Bayesian Analysis of Moment Condition Models

Demetrius Rowland

Statistical Modeling I

26 November 2019

Contents

Moment Condition Models

- Definition: A moment condition model is a set of conditions $\mathbb{E}^P[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})] = \mathbf{0}$ where

  P is an unknown probability distribution

  $\boldsymbol{X} \in \mathbb{R}^{d_x}$ is a random vector generated from P

  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is a vector of the parameters of the distribution P

  $\boldsymbol{g} : \mathbb{R}^{d_x} \times \Theta \to \mathbb{R}^d$ is a vector of known functions

- We also assume $d \geq p$.
- Example: The linear regression model can be indirectly described using

$$\mathbb{E}[y - \beta_0 - \beta_1 x] = 0$$

$$\mathbb{E}[(y - \beta_0 - \beta_1 x) * x] = 0$$

Augmentation and Misspecification

- In the case that the number of moment conditions is greater than the number of parameters ($d > p$), then it is possible that, under the true probability distribution P, not all conditions will be satisfied.
- To correct this, we introduce the parameters $\boldsymbol{V} \in \mathcal{B} \subset \mathbb{R}^d$, where the k^{th} component of $\boldsymbol{V}$ is a free parameter if the k^{th} moment condition is violated and all other components are zero.
- We formulate the augmented conditions as

$$\mathbb{E}\left[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})\right] = \boldsymbol{0}$$

$$\text{with } \boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V}) = \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{V}$$

- If, under the true probability distribution P, there do not exist $(\boldsymbol{\theta}, \boldsymbol{V}) \in (\Theta, \mathcal{B})$ that satisfy the augmented conditions, we say that the conditions are misspecified.

Bayesian Analysis under ETEL Framework

- Suppose that a set $\boldsymbol{X}_{1:n}$ of random vectors $\boldsymbol{X}_i$ are generated independently and identically from an unknown probability distribution P with parameters $\boldsymbol{\theta}$.
- A bayesian framework for determining the distribution of the parameters $\boldsymbol{\theta}, \boldsymbol{v}$ is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{v} | \boldsymbol{X}_{1:n}) \propto \pi(\boldsymbol{\theta}, \boldsymbol{v}) * \mathbb{P}(\boldsymbol{X}_{1:n} | \boldsymbol{\theta}, \boldsymbol{v})$$

where

$\boldsymbol{v}$ is a vector of the nonzero components of $\boldsymbol{V}$

$\mathbb{P}(\boldsymbol{X}_{1:n} | \boldsymbol{\theta}, \boldsymbol{v})$ is the Exponentially-Tilted Empirical Likelihood which will be defined shortly.

- The task is to choose a prior for $\boldsymbol{\theta}, \boldsymbol{v}$ and update it using the above framework.

Bayesian Analysis under ETEL Framework

- Let $p_i = \mathbb{P}(\boldsymbol{X} = \boldsymbol{X_i}|\boldsymbol{\theta}, \boldsymbol{v})$.
- The ETEL is constructed so as to minimize the KL-divergence between the probabilities assigned to each sample point $(p_1, ..., p_n)$ and the empirical probabilities $(\frac{1}{n}, ..., \frac{1}{n})$ under the constraints that $p_1, ..., p_n$ sum to 1 and the moment conditions are satisfied.
- This is a constrained optimization problem, which has solution

$$p_i = \frac{\exp(\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{v}) \cdot \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v}))}{\sum_{j=1}^{n} \exp(\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{v}) \cdot \boldsymbol{g}^A(\boldsymbol{x}_j, \boldsymbol{\theta}, \boldsymbol{v}))}$$

$$\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{v}) = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^d} \sum_{i=1}^{n} \exp(\boldsymbol{\lambda} \cdot \boldsymbol{g}^A(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{v}))$$

which is obtained by forming the Lagrangian and solving the corresponding dual problem.

Asymptotic Properties

Let $\psi = (\boldsymbol{\theta}, \boldsymbol{v})$. Then the Bayesian ETEL-framework has desirable asymptotic properties, outlined below.

