FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Seminar Paper: Bayesian Econometrics
Scalable Price Targeting
Fall, 2018

Ditlev Kiersgaard Frisch

Frederik Bjørn Melchiorsen

Supervisor: Rémi Piatek

ECTS points: 7.5

Date of submission: 30/11/2108

Keystrokes: 52,354

Abstract

We explore the use of the Weighted Likelihood Bootstrap (WLB) by Newton and Raftery (1994) to approximately simulate from a posterior distribution. Specifically, we take as a starting point the working paper by Dubé and Misra (2017) that uses the WLB method to simulate an approximate posterior of structural parameters in a demand model via a (frequentist) logistic Lasso in order to estimate a posterior distribution of own-price elasticities in a price field experiment, which they employ to set prices. We simulate a similar price experiment and compare the method to a fully Bayesian implementation, namely via the Bayesian Lasso of Park and Casella (2008) augmented with a latent-utility representation of the probit model. When N is large, we find that the Bayesian Lasso generally performs better than the WLB approach in terms of fit as its posterior implied optimal prices are much more narrowly concentrated around the true optimal prices. With low N, however, the WLB performs better as the Bayesian Lasso concentrates its implied prices too low. The posterior means and medians of implied optimal prices are more aligned, and if one employs a pricing decision based on either the posterior mean or mode, both methods perform more or less equally well independent of the sample size.

Code

All code is written in R and can be accessed on Github via the following link

https://github.com/DitlevF/Scalable-Price-Targetting

Contribution by the authors

We sincerely believe that all elements of this project are the work of both authors, where derivations and coding have been done collectively. We therefore find it most natural to be judged together. However, if required, Ditlev has been relatively more in charge of the WLB and Bayes factors (Sections 2.1 and 3), whereas Frederik has been relatively more in charge with conducting the simulation study in Section 4. Ditlev was relatively more in charge of coding up the Bayesian Lasso, whereas Frederik was relatively more in charge of the derivations in Section 2.2. Ditlev has been relatively more in charge of the introduction (Section 1), whereas the conclusion is relatively more the work of Frederik (Section 5).

Contents

1 Introduction and motivation (Ditlev) 3

1.1 High-dimensional methods: Frequentist and Bayesian . . . 3

1.2 Pricing Right in High Dimensions: A Field Study . . . 4

1.3 Set-up . . . 5

1.4 Simulating the data . . . 6

2 WLB and the Bayesian Lasso 7

2.1 Weighted Likelihood Bootstrap (Ditlev) . . . 7

2.1.1 Theory . . . 7

2.1.2 Sampling the weights: Idea and proof . . . 8

2.1.3 The Algorithm . . . 10

2.2 The Bayesian Probit-augmented Lasso (Frederik and Ditlev) . . . 10

2.2.1 The conditional posteriors . . . 11

2.2.2 The sampling scheme . . . 15

3 Bayes Factors (Ditlev) 15

3.0.1 Calculating the conditional likelihood $p(x \mid \theta^{(i)})$ . . . 17

3.0.2 Implementation in R . . . 17

4 Simulation Study (Frederik) 17

4.1 Diagnostic test . . . 18

4.2 Mean-squared error . . . 18

4.3 Profit-maximizing prices . . . 21

4.4 Bayes Factors . . . 24

4.5 Summing up . . . 24

4.6 Practical implications of our findings . . . 25

5 Conclusion (Frederik) 26

6 Appendix 26

6.1 Bayes Factor: Implementation in R . . . 26

6.2 MSE Elasticities for $N = 1,000$, signal-to-noise 3 . . . 27

1 Introduction and motivation (Ditlev)

1.1 High-dimensional methods: Frequentist and Bayesian

The dual increase in computing power and the availability of data with a high-dimensional covariate space, so-called "big data", has led to a surging literature - both frequentist and Bayesian - on how to estimate model parameters when the dimension of the covariate space p is large relative to the number of observations N. The main underlying issue in such cases is that using standard methods (such as regression) leaves model parameters to be estimated with a large variance, and in terms of prediction therefore performs poorly on new data as it "overfits" the sample.

In the frequentist literature, one prominent example of how this has been dealt with in the canonical linear regression model is the Least Absolute Shrinkage and Selection Operator (Lasso) of Tibshirani (1996) that penalizes coefficients for having too large absolute values, which forces some coefficients to be set exactly to zero thus reducing the dimension of p and the variance of the estimators at the expense of a minor bias. Introducing penalty terms into loss functions (e.g. sum of squared residuals) has since Tibshirani (1996)'s seminal paper become a fairly generic recipe for many frequentist methods to work with high-dimensional data, see e.g. Fan and Li (2001), Zou (2006), Zou and Hastie (2005). The deserved popularity of this way of working is due to the extremely efficient optimization algorithm, least angle regression (LARS), that scales very well on high-dimensional data as well as the induced variable selection that happens as coefficients are set exactly to zero due to the "kink" at zero for its penalty term.

Bayesians, on the other hand, have tackled the issue by shrinking coefficients towards zero through their prior. For instance, the Lasso's estimates can be interpreted as the posterior mode estimates when the regressors are given iid Laplace priors with mean zero and a constant scale parameter, while Ridge estimates (a frequentist estimator that similar to the Lasso penalizes coefficients for having large squared values) can be interpreted as the posterior mode under a $N(0, c)$ prior. Other prominent examples are the so-called spike-and-slab models (with no frequentist analogue) that put a mixture prior on the regressors with a point mass at zero (the "spike") and a continuous distribution (the "slab"), where weak regressors are pulled towards the spike distribution, see e.g. Mitchell and Beauchamp (1988) and George and McCulloch (1993).

In this paper, we reconstruct a price field study example where frequentist high-dimensional methods were used in conjunction with bootstrap methods to obtain a quasi-posterior distribution of the model parameters and thereby predict demand, and we compare their approach to a fully Bayesian implemention via the popular Bayesian Lasso of Park and Casella (2008). Ultimately, we are interested in assessing to what extent these bootstrap methods provide reasonable approximations to a Bayesian

posterior in a practical setting, and whether one method is prefered over the other.

1.2 Pricing Right in High Dimensions: A Field Study

In their paper, Dubé and Misra (2017), the authors face an estimation problem in a high-dimensional space. The authors conduct a pricing field experiment for the online firm Ziprecruiter.com in order to optimize prices using many characteristics of their customers to estimate the price elasticities. About Ziprecruiter, they explain:

> *"Ziprecruiter.com is an online firm that specializes in matching jobseekers to potential employees. The firm caters to a variety of potential employers across various industries that subscribe to Ziprecruiter.com to gain access to a stream of resumes of matched and qualified candidates from which they might be able to recruit. These firms pay a monthly subscription rate that they can cancel at anytime. Job applicants can use the Ziprecruiter.com platform for free. In a typical month in 2015, Ziprecruiter hosted job postings for over 40,000 registered paying firms."*

In the price experiment, the authors and the management of the firm randomly allocate prices, the monthly subscription fee, to the different firms over a one month period from which they also obtain many firm characteristics (industry, location, desired skills of the job applicant etc.). They seek to use these characteristics to estimate the willingness-to-pay conditional on firm characteristics using the exogenous variation in prices and the 0-1 decision to buy/not buy given the allocated price, where they allow for interaction effects and higher order polynomials to capture heterogeneity of the firms (who knows, perhaps medical firm from southern states looking for research staff have a higher willingness-to-pay than in the northern states). This leaves them with a high-dimensional covariate space and limited observations, wherefore they turn to regularization methods to avoid overfitting the data and in order to estimate the parameters precisely. Specifically, they set up a structural demand model where the utility from buying the product is a function of the firm characteristics and the prices, and as the observed outcome is binary, they estimate the model via a logistic lasso. The logistic lasso estimates $\hat{\beta}$ are found by minimizing the negative log-likelihood

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0 + X_i\beta) - \log(1+\exp(\beta_0 + X_i\beta)\} + \lambda \parallel \beta \parallel_1$$

where $\parallel \beta \parallel_1$ is the l_1 norm of $\beta = (\beta_1, ..., \beta_p)$ and λ is the tuning parameter that controls complexity of the model (a higher λ, a sparser model), which is usually estimated via cross-validation.

However, one caveat from using frequentist high-dimensional methods is that standard errors are not well-defined, especially on derived parameters such as the own-price elasticity. Setting a price based on one maximum-likelihood estimate without quantifying the uncertainty of the estimate may

be dangerous in applied situations. This is not a problem for Bayesians: a Bayesian analysis always provides a full posterior of the variables of interest, and it is straightforward to simulate a posterior of e.g. an elasticity given the posterior of the regressors. As the authors are not interested in doing a Bayesian analysis (as it may not scale well in very high-dimensional spaces, they argue), they therefore turn to the Bayesian bootstrap idea of Rubin (1981) and its likelihood extension in Newton and Raftery (1994), where the likelihood function (here, the logistic lasso) is augmented with a weight vector that is simulated over B bootstraps in order to approximate the posterior distribution of the model parameters, the so-called method of Weighted Likelihood Bootstrap (WLB).

