Bayesian Inference for Moment Condition Models

Demetrius Rowland
University of Texas at Austin

2019
December

Contents

1 Overview

Suppose that we are studying a d_x dimensional data set $\boldsymbol{X}_{1:n}$ with unknown distribution P. It may be difficult or computationally prohibitive to model the distribution P, so we choose instead to make use of moment conditions $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})] = \boldsymbol{0}$ that we believe describe P, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ are the parameters of P and $\boldsymbol{g}$ is a vector of d known functions. This set of conditions is known as a moment condition model. The task is to conduct bayesian inference for parameters $\boldsymbol{\theta}$ underlying random variables $\boldsymbol{X}_{1:n}$ with distribution P. To do so we need to construct a likelihood function that mimics the true likelihood, which is unknown, and that satisfies the moment conditions. We also need to construct

suitable priors for the parameters $\boldsymbol{\theta}$. Once we have done so, we can sample from the posterior distribution by way of Bayes' Theorem and with the one-block tailored Metropolis-Hastings algorithm. If we wish to compare moment condition models, we can compute the posterior distribution for the parameters $\boldsymbol{\theta}$ under each model and select the model with the highest marginal likelihood.

2 Moment Condition Models

Definition: A moment condition model is a set of conditions

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})] = \mathbf{0}$$

where $\boldsymbol{X}$ has unknown distribution P, $\boldsymbol{\theta}$ are the parameters of P, and $\boldsymbol{g}$ is a vector of known functions. Consider the circumstance in which, under the true distribution P, there do not exist a set of parameters $\boldsymbol{\theta}$ which satisfy all of the moment conditions. This is undesirable, so we introduce nuisance parameters $\boldsymbol{V} \in \mathcal{V} \subset \mathbb{R}^d$ to account for this possibility. Specifically, we subtract $\boldsymbol{V}$ from the left hand side of the moment conditions to obtain the augmented moment conditions

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})] = \mathbf{0} \\
&\text{where} \\
&\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V}) := \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) - \boldsymbol{V}
\end{aligned}$$

The k^{th} component of $\boldsymbol{V}$ is a free parameter v_k if the k^{th} moment condition is violated and zero otherwise.

We have now laid the groundwork to introduce the meaning of a misspecified model.

3 Misspecification

Definition: We say a moment condition model $\mathbb{E}[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})] = \mathbf{0}$ is misspecified iff there do not exist $\boldsymbol{\theta}, \boldsymbol{V} \in \Theta \times \mathcal{V}$ such that, under the true distribution P, the moment conditions are satisfied.

4 Exponentially Tilted Empirical Likelihood

Because we do not know the distribution P that underlies the data $\boldsymbol{X}_{1:n}$, we must approximate this distribution by way of an empirical likelihood, which is

the probability distribution that maps each $\boldsymbol{X}_i$ to the probability $\frac{1}{n}$. Specifically, we want to find a distribution $(p_1, p_2, ..., p_n)$ that mimics the empirical likelihood but satisfies the moment conditions $\mathbb{E}[\boldsymbol{g}^A(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V})] = \boldsymbol{0}$. Since the task is to mimic the empirical likelihood, we use the KL-divergence D [candidate distribution||empirical likelihood] to assess the closeness of candidate distributions to the empirical likelihood. The distribution which minimizes this KL-divergence and satisfies the moment conditions is the Exponentially Tilted Empirical Likelihood (ETEL), for reasons that will become apparent. Under this framework, we can see that the ETEL solves the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} p_i \log\{n * p_i\} \\ \text{subject to} \quad & \sum_{i=1}^{n} p_i = 1 \\ & \sum_{i=1}^{n} p_i * \boldsymbol{g}^A(\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{V}) = \boldsymbol{0} \end{aligned}$$

Observe that the objective function is the KL divergence
D $[(p_1, ..., p_n)||(1/n, ..., 1/n)]$. The first constraint is due to the fact that $(p_1, ..., p_n)$ must be a probability distribution. The rest of the constraints are due to the moment conditions. With use of Lagrange multipliers, the optimizer is

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

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

The full derivation can be found in the Appendix.

