Essays on High-Dimensional Bayesian Inference

Miquel Torrens i Dinarès

TESI DOCTORAL UPF / Year 2022

THESIS SUPERVISOR
David Rossell i Omiros Papaspiliopoulos
Department d'Economia i Empresa

Per als que encara hi són, i als que ja no.

Agraïments

M'agradaria fer un agraïment sincer, en primer lloc, als meus pares, pel seu suport permanent al llarg de tota la meva llarga etapa educativa. Sempre he sentit que podia triar el meu camí lliurement i que em farien costat en tot allò que necessités, que al cap i a la fi és el més valuós de tot.

Als meus dos directors de tesi, el David i l'Omiros, perquè si he après res durant tots aquests anys és gràcies a ells. En molts moments ha sigut dur, però sovint cal patir per poder progressar i aprendre. Ells m'han fet confiança i m'han sabut il·luminar el següent pas del camí en incomptables ocasions. Un esment especial per al David, per trobar les fonts de finançament que m'han permès completar el projecte quan el temps ha apretat.

Al Marc, per estar pendent de mi constantment i sincera durant aquests anys. L'he sentit a prop malgrat la distància física abismal que ens separa. Els amics de veritat es poden comptar amb els dits d'una mà i encara en sobren, però puc dir que, sens dubte, ell n'és un.

Al Zoel, la Maria i el Joaquim, pel seu suport, pels dinars i sopars on parlàvem de tot i de res, per les estones de desconnexió i de comunitat on el pes del món semblava més lleuger.

A la Laura, per escoltar-me i aguantar-me una i altra vegada quan compartíem pis en la pau de les nits tropicals del Poblenou.

A l'Àlex, per posar interès i bona intenció en la meva feina, especialment quan he topat amb dificultats per trobar allò que buscava.

A l'Alejandra, per ajudar-me a retrobar el camí en molts moments, un camí que ella havia traçat abans que jo. Sempre s'ha ofert a donar-me un cop de mà sense demanar mai res a canvi.

Als meus animalons, que tot i que molt probablement no els interessen el més mínim les meves pissarres atapeïdes, d'alguna manera misteriosa m'acaben arrencant un somriure plogui o faci sol.

I, sobretot, i molt especialment, a la Irina. Sense ella, res d'això no hauria esdevingut realitat. En els darrers anys ha sigut el pal de paller de tot plegat, algú meravellós en qui recolzar-se fins i tot quan anava curt d'esperança. Algun dia espero poder-li retornar ni que sigui una part de tot el que ella ha fet per mi.

Abstract

This thesis studies the capabilities of Bayesian estimators in high-dimen-sional generalised linear models, with a particular focus on treatment effect estimation. The first of four chapters provides the necessary background, advances and challenges for this thesis. In Chapter 2, I present methodological and computational contributions to tackle high-dimen-sional treatment effect estimation through confounder importance learning, a model averaging formulation based on a flexible model prior designed to mitigate problems related to over- and under-selection of controls, whose hyper-parameters are learnt by empirical Bayes through efficient gradient-based optimisation. Chapter 3 presents empirical evidence in favour of this approach, whose main application is the analysis of salary discrimination in the U.S. due to factors such as gender or race, revealing the existence of wage gaps that have not significantly improved over the last decade. Chapter 4 contributes with new theoretical properties that reinforce the use of non-local priors, showing satisfactory asymptotic results compared to other specifications.

Resum

Aquesta tesi estudia les capacitats d'estimadors bayesians en models lineals generalitzats d'alta dimensió, amb un enfocament a l'estimació dels efectes de tractament. El primer capítol proporciona context, avenços i reptes per a la tesi. Al Capítol 2, presento contribucions metodològiques i computacionals per abordar l'estimació d'efectes de tractament d'alta dimensió, a través de *confounder importance learning*, una formulació de mitjana de models basada en un prior per als model dissenyat per mitigar problemes relacionats amb la sobre- o sub-selecció de controls, i els híper-paràmetres de la qual s'aprenen a través de Bayes empíric, mitjançant optimització eficient basada en gradients. El Capítol 3 presenta evidència empírica a favor d'aquest mètode, la principal aplicació del qual és l'anàlisi de la discriminació salarial als EUA atribuïda a factors com el gènere o la raça, posant de manifest l'existència de diferències salarials que no han millorat de manera significativa en la darrera dècada. El Capítol 4 aporta noves propietats teòriques que reforcen l'ús dels priors no locals, mostrant resultats asimptòtics satisfactoris en relació a d'altres especificacions.

Executive Summary

In this work, I analyse the problem of model selection in the presence of a high-dimensional set of input features, with a particular focus on treatment effect estimation amidst a large set of potential confounding variables. The task of interest in this type of models is double: first, to discern in a large parameter space those parameters that are truly non-zero from those that are actually irrelevant, and, second, to derive valid inference on the former set. These tasks are hard because in these setups the number of parameters to be estimated can be (much) larger than the amount of observations. In its particular application to treatment effect estimation, one is only truly interested in evaluating a narrow subset of parameters which are of particular (practical) interest, leaving the rest as "control" parameters. Typically, this has been framed as the problem of identifying the effect of one single intervention, commonly known as a "treatment variable", on a response target. The added difficulty is that this intervention may be intrinsically linked to other factors that are already measured within the pool of potential variables for the model, so one must contemplate that there might exist a certain degree of confoundedness with an unknown subset of a high-dimensional set of variables, which must be adequately calibrated.

In this setup, the main issue of concern for many of the methodological approaches is the *under-selection* of controls, that is, if one omits the inclusion of any control variable that is relevant to explain the outcome that also correlates to the treatment, which leads to the well-known omitted variable bias problem, hindering inference on the parameter of interest. As a result, most proposals tend to be conservative in terms of discarding variables. This practice, however, might lead to a problem in the opposite direction, which despite being deemed as a less pressing concern by a fraction of the literature, I illustrate in Chapter 3 that its magnitude can be as detrimental to inference as the omission of relevant variables that one is trying to prevent in the first place. "Over-selection" of controls, especially of those correlating to the variables of interest, is a major driver of variance inflation for the estimate of the treatment effect, and can additionally lead to another source of bias: that related to incorporating truly irrelevant controls to the parameter estimation stage that are non-randomly chosen.

One of the main focuses of this work is on finding methodology that can prevent these two undesired effects, i.e. that can balance power and sparsity. I construct a Bayesian model averaging (BMA) estimate that provides solid foundations to avoid model over-parameterisation thanks to taking full advantage of the strong shrinking nature of parameter non-local priors, which I review later in Chapter 4. It is well-known, however, that in this scenario a regular BMA estimate built on canonical model priors can lead to low power through deterring the inclusion of variables that, despite having a non-zero effect on the response, are too correlated among themselves, as might be a treatment and a potential active confounder. To that end, I introduce a novel model prior, named confounder importance learning (CIL) prior, which can accommodate for

any number of treatments, and that can flexibly detect a priori what variables should be encouraged to be included or excluded from the model depending on their relationship with the treatments of interest. In essence, this is a separable prior that is parametrised such that correlating to a treatment does not necessarily equate to a higher prior probability of inclusion into the model: if one detects that these variables are irrelevant to the outcome, it may imply the opposite, or even no particular prior action at all. One of the main functions of this prior is to improve power in the detection of truly active parameters with respect to classical forms of BMA by tilting individual prior inclusion probabilities accordingly.

The critical parameterisation of this prior can be conducted using an empirical Bayes approach. However, since this approach can be computationally expensive, requiring the evaluation of a potentially expensive marginal likelihood in large hyper-parameter spaces, an variational algorithm based on expectation propagation is developed to achieve reliable approximations at a fraction of the computational cost, requiring the simple evaluation of marginal posterior probabilities under basic uniform priors. This drastically reduces computing time especially in the presence of a large number of treatments. These methodological contributions, including their computational implementation and corresponding software, are presented in Chapter 2.

In Chapter 3 I illustrate the performance of CIL in several scenarios. First, I contemplate a set of simulation studies where I stress its capabilities compared to other popular methods in the present literature. I find that performance of CIL is able to match the performance of every method in their best scenario, and outperform all of them in the rest, for a wide range of constructions that vary several key indicators, such as treatment effect size, model dimension, number of active controls and treatments, among others. Importantly, I show that CIL is indeed capable of attaining higher detection power with respect to conventional BMA, while deterring the inclusion of spurious parameters as well as BMA does. I also analyse potential wage discrimination in the U.S. in the 2010–2019 decade as a function of factors such as gender, race, ethnicity or place of birth. I illustrate that levels of discrimination attributed to these factors still exist, with little improvement (if any) across the decade. While CIL and BMA show similar results on the main effects, I show that CIL is capable of detecting several interaction terms that regular BMA struggles to detect, in the form of state-level deviations from the main effects, allowing for better assessment of treatment effect heterogeneity across states. As a further application, I also analyse the effect of the exposure to certain volatile organic compounds to different levels of cholesterol in the blood, with similar results in terms of performance.

Finally, in Chapter 4 I study properties of non-local priors in more detail. In particular, the interest is on finding that the so-called product moment (pMOM) non-local prior is able to attain satisfactory asymptotic properties in terms of marginal posterior inclusion probabilities, in addition to high-dimensional parameter estimation. These

posterior quantities are of capital importance in the CIL machinery, as they drive the hyper-parameter calibration algorithm that critically sets prior inclusion probabilities for each potential confounder. Here I depict their asymptotic behaviour in the sequence model, a simple but rich generalisation of the orthogonal regression setup that allows us to obtain fundamental posterior quantities in closed form. I investigate rates of convergence in high-dimensional models for marginal posterior inclusion probabilities, as well as parameter estimation. I show that for certain parameter regions and under a set of mild technical conditions, the pMOM prior can achieve stronger shrinkage to discard truly inactive parameters compared to standard Gaussian priors. Similarly good results are obtained in establishing probability bounds for the total absolute error of the BMA parameter estimates, with faster rates of convergence for truly inactive parameters relative to the Gaussian prior. Additionally, I find that for a large class of local priors in these parameter regions their concentration rates do not meaningfully improve those of the Gaussian prior as the pMOM does.

In summary, the main contributions of this thesis include a new method that can do inference on multiple simultaneous treatment effects, balancing detection power with model sparsity. This model averaging approach is based on the conjunction of a novel model prior, and a strong non-local parameter prior that introduces shrinkage. The model prior learns from data if and to what extent control inclusion or exclusion should be encouraged to inference, and comes with a computationally efficient algorithm that can render this approach practical. As for the non-local prior, it is shown that outside extreme parameter regions (i.e. too large or too close to zero) by incorporating a non-local element one is able to improve the asymptotic convergence rates of a wide set of local priors, both for marginal posterior inclusion probabilities and for parameter estimation. These results help to reinforce the validity of the prior employed by the CIL methodology.

Contents

List of Figures

Chapter 1

HIGH-DIMENSIONAL INFERENCE AND COMPUTATION

In the presence of a large set of potential predictors for a given outcome, the identification of the true subset of them that actually affect the latter, and the correct calibration of their effect, are challenging statistical problems. The main reason why they are hard is that the size of the set of covariates —or input *features*— may be systematically larger than the number of observations itself, and hence their analysis requires specific solutions that address issues not addressed by classical theory. These are the so-called high-dimensional settings, in which one wishes to extract a subset of features that truly affect some response of interest, among a large pool of potential candidates, with which one can then build a statistical model. The problem extends in fact beyond the selection of the right features: in many problems, posterior to selection, it is necessary to estimate what is the actual effect of the chosen features, as well as to quantify the level of uncertainty of both their estimates and the selection itself. One may think, for example, of the problem of policy design: not only it requires to state if a given policy has an effect on a targeted outcome (i.e. whether the predictor is actually selected) and with what certainty, but also to estimate the size of such effect, jointly with some range within which our estimation is reliable for some confidence level.

Therefore, the objective is to build a statistical model that contains only those predictors truly affecting the outcome, while simultaneously achieving valid statistical inference on the magnitude of their influence on it. This requires some method to correctly select features among the available predictors in the data. By *feature selection* we refer to a binary decision concerned with deciding if sampled evidence is strong enough to favour the inclusion of any given variable in the model, which effectively reduces to the problem of assessing correctly whether the effect of any given feature (or any set of them) is zero, or else. Selection by itself is a hard problem, and so tackling inference in such contexts is subsequently complex as well — it evolves the problem from a binary decision to a broader quantification of uncertainty. Then, estimation is not only

about point estimation or output prediction, but about being able to formally quantify confidence around these estimates.

Uncertainty quantification (UQ) in high dimensions is intrinsically linked to adequate feature selection, for achieving good selection is helpful in addressing the issue of uncertainty, especially if the inclusion or exclusion of any given feature can provide information about its UQ. As we will formalise later on, one may face some form of trade-off between proper selection and valid UQ. To gain intuition, notice for example that a parameter estimate is not a well-defined event, since the estimate is *undefined* whenever the selection of its corresponding feature does not occur. Thus, it loses mathematical meaning to attempt inference around an undefined estimate. At the same time, although including such feature would set grounds to quantify the uncertainty of its estimate, it would disregard selection altogether, were it applied uniformly to all unselected features. What is worse, any inclusion of an irrelevant feature comes with the associated costs to inadequate selection, with consequences on parameter estimation and subsequent outcome prediction, i.e. an overall increase in the mean squared error of both the model's parameters and its predictions. Building a model on a superset of the true subset of active features can inflate the variance of the sample distribution of any parameter estimate in the model, as well as induce some bias on its estimates if such *over-selection* is sample-driven, i.e. not random. Later we will illustrate that correct selection of covariates decisively determines the quality of the uncertainty estimates of interest. Several modern proposals to achieve adequate selection and UQ in both the frequentist and Bayesian literature will be reviewed, noting that no fully satisfactory uniformly valid method exists yet. Frequentist methods are generally aimed at either selection and predictive performance or at correct parameter inference and UQ, and hence can attain good results for one of these tasks separately. For example, under appropriate conditions, the LASSO (Tibshirani, 1996) can achieve good prediction albeit disregarding UQ, while the *debiased* LASSO (van de Geer et al., 2014; Javanmard and Montanari, 2014) derives valid inference while doing no discrete selection. To attempt simultaneous estimation and inference, in this work we will put especial emphasis on Bayesian methods. The Bayesian variable selection (BVS) approach incorporates parameter inclusion uncertainty into the selection stage, and model averaging into the estimation process, and hence with some careful modelling it looks as an option to potentially tackle both tasks simultaneously, although it still has theoretical room for development and faces stringent computational bottlenecks.

Computation is, in fact, a key aspect to be addressed. In practice, the existence of a reliable theoretical method is no guarantee to its feasibility. Therefore, it is worth exploring the computational complexity of relevant methods, as well as investigating efficient algorithms to put any proposed solutions into practice. This is especially relevant for high-dimensional problems, where combinatorial computations become insurmountable quickly. Modern model search and screening variable techniques, that will

be reviewed in the next section, look promising to achieve fast computation and model evaluation, although they can require strong conditions to guarantee satisfactory results. The exploration of sound algorithms that can fully exploit the advantages of powerful Bayesian constructions is another important driver of this work.

What remains of this chapter is laid out as follows. In Section 1.1, I introduce a general setup for variable selection, and present some of the most popular approaches in the literature in relation to the aspects of the respective aforementioned research areas, reviewed in Sections 1.1.1 and 1.1.2. In Section 1.2, I will enter into the particular problem of treatment effect estimation in high-dimensions within the model selection framework, which will be the main focus of the research presented in Chapter 2. I will introduce what are the issues faced in that context, and review what are the current proposals to address them present in the modern literature.

1.1 A General Framework for Variable Selection

Consider the classic linear model

$$y = X\beta^* + \varepsilon, \tag{1.1}$$

where $y = (y_1, \ldots, y_n)^\intercal$ is the output or response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $\beta^* \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}^n$ are the unobserved parameter and error vectors, respectively. In reality, we are considering the broader family of generalised linear models (GLM), but here we are focusing on the linear model for simplicity. For now, we will assume that β^* is fixed, and that the error term is homoscedastic and normally distributed $\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \phi I_n)$, with known variance $\phi \in \mathbb{R}^+$, unless stated otherwise. We are particularly interested in the case where $p \gg n$, where p may depend on n — i.e. a high-dimensional setup.

It is well known that in the classical setup the MLE

$$\hat{\beta} := (X^\intercal X)^{-1} X^\intercal y \tag{1.2}$$

is the *best linear unbiased* estimator for the parameters in (1.1), whose sampling distribution is

$$\hat{\beta} \mid y, \phi \sim \mathrm{N}(\beta^*, \phi (X^\intercal X)^{-1}).$$

The problem is that for the MLE to be unique one requires $n \leq p$, in order for the Gram matrix $X^\intercal X$ not to be rank deficient, i.e. to be invertible. Additionally, because the parameter vector β^* in (1.1) may in fact be *sparse* (that is, it may contain a large fraction of zeroes) the true dimension of the model $p^* = \sum_{j=1}^{p} \mathrm{I}(\beta_j^* \neq 0)$ may actually be $p^* < n \ll p$. If the set of active predictors $S^* := \{j : \beta_j^* \neq 0\}$ were known and of size

smaller than n, then the MLE would be computable simply screening out the unnecessary covariates. In turn, the set of variables present in the estimated model is also essential to quantify the uncertainty of the estimator. Hence, in this context the problem is two-fold. First, we desire to achieve good feature selection, as doing so would lead to accurate parameter estimation. Second, we wish to assess its uncertainty with reliable precision, since this will correctly measure the confidence of our estimation, even under imperfect selection. I will illustrate in Section 1.1.1 why achieving both objectives simultaneously is already hard from a theoretical perspective. An extra degree of difficulty is added to furthermore seek for an adequate method with good computational performance, which we examine in Section 1.1.2.

1.1.1 Selection and Uncertainty Quantification

The frequentist literature has tackled the problem of selection in high dimensions mostly using penalised likelihood (PL) methods, by looking for estimators in the form of

$$\hat{\beta}^{\text{PL}} := \underset{\beta \in \mathbb{R}^p}{\arg\max} \{ p(y \mid \beta, \phi) + h(\beta) \}, \tag{1.3}$$

where $p(y \mid \beta, \phi)$ is the likelihood function, and $h(\cdot)$ is some pre-determined function penalising parameter size[1]. The most well-studied of these estimators is the LASSO (Tibshirani, 1996), which uses an L_1-type penalty $h_{\text{L}}(\beta)$ on (1.3) as

$$h_{\text{L}}(\beta) := \lambda \|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|,$$

whose strength is controlled by a regularisation parameter $\lambda \in \mathbb{R}^+$. Along these lines, similar estimators conjugate various regularisation schemes: the classic ridge estimator, which uses an L_2-type penalty, the adaptive LASSO (Zou, 2006), which sets different weights for different parameters to the L_1 penalty, or the SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) log-concave-type penalties, conceived to tackle bias problems for large parameters. These estimators can sort out the aforementioned non-invertibility problem and are for the most part efficiently solved algorithmically thanks to its convex formulation, e.g. the LASSO via the LARS algorithm (Efron et al., 2004). The aim of these type of methods, however, has to do mostly with prediction through consistent selection. The LASSO, for example, can correctly recover the true model support, and achieve parameter consistency under some technical conditions. These conditions (see e.g. Hastie et al., 2016) have to do with (i) a ≪well-conditioned≫ design matrix, (ii) moderately low correlation between active and inactive features, (iii) some degree of sparsity of $\hat{\beta}$, (iv) lower bounds on the magnitude of $|\beta_j^*|$ to be detected, and (v) a

[1] Trivially, if $h(\beta) = c$, for any $c \in \mathbb{R}$, then $\hat{\beta}^{\text{PL}} \equiv \hat{\beta}$.

certain order for λ not to shrink too much or too little. Despite great progress in point estimation and predictive accuracy, however, UQ under PL schemes has been harder to tackle so far.

The issue of discrete selection in frequentist schemes is crucial in this aspect, because it complicates the options of building sound confidence regions for its estimates. By *discrete* we mean that, conditioning on the sample, any individual covariate is deterministically either included in or excluded from the model. Because the choice of variables to be included in the model is sample-dependent, the target of inference actually changes depending on the selected model (Berk et al., 2013), i.e. the coefficients are different according to what features are included, so even if the parameters are not random, the choice of which subset of them to include is. Thus, if one applies any selection scheme, then building a confidence interval (CI) with positive length on, say, the j^{th} feature, requires the probabilistic event of picking the j^{th} feature to occur (Lee et al., 2016). Only when selection is perfect can classical inference be applied safely, and such event is itself probabilistic — in fact, not necessarily one with high probability. As the sample grows, this problem is not relieved, as the asymptotic distribution of its estimates will include a point mass at zero for those cases in which selection does not occur (Knight and Fu, 2000), and so its resulting non-continuity in turn suggests caution on the use of any bootstrapping or subsampling schemes for UQ purposes (Dezeure et al., 2015). Furthermore, building a CI around PL point estimates even under correct selection is not straightforward either, since the introduction of some regularisation form of bias becomes an extra hurdle for such task.

A stochastic selection process also implies that classical inference *after* selection is arguably not a valid solution. At first blush, if some PL method were to deliver consistent selection, one could think of conducting decoupled inference on such selection, say using the MLE after consistent PL selection. However, any spurious effects observed in the data would damage both processes endogenously (Berk et al., 2013), effectively *overfitting* the sample. Sample splitting (Meinshausen et al., 2009) is not a uniform solution to this due to the cost of wrong selection, since subsequent inference is fairly more uncertain on a smaller sample, and in any case only valid conditional on the selected model (Fithian et al., 2017).

To address these issues, there are two main avenues of literature. The first approach involves the ≪debiasing≫ of $\hat{\beta}^{\text{PL}}$, with the idea that if such estimate is biased, debiasing it will allow to construct CIs with valid p-values around the resulting estimator for any individual predictors and so, by extension, to effectively test $\{H_{0,j} : \beta^* = 0\}$ unconditional to selection. The second approach deals with correct inference on a submodel, i.e. restricting to valid inference after the selection event, which would avoid dealing with point-mass mixtures, the so-called *post-selection* inference. This approach is focused on valid inference only on the included features, but it is conceived to be immune to selection mistakes. The idea of debiasing the LASSO (Zhang and Zhang,

2014; Javanmard and Montanari, 2014) deals with working on the LASSO's Karush-Kuhn-Tucker (KKT) conditions to find some estimator that can correct the bias of the original LASSO estimate. This debiased estimator (van de Geer et al., 2014) writes

$$\hat{\beta}^{\mathrm{D}} := \hat{\beta}^{\mathrm{L}} + \frac{1}{n}\Theta X^{\intercal}(y - X\hat{\beta}^{\mathrm{L}}), \tag{1.4}$$

where in this case $\hat{\beta}^{\mathrm{L}} \in \mathbb{R}^p$ denotes the LASSO point estimate, and $\Theta \in \mathbb{R}^{p\times p}$ is such that $\Theta X^{\intercal}X \approx I_p$ (to be determined). This estimator achieves

$$\sqrt{n}(\hat{\beta}^{\mathrm{D}} - \beta^*) = w + \delta,$$

where $w \mid X, \phi \sim \mathrm{N}(0, \phi\Theta(X^{\intercal}X)\Theta^{\intercal})$, and $\|\delta\|_\infty = o_{\mathrm{P}}(1)$, under a set of technical conditions. Its sampling distribution allows room for single-parameter CI evaluation. Such estimator relies, however, in finding an adequate way to set Θ. Two main proposals have been laid out (van de Geer et al., 2014; Javanmard and Montanari, 2014), which aim for minimising and bounding the column-wise distance $\|(X^{\intercal}X)\Theta_j^{\intercal} - e_j\|_\infty$, where $e_j \in \mathbb{R}^p$ is the j^{th} position standard unit vector. The proposal by van de Geer et al. (2014), based on a reconstruction of the covariance matrix using separate LASSO models for each feature, makes some strong sparsity assumptions on β^*. Instead, the proposal by Javanmard and Montanari (2014) tries to optimise the columnwise squared distance to the Gram matrix, subject to upper bounding the quantity $\|(X^{\intercal}X)\Theta_j^{\intercal} - e_j\|_\infty$. It makes no sparsity assumptions, but it requires some extra non-trivial parametrisation on the optimisation stage that requires further theoretical support.

The main issue with a debiased LASSO, however, is that it is in fact doing no selection, as the second term of the sum in (1.4) will not vanish. While individually one can test each predictor, it is not choosing features discretely, or stating anything about the full model. This is harmful for the parameter's MSE of the model, and for its predictive ability as an extension. This factor is especially sensitive taking into account that sparsity of β^* is relevant to its performance.

In the direction of unconditional testing, an additional proposal is related to a *decorrelated* score (Ning and Liu, 2017). This is a high-dimensional extension of Rao's score test (Rao, 1948), or LM test in econometrics. It is decorrelated because the score function for the parameter of interest is built to be uncorrelated with the score function of the rest of parameters, allowing for comfortable treatment of a high-dimensional set of them. This method aims at valid inference for oracle parameters, is applicable to small signals and does not require selection consistency. Additionally, it is extensible to joint statements. Its good behaviour, however, requires similar conditions to the debiased LASSO (most importantly, sparsity on both β^* and on the covariance between the parameter(s) of interest and nuisance parameters), and similarly is exclusively focused on inference, without presenting a formal way to attempt selection uniformly.

The second frequentist approach deals with post-selection inference. The first complete proposal on it is the PoSI method (Berk et al., 2013). Its assumption is that one has applied some prior selection procedure and that any unselected feature will not be considered, arguing that coefficients of excluded predictors ≪do not exist≫. The idea is to fit the model with the set M of all selected features, and adjust the CIs taking into account all possible models that could have been delivered by the selection scheme. The resulting least-squares estimate $\hat{\beta}_{\mathrm{M}}$ actually aims for $\beta^*_{\mathrm{M}} \in \mathbb{R}^M$, instead of β^* with p parameters, and so they try to build

$$\mathrm{CI}(\{\hat{\beta}_{\mathrm{M}}\}_j) = \{\hat{\beta}_{\mathrm{M}}\}_j + K\sqrt{\phi\,(X_{\mathrm{M}}^{\intercal}X_{\mathrm{M}})^{-1}_{j,j}},$$

which obviously requires $|M| < n$. The PoSI method's objective is to find a $K \in \mathbb{R}$ such that

$$\Pr(\{\beta^*_{\mathrm{M}}\}_j \in \mathrm{CI}(\{\hat{\beta}_{\mathrm{M}}\}_j)) \geq 1-a,$$

where $a \in (0,1)$ is the size of the test. In fact, $K := K(X,M,a)$ is conceived to be the usual t-statistic, enlarged to provide protection to wrong model specification. This procedure is robust at the cost of being quite conservative, producing very wide confidence intervals (see Hastie et al., 2016, Ch. 6.5), additionally to being computable only for models of very moderate size.

A second subsequent proposal deals with exact post-selection inference (Lee et al., 2016; Tibshirani et al., 2016). Again, because $\{\beta^* \in \mathrm{CI}(\hat{\beta}^{\mathrm{PL}}_j)\}$ is not a well-defined event, one conditions the estimate $\hat{\beta}^{\mathrm{PL}}_j$ on the selection event $\{\hat{\gamma} = 1\}$, where $\hat{\gamma}_j \in \{0,1\}$ is the estimated model inclusion indicator for feature j. The objective is to build conditional coverage intervals of the form

$$\Pr(\beta^* \in \mathrm{CI}(\mathrm{CI}(\{\hat{\beta}_{\mathrm{M}}\}_j) \mid \hat{\gamma}_j = 1) \geq 1-a,$$

for some size $a \in (0,1)$. Lee et al. (2016) show that the event $\{\hat{\gamma}_j = 1 \mid y\}$ is a union of polyhedra shaped by the KKT conditions of the LASSO problem, expressed as

$$\{A(\hat{\gamma},\hat{s})y \leq b(\hat{\gamma},\hat{s})\},$$

where $\{A,b\}$ represent the KKT conditions, and depend only on the selected model's inclusion $\hat{\gamma} \in \{0,1\}^p$ and sign $\hat{s} := \mathrm{sign}(\hat{\beta}^{\mathrm{PL}})$ vectors. Thus it suffices to study $\hat{\beta}^{\mathrm{PS}}_j \mid \{Ay \leq b\}$, where $\hat{\beta}^{\mathrm{PS}}_j := e_j^{\intercal}X^{+}_{\hat{\gamma}}y$, and $X^{+}_{\hat{\gamma}}$ is the pseudo-inverse of $X_{\hat{\gamma}}$, making $\hat{\beta}^{\mathrm{PS}}_j$ a pseudo-MLE type of estimator. The article shows that this conditional distribution is essentially a truncated Gaussian, independent of $\hat{\beta}^{\mathrm{PS}}_j$, and so a statistic $F(\hat{\beta}^{\mathrm{PS}}_j)$ exists whose conditional distribution is uniform between 0 and 1. This makes room for exact inference for a given estimate on the each separate predictor, even if selection is not perfect, as long as they are included in the model. This, however, does not solve the issue

of adequate selection, because any flaw in this arena will remain in the final model. In other words, it tolerates mistakes when tackling UQ, but relies heavily on the selected features for any other purpose, including the model's parameters MSE, and so its predictive ability is not regarded so as to outperform the LASSO. Additionally, it restricts to individual testing, falling short of making any joint statements.

The Bayesian literature, on the other hand, intends to use prior distributions to provide additional flexibility to the selection setup. Generally, the Bayesian Model Averaging (BMA) approach consists of taking into account model uncertainty to do parameter estimation, as

$$\tilde{\beta} := \mathrm{E}(\beta \mid y) = \sum_{k=1}^{2^p} \mathrm{E}(\beta \mid \gamma_k, y) p(\gamma_k \mid y), \tag{1.5}$$

a sum with 2^p terms due to combinatorial setups for the inclusion parameter vector $\gamma \in \{0,1\}^p$, for which γ_k denotes the k-th combination, with $p_k := \|\gamma_k\|_0$ active parameters. Let $\theta := \{\beta, \phi\}$, then we express posterior model probabilities by

$$p(\gamma_k \mid y) \propto p(\gamma_k) \int_\Theta p(y \mid \theta, \gamma_k) p(\theta \mid \gamma_k) \mathrm{d}\theta, \tag{1.6}$$

where here Θ denotes the parameter space. In the particular case where ϕ is known, then one can replace $p(\theta \mid \gamma_k)$ by $p(\beta \mid \phi, \gamma_k)$, and write

$$p(\gamma_k \mid y, \phi) \propto p(\gamma_k) \int_\mathrm{B} p(y \mid \beta, \phi, \gamma_k) p(\beta \mid \phi, \gamma_k) \mathrm{d}\beta.$$

Therefore, in terms of the parameters, one needs to ellicit a critical prior distribution on β, which I will discuss at length later on, and, in the case for unknown ϕ, an additional prior for the error variance[2]. Furthermore, note that in (1.6) it is necessary to also specify a prior for the model space. This prior can be furtherly decomposed as

$$p(\gamma_k) = p(\gamma_k \mid \|\gamma\|_0 = p_k) p(\|\gamma\|_0 = p_k),$$

where $\|\gamma\|_0 := \sum_{j=1}^p \gamma_j$. This prior controls the size of the model.

For the model space, mostly non-informative priors have been typically used, inducing sparsity if the dimensionality is very high. Popular choices include the *uniform* prior for moderate p, which assigns equal probability to any model, or the *binomial* prior, which gives equal prior selection probability $\rho \in (0,1)$ to each variable, thus encouraging models of size centered around ρp. In order to encourage models of smaller size, there is the *beta-binomial* prior extension, which assigns a uniform distribution on

[2]Some proposals exist to estimate the residual variance itself, both frequentist (e.g. Reid et al., 2013, for the LASSO) and Bayesian (see e.g. Moran et al., 2018, for a review), left outside of scope for now.

the model size, and then assigns equal probabilities to each model of a given size. For instance, the Beta-Binomial$(1,1)$ prior gives posterior probabilities of the form

$$p(\gamma_k) = p(\|\gamma\|_0 = p_k)p(\gamma_k \mid \|\gamma\|_0 = p_k) = \frac{1}{p}\binom{p}{p_k}^{-1}.$$

This prior specification is the product of combining $\rho \sim \text{Unif}(0,1)$ with $\|\gamma\|_0 \mid \delta \sim \text{Bin}(p,\rho)$, which does not favour mid-sized models, instead all sizes are now equally likely a prior.

The main driver for parameter estimation and UQ, however, comes in through the prior in $p(\theta \mid \gamma_k)$. Canonical choices include *conjugate* priors, which aim for closed form computation of (1.6). In our setup (under unknown ϕ) this would be the Normal-Gamma model, where

$$p(\beta, \phi \mid \gamma_k) = p(\beta \mid \phi, \gamma_k)p(\phi \mid \gamma_k)$$

with $\beta \mid \phi, \gamma_k \sim \mathcal{N}(0_p, \phi S)$ and $\phi \mid \gamma_k \sim \text{IGam}(a_\phi/2, b_\phi/2)$. This setup, however, entails setting some further hyper-parameters: canonical choices include small $a_\phi, b_\phi > 0$ (e.g. $a_\phi = b_\phi = 10^{-3}$), and $S := \tau(X^\intercal X)^{-1}$, for some $\tau > 0$, commonly known as Zellner's prior (Zellner, 1986). Setting τ is a problem itself, but a common choice is to set $g = n$ (Unit Information prior). See Kass and Wasserman (1996) for a detailed discussion on classical prior choice.

If conjugacy does not apply, in the lack of a closed expression then one will typically need to resort to numerical computation of the integrated likelihood in (1.6). A similar strategy applies to the computation of (1.5), as full model enumeration is only available for relatively small p. Gibbs sampling is simple but effective in many settings to explore the model space (see e.g. Madigan et al., 1995, for its application to model selection). This issue will be further discussed in the next subsection.

In terms of prior elicitation in the particular context of selection, some specific alternatives exist that attain good theoretical results. Intuitively, to achieve variable selection consistency it is necessary that $p(\beta \mid \gamma_k)$ is able to discard zero-elements in β^* efficiently, combined with a prior $p(\gamma_k)$ that will not encourage models to be too large with respect to the true size of the model. Then, to achieve good estimation rates, one also needs $p(\beta \mid \gamma_k)$ to have «thick» tails, so that truly non-zero elements in β^* have a significant prior probability mass. A modern specification that achieves excellent rates of convergence in terms of posterior concentration are the so-called *complexity-Laplace* priors (Castillo and van der Vaart, 2012; Castillo et al., 2015), which even though focused on recovering β^*, they are also conceived to quantify uncertainty. Their name is due to a combination of a complexity prior on model size of the type

$$p(\|\gamma\|_0) \propto c^{-\|\gamma\|_0} p^{-a\|\gamma\|_0},$$

for $a, c > 0$; with a separable product of Laplace priors $p(\beta_\gamma \mid \gamma)$, which is both computationally convenient and permissive on the tails. Heavy-tailed distributions are less strict on selection but allow for better parameter estimation, as they ≪shrink≫ less if $\gamma_j = 1$. The rest of features have a point-mass density at zero. Under compatibility conditions, this specification achieves optimal minimax posterior contraction rates for parameter estimates and prediction, while keeping model size relatively low, avoiding overshooting or charging strict supersets of the true model. On the other hand, it can run into problems to detect *weak* signals (those with a relatively small signal-to-noise ratio): their focus here is on asymptotically optimal rates but when n is finite signals that are not strong enough are dismissed with high probability.

A specification along similar lines lines is the *horseshoe* prior (van der Pas et al., 2017), a shrinkage prior which uses a mixture on Gaussians on the parameters as

$$\beta_j \mid \lambda_j, \tau \sim \mathrm{N}(0, \lambda_j^2 \tau^2),$$

with a half-Cauchy on the variance $\lambda_j \sim \mathrm{C}^+(0,1)$. Hyper-parameter $\tau \in \mathbb{R}$ is set to help the posterior contract around its true value. The authors show that with this specification UQ is ≪honest≫, but only for large $|\beta^*|$, and for $\beta^* = 0$. In terms of selection, the proportion of false positives is bounded above with probability tending to 1, and the proportion of true positives tends to one for strong signals. Under stronger conditions, even honest frequentist coverage can be achieved. This specification still suffers from the same problems, and shows that many signals are missed unless they are strong enough (see Figure 1 in van der Pas et al., 2017, for clear illustration). This is due to strong shrinkage combined with permissivity on tails, which greatly affects the ability to pick up weak (but potentially relevant) signals.

Finally, an appealing alternative that I will employ repeatedly in this work is the use of the so-called *non-local* priors, whose underlying principle is to add stronger data-induced parsimony compared to classical prior specifications, without necessarily inducing model sparsity a priori. This is achieved with a prior density that approaches zero as β_j goes to zero, whenever the parameter is considered active, i.e. that $\lim_{\beta_j \to 0} p(\beta_j \mid \gamma_j = 1) = 0$, as opposed to typical local priors, for which the opposite is true. I will discuss this class of priors at length in Chapter 2, and specially in Chapter 3.

