Moment conditions and Bayesian nonparametrics*

Luke Bornn
Department of Statistics, Harvard University and
Department of Statistics and Actuarial Science, Simon Fraser University
bornn@fas.harvard.edu

Neil Shephard
Department of Economics and Department of Statistics, Harvard University
shephard@fas.harvard.edu

Reza Solgi
Department of Statistics, Harvard University
rezasolgi@fas.harvard.edu

January 13, 2016

Abstract

Models phrased though moment conditions are central to much of modern inference. Here these moment conditions are embedded within a nonparametric Bayesian setup. Handling such a model is not probabilistically straightforward as the posterior has support on a manifold. We solve the relevant issues, building new probability and computational tools using Hausdorff measures to analyze them on real and simulated data. These new methods which involve simulating on a manifold can be applied widely, including providing Bayesian analysis of quasi-likelihoods, linear and nonlinear regression, missing data and hierarchical models.

Keywords: Decision theory; Empirical likelihood; Hausdorff measure; Markov chain Monte Carlo; Method of moments; Nonparametric Bayes; Simulation on manifolds.

1 Introduction

1.1 Overview

Much of modern inference is phrased in terms of moment conditions and analyzed using asymptotic approximations. Here we build a new methodology which dovetails with decision theory. Moment conditions are embedded within a nonparametric Bayesian setup, allowing an individual to mix moment conditions with data and scientifically informative priors to make rational decisions without the recourse to the veil of parametric assumptions or asymptotics.

Embedding moments within nonparametrics is not probabilistically straightforward. This paper spells out the issues, develops the corresponding probability theory to solve them and devises novel

*We thank Isaiah Andrews, Yang Chen, Herman van Dijk, Mikkel Plagborg-Moller and Christian Robert for their comments on an earlier draft.

strategies for simulating on a manifold to implement them in practice on simulated and real data. It covers the case where it is hard, or indeed impossible, to solve the moment equations. This allows the rational analysis of moment condition models with many solutions.

The scope of the new methods is vast. It deals with, for example, linear, nonlinear and instrumental variable regression. By thinking of the moment condition as the score of a parametric statistical model, our analysis also provides a Bayesian treatment of quasi-likelihood methods which are widely applied in statistics (e.g. Cox (1961), White (1994)). Finally, this framework provides a solid basis to deal systematically with missing data (e.g. Little and Rubin (2002)), shrink parameters (e.g. Efron (2012)) and build hierarchical models (e.g. Gelman et al. (2003)).

1.2 The conceptual challenge

It will be helpful in our discussion of the paper's contribution and to place it in the context of the literature to establish some notation; a formal statement will appear in Section 2.

Assume one has independent and identically distributed (i.i.d.) d-dimensional data Z_i, $i = 1, 2, ..., n$, taking on the known support $s_1, s_2, ..., s_J$ and having distribution function F. We then write $\mathbb{P}(Z_i = s_j|\theta, \beta) = \theta_j$ where the p-dimensional β satisfies the r-dimensional moment condition

$$\mathbb{E}_Z \{g(Z, \beta)\} = \int g(z, \beta) F(\mathrm{d}z) = \sum_{j=1}^{J} \theta_j g(s_j, \beta) = 0. \tag{1}$$

Here β is the parameter of scientific interest. We then view $\theta = (\theta_1, \theta_2, ..., \theta_{J-1})'$ (with $\theta_J = 1 - \iota'\theta$, where ι is a vector of ones of appropriate size) as nuisance parameters to be treated nonparametrically. The task is to learn $p(\beta, \theta|Z)$ or $p(\beta|Z)$, where $Z = (Z_1, Z_2, ..., Z_n)'$. A simple example of this is $g(s_j, \beta) = s_j - \beta$ which delivers the mean.

Although this problem is easy to state, it is not easily carried through, as traditional nonparametric models clash with the moment conditions, in effect overspecifying the model. Expressing this in a different way: the prior and posterior for β, θ are typically supported on a zero Lebesgue measure $(J + p - 1 - r)$-dimensional set, $\Theta_{\beta,\theta}$, in $\mathbb{R}^{J+p-1}$. As a result, traditional Markov chain Monte Carlo (MCMC) methods (or alternatives like importance sampling) for sampling from $p(\beta, \theta|Z)$ entirely collapse. This paper solves this problem in two different ways: the comparative advantages of each will depend upon the form of the moment conditions. Taken overall this paper provides a unified solution to this central problem.

1.3 Literature on classical analysis of moments

Before we detail our new approach, we will discuss how this work relates to the literature.

Moment based estimation was introduced by Pearson (1894). A relatively modern version of this procedure first estimates $\widehat{\theta}$ nonparametrically, that is F by the empirical distribution function F_n, and then plugs it into (1), yielding the function

$$\int g(z,\beta)F_n(\mathrm{d}z) = \sum_{j=1}^{J} \widehat{\theta}_j g(s_j,\beta).$$

In the $p = r$ case we move β around until this function equals a vector of zeros, delivering the method of moments estimator $\widehat{\beta}$. Extensions include, for example, Sargan (1958, 1959), Durbin (1960), Godambe (1960), Wedderburn (1974), McCullagh and Nelder (1989), Hansen (1982), Chamberlain (1987), Hansen et al. (1996), Gallant and Tauchen (1996) and Gourieroux et al. (1993). Hall (2005) gives a recent review.

An elegant implementation of moment based inference is through empirical likelihood. Motivated by Owen (1988, 1990), Qin and Lawless (1994) and Imbens et al. (1998) discussed empirical likelihood based inference in overidentified moment condition models. See also the reviews by Owen (2001), Kitamura (2007) and Lancaster and Jun (2010).

1.4 Literature on Bayesian analysis of moments

Our work is fully Bayesian. Much of our work has been inspired by Chamberlain (1987) and in particular Chamberlain and Imbens (2003). Chamberlain and Imbens (2003) place a Dirichlet prior on θ, which implies the posterior on θ is Dirichlet. These priors and posteriors are straightforward to sample from as noticed by Rubin (1981) in his Bayesian bootstrap. Chamberlain and Imbens (2003) suggest that for each posterior draw of θ they would solve the moment conditions to imply a value (or in principle a set of values) of β. Collecting a sample of such solved values provides a sample from a posterior on β. Unfortunately these authors have no control over the prior for β, the parameter of scientific interest.

Also important is Kitamura and Otsu (2011), who have two methods, both expressed in terms of Dirichlet process priors. Here we convert them into our finite framework. In their exponentially tilted case they first specify a prior $p(\beta)p(\theta)$ before finding $\theta^* = (\theta_1^*, \theta_2^*, ..., \theta_J^*)$ which minimizes $\sum_{j=1}^{J} \theta_j^* \log\left(\frac{\theta_j^*}{\theta_j}\right)$ subject to the moment constraints $\sum_{j=1}^{J} \theta_j^* g(s_j,\beta) = 0$ and the probability axioms. They then set $\mathbb{P}(Z_i = s_j|\theta,\beta) = \theta_j^*$, using this model to learn β and θ from the data. Shin (2014) carefully investigates various computational aspects of this approach. This approach has many advantages but it leaves pairs of β and θ with positive posterior probability which are not logically compatible. Kitamura and Otsu (2011) also propose a synthetic Dirichlet process (with connections to Doss (1985) and Newton et al. (1996)).

There are also many papers which provide alternative methods, including a substantial literature on the Bayesian use of moments through approximate methods. Chernozhukov and Hong (2003) specify a quadratic form in the moment conditions and use this as the basis of a log quasi-likelihood function. They then use this approximate likelihood to carry out Bayesian inference using MCMC alongside a sandwich estimator. Related work includes Yin (2009). Muller (2013) provides a Bayesian version of the asymptotic sandwich matrix commonly seen in quasi-likelihood inference and links it to decision theory.

Lazar (2003), Schennach (2005) and Yang and He (2012) provide Bayesian interpretations to empirical likelihood and study the resulting properties. Mengersen et al. (2013) look at moment conditions and empirical likelihood using approximate Bayesian computation. See also Zellner (1997) and Zellner et al. (1997), who suggested a Bayesian moment method by building a likelihood defined through the maximum entropy density consistent with the moment conditions. Related is the Bayesian work on factor and cointegration models, e.g. Strachan and van Dijk (2004).

In a series of papers Gallant and Hong (2007), Gallant et al. (2014) and Gallant (2015) develop methods which devise a prior using fiducial arguments from moment conditions. Related work includes Jaynes (2003) and Kwan (1998). Florens and Simoni (2015) have used Gaussian processes in combination with moment constraints to carry out Bayesian inference.

1.5 Computational issues

Here the prior and posterior for β, θ are supported on a zero Lebesgue measure $(J + p - 1 - r)$-dimensional set, $\Theta_{\beta,\theta}$, in $\mathbb{R}^{J+p-1}$. Hence Bayesian inference will need us to sample from a distribution defined on a zero measure set, rendering standard Monte Carlo methods useless.

In an influential paper Gelfand et al. (1992) use MCMC methods to deal with constrained parameter spaces, but in their paper the constraints do not change the dimension of the support. Hurn et al. (1999) carry out MCMC in constrained parameter spaces (sampling from a distribution $\pi(x)$ subject to a constraint $C(x) = 0$) using block updating. Golchi and Campbell (2014) carry out sampling subject to constraints using sequential Monte Carlo methods by slowly introducing the constraints. However, they do not explore the change of measure issue we discuss here. Chiu (2008) use a singular normal distribution in posterior updating for an under-identified hierarchical model. Related work includes Sun et al. (1999). Overspecified factor models also have some of these features, as discussed by West (2003). Fiorentini et al. (2004) face related but highly specialized challenges when sampling missing data in a GARCH model.

There are few recent papers on MCMC simulation from distributions defined on manifolds.

Brubaker et al. (2012) propose a Hamiltonian Monte Carlo on implicitly defined manifolds. Numeric integration of the Hamiltonian dynamics requires solving a system of $3d$ nonlinear equations for each update, where d is the dimension of the space in which the manifold is embedded (in our setting $d = J + p - 1$). Byrne and Girolami (2013) introduce a Hamiltonian Monte Carlo simulation algorithm for sampling from manifolds with known geodesic structure. They demonstrate how this algorithm can be used in order to sample from the distributions defined on hyperspheres and Stiefel manifolds of orthonormal matrices. Diaconis et al. (2013) provide a short review of concepts in geometric measure theory. They discuss algorithms for sampling from distributions defined on Riemannian manifolds that are similar to the "marginal method" that will be introduced shortly. It is this paper which has been the most helpful to us in terms of Monte Carlo methods.

1.6 Outline of the paper

In the next section of the paper we will introduce the formal model under study, and discuss how one specifies meaningful prior distributions on the parameters of interest. In Section 3 several methods for inference and their relative merits and pitfalls are discussed. Section 4 discusses mechanisms for generating priors for these models. We also draw out how to make inference when the support of the data is unknown, regarding the unseen support as missing. This is followed by Section 5 in which some illustrative examples are demonstrated. Section 6 explores several empirical studies before Section 7 concludes. An Appendix collects the proofs of the propositions stated in the paper and a collection of additional results.

2 Bayesian moment conditions models

2.1 The model

Assume the data we have available to make inference is $Z = (Z_1, ..., Z_n)$, where the Z_i are d-dimensional i.i.d. draws from an unknown distribution which has J points of known support $\{s_1, s_2, ..., s_J\} = S$ (we relax this known support condition in Section 3.6). Throughout we write

$$\mathbb{P}(Z_i = s_j|\theta, \beta) = \theta_j, \quad j = 1, 2, ..., J, \tag{2}$$

with $\theta = (\theta_1, \theta_2, ..., \theta_{J-1})' \in \Theta_\theta \subseteq \Delta^{J-1}$, where $\Delta^{J-1} = \{\theta = (\theta_1, \theta_2, ..., \theta_{J-1})'; \ \iota'\theta < 1 \text{ and } \theta_j > 0\}$ for all j and $\theta_J = 1 - \iota'\theta$, in which ι is a vector of ones. Further, the science of the problem is characterized by the values of β which solve the r unconditional moment conditions,

$$\sum_{j=1}^{J} \theta_j g(s_j, \beta) = 0, \tag{3}$$

where $\beta \in \Theta_\beta \subseteq \mathbb{R}^p$ and $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^r$. Typically the scientific conclusions will center around inferences on β, although predictive type inference may also additionally feature θ. This paper concentrates on the case of exactly identified models ($r = p$). Appendix A.7 extends to the more general case of over and under identification at the cost of more clutter but without having to generate any new ideas.

2.2 Parameter space and prior

Throughout this paper we will think of β and θ as parameters to be learned from the data, Z. We write the $J + p - 1$ parameters

$$(\beta', \theta')' \in \Theta_{\beta,\theta},$$

where $\Theta_{\beta,\theta} \subseteq \mathbb{R}^p \times \Delta^{J-1} \subset \mathbb{R}^{J+p-1}$, as the joint support for β and θ. Each point within $\Theta_{\beta,\theta}$ is a pair (β, θ) which satisfies both the moment conditions and probability axioms. The moment conditions are:

$$H_\beta \theta + g_J = 0 \quad \text{where} \quad H_\beta = (g_1, ..., g_{J-1}) - g_J \iota',$$

in which $g_j = g(s_j, \beta)$ (for $1 \leq j \leq J$). Moreover H_β is assumed to be of full row rank (we will often suppress the dependence on β and just write H). These constraints, together with the inequalities $\theta_j \geq 0$ (for $j = 1, 2, ..., J$), implicitly define the $(J-1)$-dimensional set of parameters within $\mathbb{R}^{J+p-1}$, which will be denoted by $\Theta_{\beta,\theta}$. Hence the parameter space, $\Theta_{\beta,\theta}$, depends upon the support of the data, $S = \{s_1, ..., s_J\}$, but is not data dependent. Throughout the paper, the notation Θ_λ will generically represent the parameter space of λ in which λ is a set of parameters.

The set of admissible pairs (β, θ), denoted by $\Theta_{\beta,\theta}$, is a zero measure set (with respect to Lebesgue measure) in $\mathbb{R}^{J+p-1}$. We will assume that researchers can place a prior density, $p(\beta, \theta)$, with respect to the $(J-1)-$dimensional Hausdorff measure on $\Theta_{\beta,\theta}$. Using the Hausdorff measure[1] as the base measure, we are able to assign measures to the lower dimensional subsets of $\mathbb{R}^{J+p-1}$, and therefore we can define probability density functions with respect to Hausdorff measure on manifolds (and more complex zero Lebesgue measure sets) in an Euclidean space.

[1] Assume $E \subseteq \mathbb{R}^n$, $d \in [0, +\infty)$ and $\delta \in (0, +\infty]$. The Hausdorff premeasure of E is defined as follows,

$$\mathcal{H}^d_\delta(E) = v_m \inf_{\substack{E \subseteq \cup E_j \\ d(E_j) < \delta}} \sum_{j=1}^{\infty} \left(\frac{\text{diam}(E_j)}{2} \right)^d$$

where $v_m = \frac{\Gamma(\frac{1}{2})^d}{2^d \Gamma(\frac{d}{2}+1)}$ is the volume of the unit d-sphere, and $\text{diam}(E_j)$ is the diameter of E_j. $\mathcal{H}^d_\delta(E)$ is a nonincreasing function of δ, and the d-dimensional Hausdorff measure of E is defined as its limit when δ goes to zero, $\mathcal{H}^d(E) = \lim_{\delta \to 0^+} \mathcal{H}^d_\delta(E)$. The Hausdorff measure is an outer measure. Moreover $\mathcal{H}^n$ defined on $\mathbb{R}^n$ coincide with Lebesgue measure. See Federer (1969) for more details.

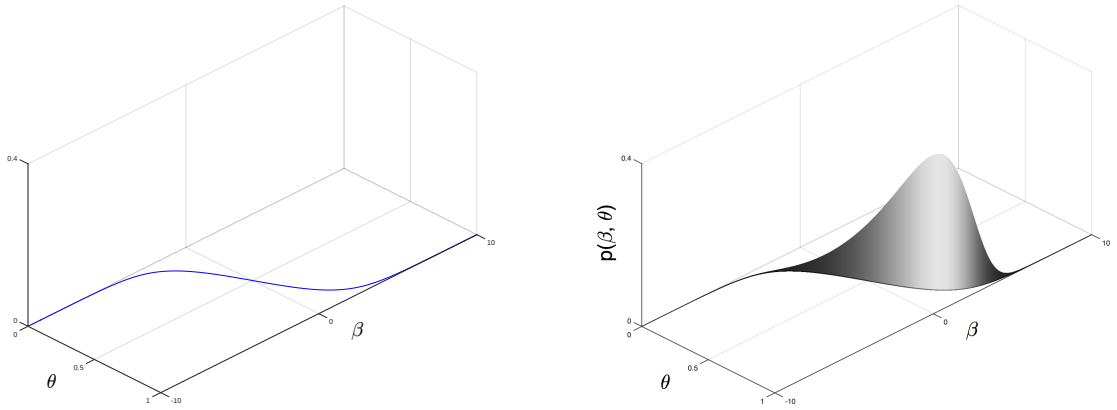

Figure 1: In the plot on the left the blue curve, $\beta = \log\left(\frac{\theta}{1-\theta}\right)$, is the parameter space of the logit model, $\Theta_{\beta,\theta}$. In the plot on the right the density of the prior $p(\beta, \theta)$ (with respect to Hausdorff measure) is depicted. This density lives on the blue curve which supports $\Theta_{\beta,\theta}$.