5 Posterior Distribution

Equipped with a likelihood function, we can now infer a posterior distribution for the parameters $\boldsymbol{\theta}, \boldsymbol{V}$ given observations $\boldsymbol{X}_{1:n}$ using Bayes' Theorem:

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

We use priors specific to the problems of interest, and many different priors are used in the Examples section. Since we now have a function proportional to the posterior distribution for the parameters $(\boldsymbol{\theta}, \boldsymbol{V})$, we can use the block MH sampler to approximate values from the true posterior.

6 MCMC Sampling

The following pseudocode describes the implementation of the block MH sampler on the posterior above:

Algorithm 1 Sample from $\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{V}|\boldsymbol{X}_{1:n})$

Define proposal distribution $q(\boldsymbol{\theta}, \boldsymbol{V}|\boldsymbol{X}_{1:n})$
for s=1,...,S **do**
  Sample $(\boldsymbol{\theta}, \boldsymbol{V}) \sim q(\boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{X}_{1:n})$
  Solve $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{V}) = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^d} \sum_{i=1}^n \exp(\boldsymbol{\lambda} \cdot \boldsymbol{g}^A(\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{V}))$
  Set $p_i = \dfrac{\exp(\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{V}) \cdot \boldsymbol{g}^A(\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{V}))}{\sum_{j=1}^n \exp(\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{V}) \cdot \boldsymbol{g}^A(\boldsymbol{X}_j, \boldsymbol{\theta}, \boldsymbol{V}))}$
  Set $\alpha = \min\{1, \frac{\mathbb{P}(\boldsymbol{\theta}, \boldsymbol{V}|\boldsymbol{X}_{1:n})}{\mathbb{P}(\boldsymbol{\theta}^{s-1}, \boldsymbol{V}^{s-1}|\boldsymbol{X}_{1:n})} \frac{q(\boldsymbol{\theta}^{s-1}, \boldsymbol{V}^{s-1}|\boldsymbol{X}_{1:n})}{q(\boldsymbol{\theta}, \boldsymbol{V}|\boldsymbol{X}_{1:n})}\}$
  Sample $u \sim \text{Unif}[0, 1]$
  if $u \leq \alpha$ **then**
    Set $(\boldsymbol{\theta}^s, \boldsymbol{V}^s) = (\boldsymbol{\theta}, \boldsymbol{V})$
  else
    Set $(\boldsymbol{\theta}^s, \boldsymbol{V}^s) = (\boldsymbol{\theta}^{s-1}, \boldsymbol{V}^{s-1})$
  end if
end for

In the examples that follow, we typically set S = 11,000 with 1,000 samples designated for burn-in and discard all but every fifth sample for thinning to obtain 2,000 samples.

7 Asymptotic Properties of Posterior

Let v the vector containing the nonzero components of V and let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{v})$. In order to establish the results that follow, we make extensive use of the theorem below. Recall that a Borel set is any set that can be formed through countable union, countable intersection, or the relative complement of open sets.

Theorem. (Bernstein von Mises' Theorem) Suppose $x_1, ..., x_n$ are sampled i.i.d. from probability distribution P with true parameters $\boldsymbol{\theta}_0$. Let B be a Borel set and let $J(\boldsymbol{\theta})$ be the Fisher information matrix for P evaluated at $\boldsymbol{\theta}$. Let $\Sigma = J(\boldsymbol{\theta}_0)^{-1}$. Then

$$\sup_B |\mathbb{P}(\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in B|\boldsymbol{x}_{1:n}) - \text{Norm}(B|\boldsymbol{0}, \Sigma)| \to_p 0$$

where $\to_p$ denotes convergence in probability and $\sup_B$ means forming the set S whose members are applications of the above expression on the left side of the arrow on the members of the Borel set B and taking the least upper bound of S.

