1 Introduction

In this note we will examine how aggregate estimators might lead to biases in the estimates of the persistence of the underlying data generating process in the face of cross-sectoral heterogeneity in the persistence of the individual price series. The note expands upon the two-sector example in the paper and allows for correlation of the errors across the cross-sectional units. We show that our results are robust. We will first look at a general case, then a special two-sector case, and then a counter-example provided by Charles Engel.

1.1 Aggregation Bias: The general case with n sectors and cross correlations

In the general n-sector case, we have, if we index sectors by i:

$$\begin{aligned} x_{it} &= \theta_i x_{it-1} + e_{it} \\ E(e_{it}^2) &= \sigma_i^2 \\ E(e_{it}e_{jt}) &= \sigma_{ij} \end{aligned}$$

where x_{it} denotes the (log) of the relative price of the i'th cross-sectional unit. This assumes an AR(1) structure but the results generalize to any AR processes. Relative to the example we provide in the main paper this allows for covariances between the innovations and for heterogeneity in the variance of the innovations.

We will assume, as is the realistic case for our purposes, that prices are positively autocorrelated. Furthermore, we order the data such that:

$$0 < \theta_1 \leq \theta_2 \leq ... \leq \theta_n$$

It follows that:

$$\begin{aligned} \sigma_{x_i}^2 &= \frac{\sigma_i^2}{1-\theta_i^2} \\ \sigma_{x_i,x_j} &= \frac{\sigma_{ij}}{1-\theta_i\theta_j} \end{aligned}$$

Let us now assume that θ_i is drawn from a distribution with mean $\overline{\theta}$ and we wish to estimate this mean persistence.

$$\overline{\theta} = \sum_{i=1}^{n} \frac{\theta_i}{n}$$

The issue is how the estimate of the "mean" persistence would be affected by cross-sectional aggregation. The cross-sectional aggregate is given by:

$$x_t = \frac{1}{n}\sum_{i=1}^{n} x_{it}$$

We will first derive the (asymptotics of the) least squares estimate of the persistence of the aggregated process. The least squares estimate of the persistence will be given by $\theta^a = \sigma_{x,x-1}/\sigma_x^2$. Straightforward algebra gives us that:

$$\begin{aligned}
\sigma_x^2 &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)\right) \\
\sigma_{x,x-1} &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\theta_i\sigma_{x_i}^2 + \sum_{j>i}^{n}(\theta_i+\theta_j)\,\sigma_{x_i,x_j}\right)\right)
\end{aligned}$$

Hence, it follows that:

$$\begin{aligned}
\theta^a &= \frac{\sum_{i=1}^{n}\left(\theta_i\sigma_{x_i}^2 + \sum_{j>i}^{n}(\theta_i+\theta_j)\,\sigma_{x_i,x_j}\right)}{\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)} \\
&= \overline{\theta} - \overline{\theta} + \frac{\sum_{i=1}^{n}\left(\theta_i\sigma_{x_i}^2 + \sum_{j>i}^{n}(\theta_i+\theta_j)\,\sigma_{x_i,x_j}\right)}{\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)} \\
&= \overline{\theta} + \frac{\sum_{i=1}^{n}\left(\left(\theta_i-\overline{\theta}\right)\sigma_{x_i}^2 + \sum_{j>i}^{n}\left[\left(\theta_i-\overline{\theta}\right)\sigma_{x_i,x_j} + \left(\theta_j-\overline{\theta}\right)\sigma_{x_i,x_j}\right]\right)}{\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)} \\
&= \overline{\theta} + \frac{\sum_{i=1}^{n}\left(\frac{\theta_i-\overline{\theta}}{1-\theta_i^2}\sigma_{\ i}^2 + \sum_{j>i}^{n}\left(\frac{\theta_i-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{\ ij} + \frac{\theta_j-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{\ ij}\right)\right)}{\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)}
\end{aligned}$$

Therefore:

$$\theta^a = \overline{\theta} + \frac{\Delta}{\sum_{i=1}^{n}\left(\sigma_{x_i}^2 + 2\sum_{j>i}^{n}\sigma_{x_i,x_j}\right)} \tag{1}$$

$$\Delta = \sum_{i=1}^{n}\left(\frac{\theta_i-\overline{\theta}}{1-\theta_i^2}\sigma_{\ i}^2 + \sum_{j>i}^{n}\left(\frac{\theta_i-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{\ ij} + \frac{\theta_j-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{\ ij}\right)\right) \tag{2}$$

The sign of the bias is the sign of Δ. Determining this sign in general depends on the covariances of the innovation terms, the pattern on heterogeneity in the persistence and the variance of the innovations. However, as we will now show, although counterexamples can be constructed, the realistic cases give rise to a positive aggregation bias.

