Factor Forecasting for Agricultural Production Processes

Wenjun Zhu
Assistant Professor
Nanyang Business School, Nanyang Technological University
wjzhu@ntu.edu.sg

Joint work with Hong Li, Ken Seng Tan, and Lysa Porth

APA 2017, Dec 1 - Dec 2, 2017, Waterloo, Canada
Outline

Motivation

Econometric Framework

Empirical Forecasting Results

Applying to Index Insurance

Conclusion & Future Work
“In the past I have talked about ‘the hikes in the spikes’; now we have to beware of ‘the bumps in the slumps’!”

Agricultural Outlook 2017-2026
Paris, France, July 10, 2017

Angel Gurría, OECD Secretary-General
Motivations

• Agricultural markets are inherently volatile, but increasingly important
  ▶ 70 percent increase in food productivity needed to feed the world’s growing population by 2050 (FAO, 2009)

• Crop yield forecasting is central to agricultural risk management at all levels
  ▶ planting decision-making; trade; policies; food security ...
Motivations

- Agricultural markets are inherently volatile, but increasingly important
  - 70 percent increase in food productivity needed to feed the world’s growing population by 2050 (FAO, 2009)
- Crop yield forecasting is central to agricultural risk management at all levels
  - planting decision-making; trade; policies; food security ...

Objectives

- Improve the agricultural yield prediction accuracy by proposing a dynamic factor model
- Achieve improved risk management approach by designing a new weather index insurance
Predicting Agricultural Yield

- Predicting yield is a very challenging task that requires more research and development efforts
  - identify the key weather variables
  - data availability and credibility
  - technological improvements
  - crop insurance program changes
Statistical Models: Advantages

- Limited reliance on experimental field data, compared to process-based model
- Straightforward and transparent
  - Clear relationship between crop yields and explanatory variables (such as weather)
- Increasing weather forecast skill over the past 40 years
  - super-computing facilities
  - satellites
Motivation

Statistical Models: Challenges

- How to determine variables included into the model
- Substantial model risks
  - too many regressors — overfitting
  - too few regressors — low predictive power
- Limited historical data: a few decades of observations
- Forecasting results are rather sensitive to the choice of regressors
- Estimation is not feasible when the dimension of regressors exceeds the number of observations.
Model I: time-series model

- $y_{i,t}$ — the yield in county $i$ at year $t$, $i = 1, \ldots, N$, and $t = 1, \ldots, T$

- $X_{i,t}$ — a $(J \times 1)$ column vector containing the regressors in county $i$ and year $t$

- Regression model for each county $i$:
  \[
  \log(y_{i,t}) = a_i + b_i t + \gamma'_i X_{i,t} + \epsilon_{i,t}.
  \]

- Can also be estimated on a state level:
  \[
  \log(y_t) = a + b t + \gamma' X_t + \epsilon_t.
  \]
Model II: cross-section model

- $y_{i,avg}$ — the average of the crop yield of county $i$ over time.
- $X_i$ — a $(K \times 1)$ column vector containing the regressors for county $i$.
- The cross-section model:

$$\log(y_{i,avg}) = a + \gamma' X_i + \varepsilon_i.$$
Dynamic Factor Approach

1. Estimate a set of latent factors through principal component analysis (PCA)

2. Follow a dynamic factor procedure to select factors that are important for yield forecasting
Dynamic Factor Approach

1. Estimate a set of latent factors through principal component analysis (PCA)

2. Follow a dynamic factor procedure to select factors that are important for yield forecasting
   - Dynamic factor approach has been successfully applied for forecasting a variety of processes
     - Macroeconomic variables: inflation (Stock and Watson 2002) and bond risk premia (Ludvigson and Ng 2009)
     - Mortality modeling (French and O’Hare 2013)
Determine Latent Factors $\hat{f}_t$

- Assume that the regressors follow a linear factor structure:

\[ x_{j,t} = \chi_j' f_t + \omega_{j,t}, \quad \forall j, \]

where $f_t$ is a $r \times 1$ vector of blue latent factors, with $r << J$.

- $\hat{f}_t$ is estimated by PCA

  (a) $\hat{f}_t$ is a linear combination of $X_t$, i.e., $\hat{f}_t = \hat{\Lambda} X_t$ for all $t$;
  
  (b) $\hat{\Lambda}$ minimizes the sum of squared residuals $\sum_{t=1}^{T} (X_t - \Lambda f_t)^2$.

- The number of PC’s in $\hat{f}_t$, $r$, is determined by the panel information criteria (IC) by Bai and Ng (2002)
Determine Latent Factors \( \hat{f}_t \)

- Assume that the regressors follow a linear factor structure:

\[
x_{j,t} = \lambda_j' f_t + \omega_{j,t}, \quad \forall j,
\]

where \( f_t \) is a \( r \times 1 \) vector of blue latent factors, with \( r << J \).