With this theorem given, we can establish an important asymptotic property of the posterior distribution for the parameters $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{v})$.

Correct Specification

Suppose the set of moment conditions is correctly specified. Let $\boldsymbol{\psi}^*$ be the true parameters of the distribution P that underlies $\boldsymbol{x}_{1:n}$. Then if we define $\Delta = \mathbb{E}[\boldsymbol{g}^A(X, \boldsymbol{\psi}^*)\boldsymbol{g}^A(X, \boldsymbol{\psi}^*)^T]$ and $\Gamma = \mathbb{E}[\frac{\partial}{\partial \boldsymbol{\psi}^*}\boldsymbol{g}^A(X, \boldsymbol{\psi}^*)]$, we must have that

$$\sup_B |\mathbb{P}(\sqrt{n}(\boldsymbol{\psi} - \boldsymbol{\psi}^*) \in B | \boldsymbol{x}_{1:n}) - \text{Norm}(B|\mathbf{0}, (\Gamma^T \Delta^{-1} \Gamma)^{-1})| \to_p 0$$

The result is that, under correct specification, the posterior for $\boldsymbol{\psi}$ converges asymptotically to a normal distribution, the center of which is the true parameter $\boldsymbol{\psi}_0$ of the probability distribution P that underlies the data $\boldsymbol{x}_{1:n}$.

8 Bayesian Model Selection

Suppose we are given a finite number of moment condition models describing a data set $\boldsymbol{x}_{1:n}$ which we will label $M_1, M_2, ..., M_K$. How can we compare the models given the data $\boldsymbol{x}_{1:n}$? One option is to compute the marginal likelihood for each model and select the model which maximizes this quantity.

The task is then to compute $\mathbb{P}(\boldsymbol{x}_{1:n}|M_i)$ for $i = 1, 2, ..., K$. Observe that by Bayes' theorem

$$\mathbb{P}(\boldsymbol{\psi}|\boldsymbol{x}_{1:n}, M_i) = \frac{\mathbb{P}(\boldsymbol{x}_{1:n}|\boldsymbol{\psi}, M_i) * \mathbb{P}(\boldsymbol{\psi}|M_i)}{\mathbb{P}(\boldsymbol{x}_{1:n}|M_i)}$$

Therefore

$$\mathbb{P}(\boldsymbol{x}_{1:n}|M_i) = \frac{\mathbb{P}(\boldsymbol{x}_{1:n}|\boldsymbol{\psi}, M_i) * \mathbb{P}(\boldsymbol{\psi}|M_i)}{\mathbb{P}(\boldsymbol{\psi}|\boldsymbol{x}_{1:n}, M_i)}$$

Notice that we can take any vector $\boldsymbol{\psi}_0$ in the support of $\boldsymbol{\psi}$ and compute the ETEL given this vector $\boldsymbol{\psi}_0$, which gives us the left side of the numerator. We obtain the right side of the numerator by evaluating the prior for $\boldsymbol{\psi}$ at $\boldsymbol{\psi}_0$. To obtain the denominator, we can sample points from the posterior for $\boldsymbol{\psi}_0$ given $\boldsymbol{x}_{1:n}$ via the MCMC algorithm detailed above. We can then assign a posterior probability to the point $\boldsymbol{\psi}_0$ equal to its relative frequency among the sampled points. We obtain the marginal likelihood $\mathbb{P}(\boldsymbol{x}_{1:n}|M_i)$ by combining these three values.

By working out the appropriate theory, we can establish that any algorithm which selects the model M_i among the finite candidate models $M_1, ..., M_K$ will satisfy the following three requirements:

1. **If all models are correctly specified:** Then the algorithm will choose the model with the maximum number of over-identifying moment conditions, i.e., the model for which d - p is maximized, where d is the number of moment conditions and p is the dimension of $\boldsymbol{\theta}$, the parameters to be estimated.