Still, a common issue with all Bayesian methods is that they are computationally demanding, and in any case some quantities may require expensive MCMC exploration, both in the model space search for the BMA, as well as in the computation of exact posterior probabilities. In the lack of a closed form expression for a high-dimensional integrated likelihood and a computationally feasible normalising constant

$$p(y) = \sum_{l=1}^{2^p} p(y \mid \gamma_l) p(\gamma_l)$$

that would complete (1.6), some form of numerical computation will be required to successfully exploit flexible Bayesian constructions. These are obstacles to be overcome for their implementation, as I review next.

1.1.2 Computational Challenges of Model Selection

Variable selection in high-dimensions is a hard problem to tackle mainly because the set of possible models given p available features is of size 2^p. Several lines of research exist in the context of large model space search.

The first strategy deals with exact computation beyond exhaustive search. A foundational idea in this direction is the *leaps-and-bounds* type of algorithms (LBA, first proposed by Furnival and Wilson, 1974), whose objective is to find the best subset of columns of X that are related to the response without considering the entire set of subsets. This works by building a regression tree with a total of 2^p nodes, one for each model, by sequentially adding or dropping single variables. As one moves down the nodes, it is possible to prune large branches by *bounding* key quantities, and then *leaping* to other branches yet to evaluate. Ideally, this requires a fast way to evaluate any given subsetted model, combined with an algorithm to keep the set of models worth visiting to the minimum. A modern approach to LBA is called *branch-and-bound* algorithm (BBA, Gatu and Kontoghiorghes, 2006), and it suggests to use quick computations to the residual sum of squares (RSS) for the first task, combined with a pruning method for the regression tree. This strategy also allows to recover the best model for any given model size, cutting computational cost polynomially on p, which is reasonable for manageable p, but whose total cost remains exponential and so becomes intractable soon. More recent advances include a flexible mixed integer programming algorithm (Bertsimas et al., 2016), which minimises the squared loss function subject to $\|\beta^*\|_0 \leq k$, for a given $k \in \mathbb{Z}^+$. Modern optimisation methods boost its ability to deal with higher dimensions while guaranteeing some fair rate of suboptimality, even under early termination. Despite its effort, however, it still falls short to address truly high-dimensional setups.

Because exact search is hard, an alternative is to attempt stochastic search. A widespread approach, especially in the Bayesian context, is to use MCMC exploration. As mentioned before, a popular option is to use a Gibbs sampling algorithm based on sampling combinations of γ. This stepwise algorithm, named MC^3 (MCMC Model Composition, as in Madigan et al., 1995), starts from an arbitrary specification of γ, and begins to update sequentially each indicator in it, conditional on the rest of them, with probability

$$p(\gamma_j = 1 \mid \gamma_{-j}, y) = \left(1 + \frac{p(y \mid \gamma_j = 0, \gamma_{-j})}{p(y \mid \gamma_j = 1, \gamma_{-j})} \frac{p(\gamma_j = 0, \gamma_{-j})}{p(\gamma_j = 1, \gamma_{-j})}\right)^{-1},$$

where $\gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p) \in \{0,1\}^{p-1}$. Critically, this updating only requires the comparison of two models —one with and one without feature j—, and so

this entails the computation of the integrated likelihood, whether exact or numerical. In each iteration, the sampler adds or drops one variable at a time, and after a large enough number of iterations one is sampling effectively from $p(\gamma \mid y)$, by encountering those models with highest posterior probability more often. However, even if asymptotically reliable, theoretically it is not clear how long should the chain be run, or at what point does it reach stationarity. These demands become harsher as model dimensionality grows. In the line of MCMC, there are additional proposed schemes in the Bayesian literature. Bayesian Adaptive Sampling (BAS, in Clyde et al., 2011) explores model sampling without replacement from the space of models, and gives conditions for it to sample "near" the Median Probability Model, i.e. the one including any predictor j such that $\Pr(\gamma_j = 1 \mid y) \geq 1/2$ (Barbieri and Berger, 2004). This approach is thus strongly focused on prediction. An appealing hybrid solution that crosses the Bayesian framework with regularisation methods is an extension of the EM algorithm (EMVS, Ročková and George, 2014), a Bayesian deterministic algorithm variant of parameter-expanded EM for posterior mode detection. This algorithm aims for fast, sparse solutions that are robust to poor initialisations. Still, these type of proposals, although faster, are focused on finding high probability models and not so much on characterising the full model space. More recently Ročková and van der Pas (2020) have explored Bayesian regression trees aimed at effective dimensionality reduction while avoiding overfitting. Under suitable priors, this method is shown to achieve near minimax rates of optimality in terms of convergence to the true posterior model probabilities.

Another popular strategy alternative to exhaustive search is variable *screening*. This a model space restriction heuristic that allows for substantial scalability. The idea is to screen out a fair share of uninteresting features prior to doing feature selection. These algorithms are designed to reduce dimensionality while preserving the true model features with high probability. *Sure independence screening* (SIS, Fan and Lv, 2008) was one of the first proposals in this direction, which required rather strong correlations between relevant predictors and the outcome, as in essence it was a determined by marginal correlations, to be thresholded by some given criterion. To overcome this, along with similar decorrelation ideas (e.g. DECO, Wang, 2016), *high-dimensional ordinary least-squares projection* was proposed (HOLP, Wang and Leng, 2016), by which a Moore-Penrose type of pseudo-inverse is used to build a pseudo-MLE estimator, of the form

$$\hat{\beta}^{\text{HOLP}} := X^{\intercal}(XX^{\intercal})^{-1}y. \tag{1.7}$$

Computationally somewhat more costly than SIS, which essentially uses $\hat{\beta}^{\text{SIS}} := X^{\intercal}y$, it can achieve better separation of relevant and irrelevant features than marginal correlations under standard high-dimensional inference assumptions. It is also backed up by the ridge estimator in terms of theory, as it is the asymptotic solution of the ridge problem as its penalty parameter $r \to 0$. With this estimator the ≪relevant≫ variables

are separable from the rest with high probability by correctly thresholding the coefficients in (1.7) when $p \gg n$. Again, this method would simply ensure that the true model is within the screened in features with high probability, even though in fact it says little about the true model size or about its specific final components. Plus, posterior to screening it is unclear how to estimate uncertainty correctly, since inference with the chosen $k \in \{1, \dots, p\}$ features in fact reports

$$p(\gamma \mid y, \gamma_{k+1} = 0, \dots, \gamma_p = 0) \neq p(\gamma \mid y).$$

This can be a relevant issue especially if k is too small and/or if β^* is not sparse enough.

Finally, another commonly invoked approach in practice is direct dimensionality reduction through orthogonalisation, e.g. using PC regression on the eigenvectors obtained in the eigendecomposition of the Gram or design matrices. This is fast, highly scalable, and potentially useful for prediction purposes, but the resulting features are a linear combination of *all* the original features, so one cannot identify the effects of any of them individually. Thus, I will not consider them in this work as we examine a wider objective scope.

1.2 Treatment Effect Estimation

1.2.1 The Endogeneity Problem

In Section 1.1, we established that the MLE is the best linear unbiased estimator for the linear model in (1.1), however recall that this is only the case under some general assumptions that need to be satisfied, namely

1. Linearity of the true model
2. Strict exogeneity, i.e. $\mathrm{E}(\varepsilon \mid X) = 0$
3. Lack of multicollinearity, since $X^\intercal X$ needs to be non-singular in order to compute the MLE in (1.2)
4. Homoscedasticity of the residuals, i.e. $\mathrm{E}(\varepsilon_i^2 \mid X) = \phi > 0, \forall i$

Linearity concerns are usually addressed by supplementing X with any necessary non-linear combinations of the original features. Multicollinearity issues can be a problem, especially as p grows: for example, when $p > n$ the MLE cannot be computed. This issue can be addressed by performing dimensionality reduction, e.g. via shrinkage operators, which I will cover later on. Homoscedasticity is also hard to fully achieve in practice, but importantly it does not affect the unbiasedness of $\hat{\beta}$, it only influences its variance. A canonical robust technique to efficiently address heteroscedasticity is the

classical Generalised Least Squares (GLS) method. Thus, generally the most delicate assumption is exogeneity. This assumption tells us that the regressors in X and the error term ε need to be marginally uncorrelated. In other words, that if the model specification omits some regressors that correlate with some of the columns in X, thereby relegating them to the error term of the regression, then the coefficient estimates become biased. To see why, note that in (1.2) one can easily show that under assumptions 1 to 4

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal y = \beta^* + (X^\intercal X)^{-1} X^\intercal \varepsilon \xrightarrow{\text{P}} \beta^*,$$

as $\text{E}(\varepsilon \mid X) = 0$ implies that $\text{E}(X^\intercal \varepsilon) = 0$ by the Law of Iterated Expectations, critically providing $\text{E}(\hat{\beta}) = \beta^*$. Thus, failing to satisfy this assumption leads to the so-called *omitted variable bias* (OVB). In order to address this problem, broader GMM approaches have classically been invoked. I briefly introduce these next.

1.2.2 GMM and The Instrumental Variable Approach

Endogenous regressors can be controlled for using the generalised method of moments (GMM). The approach is to find a matrix $Z \in \mathbb{R}^{n \times k}$ of k additional covariates that serves as a set of *instruments* to consistently estimate an endogenous variable X_j, where $j \in \{1, \ldots, p\}$ is the column index. Both the columns of Z and X_j can potentially share predictors. The key assumption here is that these instruments be uncorrelated with the error term of the initial regression, namely

$$\text{E}(Z^\intercal \varepsilon) = \text{E}(Z^\intercal (y - X^\intercal \beta^*)) = 0_k. \tag{1.8}$$

These instruments are therefore pre-determined, given that they are orthogonal to the residuals. At the same time, however, in order to be valid they need to have explanatory power over the endogenous regressors in X, which are correlated with ε. This requires to additionally assume that $\Sigma_{zx} := \text{E}(Z^\intercal X) \in \mathbb{R}^{k \times p}$ is of full column rank, i.e. $\text{rank}(\Sigma_{zx}) = p$, in order to achieve full identification of the model. If the rank condition is satisfied and $k = p$, then the model is exactly identified. This is known as the *instrumental variable* problem (IV), the idea being that each regressor in X has a corresponding intrument in Z. If instead $k > p$, then the model is over-identified, but the GMM estimator can still be computed. Beyond the scope of this study, a few other technical assumptions are required to use the GMM estimator, for which I will refer to Hayashi (2000) (Chapter 3). The GMM estimator is then obtained from minimising the sample analogue of the moment conditions in (1.8), that is

$$\hat{\beta}^{\text{GMM}} := \arg\min_{\beta \in \mathbb{R}^p} g_n(\beta) := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} Z_i (y_i - X_i^\intercal \beta), \tag{1.9}$$

where the subscript i refers to the rows, and which in its simplest version is solved by

$$\hat{\beta}^{\text{GMM}} = (X^\intercal Z (Z^\intercal Z)^{-1} Z^\intercal X)^{-1} X^\intercal Z (Z^\intercal Z)^{-1} Z^\intercal y.$$

This method can be applied to both exactly identified and over-identified models — hence the term «generalised»[3]. In the exactly identified case, this estimator collapses to the IV estimator, where the function $g_n(\beta)$ in (1.9) can be equated to zero, in line with the traditional method of moments, and still find a unique solution, since there are the as many equations as there are unknowns, given $k = p$. Then it can be solved directly as

$$g_n(\beta) = 0 \Leftrightarrow \hat{\beta}^{\text{IV}} := (Z^\intercal X)^{-1} Z^\intercal y.$$

The core underlying idea of the IV estimator is that because Z has explanatory power over X, but contrarily to X it is not correlated to ε, the instrument can be used as a predictor of X. As a result, the estimated values $\hat{X}$, obtained using Z, have no endogeneity problems and are no longer correlated with ε. Hence, despite having to pay a non-negligible price in terms of variance, the regression coefficients will now be unbiased, since

$$\mathrm{E}[\hat{\beta}^{\text{IV}}] = \mathrm{E}[(Z^\intercal X)^{-1} Z^\intercal y] = \beta^* + \mathrm{E}[(Z^\intercal X)^{-1} Z^\intercal \varepsilon] = \beta^*,$$

given that (1.8) would hold, on top of the rest of necessary assumptions previously discussed. The IV estimator is classically computed via this two-stage least squares method (2SLS), whereby in the first stage one estimates the values of the endogenous covariates in X using Z, and in the second stage one uses these estimated values to model y, obtaining unbiased estimates for the regression coefficients.

1.2.3 Treatment Effect Estimation: IV in High-dimensional Regression

In numerous applications the research question is to understand the effect on the response of a certain fraction of covariates of interest (often times, a single one), and in particular to determine whether their contribution is significantly different from zero, conditional on a set of control variables that may be related to them. A variable belonging this set of inferential interest is commonly referred to as a *treatment* variable, as it encodes the effect of inducing some specific treatment on the observation, while being potentially endogenous with the rest of controls in the model. These controls that are related to both the response and the treatments are typically referred to as *confounders*. Hence, one wishes to estimate the isolated effect of the treatment variables in a model of the following type

$$y = D\alpha^* + X\beta^* + \varepsilon, \tag{1.10}$$

[3]Note that the MLE is a special case of this setup where $Z = X$. Then the canonical normal equations that correspond to (1.1) are recovered.

where $\mathrm{E}(\varepsilon \mid D, X) = 0$, and for which $D \in \mathbb{R}^{n \times T}$ is the matrix T of treatment variables distributed columnwise, and where $\alpha^* \in \mathbb{R}^T$ is the vector of interest containing the true *treatment effects*. The treatment variables are exogenous *conditional* on the control variables X, but in practice it is rarely the case that all confounding effects can be controlled for. The omission of any relevant controls in the model that are related to the treatment variables would lead to OVB, since such omission would relegate said controls to the error term, and consequently bias the MLE, as reviewed in Section 1.2.2. There I discussed that endogeneity problems like this can be addressed by looking for valid instruments for the treatmend featured in D, and consequently the model could be consitentnly estimated using the 2SLS procedure.

However, a fundamental concern appears when the dimensionality of the model becomes excessively large, since in that case the omission of relevant controls becomes increasingly likely as the number of potential controls grows. For example, when $p \to n$, the problem becomes unstable with high probability, damaging the inferential performance of classical methodology. When $p > n$, or directly $p \gg n$, the large feature space impedes the computation of the MLE for the parameters of interest in (1.10). This situation requires a reduction in dimensionality on the set controls, at the risk of making relevant variable selection mistakes which, again, can quickly lead to OVB problems. In other words, if the set of instruments is too large, then the MLE is no longer a valid approach and the problem needs to be *regularised*. Such form of regularisation needs to achieve good selection properties, to avoid missing any relevant instruments and hinder the quality of estimation and inference of the treatment effects.

The Problem of Non-random Control Over-selection

In this work, I will mainly focus in the high-dimensional model where potentially $p \gg n$. To understand why this scenario needs specific methodology beyond generic variable selection methods, let us first review why naïve strategies provide unsatisfactory results. The use of single-equation regularisation, that is regularising the model in (1.10) directly by treating controls and treatments exchangeably, fails to achieve adequate selection, as that may induce to relevant false negatives with high probability (both on treatments and relevant controls) leading to OVB, whenever there are controls that are related to both the response and any of the treatments, and therefore present stronger correlation structures between the three sets that negatively affect the performance of generic methods. Similarly, regularising the single-equation full model while *forcing* the inclusion of the treatments will not suffice either, since then it is likely that at least some features correlated to the treatment variables will be dropped, causing OVB again if these features have a non-zero effect on the response, but are pushed to the residual term (see e.g. Belloni et al., 2014a, for a detailed discussion). As we review next, many proposals around the problem of treatment effect estimation in high-dimensions revolve around ensuring that no relevant variables are left out of the model, since OVB is a priori

the greatest concern. This typically implies encouraging inclusion of a control variable into the model if there is some evidence that connects it to either the treatments or the response. This, however, may come at a high cost related to the opposite effect: problems derived from non-random *over-selection* of controls. Since we only want to include those variables affecting the response and any confounders linked to the columns in D, the inclusion of any variable that either only affects the response through D or that is spuriously correlated to it will inflate the variance of the parameter estimates associated to the treatments. This problem becomes more acute as the dimensionality of the model grows, i.e. as the pool of potential confounders also grows. Additionally, it can also trigger a biasing effect on the estimates, since the over-selection is data-dependent. To see this effect, suppose we estimate the parameters using the MLE on a data-dependent selection of features, whose support we denote by $\hat{s}(y)$ of size less than p. Then, for a fixed design matrix X the estimates

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}((X_{\hat{s}}^{\intercal} X_{\hat{s}})^{-1} X_{\hat{s}}^{\intercal} y) = \alpha^{*} + \mathrm{E}((X_{\hat{s}}^{\intercal} X_{\hat{s}})^{-1} X_{\hat{s}}^{\intercal} \varepsilon). \tag{1.11}$$

Note that if $\hat{s}$ were a fixed superset of the truly active variables, then (under the rest of standard assumptions) $X_{\hat{s}}$ could be extracted from the expectation and the MLE would be unbiased by the fact that $\mathrm{E}(\varepsilon) = 0_n$ (even if $\hat{s}$ itself is random, independent of y, a similar argument can be made using the Law of Iterated Expectations). The problem here is that since the selection step is non-random (i.e. it depends on the observed y), the strict exogeneity assumption is violated: we have only assumed that $\mathrm{E}(\varepsilon \mid X) = 0_n$, but not that $\mathrm{E}(\varepsilon \mid X_{\hat{s}}) = 0_n$. Indeed, if $\hat{s}(y)$ is influenced by ε, then $\mathrm{E}(\varepsilon \mid X_{\hat{s}}) \neq 0_n$, and so the right term in the sum in (1.11) would not vanish. Hence, in that case the MLE would yield $\mathrm{E}(\hat{\alpha}) \neq \alpha^{*}$. These two effects will be illustrated extensively in Chapters 2 and 3.

Existing Methodological Approaches

To fix ideas, let us assume for now that D has just one column, i.e. we are in the particular case with a single treatment variable in the model, also known as the single treatment model. Literature on single treatment effect estimation has grown considerably over the last decade, with the main objective of adressing pitfalls in parameter estimation that arise under general variable selection methods. Frequentist approaches particular to the single treatment problem concentrate around varied forms of penalised likelihood (PL), which make them computationally very appealing and suitable for high-dimensional settings. One of the most popular is the Double Machine Learning (DML) approach, presented in Belloni et al. (2014a,b), a two-step procedure that resonates with the two-stage philosophy in the IV problem, by which one analises separately an *output* equation and an *exposure* equation (an analogous linear model of the treatment on the controls) with the objective of identifying which controls affect either output, to then fit a model ex-post using OLS on the model of the response on the union set of selected controls.

Its main concern is to avoid estimation bias through fully adjusting for confounding, that is, including into the model at least any control relevant to the treatment or the response. Farrell (2015) builds upon similar ideas with a doubly-robust estimator attempting to safeguard from model selection mistakes after the double selection step, in the context of multi-valued treatments, resonating also with Antonelli et al. (2018), who use matching on propensity and prognostic scores on a similar direction. Shortreed and Ertefaie (2017) employ a two-step approach as well, in this case using adaptive Lasso on the exposure equation, showing improvement on confounder-selection results. See also Ghosh et al. (2015) with shared and difference Lasso for similar multi-step contributions. Ertefaie et al. (2018) use joint likelihood L_1 penalisation of both equations instead, in order to accommodate for the information shared between the two equations. Ma et al. (2019) combine regularisation with sufficient dimension reduction into an estimator that achieves good asymptotic properties without requiring model selection consistency. Yet, even if some of these proposals do allow for inference, most of them are essentially designed with sufficient control selection in mind and tackle point estimation only after selection has been conducted, generally without a focus on quantification of uncertainty. Chernozhukov et al. (2018) ellaborate on the DML technique with a more explicit attempt at inference seeking dependence reduction on the selection step, introducing an additional debiasing operation combining Neyman-orthogonal scores on the first equation, with cross-fitting to address overfitting concerns. Other proposals exist focused on inference for PL-related methods in this particular context, see Athey et al. (2018) in the context of binary treatments, see also Dukes and Vansteelandt (2019). Generally, PL-based proposals are heavily focused on asymptotic results regarding estimation efficiency and their distributional properties, which follow from attaining guarantees of sufficient control selection, i.e. avoiding detection errors of relevant controls connected to either outcome or treatment. In contrast, they are not too concerned with excess inclusion of any other spurious variable. This can be problematic based on the aforementioned reasons, which in turn often limit their oracle performance to a super-model of the true outcome model, inflated in size with unnecessary controls that may relate to the treatment only. It is also worth mentioning that a number of these proposals are not designed for continuous or even multi-valued treatments, or sometimes outcomes.

From a Bayesian perspective, a number of different proposals also exist. A widely referenced method is Bayesian Adjustment for Confounding (BAC) as presented in Wang et al. (2012), a joint-modelling approach which essentially employs a Bayesian Model Averaging (BMA) approach under a specifically designed model space prior. This prior is separable across controls, and is governed by a hyperparameter $\omega \in [1, \infty)$ that represents the odds of including a control in the outcome model conditional on that same control being included in the exposure model. This is quite helpful since for any finite ω control inclusion on the outcome model is only encouraged and not forced,

allowing for some model flexibility that combines with the accounting of model uncertainty in the averaged point estimate. This entanglement across equations poses a series of questions, however, mainly related as to how one should set such a sensitive hyperparameter, combined with the fact that resulting marginal prior inclusion probabilities are generally high, hence encouraging model size inflation. This can hinder performance notably in high-dimensional settings as over-selection problems may enter into play. Additionally, computational demands arising from joint equation modelling can quickly become insurmountable. Further contributions to BAC include Lefebvre et al. (2014) and Wang et al. (2015), which provide some theoretical support as well as further proposals on how to set ω. Other articles build on the approach based on control selection with model averaging: Talbot et al. (2015) introduce Bayesian causal effect estimation, a similar method to BAC that incorporates informative priors aimed at deterring excess control inclusion; Antonelli et al. (2017) contribute with *guided* BAC as a generalised framework on BAC adressing treatment effect heterogeneity, as well as additional technical questions. Similar methods have also been explored in propensity score analysis, as in Cefalu et al. (2017), but more generally around model uncertainty as well, see e.g. Zigler and Dominici (2014). See also Jacobi et al. (2016) for methodological adaptations to dynamic effects in panel data. More recently, Hahn et al. (2018) preserve the joint modelling approach but move away from model averaging, by addressing BAC using hierarchical priors, in an attempt to give some prior flexibility and ease computational difficulties via posterior sampling. This is a reparametrisation technique designed to achieve debiased point-estimates using regularisation priors. Hahn et al. (2020) also extend this notion to non-parametric setups with Bayesian Causal Forests. Antonelli et al. (2019) propose a spike-and-slab prior formulation with a prior distribution that places low shrinkage to controls associated with the treatment, combined with an Empirical Bayes algorithm for hyperparameter setting. Addressing computational concerns in high-dimensional setups, Wilson and Reich (2014) propose Penalised Credible Regions (PCR), stemming from Bondell and Reich (2012). This is a decision theoretic approach that can be formulated as a PL method. It essentially uses the posterior credible region of the outcome regression parameters to form a set of possible models to choose from, and then apply L_1-type penalisation with lighter penalties to those covariates that are associated to the treatment, as a function of their strength. This relies strongly on a the quality of the posterior mean, and the fact that the treatment is always included casts some difficulty on its single parameter estimation ability as it might sacrifice precision on one parameter to favour models that aggregately perform better predictively. PCR tends to be conservative as well as to dropping variables, depending on the penalty parameter, whose setting is an open end itself. It also introduces the notion of strength of relation between controls and treatment as a determinant of relevance on the outcome equation. Additionally, it naturally allows for the inclusion of multiple simultaneous treatments, which is absent in previously reviewed methods. On the other

hand, its PL nature disallows PCR as a method of uncertainty quantification. This proposal has strong ties to the Adaptive Lasso (Zou, 2006), as well as relation to Bayesian Lasso strategies employing shrinkage priors (Park and Casella, 2008; Hans, 2010). All in all, it is worh noting that most current Bayesian proposals are also heavily concerned with omission of relevant controls, and despite extra flexibility they can run into similar problems related to over-selection, as those described under the frequentist paradigm.

Now recover the complete model (1.10), with any number of treatments $T > 1$. This is commonly referred to as the *multiple treatments* problem, where there are simultaneous interventions potentially affecting the outcome, all of which can be endogenous with the control covariates. Although this problem is recently gaining attention with the scalability of recorded data in many scientific fields, to the best of my knowledge the list of available methodology in the literature tailoired to this specific problem is short. Beyond PCR introduced before, perhaps the most notable recent contribution is ACPME by Wilson et al. (2018), a method with strong ties to BAC. This is also a BMA-based algorithm with a specific model prior that incorporates, for each feature, a measure of correlation strength between each control and the set of treatments, this time without imposing sudden jumps in prior probabilities. Similarly to BAC, the philosophy of ACPME is to tilt prior probabilities towards models fully adjusting for confounding. Again, this can put a good fraction of prior mass on super-models of the true outcome model, as feature inclusion can only be encouraged and, hence, is not designed to overcome over-selection problems. More so taking into account that marginal prior inclusion probabilities are capped below at 1/2 by construction. It is also unclear whether it can perform well in high dimensions or with a large number of treatments. Finally, in the context of treatment heterogeneity, a mention to recent work on the Debiased Orthogonal Lasso by Semenova et al. (2020) that can potentially extend the frequentist DML scheme to the multiple treatments, although it might not be the focus of its current formulation.

Chapter 2

MULTIPLE TREATMENT EFFECT INFERENCE VIA CONFOUNDER IMPORTANCE LEARNING

2.1 Motivation

This chapter addresses a problem of fundamental importance in applied research, that of evaluating the joint effect, if any, of multiple treatments on a response variable. A motivating application is the quantification of salary variation (response) due to discriminating factors (the treatments), such as gender, race and country of birth, and how this has evolved over time, using the well-known Current Population Survey data (see Chapter 3 for details). Another example (Orben and Przybylski, 2019) is to infer the effects of multiple technologies such as social media, internet and video games on teenager mental well-being while accounting for, among many others, gender, social and financial difficulties. The data used to inform public policy are often collected from observational studies hence any analysis for treatment effects is subject to selection biases and unmeasured confounding. Such biases are even stronger in "big data" that are currently routinely collected and analysed (Dunson, 2018). Thus, it is necessary to control for a large number of covariates, many of which might be correlated with the treatments. For example, gender, race and country of birth are strongly associated with access to education, occupational sector and other controls that are key determinants of salary. We refer to the covariates as *controls*, and when correlated to both the response and the treatments we call them *confounders*. Our analysis of salary variation involves more than 200 treatments and hundreds of controls.

Predicting a variable from hundreds or thousands of others, even with small training sample, can be done with good performance and even theoretical guarantees using a shrinkage and selection estimation framework such as the LASSO and Bayesian model

averaging (BMA). These approaches assume a *sparsity* structure according to which only few of the covariates are active, i.e., have non-zero regression coefficients. BMA also allows for model-averaged inference and this can be preferable when sparsity is an unwarranted assumption (Giannone et al., 2021). In the statistical nomenclature the selection aspect of these procedures is often compared to searching for needles in a haystack (Johnstone and Silverman, 2004; Castillo and van der Vaart, 2012). In our setting the "needles" correspond to the confounders. As we explain below, these predictive approaches (even BMA) are inappropriate for treatment effect inference, mainly when treatments and controls are correlated, which is typically the case. Recently, methods have been developed and already gained popularity for treatment effect inference with many controls, such as the LASSO-based Double Machine Learning (DML, Belloni et al., 2014a) and the Bayesian Adjustment for Confounding (BAC, Wang et al., 2012) that we reviewed in the previous chapter. The philosophy of these approaches is to encourage the selection of controls when they are correlated with the treatments, hence treat those as confounders.

In this chapter we highlight an overlooked problem for treatment effect inference due to the over-selection of controls, which is exacerbated as the number of controls correlated to treatments but not to the response decreases, as the number of treatments increases and as the treatment effects become smaller. This has serious implications for policy making, since these methods have reduced power to detect effects; for example they fail to find evidence of differences across states in salary discrimination due to e.g. black race in 2019 although a number of alternative analyses suggest the existence of moderate deviations from the main effects in some states (see Section 3.2.2). We address these issues with a new, simple methodology that is based on BMA but where the selection propensity (expressed in terms of prior inclusion probabilities) differs for each control and it is directly informed by the treatment and control data — in other words, we learn to tell the straws apart in the haystack. The formulation also allows relaxing the sparsity assumption discussed above, by setting high overall inclusion probabilities, when this is warranted by the data.

In more detail, we model the dependence of the response $y_i \sim p(y_i; \eta_i, \phi)$ on $t = 1, \ldots, T$ treatments $d_{i,t}$ and $j = 1, \ldots, J$ controls $x_{i,j}$, via

$$\eta_i = \sum_{t=1}^{T} \alpha_t d_{i,t} + \sum_{j=1}^{J} \beta_j x_{i,j}, i = 1, \ldots, n, \tag{2.1}$$

where $p(y_i; \eta_i, \phi)$ defines a generalised linear model with linear predictor η_i and dispersion parameter ϕ. Note that the controls, but also the treatments, might also include interaction terms and other transformations, such as polynomial or spline terms, to accurately capture the effect of treatments and controls. Whereas from an interpretational and policy making point of view the distinction between treatments and controls is clear, statistically the difference is one of priorities: we are primarily interested in inference

for the former, i.e the set of α_t's in (2.1), including uncertainty quantification such as high probability intervals for α_t, whereas the latter are included in the model to avoid omitted variable biases.

Popular frameworks for learning (2.1) such as the LASSO and BMA are useful for prediction purposes but less so for inference for the α_t's. Intuitively, when treatments and controls are correlated, the predictive model might use a subset of those really active to predict the response, resulting in significant under-selection biases when trying to infer treatment effects. Additionally, optimization-based methods such as the LASSO require subsequent analysis of their output, known as *post-selection inference*, which can result in significant loss of power to detect weaker effects. For example, in linear regression with uncorrelated controls the debiased LASSO of van de Geer et al. (2014) recovers the ordinary least-squares (OLS) inference, which is undesirable in high-dimensional settings. The more promising approach of Lee et al. (2016) is not even applicable if the treatment variables are not selected by the LASSO.

As we reviewed in Chapter 1, recent approaches have tried to address the aforementioned bias by encouraging the selection of controls that are correlated to the treatments. Among those, for reasons explained in this chapter, we concentrate on DML and BAC mentioned above. DML regresses separately the response and the treatments on the controls via penalised likelihood, typically LASSO, and in a second step fits a model like (2.1) by OLS with the controls selected in the first stage. In a similar spirit, BAC models jointly the response and treatments and uses a prior distribution that encourages controls to be simultaneously selected in the two regression models. Such methods encourage adding controls to the regression model for y_i that are correlated to the treatments, but are not conditionally related to the response, which has two effects that we highlighted in the previous chapter. The first one is a fairly obvious *over-selection variance*. This refers to an inflation of the standard errors of the treatment effects in the regression model for y_i, due to the larger co-linearity between variables in the model, which leads to a reduced power to detect weaker effects. The second one is more subtle, since the inclusion of controls in (2.1) which were screened out to be correlated with the treatments leads to biased inference for the treatment effects, a property we refer to as *control over-selection bias*. For example, suppose that a control truly has no effect on the outcome ($\beta_j = 0$) and that a data analysis method includes it into (2.1) when it has a strong observed correlation with a treatment. If the latter truly has an effect ($\alpha_t \neq 0$) then β_j is likely to be over-estimated and, since this estimate is correlated to that of α_t, the latter can become biased. Said over-selection bias and variance worsen as the number of treatments increases and as the level of confounding decreases, and it is more hurtful as treatment effects become smaller.

In Figure 2.1 we consider a single treatment simulated according to linear regression on 6 active controls ($\beta_j \neq 0$ in (2.1)), and we vary the number of controls associated to the treatment from 0 (no confounding) to 6 (full confounding, i.e., the same controls are

used for generating the response and the treatment). While LASSO and BMA perform worse the stronger the confounding due to control under-selection, DML and BAC perform well in the presence of strong confounding but poorly in the lack of it. In Figure 3.1 (described later we show that these effects can be exacerbated in the presence of an increasing number of treatments. Our proposed approach, Confounder Importance Learning (CIL), can deal successfully with both over- and under-selection, in both high and low confounding situations, and in the presence of multiple treatments. A first illustration of the merits of CIL is given in Figure 2.1, where it achieves good performance across the spectrum.

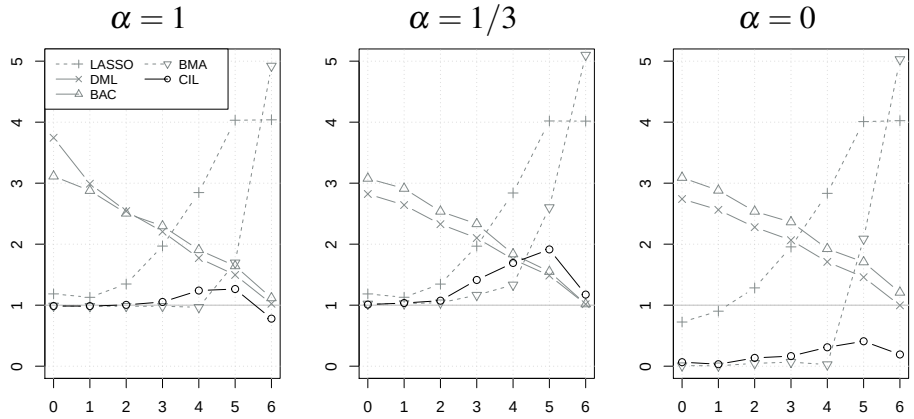

Figure 2.1: Parameter RMSE relative to an oracle OLS, for a single treatment effect ($T = 1$) averaged over 250 simulated datasets, considering strong ($\alpha = 1$), weak ($\alpha = 1/3$) and no effect ($\alpha = 0$). In all panels, $n = 100$, $J = 49$ and the response and treatment are simulated from a linear regression model based on 6 active controls each. The overlap between the two sets of active controls varies from 0 (no confounding) to 6 (full confounding). DML is double machine learning, BMA is Bayesian model averaging, BAC is Bayesian Adjustment for Confounding and CIL is confounder importance learning introduced in this paper.

CIL is based on a Bayesian shrinkage and selection framework for learning (2.1), one where the prior probability of including a control in the response model, $\pi_j = \mathrm{P}(\beta_j \neq 0)$, varies with j in a manner that is learned from data. We build a model

$$\pi_j(\theta) = \rho + (1-2\rho)\left(1+\exp\left\{-\theta_0 - \sum_{t=1}^{T}\theta_t f_{j,t}\right\}\right)^{-1} \tag{2.2}$$

which uses a variation of the logistic link function to express these probabilities in terms of *features* $f_{j,t} \geq 0$ extracted from the treatment and control data, and hyper-parameters $\rho \in (0, 1/2)$ and $\theta = (\theta_0, \theta_1, \ldots, \theta_T)$. The role of ρ is to bound the probabilities away

from 0 and 1 and, as we discuss in Section 2.2, we propose the default choice $\rho = 1/(1+J^2)$. Our method relies on a good choice of the features $f_{j,t}$ and the hyper-parameters θ.

Section 2.2 describes machine learning approaches to obtain the features, and the main idea is to obtain rough estimates of the relative impact of each control to predict each treatment. Regarding the choice of the hyper-parameters, note that with $\theta_t = 0$ for $t = 1,\ldots,T$ the inclusion probabilities are the same for all controls and the size of θ_0 determines whether they are all low or high (corresponding to making or not making a sparsity assumption). When $\theta_t > 0$, controls found to predict treatment t are encouraged to be included in the response model, and discouraged when $\theta_t < 0$. This is in contrast to methods such as DML and BAC that encourage the inclusion of any control associated with any treatment, i.e. with large $f_{j,t}$. We use the data to learn the θ_t's, specifically in Section 2.3.2 we adopt an empirical Bayes choice based on optimizing the marginal likelihood and we design a suitable computational strategy to this effect. We also propose in Section 2.3.3 a much faster alternative based on an expectation-propagation variational approximation. This approximation can be used either in isolation (our experiments suggest that it is often indistinguishable from the empirical Bayes estimator, but they also suggest when to expect differences) or as a way to initialize the optimization of the marginal likelihood.

In Chapter 3, we show some further advantages of our formulation. When considering the salary discrimination application, we illustrate how considering multiple treatments allows to portray, via posterior predictive inference, a measure of joint salary variation due to potentially discriminatory factors; this is visually depicted in Figure 3.9 both aggregated across the U.S.A. and disaggregated according to states. Our results suggest that in 2019 said variation decreased nation-wide (from 5.4% in 2010 to 1.5% in 2019) and state-wide, with lesser disparities across states compared to 2010.

