Erasmus University Rotterdam

Erasmus School of Economics

Combining Variational Inference and Importance Sampling to Efficiently Approximate Bayesian Models

A thesis submitted in partial fulfillment for the degree of
MSc Econometrics & Management Science

Stefan Lam
Student ID: 481922

Thesis supervisor: prof. dr. Richard Paap
Second assessor: dr. Kathrin Gruber
Company Supervisor: Mark den Hollander

October 7, 2022

Abstract

In this research, we extend Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2015) by combining it with Importance Sampling (IS) such that the variational density of ADVI is able to capture the posterior covariance and correlation structure of complex Bayesian logistic regression models. In particular, we introduce ADVI-IS, which uses ADVI to obtain an initial approximate posterior distribution that is iteratively improved in a repeated IS procedure. The performance of ADVI-IS is compared with ADVI and the No-U-Turn Sampler (NUTS) as a benchmark. We conduct a simulation and an empirical study on so-called Watch Effect (WE) models, which are Bayesian logistic regression models with a nonlinear Media Effect (ME) component. This ME component is used by the Nielsen Company (Nielsen, 2022) to analyze the effects of advertisement exposures on the tune-in of respondents for specific movies. The results show that ADVI and ADVI-IS are both able to outperform NUTS in terms of run time performance and scaling capabilities. Moreover, in general ADVI-IS outperforms ADVI in terms of approximating the posterior distribution and the ME component. However, in the empirical study ADVI-IS has difficulties capturing the posterior distribution of the WE model due to the complexity of this model. Nonetheless, in this case ADVI-IS is still able to accurately estimate the ME component comparable to the NUTS.

Keywords: Bayesian Inference, Variational Inference, Importance Sampling, No-U-Turn Sampler.

Contents

1 Introduction

This research is done for the Nielsen Company (Nielsen, 2022), hereafter shortened to Nielsen, which is an American media analytics firm that strives to obtain actionable insights from large amounts of data in an efficient way. Nielsen offer advice on advertisements based on a Watch Effect (WE) model (Nielsen, 2022), which is a Bayesian logistic regression model with a complex nonlinear Media Effect (ME) component. In particular, the WE model is able to estimate the probability of respondents tuning in to a specific movie on television based on their socio-demographics, viewing behavior and advertisement exposures. Moreover, the corresponding ME component is an important component that Nielsen estimates to analyze the effects of the advertisement exposure of each media channel on the tune-in of respondents for specific movies. However, estimating this complex ME component is often difficult and costs a lot of time. Especially for a large media company, such as Nielsen, who simultaneously have to process large amounts of data.

In particular, one of the core problems of Bayesian statistics is to approximate difficult-to-compute posterior probability densities. Let $\mathbf{y}$ be the observed data and $\boldsymbol{\theta}$ the parameters of interest, then the posterior probability density is defined as $p(\boldsymbol{\theta}|\mathbf{y})$. According to Bayes' theorem this probability density can be computed as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $p(\boldsymbol{\theta})$ the prior density and $p(\mathbf{y})$ the marginal likelihood function. Hence, the marginal likelihood is given by

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2}$$

which is unavailable in closed form or requires exponential time to compute for many models (Blei et al., 2017). For this reason, it is often difficult to compute the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ analytically and thus we often rely on approximate methods to compute the posterior distribution.

Currently, the main approaches to obtain posterior inferences are via approximate methods such as Markov Chain Monte Carlo (MCMC) sampling methods (Hastings, 1970). These MCMC methods are able to guarantee asymptotic reliable results, but are slow to converge and do not scale efficiently (Blei et al., 2017). In particular, Nielsen use the No-U-Turn Sampler (NUTS) (Hoffman et al., 2014), a MCMC method, to estimate their WE models. However, the computation time of MCMC methods do not scale well to large data sets and complex models. The goal for Nielsen is thus to obtain a method that is able to estimate their complex WE model in a reasonable amount of time, while achieving comparable results to the NUTS.

To achieve this goal, one could resort to Variational Inference (VI) (Blei et al., 2017; Zhang et al., 2018), which tends to be faster and easier to scale to large data sets, since VI formulates the inference problem as an optimization problem instead of a sampling problem. In short, VI tries to find the optimal approximate posterior distribution q^* from a family of variational densities $q \in \mathcal{Q}$ by minimizing the Kullback-Leibler (KL) divergence to the exact posterior distribution.

In the current literature, commonly used methods for VI require the mean-field assumption (Blei et al., 2017), which assumes that all the posterior parameters can be independently estimated, thus implying that the posterior parameters are not correlated with each other. However, in practice this assumption is often not able to hold. This causes the VI methods to have inconsistent results with the asymptotically reliable results of the MCMC methods. The VI methods are thus fast, but in practice often not accurate enough compared to the MCMC methods. For this reason, we investigate whether it is possible to improve the approximate density q^* by capturing the covariance and correlation structure between the posterior parameters using a repeated Importance Sampling (IS) (Kloek and Van Dijk, 1978) procedure. In particular, our method uses q^* as an initial importance function that is iteratively improved in a repeated IS procedure. From a theoretical point of view, the main goal is thus to combine VI and IS such that we are able to efficiently approximate the posterior distribution with comparable results to the NUTS. From a practical point of view, the need for efficient methods such as VI is predominantly the case for large media companies, such as Nielsen, who have to implement complex models using large amounts of data.

However, finding a suitable variational family $\mathcal{Q}$ for general VI methods is often hard for a non-expert. For this reason, we consider Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2015) as the main VI method, since ADVI only requires a probabilistic model and a data set to automatically derive an efficient VI algorithm. Mitigating the need to explore many variational families manually. In particular, we introduce ADVI-IS, which uses the optimal approximate posterior distribution q^* of ADVI as an initial importance function that is iteratively improved in a repeated IS procedure. Consequently, we formulate the following research question: *To what extent could a repeated IS procedure improve the optimal variational density of ADVI to achieve more accurate posterior estimates comparable to those of the NUTS?*.

The research question is answered by conducting a simulation and empirical study, where we compare the run time performance, practical performance and the approximate posterior distribution between ADVI, ADVI-IS and the NUTS on the WE models. The results show that the running time of the NUTS is more prone to the complexity and scale of the model than ADVI or ADVI-IS. Moreover, in general ADVI-IS outperforms ADVI in terms of approximating the posterior distribution. However, ADVI-IS has difficulties capturing the posterior distribution of the WE model for the empirical study due the complexity of this model. Nonetheless, in this case ADVI-IS is still able to accurately estimate the ME component comparable to the NUTS.

The remainder of this paper is structured as follows. First, in Section 2 we discuss relevant work about MCMC, IS and VI which are found in the current literature. Then, in Section 3 we describe the methods and the WE model which are used for the simulation and empirical studies described in Sections 4 and 5 respectively. Lastly, in Section 6 we provide some concluding remarks, limitations and possible extensions for future research.

2 Related Work

In this section we discuss various landmark developments and limitations of the MCMC, IS and VI methods used in the current literature. First, we discuss MCMC and several of its methods in Section 2.1. Then, we describe the basic idea of IS in Section 2.2. Lastly, we discuss the general VI problem and several VI methods in Sections 2.3 and 2.4 respectively.

2.1 Markov Chain Monte Carlo Sampling

In the current literature, MCMC sampling is an indispensable tool for the modern Bayesian statistician. In short, MCMC methods first construct an ergodic Markov chain on the unknown parameters $\boldsymbol{\theta}$, whose stationary distribution is the posterior distribution. Then, samples from the stationary distribution are collected to approximate the posterior distribution with an estimate constructed from these samples.

The first landmark developments in MCMC samplers are the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984). These methods are based on drawing from the full conditional distribution of each parameter and then iterating over all these parameters until convergence (Bishop and Nasrabadi, 2006). However, these methods may require an unreasonable long time to converge to the posterior distribution, due to the high correlation between the drawn samples and the tendency of these methods to explore the parameter space via inefficient random walks (Neal, 1993).

Alternatively, a faster and more efficient method is the Hamiltonian Monte Carlo (HMC) sampler, also known as Hybrid Monte Carlo (Duane et al., 1987). This sampler is able to suppress potential random walk behavior of the parameters by means of a clever auxiliary variable scheme that transforms the problem of sampling from a target distribution into the problem of simulating Hamiltonian dynamics (Neal et al., 2011). In particular, the HMC sampler solves a set of differential equations referred to as the Hamiltonian functions (Bishop and Nasrabadi, 2006). Consequently, HMC uses the gradient information of the parameters. This property allows for independent sampling of the parameters as discussed by Hoffman et al. (2014) and thus the problem of correlated samples is resolved.

However, the HMC sampler requires the user specification of two nuisance parameters: a step size and the number of steps taken. The tuning of these parameters can be costly and an incorrect specification can lead to a large drop in efficiency. For this reason, Hoffman et al. (2014) propose the No-U-Turn Sampler (NUTS), which can automatically determine the nuisance parameters. The NUTS uses a recursive algorithm to build a set of likely candidate points that spans a wide range of the target distribution, stopping automatically when it starts to double back and retrace its steps. Empirically, the NUTS performs at least as efficient as a well tuned HMC method according to Hoffman et al. (2014).

2.2 Importance Sampling

As introduced by Kloek and Van Dijk (1978), IS is an alternative method to sample from a target distribution. In particular, IS is a type of Monte Carlo integration, which relies on sampling from a different computational-friendly distribution to approximate samples from the target distribution. Together with MCMC sampling methods, IS has provided a foundation for simulation-based approaches to numerical integration since its introduction as a variance reduction technique in statistical physics (Tokdar and Kass, 2010). The appeal of IS lies in a simple probability result given by the following relationship:

$$\mathbb{E}_f[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int g(\boldsymbol{\theta}) w(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_h[g(\boldsymbol{\theta}) w(\boldsymbol{\theta})], \tag{3}$$

where $g(\boldsymbol{\theta})$ is a given function that is integrable with respect to the target distribution f, $w(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}$ are the importance weights, $\mathbb{E}_f$ denotes the expectation with respect to f and $\mathbb{E}_h$ denotes the expectation with respect to the computational-friendly density $h(\boldsymbol{\theta})$, which is also referred to as the importance function (Kloek and Van Dijk, 1978).

The expectation $\mathbb{E}_f$ in Equation (3) can be approximated by a weighted average of M random samples from $h(\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}^{[m]}$ and weights $w(\boldsymbol{\theta}^{[m]})$ as

$$\mathbb{E}_f[g(\boldsymbol{\theta})] \approx \frac{1}{M} \sum_{m=1}^{M} g(\boldsymbol{\theta}^{[m]}) w(\boldsymbol{\theta}^{[m]}). \tag{4}$$

Consequently, using this IS procedure we can obtain the posterior moments of $p(\boldsymbol{\theta}|\mathbf{y})$ using different choices of $g(\boldsymbol{\theta})$. For instance, the mean or covariance estimates of $p(\boldsymbol{\theta}|\mathbf{y})$ are obtained by setting $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ or $g(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[\boldsymbol{\theta}])(\boldsymbol{\theta} - \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[\boldsymbol{\theta}])^\intercal$ with $f = p(\boldsymbol{\theta}|\mathbf{y})$ in Equation (4). Moreover, if the integrating constants of f are not available, then alternatively only the kernel of f suffices to implement the IS procedure as described in Section 3.2.1.

Furthermore, if the importance function $h(\boldsymbol{\theta})$ approximates $f(\boldsymbol{\theta})$ well, that is, the probabilities or probability densities of $h(\boldsymbol{\theta})$ should be proportional to $f(\boldsymbol{\theta})$. Then, Equation (4) will converge to $\mathbb{E}_f[g(\boldsymbol{\theta})]$ for large M (Geweke, 1989). However, the main issue arises when $h(\boldsymbol{\theta})$ does not approximate $f(\boldsymbol{\theta})$ well enough. In the worst case the importance weights $w(\boldsymbol{\theta}^{[m]})$ are small with high probability and large with low probability, which happens if $f(\boldsymbol{\theta})$ has heavier tails than $h(\boldsymbol{\theta})$. For this reason, a Gaussian distribution is often not a good choice for $h(\boldsymbol{\theta})$ (Greenberg, 2012), since a Gaussian distribution tends to have relatively light tails.

2.3 The Variational Inference Optimization Problem

Although the NUTS and HMC sampler increases the computational efficiency of the traditional MCMC methods, these samplers can still take a substantial amount of time for large data sets. For this reason, VI methods are considered, since the sampling procedure of the MCMC methods

tends to be more computationally expensive than the optimization based procedure of the VI methods, as discussed by Blei et al. (2017).

In particular, the first step of VI is to specify a variational family $\mathcal{Q}$ of densities over the unknown parameters $\boldsymbol{\theta}$, where each $q(\boldsymbol{\theta}) \in \mathcal{Q}$ is a candidate density for the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. Then, the goal is to find the best approximation of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ within the family $\mathcal{Q}$. This can be done by minimizing the KL-divergence as

$$q(\boldsymbol{\theta})^* = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\, KL\big(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})\big), \tag{5}$$

where $q(\boldsymbol{\theta})^*$ is the best approximation of the posterior within the family of $\mathcal{Q}$. However, the KL-divergence given in Equation (5) can not be easily computed, since it requires the computation of the logarithm of the evidence $p(\mathbf{y})$. This can be shown if we write the KL-divergence as

$$\begin{aligned} KL\big(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})\big) &= \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(q(\boldsymbol{\theta})\big)\big] - \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\boldsymbol{\theta}|\mathbf{y})\big)\big] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(q(\boldsymbol{\theta})\big)\big] - \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\boldsymbol{\theta},\mathbf{y})\big)\big] + \log\big(p(\mathbf{y})\big), \end{aligned} \tag{6}$$

where all the expectations are taken with respect to $q(\boldsymbol{\theta})$.

For this reason, an alternative objective function called the Evidence Lower BOund (ELBO) is maximized that is equivalent to minimizing the KL-divergence up to an added constant. As discussed by Blei et al. (2017), the ELBO is defined as

$$ELBO(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\boldsymbol{\theta},\mathbf{y})\big)\big] - \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(q(\boldsymbol{\theta})\big)\big], \tag{7}$$

which can further be rewritten to be equal to the sum of the expected log-likelihood of the data and the KL-divergence between the priors $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ as

$$\begin{aligned} ELBO(q) &= \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\boldsymbol{\theta})\big)\big] + \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\mathbf{y}|\boldsymbol{\theta})\big)\big] - \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(q(\boldsymbol{\theta})\big)\big] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})}\big[\log\big(p(\mathbf{y}|\boldsymbol{\theta})\big)\big] - KL\big(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})\big), \end{aligned} \tag{8}$$

where the first term is an expected likelihood and the second term is the negative divergence between the variational density and the prior. The first and second term encourage densities that explain the observed data and densities that are close to the prior respectively, that is, this objective function reflects the balance between likelihood and prior. Moreover, the name of the ELBO can be explained if we combine Equation (6) and (7), such that

$$\log\big(p(\mathbf{y})\big) = KL\big(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})\big) + ELBO(q), \tag{9}$$

where we see that $\log\big(p(\mathbf{y})\big)$ has a lower bound equal to $ELBO(q)$, since $KL(\cdot) \geq 0$. Overall, the new optimization problem can be defined as

$$q(\boldsymbol{\theta})^* = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\max}\, ELBO\big(q(\boldsymbol{\theta}), p(\boldsymbol{\theta},\mathbf{y})\big). \tag{10}$$

2.4 Variational Inference Methods

In the current literature, commonly used VI methods are Coordinate Ascent Variational Inference (CAVI) (Bishop and Nasrabadi, 2006), Stochastic Variational Inference (SVI) (Hoffman et al., 2013) or Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017). These methods all require the mean-field assumption, that is, these methods assume that the posterior distribution can be approximated by a mean-field variational family, where the posterior parameters are mutually independent. As described by Blei et al. (2017) a generic member of the mean-field variational family can be defined as

$$q(\boldsymbol{\theta}) = \prod_{k}^{K} q_k(\theta_k), \tag{11}$$

where K are the number of unknown parameters.

In particular, CAVI iteratively optimizes each parameter of this mean-field variational density, while holding the other parameters fixed. However, this method does not easily scale, since it requires to iterate through the entire data set. For this reason, Hoffman et al. (2013) propose SVI, an extension of CAVI, which combines natural gradients (Amari, 1998) and stochastic optimization (Robbins and Monro, 1951) for a gradient-based optimization procedure to substantially scale up the VI procedure. However, manually determining a variational family might cost a lot of time, especially for non-experts. For this reason, Kucukelbir et al. (2015) introduce ADVI, which automatically determines an appropriate variational family, thus mitigating any computation costs to refine and explore many variational families manually.

3 Methodology

This section elaborates on the methods and models used in this paper. First, we describe ADVI in Section 3.1. Then, in Section 3.2 we propose how ADVI can be used in a repeated IS procedure to improve the approximate posterior distribution. In Section 3.3 we describe the WE model used by Nielsen and in Section 3.4 we provide some performance measures to compare our methods.

3.1 Automatic Differentiation Variational Inference

In general, VI is faster than the more commonly used MCMC techniques. However, manually specifying a suitable variational family for VI is often time-consuming. For this reason, Kucukelbir et al. (2017) introduce ADVI, which is able to solve this problem. The main idea of ADVI is to first transform the the unknown parameters $\boldsymbol{\theta}$ into the real coordinate space, then we posit a variational distribution to approximate the posterior distribution. Lastly, automatic differentiation and stochastic optimization are combined to optimize the variational objective. In more detail, ADVI follows these four steps:

1. The unknown parameters of the model are transformed into unconstrained real-valued variables, such that the support of these parameters encompasses all continuous variables. This is done by transforming $p(\mathbf{y}, \boldsymbol{\theta})$ into $p(\mathbf{y}, \boldsymbol{\zeta})$ with an one-to-one mapping function $T : supp\big(p(\boldsymbol{\theta})\big) \rightarrow \mathbb{R}^K$, where $\boldsymbol{\zeta} = T(\boldsymbol{\theta})$. The resulting joint distribution is defined as

$$p(\mathbf{y}, \boldsymbol{\zeta}) = p\big(\mathbf{y}, T^{-1}(\boldsymbol{\zeta}) \cdot |\det J_{T^{-1}}(\boldsymbol{\zeta})|\big), \tag{12}$$

where $|\det J_{T^{-1}}(\boldsymbol{\zeta})|$ denotes the determinant of the Jacobian J of $T^{-1}(\boldsymbol{\zeta})$. After this transformation, the transformed parameters $\boldsymbol{\zeta}$ have support in the real coordinate space $\mathbb{R}^K$, such that ADVI can use a single variational family for all models. The Gaussian distribution is commonly used as the variational family. In particular, under the mean-field assumption the unknown parameters $\boldsymbol{\theta}$ are independent of each other with variational parameters $\boldsymbol{\phi} = [\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2]$. However, re-parameterizing the mean-field Gaussian removes the constraint that the variances must always be positive (Kucukelbir et al., 2017). For this reason, we consider the logarithm of the standard deviations $\omega_k = \log(\sigma_k)$ for $k = 1, \ldots, K$, such that the support of $\boldsymbol{\omega}$ is in the real coordinate space and $\boldsymbol{\sigma}$ is always positive. The mean-field variational density can then be expressed as

$$q(\boldsymbol{\zeta}; \boldsymbol{\phi}) = q\big(\boldsymbol{\zeta}; (\boldsymbol{\mu}, \boldsymbol{\omega})\big) = \prod_{k=1}^{K} \mathcal{N}\big(\zeta_k; \mu_k, \exp(\omega_k)^2\big), \tag{13}$$

where $\mathcal{N}$ denotes the Gaussian distribution.