2. **If all models are misspecified:** Then the algorithm will choose the model which minimizes the KL-divergence $D_{KL}(P||Q)$ between the probability distribution Q for x_i implied by the model and which satisfies the moment conditions, and the true probability distribution P that underlies the data $\boldsymbol{x}_{1:n}$.

3. **If some models are correctly specified and some models are misspecified:** Then the algorithm will choose the model which is correctly specified and which maximizes the number of over-identifying moment conditions.

9 Examples

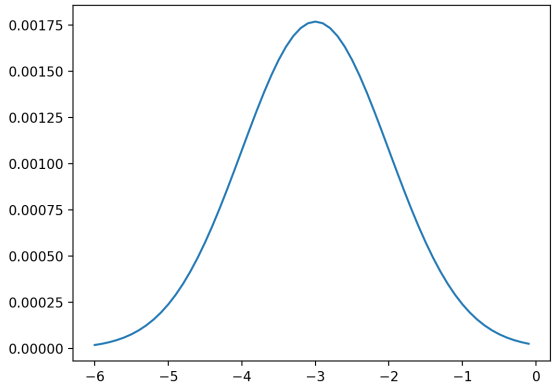
With all the theory established, we can now implement the full process on several data sets. For our original data set, I've chosen "Seismic Source Data for U.S. Below-Surface Nuclear Tests, 1946 − 1992" introduced in [1].

Toy Data Set

We generate a toy data set by sampling data points $\boldsymbol{x}_{1:1000}$ from a standard normal distribution. We then assume a model of the form

$$x = c + \epsilon$$

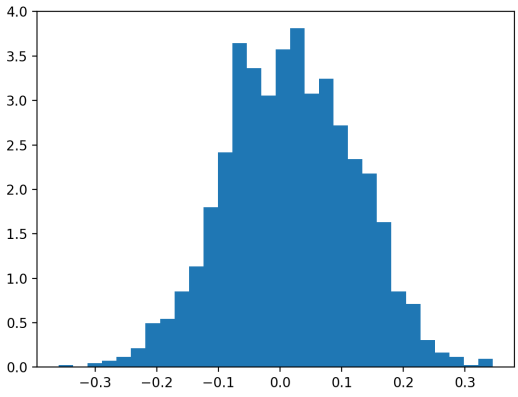
with ϵ a noise parameter, whose distribution is unknown. We then try to estimate the parameter c of the distribution of x, by first fixing a prior distribution for c that is normal with mean -3 and variance 1. It is plotted below.



We use the moment condition model

$$\mathbb{E}[\epsilon] = \mathbb{E}[x - c] = 0$$

to describe the data $\boldsymbol{x}_{1:n}$. We know that this model is correctly specified because the center of a standard normal distribution is zero. Under this model, we use MCMC block sampling to sample from the posterior for c given the generated data $\boldsymbol{x}_{1:n}$. The histogram of the sampled data points is included below.



We see that the distribution hovers very closely around the true value c = 0 and that the variance has reduced considerably, giving us hope that the method will work nicely for the data sets that follow.

Nuclear Test Data

The nuclear test data set consists of crater diameter, yield, location, etc. values for underground nuclear detonations between 1946 and 1992. Specifically, we wish to build a model which relates the crater radii to the energy of the detonations. We build two models, the first of which is the model used by Fermi during the Trinity test:

$$R = S(\frac{E * t^2}{\rho})^{1/5} + \epsilon$$

where R is the crater radius, E is the energy, ρ is atmospheric pressure, t is time until completion of the detonation, S is the coefficient to be estimated, and ϵ is noise, and we use moment condition

$$\mathbb{E}[\epsilon] = \mathbb{E}[R - S(\frac{E * t^2}{\rho})^{1/5}] = 0$$

We expect S to be positive, so we use an exponential prior $S \sim \exp(1)$.