Our motivation for this paper is two-fold: First, we are interested in investigating this quite innovative method in using high-dimensional econometric methods to a business case - in this case, optimizing the price for Ziprecruiter.com - and second, to explore how well the WLB approach approximates the posterior distribution and helps quantify the uncertainty of the estimates in an applied situation. To investigate this, we choose the popular Bayesian Lasso of Park and Casella (2008) as comparison, where we compare its posterior distribution to the WLB approximation, and investigate implied profits and fit for the two methods.

For the rest of this section we briefly describe the model and experiment set-up. In Section 2 we introduce the WLB and Bayesian Lasso and their corresponding implementation. In Section 3 we introduce the concept of Bayes factors that will be used to assess model fit. Section 4 conducts the simulation study. Section 5 concludes.

1.3 Set-up

We simulate the price experiment of Dubé and Misra (2017) where we randomly allocate prices to $N = \{1,000; 10,000\}$ customers from whom we observe a vector of characteristics that characterizes tastes and hence willingness to pay. The data-generating process (DGP) is such that the analyst observes

$$D = \{q_i, x_i, p_i\}_{i=1}^{N}$$

where q_i is a vector of purchase quantities (binary), p_i is the price offered to customer i, and $x_i \in \mathbb{R}^K$ is a high-dimensional vector of customer characteristics. We assume the following latent utility demand model

$$\Delta U_i = \alpha_i + \beta_i p_i + \epsilon_i = \alpha(x_i; \theta_\alpha) + \beta(x_i; \theta_\beta) p_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

where $q_i = 1$ if $\Delta U_i \geq 0$, 0 otherwise. The functional forms of $\alpha(\cdot)$ and $\beta(\cdot)$ are assumed linear in the covariates, such that $\alpha(x_i; \theta_\alpha) = x_i\theta_\alpha$ and $\beta(x_i; \theta_\beta) = x_i\theta_\beta$. The goal for the analyst is then

to estimate the structural parameters of the model in order to estimate price-elasticities to find an optimal uniform price that maximizes profits. As ΔU_i is latent and we only observe q_i, the natural choice in the above model is the probit model, but one could just as well assume a logistic error term and go with a logistic model.

Following Genc et al. (2016), the price elasticities are calculated discretely at price p_0 as

$$\hat{\epsilon}_{p_0} = 100 \times \sum_{i=1}^{N} \frac{\mathbf{1}(\Delta \hat{U}_i \mid_{p_0 \times 1.01} > 0) - \mathbf{1}(\Delta \hat{U}_i \mid_{p_0} > 0)}{\mathbf{1}(\Delta \hat{U}_i \mid_{p_0} > 0)} \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Due to zero marginal costs of the information good, the optimal uniform price can then be calculated as either the price p_0 for which $\epsilon_{p_0} = -1$ or simply as

$$\arg\max_{p_0} \sum_{i=1}^{N} \mathbf{1}(\Delta \hat{U}_i \mid_{p_0} > 0) p_0 \tag{3}$$

1.4 Simulating the data

We seek to approximate the real data in Dubé and Misra (2017), and therefore simulate the covariates as a mixture of four normals

$$x_i \sim \sum_{j=1}^{4} \rho_j N(\boldsymbol{\mu}_j, \Sigma), \; \sum_{j=1}^{4} \rho_j = 1 \tag{4}$$

where the location varies, but the scale is constant over the groups, and the correlation between x_i and x_j is $0.5^{|i-j|}$ as in Tibshirani (1996). For the structural parameters, we let θ_α be a 50×1 vector with 14 non-zero positive entries in the range of 0.2 to 1 and θ_β be a 50×1 vector with 10 non-zero negative entries in the range of -0.002 to -0.007. We also simulate the error terms from a $N(0, \sigma^2)$, where we vary σ^2 in order to obtain different signal-to-noise ratios[1]. As the covariance structure is not straightforward with price interactions, we approximate the signal-to-noise ratio by its empirical analogue $\hat{\text{Var}}(x\theta_\alpha + x\theta_b \boldsymbol{p})/\hat{\text{Var}}(\boldsymbol{\epsilon})$. We then randomly allocate prices for one product to N customers ranging from $p = 19$ to $p = 399$. A customer then buys the product if $\Delta U_i > 0$. Doing so, we obtain the following conversion rates in Table 1.1 - that is, the percentage of customers who buy the product at the given price - which correspond approximately to the conversion rates Dubé and Misra (2017) obtained in their field experiment.

The generic code for simulating this data set can be found here. Due to zero marginal costs, the model-free analysis would in this case put the price somewhere between 159-249 (or, perhaps, at a price above 399). However, a model-based analysis is needed for an exact value. In the following section we

[1] We also experimented with simulating the error terms from a logistic distribution, as the WLB approach uses a Logistic Lasso and the Bayesian Lasso approach is augmented with a probit model, but the results are very similar, and we therefore only present the results for normal error terms.

Table 1.1: Conversion Rates

Price	19	39	59	79	99	159	199	249	299	399
Rate	0.56	0.36	0.30	0.31	0.29	0.27	0.19	0.09	0.05	0.05

take the time to first thoroughly introduce the WLB Logistic approach employed by Dubé and Misra (2017) to sample an approximate posterior and then introduce the Bayesian Probit-augmented Lasso. Both of these methods are then used to estimate the model parameters in (1), provide predictions on $\Delta\hat{U}_i$ and subsequently the implied elasticities, prices and profits via (2) and (3).

2 WLB and the Bayesian Lasso

2.1 Weighted Likelihood Bootstrap (Ditlev)

2.1.1 Theory

Newton and Raftery (1994) developed the WLB for Bayesian inference as an alternative to other numerical integration methods, such as Markov Chain Monte Carlo (MCMC) algorithms, when optimization of the likelihood function is feasible. To sample approximately from the posterior, Newton and Raftery (1994) posit that maximization of the weighted likelihood function (if one such exists)

$$\tilde{L}(\theta) = \prod_{i=1}^{N} f_i(x_i;\theta)^{w_i},\ x_i, i = 1, ..., N \text{ are iid} \tag{5}$$

over different weights simulated e.g. from a Dirichlet distribution, produces samples from an approximate posterior which corresponds to a data-dependent prior, that is, no prior at all. To embody a fully Bayesian analysis, naturally a prior has to be specified from which it can be updated to the posterior via the likelihood. To this end, they propose to use the Sampling-Importance-Resampling (SIR) of Rubin (1987). The whole process is summarized in the following scheme

1. Sample the weight vectors $\{w_b\}_{b=1}^{B}$, where B is the number of bootstraps and $w_b = (w_{1b}, ..., w_{Nb})$.
2. Maximize (5) over all $\{w_b\}_{b=1}^{B}$ and obtain $\{\tilde{\theta}_b\}_{b=1}^{B}$.
3. Calculate a kernel density estimate of $\tilde{\theta}$, $\hat{g}(\theta)$ at each $\tilde{\theta}$, as an approximation to the marginal posterior density of θ.
4. Resample $\{\tilde{\theta}_b\}_{b=1}^{B}$ from a discrete distribution without replacement with the probability proportional to its importance ratio $\frac{\pi(\tilde{\theta}_b)L(\tilde{\theta}_b)}{\hat{g}(\tilde{\theta}_b)}$, where $\pi(\cdot)$ is the prior on θ.
   (a) That is, just like importance sampling for evaluating an integral, each observation $\tilde{\theta}_b$ is weighted according to its target distribution $\pi(\tilde{\theta}_b)L(\tilde{\theta}_b)$ (the un-normalized posterior) over

the approximate starting distribution $\hat{g}(\tilde{\theta}_b)$.

Using only the first two steps provides only an asymptotic first-order approximation to the posterior of interest, as no prior information is incorporated, leaving the method ultimately non-Bayesian. Higher order approximations involve the prior: The last two steps correct for this, and provides an approximate posterior of the parameter of interest that depends on how stable the importance weights are and the precision of the kernel density estimate.

The implementation of the WLB in Dubé and Misra (2017) only uses the first two steps, and hence no priors are specified, which alters the analysis fundamentally non-Bayesian. It still, however, provides some sort of posterior distribution of the regressors that aids them in quantifying the uncertainty around the estimates, which is also the approach we implement. In the remainder of this paper, we dub this type of posterior a "quasi-posterior" to make explicit that this posterior does not involve a prior. The open question is then how the weights should be simulated. To sample the weights, they build on ideas by Taddy et al. (2016), and in the following subsection we will introduce and derive how this is done. In effect, as we will show, the idea is to use the Dirichlet weighting proposed as a possibility in Newton and Raftery (1994) and let the prior probability of observing an observation tend to zero.

2.1.2 Sampling the weights: Idea and proof

In Dubé and Misra (2017) and Taddy et al. (2016), the DGP is represented through a probability function on a large, finite number of potential data points $\boldsymbol{D} = (D_1, ..., D_N)$

$$p(D \mid \boldsymbol{V}) = \frac{1}{\mid V_l \mid} \sum_{l=1}^{L} V_l \mathbf{1}(D = \zeta_l) \tag{6}$$

where $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_L)$ are the support points of the DGP, $\mathbf{V}$ some weights, and in the price experiment $D_i = (x_i, q_i, p_i)$. With a limiting Dirichlet-multinomial prior that puts zero prior probability mass on support points not observed in the data, it can be shown that each data point D is realized in the posterior as

$$p(D \mid \mathbf{D}) = \frac{1}{\sum_i^N V_i} \sum_{i=1}^{N} V_i \mathbf{1}(D = D_i), \qquad V_i \sim \text{Exp}(1) \tag{7}$$

and thus, in each bootstrap, we can simulate the weights for the likelihood from an Exp(1) distribution by properly scaling the draws, such that each observation is given a $V_i / \sum_i V_i$ weight.

