A non-random sample from *Experiential and social learning in firms: The case of hydraulic fracturing in the Bakken Shale* (Thomas Covert, 2013)

Liam F. Clegg

February 26, 2014

Background

- Oil wells "fracked" using a mix of sand and water
- Sand and water are costly
- Oil productivity of well varies with quantities of inputs used
- Firms must learn the optimal mix through their own, and other firms, experience.
- Econometric mission: estimate optimal mix of fracking inputs at each point in time, conditional on previous data, and compare to actual firm behavior

Well characteristics

- H_i horizontal length
- S_i sand used in fracking
- W_i water used in fracking
- lat_i, lon_i latitude and longitude of well bore

Time series of oil production

$$Y_{it} = Q_i t^{\beta} D_{it}^{\delta} \exp(\nu_{it})$$

where

- Y_{it} = production of well i in the t-th month of its life
- Q_i = baseline level of production for well i
- $\beta < 0$ parameter governing productivity decline
- D_{it} = number of days well i in operation in t-th month
- ν_{it} = mean-zero production shock

so

$$\log Y_{it} = \log Q_i + \beta \log t + \delta \log D_{it} + \nu_{it}$$

Assume

$$\mathbb{E}\left[\nu_{it} | t, H_i, D_{it}, S_i, W_i, lat_i, lon_i\right] = 0$$

Cross section of oil production

$$\log Q_i = \alpha + \eta \log H_i + f(S_i, W_i, lat_i, lon_i) + \epsilon_i$$

- Assume $\mathbb{E}[\epsilon_i | t, H_i, D_{it}, S_i, W_i, lat_i, lon_i] = 0$
- $Z_i = (S_i, W_i, lat_i, lon_i)$
- $f(S_i, W_i, lat_i, lon_i) = f(Z_i)$ captures relationship b/w baseline production and location and fracking choices
- Estimate $f(Z_i)$ non-parametrically, using Gaussian process regression (GPR)

Gaussian process regression

- Gaussian process G: distribution over continuous real functions, defined by mean function $m(Z)$ and covariance function $k(Z, Z')$, satisfying:

$$m(Z) = \int f(Z) dG(f)$$

$$k(Z, Z') = \int (f(Z) - m(Z))(f(Z') - m(Z')) dG(f)$$

- Distribution of $f(Z_1) \dots f(Z_N)$ is $\mathcal{MVN}(\mu, \Sigma)$, where

$$\mu = (m(Z_1) \dots m(Z_N))^T$$
$$\Sigma_{i,j} = k(Z_i, Z_j)$$

- For any m, k, can compute likelihood that $(g_i, Z_i)_{i=1}^N$ generated by $g = f(Z)$

GPR Likelihood

Assume $m(Z) = 0$, and $k(Z, Z')$ is MVN kernel:

$$k(Z_i, Z_j|\gamma) = \exp(2\gamma_0)\exp\left(-\frac{1}{2}\sum_{d\in\{S,W,lat,lon\}}\frac{(Z_{i,d}-Z_{j,d})^2}{\exp(2\gamma_d)}\right)$$

- γ_0 is variance of $f(Z)$
- $\gamma_S, \gamma_W, \gamma_{lat}, \gamma_{lon}$ are log-bandwidths

Log-likelihood of $(g_i, Z_i)_{i=1}^N$ is

$$\log\mathcal{L}(\gamma) = -\frac{1}{2}g^T K(\gamma)^{-1} g - \log|K(\gamma)| - \frac{N}{2}\log(2\pi)$$

where $g = (g_i \ldots g_N)^T$ and K is covariance matrix, so $K(\gamma)_{i,j} = k(Z_i, Z_j|\gamma)$

Likelihood maximization

$$\begin{aligned}\log Y_{it} &= \alpha + \beta \log t + \delta \log D_{it} + \eta \log H_i + f(Z_i) + \epsilon_i + \nu_{it} \\ &= X_{it}\theta + f(Z_i) + \epsilon_i + \nu_{it}\end{aligned}$$

- Assume $\nu_{it} \sim \mathcal{N}(0, \sigma_\nu^2)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Def. $\phi = (\sigma_\epsilon, \sigma_\nu)$
- Two-step likelihood maximization:
  1. Treat $f(Z_i)$ as observed: compute

$$\begin{aligned}g_i &= \frac{1}{T_i}\sum_{t=1}^{T_i}(\log Y_{it} - X_{it}\theta) \\ &= f(Z_i) + \epsilon_i + \frac{1}{T_i}\sum_{t=1}^{T_i}\nu_{it}\end{aligned}$$

and thence

$$\mathcal{L}(\mathbf{Y}_i, \mathbf{X}_i | g_i, \theta, \phi) = \dots$$

Likelihood maximization, cont.

$$\mathcal{L}(\mathbf{Y}_i, \mathbf{X}_i | g_i, \theta, \phi) = \dots$$

2. By GPR, vector $\mathbf{g} = (g_1 \dots g_N)^T \sim \mathcal{MVN}(0, K(\mathbf{Z}|\gamma))$
   So

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{Z} | \theta, \phi, \gamma) = \int \psi(\mathbf{g} | \mathbf{0}, K(\mathbf{Z}|\gamma)) \prod_{i=1}^{N} \mathcal{L}(\mathbf{Y}_i, \mathbf{X}_i | g_i, \theta, \phi) d\mathbf{g}$$
$$= \dots$$

Results & Discussion

- Results: see handout
- What is the benefit of this method?