Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Bayesian StoNED or Multivariate Bayesian Convex Regression with Inefficiency: Two sides of the same coin

José Luis Preciado Arreola
Andrew L. (Andy) Johnson

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Agenda

Background

- Objective
- Introduction to Bayesian Regression
- Gibbs sampling to estimate Bayesian Linear Regressions

Estimating a stochastic frontier with MBCR

- MBCR with inefficiency
- Implemented Algorithm

Status and further work

- Results, benefits and limitations
- Improvements and additions

Objective

To design a **Bayesian** method that complies with the following characteristics:

- **Semi-nonparametric** production function estimation
- **Multivariate** input vector
- Allows **Axiomatic Restrictions**
- Can model **inefficiency**

Objective

- Nonparametric or Semi-nonparametric
- Multivariate
- Concavity - Constrained
- Inefficiency

Regression Approach	Paper(s)
Least Squares	Kuosmanen and Kortelainen (2012)
Kernel	Parmeter and Racine (2013)
Bayesian	This paper

Connections with StoNED

Similarities

- Monotonic + Concave functional estimators
- No smoothness assumption
- Pointwise consistent

Differences

- MBCR obtains full posterior parameter distributions (allowing direct inference).
- StoNED uses a second stage estimator for inefficiency.
- MBCR is simulation-based, StoNED is constrained optimization.

Bayesian Regression

In Bayesian estimation, given a regression model, we attempt to compute the probability distribution of the regression coefficients *given* the observed data (*posterior* distribution)

$y = \beta x + \varepsilon$ → $p(\beta, \sigma^2 | x, y)$

Regression model

Posterior for β, σ^2

To obtain such distribution, we use Bayes' rule to relate it with two other probability distributions: The *prior* (before data is observed) distribution of the coefficients and the *likelihood function* of y.

$p(\beta, \sigma^2)$

Prior distribution on β, σ^2

$p(y | x, \beta, \sigma^2)$

Likelihood function for y

Bayesian Regression

We then rely on Bayes rule to relate these three distributions:

$$p(\beta, \sigma^2|x, y) = \frac{p(x, y|\beta, \sigma^2) \cdot p(\beta, \sigma^2)}{p(x, y)} \qquad \text{(Bayes rule)}$$

After applying a few simple proportionality and conditional probability properties we obtain :

$$p(\beta, \sigma^2|x, y) \propto p(\beta, \sigma^2) \cdot p(y|x, \beta, \sigma^2)$$

Thus, to obtain valid posterior draws we just need to sample from the distribution given by Prior · Likelihood

Bayesian Regression

A usual assumption is to assume a Normal likelihood function for the data (equivalent to assuming normally distributed noise).

Due to its *conditional conjugacy* given a Normal likelihood, a common assumption for the prior of β, σ^2 is Normal-Inverse Gamma

Prior Distributions

$$p(\beta) \sim N(\widetilde{\mu_\beta}, V_\beta)$$

$$p(\sigma^2) \sim IG(\tilde{a}, \tilde{b})$$

Likelihood Function

$$p(y|x, \beta, \sigma^2) \sim N(x'\beta, \sigma^2)$$

Posterior Distributions

$$p(\beta\,|\sigma^2, x, y) \propto N(\mu^*, V^*)$$

$$p(\sigma^2|\beta, x, y) \propto IG(a^*, b^*)$$

$\widetilde{\mu_\beta}, V_\beta, \tilde{a}, \tilde{b}$ are *prior hyperparameters.*
μ^*, V^*, a^*, b^* *are posterior hyperparameters*

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Gibbs Sampling

Once the conditional posteriors are defined, we draw from them iteratively with a Gibbs sampler:

Draw initial values for β and σ^2 from their priors

$p(\beta) \sim N(\widetilde{\mu_\beta}, V_\beta)$ and $p(\sigma^2) \sim IG(\tilde{a}, \tilde{b})$

↓

Draw values for β and σ^2 from conditional posteriors

$p(\beta \mid \sigma^2, x, y) \propto N(\mu^*, V^*)$ and $p(\sigma^2 \mid \beta, x, y) \propto IG(a^*, b^*)$

for $t_{Total} = t_{Burn-in} + t_{Stationary}$ iterations

↓

Ignore β and σ^2 draws for first $t_{Burn-in}$ iterations. Take β and σ^2 draws for remaining $t_{Stationary}$ iterations as a valid sample from the *joint* posterior distribution of β and σ^2.