Proof

Define the Dirichlet-Multinomial hierarchy as

$$\begin{aligned} D_i \mid \boldsymbol{\phi} &\sim \text{Discrete}(\phi_1, ..., \phi_L),\ i = 1, ..., N \\ \boldsymbol{\phi} &\sim \text{Dirichlet}(a_1, ..., a_L) \end{aligned} \tag{8}$$

where ϕ_l is the prior probability of observing support-point ζ_l. Due to conjugacy, the posterior of $\boldsymbol{\phi}$ is derived as

$$p(\boldsymbol{\phi} \mid \mathbf{D}) \propto p(\mathbf{D} \mid \boldsymbol{\phi}) p(\boldsymbol{\phi}) \propto \left[\phi_1^{\sum_{i=1}^N \mathbf{1}(D=\zeta_1)} ... \phi_L^{\sum_{i=1}^N \mathbf{1}(D=\zeta_L)} \right] \times \prod_{l=1}^{L} \phi_l^{a_l - 1} \propto \prod_{l=1}^{L} \phi_l^{\sum_{i=1}^N \mathbf{1}(D=\zeta_l) + a_l - 1}$$

which is clearly a Dirichlet posterior. Thus, $\boldsymbol{\phi}$ is sampled from the L-dimensional posterior

$$\boldsymbol{\phi} \mid \mathbf{D} \sim \text{Dirichlet}_L \left(a_1 + \sum_{i=1}^{N} \mathbf{1}(D = \zeta_1), ..., a_L + \sum_{i=1}^{N} \mathbf{1}(D = \zeta_L) \right) \tag{9}$$

Now, as we consider the limiting case for $a_l \to 0$ for all l, we get from (9) that for all support-points that are not observed in the data (i.e. those points ζ_q for which $\sum_{i=1}^N \mathbf{1}(D = \zeta_q) = 0$) that the posterior probability of observing it is zero, $\phi_q = 0$, which means we only need to sample from the observed data points[2], and hence (9) can equivalently be written as

$$\boldsymbol{\phi} \mid \mathbf{D} \sim \text{Dirichlet}_N(1, 1, ..., 1) \tag{10}$$

which is an N-dimensional posterior ($N \leq L$) that allows for duplet support points, $\zeta_l = \zeta_k$ for $l \neq k$. To finish exposition, we use the general result that any random vector $\mathbf{x} = (x_1, ..., x_K)$ from a Dirichlet$(\alpha_1, ..., \alpha_K)$ can be sampled from K independent draws from a Gamma distribution by first drawing $\{y_i\}_{i=1}^K \sim \text{Gamma}(\alpha_i, 1)$ and then set $x_i = y_i / \sum_{j=1}^K y_j$. Then it is clear that we can sample from (10) by simply sampling N independent draws from Gamma$(1, 1)$. Clearly, Gamma$(1, 1)$ = Exp(1), which concludes the derivation: the posterior probability of observing a data point with the priors given in (8) and $a_l \to 0$ can be sampled via Exp(1) adjusting for appropriate normalization. Thus, each data point D is realized in the posterior as written in (7), which concludes the proof.

[2]Indeed, this prior assumption is equivalent to the standard frequentist bootstrap assumption that the observed support acts as a stand-in for the population support.

2.1.3 The Algorithm

With this framework in mind, Dubé and Misra (2017) proceed to approximate the posterior distribution of the parameter estimates and elasticities via Algorithm 1.

Algorithm 1 Weighted Likelihood Bootstrap with GLM Lasso. Code here.

For each of the bootstraps $b = 1, ..., B$ (here, $B = 100$):

1) Draw weights $\{V_i^b\}_{i=1}^N \sim \text{Exp}(\mathbf{1}_N)$, and draw a weighted sample using these weights.

2) Run a Logistic Lasso using 10-fold cross-validation to determine the optimal penalty λ^{b*}

3) Retain the parameters $\hat{\Theta}^b \equiv \hat{\Theta}^b \mid_{\lambda^{b*}}$ where $\Theta \equiv (\theta_\alpha, \theta_\beta)$

For each price $p = 19, ..., 400$, calculate the own-price elasticity given $\hat{\Theta}^b \mid_{\lambda^{b*}}$ using Eq. (2)

It should be noted that as we follow Dubé and Misra (2017) and implement the WLB GLM Lasso using the R-package Gamlr, we do not incorporate the sampled weights in Step 1 directly into the maximization of the likelihood function as in (5), but instead draw a random sample using the draws in Step 1. This is simply because the Gamlr package does not allow augmentation of the likelihood function with weights. In that sense, while philosophically different, the approach is not much different from a classic bootstrap.

2.2 The Bayesian Probit-augmented Lasso (Frederik and Ditlev)

The last subsection showed how one could obtain a quasi-posterior distribution of the parameters in the demand model (1) to quantify the uncertainty around the estimates without specifying any priors on the covariates. This section introduces and derives a popular Bayesian regularization method where a proper posterior is obtained.

Tibshirani (1996) noted in his seminal paper on the Least Absolute Shrinkage and Selection Operator (Lasso) that the Lasso estimates could be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace priors. The Gibbs sampler for the Bayesian Lasso (Park and Casella (2008)) in the standard linear model continues from this idea by placing a conditional Laplace prior on the regressors

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left[\frac{-\lambda \mid \beta_j \mid}{\sqrt{\sigma^2}}\right] \tag{11}$$

and takes its starting point by noting that the Laplace distribution can be represented as a scale mixture of normals leading to the following hierarchical representation of the full model

$$\begin{aligned}
\boldsymbol{y}|\mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right) \qquad (12)\\
\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_p^2, \sigma^2 &\sim \mathcal{N}_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right), \quad \boldsymbol{D}_\tau = \operatorname{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)\\
\tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2 \tau_j^2}{2}\right\} d\tau_j^2, \quad \tau_1^2, \ldots, \tau_p^2 > 0\\
\sigma^2 &\sim p\left(\sigma^2\right) d\sigma^2\\
p(\lambda^2) &= \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp(-\delta\lambda^2), \quad \lambda^2, r, \delta > 0 \qquad (13)
\end{aligned}$$

where we for simplicity assume that the matrix of regressors is centered, i.e. $\mathbf{X}'\mathbf{1} = \mathbf{0}$. At the top layer is the canonical linear regression model. At the second layer is the scale mixture where the variable selection is induced: when a small value of τ_j is sampled, the conditional prior on β_j tends to a point-mass at zero, thereby excluding it from the model. However, τ_j^2 is never sampled as exactly 0, and thus the Bayesian Lasso does not offer variable selection in the same sense as Tibshirani (1996)'s, where some coefficients are set exactly to zero. The priors in the third layer are chosen such that if one integrates out $\tau_1^2, ..., \tau_p^2$, one obtains the desired prior in (11), and thereby also the correspondence to the frequentist Lasso. The prior λ^2 is chosen to be conjugate for analytical ease.

As the data we analyze is binary, we augment the model with a latent-utility representation of the probit model, i.e. treating $\mathbf{y}$ as unobserved and letting the binary (observed) outcome $\tilde{\boldsymbol{y}}$ be defined as

$$\tilde{y}_i = \left\{ \begin{array}{ll} 1 & y_i > 0 \\ 0 & y_i \leq 0 \end{array} \right\} \; i = 1, ..., N$$

For the rest of this section, we stick to the notation in the original paper, but it should be noted that in terms of the structural model in (1), y_i corresponds to ΔU_i, $\tilde{y}_i$ corresponds to q_i and $\boldsymbol{\beta}$ corresponds to $(\theta_\alpha, \theta_\beta)$. Further, for identification, we fix the variance such that $\sigma^2 = k \in \mathbb{R}^+$. Before we introduce the Gibbs sampling algorithm, we derive the conditional posteriors below.

