ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 1, *pp.* 5–43

Moment conditions and Bayesian non-parametrics

Luke Bornn,

Simon Fraser University, Burnaby, Canada

and Neil Shephard and Reza Solgi

Harvard University, Cambridge, USA

[Received January 2016. Final revision August 2018]

Summary. Models phrased through moment conditions are central to much of modern inference. Here these moment conditions are embedded within a non-parametric Bayesian set-up. 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 by using Hausdorff measures to analyse 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 non-linear regression, missing data and hierarchical models.

Keywords: Decision theory; Empirical likelihood; Hausdorff measure; Markov chain Monte Carlo methods; Method of moments; Non-parametric Bayes methods; Simulation on manifolds

1. Introduction

1.1. Overview

Much of modern inference is phrased in terms of moment conditions and analysed by using asymptotic approximations. Here moment conditions are embedded within a non-parametric Bayesian set-up, allowing an individual to mix moment conditions with data and informative priors to make rational decisions without the recourse to the veil of parametric assumptions or asymptotics.

Embedding moments within non-parametrics is not straightforward. This paper spells out the issues, develops the corresponding probability theory to solve them and devises strategies for simulating on a manifold to implement.

The range of the new methods is large. It deals with, for example, linear, non-linear and instrumental variable (IV) 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) and White (1994)). Finally, this framework provides a basis to deal systematically with missing data (e.g. Little and Rubin (2002)), to shrink parameters (e.g. Efron (2012)) and to build hierarchical models (e.g. Gelman *et al.* (2003)).

1.2. The conceptual challenge

To help to place this paper in the context of the literature we establish some notation; a formal

Address for correspondence: Neil Shephard, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.
E-mail: shephard@fas.harvard.edu

 1369–7412/19/81005

statement will appear in Section 2. Assume that the independent and identically distributed (IID) d-dimensional data Z_i, $i=1,2,\ldots,n$, take on the known support $s_1, s_2, \ldots, s_J$ and have distribution function F. 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 interest and is determined by θ, where $\theta=(\theta_1,\theta_2,\ldots,\theta_{J-1})'$ (with $\theta_J=1-\iota'\theta$, where ι is a vector of 1s) are non-parametric nuisance parameters. The task is to learn $p(\beta,\theta|Z)$ or $p(\beta|Z)$, where $Z=(Z_1,Z_2,\ldots,Z_n)'$.

A leading example of condition (1) would be where $g(s_j,\beta)$ is the score vector for the jth observation from a quasi-likelihood.

Although this problem is easy to state, it is not easily carried out, as traditional non-parametric models clash with the moment conditions, overspecifying the model. Expressing this in a different way, the prior and posterior for β and 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 $p(\beta,\theta|Z)$ will fail. To the best of our knowledge this computational challenge in the context of moment condition models was first raised and investigated by Kitamura and Otsu (2011). In this paper we propose a radically different solution that relies on defining a prior on the zero Lebesgue measure parameter space. The reader is referred to Morgan (2016) for additional measure theoretic background. Further, this approach naturally extends to the case where the support is unknown, which will be detailed in Section 3.7.

1.3. Literature on classical analysis of moments

Here we 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 $\hat{\theta}$ non-parametrically, i.e. F by the empirical distribution function F_n, and then plugs it into condition (1), yielding the function

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

In the $p=r$ case we move β around until this function equals a vector of 0s, delivering the method-of-moments estimator $\hat{\beta}$. 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 (1989, 1996) and Gourieroux *et al.* (1993). Hall (2005) gives a 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 Chamberlain and Imbens (2003). Chamberlain and Imbens (2003) placed a Dirichlet prior on θ, which implies that the posterior on θ is Dirichlet. These priors and posteriors are straightforward to sample from by using the Bayesian bootstrap. Chamberlain and Imbens (2003) suggested 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, which is the parameter of scientific interest. For instance in the IV example of Chamberlain and Imbens (2003), β is a vector with two elements: average earnings in the subpopulation with no schooling at all, and the return to schooling. Unfortunately they had no control over the prior for the parameter of interest β.

Also important is Kitamura and Otsu (2011), who had two methods, both expressed in terms of Dirichlet process priors. Here we convert them into our framework. In their exponentially tilted case they first specified a prior $p(\beta)p(\theta)$ before finding $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_J^*)$ which minimizes $\Sigma_{j=1}^J \theta_j^* \log(\theta_j^*/\theta_j)$ subject to the moment constraints $\Sigma_{j=1}^J \theta_j^* g(s_j, \beta) = 0$ 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 investigated various computational aspects of this approach. Kitamura and Otsu (2011) also proposed a synthetic Dirichlet process (with connections to Doss (1985) and Newton *et al.* (1996)).

Alternative methods include the Bayesian use of moments through approximate methods. Chernozhukov and Hong (2003) specified a quadratic form in the moment conditions and used this as the basis of a log-quasi-likelihood function. They then used this approximate likelihood to carry out Bayesian inference using MCMC sampling alongside a sandwich estimator. Related work includes Yin (2009). Muller (2013) provided a Bayesian version of the asymptotic sandwich matrix that is commonly seen in quasi-likelihood inference and linked it to decision theory.

Lazar (2003), Schennach (2005) and Yang and He (2012) provided Bayesian interpretations to empirical likelihood and studied the resulting properties. Mengersen *et al.* (2013) looked at moment conditions and empirical likelihood by using approximate Bayesian computation. See also Zellner (1997) and Zellner *et al.* (1997). 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) developed methods which devise a likelihood by 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.

In a similar setting to our problem, Kessler *et al.* (2015) proposed the marginally specified prior. In their model, an initially chosen non-parametric prior is modified in such a way that its β-marginal coincides with an informative prior distribution. Despite its mathematical elegance, sampling from its posterior distribution is not straightforward, unless the density function of the marginal of β for the initially chosen prior is known. They showed how an estimate (e.g. a kernel density estimate) of this density function can be employed in an approximative sampling scheme to sample from the posterior distribution. Estimating β's marginal could become challenging for moderate dimensions of β or complicated moment conditions (for instance the causal inference example that is presented in Section 6.2).

1.5. Computational issues

Here the prior and posterior for β and θ are supported on a zero Lebesgue measure set $\Theta_{\beta,\theta}$ in $\mathbb{R}^{J+p-1}$. Hence Bayesian inference will require samples from a distribution defined on a zero measure set, rendering standard Monte Carlo methods inadequate.

In an influential paper Gelfand *et al.* (1992) used MCMC methods to deal with constrained parameter spaces, but there the constraints do not change the dimension of the support. Hurn *et al.* (1999) carried out MCMC sampling in constrained parameter spaces (sampling from a distribution $\pi(x)$ subject to a constraint $C(x) = 0$) by using block updating. Golchi and Campbell

(2014) carried out sampling subject to constraints by using sequential Monte Carlo methods by slowly introducing the constraints. However, they did not explore the change-of-measure issue that we discuss here. Chiu (2008) used a singular normal distribution in posterior updating for an underidentified 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) faced related but highly specialized challenges when sampling missing data in a generalized auto-regressive conditional heteroscedasticity model.

MCMC simulation from distributions defined on manifolds have been recently studied. Byrne and Girolami (2013) introduced a Hamiltonian Monte Carlo algorithm for manifolds with known geodesic structure. They used this for the distributions defined on hyperspheres and Stiefel manifolds of orthonormal matrices. Diaconis *et al.* (2013) provided a short review of concepts in geometric measure theory. They discussed sampling from distributions defined on Riemannian manifolds that are similar to the 'marginal method' that will be introduced shortly. Brubaker *et al.* (2012) proposed a Hamiltonian Monte Carlo algorithm on implicitly defined manifolds. Numeric integration of the Hamiltonian dynamics requires solving a system of $3d$ non-linear equations for each update, where d is the dimension of the space in which the manifold is embedded (in our setting $d = J + p - 1$ and so is typically large). Statistical physicists have studied a similar problem in molecular dynamic simulations. For instance Lelièvre *et al.* (2012) developed a Hamiltonian Monte Carlo algorithm for distributions defined on submanifolds. See also Hartmann and Schütte (2005a, b), Hartmann (2008) and Leimkuhler and Matthews (2016). Implementation of this algorithm requires integrating the constrained Hamiltonian dynamics that includes solving a system of p non-linear equations (see also Leimkuhler and Reich (2004) and Lelièvre *et al.* (2010)). We detail the implementation of their algorithm to our problem. Besides that, taking advantage of the specific characteristics of the problem, we propose two other tailored solutions. The first algorithm, the marginal method, provides further insight into the intellectual problem that is studied here and suggests an importance sampling algorithm. In the second sampling algorithm, the joint method, we harness the special properties of the submanifold of the parameters. This gives us a Metropolis–Hastings algorithm that does not require solving for β.

1.6. Outline of the paper

In the next section of the paper we shall 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. We also draw out how to make inference when the support of the data is unknown in Section 3.7. Section 4 discusses mechanisms for generating priors for these models. This is followed by Section 5 in which some illustrative examples are demonstrated. Section 6 explores several empirical studies before Section 7 concludes. Appendix A collects the proofs and some additional results.

2. Bayesian analysis with moment conditions

2.1. The model

Assume that the data are $Z = (Z_1, \ldots, Z_n)$, where the Z_i are d-dimensional IID draws from an unknown distribution which has J points of known support $\{s_1, s_2, \ldots, s_J\} = S$. Throughout we write

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

with $\theta = (\theta_1, \theta_2, \ldots, \theta_{J-1})' \in \Theta_\theta \subseteq \Delta^{J-1}$, where $\Delta^{J-1} = \{\theta = (\theta_1, \theta_2, \ldots, \theta_{J-1})'; \iota'\theta < 1 \text{ and } \theta_j >$

$0\}$ for all j and $\theta_J = 1 - \iota'\theta$, in which ι is a vector of 1s. Our interest is learning β which solves 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 centre on 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.6 extends to the more general case of overidentification and underidentification at the cost of more clutter but without having to generate new ideas.

2.2. Parameter space and prior

Throughout β and θ must 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 the probability axioms. The moment conditions are

$$H_\beta \theta + g_J = 0 \qquad\qquad H_\beta = (g_1, \ldots, g_{J-1}) - g_J \iota',$$

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

$\Theta_{\beta,\theta}$ is a zero Lebesgue measure set in $\mathbb{R}^{J+p-1}$. We shall 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 as the base measure, we can 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 a Euclidean space. (Assume that $E \subseteq \mathbb{R}^n$, $d \in [0, \infty)$ and $\delta \in (0, \infty]$. The Hausdorff premeasure of E is defined as

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

where

$$v_m = \frac{\Gamma(\frac{1}{2})^d}{2^d \Gamma(d/2 + 1)}$$

is the volume of the unit d-sphere, and $\operatorname{diam}(E_j)$ is the diameter of E_j. $\mathcal{H}^d_\delta(E)$ is a non-increasing function of δ, and the d-dimensional Hausdorff measure of E is defined as its limit when $\delta \to 0$, $\mathcal{H}^d(E) = \lim_{\delta \to 0^+} \mathcal{H}^d_\delta(E)$. 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.)

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 non-parametric model against a parameter of interest.

2.3.1. Example (logistic)

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

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

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 Fig. 1(a). Of importance is that if θ moves by $\mathrm{d}\theta$ then the length along this curve will be (by Pythagoras's theorem)

$$\mathrm{d}\theta \sqrt{(1+\mathcal{J}_\theta^2)},$$
$$\mathcal{J}_\theta = \frac{\partial \beta}{\partial \theta} = \frac{\partial \log\{\theta/(1-\theta)\}}{\partial \theta}.$$

Fig. 1(b) 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 $C \subset \Theta_{\beta,\theta}$,

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

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 that, as we integrate over θ, we must multiply the density on the curve by the length of the curve.

We shall study how to transform this prior $p(\beta, \theta)$ into a posterior and to simulate from it. This will enable 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 shall return to that practical issue in Section 4.