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Agenda

Background

- Objective
- Introduction to Bayesian Regression
- Gibbs sampling to estimate Bayesian Linear Regressions

Estimating a stochastic frontier with MBCR

- MBCR with inefficiency
- Implemented Algorithm

Status and further work

- Results, benefits and limitations
- Improvements and additions

MBCR with inefficiency

Production frontier regression model

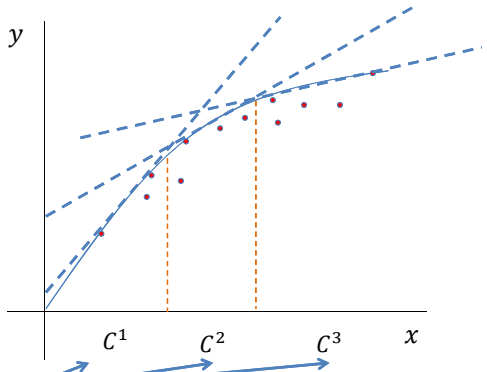
$$y_i = f(\boldsymbol{x}_i) - u_i + v_i \quad \text{where} \quad \begin{matrix} u_i \sim F \\ v_i \sim \mathrm{N}(0, \sigma^2) \end{matrix}$$

- We fit a continuously differentiable single-output, multiple-input production frontier $f(\boldsymbol{x}_i)$.
- Prior assumptions for *F*:
  - Exponential
  - Erlang
  - Gamma

MBCR with inefficiency

MBCR estimation procedure

- MBCR (Hannah and Dunson, 2011) fits functions nonparametrically by fitting dominant hyperplanes on several *basis regions*.



Basis regions

MBCR with inefficiency

- Practical use of MBCR requires hyperplane-specific heteroskedasticity.

$$y_i = \hat{f}(\boldsymbol{x}_i) + v_i \quad \text{where} \quad \hat{f}(\boldsymbol{x}_i) = \min_{k\in\{1,\ldots,K\}} \alpha_k + \beta'_k \boldsymbol{x}_i$$

K dominant hyperplanes

$$y_i \sim N(\hat{f}(\boldsymbol{x}_i), \sigma_k^2)$$

Heteroskedasticity allowed

- We refine on Hannah and Dunson's hyperplane-specific heteroskedasticity, by considering a hierarchical model on the variances, as presented in Gelman (2006)

$$y_{ij} \sim N(\hat{f}(\boldsymbol{x}_i) + (\xi\eta_j), \sigma_y^2)$$
$$\eta_j \sim N(0, \sigma_{\eta_j}^2)$$
$$\sigma_{\eta_j}^2 \sim IG(.5,2)$$
$$\sigma_y^2 \sim IG(a,b)$$

Group + Global variances

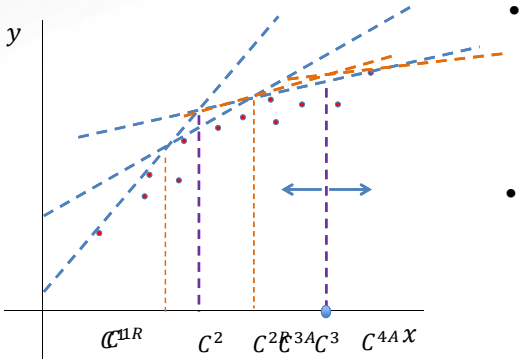
Hierarchical model on variances

13

MBCR with inefficiency

MBCR estimation procedure

- Split/combine a basis region to create hyperplane addition/removal proposals.



- For split movements, need to define number of inner knots and split up directions given a knot.
- For removals, separate basis region of removed hyperplane and assign observations to remaining hyperplanes.

MBCR with inefficiency

MBCR estimation procedure

- Determine hyperplane addition/removal/relocation according to prior for K and constant *c*:

$$b_k = c \cdot min\left\{1, \frac{p(k+1)}{p(k)}\right\} \quad d_k = c \cdot min\left\{1, \frac{p(k-1)}{p(k)}\right\} \quad r_k = 1 - b_k - d_k$$

- Select best choice in terms of likelihood chosen as candidate for the selected type of move.
- Metropolis-Hastings type probability to determine if block update is to be accepted

MBCR with inefficiency

Incorporating inefficiency

- The most natural way is to estimate u_i's and their hyperparameters *outside* of MBCR and treat MBCR as a Metropolis-within-Gibbs step to obtain an estimate of ***f(x)***.

Draw initial values for u_i's and their hyperparameters. (For Gamma distributions, let them be *P* and θ.)

↓

Draw values for β and σ^2 from conditional posteriors
Draw posterior hyperparameters for u_i's.
Draw posterior u_i values.
for $t_{Total} = t_{Burn-in} + t_{Stationary}$ iterations

↓

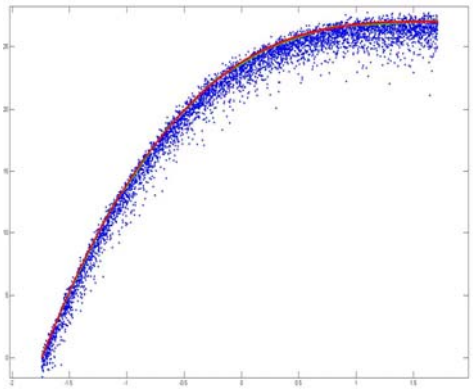
Ignore draws for first $t_{Burn-in}$ iterations. Take remaining draws for as a valid sample from the *joint* posterior distribution of all parameters.