2.3 Some examples

To cement this we have built a starkly simple example which captures most of the challenges in this problem. It faces off a nonparametric model against a scientific parameter of interest.

Example 1 ***(Logistic)*** *Assume* $Z_1|\theta \sim \text{Bernoulli}(\theta)$*, and let* $\beta = \log\left(\frac{\theta}{1-\theta}\right) = \text{logit}(\theta)$ *be the scientific parameter of interest. Jointly* β, θ *captures the inherent singularity implicit in all moment based inference. The moment condition is*

$$g(s, \beta) = s - \frac{e^\beta}{1 + e^\beta}.$$

Therefore the parameter space, $\Theta_{\beta,\theta}$*, is*

$$\Theta_{\beta,\theta} = \left\{ (\beta, \theta) \in \mathbb{R} \times [0, 1]; \beta = \log\left(\frac{\theta}{1-\theta}\right) \right\}.$$

This is shown as the blue curve sitting at ground level in the left panel of Figure 1. Of fundamental importance is that if θ *moves by* $\mathrm{d}\theta$ *then the length of the journey along this curve will be (by Pythagoras's theorem)*

$$\mathrm{d}\theta\sqrt{1 + J_\theta^2}, \quad J_\theta = \frac{\partial \beta}{\partial \theta} = \frac{\partial \log\left(\frac{\theta}{1-\theta}\right)}{\partial \theta}.$$

The right panel of Figure 1 repeats the support but now above it is a (the form of the density is not expositionally important at this point) density $p(\beta, \theta)$ *with respect to this curve, or more formally*

the one dimension Hausdorff measure on $\Theta_{\beta,\theta}$. Then for any set $C \subset \Theta_{\beta,\theta}$,

$$\Pr\{(\beta,\theta) \in C\} = \int_{C_\theta} p(\beta,\theta)\sqrt{1+\left(\frac{\partial\beta}{\partial\theta}\right)^2}\mathrm{d}\theta,$$

where C_θ is the projection of C on θ's axis (i.e. we integrate over all values of θ which imply a β such that the pair $(\beta,\theta) \in C$). This means as we integrate over θ, we must multiply the density on the curve by the length of the curve.

We will study how to transform this prior $p(\beta,\theta)$ into a posterior and simulate from it. This will allow us to learn β from the data. As with all Bayesian calculations, it is not trivial to establish a widely acceptable prior $p(\beta,\theta)$. We will return to that very practical issue in Section 4.

Before we leave this section we give a less artful example.

Example 2 ***(Mean)*** *Let Z be a scalar random variable and $g(s,\beta) = s - \beta$, so β is a mean. Then*

$$\Theta_{\beta,\theta} = \left\{ (\beta,\theta) \ ; \ \sum_{j=1}^{J} \theta_j s_j = \beta, \ \theta_j > 0 \ \text{ for all } j, \ \text{and } \iota'\theta < 1 \right\}.$$

Thus $\Theta_{\beta,\theta}$ is a region within a $(J-1)$-dimensional hyperplane in $\mathbb{R}^J$. However all elements of this set are not admissible, since θ should satisfy the probability axioms (elements of θ should be positive and $1 - \iota'\theta > 0$). Therefore the parameter space $\Theta_{\beta,\theta}$ is a convex subset on the hyperplane. Then if θ moves by $\mathrm{d}\theta_1, ..., \mathrm{d}\theta_{J-1}$ the area of the corresponding parallelogram on the hyperplane is

$$\mathrm{d}\theta_1 ... \mathrm{d}\theta_{J-1}\sqrt{1 + J_\theta J_\theta'}, \quad J_\theta = \left\{ \left(\frac{\partial\beta}{\partial\theta_1}\right), ..., \left(\frac{\partial\beta}{\partial\theta_{J-1}}\right) \right\},$$

where $\partial\beta/\partial\theta_j = s_j - s_J$, $j = 1, 2, ..., J-1$. So for any measurable set $C \subset \Theta_{\beta,\theta}$,

$$\begin{aligned} \Pr\{(\beta,\theta) \in C\} &= \int_{C_\theta} p(\beta,\theta)\sqrt{1+\sum_{j=1}^{J-1}\left(\frac{\partial\beta}{\partial\theta_j}\right)^2}\mathrm{d}\theta = \int_{C_\theta} p(\beta,\theta)\sqrt{1+\sum_{j=1}^{J-1}(s_j - s_J)^2}\mathrm{d}\theta \\ &\propto \int_{C_\theta} p(\beta,\theta)\mathrm{d}\theta. \end{aligned}$$

where C_θ is the projection of C on θ (The last proportionality is due to the fact that the Jacobian only depends on the support of the data). Thus the linearity of the moment condition (that results in a flat parameter space $\Theta_{\beta,\theta}$) translates into a somewhat trivial multiplicative correction factor and so yields a simple relationship between $\Pr\{(\beta,\theta) \in C\}$ and $p(\beta,\theta)$.

Example 3 ***(Regression)*** *The previous example can be generalized to the family of regression models. For instance consider a linear regression model, $\mathbb{E}\left(s^{(1)}|s^{(2)}\right) = \beta' s^{(2)}$, where $s = (s^{(1)}, s^{(2)})$,*

in which $s^{(1)}$ is a scalar and $s^{(2)}$ is a d-dimensional vector, and β is a p-dimensional vector of parameters. The linear regression parameters solve the following moment condition equation,

$$\mathbb{E}\left[g(s,\beta)\right] = \mathbb{E}\left[s^{(2)}(s^{(1)} - \beta' s^{(2)})\right] = 0.$$

We can also discuss the estimation of linear regression model with instrumental variables. Assume $s = (s^{(1)}, s^{(2)}, s^{(3)})$, where $s^{(1)}$ is a scalar, and $s^{(2)}$ and $s^{(3)}$ are p-dimensional vectors (independent and instrumental variables, respectively). If we define $g(s,\beta) = s^{(3)}(s^{(1)} - \beta' s^{(2)})$, then β is the solution to $\mathbb{E}\left[g(s,\beta)\right] = 0$. Moreover generalizing to the nonlinear regression model is easy. Assume $\mathbb{E}\left(s^{(1)}|s^{(2)}\right) = \mu(s^{(2)},\beta)$. Then the corresponding moment condition equation is $g(s,\beta) = s^{(2)}\left\{s^{(1)} - \mu(s^{(2)},\beta)\right\}$. For instance for a Poisson regression $g(s,\beta) = s^{(2)}\left\{s^{(1)} - \exp(\beta' s^{(2)})\right\}$.

Example 4 ***(Average treatment effect)*** *Consider a casual inference problem with the observational data $Z_j = (X_j, Y_j, W_j)$ (for $1 \leq j \leq N$), where X_j is the K-dimensional vector of the j-th unit's background variables, Y_j is its scalar outcome variable, and W_j is the binary treatment indicator. Assuming the super-population unconfoundedness, it can be shown that (Imbens and Rubin (2015)) $\mathbb{E}_{SP}\left[Y_j(1)\right] = \mathbb{E}\left[\frac{W_j Y_j}{e(X_j)}\right]$ and $\mathbb{E}_{SP}\left[Y_j(0)\right] = \mathbb{E}\left[\frac{(1-W_j)Y_j}{1-e(X_j)}\right]$, where $e(X_j)$ is the propensity score, $e(X_j) = \eta_j = \Pr(W_j = 1|X_j)$. Therefore the average treatment effect (ATE) is*

$$\tau = \mathbb{E}_{SP}\left[Y_j(1)\right] - \mathbb{E}_{SP}\left[Y_j(0)\right] = \mathbb{E}\left(\frac{W_j Y_j}{e(X_j)} - \frac{(1-W_j)Y_j}{1-e(X_j)}\right),$$

One might use a logistic regression model for the propensity score, $\eta_j = \exp(\gamma' X_j)/\left\{1 + \exp(\gamma' X_j)\right\}$, where γ is K-dimensional. Under these assumptions the model's parameters, $\beta = (\gamma, \tau)$, solve the following set of moment conditions,

$$\mathbb{E}\left[g(Z_j,\beta)\right] = \mathbb{E}\left[\begin{array}{c} X_j(Y_j - \eta_j) \\ \frac{(W_j - \eta_j)Y_j}{\eta_j(1-\eta_j)} - \tau \end{array}\right] = 0,$$

If we assume the data points are i.i.d. realizations from a discrete distribution with finite and known support $S = \{s_1, ..., s_j\}$, $\Pr(Z_i = s_j) = \theta_j$, the moment conditions are,

$$\mathbb{E}\left[g(Z_j,\beta)\right] = \left[\begin{array}{c} \sum_{j=1}^{J} \theta_j X_j(Y_j - \eta_j) \\ \sum_{j=1}^{J} \theta_j \frac{(W_j - \eta_j)Y_j}{\eta_j(1-\eta_j)} - \tau \end{array}\right] = 0.$$

Thus the propensity scores and the ATE can be estimated jointly (e.g. McCandless et al. (2009), Zigler et al. (2013) and Zigler and Dominici (2014)).

3 Inference

3.1 Likelihood and posterior

Under the assumptions formulated above, the model's likelihood is

$$L(Z|\beta,\theta) \propto \prod_{j=1}^{J} \theta_j^{n_j},$$

where $n_j = \sum_{i=1}^{N} 1(Z_i = s_j)$. Note that although β does not appear in the likelihood explicitly, due to the constraints on β and θ, the data is informative about β.

The posterior is supported on the same set as the prior, $\Theta_{\beta,\theta}$, and may be written as

$$p(\beta,\theta|Z) \propto p(\beta,\theta) \left(\prod_{j=1}^{J} \theta_j^{n_j} \right). \tag{4}$$

The terms in (4) are easy to compute for any (β,θ) in $\Theta_{\beta,\theta}$, but the support is defined implicitly.

3.2 Accessing the posterior

Inference can be carried out by sampling from the posterior distribution of the parameters. However, in this problem, traditional simulation algorithms will fail because the prior and the posterior of the model are supported on a zero Lebesgue measure set (e.g. all the proposed moves of a Metropolis-Hastings (MH) algorithm with a traditional proposal will be rejected almost surely).

Here two solutions to this problem are given. In the first approach, called the "marginal method", we will derive the density function of the marginal of θ, which has a density with respect to the Lebesgue measure $p(\theta)$ and therefore can be processed by conventional Monte Carlo methods. Examples include standard MCMC algorithm and importance sampling. This is simple but comes at the cost of having to solve for β for each proposal. If finding β (or indeed all the values of β which solve given θ) is cheap then this provides a very solid solution to the problem.

In the second approach, called the "joint method", we define a proposal in the space of (β,θ) that assigns positive probability to $\Theta_{\beta,\theta}$ (so, with positive probability, the proposed moves remain on the manifold $\Theta_{\beta,\theta}$ and will be accepted). An MH algorithm with this proposal is able to efficiently move in the space. This does not require us to solve the moment conditions at all, which is extremely attractive for difficult to solve moment condition models.

3.3 Marginal method

Let $p(\beta,\theta)$ be the density function of the model's prior or posterior with respect to Hausdorff measure on $\Theta_{\beta,\theta}$. Proposition 1 gives the marginal density of θ with respect to Lebesgue measure.

This implies that standard Monte Carlo methods (e.g. MCMC, importance sampling, sequential importance sampling and Hamiltonian Monte Carlo) can be used[2].

Proposition 1 *Let $p(\beta, \theta)$ be the density function of the prior or posterior with respect to Hausdorff measure supported on $\Theta_{\beta,\theta}$. Moreover, assume $p = r$ (the "just identified" case) and β is uniquely determined by θ, i.e. $\beta = \beta(\theta)$. Then the density function of θ with respect to Lebesgue measure is*

$$p(\theta) = \sqrt{\left|J_\theta J_\theta' + I_p\right|}\; p(\beta, \theta), \tag{5}$$

where

$$J_\theta = \frac{\partial \beta}{\partial \theta'} = -\left\{\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right)\right\}^{-1} H_\beta, \quad \textit{where} \quad H_\beta = (g_1, ..., g_{J-1}) - g_J \iota', \tag{6}$$

with ι being a $(J-1)$-vector of ones and

$$\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right) = \sum_{j=1}^{J} \theta_j \frac{\partial g(s_j, \beta)}{\partial \beta'}.$$

This proposition is a direct result of the "area formula" of Federer (1969) (see also Diaconis et al. (2013)) and it can be generalized straightforwardly to the cases where for some values of θ there exist more than one β by summing over the right hand side for each solution in β.

The Jacobian[3] term $\sqrt{\left|J_\theta J_\theta' + I_p\right|}$ depends on the geometry of the parameter space $\Theta_{\beta,\theta}$ (in other words, it only depends on the moment conditions) and is independent of $p(\beta, \theta)$. To compute this term we need to invert a $p \times p$ matrix and evaluate the determinant of a $p \times p$ matrix. However, p is usually small, in which case the computational cost of these operations is negligible.

Importantly knowledge of the functional form of β as a function of θ is not needed, since the partial derivatives can be obtained by the implicit function theorem. However, in order to evaluate this density function for a given θ, we need its corresponding β. Although in some problems β has a known analytic form as a function of θ, in many other situations it can be obtained through a numeric optimization. We now return to the examples introduced in Section 2.

Example 1 (Continued) *The density of θ in the logistic model is*

$$p(\theta) = p(\beta, \theta)\sqrt{1 + \left(\frac{\partial \log\left(\frac{\theta}{1-\theta}\right)}{\partial \theta}\right)^2}. \tag{7}$$

[2]We sample from the unconstrained $p(\eta)$, where $\eta_j = \log(\theta_{j+1}/\theta_j)$, for $j = 1, ..., J-1$, with $|\partial\theta/\partial\eta| = \prod_{j=1}^{J} \theta_j$.

[3]A Jacobian correction terms also appears in reversible jump MCMC (e.g. Green (1995)), when the chain is allowed to jump between models with different number of parameters. However there the (one-to-one) transformations are operating between spaces of the same dimension, and the distributions in both spaces have densities w.r.t. Lebesgue measure. On the other hand, the Jacobian in Proposition 1 corrects for a one-to-one mapping between spaces of different dimensions and relates two densities that are defined w.r.t. different reference measures.

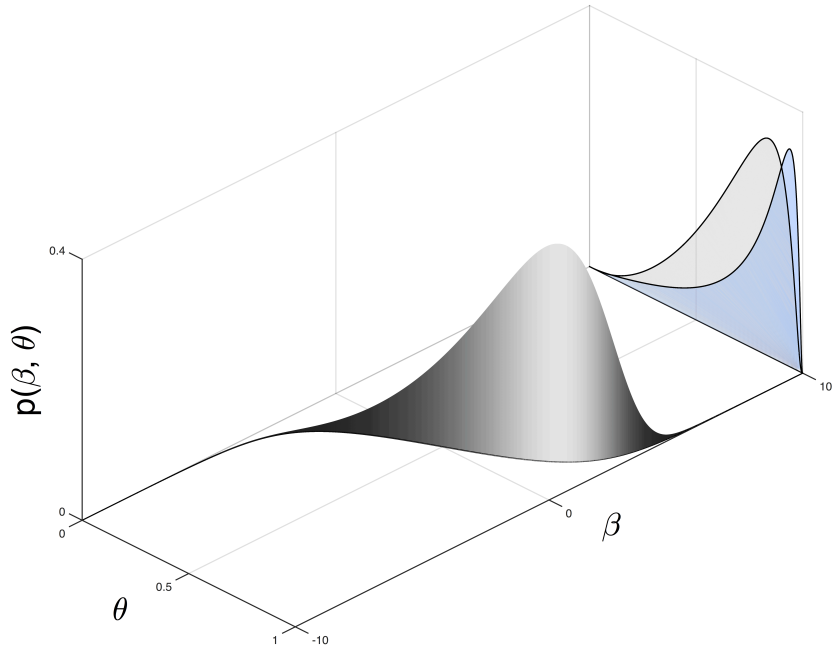

Figure 2: Projection to the marginal density for θ. Blue density is the correct marginal density $p(\theta)$, given in (7), with respect to Lebesgue measure. The grey density is the naive density $p\left[\beta = \log\left\{\theta/\left(1-\theta\right)\right\}, \theta\right]$ which ignores the corresponding length of the support.