2.2.1 The conditional posteriors

The likelihood

Write the likelihood as

$$p(y \mid \sigma^2, \boldsymbol{\beta}, \mu, \mathbf{X}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{1}\mu - \boldsymbol{X\beta})'(I\sigma^2)^{-1}(\mathbf{y} - \mathbf{1}\mu - \boldsymbol{X\beta})\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_c + \mathbf{1}\bar{y} - \mathbf{1}\mu - \boldsymbol{X\beta})'(\mathbf{y}_c + \mathbf{1}\bar{y} - \mathbf{1}\mu - \boldsymbol{X\beta})\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}_c - \boldsymbol{X\beta})'(\mathbf{y}_c - \boldsymbol{X\beta}) + (\mathbf{1}\bar{y})'(\mathbf{1}\bar{y}) - (\mathbf{1}\bar{y})'(\mathbf{1}\mu) - ....\right]\right\} =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}_c - \boldsymbol{X\beta})'(\mathbf{y}_c - \boldsymbol{X\beta}) + N(\bar{y} - \mu)^2\right]\right\} \quad (14)$$

where $\boldsymbol{y}_c \equiv \boldsymbol{y} - \bar{y}$, and we have used that $\mathbf{X'1} = \mathbf{0}$. Then, given the Jeffreys prior on μ, $p(\mu) \propto 1$, it is evident from (14) that

$$\mu \mid \mathbf{y}, \sigma^2 \sim N(\bar{y}, \frac{\sigma^2}{n}) \quad (15)$$

which is of interest in our case as we need to sample from the latent data. We follow the original paper and integrate out μ when sampling the remaining parameters

$$p(\mathbf{y} \mid \sigma^2, \boldsymbol{\beta}, \boldsymbol{X}) = \int p(\mathbf{y} \mid \sigma^2, \mu, \boldsymbol{\beta}, \boldsymbol{X}) p(\mu) d\mu =$$
$$(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y_c} - \mathbf{X\beta})'(\mathbf{y_c} - \mathbf{X\beta})\right) \int \exp\left\{-\frac{N}{2\sigma^2}\left[(\bar{y} - \mu\right]^2\right\} d\mu =$$
$$(2\pi\frac{\sigma^2}{N})^{\frac{1}{2}} \times (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y_c} - \mathbf{X\beta})'(\mathbf{y_c} - \mathbf{X\beta})\right) \int (2\pi\frac{\sigma^2}{N})^{-\frac{1}{2}} \exp\left\{-\frac{N}{2\sigma^2}\left[(\mu - \bar{y}\right]^2\right\} d\mu =$$
$$\frac{1}{\sqrt{N}(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y_c} - \mathbf{X\beta})'(\mathbf{y_c} - \mathbf{X\beta})\right) \quad (16)$$

Conditional posterior of $\boldsymbol{\beta}$

Using (16), we get

$$
\begin{aligned}
p\left(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2,\tau\right) \propto & p\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\tau\right)p\left(\boldsymbol{\beta}|\sigma^2,\tau\right) \propto \\
& \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{y_c}-\boldsymbol{X\beta}\right)'\left(\boldsymbol{y_c}-\boldsymbol{X}\,\boldsymbol{\beta}\right)\right]\right\}\times\prod_{j=1}^{p}\exp\left\{-\frac{1}{2\sigma^2\tau_j}\beta_j^2\right\}\propto \\
& \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{y_c}-\boldsymbol{X\beta}\right)'\left(\boldsymbol{y_c}-\boldsymbol{X}\,\boldsymbol{\beta}\right)\right]\right\}\times\exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}'D_\tau^{-1}\boldsymbol{\beta}\right\}\propto \\
& \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}'(D_\tau^{-1}+\mathbf{X}'\mathbf{X})\boldsymbol{\beta}+-2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{y}_c)\right]\right\} \\
& \Rightarrow \boldsymbol{\beta}|y,\sigma^2,\tau\sim\mathcal{N}\left(\boldsymbol{a}_N,\sigma^2\boldsymbol{A}_N\right)
\end{aligned}
\tag{17}
$$

where

$$
\begin{aligned}
\boldsymbol{A}_N &= \left(\boldsymbol{X}'\boldsymbol{X}+\mathbf{D}_\tau^{-1}\right)^{-1} \\
\boldsymbol{a}_N &= \boldsymbol{A}_N\boldsymbol{X}'\boldsymbol{y}_c
\end{aligned}
$$

Conditional posterior of τ_j^2 (sampling via τ_j^{-2})

Sampling the τ_j^2's is the trickiest part of the derivations, as we need to define an auxiliary variable to sample from. The first step is to write out the posterior kernel

$$
\begin{aligned}
p\left(\tau_j^2|\sigma^2,\beta_j\right) \propto & p\left(\beta_j|\sigma^2,\tau_j^2,\lambda^2\right)p\left(\tau_j^{"}\right)\propto \\
& \frac{1}{\left(2\pi\sigma^2\tau_j^2\right)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\}\frac{\lambda^2}{2}\exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\}\propto\left(\tau_j^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{\beta_j^2}{\sigma^2\tau_j^2}+\lambda^2\tau_j^2\right]\right\}
\end{aligned}
$$

Now, defining the auxiliary variable $\theta_j^2\equiv\tau_j^{-2}$ and using the change of variables formula

$$
f_y\left(y\right)=-f_x\left(v\left(y\right)\right)\times v'\left(y\right)
$$

where

$$
\theta_j^2=\tau_j^{-2}\Leftrightarrow\tau_j^2=v\left(\theta_j^2\right)=\left(\theta_j^2\right)^{-1}\Rightarrow v'\left(\theta_j^2\right)=-\left(\theta_j^2\right)^{-2}=-\theta_j^{-4}
$$

we get the posterior of θ_j^2 as

$$p\left(\theta_j^2|\sigma^2,\beta_j\right) \propto -\left(\theta_j^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{\beta_j^2}{\sigma^2}\theta_j^2+\frac{\lambda^2}{\theta_j^2}\right]\right\} \times -\theta_j^{-4} \propto \left(\theta_j^2\right)^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left[\frac{\beta_j^2}{\sigma^2}\theta_j^2+\frac{\lambda^2}{\theta_j^2}\right]\right\} \propto$$
$$\left(\theta_j^2\right)^{-\frac{3}{2}} \exp\left\{-\frac{\beta_j^2\left(\theta_j^2\right)^2+\lambda^2\sigma^2}{2\sigma^2\theta_j^2}\right\} \propto \left(\theta_j^2\right)^{-\frac{3}{2}} \exp\left\{-\frac{\beta_j^2\left[\left(\theta_j^2\right)^2+\frac{\lambda^2\sigma^2}{\beta_j^2}\right]}{2\sigma^2\theta_j^2}\right\} \propto$$
$$\left(\theta_j^2\right)^{-\frac{3}{2}} \exp\left\{-\frac{\beta_j^2\left[\left(\theta_j^2-\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}\right)^2\right]}{2\sigma^2\theta_j^2}\right\} \tag{18}$$

Employing the parameterization of the Inverse Gaussian density

$$f(x) = \sqrt{\frac{\lambda'}{2\pi}} x^{-\frac{3}{2}} \exp\left\{-\frac{\lambda'\left(x-\mu'\right)^2}{2\left(\mu'\right)^2 x}\right\}$$

it is evident from (18) that $\theta_j^2 \sim$ Inverse Gaussian$(\mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j}}, \lambda' = \lambda^2)$. Thus, in each iteration of the Gibbs sampler, $1/\tau_j^2$ is sampled via this Inverse Gaussian density.

Conditional posterior of λ^2

The prior on λ^2 is chosen such that it delivers a Gamma posterior

$$p\left(\lambda^2|\tau_1^2,\ldots,\tau_p^2\right) \propto p\left(\tau_1^2,\ldots,\tau_k^2|\lambda\right) p(\lambda) \propto \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_j^2}{2}\right\} \frac{\delta^r}{\Gamma(r)} \left(\lambda^2\right)^{r-1} \exp\left\{-\delta\lambda^2\right\} \propto$$
$$\left(\lambda^2\right)^p \left(\lambda^2\right)^{r-1} \exp\left\{-\sum_{j=1}^{p} \frac{\lambda^2\tau_j^2}{2} - \delta\lambda^2\right\} \propto \left(\lambda^2\right)^{p+r-1} \exp\left\{-\lambda^2\left(\frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + \delta\lambda^2\right)\right\} \Rightarrow$$
$$\lambda^2 \sim \text{Gamma}(p+r, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + \delta\lambda^2) \tag{19}$$

Conditional posterior of the latent data y given the binary data $\tilde{y}$

We follow the Bayesian Econometrics summer school course (or refer the reader to Lynch (2007) pp. 195-203), and simulate $\mathbf{y}$ from a truncated normal given $\tilde{\boldsymbol{y}}$, μ and $\boldsymbol{\beta}$ fixing $\sigma^2 = k$

$$y_i \mid \tilde{y}_i, X_i, \beta \sim \begin{cases} TN_{(0,\infty)}(\mu\mathbf{1}_n + X_i'\boldsymbol{\beta}, k), \text{ if } \tilde{y}_i = 1 \\ TN_{(-\infty,0)}(\mu\mathbf{1}_n + X_i'\boldsymbol{\beta}, k), \text{ if } \tilde{y}_i = 0 \end{cases} \tag{20}$$

2.2.2 The sampling scheme

The conditional posteriors above suggest the following Gibbs sampler

Algorithm 2 Bayesian Probit-augmented Lasso. Code here.

Initialize $\boldsymbol{\beta}^{(0)} = (\beta_1, ..., \beta_K)^{(0)}$, $\boldsymbol{\tau}^{(0)} = (\tau_1, ..., \tau_K)^{(0)}$, $\mu^{(0)}$ and fix $\sigma^2 = k$.