All additional empirical results are extensively analysed in Chapter 3, while technical results are available in Section 2.4.

2.2 Modelling Framework

2.2.1 Sparse Treatment and Control Selection

We model the dependence of the response y_i on treatments $d_i = (d_{i,1},\ldots,d_{i,T})$ and controls $x_i = (x_{i,1},\ldots,x_{i,J})$, according to (2.1), where $\phi > 0$ is a dispersion parameter. We are primarily interested in inference for $\alpha = (\alpha_1,\ldots,\alpha_T)$, i.e. the *treatment effects*. We call *single treatment* to the special case $T = 1$, while for $T > 1$ we have *multiple treatments*.

We adopt a Bayesian framework where we use variable inclusion indicators $\gamma_j =$

$\mathrm{I}(\beta_j \neq 0)$ and $\delta_t = \mathrm{I}(\alpha_t \neq 0)$, and define a model prior

$$p(\alpha, \beta, \delta, \gamma, \phi \mid \theta) = p(\alpha, \beta \mid \delta, \gamma, \phi) p(\gamma \mid \theta) p(\delta) p(\phi), \tag{2.3}$$

where θ are the hyper-parameters in (2.2), and $p(\phi)$ is dropped for models with known dispersion parameter (e.g. logistic or Poisson regression). For the regression coefficients, we assume prior independence,

$$p(\alpha, \beta \mid \delta, \gamma, \phi) := \prod_{t=1}^{T} p(\alpha_t \mid \delta_t, \phi) \prod_{j=1}^{J} p(\beta_j \mid \gamma_j, \phi),$$

and adopt the so-called product moment (pMOM) non-local prior of Johnson and Rossell (2012), according to which $\alpha_t = 0$ if $\delta_t = 0$, and

$$p(\alpha_t \mid \delta_t = 1, \phi) = \frac{\alpha_t^2}{\tau\phi} \mathrm{N}(\alpha_t; 0, \tau\phi),$$

with the analogous setting for every β_j. Figure 2.2 illustrates the density of the product MOM non-local prior. This prior involves a hyper-parameter $\tau > 0$, that we set to $\tau = 0.348$, following Johnson and Rossell (2010), so that the prior signal-to-noise ratio $|\alpha_t|/\sqrt{\phi}$ is greater than 0.2 with probability 0.99. Non-local priors consistently learn which α_t's and β_j's are non-zero as the sample size $n \to \infty$ on a range of high-dimensional linear and generalised linear regression models and play an important role in helping discard spurious predictors (Johnson and Rossell, 2012; Wu, 2016; Shin et al., 2018; Rossell, 2021). We dedicate Chapter 4 to provide some further asymptotic properties attained by the pMOM prior, compared to those available for a wide class of classical local priors. For models with a dispersion parameter, such as linear regression, we place a standard $\phi \sim \mathrm{IGam}(a_\phi = 0.01, b_\phi = 0.01)$ prior, see e.g. Gelman (2006).

For the inclusion indicators, we also assume prior independence, and set

$$p(\delta) = \prod_{t=1}^{T} \mathrm{Bern}(\delta_t; 1/2), \tag{2.4}$$

$$p(\gamma \mid \theta) = \prod_{j=1}^{J} \mathrm{Bern}(\gamma_j; \pi_j(\theta)). \tag{2.5}$$

All treatments get a fixed marginal prior inclusion probability $\mathrm{P}(\delta_t = 1) = 1/2$, as we do not want to favour their exclusion a priori, considering that there is at least some suspicion that any given treatment has an effect. This choice is a practical default when the number of treatments T is not too large, else one may set $\mathrm{P}(\delta_t = 1) < 1/2$ to avoid false positive inflation due to multiple hypothesis testing (Scott and Berger, 2010; Rossell, 2021). Our software allows the user to set any desired $\mathrm{P}(\delta_t = 1)$.

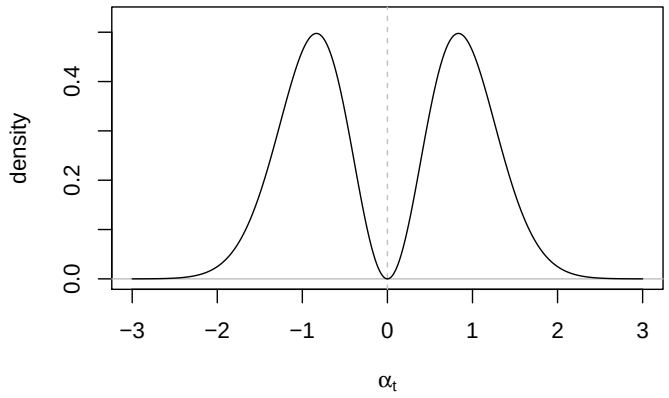

Figure 2.2: Prior density $p(\alpha_t \mid \delta_t = 1, \phi = 1)$ of the MOM non-local prior, with $\tau = 0.348$.

The main modelling novelty in the work in this chapter is the choice of $\pi_j(\theta)$, which we set according to (2.2). A key part of the construction is the choice of features $f_{j,t}$. Our generic approach is to take $f_{j,t} = |w_{j,t}|$, where $w_t = (w_{1,t}, \ldots, w_{J,t})$ are regression coefficients obtained via a high-dimensional regression of d_t on the controls. We highlight two possibilities. First, a LASSO regression,

$$w_t := \underset{(v_{t,1},\ldots,v_{t,J})}{\arg\min} \left\{ \sum_{i=1}^{n} \log p\left(d_{i,t}; \sum_{j=1}^{J} x_{i,j} v_{t,j} \right) + \lambda \sum_{j=1}^{J} |v_{t,j}| \right\}, \tag{2.6}$$

where $\lambda > 0$ is a regularization parameter, which we set by minimizing the BIC (we obtained similar results when using cross-validation). The choice in (2.6) balances speed with reasonable point estimate precision, and is the option that we used in all our examples. A second option, available when dealing with continuous treatments, is to use the minimum norm ridge regression,

$$w_t = (X^\intercal X)^+ d_t, \tag{2.7}$$

where $(X^\intercal X)^+$ is the Moore-Penrose pseudo-inverse, and X the $n \times J$ design matrix. For $J < n$ this is the familiar OLS estimator, but (2.7) is also well-defined when $J > n$, and it has been recently investigated in terms of the so-called benign over-fitting property in Bartlett et al. (2020).

The scalar ρ in (2.2) ensures that prior probabilities are bounded away from 0 and 1. In particular, we set it to $\rho = 1/(J^2+1)$. This sets a lower-bound $\mathrm{P}(\beta_j \neq 0) \geq 1/(J^2+1)$ that is of the same order as J grows as the Complexity priors in (Castillo and van der Vaart, 2012), which are sufficiently sparse to discard irrelevant predictors and attain minimax estimation rates (Castillo et al., 2015; Rossell, 2021).

The final element in (2.2) are the hyper-parameters θ_t, which can encourage the inclusion or exclusion of controls associated to the treatment t. Figure 2.3 illustrates $\pi_j(\theta)$ for three different values of θ_1. Setting θ is critical for the performance of our inferential paradigm, and in Section 2.3 we introduce data-driven criteria and algorithm for its choice.

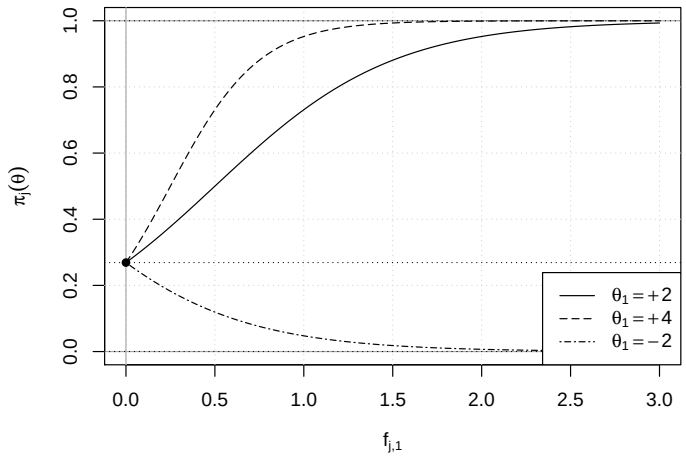

Figure 2.3: Prior inclusion probability in (2.2) as a function of $f_{j,1}$, a feature measuring correlation between control j and treatment $t = 1$, for $\theta_0 = -1$, $\rho = (J^2+1)^{-1}$, and $J = 99$ controls. Top and bottom dotted lines show the upper and lower bounds, $1-\rho$ and ρ, respectively. The dotted line in the middle corresponds to $\theta_1 = 0$.

2.2.2 Connections to The Literature

As reviewed in Section 1.2.3, the main idea in both frequentist and Bayesian literatures is to encourage the inclusion of confounders in (2.1) to mitigate under-selection bias. Farrell (2015) adapted the DML framework of Belloni et al. (2014a) by using a robust estimator to safeguard from mistakes in the double selection step, Shortreed and Ertefaie (2017) employed a two-step adaptive LASSO approach, Antonelli et al. (2018) used propensity matching, and Chernozhukov et al. (2018) extended DML by introducing a de-biasing step, and cross-fitting to ameliorate false positive inclusion of controls. An an alternative to these two-step approaches, Ertefaie et al. (2018) used a joint likelihood L_1 penalization on the outcome and treatment regressions.

Within a Bayesian framework, a natural approach is to build a joint model

$$p(y_i, d_i \mid x_i) = p(y_i \mid d_i, x_i)\, p(d_i \mid x_i), \tag{2.8}$$

where $p(y_i \mid d_i, x_i)$ is as in (2.1) and $p(d_i \mid x_i)$ adds $T \times J$ inclusion indicators ξ_{tj} describing the dependence between each treatment t and control j. BAC (Wang et al., 2012)

considers this approach only for $T = 1$, setting a prior for γ_j where each control has two potential prior inclusion probabilities. If a control j is associated to the single treatment $t = 1$ ($\xi_{tj} = 1$), the prior inclusion probability $\mathrm{P}(\gamma_j = 1)$ increases by a factor determined by a hyper-parameter ω that is set by the user. Lefebvre et al. (2014) and Wang and Leng (2016) provided some theoretical support and proposals to set ω, and Wilson et al. (2018) proposed a multiple treatment extension of BAC. Talbot et al. (2015) introduced Bayesian causal effect estimation, which incorporates informative priors to deter excess control inclusion, and Antonelli et al. (2017) generalised BAC to address treatment effect heterogeneity. From a practical point of view, (2.8) multiplies the size of the model space by a factor of 2^{JT}, rendering the framework impractical even for moderate values of T.

In a different thread, Hahn et al. (2018) proposed a shrinkage prior framework based on re-parameterizing a joint outcome and treatment regression, designed to improve estimation biases, and Hahn et al. (2020) considered non-parametric Bayesian causal forests. Antonelli et al. (2019) proposed a spike-and-slab Laplace prior on the controls that shrinks less those controls that are associated to the treatment, and an empirical Bayes algorithm for hyper-parameter setting.

Our main contributions are of an applied, but relevant, nature: replacing the joint model (2.8) by extracting features derived from $p(d_i \mid x_i)$ to render computations practical, and learning from data whether confounder inclusion should be encouraged, discouraged, or neither, to avoid over-selection issues. In Figure 2.4 we illustrate how this approach compared to other popular available methods discussed before. Another contribution is considering the multiple treatments problem ($T > 1$), which has been considerably less studied.

2.3 Computational Methodology

2.3.1 Bayesian Model Averaging

All expressions in this section are conditional on the observed (x_i, d_i), we drop them from the notation for simplicity. Inference for our approach relies on posterior model probabilities

$$p(\gamma, \delta \mid y, \theta) \propto p(y \mid \gamma, \delta) p(\gamma \mid \theta) p(\delta)$$

where

$$p(y \mid \gamma, \delta) = \int p(y \mid \alpha, \beta, \phi, \delta, \gamma) p(\alpha, \beta \mid \delta, \gamma, \phi) p(\phi) \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\phi \tag{2.9}$$

is the marginal likelihood of model (γ, δ). We set the hyper-parameter θ to a point estimate $\hat{\theta}$ described in the next section. Conditional on θ, our model prior $p(\gamma \mid \theta)$ is

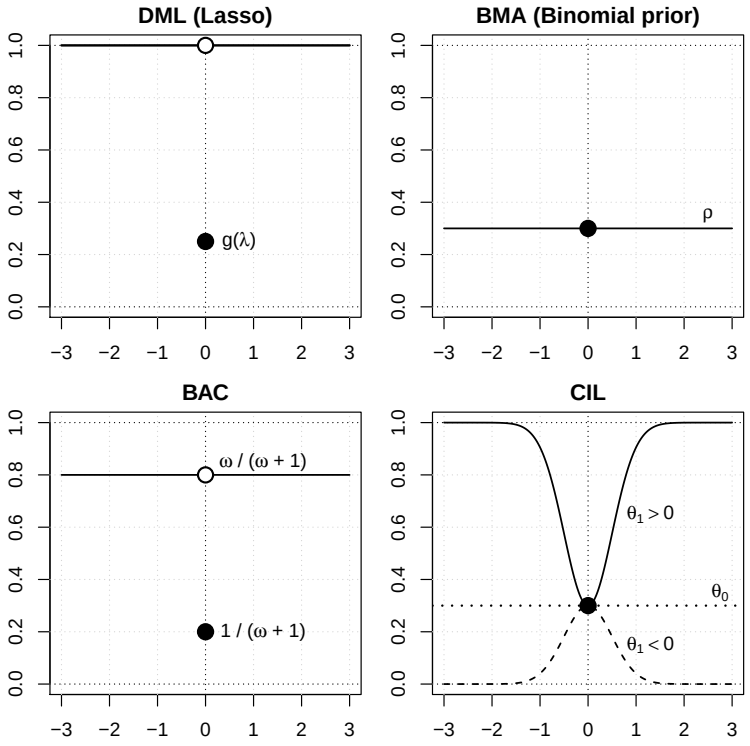

Figure 2.4: Artificial illustration to compare (prior) inclusion probabilities (y-axis) as set by different depicted methods for control x_j on the single treatment model. The x-axis shows the coefficient measuring the strength of the relationship between the treatment and control x_j, in the case of CIL obtained via $v_{i,j}$ in expression (2.6). If this coefficient is zero (no relationship), the value set by DML depends on the Lasso penalty parameter $(\lambda \geq 0)$ via the corresponding function $g(\cdot)$, that of BMA on the Binomial parameter (ρ), and that of BAC on the hyper-parameter $\omega \geq 1$. For CIL, here we have set its paramter $\rho = 0$ in expression (2.2) just for illustration purposes.

a product of independent Bernouilli's with asymmetric success probabilities defined by (2.2). As a simple variation of standard BMA, one can exploit existing computational algorithms, which we outline next.

Outside particular cases such as Gaussian regression under Gaussian priors, (2.9) does not have a closed-form expression. To estimate (2.9) under our pMOM prior we adopt the approximate Laplace approximations of Rossell et al. (2021), see Section 2.3.4.

We obtain point estimates using BMA,

$$\tilde{\alpha} := \sum_{\gamma,\delta} \mathrm{E}(\alpha \mid y, \gamma, \delta) p(\gamma, \delta \mid y, \theta), \tag{2.10}$$

and similarly employ the BMA posterior density $p(\alpha \mid y, \theta)$ to provide posterior credible intervals. To this end we use posterior samples from this density using a latent truncation representation described by Rossell and Telesca (2017). Expression (2.10) is a sum across 2^{T+J} models, which is unfeasible to obtain when $T+J$ is large, then we use Markov Chain Monte Carlo methods, see e.g. Clyde and Ghosh (2012) for a review.

We used all the algorithms described above as implemented by the `modelSelection` function in the `R` package `mombf` (Rossell et al., 2022), for which a new module has been added to accomodate for the `cil` function. This function can be used to implement the methodology just introduced to any requiring treatment effects problem.

2.3.2 Confounder importance learning via Marginal Likelihood

Our main computational contribution is a strategy to learn the hyper-parameter θ, which plays a critical role by determining prior inclusion probabilities. We devised an empirical Bayes approach maximizing the marginal likelihood, with

$$\theta^{\mathrm{EB}} := \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max}\, p(y \mid \theta) = \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{(\delta,\gamma)} p_u(\delta, \gamma \mid y) p(\delta, \gamma \mid \theta) \tag{2.11}$$

where the right-hand side follows easily, denoting by $p_u(\delta, \gamma \mid y)$ the posterior probabilities under a uniform model prior $p_u(\delta, \gamma) \propto 1$. The use of empirical Bayes for hyper-parameter learning in variable selection has been well-studied, see George and Foster (2000); Scott and Berger (2010); Petrone et al. (2014).

A major challenge is that one must evaluate the costly sum in (2.11) for each value of θ considered by an optimization algorithm. Note that $p_u(y \mid \gamma, \delta)$ does not depend on θ, and hence can be re-used to evaluate (2.11) for any number of θ values. In fact, by Proposition 2.1 below, this provides grounds to use stochastic gradient methodology to maximize (2.11).

Proposition 2.1. *If* $p(y \mid \gamma, \delta, \theta) = p(y \mid \gamma, \delta)$*, then*

$$\nabla_\theta \log p(y \mid \theta) = \sum_{(\delta,\gamma)} p(\gamma, \delta \mid y, \theta) \nabla_\theta \log p(\gamma, \delta \mid \theta).$$

If, additionally, the model prior is separable such that

$$p(\gamma, \delta \mid \theta) = \prod_{t=1}^{T} p(\delta_t) \prod_{j=1}^{J} p(\gamma_j \mid \theta),$$

then

$$\nabla_\theta \log p(y \mid \theta) = \sum_{j=1}^{J} \mathrm{E}\left[\nabla_\theta \log p(\gamma_j \mid \theta) \mid y\right], \tag{2.12}$$

where the expectation is with respect to γ_j*.*

Corollary 2.2. *Under the model prior in* (2.4) *and* (2.5)*, and with* $\pi_j(\theta)$ *as defined by* (2.2)*,*

$$\nabla_\theta \log p(y \mid \theta) = (1 - 2\rho) \sum_{j=1}^{J} f_j \left[\mathrm{P}(\gamma_j = 1 \mid y, \theta) - \pi_j(\theta)\right], \tag{2.13}$$

where $f_j = (1, f_{j,1}, \ldots, f_{j,T})^\intercal$*.*

Expressions (2.12) and (2.13) evaluate the gradient with a sum of J terms, relative to the 2^{J+T} terms in (2.11). Further, (2.13) only depends on y via marginal inclusion probabilities $\mathrm{P}(\gamma_j = 1 \mid y, \theta)$, which can typically be estimated more accurately than the joint model probabilities in (2.11). However, two problems remain unaddressed. First, one must compute $\mathrm{P}(\gamma_j = 1 \mid y, \theta)$ for every considered θ, which is cumbersome. Second, $\log p(y \mid \theta)$ can have multiple optima. Hence, standard algorithms may converge to low-quality local optima if θ is poorly initialised. Figure 2.5 (left) shows an example of a multi-modal $p(y \mid \theta)$. We next describe an Expectation Propagation approximation which, as illustrated in Figure 2.5, typically provides a good approximation to the global mode.

2.3.3 Confounder importance learning by Expectation-Propagagation

The use of Expectation Propagation (Minka, 2001a,b) is common in Bayesian machine learning, including in variable selection (Seeger et al., 2007; Hernández-Lobato et al., 2013; Xu et al., 2014). We propose a computationally tractable approximation to (2.11), which can also serve as an initialization point for an algorithm to solve (2.11) exactly, if so desired.

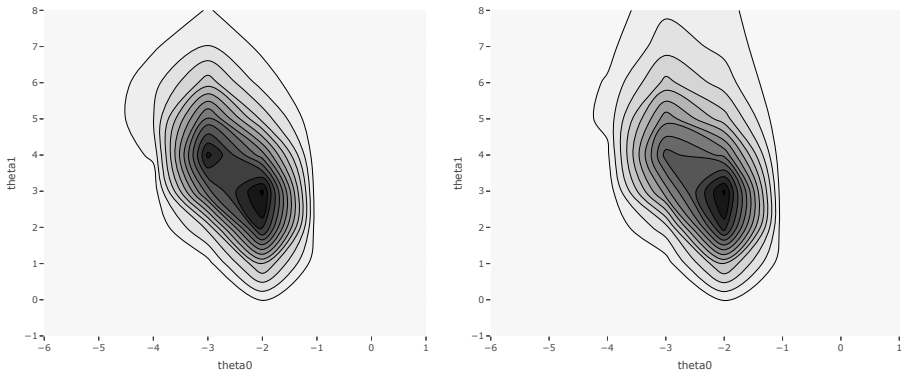

Figure 2.5: Empirical Bayes (left) and Expectation-Propagation (right) objective functions (2.11) and (2.15) in the single treatment case ($T = 1$). Here, $\hat{\theta}^{\text{EB}} = (-2.43, 3.19)$ and $\hat{\theta}^{\text{EP}} = (-2.34, 3.09)$, for $n = 100$ and $J = 49$, for the first data realization for the simulation design displayed in the centre-left panel of Figure 2.1 with three confounders. See Section 3.1 for further details on the simulation setup.

We consider a mean-field approximation to the posterior probabilities in (2.11),

$$\hat{p}_u(\delta, \gamma \mid y) = \prod_{t=1}^{T} \text{Bern}(\delta_t; s_t) \prod_{j=1}^{J} \text{Bern}(\gamma_j; q_j). \tag{2.14}$$

where $s = (s_1, \ldots, s_T)$ and $q = (q_1, \ldots, q_J)$ are given in Proposition 2.3 to optimally approximate $p(\delta, \gamma \mid y)$. By Proposition 2.3 below, (2.14) leads to replacing (2.11) by a new objective function (2.15) that only requires an inexpensive product across J terms. These only depend on y via posterior inclusion probabilities $q_j = \text{P}(\gamma_j = 1 \mid y, \theta = 0_{T+1})$ that can be pre-computed prior to conducting the optimization exercise.

Proposition 2.3. *Let s_t, q_j and $\hat{p}_u(\delta, \gamma \mid y)$ be as defined in* (2.14)*. Then, $s_t^{\text{EP}} = \text{P}(\delta_t = 1 \mid y, \theta = 0_{T+1})$ and $q_j^{\text{EP}} = \text{P}(\gamma_j = 1 \mid y, \theta = 0_{T+1})$ minimize Kullback-Leibler divergence from $p_u(\delta, \gamma \mid y)$ to $\hat{p}_u(\delta, \gamma \mid y)$. Further*

$$\begin{aligned} \theta_u^{\text{EP}} &:= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{\delta, \gamma} \hat{p}_u(\delta, \gamma \mid y) p(\delta, \gamma \mid \theta) \\ &= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{j=1}^{J} \log\left(q_j^{\text{EP}} \pi_j(\theta) + (1 - q_j^{\text{EP}})(1 - \pi_j(\theta))\right). \end{aligned} \tag{2.15}$$

The gradient of the objective function (2.15) is in Section 2.3.4. Since this function may have multiple maxima, we conduct an initial grid search and subsequently use a quasi-Newton BFGS algorithm. See Section 2.3.4 and Algorithm 1 therein for a full description of our algorithm to obtain θ^{EB} and θ^{EP}. In most our examples θ^{EB} and

θ^{EP} provided virtually indistinguishable inference, the latter incurring a significantly lower computational cost, but the exact θ^{EB} did provide slight advantages in some high-dimensional settings (see Section 3.1).

2.3.4 Computational methods

Numerical computation of the marginal likelihood for non-local priors

Briefly, denote by $p_{\text{N}}(\alpha_t \mid \delta_t = 1, \phi) = \text{N}(\alpha_t; 0, \tau\phi)$ independent Gaussian priors for $t = 1, .., T$, and similarly $p_{\text{N}}(\beta_j \mid \gamma_j = 1, \phi) = \text{N}(\beta_j; 0, \tau\phi)$ for $j = 1, \ldots, J$. Proposition 1 in Rossell and Telesca (2017) shows that the following identity holds exactly

$$p(y \mid \gamma, \delta) = p^{\text{N}}(y \mid \gamma, \delta)\text{E}_{\text{N}}\left[\prod_{t=1}^{T} \frac{\alpha_t^2}{\tau\phi} \prod_{j=1}^{J} \frac{\beta_j^2}{\tau\phi} \mid y, \gamma, \delta\right]$$

where $p^{\text{N}}(y \mid \gamma, \delta)$ is the integrated likelihood under $p^{\text{N}}(\alpha, \beta)$, and $\text{E}^{\text{N}}[\cdot]$ denotes the posterior expectation under $p^{\text{N}}(\alpha, \beta \mid y, \gamma, \delta)$. To estimate $p^{\text{N}}(y \mid \gamma, \delta)$ for non-Gaussian outcomes we use a Laplace approximation. Regarding the second term, we approximate it by a product of expectations, which Rossell et al. (2021) showed leads to the same asymptotic properties and typically enjoys better finite-n properties than a Laplace approximation.

Numerical optimization in search of $\hat{\theta}^{\text{EB}}$ and $\hat{\theta}^{\text{EP}}$

Algorithm 2.1 describes our method to estimate $\hat{\theta}^{\text{EP}}$ and $\hat{\theta}^{\text{EB}}$. We employ the quasi-Newton BFGS algorithm to optimize the objective function. For $\hat{\theta}^{\text{EB}}$, we use the gradients from Corollary 2.2, while the Hessian is evaluated numerically using line search, with the `R` function `nlminb`. Note, however, that obtaining $\hat{\theta}^{\text{EB}}$ requires sampling models from their posterior distribution for each θ, which is impractical, to then obtain posterior inclusion probabilities required by (2.13). Instead, we restrict attention to the models M sampled for either $\theta = 0_{T+1}$ or $\theta = \hat{\theta}^{\text{EP}}$ in order to avoid successive MCMC runs at every step, relying on the relative regional proximity between the starting point $\hat{\theta}^{\text{EP}}$ and $\hat{\theta}^{\text{EB}}$. This proximity would ensure that M contains the large majority of models with non-negligible posterior probability under $\hat{\theta}^{\text{EB}}$. For $\hat{\theta}^{\text{EP}}$, we use employ the same BFGS strategy using gradient computed in 2.3.4, with numerical evaluation of the Hessian. This computation requires only one MCMC run at $\theta = 0_{T+1}$, which allows us to use grid search to avoid local optima. As for the size of the grid, we let the user specify what points are evaluated. For K points in the grid one must evaluate the log objective function K^{T+1} times, so we recommend to reduce the grid density as T grows. By default, we evaluate every integer in the grid assuming T is not large, but preferably we avoid coordinates greater than 10 in absolute value, as in our experiments it is very

unlikely that any global posterior mode far from zero is isolated, i.e. not reachable by BFGS by starting to its closest point in the grid. Additionally, even if that were the case, numerically it makes no practical difference, considering that marginal inclusion probabilities are bounded away from zero and one regardless.

Algorithm 2.1 Obtaining θ^{EP} and θ^{EB}

Output: $\hat{\theta}^{\text{EP}}$ and $\hat{\theta}^{\text{EB}}$
1: Obtain B posterior samples $(\gamma,\delta)^{(b)} \sim p(\gamma,\delta \mid y, \theta = 0_{T+1})$ for $b = 1,\ldots,B$. Denote by $M^{(0)}$ the corresponding set of unique models.
2: Compute $s_t = \text{P}(\delta_t = 1 \mid y, \theta = 0_{T+1})$ and $q_j = \text{P}(\gamma_j = 1 \mid y, \theta = 0_{T+1})$.
3: Conduct a grid search for $\hat{\theta}^{\text{EP}}$ around $\theta = 0_{T+1}$. Optimize (2.15) with the BFGS algorithm initialised at the grid's optimum.
4: Obtain B posterior samples $(\gamma,\delta)^{(b)} \sim p(\gamma,\delta \mid y, \theta = \hat{\theta}^{\text{EP}})$. Denote by $M^{(1)}$ the corresponding set of unique models. Set $M = M^{(0)} \cup M^{(1)}$.
5: Initialize search for $\hat{\theta}^{\text{EB}}$ at $\hat{\theta}^{\text{EP}}$. Use the BFGS algorithm to optimize (2.11), restricting the sum to $(\delta,\gamma) \in M$.

Gradient of the function optimised in (2.15) **in Proposition 2.3**

From (2.15), for a given set of q_j we have

$$\theta^{\text{EP}} = \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{j=1}^{J} \log\left(q_j^{\text{EP}} \pi_j(\theta) + (1 - q_j^{\text{EP}})(1 - \pi_j(\theta))\right). \tag{2.16}$$

We are interested in computing the gradient of the function being optimised in (2.16). Denote $h_j(\theta) := q_j \pi_j(\theta) + (1 - q_j)(1 - \pi_j(\theta))$ for short. Simple algebra provides

$$\nabla_\theta h_j(\theta) = (2q_j - 1)\nabla_\theta \pi_j(\theta).$$

From (2.20) we recover the remaining gradient in the last expression and derive

$$\nabla_\theta \log h_j(\theta) = \frac{\nabla_\theta h_j(\theta)}{h_j(\theta)} = \frac{2q_j - 1}{h_j(\theta)} \left[(1 - 2\rho) f_j \pi_j(\theta)(1 - \pi_j(\theta))\right],$$

where $f_j = (1, f_{j,1}, \ldots, f_{j,T})^\intercal$, and so the gradient for the expression in (2.16) is simply

$$\nabla_\theta \sum_{j=1}^{J} \log h_j(\theta) = (1 - 2\rho) \sum_{j=1}^{J} f_j \frac{\pi_j(\theta)(1 - \pi_j(\theta))}{h_j(\theta)}.$$

■

2.4 Technical Appendix

2.4.1 Proof of Proposition 2.1

Let $p(y \mid \gamma, \delta, \theta) = p(y \mid \gamma, \delta)$, then

$$\begin{aligned}
\nabla_\theta \log p(y \mid \theta) &= \frac{\nabla_\theta p(y \mid \theta)}{p(y \mid \theta)} \\
&= \frac{\nabla_\theta \sum_{(\gamma,\delta)} p(y \mid \gamma, \delta, \theta) p(\gamma, \delta \mid \theta)}{p(y \mid \theta)} \\
&= \frac{\sum_{(\gamma,\delta)} p(y \mid \gamma, \delta) \nabla_\theta p(\gamma, \delta \mid \theta)}{p(y \mid \theta)} \\
&= \sum_{(\gamma,\delta)} \frac{p(y \mid \gamma, \delta, \theta)}{p(y \mid \theta)} \frac{p(\gamma, \delta \mid \theta)}{p(\gamma, \delta \mid \theta)} \nabla_\theta p(\gamma, \delta \mid \theta) \\
&= \sum_{(\gamma,\delta)} \frac{\nabla_\theta p(\gamma, \delta \mid \theta)}{p(\gamma, \delta \mid \theta)} p(\gamma, \delta \mid y, \theta) \\
&= \sum_{(\gamma,\delta)} p(\gamma, \delta \mid y, \theta) \nabla_\theta \log p(\gamma, \delta \mid \theta). \qquad (2.17)
\end{aligned}$$

If, further, the model prior satisfies $p(\gamma, \delta \mid \theta) = \prod_{t=1}^{T} p(\delta_t) \prod_{j=1}^{J} p(\gamma_j \mid \theta)$, then

$$\nabla_\theta \log p(\gamma, \delta \mid \theta) = \sum_{j=1}^{J} \nabla_\theta \log p(\gamma_j \mid \theta),$$

and so

$$\begin{aligned}
\nabla_\theta \log p(y \mid \theta) &= \sum_{j=1}^{J} \sum_{(\gamma,\delta)} \nabla_\theta \log p(\gamma_j \mid \theta) p(\gamma, \delta \mid y, \theta) \\
&= \sum_{j=1}^{J} \mathrm{E}\left[\nabla_\theta \log p(\gamma_j \mid \theta) \mid y, \theta \right].
\end{aligned}$$

■

2.4.2 Proof of Corollary 2.2

The empirical Bayes estimate defined by (2.11) writes

$$\theta^{\mathrm{EB}} = \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \log p(y \mid \theta) = \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \log \sum_{(\gamma,\delta)} p(y \mid \gamma, \delta) p(\gamma, \delta \mid \theta).$$

For short, denote $H(\theta) = p(y \mid \theta)$ and $h_j(\theta) = p_j(\gamma_j \mid \theta)$, where generically $\nabla_\theta \log H(\theta) = \nabla_\theta H(\theta)/H(\theta)$. Under the assumptions of Corollary 2.2

$$\begin{aligned}\nabla_\theta H(\theta) &= \sum_{(\gamma,\delta)} p(y \mid \gamma,\delta) p(\delta) \nabla_\theta \prod_{j=1}^{J} h_j(\theta) \\ &= \sum_{(\gamma,\delta)} p(y \mid \gamma,\delta) p(\delta) \sum_{j=1}^{J} \left(\nabla_\theta h_j(\theta) \prod_{j \neq l} h_l(\theta) \right). \end{aligned} \tag{2.18}$$

Denoting $f_j = (1, f_{j,1}, \ldots, f_{j,T})^\intercal$, direct algebra gives

$$\begin{aligned}\nabla_\theta h_j(\theta) &= \nabla_\theta \left\{ \pi_j(\theta)^{\gamma_j} (1 - \pi_j(\theta))^{1-\gamma_j} \right\} \\ &= (1 - 2\rho) f_j (\gamma_j - \pi_j(\theta)) h_j(\theta), \end{aligned} \tag{2.19}$$

since

$$\nabla_\theta \pi_j(\theta) = (1 - 2\rho) f_j \pi_j(\theta)(1 - \pi_j(\theta)). \tag{2.20}$$

Then, replacing (2.19) into (2.18)

$$\begin{aligned}&\nabla_\theta H(\theta) = \sum_{(\gamma,\delta)} p(y \mid \gamma,\delta) p(\delta) \sum_{j=1}^{J} (1 - 2\rho) f_j (\gamma_j - \pi_j(\theta)) \prod_{j=1}^{J} f_j(\theta) \\ &= \sum_{j=1}^{J} (1 - 2\rho) f_j \sum_{(\gamma,\delta)} (\gamma_j - \pi_j(\theta)) p(y \mid \gamma,\delta) p(\delta,\gamma \mid \theta) \\ &= \sum_{j=1}^{J} (1 - 2\rho) f_j \left[(1 - \pi_j(\theta)) \sum_{(\gamma,\delta):\gamma_j=1} p(y,\delta,\gamma \mid \theta) - \pi_j(\theta) \sum_{(\gamma,\delta):\gamma_j=0} p(y,\delta,\gamma \mid \theta) \right]. \end{aligned}$$

Finally

$$\begin{aligned}&\nabla_\theta \log H(\theta) = \frac{\nabla_\theta H(\theta)}{H(\theta)} \\ &= \sum_{j=1}^{J} (1 - 2\rho) f_j \left[(1 - \pi_j(\theta)) \frac{\sum_{(\gamma,\delta):\gamma_j=1} p(y,\delta,\gamma \mid \theta)}{\sum_{(\gamma,\delta)} p(y,\delta,\gamma \mid \theta)} - \pi_j(\theta) \frac{\sum_{(\gamma,\delta):\gamma_j=0} p(y,\delta,\gamma \mid \theta)}{\sum_{(\gamma,\delta)} p(y,\delta,\gamma \mid \theta)} \right] \\ &= \sum_{j=1}^{J} (1 - 2\rho) f_j \left[(1 - \pi_j(\theta)) P(\gamma_j = 1 \mid y,\theta) - \pi_j(\theta)(1 - P(\gamma_j = 1 \mid y,\theta)) \right] \\ &= (1 - 2\rho) \sum_{j=1}^{J} f_j \left[P(\gamma_j = 1 \mid y,\theta) - \pi_j(\theta) \right]. \end{aligned}$$

■

2.4.3 Proof of Proposition 2.3

Consider the right-hand side in (2.11),

$$\underset{\theta\in\mathbb{R}^{T+1}}{\arg\max} \sum_{(\delta,\gamma)} p_u(\delta,\gamma \mid y) p(\delta,\gamma \mid \theta) \tag{2.21}$$

where $p_u(\delta,\gamma \mid y)$ are the posterior probabilities under a uniform prior $p_u(\delta,\gamma) \propto 1$.