Correctly Specified:

- Let ψ^* denote the true parameters for the population distribution P and let $\Delta = \mathbb{E}^P[\boldsymbol{g}^A(\boldsymbol{X}, \psi^*)\boldsymbol{g}^A(\boldsymbol{X}, \psi^*)^T]$, $\Gamma = \mathbb{E}^P[\frac{\partial}{\partial \psi}\boldsymbol{g}^A(\boldsymbol{X}, \psi^*)]$
- Then $\sqrt{n}(\psi - \psi^*)$ converges in probability to a normal distribution with center $\mathbf{0}$ and covariance matrix $\Gamma^T \Delta^{-1} \Gamma$

Misspecified:

- Since there is no true population parameter ψ that satisfies the moment conditions, we define a pseudo-true value $\psi = \psi_0$ for which KL-divergence between the probability $Q(\boldsymbol{X}|\psi)$ and the population distribution P is minimized.
- Then $\sqrt{n}(\psi - \psi_0)$ converges in probability to a normal distribution with finite mean and variance.

Model Selection

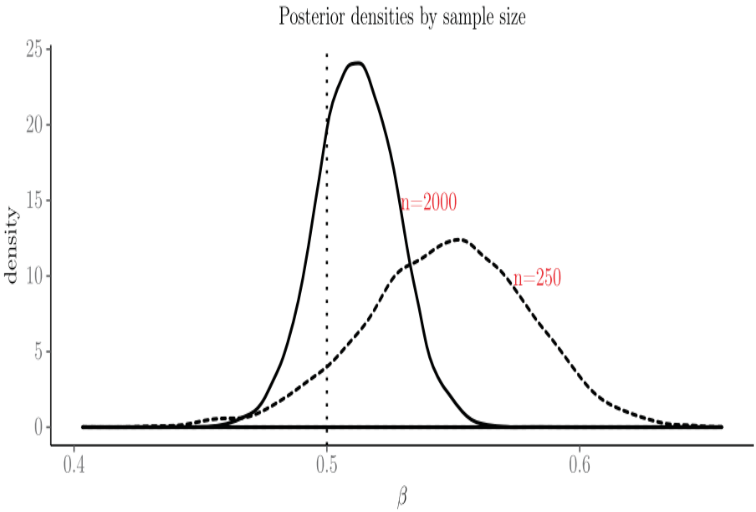
- Consider a set of moment condition models $\mathbb{E}[\boldsymbol{g}^{\ell}(\boldsymbol{X}, \boldsymbol{\theta}^{\ell})] = \mathbf{0}$ that describe a data set $\boldsymbol{X}_{1:n}$ generated from a distribution P, with $\ell = 1, 2, \ldots$
- The task is to develop criteria for selecting the best model and choose accordingly.
- This is accomplished by setting a prior for each ψ_{ℓ}, forming the marginal likelihood $m(\boldsymbol{X}_{1:n}|M_{\ell})$ and choosing the model with the highest marginal likelihood.
- With some work, we can guarantee the following results:
  - If all models are correctly specified, the model with the maximum number of overidentifying moment conditions is chosen.
  - If all models are misspecified, the model that gives a probability distribution Q for $\boldsymbol{X}$ for which $D_{KL}(P||Q)$ is minimized is chosen.
  - If some models are correctly specified and some are not, the model that is correctly specified and has the maximum number of overidentifying moment conditions is chosen.

Example

IV Regression:

- Consider a model of the form $y = \alpha + \beta x + \delta w + \epsilon, \quad \mathbb{E}[\epsilon] = 0$
- Suppose that there exist instrumental variables z_1 and z_2 uncorrelated with ϵ but correlated with x.
- Then we generate n data points of the form (y, x, z_1, z_2) with
  $y = 1 + .5x + .7w + \epsilon$
  $x = z_1 + z_2 + w + u$
  $z_i \sim \mathcal{N}(0, .5)$
  $w \sim \text{Unif}(0, 1)$
  (ϵ, u) are drawn from a Gaussian copula with covariance diagonal $[1, 1]$ and non-diagonal entry .7
- Then implementing the Bayesian-ETEL framework on this generated data for n $=$ 200 and n $=$ 2000 yields the following posteriors for β given a prior $\beta \sim t_{2.5}(0, 5^2)$.

Example


Posterior densities by sample size
density
25
20
15
10
5
0
0.4
0.5
0.6
β
n=2000
n=250