $\epsilon \sim \text{Gamma}(1 + 2|x_1|, 1 + 2|x_1|)$.

- [**Bimodal**] $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-1.5, 1.5)$, and $Y|\mathbf{x} \sim 0.5\text{N}(f(\mathbf{x}) - g(\mathbf{x}), \sigma^2(\mathbf{x})) + 0.5\text{N}(f(\mathbf{x}) + g(\mathbf{x}), \sigma^2(\mathbf{x}))$, with $f(\mathbf{x}) = (x_1 - 1)^2(x_1 + 1)$, $g(\mathbf{x}) = 2\mathbb{I}(x_1 \geq -0.5)\sqrt{x_1 + 0.5}$, and $\sigma^2(\mathbf{x}) = 1/4 + |x_1|$. This is the example from Lei and Wasserman (2014) with $d - 1$ additional irrelevant variables.

- [**Heteroscedastic**] $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} \sim \text{N}(x_1, 1 + |x_1|)$.

- [**Homoscedastic**] $\mathbf{X} = (X_1, \ldots, X_d)$, with $X_i \overset{\text{iid}}{\sim} \text{Unif}(-5, 5)$, and $Y|\mathbf{x} \sim \text{N}(x_1, 1)$

We compare the performance of the following methods:

- [**Reg-split**] The regression-split method (LEI *et al.*, 2018), based on the conformal score $|Y_i - \widehat{r}(\mathbf{x}_i)|$, where $\widehat{r}$ is an estimate of the regression function.

- [**Local Reg-split**] The local regression-split method (LEI *et al.*, 2018), based on the conformal score $\frac{|Y_i - \widehat{r}(\mathbf{x}_i)|}{\widehat{\rho}(\mathbf{x}_i)}$, where $\widehat{\rho}$ is an estimate of the conditional mean absolute deviation of $(Y_i - r(\mathbf{x}_i))|\mathbf{x}_i$.

- [**Quantile-split**] The conformal quantile regression method (ROMANO; PATTERSON; CANDÈS, 2019; SESIA; CANDÈS, 2019), based on conformalized quantile regression.

- [**Dist-split**] From chapter 3.

- [**CD-split**] From chapter 4 with $\lceil \frac{n}{100} \rceil$ partitions.

Each experiment is performed with comparable settings. Each experiment uses a coverage level of $1 - \alpha = 90\%$ and is run 5,000 times. Also, random forests (BREIMAN, 2001) are used to estimate all quantities needed in each method, namely: the regression function in Reg-split, the conditional mean absolute deviation in Local Reg-split, the conditional quantiles via quantile forests (MEINSHAUSEN, 2006) in Quantile-split, and the conditional density via FlexCode (IZBICKI; LEE, 2017) in `Dist-split` and `CD-split`. A conditional cumulative distribution estimate, $\widehat{F}(y|\mathbf{x})$ is obtained by integrating the conditional density estimate: $\widehat{F}(y|\mathbf{x}) = \int_{-\infty}^{y} \widehat{f}(y|\mathbf{x})dy$. The tuning parameters of all methods were set to be the default values of the packages that were used.

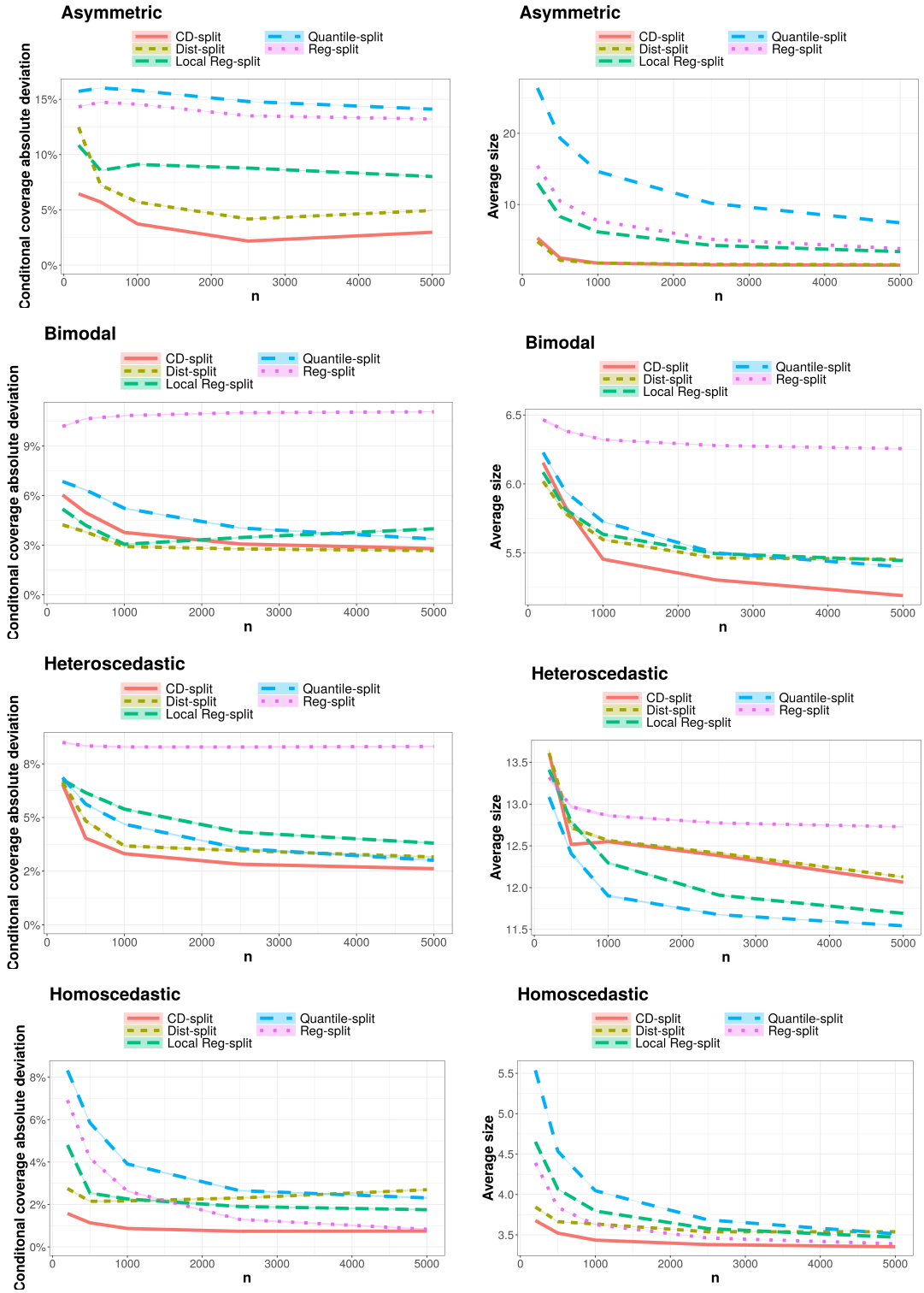


Figure 5 – Performance of each conformal method as a function of the sample size. Left panels show how much the conditional coverage varies with $\mathbf{x}$; right panels display the average size of the prediction bands.

Figure 5 shows the performance of each method as a function of the sample size. While the left side figures display how well each method controls conditional coverage, the right side displays the average size of the regions that are obtained. The control of the conditional coverage is measured through the conditional coverage absolute deviation, that is, $\mathbb{E}[|\mathbb{P}(Y^* \in C(\mathbf{X}^*)|\mathbf{X}^*) - (1-\alpha)|]$. Since all of the methods obtain marginal coverage very close to the nominal 90% level, this information is not displayed in the figure. Figure 5 shows that, in all settings, `CD-split` is the method which best controls conditional coverage. Also, in most cases its prediction bands also have the smallest size. Similarly, `Dist-split` frequently is the second method with both highest control of conditional coverage and also smallest prediction bands.

We also apply `CD-split` to a classification setting. We consider $\mathbf{X} = (X_1, \ldots, X_{20})$, with $X_i \overset{\text{iid}}{\sim} N(0,1)$ and $Y|\mathbf{X}$ follows the logistic model, $\mathbb{P}(Y = i|\mathbf{x}) \propto \exp\{\beta_i \cdot x_1\}$, where $\boldsymbol{\beta} = (-6, -5, -1.5, 0, 1.5, 5, 6)$. We compare `CD-split` to Probability-split, the method described in Sadinle, Lei and Wasserman (2019, Sec. 4.3), which has the goal of controlling global coverage. Probability-split is a particular case of `CD-split`: it corresponds to applying `CD-split` with $J = 1$ partitions. Figure 6 shows the results. `CD-split` better controls conditional coverage. On the other hand, its prediction bands are, on average, larger than those of Probability-split.

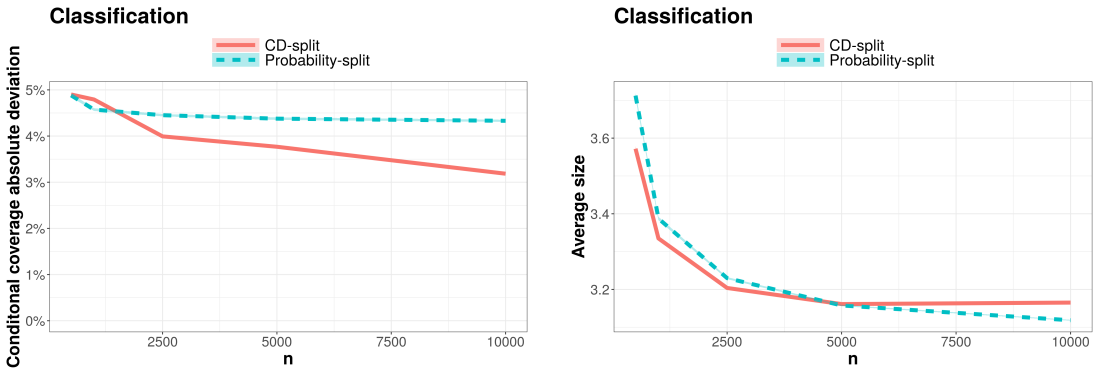


Figure 6 – Performance of each conformal method as a function of the sample size. Left panel shows how much the conditional coverage vary with $\mathbf{x}$; right panel displays the average size of the prediction bands.

6 FINAL REMARKS

We introduce `Dist-split` and `CD-split`, which obtain asymptotic conditional coverage and converge to optimal oracle bands, even in high-dimensional feature spaces. These results do not require assumptions about the dependence between the target variable and the features. Both methods are based on estimating conditional densities. While `Dist-split` necessarily leads to intervals, which are easier to interpret, `CD-split` leads to smaller prediction regions. A simulation study shows that both methods yield smaller prediction bands and better control of conditional coverage than other methods in the literature under a variety of settings. We also show that `CD-split` leads to good results in classification problems.

`CD-split` is based on a novel data-driven metric on the feature space that is appropriate for defining neighborhoods for conformal methods, in particular in high-dimensional settings. It might be possible to use this metric with other conformal methods to obtain asymptotic conditional coverage.

R code for implementing `Dist-split` and `CD-split` is available at https://github.com/rizbicki/predictionBands.

Part II

7 INTRODUCTION

Unsupervised machine learning methods allow the analysis of multivariate data sets in which no response variable is available. This type of analysis is especially useful as the amount of unstructured information grows (in the form of texts, for example), enabling the unveiling of latent structure in the data. In particular, the Latent Dirichlet Allocation (LDA) method is an unsupervised technique that focuses on identifying unobservable groups. This method is different from traditional unsupervised methods such as hard clustering, where sampling units can only be classified into a single group. In LDA, soft clustering is performed, that is, a sample unit can belong to several groups at the same time.