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

2.3.2. Example 2 (mean)

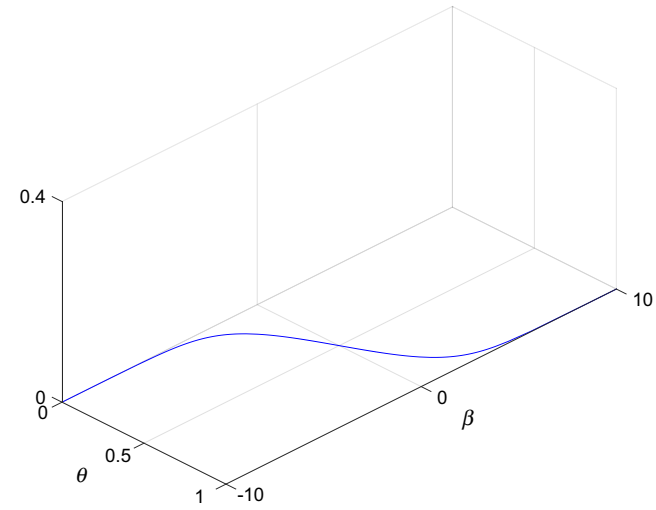
Let Z be a scalar random variable and $g(s, \beta) = s - \beta$, so β is a mean. Then

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

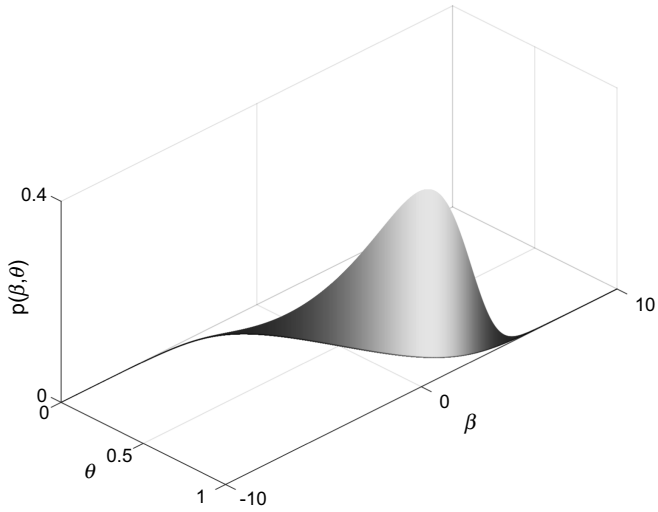
Thus $\Theta_{\beta,\theta}$ is a region within a $(J-1)$-dimensional hyperplane in $\mathbb{R}^J$. However, all elements of this set are not admissible, since θ should satisfy the probability axioms (elements of θ should be positive and $1 - \iota'\theta > 0$). Therefore the parameter space $\Theta_{\beta,\theta}$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

$$\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{J-1} \sqrt{(1 + \mathcal{J}_\theta \mathcal{J}_\theta')}, \qquad \mathcal{J}_\theta = \left\{ \left(\frac{\partial \beta}{\partial \theta_1}\right), \ldots, \left(\frac{\partial \beta}{\partial \theta_{J-1}}\right) \right\},$$

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



(a)



(b)

Fig. 1. (a) $\beta = \log\{\theta/(1-\theta)\}$ (——) is the parameter space of the logit model, $\Theta_{\beta,\theta}$; (b) density of the prior $p(\beta,\theta)$ (with respect to Hausdorff measure); this density lives on the blue curve which supports $\Theta_{\beta,\theta}$

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

where C_θ is the projection of C on θ. (The last proportionality is because the Jacobian depends only 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 $p(\beta,\theta)$.

2.3.3. Example 3 (regression)

The previous example can be generalized to the family of regression models. For instance consider a linear regression model $\mathbb{E}[s^{(1)}|s^{(2)}]=\beta' s^{(2)}$, 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:

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

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

2.3.4. Example 4 (average treatment effect)

Consider a causal inference problem with the observational data $Z_j=(X_j,Y_j,W_j)$ (for $1\leqslant j\leqslant N$), where X_j is the K-dimensional vector of the jth unit's background variables, Y_j is its scalar outcome variable and W_j is the binary treatment indicator. Assuming superpopulation unconfoundedness, it can be shown that (Imbens and Rubin, 2015) $\mathbb{E}_{\mathrm{SP}}[Y_j(1)]=\mathbb{E}[W_jY_j/e(X_j)]$ and $\mathbb{E}_{\mathrm{SP}}[Y_j(0)]=\mathbb{E}[(1-W_j)Y_j/\{1-e(X_j)\}]$, 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

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

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

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

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

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

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 that were formulated above, the model's likelihood is

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

where $n_j = \Sigma_{i=1}^{N} \mathbf{1}(Z_i = s_j)$. Although β does not appear in the likelihood explicitly, because of the constraints on β and θ, the data are informative about β.

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

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

The terms in expression (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, the prior and the posterior of the model are supported on a zero Lebesgue measure set, which makes the sampling problem challenging.

Here three solutions to this problem are given.

The first approach, which is called the 'marginal method', derives 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 algorithms 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.

The second approach, which is called the 'joint method', defines 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 submanifold $\Theta_{\beta,\theta}$ and will be accepted). A Metropolis–Hastings algorithm with this proposal can 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.

The third approach is the application of the constrained Hamiltonian Monte Carlo algorithm of Lelièvre *et al.* (2012). Integrating the constrained Hamiltonian dynamics requires, at each step, solving a system of non-linear equations.

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 methods) can be used. (We sample from the unconstrained $p(\eta)$, where $\eta_j = \log(\theta_{j+1}/\theta_j)$, for $j = 1,\ldots,J-1$, with $|\partial\theta/\partial\eta| = \Pi_{j=1}^{J}\theta_j$.)

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 that $p = r$ (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

$$p(\theta) = \sqrt{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|}\; p(\beta, \theta), \tag{5}$$

where

$$\mathcal{J}_\theta = \frac{\partial \beta}{\partial \theta'} = -\mathbb{E}_\theta\left[\frac{\partial g}{\partial \beta'}\right]^{-1} H_\beta, \qquad H_\beta = (g_1, \ldots, g_{J-1}) - g_J \iota', \tag{6}$$

with ι being a $(J-1)$-vector of 1s 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 are more than one β by summing over the right-hand side for each solution in β.

The Jacobian term $\sqrt{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|}$ depends on the geometry of the parameter space $\Theta_{\beta,\theta}$ (in other words, it depends only 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 to evaluate the determinant of a $p \times p$ matrix. However, p is usually modest, in which case the computational cost of these operations is negligible. Similar correction terms appear in reversible jump MCMC (e.g. Green (1995)) and compressible generalized hybrid Monte Carlo methods (in which the dynamics need not be volume preserving; see for instance Fang *et al.* (2014)). (In reversible jump MCMC sampling the chain is allowed to jump between models with a different number of parameters. However, there are (one-to-one) transformations operating between spaces of the same dimensions, and the distributions in both spaces have densities with respect to Lebesgue measure. In contrast, the Jacobian in proposition 1 corrects for a one-to-one mapping between two different spaces and relates two densities that are defined with respect to different reference measures.)

Importantly, knowledge of the functional form of β as a function of θ is not needed, since the partial derivatives can be obtained by thinking of the moment condition $g\{\theta, \beta(\theta)\} = \Sigma_{j=1}^{J} \theta_j\, g(s_j, \beta) = 0$ and then using the implicit function theorem

$$\frac{\partial g}{\partial \theta'} + \frac{\partial \beta}{\partial \theta'} \frac{\partial g}{\partial \beta'} = 0.$$

However, 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 that were introduced in Section 2.

3.3.1. Example 5 (continues example 1)
The density of θ in the logistic model is

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

This moment condition impacts the marginal prior on θ. Fig. 2 shows the function $p(\theta)$, which is the blue shaded area below the curve, together with the naive $p[\beta = \log\{\theta/(1-\theta)\}, \theta]$, which is the grey shaded area. We can see that 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.

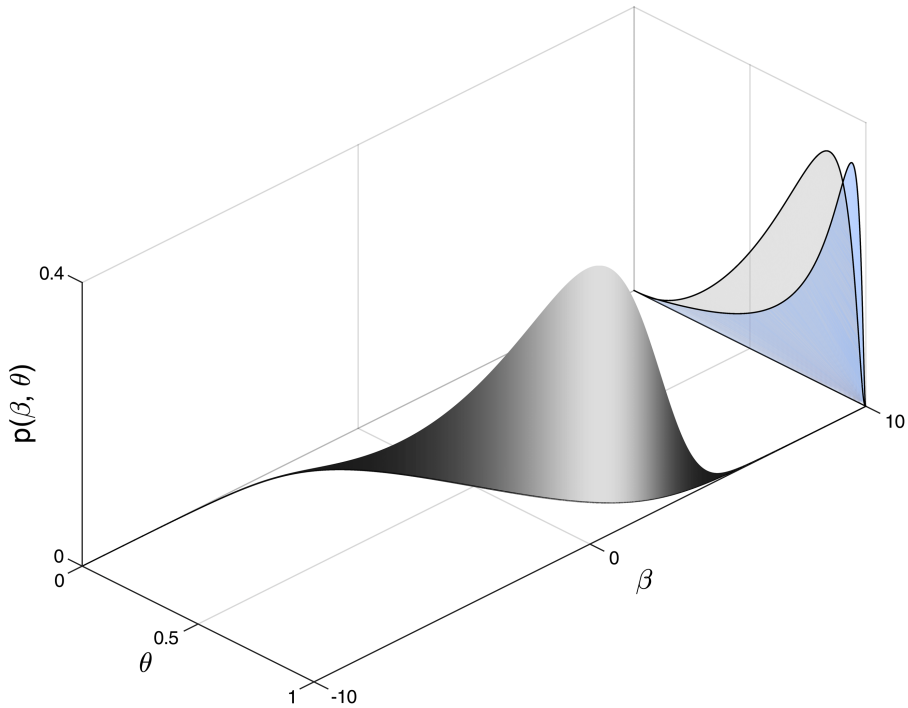


Fig. 2. Projection to the marginal density for θ: ■, correct marginal density $p(\theta)$, given in equation (7), with respect to Lebesgue measure; ■, naive density $p[\beta = \log\{\theta/(1-\theta)\}, \theta]$ which ignores the corresponding length of the support

3.3.2. Example 6 (continues example 2)
The density of θ in the mean model is

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

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

3.3.3. Example 7 (continues example 3)
For the regression model write $g_j = g(s_j, \beta)$ for $1 \leqslant j \leqslant J$. Therefore

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} s_j^{(2)'},$$

and

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

Moreover

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

Similarly for the linear regression model with IVs we have

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(3)} s_j^{(2)'},$$

and

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

and therefore

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

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

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} \frac{\partial \mu_j}{\partial \beta'},$$

and

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

which implies that

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

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

$$\frac{\partial g_j}{\partial \beta'} = -s_j^{(2)} \exp(\beta_j'^{(2)}) s_j^{(2)'},$$

and

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

and hence

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

3.3.4. Example 8 (continues example 4)

For the causal inference problem write $g_j = g(s_j, \beta)$, for $1 \leqslant j \leqslant J$. Then

$$\frac{\partial g_j}{\partial \beta'} = \begin{pmatrix} s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)'} & \mathbf{0}_{K \times 1} \\ \mathbf{0}_{1 \times K} & -1 \end{pmatrix},$$

and

$$\frac{\partial \beta}{\partial \theta_i} = \begin{pmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1 - \eta_j) s_j^{(1)'} & \mathbf{0}_{K \times 1} \\ \mathbf{0}_{1 \times K} & -1 \end{pmatrix}^{-1} (g_i - g_J),$$

which implies that

$$\mathcal{J}_\theta \mathcal{J}_\theta' = \begin{pmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1-\eta_j) s_j^{(1)'} & \mathbf{0}_{K\times 1} \\ \mathbf{0}_{1\times K} & -1 \end{pmatrix}^{-1} \left\{ \sum_{i=1}^{J} (g_i - g_J)(g_i - g_J)' \right\} \times \begin{pmatrix} \sum_{j=1}^{J} \theta_j s_j^{(1)} \eta_j (1-\eta_j) s_j^{(1)'} & \mathbf{0}_{K\times 1} \\ \mathbf{0}_{1\times K} & -1 \end{pmatrix}^{-1}.$$

An immediate consequence of proposition 1 is that, if we reparameterize the scientific parameters of interest $\psi = \psi(\beta)$ by 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 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 1. Therefore 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 (to compute the acceptance probability) we should be able to evaluate its density function. This subsection will explain how this can be achieved. The idea that we have employed here for constructing a proposal distribution is similar to the algorithm that was used by Kitamura and Otsu (2011) to define their prior distribution; however, we have applied this idea only as a computational tool (building a proposal distribution for the Metropolis–Hastings algorithm, and not as part of our model).

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 enables us to define a suitable proposal distribution for (β, θ). Assume that the current state of the MCMC algorithm is $(\beta^{(t)}, \theta^{(t)})$. First we explain how a random sample from the proposal can be drawn, and then we shall show how the density of this proposal can be evaluated. To draw a random sample from $q(\cdot, \cdot | \beta^{(t)}, \theta^{(t)})$ requires the following steps.