We seek to set the parameters s_t and q_j in the approximation

$$\hat{p}_u(\delta,\gamma \mid y) = \prod_{t=1}^{T} \mathrm{Bern}(\delta_t; s_t) \prod_{j=1}^{J} \mathrm{Bern}(\gamma_j; q_j)$$

using Expectation Propagation. That is, setting and $q = (q_1,\ldots,q_J)$ such that

$$q^{\mathrm{EP}} = \underset{q\in[0,1]^J}{\arg\max} \sum_{(\gamma,\delta)} p_u(\delta,\gamma \mid y) \log\left(\prod_{t=1}^{T} s_t^{\delta_t}(1-s_t)^{1-\delta_t} \prod_{j=1}^{J} q_j^{\gamma_j}(1-q_j)^{1-\gamma_j} \right).$$

and analogously for $s = (s_1,\ldots,s_T)$. Proceeding elementwise, we derive

$$\begin{aligned}
q_j^{\mathrm{EP}} &:= \underset{q_j\in[0,1]}{\arg\max} \sum_{(\gamma,\delta)} p_u(\delta,\gamma \mid y) \times \\
&\times \left(\sum_{j=1}^{J} [\gamma_j \log q_j + (1-\gamma_j)\log(1-q_j)] + \sum_{t=1}^{T} [\delta_t \log s_j + (1-\delta_t)\log(1-s_t)] \right) \\
&= \underset{q_j\in[0,1]}{\arg\max} \sum_{(\gamma,\delta)} p_u(\delta,\gamma \mid y) \left(\sum_{j=1}^{J} [\gamma_j \log q_j + (1-\gamma_j)\log(1-q_j)] \right) \\
&= \arg\max_{q_j\in[0,1]} \sum_{j=1}^{J} \sum_{(\gamma,\delta)} p_u(\delta,\gamma \mid y) \left[\gamma_j \log q_j + (1-\gamma_j)\log(1-q_j) \right].
\end{aligned}$$

Optimizing this expression yields

$$\begin{aligned}
\frac{\partial}{\partial q_j} = 0 &\Leftrightarrow \sum_{(\gamma,\delta)} p_u(\delta,\gamma \mid y) \left(\frac{\gamma_j}{q_j^{\mathrm{EP}}} - \frac{1-\gamma_j}{1-q_j^{\mathrm{EP}}} \right) = 0 \\
&\Leftrightarrow \frac{1}{q_j^{\mathrm{EP}}} \sum_{(\gamma,\delta):\gamma_j=1} p_u(\delta,\gamma \mid y) - \frac{1}{1-q_j^{\mathrm{EP}}} \sum_{(\gamma,\delta):\gamma_j=0} p_u(\delta,\gamma \mid y) = 0 \\
&\Leftrightarrow \frac{P_u(\gamma_j=1 \mid y)}{q_j^{\mathrm{EP}}} - \frac{P_u(\gamma_j=0 \mid y)}{1-q_j^{\mathrm{EP}}} = 0 \\
&\Leftrightarrow q_j^{\mathrm{EP}} = \mathrm{P}_u(\gamma_j=1 \mid y) = \mathrm{P}(\gamma_j=1 \mid y, \theta = 0_{T+1}).
\end{aligned} \tag{2.22}$$

With the same exact procedure one analogously obtains $s_t^{\mathrm{EP}} = \mathrm{P}_u(\delta_t = 1 \mid y)$. Let

$$\begin{aligned} h(\delta) &:= \prod_{T=1}^{T} \mathrm{Bern}(\delta_t; s_t^{\mathrm{EP}}) \prod_{j=1}^{J} \mathrm{Bern}(\delta_t; \pi_t) \\ &= \prod_{t=1}^{T} \left[s_t^{\mathrm{EP}} \pi_t\right]^{\delta_t} \left[(1 - s_t^{\mathrm{EP}})(1 - \pi_t)\right]^{1-\delta_t}, \end{aligned}$$

which is independent of θ, and where π_t is the marginal prior inclusion probability within our framework. Then, implementing the approximation (2.22) into (2.21) gives

$$\begin{aligned} \theta^{\mathrm{EP}} &:= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{(\gamma,\delta)} h(\delta) \prod_{j=1}^{J} \mathrm{Bern}(\gamma_j; q_j^{\mathrm{EP}}) \prod_{j=1}^{J} \mathrm{Bern}(\gamma_j; \pi_j(\theta)) \\ &= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{(\gamma,\delta)} h(\delta) \prod_{j=1}^{J} \left[q_j^{\mathrm{EP}} \pi_j(\theta)\right]^{\gamma_j} \left[(1 - q_j^{\mathrm{EP}})(1 - \pi_j(\theta))\right]^{1-\gamma_j}. \end{aligned} \tag{2.23}$$

Note that the product in the RHS of (2.23) defines a probability distribution on $(\delta_1, \ldots, \delta_T, \gamma_1, \ldots, \gamma_J)$ with independent components, hence the sum is the normalizing constant of such distribution. Thus, this constant is just the product of the univariate normalizing constants. The univariate normalizing constant of each Bernouilli is then

$$q_j^{\mathrm{EP}} \pi_j(\theta) + (1 - q_j^{\mathrm{EP}})(1 - \pi_j(\theta))$$

for every q_j, and similarly $s_t^{\mathrm{EP}} \pi_t + (1 - s_t^{\mathrm{EP}})(1 - \pi_t)$ for every s_t. Hence, replacing into (2.23) we obtain

$$\begin{aligned} \theta^{\mathrm{EP}} &:= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \prod_{j=1}^{J} \left\{q_j^{\mathrm{EP}} \pi_j(\theta) + (1 - q_j^{\mathrm{EP}})(1 - \pi_j(\theta))\right\} \prod_{t=1}^{T} \left\{s_t^{\mathrm{EP}} \pi_t + (1 - s_t^{\mathrm{EP}})(1 - \pi_t)\right\}. \\ &= \underset{\theta \in \mathbb{R}^{T+1}}{\arg\max} \sum_{j=1}^{J} \log\left(q_j^{\mathrm{EP}} \pi_j(\theta) + (1 - q_j^{\mathrm{EP}})(1 - \pi_j(\theta))\right). \end{aligned}$$

■

Chapter 3

APPLICATIONS TO CONFOUNDER IMPORTANCE LEARNING

In this chapter we review a number of synthetic and real applications that highlight similarities and differences between Confounder Importance Learning (CIL) introduced in Chapter 2 and other relevant and popular methodology frequently employed in the literature. In particular, we compare our CIL approach (under the EP approximation) to three methods: OLS under the full model, DML based on the LASSO (Belloni et al., 2014a), and standard BMA with a Beta-Binomial$(1,1)$ model prior and the pMOM prior in Section 2.2.1.

In Section 3.1 using simulated data we also compare to BAC (Wang et al., 2012), which was computationally unfeasible to apply to the salary data. We set its hyper-parameter to $\omega = +\infty$, which encourages the inclusion of confounders relative to standard BMA. For completeness, we also considered a standard LASSO regression on the outcome equation (2.1), setting the penalization parameter via cross-validation. We compared these methods to the oracle OLS, i.e. based on the subset of controls truly featuring in (2.1). These methods are implemented in `R` packages `glmnet` (Friedman et al., 2010) for LASSO, `mombf` for BMA and CIL, `hdm` (Chernozhukov et al., 2016) for DML and `BACprior` (Talbot et al., 2014) for BAC.

3.1 Simulation Studies

3.1.1 Main numerical results

To illustrate issues associated to under- and over-selection of controls, a key factor we focus on is the *level of confounding*. Our scenarios range from no confounding (no controls affect both y and d) to complete confounding (any control affecting y also affects d, and vice versa). We also considered the effect of dimensionality, treatment

effect sizes α, and true level of sparsity (number of active controls).

We considered a Gaussian outcome and a single treatment ($T = 1$), and an error variance $\phi = 1$. The controls were obtained as independent Gaussian draws $x_i \sim \mathrm{N}(0_J, \mathrm{I})$, and any active control had a coefficient $\beta_j = 1$. The treatment d was also Gaussian with its mean depending linearly on the controls, unit error variance, and any control having an effect on d had unit regression coefficient. Throughout, the number of controls that truly had an effect on d was set equal to the number of controls that affect the outcome y. We measured the root mean squared error (RMSE) of the estimated $\hat{\alpha}$.

Figures 2.1, 3.3, 3.2 and 3.1 summarize the results. Figure 2.1 shows that the RMSE of BMA and LASSO worsens as confounding increases, this was due to a lower power to select all relevant confounders (see Figure 3.2 for model selection diagnostics), i.e. an omitted variable bias. These effects have been well studied. Methods such as DML and BAC were designed to prevent omitted variables, but as shown in Figure 2.1 they can run into over-selection when there truly are few confounders. In contrast, our CIL performed well at all levels of confounding. The Empirical Bayes and the Expectation Propagation versions of CIL provide nearly indistinguishable results (not shown). It is worth noting that, when the treatment truly had no effect ($\alpha = 0$), CIL provided a strong shrinkage that delivered a significantly lower RMSE than other methods.

Figure 3.3 extends the analysis to consider a growing number of covariates, under a strong treatment effect ($\alpha = 1$). As dimensionality grew, standard LASSO and BMA incurred a significantly higher RMSE under strong confounding. Our CIL generally provided marked improvements over BMA, except for the larger $J + T = 200$. Here we observed the only perceptible differences between the EB and EP approximations, with the former attaining better results, pointing to advantages of the EB approach in higher dimensions. Figure 3.4 further extends the analysis to less sparse settings, with $\|\gamma\|_0 = 6$, 12 and 18 active parameters. Overall, the results were similar to Figures 2.1 and 3.3.

A focus in this paper is to understand over-selection issues in multiple treatment inference. To this end, we added a multiple treatments design with an increasing number of treatments, with a maximum of $T = 5$. There, every present treatment was active, setting $\alpha_t = 1$ on all treatments. For all levels of T, we set $\beta_j = 1$ for $j = 1, \ldots, 20$, denoting the set of active controls by $x_{1:20}$, and $\beta_j = 0$ for the rest of controls $x_{21:J}$. Regarding the association between treatments and controls, $x_{1:20}$ were divided into five disjoint subsets with four variables each, and each of these subsets was linearly associated to a different treatment. Additionally, each treatment also depended on a further subset of controls in the set $x_{21:J}$. In this case, the size of such subset was increasing by four with each added treatment: treatment 1 was associated to $x_{21:24}$, treatment 2 was associated to $x_{21:28}$, etc., up to treatment 5, which correlated to $x_{21:40}$. All controls affecting a treatment had a unit linear coefficient. The rest of the design is akin to that in Figure 2.1. We also replaced BAC with the ACPME method of (Wilson et al., 2018),

an extension of BAC for multiple treatments.

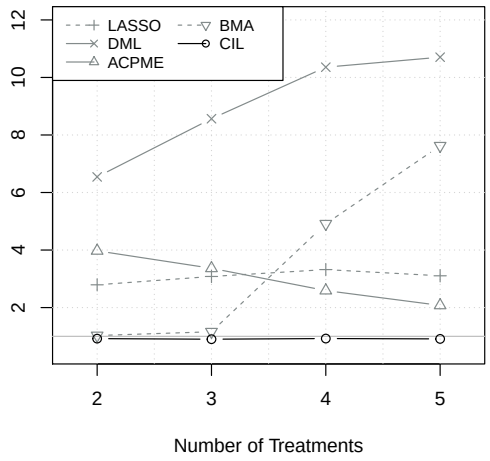

Figure 3.1: Treatment parameter RMSE (relative to oracle least-squares) based on $R = 250$ simulated datasets at every value of T, for $n = 100$, $J = 95$, and $T \in \{2,3,4,5\}$. For every T (x-axis), we show the average RMSE across Treatments $1,\ldots,T$.

Figure 3.1 shows the estimation results on α for the different values of T, akin to Figure 2.1. We observe similar trends as before. DML included too many controls, particularly for larger T. On the other hand, under-selection (here suffered by BMA) was also problematic, as for larger T the model became highly confounded, as a subset of the controls accounted for a larger proportion of the variance in the outcome, as well as for that of the treatment(s). This led to BMA discarding with high probability active but highly correlated variables between treatments and confounders. We also observed a stable performance of ACPME that improved slightly as T grew, but even for $T = 5$ its RMSE more than doubled that of oracle OLS. On the other end, our CIL proposal was able to achieve oracle-type performance for every examined value of T.

3.1.2 Supplementary numerical results

Figure 3.2 summarises model selection results for the simulations described in Figure 2.1.

Figure 3.3 studies the effect of growing number of covariates on inference, specifically for $J+T = 25$, 100 and 200.

Figure 3.4 shows the effect of having various amounts of active confounders. The results look consistent to the effects reported in Figures 2.1 and 3.3, which were magnified for large amounts of active confounders. These are really challenging situations to tackle since the tested methods aim at model sparsity, while the true model size is relatively large. Although our method still performed at oracle rates in low-confounding

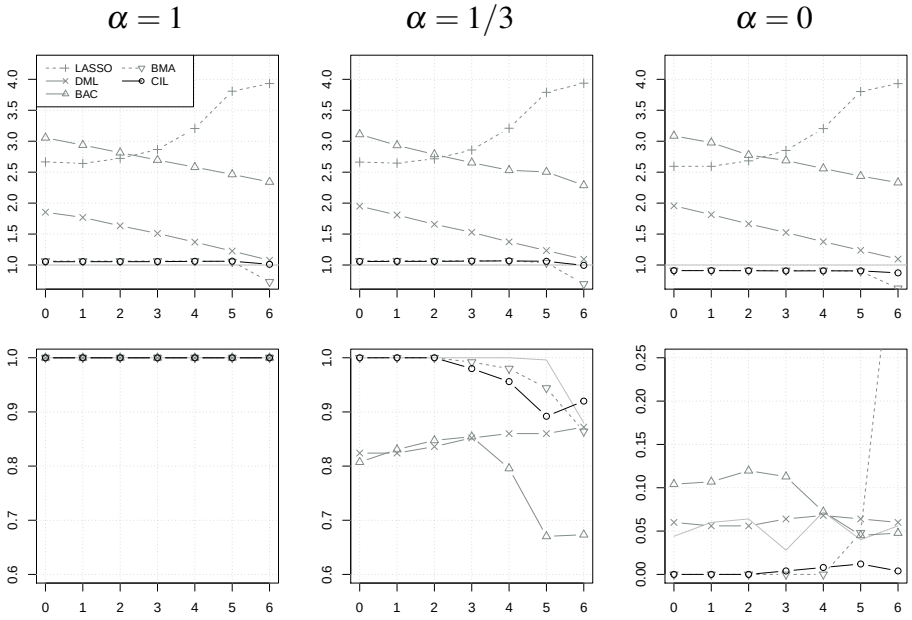

Figure 3.2: To be read vertically in relation to Fig. 2.1. The top panels show the average outcome model size across levels of confounding, divided by the true model size (i.e. 1 indicates that it matches the true model size). The bottom panels show the probability of selecting the treatment using a 0.05 P-value cut-off for DML, and for Bayesian methods the treatment is included when marginal posterior inclusion probability is $>1/2$. LASSO does not appear in these panels as its not designed for inference.

scenarios, its relative performance was compromised for the highest levels of confounding. This occurred in part because accurate point estimation in (2.6) became increasingly harder as the correlation between covariates strengthened, which in turn influenced the ability of the algorithm to calibrate θ reliably. Even in these hard cases, however, its performance was not excessively far to the best competing method, while it clearly outperformed BMA on all of them.

3.2 Wage Discrimination on The Current Population Survey

We studied the association between belonging to certain social groups and the hourly salary, and its evolution over the last decade (prior to the COVID-19 pandemic), to assess progress in wage discrimination. We analysed the U.S.A. Current Population Survey (CPS) microdata (Flood et al., 2020), which records many social, economic and job-related factors. The outcome is the individual log-hourly wage, re-scaled by the

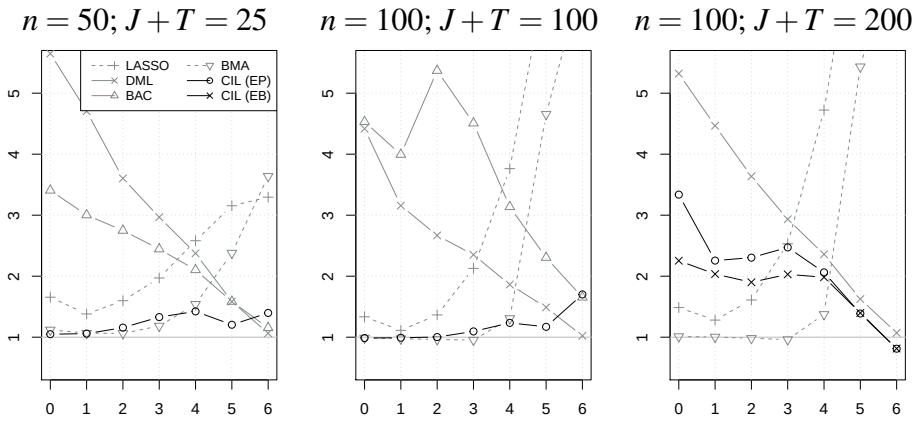

Figure 3.3: Single treatment parameter RMSE (relative to Oracle OLS) based on $R = 250$ simulated datasets for each level of confounding. In all panels, $\alpha = 1$ and $\|\gamma\|_0 = 6$. We show the empirical Bayes version CIL only in the right panel, for the other panels results are indistinguishable relative to EP.

consumer price index of 1999, and we considered four treatments: gender, black race, Hispanic ethnicity and Latin America as place of birth. These treatments are highly correlated to sociodemographic and job characteristics that can impact salary, i.e. there are many potential confounders.

Section 3.2.1 describes the data and Section 3.2.2 contains results on the treatment effects, both individually and in terms of a composite score measuring their joint association with salary. These results support that methods designed for treatment effect inference may run into over-selection, whereas naive methods may run into under-selection. To provide further insight, Section 3.1 shows simulation studies, with particular attention on how the presence/absence of confounders affects each method.

3.2.1 Data

We downloaded data from 2010 and 2019 and analysed each year separately. We selected individuals aged over 18, with a yearly income over \$1,000 and working 20 to 60 hours per week, giving $n = 64,380$ and $n = 58,885$ in 2010 and 2019, respectively. The controls included characteristics of the place of residence, education, labor force status, migration status, household composition, housing type, health status, financial and tax records, reception of subsidies, and sources of income (beyond wage). Overall, there were $J = 229$ controls, after adding 50 binary indicators for state.

Since every state has its own regulatory, sociodemographic and political framework, we captured state effects by adding interactions for each pair of treatment and state. On these interactions, we applied a sum to zero constraint, so that the coefficients associated to the four treatments remain interpretable as average effects across the USA, and the

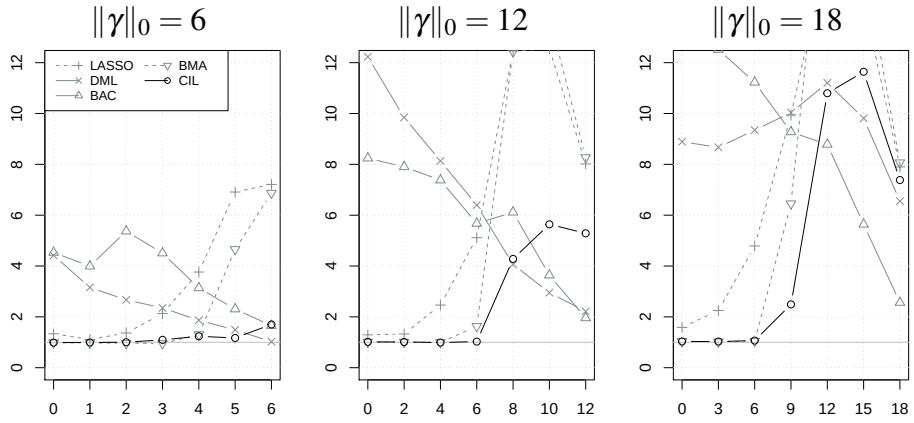

Figure 3.4: Single treatment parameter RMSE (relative to Oracle OLS) based on $R = 250$ simulated datasets for each level of confounding reported, as described in Figure 2.1. In all panels, $n = 100$, $J + T = 100$ and $\alpha = 1$. Sudden general improvement at the right end of centre and right panels is due to a sharper deterioration of oracle OLS RMSE at complete confounding relative to other methods.

interactions as deviation from the average. Hence, overall, we have $T = 4 + 4 \times 50 = 204$ treatments, our main interest being in the first four. In our CIL prior we assumed a common θ_t shared between each main treatment and all its interactions with state, so that $\dim(\theta) = 5$.

To study issues related to under- and over-selection, we analysed the original data and two augmented datasets where we added artificial controls correlated with the treatments but not the outcome. The augmented data incorporated 100 artificial controls in the first scenario, and 200 in the second one, which were split into four subsets of size 25 and 50, respectively. Each of these subsets was designed to correlate to one of the four main treatments. The simulation procedure worked as follows. For both amounts $K_1 = 100$ and $K_2 = 200$ of artificial predictors, the simulation protocol was the same. Every artificial control $z_k \in \mathbb{R}^n$, for $k = 1, \ldots, 100$ or $k = 1, \ldots, 200$ respectively, was simulated to correlate to one individual treatment, according to which subset said control was assigned to, correlating only indirectly to the rest of treatments. In particular, we drew elements of z_k from $z_{i,k} \mid d_{i,t} = 1 \sim \mathrm{N}(1.5, 1)$, and $z_{i,k} \mid d_{i,t} = 0 \sim \mathrm{N}(-1.5, 1)$, where d_t denotes the corresponding column in the treatment matrix associated to the given z_k. The resulting average correlation between gender and its associated artificial variables was of 0.83, and analogously of 0.69, 0.76 and 0.67 for black race, ethnicity and place of birth with their corresponding correlated variables, respectively.

3.2.2 Salary survey results

In Figure 3.5 we focus on the gender and black race indicators. All considered methods show that the average log-salary is reduced for women, and that this gap is similarly pronounced in 2019 relative to 2010. However, the methods differ in their conclusions for the black race. To understand better what drives these differences, we added 100 and 200 simulated controls that are dependent on the treatments but conditionally independent of the response. The figure shows a marked robustness of Bayesian methods to the addition of said controls, whereas other methods lose their ability to detect the weaker effects (e.g. race in both 2010 and 2019). We see this as empirical evidence that previous methods fail to detect the effect for black race due to existence of many controls correlated to the treatments, an inability that it is only exacerbated when adding the simulated controls.

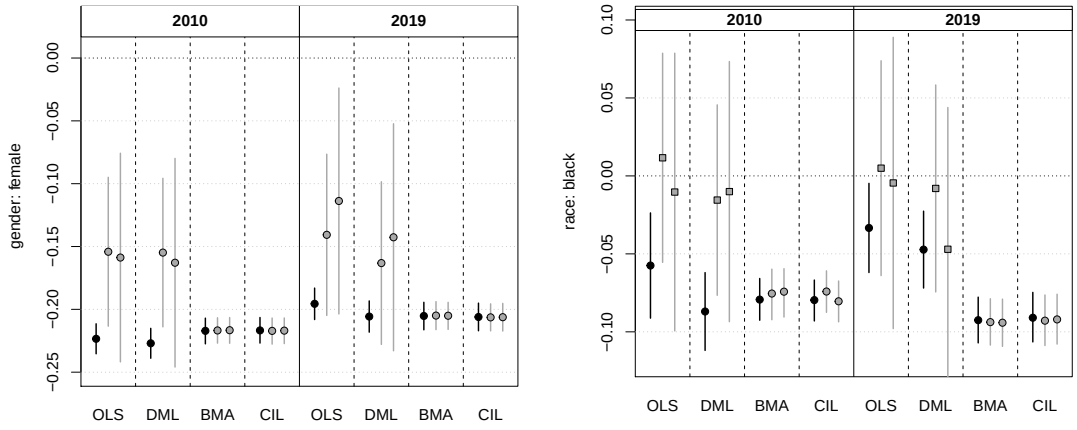

Figure 3.5: Inference for treatments "female" (left) and "black" (right) in 2010 and 2019; see Section 3.2. We analyze Current Population Survey data with $J = 229$ controls (left black point and bar in each panel) but also adding 100 (middle) and 200 (right) artificial controls correlated with the treatments and conditionally independent of the response. Names of methods as in the caption of Figure 2.1.

The treatment effect for gender is picked up by all methods in both years with similar point estimates. No method finds any remarkable decrease of this effect in 2019, only OLS detects a very moderate significant reduction in this gap. When adding the artificial controls, the confidence intervals for OLS and DML became notably wider, which can potentially lead to a loss of statistical significance. This points towards a relevant loss in power due to over-selecting irrelevant controls, with the associated variance inflation. The BMA and CIL results were particularly robust to the artificial controls.

As for race, we obtained a similar pattern of results results. In 2010, All methods found a significant negative association between black race and salary. In 2019, OLS

and DML found a somewhat smaller association, while both Bayesian methods found a negative relation similar to that established in 2010, with a mildly more negative point estimate. Once we introduced the artificial controls, we observed that OLS and DML suffered a large variance inflation. On the other end, BMA and CIL experience no perceptible change to adding the artificial controls. These results seem to suggest that they have sufficient power to detect the difference that other methods can underestimate in the original data in 2019.

As argued earlier, the differences in results between methods are due to a different incorporation of controls into (2.1). For example, on the original data OLS found that in the full model 165 and 176 predictors had an associated p-value smaller than 0.05 in 2010 and 2019, respectively, while DML selected 138 and 135 variables, i.e. showing almost no model size reduction across years compared to OLS. BMA had a posterior model size of 85 in both years, comparably deterring the inclusion of a large number of covariates in the model, similarly to CIL that included 85 and 86 predictors on average. Model sizes are altered in the augmented datasets: with the 200 artificial predictors, in 2010 OLS grows the model up to 172 variables, and DML grows up to 146 variables. Differences between BMA and CIL are not more pronounced in the augmented datasets for 2010, staying in a range between 84 to 87 variables for either method, year, and augmented dataset.

Figure 3.6 follows Figure 3.5 by showing the results for the other two treatments: Hispanic ethnicity, and birthplace in Latin America. In the case of ethnicity, OLS and DML pick up a small but negative effect on log-salary, which is barely significant in 2010, and only maintained by DML in 2019. Neither of the Bayesian methods detect any effect in these years. The augmented datasets pose the same problems as in the previous figure in relation to variance inflation and point estimate instability of OLS and DML. We also observe minor sensibility of Bayesian methods to this intervention: interval lengths widen for BMA in 2010 and, more moderately, for CIL in 2019. In both cases, however, negative associations remain insignificant for them. As for place of birth, no method detects any association to salary in 2010, but this changes notably in 2019, with the appearence of a negative effect of being born in Latin America on log-hourly wage according to all methods. There is a significant difference in point estimation between frequentist and Bayesian methods tested, close to double the effect for the latter group, although intervals are considerably wider, once more. In any case, the negative effects in 2019 are significant to all. Additionally, on this treatment CIL seems to attain narrower intervals compared to BMA, as in this aspect BMA also looks moderately more sensitive to the noise of augmented datasets in both years, altering the significance of the point estimates in 2010.

Contrary to the simulations of Section 3.1, not too many differences between BMA and CIL appear in the main effects of the CPS application. We briefly depict their respective performance next. In Figure 3.7 we appreciate that on the original dataset the

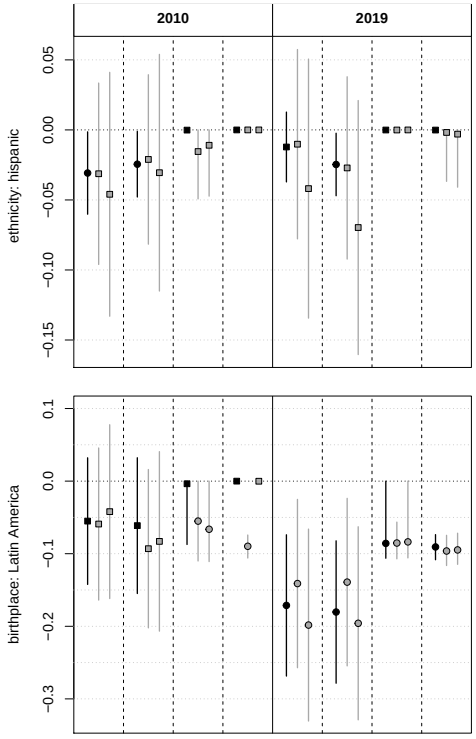

Figure 3.6: Inference for treatment variables “hispanic” (top) and “born in Latin America” (bottom) in 2010 and 2019; see Section 3.2. Read caption to Figure 3.5 to read this figure.

estimated values for the elements of $\hat{\theta}$ are quite close to zero, especially in 2010, which leads to marginal prior inclusion probabilities close to $1/2$. Since the marginal prior inclusion probabilities under the BetaBin$(1,1)$ model prior are precisely $1/2$, it seems consistent that no meaningful differences between the two methods are spotted in the original data. This points to this being a "low-confounding" scenario as defined in the simulations of Section 3.1, situations where both methods were reported to perform similarly well on average. This is not so much the case for the augmented datasets, where we the estimated $\hat{\theta}$ elements of $\hat{\theta}$ are incrementally smaller and further away from zero, leading to smaller marginal prior inclusion probabilities for CIL on average, influenced by the large amount of spurious regressors intentionally introduced in the augmented datasets. However, we do observe some differences if we consider treatment effect heterogeneity, i.e. if we analyse the set of interactions between the main treatments and the state-level binary variables. In Figure 3.8 we illustrate the treatment "black" in 2019. As observed in Figure 3.5, the main effect is persistent in magnitude for both BMA and CIL in 2019, however, some improvement in power is attained by CIL with respect to BMA when trying to capture state-level heterogeneity. CIL is able to detect that a fair number of states deviate from the national average, including for example Hawai'i and Wyoming, which actually show no significant discrimination to "black" (zero is within the posterior interval), arising from signal undetected by BMA. Similarly, some states show a negative deviation from an already negative average, effects only captured by CIL.

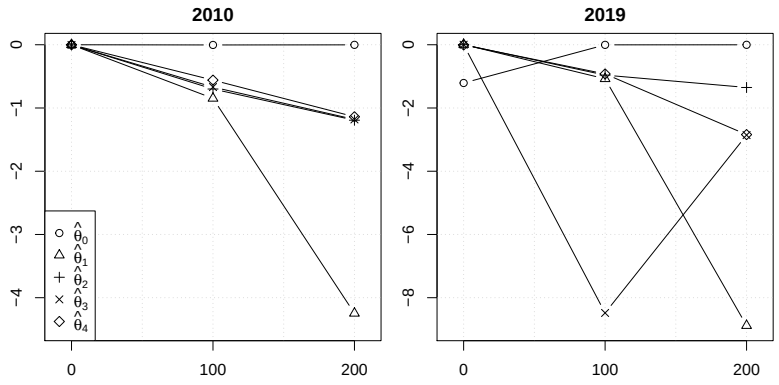

Figure 3.7: Evolution of the elements of $\hat{\theta}$ as estimated by CIL (EP version) on the CPS data for 2010 and 2019. In the x-axis we represent the amount of artificial predictors included in the model, where "0" reports the values for the original dataset.

The full scope of our proposed approach is materialised when considering more complex functions of the parameters. We study a measure of overall contribution of the four treatments to deviations from the average salary. For a new observation $n+1$, with

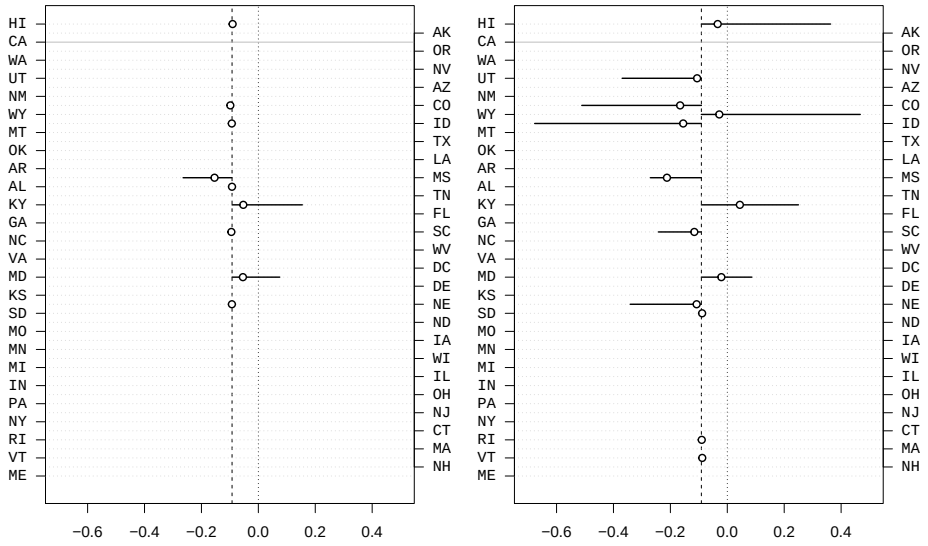

Figure 3.8: State level deviations from main effects for treatment "black" in 2019 for the CPS data. The left panel shows the results for BMA, while the right panel does so for CIL. In both panels, the x-axis shows the magnitude of the coefficients, and the y-axis (two-sided) depicts every state intercalated, by order of appearence in the data. The vertical dashed line represents the main effect point estimate as established by BMA (left) and CIL (right). White dots indicate point estimates, while solid bars represent the corresponding 95% posterior intervals. California, as the reference category, shows no data.

observed treatments d_{n+1} and controls x_{n+1}, let

$$\begin{aligned} &h_{n+1}(d_{n+1},\alpha,x_{n+1}) = \\ &|\mathrm{E}(y_{n+1} \mid d_{n+1},x_{n+1},\alpha,\beta) - \mathrm{E}(y_{n+1} \mid x_{n+1},\alpha,\beta)| = \\ &|[d_{n+1} - \mathrm{E}(d_{n+1} \mid x_{n+1})]^{\intercal}\alpha| \end{aligned} \tag{3.1}$$

be its expected salary minus the expected salary averaged over possible d_{n+1}, given equal control values x_{n+1}. Since y_{n+1} is a log-salary, we examine the posterior predictive distribution of $\exp\{h_{n+1}(d_{n+1},\alpha,x_{n+1})\}$ as a measure of salary variation associated to the treatments. A value of 1 indicates no deviation from the average salary, relative to another individual with the same controls x_{n+1}.

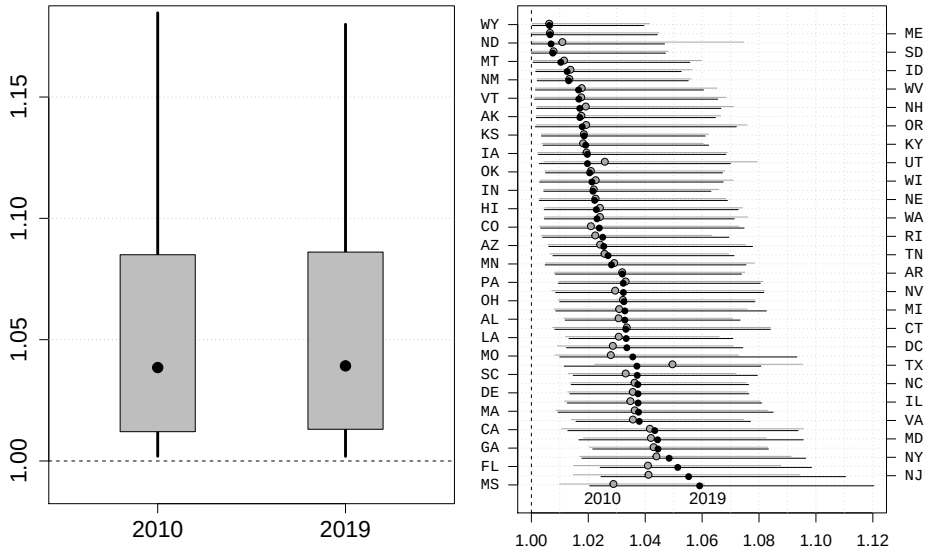

Figure 3.9: The left panel shows the posterior predictive distribution of deviations from average salary as given by $\exp\{h_{n+1}(d_{n+1},\alpha,x_{n+1})\}$ in (3.1), for 2010 and 2019. The gray boxes represent 50% posterior intervals and the black lines are 90% intervals. The black dot is the posterior median. The right panel shows the posterior median of these deviations for every U.S. state in 2010 and 2019 on the horizontal axis, ordered by their value in 2019, with the corresponding 50% posterior intervals for both years.

To evaluate the posterior predictive distribution of (3.1) given y, the observed d and the set of controls, we obtained posterior samples from the model averaged posterior $p(\alpha \mid y)$ associated to CIL (Section 2.3.1). Given that we do not have an explicit model for (d_{n+1},x_{n+1}), we sampled pairs (d_{n+1},x_{n+1}) from their empirical distribution, and estimated $\mathrm{E}(d_{n+1} \mid x_{n+1})$ from a logistic regression of d on the set of controls. Figure 3.9 shows the results. In 2010, joint variation in the treatments was associated to an average 5.8% salary variation (90% predictive interval [0.1%, 18.5%]). The posterior

mean in 2019 was almost identical, with the same average at 5.8% and a 90% predictive interval at [0.1%, 18.0%]. This reinforces the notion that the treatments played a very similar role in the 2019 average relative to 2010, with almost identical levels in terms of inequality, e.g. individuals whose salary was furthest from the average.