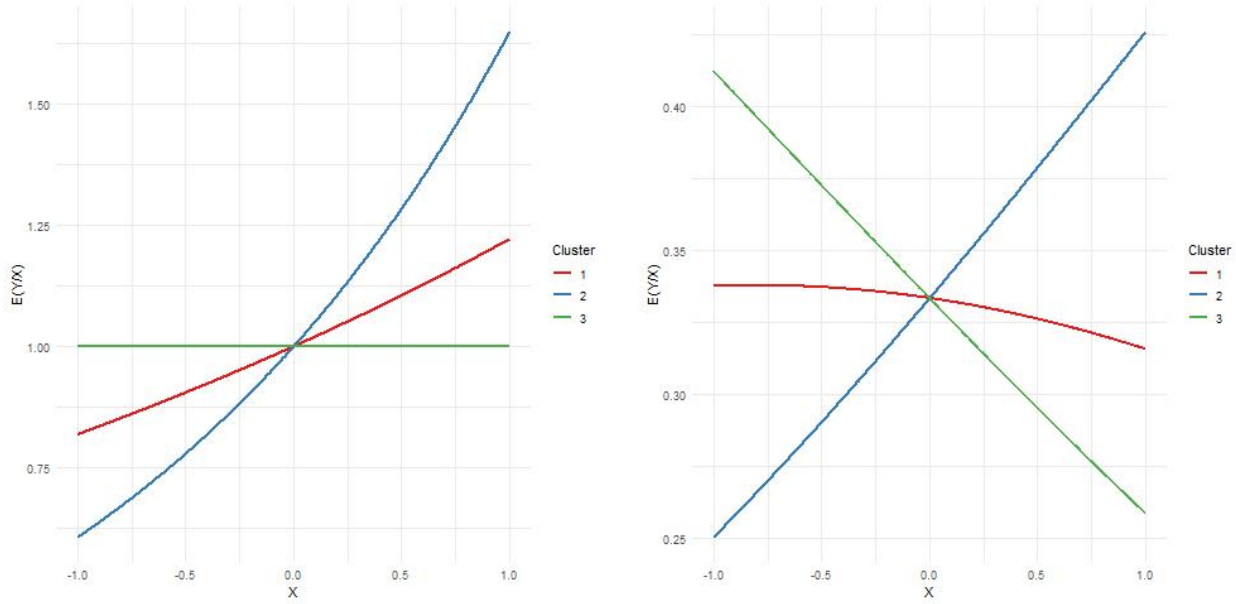
The LDA model was originally proposed by Pritchard, Stephens and Donnelly (2000) in the context of population genetics, but it became popular in the context of machine learning through the work of Blei, Ng and Jordan (2003) on text-mining applications. In these text-mining applications, the goal is to discover topics that are present in each document based on the words that appear on these documents. This model has been applied to several areas of knowledge. For example, Lukins, Kraft and Etzkorn (2010) used LDA to understand software bug reports. In another application, Lienou, Maitre and Datcu (2009) applied this model to create annotation of satellite imagery. LDA was also used by Xing and Girolami (2007) to detect fraudulent calls in the telecommunications industry based on the patterns found for each customer. Finally, Valle *et al.* (2014) used LDA on biodiversity data to describe groups of trees.

Several variations of LDA exist. For instance, Mcauliffe and Blei (2008) introduced the supervised LDA model where documents are labeled with continuous or discrete response variables. Wang and Grimson (2008) considered a spatial structure to group spatially close elements (such as words that are close in the text). Blei and Lafferty (2006) analyzed the evolution of topics over time through a family of probabilistic time series models. Albuquerque, Valle and Li (2019) adapted the LDA model for different types of data (multinomial, binomial and bernoulli) and used a special prior called truncated stick-breaking (TSB) prior to identify the optimal number of groups.

In many problems, one also has access to additional information about instances that comes in the form of features (covariates). For example, a company may have socioeconomic information about its customers, such as age or income, that can help in understanding customer interactions via chat. In these cases, it can be useful to explore the relationship between these covariates and the identified groups. Roberts, Stewart and Airoldi (2016) developed Structural Topic Models (STM), in which covariates were incorporated into LDA through a Multinomial regression model so that the probability of each topic in a

given instance is allowed to depend on covariates. The focus on the probability of each topic, instead of the abundance of each topic, is an important limitation. For example, the type of scientific article (e.g., a commentary or a review article) can significantly change the number of words associated with each topic. However, STM's might fail to identify this effect if the proportion of the different topics remains the same.

In this work, we propose a new formulation to the LDA model where we use covariates to explain the number of elements (e.g., number of words) in each group, rather than the proportion of each group. In a sense, our model is more general than STM's because the probabilities of each group can also be derived from it. Another important advantage of our approach is that the covariate coefficients (i.e., the slope parameters) can be interpreted more easily through the logarithmic link function of the Negative-Binomial regression rather than the logistic link function used within the Multinomial regression in STM's. The log link function allows a straightforward interpretation of the coefficients in the sense that a positive (negative) coefficient describes a positive (negative) relationship between the corresponding covariate and the abundance of each group. On the other hand, as illustrated in Fig. 7, a positive coefficient in the multinomial logistic link function can imply a positive or negative relationship between the proportion of the group and the corresponding covariate.



(a) Negative-binomial log link

(b) Multinomial logistic link

Figure 7 – Illustration of the difference between the logarithmic (left panel) and multinomial logistic (right panel) link functions considering 3 groups and 1 covariate (without intercept) with $\beta_1 = 0.2$, $\beta_2 = 0.5$ and $\beta_3 = 0.0$. Notice that, despite the positive coefficient for group 1, there is a negative relationship in panel (b) between the covariate and the expected value of y given x.

In chapter 8 we present the proposed Bayesian model and the full conditional

distributions required for our Gibbs sampler. In chapter 9 we describe the estimation method and the software used. Then, in chapter 10, we apply our model to simulated data sets to demonstrate its effectiveness in providing inferences on the parameters of interest. In chapter 11, we illustrate the versatility of our model by applying it to data sets from different fields. Chapter 12 compares our model to STMs. We conclude with a discussion of the advantages / disadvantages of the method and suggestions for future research.

8 MODEL

Here we introduce the proposed statistical model. Let L denote the total number of instances, K be the total number of groups/clusters and S be the total number of categories (states) each element from each instance can belong to. For example, in text analysis we can have K topics, a vocabulary of S distinct words and L documents. In this case, the words in each document are the elements. We denote by $n_{l,*,*} := \sum_{s,k} n_{l,s,k}$ the total number of elements at instance l, where $n_{l,s,k}$ is a latent variable representing the total number of elements of category s and cluster k in instance l.

The data that we observe consist of

- $y_{i,l} \in \{1, \dots, S\}$, the category of the i-th element of instance l, $i = 1, \dots, n_{l,*,*}$ and $l = 1, \dots, L$.
- $\mathbf{x}_l$: a d-dimension vector with the features (covariates) associated to instance l, $l = 1, \dots, L$.

The data $y_{i,l}$ are often summarized as an instance-by-category abundance matrix. More specifically, each cell in this matrix contains the total number of elements of category s on instance l, given by $w_{l,s} := \sum_i \mathbb{I}(y_{i,l} = s)$, $l = 1, \dots, L$ and $s = 1, \dots, S$.

In our model, the link between the covariates and the abundance of each cluster in each instance $n_{l,*,k} = \sum_{s=1}^{S} n_{l,s,k}$ is given by a Negative-Binomial regression:

$$n_{l,*,k} \mid \boldsymbol{\beta}_k, N \sim \text{NegBinom}(\lambda_{l,k}, N)$$

where $\boldsymbol{\beta}_k$ is a d-dimension vector, N is the overdispersion parameter and $\lambda_{l,k} := E[n_{l,*,k}] = \exp(\mathbf{x}_l^T \boldsymbol{\beta}_k)$. Notice that $n_{l,*,k}$ are latent variables (and thus are not observed), and therefore we need to estimate the coefficients of the regression function *at the same time* we estimate $n_{l,*,k}$ (see details in Section 8.1).

The model also assumes that

$$(n_{l,1,k}, \dots, n_{l,S,k}) \mid n_{l,*,k}, \boldsymbol{\phi}_k \sim \text{Multinomial}(n_{l,*,k}, \boldsymbol{\phi}_k),$$

where $\boldsymbol{\phi}_k \in \mathbb{R}^S$ is a vector on a simplex that represents the composition of categories inside cluster k. We call Φ the matrix with elements $\phi_{k,s}$. Furthermore, note that within the standard LDA model, a parameter of primary interest is $\theta_{l,k}$, which is the proportion of cluster k in instance l. This parameter can be easily retrieved based on the $n_{l,*,k}$ results from our model by calculating $\theta_{l,k} = \frac{n_{l,*,k}}{\sum_{c=1}^{K} n_{l,*,c}}$. We call Θ the matrix with elements $\theta_{l,k}$.

We use the following prior distributions:

$$N \sim \text{Unif}(0, N_0),$$

$$\boldsymbol{\phi}_k \mid \boldsymbol{\gamma} \sim \text{Dirichlet}(\boldsymbol{\gamma}), \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S),$$

and

$$\boldsymbol{\beta}_k \sim N_d(\mathbf{0}, \mathbf{T}),$$

where $\mathbf{T}$ is a diagonal matrix. The hyper parameters N_0, $\boldsymbol{\gamma}$ and $\mathbf{T}$ are a priori set by the modeler.

The joint density function induced by the likelihood function and prior distributions is given by

$$p(\{n_{l,s,k}\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{\beta}_k\} | \{w_{l,s}\}, \{x_l\}) \propto$$
$$\left[\prod_{l=1}^{L}\prod_{k=1}^{K}\left[\text{Multinomial}([n_{l,1,k}, \ldots, n_{l,S,k}] \mid n_{l,*,k}, \boldsymbol{\phi}_k)\text{NegBinom}(n_{l,*,k} \mid \exp(\mathbf{x}_l^T\boldsymbol{\beta}_k), N)\right]^{\mathbb{I}(w_{l,s}=\sum_{k=1}^{K} n_{l,s,k})}\right] \times$$
$$\times \left[\prod_{k=1}^{K}\text{Dirichlet}(\boldsymbol{\phi}_k \mid \boldsymbol{\gamma})\right]\left[\prod_{k=1}^{K} N_d(\boldsymbol{\beta}_k \mid \mathbf{0}, \mathbf{T})\right]\text{Unif}(N \mid 0, N_0).$$

8.1 Full Conditional Distributions

We can obtain samples from the posterior distribution by using a Gibbs sampler (GEMAN; GEMAN, 1984). In order to do that, we first derive the full conditional distributions for the parameters in our model. First, we derive the conditional distribution of each $\boldsymbol{\phi}_k$ given all the other quantities:

$$\begin{aligned} p(\boldsymbol{\phi}_k \mid \ldots) &\propto \left[\prod_{i=1}^{n_{l,*,*}}\prod_{l=1}^{L}\text{Categorical}(y_{il} \mid \boldsymbol{\phi}_k)^{\mathbb{I}(z_{il}=k)}\right]\text{Dirichlet}(\boldsymbol{\phi}_k \mid \boldsymbol{\gamma}) \\ &\propto \left[\prod_{i=1}^{n_{l,*,*}}\prod_{l=1}^{L}\phi_{k,1}^{\mathbb{I}(y_{il}=1,z_{il}=k)} \times \cdots \times \phi_{k,S}^{\mathbb{I}(y_{il}=S,z_{il}=k)}\right]\phi_{k,1}^{\gamma_1-1} \times \cdots \times \phi_{k,S}^{\gamma_S-1} \\ &\propto \phi_{k,1}^{n_{*,1,k}+\gamma_1-1} \times \cdots \times \phi_{k,S}^{n_{*,S,k}+\gamma_S-1}. \end{aligned}$$

Thus,

$$\boldsymbol{\phi}_k \mid \cdots \sim \text{Dirichlet}\left([n_{*,1,k} + \gamma_1, \ldots, n_{*,S,k} + \gamma_S]\right),$$

which is straightforward to sample from.