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

$$p(\beta, \theta) = \frac{|\mathcal{J}_\theta \mathcal{J}_\theta'|^{1/2}}{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|^{1/2}}\ p(\beta) p(\theta|\beta), \qquad \mathcal{J}_\theta = \frac{\partial \beta}{\partial \theta'}. \tag{10}$$

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 $\theta_J^* = 1 - \iota'\theta^* \leqslant 0$). Obviously in these cases the proposal is rejected (since the posterior is 0), the MCMC algorithm sticks, and the proposal's density need not 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.

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 $J \to \infty$, the Jacobian term in expression (10) converges to 1. To see this assume that the data-generating process is a continuous distribution or a discrete distribution with infinite support, $s_j \sim H$. Then, as

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

and

$$\frac{1}{J}H_\beta H_\beta' = \frac{1}{J}\sum_{j=1}^{J}(g_j - g_J)(g_j - g_J)',$$

so if the s_j are IID then

$$\frac{|\mathcal{J}_\theta \mathcal{J}_\theta'|}{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|} = \frac{|(1/J)\mathcal{J}_\theta \mathcal{J}_\theta'|}{|(1/J)\mathcal{J}_\theta \mathcal{J}_\theta' + (1/J) I_p|} \to 1,$$

with probability 1 as $J \to \infty$. 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. Constrained Hamiltonian Monte Carlo sampling

Following Lelièvre *et al.* (2012), let $q = (\beta, \theta)$ be the vector of constrained parameters and, similarly to the classical Hamiltonian Monte Carlo algorithm, introduce $p + J - 1$ auxiliary moment variables, $u = (p_\beta, p_\theta)$. The Hamiltonian of the system is defined to be

$$H(\beta, \theta, p_\beta, p_\theta) = u' M^{-1} u - \log\{p\beta, \theta | Z\},$$

subject to the constraints $\xi(q) = \Sigma_{j=1}^{J} \theta_j g(s_j, \beta) = 0$, and M is a strictly positive symmetric mass matrix (usually a diagonal matrix). The equilibrium distribution of the Hamiltonian dynamics,

$$\mathrm{d}q_t = M^{-1} u_t \, \mathrm{d}t,$$

$$\mathrm{d}u_t = \left\{ \nabla p(\beta, \theta | Z) - \gamma(q_t) M^{-1} u_t \right\} \mathrm{d}t + \sigma(q_t) \, \mathrm{d}W_t + \nabla \xi(q_t) \, \mathrm{d}\lambda_t,$$

with $\sigma(q_t)\sigma(q_t)' = 2\gamma(q_t)$, and subject to the constraint $\xi(q_t) = 0$, has a density that is proportional to $\exp\{-H(\beta, \theta, p_\beta, p_\theta)\}$. Here λ_t is the Lagrange multiplier process that is associated with the moment conditions. Lelièvre *et al.* (2012) proposed a numerical scheme that relies on a splitting strategy (see also Bou-Rabee and Owhadi (2010)). They discussed a special choice of parameters M and γ (the overdamped Langevin dynamics), in which the numerical scheme is an Euler discretization of the dynamics with a projection that is associated with the constraints. A Metropolis–Hastings step corrects the discretization error, so that the stationary distribution of the chain is equal to $p(\beta, \theta | Z)$ supported on the submanifold of the parameters. The discretization of the constrained Hamiltonian relies on an explicit integrator ('rattle'; see Anderson (1983)) that involves solving a system of p non-linear equations for λ of the form $\xi\{q^n + 2u^{n+1/4} + \delta t \nabla p(q^n | Z) + 2\nabla \xi(q^n)\lambda\} = 0$. This may be computationally expensive for moderate p and complicated moment conditions and could slow down the algorithm. This numerical scheme is described in Appendix A.6.5. For theoretical details see Lelièvre *et al.* (2012).

3.6. 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 that β can be solved given θ.

Our generalization of Chamberlain and Imbens (2003) starts with the Dirichlet prior $\pi^*(\theta) \propto \Pi_{j=1}^{J} \theta_j^{\alpha-1}$, $\alpha > 0$. 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 that the researcher does this M times, writing the draws as $\{\theta^{(k)}\}_{k=1,2,\ldots,M}$. For each $\theta^{(k)}$ we assume that there is a unique $\beta^{(k)}$ which solves the corresponding moment conditions. Chamberlain and Imbens (2003) stopped 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

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

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

$$w^{(k)} = \frac{p(\beta^{(k)}, \theta^{(k)}) |\mathcal{J}_\theta^{(k)} \mathcal{J}_\theta^{(k)\prime} + I_p|^{1/2}}{\prod_{j=1}^{J} (\theta_j^{(k)})^{\alpha-1}}, \qquad k = 1, 2, \ldots, M \tag{14}$$

(where $\mathcal{J}_\theta^{(k)}$ is equal to $\mathcal{J}_\theta$ evaluated at $(\beta^{(k)}, \theta^{(k)})$) which normalize as $w^{(k)*} = w^{(k)} / \Sigma_{k=1}^{M} w^{(k)}$. An encouraging aspect of this weight is that its functional form is free of data (n_j, for $j = 1, \ldots, J$), since the models have a common likelihood.

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

$$w^{(k)} \propto \pi(\beta^{(k)}) |\mathcal{J}_\theta^{(k)} \mathcal{J}_\theta^{(k)\prime} + I_p|^{1/2}, \qquad k = 1, 2, \ldots, M. \tag{15}$$

We can use these weights to estimate $E[h(\beta)|Z] \simeq (1/M)\Sigma_{k=1}^{M} w^{(k)*} h(\beta^{(k)})$. This is importance sampling, e.g. Marshall (1956), Geweke (1989) and Liu (2001). An alternative is to resample with probability proportional to the weight $w^{(k)}$, which delivers sampling–importance resampling (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 that is obtained from the Bayesian bootstrap.

3.7. Unknown support

So far we have assumed that the support of the data is known. Here we deal with the case where the support is unknown but finite. Suppose that the support has J elements, $S = (s_1, \ldots, s_J)$. Let θ be the vector of the probabilities of the elements of S.

We assume that the support is IID draws from F_S, $s_j \sim^{\text{IID}} F_S$ for $j = 1, \ldots, J$, with density f_S with respect to Lebesgue measure.

Then the moment conditions are

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

whereas the posterior is

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

where $n_j = \Sigma_{i=1}^{N} \mathbf{1}(Z_i = s_j)$.

Assume that the researcher expresses a prior on $(\beta, \theta)|S$ with respect to the Hausdorff measure, $p(\beta, \theta|S)$. Then

$$p(\beta, \theta, S) = \left\{ \prod_{j=1}^{J} f_S(s_j) \right\} p(\beta, \theta|S). \tag{16}$$

Given θ and S, β is uniquely determined. Therefore the core result that we need to do inference is a generalization of proposition 1: the density of the probabilities and the support with respect to the Lebesgue measure is

$$p(\theta, S) = |\mathcal{J}_\theta \mathcal{J}_\theta' + \mathcal{J}_S \mathcal{J}_S' + I_p|^{1/2} \left\{ \prod_{j=1}^{J} f_S(s_j) \right\} p(\beta, \theta|S), \tag{17}$$

where

$$\mathcal{J}_\theta = \frac{\partial \beta}{\partial \theta} = \left(\sum_{j=1}^{J} \theta_j \frac{\partial g_j}{\partial \beta'} \right)^{-1} H_\beta,$$

$$\mathcal{J}_S = \frac{\partial \beta}{\partial S} = \left(\sum_{j=1}^{J} \theta_j \frac{\partial g_j}{\partial \beta'} \right)^{-1} M,$$

$$M = \left\{ \theta_1 \left(\frac{\partial g_1}{\partial s_1'} \right), \ldots, \theta_J \left(\frac{\partial g_J}{\partial s_J'} \right) \right\}. \tag{18}$$

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

$$p(\beta, \theta, S) = \frac{|\mathcal{J}_\theta \mathcal{J}_\theta' + \mathcal{J}_S \mathcal{J}_S'|^{1/2}}{|\mathcal{J}_\theta \mathcal{J}_\theta' + \mathcal{J}_S \mathcal{J}_S' + I_p|^{1/2}} \, p(\beta|S) \, p(\theta|\beta, S) \prod_{j=1}^{J} f_S(s_j). \tag{19}$$

This paper leaves open what happens to this analysis as $J \to \infty$ for further research.

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 non-parametric standpoint it is natural to build a prior from $p(\theta)$, e.g. Dirichlet. Then proposition 1 implies that there is a unique joint prior

$$p(\beta, \theta) = \frac{p(\theta)}{\sqrt{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|}}, \tag{20}$$

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

$$\Pr\{(\beta, \theta) \in C\} = \int_{C_\theta} p(\theta) \mathrm{d}\theta. \tag{21}$$

The Dirichlet special case (20) is the implicit Chamberlain and Imbens (2003) prior on $p(\beta, \theta)$.

4.2. A prior on β

Proposition 2 says that

$$p(\beta, \theta) = \frac{|\mathcal{J}_\theta \mathcal{J}_\theta'|^{1/2}}{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|^{1/2}} \, p(\beta) \, p(\theta|\beta). \tag{22}$$

If we place a prior on β, with density $p(\beta)$ with respect to Lebesgue measure, then we can form a scientifically centred 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 becomes large the Jacobian in equation (22) will become unimportant in practice.

4.3. Ad hoc 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 that 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 prior (20), setting

$$p(\beta,\theta) \propto \frac{\eta(\beta,\theta)}{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|^{1/2}} \mathbf{1}_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{23}$$

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

The simple case of $\eta(\beta,\theta) = \eta(\beta)\eta(\theta)$ would imply under expression (23) that

$$p(\theta) \propto \eta(\beta)\eta(\theta)\mathbf{1}_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{24}$$

The case where $\eta(\theta)$ is Dirichlet is important. Then the Bayesian bootstrap weights (23) would become the rather simple

$$w_j \propto \eta(\beta^{(j)}), \qquad j = 1,2,\ldots,M. \tag{25}$$

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

An alternative to expression (23) is to put no mass on inadmissible combinations of β and θ. We call this the 'truncated prior'

$$p(\beta,\theta) \propto \eta(\beta,\theta)\mathbf{1}_{\Theta_{\beta,\theta}}(\beta,\theta) \tag{26}$$

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 $C \in \mathbb{R}^{J-1+p}$ that

$$\begin{aligned} \Pr\{(\beta,\theta) \in C\} &= \int_{C_\theta} p(\beta,\theta)\sqrt{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|}\mathrm{d}\theta \\ &\propto \int_{C_\theta} \eta(\beta,\theta)\sqrt{|\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|}\mathrm{d}\theta. \end{aligned} \tag{27}$$

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

4.3.1. *Example 9 (continuing logistic example 1)*

Assume the initial prior

$$\eta(\beta,\theta) \propto \theta^{0.01-1}(1-\theta)^{0.01-1} \exp\left\{-\tfrac{1}{2}(\beta-1)^2\right\}, \tag{28}$$

which is a relatively ignorant Dirichlet prior on the probabilities and an informative Gaussian prior for β centred on one. This is depicted in Fig. 3. With this initial prior and using the class of priors (26), the density with respect to the univariate Hausdorff measure is

$$p(\beta,\theta) \propto \eta(\beta,\theta)\mathbf{1}_{\Theta_{\beta,\theta}}(\beta,\theta). \tag{29}$$

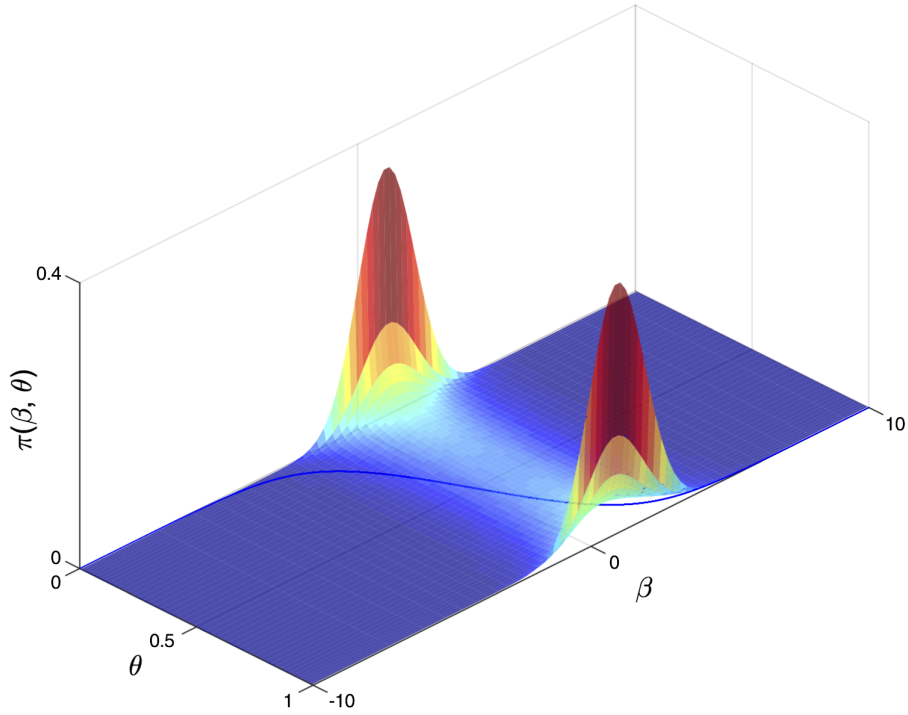


Fig. 3. Parameter space (——) $\Theta_{\beta,\theta}$ and the initial prior $\pi(\beta, \theta)$: Fig. 1 shows the implied $p(\beta, \theta)$