2. Recast the gradient of the variational objective function as an expectation over $q(\boldsymbol{\zeta}; \boldsymbol{\phi})$, since it is possible to approximate it with Monte Carlo methods by expressing the objective function as an expectation (Ranganath et al., 2014). In particular, the ELBO in the real coordinate space as derived by Kucukelbir et al. (2017) can be written as

$$\text{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{\zeta};\boldsymbol{\phi})}\left[\log\left(p\left(\mathbf{y}, T^{-1}(\boldsymbol{\zeta})\right) + \log|\det J_{T^{-1}}(\boldsymbol{\zeta})|\right)\right] + \mathbb{H}[q(\boldsymbol{\zeta}; \boldsymbol{\phi})], \tag{14}$$

where $\mathbb{H}[q(\boldsymbol{\zeta}; \boldsymbol{\phi})] = \mathbb{E}_{q(\boldsymbol{\zeta};\boldsymbol{\phi})}[\log q(\boldsymbol{\zeta}; \boldsymbol{\phi})]$ is defined as the entropy (Kucukelbir et al., 2017), such that the ELBO is a function of the variational parameters $\boldsymbol{\phi}$ and the entropy $\mathbb{H}$.

3. Reparamaterize the objective function in terms of a standard Gaussian, since we can not use automatic differentiation on the ELBO in Equation (14) due to an intractable expectation. Kucukelbir et al. (2017) suggest to first standardize the parameters with

$$\boldsymbol{\eta} = S_{\boldsymbol{\phi}}(\boldsymbol{\zeta}) = \text{diag}\big(\exp(\boldsymbol{\omega})\big)^{-1}(\boldsymbol{\zeta} - \boldsymbol{\mu}), \tag{15}$$

where $\text{diag}\big(\exp(\boldsymbol{\omega})\big)$ denotes the diagonal matrix of $\exp(\boldsymbol{\omega})$. This standardization produces the fixed variational density

$$q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}; \mathbf{0}_K, \mathbf{I}_K) = \prod_{k=1}^{K} \mathcal{N}(\eta_k; 0, 1), \tag{16}$$

which transforms the ELBO given in Equation (14) into

$$\text{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log\left(p\Big(\mathbf{y}, T^{-1}\big(S_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\big)\Big) + \log\left|\det J_{T^{-1}}\big(S_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\eta})\big)\right|\right)\right] + \mathbb{H}[q(\boldsymbol{\zeta};\boldsymbol{\phi})]. \quad (17)$$

The transformation enables ADVI to efficiently compute Monte Carlo approximations as it only needs to sample from a standard Gaussian (Kingma and Welling, 2014).

4. Optimize the ELBO defined in Equation (17) with the gradient by computing the gradient of the terms inside the expectation and the entropy with automatic differentiation. Then, only the intractable expectation has to be computed. This in turn can be approximated with Monte Carlo integration, that is, draw M samples from the standard Gaussian and evaluate the empirical mean of the relevant gradients inside the expectation. This results in noisy unbiased gradients derived by Kucukelbir et al. (2017) as

$$\nabla_{\boldsymbol{\mu}} ELBO = \mathbb{E}_{q(\boldsymbol{\eta})}\left[\left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y},\boldsymbol{\theta}) \nabla_{\boldsymbol{\zeta}} T^{-1}(\boldsymbol{\zeta}) + \nabla_{\boldsymbol{\zeta}} \log\left|\det J_{T^{-1}}(\boldsymbol{\zeta})\right|\right)\right] \quad \text{and} \quad (18)$$

$$\nabla_{\boldsymbol{\omega}} ELBO = \mathbb{E}_{q(\boldsymbol{\eta})}\left[\left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y},\boldsymbol{\theta}) \nabla_{\boldsymbol{\zeta}} T^{-1}(\boldsymbol{\zeta}) + \nabla_{\boldsymbol{\zeta}} \log\left|\det J_{T^{-1}}(\boldsymbol{\zeta})\right|\right)\boldsymbol{\eta}^{\intercal} \operatorname{diag}(\exp(\boldsymbol{\omega}))\right] + \mathbf{1}. \quad (19)$$

These gradients can then be used in a stochastic optimization routine to automate variational inference as shown in Algorithm 1.

Algorithm 1 Automatic Differentiation Variational Inference (ADVI)

Require: Initial learning rate $\rho^{(0)}$, number of iterations I and model $p(\mathbf{y},\boldsymbol{\theta})$

1: Initialize $\boldsymbol{\mu}^{(0)} = \mathbf{0}$ and $\boldsymbol{\omega}^{(0)} = \mathbf{0}$
2: $i \leftarrow 0$
3: **for** $i \leq I$ **do**
4: Draw $\eta_m \sim \mathcal{N}(0,1)$ for $m = 1,\ldots,M$
5: Approximate $\nabla_{\boldsymbol{\mu}} ELBO$ using Monte Carlo Integration and η_m for $m = 1,\ldots,M$
6: Approximate $\nabla_{\boldsymbol{\omega}} ELBO$ using Monte Carlo Integration and η_m for $m = 1,\ldots,M$
7: Calculate the step-size parameter $\rho^{(i+1)}$ ▷ Using Adam (Kingma and Ba, 2014).
8: $\boldsymbol{\mu}^{(i+1)} \leftarrow \boldsymbol{\mu}^{(i)} + \operatorname{diag}(\rho^{(i)}) \cdot \nabla_{\boldsymbol{\mu}} ELBO$
9: $\boldsymbol{\omega}^{(i+1)} \leftarrow \boldsymbol{\omega}^{(i)} + \operatorname{diag}(\rho^{(i)}) \cdot \nabla_{\boldsymbol{\omega}} ELBO$
10: $i \leftarrow i + 1$
11: **end for**
12: $\boldsymbol{\mu}^* \leftarrow \boldsymbol{\mu}^{(I)}$ and $\boldsymbol{\omega}^* \leftarrow \boldsymbol{\omega}^{(I)}$
13: **return** $\boldsymbol{\mu}^*$, $\boldsymbol{\omega}^*$

3.2 Combining ADVI and IS to approximate the Posterior Distribution

In this section we describe how we can combine the optimal variational density $q^*(\boldsymbol{\theta})$ from ADVI with a repeated IS procedure to obtain a density, which is able to approximate the posterior

covariance and correlation structure. First, in Section 3.2.1 we describe how to implement an IS procedure if the target distribution $f(\boldsymbol{\theta})$ is unknown. In Section 3.2.2 we describe how to initialize a repeated IS procedure using ADVI.

3.2.1 Importance Sampling from an Unknown Posterior Distribution

In a Bayesian approach the goal is to sample from the posterior distribution. It thus holds that the target distribution $f(\boldsymbol{\theta})$ for an IS procedure is equal to the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. In practice, we are often not able to compute $w(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\mathbf{y})}{h(\boldsymbol{\theta})}$, since we do not always know the exact density $p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}$ with its integrating constants.

Alternatively, we can use the posterior kernel $p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$ to calculate $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[g(\boldsymbol{\theta})]$. In particular, using Bayes' theorem defined in Equation (1) we can rewrite $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[g(\boldsymbol{\theta})]$ as

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{\theta}|y)}[g(\boldsymbol{\theta})] &= \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \frac{\int g(\boldsymbol{\theta})p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{p(\mathbf{y})} \\ &= \frac{\int g(\boldsymbol{\theta})p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\int g(\boldsymbol{\theta})w_{post}(\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int w_{post}(\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\mathbb{E}_h[g(\boldsymbol{\theta})w_{post}(\boldsymbol{\theta})]}{\mathbb{E}_h[w_{post}(\boldsymbol{\theta})]},\end{aligned} \tag{20}$$

where the posterior importance weights $w_{post}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{h(\boldsymbol{\theta})}$ are calculated using the posterior kernel. Consequently, we can approximate $\mathbb{E}_{p(\boldsymbol{\theta}|y)}[g(\boldsymbol{\theta})]$ with parameters $\boldsymbol{\theta}^{[m]}$ drawn from $h(\boldsymbol{\theta})$ as

$$\mathbb{E}_{p(\boldsymbol{\theta}|y)}[g(\boldsymbol{\theta})] \approx \frac{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})g(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})}, \tag{21}$$

3.2.2 Initializing a Repeated IS Procedure using ADVI

The optimal variational density $q^*(\boldsymbol{\theta}) = q(\boldsymbol{\theta}; \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*})$ obtained from ADVI is not able to estimate the posterior covariance and correlation structure due to the mean-field property, since it is defined as

$$q(\boldsymbol{\theta}; \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\theta}_k; \mu_k^*, \sigma_k^{2*}). \tag{22}$$

For this reason, we implement the IS procedure to find a density able to estimate the posterior covariance and correlation structure. In particular, the posterior mean $\boldsymbol{\mu}_{IS}$ and posterior covariance matrix $\boldsymbol{\Sigma}_{IS}$ can be obtained using the IS procedure with $h(\boldsymbol{\theta}) = q^*(\boldsymbol{\theta})$ as the initial importance function. Then, $\boldsymbol{\mu}_{IS}$ and $\boldsymbol{\Sigma}_{IS}$ can be plugged into a distribution, which does have a covariance structure. We decide to use a Multivariate Gaussian distribution, since the mean-field distribution of ADVI is originally a Gaussian distribution. Thus, we define the new approximate posterior distribution as

$$q_{IS}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS}, \boldsymbol{\Sigma}_{IS}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS}, \boldsymbol{\Sigma}_{IS}), \tag{23}$$

where we consider two approaches to improve this new density q_{IS} as shown in Algorithm 2.

Algorithm 2 Repeated IS using ADVI as initialization for an unknown posterior distribution

Require: The model $p(\mathbf{y}, \boldsymbol{\theta})$, number of draws M, number of IS repeats S, blowup parameter λ, optimal variational parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}^{2*}$ from ADVI.

Draw $\mathbf{z}^{[m]} \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ for $m = 1, \ldots, M$

Set $\boldsymbol{\theta}^{[m]} \leftarrow \mathbf{z}^{[m]} \odot \lambda\boldsymbol{\sigma}^* + \boldsymbol{\mu}^*$ for $m = 1, \ldots, M$ ▷ Initial draws from $q(\boldsymbol{\theta}; \boldsymbol{\mu}^*, \lambda^2\boldsymbol{\sigma}^{2*})$

$h(\boldsymbol{\theta}) \leftarrow q^*(\boldsymbol{\theta}; \boldsymbol{\mu}^*, \lambda^2\boldsymbol{\sigma}^{2*})$ ▷ Initial importance function from ADVI

$s \leftarrow 1$

for s $\leq$ S **do**

  $w_{post}(\boldsymbol{\theta}^{[m]}) \leftarrow \frac{p(\boldsymbol{\theta}^{[m]})p(\mathbf{y}|\boldsymbol{\theta}^{[m]})}{h(\boldsymbol{\theta}^{[m]})}$ for $m = 1, \ldots, M$

  $\boldsymbol{\mu}_{IS,s} \leftarrow \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[\boldsymbol{\theta}]$ ▷ Using Equation (21)

  $\boldsymbol{\Sigma}_{IS,s} \leftarrow \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}[(\boldsymbol{\theta} - \boldsymbol{\mu}_{IS,s})(\boldsymbol{\theta} - \boldsymbol{\mu}_{IS,s})^\intercal]$ ▷ Using Equation (21)

  $\mathbf{L}_s\mathbf{L}_s^\intercal \leftarrow \text{cholesky}(\boldsymbol{\Sigma}_{IS,s})$ ▷ The Cholesky decomposition of $\boldsymbol{\Sigma}_{IS,s}$

  Set $\boldsymbol{\theta}^{[m]} \leftarrow \mathbf{L}_s\mathbf{z}^{[m]} + \boldsymbol{\mu}_{IS,s}$ for $m = 1, \ldots, M$ ▷ Draws from $q_{IS}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS,s}, \boldsymbol{\Sigma}_{IS,s})$

  $h(\boldsymbol{\theta}) \leftarrow q_{IS}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS,s}, \boldsymbol{\Sigma}_{IS,s})$

  $s \leftarrow s + 1$

end for

return $q_{IS}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS,S}, \boldsymbol{\Sigma}_{IS,S})$

In particular, the first approach varies the heaviness of the tails by increasing or decreasing the variance of each parameter of the initial importance function $q^*(\boldsymbol{\theta})$ from ADVI by multiplying σ_k^* with a blowup parameter λ before the IS procedure for $k = 1, \ldots, K$. In general, a heavy tailed importance function is preferred, since the target function should not have heavier tails than the importance function as discussed in Section 2.2.

The second approach repeats the IS procedure S times by plugging the importance weighted posterior means $\boldsymbol{\mu}_{IS,s}$ and variances $\boldsymbol{\Sigma}_{IS,s}$ into the density q_{IS} and then performing the IS procedure again using draws from $q_{IS}(\boldsymbol{\theta}; \boldsymbol{\mu}_{IS,s}, \boldsymbol{\Sigma}_{IS,s})$ for each subsequent IS repetition $s = 1, \ldots, S$. The goal is to bring the mean and covariance matrix of $q_{IS}(\boldsymbol{\theta})$ closer to the actual posterior mean and covariance matrix with each subsequent IS repetition.

3.3 Watch Effect Model

The WE models of Nielsen (2022) offer an extensive analysis of advertisement effects on respondents by estimating the effect of advertisement exposure on the *tune-in* of respondents to specific movies. In particular, Nielsen denotes the *tune-in* of respondent i as the dependent variable $\mathbf{y}_i$ for $i = 1, \ldots, N$. This dependent variable $\mathbf{y}_i$ is then estimated in a Bayesian logistic regression model with a ME component using socio-demographical and viewing behavioral variables as the control variables $\mathbf{x}^{ctrl}$, and the advertisement exposures of different media channels as the exposure variables $\mathbf{x}^{exp}$. The proposed Bayesian logistic regression model of Nielsen is defined as

$$P(y_i = 1|\alpha, \boldsymbol{\beta}^{ctrl}, \boldsymbol{\beta}^{spd}, \boldsymbol{\beta}^{pot}) = \Lambda\left(\alpha + x_i^{ctrl\intercal}\boldsymbol{\beta}^{ctrl} + \underbrace{\sum_{l=1}^{L} \beta_l^{pot} \cdot \tanh\left(\beta_l^{spd} \cdot \frac{x_{i,l}^{exp}}{2}\right)}_{\text{ME component}}\right) \quad \forall i = 1, \ldots, N, \tag{24}$$

where $\Lambda(\cdot)$ denotes the logistic function, α denotes the intercept, $\boldsymbol{\beta}^{ctrl}$ denotes the base effect of the control variables, the potential parameter $\boldsymbol{\beta}^{pot}$ denotes the maximum contribution of the exposures on the *tune-in* and the speed parameter $\boldsymbol{\beta}^{spd}$ denotes the rate at which this maximum contribution is achieved.

The estimated potential $\boldsymbol{\beta}^{pot}$ and speed $\boldsymbol{\beta}^{spd}$ parameters obtained by estimating the WE model defined in Equation 24 can be used to construct ME curves for each media channel. These curves show the effect of advertisement exposure on the *tune-in* of respondents, where the hyperbolic tangent $\tanh(\cdot)$ ensures that the effect of the advertisement exposure exhibits diminishing returns. In particular, the ME of media channel l for a certain level of advertisement exposure $x^{exp} \in \mathbb{R}^+$ can be calculated as

$$\text{ME}_l = \beta_l^{pot} \cdot \tanh\left(\beta_l^{spd} \cdot \frac{x^{exp}}{2}\right) \quad \forall l = 1, \ldots, L. \tag{25}$$

3.3.1 Prior Specification Watch Effect Model

The priors for the WE model are specified similarly as the specification that Nielsen uses in their previous projects. We consider these prior specifications, since Nielsen has already done substantial research regarding these priors and the goal of this research is to improve upon the current setting of Nielsen.

In particular, the control variables are centered and scaled in such a way that the intercept α can be interpreted as the baseline tune-in probability. Consequently, due to this centering and scaling the effects of the control variables is expected to be centered around zero with some small deviation. Nielsen thus specifies a standard normal prior distribution for the C control parameters $\boldsymbol{\beta}^{ctrl}$ as

$$\boldsymbol{\beta}^{ctrl} \sim \mathcal{N}(\mathbf{0}_C, \mathbf{I}_C), \tag{26}$$

where $\mathbf{0}_C$ is a $C \times 1$ vector of zeros and $\mathbf{I}_C$ denotes the $C \times C$ identity matrix.

Next, the prior distribution for the intercept α is also considered to be a Gaussian. However, a standard Gaussian distribution would not seem to be reasonable, since we can interpret α as the baseline for the probability of *tune-in* due to the specified prior for $\boldsymbol{\beta}^{ctrl}$. In particular, if the prior for α is standard normal and if there are no explanatory variables, then the modal baseline probability of *tune-in* for a respondent is $\Lambda(0) = \frac{\exp(0)}{1+\exp(0)} = 0.5$, which is a rather high probability. For this reason, Nielsen specifies the prior distribution for α as

$$\alpha \sim \mathcal{N}(-5, 1). \tag{27}$$

This corresponds to a more reasonable modal baseline probability of $\Lambda(-5) = \frac{\exp(-5)}{1+\exp(-5)} = 0.0067$, which implies that roughly 1 in 150 respondents watch a specific movie, if they are not exposed to any advertisements.