The conditional distribution of $\boldsymbol{\beta}_k$ given all of the other quantities is

$$p(\boldsymbol{\beta}_k \mid \ldots) \propto \left[\prod_{l=1}^{L}\text{NegBinom}\left(n_{l,*,k} \mid \exp(\mathbf{x}_l^T\boldsymbol{\beta}_k), N\right)\right] N_d(\boldsymbol{\beta}_k \mid \mathbf{0}, \tau^2 I_d).$$

Because of lack of conjugacy, we rely on a slice-sampler algorithm (see Appendix E) to sample from this FCD.

For the parameter N, we obtain

$$p(N \mid \ldots) \propto \left[\prod_{k=1}^{K}\prod_{l=1}^{L}\text{NegBinom}\left(n_{l,*,k} \mid \exp(\mathbf{x}_l^T\boldsymbol{\beta}_k), N\right)\right] Unif(N|0, N_0).$$

Again, we rely on a slice-sampler algorithm to sample from this FCD.

Finally, we obtain the conditional distribution of $z_{i,l}$ (the latent group membership of the i-th element in the l-th instance):

$$
\begin{aligned}
&p(z_{i,l'} = k \mid y_{i,l'} = s', \dots) \\
&\propto \prod_{l=1}^{L} \prod_{k=1}^{K} [\text{Multinomial}([n_{l,1,k}, \dots, n_{l,S,k}] \mid n_{l,*,k}, \boldsymbol{\phi}_k) \text{NegBinom}(n_{l,*,k} \mid \lambda_{l,k}, N)]
\end{aligned}
$$

After integrating $\boldsymbol{\phi_k}$ out and simplifying this expression (a detailed derivation of these results is provided in Appendix B and C), we obtain:

$$
p(z_{i,l'} = k \mid y_{i,l'} = s', \dots) \propto \frac{(n_{l',*,k} + N)(n_{*,s',k} + \gamma_{s'})}{(n_{l',s',k} + 1)(n_{*,*,k} + \sum_s \gamma_s)} (1 - p_{l',k}).
$$

where $p_{l',k} = \frac{N}{N + \lambda_{l',k}}$. Thus,

$$
z_{i,l} \mid y_{i,l} = s, \cdots \sim \text{Categorical}\left(\left[\frac{\frac{(n_{l,*,1}+N)(n_{*,s,1}+\gamma_s)}{(n_{l,s,1}+1)(n_{*,*,1}+\sum_s \gamma_s)}(1-p_{l,1})}{\sum_{k=1}^{K} \frac{(n_{l,*,k}+N)(n_{*,s,k}+\gamma_s)}{(n_{l,s,k}+1)(n_{*,*,k}+\sum_s \gamma_s)}(1-p_{l,k})}, \dots, \frac{\frac{(n_{l,*,K}+N)(n_{*,s,K}+\gamma_s)}{(n_{l,s,K}+1)(n_{*,*,K}+\sum_s \gamma_s)}(1-p_{l,K})}{\sum_{k=1}^{K} \frac{(n_{l,*,k}+N)(n_{*,s,k}+\gamma_s)}{(n_{l,s,k}+1)(n_{*,*,k}+\sum_s \gamma_s)}(1-p_{l,k})}\right]\right),
$$

which is easy to sample from.

9 ESTIMATION AND SOFTWARE

In order to fit our model, we first need to define K, the number of clusters that will be used. We do this by using the LDA model proposed by Albuquerque, Valle and Li (2019). Although this model does not include covariates, it uses a truncated stick-breaking prior distribution that identifies the optimal number of clusters. We use the following values for the hyperparameters: $N_0 = 1000$, $\gamma_1 = \cdots = \gamma_S = 0.1$ and $\mathbf{T}$ is a diagonal matrix where the diagonal elements are equal to 10.

In our experiments, we use the Gibbs sampler implementation based on the conditional distributions described in Section 8.1 with one exception: the samples from Φ were generated using the model without covariates described in Albuquerque, Valle and Li (2019). We took this approach because preliminary results revealed that this model had difficulty estimating the Φ matrix even in situations where the model without covariates estimated this matrix well. This problem arises because, differently from a standard regression in which the response variable is observed, the response variable here is latent and has to be estimated together with the regression parameters. As a result, a misspecified regression model can negatively impact the latent response variable, potentially mischaracterizing the identified clusters. Our simulated data studies reveal that this two-stage estimation results in good estimates for the parameters of interest.

In each experiment, we assessed convergence by visually evaluating trace-plots of the generated MCMC chains. Part of our algorithm was developed in R while part of the code was made in C++ using the Rcpp package (EDDELBUETTEL *et al.*, 2011). An R package and a tutorial on how to use the model can be found at https://github.com/gilsonshimizu/ldacov.

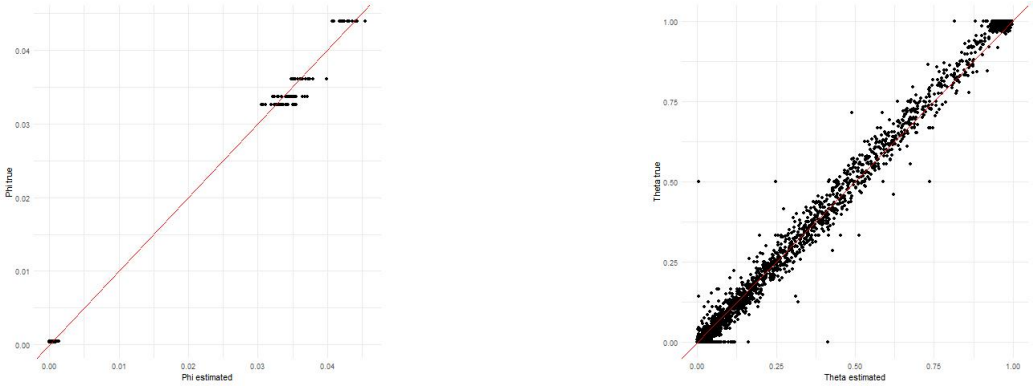
10 SIMULATED EXPERIMENTS

First, we apply our model to two simulated data sets where all parameters are known, enabling the assessment of whether the model is estimating the parameters of interest well.

10.1 Simulation set 1

The first simulated dataset consists of 1000 instances, 100 categories and four clusters. Four covariates are also used, where each covariate explains only one of the four clusters, that is, the matrix with the regression slope coefficients is an identity matrix. We also choose covariate values such that some elements of the Θ are equal to 1 (i.e., some instances have elements from only one cluster). Similarly, we assume that some categories are only present in a single cluster. We do this to help model identifiability.

Figure 8 shows a high correlation between the true and estimated elements of Φ and Θ, indicating that the true parameter values can be recovered from the model when it is estimated using the strategy described in Chapter 9.



(a) Φ (b) Θ

Figure 8 – Scatter plots of true and estimated values of the parameters Φ and Θ for the simulated data set 1 using new LDA formulation.

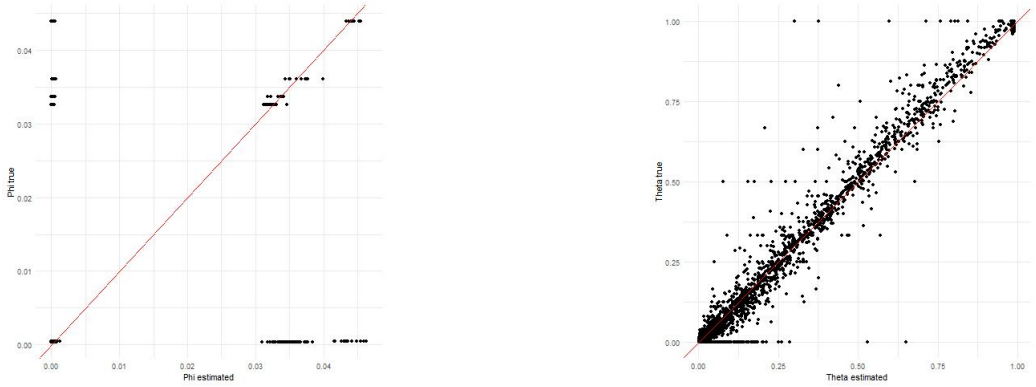
Table 2 shows the posterior means of the regression parameters $\boldsymbol{\beta}_k$, as well as an indicator (*) of whether their respective 99% credible intervals did not contain the value 0. In all cases, the true parameter values are contained in the corresponding credible intervals, demonstrating that our model estimates well the parameters of interest. We will omit here and in the other examples, but trace-plots of the log likelihood and model parameters (see Appendix D) demonstrated the convergence of the algorithm.

Figure 9 shows that the model proposed by Roberts, Stewart and Airoldi (2016) manages to estimate the Θ matrix well but has difficulty in estimating the Φ matrix.

Table 2 – Posterior mean for the regression parameters of the simulated dataset 1.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	True	Estimated	True	Estimated	True	Estimated	True	Estimated
Intercept	1.592	2.131*	1.872	2.055*	1.755	2.127*	1.860	1.827*
Var 1	1.000	0.787*	0.000	−0.055	0.000	−0.044	0.000	−0.027
Var 2	0.000	0.021	0.000	0.008	1.000	0.883*	0.000	0.072
Var 3	0.000	−0.030	1.000	0.870*	0.000	−0.043	0.000	−0.040
Var 4	0.000	−0.018	0.000	0.018	0.000	0.054	1.000	1.015*

* "Statistically significant" results, defined as parameters for which the 99% credible intervals did not overlap zero.



(a) Φ (b) Θ

Figure 9 – Scatter plots of true and estimated values of the parameters Φ and Θ for the simulated data set 1 using STM.

Our method can be used to make predictions for the abundance matrix on the data samples using the information given by the covariates. Figure 10 shows that the method leads to high prediction accuracy on a hold-out set with 1000 instances.

10.2 Simulation set 2

The second set of simulated data is similar to the one described previously. However, instead of using the correct set of covariates, we relied on randomly generated covariates. As result, these covariates were independent of the number of individuals in each cluster. The purpose of this data set is to verify whether the model is able to infer when none of the covariates are relevant.

Figure 11 shows that both Φ and Θ were well estimated.

Table 3 shows the posterior means for the regression parameters. All 99% credible intervals for β's contain the value zero, which are the correct values given that the covariates were independent of the number of elements in each cluster.

Again, the model proposed by Roberts, Stewart and Airoldi (2016) manages to estimate the Θ matrix well but has difficulties in estimating the Φ matrix (Figure 12).