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

$$p(\theta) \propto \eta(\beta, \theta) \sqrt{\left\{1 + \left(\frac{\partial \beta}{\partial \theta}\right)^2\right\}} \mathbf{1}_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{30}$$

with respect to the Lebesgue measure. With the alternative prior (23), then

$$\begin{aligned} p(\beta, \theta) &\propto \frac{\eta(\beta, \theta)}{\sqrt{\{1 + (\partial \beta / \partial \theta)^2\}}} \mathbf{1}_{\Theta_{\beta,\theta}}(\beta, \theta), \\ p(\theta) &\propto \eta(\beta, \theta) \mathbf{1}_{\Theta_{\beta,\theta}}(\beta, \theta). \end{aligned} \tag{31}$$

5. Illustrative examples

In this section we present some illustrative examples and simulation studies. Since the MCMC results obtained by the marginal, joint and constrained Hamiltonian Monte Carlo 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 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 two-dimensional Hausdorff prior as

$$p(\beta, \theta) \propto \exp(-2|\beta - m|) \theta_1^{\alpha-1} \theta_2^{\alpha-1} (1 - \theta_1 - \theta_2)^{\alpha-1} \mathbf{1}(\min\{\theta_1, \theta_2, 1 - \theta_1 - \theta_2\} \geqslant 0). \tag{32}$$

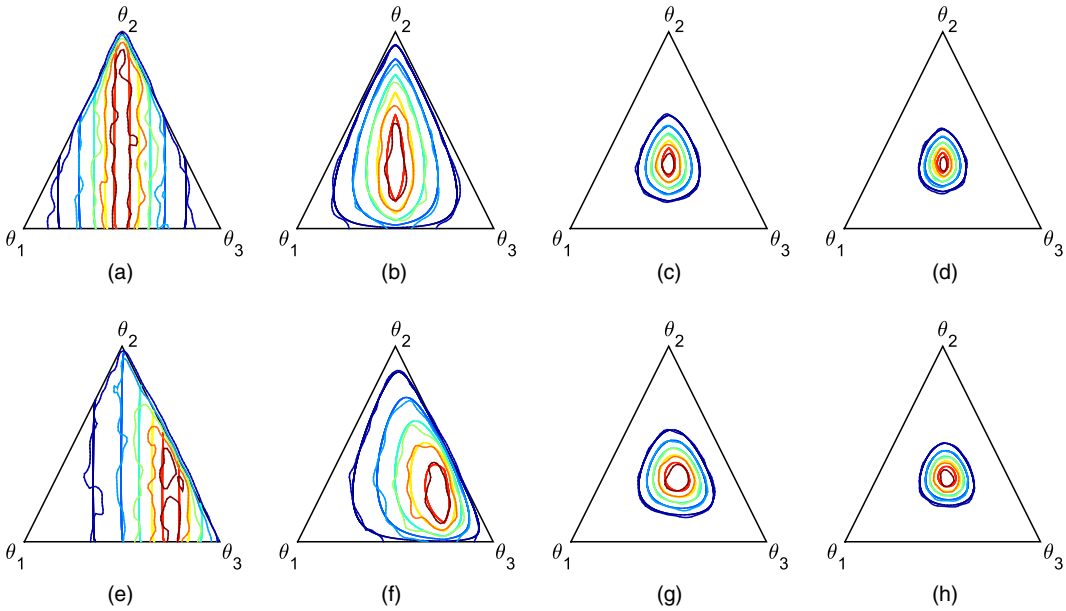


Fig. 4. Equiprobability contours implied by the Laplace–Dirichlet prior on $p(\beta, \theta)$ with respect to the Hausdorff measure (plotted is the marginal $p(\theta_1, \theta_3)$ for several values of m and α, with θ_2 implied as $\theta_2 = 1 - \theta_1 - \theta_3$; this case has $J = 3$ points of support ($s_1 = -1$, $s_2 = 0$ and $s_3 = 1$) and $r = 1$ moment constraints (the mean)): (a) $\alpha = 0.01$, $m = 0$; (b) $\alpha = 0.5$, $m = 0$; (c) $\alpha = 5$, $m = 0$; (d) $\alpha = 10$, $m = 0$; (e) $\alpha = 0.01$, $m = 0.5$; (f) $\alpha = 0.5$, $m = 0.5$; (g) $\alpha = 5$, $m = 0.5$; (h) $\alpha = 10$, $m = 0.5$

We call this a 'Laplace–Dirichlet' distribution, where β is centred on m and the Dirichlet part is indexed by α.

By the marginal method:

$$p(\theta) \propto \exp(-2|1 - 2\theta_1 - \theta_2 - m|)\theta_1^{\alpha-1}\theta_2^{\alpha-1}(1 - \theta_1 - \theta_2)^{\alpha-1}\mathbf{1}(\min\{\theta_1, \theta_2, 1 - \theta_1 - \theta_2\} \geqslant 0). \tag{33}$$

Fig. 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 that 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.

Fig. 5 draws the prior for β. Here the support of the data means that 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 β's prior centred at a positive value results in a prior for θ tilted towards 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.

5.2. *Linear regression*

Recall the linear regression of example 3. Assume that the observed data are $Z = \{(1, 1), (2, 4), (3, 9)\}$. Earlier we have seen that the parameter space $\Theta_{\beta,\theta}$ is a non-flat surface in $\mathbb{R}^3$. Fig. 6 demonstrates the posterior distribution of the parameters defined on this surface (the prior parameters are $\alpha = 0.5$ and $m = 3$).

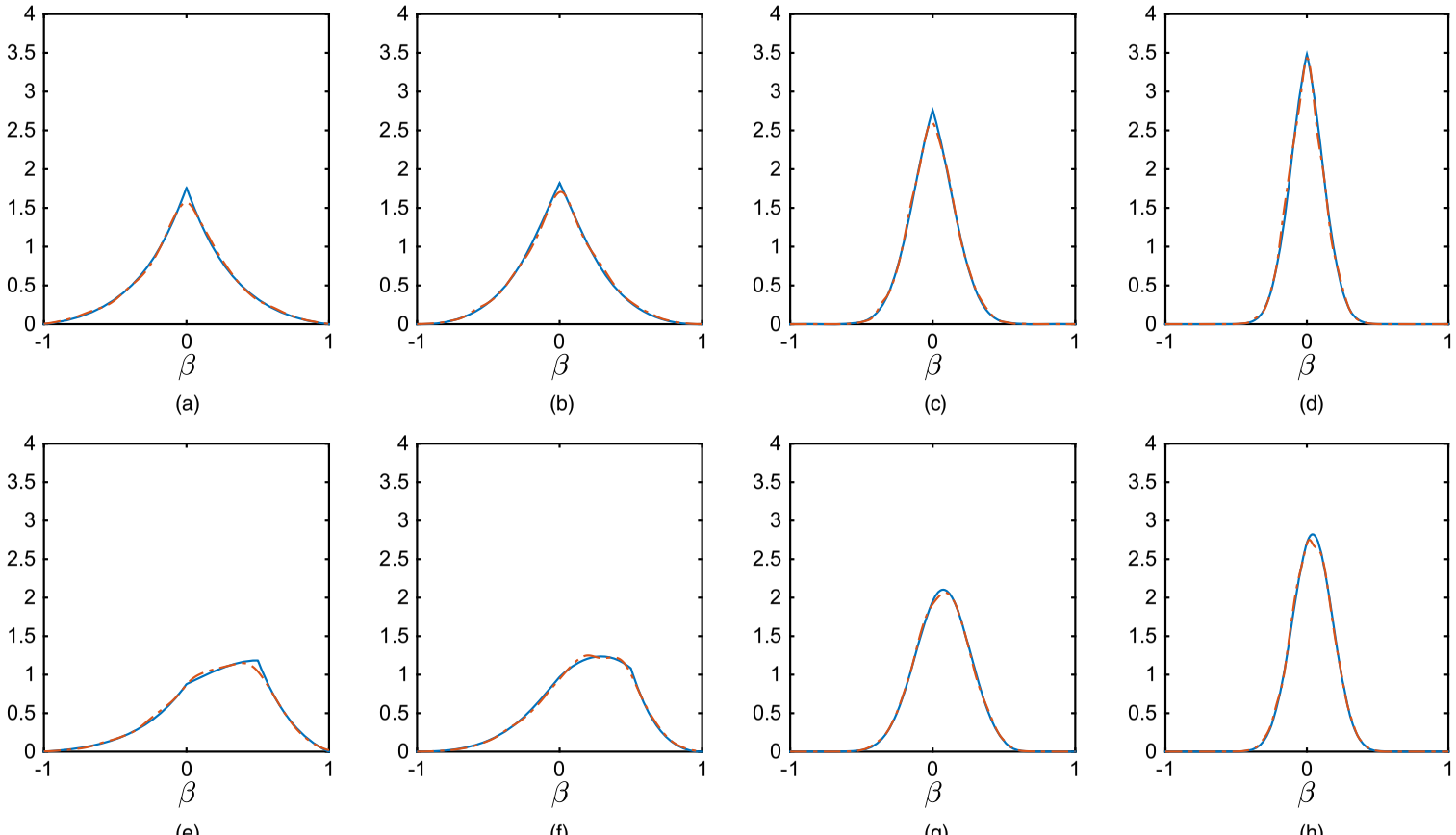


Fig. 5. Illustrating example 1 (estimating the mean), $p(\beta)$ for several values of m and α (this case has $J=3$ points of support ($s_1=-1, s_2=0$ and $s_3=1$) and $r=1$ moment; the initial prior for β is Laplace centred at m and the initial prior for θ is symmetric Dirichlet with parameter α): (a) $\alpha=0.01, m=0$; (b) $\alpha=0.5, m=0$; (c) $\alpha=5, m=0$; (d) $\alpha=10, m=0$; (e) $\alpha=0.01, m=0.5$; (f) $\alpha=0.5, m=0.5$; (g) $\alpha=5, m=0.5$; (h) $\alpha=10, m=0.5$

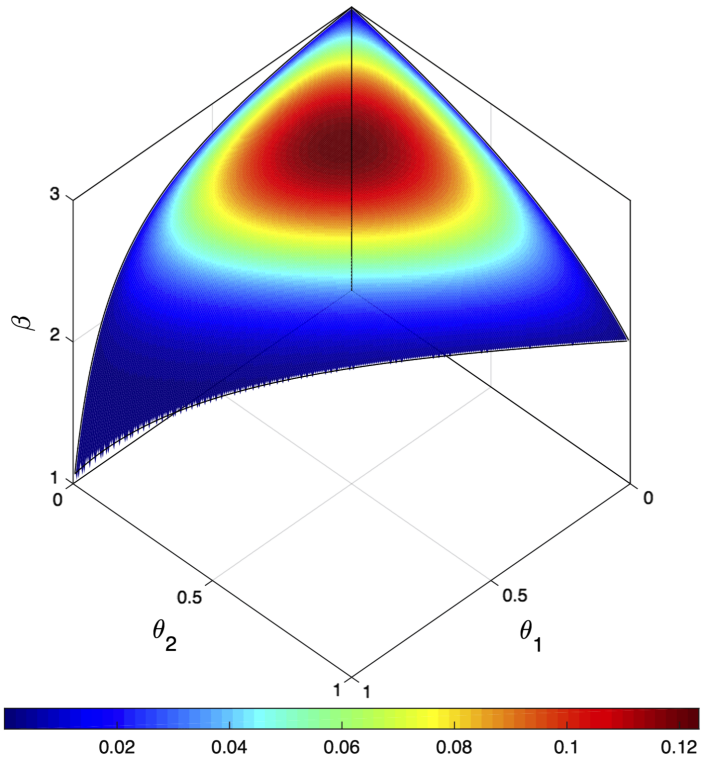


Fig. 6. Posterior distribution of the linear regression model with data $Z = \{(1,1),(2,4),(3,9)\}$: the prior parameters are $\alpha = 0.5$ and $m = 3$

Following the suggested MCMC simulation algorithms we draw 100000 samples from the posterior distribution of the parameters. In Fig. 7 we have drawn the contour plots of the posterior distribution of the probabilities. Analytical results have been compared with the estimates that were obtained by a kernel density estimator using the MCMC draws.