Thus moment condition impacts the marginal prior on θ. Figure 2 shows the function $p(\theta)$, which has blue shade below the curve, together with the naive $p\left(\beta = \log\left\{\theta/\left(1-\theta\right)\right\}, \theta\right)$, which has grey shade. We can see the correct density is higher for high values of θ as there are more dense values of β compatible with high values of θ than when θ is close to 0.5.

Example 2 (Continued) *The density of θ in the mean model is*

$$p(\theta) = p(\beta, \theta)\sqrt{1 + \sum_{j=1}^{J-1}\left(s_j - s_J\right)^2} \propto p(\beta, \theta).$$

Hence in this case the geometry of moment condition does not impact the prior on θ. This will be the case generally when the parameter space, $\Theta_{\beta,\theta}$, is flat.

Example 3 (Continued) *For the regression model write $g_j = g(s_j, \beta)$ for $1 \leq j \leq J$. Therefore*

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} s_j^{(2)\prime}, \quad \textit{and} \quad \frac{\partial \beta}{\partial \theta_i} = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} s_j^{(2)\prime}\right)^{-1} \left(g_i - g_J\right).$$

Moreover

$$J_\theta J_\theta' = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} s_j^{(2)\prime}\right)^{-1} \left\{\sum_{i=1}^{J} \left(g_i - g_J\right)\left(g_i - g_J\right)'\right\} \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} s_j^{(2)\prime}\right)^{-1}.$$

Similarly for the linear regression model with instrumental variables we have,

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(3)} s_j^{(2)\prime}, \quad \textit{and} \quad \frac{\partial \beta}{\partial \theta_i} = \left(\sum_{j=1}^{J} \theta_j s_j^{(3)} s_j^{(2)\prime} \right)^{-1} (g_i - g_J),$$

and therefore

$$J_\theta J_\theta' = \left(\sum_{j=1}^{J} \theta_j s_j^{(3)} s_j^{(2)\prime} \right)^{-1} \left\{ \sum_{i=1}^{J} (g_i - g_J)(g_i - g_J)' \right\} \left(\sum_{j=1}^{J} \theta_j s_j^{(3)} s_j^{(2)\prime} \right)^{-1}.$$

Again generalizing to nonlinear regression models is straightforward. If we define $\mu_j = \mu(\beta, s_j^{(2)})$, *then*

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} \frac{\partial \mu_j}{\partial \beta'}, \quad \textit{and} \quad \frac{\partial \beta}{\partial \theta_i} = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \frac{\partial \mu_j}{\partial \beta'} \right)^{-1} (g_i - g_J),$$

which implies

$$J_\theta J_\theta' = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \frac{\partial \mu_j}{\partial \beta'} \right)^{-1} \left\{ \sum_{i=1}^{J} (g_i - g_J)(g_i - g_J)' \right\} \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \frac{\partial \mu_j}{\partial \beta'} \right)^{-1}.$$

For instance for $\mu(\beta, s^{(2)}) = \exp(\beta'^{(2)})$ *we have,*

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} \exp(\beta_j'^{(2)}) s_j^{(2)\prime}, \quad \textit{and} \quad \frac{\partial \beta}{\partial \theta_i} = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \exp(\beta_j'^{(2)}) s_j^{(2)\prime} \right)^{-1} (g_i - g_J),$$

and hence

$$J_\theta J_\theta' = \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \exp(\beta_j'^{(2)}) s_j^{(2)\prime} \right)^{-1} \left\{ \sum_{i=1}^{J} (g_i - g_J)(g_i - g_J)' \right\} \left(\sum_{j=1}^{J} \theta_j s_j^{(2)} \exp(\beta_j'^{(2)}) s_j^{(2)\prime} \right)^{-1}.$$

Example 4 (Continued) *For the casual inference problem write* $g_j = g(s_j, \beta)$, *for* $1 \leq j \leq J$. *Then*

$$\frac{\partial g_j}{\partial \beta'} = \begin{bmatrix} s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)\prime} & 0_{K \times 1} \\ 0_{1 \times K} & -1 \end{bmatrix}, \quad \textit{and} \quad \frac{\partial \beta}{\partial \theta_i} = \left(\begin{bmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)\prime} & 0_{K \times 1} \\ 0_{1 \times K} & -1 \end{bmatrix} \right)^{-1} (g_i - g_J),$$

which implies

$$\begin{aligned} J_\theta J_\theta' \;=\; & \left(\begin{bmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)\prime} & 0_{K \times 1} \\ 0_{1 \times K} & -1 \end{bmatrix} \right)^{-1} \left\{ \sum_{i=1}^{J} (g_i - g_J)(g_i - g_J)' \right\} \\ & \left(\begin{bmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)\prime} & 0_{K \times 1} \\ 0_{1 \times K} & -1 \end{bmatrix} \right)^{-1}. \end{aligned}$$

An immediate consequence of Proposition 1 is that if we reparametrize the scientific parameters of interest $\psi = \psi(\beta)$ using a one to one transform, then

$$p(\psi, \theta) = \frac{\sqrt{\left| \frac{\partial \beta}{\partial \theta'} \frac{\partial \beta}{\partial \theta'}' + I_p \right|}}{\sqrt{\left| \frac{\partial \psi}{\partial \theta'} \frac{\partial \psi}{\partial \theta'}' + I_p \right|}} p(\beta, \theta), \tag{8}$$

where $p(\psi, \theta)$ and $p(\beta, \theta)$ are densities with respect to Hausdorff measures.

3.4 Joint method

Alternatively, we may draw random samples directly from the posterior of (β, θ). This distribution is supported on a zero Lebesgue measure set, $\Theta_{\beta,\theta}$, with the density function (with respect to Hausdorff measure) $p(\beta, \theta)$. If we ignore this and propose moves from a continuous proposal distribution in $\mathbb{R}^{J+p-1}$ (for instance a Gaussian proposal), the proposed moves are off the support of $p(\beta, \theta)$ almost surely, and they will be rejected with probability one. Therefore in order to sample from $p(\beta, \theta)$ we must find a proposal distribution that assigns positive probability to $\Theta_{\beta,\theta}$. Drawing random samples from this proposal should be easy and fast and (in order to compute the acceptance probability) we should be able to evaluate its density function. This subsection will explain how this can be achieved.

For a given value of β, the moment conditions imply the affine constraints on θ:

$$H_\beta \theta + g_J = 0. \tag{9}$$

Therefore $\Theta_{\theta|\beta}$ is a $(J-1)$-hyperplane in $\mathbb{R}^{J+p-1}$. This property allows us to define a suitable proposal distribution for (β, θ). Assume the current state of the MCMC is $(\beta^{(t)}, \theta^{(t)})$. First we explain how a random sample from the proposal can be drawn, and then will show how the density of this proposal can be evaluated. In order to draw a random sample from $q(\cdot, \cdot|\beta^{(t)}, \theta^{(t)})$,

1. Draw $\beta^*|\beta^{(t)}, \theta^{(t)}$ from an (almost) arbitrary proposal $q(\cdot|\beta^{(t)}, \theta^{(t)})$.

2. Draw θ^* from a singular distribution supported on the hyperplane $\mathcal{P}^* = \{\lambda \in \mathbb{R}^{J-1}; H^*_{\beta^*}\lambda + g^*_J = 0\}$. We denote the density of this distribution (with respect to the Hausdorff measure) by $q(\cdot|\beta^{(t)}, \theta^{(t)}, \beta^*)$. Moreover we assume the density can be easily evaluated at any θ^*. A singular Normal distribution supported on $\mathcal{P}^*$ is one suitable choice (see Khatri (1968)). In the Appendix A.3 we provide a way to determine the parameters of a singular Normal distribution that can be used to propose for $\theta^*|\beta^{(t)}, \theta^{(t)}, \beta^*$.

So far we have shown how a random proposal can be generated from $q(\cdot,\cdot|\beta^{(t)},\theta^{(t)})$. The following propositions demonstrates how the density of this proposal can be evaluated when $p = r$.

Proposition 2 *Let $p(\beta,\theta)$ be the density of (β,θ) with respect to $(J-1)$-dimensional Hausdorff measure on $\Theta_{\beta,\theta}$. Moreover assume the density of β with respect to Lebesgue measure is $p(\beta)$, and the density of $\theta|\beta$ with respect to Hausdorff measure is $p(\theta|\beta)$ on $\Theta_{\theta|\beta}$, where $\Theta_{\theta|\beta}$ is a hyperplane. Then*

$$p(\beta,\theta) = \frac{|J_\theta J_\theta'|^{\frac{1}{2}}}{\left|J_\theta J_\theta' + I_p\right|^{\frac{1}{2}}}\, p(\beta)p(\theta|\beta), \quad \textit{where} \quad J_\theta = \frac{\partial \beta}{\partial \theta'}. \tag{10}$$

The proposed pairs (β^*,θ^*) satisfy the moment conditions, however the probabilities may not satisfy the probability axioms (as some of θ^* may be negative or $\theta_J^* = 1 - \iota'\theta^* \leq 0$). Obviously in these cases the proposal is rejected (since the posterior is zero), the MCMC algorithm sticks, and the proposal's density need not to be evaluated. If the proposal is valid, then the move $(\beta,\theta) \to (\beta^*,\theta^*)$ is accepted with probability

$$\min\left\{1, \frac{p(\beta^*,\theta^*|Z)}{p(\beta,\theta|Z)}\frac{q(\beta,\theta|\beta^*,\theta^*)}{q(\beta^*,\theta^*|\beta,\theta)}\right\}. \tag{11}$$

The terms inside this acceptance probability are straightforward to compute up to proportionality.

Note that in the joint method we do not need to solve for β in each iteration of the simulation, because our proposed moves are elements of the parameter space $\Theta_{\beta,\theta}$. Moreover, when J goes to infinity, the Jacobian term in (10) converges to 1. To see this assume the data generating process is a continuous distribution or a discrete distribution with infinite support, $s_j \sim H$. Then, with probability one, just using a strong law of large numbers,

$$\frac{1}{J}J_\theta J_\theta' = \frac{1}{J}\left\{\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right)\right\}^{-1} H_\beta H_\beta' \left\{\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right)'\right\}^{-1} \to \mathcal{J}.$$

where $\mathcal{J} = \left\{\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right)\right\}^{-1} \mathbb{E}_H(g_j g_j') \left\{\mathbb{E}_\theta\left(\frac{\partial g}{\partial \beta'}\right)'\right\}^{-1}$. Therefore

$$\frac{|J_\theta J_\theta'|}{\left|J_\theta J_\theta' + I_p\right|} = \frac{\left|\frac{1}{J}J_\theta J_\theta'\right|}{\left|\frac{1}{J}J_\theta J_\theta' + \frac{1}{J}I_p\right|} \to 1,$$

with probability one as J goes to infinity. This asymptotic approximation could be used to simplify the computation of the acceptance probability, but otherwise does not change the substance of this section, as proposals will be made in the same way — directly on the manifold.

3.5 Relationship to the Bayesian bootstrap

The Rubin (1981) "Bayesian bootstrap" is at the core of Chamberlain and Imbens (2003). We can implement our Proposition 1 by using their Bayesian bootstrap as a proposal which can be reweighted to allow for informative priors on β. Throughout we assume β can be solved given θ.

Our generalization of Chamberlain and Imbens (2003) starts with the Dirichlet prior $\pi^*(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1}$, $\alpha > 0$. The Bayesian bootstrap then simulates from the proposal density,

$$g(\theta|Z) \propto \prod_{j=1}^{J} \theta_j^{n_j+\alpha-1}. \tag{12}$$

We assume the researcher does this M times, writing the draws as $\left\{\theta^{(k)}\right\}_{k=1,2,...,M}$. For each $\theta^{(k)}$ we assume there is a unique $\beta^{(k)}$ which solves the corresponding moment conditions. Chamberlain and Imbens (2003) stop at this point, using this sample as a Monte Carlo estimate of the posterior.

Correcting for the geometry of the problem, the actual posterior is

$$p(\theta|Z) \propto p(\beta,\theta) \left(\prod_{j=1}^{J} \theta_j^{n_j}\right) \left|J_\theta J_\theta' + I_p\right|^{\frac{1}{2}}. \tag{13}$$

The resulting weights from the true posterior density with respect to the Lebesgue measure dividing by the density from the proposal are

$$w^{(k)} = \frac{p(\beta^{(k)},\theta^{(k)}) \left|J_\theta^{(k)} J_\theta^{(k)\prime} + I_p\right|^{\frac{1}{2}}}{\prod_{j=1}^{J} \left(\theta_j^{(k)}\right)^{\alpha-1}}, \quad k = 1, 2, ..., M, \tag{14}$$

(where $J_\theta^{(k)}$ is equal to J_θ evaluated at $(\beta^{(k)},\theta^{(k)})$) which normalize as $w^{(k)*} = w^{(k)} / \left(\sum_{k=1}^{M} w^{(k)}\right)$. An encouraging aspect of this weight is that it does not depend on the data.

In the special case where $p(\beta,\theta) \propto \pi(\beta)\pi^*(\theta)$, the weights may be simply evaluated as

$$w^{(k)} \propto \pi(\beta^{(k)}) \left|J_\theta^{(k)} J_\theta^{(k)\prime} + I_p\right|^{\frac{1}{2}}, \quad k = 1, 2, ..., M. \tag{15}$$

We can use these weights to estimate $\mathrm{E}\left(h(\beta)|Z\right) \simeq \frac{1}{M}\sum_{k=1}^{M} w^{(k)*} h(\beta^{(k)})$. This is importance sampling, e.g. Marshall (1956), Geweke (1989), Liu (2001). An alternative is to resample with probability proportional to the weight $w^{(k)}$, which delivers sampling importance resampling (SIR, see Rubin (1988)). As with all importance samplers, the weights may become uneven although the simplicity of the structure of the weights is encouraging. This sampling strategy becomes appealing in the models where the β can be computed easily for any θ, and the prior distribution of β is not too far from the posterior obtained from the Bayesian bootstrap.

3.6 Missing support

So far we have assumed the support of the data is known. Here we extend this to assume the support has $J^* > J$ elements, $S^* = (s_1, ..., s_{J^*})$, where its first $\tilde{J} = J^* - J$ elements, $\tilde{S} = (s_1, ..., s_{\tilde{J}})$, have not been observed in the sample, while the rest of its elements $S = (s_{\tilde{J}+1}, ..., s_J)$ have been observed at least once. Moreover let $\theta^* = (\tilde{\theta}, \theta)$, where $\tilde{\theta}$ and θ are the vector of the probabilities of the elements of $\tilde{S}$ and $(s_{\tilde{J}+1}, ..., s_{J-1})$, and define $\theta_{J^*} = 1 - \sum_{j=1}^{J^*-1} \theta_j^*$. We assume the missing elements of the support are i.i.d. draws from F_S, $s_j \overset{iid}{\sim} F_S$ for $j = 1, ..., \tilde{J}$, with density f_S with respect to Lebesgue measure. The moment conditions are then

$$\sum_{j=1}^{J^*} \theta_j g(s_j, \beta) = 0, \tag{16}$$

while the posterior is

$$p(\beta, \theta^*, \tilde{S}|Z) \propto p(\beta, \theta^*, \tilde{S}) \left(\prod_{j=1}^{J^*} \theta_j^{n_j} \right),$$

where $n_j = \sum_{i=1}^{N} 1(Z_i = s_j)$. Note that $n_j = 0$ for $1 \leq j \leq \tilde{J}$, while n_j is positive for $\tilde{J}+1 \leq j \leq J^*$. Assume the researcher expresses a prior on $(\beta, \theta^*)|\tilde{S}$ with respect to the Hausdorff measure (suppressing the conditioning on S for notational convenience),

$$p(\beta, \theta^*|\tilde{S}). \tag{17}$$

Therefore

$$p(\beta, \theta^*, \tilde{S}) = \left(\prod_{j=1}^{\tilde{J}} f_S(\tilde{s}_j) \right) p(\beta, \theta^*|\tilde{S}) \tag{18}$$

Given θ^* and $\tilde{S}$ (and S), β is uniquely determined. Therefore the core result we need to do inference is a generalization of Proposition 1: the density of the probabilities and the missing support with respect to the Lebesgue measure is

$$p(\theta^*, \tilde{S}) = \left| J_{\theta^*} J'_{\theta^*} + J_{\tilde{S}} J'_{\tilde{S}} + I_p \right|^{\frac{1}{2}} \left(\prod_{j=1}^{\tilde{J}} f_S(\tilde{s}_j) \right) p(\beta, \theta^*|\tilde{S}), \tag{19}$$

where

$$\begin{aligned} J_{\theta^*} &= \frac{\partial \beta}{\partial \theta^*} = \left(\sum_{j=1}^{J^*} \theta_j \frac{\partial g_j}{\partial \beta'} \right)^{-1} H_\beta^*, \quad J_{\tilde{S}} = \frac{\partial \beta}{\partial \tilde{S}} = \left(\sum_{j=1}^{J^*} \theta_j \frac{\partial g_j}{\partial \beta'} \right)^{-1} \tilde{M}, \\ \tilde{M} &= \left\{ \theta_1 \left(\frac{\partial g_1}{\partial s_1'} \right), \cdots, \theta_{\tilde{J}} \left(\frac{\partial g_{\tilde{J}}}{\partial s_{\tilde{J}}'} \right) \right\}. \end{aligned} \tag{20}$$

Again this result follows from the area formula. Proposition 2 generalizes in the same way delivering

$$p(\beta,\theta^*,\tilde{S}) = \frac{\left|J_{\theta^*}J'_{\theta^*} + J_{\tilde{S}}J'_{\tilde{S}}\right|^{\frac{1}{2}}}{\left|J_{\theta^*}J'_{\theta^*} + J_{\tilde{S}}J'_{\tilde{S}} + I_p\right|^{\frac{1}{2}}}\, p(\beta)p(\theta^*|\beta,\tilde{S}) \left(\prod_{j=1}^{\tilde{J}} f_S(\tilde{s}_j)\right). \tag{21}$$

Again the Jacobian will be close to one if J^* is large. The ratio of J to J^* does not make any difference to this approximation.