It is also of interest to study differences between states. This is possible in our model, which features 200 interaction terms for the 4 treatments and 50 states. Figure 3.9 (right) shows the results. The most salient feature is a larger heterogeneity across states in 2019 relative to 2010. The three states whose median improved the most were Texas (reducing it by 1.3%), Utah (0.6%) and North Dakota (0.4%), while those with a largest increase in the gap were Mississipi (3.0%), New Jersey (1.4%) and Florida (1.1%), ranked with the highest meadian gaps in 2019. This points against any gradual catch-up effect across U.S. states, although the intervals still show some variability within states.

3.3 Factors Involved in Explaining Cholesterol Levels in The Blood

We analyse the data produced by the National Health and Nutrition Examination Survey (NHANES), published by the Centers for Disease Control and Prevention in the U.S., which features a series of measurements on individuals that result from body examination (e.g. body measurements, physical activity), laboratory results (bloodwork, exposure to certain chemicals, etc.), and health behaviour questionnaires (nutrition, smoking activity, etc.). In particular, the dataset used in this section was the same analysed by Wilson et al. (2018) in the context of multiple treatment effects, and was first developed in Patel et al. (2016). The main objective of this exercise is to determine the treatment effect of the exposure to certain *volatile organic compounds* (VCs) on the observed levels of cholesterol in the blood, while accounting for a large number of controls that can affect both cholesterol results and the levels of exposure themselves.

The composition of the dataset is as follows. We have $n = 172$ individual observations, on which we observe three different measurements of cholesterol in the blood (i.e. three different response variables): levels of LDL (Low-density lipoprotein), HDL (High-density lipoprotein) and trygliceride. In terms of VCs, the exposure to ten distinct compounds are contemplated, we refer to Patel et al. (2016) for technical details on this selection, as well as a complete list of measured chemical compounds. Similarly, there are a total of 82 potential control variables that are analysed, all of which could be affecting at least one of the outcomes or treatments. Thus, we build three different models, one for each outcome, featuring $T = 10$ simultaneous treatment variables, and $p = 82$ controls. Since in this situation $n > p$, we are able to compare our CIL to full OLS, to which we add DML and regular BMA for benchmark. Results for the treatment effect estimates on the three models can be found in Figures 3.10 (LDL), 3.11 (HDL) and 3.12

(triglyceride), respectively. In all figures, we portray point estimates for all methods, as well as 95% (posterior) confidence intervals, with the corresponding *p*-values (OLS, DML) and marginal posterior inclusion probabilities (BMA, CIL) at the bottom of the figures.

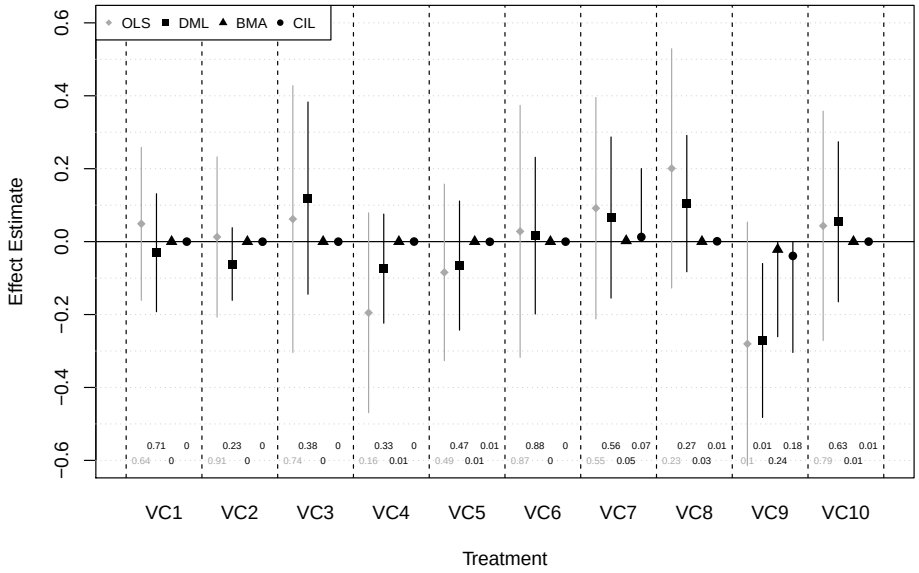

Figure 3.10: Treatment effects of exposure to selected VCs on LDL levels in the blood.

Results in Figure 3.10 point to no significant treatment effect of any of the VCs on LDL levels, as per every method tested, with the exception of VC 9, that we adress next. As expected, we observe shorter intervals for DML with respect to full OLS, whereas both Bayesian methods have zero interval length, as they are able to discard these with overwhelming posterior probability. As mentioned, the only exception is VC 9 (*o*-Xylene, measured in ng/mL), which only DML picks up as significant at the 95% level, with a negative effect on LDL, and a similar point estimate as to that of OLS. Bayesian methods give non-negligible posterior probability to this VC (0.24 for regular BMA, and 0.18 for CIL), but not enough to grant a sizeable point estimate. BMA and CIL perform almost identically on this outcome for all VCs, although upon visual inspection CIL seems to be somewhat more conservative on the inclusion of controls, with moderately lower marginal posterior inclusion probabilities in general.

As for effects on HDL, in Figure 3.11 we observe similar activity to that of LDL. In this case, the effect of VC 9 is positive and, even though *p*-values and MPIPs are not overwhelming, it is picked up by all methods tested. Regular BMA is more confident than CIL with a marginal posterior probability of inclusion above 0.99, compared to a more conservative 0.7 obtained by CIL. However, both methods provide clearly positive point estimates, with a narrower interval length for BMA. Thus, our method shows to

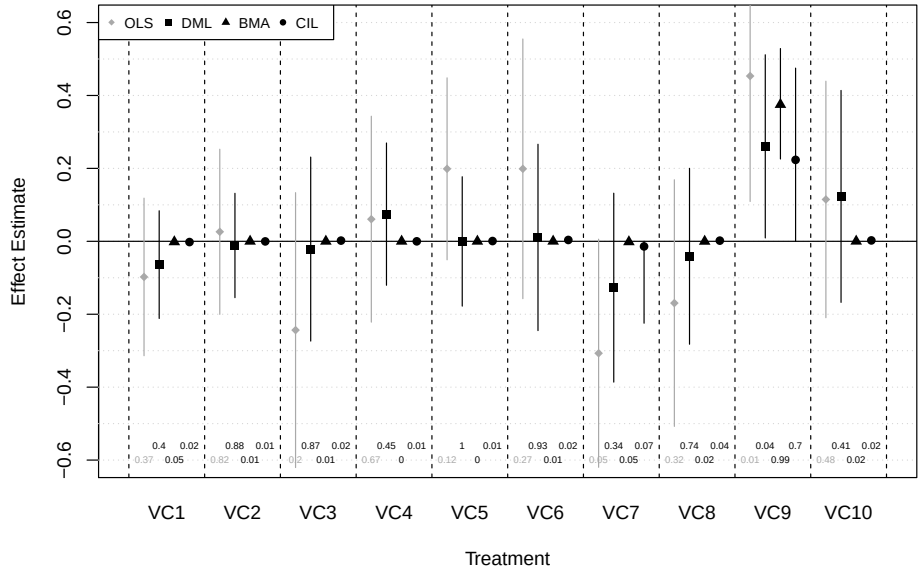

Figure 3.11: Treatment effects of exposure to selected VCs on HDL levels in the blood.

be a bit more conservative in the inclusion of controls also for the LDL model. The rest of VCs show no significant effect on LDL levels.

Finally, in Figure 3.12 we present the results obtained in the model for triglyceride levels. In this model DML detects a couple more signals, those for VC 1 (Tetrachloroethene ng/mL) and VC 4 (Benzene ng/mL), which do not stand out according to the rest of methods. DML results in the three models actually look similar those of BAC analysed in Chapter 9 of Tadesse and Vannucci (2021), where the same data is used, which is coherent with the philosophy of the two methods to strongly emphasise avoiding under-selection. As for the rest of methods, in this model we observe no signal for any VC on any of the other methods tested, with similarly low MPIPs for both Bayesian methods displayed.

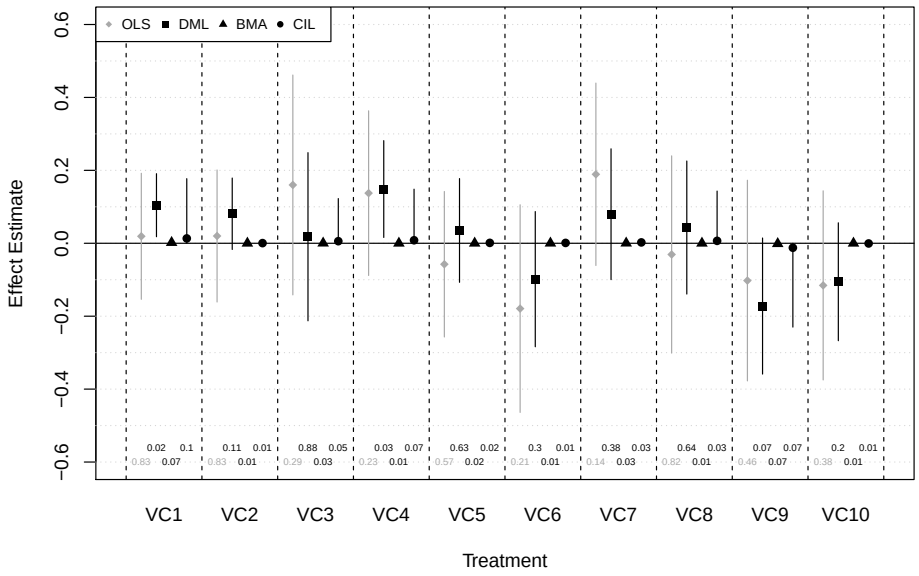

Figure 3.12: Treatment effects of exposure to selected VCs on triglyceride levels in the blood.

Chapter 4

ASYMPTOTIC THEORY FOR NON-LOCAL PRIORS ON THE SEQUENCE MODEL

In previous chapters we developed a Bayesian model selection and averaging framework for treatment effect inference along with efficient computational methods, with an emphasis on high-dimensional settings. The approach used a combination of product MOM (pMOM) priors on the parameters and prior inclusion probabilities that were estimated via empirical Bayes. In this chapter we study the theoretical properties of the pMOM prior under a simplified setting, the so-called sequence model, and compare them to those of a wide family of alternative prior distributions. We provide results on both model selection recovery and parameter estimation accuracy. In particular, the former describe the behaviour of marginal posterior inclusion probabilities for individual parameters, which were an important building block for the Expectation-Propagation algorithm in Chapter 2.

Our main finding is that pMOM priors require less stringent conditions to attain model selection consistency in terms of how the number of non-zero parameters and the total number of parameters are related to the sample size, and that consistency occurs at a faster rate. These results extend those in Johnson and Rossell (2012) and Rossell (2021), who focused on model selection in regression settings, and those in Rossell and Telesca (2017), who considered estimation in orthogonal regression.

4.1 Scope and Contributions

There are several lines of literature highlighting theoretical properties attained by non-local priors in different contexts. Pioneering work of Johnson and Rossell (2010, 2012) focused on model selection properties, showing that asymptotic rates of convergence

of posterior model probabilities improved the imbalance attributed to most Bayesian tests through the study of Bayes factors, including the analysis of the particular case of the linear model also analysed in this work, showing consistency of these posterior quantities. They additionally provide diverse numerical evidence to support that non-local priors can outperform popular penalised likelihood methods as a method to do model selection. Most results, however, are presented for the case were $n \geq p$. In subsequent work, Rossell and Telesca (2017) derive high-dimensional model selection consistency results, and extend their analysis on parameter estimation for finitely many parameters. They show that selection priors can be a valid choice to make estimation in high-dimensions, using NLPs to compute BMA estimates for the linear model, and showing spurious parameter shrinkage at fast polynomial or quasi-exponential rates, depending on the prior density employed, without negatively impacting the estimation truly active parameters.

More abundant general results also exist for BMS and BMA, for example theoretical work by Castillo and van der Vaart (2012); Castillo et al. (2015) shows that a set local priors exist that attain optimal minimax rates of posterior concentration for parameter estimation, introducing the so-called *complexity* priors, which form a particularly set of strong shrinkage priors. In terms of model selection, see Rossell (2021) for a detailed exposition of results currently available on model selection consistency in high-dimensional regression, in which a variety of critical factors are weighed in, namely sample size, signal-to-noise ratio, model dimension and true parameter sparsity. There he exposes that even though there are asymptotically optimal specifications, in many setting less sparsity may still achieve consistency while improving significantly finite-sample performance.

Marginal posterior inclusion probabilities under non-local priors, however, have been less studied. In this work, we try to contribute in understanding how these quantities behave using NLPs, in line with understanding the underlying properties of CIL presented in Chapter 2, which critically makes use of these quantities to assess whether each individual treatment has a non-zero effect on the outcome of interest. Our contribution is two-fold. First, we complement model selection consistency results for NLPs with convergence rates of marginal posterior inclusion probabilities. Second, we analyse estimation properties of BMA estimates under non-local priors in high-dimensions. This has yet not been well studied in this type of regimes, as previous work has mostly focused on finite dimensions. We show that even though Normal tails are suboptimal in the minimax sense (as pointed out by the aforementioned cited work), in certain parameter regions they do not differ decisively to those of popular choices like the double exponential, and that the non-local element of NLPs can contribute to improve on a large class of local priors in this regions. We portray that this occurs for any parameter values that are not too extreme (i.e. too large, or to small but different than zero). Results are presented in Section 4.5, which hold under the corresponding modelling conditions

introduced in Section 4.4.

4.2 Inference For The Sequence Model

The sequence model as parameterised by Wainwright (2019) (Section 7.1.2) assumes that

$$y_i = \sqrt{n}\beta_i^* + \varepsilon_i, \quad \text{with } \varepsilon_i \sim \mathrm{N}(0, \phi_i), \tag{4.1}$$

with independence across $i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ is the observed scalar outcome, β_i^* the unknown parameter of interest, and $\phi_i \in \mathbb{R}^+$ is the known error variance. We denote by $\beta^* = (\beta_1^*, \ldots, \beta_p^*)$ and by p^* the number of non-zero entries in β^*. The maximum likelihood estimator (MLE) of this model is simply

$$\hat{\beta}_i = y_i/\sqrt{n} \sim \mathrm{N}(\beta_i^*, \phi_i/n).$$

Model (4.1) is commonly used to study high-dimensional estimation and model selection problems in a simplified setting where there is independence across parameters. For example, in their pioneering work, Donoho et al. (1992) showed that the optimal minimax estimation rate for β^* under the L_2 loss is $\sqrt{p^* \log(ep/p^*)}$. Johnstone and Silverman (2004) studied the sequence model in a Bayesian setting, exploring estimation rates under a spike-and-slab type of prior with a heavy-tailed slab component, and where the mixing weight was set by empirical Bayes. They showed that using the posterior median as the parameter estimator can achieve optimal estimation rates. Later, Castillo and van der Vaart (2012) found that to attain minimax optimal rates one needs to set priors with thicker than Gaussian tails, and that certain conditions on prior sparsity guarantee a control of false positives. See also Comminges et al. (2021) who provided estimation rates when the error variance is unknown, and also when errors are non-Gaussian.

Model (4.1) can be thought of as a generalization of the Gaussian linear regression model where the covariates are orthogonal, and y_i is a $\sqrt{n}$-scaled version of the least-squares coefficient. Specifically, denote the true linear regression mean by $X\beta^*$ and the true error variance by ϕ, and consider a design matrix such that $X^\intercal X = nI_p$ (e.g. X has zero column means and unit sample variance). Then the sampling distribution of the least-squares estimator is $\hat{\beta} \sim \mathrm{N}(\beta^*, \phi/nI_p)$. Defining $y_i = \sqrt{n}\hat{\beta}_i$, we obtain

$$y_i = \sqrt{n}\hat{\beta}_i \sim \mathrm{N}(\sqrt{n}\beta_i^*, \phi)$$

as in (4.1), for the homoskedastic case where $\phi_i = \phi$ for all $i = 1, \ldots, p$. Without loss of generality one may set $\phi_i = 1$ by dividing both sides of (4.1) by $\sqrt{\phi_i}$, effectively defining a new outcome $y_i/\sqrt{\phi_i} \sim \mathrm{N}(\sqrt{n/\phi_i}\beta_i^*, 1)$. Here we keep the general ϕ_i so that the role of the noise variance in our results can be more easily appreciated.

For concreteness, consider first a standard setting where the β_i^*'s are fixed, i.e. they do not change with n or p. Then (4.1) implies that the expectation of y_i grows at a $\sqrt{n}$ rate. Interestingly, the framework also allows one to consider small signals where β_i^* decreases with n, so that the expectation of y_i grows at a rate slower than $\sqrt{n}$.

Our primary interest in this chapter is two-fold. First, we examine the Bayesian model selection (BMS) recovery rates to identify the p^* non-zero elements in β^* under two prior distributions that belong to the local and non-local prior families (Johnson and Rossell, 2010, 2012). We introduce these below. Second, we study the estimation accuracy of the corresponding Bayesian model averaging (BMA) estimators of β^*.

BMS and BMA allow to tackle both selection and estimation. Let $\gamma_i := I(\beta_i \neq 0)$, for $i = 1, \ldots, p$, be parameter inclusion indicators and $\pi(\beta)$ a given prior density, which for simplicity we assume to factor across i. The prior density on β_i is defined in the Radon-Nikodym sense with respect to the sum of the Lebesgue measures plus a point mass at zero, so that the prior probability $\mathrm{P}_\pi(\beta_i \neq 0) = \mathrm{P}_\pi(\gamma_i = 1) > 0$. We denote the corresponding posterior inclusion probability by

$$\begin{aligned} \mathrm{P}_\pi(\gamma_i = 1 \mid y_i) &= \left(1 + \frac{p_\pi(y_i \mid \gamma_i = 0)}{p_\pi(y_i \mid \gamma_i = 1)} \frac{\mathrm{P}(\gamma_i = 0)}{\mathrm{P}(\gamma_i = 1)}\right)^{-1} \\ &= \left(1 + B^\pi_{i,01} \frac{\mathrm{P}(\gamma_i = 0)}{\mathrm{P}(\gamma_i = 1)}\right)^{-1}, \end{aligned} \tag{4.2}$$

where the equality follows from Bayes theorem and denoting by

$$B^\pi_{i,01} = \frac{p_\pi(y_i \mid \gamma_i = 0)}{p_\pi(y_i \mid \gamma_i = 1)}$$

the so-called Bayes Factor. BMS is typically based on evaluating $\mathrm{P}_\pi(\gamma_i = 1 \mid y_i)$, for which large values of $B^\pi_{i,01}$ favour setting β_i to 0, and vice versa.

As we have seen in previous chapters, BMA uses the point estimate

$$\tilde{\beta}_i := \mathrm{E}_\pi(\beta_i \mid y_i) = \mathrm{E}_\pi(\beta_i \mid \gamma_i = 1, y_i)\mathrm{P}_\pi(\gamma_i = 1 \mid y_i),$$

where the right-hand side follows from the law of total expectation. The BMA estimate involves evaluating two terms: the posterior inclusion probability of the parameter and its expected value conditional on inclusion

$$\mathrm{E}_\pi(\beta_i \mid y_i, \gamma_i = 1) = \int \beta_i p_\pi(\beta_i \mid y_i, \gamma_i = 1) \mathrm{d}\beta_i. \tag{4.3}$$

Computing these posterior quantities requires specifying two priors. First, a model prior $p(\gamma)$ determining prior (marginal) inclusion probabilities, which for simplicity we take to be independent and identicaly distributed across $\gamma_1, \ldots, \gamma_p$. Specifically, we denote by

$$\rho := \mathrm{P}(\gamma_i = 1)$$

the common prior inclusion probability across $i = 1, \ldots, p$. And second, a prior distribution on the included parameters

$$p_\pi(\beta \mid \gamma) = \prod_{i:\gamma_i=1} p_\pi(\beta_i \mid \gamma_i = 1) \tag{4.4}$$

with the convention to set $\beta_j = 0$ with probability 1 for excluded parameters (i.e. with associated $\gamma_i = 0$).

We consider two classes of parameter priors: local and non-local priors. The main principle behind non-local priors (NLP, Johnson and Rossell, 2010, 2012) is model separation. For example, if one contemplates two models that are nested, it can be hard to discern them as to which one of them (if any) represents the data-generating truth. Even though posterior model probabilities will eventually favour the true model, we have no guarantee that the degree of parsimony they provide is favourable enough, especially in high dimensions. NLPs are designed to address this issue by inducing a probabilistic separation between the two models, leading to a stronger data-dependent parsimony compared to more conventional prior specifications, without inducing sparse model probabilities. For that, NLPs esentially entail that the prior density vanishes for any active parameter as it approaches to zero, i.e. that $\lim_{\beta_i \to 0} p_\pi(\beta_i \mid \gamma_i = 1) = 0$, as opposed to tradititional local priors, for which the opposite is true. To illustrate the behaviour of local priors, here we will use the Gaussian prior, in particular

$$\beta_i \mid \phi_i, \tau \sim \mathrm{N}(0, \tau\phi_i), \tag{4.5}$$

where $\tau > 0$ is a user-defined prior dispersion parameter. Many default choices for τ set it to a constant value not depending on n or p, e.g. $\tau = 1$ gives the Unit Information prior (Schwarz, 1978), which is connected to the Bayesian information criterion. However, it is also possible to let τ depend on n, for instance τ growing with n has been shown to favour sparser models. For example, $\tau = \max\{1, p^2/n\}$ was advocated by Foster and George (1994), whereas the shrinking and diffusing priors of Narisetty and He (2014) advocated a certain $\tau > \max\{1, p^2/n\}$.

As for the set of non-local priors, we focus on the product moment MOM prior (pMoM, Johnson and Rossell, 2010). The pMOM represents a choice that only encourages sparsity mildly, other choices exist that embed a stronger sparsity enforcement at the potential cost of reduced power to find strong signals (see Rossell and Telesca, 2017). The pMOM prior is given by

$$\beta_i \mid \phi_i, \tau, \gamma_i = 1 \sim \frac{\beta_i^2}{\tau\phi_i}\mathrm{N}(0, \tau\phi_i). \tag{4.6}$$

The rest of the chapter is structured as follows. Section 4.3 provides expressions for Bayes factors and posterior distributions under the Gaussian and pMOM priors. Section 4.4 discusses technical conditions required by our main results. Section 4.5 describes

the theoretical properties of these two priors for model selection and estimation, and generalize these properties for a much broader class of local priors beyond the Gaussian. Sections 4.6.1 and 4.6.2 contain auxiliary results and the proofs of all theorems, respectively.

For the remaining sections, we also establish the following notation. For two positive sequences a_n and b_n, we denote $a_n \asymp b_n$ if for two constants $0 < c_1 \leq c_2 < +\infty$ it holds that $\lim_{n\to+\infty} a_n/b_n \in [c_1, c_2]$. Further, we denote $a_n \preceq b_n$ if for some constant $c > 0$ it holds that $\lim_{n\to+\infty} a_n/b_n \leq c$. If instead $\lim_{n\to+\infty} a_n/b_n = 0$, we will write $a_n \ll b_n$. Bayesian densities are expressed with $p(\cdot)$, while frequentist densities are represented with $p_{\theta^*}(\cdot)$, under the corresponding data-generating value θ^*, and analogously for Bayesian and frequentist probabilities, expressed with $\mathrm{P}(\cdot)$ and $\mathrm{P}_{\theta^*}(\cdot)$, respectively. Similarly, the stochastic order operators O_P and o_P refer to asymptotic probability statements under P_{θ^*}.

4.3 BMS and BMA expressions under the Gaussian and pMOM priors

For the Gaussian prior in (4.5), standard algebra shows that the Bayes Factor writes

$$B^{\mathrm{N}}_{i,01} = (\tau n + 1)^{1/2} \exp\left\{ -\frac{1}{2} \frac{\tau n}{\tau n + 1} \frac{n}{\phi_i} \hat{\beta}_i^2 \right\}, \tag{4.7}$$

and consequently

$$\mathrm{P}_{\mathrm{N}}(\gamma_i = 1 \mid y_i) = \left(1 + \frac{1-\rho}{\rho} (\tau n + 1)^{1/2} \exp\left\{ -\frac{1}{2} \frac{\tau n}{\tau n + 1} \frac{n}{\phi_i} \hat{\beta}_i^2 \right\} \right)^{-1}. \tag{4.8}$$

Intuitively, large values of $\hat{\beta}_i^2$, i.e. if the MLE is far away from zero, provide large inclusion probabilities, and vice versa. Further, the posterior distribution of each parameter β_i is given by

$$\beta_i \mid y_i, \phi_i, \gamma_i = 1 \sim \mathrm{N}\left(\frac{\tau n}{\tau n + 1} \hat{\beta}_i, \frac{\tau}{\tau n + 1} \phi_i \right) \tag{4.9}$$

independently across $i = 1, \ldots, p$. Hence, the corresponding BMA estimate is

$$\tilde{\beta}_i^{\mathrm{N}} = \hat{\beta}_i \frac{\tau n}{\tau n + 1} \left(1 + \frac{1-\rho}{\rho} (\tau n + 1)^{1/2} \exp\left\{ -\frac{1}{2} \frac{\tau n}{\tau n + 1} \frac{n}{\phi_i} \hat{\beta}_i^2 \right\} \right)^{-1}. \tag{4.10}$$

That is, the BMA estimate is equal to the MLE times two factors. The first is a linear shrinkage factor determined by τn, and which converges to 1 as $n \to \infty$ for all default

τ considered in the literature. The second is a non-linear shrinkage factor such that, as the MLE $\hat{\beta}_j$ approaches 0, the BMA estimate converges to 0 at an exponential rate.

We remark that, although (4.7) provides Bayes factors under the Gaussian prior, Result 4.11 in Section 4.6.1 provides a general representation of the Bayes Factor under any parameter prior where (4.7) is the leading term, under certain conditions. In particular, for the pMOM prior from (4.6) the Bayes factor can be obtained applying Result 4.11, giving

$$B^{\mathrm{M}}_{i,01} = (\tau n+1)\left[1+\frac{\tau n}{\tau n+1}\frac{n}{\phi_i}\hat{\beta}_i^2\right]^{-1} B^{\mathrm{N}}_{i,01}. \tag{4.11}$$

Consequently, the posterior inclusion probability is

$$\mathrm{P_M}(\gamma_i = 1 \mid y_i) = \left(1+\frac{1-\rho}{\rho}(\tau n+1)\left[1+\frac{\tau n}{\tau n+1}\frac{n}{\phi_i}\hat{\beta}_i^2\right]^{-1} B^{\mathrm{N}}_{i,01}\right)^{-1}. \tag{4.12}$$

The posterior distribution of β_i for this non-local prior is expressed by

$$\beta_i \mid y_i, \phi_i, \tau \sim \left[\frac{\beta_i^2}{\hat{\beta}_i^2+\phi_i/n}\right] \mathrm{N}\left(\beta_i; \hat{\beta}_i, \phi_i/n\right). \tag{4.13}$$

Using these expressions, simple derivations give that the BMA estimate under the pMOM prior is

$$\tilde{\beta}^{\mathrm{M}}_i = \hat{\beta}_i\left[\frac{\frac{n}{\phi_i}\hat{\beta}_i^2+3}{\frac{n}{\phi_i}\hat{\beta}_i^2+1}\right]\left(1+\frac{1-\rho}{\rho}(\tau n+1)\left[1+\frac{\tau n}{\tau n+1}\frac{n}{\phi_i}\hat{\beta}_i^2\right]^{-1} B^{\mathrm{N}}_{i,01}\right)^{-1}. \tag{4.14}$$

4.4 Technical conditions

We state and discuss a set of conditions required by the results presented in Section 4.5.

(C1) Lower limit on prior dispersion: $\tau n \to +\infty$ as $n \to +\infty$.

(C2) The marginal prior inclusion probability ρ is non-increasing in n.

(C3) Prior odds for exclusion (Gaussian)

$$\frac{1-\rho}{\rho} \ll (\tau n)^{-1/2} \exp\left\{\frac{1}{2}\frac{n}{\log n}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}.$$

(C4) Prior odds for exclusion (pMOM):

$$\frac{1-\rho}{\rho} \ll (\tau n)^{-3/2} \exp\left\{\frac{1}{2}\frac{n}{\log n}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\left(1+\frac{n}{\log n}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right).$$

(C5) Upper limit on the growth of the proportion of truly active features: $p^*/p \ll 1-1/p$, or re-arraging, that $1/p \ll 1-p^*/p$.

(C6) The true model size p^* satisfies

$$p^* \ll \exp\left\{\frac{n}{2}\left(1-\frac{1}{\sqrt{\log n}}\right)^2 \min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i} - \frac{1}{4}\log\log n\right\}.$$

(C7) Beta-min condition (strong): $\min_{i:\beta_i^*\neq 0}\frac{|\beta_i^*|}{\sqrt{\phi_i}} \gg \frac{1}{\sqrt{n}}\left(1-\frac{1}{\sqrt{\log n}}\right)$.

(C8) Beta-min condition (weak): $\min_{i:\beta_i^*\neq 0}|\beta_i^*| \gg \frac{\log n}{n}$.

(C9) Number of spurious parameters (Gaussian)

$$(p-p^*)\log(p-p^*) \ll \left(\frac{1-\rho}{\rho}\right)(\tau n)^{1/2}.$$

(C10) No. of spurious parameters (pMOM):

$$(p-p^*)\log(p-p^*)\left[\log(p-p^*)+\log\log(p-p^*)\right] \ll \left(\frac{1-\rho}{\rho}\right)(\tau n)^{3/2}.$$

Conditions (C1)–(C4) can be primarily viewed as conditions on the prior. (C1) is a minimal condition that the prior dispersion $\tau \gg 1/n$ is not too small. Recall from Section 4.2 that default choices like $\tau = 1$ or $\tau = \min\{1, p^2/n\}$ satisfy (C1). (C2) is also a mild condition that the prior on the model does not become less sparse as n grows, again this is satisfied by all default prior choices in the literature. (C3) and (C4) are the counterpart (for Gaussian and pMOM priors, respectively) that the prior cannot become too sparse as n grows. Again, both are very mild, e.g. for fixed β^* the prior odds for exclusion can grow almost exponentially in n.

Conditions (C5)–(C10) involve to a larger extent the data-generating truth. (C5) is a minimal condition to ensure that the proportion of truly active parameters is not growing excessively as the dimension of the model increases, in other words, that $p-p^* \gg 1$, so that $p-p^* \to \infty$ as $n \to \infty$. This relates to (C6), which states that the number of truly active parameters cannot be too large. Up to lower-order terms, the condition requires

$$p^* \ll \exp\left\{\frac{n}{2}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\},$$

which is again a mild condition. Condition (C7) is a really minimal beta-min condition on the size of truly active parameters. In fact, Wainwright (2009) showed that a necessary condition for any method to consistently recover the true model is that

$$\min_j \frac{|\beta_j^*|}{\sqrt{\phi_j}} > \sqrt{\frac{\log(p/p^*)}{n}}$$

which is more stringent than (C7) when p/p^* grows with n. (C8) is a potentially even milder condition that suffices for some of the results below.

Finally, (C9) and (C10) provide upper bounds for the amount of spurious parameters under each prior, and reflect the advantage of adopting a pMOM over the Gaussian. Clearly, (C10) required by the pMOM prior is milder than (C9) required by the Gaussian, the main difference being in the right-hand side of the respective expressions. For example, if one were to set both constant prior dispersion τ and prior inclusion probability ρ, then (C9) implies that the total number of parameters must be $p \ll \sqrt{n}$, whereas (C10) implies $p \ll n^{3/2}$, paying a multiplicative log cost. More generally, in high-dimensional settings it is common to either set τ to grow with n (Foster and George, 1994; Fernandez et al., 2001; Narisetty and He, 2014) or ρ to decrease with n (Castillo and van der Vaart, 2012). For example, taking $\tau = \max\{1, p^2/n\}$ as in Foster and George (1994) and constant ρ gives that (C10) is satisfied but (C9) is not.

4.5 Results

Before stating the results we outline the main findings. Our first main result is Theorem 4.1, which portrays model selection properties under the Gaussian and pMOM priors. Parts (i) and (ii) bound the marginal posterior inclusion probabilities for truly inactive parameters. For the Gaussian prior, up to lower-order terms these are at most

$$O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{p\log p}{\sqrt{\tau n}}\right).$$

In contrast, for the pMOM prior we obtain the faster rate

$$O_{\mathrm{P}}\left(\frac{\log p}{\tau n}\frac{\rho}{1-\rho}\frac{p\log p}{\sqrt{\tau n}}\right).$$

That is, the pMOM attains better rates to discard truly zero parameters. Parts (iii) and (iv) bound the posterior inclusion probabilities for truly inactive parameters. In this case, the Gaussian prior roughly attains that they are at most

$$O_{\mathrm{P}}\left(\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{\lambda_{\min}}{2}\frac{n}{\log n}\right\}\right),$$

for $\lambda_{\min} := \min_{i:\beta_i^* \neq 0} \beta_i^{*2}/\phi_i$, whereas the pMOM prior achieves at most

$$O_{\mathrm{P}}\left(\frac{\tau \log n}{\lambda_{\min}} \frac{1-\rho}{\rho} (\tau n)^{1/2} \exp\left\{-\frac{\lambda_{\min}}{2} \frac{n}{\log n}\right\}\right).$$

These rates to pick up the hardest active parameters are quite similar given the exponential term, and moderately better for the non-local prior only if it holds that $\tau \log n \ll \lambda_{\min}$, i.e. if the smallest of the truly active parameters grows at least logarithmically with n, conditional on the specified τ.

Our second result is Theorem 4.2, which gives results for the total absolute error associated to the estimated $\tilde{\beta}$ under the two prior specifications. We essentially examine the probability bounds for the L_∞-norm of the BMA estimate's error $|\tilde{\beta} - \beta^*|$, for the set of truly active and inactive parameters. Our main finding here is that, for the truly inactive parameters, the BMA estimate corresponding to the pMOM prior can achieve a faster rate of convergence to zero relative to that of the Gaussian prior. This extra shrinking factor is roughly of order $\log(p-p^*)/\tau n$. With respect to those parameters that are truly active, both estimators converge at similar speeds to the rates of the MLE. In this case, the Gaussian prior converges to the MLE at a rate of $(\tau n)^{-1}$, and the pMOM at a rate of $\log n/n$, up to a lower order factor.

Our third main result is embodied in Theorem 4.3, which essentially states that for a wide class of local priors, Bayes Factors can be bounded by the regular Gaussian prior Bayes Factor in 4.7 times a prior-dependent constant, plus a vanishing term of order $o_{\mathrm{P}}(1)$.

Corollary 4.4 follows Theorem 4.3, so as to extend the results Theorem 4.1 for the Gaussian prior to a wider class of local priors. The result includes priors with tails both lighter or heavier than those of the Gaussian prior, and shows that the rates for such prior do not differ significantly from those for the Gaussian prior, up to lower-order terms, and under certain mild conditions described in Theorem 4.3. This implies that, for model selection purposes, the pMOM prior holds the advantages shown with respect to the Gaussian prior, but now for a wider class of local priors. Finally, Corollary 4.5 also stems from Theorem 4.3, and similarly extends the results of Theorem 4.2 to local priors beyond the Gaussian prior, again preserving the advantages displayed by the pMOM prior.

Theorem 4.1 (Marginal Posterior Inclusion Probabilities)**.** *Consider the Gaussian and pMOM priors introduced in* (4.5) *and* (4.6)*, respectively. Assume that (C1) and (C2) hold. Then*

i. for the set of truly inactive coefficients, under (C5) and (C9), the Gaussian prior achieves

$$\max_{i:\beta_i^*=0} \mathrm{P}_{\mathrm{N}}(\gamma_i = 1 \mid y_i) = O_{\mathrm{P}}\left(\frac{\rho}{1-\rho} \frac{(p-p^*)}{(\tau n)^{1/2}} \log(p-p^*)\right),$$

ii. while under (C5) and (C10), the pMOM prior achieves

$$\max_{i:\beta_i^*=0} \mathrm{P_M}(\gamma_i = 1 \mid y_i) = O_{\mathrm{P}}\left(\frac{\log(p-p^*)}{\tau n}\frac{\rho}{1-\rho}\frac{(p-p^*)}{(\tau n)^{1/2}}\log(p-p^*)\right).$$

Further, if conditions (C7) and (C6) hold,

iii. for the set of non-zero coefficients, under (C3), the Gaussian prior attains

$$\max_{i:\beta_i^*\neq 0} \mathrm{P_N}(\gamma_i = 0 \mid y_i) = O_{\mathrm{P}}\left(\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\right),$$

iv. while under (C4), the pMOM prior attains

$$\max_{i:\beta_i^*\neq 0} \mathrm{P_M}(\gamma_i = 0 \mid y_i) = O_{\mathrm{P}}\left(\frac{\tau\log n}{\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\right).$$

In Theorem 4.2 we focus on the BMA estimation error under the two examined priors. We provide probability bounds for the L_1-norm of these estimates with respect to the truth. These bounds correspond to the Mean Absolute Error in the simplest model with one parameter, and in high dimensions they are essentially the bounds to the L_∞-norm of the absolute error between $\tilde{\beta}$ and β^*. In parts (i) and (ii) we give these rates for those parameters that are truly inactive. We find that the pMOM prior is able to discard spurious variables at a faster rate of convergence than the Gaussian prior by a factor of order roughly $\log(p-p^*)/\tau n$. In parts (iii) and (iv) we provide the rates for the truly active variables. We show that for both priors the rates converge to those of the MLE, and that said speed towards the MLE is of order $1/n$, for a given pre-specified τ, and where the pMOM pays an extra logarithmic price, that also depends on the size of the smallest parameters.