Gibbs Sampling. At each iteration $t = 1, ..., T$, run through

1) Sample $y_i^{(t)}$ from $p(y_i \mid \tilde{y}_i, X_i, \boldsymbol{\beta}^{(t-1)}, \sigma^2)$ for all $i = 1, ..., N$, see Eq. (20)

2) Sample $\boldsymbol{\beta}^{(t)}$ from $p(\boldsymbol{\beta} \mid \boldsymbol{y}^{(t)}, \sigma^2, \boldsymbol{\tau}^{(t-1)})$, see Eq. (17)

3) Sample $(\lambda^2)^{(t)}$ from $p(\lambda^2 \mid \boldsymbol{\tau}^{(t-1)})$, see Eq. (19)

4) Sample $(\tau_j{}^2)^{(t)}$ from $p((\tau_j{}^2)^{(t)} \mid \boldsymbol{\beta}^{(t)}, (\sigma^2)^{(t)}, \lambda^{(t)})$, see the discussion under Eq. (18)

5) Sample $\mu^{(t)}$ from $p(\mu \mid \boldsymbol{y}^{(t)}, \sigma^2)$ see Eq. (15)

For each price $p = 19, ..., 400$, use the final M draws $\{\boldsymbol{\beta}^{(t)}\}_{t=T-M}^{T}$ to calculate the own-price elasticity of demand. Note, that we have stuck with notation in the original Bayesian Lasso paper, and naturally the posterior draws $\{\boldsymbol{\beta}^{(t)}\}_{t=T-M}^{T}$ correspond to the parameters in the latent utility model (1), θ_α and θ_β. Here, M is chosen sufficiently small and T sufficiently large such that the Markov chain converges to its equilibrium distribution.

3 Bayes Factors (Ditlev)

In Section 4's simulation study, we make use of several metrics to evaluate and compare the performance of the two models. One of them is the so-called Bayes factor. The Bayes factor is a means to compare performance of two (or more) models, which involves specifying prior probabilities of models $p(M_j)$ and calculating the posterior probability given the data x of the model via Bayes's theorem

$$p(M_j \mid x) = \frac{p(x \mid M_j)p(M_j)}{\sum_j p(x \mid M_j)p(M_j)} \tag{21}$$

In the case of comparing two models, we obtain from (21)

$$\frac{p(M_1 \mid x)}{p(M_2 \mid x)} = \frac{p(x \mid M_1)}{p(x \mid M_2)} \times \frac{p(M_1)}{p(M_2)} = B_{12} \times \frac{p(M_1)}{p(M_2)} \Rightarrow B_{12} = \frac{p(x \mid M_1)}{p(x \mid M_2)}$$

where B_{12} is the Bayes factor. Usually, one puts equal prior probability on both models, and therefore tend to favor Model 1 if $B_{12} > 1$, Model 2 otherwise. Jeffreys (1961), App. B, provides some metrics to evaluate the Bayes factors, where e.g. $B_{12} \in [1, 3.2]$ is said to hold no evidence against either of the

models, and $B_{12} > 10$ is strong evidence in favor of Model 1 (vice versa for Model 2).

Computing B_{12} involves evaluating the integral

$$p(x \mid M_j) = \int p(x \mid M_j, \theta_j) p(\theta_j \mid M_j) d\theta_j \tag{22}$$

where θ_j are model parameters of the j'th model. One simple way to evaluate the integral is via a basic Monte Carlo estimate

$$\hat{p}_1(x) = \frac{1}{m} \sum_{i=1}^{m} p(x \mid \theta^{(i)}) \tag{23}$$

where $\theta^{(i)}$ is drawn iid from its prior distribution, and the notational dependence on M_j has been dropped for ease of notation. A major caveat of this estimator is that most of the $\theta^{(i)}$ will have small likelihood values when the posterior is concentrated relative to the prior thus leading to an inefficient simulation process. Further, this also causes the estimator to be dominated by a few large values of the likelihood leading to a large variance. Instead, we follow Kass and Raftery (1995) and implement an importance sampling scheme:

$$\hat{p}_2(x) = \frac{\sum_{i=1}^{m} w_i p(x \mid \theta^{(i)})}{\sum_{i=1}^{m} w_i}, \; w_i = \frac{p(\theta^{(i)})}{p(\theta^{(i)} \mid x)} \tag{24}$$

where the importance sampling function is the posterior $p(\theta \mid x)$ from which we simulate θ. Eq. (24) can be simplified by noting that

$$w_i = \frac{p(\theta^{(i)})}{p(\theta^{(i)} \mid x)} = \frac{p(\theta^{(i)}) p(x)}{p(x \mid \theta^{(i)}) p(\theta^{(i)})} = \frac{p(x)}{p(x \mid \theta^{(i)})}$$

leading to

$$\hat{p}_2(x) = \frac{\sum_{i=1}^{m} \frac{p(x)}{p(x|\theta^{(i)})} p(x \mid \theta^{(i)})}{\sum_{i=1}^{m} \frac{p(x)}{p(x|\theta^{(i)})}} = \frac{mp(x)}{p(x) \sum_{i=1}^{m} \frac{1}{p(x|\theta^{(i)})}} = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{1}{p(x \mid \theta^{(i)})} \right]^{-1} \tag{25}$$

That is, the harmonic mean estimator by Newton and Raftery (1994), where $\theta^{(i)}$ are drawn from the posterior. This estimator is also known to be unstable, but gives reasonable results in practice, see Kass and Raftery (1995) and Carlin and Chib (1995). Therefore, we will use this estimator in our simulation studies. Although it may be a bit of a stretch to evaluate the WLB by a Bayes factor by drawing from its quasi-posterior, $\hat{p}_2(x)$ actually enables this calculation as the prior distribution is not needed for calculation, and we see it as an interesting exercise to evaluate the models fully Bayesian.

3.0.1 Calculating the conditional likelihood $p(x \mid \theta^{(i)})$

For both the WLB Logit Lasso and the Bayesian Probit-augmented Lasso, we posit that the log-likelihood is

$$\log L = \sum_{i=1}^{n} \tilde{y}_i \log F(x_i'\Theta) + (1 - \tilde{y}_i) \log(1 - F(x_i'\Theta) \tag{26}$$

where $F(\cdot)$ is either the logistic cdf or the standard normal cummulative distribution. This follows from the fact that in WLB, the estimator is defined as the minimizer of the constrained negative log-likelihood

$$-\sum_{i=1}^{n} \tilde{y}_i \log F(x_i'\Theta) + (1 - \tilde{y}_i) \log(1 - F(x_i'\Theta) \text{ s.t. } \sum_{j=1}^{p} \mid \Theta_j \mid \leq \lambda$$

and for the Bayesian Probit-augmented Lasso, the observed likelihood conditional on the regressors is simply a probit model.

3.0.2 Implementation in R

Due to the very low values of the likelihood function in the high-dimensional $N \times p$ space, we modify the calculation of the Bayes factor, which can be seen in our code and in Appendix 6.1.

4 Simulation Study (Frederik)

To evaluate the performance of the two methods in the price experiment, we simulate the data as described in Section 1.2 over $n = 2$ seeds, with varying signal-to-noise ratios ($\sim 1, 3, 10$) and varying sample sizes ($N = 1,000, 10,000$), but with a constant number of regressors $p = 100$. We have also experimented with different demand curves (with different degrees of kinks), but the results are very similar and we choose to only present the results from the demand curve described in Table 1.1 on page 7. As the implications of our results do not change qualitatively over the different seeds, we choose only to present graphs from one seed with $N = 10,000$, and we average the results over two different seeds[3]. As mentioned, we simulate a Markov chain with $M = 20,000$ iterations. For the WLB, we follow Dubé and Misra (2017) and only use $B = 100$ bootstraps. As the bootstrap draws are independent, the WLB can be parallelized, which speeds up computational time considerably. With parallelization on 4 cores, the WLB takes roughly 15 minutes to run and the Gibbs sampler takes roughly 40 minutes on a Macbook Pro with a 2.3 GHz Intel Core i5 processor.

[3]Naturally, the robustness of the analysis would benefit from a higher n, say, $n = 100$, but this is computationally time-consuming, and we focus on the implications that seem robust over the different seeds.

As starting values for the Bayesian Lasso, we set $\mu^{(0)} = 0$, draw $\left(1/\tau_j^2\right)^{(0)}$ from an InverseGaussian$(1, 1)$ for all j, set $\beta_j^{(0)}$ to its probit estimate, and fix $\sigma^2 = 1$. Before turning to the results from fitting the model, we briefly touch some diagnostics tests for the Bayesian Lasso to assess that we are in fact drawing from its posterior distribution.

4.1 Diagnostic test

We assess practical convergence and mixing by the trace plots. Figure 4.1 shows trace plots for the last $2,000$ draws of the MCMC for $\beta_2, \tau_2^2, \beta_{62}, \tau_{62}^2$ for the $N = 10,000$ case, where β_2 is truly non-zero and β_{62} is truly zero. The same picture emerges for the $N = 1,000$ case, but with higher autocorrelations in the chains for β_1 and β_{62}. As mentioned in Section 2.2, τ^2 effectively performs the regularization. Hence, when τ_j^2 is small, the conditional prior on β_j tends to a point mass of zero. Thus, we would expect that τ_{62}^2 would be more concentrated around 0 than τ_2^2, which is also the case. The trace plots indicate proper convergence and mixing (bear in mind that τ^2 is truncated at zero), and that the significant variable β_2 is centered away from zero, whereas β_{62} is more or less zero.

However, there is some stickiness in the chain, which we explore in Table 4.1 where ineffiency factors and effective sample size for the final $2,000$ draws of the chain are calculated (via the R coda package). For $N = 10,000$, we get an effective sample size for the parameters of interest of around 100-200, which is comparable to the $B = 100$ bootstraps in the WLB. For $N = 1,000$, however, we only obtain a sample size in the order of 20-70 because of poorer mixing. This is in line with convergence results from Rajaratnam and Sparks (2015) who show that in many standard high-dimensional cases, the autocorrelation of the chain is increasing in p/N. This suggests that we use a larger sample from the posterior, and in the rest of the study we therefore use the final $6,000$ draws from the posterior when $N = 1,000$ and $2,000$ draws when $N = 10,000$.