Let us call i_0 the index such that for all $i < i_0$, $\theta_i < \overline{\theta}$ and $\theta_{i_0} \geq \overline{\theta}$. This index exists since $\overline{\theta}$ is a convex combination of the $\{\theta_i\}_{i=1}^n$. We can now show that:

$$\begin{aligned}\Delta &\geq 0 \Leftrightarrow \\ &\sum_{i=i_0}^{n}\left(\frac{\theta_i-\overline{\theta}}{1-\theta_i^2}\sigma_i^2+\sum_{j>i}\left[\frac{\theta_i-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{ij}+\frac{\theta_j-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{ij}\right]\right) \\ \geq &\sum_{i=1}^{i_0-1}\left(\frac{\overline{\theta}-\theta_i}{1-\theta_i^2}\sigma_i^2+\sum_{i_0>j>i}\left[\frac{\overline{\theta}-\theta_i}{1-\theta_i\theta_j}\sigma_{ij}+\frac{\overline{\theta}-\theta_j}{1-\theta_i\theta_j}\sigma_{ij}\right]\right)\end{aligned}$$

We can spell out *sufficient* conditions for the bias to be positive:
If $\sigma_i^2 \backsimeq \sigma^2$ and $\sigma_{ij} \backsimeq \chi \geq 0$ then the bias is positive.

Proof:
Since $\theta_1 \leq \theta_2 \leq ... \leq \theta_n$, we have $\frac{1}{1-\theta_i^2} \geq \frac{1}{1-\theta_j^2} > 0$ whenever $i > j$.
Hence $\sum_{i=i_0}^n \frac{\theta_i-\overline{\theta}}{1-\theta_i^2} \geq \sum_{i=1}^{i_0-1} \frac{\overline{\theta}-\theta_i}{1-\theta_i^2}$ since $\sum_{i=1}^n \left(\theta_i - \overline{\theta}\right) = 0$ by definition of $\overline{\theta}$.
Similarly $\sum_{\substack{i<j \\ i\geq i_0}} \left[\frac{\theta_i-\overline{\theta}}{1-\theta_i\theta_j} + \frac{\theta_j-\overline{\theta}}{1-\theta_i\theta_j}\right] \geq \sum_{\substack{i<j \\ j<i_0}} \left[\frac{\overline{\theta}-\theta_i}{1-\theta_i\theta_j} + \frac{\overline{\theta}-\theta_j}{1-\theta_i\theta_j}\right]$ for the same reason. Therefore $\Delta \geq 0$.

These conditions on the variance covariance matrix of the sectoral errors are close to what we observe in the actual sectoral price data. This case is therefore the empirically relevant case and the aggregation bias is positive.

Note that for the bias to be negative, we would need strong and systematic asymmetries in the price data.
Necessary (and NOT sufficient) conditions would be: a) either that $\frac{\overline{\theta}-\theta_i}{1-\theta_i\theta_j}\sigma_{ij} > \frac{\overline{\theta}-\theta_k}{1-\theta_k\theta_l}\sigma_{lk}$ at least for some $i, j \in \{1, .., i_0 - 1\}$ and for some $k, l \in \{i_0, .., n\}$ and $\sigma_{ij} > 0$ and $\sigma_{lk} > 0$; b) or that $\frac{\overline{\theta}-\theta_i}{1-\theta_i\theta_j}\sigma_{ij} < \frac{\overline{\theta}-\theta_k}{1-\theta_k\theta_l}\sigma_{lk}$ at least for some $i, j \in \{1, .., i_0 - 1\}$ and for some $k, l \in \{i_0, .., n\}$ and $\sigma_{ij} < 0$.

1.2 Special Cases

1.2.1 No cross-correlation of the innovation errors

Assuming that the errors are uncorrelated across sectors gives us immediately that:

$$
\begin{aligned}
\theta^a &= \bar{\theta} + \frac{\sum_{i=1}^n \left(\frac{\theta_i - \bar{\theta}}{1-\theta_i^2} \sigma^2_{\ i} + \sum_{j>i}^n \left(\frac{\theta_i - \bar{\theta}}{1-\theta_i\theta_j} \sigma_{\ ij} + \frac{\theta_j - \bar{\theta}}{1-\theta_i\theta_j} \sigma_{\ ij} \right) \right)}{\sum_{i=1}^n \left(\sigma^2_{x_i} + 2 \sum_{j>i}^n \sigma_{x_i, x_j} \right)} \\
&= \bar{\theta} + \frac{\sum_{i=1}^n \left(\frac{\theta_i - \bar{\theta}}{1-\theta_i^2} \sigma^2_{\ i} \right)}{\sum_{i=1}^n \left(\sigma^2_{x_i} \right)} \geq \bar{\theta}
\end{aligned}
$$

where strict equality holds whenever there is cross sectoral heterogeneity. This follows simply because, as explained above, the more persistent components receive a higher weight in the averaged estimaor.