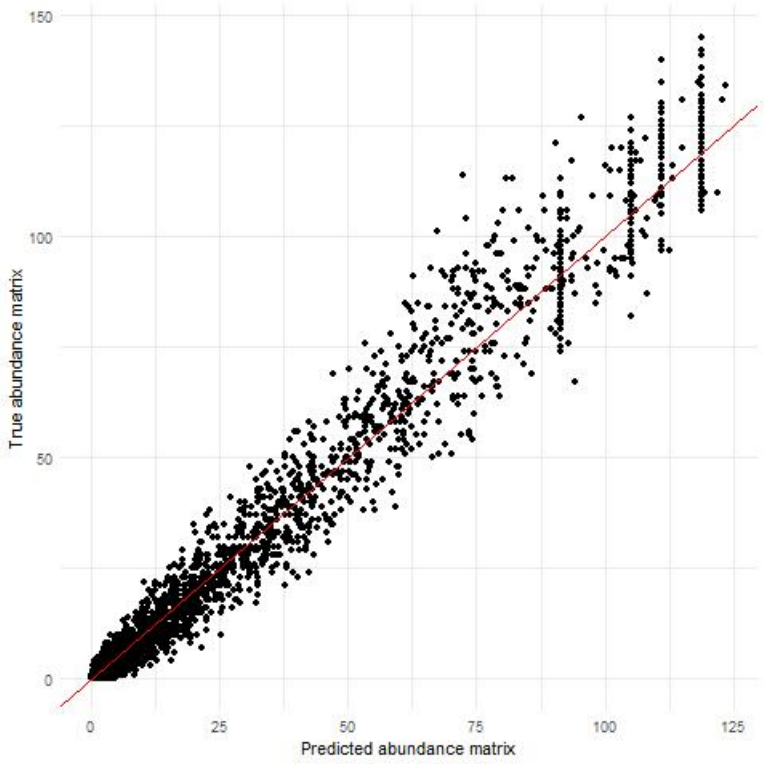
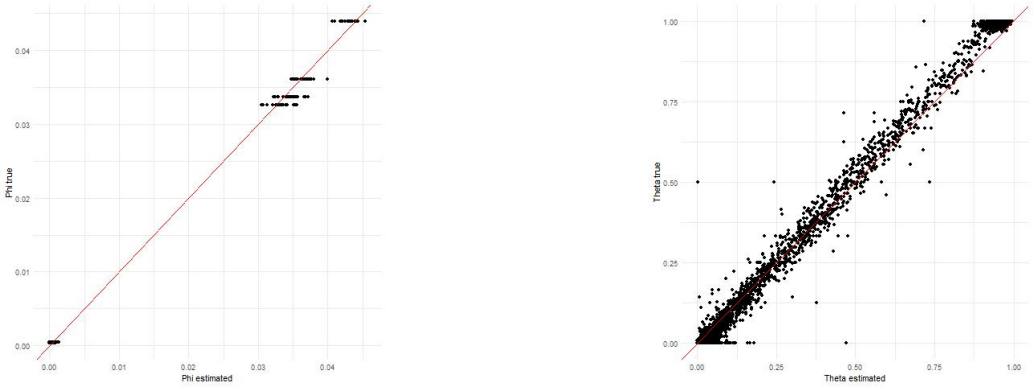


Figure 10 – Scatter plot of the predicted abundance matrix versus true abundance matrix for the simulated data set 1.

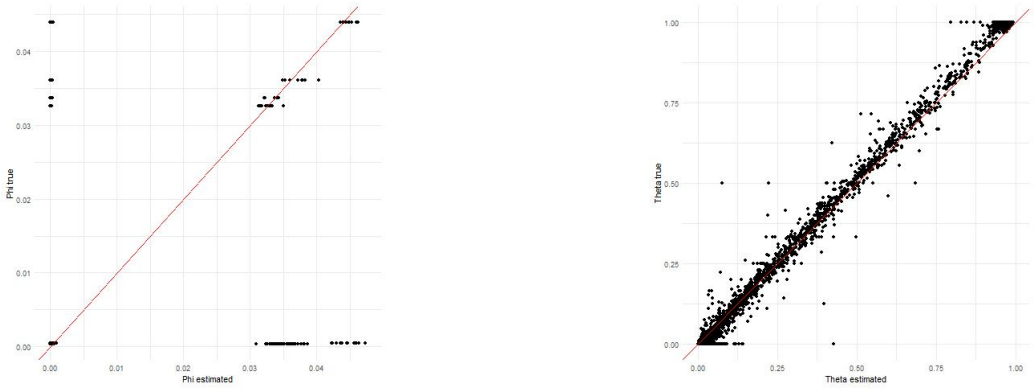


(a) Φ (b) Θ

Figure 11 – Scatter plots of true and estimated values of the parameters Φ and Θ for the simulated data set 2.

Table 3 – Posterior mean for the regression parameters of the simulated dataset 2.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	True	Estimated	True	Estimated	True	Estimated	True	Estimated
Var 1	0.000	0.046	0.000	-0.052	0.000	-0.009	0.000	-0.002
Var 2	0.000	0.018	0.000	-0.059	0.000	0.028	0.000	-0.018
Var 3	0.000	0.110	0.000	0.082	0.000	-0.102	0.000	0.012
Var 4	0.000	0.047	0.000	-0.028	0.000	0.020	0.000	0.022



(a) Φ (b) Θ

Figure 12 – Scatter plots of true and estimated values of the parameters Φ and Θ for the simulated data set 2 using STM.

Taken together, these results reveal that our model is able to estimate the matrices Φ and Θ as well as identify the relevant variables to explain the quantities in each cluster.

11 APPLICATIONS

To demonstrate the model's effectiveness and flexibility, we applied it to three real data sets from different areas:

- [**Covid Articles**] This dataset, available on Kaggle (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge), contains 134,000 articles on Covid-19 and other coronavirus. A sample of size 2,000 was extracted for analysis. We use a bag-of-words representation of the abstract in which we remove stop words, numbers and words that appear in less than 6% of abstracts. In this way, we end up with 211 words as tokens. We use the year in which the paper was published and keywords of the respective journal as covariates. The following keywords were used: virology, chemistry, infectious diseases, microbiology, veterinary, vaccine, immunology, medicine, public health and bioinformatics.

- [**Grocery Shopping**] This dataset is also available in Kaggle (https://www.kaggle.com/karthickveerakumar/orders-data) and contains information about 5,000 customer purchases of 99 products at a supermarket. A sample of size 2,000 was used. We use the day of the week of the purchase and the number of days since the last purchase as covariates.

- [**Barro Colorado Island**] (HARMS *et al.*, 2001) We evaluated the spatial patterns in the tree composition of the moist lowland 50-ha forest dynamic plot (FDP) on the Barro Colorado Island (BCI), Panama. FDP on BCI was established in 1981 and all free-standing woody plants with diameter at breast height (dbh) greater or equal to 1 cm were measured in 1982-93, 1985, 1990, 1995, 2000, 2005, 2010 and 2015. Annual rainfall averages 2600 mm, with a four-month dry season between December and April, while mean annual temperature is 27^{o}C. The total number of species identified at BCI is 326. For our analysis, we only utilized data from the last survey (2015) and we divided the FDP into 200 quadrats of size (50 m $\times$ 50 m); this was deemed the most appropriate scale to identify the spatial structure in biodiversity in BCI. We then aggregated the 2015 BCI census data by calculating the abundance of each species at each of the 200 quadrants. Before analyzing these data, we removed species that were extremely rare (defined as those species with less than 10 trees across the entire 50-ha plot). Our criteria resulted in the removal of 70 (21%) species, representing less than 0.1% of the total number of trees in our dataset.

11.1 Covid Articles

The number of monthly articles about coronavirus practically doubled in the first quarter of 2020 in relation to the monthly average of publications in 2019. Given this significant increase and relevance of the subject, our goal here is to find and understand the differences between possible clusters of articles.

In this text-mining application, we follow the literature in referring to topics instead of clusters. We set the maximum number of topics to 10 and use the TSB prior model proposed by Albuquerque, Valle and Li (2019) to identify the optimal number of topics. This analysis identifies that the optimal number of topics was equal to 5 for this dataset. Tables 4 and 5 present the relevant words for each topic and the estimates of the regression parameters, respectively.

Topics 2 and 3 are more related to Covid-19 and, as expected, are strongly associated with diagnostic tests, symptoms, public health and prevention since, at the time this dataset was retrieved (april/2020), there were still no vaccines or in-depth genetic studies on Covid-19. Topics 1 and 4 are related to older articles and are focused on types of studies that had not been conducted for Covid-19 at the time these data were gathered: vaccines, animal tests, and genetic studies. Topic 5 is focused on other viruses.

Although it is natural that there is a strong relationship between keywords and topics, we emphasize that the use of these keywords as covariates allowed an easier interpretation of topics than just analyzing the relevant words of each topic.

11.2 Grocery Shopping

Our goal is to find clusters of grocery shopping baskets while also identifying how these clusters are associate with day of week and days since last purchase. Although this type of data is not usually analyzed with LDA, we believe that LDA could be useful in identifying the main types of shopping baskets that are made and how these purchases are influenced by covariates.

After running the model without covariates, we obtain an ideal number of clusters equal to 4. Tables 6 and 7 show the relevant products of each cluster and the estimates of the regression parameters, respectively.

Below we describe the clusters that were found:

- Cluster 1: This cluster has many products with purchases made on any day of the week and with varying frequencies. The products are of daily use like breads, cereals, coffee and also cleaning products.

- Cluster 2: This cluster contains herbs, spices, vegetables, poultry, etc. This type of

Table 4 – Relevant words in topics of the Covid dataset.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
protein	health	patients	mice	influenza
rna	public	symptoms	vaccine	viruses
expression	china	positive	responses	human
mechanism	countries	lower	levels	virus
sequence	outbreak	acute	evaluated	assay
replication	epidemic	age	groups	strains
target	prevention	collected	animals	highly
genes	emerging	samples	group	detection
species	diseases	confirmed	induced	
host	transmission	without	response	
genome	spread	common	immune	
mechanisms	research	respectively	increased	
molecular	future	tested	antibodies	
antiviral	information	performed	significantly	
small	population	severe	effects	
shown	infectious	patient	higher	
revealed	use	associated	caused	
function	strategies	hospital	type	
furthermore	will	detected	observed	
involved	current	total	compared	
previously	effective	clinical	effect	
cell	since	syndrome	significant	
specific	care	among		
thus	number	diagnosis		
	evidence	sars		
	review	rate		
	pathogen			
	available			
	data			
	new			
	risk			

purchase might be associated with the preparation of a special meal. This purchase is usually made on a Saturday.

- Cluster 3: This cluster contains frozen meals and prepared soups.
- Cluster 4: This cluster is very peculiar with baby formulas, beers and wines. This cluster is reflective of the classic example of basket analysis, where parents go to buy baby diapers and take the opportunity to buy beer and wine. These purchases are made less frequently than other clusters and are generally not made on Saturdays.

We note here that a customer can buy more than one basket (cluster) at the same time, which is a very interesting feature of the LDA models (compared to traditional

Table 5 – Estimated Regression parameters of the covid dataset.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Intercept	2.876^{*}	2.078^{*}	1.510^{*}	0.041	-0.256^{*}
Year 2020	-0.658^{*}	0.468^{*}	0.475^{*}	-1.044^{*}	-3.227^{*}
Virology	0.462^{*}	-0.823^{*}	−0.168	0.549^{*}	0.835^{*}
Chemistry	0.093	-1.055^{*}	-0.688^{*}	-3.455^{*}	−0.166
Infect disease	-0.468^{*}	0.410^{*}	0.910^{*}	−0.281	0.814^{*}
Microbiology	0.250^{*}	-0.302^{*}	−0.175	−0.031	0.525^{*}
Veterinary	−0.065	0.169	0.206	0.535^{*}	-0.607^{*}
Vaccine	0.054	−0.109	−0.178	1.602^{*}	0.697^{*}
Immunology	0.135	-0.643^{*}	-0.312^{*}	0.825^{*}	-0.931^{*}
Medicine	-0.436^{*}	0.501^{*}	0.507^{*}	-0.589^{*}	0.638^{*}
Public Health	-0.969^{*}	1.066^{*}	0.137	−0.223	-0.722^{*}
Comput bioinformatic	0.049^{*}	0.586^{*}	-3.765^{*}	-3.037^{*}	−0.749

* “Statistically significant” results, defined as parameters for which the 95% credible intervals did not overlap zero.