- \( \hat{f}_t \) is estimated by PCA
  
  (a) \( \hat{f}_t \) is a linear combination of \( X_t \), i.e., \( \hat{f}_t = \hat{\Lambda} X_t \) for all \( t \);
  
  (b) \( \hat{\Lambda} \) minimizes the sum of squared residuals \( \sum_{t=1}^{T} (X_t - \Lambda f_t)^2 \).

- The number of PC’s in \( \hat{f}_t \), \( r \), is determined by the panel information criteria (IC) by Bai and Ng (2002)
Determine Optimal $\hat{F}_{k,t}$

For $k = 1, \ldots, 2^r$:

1. Construct candidate $\hat{F}_{k,t}$, as a subset of $\hat{f}_t$;
2. Estimate the following regression:
   \[
   \log(y_t) = a + bt + \theta' \hat{F}_{k,t} + \epsilon_t, \tag{1}
   \]
3. Pick the optimal factor $\hat{F}_{k^*,t}$ that gives the minimal BIC.
Data

- Corn, soybean, and winter wheat in Illinois
  - County-level & State-level, 1981-2016
  - National Agricultural Statistics Service (NASS) crops survey data
Data

- Corn, soybean, and winter wheat in Illinois
  - County-level & State-level, 1981-2016
  - National Agricultural Statistics Service (NASS) crops survey data
- In total, take more than 80% of cropland coverage
Meteorological & Soil Information

- Monthly average temperature and accumulative precipitation
  - PRISM Climate Group
- Soil information
  - USDA Natural Resources Conservation Service
- Define growing seasons according to USDA (1997)

<table>
<thead>
<tr>
<th>Crop</th>
<th>Growing Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>May - August</td>
</tr>
<tr>
<td>Soybeans</td>
<td>May - August</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>October - June (next year)</td>
</tr>
</tbody>
</table>

- Final design matrix with 104 explanatory variables
Benchmark: Lobell and Burke (2010)

- Time-series specification:
  \[ X_{i,t} = (T_{i,t}, P_{i,t})', \]  
  \[ (2) \]

- Cross-sectional specification:
  \[ X_{i,avg} = (T_{i,avg}, P_{i,avg}, T_{i,avg}^2, P_{i,avg}^2)', \]  
  \[ (3) \]
Model Fitness

- Select the optimal factors $\hat{F}_t$
  - Full Factor: selecting $m$ applying the dynamic factor procedure
  - Constrained Factor: restricted $m$ to be 2 in time-series models and to be 4 in the cross-section model
Model Fitness

- Select the optimal factors $\hat{F}_t$
  - **Full Factor**: selecting $m$ applying the dynamic factor procedure
  - **Constrained Factor**: restricted $m$ to be 2 in time-series models and to be 4 in the cross-section model
Model Fitness

- Select the optimal factors $\hat{F}_t$
  - Full Factor: selecting $m$ applying the dynamic factor procedure
  - Constrained Factor: restricted $m$ to be 2 in time-series models and to be 4 in the cross-section model

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>51%</td>
<td>63%</td>
<td>85%</td>
<td>8.83</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Soybean</td>
<td>56%</td>
<td>65%</td>
<td>76%</td>
<td>4.72</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>32%</td>
<td>46%</td>
<td>53%</td>
<td>3.41</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>
Model Fitness

- Select the optimal factors $\hat{F}_t$
  - Full Factor: selecting $m$ applying the dynamic factor procedure
  - Constrained Factor: restricted $m$ to be 2 in time-series models and to be 4 in the cross-section model

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>51%</td>
<td>63%</td>
<td>85%</td>
<td>8.83</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Soybean</td>
<td>56%</td>
<td>65%</td>
<td>76%</td>
<td>4.72</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>32%</td>
<td>46%</td>
<td>53%</td>
<td>3.41</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>
## Model Fitness

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>68%</td>
<td>79%</td>
<td>94%</td>
<td>12</td>
</tr>
<tr>
<td>Soybean</td>
<td>71%</td>
<td>79%</td>
<td>93%</td>
<td>13</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>61%</td>
<td>69%</td>
<td>75%</td>
<td>6</td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>58%</td>
<td>78%</td>
<td>86%</td>
<td>10</td>
</tr>
<tr>
<td>Soybean</td>
<td>67%</td>
<td>81%</td>
<td>89%</td>
<td>8</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>33%</td>
<td>85%</td>
<td>99%</td>
<td>10</td>
</tr>
</tbody>
</table>
## Cross-Validation