Example 2 (Continued) *Now add a single point of missing support. Then* $J^* = 4$, $\theta^* = (\theta_1,\theta_2,\theta_3)'$, $J = 3$ *and* $S^* = \{s_1,s_2,s_3,s_4\} = \{-1,0,1,s_4\}$. *Then* $\beta = \theta_3 - \theta_1 + s_4\theta_4 = \theta_3 - \theta_1 + s_4(1-\theta_1-\theta_2-\theta_3)$. *For this model*

$$J_{\theta^*} = \frac{\partial\beta}{\partial\theta^*} = (-1-s_4, -s_4, 1-s_4) \quad \textit{and} \quad J_{\tilde{S}} = \frac{\partial\beta}{\partial\tilde{S}} = \theta_4, \tag{22}$$

and so $J_\theta J'_\theta = 2 + 3s_4^2$ *and* $J_{\tilde{S}}J'_{\tilde{S}} = \theta_4^2$. *Hence, writing* $\theta_4 = 1-\theta_1-\theta_2-\theta_3$,

$$p(s_4,\theta^*) = \left\{\sqrt{\left(2+3s_4^2\right) + \theta_4^2 + 1}\right\} f_S(s_4)p(\beta,\theta^*|s_4). \tag{23}$$

4 Some potential priors

So far we have discussed working with any prior $p(\beta,\theta)$ which is defined with respect to lower dimensional Hausdorff measure supported on $\Theta_{\beta,\theta}$. In this section we discuss potential ways of selecting $p(\beta,\theta)$. As with all prior selection there is no uniquely good way of carrying this out.

4.1 A non-science prior

From a nonparametric standpoint it is natural to build a prior from $p(\theta)$, e.g. Dirichlet. Then Proposition 1 implies there is a unique joint prior

$$p(\beta,\theta) = \frac{p(\theta)}{\sqrt{\left|J_\theta J'_\theta + I_p\right|}}, \tag{24}$$

which achieves this. The right hand side $p(\theta)$ is the density of θ with respect to Lebesgue measure, while $p(\beta,\theta)$ is the density of (β,θ) with respect to Hausdorff measure. This implies

$$\Pr\left\{(\beta,\theta)\in C\right\} = \int_{C_\theta} p(\theta)\mathrm{d}\theta. \tag{25}$$

The Dirichlet special case (24) is the implicit Chamberlain and Imbens (2003) prior on $p(\beta,\theta)$.

4.2 A prior on the science

Proposition 2 says that

$$p(\beta,\theta)=\frac{\left|J_\theta J_\theta'\right|^{1/2}}{\left|J_\theta J_\theta'+I_p\right|^{1/2}}p(\beta)p(\theta|\beta). \tag{26}$$

If we place a prior on the science $p(\beta)$ with respect to the Lebesgue measure, then we can form a scientifically centered prior on $p(\beta,\theta)$ by specifying a prior on $p(\theta|\beta)$ with respect to the $J-1-p$ dimensional Hausdorff measure. This prior sits on the hyperplane $\theta|\beta$ satisfying the linear constraints (9) and the probability axioms. One such prior is Dirichlet subject to the constraints. Again if J gets large the Jacobian in (26) will become unimportant in practice.

4.3 Adhoc priors

A more brutal approach to building a prior is to define an "initial" prior (with respect to Lebesgue measure) for β and θ which ignores the moment condition $\eta(\beta,\theta)$ where the implied initial marginal prior on β, $\eta(\beta)$, could be our substantive initial prior. From the Borel paradox (Kolmogorov (1956)) we know there are many ways of building a $p(\beta,\theta)$ from $\eta(\beta,\theta)$ (conditioning on satisfying the moment condition is not enough) but here we discuss various plausible methods.

This line of thinking leads to a generalization of (24), setting

$$p(\beta,\theta)\propto\frac{\eta(\beta,\theta)}{\left|J_\theta J_\theta'+I_p\right|^{1/2}}1_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{27}$$

This prior scales the initial prior to countereffect the length of the curve mapping out the relationship between θ and β implied by the moment condition. This prior has the property that $p(\theta)\propto\eta(\beta,\theta)1_{\Theta_{\beta,\theta}}(\beta,\theta)$, with respect to the Lebesgue measure.

The simple case of $\eta(\beta,\theta)=\eta(\beta)\eta(\theta)$, would imply under (27)

$$p(\theta)\propto\eta(\beta)\eta(\theta)1_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{28}$$

The case where $\eta(\theta)$ is Dirichlet is important. Then the Bayesian bootstrap weights (27) would become the rather simple

$$w_j\propto\eta(\beta^{(j)}),\quad j=1,2,...,M. \tag{29}$$

This is a minimally informative generalization of Chamberlain and Imbens (2003).

An alternative to (27) is to put no mass on inadmissible combinations of β,θ. We call this the "truncated prior"

$$p(\beta,\theta)\propto\eta(\beta,\theta)1_{\Theta_{\beta,\theta}}(\beta,\theta) \tag{30}$$

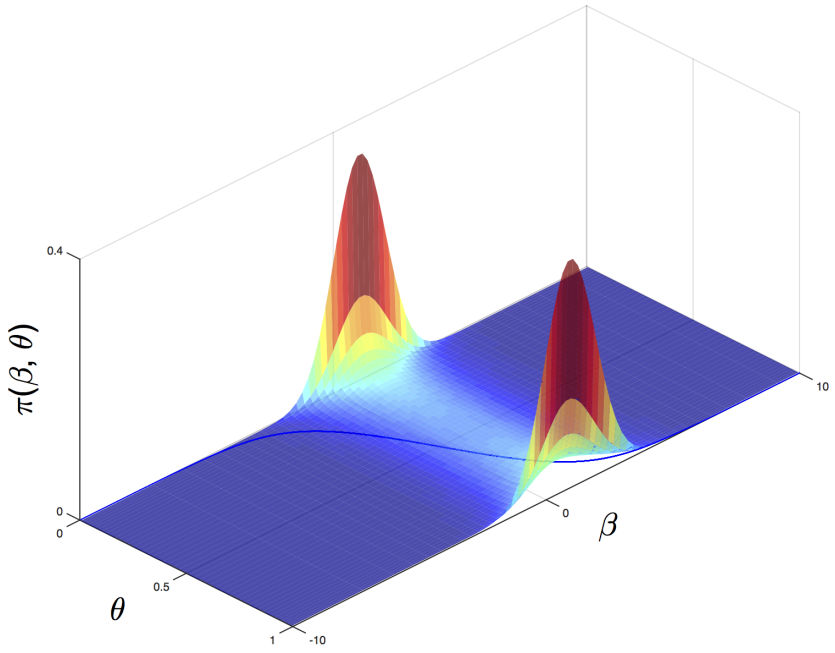

Figure 3: The parameter space (the blue curve) $\Theta_{\beta,\theta}$ and the initial prior $\pi(\beta, \theta)$ have been depicted. Figure 1 shows the implied $p(\beta, \theta)$.

in which $p(\beta, \theta)$ is the density of the prior with respect to the $(J-1)-$dimension Hausdorff measure in $\mathbb{R}^{J-1+p}$. This would imply for any set $C \in \mathbb{R}^{J-1+p}$

$$\begin{aligned} \Pr\{(\beta, \theta) \in C\} &= \int_{C_\theta} p(\beta, \theta)\sqrt{|J_\theta J_\theta' + I_p|}\mathrm{d}\theta \\ &\propto \int_{C_\theta} \eta(\beta, \theta)\sqrt{|J_\theta J_\theta' + I_p|}\mathrm{d}\theta. \end{aligned} \tag{31}$$

Obviously it implies $p(\theta) \propto \eta(\beta, \theta)\sqrt{|J_\theta J_\theta' + I_p|}1_{\Theta_{\beta,\theta}}(\beta, \theta)$, with respect to the Lebesgue measure.

Example 5 *(continuing logistic Example 1). Assume the initial prior*

$$\eta(\beta, \theta) \propto \theta^{0.01-1}(1-\theta)^{0.01-1}e^{-\frac{1}{2}(\beta-1)^2}, \tag{32}$$

which is a relatively ignorant Dirichlet prior on the probabilities and an informative Gaussian prior for β centered on one. This is depicted in Figure 3. With this initial prior and using the class of priors (30), the density with respect to the univariate Hausdorff measure is

$$p(\beta, \theta) \propto \eta(\beta, \theta)1_{\Theta_{\beta,\theta}}(\beta, \theta). \tag{33}$$

Figure 1 shows the corresponding $p(\beta, \theta)$ living on the manifold. In this case

$$p(\theta) \propto \eta(\beta, \theta)\sqrt{1 + \left(\frac{\partial\beta}{\partial\theta}\right)^2}1_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{34}$$

with respect to the Lebesgue measure. With the alternative (27) prior, then

$$p(\beta,\theta) \propto \frac{\eta(\beta,\theta)}{\sqrt{1+\left(\frac{\partial\beta}{\partial\theta}\right)^2}} 1_{\Theta_{\beta,\theta}}(\beta,\theta), \quad \textit{and} \quad p(\theta) \propto \eta(\beta,\theta) 1_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{35}$$

5 Illustrative examples

In this section we present some illustrative examples and simulation studies. Since the MCMC results obtained by the marginal and joint methods are indistinguishable we present only one of them. At the end of the section we study how the methods scale.

5.1 The mean

Recall inference on the mean studied in Example 2. Now focus on $J = 3$ and $S = (-1, 0, 1)$, so $\beta = \theta_3 - \theta_1 = 1 - 2\theta_1 - \theta_2$. Here we have taken the 2 dimensional Hausdorff prior as

$$p(\beta,\theta) \propto e^{-2|\beta-m|} \theta_1^{\alpha-1} \theta_2^{\alpha-1} (1-\theta_1-\theta_2)^{\alpha-1} 1(\min\{\theta_1,\theta_2,1-\theta_1-\theta_2\} \geq 0). \tag{36}$$

We call this a "Laplace-Dirichlet" distribution on (β,θ), where β is centered around m and the Dirichlet part is indexed by α.

By the marginal method:

$$p(\theta) \propto e^{-2|1-2\theta_1-\theta_2-m|} \theta_1^{\alpha-1} (1-\theta_1-\theta_2)^{\alpha-1} \theta_2^{\alpha-1} 1(\min\{\theta_1,\theta_2,1-\theta_1-\theta_2\} \geq 0), \tag{37}$$

Figure 4 shows the contours of $p(\theta)$ for various values of m and α. We have plotted these contours against $(\theta_1,\theta_2,\theta_3)'$ so the reader can compare θ_1 and θ_3.

If the Laplace-Dirichlet distribution has $m = 0$ then the density is symmetric with respect to θ_1 and θ_3. When the location parameter of $p(\beta)$ is positive θ_1 is on average smaller than θ_3. Moreover as α increases, the variability of $p(\theta)$ decreases.

Figure 5 draws the prior for β. Here the support of the data means β is restricted to the real line, after observing the support of the data its prior is restricted to $[-1, 1]$. As α increases the variance of β decreases. For instance the β's prior centered at a positive value results in a prior for θ tilted toward θ_3, even if the prior of θ is symmetric. In the same way, a more informative initial prior for θ yields a more peaked prior for β.

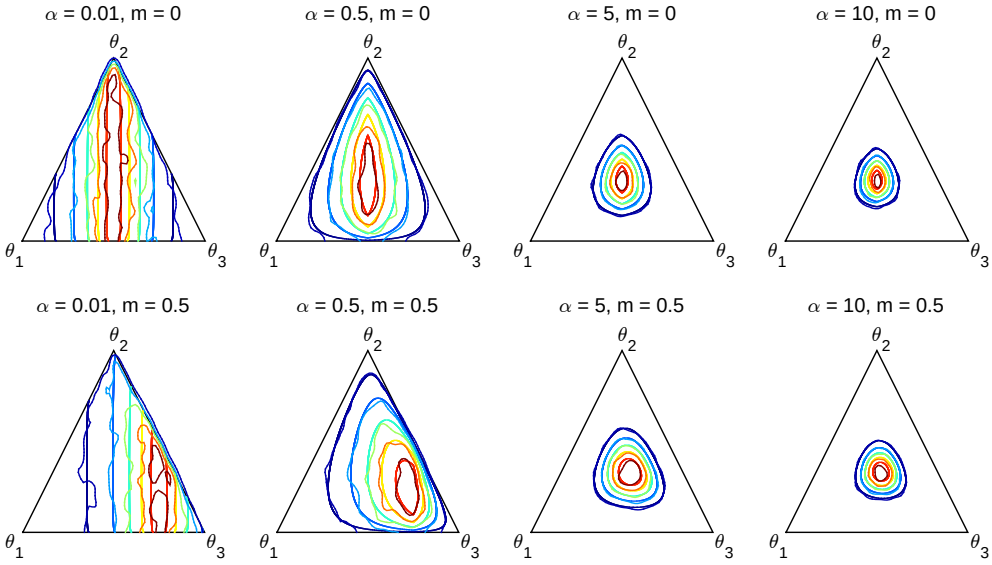

Figure 4: Equiprobability contours implied by the Laplace-Dirichlet prior on $p(\beta, \theta)$ with respect to the Hausdorff measure. Plotted is the marginal $p(\theta_1, \theta_3)$ for several values of m and α, with θ_2 implied as $\theta_2 = 1 - \theta_1 - \theta_3$. This case has $J = 3$ points of support $(s_1 = -1, s_2 = 0, s_3 = 1)$ and $r = 1$ moment constraints (the mean).

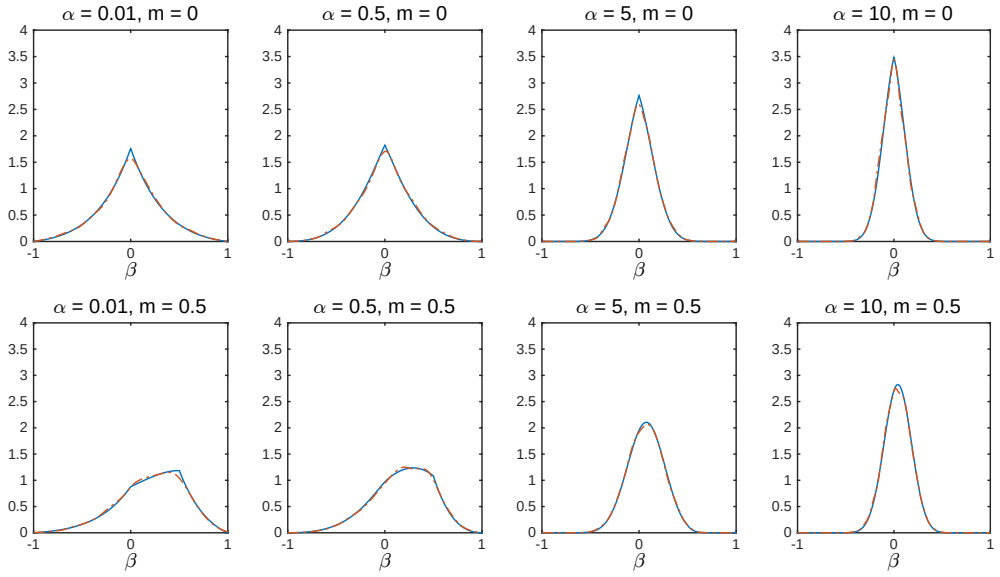

Figure 5: Illustrating Example 1 (estimating the mean). Plot of $p(\beta)$ for several values of m and α. This case has $J = 3$ points of support $(s_1 = -1, s_2 = 0, s_3 = 1)$ and $r = 1$ moment constraints (the mean). Initial prior for β is Laplace centered at m and the initial prior for θ is symmetric Dirichlet with parameter α.