The second model we use describes a simple linear relationship between the energy E and the radius R:

$$R = S * E + \epsilon$$

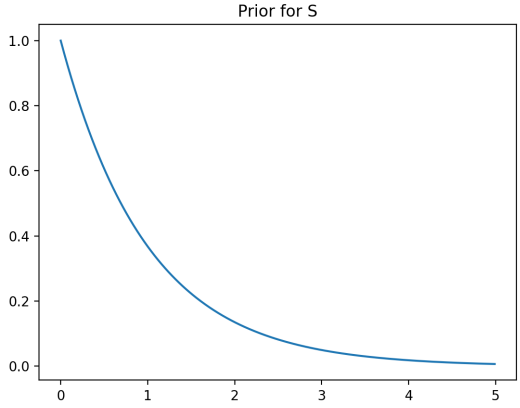
where S is the coefficient to be estimated and ϵ is noise. The moment condition we use is

$$\mathbb{E}[\epsilon] = \mathbb{E}[R - S * E] = 0$$

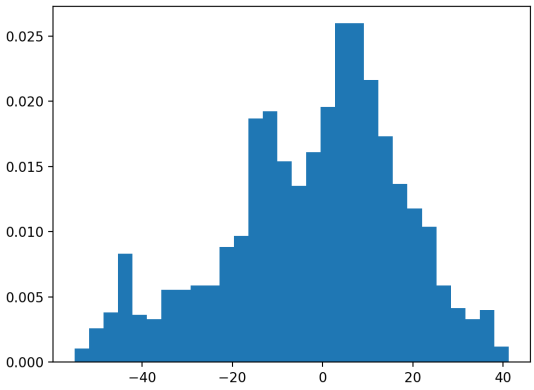
We also use an exponential prior $S \sim \exp(1)$.

Below we plot the priors and histograms of the posteriors for S under each model.

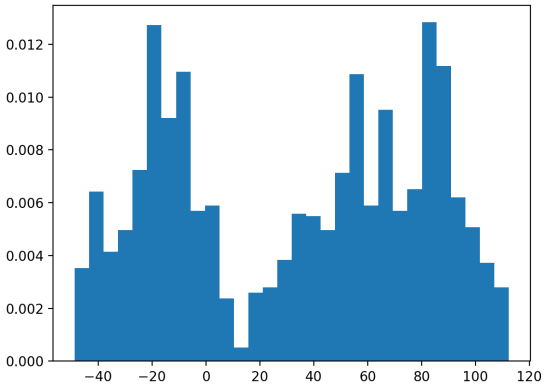
Prior



Model 1



Model 2



We can now use the Bayesian Model Selection methodology to compute the marginal likelihood for Model 1 and Model 2 and choose the one which maximizes the marginal likelihood of the observations $\boldsymbol{x}_{1:n}$. If we let m_1, m_2 denote the marginal likelihoods under Model 1 and 2 respectively, then we have that, by the formula

$$\mathbb{P}(\boldsymbol{x}_{1:n}|M_i) = \frac{\mathbb{P}(\boldsymbol{x}_{1:n}|\boldsymbol{\psi}, M_i) * \mathbb{P}(\boldsymbol{\psi}|M_i)}{\mathbb{P}(\boldsymbol{\psi}|\boldsymbol{x}_{1:n}, M_i)}$$

$(m_1, m_2) = (.026, .013)$. So we say that Model 1 is the better choice for a description of the relationship between the energy yield and the radius of the nuclear blast.