Lastly, the speed $\boldsymbol{\beta}^{spd}$ and potential $\boldsymbol{\beta}^{pot}$ parameters are assumed to be non-negative. For this reason, Nielsen specifies the priors for these parameters to be a lognormal distribution expressed as

$$\boldsymbol{\beta}^{spd} \sim \log \mathcal{N}(-0.5 \cdot \mathbf{1}_L, 0.8 \cdot \mathbf{I}_L) \quad \text{and} \quad \boldsymbol{\beta}^{pot} \sim \log \mathcal{N}(\mathbf{0}_L, 0.05 \cdot \mathbf{I}_L), \tag{28}$$

where $\mathbf{1}_L$ is a $L \times 1$ vector of ones and $\mathbf{I}_L$ denotes the $L \times L$ identity matrix for the L media channels. These values work well according to previous researches done by Nielsen, which results in a mean and standard deviation equal to 0.61 and 0.80 respectively for $\boldsymbol{\beta}^{spd}$ and a mean and standard deviation equal to 0.15 and 0.08 respectively for $\boldsymbol{\beta}^{pot}$. These prior distributions imply that on average the maximum contribution of the exposures is around 0.15, while the rate is on average 0.61 at which this maximum contribution is achieved.

3.3.2 Reparameterizing constrained Parameters

Currently, the $\boldsymbol{\beta}^{spd}$ and $\boldsymbol{\beta}^{pot}$ parameters specified in the Bayesian logistic regression model live in a constrained parameter space, since these parameters have to be non-negative. However, the VI methods that are used to estimate this model can be more efficiently implemented if all the parameters are unconstrained. This is done by taking the logarithm of $\boldsymbol{\beta}^{spd}$ and $\boldsymbol{\beta}^{pot}$ as

$$\boldsymbol{\omega}^{spd} = \log \boldsymbol{\beta}^{spd} \quad \text{and} \quad \boldsymbol{\omega}^{pot} = \log \boldsymbol{\beta}^{pot}.$$

Consequently, the unconstrained logistic regression model is defined as

$$P(y_i = 1 | \alpha, \boldsymbol{\beta}^{ctrl}, \boldsymbol{\omega}^{spd}, \boldsymbol{\omega}^{pot}) = \Lambda \left(\alpha + x_i^{ctrl\intercal} \boldsymbol{\beta}^{ctrl} + \underbrace{\sum_{l=1}^{L} \exp(\omega_l^{pot}) \cdot \tanh\left(\exp(\omega_l^{spd}) \cdot \frac{x_{i,l}^{exp}}{2} \right)}_{\text{ME component}} \right). \tag{29}$$

It holds that the logarithm of a lognormal random variable is Gaussian distributed (Bain and Engelhardt, 1992). Thus, the prior specifications for $\boldsymbol{\omega}^{spd}$ and $\boldsymbol{\omega}^{pot}$ are defined as

$$\boldsymbol{\omega}^{spd} \sim \mathcal{N}(-0.5 \cdot \mathbf{1}_L, 0.8 \cdot \mathbf{I}_L) \quad \text{and} \quad \boldsymbol{\omega}^{pot} \sim \mathcal{N}(\mathbf{0}_L, 0.05 \cdot \mathbf{I}_L). \tag{30}$$

In this research our methods estimate the unconstrained logistic regression model defined in Equation (29). Moreover, the IS repetitions of ADVI-IS also use the reparameterized $\boldsymbol{\omega}^{spd}$ and $\boldsymbol{\omega}^{pot}$ parameters to compute the importance weights. In particular, after estimating the unconstrained model we take the exponents of $\boldsymbol{\omega}^{spd}$ and $\boldsymbol{\omega}^{pot}$ to obtain inferences about $\boldsymbol{\beta}^{spd}$ and $\boldsymbol{\beta}^{pot}$.

3.4 Performance Measures

This research aims to obtain results of ADVI-IS, which are comparable to the results of the NUTS. For this reason, we assume that the NUTS is able to produce samples from the true posterior distribution. Moreover, we describe three main performance measures, alongside the ME curves, which we use to compare and evaluate the implemented estimation methods.

First, similar to Kucukelbir et al. (2015), we consider the predictive likelihood against the running time to compare the run time performance and scaling capabilities of the estimation methods on a common scale. The predictive likelihood can be interpreted as the probability of the held-out test data conditional on the training data under the given estimation method as described in Section 3.4.1. This measure evaluates to what extent ADVI-IS improves the run time performance and scaling capabilities compared to the NUTS, and whether ADVI-IS is able to achieve similar predictive performances as the NUTS.

Secondly, the main problem of ADVI is that it is not able to capture any posterior covariance or correlation structure due to the mean-field assumption. On the contrary, the NUTS is able to capture this structure. For this reason, we compare the posterior covariance and correlation of ADVI-IS to that of the NUTS. In particular, we use the Frobenius norm as described in Section 3.4.2 to measure the distance of the covariance or correlation matrices between ADVI(-IS) and the NUTS, where we assume that the NUTS is able to produce the true posterior covariance and correlation structure.

Lastly, in Section 3.4.3 we define a quality measure to measure the similarity between the approximate posterior distribution of ADVI(-IS) and the actual posterior distribution using the posterior importance weights w_{post}. This quality measure can then be used to select the optimal number of repeats S and blowup λ for ADVI-IS, since we aim to produce the approximate posterior distribution closest to the actual posterior distribution.

3.4.1 Log Predictive Likelihood

The predictive likelihood is used as a predictive performance measure for a specific estimation method. In particular, this likelihood is defined as

$$p(\mathbf{y}_{test}|\mathbf{y}_{train}) = \prod_{i=1}^{N} p(y_{test,i}|\mathbf{y}_{train}) = \prod_{i=1}^{N} \int p(y_{test,i}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{train}) d\boldsymbol{\theta} \tag{31}$$

where $\mathbf{y}_{test}$ is the held-out test data and $\mathbf{y}_{train}$ is the training data. Moreover, the integral in Equation (31) can be computed using Monte Carlo estimation as

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y}_{train})}[p(y_{test,i}|\boldsymbol{\theta})] &= \int p(y_{test,i}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{train}) d\boldsymbol{\theta} \\ &\approx \frac{1}{M} \sum_{m=1}^{M} p(y_{test,i}|\boldsymbol{\theta}^{[m]}) \quad \text{for} \quad i = 1, \ldots, N, \end{aligned} \tag{32}$$

where in the cases of ADVI, ADVI-IS and NUTS the samples $\boldsymbol{\theta}^{[m]}$ are drawn from $q^*(\boldsymbol{\theta})$, $q_{IS}(\boldsymbol{\theta})$, and the Markov chain after the NUTS warmup phase respectively. Consequently, the log predictive likelihood can be estimated as

$$\log\left(p(\mathbf{y}_{test}|\mathbf{y}_{train})\right) \approx \sum_{i=1}^{N} \log\left(\sum_{m=1}^{M} p(y_{test,i}|\boldsymbol{\theta}^{[m]})\right) - \log(M). \tag{33}$$

3.4.2 Frobenius Norm

The Frobenius norm, also called the Euclidean norm, is used to measure the distance between two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ as

$$||\mathbf{A} - \mathbf{B}||_F = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}(a_{ij} - b_{ij})^2}. \tag{34}$$

Consequently, this distance measure can then be used to measure the similarity between the NUTS and ADVI(-IS) in terms of the posterior covariance or posterior correlation matrix of the posterior distribution, which we denote by F_{cov} and F_{corr} respectively.

3.4.3 Quality Measure for the Approximate Posterior Distribution

The importance function $h(\boldsymbol{\theta})$ used in Equation (3) is optimal, if $w(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} = 1$ for $m = 1, \ldots, M$, since this implies that the importance function $h(\boldsymbol{\theta})$ resembles the target distribution $f(\boldsymbol{\theta})$ perfectly. Thus, if $h(\boldsymbol{\theta})$ is optimal, then

$$\mathbb{E}_f^{optimal}[g(\boldsymbol{\theta})] \approx \frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{\theta}^{[m]})w(\boldsymbol{\theta}^{[m]}) = \frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{\theta}^{[m]}), \tag{35}$$

where $\boldsymbol{\theta}^{[m]}$ is drawn from $h(\boldsymbol{\theta})$.

Similarly, this optimal approximation should also hold when implementing the IS procedure for an unknown posterior distribution as

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})}^{optimal}[g(\boldsymbol{\theta})] \approx \frac{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})g(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})} = \frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{\theta}^{[m]}), \tag{36}$$

where $\boldsymbol{\theta}^{[m]}$ is drawn from $h(\boldsymbol{\theta})$. Now, from Equation (36) we can see that the optimal approximation is obtained, if

$$\frac{w_{post}(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})} = \frac{1}{M} \Leftrightarrow \frac{M w_{post}(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})} = 1 \quad \forall m = 1, \ldots, M. \tag{37}$$

Using this property we can compute the average squared posterior deviation σ^2_{post} between $\frac{M w_{post}(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})}$ and 1 as

$$\sigma^2_{post} = \frac{1}{M}\sum_{m=1}^{M}\left(1 - \frac{M w_{post}(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})}\right)^2, \tag{38}$$

where $\boldsymbol{\theta}^{[m]}$ is drawn from the approximate posterior distributions $q^*(\boldsymbol{\theta})$ or $q_{IS}(\boldsymbol{\theta})$. The posterior deviation σ_{post} can then be used as a quality measure to measure the similarity between the approximate posterior distribution and the actual posterior distribution.

3.5 Experimental Setup and Algorithm Configurations

In this research, we first conduct a simulation study in Section 4 to obtain a better understanding on the characteristics and performances of ADVI, ADVI-IS and the NUTS. Then, in Section 5 we implement ADVI, ADVI-IS and the NUTS on an empirical data set to evaluate how these methods perform on a real world data set.

Furthermore, in our studies, ADVI is implemented using the Adam optimizer for 2000 training iterations with an initial learning rate of 0.1. The NUTS is implemented using 500 warmup samples, an acceptance probability of 0.8, a maximum tree depth of 10 and an initial step size of 1.0. On the contrary, we use the posterior deviation σ_{post} of the optimal importance function to determine the optimal number of repeats S and blowup λ for ADVI-IS, since we aim to determine whether ADVI-IS with an optimal configuration is able to obtain similar results as a the NUTS with a simple configuration.

Lastly, in the simulation and the empirical studies, a Dell notebook with an i9 processor and 32GB of RAM is used to implement the methods and models in Python. In particular, the NumPyro package (Phan et al., 2019; Bingham et al., 2019), a probabilistic programming library, is used to implement the NUTS and ADVI, where the repeated IS procedure is implemented from scratch. A brief description of the Python code can be found in Appendix C.

4 Simulation Study

In this section we investigate the performance and properties of our estimation methods using multiple simulated data sets. The simulated models are specified in Section 4.1 and in Section 4.2 the results of our simulation study are discussed.

4.1 Model Specifications

We consider three simulated models. First, we consider a simple linear regression model as specified in Section 4.1.1 to obtain a better understanding of our estimation methods. Then, we consider a logistic regression model as specified in Section 4.1.2, since this is a starting point for the WE models of Nielsen. Lastly, in Section 4.1.3 a ME component is added to the previous specified logistic regression model, since the ME component is the most important part of the WE model, which is often difficult to estimate due to the high nonlinearity in the parameters.

4.1.1 Linear Regression Model

The data generating process and model specification for the linear regression model can be summarized as

$$\begin{aligned} y_i &= \alpha + \sum_{c=1}^{6} x_{i,c}^{ctrl}\beta_c^{ctrl} + \epsilon_i, \quad \text{where} \quad \alpha = -1, \quad \boldsymbol{\beta}^{ctrl} = [-3,-2,-1,1,2,3]^\intercal, \\ x_i^{ctrl} &\sim \mathcal{N}(\mathbf{0}_6, \boldsymbol{\Sigma}^{ctrl}), \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(0,2) \quad \forall i = 1,\dots,N. \end{aligned} \tag{39}$$

In this case, $\boldsymbol{\Sigma}^{ctrl}$ is the covariance matrix of the control variables for which all variances and covariances are equal to 2 and 1 respectively, such that we incorporate some dependence between the posterior distribution of the parameters. The priors for α and $\boldsymbol{\beta}^{ctrl}$ are all set to be standard normal.

4.1.2 Logistic Regression Model

For a logistic regression model we extend on the linear regression model summarized in Equation (39). In particular, we set α and $\boldsymbol{\beta}^{ctrl}$, and generate $\mathbf{x}^{ctrl}$ similarly as the linear model described in Section 4.1.1. Then, samples from the logistic regression model can be generated as

$$y_i \sim Bernoulli\left(\Lambda(\alpha + \sum_{c=1}^{6} x_{i,c}^{ctrl}\beta_c^{ctrl})\right) \quad \forall i = 1,\dots,N, \tag{40}$$

where $\Lambda(\cdot)$ is the logistic function, and the priors for α and $\boldsymbol{\beta}^{ctrl}$ are set to be standard normal.

4.1.3 Logistic Regression Model with Media Effect

Lastly, we add a ME component to our logistic regression model from Section 4.1.2 as

$$\begin{aligned} y_i &\sim Bernoulli\left(\Lambda\left(\alpha + \sum_{c=1}^{6} x_{i,c}^{ctrl}\beta_c^{ctrl} + \underbrace{\sum_{l=1}^{5} \beta_l^{pot} \tanh\left(\beta_l^{spd} \cdot \frac{x_{i,l}^{exp}}{2}\right)}_{\text{ME component}}\right)\right), \\ &\text{where} \quad \boldsymbol{\beta}^{pot} = [0.1, 0.3, 0.5, 0.7, 0.9]^\intercal, \quad \boldsymbol{\beta}^{spd} = [1.0, 0.8, 0.6, 0.4, 0.2]^\intercal \\ &\text{and} \quad x_{i,l}^{exp} \sim \log\mathcal{N}(1,2) \quad \forall i = 1,\dots,N \quad \text{and} \quad \forall l = 1,\dots,5. \end{aligned} \tag{41}$$

Here, we note that the exposure variables $\mathbf{x}^{exp}$ has to be non-negative and in practice most people are only exposed a few times. For this reason, we sample $\mathbf{x}^{exp}$ from a lognormal distribution with the location and scale parameters equal to 1 and 2 respectively, since this distribution has high probability mass for small exposures and low probability mass for large exposures.

Furthermore, we specify a standard normal prior for both α and $\boldsymbol{\beta}^{ctrl}$, while a lognormal distribution is specified for $\boldsymbol{\beta}^{pot}$ and $\boldsymbol{\beta}^{spd}$ as

$$\boldsymbol{\beta}^{pot} \sim \log\mathcal{N}(\mathbf{0}_5, 0.7 \cdot \mathbf{I}_5)) \quad \text{and} \quad \boldsymbol{\beta}^{spd} \sim \log\mathcal{N}(\mathbf{0}_5, 0.7 \cdot \mathbf{I}_5), \tag{42}$$

since these potential and speed parameters are assumed to be non-negative. Then, the model defined in Equation (41) can be efficiently estimated using the same parameter reparameterization as described in Section 3.3.2. This results in the following unconstrained logistic model with ME

$$y_i \sim Bernoulli\left(\Lambda\left(\alpha + \sum_{c=1}^{6} x_{i,c}^{ctrl}\,\beta_c^{ctrl} + \sum_{l=1}^{5} \exp\left(\omega_l^{pot}\right) \tanh\left(\exp\left(\omega_l^{spd}\right) \cdot \frac{x_{i,l}^{exp}}{2}\right)\right)\right), \tag{43}$$

where the prior specifications for $\boldsymbol{\omega}^{spd}$ and $\boldsymbol{\omega}^{pot}$ are defined as

$$\boldsymbol{\omega}^{pot} \sim \mathcal{N}(\mathbf{0}_5, 0.7 \cdot \mathbf{I}_5) \quad \text{and} \quad \boldsymbol{\omega}^{spd} \sim \mathcal{N}(\mathbf{0}_5, 0.7 \cdot \mathbf{I}_5). \tag{44}$$

Note that our methods estimate the unconstrained model defined in Equation (43). Moreover, the IS repetitions of ADVI-IS use the reparameterized $\boldsymbol{\omega}^{pot}$ and $\boldsymbol{\omega}^{spd}$ parameters, whereafter we take exponents of $\boldsymbol{\omega}^{pot}$ and $\boldsymbol{\omega}^{spd}$ to obtain inferences about $\boldsymbol{\beta}^{pot}$ and $\boldsymbol{\beta}^{spd}$.

4.2 Simulation Results

In this section we discuss the simulation results of the three simulated models, where we set the number of observations $N = 20,000$ and the number of posterior samples $M = 1,000$. These results consist of three main analyses.

First, in Section 4.2.1 we compare the run time performances and scaling capabilities of the estimation methods for different number of observations $N \in \{20,000; 100,000; 200,000\}$ and number of repeats $S \in \{1, 5, 10, 20, 30\}$ to determine whether it is possible to efficiently replace the NUTS with ADVI-IS. Moreover, we set the blowup $\lambda = 1$, since different values of λ do not influence the time performance or scaling capabilities substantially.

Secondly, in Section 4.2.2 we analyze the effects of the number of repeats S and the blowup λ on the approximate posterior distribution of ADVI-IS to determine whether different values of S and λ are able to improve this approximate posterior distribution. Moreover, in this section we also determine the optimal number of repeats S^* and blowup λ^* based on the lowest posterior deviation σ_{post} between the approximate posterior distribution and the actual posterior distribution.

Lastly, in Section 4.2.3 we use S^* and λ^* to compare the practical performance between ADVI, the optimal ADVI-IS and the NUTS to determine whether the optimal ADVI-IS could efficiently replace the NUTS.

4.2.1 Run Time Performance and Scaling Capabilities

Figure 1 shows us the convergence rate of the average log predictive likelihood for the different estimation methods. In particular, the line markers show the speed at which ADVI and ADVI-IS converge to the same predictive likelihood as the NUTS.