Table 6 – Relevant products in clusters of the grocery dataset.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
cereal	fresh herbs	frozen vegan vegetarian	red wines
ice cream ice	canned jarred vegetables	frozen meals	baby food formula
water seltzer sparkling water	spices seasonings	frozen breakfast	beers coolers
candy chocolate	fresh vegetables	tofu meat alternatives	
refrigerated	poultry counter	frozen pizza	
frozen appetizers sides	asian foods	fresh dips tapenades	
tea	grains rice dried goods	energy granola bars	
packaged produce	canned meals beans	prepared soups salads	
laundry	oils vinegars		
paper goods	specialty cheeses		
chips pretzels	pickled goods olives		
coffee	dry pasta		
frozen meat seafood	packaged poultry		
bread			
cleaning products			
lunch meat			
spreads			
soap			
dish detergents			
soft drinks			
...			

Table 7 – Estimated regression parameters of the grocery dataset.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Intercept	1.864^{*}	0.562^{*}	-1.800^{*}	-3.386^{*}
Last order >= 6 days	0.219	0.051	−0.070	0.974^{*}
Saturday	0.152	0.376^{*}	0.330	-0.671^{*}

* "Statistically significant" results, defined as parameters for which the 95% credible intervals did not overlap zero.

Table 8 – Estimated regression parameters of the BCI dataset.

	Groups										
Covariates	1	2	3	4	5	6	7	8	9	10	11
Intercept	5.474^{*}	5.524^{*}	5.238^{*}	4.926^{*}	4.678^{*}	4.974^{*}	4.746^{*}	4.829^{*}	4.529^{*}	4.541^{*}	4.645^{*}
Elevation	0.107	0.13	-0.337^{*}	−0.114	0.296^{*}	0.016	0.008	0.227^{*}	0.054	0.203^{*}	0.009
Slope	−0.060	−0.093	-0.239^{*}	0.028	0.125	0.156^{*}	0.274^{*}	0.032	0.143^{*}	0.337^{*}	−0.073
Convexity	−0.088	−0.03	0.138^{*}	0.094^{*}	−0.043	0.086	−0.038	-0.129^{*}	0.101^{*}	-0.153^{*}	-0.131^{*}
Al	−0.091	0.06	0.129	−0.009	−0.118	−0.090	−0.095	−0.129	0.021	-0.187^{*}	0.043
Mn	0.186^{*}	−0.027	-0.150^{*}	0.021	−0.047	0.084	0.303^{*}	0.140^{*}	−0.137	0.168^{*}	-0.245^{*}
Zn	−0.094	−0.016	0.013	−0.116	0.232^{*}	-0.217^{*}	-0.259^{*}	-0.206^{*}	0.451^{*}	−0.041	0.105
N	0.039	0.005	0.071	0.030	−0.132	-0.145^{*}	0.114	0.000	−0.011	-0.226^{*}	−0.029
pH	−0.126	0.080	0.009	−0.034	0.194	−0.099	0.022	-0.277^{*}	−0.136	−0.002	−0.071

* "Statistically significant" results, defined as parameters for which the 95% credible intervals did not overlap zero.

cluster analysis) and more realistic for this data set. In a conventional cluster analysis a customer would be classified into just one cluster.

11.3 Barro Colorado Island

We ran the LDA model without covariates and, from a total of 20 potential groups, found 11 dominant groups that together comprised approximately 91% of all individuals. We find relatively strong spatial patterns in the distribution of these groups (13). For instance, group 10 is clearly restricted to the areas with steep slopes while group 3 has much higher abundance in flat areas. Interestingly, several of the groups identified here seems to closely correspond to the BCI habitat classification proposed by Harms *et al.* (2001). For example, group 11 seems to match the "old forest, swamp" class while group 10 seems to match the "Old forest, Streamside". However, different from this discrete classification of habitats, we find spatial patterns that reflect substantial mixed membership.

Similar results could have been obtained from the LDA model without covariates. The novelty of the proposed model is the ability to make formal inference on the effect of covariates (8). We find that all groups, except for group 2, were strongly associated with one or more covariates. For example, as expected, groups 10 and 3 were positively and negatively associated with slope, respectively. The variables that tended to influence a large number of groups were slope, convexity, magnesium and zinc.

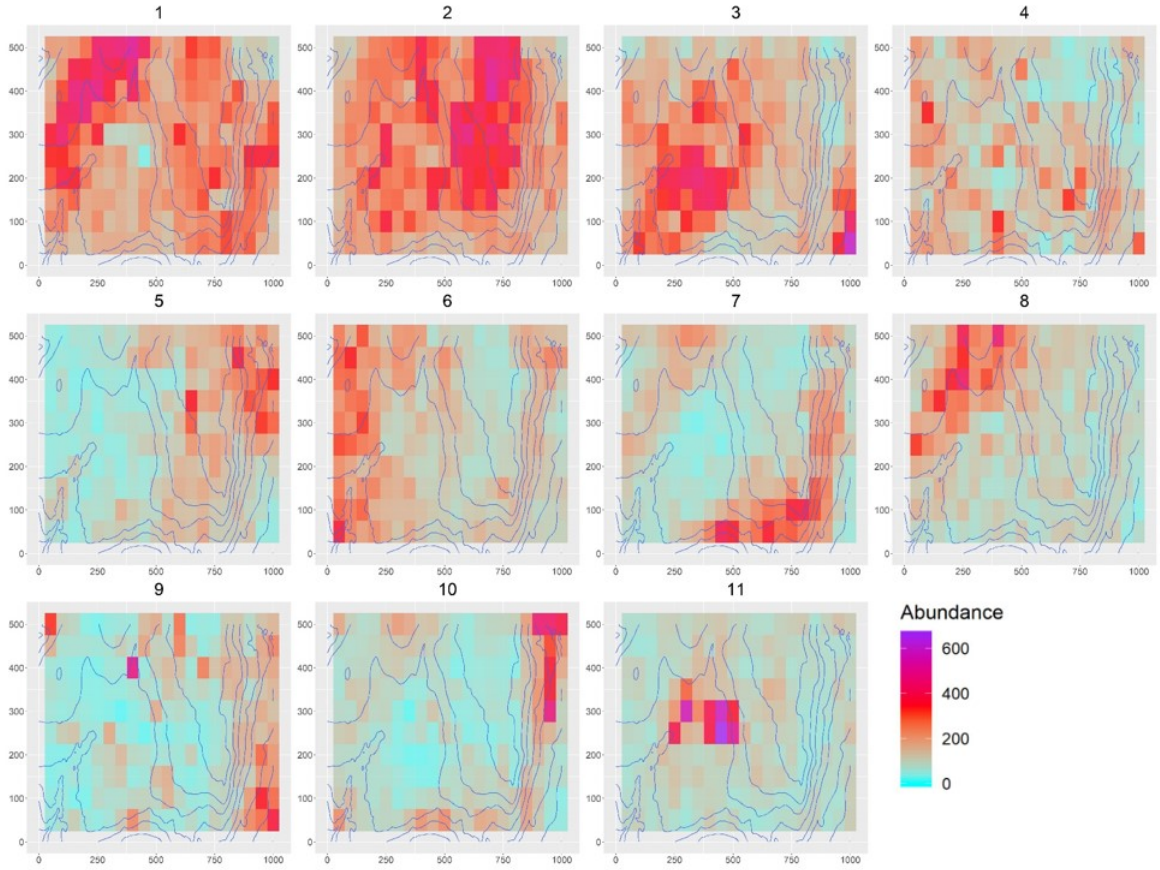


Figure 13 – Spatial distribution of the groups identified by our model. Each panel displays the results for a given group. Hotter colors indicate higher abundance. Elevation is shown with level curves, shown at 5-m intervals.

12 MODEL COMPARISON USING PROBABILISTIC COHERENCE

Next, we compare the results of our model with the STM model Roberts, Stewart and Airoldi (2016). In order to do that, we used the textmineR package (JONES, 2019) in R. In particular, we use *probabilistic coherence* to compare the estimated Φ matrices for each method. In the context of text mining, the probabilistic coherence calculates for each pair of words the measure $P(s_1|s_2) - P(s_1)$, where the word s_1 is more likely than the word s_2 in the focus topic. As a result, probabilistic coherence measures how strongly associated are words s_1 and s_2. A well delineated topic with a high frequency of these words should have a high probabilistic coherence. We consider only the 5 most frequent words in the topics and use the sum of the probabilistic coherence of all topics as a measure of the quality of the estimated Φ matrix.

Table 9 shows these measures for all datasets analyzed in this work. In all cases, the proposed model obtained better probabilistic coherence measures than the STM model, indicating that the topics that are found with our method are more coherent.

Table 9 – Probabilistic coherence for all datasets comparing LDA with covariates and STM. Best values are in bold.

	Dataset				
Method	Simulation set 1	Simulation set 2	Covid Articles	Grocery Shopping	BCI
LDA Covariates	**1.554**	**1.550**	**0.713**	**0.370**	**0.021**
STM	0.395	0.455	0.592	0.235	0.011

13 DISCUSSION

We propose a new formulation for the LDA model that allows the incorporation of covariates. This model differs from other LDA methods because it models how covariates affects the number of elements of each cluster rather than the proportions of the clusters. Because these proportions can be derived from the number of elements, our model generalizes existing LDA models that also incorporate covariates.

The main advantage of our model is that it enables a much more straight-forward interpretation of the regression coefficients. This is due to the use of the logarithmic link function on the quantities in each cluster instead of a multinomial logistic function on the proportions. Furthermore, by more faithfully representing uncertainty, it is possible that the inference on the regression coefficients is better with our Gibbs sampler algorithm than when using approximate variational estimation methods.

In our simulated examples, we are able to show that our model estimates well the Φ and Θ matrices with or without relevant covariates. We also illustrate the model's ability to make inferences about regression coefficients through credible intervals. Importantly, our examples with real data sets demonstrate the flexibility of the model to be applied in different areas and for different types of data.

The dataset about Covid articles, for example, is a traditional text mining data set. The use of covariates, together with the main words of each topic, enabled us to determine how the focus of these articles has changed as the new coronavirus spurs a pandemic across the world. The data set on supermarket purchases is not commonly analyzed with this type of model. Nevertheless, our model was able to create clusters and relate them to the time and day of the week covariates. We believe that this tool is very useful to segment customers and also to optimize the layout of products within a grocery store. Finally, when applied to the BCI dataset, our model was able to find clusters with distinct spatial patterns and at the same time relate these patterns to some of the soil and topography features.

A disadvantage of our method is the computational cost (especially in large datasets with many sample units and categories) compared to models using variational Bayes methods. A possible future work would be to consider a version of this model with a variational inference approach.

REFERENCES

ALBUQUERQUE, P. H.; VALLE, D. R. do; LI, D. Bayesian lda for mixed-membership clustering analysis: The rlda package. **Knowledge-Based Systems**, Elsevier, v. 163, p. 988–995, 2019.

ARTHUR, D.; VASSILVITSKII, S. k-means++: The advantages of careful seeding. *In*: SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS. **Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms**. [*S.l.: s.n.*], 2007. p. 1027–1035.

BARBER, R. F. *et al.* The limits of distribution-free conditional predictive inference. **arXiv preprint arXiv:1903.04684**, 2019.

BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. *In*: **Proceedings of the 23rd international conference on Machine learning**. [*S.l.: s.n.*], 2006. p. 113–120.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

DALMASSO, N. *et al.* Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. **arXiv preprint arXiv:1908.11523**, 2019.

DAMLEN, P.; WAKEFIELD, J.; WALKER, S. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 61, n. 2, p. 331–344, 1999.

EDDELBUETTEL, D. *et al.* Rcpp: Seamless r and c++ integration. **Journal of Statistical Software**, v. 40, n. 8, p. 1–18, 2011.

FREEMAN, P. E.; IZBICKI, R.; LEE, A. B. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 468, n. 4, p. 4556–4565, 2017.

FRIEDMAN, J. *et al.* **The elements of statistical learning**. [*S.l.: s.n.*]: Springer series in statistics New York, 2001. v. 1.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 6, p. 721–741, 1984.

GUAN, L. Conformal prediction with localization. **arXiv preprint arXiv:1908.08558**, 2019.

HARMS, K. E. *et al.* Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. **Journal of Ecology**, Wiley Online Library, v. 89, n. 6, p. 947–959, 2001.

IZBICKI, R.; LEE, A. B. Nonparametric conditional density estimation in a high-dimensional regression setting. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 25, n. 4, p. 1297–1316, 2016.

IZBICKI, R.; LEE, A. B. Converting high-dimensional regression to high-dimensional conditional density estimation. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 11, n. 2, p. 2800–2831, 2017.

IZBICKI, R.; SANTOS, T. M. dos. Machine learning sob a ótica estatística. **Ufscar/Insper**, 2018.

IZBICKI, R.; SHIMIZU, G.; STERN, R. B. **CD-split and HPD-split: efficient conformal regions in high dimensions**. 2021.

JAMES, G. *et al.* **An introduction to statistical learning**. [*S.l.: s.n.*]: Springer, 2013. v. 112.

JONES, T. **textmineR: Functions for Text Mining and Topic Modeling**. [*S.l.*], 2019. R package version 3.0.4. Available at: https://CRAN.R-project.org/package=textmineR.

LEI, J. *et al.* Distribution-free predictive inference for regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 523, p. 1094–1111, 2018.

LEI, J.; WASSERMAN, L. Distribution-free prediction bands for non-parametric regression. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 76, n. 1, p. 71–96, 2014.

LIENOU, M.; MAITRE, H.; DATCU, M. Semantic annotation of satellite images using latent dirichlet allocation. **IEEE Geoscience and Remote Sensing Letters**, IEEE, v. 7, n. 1, p. 28–32, 2009.

LUECKMANN, J.-M. *et al.* Flexible statistical inference for mechanistic models of neural dynamics. *In*: **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*], 2017. p. 1289–1299.

LUKINS, S. K.; KRAFT, N. A.; ETZKORN, L. H. Bug localization using latent dirichlet allocation. **Information and Software Technology**, Elsevier, v. 52, n. 9, p. 972–990, 2010.

MCAULIFFE, J. D.; BLEI, D. M. Supervised topic models. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2008. p. 121–128.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, n. Jun, p. 983–999, 2006.

NETER, J. *et al.* **Applied linear statistical models**. [*S.l.: s.n.*]: Irwin Chicago, 1996. v. 4.

PAPADOPOULOS, H. Inductive conformal prediction: Theory and application to neural networks. *In*: **Tools in artificial intelligence**. [*S.l.: s.n.*]: IntechOpen, 2008.

PAPAMAKARIOS, G.; PAVLAKOU, T.; MURRAY, I. Masked autoregressive flow for density estimation. *In*: **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*], 2017. p. 2338–2347.

PARMIGIANI, G.; INOUE, L. **Decision theory: Principles and approaches**. [*S.l.: s.n.*]: John Wiley & Sons, 2009. v. 812.

POSPISIL, T.; LEE, A. B. (f) rfcde: Random forests for conditional density estimation and functional data. **arXiv preprint arXiv:1906.07177**, 2019.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Genetics Soc America, v. 155, n. 2, p. 945–959, 2000.

ROBERTS, M. E.; STEWART, B. M.; AIROLDI, E. M. A model of text for experimentation in the social sciences. **Journal of the American Statistical Association**, Taylor & Francis, v. 111, n. 515, p. 988–1003, 2016.

ROMANO, Y.; PATTERSON, E.; CANDÈS, E. J. **Conformalized Quantile Regression**. 2019.

SADINLE, M.; LEI, J.; WASSERMAN, L. Least ambiguous set-valued classifiers with bounded error levels. **Journal of the American Statistical Association**, Taylor & Francis, v. 114, n. 525, p. 223–234, 2019.

SESIA, M.; CANDÈS, E. J. A comparison of some conformal quantile regression methods. **arXiv preprint arXiv:1909.05433**, 2019.

VALLE, D. *et al.* Decomposing biodiversity data using the latent dirichlet allocation model, a probabilistic multivariate statistical method. **Ecology letters**, Wiley Online Library, v. 17, n. 12, p. 1591–1601, 2014.

VOVK, V. Conditional validity of inductive conformal predictors. *In*: **Asian conference on machine learning**. [*S.l.: s.n.*], 2012. p. 475–490.

VOVK, V. *et al.* On-line predictive linear regression. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 37, n. 3, p. 1566–1590, 2009.

VOVK, V. *et al.* **Algorithmic learning in a random world**. [*S.l.: s.n.*]: Springer Science & Business Media, 2005.

WANG, X.; GRIMSON, E. Spatial latent dirichlet allocation. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2008. p. 1577–1584.

XIAO, H.; RASUL, K.; VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. **arXiv preprint arXiv:1708.07747**, 2017.

XING, D.; GIROLAMI, M. Employing latent dirichlet allocation for fraud detection in telecommunications. **Pattern Recognition Letters**, Elsevier, v. 28, n. 13, p. 1727–1734, 2007.

APPENDIX

APPENDIX A – PROOFS FROM PART I

Definition A.1. Whenever $\widehat{F}$ is a cdf, $\widehat{F}^{-1}$ refers to the generalized inverse of $\widehat{F}$.

Definition A.2. $U_{\lfloor \alpha \rfloor}$ and $U_{\lceil \alpha \rceil}$ are the $n^{-1}\lfloor (n\alpha) \rfloor$ and $n^{-1}\lceil (n\alpha) \rceil$ empirical quantiles of $U_1, \ldots, U_n$,

Related to `Dist-split`

Proof of Theorem 3.2. Let $U_i = \widehat{F}(Y_i|\mathbf{X}_i)$. Since $(\mathbf{X}_i, Y_i)$ are i.i.d. continuous random variables and $\widehat{F}$ is continuous, obtain that U_i are i.i.d. continuous random variables. Note that by exchangeability, the rank of U_{n+1} among $\{U_1, U_2, \ldots, U_{n+1}\}$ is uniformly distributed over the set $\{1, 2, \ldots, n+1\}$. Using the cumulative distribution of the discrete uniform and its symmetry property then:

$$1 - \alpha \leq \mathbb{P}\left(U_{n+1} \in [U_{\lfloor 0.5\alpha \rfloor}; U_{\lceil 1-0.5\alpha \rceil}]\right) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \leq 1 - \alpha + (n+1)^{-1}.$$

The conclusion follows from noticing that

$$\begin{aligned} &\mathbb{P}\left(U_{n+1} \in [U_{\lfloor 0.5\alpha \rfloor}; U_{\lceil 1-0.5\alpha \rceil}]\right) \\ =&\mathbb{P}\left(Y_{n+1} \in [\widehat{F}^{-1}(U_{\lfloor 0.5\alpha \rfloor}|\mathbf{X}_{n+1}); \widehat{F}^{-1}(U_{\lceil 1-0.5\alpha \rceil}|\mathbf{X}_{n+1})]\right) \\ =&\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1})) \end{aligned}$$

□

Lemma A.3. *Let* $I_1 = \left\{i \leq n : |\widehat{F}(Y_i|\mathbf{X}_i) - F(Y_i|\mathbf{X}_i)| < \eta_n^{1/3}\right\}$ *and* $I_2 = \{1, \ldots, n\} - I_1$. *Under Assumption 3.3,* $|I_2| = o_P(n)$ *and* $|I_1| = n + o_P(n)$.

Proof. Let $A_n = \left\{\mathbb{E}\left[\sup_{y \in \mathcal{Y}} \left(\widehat{F}(y|\mathbf{X}) - F(y|\mathbf{X})\right)^2 \Big| \widehat{F}\right] \geq \eta_n\right\}$
and $B_n = \left\{|\widehat{F}(Y|\mathbf{X}) - F(Y|\mathbf{X})| \geq \eta_n^{1/3}\right\}$. Using Markov's inequality then

$$\begin{aligned} \mathbb{P}(B_n) &= \mathbb{E}[\mathbb{I}(B_n)] = \mathbb{E}[\mathbb{E}[\mathbb{I}(B_n)|\widehat{F}]] = \mathbb{E}[\mathbb{P}(\mathbb{I}(B_n)|\widehat{F})] \\ &= \mathbb{E}[\mathbb{P}(B_n|\widehat{F})\mathbb{I}(A_n)] + \mathbb{E}[\mathbb{P}(B_n|\widehat{F})\mathbb{I}(A_n^c)] \\ &\leq \mathbb{P}(A_n) + \mathbb{E}\left[\frac{\mathbb{E}[(\widehat{F}(Y|\mathbf{X}) - F(Y|\mathbf{X}))^2|\widehat{F}]}{\eta_n^{2/3}}\mathbb{I}(A_n^c)\right] \\ &\leq \rho_n + \eta_n^{1/3} = o(1) \end{aligned}$$

Note that $|I_2| \sim \text{Binomial}(n, \mathbb{P}(B_n))$. Since $\mathbb{P}(B_n) = o(1)$, conclude that $|I_2| = o_P(n)$. That is, $|I_1| = n + o_P(n)$. □

Lemma A.4. *Under Assumption 3.3, If* $U_i = \widehat{F}(Y_i|\mathbf{X}_i)$*, then for every* $\alpha \in (0,1)$*,* $U_{\lfloor \alpha \rfloor} = \alpha + o_P(1) = U_{\lceil \alpha \rceil}$.