5.3. 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 $1 \leqslant j \leqslant J$, are 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 that 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 maximum likelihood estimators, and Σ_0 is equal to the asymptotic variance of the maximum likelihood 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 50000 samples from the posterior after a 5000-sample burn-in (the chain's trace has been thinned with a factor of 100 and so has been iterated 5 million times). The scatter plot of the sample is depicted in Fig. 8(a). 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 Fig. 8(b) the correlogram (auto-correlation function of the chains of β and 10 elements of θ have been presented (the red broken curves and the blue dotted curves correspond to β and θ respectively.) The auto-correlation functions demonstrate that the Markov chain is mixing sufficiently well. In Fig. 8(c) the contour plot of the posterior distribution of β has been compared with that obtained by the Bayesian

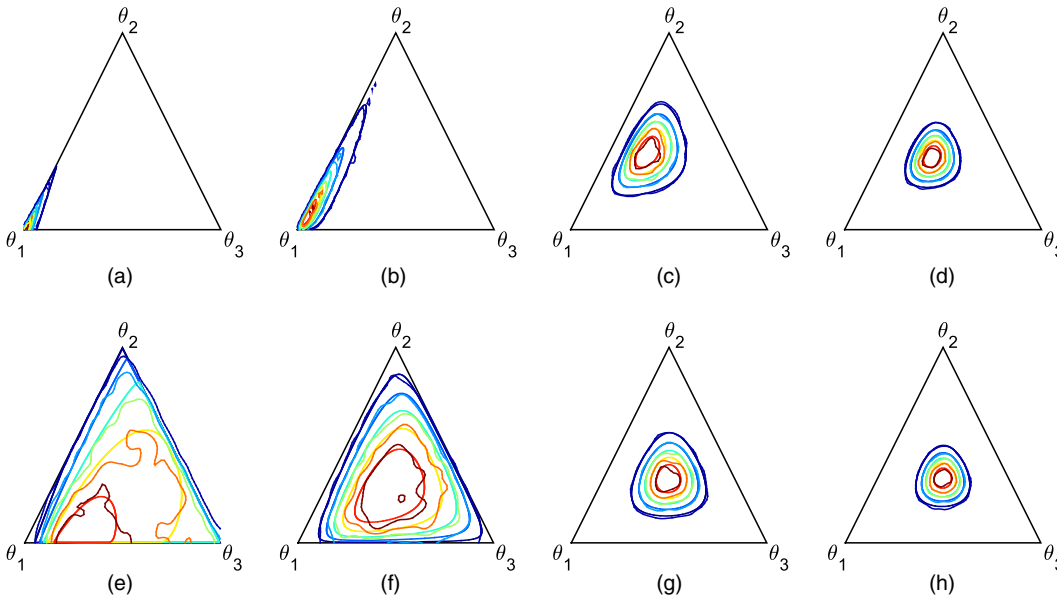


Fig. 7. Posterior distribution of θ in the linear regression model with data $Z = \{(1,1),(2,4),(3,9)\}$ (analytical results and the estimates obtained by a kernel density estimator by using 100000 MCMC draws): (a) $\alpha = 0.01, m = 1$; (b) $\alpha = 0.05, m = 1$; (c) $\alpha = 5, m = 1$; (d) $\alpha = 10, m = 1$; (e) $\alpha = 0.01, m = 3$; (f) $\alpha = 0.5, m = 3$; (g) $\alpha = 5, m = 3$; (h) $\alpha = 10, m = 3$

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. Fig. 8(d) 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 IV-based estimator; the second looks at estimating the ATE from an experiment.

6.1. Instrumental variables

We use a subsample of the earnings and schooling data set that was studied in Chamberlain and Imbens (2003). This data set is a subset of the data that were studied in Angrist and Krueger (1991) and consist of the self-reported weekly log-earnings (self-reported annual earnings divided by 52) of 162512 male subjects who reported positive annual wages in 1979 along with 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) discussed the myriad of problems of self-report income data but we do not address that issue here.

Chamberlain and Imbens (2003) studied the dependence of earnings on the level of schooling by using a linear additive treatment effect model (e.g. Imbens and Rubin (2015)). They modelled 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 are modelled here as $E[Y_X|X, Y_0] = Y_0 + \beta_1 X$, 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.

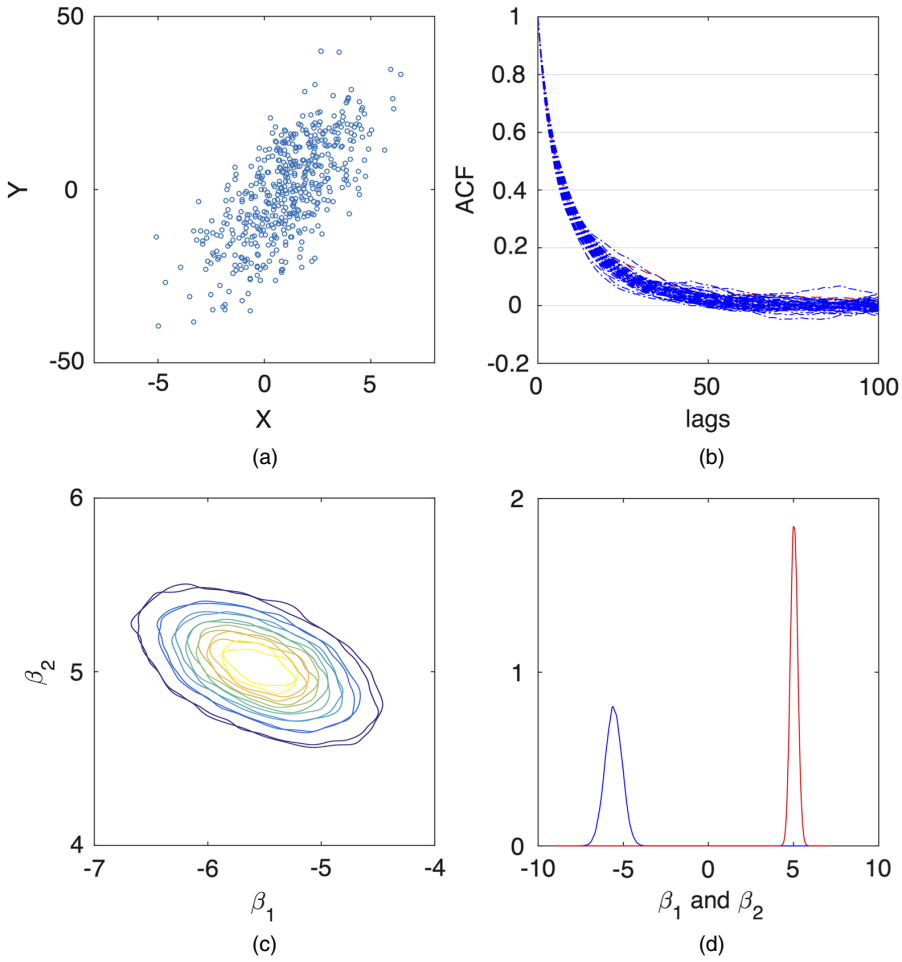


Fig. 8. Inference in a linear regression model with $J = 500$: (a) sample and the posterior probabilities (circles whose radius is the posterior expectation of the probabilities given the data, $E[\theta_j|Z]$); (b) correlogram for the thinned draws of the elements of β and 10 elements of θ; (c) estimated contour and (d) marginal densities of the resulting posterior (——, β_1; ——, β_2)

To estimate the unknown parameters, $\beta = (\beta_0, \beta_1)$, we follow Angrist and Krueger (1991) and Chamberlain and Imbens (2003) and use an 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 IV W is correlated with the regressor X and thought by the researchers to be uncorrelated with the errors.

We obtain the classical IV estimates of β by 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 1000 times. Our aim will be to compare different estimators by using these smaller samples.

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

$$p(\beta, \theta) \propto \frac{1}{\sqrt{(\mathcal{J}_\theta \mathcal{J}_\theta' + I_2)}} \eta(\beta)\eta(\theta)\mathbf{1}_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{34}$$

where $\eta(\beta) = \varphi(\beta_0; 5, 4)\varphi(\beta_1; 0, 0.2)$, and $\varphi(\cdot; \mu, \sigma^2)$ is the Gaussian density with mean μ and variance σ^2. The intercept is centred at 5 with variance 4, implying that the mean annual income for those with no schooling is equal to \$7717 (with 95% confidence interval [\$153, \$388965]) with 0 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 $\eta(\theta) \propto \Pi_{j=1}^{J} \theta_j^{\alpha-1}$, where $\alpha = 10^{-6}$ (we also tried $\alpha = J^{-1}$, with no substantial change in the results).

For 1000 replications, a random sample of size J has been drawn with replacement from the 162512 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 1000 cases. Fig. 9 shows the pointwise 95% confidence intervals of the marginal prior distributions over these 1000 cases, for $J = 100$, 1000, 5000, 10000. It shows that the prior is modestly informative and only mildly depends on 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 classical IV estimates of β and the Chamberlain and Imbens (2003) Bayesian bootstrapping estimates obtained by 10000 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 that was described earlier).

The Bayesian estimates are obtained by the following resampling method. Initially a sample of size 10000 is drawn from a Dirichlet distribution with parameter $(n_1 + \alpha - 1, \ldots, n_J + \alpha - 1)$, and the importance sampling weights are computed as $w^{(k)} \propto \eta(\beta^{(k)})$. 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 are reported here. For $J = 10$, 100, 1000, 5000, 10000, 40000, 100000 the effective sample size divided by J (Liu (2001), page 35) was 0.620, 0.576, 0.607, 0.719, 0.819, 0.978 and 0.997 respectively. This suggests that this is a reasonable method for this problem.

In Fig. 10 the sampling distribution of these five estimators have been plotted. The blue 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 also (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 are reported in Table 1. Although the Bayesian estimators are slightly biased, thanks to their small variances they have lower RMSE. In Table 1 and Fig. 11 we have also reported the length of the 95% confidence intervals of the sampling distribution of the estimators (over the 1000 replications) of β_0 and β_1 for sample sizes $J = 10, 100, 1000, 5000, 10000, 40000, 100000$. 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 100000 the old methods catch up 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, cuts off the tails of the posterior corresponding to these implausible values. Because of the ridge-like posterior that is induced by the weakly informative likelihood, the posterior contracts onto a manifold, rather than a single point. As such, having a prior which constrains the feasible support provides value.