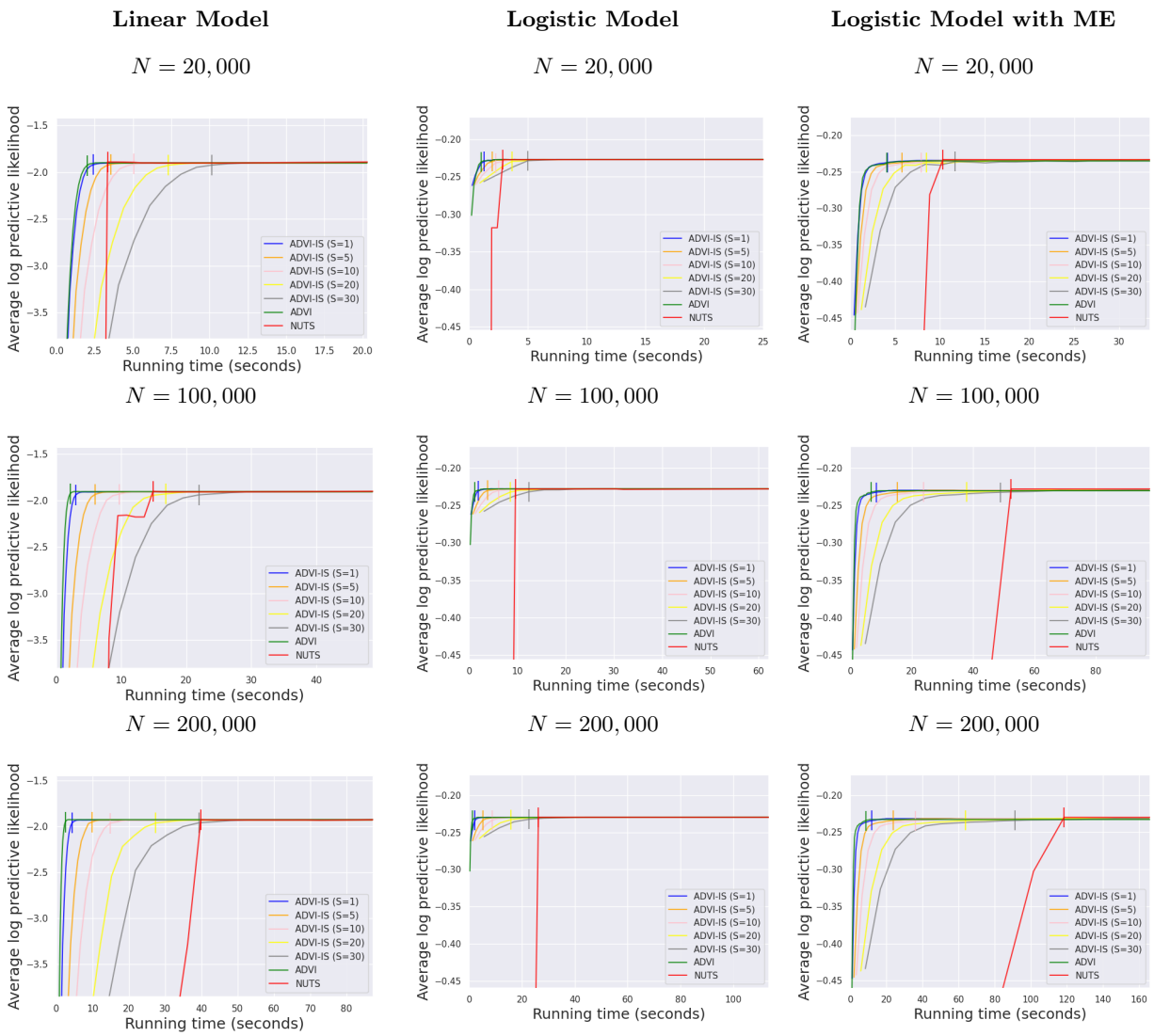

Figure 1: The average log predictive likelihood against the running time in seconds of the NUTS and ADVI-IS, where the line markers indicate that ADVI-(IS) has converged within $\pm$ 2% of the highest average log predictive likelihood of the NUTS. Note that the predictive likelihood is calculated using 70% of the N observations as training data and the remaining 30% as test data. Moreover, we run the models and methods three times using the same seed to obtain an accurate representation of the running time.

Moreover, the graphs in Figure 1 show that the convergence rate of ADVI-IS becomes relatively faster than the convergence rate of the NUTS as the number of observations increases or as the complexity of the model increases, where we consider the linear model to be the simplest and the logistic model with ME to be the most complex. This indicates that ADVI-IS could potentially be used as an efficient replacement of the NUTS for cases with large scale data sets or complex models.

4.2.2 Effects of the number of IS repeats S and the blowup λ on ADVI-IS

We first determine the effect of the number of IS repeats S and the blowup λ on the covariance and correlation structure of the approximate posterior distribution of ADVI-IS. This is done by

analyzing the Frobenius norms of the posterior covariance and correlation matrices between ADVI-IS and the NUTS for different values of S and λ as shown in Figure 2, where we assume that the NUTS produces the true posterior covariance and correlation structure. This figure shows us three important observations.

Firstly, we notice that the Frobenius norms of the logistic model with ME requires the highest number of IS repeats to converge. This is not surprising as the ME component adds some nonlinearity between the parameters of the posterior distribution, which increases the difficulty to converge to the true posterior covariance and correlation structure.

Secondly, we see that all Frobenius norms are able to converge to a smaller value than ADVI for at least one combination of S and λ, which implies that ADVI-IS is able to improve the posterior covariance and correlation structure with respect to ADVI.

Thirdly, we see in almost all cases that the frobenius norms without blowup (λ=1) or a shrinkage (λ=0.5) of the variance of the initial importance function has difficulties to converge. On the other hand, we do see that in all cases a slight blowup (λ=1.5 or $\lambda = 2.0$) of the variance of the initial approximate posterior distribution improves the convergence rate of the Frobenius norms. This implies that a slight blowup is able to improve the initial approximate posterior distribution for the repeated IS procedure by blowing up the tails of this approximate distribution.

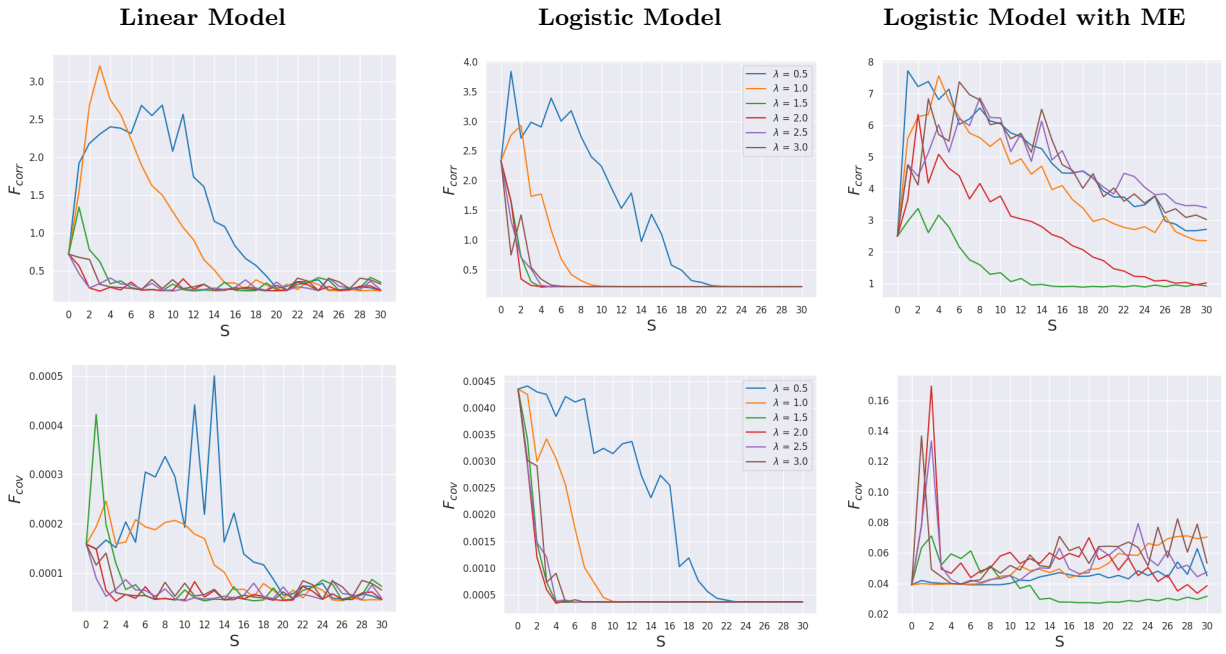

Figure 2: The first and second rows contain the Frobenius norms of the correlation and covariance matrices between the NUTS and ADVI-IS respectively, where the NUTS is assumed to produce the true covariance and correlation matrices. Note that $S = 0$ corresponds to ADVI without blowup.

Although the frobenius norms are able to give a good initial indication of the quality of the approximate posterior distribution, we can not use it as a reliable measure to choose the optimal number of repeats S^* and blowup λ^*, since this measure only considers the covariance and cor-

relation structure between the samples from ADVI-IS and NUTS. For this reason, we choose S^* and λ^* based on the lowest posterior deviation σ_{post} between the approximate posterior distribution and the actual posterior distribution, since we aim to produce the approximate posterior distribution closest to this actual posterior distribution. In particular, Figure 3 shows the effects of S and λ on σ_{post}, where the cross markers indicate the lowest posterior deviation between the approximate posterior distribution of ADVI-IS and the actual posterior distribution.

These results show us that the approximate posterior distribution of ADVI-IS approaches the actual posterior distribution for all models as S increases, which is also visualized with the posterior importance weights w_{post} shown in Figures 17, 18 and 19 of Appendix A.1. Moreover, similarly to the results of Figure 2 we notice that σ_{post} for the linear and logistic models converge faster compared to the logistic model with ME due to the added complexity of the nonlinear ME component. The optimal number of IS repeats and blowup parameter shown in Table 1 are based on the lowest posterior deviation. These optimal parameter values are used to compare the practical performance of the optimal ADVI-IS with ADVI and the NUTS in Section 4.2.3.

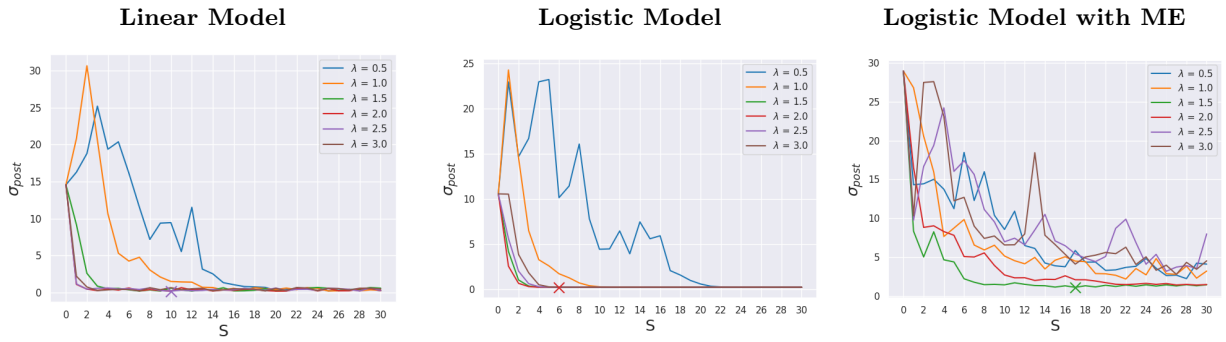

Figure 3: The posterior deviation σ_{post} between the approximate posterior distribution of ADVI-IS and the actual posterior distribution for different blowups and number of IS repeats, where the cross marker indicates the lowest deviation. Note that $S = 0$ corresponds to ADVI without blowup.

Table 1: Optimal number of IS repeats S^* and blowup λ^* for the linear, logistic and logistic with ME models based on the lowest posterior deviation σ_{post} between the approximate posterior distribution and the actual posterior distribution.

	Model		
	Linear	**Logistic**	**Logistic with ME**
S^*	10	6	17
λ^*	2.5	2.0	1.5

4.2.3 Practical Performance

In Figure 4 we show the posterior correlation and covariance of ADVI and the optimal ADVI-IS against those of the NUTS to visualize to what extent ADVI-IS improves the posterior covariance and correlation structure compared to ADVI. In particular, these results show us that ADVI is not able to capture any covariance or correlation structure of the posterior distribution for the

three simulated models. On the other hand, ADVI-IS is able to capture a posterior covariance and correlation structure, which are close to the posterior covariance and correlations structure of the samples from NUTS for the three simulated models. This is especially the case for the logistic model, where the correlations and covariances are almost all on the 45 degrees line.

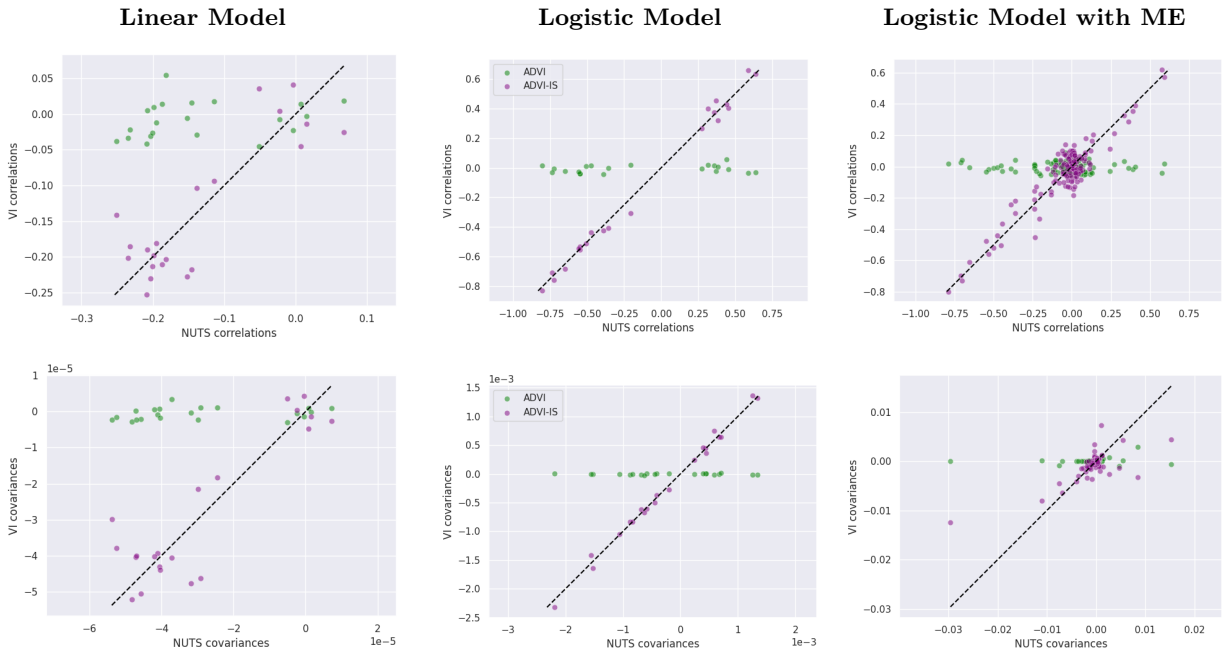

Figure 4: The first row contains the scatterplot of the posterior VI correlations against the NUTS correlations and the second row contains the scatterplot of the posterior VI covariances against the NUTS covariances, where the NUTS is assumed to produce the true covariance and correlation matrices. The dashed black line corresponds to a 45 degrees line through (0,0).

Next, we show the posterior sample means of all the parameters for each simulated model in Figure 5. This figure shows us that all the methods produce posterior sample means close to the actual intercept and control parameters for the three simulated models. Moreover, for the logistic model and logistic model with ME we see that posterior samples from ADVI do not vary much compared to posterior samples from ADVI-IS and the NUTS as can be seen by the more established 5th and 95th quantiles of the latter two methods. In particular, the similar quantiles and mean of ADVI-IS and the NUTS indicate that these two methods produce posterior samples from a similar distribution.

Furthermore, in Figure 6 we show caterpillar plots of the posterior sample mean of the potential and speed parameters for the logistic model with ME. This figure shows us that the three methods often have difficulty estimating the actual values of the potential and speed parameters. Nonetheless, the actual values of the speed and potential parameters are always within the 5th and 95th quantiles of the parameter estimates of ADVI-IS and the NUTS, while this is not the case for the parameter estimates of ADVI for β_2^{pot}, β_3^{pot}, β_4^{pot}, β_2^{spd} and β_3^{spd}. This indicates that ADVI has difficulties capturing the ME component due to the lack of a variance structure in its

parameter estimates. In general, ADVI-IS and the NUTS produce similar posterior sample means of the potential and speed parameters. Except for β_1^{spd}, where we clearly see that ADVI-IS produces better sample means than the NUTS. This indicates that ADVI-IS outperforms the NUTS in terms of parameter estimation. Moreover, ADVI-IS produces similar quantiles as the NUTS, indicating again that these two methods produce posterior samples from a similar distribution for the potential and speed parameters.

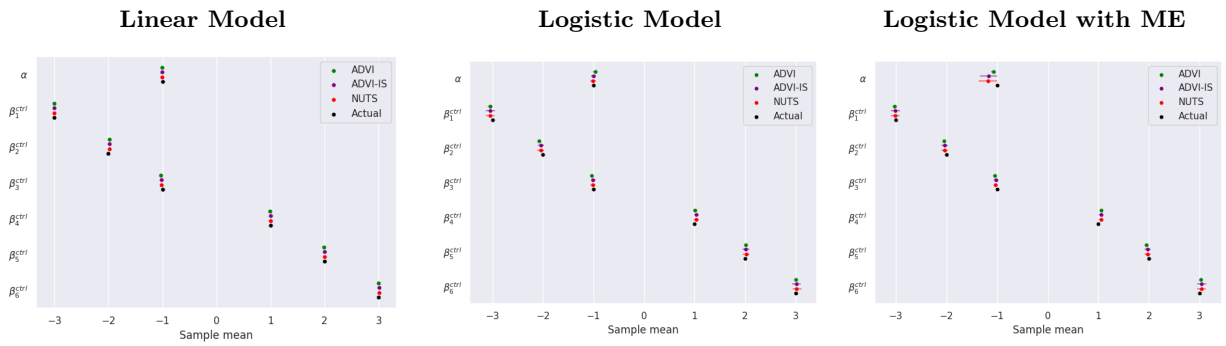

Figure 5: Catterpillar plots of the posterior samples means for the intercept α and the control parameters $\boldsymbol{\beta}^{ctrl}$, where the error lines indicate the 5th and the 95th quantiles.