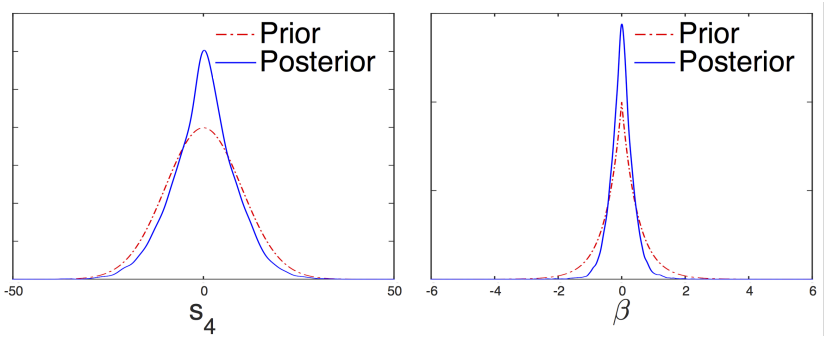

Figure 6: Illustrating Example 1 with the partially observed support: inferring the mean β. The prior and the posterior of the missing element of the support s_4 (left panel) and mean β (right panel).

5.2 Missing support and the mean

In the previous section, the finite support of β is caused by the known support of the data. We now extend this to cover Example 2 where we have a single missing datapoint

$$s_4 \sim N(0, 10^2), \tag{38}$$

all other features of the problem are unchanged. An adaptive MH algorithm has been used in order to draw $10,000$ samples from the joint distribution. For the sake of brevity we present the results only for the case of $m = 0$ and $\alpha = 0.5$.

The means and standard error of the probabilities θ^* are $(0.3065, 0.3335, 0.3079, 0.0521)$ and $(0.0026, 0.0030, 0.0024, 0.0010)$, respectively. The left panel of Figure 6 shows the initial prior (38) and the implied marginal distribution of the missing element of the support s_4 from the joint prior. The variance of the implied marginal is smaller than the prior's variance, because the prior distribution of β is informative about the support of the data. The right panel of Figure 6 shows the Laplace element of the prior $e^{-2|\beta - m|}$ and the full marginal prior for β. The full marginal prior is not the same as the Laplace distribution due to the informative priors on the probabilities.

5.3 Linear regression

Recall the linear regression of Example 3. Assume the observed data is $Z = \{(1,1),(2,4),(3,9)\}$. Earlier we have seen that the parameter space, $\Theta_{\beta,\theta}$ is a non-flat surface in $\mathbb{R}^3$. Figure 7 demonstrates the posterior distribution of the parameters defined on this surface (the prior parameters are $\alpha = 0.5$ and $m = 3$).

Following the suggested MCMC simulation algorithms we draw $100,000$ samples from the poste-

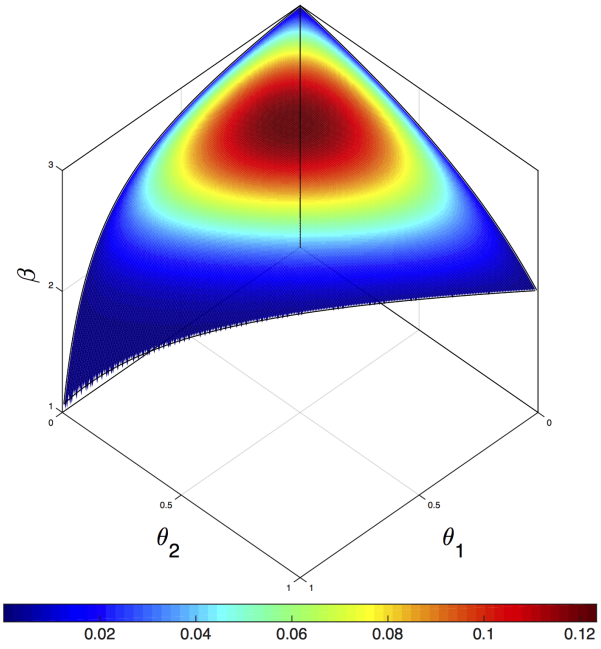

Figure 7: The posterior distribution of the linear regression model with data $Z = \{(1,1),(2,4),(3,9)\}$. The prior parameters are $\alpha = 0.5$ and $m = 3$.

rior distribution of the parameters. In the Figure 8 we have drawn the contour plots of the posterior distribution of the probabilities. Analytical results have been compared with the estimates obtained by a kernel density estimator using the MCMC draws.

5.4 Simulation study

To demonstrate the scalability of the algorithms we consider a linear regression model with sample size $J = 500$. The data $Z_j = (Y_j, X_j)$, for $1 \leq j \leq J$, is generated according to $X_j \sim \mathcal{N}(1, 2^2)$, $Y_j|X_j \sim \mathcal{N}(2 + 5X_j, 10^2)$. We assume the substantive prior of β is $\beta \sim \mathcal{N}(\mu_0, \Sigma_0)$, where the elements of μ_0 are equal to the 25% quantiles of the asymptotic MLE estimators, and Σ_0 is equal to the asymptotic variance of the MLE estimator multiplied by 100 (see Appendix A.5 for the results with a different prior). The initial prior of θ is a symmetric Dirichlet distribution with parameter $\alpha = 0.01$. We have drawn $50,000$ samples from the posterior after a $5,000$ sample burn-in (the chain's trace has been thinned with a factor of 100, so has been iterated $5,000,000$ times). The scatter plot of the sample is depicted in the top-left panel of Figure 9. Each circle represents a data point in our sample and its radius is proportional to the expected value of its posterior probability, i.e. $\mathbb{E}(\theta_j|Z)$. In the top-right panel the correlogram (ACF) of the chains of β and 10 elements of θ have been presented (the red dashed lines and the blue dotted lines are

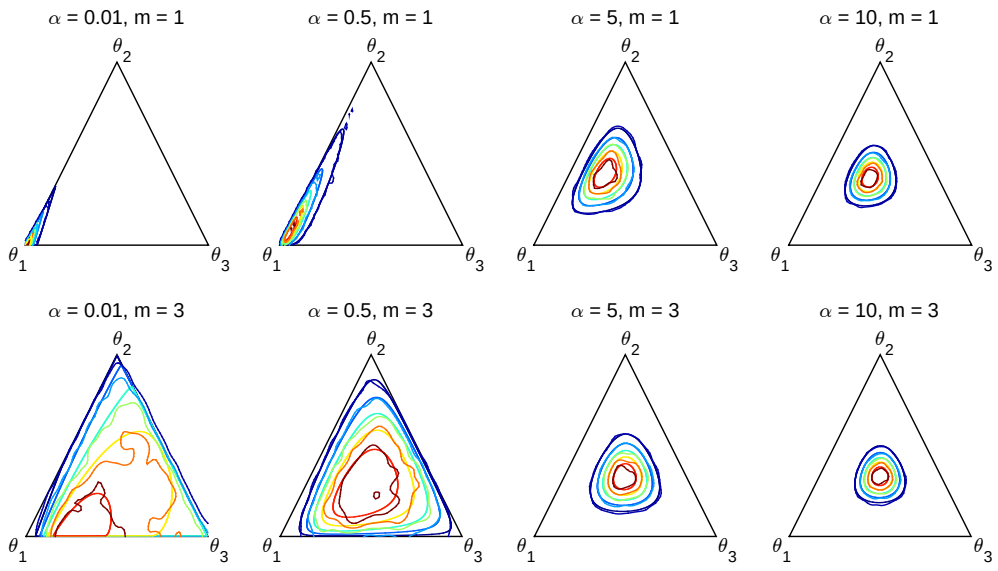

Figure 8: The posterior distribution of θ in the linear regression model with data $Z = \{(1,1),(2,4),(3,9)\}$ (analytical results and the estimates obtained by a kernel density estimator using $100,000$ MCMC draws). The prior parameters are $\alpha = 0.5$ and $m = 3$.

corresponding to β and θ, respectively.) The ACFs demonstrate that the Markov chain is mixing sufficiently well. In the bottom-left panel the contour plot of the posterior distribution of β has been compared to the one obtained by the Bayesian bootstrapping of Chamberlain and Imbens (2003). The posterior distributions are very close, because the prior's information is roughly 1% of the information content of the sample. The bottom-right panel shows a histogram of the samples from the posterior distribution of β.

6 Empirical studies

In this section we study two empirical examples. The first focuses on an instrumental variable based estimator, the second looks at estimating the average treatment effect from an experiment.

6.1 Instrumental variables

In this section we demonstrate the applicability and scalability of the methodology developed in this paper to a real dataset. We use a subsample of the earnings and schooling dataset studied in Chamberlain and Imbens (2003). This dataset is a subset of the data studied in Angrist and Krueger (1991) and consist of the self-reported weekly log-earnings (self-reported annual earnings divided by 52) of $162,512$ male subjects who reported positive annual wages in 1979 along with

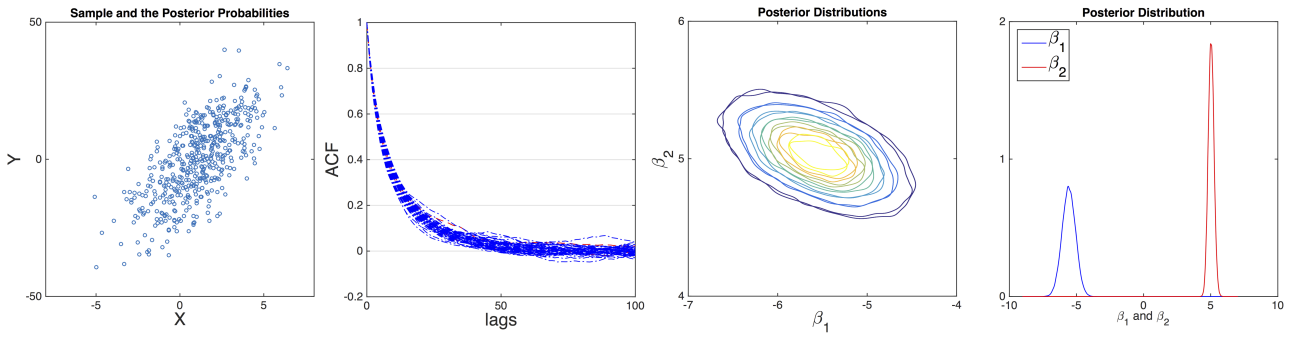

Figure 9: Inference in linear regression model with $J = 500$. Top left shows circles who's radius is the posterior expectation of the probabilities given the data: $E(\theta_j|Z)$. Top right is the correlogram for the thinned draws of the elements of β and ten elements of θ. The bottom graphs show the estimated contour and marginal densities of the resulting posterior.

their number of years of education and their quarter of birth date. In turn this is a 5% random sample from the 1980 Public Use Census Data. Bound et al. (2001) discuss the myriad of problems of self-report income data but we do not address that issue here. For example, Britton et al. (2015) compared UK self-reported income with tax based administrative data finding high income earners significantly under self-recorded their incomes compared to that seen in administrative data.

Chamberlain and Imbens (2003) studied the dependence of earnings on the level of schooling using a linear additive treatment effect model (e.g. Imbens and Rubin (2015)). They model schooling levels as being determined by rational agents' optimization of their lifetime expected utility. Since the utility is a function of the earnings they needed to estimate the distribution of earnings as a function of the schooling level.

The expected log-earnings Y_X with schooling level X is modeled here as $\mathrm{E}(Y_X|X, Y_0) = Y_0 + \beta_1 X$, where X is the schooling level, β_1 is the unknown return to education, and Y_0 is the earnings level with no schooling at all. Let β_0 be the expected value of Y_0, so $Y_0 - \beta_0$ has a zero mean.

In order to estimate the unknown parameters, $\beta = (\beta_0, \beta_1)$, we follow Angrist and Krueger (1991) and Chamberlain and Imbens (2003) and use an instrumental variable (IV) W that is a binary indicator: $W = 0$ if the subject was born in the first three quarters of the year and $W = 1$ otherwise. The instrumental variable W is correlated with the regressor X and thought by the researchers to be uncorrelated with the errors.

We obtain the classic IV estimate of β using the full sample, and treat them as the "true" values of β. Then we draw random samples with replacement of size J from the original data $1,000$ times. Our aim will be to compare different estimators using these smaller samples.

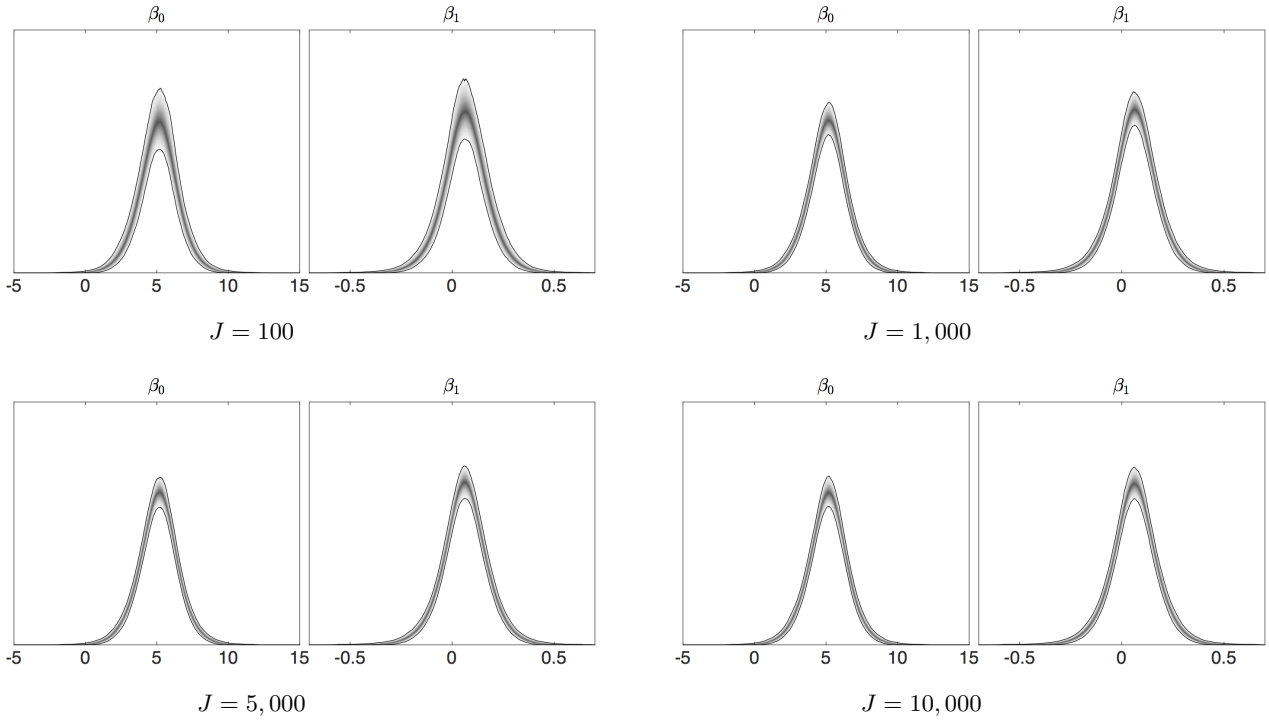

Figure 10: 95% pointwise confidence regions for the marginal prior for β for $J = 10$, $J = 1,000$, $J = 10,000$ and $J = 100,000$ points of random support in top-left, top-right, bottom-left and bottom-right, respectively. The confidence region is computed over 1,000 replications.

Our prior distribution, which is specified to be weakly informative, is

$$p(\beta, \theta) \propto \frac{1}{\sqrt{J_\theta J_\theta' + I_2}} \eta(\beta)\eta(\theta) 1_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{39}$$

where $\eta(\beta) = \varphi(\beta_0; 5, 4)\, \varphi(\beta_1; 0, 0.2)$, and $\varphi\left(\cdot; \mu, \sigma^2\right)$ is the Gaussian density with mean μ and variance σ^2. The intercept is centered at 5 with variance 4, implying that the mean annual income for those with no schooling is equal to \$7,717 (with 95% confidence interval [\$153, \$388965]) with zero years of schooling. Moreover the prior of β_1 has zero mean (no effect of number of schooling years on income) with 95% interval $[-0.88, 0.88]$ (that is equivalent to $[-0.41\%, 241\%]$ income increment for each additional year of schooling.) The probabilities θ are taken as a mildly informative Dirichlet prior $\eta(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1}$, where $\alpha = 10^{-6}$ (we also tried $\alpha = J^{-1}$, with no substantial change in the results).