For the auxiliary parameters, Figure 4.2 shows trace plots for μ and λ for the entire chain. λ converges fast to its target at around 25 where the mixing indicates that it explores its posterior quite efficiently, whereas it takes $5,000$ iterations for μ to settle on values around -5 with high autocorrelations. Recall that μ is the mean in the linear regression and if the fit $\boldsymbol{X\beta}$ does not change much from iteration to iteration, then one would also expect it to exhibit high autocorrelations as it is drawn from a normal with mean $\bar{y}$.

4.2 Mean-squared error

As a first test of fit, we compare the mean-squared error (MSE) for the elasticities in Table 4.2 for $N = 10,000$ and for $N = 1,000$ in Table 4.3 over different intervals of prices and different signal-to-noise ratios, all averaged over two seeds.

Figure 4.1: Trace plots of β_1, β_{62} and τ_1^2, τ_{62}^2

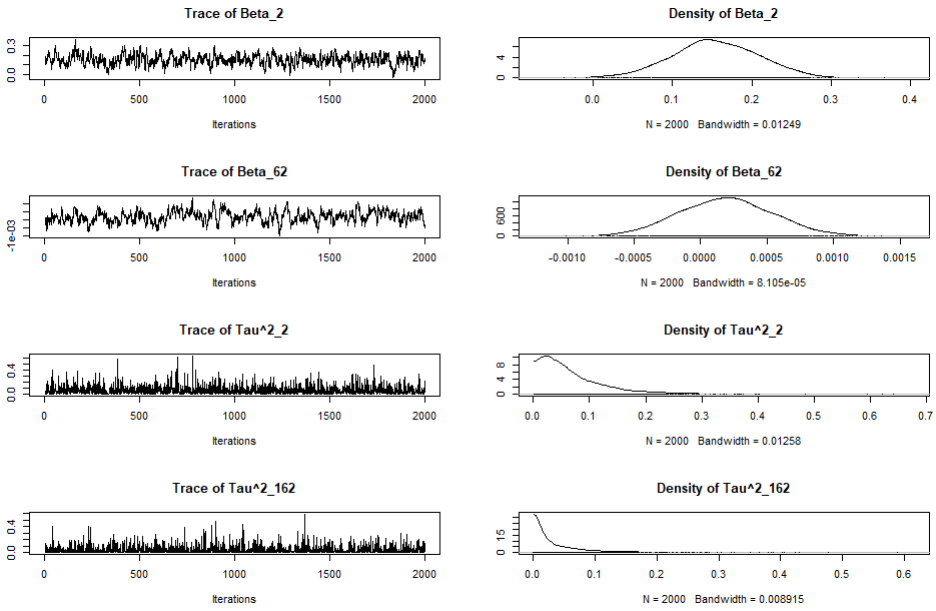


Table 4.1: Effective sample size and inefficiency factors

	β_2	β_{62}	τ_2^2	τ_{62}^2	μ	λ
10,000 Observations, Final 2,000 draws						
Effective sample size	209.78	112.35	1750.70	1594.35	3.08	602.69
Inefficiency factor	9.53	17.80	1.14	1.25	649.35	3.32
1,000 Observations, Final 2,000 draws						
Effective sample size	71.04	19.75	855.08	2187.50	10.95	215.00
Inefficiency factor	28.15	101.27	23.39	0.914	182.65	9.30

Figure 4.2: Traceplot of λ and μ

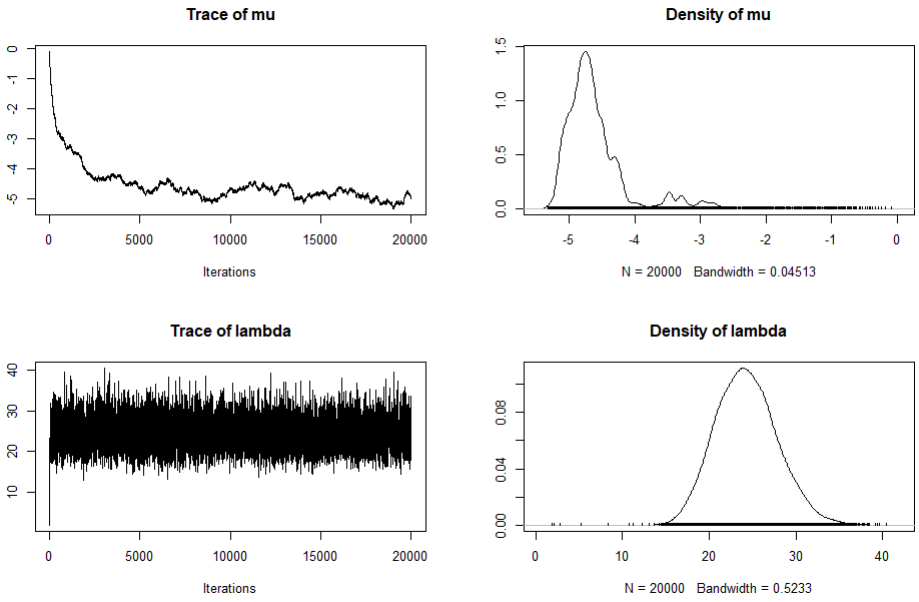


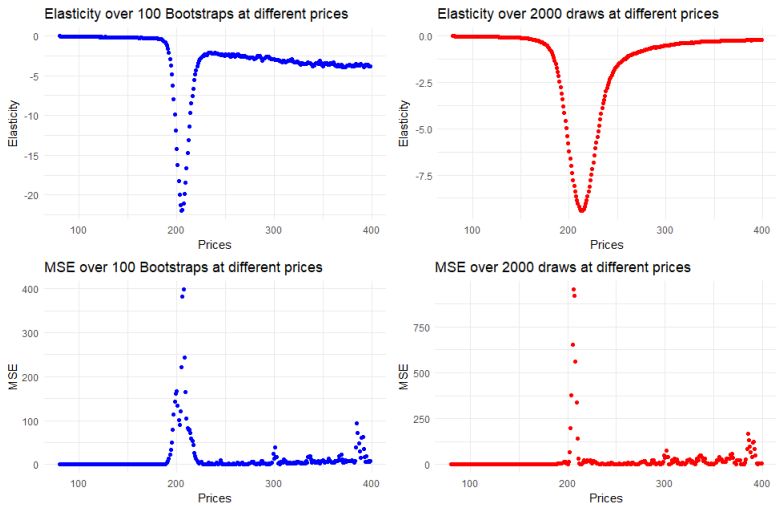
Table 4.2: MSE for the elasticities for $N = 10,000$

	WLB			Bayesian Lasso		
	StN < 1	StN ≈ 3	StN ≈ 10	StN < 1	StN ≈ 3	StN ≈ 10
Total MSE	36.19	14.61	17.18	31.43	24.00	10.96
			Intervals			
Price = [89 : 139]	0.54	0.01	0.01	0.03	0.00	0.00
Price = [139 : 189]	33.35	0.02	0.02	0.61	0.04	0.02
Price = [189 : 239]	156.44	65.19	80.40	144.79	90.20	21.21
Price = [239 : 289]	4.65	2.27	2.24	3.47	5.09	4.18
Price = [289 : 339]	8.55	5.35	5.64	11.24	14.46	10.44
Price = [339 : 389]	11.44	8.07	9.04	18.22	21.03	15.76

Table 4.3: MSE for the elasticities for $N = 1,000$

	WLB			Bayesian Lasso		
	StN < 1	StN ≈ 3	StN ≈ 10	StN < 1	StN ≈ 3	StN ≈ 10
Total MSE	65.02	56.15	53.63	55.76	52.44	39.45
			Intervals			
Price = [89 : 139]	0.04	0.05	0.04	0.19	0.07	0.37
Price = [139 : 189]	4.05	4.18	0.14	1.00	1.15	0.15
Price = [189 : 239]	180.88	148.73	142.30	165.90	145.19	60.84
Price = [239 : 289]	57.20	23.32	27.42	17.88	15.69	19.08
Price = [289 : 339]	47.73	47.25	45.26	42.51	42.06	44.68
Price = [339 : 389]	44.53	54.05	49.51	43.21	44.43	43.81

Figure 4.3: Estimated elasticities and MSE over all prices



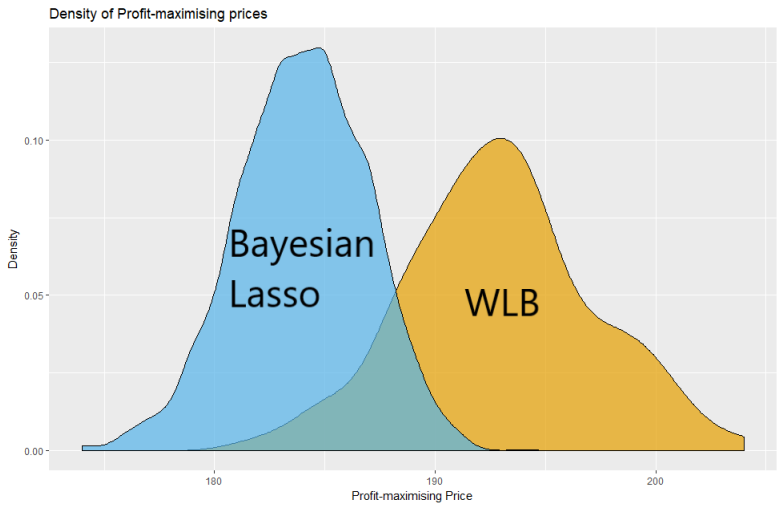
In Figure 4.3, the same results are exemplified for $N = 10,000$ and a signal-to-noise ratio of 3, which shows how both estimators have difficulties in getting the elasticity right at prices in the interval $[190 : 210]$. This makes intuitive sense, as this is where there is a large kink in the demand curve, which makes it harder for the two models to estimate the elasticities precisely. However, the MSEs do not give a clear picture as to whether one model outperforms the other: for a signal-to-noise ratio of ~ 3, the WLB seems to perform better, whereas the Bayesian Lasso fares better in the other cases. For $N = 1,000$, however, the Bayesian Lasso performs better than the WLB in all cases, which is because it identifies the elasticities of approximately zero a lot better for prices over 200, which is illustrated in Appendix 6.2.