Proof. Let I_1 and I_2 be such as in lemma A.3. Also, let $\hat{G}_1^{-1}$, G_1^{-1} and G_0^{-1} be, the empirical quantiles of, respectively, $\{U_i : i \in I_1\}$, $\{F(Y_i|\mathbf{X}_i) : i \in I_1\}$, and $\{F(Y_i|\mathbf{X}_i) : i \leq n\}$. By definition of I_1, for every $\alpha^* \in [0,1]$, $\hat{G}_1^{-1}(\alpha^*) = G_1^{-1}(\alpha^*) + o(1)$. Also, $G_0^{-1}(\alpha^*) = \alpha^* + o_P(1)$. Therefore, since

$$G_0^{-1}\left(\frac{|I_1|\alpha^*}{n}\right) \leq G_1^{-1}(\alpha^*) \leq G_0^{-1}\left(\frac{|I_1|\alpha^* + |I_2|}{n}\right),$$

conclude that $\hat{G}_1^{-1}(\alpha^*) = \alpha^* + o_P(1)$. Finally, since

$$\hat{G}_1^{-1}\left(\frac{n\alpha - |I_2|}{|I_1|}\right) \leq U_{\lfloor \alpha \rfloor} \leq U_{\lceil \alpha \rceil} \leq \hat{G}_1^{-1}\left(\frac{n\alpha}{|I_1|}\right),$$

Conclude that $U_{\lfloor \alpha \rfloor} = \alpha + o_P(1) = U_{\lceil \alpha \rceil}$. □

Lemma A.5. *Let* $U_i = \hat{F}(Y_i|\mathbf{X}_i)$*. Under Assumptions 3.3 and 3.4,*

$$\widehat{F}^{-1}(U_{[0.5\alpha]}|\mathbf{X}_{n+1}) = F^{-1}(0.5\alpha|\mathbf{X}_{n+1}) + o_P(1)$$
$$\widehat{F}^{-1}(U_{[1-0.5\alpha]}|\mathbf{X}_{n+1}) = F^{-1}(1-0.5\alpha|\mathbf{X}_{n+1}) + o_P(1)$$

Proof. In order to prove the first equality, it is enough to show that $F^{-1}(U_{[0.5\alpha]}|\mathbf{X}_{n+1}) = F^{-1}(0.5\alpha|\mathbf{X}_{n+1}) + o_P(1)$ and that $\widehat{F}^{-1}(U_{[0.5\alpha]}|\mathbf{X}_{n+1}) = F^{-1}(U_{[0.5\alpha]}|\mathbf{X}_{n+1}) + o_P(1)$. The first part follows from lemma A.4 and the continuity of $F(y|\mathbf{x})$ (Assumption 3.4). For the second part, note that, if $\sup_y |\hat{F}(y|\mathbf{x}) - F(y|\mathbf{x})| < \eta_n$ and using the mean value theorem, then, for every α^*, $|\hat{F}^{-1}(\alpha^*) - F^{-1}(\alpha^*)| \leq \eta_n \left(\inf_y \frac{dF(y|\mathbf{x})}{dy}\right)^{-1}$. Using this observation, the proof of the second part follows from Assumption 3.4, and observing that $U_{[0.5\alpha]} = 0.5\alpha + o_P(1)$ (lemma A.4) and $\mathbb{P}(\sup_y |\hat{F}(y|\mathbf{x}) - F(y|\mathbf{x})| \geq \eta_n) = o(1)$ (Assumption 3.3).

The proof for the $1 - .5\alpha$ quantile is analogous to the one for the $.5\alpha$ quantile. □

Proof of theorem 3.5. Solving $\frac{\partial \mathbb{E}[L((a,b),Y_{n+1})]}{\partial a} = \frac{\partial \mathbb{E}[L((a,b),Y_{n+1})]}{\partial b} = 0$ we have that $C^*(\mathbf{X}_{n+1}) = [F^{-1}(0.5\alpha|\mathbf{X}_{n+1}); F^{-1}(1-0.5\alpha|\mathbf{X}_{n+1})]$. So the result follows directly from lemma A.5. We deduce the corollary 3.6 remembering that obtaining marginal coverage for $\widehat{F}(Y_{n+1}|\mathbf{X}_{n+1})$ is close to obtaining conditional coverage. □

Related to `CD-split`

Proof theorem 4.3. Let $\{i_1, \ldots, i_{n_j}\} = \{i : \mathbf{X}_i \in A(\mathbf{x}_{n+1})\}$, $U_l = \widehat{f}(Y_{i_l}|\mathbf{X}_{i_l})$, for $l = 1, \ldots, n_j$, and $U_{n_j+1} = \widehat{f}(Y_{n+1}|\mathbf{X}_{n+1})$. Since $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n_j}, Y_{n_j}), (\mathbf{X}_{n+1}, Y_{n+1})$ are i.i.d. random variables, obtain that U_i are i.i.d. random variables conditional on the event $\mathbf{X}_{n+1} \in A(\mathbf{x}_{n+1})$ and on $i_1, \ldots, i_{n_j}$. Therefore,

$$1 - \alpha \leq \mathbb{P}\left(U_{m+1} \geq U_{\lfloor \alpha \rfloor} | \mathbf{X}_{n+1} \in A(\mathbf{x}_{n+1}), i_1, \ldots, i_{n_j}\right)$$

The conclusion follows from the fact that $Y_{n+1} \in C(\mathbf{X}_{n+1}) \iff U_{m+1} \geq U_{\lfloor \alpha \rfloor}$ and because this holds for every sequence $i_1, \ldots, i_{n_j}$.

Note that if $\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1}) | \mathbf{X}_{n+1} \in A_j) \geq 1 - \alpha$ for every j then by the law of total probability $\mathbb{P}(Y_{n+1} \in C(\mathbf{X}_{n+1})) \geq 1 - \alpha$. □

Proof of Theorem 4.8. Let $U_i := f(Y_i|\mathbf{x}_i)$, $i = 1, \ldots, m$, $U_{n+1} := f(Y_{n+1}|\mathbf{x}_{n+1})$, and $W := (\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}_{n+1})$. If $g_{\mathbf{x}_i} = g_{\mathbf{x}_{n+1}}$ for every $i = 1, \ldots, m$, then $U_1, \ldots, U_m, U_{n+1}$ are i.i.d. conditional on W. Indeed, for every $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(U_i \geq t|W) &= \mathbb{P}(f(Y_i|\mathbf{x}_i) \geq t|\mathbf{x}_i) \\ &= \mathbb{P}(f(Y_{n+1}|\mathbf{x}_{n+1}) \geq t|\mathbf{x}_{n+1}) \\ &= \mathbb{P}(U_{n+1} \geq t|\mathbf{x}_{n+1}), \end{aligned}$$

where the next-to-last equality follows from the definition of the profile of the density.

For every $K \in \mathbb{R}$, let $Q(K) := |\{i : f(Y_i|\mathbf{x}_i) \geq K\}|$. Because U_i's are conditionally independent and identically distributed, then $Q(K)|W \sim \text{Binomial}(m, \mathbb{P}(f(Y_1|\mathbf{x}_1) \geq K))$. By the strong law of large numbers, it follows that $Q(K)/m \xrightarrow[a.s.]{m \longrightarrow \infty} \mathbb{P}(f(Y_1|\mathbf{x}_1) \geq K)$. In particular, $Q(t^*)/m \xrightarrow[a.s.]{m \longrightarrow \infty} 1 - \alpha$. Now, by definition $Q(T_m)/m \xrightarrow[a.s.]{m \longrightarrow \infty} 1 - \alpha$. Conclude by contradiction that $T_m \xrightarrow[a.s.]{m \longrightarrow \infty} t^*$.

□

Proof of Theorem 4.9. Item (i) was already shown as part of the proof of Theorem 4.8. To show (ii), assume that $t^*(\mathbf{x}_a, \alpha) = t^*(\mathbf{x}_b, \alpha)$ for every $\alpha \in (0, 1)$. Now, notice that $t^*(\mathbf{x}_a, \alpha)$ is such that $g_{\mathbf{x}_a}(t^*(\mathbf{x}_a, \alpha)) = 1 - \alpha$. Conclude that $g_{\mathbf{x}_a}(t^*(\mathbf{x}_a, \alpha)) = g_{\mathbf{x}_b}(t^*(\mathbf{x}_b, \alpha))$ for every $\alpha \in (0, 1)$. Now, because $\widehat{f}$ is continuous, $\{t^*(\mathbf{x}_a, \alpha) : \alpha \in (0, 1)\} = \text{Im}(\widehat{f}(\cdot|\mathbf{x}_a))$. Thus, $g_{\mathbf{x}_a} = g_{\mathbf{x}_b}$, and therefore $\mathbf{x}_a \sim \mathbf{x}_b$. □

APPENDIX B – FULL CONDITIONAL DISTRIBUTION OF $z_{i,l}$

For simplicity we consider only 2 communities ($K = 2$) and suppose that our focus is on $l = l'$ and $s = s'$. It is also assumed that after removing the i-th element we have $[n_{l',*,1}, n_{l',*,2}]$, $[n_{l',1,1}, \ldots, n_{l',S,1}]$ and $[n_{l',1,2}, \ldots, n_{l',S,2}]$. We consider $\lambda_{l,k} = \exp(\mathbf{x}_l^T \boldsymbol{\beta}_k)$ and $p_{l,k} = \frac{N}{N+\lambda_{l,k}}$. Then integrating out $\boldsymbol{\phi_k}$ we have that

$$
\begin{aligned}
&p(z_{i,l'} = 1 \mid y_{i,l'} = s', \ldots) \\
&\propto \prod_{k=1}^{K} \left[NB(n_{l',*,k} \mid \lambda_{l',k}, N) \int \left(\prod_{l=1}^{L} Multinomial([n_{l,1,k}, \ldots, n_{l,S,k}] \mid n_{l,*,k}, \boldsymbol{\phi}_k) \right) \text{ Dirichlet } (\boldsymbol{\phi}_k \mid \gamma)\, d\boldsymbol{\phi}_k \right].
\end{aligned}
$$

The integral involving $\boldsymbol{\phi}_k$ is available in closed form (see Appendix C). Furthermore, several elements in the equation above can be eliminated because they are constants. As a result, we obtain the following expression:

$$
\begin{aligned}
&p(z_{i,l'} = 1 \mid y_{i,l'} = s', \ldots) \\
&\propto \left[\frac{\Gamma(n_{l',*,1} + 1 + N)\, p_{l',1}^{N} (1 - p_{l',1})^{(n_{l',*,1}+1)}}{\Gamma(N)\,(n_{l',*,1}+1)!} \times \frac{\Gamma(n_{l',*,2} + N)\, p_{l',2}^{N} (1 - p_{l',2})^{(n_{l',*,2})}}{\Gamma(N) n_{l',*,2}!} \right] \\
&\times \left(\prod_{l \neq l'} \frac{n_{l,*,1}!}{n_{l,1,1}! \ldots n_{l,S,1}!} \right) \left(\frac{(n_{l',*,1}+1)!}{n_{l',1,1}! \ldots (n_{l',s,1}+1)! \ldots n_{l',S,1}!} \right) \frac{(n_{*,s',1} + 1 + \gamma_{s'}) \prod_{s \neq s'} (n_{*,s,1} + \gamma_s)}{\Gamma(n_{*,*,1} + 1 + \sum_s \gamma_s)} \\
&\times \left(\prod_{l} \frac{n_{l,*,2}!}{n_{l,1,2}! \ldots n_{l,S,2}!} \right) \frac{\prod_{s=1} (n_{*,s,2} + \gamma_s)}{\Gamma(n_{*,*,2} + \sum_s \gamma_s)}.
\end{aligned}
$$