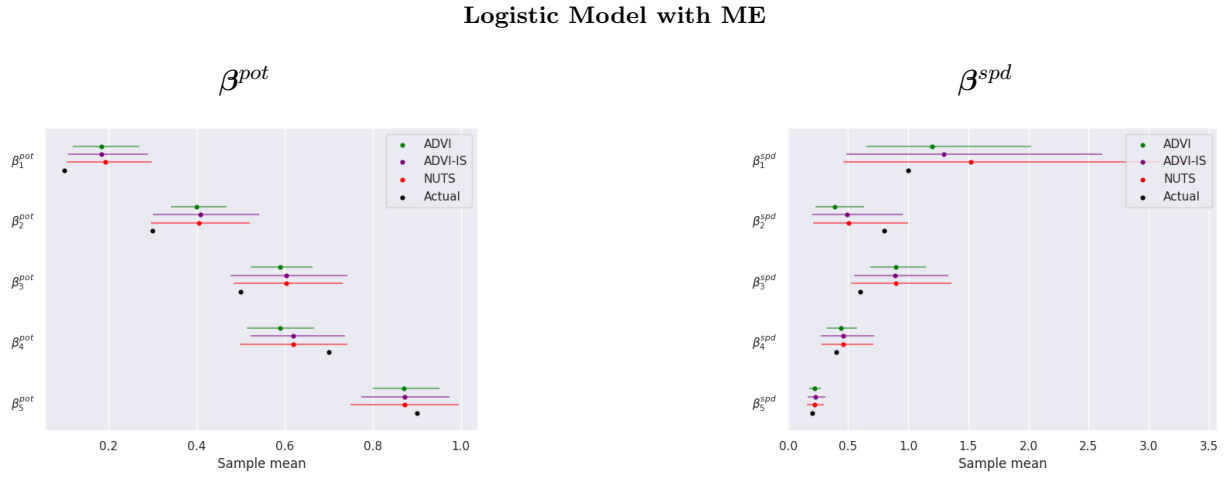

Figure 6: Catterpillar plots of the posterior samples means for the potential $\boldsymbol{\beta}^{pot}$ and speed $\boldsymbol{\beta}^{spd}$ parameters of the logistic model with ME, where the error lines indicate the 5th until the 95th quantiles.

In Figure 7 we visualize the potential and speed parameters as a joint distribution per media channel to further analyze the ME component. This figure clearly shows that the three methods are able to capture similar posterior sample means. However, we see a clear difference in the posterior covariance structure between the methods. In particular, ADVI-IS is able to capture similar dependencies as the NUTS, which can be seen by the similar contour lines. On the contrary, ADVI has smaller dependencies between the potentials and speeds compared to ADVI-IS and the NUTS, as can be see by the denser contour lines. These results thus imply that ADVI-IS is able to replicate the joint distribution between the potential and speed parameters of the NUTS.

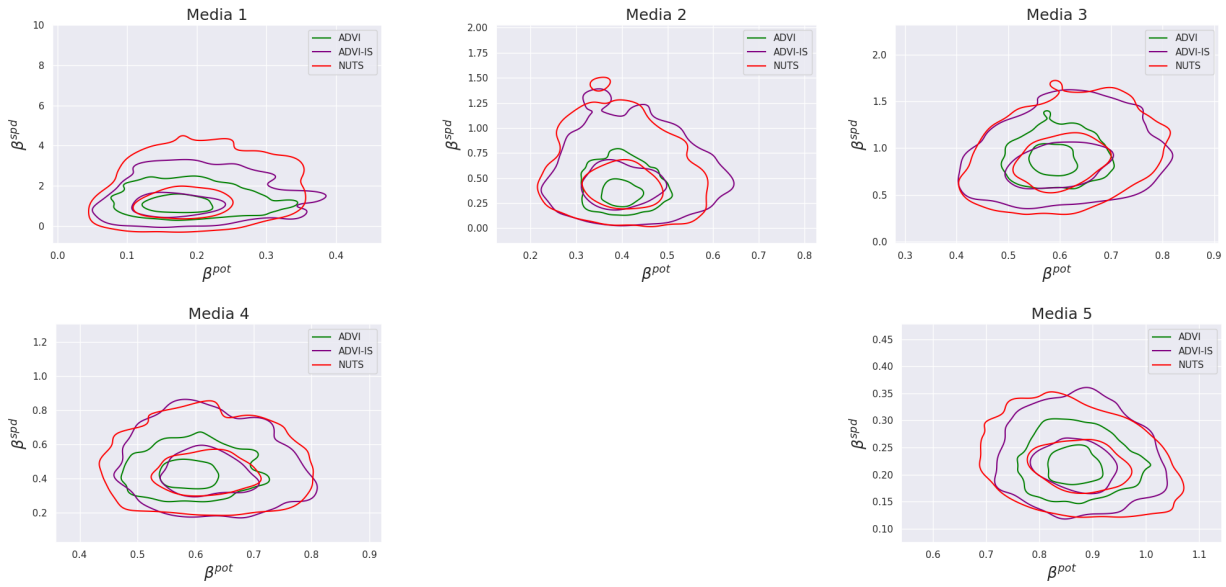

Figure 7: Contour plots of the joint distribution between the potential $\boldsymbol{\beta}^{pot}$ and $\boldsymbol{\beta}^{spd}$ parameters for the logistic model with ME, where the NUTS is assumed to produce the true joint distributions.

Lastly, in Figure 8 we show the ME curves of each media channel. In particular, this figure shows that ADVI produces worse ME curves for media 2 and media 4 compared to ADVI-IS and the NUTS, while producing similar ME curves for the other medias. On the other hand, ADVI-IS is able to produce a better ME curve for media 1, while producing similar ME curves for the other media compared to the NUTS. These results thus imply that ADVI-IS produces the best ME curves, which means that ADVI-IS is able to outperform the NUTS in terms of estimating the ME component.

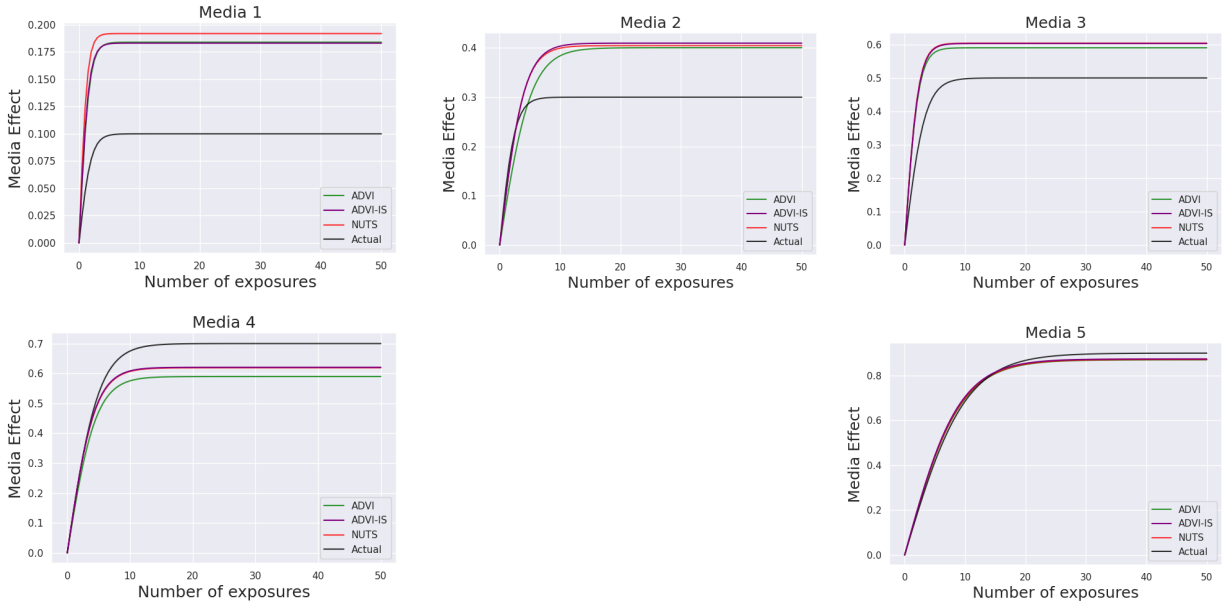

Figure 8: Media effect curves for the five different simulated media channels.

In short, these results shows us that ADVI is not able to capture any posterior covariance

and correlation structure, while ADVI-IS is able to capture a similar structure as the NUTS. The lack of the covariance structure in the parameter estimates of ADVI causes it to have difficulties estimating the ME component. Furthermore, the similar posterior quantiles between ADVI-IS and the NUTS imply that these two methods produce posterior samples from a similar distribution, which shows us that ADVI-IS can be used a replacement of NUTS. Moreover, ADVI-IS produces similar ME curves as the NUTS. Except for Media 1, where we see that ADVI-IS produces a better ME curve than the NUTS. This means that ADVI-IS could replace and even outperform the NUTS in terms of estimating the ME component. This is especially useful, since ADVI-IS scales better for complex models or large scale data sets.

5 Empirical Study

The simulation study in Section 4 shows us that ADVI-IS is able to outperform NUTS in terms of run time performance and scaling capabilities, while producing similar or even better posterior samples. ADVI-IS thus seems to be a viable alternative to the NUTS. To validate whether ADVI-IS is indeed a viable alternative to the NUTS, we consider an empirical study, where the data is provided by Nielsen (2022). The provided data is considered to be reliable as Nielsen is a large global data analytics conglomerate that focuses on media research and measurements.

In particular, Nielsen provides this research with their Nielsen People Meter (NPM) data, where households participate in a panel for a small compensation. This allows Nielsen to install a smart meter in their households, which obtains accurate data of the viewing behavior of the respondents from all the households. The pre-processing of the used empirical data set is further described in Section 5.1, where the results of the empirical study are discussed in Section 5.2.

5.1 Empirical Data Set

The NPM data contains information of multiple television channels in the United States. However, this research only focuses on the Hallmark channel, since it would be computationally infeasible to consider the data of all the channels for the implemented estimation methods.

In particular, we investigate the viewing behavior of households for 10 different Christmas movies, which are aired in the time period of November 2020 until December 2020. This part of the NPM data set contains 66 variables and 170,750 observations after the respondent and variable selections described in Sections 5.1.1 and 5.1.2 respectively. Moreover, the 170,750 observations consists of 17,075 respondents each with one observation per movie.

5.1.1 Respondent selection

In practice, the NPM data set is continuously changing with respondents dropping in and out at any given time. This means that not all respondents in the NPM data set are relevant, since individuals that were only part of the panel for a few days are not likely to be representative for the entire sample due to the lack of information about these respondents.

For this reason, Nielsen determines a subset of the NPM data set with only respondents that are representative for the entire population. This process is called *unification*, where Nielsen unifies the data by only considering respondents that are present in the panel for at least 75% of the complete measurement period.

5.1.2 Variable Selection and Preprocessing

The *tune-in* is the dependent variable for our models, which is defined as a binary variable equal to 1 if the duration of watching a movie is longer than 6 minutes and equal to 0 otherwise. Furthermore, a detailed description of the explanatory variables from the Hallmark data set is described in Appendix B. These variables of the Hallmark data set were selected by Nielsen through the evaluation of partial dependence plots obtained with Gradient Boosting Machines (GBM) (Natekin and Knoll, 2013). In particular, we can categorize the explanatory variables intro three main groups:

- **Social Demographics ($\mathbf{x}^{sd}$):** this group contains 17 variables about the characteristics of the respondents, such as the age or ethnicity.
- **Advertisement Exposures ($\mathbf{x}^{exp}$):** this group contains 24 variables about the number of advertisement exposures that a respondent receives regarding a movie. This number is based on the media channel, promotion type and the recency, where the recency indicates how long ago a respondent got exposed to a promotion relative to the day of the premiere of the movie.
- **Viewing Behavior ($\mathbf{x}^{vb}$):** this group contains 25 variables about the viewing behavior of respondents, such as the duration of watching a certain channel, show or genre.

However, the advertisement exposures have many zero valued columns and rows in $\mathbf{x}^{exp}$ for each movie as can be seen in Table 2. In particular, we group the exposure variables of each promotion type and recency per media channel to resolve the sparsity in the columns. This results in a total of six combined advertisement exposure variables without empty columns corresponding to six different media channels: *cross*, *dd*, *bcsimul*, *comcast*, *locals* and *on*. Each of these media channels correspond to different networks, who promote the specific advertisements. The groupings and networks for each media channel are described in Tables 7 and 8 of Appendix B respectively.

One of the main goals of Nielsen is to estimate the ME curves properly, which are estimated using the advertisement exposures $\mathbf{x}^{exp}$. In table 2 we can see that the majority of the respondents who tune in are exposed to advertisements, hence the low percentage of empty rows. On the contrary, the majority of the respondents who do not tune in are not exposed to any advertisements, hence the relatively high percentage of empty rows. Thus, the respondents who tune in have the most information regarding the ME curves, as they have the least sparse advertisement exposures. However, on average only 1.72% of the respondents tune in to the specific movie, which implies that the advertisement exposure variables are highly sparse on average.

The sparsity of the advertisement exposures could cause the estimation of the ME component to be difficult for our methods due to the lack of information. For this reason, we extend the data set by bootstrapping respondents who tune in to movies until there is an equal proportion between the two classes. It has to be noted that this bootstrapped data set is not used to obtain actual inferences about the effects of advertisement exposures on the tune-in of respondents, since now it has a bias for respondents who tune-in to a specific movie. Instead, this bootstrapped data is used such that we can determine whether the sparsity of the advertisement exposures has a negative effect on the estimation of the ME curves by our methods.

Table 2: The missing values of the 24 Advertisement Exposures $\mathbf{x}^{exp}$ variables for each movie.

	Movie ID										
	1	3	13	17	19	22	26	28	29	35	Average
Number of empty columns in $\mathbf{x}^{exp}$	10	10	6	7	7	8	7	4	4	7	7
% Respondents with *tune-in*=1	1.57%	1.50%	1.78%	1.93%	1.75%	1.21%	1.75%	2.35%	1.62%	1.69%	1.72%
% Respondents with *tune-in*=0	98.43%	98.50%	98.22%	98.07%	98.25%	98.79%	98.25%	97.65%	98.38%	98.31%	98.29%
% Empty rows in $\mathbf{x}^{exp}$ if *tune-in*=1	10.82%	5.08%	4.93%	2.43%	5.35%	3.40%	3.34%	1.24%	1.81%	4.15%	4.26%
% Empty rows in $\mathbf{x}^{exp}$ if *tune-in*=0	84.95%	81.12%	74.34%	77.15%	76.81%	76.37%	76.10%	68.07%	73.19%	72.85%	76.10%

5.2 Empirical Results

In this section we only present the results for the movie with ID 28, since the results of other movies are similar to each other. Moreover, this movie has the lowest percentage of empty rows if a respondent tunes in and it also has the highest tune-in percentage. This movie thus requires the least number of bootstrap samples for the bootstrapped Hallmark WE model. In particular, the regular Hallmark WE model has 17,075 observations, while the bootstrapped Hallmark WE model has 33,346 observations.

Furthermore, we set the number of posterior samples $M = 1.000$ and we perform three main analyses on these Hallmark models similarly as the simulation results in Section 4.2. Specifically, in Section 5.2.1 we evaluate the run time performance of the Hallmark models, in Section 5.2.2 we analyze the effects of the number of IS repeats S and the blowup λ on ADVI-IS and in Section 5.2.3 we compare the practical performances between ADVI, the optimal ADVI-IS and the NUTS.

5.2.1 Run Time Performance

Figure 9 shows that ADVI and ADVI-IS are able to converge faster than the NUTS for both Hallmark models. However, ADVI and ADVI-IS are not able to converge to the same predictive likelihood as the NUTS for the Hallmark WE model, while they do converge to the same predictive likelihood as the NUTS for the bootstrapped Hallmark WE model.

These results indicate that ADVI and ADVI-IS have difficulties estimating the ME component of the Hallmark WE model, since the corresponding advertisement exposures are highly sparse. Thus, ADVI-IS could be a more efficient alternative than the NUTS for the bootstrapped Hallmark WE model in terms of predictive performance, while the Hallmark WE model should be further analyzed to determine whether ADVI-IS could be used as an alternative to the NUTS.

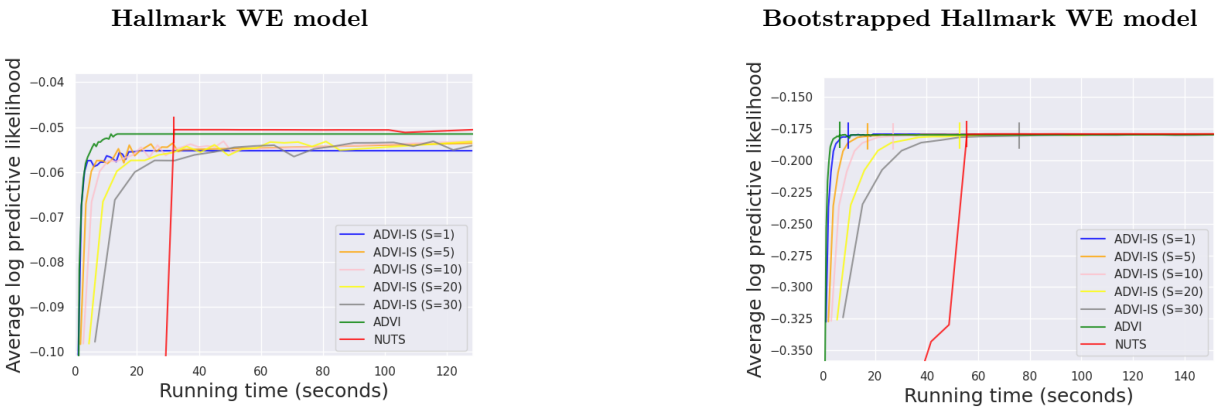

Figure 9: The average log predictive likelihood against the running time in seconds of NUTS and the ADVI-IS, where the line markers indicate that ADVI-(IS) has converged within $\pm$ 2% of the highest average log predictive likelihood of the NUTS. Note that predictive likelihood is calculated using 70% of the observations are used as training data and the remaining 30% as test data. Moreover, we run the models and methods three times using the same seed to obtain an accurate representation of the running time.

5.2.2 Effects of the number of IS repeats S and the blowup λ on ADVI-IS

First, we analyze the posterior covariance and correlation structure to determine whether ADVI-IS improves the posterior covariance and correlation structure of ADVI for the Hallmark models. In Figure 10 we show that the covariance and correlation Frobenius norms are not improved as we increase the number of repeats or the blowup.