For $1,000$ iterations, a random sample of size J has been drawn with replacement from the $162,512$ population. For each replication the resulting marginal prior distributions of β_0 and β_1 depend on the draws which generate the support and so vary over the $1,000$ samples. Figure 10 shows the pointwise 95% confidence intervals of the marginal prior distributions over these $1,000$

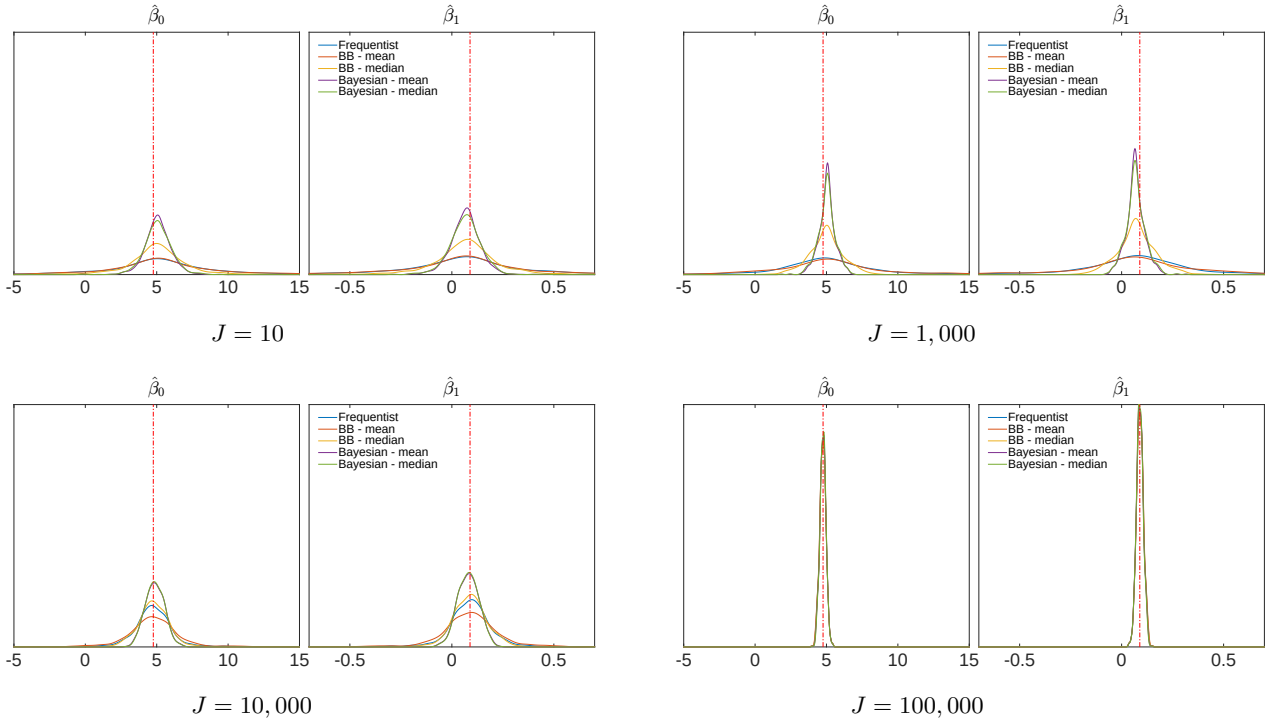

Figure 11: The sampling distribution of classic IV (denoted frequentist), Bayesian bootstrapping and Bayesian estimators of β in the linear regression model with the instrumental variable employing sample sizes $J = 10$, $J = 1,000$, $J = 10,000$ and $J = 100,000$.

replications, for $J = 100$, $1,000$, $5,000$ and $10,000$. It shows the information content of the prior is modest and only mildly depends upon the random support and J, with less variation across replications in the prior density as J increases. Similar results have been obtained for other sample sizes J.

For each random sample, we compute the classic IV estimates of β and the Chamberlain and Imbens (2003) Bayesian bootstrapping estimates obtained by $10,000$ draws. For the latter we report both the means and the medians as the estimators. These estimators are compared with the weakly informative Bayesian estimators (using the prior described earlier).

The Bayesian estimates are obtained by the following resampling method. Initially a sample of size $10,000$ is drawn from a Dirichlet distribution with parameter $(n_1 + \alpha - 1, ..., n_J + \alpha - 1)$, and the importance sampling weights are computed $w^{(k)} \propto \eta(\beta^{(k)})$. Then a sample from the posterior can be obtained by resampling using the normalized weights. Estimators of the mean and the median of the posterior have been reported here. For $J = 10$, 100, $1,000$, $5,000$, $10,000$, $40,000$ and $100,000$ the effective sample size divided by J (Liu, 2001, p. 35) was 0.620, 0.576, 0.607, 0.719, 0.819, 0.978 and 0.997, respectively. This suggests this is a reasonable method for this problem.

In Figure 11 the sampling distribution of these five estimators have been plotted. The blue

	β_0											
	Bias of mean			Bias of median			RMSE			95% CR length		
Sample size J	10	1,000	10,000	10	1,000	10,000	10	1,000	10,000	10	1,000	10,000
Classic IV	-0.104	0.214	-0.015	0.285	0.144	-0.009	14.27	35.41	0.216	42.83	44.52	0.832
BB $E(\theta\|Z)$	-0.174	0.909	-0.020	0.347	0.259	-0.015	50.01	27.32	0.221	42.21	45.36	0.851
BB $med(\theta\|Z)$	0.240	0.190	-0.015	0.287	0.227	-0.009	2.369	1.247	0.216	9.491	5.137	0.834
$E(\theta\|Z)$	0.323	0.269	-0.007	0.324	0.292	-0.003	0.979	0.640	0.211	3.667	2.447	0.815
$med(\theta\|Z)$	0.324	0.261	-0.002	0.326	0.290	0.003	1.034	0.669	0.207	3.837	2.572	0.803
	β_1											
Classic IV	0.007	-0.016	0.001	-0.017	-0.011	0.001	1.100	2.783	0.017	3.398	3.496	0.065
BB $E(\theta\|Z)$	-0.001	-0.072	0.002	-0.019	-0.020	0.001	3.940	2.151	0.017	3.402	3.546	0.067
BB $med(\theta\|Z)$	-0.017	-0.015	0.001	-0.018	-0.017	0.001	0.186	0.098	0.017	0.725	0.404	0.066
$E(\theta\|Z)$	-0.023	-0.021	0.001	-0.020	-0.022	0.000	0.077	0.050	0.017	0.295	0.193	0.064
$med(\theta\|Z)$	-0.023	-0.020	0.000	-0.021	-0.022	0.000	0.081	0.052	0.016	0.307	0.200	0.063

Table 1: Results for the linear regression with an instrument using $1,000$ replications sampling with replacement. The bias of the mean is the difference of the mean of the replications and the true value (using all 162,512 data points). The bias of the median is the median of the replications minus the true value. The 95% confidence region (CR) length is the length of 95% of the replications placing 2.5% of the mass in each tail. RMSE is the root mean square error over the replications. BB denotes the (non-informative) Bayesian bootstrap. med denotes median. The last two rows are posterior mean and posterior median of the Bayesian model with weakly informative prior.

curves correspond to the classical IV estimator. They exhibit a very imprecise estimator and assign significant probabilities to economically irrelevant values of β (this is a well known disappointing property of this estimator, e.g. Bound et al. (1995)). The mean of the Bayesian bootstrapping estimator of Chamberlain and Imbens (2003) has a very large variance too (the orange curves), but its median is more precise (the yellow curves). The Bayesian estimators (that are the mean and the median of the posterior) are the most precise estimators.

The bias (with its standard error) and the root mean square error (RMSE) of the estimators have been reported in Table 1. Although the Bayesian estimators are slightly biased, thanks to their small variances they have lower RMSEs. In the Table and Figure 12 we have also reported the length of the 95% confidence intervals of the sampling distribution of the estimators (over the 1,000 replications) of β_0 and β_1 for different sample sizes $J = 10, 100, 1,000, 5,000, 10,000, 40,000$ and $100,000$. This shows that the Bayesian estimators are far more accurate than the classical IV estimator and Bayesian bootstrapping for most sample sizes. However, when J hits around $100,000$ the old methods catchup to our techniques.

Why does our method do better? For weakly identified models even a very modestly informative prior, which downweights economically implausible values of the parameter space, has the trait of cutting off the tails of the posterior corresponding to these implausible values. Because of the ridge-like posterior induced by the weakly informative likelihood, the posterior contracts onto a

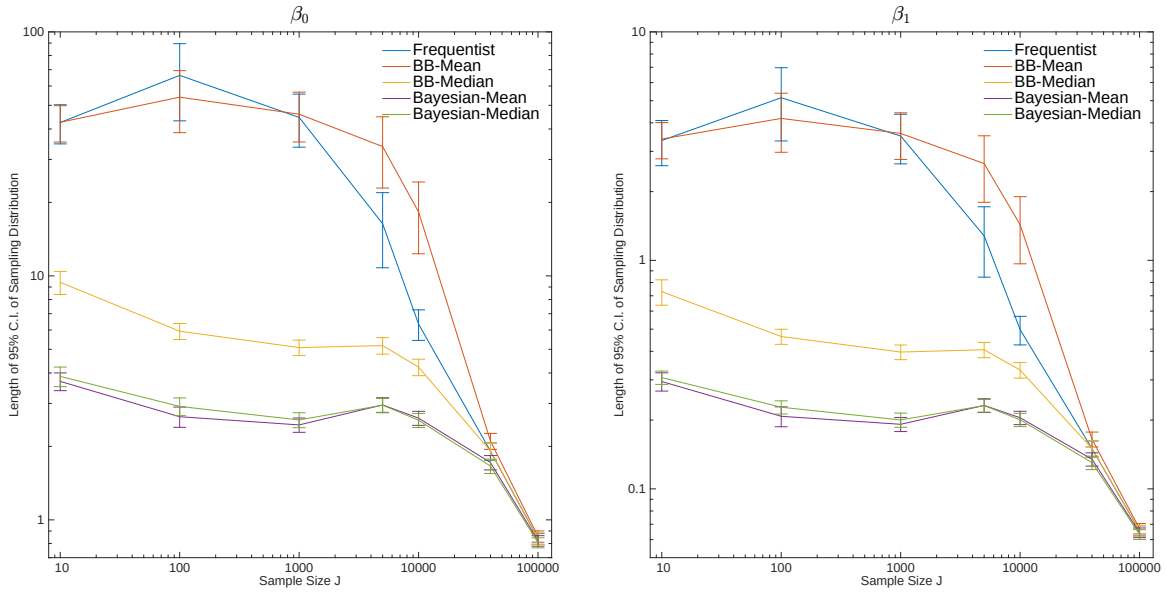

Figure 12: The length of the 95% confidence intervals of the sampling distribution of the parameters β_0 and β_1 for different sample sizes $J = 10, 100, 1,000, 5,000, 10,000, 40,000$ and $100,000$, and for classical IV estimator, Bayesian bootstrapping (mean and median) and Bayesian (mean and median). The bars denote our estimated 95% confidence intervals estimates of the lengths.

manifold, rather than a single point. As such, having a prior which constrains the feasible support provides significant value.

In the Appendix A.6 we have relaxed the assumption that the support of Z is fully observed in our sample. It can be seen that the estimates would not change significantly as long as α, the parameter of the Dirichlet distribution in the prior of θ^*, is small. It can be shown that, when $\alpha \to 0^+$, the marginal posterior distribution of θ and β of both models coincide.

6.2 Causal Inference

In this example we analyze a dataset originally collected and studied in Imbens et al. (2001). The dataset contains socioeconomic variables of 496 individuals who had won monetary prizes in the Massachusetts lottery. Following Imbens and Rubin (2015), we call the individuals who won large sums of money "the winners" (237 observations), and the ones who won only small amounts "the losers" (259 observations). The goal is to study the effect of unearned income on the economic behavior of the subjects, more specifically, on their average labor income over the first six years

following the year in which they had won the lottery. For each individual the treatment indicator, W_i, is equal to one for the winners and zero for the losers. The uncontroversial assumption behind this study is the random treatment assignment, however one may argue that the sample is not representative of the population. For instance in the literature it is well documented that the lottery players are slightly more likely to be male and middle-aged, with lower income and less education (see Clotfelter and Cook (1989), Farrel and Walker (1999) and Ariyabuddhiphongs (2011), among others).

The dataset includes the year in which the winning lottery ticket is purchased (`YW`), the number of tickets purchased in a typical week (`TB`), the individual's age (`Age`), gender (`G`) and years of schooling (`YS`), an indicator showing whether she has been working during the year the winning ticket is purchased (`WT`), and the annual social security earnings from 6 years prior to the year in which the winning ticket is purchased (`EYB1` to `EYB6`) to 6 years after that (`EYA1` to `EYA6`), all converted to 1986 dollars. The authors argue, perhaps optimistically, that the social security income is potentially the most reliable measure of income in long run, although it is capped to the maximum taxable earning ($\$42,000$ in 1986).

In order to improve the overlap of the background variables, following the recommendation of Imbens and Rubin (2015), initially we model the propensity scores using a logistic regression model, and estimate the model's parameter using the Bayesian bootstrapping of Chamberlain and Imbens (2003). The covariates of the model are a constant, the linear terms `TB`, `YS`, `WT`, `EYB1`, `Age`, `YW`, the indicator for the positiveness of the earning 5 years before winning the lottery (`SEYB5`), `G`, and the quadratic terms `YW` $\times$ `YW`, `EYB1` $\times$ `G`, `TB` $\times$ `TB`, `TB` $\times$ `WT`, `YS` $\times$ `YS`, `YS` $\times$ `EYB1`, `TB` $\times$ `YS`, `EYB1` $\times$ `Age`, `Age` $\times$ `Age`, and `YW` $\times$ `G`. We discard the observations with too small (< 0.0891) or too large (> 0.9109) estimates of propensity scores. This results in a sample of size $N = 295$ (142 winners and 153 losers). In the proposed model the propensity score is regressed on 13 covariates using a logistic regression. The vector of covariates is denoted by X_i, and include a constant, the linear terms `TB`, `YS`, `WT`, `EYB1`, `Age`, `SEYB5`, `YW`, `EYB5`, and the quadratic terms `YW` $\times$ `YW`, `TB` $\times$ `YW`, `TB` $\times$ `TB`, and `WT` $\times$ `YW`. For details on the variable selection see Imbens and Rubin (2015). The outcome, Y_i, is the average of the individual's income averaged over the first 6 years after purchasing the winning lottery ticket. Therefore the parameters of the logistic regression model, γ, and the ATE, τ, satisfy the following moment conditions,

$$\mathbb{E}\left[g(Z_i, \beta)\right] = 0, \tag{40}$$

in which, $Z_i = (X_i, Y_i, W_i)$, $\beta = (\gamma, \tau)$, and,

$$g(Z_i, \beta) = \begin{bmatrix} X_i(Y_i - \eta_i) \\ \frac{(W_i - \eta_i)Y_i}{\eta_i(1-\eta_i)} - \tau \end{bmatrix}, \tag{41}$$

where $\eta_i = \frac{\exp(\gamma' X_i)}{1+\exp(\gamma' X_i)}$. If we assume Z_is are i.i.d. draws from a discrete distribution supported on $\{s_1, ..., s_J\}$, with $\mathbb{P}(Z_i = s_j) = \theta_j$, the parameters (β, θ) will satisfy the following system of equations,

$$\begin{bmatrix} \sum_{j=1}^{J} \theta_j x_j (y_j - \eta_j) \\ \sum_{j=1}^{J} \theta_j \frac{(w_j - \eta_j) y_j}{\eta_j (1-\eta_j)} - \tau \end{bmatrix} = 0. \tag{42}$$

We let the prior of (β, θ) be

$$p(\beta, \theta) \propto \frac{1}{\sqrt{J_\theta J_\theta' + I_{14}}} \eta(\gamma)\eta(\tau)\eta(\theta) 1_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{43}$$

in which the initial prior of the regression coefficients, $\eta(\gamma)$, is a normal distribution centered at their estimates obtained from the Bayesian bootstrap of Chamberlain and Imbens (2003) and its covariance matrix is equal to the covariance matrix of estimates scaled by a factor of 100, and the initial prior of ATE is a zero mean normal distribution with variance equal to 100. Moreover we use a symmetric Dirichlet distribution with parameter $\alpha = 10^{-6}$ as the initial prior on θ.