Theorem 4.2 (L_1 Estimation Error)**.** *Consider the same set of priors as in Theorem 4.2. Assume that (C1) and (C2) hold. Then*

i. for the Gaussian prior, with corresponding estimate $\tilde{\beta}^{\mathrm{N}}$, and if conditions (C5) and (C9) hold,

$$\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i^{\mathrm{N}} - \beta_i^*| = O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{1/2}}\sqrt{\frac{[\log(p-p^*)]^3}{n}}\max_{i:\gamma_i^*=0}\phi_i^{1/2}\right),$$

ii. whereas for the pMOM prior, with estimate $\tilde{\beta}^{\mathrm{M}}$, and if conditions (C5) and (C9) hold,

$$\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i^{\mathrm{M}} - \beta_i^*| = O_{\mathrm{P}}\left(\frac{k_n}{\tau n}\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{1/2}}\sqrt{\frac{[\log(p-p^*)]^3}{n}}\max_{i:\gamma_i^*=0}\phi_i^{1/2}\right),$$

where $k_n = \log(p-p^*) + \log\log(p-p^*)$.

In turn, if conditons (C7) and (C6) hold, then

iii. for the Gaussian prior, under (C3),

$$\sum_{i:\gamma_i^*=1} |\tilde{\beta}_i^{\mathrm{N}} - \beta_i^*| = O_{\mathrm{P}}\left(p^*\left[\left(1+\frac{1}{n\tau}\right)\sqrt{\frac{\log p^*}{\tau n}}\max_{i:\gamma_i^*\neq 0}\phi_i^{1/2} + \frac{1}{n\tau}\max_{i:\gamma_i^*\neq 0}|\beta_i^*|\right]\right);$$

iv. while for the pMOM prior, under (C4) and (C8),

$$\sum_{i:\gamma_i^*=1} |\tilde{\beta}_i^{\mathrm{M}} - \beta_i^*| = O_{\mathrm{P}}\left(p^*\left[\left(1+\frac{\log n}{n}\frac{1}{\min_{i:\gamma_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\right)\sqrt{\frac{\log p^*}{n}}\max_{i:\gamma_i^*\neq 0}\phi_i^{1/2} + \frac{\log n}{n}\frac{\max_{i:\gamma_i^*\neq 0}|\beta_i^*|}{\min_{i:\gamma_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\right]\right).$$

In Theorem 4.3 we show that, for a wide class of local priors, the Bayes Factor $B_{i,01}^{\pi}$ corresponding to the truly active (inactive) parameters is bounded below (above) by that of the Gaussian prior in (4.7) weighted by a constant that depends on the elicited prior, plus a vanishing term of order $o_{\mathrm{P}}(1)$. This result holds for priors that are light- or heavy-tailed, within some mild constraints. This result provides a foundation to extend the results in Theorem 4.1 and Theorem 4.2 from the Gaussian prior to this augmented local prior class, which is addressed in two subsequent corollaries.

Theorem 4.3 (Convergence of Bayes Factors for Other Local Priors)**.** *Let π be a local prior that is bounded, and centreed and symmetric around zero. If the tails of π are lighter than those of the Gaussian prior in* (4.5)*, assume that the density of π is at most $\pi(\beta_i) \propto \exp\left\{-\frac{1}{2}\frac{\beta_i^{2k}}{\tau\phi_i}\right\}$, for any fixed integer $k > 1$. If the opposite holds, then assume that the density of π is at most Cauchy, with scale parameter $\nu \propto \sqrt{\tau\pi_i}$. Then, under the corresponding conditions in Theorem 4.1 for the Gaussian prior*

i. for $\beta_i^* = 0$, the Bayes Factor $B_{i,01}^{\pi}$ satisfies

$$B_{i,01}^{\pi} \leq c_{\pi} B_{i,01}^{\mathrm{N}} + o_{\mathrm{P}}(1),$$

ii. while for $\beta_i^* \neq 0$, it holds that

$$B_{i,01}^{\pi} \geq \tilde{c}_{\pi} B_{i,01}^{\mathrm{N}} + o_{\mathrm{P}}(1),$$

for some fixed pair of prior-dependent constants $0 < c_{\pi}, \tilde{c}_{\pi} < \infty$.

The result in Theorem 4.3 implies that, for any prior in this class, the Bayes Factor essentially does not improve the probability bounds of the Gaussian prior, and thus by extension neither will it improve the posterior inclusion probabilities featured in Theorem 4.1, in parts (i) and (iii). This is shown in Corollary 4.4. A direct consequence of this is that the advantages shown by the pMOM prior in terms of convergence rates for the inclusion probabilities, for active and inactive parameters, are maintained with respect to the entire class of local priors featured in Theorem 4.3.

Corollary 4.4. *For any prior π satisfying the requirements in Theorem 4.3, and under the appropriate conditions established by Theorem 4.1 relative to the Gaussian prior, it holds that the probability bounds for $\max_{i:\beta_i^*=0} \mathrm{P}_{\pi}(\gamma_i = 1 \mid y_i)$ and for $\max_{i:\beta_i^*\neq 0} \mathrm{P}_{\pi}(\gamma_i = 0 \mid y_i)$ do not improve those for $\max_{i:\beta_i^*=0} \mathrm{P}_{\mathrm{N}}(\gamma_i = 1 \mid y_i)$ and $\max_{i:\beta_i^*\neq 0} \mathrm{P}_{\mathrm{N}}(\gamma_i = 0 \mid y_i)$, respectively, presented in Theorem 4.1.*

Similarly, in Corollary 4.5 we extend the results from Theorem 4.2 beyond the Gaussian prior. The basic result is that if the corresponding BMA estimate $\tilde{\beta}_i^{\pi}$ has a shrinking factor that converges to 1, then the Gaussian rates of convergence of the L_1-error apply to $\tilde{\beta}_i^{\pi}$, with a minor adjustment that will only depend on the rate of convergence of said shrinkage factor.

Corollary 4.5. *For any prior π satisfying the requirements in Theorem 4.3, with an associated shrinking factor $f_{\pi} \to 1$ such that for every individual parameter β_i*

$$|f_{\pi} - 1| = O_{\mathrm{P}}(s_n),$$

for some well-defined positive sequence $s_n \to 0$, and under the appropriate conditions established by Theorem 4.2 relative to the Gaussian prior, it holds that the probability bound for $\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i^{\pi} - \beta_i^*|$ does not improve the corresponding bound for $\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i^{\mathrm{N}} - \beta_i^*|$, introduced in Theorem 4.2. Similarly, for the set of truly active parameters probability bounds under prior π cannot improve

$$\sum_{i:\gamma_i^*=1} |\tilde{\beta}_i^{\pi} - \beta_i^*| = O_{\mathrm{P}}\left(p^* \left[(1+s_n) \sqrt{\frac{\log p^*}{\tau n}} \max_{i:\gamma_i^*\neq 0} \phi_i^{1/2} + s_n \max_{i:\gamma_i^*\neq 0} |\beta_i^*| \right] \right).$$

Therefore, corollaries 4.4 and 4.5 show that in essence any local prior within the set established in Theorem 4.3 will converge to the standard Gaussian prior as n grows in terms of convergence rates, up to lower-order terms. Consequently, the advantages shown by the pMOM non-local prior in terms of rates for posterior inclusion probabilities and L_1-error are preserved with relation to this prior set.

4.6 Technical Appendix

4.6.1 Auxiliary Results

Result 4.6. *Let $Z \geq 0$ be a random variable such that $\mathrm{E}(Z) \leq a_n$, for some sequence $a_n > 0$. Then $Z = O_{\mathrm{P}}(a_n)$. Further, for any $f_n \asymp a_n$, $Z = O_{\mathrm{P}}(f_n)$.*

Proof. To prove the first statement, note that since $\mathrm{E}(Z) \leq a_n$, for any fixed $\varepsilon > 0$

$$P(Z > a_n/\varepsilon) \leq \frac{a_n}{a_n/\varepsilon} = \varepsilon,$$

by Markov's inequality. Hence, for every $\varepsilon > 0$, some $\delta > 0$ and n_0 exist such that $P(Z/a_n > \delta) \leq \varepsilon$ for every $n \geq n_0$, in particular $\delta = 1/\varepsilon$ and $n_0 = 1$. Thus, $Z = O_{\mathrm{P}}(a_n)$. For the second statement, $f_n \asymp a_n$ implies that some $\lambda > 0$ exists such that $f_n \leq \lambda a_n$ for sufficiently large n. Hence, again using Markov's inequality, for every $\varepsilon > 0$

$$P(Z > f_n/\varepsilon) \leq \frac{f_n}{a_n/\varepsilon} \overset{n\to\infty}{\leq} \frac{\lambda a_n}{a_n/\varepsilon} = \varepsilon\lambda,$$

and so, for every $\varepsilon > 0$, some $\delta > 0$ and some sufficiently large n exist such that $P(Z/f_n > \delta) \leq \varepsilon\lambda$, in particular $\delta = 1/\varepsilon$. Hence $Z = O_{\mathrm{P}}(f_n)$. □

Result 4.7. *Let $X_1, \ldots, X_p$ be a set of i.i.d. random variables where $X_i \sim N(m_i, v_i)$. Then,*

$$\max_{i\in\{1,\ldots,p\}} |X_i| = O_{\mathrm{P}}\left(\max_i |m_i| + \sqrt{\log(p) \max_{i\in\{1,\ldots,n\}} v_i}\right).$$

Proof. The proof strategy is to first show that

$$\max_i X_i = O_{\mathrm{P}}\left(\sqrt{\log(p) \max_i v_i}\right),$$

which by symmetry also implies that $-\min_i X_i = O_{\mathrm{P}}\left(\sqrt{\log(p)\max_i v_i}\right)$, and finally show that these two results imply that

$$\max_i |X_i| = O_{\mathrm{P}}\left(\sqrt{\log(p) \max_i v_i}\right).$$

To prove that $\max_i X_i = O_{\mathrm{P}}\left(\sqrt{\log(p)\max_i v_i}\right)$ we show that $\mathrm{E}[\max_i X_i] \leq \sqrt{2\log(p)\max_i v_i}$. and apply Markov's inequality. Consider an arbitrary $s > 0$. Using Jensen's inequality

$$\begin{aligned}
\exp\left\{s\mathrm{E}\left[\max_i X_i\right]\right\} &\leq \mathrm{E}\left[\exp\left\{s\max_i X_i\right\}\right] = \mathrm{E}\left[\max_i e^{sX_i}\right] \\
&= \int_0^\infty P\left(\max_i e^{sX_i} > t\right) \mathrm{d}t \leq \int_0^\infty \sum_i P\left(e^{sX_i} > t\right) \mathrm{d}t \\
&= \sum_i \mathrm{E}\left[e^{sX_i}\right].
\end{aligned}$$

Applying logs on both sides and re-arranging, we have that for any $s > 0$

$$\mathrm{E}\left[\max_i X_i\right] \leq \frac{1}{s}\log\left(\sum_i \mathrm{E}\left[e^{sX_i}\right]\right). \tag{4.15}$$

Since $X_i \sim \mathrm{N}(0, v_i)$, the corresponding MGF is given by $\mathrm{E}\left[e^{sX_i}\right] = e^{sm_i + v_i s^2/2}$, and (4.15) becomes

$$\mathrm{E}\left[\max_i X_i\right] \leq \frac{1}{s}\log\left(pe^{s\max_i m_i + \max_i v_i s^2/2}\right) = \max_i m_i + \frac{1}{s}\log p + \frac{s}{2}\max_i v_i.$$

Minimizing this RHS over $s > 0$ provides the optimum at $s = \sqrt{2\log p / \max_i v_i}$ and

$$\mathrm{E}[\max_i X_i] \leq \max_i m_i + \sqrt{2\log(p)\max_i v_i}. \tag{4.16}$$

This immediately implies that $\max_i X_i = O_{\mathrm{P}}(\max_i m_i + \sqrt{\log(p)\max_i v_i})$. To see why, by Markov's inequality, for any fixed $\varepsilon > 0$ we have

$$p\left(\max_i X_i > \frac{1}{\varepsilon}\left[\max_i m_i + \sqrt{2\log(p)\max_i v_i}\right]\right) \leq \varepsilon,$$

which by definition gives that $\max_i X_i = O_{\mathrm{P}}(\max_i |m_i| + \sqrt{\log(p)\max_i v_i})$.

By symmetry of the $\mathrm{N}(0, v_j)$ distribution, (4.16) also gives

$$\mathrm{E}\left(\min_i X_i\right) = -\mathrm{E}\left(\max_i X_i\right) \geq -\left[\max_i m_i + \sqrt{2\log p \max_i v_i}\right],$$

and so

$$-\min_i X_i = O_{\mathrm{P}}\left(\max_i m_i + \sqrt{2\log p \max_i v_i}\right).$$

Lastly, let $a_p = \max_i m_i + \sqrt{2\log(p)\max_i v_i}$, then

$$\begin{aligned} P\left(\max_i |X_i| > \frac{a_p}{\varepsilon/2}\right) &= P\left(\left\{\min_i X_i < -\frac{a_p}{\varepsilon/2}\right\} \bigcup \left\{\max_i X_i > \frac{a_p}{\varepsilon/2}\right\}\right) \\ &\leq P\left(\min_i X_i < -\frac{a_p}{\varepsilon/2}\right) + P\left(\max_i X_i > \frac{a_p}{\varepsilon/2}\right) \leq \varepsilon, \end{aligned}$$

for any $\varepsilon > 0$, proving that $\max_i |X_i| = O_{\mathrm{P}}\left(\max_i m_i + \sqrt{2\log p \max_i v_i}\right)$. □

Result 4.8. *Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\intercal}$ denote the MLE of β^*. Then*

$$\max_{i:\beta_i^*=0} |\hat{\beta}_i| = O_{\mathrm{P}}\left(\sqrt{\frac{\log(p-p^*)}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right).$$

Proof. Note that $\hat{\beta}_i = y_i/\sqrt{n}$, and so that for $\beta_i^* = 0$, $\hat{\beta}_i \sim \mathrm{N}(0, \phi_i/n)$. Then apply Result 4.7 to the set $\{i : \beta_i^* = 0\}$, containing $(p - p^*)$ elements. □

Result 4.9. *Let $X_1, \ldots, X_p$ be a set of i.i.d. random variables such that $X_i \sim \chi^2_1(\lambda_i)$, $\forall i \in \{1, \ldots, p\}$, where $\lambda_i \geq 0$ denotes the non-centrality parameter, and let $\lambda_{\min} := \min_{i \in \{1,\ldots,p\}} \lambda_i$. Assume that $\lambda_{\min} \to +\infty$ as $p \to +\infty$. Then, for any sequence $0 < a_p \ll \lambda_{\min}$,*

$$p\left(\max_{i\in\{1,\ldots,p\}} \frac{1}{X_i} > a_p^{-1}\right) \leq 1 - \exp\left\{-pe^{-\frac{\lambda_{\min}}{2}\left[1-\left(\frac{a_p}{\lambda_{\min}}\right)^{1/2}\right]^2}\left(\frac{a_p}{\lambda_{\min}}\right)^{1/4}\right\}. \quad (4.17)$$

Further, if $p \ll \exp\left\{\frac{\lambda_{\min}}{2}\left(1 - [\log(1+\lambda_{\min})]^{-1/2}\right)^2\right\} + [\log(1+\lambda_{\min})]^{1/4}$ then

$$\max_{i\in\{1,\ldots,p\}} \frac{1}{X_i} = o_{\mathrm{P}}\left(\frac{\log(1+\lambda_{\min})}{\lambda_{\min}}\right). \quad (4.18)$$

Proof. To prove (4.17), one needs to find an upper bound to $p(\max_i 1/X_i > a_p^{-1}) = p(1/\min_i X_i > a_p^{-1}) = p(\min_i X_i < a_p)$. Assuming independence we have that for any sequence $a_p > 0$

$$p(\min_i X_i < a_p) = 1 - p(\min_i X_i > a_p) = 1 - \prod_{i=1}^{p} p(X_i > a_p) \leq 1 - p(X_k > a_p)^p,$$

where $k := \arg\min_i \lambda_i$. It then suffices to find a lower bound to $p(X_k > a_p)^p$. In the next inequality we apply Lemma S2 from Rossell (2021), and write that for any $a_p < \lambda_k$ we have that

$$p(X_k > a_p)^p = [1 - p(X_k < a_p)]^p \geq \left[1 - \exp\left\{-\frac{\lambda_k}{2}\left(1 - \sqrt{\frac{a_p}{\lambda_k}}\right)^2\right\}\left(\frac{a_p}{\lambda_k}\right)^{1/4}\right]^p.$$

Now consider such $a_p \ll \lambda_k$, for which

$$\begin{aligned} p(\min_i X_i < a_p) &\leq 1 - \left[1 - e^{-\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2}\left(\frac{a_p}{\lambda_k}\right)^{1/4}\right]^p \\ &= 1 - \exp\left\{p\log\left(1 - e^{-\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2}\left(\frac{a_p}{\lambda_k}\right)^{1/4}\right)\right\} \\ &\leq 1 - \exp\left\{p\left[1 - e^{\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2}\left(\frac{\lambda_k}{a_p}\right)^{1/4}\right]^{-1}\right\}, \end{aligned}$$

where in the last inequality we used that for $x \in (0,1)$, $\log(1-x) \geq (1-1/x)^{-1}$. Note that as $\lambda_k \to +\infty$ we have that

$$\left[1 - e^{\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2}\left(\frac{\lambda_k}{a_p}\right)^{1/4}\right]^{-1} \overset{\lambda_k \to +\infty}{\to} -\exp\left\{-\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2\right\}\left(\frac{a_p}{\lambda_k}\right)^{1/4},$$

and so for sufficiently large dimension we have that

$$p(\max_i 1/X_i > a_p^{-1}) = p(\min_i X_i < a_p) \leq 1 - \exp\left\{-pe^{-\frac{\lambda_k}{2}\left(1-\sqrt{\frac{a_p}{\lambda_k}}\right)^2}\left(\frac{a_p}{\lambda_k}\right)^{1/4}\right\},$$

which proves (4.17). To prove (4.18), choose $a_p = \lambda_k/\log(1+\lambda_k)$, and so the previous bound becomes

$$p(\max_i 1/X_i > a_p^{-1}) \leq 1 - \exp\left\{-pe^{-\frac{\lambda_k}{2}\left(1-\sqrt{\frac{1}{\log(1+\lambda_k)}}\right)^2}\left(\frac{1}{\log(1+\lambda_k)}\right)^{1/4}\right\},$$

Note that the RHS of this bound can be made arbitrarily small as $p \to +\infty$ as long as

$$\frac{\lambda_k}{2}\left(1-\sqrt{\frac{1}{\log(1+\lambda_k)}}\right)^2 + \frac{1}{4}\log\log(1+\lambda_k) - \log p \to +\infty,$$

for which (assuming $\lambda_k \to +\infty$) it suffices that $p < e^{\frac{\lambda_k}{2}\left(1-\sqrt{\frac{1}{\log(1+\lambda_k)}}\right)^2} + [\log(1+\lambda_k)]^{1/4}$. Thus, if such condition holds then a sufficiently large p exists such that

$$p\left(\max_i \frac{1}{X_i} > \delta a_p^{-1}\right) \leq 1 - \varepsilon,$$

for every fixed $\varepsilon > 0$ and $\delta > 0$, where $a_p^{-1} = \log(1+\lambda_k)/\lambda_k$. Hence, $\max_i 1/X_i = o_{\mathrm{P}}(\log(1+\lambda_k)/\lambda_k)$, proving (4.18). □

Result 4.10. *Let $X > 0$ be a random variable such that $X^{-1} = o_{\mathrm{P}}(f_n)$, for some well-defined sequence $f_n > 0$. Then*

$$\frac{1}{X+1} = O_{\mathrm{P}}\left(\frac{1}{f_n^{-1}+1}\right).$$

Proof. Since $X^{-1} = o_{\mathrm{P}}(f_n)$, we know for every $\varepsilon, \delta > 0$ and $n > n_0$, $P(X^{-1} > \delta f_n) < \varepsilon$ for some sufficiently large n_0. Note then that

$$\begin{aligned} P_{\theta^*}(X^{-1} > \delta f_n) = P_{\theta^*}(X < (\delta f_n)^{-1}) &= P_{\theta^*}\left(\frac{1}{X+1} < \frac{1}{(\delta f_n)^{-1}+1}\right) \\ &= P_{\theta^*}\left(\frac{1}{X+1} < \delta\frac{1}{f_n^{-1}+\delta}\right) < \varepsilon. \end{aligned}$$

Since this holds for every $\delta > 0$, the following is also true

$$P_{\theta^*}\left(\frac{1}{X+1} < \delta \frac{1}{f_n^{-1}+1}\right) \leq \varepsilon,$$

for any $\varepsilon > 0$ and large enough n_0, and so by definition $\frac{1}{X+1} = O_{\mathrm{P}}\left(\frac{1}{f_n^{-1}+1}\right)$. □

Result 4.11. *Let the Bayes Factor under the Gaussian prior $B^{\mathrm{N}}_{i,01}$ be defined as in* (4.7)*. Then, for any parameter prior π*

$$B^{\pi}_{i,01} = B^{\mathrm{N}}_{i,01} \times \left[\int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} p_{\mathrm{N}}(\beta_i \mid y_i) d\beta_i\right]^{-1}.$$

Proof. Note that

$$\begin{aligned} p_\pi(y_i \mid \beta_i \neq 0) &= \int_{\beta_i \neq 0} p(y_i \mid \beta_i)\pi(\beta_i)\mathrm{d}\beta_i \\ &= p_{\mathrm{N}}(y_i \mid \beta_i \neq 0) \int \frac{p(y_i \mid \beta_i)\mathrm{N}(\beta_i; 0, \tau\phi_i)}{p_{\mathrm{N}}(y_i \mid \beta_i \neq 0)} \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} \mathrm{d}\beta_i \\ &= p_{\mathrm{N}}(y_i \mid \beta_i \neq 0) \int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} p_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \end{aligned}$$

Then, replacing this expression into the Bayes Factor, we obtain

$$\begin{aligned} B^{\pi}_{i,01} &= \frac{p(y_i \mid \beta_i = 0)}{p_\pi(y_i \mid \beta_i \neq 0)} \\ &= \frac{p(y_i \mid \beta_i = 0)}{p_{\mathrm{N}}(y_i \mid \beta_i \neq 0)} \left[\int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} p_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i\right]^{-1} \\ &= B^{\mathrm{N}}_{i,01} \left[\int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} p_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i\right]^{-1}. \end{aligned}$$

□

Result 4.12. *Let $k_n, g_n > 0$ be sequences such that $k_n \to 0$ and $k_n g_n \to 0$ as $n \to \infty$. Then, for any event A_n such that*

$$\mathrm{P}_{\theta^*}(A_n) \leq 1 - (1 - k_n)^{g_n},$$

it holds that

$$\mathrm{P}_{\theta^*}(A_n) \leq k_n g_n (1 + \omega_n),$$

for some $\omega_n \to 0$.

Proof. Start by re-expressing

$$\mathrm{P}_{\theta^*}(A_n) \leq 1-(1-k_n)^{g_n} = 1-\left[\left(1-\frac{1}{k_n^{-1}}\right)^{k_n^{-1}}\right]^{k_n g_n}.$$

Given that $k_n \to 0$ (i.e. $k_n^{-1} \to \infty$), we know that $\lim_{n\to\infty}(1-1/k_n^{-1})^{k_n^{-1}} = e^{-1}$, and so we have

$$1-\left[\left(1-\frac{1}{k_n^{-1}}\right)^{k_n^{-1}}\right]^{k_n g_n} = 1-\left(\left[\frac{\left(1-\frac{1}{k_n^{-1}}\right)^{k_n^{-1}}}{e^{-1}}\right]e^{-1}\right)^{k_n g_n} \to 1-e^{-k_n g_n}.$$

This last convergence comes from the fact that

$$\lim_{n\to\infty}\left[\frac{\left(1-\frac{1}{k_n^{-1}}\right)^{k_n^{-1}}}{e^{-1}}\right]^{k_n g_n} = \lim_{n\to\infty}\exp\left\{k_n g_n \log\left(\frac{\left(1-\frac{1}{k_n^{-1}}\right)^{k_n^{-1}}}{e^{-1}}\right)\right\} = 1,$$

which in turn comes from

$$\lim_{n\to\infty} k_n g_n \log\left(\frac{\left[1-\frac{1}{k_n^{-1}}\right]^{k_n^{-1}}}{e^{-1}}\right) = \lim_{n\to\infty} k_n g_n\left[\log\left(\left[1-\frac{1}{k_n^{-1}}\right]^{k_n^{-1}}\right)+1\right] = 0,$$

using the fact that $\lim_{n\to\infty}\log\left(\left[1-\frac{1}{k_n^{-1}}\right]^{k_n^{-1}}\right) = -1$ and that we assumed $\lim_{n\to\infty} k_n g_n = 0$. Note then that using $\lim_{x\to 0}\frac{e^{ax}-1}{bx} = a/b$ for $a,b \in \mathbb{R}$ we have that

$$\lim_{x\to 0}\frac{1-e^{-x}}{x} = -\lim_{x\to 0}\frac{e^{-x}-1}{x} = -\frac{(-1)}{1} = 1,$$

and hence

$$1-e^{-k_n g_n} \to k_n g_n.$$

Therefore, we have that under the above conditions

$$\mathrm{P}_{\theta^*}(A_n) \leq k_n g_n(1+\omega_n),$$

where $\omega_n \to 0$. □

4.6.2 Proofs

Proof of Theorem 4.1

Proof. Theorem 4.1; Part (i). The proof strategy for parts (i) and (ii) is to notice that the posterior inclusion probabilities $\mathrm{P_N}(\gamma_j = 1 \mid y_i)$ and $\mathrm{P_M}(\gamma_j = 1 \mid y_i)$ are monotone functions of a random variable that is χ^2_1-distributed for every $\beta^*_i = 0$. Hence, to bound the maximum posterior inclusion probability it suffices to bound in probability the maximum of χ^2_1-distributed variables.

For the Gaussian and pMOM prior, $\mathrm{P}(\gamma_i = 1 \mid y_i)$ is a monotonically increasing function of $W_i := n\hat{\beta}^2_i/\phi_i$, denoted by $h_1(W_i)$. Thus,

$$\max_{i:\beta^*_i=0} \mathrm{P_N}(\gamma_i = 1 \mid y_i) = h_1(Z),$$

where $Z := \max_{i:\beta^*_i=0} W_i$ for short, and $W_i \sim \chi^2_1(0)$ independently for every $i : \beta^*_i = 0$. In particular, from (4.8) we have that $h_1(Z) = (1 + c_n e^{-b_n Z})^{-1}$, where for brevity we will denote by $b_n = \frac{1}{2}\frac{\tau n}{\tau n+1} \geq 0$ and by $c_n = \frac{1-\rho}{\rho}(\tau n + 1)^{1/2} \geq 0$.

Our goal is to show that a sequence $f_n > 0$ exists, such that $\forall \varepsilon > 0$, $\exists \delta > 0$ and $n_0 \in \mathbb{Z}^+$ for which $\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) \leq \varepsilon$, for every $n \geq n_0$. By definition, this would imply that $h_1(Z) = O_{\mathrm{P}}(f_n)$.

We start by re-expressing

$$\begin{aligned}
&\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) = \mathrm{P}_{\beta^*}\left(Z > h_1^{-1}(\delta f_n)\right) = \mathrm{P}_{\beta^*}\left(\max_i W_i > h_1^{-1}(\delta f_n)\right) \\
&= 1 - \mathrm{P}_{\beta^*}\left(\max_i W_i \leq h_1^{-1}(\delta f_n)\right) = 1 - \mathrm{P}_{\beta^*}\left(\cap_i \left\{W_i \leq h_1^{-1}(\delta f_n)\right\}\right) \\
&= 1 - \left[\mathrm{P}_{\beta^*}\left(W_i \leq h_1^{-1}(\delta f_n)\right)\right]^{p-p^*} = 1 - \left[1 - \mathrm{P}_{\beta^*}\left(W_i > h_1^{-1}(\delta f_n)\right)\right]^{p-p^*}. \qquad (4.19)
\end{aligned}$$

where the third line follows from the W_i's being i.i.d. In the first equality we used the monotonicity of h_1 to preserve the direction of the inequality. Thus, to bound $\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n)$, it suffices to upper bound the central χ^2_1 right-tail probability $\mathrm{P}_{\beta^*}\left(W_i > h_1^{-1}(\delta f_n)\right)$. From Lemma S1 in Rossell (2021) we use the Chernoff bound as follows. For any $h_1^{-1}(\delta f_n) > 1$, it holds that

$$\begin{aligned}
\mathrm{P}_{\beta^*}\left(W_i > h_1^{-1}(\delta f_n)\right) &\leq \left(e h_1^{-1}(\delta f_n)\right)^{1/2} \exp\left\{-\frac{1}{2} h_1^{-1}(\delta f_n)\right\} \\
&= e^{1/2}\left[-\frac{1}{b_n}\log\left(\frac{\frac{1}{\delta f_n} - 1}{c_n}\right)\right]^{1/2}\left(\frac{\frac{1}{\delta f_n} - 1}{c_n}\right)^{\frac{1}{2b_n}}, \qquad (4.20)
\end{aligned}$$

where in (4.20) we simply replaced $h_1^{-1}(\delta f_n) = -\frac{1}{b_n}\log\left(\frac{1}{c_n}\left[\frac{1}{\delta f_n} - 1\right]\right)$ and re-arranged. Now we look at the condtions that f_n needs to satisfy for this bound to converge to zero.

Note that for (4.20) to vanish as $n \to \infty$, it suffices that $h_1^{-1}(\delta f_n) \gg 1$, i.e. that

$$\delta f_n \gg (1 + c_n e^{-b_n})^{-1} \asymp (1 + c_n)^{-1} \asymp c_n^{-1} = \frac{\rho}{1-\rho}(\tau n + 1)^{-1/2}, \tag{4.21}$$

since $b_n \to 1/2$, and where $c_n \to \infty$ under (C1) and (C2). For brevity, denote the sequence in the RHS of (4.20) by s_n, and assume that the choice of f_n satisfies (4.21), and thus it holds that $s_n \to 0$ (i.e. $s_n^{-1} \to \infty$) as $n \to \infty$. Then, if $s_n(p - p^*) \to 0$ with $n \to \infty$, we can apply Result 4.12, which gives that

$$\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) \leq s_n(p - p^*)(1 + \omega_n), \tag{4.22}$$

for some $\omega_n \to 0$. Since s_n depends of the choice of f_n, we will need to check later that in fact $s_n(p - p^*) \to 0$ for our choice of f_n. Importantly, this condition would automatically imply that the bound converges to zero.

Now, choose

$$\delta f_n = \left(1 + c_n e^{-b_n a_n}\right)^{-1}, \tag{4.23}$$

for some $a_n \gg 1$, which then satisfies (4.21). In particular, set

$$a_n = 2\left[\log(p - p^*) + \log\log(p - p^*)\right]. \tag{4.24}$$

This implies that the bound in (4.22) can be expressed as

$$\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) \leq \tag{4.25}$$

$$\leq (1 + \omega_n)(p - p^*)\left[-\frac{e}{b_n}(1 + \omega_n)\left[\log\left(e^{-b_n a_n}\right)\right]\left(e^{-b_n a_n}\right)^{1/b_n}\right]^{1/2} \tag{4.26}$$

$$= (1 + \omega_n)(p - p^*)\left[e a_n\right]^{1/2} e^{-a_n/2}$$

$$= (1 + \omega_n)(p - p^*)\left[2e(\log(p - p^*) + \log\log(p - p^*))\right]^{1/2}\left[\frac{1}{(p - p^*)\log(p - p^*)}\right]$$

$$= (1 + \omega_n)\sqrt{2e}\left[\frac{\log(p - p^*)}{[\log(p - p^*)]^2} + \frac{\log\log(p - p^*)}{[\log(p - p^*)]^2}\right]^{1/2}$$

$$= (1 + \omega_n)\sqrt{2e}\left[\frac{1}{\log(p - p^*)} + \frac{\log\log(p - p^*)}{[\log(p - p^*)]^2}\right]^{1/2}. \tag{4.27}$$

Hence, for the respective choices of δf_n and a_n in (4.23) and (4.24), it suffices that $p - p^* \to \infty$ to achieve $s_n(p - p^*) \to 0$, which is displayed in (4.26) and (4.27), and so in that case Result 4.12 is applicable as in (4.22). Thus, under condition (C5) the bound in (4.27) can be made arbitrarily small for a sufficiently large $p - p^*$, i.e. for sufficiently large n. That is, for every $\varepsilon > 0$, one can find a sufficiently large integer

n_0 and corresponding $\delta > 0$ such that $\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) \leq \varepsilon$, for every $n_0 > n$. Hence, $h_1(Z) = O_{\mathrm{P}}(f_n)$, for f_n defined as in (4.23) and some fixed δ.