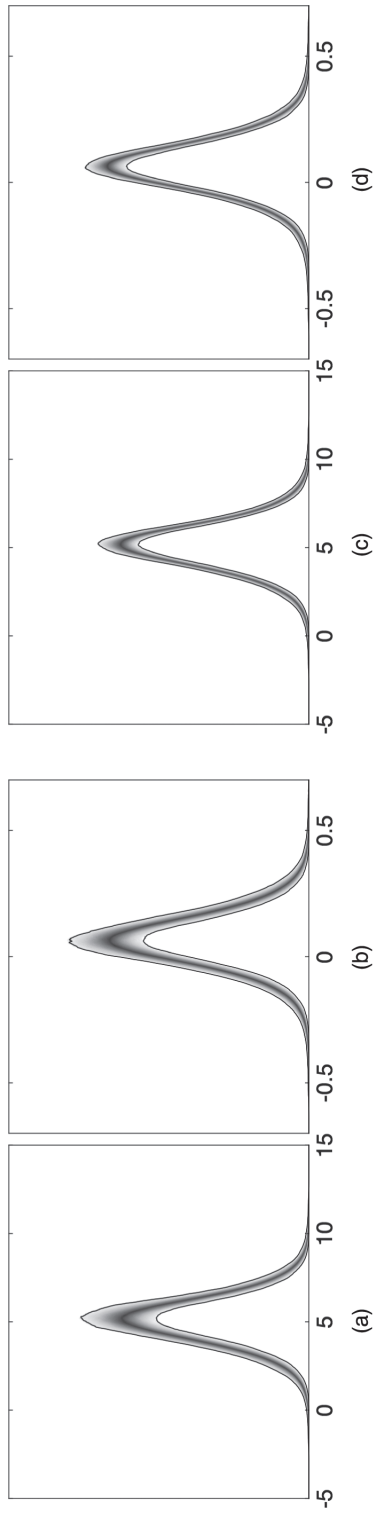


Fig. 9. 95% pointwise confidence regions for the marginal prior for β for (a), (b) $J = 100$, (c), (d) $J = 1000$, (e), (f) $J = 5000$ and (g), (h) $J = 10000$ points of random support (the confidence region is computed over 1000 replications): (a), (c), (e), (g) β_0; (b), (d), (f), (h) β_1

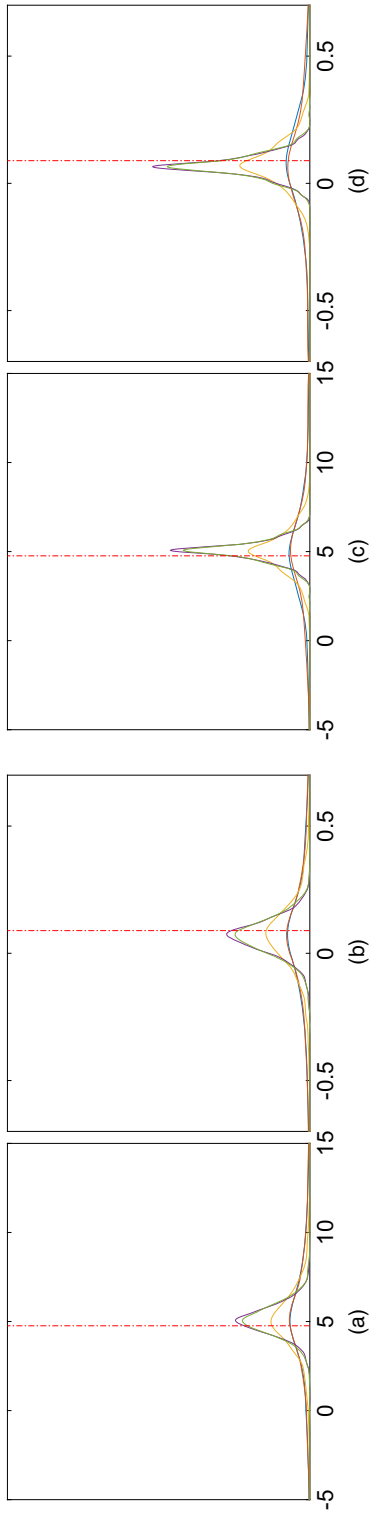


Fig. 10. Sampling distribution of classical IV (——) (denoted frequentist), Bayesian bootstrapping (——, mean; ——, median) and Bayesian estimators (——, mean; ——, median) of β in the linear regression model with the IV employing sample sizes (a), (b) $J = 10$, (c), (d) $J = 1000$, (e), (f) $J = 10000$ and (g), (h) $J = 100000$

Table 1. Results for the linear regression with an instrument using 1000 replications sampling with replacement†

	Biases of mean for the following sample sizes J:			*Biases of median for the following sample sizes J:*			*RMSEs for the following sample sizes J:*			*95% confidence region lengths for the following sample sizes J:*		
	10	*1000*	*10000*	*10*	*1000*	*10000*	*10*	*1000*	*10000*	*10*	*1000*	*10000*
β_0												
Classical 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
Bayesian bootstrap $E[\theta\|Z]$	−0.174	0.909	−0.020	0.347	0.259	−0.015	50.01	27.32	0.221	42.21	45.36	0.851
Bayesian bootstrap 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												
Classical 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
Bayesian bootstrap $E[\theta\|Z]$	−0.001	−0.072	0.002	−0.019	−0.020	0.001	3.940	2.151	0.017	3.402	3.546	0.067
Bayesian bootstrap 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

†The bias of the mean is the difference of the mean of the replications and the true value (using all 162512 data points). The bias of the median is the median of the replications minus the true value. The 95% confidence region 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. The Bayesian bootstrap is the (non-informative) Bayesian bootstrap; med denotes median. The last two rows are the posterior mean and posterior median of the Bayesian model with weakly informative prior.

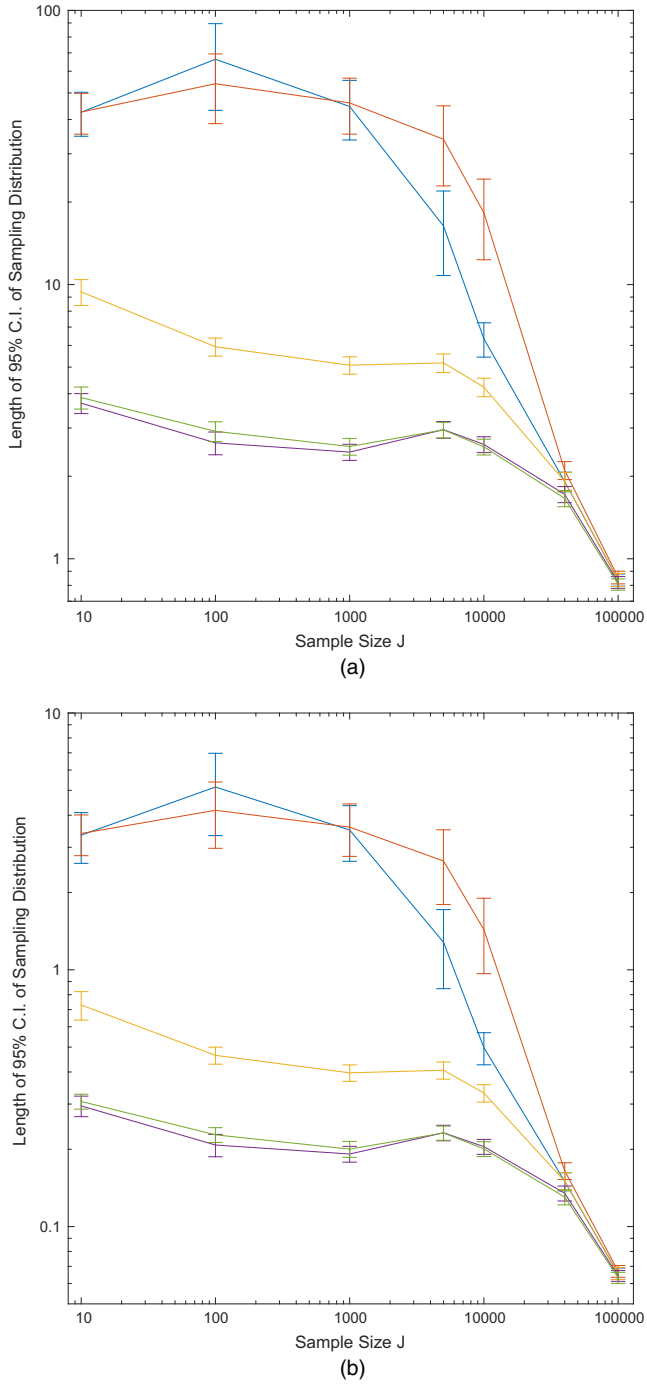


Fig. 11. Length of the 95% confidence intervals of the sampling distribution of the parameters β_0 and β_1 for various sample sizes J, and for the classical IV estimator (——), Bayesian bootstrapping (——, mean; ——, median) and Bayesian method (——, mean; ——, median) ($\mathbb{I}$, our estimated 95% confidence interval estimates of the lengths): (a) β_0; (b) β_1

6.2. Causal inference

In this example we analyse the data set of Imbens *et al.* (2001). The data set contains socio-economic 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 those who won only small amounts 'the losers' (259 observations). The goal is to study the effect of unearned income on the economic behaviour of the subjects, more specifically, on their average labour income over the first 6 years following the year in which they had won the lottery. For each individual the treatment indicator W_i is equal to 1 for the winners and 0 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 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 data set 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, which is an indicator showing whether she has been working during the year that the winning ticket is purchased, WT, and the annual social security earnings from 6 years before the year in which the winning ticket is purchased, EYB1, ..., EYB6, to 6 years after that, EYA1, ..., EYA6, all converted to 1986 dollars. The authors argued, perhaps optimistically, that social security income is potentially the most reliable measure of income in the long run, although it is capped to the maximum taxable earning ($42000 in 1986).

To improve the overlap of the background variables, following the recommendation of Imbens and Rubin (2015), initially we model the propensity scores by using a logistic regression model, and estimate the model's parameter by 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 and YW, the indicator for the positiveness of the earnings 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 (less than 0.0891) or too large (greater than 0.9109) estimates of propensity scores. This results in a sample of size $N = 295$ (142 winners and 153 losers). In the model proposed the propensity score is regressed on 13 covariates by using a logistic regression. The vector of covariates is denoted by X_i and includes a constant, the linear terms TB, YS, WT, EYB1, Age, SEYB5, YW and 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 moment conditions,

$$\mathbb{E}[g(Z_i, \beta)] = 0, \qquad g(Z_i, \beta) = \begin{pmatrix} X_i(Y_i - \eta_i) \\ (W_i - \eta_i)Y_i / \{\eta_i (1 - \eta_i)\} - \tau \end{pmatrix}, \tag{35}$$

in which, $Z_i = (X_i, Y_i, W_i)$, $\beta = (\gamma, \tau)$ and $\eta_i = \exp(\gamma' X_i)/\{1 + \exp(\gamma' X_i)\}$. If we assume that the Z_is are IID draws from a discrete distribution supported on $\{s_1, \ldots, s_J\}$, with $\mathbb{P}(Z_i = s_j) = \theta_j$, the parameters (β, θ) will satisfy the following system of equations:

$$\begin{pmatrix} \sum_{j=1}^{J} \theta_j x_j (y_j - \eta_j) \\ \sum_{j=1}^{J} \theta_j (w_j - \eta_j) y_j / \{\eta_j (1 - \eta_j)\} - \tau \end{pmatrix} = 0. \tag{36}$$

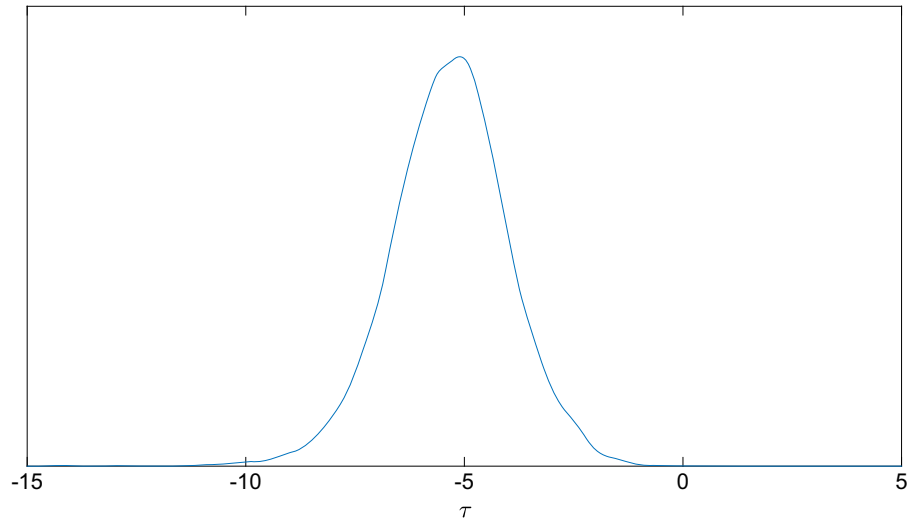


Fig. 12. Posterior distribution of the ATE on subsequent annual earnings of a substantial lottery win for the lottery data set

We let the prior of (β, θ) be

$$p(\beta, \theta) \propto \frac{1}{\sqrt{(\mathcal{J}_\theta \mathcal{J}_\theta' + I_{14})}} \eta(\gamma)\eta(\tau)\eta(\theta)\mathbf{1}_{\Theta_{\beta,\theta}}(\beta, \theta), \tag{37}$$

in which the initial prior of the regression coefficients $\eta(\gamma)$ is a normal distribution centred 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 the 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 10000 independent draws from the posterior of our model. An estimate of the posterior distribution of the ATE is depicted in Fig. 12. *A posteriori* the expected value of the ATE is −\$5346 (with 95% credible interval [−\$8069, −\$2720]). This indicates that the average income of the winners of the lotteries, in the years after winning the prize, tend to decrease slightly. Our estimate of the ATE is only slightly different from the frequentist estimate.