By reweighting draws from the posterior distribution of the Bayesian bootstrap of Chamberlain and Imbens (2003), we obtain $10,000$ independent draws from the posterior of our model. An estimate of the posterior distribution of the ATE is depicted in Figure 13. *A posteriori* the expected value of ATE is $-\$5,346$ (with 95% credible interval of $[-\$8,069, -\$2,720]$). This indicates that the average income of the winners of the lotteries, in the years after winning the prize, tend to slightly decrease. Our estimate of ATE is only slightly different from the frequentist estimate.

7 Conclusions

In this paper we have provided a coherent Bayesian calculus for rational nonparametric moment based estimators, allowing users to specify scientifically meaningful priors. At the core of our analysis is a prior density placed on the Hausdorff measure whose support is generated by the scientific parameters of interest and the nonparametric probabilities. We show how to transform this prior into a posterior density.

Much moment based analysis favoured in the literature delivers weakly identified parameters. The use of very modest priors can dramatically improve estimation by downweighting vast regions of economically implausible parameter values. Such weak priors play little role when the data is informative but provide a safety net when this is not the case.

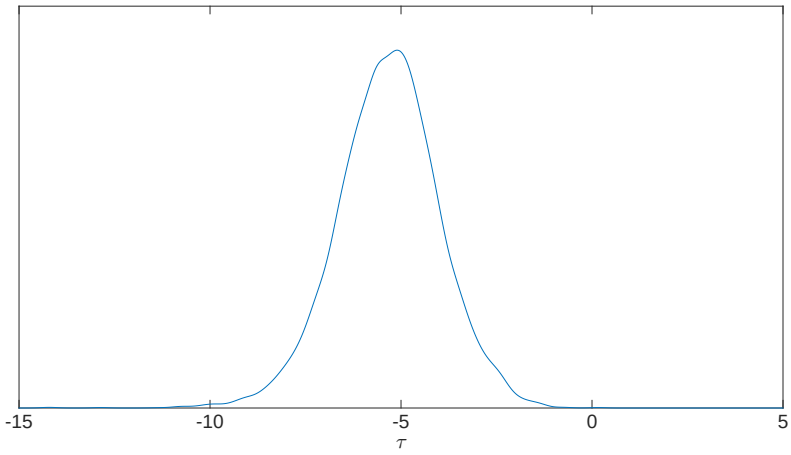

Figure 13: The posterior distribution of the average treatment effect (ATE) on subsequent annual earnings of a substantial lottery win for the lottery data set.

To harness these gains, at the center of our paper are the marginal method and the joint method. The first is based on finding the density of the probabilities with respect to a Lebesgue measure. This allows for the use of conventional simulation methods such as MCMC, importance sampling and Hamiltonian Monte Carlo. It is convenient to use where the moment conditions can be solved analytically or numerically very fast.

Our joint method is somewhat harder to code but has the virtue of never having to solve the moment equations. This has some speed advantages but more fundamentally allows the rational analysis of moment condition models with many solutions. As a side product our method provides a novel way of generically simulating on a wide class of manifolds, which may be useful in other areas of science.

References

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*, 979–1014.

Antoine, A., H. Bonnal, and E. Renault (2007). On the efficient use of the informational content of estimating equations: implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics 138*, 461–487.

Ariyabuddhiphongs, V. (2011). Lottery gambling: A review. *Journal of Gambling Studies 27*, 15–33.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. In J. J.

Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 5*, pp. 3705–3843. North Holland.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*, 443–450.

Britton, J., N. Shephard, and A. Vignoles (2015). Comparing sample survey measures of English earnings of graduates with administrative data. Unpublished paper: Department of Economics, Harvard University.

Brubaker, M., M. Salzmann, and R. Urtasun (2012). A family of MCMC methods on implicitly defined manifolds. *In: JMLR Workshop and Conference Proceedings 20*, 161–172.

Byrne, S. and M. Girolami (2013). Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*, 825–845.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–334.

Chamberlain, G. and G. Imbens (2003). Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics 21*, 12–18.

Chernozhukov, V. and H. Hong (2003). An MCMC approach to classical inference. *Journal of Econometrics 115*, 293–346.

Chiu, G. S. (2008). On identifiability of covariance components in hierarchical generalized analysis of covariance models. Unpublished paper: Department of Statistics and Actuarial Science, University of Waterloo.

Clotfelter, C. and P. Cook (1989). *Selling hope: State lotteries in America*. Cambridge, MA: Harvard University Press.

Cox, D. R. (1961). Tests of seperate families of hypotheses. *Proceedings of the Berkeley Symposium 4*, 105–123.

Diaconis, P., S. Holmes, and M. Shahshahani (2013). Sampling from a manifold. In G. Jones and X. Shen (Eds.), *Advances in Modern Statistical Theory and Applications*. Institute of Mathematical Statistics.

Doss, H. (1985). Bayesian nonparametric estimation of the median; part i: Computation of the estimates. *The Annals of Statistics 13*, 1432–1444.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society, Series B 22*, 139–153.

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Farrel, L. and I. Walker (1999). The welfare effects of lotto: Evidence from the UK. *Journal of Public Economics 72*, 99–120.

Federer, H. (1969). *Geometric Measure Theory*. New York: Springer–Verlag.

Fiorentini, G., E. Sentana, and N. Shephard (2004). Likelihood-based estimation of latent generalised ARCH structures. *Econometrica 72*, 1481–1517.

Florens, J. and A. Simoni (2015). Gaussian processes and Bayesian moment estimation. Unpublished paper: CREST.

Gallant, A. R. (2015). Reflections on the probability space induced by moment conditions with implications for Bayesian inference. *Journal of Financial Econometrics*. Forthcoming.

Gallant, A. R., R. Giacomini, and G. Ragusa (2014). Generalized method of moments with latent variables. Unpublished paper: Department of Economics, University College London.

Gallant, A. R. and H. Hong (2007). A statistical inquiry into the plausibility of recursive utility. *Journal of Financial Econometrics 5*, 523–559.

Gallant, A. R. and G. Tauchen (1996). Which moments to match. *Econometric Theory 12*, 657–81.

Gelfand, A. E., A. F. M. Smith, and T.-M. Lee (1992). Bayesian analysis of constrained parameter and truncated data problems. *Journal of the American Statistical Association 87*, 523–532.

Gelman, A., J. B. Carlin, D. B. D. Hal S Stern, A. Vehtari, and D. B. Rubin (2003). *Bayesian Data Analysis* (3 ed.). London: Chapman & Hall.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–39.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics 31*, 1208–1212.

Golchi, S. and D. A. Campbell (2014). Sequentially constrainted Monte Carlo. Unpublished paper: Department of Statistics and Actuarial Science, Simon Fraser University.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*, S85–S118.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika 82*, 711–32.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford: Oxford University Press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–54.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics 14*, 262–280.

Hurn, M. A., H. Rue, and N. A. Sheehan (1999). Block updating in constrained Markov chain Monte Carlo sampling. *Statistics and Probability Letters 41*, 353–361.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University.

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *The American Economic Review 91*, 778–794.

Imbens, G. W., R. H. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*, 333–358.

Jaynes, E. (2003). *Probability Theory*. Cambridge University Press.

Khatri, C. G. (1968). Some results for the singular normal multivariate regression models. *Sankhya: The Indian Journal of Statistics, Series A 30*(3), 267–280.

Kitamura, Y. (2007). Empirical likelihood methods in econometrics: Theory and practice. In R. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress Volume 3*, pp. 174–237. Cambridge: Cambridge University Press.

Kitamura, Y. and T. Otsu (2011). Bayesian analysis of moment restriction models using nonparametric priors. Unpublished paper: Department of Economics, Yale University.

Kolmogorov, A. (1956). *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company.

Kwan, Y. K. (1998). Asymptotic Bayesian analysis based on a limited information estimator. *Journal of Econometrics 88*, 99–121.

Lancaster, T. and S. J. Jun (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics 25*, 287–307.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika 90*, 319–326.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis of Missing Data* (2 ed.). New York: Wiley.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Marshall, A. (1956). The use of multi-stage sampling schemes in Monte Carlo computations. In M. Meyer (Ed.), *Symposium on Monte Carlo Methods*, pp. 123–140. New York: Wiley.

McCandless, L. C., P. Gustafson, and P. C. Austin (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine 28*, 94–112.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

Mengersen, K., P. Pudlo, and C. P. Robert (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences 110*, 1321–1326.

Muller, U. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica 81*, 1805–1849.

Newton, M., C. Czado, and R. Chappell (1996). Semiparametric bayesian inference for binary regression. *The Journal of the American Statistical Association 91*, 142–153.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*(1), 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics 19*, 1725–1747.

Owen, A. (2001). *Empirical Likelihood*. London: Chapman and Hall.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, Series A 185*, 71–110.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*, 300–325.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics 9*, 130–134.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 395–402. Oxford: Oxford University Press Press.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica 26*, 393–415.

Sargan, J. D. (1959). The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society, Series B 21*, 91–105.

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika 92*, 31–46.

Shin, M. (2014). Bayesian GMM. Unpublished paper: Department of Economics, University of Pennsylvania.

Strachan, R. W. and H. K. van Dijk (2004). Valuing structure, model uncertainty and model averaging in vector autoregressive processes. Econometric Institute Report EI 2004-23.

Sun, D., P. L. Speckman, and R. K. Tsutakawa (1999). Random effects in generalized linear mixed models. Unpublished paper: National Institute of Statistical Sciences.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton methods. *Biometrika 61*, 439–47.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7*, pp. 733–742. Oxford: Oxford University Press.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

Yang, Y. and X. He (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics 40*, 1102–1131.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis 4*, 191–208.

Zellner, A. (1997). The Bayesian method of moments (BMOM): theory and applications. *Advances in Econometrics 12*, 85–105.

Zellner, A., J. Tobias, and H. Ryu (1997). Bayesian method of moments (BMOM) analysis of parametric and semiparametric regression models. *In Proceedings of the Section on Bayesian Statistical Science, Alexandria, Virginia: American Statistical Association*, 211–216.

Zigler, C. M. and F. Dominici (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *Journal of American Statistical Association 109*, 95–107.

Zigler, C. M., K. Watts, R. W. Yeh, Y. Wang, B. A. Coull, and F. Dominici (2013). Model feedback in Bayesian propensity score estimation. *Biometrics 69*, 263–273.

A Appendices

A.1 Proof of proposition 1

Since corresponding to every $\theta \in \Theta_\theta$ there is a unique β, there exist a one-to-one mapping between $\Theta_{\beta,\theta}$ and Θ_θ: $(\beta, \theta) = \{\beta(\theta), \theta)\} = F(\theta)$. Now let A be a measurable set on $\Theta_{\beta,\theta}$, and assume $S_\theta(A)$ is its projection on Θ_θ. Therefore

$$\mathbb{P}(S_\theta(A)) = \mathbb{P}(A) = \int_A p(\beta, \theta) dA = \int_{S_\theta(A)} \|v_1 \wedge \cdots \wedge v_{J-1}\| \, p(\beta, \theta) dS$$

where $v_j = \frac{\partial F}{\partial \theta_j}$ (for $1 \leq j \leq J-1$). Therefore $\|v_1 \wedge \cdots \wedge v_{J-1}\| \, p(\beta, \theta)$ is the density of θ with respect to Lebesgue measure. Moreover,

$$\|v_1 \wedge \cdots \wedge v_{J-1}\| = [Gram(v_1, ..., v_{J-1})]^{\frac{1}{2}} = \left|J_\theta' J_\theta + I_{J-1}\right|^{\frac{1}{2}} = \left|J_\theta J_\theta' + I_p\right|^{\frac{1}{2}}$$

where $Gram(\cdot)$ is the Gramian determinant and $J_\theta = \partial\beta / \partial\theta^{'}$.

A.2 Proof of proposition 2

Let $p(\beta)$ be the density of β. Then, given β, the vector of probabilities θ lives on a $J-1-p$ dimensional hyperplane in $\mathbb{R}^{J-1}$ defined by $H\theta + g_J = 0$. This system of equation can be solved for p elements of the variables $\theta_{J-p:J-1} = -H_2^{-1}(H_1\theta_{1:J-p-1} - g_J)$, where $H_1 = [h_1 \ ... \ h_{J-p-1}]$ and $H_2 = [h_{J-p} \ ... \ h_{J-1}]$. Therefore, $\partial\theta_{J-p:J-1} / \partial\theta_{1:J-p-1} = -H_2^{-1}H_1$ and so

$$\begin{aligned} p(\theta_{1:J-p-1}|\beta) &= \left|H_2^{-1}H_1H_1'H_2^{'-1} + I_p\right|^{\frac{1}{2}} p(\theta|\beta), \\ p(\theta_{1:J-p-1}, \beta) &= \left|H_2^{-1}H_1H_1'H_2^{'-1} + I_p\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta). \end{aligned}$$

Therefore the density of θ is

$$\begin{aligned}
p(\theta) &= \left|\frac{\partial(\theta_{1:J-p-1},\beta)}{\partial(\theta)}\right| p(\theta_{1:J-p-1},\beta) = \left|\frac{\partial\beta}{\partial\theta_{J-p:J-1}}\right| p(\theta_{1:J-p-1},\beta) \\
&= \left|\left[\mathbb{E}\left(\frac{\partial g}{\partial\beta'}\right)\right]^{-1} H_2\right| p(\theta_{1:J-p-1},\beta) \\
&= \left|\left[\mathbb{E}\left(\frac{\partial g}{\partial\beta'}\right)\right]^{-1} H_2\right| \left|H_2^{-1}H_1H_1'H_2'^{-1} + I_p\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta) \\
&= \left|\left[\mathbb{E}\left(\frac{\partial g}{\partial\beta'}\right)\right]^{-1}\right| \left|H_1H_1' + H_2H_2'\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta) = \left|\left[\mathbb{E}\left(\frac{\partial g}{\partial\beta'}\right)\right]^{-1}\right| \left|HH'\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta) \\
&= \left|\left[\mathbb{E}\left(\frac{\partial g}{\partial\beta'}\right)\right]^{-1} HH' \left[\mathbb{E}\left(\frac{\partial g'}{\partial\beta}\right)\right]^{-1}\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta) = \left|\frac{\partial\beta}{\partial\theta'}\frac{\partial\beta'}{\partial\theta}\right|^{\frac{1}{2}} p(\beta)p(\theta|\beta).
\end{aligned}$$

Therefore:

$$p(\beta,\theta) = \frac{\left|\frac{\partial\beta}{\partial\theta'}\frac{\partial\beta'}{\partial\theta}\right|^{\frac{1}{2}}}{\left|\frac{\partial\beta}{\partial\theta'}\frac{\partial\beta'}{\partial\theta} + I_p\right|^{\frac{1}{2}}} p(\beta)p(\theta|\beta).$$

A.3 Joint method proposal

In order to generate a proposal value for θ^*, we can first draw π^* from $\mathcal{N}(\theta,\Sigma_Q)$, and let θ^* be the closest point to π^* in the hyperplane $\mathcal{P}^* = \{\lambda\in\mathbb{R}^{J-1}; H^*\lambda + g_J^* = 0\}$, where we measure the distance between π^* and θ^* with the squared Euclidean norm:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \frac{1}{2}\left\|\pi^*-\theta\right\|_2^2 + \frac{1}{2}\left(\iota'\pi^* - \iota'\theta\right)^2.$$

The quadratic penalty is certainly inelegant (e.g. compared to the log-likelihood of the multinomial model, but see, for example, Owen (1991) and Antoine et al. (2007) who use it for their Euclidean empirical likelihood) as the resulting θ^* can have negative elements or may result in $\theta_J^* = 1-\iota'\theta^* \leq 0$. However, by using a quadratic penalty, θ^* becomes the solution to a quadratic optimization problem subject to p equality constraints, and so has an analytic solution $\theta^* = a^* + B^*\pi^*$.