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.025</td>
<td>0.022</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.023</td>
<td>0.021</td>
<td>0.009</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.031</td>
<td>0.028</td>
<td>0.023</td>
</tr>
<tr>
<td><strong>County level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.044</td>
<td>0.038</td>
<td>0.019</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.035</td>
<td>0.030</td>
<td>0.022</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.045</td>
<td>0.039</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.016</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.019</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.014</td>
<td>0.007</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Cross-Validation

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.025</td>
<td>0.022</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.023</td>
<td>0.021</td>
<td>0.009</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.031</td>
<td>0.028</td>
<td>0.023</td>
</tr>
<tr>
<td><strong>County level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.044</td>
<td>0.038</td>
<td>0.019</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.035</td>
<td>0.030</td>
<td>0.022</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.045</td>
<td>0.039</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.016</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.019</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.014</td>
<td>0.007</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
## Cross-Validation

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Constrained Factor</th>
<th>Full Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.025</td>
<td>0.022</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.023</td>
<td>0.021</td>
<td>0.009</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.031</td>
<td>0.028</td>
<td>0.023</td>
</tr>
<tr>
<td><strong>County level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.044</td>
<td>0.038</td>
<td>0.019</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.035</td>
<td>0.030</td>
<td>0.022</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.045</td>
<td>0.039</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>0.016</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.019</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>Winter Wheat</td>
<td>0.014</td>
<td>0.007</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Index Insurance Design

- Indemnity-based Insurance
- On-site Assessment
- Adverse Selection
- Moral Hazard

- Index Insurance
- Basis risk
Index Insurance Design

Indemnity-based Insurance

- On-site Assessment
- Adverse Selection
- Moral Hazard

Index Insurance

- Basis risk

Improved Yield Forecasting Model
Basis Risk: Frequency and Severity

- **Type I Error:**
  - True $H_0$ is rejected
  - the insurer fails to pay the producers

- **Type II Error:**
  - False $H_0$ is accepted
  - the insurer incorrectly pays the producers
Index Insurance Design

- The optimal model selected from dynamic factor procedure

\[
\hat{y}_t = \mathcal{M}^*(\hat{\Lambda}_*, \hat{F}_t)
\]

- Index insurance payoff

\[
P(\hat{\Lambda}_*, \hat{F}_t) = \text{Area} \times \text{Price} \times \max(K - \mathcal{M}^*(\hat{\Lambda}_*, \hat{F}_t), 0)
\]

- Backtesting with MPCI, 2001-2016
### Basis Risk Analysis

<table>
<thead>
<tr>
<th>Crop</th>
<th>Corn</th>
<th>Soybean</th>
<th>Winter Wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Summary of Actual Losses Based on MPCI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss Prob.</td>
<td>34.14%</td>
<td>33.60%</td>
<td>35.00%</td>
</tr>
<tr>
<td>Loss Mean</td>
<td>7.41</td>
<td>1.45</td>
<td>2.23</td>
</tr>
<tr>
<td>Loss Std.</td>
<td>16.96</td>
<td>2.95</td>
<td>4.30</td>
</tr>
</tbody>
</table>

### Basis Risk Analysis

<table>
<thead>
<tr>
<th></th>
<th>Factor</th>
<th>Benchmark</th>
<th>Factor</th>
<th>Benchmark</th>
<th>Factor</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>10.30%</td>
<td>28.52%</td>
<td>18.84%</td>
<td>48.95%</td>
<td>23.72%</td>
<td>31.22%</td>
</tr>
<tr>
<td>Type II Error</td>
<td>6.50%</td>
<td>20.91%</td>
<td>11.66%</td>
<td>14.30%</td>
<td>17.59%</td>
<td>41.40%</td>
</tr>
<tr>
<td>Mismatch Prob.</td>
<td>89.70%</td>
<td>71.48%</td>
<td>81.16%</td>
<td>51.05%</td>
<td>76.28%</td>
<td>68.78%</td>
</tr>
<tr>
<td>Mismatch Mean</td>
<td>-0.12</td>
<td>2.12</td>
<td>0.04</td>
<td>0.77</td>
<td>0.33</td>
<td>0.88</td>
</tr>
<tr>
<td>Mismatch Std.</td>
<td>4.46</td>
<td>10.02</td>
<td>1.17</td>
<td>2.44</td>
<td>2.29</td>
<td>3.79</td>
</tr>
</tbody>
</table>
Basis Risk Analysis

Dynamic Factor

Benchmark
Conclusion & Future Work

- We propose a dynamic factor approach to construct a robust and accurate yield forecasting model
  - allow high dimensional matrix of regressors
  - efficient dimension reduction and variable selection
- A new index insurance is designed and is shown to be able to reduce basis risk of both severity and frequency
- Include remote sensing data into the analysis
References


Thank you for Attentions!