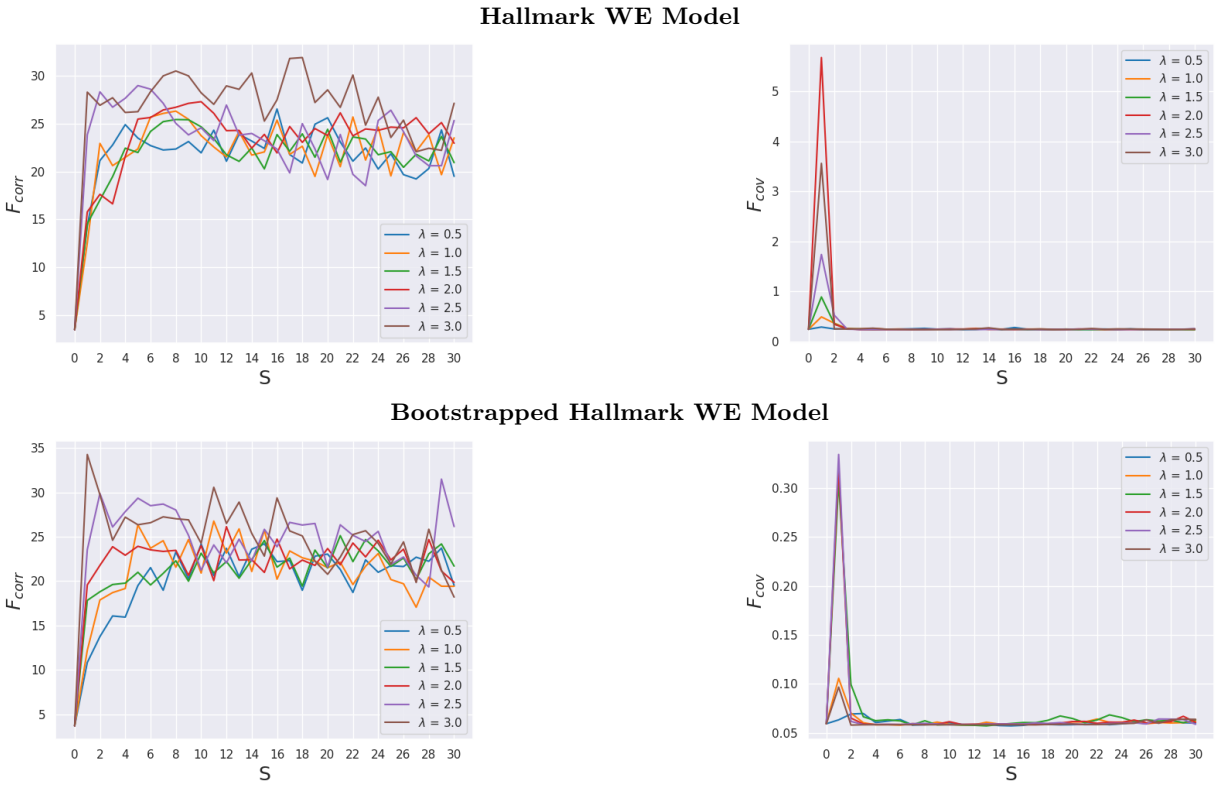

Figure 10: The Frobenius norms of the correlation and covariance matrices between the NUTS and ADVI-IS respectively, where the NUTS is assumed to produce the true covariance and correlation matrices. Note that $S = 0$ corresponds to ADVI without blowup.

In particular, the covariance Frobenius norms shows a high peak for the first IS repetition, whereafter it converges quickly to the same value as ADVI. This indicates that the first repetition causes too high covariances, which is due to first repetition having the most change in the approximate posterior distribution, since it changes from multiple independent Gaussians distributions to a single multivariate Gaussian distribution.

Table 3 shows the optimal blowup and number of repeats for the Hallmark WE models based on the lowest posterior deviation between the approximate posterior distribution and the actual posterior distribution, which are shown in Figure 11. This figure shows that the approximate posterior distribution of ADVI-IS is not able to converge to the actual posterior distribution.

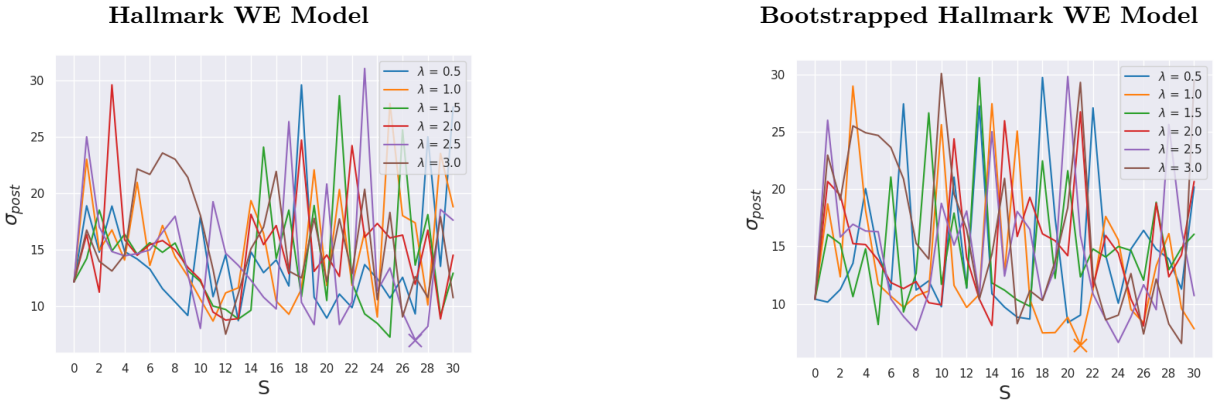

Figure 11: The posterior deviation σ_{post} between the importance function of ADVI-IS and the posterior density function for different blowups and number of repeats, where the cross marker indicates the lowest deviation. Note that $S = 0$ corresponds to ADVI without blowup.

Table 3: Optimal number of IS repeats S^* and blowup λ^* for the Hallmark and bootstrapped Hallmark WE models based on the lowest posterior deviation σ_{post}.

	Model	
	Hallmark WE Model	**Bootstrapped Hallmark WE Model**
S^*	17	21
λ^*	2.5	1.0

The multivariate Gaussian distribution of ADVI-IS is thus not flexible enough to capture the actual posterior distribution of both the bootstrapped and original Hallmark WE models, which can also be seen in Figure 12. Optimally all the normalized weights are in the $[0.8, 1.0]$ bin in the, but almost no normalized weights are in the $[0.8, 1.0]$ bin, while the majority of the weights are in the $[0.0, 0.4]$ bin and a few weights are in the $[2+]$ bin. The approximate posterior distribution thus produces weights with a low probability for large weights and a high probability for small weights, which implies that the tails of the multivariate Gaussian distribution are too light compared to the actual posterior distribution as is discussed by Greenberg (2012). These weights cause the posterior moments obtained from the IS procedure to be mainly based on the few posterior samples with the highest weights, which results in highly inefficient IS repetitions that are not able to converge to the actual posterior distribution.

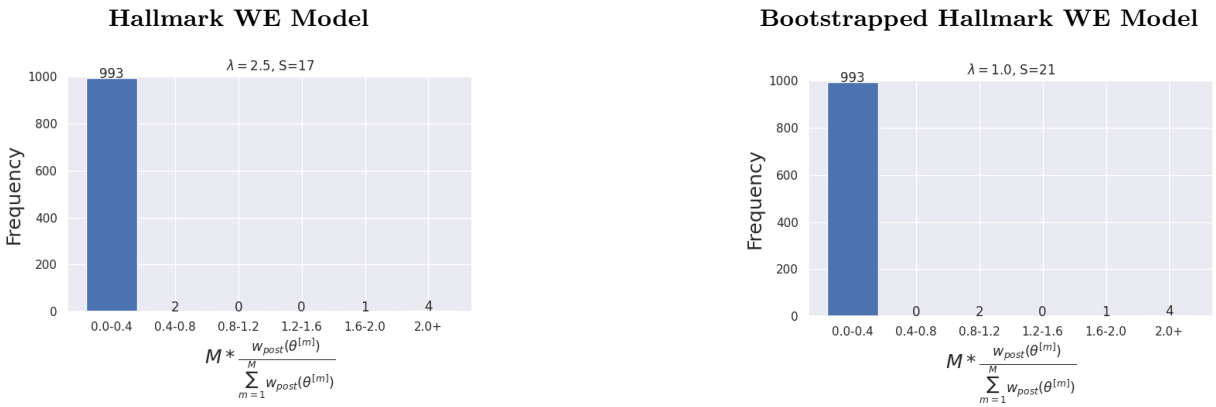

Figure 12: Barplots of normalized weights times the number of posterior samples $M = 1.000$ obtained from the approximate posterior distribution of the optimal ADVI-IS.

5.2.3 Practical performance

Although the approximate posterior distribution of ADVI-IS is not able to capture the actual posterior distribution, we can still show how the optimal ADVI-IS would perform in practice. In particular, first we show in Figure 13 the posterior correlation and covariance of ADVI and the optimal ADVI-IS against those of the NUTS.

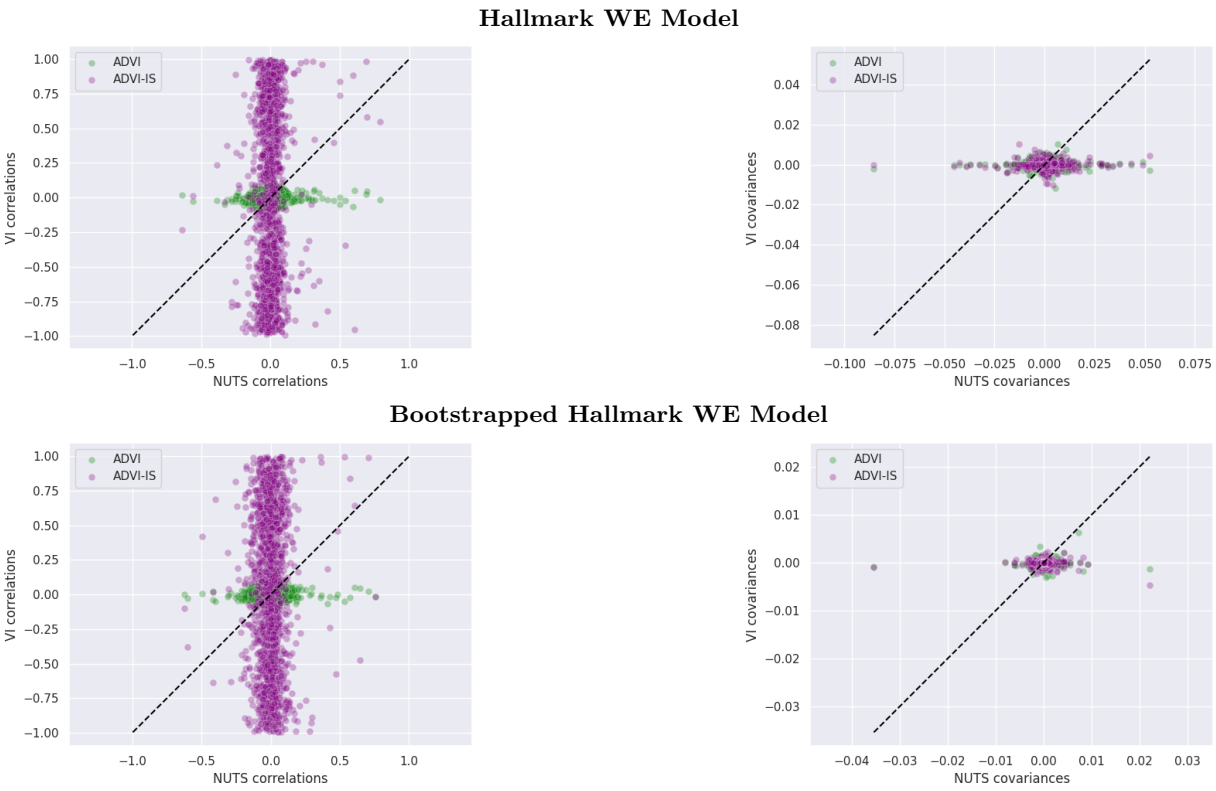

Figure 13: The comparison between the VI and NUTS correlations and the comparison between the VI and NUTS covariances, where the NUTS is assumed to produce the true covariance and correlation matrices. The dashed black line corresponds to a 45 degrees line through (0,0).

Figure 13 shows one of the main reasons that causes ADVI-IS to be unable to capture the posterior distribution, which is the inability to capture the posterior covariance correlation structure of the Hallmark models. These results show that similarly to ADVI, ADVI-IS almost does not have any covariance structure. Moreover, ADVI-IS has some correlation structure, but it tends to overestimate this structure, which causes the relatively high correlation Frobenius norms. Nonetheless, the inability to capture the posterior covariance structure of ADVI and ADVI-IS does not influence the log predictive performance much as seen in Figure 9, since the sample means of all the parameter estimates are still relatively close to the sample means of the NUTS as can be seen in the potential and speed catterpillar plots of Figure 14 and the control catterpillar plots of Figure 20 in Appendix A.2. Moreover, although the potential and speed estimates of ADVI-IS lack a variance structure, we do see that the estimates of ADVI-IS are substantially closer to the estimates of the NUTS compared to the estimates of ADVI. This indicates that ADVI-IS is able to estimate the ME component better than ADVI.

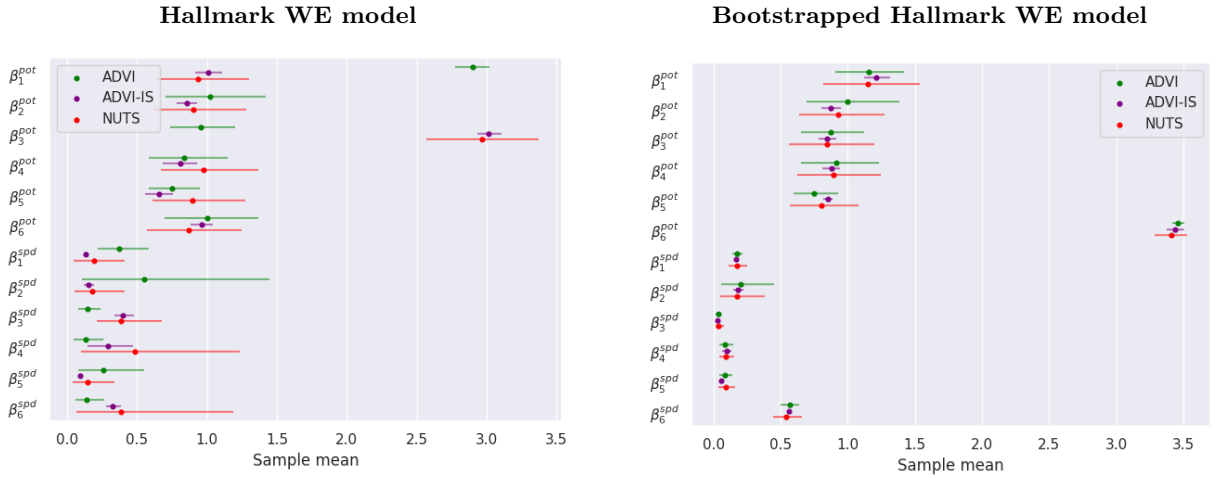

Figure 14: Catterpillar plots of the posterior samples means for the social demographic potential $\boldsymbol{\beta}^{pot}$ and speed $\boldsymbol{\beta}^{spd}$ parameters, where we assume that the NUTS produces the true posterior sample means and the error lines indicate the 5th and 95th quantiles

Furthermore, we analyze the ME component to determine whether ADVI-IS is preferred over ADVI or the NUTS for the Hallmark models, since the ME component is the most difficult and important part to estimate. The catterpillar plots in Figure 14 show three important observations. First, ADVI-IS has little variance for the potential and speed parameters compared to ADVI and the NUTS, which indicates that ADVI-IS focuses on estimating the correct posterior mean, rather than the posterior covariance and correlation structure. Secondly, ADVI has difficulty estimating the potential and speed parameters for the original Hallmark WE model, while ADVI-IS is able to estimate similar values as the NUTS. Thirdly, ADVI, ADVI-IS and the NUTS are all able to estimate similar values for the bootstrapped Hallmark model. The latter two results imply that ADVI has difficulty estimating the ME curves, if the corresponding advertisement exposure variables are sparse. In this case, the repeated IS procedure of ADVI-IS is able to improve the

posterior parameter estimates to be closer to the NUTS estimates.

The potential and speed parameters can be visualized as contour plots as shown in Figure 15. These contour plots clearly show that ADVI-IS focuses on estimating the true posterior mean, while ADVI has difficulty estimating this mean for the *comcast* and *dd* media channels of the original Hallmark model. On the contrary, ADVI is able to find the true posterior mean for the bootstrapped Hallmark model. In particular, ADVI has less dense contour lines than ADVI-IS for the bootstrapped Hallmark WE model, which indicates that the IS repeats cause the approximate posterior distribution to trade off its covariance structure for a more concentrated mean.