The Lagrangian of the optimization is,

$$E(\theta,\lambda) = \left\|\pi^*-\theta\right\|_2^2 + \left(\iota'\pi^* - \iota'\theta\right)^2 + \lambda'^*\theta + g_J^*)$$

and the first order conditions are:

$$\frac{\partial E}{\partial\theta} = (I + \iota\iota'^* - \pi^*) + H^{*'}\lambda = 0, \quad \frac{\partial E}{\partial\lambda} = H^*\theta^* + g_J^* = 0.$$

Solving them for θ^* and λ results in,

$$\theta^* = \pi^* - (I + \iota\iota'^{-1}H^{*'}\left[H^*(I + \iota\iota'^{-1}H^{*'}\right]^{-1}(H^*\pi^* + g_J^*)$$

$$\lambda = \left[H^*(I+\iota\iota'^{-1}H^{*'}\right]^{-1}(H^*\pi^*+g_J^*).$$

Therefore θ^* is an affine transformation of π^*: $\theta^* = a^* + B^*\pi^*$, where

$$\begin{aligned} a^* &= -(I+\iota\iota'^{-1}H^{*'}\left[H^*(I+\iota\iota'^{-1}H^{*'}\right]^{-1}g_J^* \\ B^* &= I-(I+\iota\iota'^{-1}H^{*'}\left[H^*(I+\iota\iota'^{-1}H^{*'}\right]^{-1}H^*. \end{aligned}$$

This transformation from π^* to θ^* is a many-to-one affine transformation. Consequently, $\theta^*|\beta^*, \beta^{(t)}, \theta^{(t)}$ is a singular normal distribution with mean $a^*+B^*\theta^{(t)}$ and variance matrix $B^*\Sigma_Q B^*$.

A singular normal distribution with mean μ and (singular) variance matrix Σ has a density on the range of the covariance matrix (e.g. Khatri (1968)), given by

$$(2\pi)^{-\frac{1}{2}rank(\Sigma)}|\Sigma|_{rank(\Sigma)}^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^+(x-\mu)\right\},$$

where $|\Sigma|_{rank(\Sigma)}$ is the product of non-zero eigenvalues of Σ and Σ^+ is its Moore-Penrose inverse.

In our algorithm, Σ_Q and the parameters inside $q(\cdot|\beta^{(t)}, \theta^{(t)})$ are the tuning parameters. We may either adapt them in the course of simulation, or they can be set to some fixed values obtained from an estimate of the posterior's distribution. Here we document how we have carried this out for our simulation and empirical work. A simple to calculate candidate for the covariance of β's proposal is $\Sigma_\beta = \left(\Sigma_{0_{\boldsymbol{\beta}}}^{-1} + \Sigma_{BB_{\boldsymbol{\beta}}}^{-1}\right)^{-1}$, where $\Sigma_{0_{\boldsymbol{\beta}}}^{-1}$ is the prior's covariance and $\Sigma_{BB_{\boldsymbol{\beta}}}^{-1}$ is the covariance of the estimates of β obtained by Bayesian bootstrapping of Chamberlain and Imbens (2003) (As an alternative we may use the asymptotic covariance of the least squares or GMM estimators). Moreover a suitable candidate for Σ_Q is $diag(\hat{\theta}_1^2, ..., \hat{\theta}_{J-1}^2)$ where:

$$\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_{J-1}) = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{J} n_j \ln\theta_j \quad \text{subject to} \quad \hat{H}\theta + \hat{g}_J = 0, \tag{44}$$

in which $\hat{H} = (\hat{g}_1, ..., \hat{g}_{J-1}) - \hat{g}_J\iota'$, $\hat{g}_j = g(\hat{\beta}, s_j)$ and $\hat{\beta} = \left(\Sigma_{0_\beta} + \Sigma_{BB_\beta}\right)^{-1}\left(\Sigma_{0_\beta}\mu_{BB_\beta} + \Sigma_{BB_\beta}\mu_{0_\beta}\right)$.

A.4 Large support

An apparent drawback of the joint method is that in each evaluation of the proposal's density, the Moore-Penrose inverse of the $(J-1)\times(J-1)$ matrix $B^*\Sigma_Q B^{*\prime}$ should be computed. In general this costs $O(J^3)$ computational operations. This type of challenge is very common in Bayesian analysis and a standard approach to this problem is to make proposals to update a block of $K \ll J$ elements of θ, with cost $O(K^3)$.

Let the $K\times 1$ vector u be a randomly (without replacement) selected subset of the indices $\{1, ..., J-1\}$ and the $(J-K-1)\times 1$ vector v be its complement. Moreover let $\tilde{\theta} = (\theta_{u_1}, ..., \theta_{u_K})$

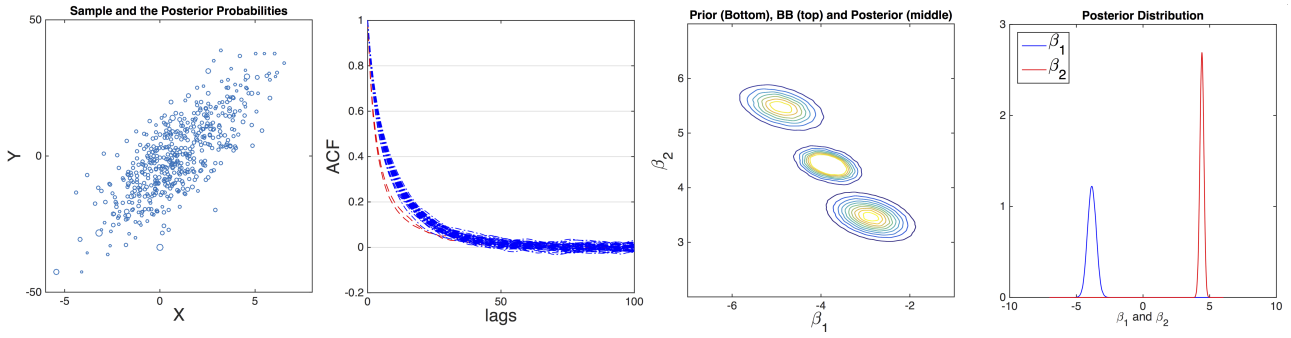

Figure 14: Inference in linear regression model with $J = 500$ and an informative prior.

and $\bar{\theta} = (\theta_{v_1}, ..., \theta_{v_{J-K-1}})$. The proposal's vector of probabilities, θ^*, is equal to θ except for the K elements with indices in u, $\tilde{\theta}^* = (\theta^*_{u_1}, ..., \theta^*_{u_K})$, that is obtained by solving:

$$\tilde{\theta}^* = \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{2}\|\tilde{\theta} - \tilde{\pi}^*\| + \frac{1}{2}\left(\iota'\tilde{\theta} - \iota'\tilde{\pi}^*\right) \quad \text{subject to} \quad \bar{H}^*\bar{\theta}^{(t)} + \tilde{H}^*\tilde{\theta} + g^*_J = 0, \tag{45}$$

where $\tilde{H}^* = \left(g^*_{u_1}, ..., g^*_{u_K}\right) - g^*_J\iota'$, $\bar{H}^* = \left(g^*_{v_1}, ..., g^*_{v_{J-K-1}}\right) - g^*_J\iota'$, and $\tilde{\pi}^*$ is a random draw from $N(\tilde{\theta}, \Sigma_{\tilde{\theta}})$. Again this is a quadratic optimization problem subject to a set of equality constraints with the following solution: $\tilde{\theta}^* = \tilde{a}^* + \tilde{B}^*\tilde{\pi}^*$, where

$$\begin{aligned}
\tilde{a}^* &= -(I + \iota\iota'^{-1}\tilde{H}^{*\prime}\left[\tilde{H}^*(I + \iota\iota'^{-1}\tilde{H}^{*\prime}\right]^{-1}\left(\bar{H}^*\bar{\theta}^{(t)} + g^*_J\right) \\
\tilde{B}^* &= I - (I + \iota\iota'^{-1}\tilde{H}^{*\prime}\left[\tilde{H}^*(I + \iota\iota'^{-1}\tilde{H}^{*\prime}\right]^{-1}\tilde{H}^*.
\end{aligned}$$

A.5 Linear regression with an informative prior

Here we report the results for the linear regression model with sample size $J = 500$, and an informative prior for β. We place a normal prior on β with the mean equal to $\hat{\beta}_{\text{MLE}} + (5, -5)'$ and the variance equal to the asymptotic variance of $\hat{\beta}_{\text{MLE}}$. Therefore the prior is as informative as the data, however centered at a significantly different point.

Figure 14's top left panel shows a scatter plot of the sample. Each circle represents a data point and its radius is proportional to $\mathbb{E}(\theta_j|Z)$. In the top-right the ACF of the chains of β and 50 elements of θ have been presented (the red dashed lines and the blue dotted lines correspond to β and θ, respectively.) These show that the Markov chain is mixing sufficiently well. In the bottom-left panel the contour plot of the prior distribution (bottom), posterior distribution of β using Bayesian bootstrapping, and the posterior distribution of β considering the informative prior (middle) have been depicted. In the bottom-right panel the histogram of the samples from the posterior of β can be seen.

A.6 Instrumental variables with partially observed support

Now we assume the support of $S = (X, Y, W)$ has other J missing elements (not observed in the data), therefore $J^* = 2J$. The density of our prior for the missing elements of the support is,

$$f_S(\tilde{s}) = f_{S^{(1)}}(\tilde{s}^{(1)}) f_{S^{(2)}}(\tilde{s}^{(2)}) f_{S^{(3)}}(\tilde{s}^{(3)}) \tag{46}$$

in which $f_{S^{(1)}}(\tilde{s}^{(1)})$ and $f_{S^{(3)}}(\tilde{s}^{(3)})$ are the density of a uniform distribution on $\{0, 1, ..., 20\}$ and $\{0, 1\}$, respectively, and $f_{S^{(2)}}(\tilde{s}^{(2)})$ is a normal density with mean 6 and standard deviation 3. Moreover we assume,

$$p(\beta, \theta^* | \tilde{S}) \propto \frac{1}{\sqrt{J_{\theta^*} J'_{\theta^*} + J_{\tilde{S}} J'_{\tilde{S}} + I_2}} \eta(\beta) \eta(\theta^*) 1_{\Theta_{\beta, \theta^*, \tilde{S}}}(\beta, \theta^*, \tilde{S}), \tag{47}$$

where $\eta(\beta) = \varphi(\beta_0; 5, 4)\, \varphi(\beta_1; 0, 0.2)$ (similar to the previous case), and $\eta(\theta^*)$ is the density a symmetric Dirichlet distribution with $\alpha = 10^{-6}$. Hence the posterior distribution of $(\theta^*, \tilde{S})$ is:

$$p(\theta^*, \tilde{S} | Z) \propto \eta(\beta) \left(\prod_{j=1}^{\tilde{J}} f_S(\tilde{s}_j) \right) \left(\prod_{j=1}^{J^*} \theta_j^{*\, n_j + \alpha - 1} \right). \tag{48}$$

To sample from this distribution we can reweight random draws from the following proposal,

$$g(\theta^*, \tilde{S} | Z) \propto \left(\prod_{j=1}^{\tilde{J}} f_S(\tilde{s}_j) \right) \left(\prod_{j=1}^{J^*} \theta_j^{*\, n_j + \alpha - 1} \right), \tag{49}$$

with he weights proportional to $\eta(\beta)$. Now we set $J = 10$, and for 1000 times we draw a random sample from our dataset. Then we compare the posterior distribution of the parameters under two assumptions. In the first model we assume the support of S is fully observed in the data (similar to the previous section of this example), while in the second model we assume the data has $\tilde{J} = 10$ more elements that are not observed in our sample. Since the prior of the probabilities and β is barely informative the posterior distributions of β are almost indistinguishable under these two assumptions.

A.7 Not the just identified case

A.7.1 Abstract expression of the problem

Collect all the parameters in the model and constraints as

$$\psi = (\theta_1, ..., \theta_{J-1}, \beta_1, ..., \beta_p)', \quad g(\psi) = 0_r.$$

Then resulting constrained support is $\psi \in \Theta_\psi$. Write $\lambda = \psi_{\mathcal{I}}$, $\phi = \psi_{\mathcal{I}^c}$, where $\mathcal{I}$ selects distinct indexes of ψ and $\mathcal{I}^c$ is the complement, so $\mathcal{I} \cup \mathcal{I}^c = \{1, 2, ..., p + J - 1\}$. Throughout we take $\dim(\phi) = r$ and consequently $\dim(\lambda) = J - m$, where $m = r - p + 1$.

Given the freedom to build $\mathcal{I}$ we make the following assumption.

Assumption A. *Under $g(\psi) = 0$ knowledge of λ reveals ϕ, so there exists a unique $\phi = t(\lambda)$.*

A.7.2 Marginal method

Under Assumption A, the area formula implies that $p(\lambda) = p(\psi)\sqrt{\left|I_r + J_{\phi\lambda}J'_{\phi\lambda}\right|}$, $J_{\phi\lambda} = \partial\phi/\partial\lambda'$, where $p(\psi)$ is a density with respect to the $(J - 1 + p - r)$-dimensional Hausdorff measure on Θ_ψ, while $p(\lambda)$ is a density with respect to the $J - m$-dimensional Lebesgue measure.

A.7.3 Underidentification

Definition 1 *If $r < p$ (so $m \leq 0$) then the system is called underidentified.*

We split $\beta = \left(\beta_1, ..., \beta_p\right)'$ as $\beta_{[1]} = \beta_{\mathcal{J}}$, $\beta_{[2]} = \beta_{\mathcal{J}^c}$, where $\mathcal{J} \cup \mathcal{J}^c = \{1, 2, ..., p\}$, $\dim(\mathcal{J}) = p - r$ and $\dim(\mathcal{J}^c) = r$, and build $\lambda = (\theta_1, ..., \theta_{J-1}, \beta'_{[1]})'$, $\phi = \beta_{[2]}$. Hence λ augments θ with $p - r$ elements from β. Assumption A holds if $\mathcal{J}$ can be found such that $\beta_{[2]} = t(\theta_1, ..., \theta_{J-1}, \beta_{[1]})$.

Example 6 *Consider instrumental variables problem $g(s, \beta) = s^{(3)}\left\{s^{(1)} - \beta' s^{(2)}\right\}$, $\dim(s^{(2)}) = p$, $\dim(s^{(3)}) = r$. If $p > r$ then split $\beta = \left(\beta'_{[1]}, \beta'_{[2]}\right)'$, where $\dim(\beta_{[1]}) = r - p$ and $\dim(\beta_{[2]}) = r$. Write $s_j = \left(s'_{j,[1]}, s'_{j,[2]}\right)'$, then*

$$\sum_{j=1}^{J} \theta_j s_j^{(3)} \left\{\left(s_j^{(1)} - \beta'_{[1]} s_{j,[1]}^{(2)}\right) - \beta'_{[2]} s_{j,[2]}^{(2)}\right\} = 0_r.$$

Knowledge of $\beta_{[1]}$ puts us back to the just identified, so Assumption A holds under weak assumptions and so $p(\theta, \beta_{[2]})$ can be computed using the area formula.

A.7.4 Overidentification

Definition 2 *If $r > p$ so $m \geq 1$ (e.g. $r = 2$,$p = 1$, $m = 2$) then the system is called overidentified.*

We split $\theta = (\theta_1, ..., \theta_{J-1})'$ as $\theta_{[1]} = \theta_{\mathcal{J}}$, $\theta_{[2]} = \theta_{\mathcal{J}^c}$,where $\mathcal{J} \cup \mathcal{J}^c = \{1, 2, ..., J - 1\}$, $\dim(\mathcal{J}) = J - m$ and $\dim(\mathcal{J}^c) = m - 1$, and build $\lambda = \theta'_{[1]}$, $\phi = \left(\theta'_{[2]}, \beta'\right)'$. Hence λ is a subset of θ with $J - m$ elements, while ϕ contains all the other probabilities and the entire β. Then Assumption A holds if we can find a $\mathcal{J}$ such that $\left(\theta'_{[2]}, \beta'\right)' = t(\theta'_{[1]})$.

Example 7 *Again consider* $g(s,\beta) = s^{(3)}\left\{s^{(1)} - \beta' s^{(2)}\right\}$, $\dim(s^{(2)}) = p$, $\dim(s^{(3)}) = r$. *If* $p < r$ *then split* $\theta = \left(\theta'_{[1]}, \theta'_{[2]}\right)'$, *where* $\dim(\theta'_{[2]}, \beta') = r$, *so there are* r *moment conditions and* r *unknowns. Given* $\theta_{[1]}$, *we can then solve for the extended set of parameters* $\left(\theta'_{[2]}, \beta'\right)$, *where*

$$\sum_{j=1}^{\dim(\theta_{[1]})} \theta_j s_j^{(3)} \left\{s_j^{(1)} - \beta' s_j^{(2)}\right\} + \sum_{j=\dim(\theta_{[1]})+1}^{J} \theta_j s_j^{(3)} \left\{s_j^{(1)} - \beta' s_j^{(2)}\right\} = 0_r.$$

This is typically exactly identified, but non-linear due to the $\theta_j\beta$ *terms for* $j = \dim(\theta_{[1]}) + 1, ..., J$.