We drop additional terms that are constants to obtain:

$$
\begin{aligned}
&p(z_{i,l'} = 1 \mid y_{i,l'} = s', \ldots) \\
&\propto \left[\frac{\Gamma(n_{l',*,1} + 1 + N)\, (1 - p_{l',1})^{(n_{l',*,1}+1)}}{(n_{l',*,1}+1)!} \times \frac{\Gamma(n_{l',*,2} + N)\, (1 - p_{l',2})^{(n_{l',*,2})}}{n_{l',*,2}!} \right] \\
&\times \left(\frac{(n_{l',*,1}+1)!}{n_{l',1,1}! \ldots (n_{l',s,1}+1)! \ldots n_{l',S,1}!} \right) \frac{(n_{*,s',1} + 1 + \gamma_{s'})}{\Gamma(n_{*,*,1} + 1 + \sum_s \gamma_s)} \\
&\times \left(\frac{n_{l',*,2}!}{n_{l',1,2}! \ldots n_{l',s',2}! \ldots n_{l',S,2}!} \right) \frac{(n_{*,s',2} + \gamma_{s'})}{\Gamma(n_{*,*,2} + \sum_s \gamma_s)} \propto b_1 a_1
\end{aligned}
$$

where:

$$
\begin{aligned}
a_1 = &\left(\frac{(n_{l',*,1}+1)!}{n_{l',1,1}! \ldots (n_{l',s,1}+1)! \ldots n_{l',S,1}!} \right) \frac{(n_{*,s',1} + 1 + \gamma_s)}{\Gamma(n_{*,*,1} + 1 + \sum_s \gamma_s)} \\
&\times \left(\frac{n_{l',*,2}!}{n_{l',1,2}! \ldots n_{l',s',2}! \ldots n_{l',S,2}!} \right) \frac{(n_{*,s',2} + \gamma_{s'})}{\Gamma(n_{*,*,2} + \sum_s \gamma_s)} \text{ and}
\end{aligned}
$$

$$b_1 = \left[\frac{\Gamma\left(n_{l',*,1} + 1 + N\right)\left(1 - p_{l',1}\right)^{\left(n_{l',*,1}+1\right)}}{\left(n_{l',*,1}+1\right)!} \times \frac{\Gamma\left(n_{l',*,2} + N\right)\left(1 - p_{l',2}\right)^{\left(n_{l',*,2}\right)}}{n_{l^*,*,2}!} \right].$$

Similarly, it can be shown that

$$p(z_{i,l'} = 2 \mid y_{i,l'} = s', \dots) \propto b_2 a_2$$

where

$$\begin{aligned} a_2 = &\left(\frac{n_{l',*,1}!}{n_{l',1,1}! \dots n_{l',s',1}! \dots n_{l',S,1}!} \right) \frac{\left(n_{*,s',1} + \gamma_{s'}\right)}{\Gamma\left(n_{*,*,1} + \sum_s \gamma_s\right)} \\ &\times \left(\frac{\left(n_{l',*,2}+1\right)!}{n_{l',1,2}! \dots \left(n_{l',s,2}+1\right)! \dots n_{l',S,2}!} \right) \frac{\left(n_{*,s',2} + 1 + \gamma_s\right)}{\Gamma(n_{*,*,2} + 1 + \sum_s \gamma_s)} \text{ and} \end{aligned}$$

$$b_2 = \left[\frac{\Gamma\left(n_{l',*,1} + N\right)\left(1 - p_{l',1}\right)^{\left(n_{l',*,1}\right)}}{n_{l^*,*,1}!} \times \frac{\Gamma\left(n_{l',*,2} + 1 + N\right)\left(1 - p_{l',2}\right)^{\left(n_{l'*,2}+1\right)}}{\left(n_{l',*,2}+1\right)!} \right].$$

Because $z_{i,l'}$ is either equal to 1 or 2, we can divide both sizes by $b_1 a_1 + b_2 a_2$ and, using factorial and gamma function rules, we obtain:

$$\begin{aligned} p(z_{i,l'} = 1 \mid y_{i,l'} = s', \dots) \propto &\left(\frac{\frac{\left(n_{l',*,1}+1\right)\left(n_{*,s',1}+\gamma_{s'}\right)}{\left(n_{l',s',1}+1\right)\left(n_{*,*,1}+\sum_s \gamma_s\right)}}{\frac{\left(n_{l',*,1}+1\right)\left(n_{*,s',1}+\gamma_{s'}\right)}{\left(n_{l',s',1}+1\right)\left(n_{*,*,1}+\sum_s \gamma_s\right)} + \frac{\left(n_{l',*,2}+1\right)\left(n_{*,s',2}+\gamma_{s'}\right)}{\left(n_{l',s',2+1}\right)\left(n_{*,*,2}+\sum_s \gamma_s\right)}} \right) \\ &\times \left(\frac{\frac{\left(n_{l',*,1}+N\right)\left(1-p_{l',1}\right)}{\left(n_{l',*,1}+1\right)}}{\frac{\left(n_{l',*,1}+N\right)\left(1-p_{l',1}\right)}{\left(n_{l',*,1}+1\right)} + \frac{\left(n_{l',*,2}+N\right)\left(1-p_{l',2}\right)}{\left(n_{l',*,2}+1\right)}} \right) \\ &\propto \frac{\left(n_{l',*,1} + N\right)\left(n_{*,s',1} + \gamma_{s'}\right)}{\left(n_{l',s',1} + 1\right)\left(n_{*,*,1} + \sum_s \gamma_s\right)} \left(1 - p_{l',1}\right). \end{aligned}$$

And finally we have that

$$z_{i,l} \mid y_{i,l} = s, \dots \sim Cat\left(\left[\frac{\frac{\left(n_{l,*,1}+N\right)\left(n_{*,s,1}+\gamma_s\right)}{\left(n_{l,s,1}+1\right)\left(n_{*,*,1}+\sum_s \gamma_s\right)}\left(1 - p_{l,1}\right)}{\sum_{k=1}^{K} \frac{\left(n_{l,*,k}+N\right)\left(n_{*,s,k}+\gamma_s\right)}{\left(n_{l,s,k}+1\right)\left(n_{*,*,k}+\sum_s \gamma_s\right)}\left(1 - p_{l,k}\right)}, \dots, \frac{\frac{\left(n_{l,*,K}+N\right)\left(n_{*,s,K}+\gamma_s\right)}{\left(n_{l,s,K}+1\right)\left(n_{*,*,K}+\sum_s \gamma_s\right)}\left(1 - p_{l,K}\right)}{\sum_{k=1}^{K} \frac{\left(n_{l,*,k}+N\right)\left(n_{*,s,k}+\gamma_s\right)}{\left(n_{l,s,k}+1\right)\left(n_{*,*,k}+\sum_s \gamma_s\right)}\left(1 - p_{l,k}\right)} \right] \right).$$

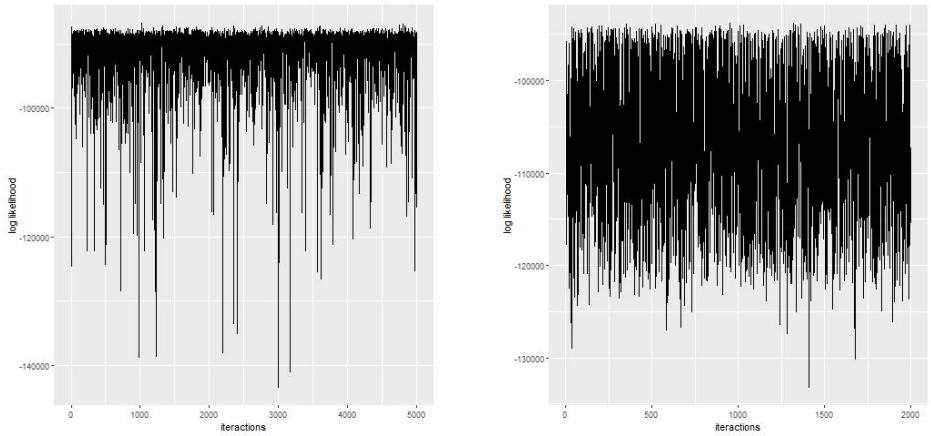
APPENDIX C – MULTINOMIAL INTEGRATION IN ϕ_k

To simplify the calculation of the conditional distribution of $z_{i,l}$ we can integrate out $\boldsymbol{\phi}_k$ as shown below.

$$
\begin{aligned}
&\prod_{k=1}^{K} \int \left[\prod_{l=1}^{L} \text{Multinomial}\left([n_{l,1,k}, \ldots, n_{l,S,k}] \mid n_{l,*,k}, \boldsymbol{\phi}_k \right) \right] \text{ Dirichlet } \left(\boldsymbol{\phi}_k \mid \gamma \right) d\boldsymbol{\phi}_k \\
&\propto \prod_{k=1}^{K} \int \left[\prod_{l=1}^{L} \frac{n_{l,*,k}!}{n_{l,1,k}! \ldots n_{l,S,k}!} \phi_{k,1}^{n_{l,1,k}} \ldots \phi_{k,S}^{n_{l,S,k}} \right] \phi_{k,1}^{\gamma_1 - 1} \ldots \phi_{k,S}^{\gamma_S - 1} d\boldsymbol{\phi}_k \\
&\propto \prod_{k=1}^{K} \left(\prod_{l=1}^{L} \frac{n_{l,*,k}!}{n_{l,1,k}! \ldots n_{l,S,k}!} \right) \int \phi_{k,1}^{n_{*,1,k} + \gamma_1 - 1} \ldots \phi_{k,S}^{n_{*,S,k} + \gamma_S - 1} d\boldsymbol{\phi}_k \\
&\propto \prod_{k=1}^{K} \left(\prod_{l=1}^{L} \frac{n_{l,*,k}!}{n_{l,1,k}! \ldots n_{l,S,k}!} \right) \frac{\prod_{s=1}^{S} \left(n_{*,s,k} + \gamma_s \right)}{\Gamma\left(n_{*,*,k} + \sum_{s=1}^{S} \gamma_s \right)}
\end{aligned}
$$

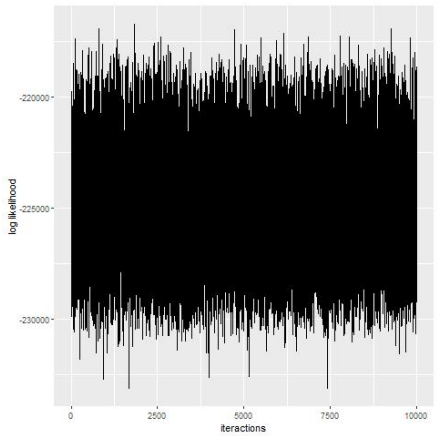
APPENDIX D – MCMC CONVERGENCE DIAGNOSTICS

Figure 14 shows the convergence diagnosis of the maximum likelihood function for all analyzed data sets.

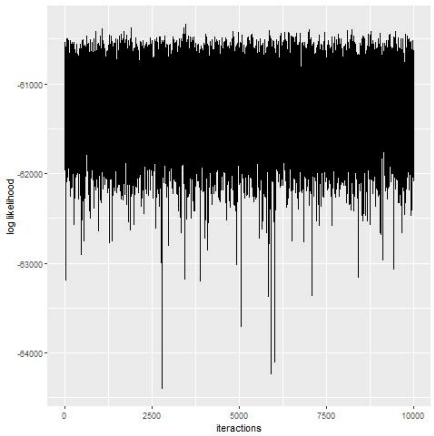


(a) Simulation 1

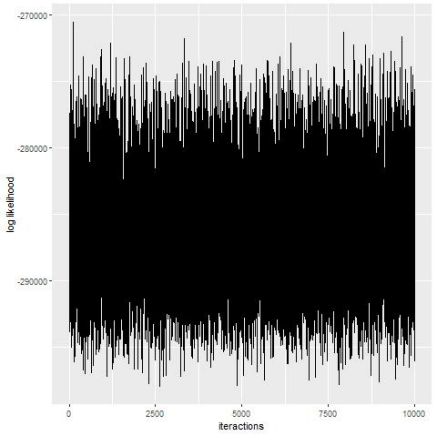
(b) Simulation 2



(c) Covid



(d) Grocery



(e) BCI

Figure 14 – MCMC convergence diagnostics of simulated and real data.

APPENDIX E – SLICE SAMPLING

Algorithm 3 Slice Sampling (DAMLEN; WAKEFIELD; WALKER, 1999)

1: Choose an initial value x_0 for which $f(x_0) > 0$.
2: Sample a value of y uniformly between 0 and $f(x_0)$.
3: Draw a horizontal line through the curve at this y position.
4: Sample a point (x, y) from the line inside the curve.
5: Repeat from step 2 using the new value of x.