7. Conclusions

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

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

To harness these gains, at the centre 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 enables the use of conventional simulation methods such as MCMC, importance sampling and Hamiltonian Monte Carlo methods. It is convenient to use where the moment conditions can be solved analytically or numerically very fast.

Our joint method is somewhat more difficult 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.

Acknowledgements

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

Appendix A

A.1 Proof of proposition 1

Since corresponding to every $\theta \in \Theta_\theta$ there is a unique β, there is 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 that $S_\theta(A)$ is its projection on Θ_θ. Therefore

$$\mathbb{P}\{S_\theta(A)\} = \mathbb{P}(A) = \int_A p(\beta, \theta)\, \mathrm{d}A = \int_{S_\theta(A)} \|v_1 \wedge \ldots \wedge v_{J-1}\| p(\beta, \theta)\, \mathrm{d}S$$

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

$$\|v_1 \wedge \ldots \wedge v_{J-1}\| = \mathrm{Gram}(v_1, \ldots, v_{J-1})^{1/2} = |\mathcal{J}_\theta \mathcal{J}_\theta' + I_{J-1}|^{1/2} = |\mathcal{J}_\theta \mathcal{J}_\theta' + I_p|^{1/2}$$

where $\mathrm{Gram}(\cdot)$ is the Gramian determinant and $\mathcal{J}_\theta = \partial \beta / \partial \theta'$.

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 equations 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 $H_1 = (h_1 \ldots h_{J-p-1})$ and $H_2 = (h_{J-p} \ldots h_{J-1})$. Therefore, $\partial \theta_{J-p:J-1} / \partial \theta_{1:J-p-1} = -H_2^{-1} H_1$ and so

$$\begin{aligned} p(\theta_{1:J-p-1}|\beta) &= |H_2^{-1} H_1 H_1' H_2^{'-1} + I_p|^{1/2} p(\theta|\beta), \\ p(\theta_{1:J-p-1}, \beta) &= |H_2^{-1} H_1 H_1' H_2^{'-1} + I_p|^{1/2} p(\beta) p(\theta|\beta). \end{aligned}$$

Therefore the density of θ is

$$\begin{aligned} p(\theta) &= \left| \frac{\partial(\theta_{1:J-p-1}, \beta)}{\partial(\theta)} \right| p(\theta_{1:J-p-1}, \beta) = \left| \frac{\partial \beta}{\partial \theta_{J-p:J-1}} \right| p(\theta_{1:J-p-1}, \beta) \\ &= \left| \mathbb{E}\left[\frac{\partial g}{\partial \beta'} \right]^{-1} H_2 \right| p(\theta_{1:J-p-1}, \beta) \\ &= \left| \mathbb{E}\left[\frac{\partial g}{\partial \beta'} \right]^{-1} H_2 \right| |H_2^{-1} H_1 H_1' H_2^{'-1} + I_p|^{1/2} p(\beta) p(\theta|\beta) \end{aligned}$$

$$= \left| \mathbb{E}\left[\frac{\partial g}{\partial \beta'}\right]^{-1} \right| |H_1 H_1' + H_2 H_2'|^{1/2} p(\beta) p(\theta|\beta) = \left| \mathbb{E}\left[\frac{\partial g}{\partial \beta'}\right]^{-1} \right| |HH'|^{1/2} p(\beta) p(\theta|\beta)$$

$$= \left| \mathbb{E}\left[\frac{\partial g}{\partial \beta'}\right]^{-1} HH' \, \mathbb{E}\left[\frac{\partial g'}{\partial \beta}\right]^{-1} \right|^{1/2} p(\beta) p(\theta|\beta) = \left| \frac{\partial \beta}{\partial \theta'} \frac{\partial \beta'}{\partial \theta} \right|^{1/2} p(\beta) p(\theta|\beta).$$

Therefore

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

A.3. Joint method proposal

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:

$$\theta^* = \arg\min_{\theta} \tfrac{1}{2} \|\pi^* - \theta\|_2^2 + \tfrac{1}{2} (\iota' \pi^* - \iota' \theta)^2.$$

The quadratic penalty is certainly inelegant (e.g. compared with the log-likelihood of the multinomial model, but see, for example, Owen (1991) and Antoine *et al.* (2007) who used it for their Euclidean empirical likelihood) as the resulting θ^* can have negative elements or may result in $\theta_J^* = 1 - \iota' \theta^* \leqslant 0$. 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

$$E[\theta, \lambda] = \tfrac{1}{2} \|\pi^* - \theta\|_2^2 + \tfrac{1}{2} (\iota' \pi^* - \iota' \theta)^2 + \lambda' (H^* \theta + g_J^*)$$

and the first-order conditions are

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

$$\frac{\partial E}{\partial \lambda} = H^* \theta^* + g_J^* = 0.$$

Solving them for θ^* and λ results in

$$\theta^* = \pi^* - (I + \iota\iota')^{-1} H^{*'} \{H^* (I + \iota\iota')^{-1} H^{*'}\}^{-1} (H^* \pi^* + g_J^*),$$

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

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

$$a^* = -(I + \iota\iota')^{-1} H^{*'} \{H^* (I + \iota\iota')^{-1} H^{*'}\}^{-1} g_J^*,$$

$$B^* = I - (I + \iota\iota')^{-1} H^{*'} \{H^* (I + \iota\iota')^{-1} H^{*'}\}^{-1} H^*.$$

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

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

where $|\Sigma|_{\mathrm{rank}(\Sigma)}$ 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 $\Sigma_\beta=(\Sigma_{0_\beta}^{-1}+\Sigma_{\mathrm{BB}_\beta}^{-1})^{-1}$, where $\Sigma_{0_\beta}^{-1}$ is the prior's covariance and $\Sigma_{\mathrm{BB}_\beta}^{-1}$ is the covariance of the estimates of β obtained by the Bayesian bootstrapping of Chamberlain and Imbens (2003). (As an alternative we may use the asymptotic covariance of the least squares or generalized method-of-moments estimators.) Moreover a suitable candidate for Σ_Q is $\mathrm{diag}(\hat{\theta}_1^2,\ldots,\hat{\theta}_{J-1}^2)$ where

$$\hat{\theta}=(\hat{\theta}_1,\ldots,\hat{\theta}_{J-1})=\arg\max_{\theta}\sum_{j=1}^{J}n_j\ln(\theta_j)\qquad\text{subject to }\hat{H}\theta+\hat{g}_J=0,\tag{38}$$

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

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

$$\tilde{\theta}^{*}=\arg\min_{\tilde{\theta}}\tfrac{1}{2}\|\tilde{\theta}-\tilde{\pi}^{*}\|+\tfrac{1}{2}(\iota'\tilde{\theta}-\iota'\tilde{\pi}^{*})\qquad\text{subject to }\bar{H}^{*}\bar{\theta}^{(t)}+\tilde{H}^{*}\tilde{\theta}+g_J^{*}=0,\tag{39}$$

where $\tilde{H}^{*}=(g_{u_1}^{*},\ldots,g_{u_K}^{*})-g_J^{*}\iota'$, $\bar{H}^{*}=(g_{v_1}^{*},\ldots,g_{v_{J-K-1}}^{*})-g_J^{*}\iota'$, 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 solution $\tilde{\theta}^{*}=\tilde{a}^{*}+\tilde{B}^{*}\tilde{\pi}^{*}$, where

$$\tilde{a}^{*}=-(I+\iota\iota')^{-1}\tilde{H}^{*'}\{\tilde{H}^{*}(I+\iota\iota')^{-1}\tilde{H}^{*'}\}^{-1}(\bar{H}^{*}\bar{\theta}^{(t)}+g_J^{*}),$$

$$\tilde{B}^{*}=I-(I+\iota\iota')^{-1}\tilde{H}^{*'}\{\tilde{H}^{*}(I+\iota\iota')^{-1}\tilde{H}^{*'}\}^{-1}\tilde{H}^{*}.$$

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 $\hat{\beta}_{\mathrm{MLE}}+(5,-5)'$ and the variance equal to the asymptotic variance of $\hat{\beta}_{\mathrm{MLE}}$. Therefore the prior is as informative as the data, however, centred at a significantly different point.

Fig. 13(a) 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 Fig. 13(b) the auto-correlation function ACF of the chains of β and 50 elements of θ have been presented (the red broken curves and the blue dotted curves correspond to β and θ respectively). These show that the Markov chain is mixing sufficiently well. In Fig. 13(c) the contour plot of the prior distribution (bottom), posterior distribution of β by using Bayesian bootstrapping and the posterior distribution of β considering the informative prior (middle) have been depicted. In Fig. 13(d) the histogram of the samples from the posterior of β can be seen.

A.6. Not the just-identified case

A.6.1. Abstract expression of the problem

Collect all the parameters in the model and constraints as $\psi=(\theta_1,\ldots,\theta_{J-1},\beta_1,\ldots,\beta_p)'$, and $g(\psi)=0_r$.

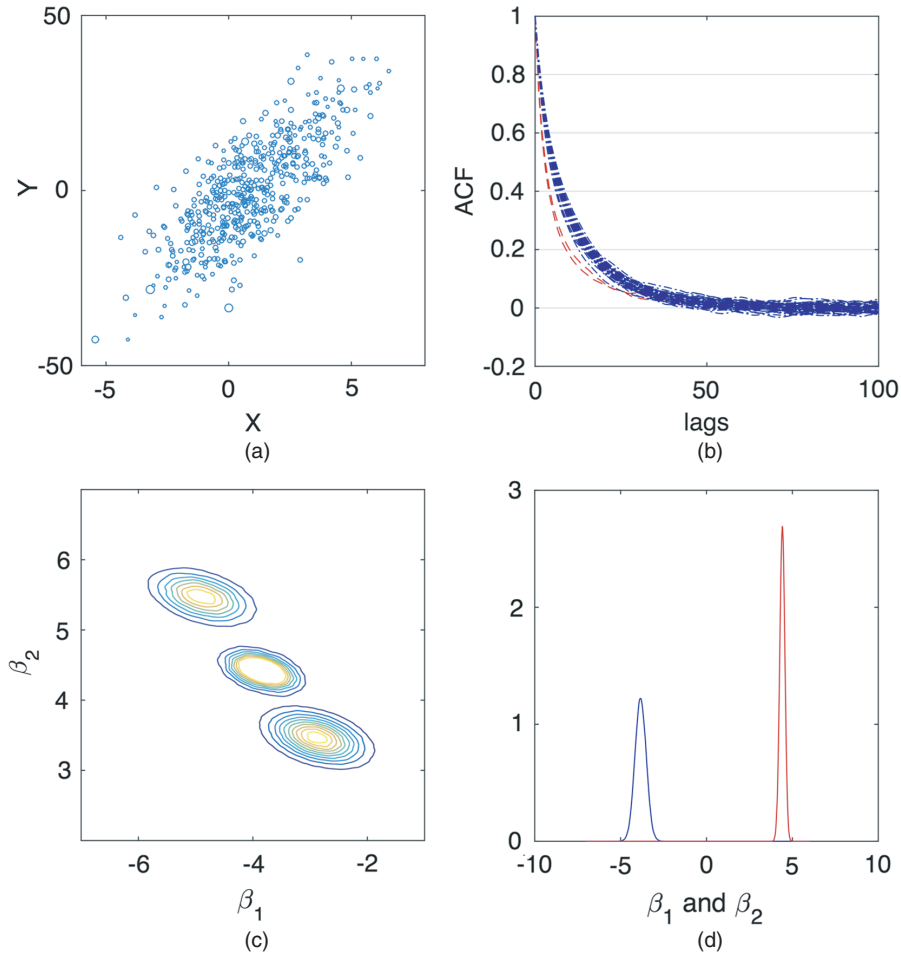


Fig. 13. Inference in the linear regression model with $J = 500$ and an informative prior: (a), (b) sample and the posterior probabilities; (c) prior (bottom), Bayesian bootstrap (top) and posterior (middle); (d) posterior distribution (——, β_1; ——, β_2)

The resulting constrained support is $\psi \in \Theta_\psi$. Write $\lambda = \psi_{\mathcal{I}}$ and $\phi = \psi_{\mathcal{I}^c}$, where $\mathcal{I}$ selects distinct indices of ψ and $\mathcal{I}^c$ is the complement, so $\mathcal{I} \cup \mathcal{I}^c = \{1, 2, \ldots, p + J - 1\}$. 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 1. Under $g(\psi) = 0$ knowledge of λ reveals ϕ, so there is a unique $\phi = t(\lambda)$.