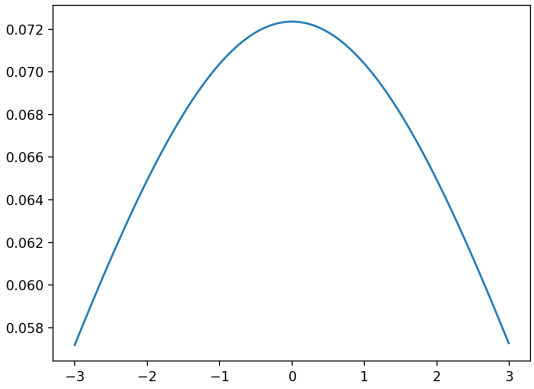
Count Regression

We have a model of the form

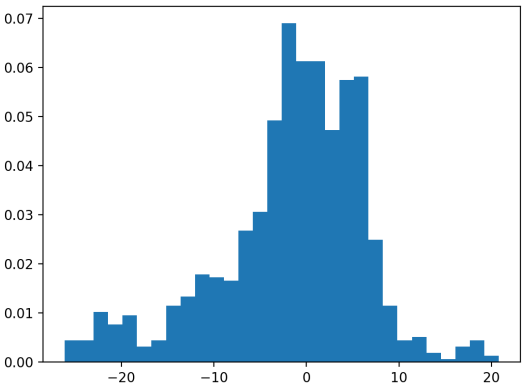
$$y_i \sim \text{Neg-Bin}(\frac{p}{1-p}u_i, p) u_i = \exp\{x_i\beta\}$$

For simulation we set the true value of β to be 1. We generate x_i according to $x_i \sim \text{Norm}(.4, 1/9)$ and we set $p = 1/2$.

We now wish to infer β given the prior $\beta \sim t_{2.5}(0, 5)$. The prior is plotted below.



We generate 200 points according to the above scheme and sample 2000 burned-in and thinned points from the posterior to obtain the histogram below.



We see that the posterior hovers around the true value $\beta = 1$.

IV Regression

We have a model of the form

$$y = 1 + \beta x + .7w + \epsilon$$

where we use the moment conditions

$$\mathbb{E}[\epsilon] = \mathbb{E}[y - 1 - \beta x - .7w] = 0$$

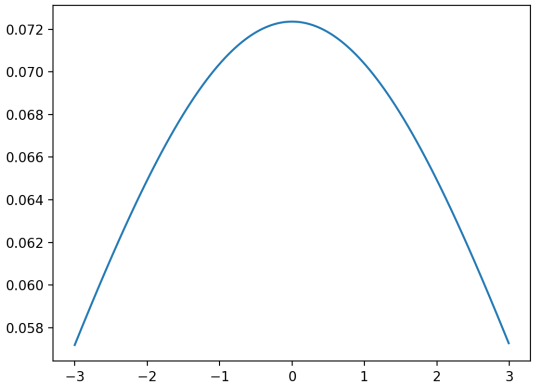
For simulation, we set the true value $\beta = .5$ and we have

$$x = z_1 + z_2 + w + u$$

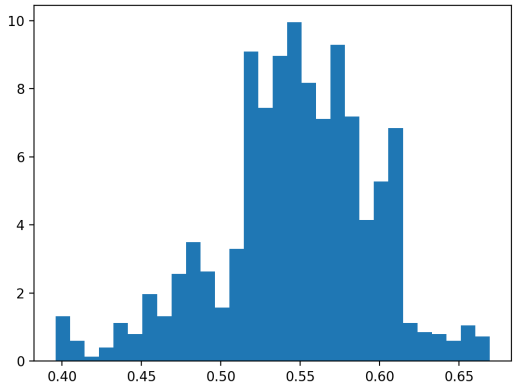
where $z_1, z_2 \sim \text{Norm}(.5, 1)$ and $w \sim \text{Unif}[0, 1]$.

Meanwhile ϵ, u are drawn from a Gaussian copula with covariance matrix $\begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$

The task is to estimate β. We choose a prior $\beta \sim t_{2.5}(0, 5)$. It is plotted below.



We sample n = 200 data points and include a histogram of the posterior for β below.