Additionally, if $c_n e^{-b_n a_n} \to \infty$ as $n \to \infty$, then $h_1(Z) = O_{\mathrm{P}}(o(1)) = o_{\mathrm{P}}(1)$, which by definition would entail $h_1(Z) \xrightarrow{P} 0$. We now show that $c_n e^{-b_n a_n} \to 0$ under a set of conditions, i.e. we need to show that $-b_n a_n + \log c_n \to +\infty$, or since $b_n \to 1/2$ and $\tau n + 1 \to \tau n$,

$$-\log(p-p^*) - \log\log(p-p^*) + \frac{1}{2}\log(\tau n) + \log\left(\frac{1-\rho}{\rho}\right) \to +\infty$$

as $n \to \infty$. This condition holds assuming (C1), (C2) and (C9). Then, from the expression in (4.23) we have that $\lim_{n\to\infty} \frac{f_n}{c_n^{-1} e^{a_n b_n}} = 1$, and so

$$h_1(Z) = O_{\mathrm{P}}\left(c_n^{-1} e^{a_n b_n}\right) = O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)}{(\tau n)^{1/2}}\log(p-p^*)\right),$$

which completes the proof. □

Proof. Theorem 4.1; Part (ii). We adapt and apply the same proof strategy as for Part (i). Consider the pMoM prior in (4.6), for which $h_1(Z) = \left(1 + c_n \frac{e^{-\frac{1}{2}b_n Z}}{1+b_n Z}\right)^{-1}$ following (4.12), where in this case we denote by $b_n = \frac{\tau n}{\tau n+1} \geq 0$ and by $c_n = \frac{1-\rho}{\rho}(\tau n+1)^{3/2} \geq 0$. Then, for some fixed $\delta > 0$

$$\begin{aligned}
\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) &= \mathrm{P}_{\beta^*}\left(\left[1 + c_n \frac{e^{-\frac{1}{2}b_n Z}}{1+b_n Z}\right]^{-1} > \delta f_n\right) \\
&= \mathrm{P}_{\beta^*}\left(\frac{e^{-\frac{1}{2}b_n Z}}{1+b_n Z} < c_n^{-1}((\delta f_n)^{-1} - 1)\right) \\
&= \mathrm{P}_{\beta^*}\left(e^{\frac{1}{2}b_n Z}(1+b_n Z) > c_n((\delta f_n)^{-1} - 1)^{-1}\right) \\
&\leq \mathrm{P}_{\beta^*}\left(e^{\frac{1}{2}b_n Z} > \frac{c_n((\delta f_n)^{-1} - 1)^{-1}}{q_n}\right) + \mathrm{P}_{\beta^*}(1 + b_n Z > q_n),
\end{aligned} \tag{4.28}$$

for any $q_n > 0$, where in the last step we used the fact that for two random variables $X, Y > 0$, and $a, b > 0$, we have that $\mathrm{P}(XY > a) \leq \mathrm{P}(X > a/b) + \mathrm{P}(Y > b)$. Now we follow the steps in (4.19) and then apply the Chernoff bound analogously to the proof

of Part (i). In the first term in the RHS of (4.28) we have

$$\begin{aligned}
&\mathrm{P}_{\beta^*}\left(e^{\frac{1}{2}b_n Z} > \frac{c_n((\delta f_n)^{-1}-1)^{-1}}{q_n}\right) = \\
&\quad = \mathrm{P}_{\beta^*}\left(Z > \frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right)\right) \\
&\quad = 1 - \left[1 - \mathrm{P}_{\beta^*}\left(W_i > \frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right)\right)\right]^{p-p^*},
\end{aligned}$$

upon which using the Chernoff bound in Lemma S2 of Rossell (2021) we can establish that, for $\frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right) > 1$,

$$\begin{aligned}
&\mathrm{P}_{\beta^*}\left(W_i > \frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right)\right) \\
&\leq \left(\frac{2e}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right)\right)^{1/2} \exp\left\{-\frac{1}{2}\frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right)\right\}.
\end{aligned} \tag{4.29}$$

For this bound to converge to zero, it suffices that $\frac{2}{b_n}\log\left(\frac{c_n}{q_n}\left[(\delta f_n)^{-1}-1\right]^{-1}\right) \gg 1$, i.e. that

$$\delta f_n \gg \left(1 + \frac{c_n}{q_n}e^{-b_n/2}\right)^{-1}. \tag{4.30}$$

Similarly, in the second term of (4.28), and following (4.19), we express

$$\mathrm{P}_{\beta^*}(1 + b_n Z > q_n) = \mathrm{P}_{\beta^*}\left(Z > \frac{q_n - 1}{b_n}\right) = 1 - \left[1 - \mathrm{P}_{\beta^*}\left(W_i > \frac{q_n-1}{b_n}\right)\right]^{p-p^*}, \tag{4.31}$$

on which we can use the same Chernoff bound, and write that for $\frac{q_n-1}{b_n} > 1$

$$\mathrm{P}_{\beta^*}\left(W_i > \frac{q_n-1}{b_n}\right) \leq \left(e\frac{q_n-1}{b_n}\right)^{1/2} \exp\left\{-\frac{1}{2}\frac{q_n-1}{b_n}\right\}.$$

Once more this bound will converge to zero for $\frac{q_n-1}{b_n} \gg 1$, i.e. for

$$q_n \gg 1 + b_n. \tag{4.32}$$

Denote for brevity the sequence in the RHS of (4.29) by s_n, and that in the RHS of (4.31) by t_n. Note that under (4.30) and (4.32), $s_n \to 0$ and $t_n \to 0$, and thus we can apply Result 4.12 to each of the terms of (4.28), and re-write

$$\begin{aligned} \mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) &\leq \left[1 - \left(1 - \frac{1}{s_n^{-1}}\right)^{p-p^*}\right] + \left[1 - \left(1 - \frac{1}{t_n^{-1}}\right)^{p-p^*}\right] \\ &\leq (p-p^*)(s_n + t_n)(1+\omega_n), \end{aligned}$$

for some $\omega_n \to 0$. The last inequality also requires that $s_n(p-p^*) \to 0$ and that $t_n(p-p^*) \to 0$, which we will check below.

Now, recover a_n from (4.24) in Part (i), and set

$$q_n = 1 + a_n b_n,$$

and

$$\delta f_n = \left(1 + \frac{c_n}{q_n} e^{-\frac{b_n a_n}{2}}\right)^{-1} = \left(1 + c_n \frac{e^{-\frac{b_n a_n}{2}}}{1 + a_n b_n}\right)^{-1}, \tag{4.33}$$

which clearly satisfy (4.32) and (4.30), respectively. Replacing these expressions into the bounds in (4.29) and (4.31) provides that

$$s_n = t_n = (e a_n)^{1/2} e^{-a_n/2}.$$

Therefore, incorporating the algebra derived in (4.26) of Part (i), we have that

$$\begin{aligned} \mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) &\leq 2 s_n (p-p^*)(1+\omega_n) \\ &= \sqrt{8e}\left[\frac{1}{\log(p-p^*)} + \frac{\log\log(p-p^*)}{[\log(p-p^*)]^2}\right]^{1/2}(1+\omega_n), \end{aligned}$$

for $\omega_n \to 0$. This bound converges to zero as long as $p - p^* \to \infty$, i.e. if (C5) holds, implying then that $s_n(p-p^*) \to 0$ and that $t_n(p-p^*) \to 0$, and thus that the use of Result 4.12 was valid. Hence, we have that for any $\varepsilon > 0$ one may find a sufficiently large integer n_0 such that $\mathrm{P}_{\beta^*}(h_1(Z) > \delta f_n) \leq \varepsilon$, for some corresponding fixed $\delta > 0$ and every $n > n_0$. Thus, $h_1(Z) = O_{\mathrm{P}}(f_n)$, with f_n defined as in (4.33).

Finally, as in Part (i), let us also show that $h_1(Z) \xrightarrow{P} 0$, for which we require that $c_n \frac{e^{-\frac{1}{2}a_n b_n}}{1+a_n b_n} \to +\infty$. This holds if $-\frac{1}{2}a_n b_n + \log c_n - \log(1 + a_n b_n) \to +\infty$ as $n \to \infty$, i.e., given that $b_n \to 1$, if

$$\begin{aligned} &-\log(p-p^*) - \log\log(p-p^*) + \frac{3}{2}\log(\tau n + 1) + \\ &+ \log\left(\frac{1-\rho}{\rho}\right) - \log(\log(p-p^*) + \log\log(p-p^*)) \to +\infty \end{aligned}$$

as n grows. This condition holds assuming that (C1), (C2) and (C10) are satisfied. Then, using (4.33), we have that $\lim_{n\to\infty} \frac{f_n}{c_n^{-1}(1+a_n b_n)e^{\frac{1}{2}b_n a_n}} = 1$, and so

$$
\begin{aligned}
h_1(Z) &= O_{\mathrm{P}}\left(c_n^{-1}(1+a_n b_n)e^{\frac{1}{2}b_n a_n}\right) \\
&= O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)\log(p-p^*)}{(\tau n)^{3/2}}\left(1+2\left[\log(p-p^*)+\log\log(p-p^*)\right]\right)\right) \\
&= O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)[\log(p-p^*)]^2}{(\tau n)^{3/2}}\right)
\end{aligned}
$$

completes the proof. □

Proof. Theorem 4.1; Part (iii). The proof strategy is to find a valid sequence to bound this maximum in probability using (4.17) from Result 4.9. We establish the requirements to satisfy the conditions required therein, leading to a valid sequence. Once we have it, we look for the conditions that ensure that this bound converges to zero, which allow for the bound in probability to hold. Finally, we analyze the necessary conditions to make this maximum converge to zero.

Note that $\frac{n}{\phi_i}\hat{\beta}_i^2 \sim \chi_1^2\left(\frac{n}{\phi_i}\beta_i^{*2}\right)$. Let $Z := \min_{i:\beta_i^*\neq 0}\frac{n}{\phi_i}\hat{\beta}_i$, for which $Z^{-1} = \max_{i:\beta_i^*\neq 0}\frac{1}{\frac{n}{\phi_i}\hat{\beta}_i}$. Further, let $\lambda := \min_{i:\beta_i^*\neq 0}\frac{n}{\phi_i}\beta_i^{*2}$ be the smallest of the non-centrality parameters, where $\lambda \to +\infty$ as $n \to +\infty$. Recover the expressions for h_1, b_n and c_n from the proof of Part (i). Then

$$
\max_{i:\beta_i^*\neq 0} \mathrm{P}_{\mathrm{N}}(\gamma_i = 0 \mid y_i) = 1 - h_1(Z) = h_0(Z),
$$

since both h_1 and h_0 are monotonic functions. Then for $\delta > 0$ and a sequence $f_n > 0$

$$
\begin{aligned}
\mathrm{P}_{\beta^*}(h_0(Z) > \delta f_n) &= \mathrm{P}_{\beta^*}(1 - h_1(Z) > \delta f_n) = \mathrm{P}_{\beta^*}(h_1(Z) < 1 - \delta f_n) \\
&= \mathrm{P}_{\beta^*}(Z < h_1^{-1}(1-\delta f_n)) = \\
&= \mathrm{P}_{\beta^*}(Z^{-1} > \left[h_1^{-1}(1-\delta f_n)\right]^{-1}).
\end{aligned}
$$

Our objective is to find some δf_n such that this probability can be bounded by an arbitrarily small number. Denote $q_n := h_1^{-1}(1-\delta f_n)$, and notice that if $q_n \ll \lambda$, then we can use (4.17) from Result 4.9. For this condition to hold, we only need that $h_1^{-1}(1-\delta f_n) \ll \lambda$, i.e. that $\delta f_n \gg h_0(\lambda)$. In particular, pick $\delta f_n = h_0(\lambda/\log(n+1)) \gg h_0(\lambda)$, that is

$$
\delta f_n = 1 - \left(1 + c_n e^{-b_n \lambda/\log(n+1)}\right)^{-1}, \tag{4.34}
$$

which satisfies said condition thanks to the extra $\log(n+1)$ fraction in the exponent. Then, applying (4.17)

$$p_{\beta^*}\left(Z^{-1} > q_n^{-1}\right) \leq 1 - \exp\left\{-p^* e^{-\frac{\lambda}{2}\left[1-(q_n/\lambda)^{1/2}\right]^2} (q_n/\lambda)^{1/4}\right\}$$
$$= 1 - \exp\left\{-p^* e^{-\frac{\lambda}{2}\left[1-\left(\frac{1}{\log(n+1)}\right)^{1/2}\right]^2} \left(\frac{1}{\log(n+1)}\right)^{1/4}\right\}, \quad (4.35)$$

where in the last expression we replaced δf_n from (4.34). To bound $h_0(Z)$ in probability, we need that the RHS of this last expression vanishes for large enough n, which will happen if

$$p^* e^{-\frac{\lambda}{2}\left[1-\left(\frac{1}{\log(n+1)}\right)^{1/2}\right]^2} \left(\frac{1}{\log(n+1)}\right)^{1/4} \to 0 \quad (4.36)$$

as $n \to +\infty$. This will occur under two conditions: first, that asymptotically

$$\frac{\lambda}{2}\left[1-\left(\frac{1}{\log(n+1)}\right)^{1/2}\right]^2 \to \infty \Rightarrow \frac{|\beta_k^*|}{\sqrt{\phi_k}} \gg \frac{\sqrt{2}}{\sqrt{n}}\left(1-[\log(n+1)]^{-1/2}\right),$$

i.e. under (C7), and second that following (4.36)

$$p^* \ll e^{\frac{n}{2}\frac{\beta_k^{*2}}{\phi_k}\left(1-[\log(n+1)]^{-1/2}\right)^2 - \frac{1}{4}\log\log(n+1)},$$

i.e. under (C6), where $k := \arg\min_{i:\beta_i^* \neq 0} \beta_i^{*2}/\phi_i$. Hence, under (C7) and (C6), retaking (4.35) we can write

$$p_{\beta^*}(h_0(Z) > \delta f_n) \leq 1 - \exp\left\{-p^* e^{-\frac{\lambda}{2}\left[1-\left(\frac{1}{\log(n+1)}\right)^{1/2}\right]^2} \left(\frac{1}{\log(n+1)}\right)^{1/4}\right\}.$$

Importantly, then, this bound can be made arbitrarily small by having a sufficiently large n. So, for any $\varepsilon > 0$, one can find a sufficiently large n_0 such that $p_{\beta^*}(h_0(Z) > \delta f_n) \leq \varepsilon$, for every $n > n_0$, and some fixed $\delta > 0$. Therefore, $h_0(Z) = O_{\mathrm{P}}(f_n)$ as defined by (4.34), i.e.

$$h_0(Z) = O_{\mathrm{P}}\left(1 - \left(1 + c_n e^{-b_n \lambda / \log(n+1)}\right)^{-1}\right).$$

We also want to show that $h_0(Z) \to 0$, which requires that $c_n e^{-b_n\lambda/\log(n+1)} \to 0$, i.e. that provided $b_n \to 1/2$ and $\tau n + 1 \to \tau n$,

$$\begin{aligned} & b_n \frac{\lambda}{\log(n+1)} - \log c_n \to \infty \\ & \Rightarrow \frac{1}{2}\frac{n}{\log(n)}\frac{\beta_k^{*2}}{\phi_k} - \log\left(\frac{1-\rho}{\rho}\right) - \frac{1}{2}\log(\tau n) \to \infty. \end{aligned}$$

This condition holds if we assume (C3), and so we have that $\lim_{n\to\infty} \frac{f_n}{c_n e^{-b_n\lambda/\log(n+1)}} = 1$. Thus, we have that

$$\begin{aligned} h_0(Z) &= O_{\mathrm{P}}\left(c_n e^{-b_n\lambda/\log(n+1)}\right) \\ &= O_{\mathrm{P}}\left(\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\right), \end{aligned}$$

which completes the proof. □

Proof. Theorem 4.1; Part (iv). We proceed analogously to the proof of Part (iii). In this case, recover λ from said proof, and then h_1, b_n and c_n from the proof of Part (ii) instead. As in Part (iii), we have

$$\max_{i:\beta_i^*\neq 0} \mathrm{P}_{\mathrm{M}}(\gamma_i = 0 \mid y_i) = 1 - h_1(Z) = h_0(Z).$$

Now we replicate the steps leading to (4.28) from the proof of Part (ii), where for $\delta > 0$, some sequence $f_n > 0$ and any $q_n > 0$, we have

$$\begin{aligned} \mathrm{P}_{\beta^*}(h_0(Z) > \delta f_n) &= \mathrm{P}_{\beta^*}(h_1(Z) < 1 - \delta f_n) \\ &= \mathrm{P}_{\beta^*}\left(e^{\frac{1}{2}b_n Z}(1+b_n Z) < c_n((1-\delta f_n)^{-1} - 1)^{-1}\right) \\ &\leq \mathrm{P}_{\beta^*}\left(e^{\frac{1}{2}b_n Z} < \frac{c_n((1-\delta f_n)^{-1} - 1)^{-1}}{q_n}\right) + p_{\beta^*}\left((1+b_n Z) < q_n\right) \qquad (4.37) \\ &= \mathrm{P}_{\beta^*}\left(Z^{-1} > \left[\frac{2}{b_n}\log\left(\frac{c_n((1-\delta f_n)^{-1} - 1)^{-1}}{q_n}\right)\right]^{-1}\right) + \\ &+ p_{\beta^*}\left(Z^{-1} > \left[\frac{q_n - 1}{b_n}\right]^{-1}\right), \qquad (4.38) \end{aligned}$$

where in (4.37) we employed the analogous inequality to that of Part (ii). Our objective is again to find such f_n, for which we will employ Result 4.9 again. In order to use it,

we need that two conditions hold: first, that

$$\frac{q_n-1}{b_n} \ll \lambda \Rightarrow q_n \ll 1+b_n\lambda,$$

thus let us pick $q_n = 1+b_n\lambda/\log(n+1)$, which satisfies this condition; and second, that

$$\frac{2}{b_n}\log\left(\frac{c_n((1-\delta f_n)^{-1}-1)^{-1}}{q_n}\right) \ll \lambda \Rightarrow \delta f_n \gg 1-\left(1+c_n\frac{e^{-\frac{1}{2}b_n\lambda}}{q_n}\right)^{-1},$$

for which we can set

$$\delta f_n = 1-\left[1+c_n\frac{\exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}\right\}}{1+b_n\frac{\lambda}{\log(n+1)}}\right]^{-1}.$$

Note that this implies that the bound for each of the two terms in (4.38) is that of (4.35), and so

$$\mathrm{P}_{\beta^*}(h_0(Z)>\delta f_n) \leq 2\left[1-\exp\left\{-p^*e^{-\frac{\lambda}{2}\left[1-\left(\frac{1}{\log(n+1)}\right)^{1/2}\right]^2}\left(\frac{1}{\log(n+1)}\right)^{1/4}\right\}\right].$$

As in Part (iii), to make this bound vanish for sufficiently large n, we need that conditions (C7) and (C6) hold, which they do by assumption. In that case, this bound can be made arbitrarily small for a large enough n, and so for any $\varepsilon>0$ a sufficiently large n_0 exists such that $\mathrm{P}_{\beta^*}(h_0(Z)>\delta f_n)\leq\varepsilon$, for every $n>n_0$, and some fixed $\delta>0$. Therefore, $h_0(Z)=O_{\mathrm{P}}(f_n)$, i.e.

$$h_0(Z) = O_{\mathrm{P}}\left(1-\left[1+c_n\frac{\exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}\right\}}{1+b_n\frac{\lambda}{\log(n+1)}}\right]^{-1}\right).$$

Once more, we want to show that $h_0(Z)\to 0$ as $n\to\infty$, this would require that

$$\begin{aligned} &c_n\frac{\exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}\right\}}{1+b_n\frac{\lambda}{\log(n+1)}} = \\ &= \exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}+\log c_n-\log\left(1+b_n\frac{\lambda}{\log(n+1)}\right)\right\}\to 0 \end{aligned}$$

as $n\to\infty$, i.e. that given $b_n\to 1/2$ and $\tau n+1\to\tau n$

$$\frac{1}{2}\frac{\lambda}{\log(n+1)}+\log\left(1+\frac{\tau n}{\tau n+1}\frac{\lambda}{\log(n+1)}\right)-\log\left(\frac{1-\rho}{\rho}\right)-\frac{3}{2}\log(\tau n)\to+\infty.$$

This will happen as long as we assume (C4). In such case, $\lim_{n\to\infty} \frac{f_n}{c_n \frac{\exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}\right\}}{1+b_n\frac{\lambda}{\log(n+1)}}} = 1$,

and thus we can write

$$
\begin{aligned}
h_0(Z) &= O_{\mathrm{P}}\left(c_n \frac{\exp\left\{-\frac{b_n}{2}\frac{\lambda}{\log(n+1)}\right\}}{1+b_n\frac{\lambda}{\log(n+1)}}\right) \\
&= O_{\mathrm{P}}\left(\frac{1-\rho}{\rho}(\tau n)^{3/2}\frac{\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}}{1+\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\right) \\
&= O_{\mathrm{P}}\left(\frac{\tau n}{\frac{n}{\log n}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\right) \\
&= O_{\mathrm{P}}\left(\frac{\tau \log n}{\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\frac{1-\rho}{\rho}(\tau n)^{1/2}\exp\left\{-\frac{1}{2}\frac{n}{\log(n+1)}\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}\right\}\right),
\end{aligned}
$$

to complete the proof. □

Proof of Theorem 4.2

Proof. Theorem 4.2; Part (i) and Part (ii). The strategy of the proof is to decompose the sum into its individual terms, and then upper bounding these elements using the maxima and minima of their respective sets. Then we combine the probability bounds of these extrema, e.g. re-using those identified in Theorem 4.1, to produce a probability bound for the entire expression, separately for the Gaussian prior in Part (i) and pMOM prior in Part (ii).

Denote $Z_i = n\hat{\beta}_i^2/\phi_i$ for short, and note that $Z_i \sim \chi_1^2(n\hat{\beta}_i^*/\phi_i)$, i.e. with a non-centrality parameter that grows with n whenever $\beta_i^* \neq 0$. Following the notation from the proof of Theorem 4.1, denote $h_1(Z_i) := \mathrm{P}(\gamma_i = 1 \mid y_i)$ and $h_0(Z_i) := \mathrm{P}(\gamma_i = 0 \mid y_i)$ for short. We have that

$$
\begin{aligned}
\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i - \beta_i^*| = \sum_{i:\gamma_i^*=0} |\tilde{\beta}_i| &= \sum_{i:\gamma_i^*=0} \left| h_1(Z_i) f(Z_i) \hat{\beta}_i \right| \\
&\leq (p-p^*) h_1\left(\max_{i:\beta_i^*=0} Z_i\right) f\left(\min_{i:\beta_i^*=0} Z_i\right) \max_{i:\beta_i^*=0} |\hat{\beta}_i|, \quad (4.39)
\end{aligned}
$$

where $f(\cdot) > 0$ and $h_1(\cdot) > 0$ are monotonically non-increasing and increasing functions, respectively. For the pMOM prior, we define $f(Z_i) = \frac{Z_i+3}{Z_i+1}$ following (4.14). For

the Gaussian prior, f is independent of Z_i and is simply $f(Z_i) = \frac{\tau n}{\tau n+1}$, following (4.10). For brevity, denote these by f_{M} and f_{N} respectively, and note that in both cases $f_{\mathrm{M}} \to 1$ and $f_{\mathrm{N}} \to 1$, and also that $f_{\mathrm{M}} = O_{\mathrm{P}}(1)$ and $f_{\mathrm{N}} = O(1)$.

Additionally, in Part (i) and Part (ii) of Theorem 4.1 we established the probability bounds of $h_1\left(\max_{i:\beta_i^*=0} Z_i\right)$ for both priors. Thus, to bound (4.39) in probability we are left to bound $\max_{i:\beta_i^*=0} |\hat{\beta}_i|$. To bound this term, it suffices to use Result 4.7 on the set of $\hat{\beta}_i \sim \mathrm{N}(0, \phi_i/n)$ estimates (with $p - p^*$ elements) to obtain

$$\max_{i:\beta_i^*=0} |\hat{\beta}_i| = O_{\mathrm{P}}\left(\sqrt{\frac{2\log(p-p^*)}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right).$$

Therefore, to prove Part (i), we combine the bound in Part (i) of Theorem 4.1 (under its corresponding conditions) with this last result, and establish that

$$\begin{aligned}\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i - \beta_i^*| &= O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{1/2}}\sqrt{\frac{2[\log(p-p^*)]^3}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right) \\ &= O_{\mathrm{P}}\left(\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{1/2}}\sqrt{\frac{[\log(p-p^*)]^3}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right).\end{aligned}$$

To prove Part (ii), we analogously use the bound in Part (ii) of Theorem 4.1, and write

$$\begin{aligned}\sum_{i:\gamma_i^*=0} |\tilde{\beta}_i - \beta_i^*| &= O_{\mathrm{P}}\left((1+2k_n)\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{3/2}}\sqrt{\frac{2[\log(p-p^*)]^3}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right) \\ &= O_{\mathrm{P}}\left(\frac{k_n}{\tau n}\frac{\rho}{1-\rho}\frac{(p-p^*)^2}{(\tau n)^{3/2}}\sqrt{\frac{[\log(p-p^*)]^3}{n}} \max_{i:\beta_i^*=0} \phi_i^{1/2}\right),\end{aligned}$$

where $k_n = \log(p-p^*) + \log\log(p-p^*)$. □

Proof. Theorem 4.2; Part (iii) and Part (iv). We replicate the same proof strategy of Parts (i) and (ii). Let us recover the same notation as in said proof.

To upper bound the target expression, we use the triangle inequality

$$\begin{aligned}\sum_{i:\gamma_i^*=1} |\tilde{\beta}_i - \beta_i^*| &\leq \sum_{i:\gamma_i^*=1} \left[|\tilde{\beta}_i - \hat{\beta}_i| + |\hat{\beta}_i - \beta_i^*|\right] \\ &\leq p^* \left[\max_{i:\beta_i^*\neq 0} |\tilde{\beta}_i - \hat{\beta}_i| + \max_{i:\beta_i^*\neq 0} |\hat{\beta}_i - \beta_i^*|\right]. \qquad (4.40)\end{aligned}$$

Note that $\hat{\beta}_i - \beta_i^* \sim \mathrm{N}(0, \phi_i/n)$, and so we can apply Result 4.8 to the RHS of (4.40) and get

$$\max_{i:\beta_i^*\neq 0} |\hat{\beta}_i - \beta_i^*| = O_{\mathrm{P}}\left(\sqrt{\frac{2\log p^*}{n}} \max_{i:\beta_i^*\neq 0} \phi_i^{1/2}\right). \tag{4.41}$$

As for the first term of the RHS of (4.40), we use the triangle inequality

$$\begin{aligned}
\max_{i:\beta_i^*\neq 0} |\tilde{\beta}_i - \hat{\beta}_i| &= \max_{i:\beta_i^*\neq 0} \left| h_1(Z_i) f(Z_i) \hat{\beta}_i - \hat{\beta}_i \right| \\
&= \max_{i:\beta_i^*\neq 0} \left| (h_1(Z_i) - 1)\, f(Z_i) \hat{\beta}_i + (f(Z_i) - 1)\, \hat{\beta}_i \right| \\
&\leq \max_{i:\beta_i^*\neq 0} \left[\left| -h_0(Z_i) f(Z_i) \hat{\beta}_i \right| + \left| (f(Z_i) - 1)\, \hat{\beta}_i \right| \right] \\
&= \max_{i:\beta_i^*\neq 0} \left[h_0(Z_i) f(Z_i) |\hat{\beta}_i| + |f(Z_i) - 1|\, |\hat{\beta}_i| \right] \\
&\leq \left[f\left(\min_{i:\beta_i^*\neq 0} Z_i \right) h_0\left(\min_{i:\beta_i^*\neq 0} Z_i \right) + \left| f\left(\min_{i:\beta_i^*\neq 0} Z_i \right) - 1 \right| \right] \max_{i:\beta_i^*\neq 0} |\hat{\beta}_i|.
\end{aligned} \tag{4.42}$$

First, notice given that $\hat{\beta}_i \sim \mathrm{N}(\beta_i^*, \phi_i/n)$ we can apply Result 4.7, by which

$$\begin{aligned}
\max_{i:\beta_i^*\neq 0} |\hat{\beta}_i| &= O_{\mathrm{P}}\left(\max_{i:\beta_i^*\neq 0} |\beta_i^*| + \sqrt{\frac{2\log p^*}{n}} \max_{i:\beta_i^*= 0} \phi_i^{1/2} \right) \\
&= O_{\mathrm{P}}\left(\max_{i:\beta_i^*\neq 0} |\beta_i^*| + \sqrt{\frac{\log p^*}{n}} \max_{i:\beta_i^*= 0} \phi_i^{1/2} \right).
\end{aligned}$$

Our interest in (4.42), beyond the term $\max_{i:\beta_i^*\neq 0} |\hat{\beta}_i|$, goes to the second term $\left| f\left(\min_{i:\beta_i^*\neq 0} Z_i\right) - 1 \right|$, since this will be the dominating term of this bound, given that from Theorem 4.1 we know that $h_0\left(\min_{i:\beta_i^*\neq 0} Z_i\right)$ will in any case decrease exponentially in n, up to lower order factors. We analyse this quantity for the two priors.

For the Gaussian prior, this term is simply

$$\left| f_{\mathrm{N}}\left(\min_{i:\beta_i^*\neq 0} Z_i \right) - 1 \right| = \frac{1}{\tau n + 1} = O\left((\tau n)^{-1}\right).$$

Therefore, recovering (4.42), for said prior we obtain

$$\max_{i:\beta_i^*\neq 0} |\tilde{\beta}_i^{\mathrm{N}} - \hat{\beta}_i| = O_{\mathrm{P}}\left(\frac{1}{\tau n} \left[\sqrt{\frac{\log p^*}{\tau n}} \max_{i:\beta_i^*= 0} \phi_i^{1/2} + \max_{i:\beta_i^*\neq 0} |\beta_i^*| \right] \right),$$

and so combining this with (4.41), and recovering (4.40), we obtain

$$\sum_{i:\gamma_i^*=1} |\tilde{\beta}_i^{\mathrm{N}} - \beta_i^*| = O_{\mathrm{P}}\left(p^*\left[\left(1+\frac{1}{\tau n}\right)\sqrt{\frac{\log p^*}{\tau n}}\max_{i:\beta_i^*=0}\phi_i^{1/2} + \frac{1}{\tau n}\max_{i:\beta_i^*\neq 0}|\beta_i^*|\right]\right). \quad (4.43)$$

which proves the statement in Part (iii).

Instead, for the pMOM prior

$$\left| f_{\mathrm{M}}\left(\min_{i:\beta_i^*\neq 0} Z_i\right) - 1\right| = \frac{2}{\min_{i:\beta_i^*\neq 0} Z_i + 1}. \quad (4.44)$$

Since $\left(\min_{i:\beta_i^*\neq 0} Z_i\right)^{-1} = \max_{i:\beta_i^*\neq 0}\frac{1}{Z_i} = o_{\mathrm{P}}\left(\frac{\log(1+\lambda_{\min})}{\lambda_{\min}}\right)$, for

$$\lambda_{\min} = n\min_{i:\beta_i^*\neq 0}(\beta_i^{*2}/\phi_i),$$

we can apply Result 4.10 to the RHS of (4.44), by which

$$\begin{aligned}\frac{2}{\min_{i:\beta_i^*\neq 0} Z_i + 1} &= O_{\mathrm{P}}\left(\frac{1}{\frac{\lambda_{\min}}{\log(1+\lambda_{\min})}+1}\right)\\ &= O_{\mathrm{P}}\left(\left(1+\frac{n\min_{i:\beta_i^*\neq 0}(\beta_i^{*2}/\phi_i)}{\log(1+n\min_{i:\beta_i^*\neq 0}(\beta_i^{*2}/\phi_i))}\right)^{-1}\right).\end{aligned}$$

As $n\to\infty$ this probability bound becomes

$$\left| f_{\mathrm{M}}\left(\min_{i:\beta_i^*\neq 0} Z_i\right) - 1\right| = O_{\mathrm{P}}\left(\frac{\log n}{n\min_{i:\beta_i^*\neq 0}\beta_i^{*2}/\phi_i}\right) = O_{\mathrm{P}}\left(\frac{\log n}{n}\frac{1}{\min_{i:\beta_i^*\neq 0}\beta_i^{*2}/\phi_i}\right),$$

since for any $x_n\to\infty$ it holds $\lim_{n\to\infty}\frac{\frac{\log(1+x_n)}{x_n}}{\left(1+\frac{x_n}{\log(1+x_n)}\right)^{-1}} = 1$. In this case, the bound in (4.42) becomes

$$\max_{i:\beta_i^*\neq 0}|\tilde{\beta}_i^{\mathrm{M}} - \hat{\beta}_i| = O_{\mathrm{P}}\left(\frac{\log n}{\min_{i:\beta_i^*\neq 0}\beta_i^{*2}/\phi_i}\sqrt{\frac{2\log p^*}{n^3}}\max_{i:\beta_i^*=0}\phi_i^{1/2}\right).$$

Again, we combine this with (4.41), and the probability bound for (4.40) writes

$$\sum_{i:\gamma_i^*=1}|\tilde{\beta}_i^{\mathrm{M}} - \beta_i^*| =$$

$$O_{\mathrm{P}}\left(p^*\left[\left(1+\frac{\log n}{n}\frac{1}{\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\right)\sqrt{\frac{\log p^*}{n}}\max_{i:\beta_i^*=0}\phi_i^{1/2} + \frac{\log n}{n}\frac{\max_{i:\beta_i^*\neq 0}|\beta_i^*|}{\min_{i:\beta_i^*\neq 0}\frac{\beta_i^{*2}}{\phi_i}}\right]\right),$$

which completes the proof for Part (iv). If (C8) holds, then $\to 0$, similarly to the bound in Part (iii). □

Proof of Theorem 4.3

Proof. Theorem 4.3; Part (i). The strategy for the proof is use the expression of Result 4.11, and then upper bound the integral therein to obtain a bound for $B^{\pi}_{i,01}$. To obtain this bound we partition the integral by different relevant intervals, and find simpler bounds on each of these parts, which then added up provide the presented result.

To start, let us state two well-known tail bounds for the standard Normal distribution that will be used throughout, namely if $X \sim \mathrm{N}(0,1)$, then for any $t > 0$

$$\mathrm{P}(X > t) \leq e^{-t^2/2}, \tag{4.45}$$

which is the Chernoff bound, and

$$\mathrm{P}(X > t) \geq \frac{\phi(t)}{t}\left(1 - \frac{1}{t^2}\right), \tag{4.46}$$

where $\phi(t)$ is the PDF of the standard normal evaluated at $t > 0$.

To ease notation henceforth, denote $w_{\pi}(\beta_i) := \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i;0,\tau\phi_i)}$, and $\hat{m}_i := \frac{\tau n}{\tau n+1}\hat{\beta}_i$, with corresponding $m^*_i = \frac{\tau n}{\tau n+1}\beta^*_i$, where clearly $\hat{m}_i \to m^*_i$. For Part (i), $m^*_i = 0$. Denote also $s_i = \frac{\tau}{\tau n+1}\phi_i$, which is known. Hence, $\beta_i \mid y_i \sim \mathrm{N}(\hat{m}_i, s_i)$.

Further, define $\beta_0 := \max_{\beta_i}\left\{\arg\min_{\beta_i} |w_{\pi}(\beta_i) - 1|\right\}$. In plain terms, β_0 is the largest value at which the two densities intersect. Let also $c_A := \min_{|\beta_i|<\beta_0} w_{\pi}(\beta_i)$ and $c_B := \min_{|\beta_i|<\hat{m}_i} w_{\pi}(\beta_i)$, and finally $c_1 := \min\{c_A, c_B\}$. These are simply the smallest values of the density ratio in the central parts of the distribution.