A.6.2. Marginal method

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

A.6.3. Underidentification

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

We split $\beta = (\beta_1, \ldots, \beta_p)'$ as $\beta_{[1]} = \beta_{\mathcal{G}}$ and $\beta_{[2]} = \beta_{\mathcal{G}^c}$, where $\mathcal{G} \cup G^c = \{1, 2, \ldots, p\}$, $\dim(\mathcal{G}) = p - r$ and

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

A.6.3.1. Example 10. Consider the IV problem $g(s, \beta) = s^{(3)}(s^{(1)} - \beta' s^{(2)})$, $\dim(s^{(2)}) = p$ and $\dim(s^{(3)}) = r$. If $p > r$ then split $\beta = (\beta'_{[1]}, \beta'_{[2]})'$, where $\dim(\beta_{[1]}) = r - p$ and $\dim(\beta_{[2]}) = r$. Write $s_j = (s'_{j,[1]}, s'_{j,[2]})'$; then

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

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

A.6.4. Overidentification

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

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

A.6.4.1. Example 11. Again consider $g(s, \beta) = s^{(3)}(s^{(1)} - \beta' s^{(2)})$, $\dim(s^{(2)}) = p$ and $\dim(s^{(3)}) = r$. If $p < r$ then split $\theta = (\theta'_{[1]}, \theta'_{[2]})'$, 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 $(\theta'_{[2]}, \beta')$, where

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

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

A.6.5. Constrained Hamiltonian Monte Carlo sampling algorithm

Let $q^n = (\beta^n, \theta^n)$ be the current state of the Markov chain. Following Lelièvre *et al.* (2012), by setting $M = (\Delta t/2) I$, $\gamma = 2I$ and $\sigma\sigma' = 4I$, one can obtain the following sampling scheme.

Step 1: draw the new momentum

$$u^{n+1/4} = \sqrt{\left(\frac{\Delta t}{8}\right)} [I - \nabla\xi(q^n)\{\nabla\xi(q^n)'^n\}^{-1} \nabla\xi(q^n)'] \sigma G^n,$$

where G^n is a vector of independent and identically standard Gaussian random variables.
Step 2: integrate the constrained Hamiltonian,

(a) Obtain $\lambda^{n+1/2}$ by solving the following system of p non-linear equations,

$$\xi\{q^n + 2u^{n+1/4} - \Delta t \nabla V(q^n) + 2\nabla\xi(q^n)\lambda^{n+1/2}\} = 0,$$

and
(b) set

$$u^{n+1/2} = u^{n+1/4} - \frac{\Delta t}{2} \nabla V(q^n) + \nabla\xi(q^n)\lambda^{n+1/2},$$

$$q^{n+1} = q^n + 2u^{n+1/4} - \Delta t \nabla V(q^n) + 2\nabla\xi(q^n)\lambda^{n+1/2},$$

$$u^{n+3/4} = u^{n+1/2} - \frac{\Delta t}{2} \nabla V(q^{n+1}) + \nabla\xi(q^{n+1})\{\nabla\xi(q^{n+1})'^{n+1}\}^{-1} \nabla\xi(q^{n+1})' \left\{\frac{\Delta t}{2} \nabla V(q^{n+1}) - p^{n+1/2}\right\}.$$

Step 3: accept the new state q^{n+1}, with probability

$$p_{\text{acc}} = \exp\{-H(q^{n+1}, u^{n+3/4}) + H(q^n, u^{n+1/4})\} \wedge 1.$$

Otherwise, we set $q^{n+1} = q^n$.

References

Anderson, H. C. (1983) Rattle: a velocity version of the shake algorithm for molecular dynamics calculations. *J. Computnl Phys.*, **52**, 24–34.

Angrist, J. D. and Krueger, A. B. (1991) Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.*, **106**, 979–1014.

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

Ariyabuddhiphongs, V. (2011) Lottery gambling: a review. *J. Gmblng Stud.*, **27**, 15–33.

Bound, J., Brown, C. and Mathiowetz, N. (2001) Measurement error in survey data. In *Handbook of Econometrics*, vol. 5 (eds J. J. Heckman and E. Leamer), pp. 3705–3843. Amsterdam: North-Holland.

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

Bou-Rabee, N. and Owhadi, H. (2010) Long-run behavior of variational integrators in the stochastic context. *SIAM J. Numer. Anal.*, **48**, 278–297.

Brubaker, M., Salzmann, M. and Urtasun, R. (2012) A family of MCMC methods on implicitly defined manifolds. *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **20**, 161–172.

Byrne, S. and Girolami, M. (2013) Geodesic Monte Carlo on embedded manifolds. *Scand. J. Statist.*, **40**, 825–845.

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

Chamberlain, G. and Imbens, G. (2003) Nonparametric applications of Bayesian inference. *J. Bus. Econ. Statist.*, **21**, 12–18.

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

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

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

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. Berkeley Symp. Mathematical Statistics*, vol. 4 (ed. J. Neyman), pp. 105–123. Berkeley: University of California Press.

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

Doss, H. (1985) Bayesian non-parametric estimation of the median; part i: Computation of the estimates. *Ann. Statist.*, **13**, 1432–1444.

Durbin, J. (1960) Estimation of parameters in time-series regression models. *J. R. Statist. Soc.* B, **22**, 139–153.

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

Fang, Y., Sanz-Serna, J. M. and Skeel, R. D. (2014) Compressible generalized hybrid Monte Carlo. *J. Chem. Phys.*, **140**, article 174108.

Farrel, L. and Walker, I. (1999) The welfare effects of lotto: evidence from the UK. *J. Publ. Econ.*, **72**, 99–120.

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

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

Florens, J. and Simoni, A. (2015) Gaussian processes and Bayesian moment estimation. *Unpublished*. Centre de Recherche en Economie et Statistique, Paris.

Gallant, A. R. (2015) Reflections on the probability space induced by moment conditions with implications for Bayesian inference. *J. Finan. Econmetr.*, **14**, 229–247.

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

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

Gallant, A. R. and Tauchen, G. (1989) Semi non-parametric estimation of conditionally constrained heterogeneous processes. *Econometrica*, **57**, 1091–1120.

Gallant, A. R. and Tauchen, G. (1996) Which moments to match. *Econmetr. Theory*, **12**, 657–681.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, **87**, 523–532.

Gelman, A., Carlin, J. B., Dunson, D. B., Stern, H. S., Vehtari, A. and Rubin, D. B. (2003) *Bayesian Data Analysis*, 3rd edn. Boca Raton: Chapman and Hall.
Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317–1339.
Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1212.
Golchi, S. and Campbell, D. A. (2014) Sequentially constrained Monte Carlo. *Unpublished*. Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby.
Gourieroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *J. Appl. Econmetr.*, **8**, S85–S118.
Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
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**, 1029–1054.
Hansen, L. P., Heaton, J. and Yaron, A. (1996) Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Statist.*, **14**, 262–280.
Hartmann, C. (2008) An ergodic sampling scheme for constrained Hamiltonian systems with applications to molecular dynamics. *J. Statist. Phys.*, **130**, 687–711.
Hartmann, C. and Schütte, C. (2005a) A constrained hybrid Monte-Carlo algorithm and the problem of calculating the free energy in several variables. *J. Appl. Math. Mech.*, **85**, 700–710.
Hartmann, C. and Schütte, C. (2005b) A geometric approach to constrained molecular dynamics and free energy. *Communs Math. Sci.*, **3**, 1–20.
Hurn, M. A., Rue, H. and Sheehan, N. A. (1999) Block updating in constrained Markov chain Monte Carlo sampling. *Statist. Probab. Lett.*, **41**, 353–361.
Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.
Imbens, G. W., Rubin, D. B. and Sacerdote, B. I. (2001) Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. *Am. Econ. Rev.*, **91**, 778–794.
Imbens, G. W., Spady, R. H. and Johnson, P. (1998) Information theoretic approaches to inference in moment condition models. *Econometrica*, **66**, 333–358.
Jaynes, E. (2003) *Probability Theory*. New York: Cambridge University Press.
Kessler, D. C., Hoff, P. D. and Dunson, D. B. (2015) Marginally specified priors for non-parametric Bayesian estimation. *J. R. Statist. Soc.* B, **77**, 35–58.
Khatri, C. G. (1968) Some results for the singular normal multivariate regression models. *Sankhya* A, **30**, 267–280.
Kitamura, Y. (2007) Empirical likelihood methods in econometrics: theory and practice. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, vol. 3 (eds R. Blundell, W. K. Newey and T. Persson), pp. 174–237. Cambridge: Cambridge University Press.
Kitamura, Y. and Otsu, T. (2011) Bayesian analysis of moment restriction models using nonparametric priors. *Unpublished*. Department of Economics, Yale University, New Haven.
Kolmogorov, A. (1956) *Foundations of the Theory of Probability*. New York: Chelsea Publishing.
Kwan, Y. K. (1998) Asymptotic Bayesian analysis based on a limited information estimator. *J. Econmetr.*, **88**, 99–121.
Lancaster, T. and Jun, S. J. (2010) Bayesian quantile regression methods. *J. Appl. Econmetr.*, **25**, 287–307.
Lazar, N. A. (2003) Bayesian empirical likelihood. *Biometrika*, **90**, 319–326.
Leimkuhler, B. and Matthews, C. (2016) Efficient molecular dynamics using geodesic integration and solvent–solute splitting. *Proc. R. Soc. Lond.* A, **472**, article 0138.
Leimkuhler, B. and Reich, S. (2004) *Simulating Hamiltonian Dynamics*. New York: Cambridge University Press.
Lelièvre, T., Rousset, M. and Stoltz, G. (2010) *Free Energy Computations: a Mathematical Perspective*. London: Imperial College Press.
Lelièvre, T., Rousset, M. and Stoltz, G. (2012) Langevin dynamics with constraints and computation of free energy differences. *Math. Comp.*, **81**, 2071–2125.
Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis of Missing Data*, 2nd edn. 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 *Proc. Symp. Monte Carlo Methods* (ed. M. Meyer), pp. 123–140. New York: Wiley.
McCandless, L. C., Gustafson, P. and Austin, P. C. (2009) Bayesian propensity score analysis for observational data. *Statist. Med.*, **28**, 94–112.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Mengersen, K., Pudlo, P. and Robert, C. P. (2013) Bayesian computation via empirical likelihood. *Proc. Natn. Acad. Sci. USA*, **110**, 1321–1326.
Morgan, F. (2016) *Geometric Measure Theory: a Beginner's Guide*. New York: Academic Press.

Muller, U. (2013) Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, **81**, 1805–1849.
Newton, M., Czado C. and Chappell, R. (1996) Semiparametric Bayesian inference for binary regression. *J. Am. Statist. Ass.*, **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. *Ann. Statist.*, **18**, 90–120.
Owen, A. (1991) Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725–1747.
Owen, A. (2001) *Empirical Likelihood*. London: Chapman and Hall.
Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond.* A, **185**, 71–110.
Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Rubin, D. B. (1988) Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 395–402. Oxford: Clarendon.
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. *J. R. Statist. Soc.* B, **21**, 91–105.
Schennach, S. M. (2005) Bayesian exponentially tilted empirical likelihood. *Biometrika*, **92**, 31–46.
Shin, M. (2014) Bayesian GMM. *Unpublished*. Department of Economics, University of Pennsylvania, Philadelphia.
Strachan, R. W. and van Dijk, H. K. (2004) Valuing structure, model uncertainty and model averaging in vector autoregressive processes. *Report EI 2004-23*. Econometric Institute Erasmus Institute, Amsterdam.
Sun, D., Speckman, P. L. and Tsutakawa, R. K. (1999) Random effects in generalized linear mixed models. *Unpublished*. National Institute of Statistical Sciences, Washington DC.
Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton methods. *Biometrika*, **61**, 439–447.
West, M. (2003) Bayesian factor regression models in the large p, small n paradigm. In *Bayesian Statistics 7* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 733–742. Oxford: Oxford University Press.
White, H. (1994) *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.
Yang, Y. and He, X. (2012) Bayesian empirical likelihood for quantile regression. *Ann. Statist.*, **40**, 1102–1131.
Yin, G. (2009) Bayesian generalized method of moments. *Baysn Anal.*, **4**, 191–208.
Zellner, A. (1997) The Bayesian method of moments (BMOM): theory and applications. *Adv. Econmetr.*, **12**, 85–105.
Zellner, A., Tobias, J. and Ryu, H. (1997) Bayesian method of moments (BMOM) analysis of parametric and semiparametric regression models. *Proc. Baysn Statist. Sci. Sect. Am. Statist. Ass.*, 211–216.
Zigler, C. M. and Dominici, F. (2014) Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *J. Am. Statist. Ass.*, **109**, 95–107.
Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A. and Dominici, F. (2013) Model feedback in Bayesian propensity score estimation. *Biometrics*, **69**, 263–273.