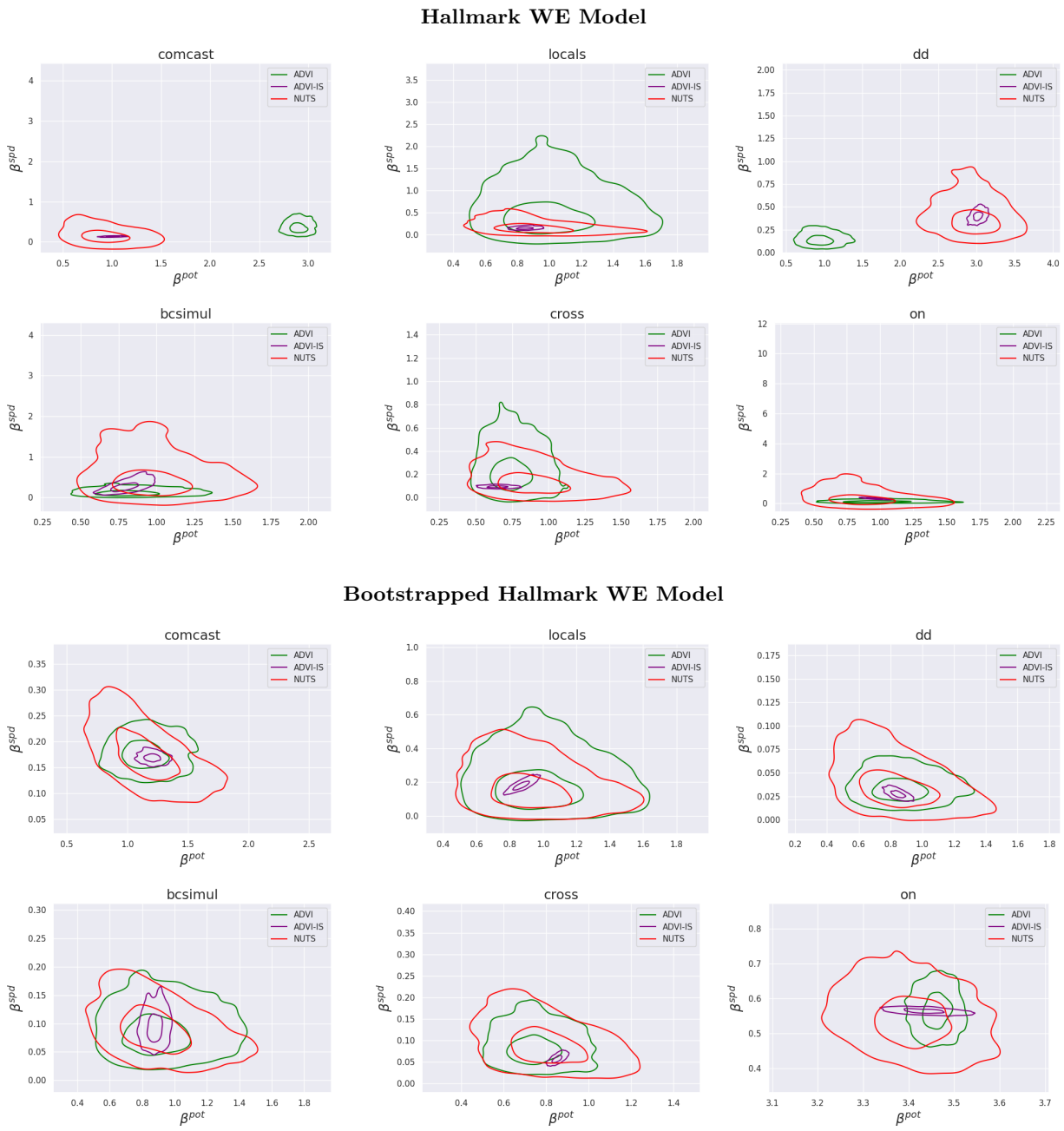

Figure 15: Contour plots of the joint distribution between the potential $\boldsymbol{\beta}^{pot}$ and $\boldsymbol{\beta}^{spd}$ parameters for the six different media channels: *comcast*, *locals*, *dd*, *bcsimul*, *cross* and *on*, where the NUTS is assumed to produce the true joint distributions.

In Figure 16 we show the ME curves of each media channel for the Hallmark models. This figure shows that ADVI is not able to estimate similar ME curve as the NUTS for every media channel of the Hallmark WE model, while ADVI-IS is able to estimate similar ME curve as the NUTS for all media channels except for the *cross* and *bcsimul* media channels. Thus, ADVI-IS outperforms ADVI in terms of estimating the ME curve for the Hallmark WE model. On the other hand, we see that both ADVI and ADVI-IS produce similar ME curves as the NUTS for the bootstrapped Hallmark WE model, where ADVI-IS performs better than ADVI for the *locals* and *cross* media channels. These results imply that the repeated IS procedure substantially improves the estimated ME curves of ADVI, especially if the advertisement exposures are sparse.

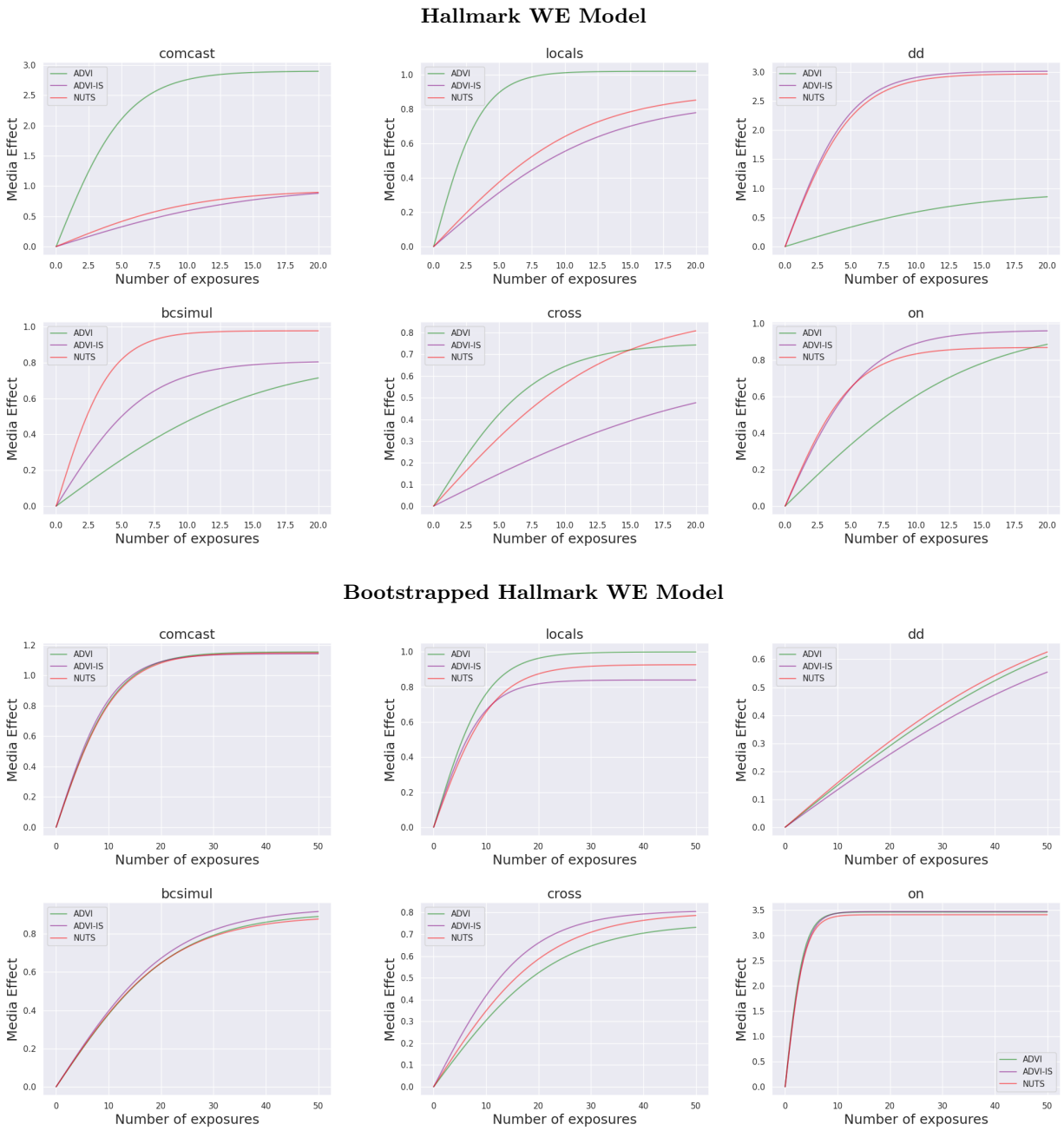

Figure 16: Media effect curves for the six different media channels of the Hallmark and bootstrapped Hallmark WE model, where the NUTS is assumed to produce the true ME curves.

In short, ADVI and ADVI-IS are scalable and could outperform NUTS in terms of convergence rate. However, the Multivariate Gaussian used in ADVI-IS is not complex enough to capture the actual posterior covariance and correlation structure of the Hallmark models. Nonetheless, ADVI-IS is able to reliably estimate the posterior potential and speed parameter estimates. Especially when the corresponding advertisement exposures are sparse, since in this case the repeated IS procedure substantially improves the initial posterior mean of ADVI to be closer to the posterior mean of NUTS, which can also be seen by the superior ME curves of ADVI-IS compared to ADVI for the original Hallmark WE model.

6 Concluding Remarks

In this section we provide the findings of our research. First, a summary of the research and the concluding remarks are provided in Section 6.1. Then limitations and possible extensions of our methods for future research are discussed in Section 6.2

6.1 Conclusion

In this research we investigated whether the optimal approximate posterior distribution of ADVI can be improved by plugging it into a repeated IS procedure. In particular, the main goal is to obtain a method that is faster than the NUTS, while having comparable results to the NUTS. To investigate this we performed a simulation and empirical study to answer the following research question: *To what extent could a repeated IS procedure improve the optimal variational density of ADVI to achieve more accurate posterior estimates comparable to those of the NUTS?*.

The results of the simulation and empirical study show that the running time of the NUTS is more prone to the complexity and scale of the model than ADVI-IS. This implies that ADVI-IS is thus a good potential replacement of the NUTS depending on the performance of the approximate posterior distribution of ADVI-IS.

Moreover, the simulation study shows that increasing the number of IS repeats or slightly increasing the blowup parameter improves the quality of the approximate posterior distribution. In particular, the optimal approximate distribution of ADVI-IS is able to obtain comparable posterior estimates, covariances and correlations to the NUTS for all the simulated models, while ADVI fails to capture the posterior covariance correlation structure due to the mean-field property. The simulation results further show that ADVI-IS outperforms ADVI in terms of estimating the ME component. Specifically, ADVI-IS estimates the ME component comparable to the NUTS. In these simulated models ADVI-IS could thus be used as an efficient replacement of the NUTS.

However, the empirical study shows that ADVI-IS is not able to capture the posterior distribution for both the original Hallmark WE model and the bootstrapped Hallmark WE model. This can be explained by the too light tails of the multivariate Gaussian distribution, which causes large importance weights with low probability and small importance weights with high probability. This implies that the IS repetitions practically only use the few posterior samples with the highest

weights to obtain the posterior moments, which is highly inefficient and thus cause ADVI-IS to not be able to capture the actual posterior distribution.

Nonetheless, the empirical results do show that ADVI-IS estimates the potential and speed parameters of the ME component substantially better than ADVI for the original Hallmark WE model. In particular, ADVI-IS produces ME curves almost identical to the ME curves of the NUTS for all the media channels, except for the *bcsimul* and *cross* media channels. Thus, although ADVI-IS is not able to capture the covariance and correlation structure of the Hallmark WE model, it is at least able to capture the ME component reasonably well.

In conclusion, the repeated IS procedure improves the optimal variational distribution of ADVI by achieving posterior means, covariances and correlations more comparable to the NUTS. In particular, the simulation study shows that ADVI-IS is able to capture the complex posterior distribution of the WE models, while the empirical study shows that ADVI-IS handles sparse advertisement exposures better than ADVI. However, ADVI-IS has difficulties capturing the posterior distribution of the empirical WE models due to the high complexity of these models. Nonetheless, in this case ADVI-IS is still preferred over ADVI, since ADVI-IS is able to accurately estimate the ME component of the WE model. Thus, ADVI-IS could be used to obtain fast preliminary results of the ME component for complex and large scale models, since the running time of ADVI-IS scales better to the complexity and scale of the model compared to the NUTS.

6.2 Limitations and Future Research

This research focuses on efficient estimation methods for Bayesian logistic models, specifically, the Hallmark WE model of Nielsen (Nielsen, 2022). A limitation of this model is the corresponding data set. Although the Hallmark data set is fairly large, the advertisement exposure variables are highly sparse. This sparsity cause the estimation methods to have difficulty learning the nonlinear ME component of the model. To possibly mitigate this problem, we suggest to incorporate a hierarchical structure for each movie of the data set into the logistic model, since this results in more information in the data regarding the ME curve.

Furthermore, the results have shown that the used multivariate Gaussian distribution of ADVI-IS is not flexible enough to capture the actual posterior distribution in the empirical study. We thus propose to further investigate this repeated IS procedure with a more flexible distribution such as the Student-T distribution, which has heavier tails than the Gaussian distribution. Moreover, it could be worthwhile to investigate the use of mixtures of Student-t distributions in this repeated IS procedure, since the target distribution could be multimodal. In particular, Ardia et al. (2009) have already done research on using mixtures of Student-t Distributions in a regular IS procedure.

Acknowledgements

I would like to thank both my company supervisor, Mark den Hollander, and my university supervisor, prof. dr. Paap, for helpful discussions and suggestions during this research.

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Ardia, D., Hoogerheide, L. F., and Van Dijk, H. K. (2009). Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package AdMit. *Journal of Statistical Software*, 29(3):1–32.

Bain, L. J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*, volume 4. Duxbury Press Belmont, CA.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20:28:1–28:6.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339.

Greenberg, E. (2012). *Introduction to Bayesian Econometrics*. Cambridge University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2014). Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, page 121.

Kloek, T. and Van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in Stan. *Advances in Neural Information Processing Systems*, 28.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Nielsen (2022). *Nielsen advertising effectiveness*. `https://www.nielsen.com/apac/en/solutions/advertising-effectiveness` [Accessed: 2022-03-03].

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2018). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026.

A Additional Results

A.1 Simulation Study: Barplots of the Importance Weights

In Figures 17, 18 and 19 we present the normalized importance weights times M for the linear, logistic and logistic with ME models respectively. If the importance function performs well, then the bar in the range [0.8, 1.2] should be the highest, since an optimal importance function implies that $\frac{M w_{post}(\boldsymbol{\theta}^{[m]})}{\sum_{m=1}^{M} w_{post}(\boldsymbol{\theta}^{[m]})} = 1 \; \forall m = 1, \ldots, M$ as discussed in Section 3.4.3.

In particular, these figures show that frequency of and around the [0.8, 1.2] bar increases, as the number of repeats S increases. This implies that the importance function improves as the number of repeats increases. These results thus visualize how the posterior deviation σ_{post} in Figure 3 from Section 4.2.2 decreases, as S increases.

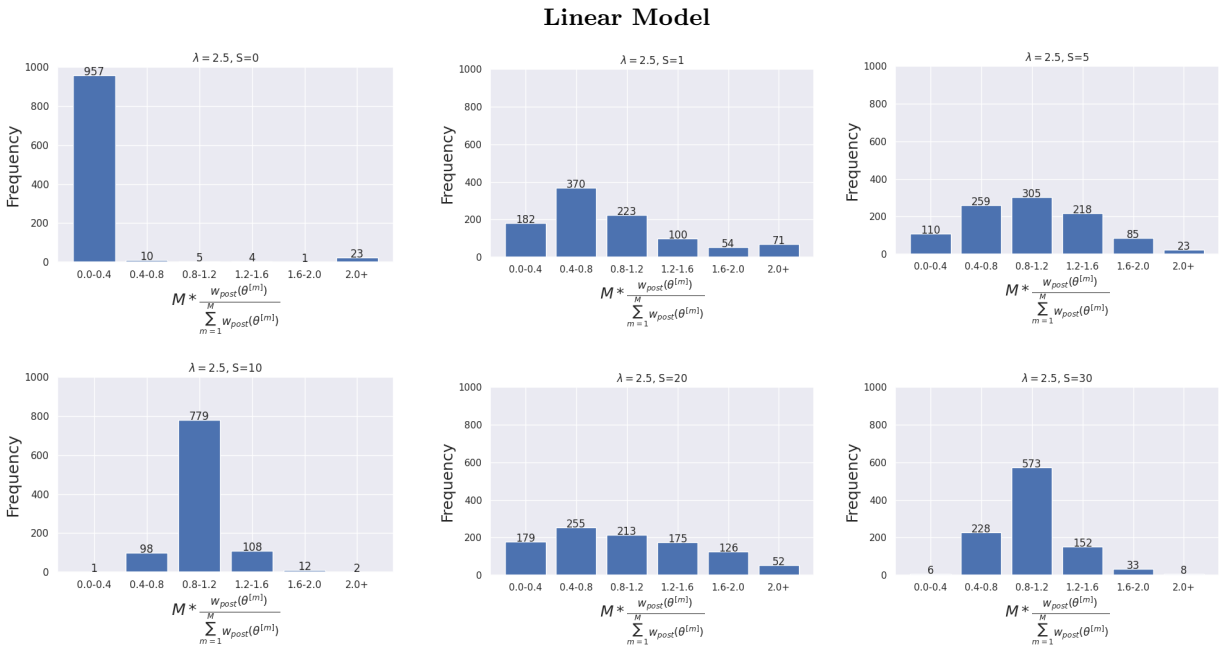

Figure 17: Barplot of the normalized importance weights times the number of posterior samples for the linear model.

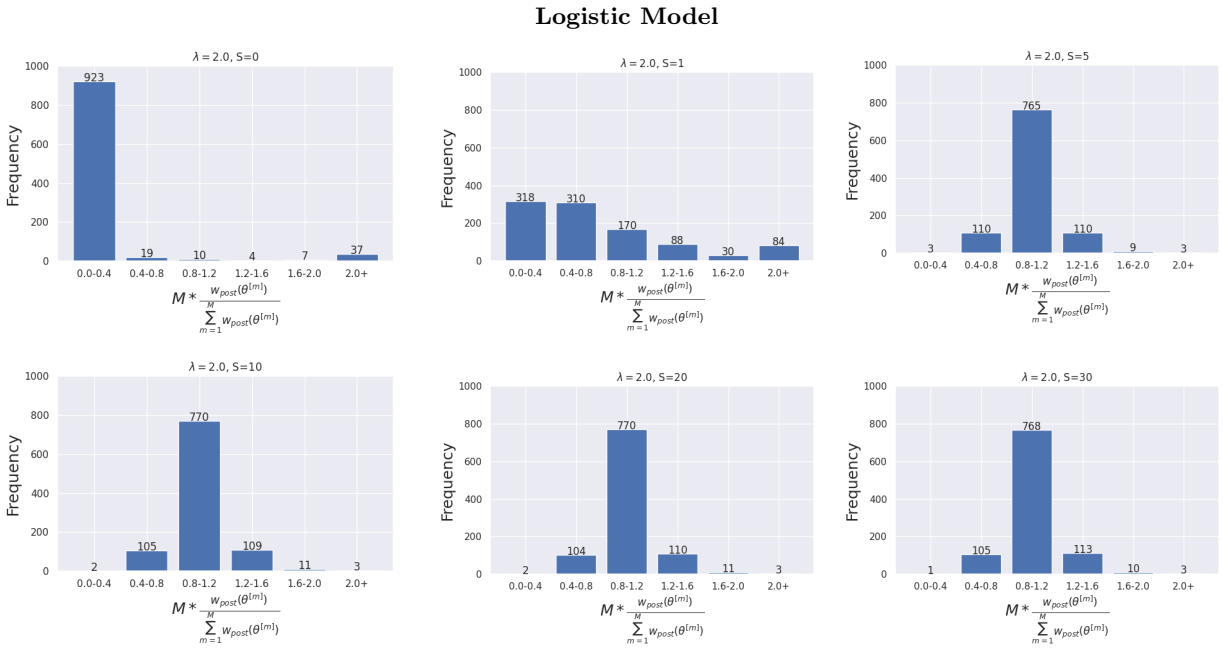

Figure 18: Barplot of the normalized importance weights times the number of posterior samples for the logistic model.