Recover the expression in Result 4.11 for our prior π. Since we are interested in checking if $B^{\pi}_{i,01} \to \infty$ faster than the Gaussian, we are looking for an upper bound on $B^{\pi}_{i,01}$, and thus for a lower bound on the integral of Result 4.11. We partition said integral as

$$\begin{aligned}\int w_{\pi}(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i = &\int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} w_{\pi}(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i + \\ &+ \int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_{\pi}(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i + \\ &+ \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} w_{\pi}(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i.\end{aligned} \tag{4.47}$$

We now address each of these partitioned integrals separately, and in order of appearence. For the first two, we do not need to assume anything on π beyond being

bounded, and centreed and symmetric around zero. Then

$$\begin{aligned}
\int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i &\geq \int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} c_1\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_1\mathrm{P}(|\beta_i| < \min\{|\hat{m}_i|,\beta_0\} \mid y_i) \\
&= c_1\left[\mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1)\right] \\
&= c_1\left[\left(\underbrace{\mathrm{P}(|\beta_i| < |m_i^*| \mid y_i, |\hat{m}_i| < \beta_0)}_{=0} + o_\mathrm{P}(1)\right) + o_\mathrm{P}(1)\right] \\
&= c_1 o_\mathrm{P}(1),
\end{aligned} \tag{4.48}$$

where in the third step we used the fact that here for any event A_n

$$\begin{aligned}
\mathrm{P}(A_n) &= \mathrm{P}(A_n \mid |\hat{m}_i| < \beta_0)\underbrace{\mathrm{P}(|\hat{m}_i| < \beta_0)}_{\to 1} + \underbrace{\mathrm{P}(A_n \mid |\hat{m}_i| > \beta_0)}_{\leq 1}\underbrace{\mathrm{P}(|\hat{m}_i| > \beta_0)}_{\to 0} \\
&= \mathrm{P}(A_n \mid |\hat{m}_i| < \beta_0)\left[1 - o_\mathrm{P}(1)\right] + o_\mathrm{P}(1) \\
&= \mathrm{P}(A_n \mid |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1),
\end{aligned} \tag{4.49}$$

given that $m_i^* < \beta_0$ by construction; and in the fourth step we used the fact that $\hat{m}_i \to m_i^*$. As for the second integral, let $X \sim \mathrm{N}(0,1)$, and write

$$\begin{aligned}
&\int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&\geq \int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} c_1\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_1\left[\mathrm{P}(|\beta_i| < \max\{|\hat{m}_i|,\beta_0\} \mid y_i) - \mathrm{P}(|\beta_i| < \min\{|\hat{m}_i|,\beta_0\} \mid y_i)\right] \\
&= c_1\left[\mathrm{P}(|\beta_i| < \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) - \mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1)\right] \\
&= c_1\left[1 - \mathrm{P}(|\beta_i| > \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) - \mathrm{P}(|\beta_i| < |m_i^*| \mid y_i, |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1)\right] \\
&= c_1\left[1 - \mathrm{P}(\beta_i > \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) - \mathrm{P}(-\beta_i < -\beta_0 \mid y_i, |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1)\right] \\
&= c_1\left[1 - P\left(X > \frac{\beta_0 - \hat{m}_i}{\sqrt{s_i}}\right) - P\left(X > \frac{\beta_0 + \hat{m}_i}{\sqrt{s_i}}\right) + o_\mathrm{P}(1)\right] \\
&\geq c_1\left[1 + o_\mathrm{P}(1) \underbrace{- O_\mathrm{P}\left(e^{-\frac{n}{2}\frac{\beta_0^2}{\phi_i}}\right) - O_\mathrm{P}\left(e^{-\frac{n}{2}\frac{\beta_0^2}{\phi_i}}\right)}_{=o_\mathrm{P}(1)}\right] = c_1 + o_\mathrm{P}(1),
\end{aligned} \tag{4.50}$$

where in the second-to-last step we used the Chernoff bound from (4.45) on the two

terms, as in

$$P\left(X > \frac{\beta_0 - \hat{m}_i}{\sqrt{s_i}}\right) \leq \exp\left\{-\frac{1}{2}\frac{(\beta_0 - \hat{m}_i)^2}{s_i}\right\} = \exp\left\{-\frac{1}{2}\frac{(\beta_0 - m_i^*)^2}{s_i}\right\} + o_{\mathrm{P}}(1)$$
$$= \exp\left\{-\frac{1}{2}\frac{\tau n + 1}{\tau}\frac{\beta_0^2}{\phi_i}\right\} + o_{\mathrm{P}}(1) = O_{\mathrm{P}}\left(e^{-\frac{n}{2}\frac{\beta_0^2}{\phi_i}}\right) + o_{\mathrm{P}}(1), \quad (4.51)$$

and analogously for the second term. Finally, for the third integral, there are two possibilities: that the tails are thinner, or thicker than those of the Gaussian prior. If the tails are thinner, we have assumed that they are at most $\pi(\beta_i) \propto \exp\left\{-\frac{1}{2}\frac{\beta_i^{2k}}{\tau\phi_i}\right\}$, for some fixed integer $k > 1$. Let c_2 be the ratio of normalising constants between π and the Gaussian, and so

$$\begin{aligned}
&\int_{|\beta_i| > \max\{|\hat{m}_i|, \beta_0\}} w_\pi(\beta_i) \mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i) \mathrm{d}\beta_i \\
&\geq \int_{|\beta_i| > \max\{|\hat{m}_i|, \beta_0\}} c_2 \exp\left\{-\frac{1}{2}\frac{\beta_i^{2k}}{\tau\phi_i}\right\} \mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i) \mathrm{d}\beta_i \\
&= c_2 \int_{|\beta_i| > \max\{|\hat{m}_i|, \beta_0\}} (2\pi s_i)^{-1/2} \exp\left\{-\frac{1}{2s_i}\left[\underbrace{(\beta_i - \hat{m}_i)^2 - \frac{\beta_i^{2k}}{\tau n + 1}}_{(\beta_i - \hat{m}_i)^2 + o(1)}\right]\right\} \mathrm{d}\beta_i \\
&= c_2 \left[\mathrm{P}(|\beta_i| > \max\{|\hat{m}_i|, \beta_0\} \mid y_i) + o(1)\right] \\
&= c_2 \left[\mathrm{P}(|\beta_i| > \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) + o_{\mathrm{P}}(1) + o(1)\right] \\
&= c_2 \left[\mathrm{P}(\beta_i > \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) + \mathrm{P}(-\beta_i < -\beta_0 \mid y_i, |\hat{m}_i| < \beta_0) + o_{\mathrm{P}}(1)\right] \\
&= c_2 \left[P\left(X > \frac{\beta_0 - \hat{m}_i}{\sqrt{s_i}}\right) + P\left(X > \frac{\beta_0 + \hat{m}_i}{\sqrt{s_i}}\right) + o_{\mathrm{P}}(1)\right] \\
&\geq c_2 \left[\underbrace{O_{\mathrm{P}}\left(n^{-1/2} e^{-\frac{n}{2}\frac{\beta_0^2}{\phi_i}}(1 - 1/n)\right) + O_{\mathrm{P}}\left(n^{-1/2} e^{-\frac{n}{2}\frac{\beta_0^2}{\phi_i}}(1 - 1/n)\right)}_{= o_{\mathrm{P}}(1)} + o_{\mathrm{P}}(1)\right] \\
&= c_2 o_{\mathrm{P}}(1), \qquad (4.52)
\end{aligned}$$

where in the second to last step we used (4.46) and rearranged in a similar fashion as in (4.51). If, instead, the tails are thicker than the Gaussian, we have assumed that they are at most Cauchy, centreed at zero, with a scale parameter $\nu = c_3^{-1/2}\sqrt{\tau\phi_i}$, for some

constant parameter $c_3 > 0$. In that case

$$\begin{aligned}
&\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&\geq \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} \pi_{\text{Cauchy}}(\beta_i; 0, c_3^{-1/2}\sqrt{\tau\phi_i})\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_3^{-1/2} \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} (2\pi s_i)^{-1/2} \times \\
&\times \exp\left\{ -\frac{1}{2s_i} \left[\underbrace{(\beta_i - \hat{m}_i)^2 + \frac{2\tau\phi_i}{\tau n + 1} \log\left(1 + c_3 \frac{\beta_i^2}{\tau\phi_i}\right)}_{(\beta_i-\hat{m}_i)^2 + o(1)} \right] \right\} \mathrm{d}\beta_i \\
&= c_3^{-1/2} \left[\mathrm{P}(|\beta_i| > \max\{|\hat{m}_i|, \beta_0\} \mid y_i) + o(1)\right] \\
&\geq c_3^{-1/2} o_{\mathrm{P}}(1),
\end{aligned} \tag{4.53}$$

where in the last step we have proceeded identically to the corresponding step of thin tails case.

In summary, gathering the results in (4.48), (4.50), (4.52) and (4.53), combined with Result 4.11 and the conditions in Theorem 4.1 for the Gaussian prior, we have that

$$\begin{aligned}
B_{i,01}^{\pi} &= B_{i,01}^{\mathrm{N}} \left[\int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i; 0, \tau\phi_i)} \mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \right]^{-1} \\
&\leq B_{i,01}^{\mathrm{N}} [c_1 + o_{\mathrm{P}}(1)]^{-1} = c_1^{-1} B_{i,01}^{\mathrm{N}} + o_{\mathrm{P}}(1).
\end{aligned}$$

By letting $c_\pi = c_1^{-1}$, the statement of Part (i) is proved. □

Proof. Theorem 4.3; Part (ii). We proceed analogously to the proof of Part (i). Recover the notation therein, where now $m_i^* \neq 0$. Assume without loss of generality that $m_i^* > 0$. Replace $c_A := \max_{|\beta_i|<\beta_0} w_\pi(\beta_i)$, $c_B := \max_{|\beta_i|<\hat{m}_i} w_\pi(\beta_i)$, and finally $c_4 := \max\{c_A, c_B\}$.

Since we look for a large posterior inclusion probability, we are now interested in checking if $B_{i,01}^{\pi} \to 0$ faster than the Gaussian, i.e. we are looking for a lower bound on $B_{i,01}^{\pi}$, and so for an upper bound on the integral of Result 4.11. The integral of interest here can be expressed as

$$\begin{aligned}
\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i = \\
&\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \mathrm{I}(|m_i^*| < \beta_0) + \\
&+ \int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i (1 - \mathrm{I}(|m_i^*| < \beta_0)).
\end{aligned} \tag{4.54}$$

We shall analyse the two cases of this indicator function separately.

Let us begin with the heavier-than-Gaussian tails case. First, notice that the ratio of densities with a Cauchy density

$$\frac{\pi_{\mathrm{C}}(\beta_i;0,c_3^{-1/2}\sqrt{\tau\phi_i})}{\pi_{\mathrm{N}}(\beta_i;0,\tau\phi_i} = \frac{(\pi c_3^{1/2}\sqrt{\tau\phi_i})^{-1}\left(1+\frac{\beta_i^2}{c_3^{-1}\tau\phi_i}\right)^{-1}}{(2\pi\tau\phi_i)^{-1/2}\exp\left\{-\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}}$$
$$= c_3^{1/2}\frac{\sqrt{2\pi}}{\pi}\frac{\exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}}{1+\frac{\beta_i^2}{c_3^{-1}\tau\phi_i}} \leq \tilde{c}_3\exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\},$$

for $\tilde{c}_3 = c_3(2/\pi)^{1/2}$. Now recover the partition introduced in (4.47). We first tackle the case in which $\beta_0 > |m_i^*|$ Next, we bound the partitioned integrals. We start by the tails

$$\begin{aligned}
&\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&\leq \tilde{c}_3\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} \exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= \tilde{c}_3\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} (2\pi s_i)^{-1/2}\exp\left\{-\frac{1}{2s_i}\left[\underbrace{(\beta_i-\hat{m}_i)^2 - \frac{\beta_i^2}{\tau n+1}}_{(\beta_i-\hat{m}_i)^2+o(1)}\right]\right\}\mathrm{d}\beta_i \\
&= \tilde{c}_3\left[P\left(X > \frac{\beta_0-\hat{m}_i}{\sqrt{s_i}}\right) + P\left(X > \frac{\beta_0+\hat{m}_i}{\sqrt{s_i}}\right) + o_{\mathrm{P}}(1)\right] \\
&\leq \tilde{c}_3\left[\underbrace{O_{\mathrm{P}}\left(e^{-\frac{n}{2}\frac{(\beta_0-m_i^*)^2}{\phi_i}}\right) + O_{\mathrm{P}}\left(e^{-\frac{n}{2}\frac{(\beta_0+m_i^*)^2}{\phi_i}}\right)}_{=o_{\mathrm{P}}(1)} + o_{\mathrm{P}}(1)\right] \\
&= \tilde{c}_3 o_{\mathrm{P}}(1),
\end{aligned} \tag{4.55}$$

where the missing steps are identical to the derived in (4.52), and in the last step we

used Chernoff analogously to (4.51). In the central part of the integral

$$\begin{aligned}\int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i &\leq \int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\ &= c_4\left[\underbrace{\mathrm{P}(|\beta_i|<|\hat{m}_i| \mid y_i,|m_i^*|<\beta_0)}_{=1/2+o_\mathrm{P}(1)}+o_\mathrm{P}(1)\right] \\ &= c_4/2+o_\mathrm{P}(1), \end{aligned} \tag{4.56}$$

where the second step follows follows from the fact that $\beta_i \mid y_i$ is centreed around $\hat{m}_i$, and so there is half of the probability mass on each side of $\hat{m}_i$, a plus an $o_\mathrm{P}(1)$ term bounded by a tail probability as before. Similarly,

$$\begin{aligned}&\int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\ &\leq \int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\ &= c_4\left[(\mathrm{P}(|\beta_i|<\beta_0 \mid y_i,|m_i^*|<\beta_0)+o_\mathrm{P}(1))-(\mathrm{P}(|\beta_i|<|\hat{m}_i| \mid y_i,|m_i^*|<\beta_0)+o_\mathrm{P}(1))\right] \\ &= c_4\left[1-\underbrace{\mathrm{P}(|\beta_i|>\beta_0 \mid y_i,|m_i^*|<\beta_0)}_{=2\mathrm{P}(\beta_i>\beta_0|y_i,|m_i^*|<\beta_0)}-\underbrace{\mathrm{P}(|\beta_i|<|\hat{m}_i| \mid y_i,|m_i^*|<\beta_0)}_{=1/2+o_\mathrm{P}(1)}+o_\mathrm{P}(1)\right] \\ &= c_4\left[1/2-2O_\mathrm{P}\left(n^{-1/2}e^{-\frac{n}{2}\frac{(\beta_0-m^*)^2}{\phi_i}}(1-1/n)\right)+o_\mathrm{P}(1)\right] \\ &= c_4/2+o_\mathrm{P}(1), \end{aligned} \tag{4.57}$$

where in the second-to-last step we again used the tail bound in (4.46) and re-arranged. Adding up the three partitions, we obtain that for the thick-tails case

$$\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i\mathrm{I}(|m_i^*|<\beta_0) = [c_4+o_\mathrm{P}(1)]\,\mathrm{I}(|m_i^*|<\beta_0).$$

In the opposite case where $\beta_0 < |\hat{m}_i^*|$, we skip many steps already derived before to

express

$$\begin{aligned}\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} & w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq \\ & \leq \tilde{c}_3 \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} \exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\ & = \tilde{c}_3[\mathrm{P}(|\beta_i| > \max\{\beta_0, |\hat{m}_i|\} \mid y_i) + o(1)] \\ & = \tilde{c}_3[\mathrm{P}(|\beta_i| > |\hat{m}_i| \mid y_i, |\hat{m}_i| > \beta_0) + o_\mathrm{P}(1)] \\ & = \tilde{c}_3\left[(1/2 + o_\mathrm{P}(1)) + o_\mathrm{P}(1)\right] \\ & = \tilde{c}_3/2 + o_\mathrm{P}(1). \end{aligned} \tag{4.58}$$

The other two partitioned integrals, we merge them back in this scenario

$$\begin{aligned}\int_{|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i & \leq \int_{|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\ & = c_4\left[\mathrm{P}(|\beta_i| < \max\{|\hat{m}_i|, \beta_0\} \mid y_i) + o(1)\right] \\ & = c_4\left[\underbrace{\mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, \beta_0 < |m_i^*|)}_{=1/2+o_\mathrm{P}(1)} + o_\mathrm{P}(1)\right] \\ & = c_4/2 + o_\mathrm{P}(1), \end{aligned} \tag{4.59}$$

These last two results provide

$$\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i(1 - \mathrm{I}(|m_i^*| < \beta_0)) = \left[\frac{c_4 + \tilde{c}_3}{2} + o_\mathrm{P}(1)\right](1 - \mathrm{I}(|m_i^*| < \beta_0)).$$

Hence, letting $c_\pi^{-1} := \max\{c_4, \tilde{c}_3\}$, we obtain that for the thick tails case

$$\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq c_\pi^{-1} + o_\mathrm{P}(1).$$

Let us turn now to the thin tails case. Recover the scenario in which $\beta_0 > |\hat{m}_i|$, and the partition in (4.47). Again, we cover the three parts separately. Given that

$\pi_{\mathrm{C}}(\beta_i)/\pi_{\mathrm{N}}(\beta_i) < 1/\pi_{\mathrm{N}}(\beta_i)$, we know that

$$
\begin{aligned}
&\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&\leq \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} c_2 \exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_2 \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} (2\pi s_i)^{-1/2}\exp\left\{-\frac{1}{2s_i}\left[(\beta_i-\hat{m}_i)^2 - \frac{\beta_i^2}{\tau n+1}\right]\right\}\mathrm{d}\beta_i \\
&= c_2\left[\mathrm{P}(|\beta_i| > \max\{|\hat{m}_i|,\beta_0\} \mid y_i) + o(1)\right] \\
&= c_2\left[\mathrm{P}(|\beta_i| > \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) + o_{\mathrm{P}}(1)\right] \\
&= c_2\left[P\left(Z > \frac{\beta_0-\hat{m}_i}{\sqrt{s_i}} \mid y_i, |\hat{m}_i| < \beta_0\right) + P\left(Z > \frac{\beta_0+\hat{m}_i}{\sqrt{s_i}} \mid y_i, |\hat{m}_i| < \beta_0\right) + o_{\mathrm{P}}(1)\right] \\
&\leq c_2\left[O_{\mathrm{P}}\left(e^{-\frac{n}{2}\frac{(\beta_0-m_i^*)^2}{\phi_i}}\right) + O_{\mathrm{P}}\left(e^{-\frac{n}{2}\frac{(\beta_0+m_i^*)^2}{\phi_i}}\right) + o_{\mathrm{P}}(1)\right] \\
&= c_2 o_{\mathrm{P}}(1).
\end{aligned} \tag{4.60}
$$

Then

$$
\begin{aligned}
&\int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq \int_{|\beta_i|<\min\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P}_{\mathrm{N}}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_4\left[\mathrm{P}(|\beta_i| < \min\{|\hat{m}_i|,\beta_0\} \mid y_i) + o(1)\right] \\
&= c_4\left[\underbrace{\mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, |\hat{m}_i| < \beta_0)}_{=\mathrm{P}(|\beta_i|<\hat{m}_i|y_i,0<\hat{m}_i<\beta_0)+o_{\mathrm{P}}(1)} + o_{\mathrm{P}}(1)\right] \\
&= c_4\left[\mathrm{P}(|\beta_i| < \hat{m}_i \mid y_i, 0 < \hat{m}_i < \beta_0) + o_{\mathrm{P}}(1)\right] \\
&= c_4\left[\mathrm{P}(\beta_i < \hat{m}_i \mid y_i, 0 < \hat{m}_i < \beta_0) - \underbrace{\mathrm{P}(\beta_i < -\hat{m}_i \mid y_i, 0 < \hat{m}_i < \beta_0)}_{=\mathrm{P}(Z>2m_i^*/\sqrt{s_i})+o_{\mathrm{P}}(1)} + o_{\mathrm{P}}(1)\right] \\
&\leq c_4\left[1/2 - O_{\mathrm{P}}\left(n^{-1/2}e^{-\frac{n}{2}\frac{(2m_i^*)^2}{\phi_i}}(1-1/n)\right) + o_{\mathrm{P}}(1)\right] \\
&= c_4/2 + o_{\mathrm{P}}(1),
\end{aligned} \tag{4.61}
$$

where the fourth step follows from the fact that $\hat{m}_i \to m_i^* > 0$, and thus asymptotically

$\hat{m}_i > 0$. Finally

$$
\begin{aligned}
&\int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq \\
&\leq \int_{\min\{|\hat{m}_i|,\beta_0\}<|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_4\left[\mathrm{P}(|\beta_i| < \max\{|\hat{m}_i|,\beta_0\} \mid y_i) - \mathrm{P}(|\beta_i| < \min\{|\hat{m}_i|,\beta_0\} \mid y_i) + o(1)\right] \\
&= c_4\left[\mathrm{P}(|\beta_i| < \beta_0 \mid y_i, |\hat{m}_i| < \beta_0) - \mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, |\hat{m}_i| < \beta_0) + o_\mathrm{P}(1)\right] \\
&= c_4\left[\mathrm{P}(|\beta_i| < \beta_0 \mid y_i, 0 < \hat{m}_i < \beta_0) - \underbrace{\mathrm{P}(|\beta_i| < \hat{m}_i \mid y_i, 0 < \hat{m}_i < \beta_0)}_{=1/2+O_\mathrm{P}\left(e^{-\frac{n}{2}\frac{(2m_i^*)^2}{\phi_i}}\right)+o_\mathrm{P}(1)} + o_\mathrm{P}(1)\right] \\
&= c_4\left[(1-\mathrm{P}(|\beta_i| > \beta_0 \mid y_i, 0 < \hat{m}_i < \beta_0)) - (1/2 + o_\mathrm{P}(1)) + o_\mathrm{P}(1)\right] \\
&= c_4\left[1/2 - (\mathrm{P}(\beta_i > \beta_0 \mid y_i, 0 < \hat{m}_i < \beta_0) - \mathrm{P}(\beta_i < -\beta_0 \mid y_i, 0 < \hat{m}_i < \beta_0)) + o_\mathrm{P}(1)\right] \\
&\leq c_4\left[1/2 - O_\mathrm{P}\left(n^{-1/2}e^{-\frac{n}{2}\frac{(\beta_0-m_i^*)^2}{\phi_i}}(1-1/n)\right) + O_\mathrm{P}\left(e^{-\frac{n}{2}\frac{(\beta_0+m_i^*)^2}{\phi_i}}\right) + o_\mathrm{P}(1)\right] \\
&= c_4/2 + o_\mathrm{P}(1). \qquad (4.62)
\end{aligned}
$$

Thus, results (4.60), (4.61) and (4.62) provide that also in the thin-tailed case

$$\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \mathrm{I}(|m_i^*| < \beta_0) = [c_4 + o_\mathrm{P}(1)]\,\mathrm{I}(|m_i^*| < \beta_0).$$

We only have left to look at the scenario where $|m_i^*| > \beta_0$ for thin-tails.

$$
\begin{aligned}
&\int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq \\
&\qquad \leq \int_{|\beta_i|>\max\{|\hat{m}_i|,\beta_0\}} c_2 \exp\left\{\frac{1}{2}\frac{\beta_i^2}{\tau\phi_i}\right\}\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&\qquad = \tilde{c}_2[\mathrm{P}(|\beta_i| > \max\{\beta_0, |\hat{m}_i|\} \mid y_i) + o(1)] \\
&\qquad = \tilde{c}_2[\mathrm{P}(|\beta_i| > |\hat{m}_i| \mid y_i, |\hat{m}_i| > \beta_0) + o_\mathrm{P}(1)] \\
&\qquad = \tilde{c}_2[(\mathrm{P}(|\beta_i| > \hat{m}_i \mid y_i, |\hat{m}_i| > \beta_0, \hat{m}_i > 0) + o_\mathrm{P}(1)) + o_\mathrm{P}(1)] \\
&\qquad = \tilde{c}_2\left[1/2 o_\mathrm{P}(1)\right] \\
&\qquad = \tilde{c}_2/2 + o_\mathrm{P}(1). \qquad (4.63)
\end{aligned}
$$

We merge the other two integrals analogously to the thick-tailed case.

$$
\begin{aligned}
&\int_{|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq \int_{|\beta_i|<\max\{|\hat{m}_i|,\beta_0\}} c_4\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \\
&= c_4\left[\mathrm{P}(|\beta_i| < \max\{|\hat{m}_i|,\beta_0\} \mid y_i) + o(1)\right] \\
&= c_4\left[\mathrm{P}(|\beta_i| < |\hat{m}_i| \mid y_i, \beta_0 < |m_i^*|) + o_\mathrm{P}(1)\right] \\
&= c_4\left[\mathrm{P}(|\beta_i| < \hat{m}_i \mid y_i, \hat{m}_i > 0, \beta_0 < |m_i^*|) + o_\mathrm{P}(1)\right] \\
&= c_4\left[\mathrm{P}(\beta_i < \hat{m}_i \mid y_i, \hat{m}_i > 0, \beta_0 < |m_i^*|) - \mathrm{P}(\beta_i < -\hat{m}_i \mid y_i, \hat{m}_i > 0, \beta_0 < |m_i^*|) + o_\mathrm{P}(1)\right] \\
&\leq c_4\left[1/2 - O_\mathrm{P}\left(n^{-1/2}e^{-\frac{n}{2}\frac{(2m_i^*)^2}{\phi_i}}(1-1/n)\right) + o_\mathrm{P}(1)\right] \\
&= c_4/2 + o_\mathrm{P}(1),
\end{aligned}
\tag{4.64}
$$

These last two results provide

$$
\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i(1-\mathrm{I}(|m_i^*| < \beta_0)) = \left[\frac{c_2+c_4}{2} + o_\mathrm{P}(1)\right](1-\mathrm{I}(|m_i^*| < \beta_0)).
$$

Hence, letting $c_\pi^{-1} := \max\{c_4, c_2\}$, we obtain that for the thick tails case

$$
\int w_\pi(\beta_i)\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i \leq c_\pi^{-1} + o_\mathrm{P}(1).
$$

In both cases, the result is the same, namely that under the conditions stablished in the Part (i) of the proof, we have that

$$
\begin{aligned}
B_{i,01}^\pi = B_{i,01}^\mathrm{N}\left[\int \frac{\pi(\beta_i)}{\mathrm{N}(\beta_i;0,\tau\phi_i)}\mathrm{P_N}(\beta_i \mid y_i)\mathrm{d}\beta_i\right]^{-1} \leq \\
\leq B_{i,01}^\mathrm{N}[c_\pi^{-1} + o_\mathrm{P}(1)]^{-1} = c_\pi B_{i,01}^\mathrm{N} + o_\mathrm{P}(1),
\end{aligned}
$$

which completes the proof of Part (ii). □

Remaining proofs

Proof. Corollary 4.4. For the case where $\beta_i^* = 0$, by Theorem 4.3 we have that for every parameter β_i, π satisfies

$$
\begin{aligned}
\mathrm{P}_\pi(\gamma_i = 1 \mid y_i) &= \left(1 + \frac{1-\rho}{\rho}B_{i,01}^\pi\right)^{-1} \\
&\geq \left(1 + \frac{1-\rho}{\rho}\tilde{c}_\pi B_{i,01}^N + \frac{1-\rho}{\rho}o_\mathrm{P}(1)\right)^{-1} \\
&\geq \left(1 + \frac{1-\rho}{\rho}\tilde{c}_\pi B_{i,01}^N\right)^{-1}.
\end{aligned}
$$

Thus, since $\tilde{c}_\pi$ is a prior-dependent constant, we find that no individual $\mathrm{P}_\pi(\gamma_i = 1 \mid y_i)$ improves the probability bounds of their Gaussian counterpart, and so $\max_{i:\beta_i^*=0} \mathrm{P}_\pi(\gamma_i = 1 \mid y_i)$ cannot improve the rates of $\max_{i:\beta_i^*=0} \mathrm{P}_\mathrm{N}(\gamma_i = 1 \mid y_i)$ presented in Theorem 4.1, given that the estimation of each parameter β_i is decoupled from each other.

For the case where $\beta_i^* \neq 0$, we similarly write

$$\begin{aligned}\mathrm{P}_\pi(\gamma_i = 0 \mid y_i) &= 1 - \mathrm{P}_\pi(\gamma_i = 1 \mid y_i) \\ &\geq 1 - \left(1 + \frac{1-\rho}{\rho} c_\pi B_{i,01}^N + \frac{1-\rho}{\rho} o_\mathrm{P}(1)\right)^{-1},\end{aligned}$$

by Theorem 4.3. Note that under assumptions (C1), (C3) and (C7) combined (assumed by Theorem 4.1) we have that $\frac{1-\rho}{\rho} = o_\mathrm{P}(1)$, and so the last term inside the sum of this equation is entirely $o_\mathrm{P}(1)$. Then, by the same argument as for the $\beta_i^* = 0$ case, and given that c_π is a prior-dependent constant, we find that $\max_{i:\beta_i^*\neq 0} \mathrm{P}_\pi(\gamma_i = 0 \mid y_i)$ cannot improve the probability bounds of $\max_{i:\beta_i^*\neq 0} \mathrm{P}_\mathrm{N}(\gamma_i = 0 \mid y_i)$ introduced in Theorem 4.1. □

Proof. Corollary 4.5. The first statement follows directly combining the probability bounds established in Corollary 4.4 with the assumption $f_\pi \to 1$, and then replicating the Proof for Theorem 4.2, Part (i), simply replacing terms.

For the second statement, note that if π satisfies the conditions in Theorem 4.3, the O_P-rates for the posterior inclusion probabilities are not better than the Gaussian, up to an $o_\mathrm{P}(1)$ factor. Re-taking the notation from the proof for Theorem 4.2, Since by assumption we know that $|f_\pi(\min_{i:\beta_i^*\neq 0} Z_i) - 1| = O_\mathrm{P}(s_n)$, we can use expression (4.42) and replicate the steps of the proof for the Gaussian case. In that case, $s_n = (\tau n)^{-1}$, whose O_P-rates are expressed in (4.43), and so the result stated in Corollary 4.5 follows from there. □

Conclusions and Further Work

In this project, I have analysed the problem of high-dimensional inference mainly from the perspective of treatment effect estimation, to which I have tried to address a number of problems that commonly appear in this context, in particular those related to under- and over-selection, multiple treatments, treatment effect heterogeneity, and computation.

There are two main ingredients in the CIL proposal. First, learning from data whether and to what extent control inclusion or exclusion should be encouraged to improve multiple treatment inference, which includes learning the overall degree of sparsity. Second, a computational strategy to render the approach practical. This is in contrast to other literature, which either imposes a sparsity assumption that results in under-selection biases under strong confounding, or encourages the inclusion of certain controls to avoid under-selection but can run into serious over-selection issues, as it has been illustrated. The CIL framework learns the relative importance of each potential confounders, which helps bypass said over- and under-selection issues. By learning the relative importance of potential confounders, as the CIL framework proposes, one may bypass this problem.

These issues are practically relevant, e.g. in the salary data one may fail to detect small but relevant associations between e.g. black race and salary in many particular states. Further, the proposed Bayesian framework naturally allows for posterior predictive inference on functions that depend on multiple parameters, such as the variation in salary jointly associated with multiple treatments to measure overall discrimination. Interestingly, our analyses revealed that association between salary and discriminatory factors such as gender or race in 2019 relative to 2010 has practically not been reduced, as well as the heterogeneity across states. These results are conditional on controls that include education, employment and other characteristics that affect salary. That is, our results reveal similar salary discrepancies in 2019 between races/genders, provided that two individuals have the same characteristics (and that they were hired in the first place). This analysis offers a complementary view to analyses that are unadjusted by controls, and which may reveal equally interesting information. For example, if females migrated towards higher-paying occupational sections in 2019 and received a higher salary as a consequence, this would not be detected by our analysis, but would be revealed by an

unadjusted analysis.

To keep our exposition simple, I used to the term *confounders* generally to refer to potentially confounding variables one controls for, although in many practical applications one should differentiate between *confounders* and *mediators*, both of which need to be adjusted for. The difference between the two lies in the fact that a mediator variable may be affected by the treatment in the first place, and then reinforce the observed treatment as a result. For example, if because of a socioeconomic background a given *treatment* leads to lower levels of education on average, which then cause lower salary levels, then the education variable is a *mediator*. Contrarily, a confounder will trigger the treatment without necessarily engaging feedback effects. However, this distinction should have no major impact in the analysis here presented as the goal of the exercise was to find out if any degree of discrimination remains, even after accounting for education.

Also for simplicity, I focused the discussion on linear treatment effects, but it is possible to extend the framework to non-linear effects and interactions between treatments, e.g. via splines or other suitable non-linear bases. Incorporating non-linearities, complex interactions among treatment and control variables, and treatment effect heterogeneity is actually a relevant line of research to adapt the CIL machinery. For example, how to include interactions on a general problem, given that the size of the set of potential interactions is much larger than the number of controls itself. The ones in our salary example are specified a priori, but this need not be the general case. Additionally, in the case of interactions the CIL prior has been simplified here by assigning the same hyper-parameter to every interaction of a main effect, so as to avoid major inflation of computational costs. In the presence of very large of interaction sets it may be desirable to look for a more general method. Similarly, the method should be adjusted to account for non-linearities, i.e. for covariates being associated to several columns in the design matrix. It could be worth investigating how the design matrix can be re-parameterised to that end.

A further interesting line of research, in both Chapter 2 and 4, would be to study further specifications of non-local priors, and investigate whether their properties can match or improve those of the pMOM prior that I dealt with, when analytically feasible. This remains an open end for future work.

Bibliography

Antonelli, J., Cefalu, M., Palmer, N., and Agniel, D. (2018). Double robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179.

Antonelli, J., Parmigiani, G., and Dominici, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, 14(3):805–828.

Antonelli, J., Zigler, C., and Dominici, F. (2017). Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics*, 18(3):553–568.

Athey, S., Imbens, G., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:597–623.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81:608–650.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624. PMID: 23482517.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

Cefalu, M., Dominici, F., Arvold, N., and Parmigiani, G. (2017). Model averaged double robust estimation. *Biometrics*, 73(2):410–421.

Chernozhukov, V., Hansen, C., and Spindler, M. (2016). hdm: High-dimensional metrics. *R Journal*, 8(2):185–199.

Chernozhukov, V., Newey, W. K., and Singh, R. (2018). Automatic debiased machine learning of causal and structural effects. In *arXiv:1809.05224*.

Clyde, M. A. and Ghosh, J. (2012). Finite population estimators in stochastic search variable selection. *Biometrika*, 99(4):981–988.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2021). Adaptive robust estimation in sparse vector model. *The Annals of Statistics*, 49(3):1347–1377.

Dezeure, R., Bhlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and `R`-software `hdi`. *Statistical Science*, 30(4):533–558.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):41–81.

Dukes, O. and Vansteelandt, S. (2019). Uniformly valid confidence intervals for conditional treatment effects in misspecified high-dimensional models. In *arXiv:1410.2597v4*.

Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, 136(C):4–9.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Ertefaie, A., Asgharian, M., and Stephens, D. A. (2018). Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, 6(1):20170010.

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Methodological)*, 70(5):849–911.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189:1–23.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Fithian, W., Sun, D. L., and Taylor, J. (2017). Optimal inference after model selection. In *arXiv:1410.2597v4*.

Flood, S., King, M., Rodgers, R., Ruggles, S., and Warren, J. R. (2020). Integrated public use microdata series, current population survey: Version 8.0 [dataset]. https://doi.org/10.18128/D030.V8.0. Minneapolis, MN: IPUMS. Accessed: 2021-01-13.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, 16(4):499–511.

Gatu, C. and Kontoghiorghes, E. J. (2006). Branch-and-bound algorithms for computing the best-subset regression models. *Journal of Computational and Graphical Statistics*, 15(1):139–156.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

Ghosh, D., Zhu, Y., and Coffman, D. L. (2015). Penalized regression procedures for variable selection in the potential outcomes framework. *Statistics in Medicine*, 34:1645–1658.

Giannone, D., Lenza, M., and Primiceri, G. (2021). Economic predictions with Big Data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Hahn, P., Murray, J., and Carvalho, C. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163–182.

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20:221–229.

Hastie, T., Tibshirani, R., and Wainwright, M. (2016). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14(23):1891–1945.

Jacobi, L., Wagner, H., and Frühwirth-Schnatter, S. (2016). Bayesian treatment effects models with variable selection for panel outcomes with an application to earnings effects of maternity leave. *Journal of Econometrics*, 193(1):234–250.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, 72(2):143–170.

Johnson, V. E. and Rossell, D. (2012). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the American Statistical Association*, 24(498):649–660.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Ann. Statist*, 32:1594–1649.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lefebvre, G., Atherton, J., and Talbot, D. (2014). The effect of the prior distribution in the bayesian adjustment for confounding algorithm. *Computational Statistics & Data Analysis*, 70:227–240.

Ma, S., Zhu, L., Zhang, Z., Tsai, C.-L., and Carroll, R. J. (2019). A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *The Annals of Statistics*, 47(3):1505 – 1535.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Minka, T. P. (2001a). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Minka, T. P. (2001b). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, USA. AAI0803033.

Moran, G. E., Ročková, V., and George, E. I. (2018). On variance estimation for bayesian variable selection. In *arXiv:1801.03019*.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Orben, A. and Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 3(2):173–182.

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

Patel, C., Pho, N., Mcduffie, M., Easton-Marks, J., Kothari, C., Kohane, I., and Avillach, P. (2016). A database of human exposomes and phenomes from the us national health and nutrition examination survey. *Scientific Data*, 3(1).

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika*, 101(2):285–302.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 10(2):159–203.

Reid, S., Tibshirani, R., and Friedman, J. (2013). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67.

Rossell, D. (2021). Concentration of posterior model probabilities and normalized l0 criteria. *Bayesian Analysis*, 1(1):1–27.

Rossell, D., Abril, O., and Bhattacharya, A. (2021). Approximate laplace approximations for scalable model selection. *Journal of the Royal Statistical Society: Series B (Methodological)*, (in press).

Rossell, D., Cook, J. D., Telesca, D., Roebuck, P., Abril, O., and Torrens i Dinarès, M. (2022). *mombf: Bayesian Model Selection and Averaging for Non-Local and Local Priors*. R package version 3.1.2.

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265. PMID: 29881129.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Seeger, M., Gerwinn, S., and Bethge, M. (2007). Bayesian inference for sparse generalized linear models. In Kok, J. N., Koronacki, J., Mantaras, R. L. d., Matwin, S., Mladenič, D., and Skowron, A., editors, *Machine Learning: ECML 2007*, pages 298–309, Berlin, Heidelberg. Springer Berlin Heidelberg.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2020). Estimation and inference on heterogeneous treatment effects in high-dimensional dynamic panels. In *arXiv:1712.09988v4*.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable bayesian variable selection using non-local prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053–1078.

Shortreed, S. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Tadesse, M. G. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman and Hall/CRC.

Talbot, D., Lefebvre, G., and Atherton, J. (2014). *BACprior: Choice of the Hyperparameter Omega in the Bayesian Adjustment for Confounding (BAC) Algorithm*. R package version 2.0.

Talbot, D., Lefebvre, G., and Atherton, J. (2015). The bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2):207–236.

Tibshirani, R. (1996). Regression shrinkage selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe. *Bayesian Analysis*, 12(4):1221–1274.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741.

Wang, C., Dominici, F., Parmigiani, G., and Zigler, C. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models: Accounting for uncertainty in confounder and effect modifier selection when estimating aces in glms. *Biometrics*, 71(3):654–665.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–686.

Wang, X. (2016). *Distributed Feature Selection in Large n and Large p Regression Problems*. PhD thesis, Duke University.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 78(3):589–611.

Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861.

Wilson, A., Zigler, C., Patel, C., and Dominici, F. (2018). Model-averaged confounder adjustment for estimating multivariate exposure effects with linear regression: Model-averaged confounder adjustment for estimating multivariate exposure effects. *Biometrics*, 74(3):1034–1044.

Wu, H.-H. (2016). *Nonlocal priors for Bayesian variable selection in generalized linear models and generalized linear mixed models and their applications in biology data*. PhD thesis, University of Missouri–Columbia.

Xu, M., Lakshminarayanan, B., Teh, Y. W., Zhu, J., and Zhang, B. (2014). Distributed bayesian posterior sampling via moment sharing. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3356–3364, Cambridge, MA, USA. MIT Press.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76(1):217–242.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

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