4.3 Profit-maximizing prices

For each draw in the posterior, we calculate the profit-maximizing price using (3), which provides us with a distribution over the optimal prices. This is illustrated in Figure 4.4 for one particular seed and a signal-to-noise ratio of ~ 3. Clearly, the Bayesian Lasso is more concentrated than the WLB, which is actually always the case. To rule out that this it due to there being more MCMC draws compared to $B = 100$ bootstraps, we run a test for the WLB over $2,000$ bootstraps and find the same results.

Having these distributions, the convenient choice of pricing would then be the posterior median or mean (both more or less coincide in this case) resulting in a price of $p_{BL} = 184$ and $p_{WLB} = 190$ when using the posterior mean, which in this particular case would lead to higher profits for the WLB as illustrated in Figure 4.5 . That is, in this case the WLB identifies the optimal price better than the

Figure 4.4: Posterior density of profit-maximizing prices



Bayesian Lasso.

However, evaluated over different seeds and signal-to-noise ratios the picture is blurred. For instance, the Bayesian Lasso outperforms the WLB in identifying the true optimal price when using the posterior mean as an estimate for $N = 10,000$ for a low signal-to-noise ratio, which can be seen from Table 4.4. However, for $N = 1,000$ the WLB performs relatively better in most cases. As a robustness check to potential overfitting, Tables 4.4 and 4.5 also include a "cross-validation" comparison where we simulate new data with the same DGP to see how the implied prices perform on this new data. As the prices in this case do not change by much, the conclusions remain the same: both approaches identify the prices better in some cases, and worse in other cases.

The tables also include as comparison the implied prices by using a standard WLB logistic regression (with no regularization) with $B = 100$ and using the quasi-posterior mean as the price. For $N = 10,000$, both the WLB and Bayesian Lasso generally identifies the optimal price better, but not strikingly. For $N = 1,000$ the WLB logistic regression degrades into nonsense prices above of above 399, whereby it is greatly outperformed by both regularization methods. This makes intuitive sense, as when $N/p \to 0$, overfitting becomes a bigger issue thereby making regularization methods preferable. This is a nice example of the usefulness of applying proper high-dimensional methods, both frequentist and Bayesian, to high-dimensional data.

Figure 4.5: Implied profits at the different prices

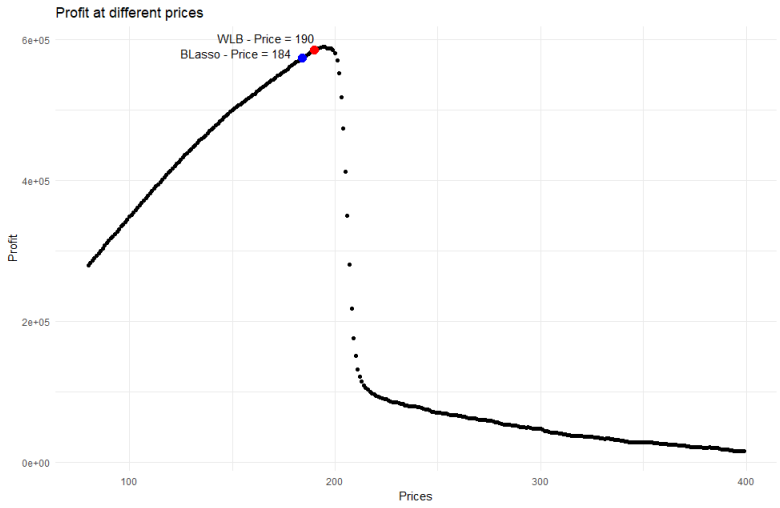


Table 4.4: Optimal Prices and Implied Profits at $N = 10.000$

	StN < 1			StN ≈ 3			StN ≈ 10		
	Seed 1	**Seed 2**	**Diff**	**Seed 1**	**Seed 2**	**Diff**	**Seed 1**	**Seed 2**	**Diff**
True Price	\$195	\$196		\$195	\$196		\$195	\$196	
BLasso Price	\$161	\$156	-21.7%	\$184	\$181	-6.5%	\$195	\$193	-0,5%
WLB Prices	\$148	\$142	-32.4%	\$190	\$188	-3.2%	\$196	\$193	0.0%
Logistic Prices	\$167	\$160	-17.4%	\$181	\$172	-8.3%	\$192	\$192	-2.1%
True Profit	\$590,460	\$592,116		\$590,460	\$592,116		\$590,460	\$592,116	
BLasso Profit	\$523,572	\$515,424	-12.8%	\$574,080	\$555,711	-3,1%	\$590,460	\$590,387	-0.5%
WLB Profit	\$495,060	\$480,670	-19.3%	\$585,390	\$580,544	-1,1%	\$588,392	\$590,387	-0.6%
Logistic Prices	\$538,409	\$524,960	-9.7%	\$568,340	\$549,712	-4.2%	\$588,480	\$588,480	-4.2%

Diff is calculated as $\frac{(\text{Simulated}_{\text{Seed 1}} - \text{True}_{\text{Seed 2}})}{\text{Simulated}_{\text{Seed 1}}}$

Table 4.5: Optimal Prices and Implied Profits at $N = 1.000$

	StN ≈ 1			StN ≈ 3			StN ≈ 10		
	Seed 1	**Seed 2**	**Diff**	**Seed 1**	**Seed 2**	**Diff**	**Seed 1**	**Seed 2**	**Diff**
True Price	\$197	\$198		\$197	\$198		\$197	\$198	
BLasso Price	\$157	\$162	-26.1%	\$161	\$162	-23.0%	\$192	\$196	-3.1%
WLB Prices	\$182	\$196	-8.8%	\$178	\$187	-11.2%	\$189	\$202	-4.8%
Logistic Prices	\$399	\$399	50.4%	\$399	\$399	50.4%	\$399	\$399	50.4%
True Profit	\$59,888	\$59,994		\$59,888	\$59,994		\$59,888	\$59,994	
BLasso Profit	\$51,967	\$52,650	-15.2%	\$53,130	\$52,650	-12.9%	\$58,944	\$59,584	-1.8%
WLB Profit	\$1,197	\$1,197	-4,912%	\$1,197	\$1,197	-4,912%	\$1,197	\$1,197	-4,912%

Diff is calculated as $\frac{(\text{Simulated}_{\text{Seed 1}} - \text{True}_{\text{Seed 2}})}{\text{Simulated}_{\text{Seed 1}}}$

4.4 Bayes Factors

Evaluating the performance of the two models via the profitability is based on comparing one estimate for each model (the posterior mean). These figures did not provide a simple answer: the Bayesian Lasso seemed to do relatively better for large N cases, and vice versa for the WLB, but not by much and not in all cases. As an alternative, the Bayes factor takes into account the entire posterior distribution when comparing the two models, as it averages the fit for all configurations of the parameters. We proceed by calculating the conditional likelihood via the harmonic mean estimator as described in Section 3, where we use draws from the posterior (or quasi-posterior for the WLB) to estimate the marginal likelihood of the model. Table 4.6 presents the logarithm of the Bayes factor for the three signal-to-noise ratios averaged over the two seeds. The code can be accessed here.

The results are quite extreme leaning either heavily in favor of one or the other, and most of the conclusions mirror those of the previous version: the Bayesian Lasso generally performs relatively well on the $N = 10,000$ data, whereas the WLB perform relatively better on the $N = 1,000$ data. One interesting insight, however, is that the implied profits were higher for WLB for a signal-to-noise ratio of 3 in the $N = 10,000$ case, but the Bayes factor analysis favors the Bayesian Lasso here. This is due to the WLB having a wider quasi-posterior distribution that covers parameter estimates that fits to the data poorly. This is exemplified in Figure 4.4 for the first seed, where the support of the quasi-posterior distribution for WLB covers prices above 200 that would lead to low profits due to the kink in the demand curve, see Figure 4.5. In that sense, the Bayes factor provides a more holistic evaluation of the model performance than based on a single point estimate, and it thus clearly seperates the performance of the two approaches conditional on the number of observations: WLB is prefered with low N, Bayesian Lasso is prefered with high N.