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

MBCR with inefficiency

Incorporating inefficiency

- ✓ Good fit against the true frontier
- ✓ Good estimates for $\beta, \sigma \; and \; \theta$.
- ✓ Statistically consistent residual noise distribution.
- ✓ Maintains MBCR ability to estimate thousands of observations.



MBCR fit (green) with n=5000

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

SFA with MBCR

Some results

Using MBCR with no inefficiency on Linear function					
Iterations	n	meanudraw	std(meanudraw)	meanineffdraw	Bias
400, 100 burnin	500	-	-	-	-0.017
400, 100 burnin	1000	-	-	-	-0.008
Using MBCR with inefficiency on Linear function					
Iterations	n	meanudraw	std(meanudraw)	meanineffdraw	Bias
400, 100 burnin	500	0.7561	0.0263	0.7388	0.0035
400, 100 burnin	1000	0.751	0.0145	0.7268	0.0298
Using MBCR with inefficiency on cubic function (Example 1)					
Iterations	n	meanudraw	std(meanudraw)	meanineffdraw	Bias
400, 100 burnin	1000	0.6865	0.027	0.7157	-0.0187
400, 100 burnin	2000	0.6578	0.0173	0.6979	-0.0462
400, 100 burnin	3000	0.6393	0.017	0.6789	-0.0321
400, 100 burnin	3000	0.6738	0.016	0.6884	-0.0229
400, 100 burnin	5000	0.6716	0.0133	0.6983	-0.0342

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

SFA with MBCR

Some results, including runtimes

Using MBCR with inefficiency on cubic function (Example 1)								
Iterations	n	meanudraw	std(meanudraw)	meanineffdraw	Bias	%Bias	MSE frontier	Total time (minutes)
400, 100 burnin	250	0.6533	0.05	0.665	-0.0178	18.00%	0.0425	48.19
400, 100 burnin	500	0.6919	0.0351	0.7214	-0.0353	10.24%	0.01	51.74
400, 100 burnin	1000	0.7029	0.026	0.7058	-0.0124	1.68%	0.0395	47.83
400, 100 burnin	2000	0.6738	0.0188	0.6791	-0.0112	1.28%	0.0051	54.40
400, 100 burnin	3000	0.6612	0.0163	0.6817	-0.0197	1.67%	0.0058	59.56
400, 100 burnin	4000	0.6955	0.0139	0.6936	0.001	1.39%	0.0036	63.02
400, 100 burnin	5000	0.648	0.0131	0.6768	-0.0335	0.44%	0.0036	65.5

Using MBCR with inefficiency on quadratic, bivariate function (Example 2)								
Iterations	n	meanudraw	std(meanudraw)	meanineffdraw	Bias	%Bias	MSE frontier	Total time (minutes)
400, 100 burnin	250	0.6559	0.0491	0.6766	-0.0466	4.06%	0.0234	NA
400, 100 burnin	500	0.6832	0.0461	0.7632	-0.0965	8.35%	0.0179	48.59
400, 100 burnin	1000	0.6798	0.0244	0.6801	0.0135	1.18%	0.0049	47.94
400, 100 burnin	2000	0.6784	0.0191	0.6926	-0.0321	2.75%	0.0059	76.31

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Agenda

Background

- Objective
- Introduction to Bayesian Regression
- Gibbs sampling to estimate Bayesian Linear Regressions

Estimating a stochastic frontier with MBCR

- SFA with MBCR
- Implemented Algorithm

Status and further work

- Results, benefits and limitations
- Improvements and additions

Current limitations

- MBCR has larger estimation errors for samples under 500 observations.
- Have not yet explored more general Erlang cases than Exponential for the inefficiency distribution.

Further steps

- Currently working on incorporating greater prior information to the model by means of the *smoothing* hierarchical prior on the σ_k^2's, presented here. Will potentially help with small sample sizes.
- Stopping rules need to be defined for both the Metropolis-within-Gibbs MBCR and the overall algorithm.

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Questions?

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Possible Questions

Why are Bayesian methods easier to solve?

Because there is nothing to optimize. The posterior distributions of the coefficients are known from the beginning and the complexity of the method lies only in the computation of the posterior hyperparameters and the ability to draw efficiently from the posterior distributions.

"If we can only solve for 1,000 observations with CNLS, why can MBCR handle 5000?“

The key idea behind MBCR is that we fit minimums of hyperplanes to a concave function. The set of piecewise linear functions formed by minimums of hyperplanes automatically complies with concavity so no further constraints are needed. Also, as unless the function to be fitted is linear, we never need to take matrix inverses of the full dataset, only of subsets of the size defined by each basis region.