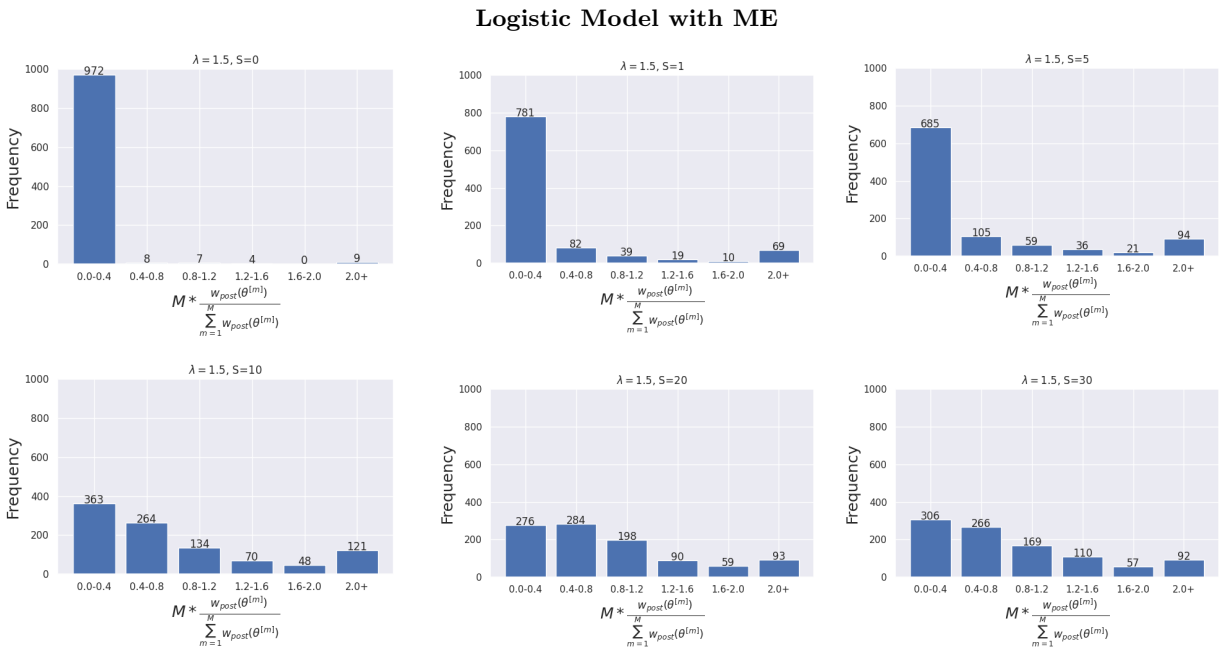

Figure 19: Barplot of the normalized importance weights times the number of posterior samples for the logistic model with ME.

A.2 Empirical Study: Catterpillar Plots of the Control Parameters

The lack of the posterior covariance structure produced by the approximate posterior distribution of ADVI-IS for the Hallmark WE models can be seen in Figure 20, where we barely see any variance of the posterior sample means for the social demographical and viewing behavior parameters.

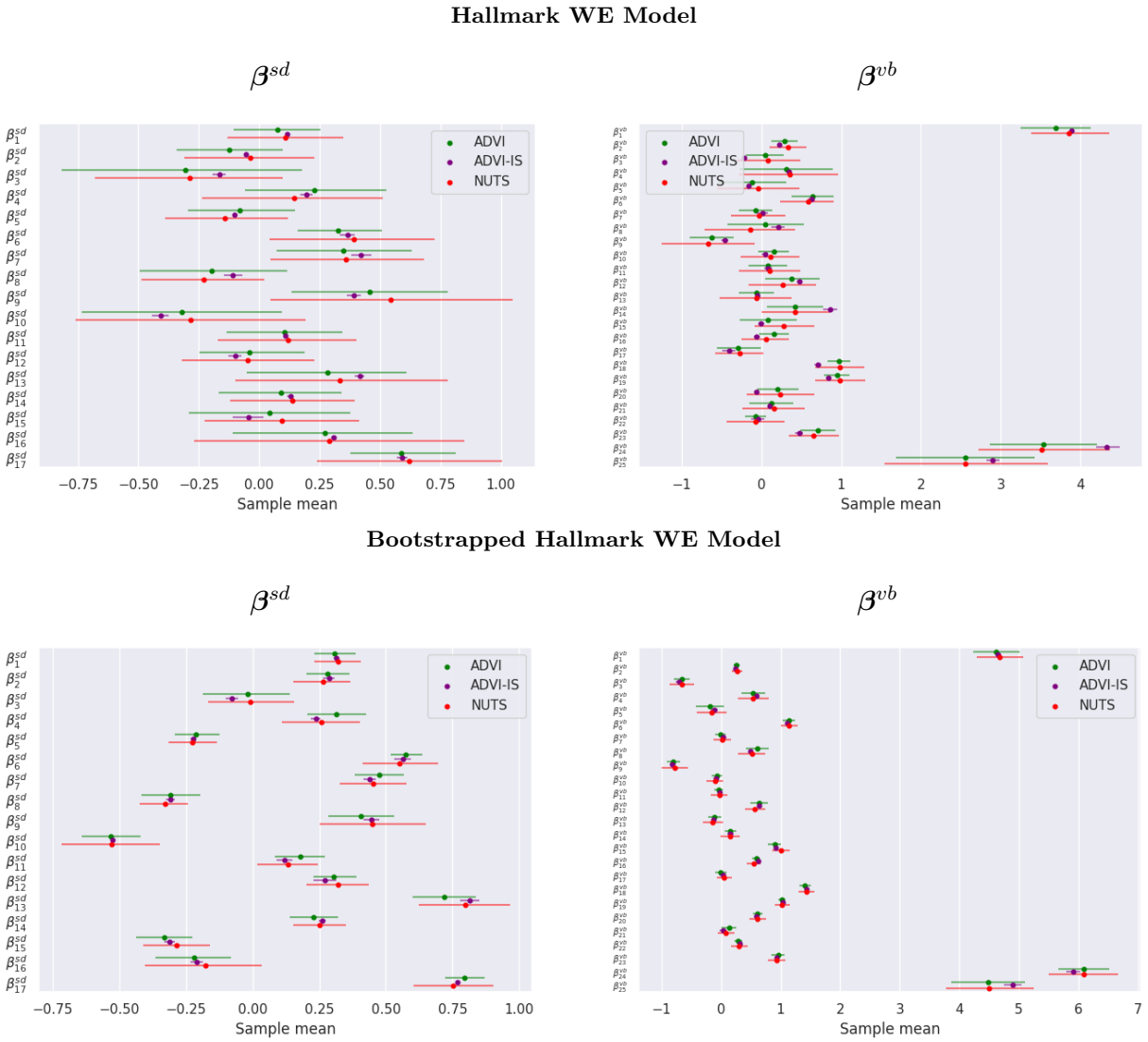

Figure 20: Catterpillar plots of the samples means for the social demographic $\boldsymbol{\beta}^{sd}$ (left column) and viewing behavior $\boldsymbol{\beta}^{vb}$ (right column) parameters, where the error lines indicate the 5th and 95th quantiles.

B Data Description

Table 4: Description of the variables in the social demographics data $\mathbf{x}^{sd}$ in the Hallmark data set.

Social Demographics	
Variable	**Description**
sd_age_[43,54]	Dummy for age between 43 and 54
sd_education_[other]	Dummy for education: other
sd_countysize_[b]	Dummy for the country size: b
sd_countysize_[c]	Dummy for the country size: c
sd_countysize_[d]	Dummy for the country size: d
sd_hhincome_[100k,)	Dummy for a household income of 100.000 or more.
sd_hhincome_[50k,100k]	Dummy for a household income between 50.000 and 100.000
sd_hhsize_[2,4]	Dummy for a household size between 2 and 4
sd_languagespoken_[onlyenglish]	Dummy for the language spoken: only English
sd_languagespoken_[somespanish]	Dummy for the language spoken: some Spanish
sd_majorterritory_[east]	Dummy for major territory: east
sd_majorterritory_[south]	Dummy for major territory: south
sd_majorterritory_[metropolitan]	Dummy for major territory: metropolitan
sd_occupation_[professional/managerial]	Dummy for occupation: professional/managerial
sd_occupation_[unemployed]	Dummy for occupation: unemployed
sd_resprace_[other]	Dummy for ethnicity: other
sd_resprace_[white]	Dummy for ethnicity: white

Table 5: Description of the variables in the viewing behavior data $\mathbf{x}^{vb}$ in the Hallmark data set.

Viewing Behavior	
Variable	**Description**
vb_on_inherited_viewing	Total watching minutes, 15 minutes prior to the premiere on the *Hallmark* channel
vb_put_inherited_viewing	Total watching minutes, 15 minutes prior to the premiere
vb_tercile_group_[midheavy]	Dummy for total watching minutes for the *midheavy* tercile group
vb_total_hallmark_drama_group_[low]	Dummy for total watching minutes of the *Hallmark Drama* channel in the *low* group
vb_total_hallmark_drama_group_[high]	Dummy for total watching minutes of the *Hallmark Drama* channel in the *high* group
vb_total_hmm_group_[low]	Dummy for total watching minutes of the *Hallmark Movies & Mysteries* channel in the *low* group
vb_total_hmm_group_[high]	Dummy for total watching minutes of the *Hallmark Movies & Mysteries* channel in the *high* group
vb_total_dish_group_[low]	Dummy for total watching minutes of the *Dish* channel in the *low* group
vb_total_dish_group_[high]	Dummy for total watching minutes of the *Dish* channel in the *high* group
vb_total_broadcast_cable_group_[low]	Dummy for total watching minutes of the *Broadcast & Cable* channel in the *low* group
vb_total_broadcast_cable_group_[high]	Dummy for total watching minutes of the *Broadcast & Cable* channel in the *high* group
vb_total_directv_group_[low]	Dummy for total watching minutes of the *DirectTV* channel in the *low* group
vb_total_directv_group_[high]	Dummy for total watching minutes of the *DirectTV* channel in the *high* group
vb_total_comcast_group_[low]	Dummy for total watching minutes of the *Comcast* channel in the *low* group
vb_total_comcast_group_[high]	Dummy for total watching minutes of the *Comcast* channel in the *high* group
vb_total_simulmedia_group_[low]	Dummy for total watching minutes of the *Simulmedia* channel in the *low* group
vb_total_simulmedia_group_[high]	Dummy for total watching minutes of the *Simulmedia* channel in the *high* group
vb_total_on_prime_prop_group_[low]	Dummy for the proportion of total watching minutes on the *Hallmark* channel during prime time in the *low* group
vb_total_on_prime_prop_group_[high]	Dummy for the proportion of total watching minutes on the *Hallmark* during prime time in the *high* group
vb_put_weekday_prime_group_[low]	Dummy for the proportion of total watching minutes in weekdays during prime time in the *low* group
vb_put_weekday_prime_group_[mid]	Dummy for the proportion of total watching minutes in weekdays during prime time in the *mid* group
vb_put_weekday_prime_group_[high]	Dummy for the proportion of total watching minutes in weekdays during prime time in the *high* group
vb_other_dummy_hmm	Dummy for watching*other* category on the *Hallmark Movies & Mysteries* channel
vb_prior_pct_hallmark_ln	Prior percentile of watching the *Hallmark* channel for the *lightnon* tercile
vb_prior_pct_hallmark_mh	Prior percentile of watching the *Hallmark* channel for the *mediumheavy* tercile

Table 6: Description of the variables in the advertisement exposures data $\mathbf{x}^{exp}$ of the Hallmark data set.

Advertisement Exposures	
Variable	**Description**
exp_cross_before_general	Exposures for the Hallmark channel via the *cross* media channel during the whole campaign period
exp_cross_before_specific	Exposures for the specific movie via the *cross* media channel during the whole campaign period
exp_cross_premiereweek_general	Exposures for the Hallmark channel via the *cross* media channel a week before the premier of the movie
exp_cross_premiereweek_specific	Exposures for the specific movie via the *cross* media channel a week before the premier of the movie
exp_dd_before_general	Exposures for the Hallmark channel via the *dd* media channel during the whole campaign period
exp_dd_before_specific	Exposures for the specific movie via the *dd* media channel during the whole campaign period
exp_dd_premiereweek_general	Exposures for the Hallmark channel via the *dd* media channel a week before the premier of the movie
exp_dd_premiereweek_specific	Exposures for the specific movie via the *dd* media channel a week before the premier of the movie
exp_bcsimul_before_general	Exposures for the Hallmark channel via the *bcsimul* media channel during the whole campaign period
exp_bcsimul_premiereweek_general	Exposures for the Hallmark channel via the *bcsimul* media channel a week before the premier of the movie
exp_bcsimul_premiereweek_specific	Exposures for the specific movie via the *bcsimul* media channel a week before the premier of the movie
exp_comcast_before_general	Exposures for the Hallmark channel via the *comcast* media channel during the whole campaign period
exp_comcast_premiereweek_general	Exposures for the Hallmark channel via the *comcast* media channel a week before the premier of the movie
exp_comcast_premiereweek_specific	Exposures for the specific movie via the *comcast* media channel a week before the premier of the movie
exp_locals_before_general	Exposures for the Hallmark channel via the *locals* media channel during the whole campaign period
exp_locals_before_specific	Exposures for the specific movie via the *locals* media channel during the whole campaign period
exp_locals_premiereweek_general	Exposures for the Hallmark channel via the *locals* media channel a week before the premier of the movie
exp_locals_premiereweek_specific	Exposures for the specific movie via the *locals* media channel a week before the premier of the movie
exp_on_before_general	Exposures for the Hallmark channel via the *on* media channel during the whole campaign period
exp_on_before_specific	Exposures for the specific movie via the *on* media channel during the whole campaign period
exp_on_premiereday_general	Exposures for the Hallmark channel via the *on* media channel a day before the premier of the movie
exp_on_premiereday_specific	Exposures for the specific movie via the *on* media channel a day before the premier of the movie
exp_on_premiereweek_general	Exposures for the Hallmark channel via the *on* media channel a week before the premier of the movie
exp_on_premiereweek_specific	Exposures for the specific movie via the *on* media channel a week before the premier of the movie

Table 7: Description of the combined advertisement exposures variables.

Combined Advertisement Exposures	
Variable	**Applied Transformation**
exp_cross	exp_cross_before_general + exp_cross_before_specific + exp_cross_premiereweek_general + exp_cross_premiereweek_specific
exp_dd	exp_dd_before_general + exp_dd_before_specific + exp_dd_premiereweek_general + exp_dd_premiereweek_specific
exp_bcsimul	exp_bcsimul_before_general + exp_bcsimul_premiereweek_general + exp_bcsimul_premiereweek_specific
exp_comcast	exp_comcast_before_general + exp_comcast_premiereweek_general + exp_comcast_premiereweek_specific
exp_locals	exp_locals_before_general + exp_locals_before_specific + exp_locals_premiereweek_general + exp_locals_premiereweek_specific
exp_on	exp_on_before_general + exp_on_before_specific + exp_on_premiereday_general + exp_on_premiereday_specific + exp_on_premiereweek_general + exp_on_premiereweek_specific

Table 8: The different networks that each media channels represents.

Media Channel	**Networks**
comcast	The Comcast Corporation
locals	All the local networks in its region
dd	Dish media or DirecTV
bcsimul	Broadcast cable or Simulmedia
cross	The Hallmark Mysteries & Movies channel or the Hallmark Drama channel
on	The Hallmark channel

C Code Description

Table 9: Python classes for the implemented methods and NumPyro models

Methods	**Description**
numpyro_base.py	Base class which all the method classes inherit from
numpyro_nuts.py	Class that implements the NUTS via NumPyro
numpyro_advi.py	Class that implements the ADVI via NumPyro
numpyro_advi_is_multi_normal.py	Class that inherits **numpyro_advi.py** by adding the repeated IS procedure for a Multivariate Gaussian
NumPyro Models	**Description**
linear_covariance.py	Simulated linear model in NumPyro
logit_covariance.py	Simulated logistic model in NumPyro
logit_covariance_me.py	Simulated logistic model with ME in NumPyro
hallmark_WE.py	The Hallmark WE model with empirical data in NumPyro
models.py	Class that initializes all the above models in classes

Table 10: Main and run classes to configure and run the methods, models and plots.

Main Classes	**Description**
main.py	Main class to implement the methods on all models
main_time.py	Main class to measure the time performances of the methods for all models
main_computation.py	Main class to obtain the results from the trained models from **main.py**
main_plot.py	Main class to obtain the relevant plots using the results from **main_computation.py**
Run Classes	**Description**
run.py	Class to run **main.py**
run_computation.py	Class to run **main_computation.py**
run_plot.py	Class to run **main_plot.py**
run_time.py	Class to run **main_time.py**

Table 11: Utility classes.

Utility Classes	**Description**
utils.py	Class containing helper functions for everything
utils_plot.py	Class containing helper functions for plotting purposes

D Abbreviations

Table 12: Description of the abbreviations used in this research.

Abbreviation	**Definition**
Nielsen	The Nielsen Company
NPM	Nielsen People Meter
ADVI	Automatic Differentiation Variational Inference
CAVI	Coordinate Ascent Variational Inference
SVI	Stochastic Variational Inference
KL	Kullback-Leibler
ELBO	Evidence Lower BOund
MCMC	Markov Chain Monte Carlo
HMC	Hamiltonian Monte Carlo
NUTS	No-U-Turn Sampler
IS	Importance Sampling
WE	Watch Effect
ME	Media Effect