We see that the posterior centers asymptotically around the true value $\beta = .5$.

10 Appendix

Solving the Dual Problem

We wish to solve the constrained optimization problem

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} p_i \log\{n * p_i\} \\ \text{subject to} \quad & \sum_{i=1}^{n} p_i = 1 \\ & \sum_{i=1}^{n} p_i * \boldsymbol{g}^A(\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{V}) = \boldsymbol{0} \end{aligned}$$

So we form the Lagrangian

$$L = \sum_{i=1}^{n} p_i \log\{np_i\} + \lambda_0(\sum_{i=1}^{n} p_i - 1) + \lambda_1 \sum_{i=1}^{n} p_i g_1^A(X_i, \boldsymbol{\theta}, \boldsymbol{V}) + ... + \lambda_d \sum_{i=1}^{n} p_i g_d^A(X_i, \boldsymbol{\theta}, \boldsymbol{V})$$

where $\lambda_0, \lambda_1, ..., \lambda_d$ are the unknown Lagrange multipliers. The KKT conditions tell us that the solution to the above optimization is a saddle point of the Lagrangian, i.e. it is a minimizer for L in the direction of $p_1, ..., p_n$ and a maximizer in the direction of the Lagrange multipliers $\lambda_0, ..., \lambda_d$. Because the optimal $p_1, p_2, ..., p_n, \lambda_0, ..., \lambda_d$ must be a critical point of L, we must differentiate L in each p_k to find p_k in terms of $\lambda_0, ..., \lambda_d$. This gives

$$\begin{aligned} \frac{\partial}{\partial p_k} L &= \log\{np_k\} + \frac{1}{n} + \lambda_0 + \boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V}) \\ \therefore p_k &= \frac{1}{n} \exp\{-\frac{1}{n} - \lambda_0 - \boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V})\} \end{aligned}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$. We know that the probabilities p_k must sum to one, so we have

$$\begin{aligned} \sum_{k=1}^{n} \frac{1}{n} \exp\{-\frac{1}{n} - \lambda_0 - \boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V})\} &= 1 \\ \exp\{-\lambda_0 - \frac{1}{n}\} \sum_{k=1}^{n} \exp\{-\boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V})\} &= n \\ \therefore \exp\{-\lambda_0 - \frac{1}{n}\} &= \frac{n}{\sum_{k=1}^{n} \exp\{-\boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V})\}} \end{aligned}$$

Substituting this into the expression for p_k we obtain

$$p_k = \frac{\exp\{-\boldsymbol{\lambda}^T \boldsymbol{g}^A(X_k, \boldsymbol{\theta}, \boldsymbol{V})\}}{\sum_{j=1}^{n} \exp\{-\boldsymbol{\lambda}^T \boldsymbol{g}^A(X_j, \boldsymbol{\theta}, \boldsymbol{V})\}}$$

We can substitute this expression for p_k into the Lagrangian and with some tedious computations we obtain

$$L = -\frac{1}{n}\sum_{i=1}^{n} \exp\{\boldsymbol{\lambda}^T \boldsymbol{g}^A(X_i, \boldsymbol{\theta}, \boldsymbol{V})$$

The optimal $\boldsymbol{\lambda}$, i.e. $\hat{\boldsymbol{\lambda}}$, is a maximizer for the above expression. So by maximizing L with respect to $\boldsymbol{\lambda}$ we obtain our desired expression for optimal $p_1, ..., p_n$.

References

[1] Donald L. Springer, Gayle A. Pawloski, Janet L. Ricca, Robert F. Rohrer, and David K. Smith. *Seismic Source Summary for All U.S. Below-Surface Nuclear Explosions*. Bulletin of the Seismological Society of America, 2002.

[2] Siddhartha Chib, Minchul Shin and Anna Simoni. *Bayesian Estimation and Comparison of Moment Condition Models*. Journal of the American Statistical Association, 2018.