Table 4.6: Bayes Factors

	StN ≈ 1		StN ≈ 3		StN ≈ 10	
	$N = 1.000$	$N = 10.000$	$N = 1.000$	$N = 10.000$	$N = 1.000$	$N = 1.000$
$\log(\text{BF})$	-21.27	41.15	-79.93	109.74	-100.52	272.13

4.5 Summing up

For the $N = 10,000$ case, the MSE elasticities were similar, while the Bayesian Lasso outperformed the WLB in getting the elasticities right for $N = 1,000$, which is mainly driven by it being better at identifying the "kink" in the demand curve at prices around 200. The Bayesian Lasso outperformed the WLB in terms of Bayesian support for the model fit as evaluated by Bayes factors for $N = 10,000$, but profitability-wise there was no clear picture. The reverse was true for $N = 1,000$. Thus, the extent to which the Bayesian Lasso is preferable to the WLB implementation is dubious, as the WLB

performs well on low N/p ratios, and covers reasonable values in its approximate posterior for high N/p values. The quasi-Bayesian WLB therefore seems as a useful approximation in settings where one is not interested in using fully Bayesian methods, which could be due to problems with convergence, mixing or speed. Indeed, as discussed in Hastie et al. (2015), the complexity/speed of the Bayesian Lasso seems to scale at $\mathcal{O}(p^2)$, whereas the bootstrap method scales at order $\mathcal{O}(p)$, and hence in very large p settings, the Bayesian implementation may simply be infeasible. Based on the results in this report, we conclude that the WLB is a competitive alternative in low N/p settings when one is not interested in applying much prior knowledge to the analysis, and that even as N grows, it still provides a rather good approximation to the posterior distribution as benchmarked by the Bayesian Lasso.

Possible improvements and critiques

Of course, the above results would benefit tremendously from a more thorough investigation of the performance in different settings and averaged over more data sets, which we have not had the computational time to do yet. Especially, the analysis would benefit from being replicated on a larger number of N/p ratios, where the performance of regularization methods naturally perform relatively better than standard methods as was also indicated for the profit analysis in the $N = 1,000$ case.

The Bayes factor results are also surprisingly extreme, which may be due to the instability of the harmonic mean estimator and/or because it evaluates the draws from the WLB as if they came from a true posterior, which is not the case and thus may affect the stability of the estimator. We believe the Bayes factor results should be taken with a grain of salt, but we still think it gives a valuable input into conditioning the conclusion of the relative performance of the WLB over the Bayesian Lasso on the sample size.

4.6 Practical implications of our findings

We also take the time to evaluate the practical implications of our findings. In Dubé and Misra (2017), they mention that the managers of Ziprecruiter take a conservative approach and implement a price that is slightly lower than the one recommended from their analysis based on finding the price with an own-price elasticity of -1. From our analysis, we believe this is a good approach to take, as the posterior mean of WLB's optimal prices can sometimes overestimate the true optimal price, as evidenced in Table 4.5, which can lead to quite a substantial fall in profits when there is a kink in the demand curve. Both conversion rate tables on our simulated data and in Dubé and Misra (2017) suggest this kink, and it is hence a clever idea to take a conservative approach to pricing in this case.

For the Bayesian Lasso, on the other hand, we already see that it takes a conservative approach while also having a smaller range of optimal prices. Thus, it seems recommendable to use the posterior mean as the price if one were to use the Bayesian Lasso, but be more conservative when using the WLB.

5 Conclusion (Frederik)

We have presented the theory and implementation of two competing methods that can be used to estimate structural parameters in a high-dimensional demand models, and thereby help firms in pricing better. We find that the Bayesian Lasso's posterior distribution of optimal prices is more concentrated than the WLB's quasi-posterior, but does not convincingly fit the data much better. If one uses the posterior mean or mode as a tool to set prices, the Bayesian Lasso and WLB are comparable and both methods outperform standard ones such as logistic regression, especially in low N/p cases. Additionally, the WLB also provides reasonable, although larger, quasi-credible intervals for uncertainty quantification when executing a managerial pricing decisions.

6 Appendix

6.1 Bayes Factor: Implementation in R

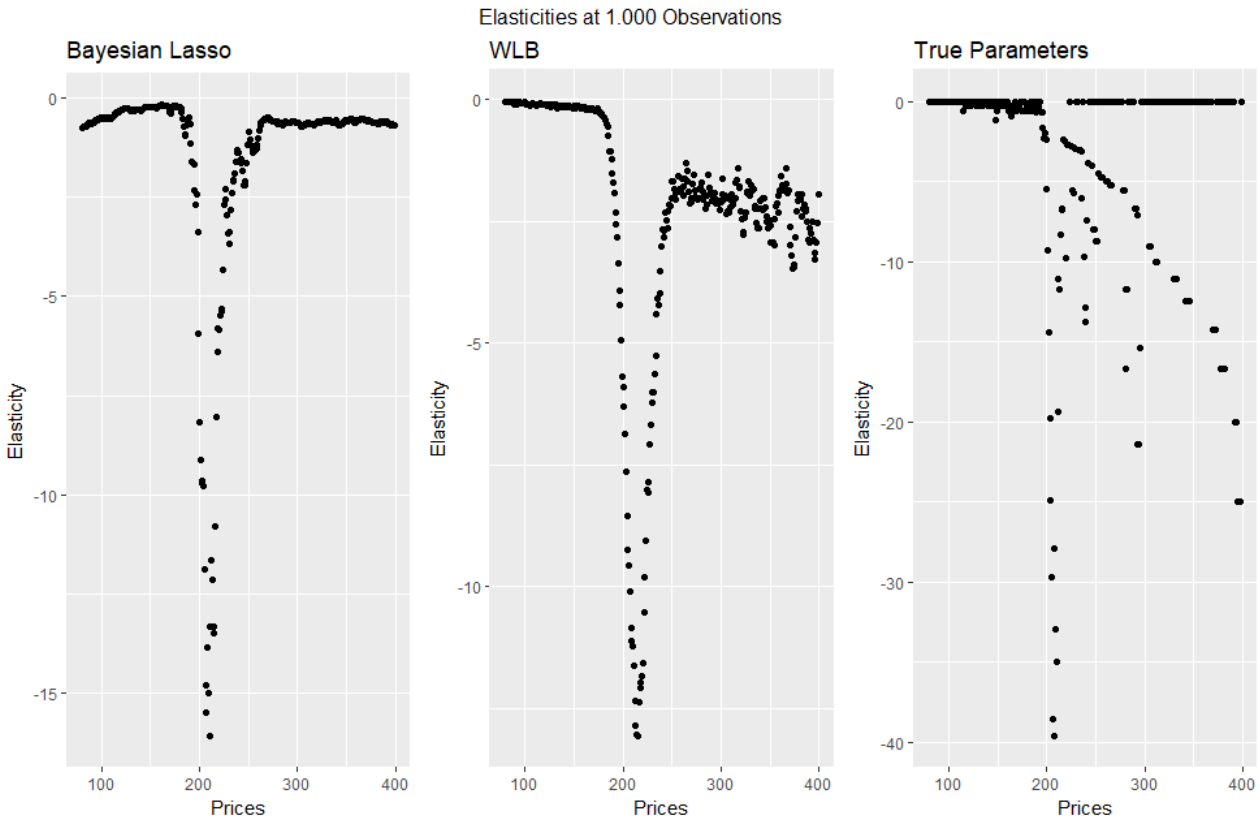
For the harmonic mean estimator, we augment the Bayes factor calculation by adding a constant term to the exponentials

$$B_{2,BL,WLB} = \frac{m_{BL}\left(\sum_{i=1}^{m_{BL}} p(x \mid \theta_{BL}^{(i)})^{-1}\right)^{-1}}{m_{WLB}\left(\sum_{i=1}^{m_{WLB}} p(x \mid \theta_{WLB}^{(i)})^{-1}\right)^{-1}} = \frac{m_{BL}}{m_{WLB}} \times \frac{\left(\sum_{i=1}^{m_{WLB}} p(x \mid \theta_{WLB}^{(i)})^{-1}\right)}{\left(\sum_{i=1}^{m_{BL}} p(x \mid \theta_{BL}^{(i)})^{-1}\right)} =$$

$$\frac{m_{BL}}{m_{WLB}} \times \frac{\sum_{i=1}^{m_{WLB}} \exp\left(-\log L(\theta_{WLB}^{(i)} \mid x) + \gamma\right)}{\sum_{i=1}^{m_{BL}} \exp\left(-\log L(\theta_{BL}^{(i)} \mid x) + \gamma\right)}$$

where $\gamma = \min\left\{\log L(\theta_{WLB}^{(i)} \mid x)\right\}$, BL is subscript for the Bayesian Lasso. The introduction of γ is made for R to be able to calculate the value that would otherwise be approximated to exactly zero.

6.2 MSE Elasticities for $N = 1,000$, signal-to-noise 3



References

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–484.

Dubé, J.-P. and Misra, S. (2017). Scalable price targeting.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Genc, M. et al. (2016). Empirical estimation of elasticities and their use. Technical report, EcoMod.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Jeffreys, H. (1961). Theory of probability.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Rajaratnam, B. and Sparks, D. (2015). Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *arXiv preprint arXiv:1508.00947*.

Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, pages 130–134.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):543–546.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.