1.2.2 Two sectors

When there are just two sectors we can show immediately that the sign of the bias is independent of the correlation of the innovations because:

$$
\begin{aligned}
\theta^a &= \bar{\theta} + \frac{\sum_{i=1}^n \left(\frac{\theta_i - \bar{\theta}}{1-\theta_i^2} \sigma^2_{\ i} + \sum_{j>i}^n \left(\frac{\theta_i - \bar{\theta}}{1-\theta_i\theta_j} \sigma_{\ ij} + \frac{\theta_j - \bar{\theta}}{1-\theta_i\theta_j} \sigma_{\ ij} \right) \right)}{\sum_{i=1}^n \left(\sigma^2_{x_i} + 2 \sum_{j>i}^n \sigma_{x_i, x_j} \right)} \\
&= \bar{\theta} + \frac{\left(\frac{\theta_1 - \bar{\theta}}{1-\theta_1^2} \sigma^2_1 + \frac{\theta_2 - \bar{\theta}}{1-\theta_2^2} \sigma^2_2 + \left(\frac{\theta_1 - \bar{\theta}}{1-\theta_1\theta_2} \sigma_{12} + \frac{\theta_2 - \bar{\theta}}{1-\theta_1\theta_2} \sigma_{12} \right) \right)}{\sum_{i=1}^n \left(\sigma^2_{x_i} + 2 \sum_{j>i}^n \sigma_{x_i, x_j} \right)} \\
&= \bar{\theta} + \frac{\frac{\theta_1 - \bar{\theta}}{1-\theta_1^2} \sigma^2_1 + \frac{\theta_2 - \bar{\theta}}{1-\theta_2^2} \sigma^2_2}{\sum_{i=1}^n \left(\sigma^2_{x_i} + 2 \sum_{j>i}^n \sigma_{x_i, x_j} \right)}
\end{aligned}
$$

Furthermore, the sign of the bias is positive as soon as the more persistent sectoral components are at least as volatile as the less persistent components (*sufficient* condition).

1.2.3 A counter-example

In what follows, we reproduce a counter-example suggested to us by Charles Engel. It goes as follows. There are three processes, x_1, x_2, and x_3. Assume $x_1 = -x_2$, and both are AR(1), which means:

$$
\begin{aligned}
x_{1t} &= \theta_1 x_{1t-1} + \varepsilon_t \\
x_{2t} &= \theta_1 x_{2t-1} - \varepsilon_t
\end{aligned}
$$

and $x_{10} = -x_{20}$. In this case, $x_1 + x_2 = 0$. x_3, in turn, is given by:

$$
x_{3t} = \theta_3 x_{3t-1} + \eta_t
$$

with arbitrarily low θ. In this case, it is possible that cross-sectional aggregation gives rise to a *negative* bias. With similar notations to the ones introduced earlier, we have

$$\begin{aligned} x_t &= \frac{1}{3}\left(x_{1t}+x_{2t}+x_{3t}\right) \\ &= \frac{1}{3}x_{3t} \end{aligned}$$

Thus, we find immediately that the OLS estimate of the aggregate persistence is given as:

$$\theta^a = \theta_3$$

Since $\overline{\theta} = (2\theta_1 + \theta_3)/3$ we then get that:

$$\begin{aligned} \theta^a - \overline{\theta} &= \\ \theta_3 - (2\theta_1+\theta_3)/3 &= 2\left(\frac{\theta_3-\theta_1}{3}\right) \end{aligned}$$

Therefore

$$\theta^a < \overline{\theta} \text{ as soon as } \theta_3 < \theta_1$$

If we go back to the general proof we note that this case corresponds to a very particular case with:

$n = 3,\ \theta_1 = \theta_2 > \theta_3;\ \sigma_{12} = -\sigma_1^2 < 0;\ \sigma_{23} = 0; \sigma_{13} = 0.$

$$\theta^a - \overline{\theta} = \frac{\Delta}{\sum_{i=1}^n \sigma_{x_i}^2 + 2\sum_{i<j}^n \sigma_{x_i,x_j}}$$

with $\Delta = \sum_{i=1}^n \frac{\theta_i-\overline{\theta}}{1-\theta_i^2}\sigma_i^2 + \sum_{i<j}^n \left(\frac{\theta_i-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{ij} + \frac{\theta_j-\overline{\theta}}{1-\theta_i\theta_j}\sigma_{ij}\right)$

$\Delta = 2\frac{\theta_1-\overline{\theta}}{1-\theta_1^2}\sigma_1^2 + \frac{\theta_3-\overline{\theta}}{1-\theta_3^2}\sigma_3^2 + 2\frac{\theta_1-\overline{\theta}}{1-\theta_1\theta_1}\sigma_{12} = \frac{\theta_3-\overline{\theta}}{1-\theta_3^2}\sigma_3^2$

So

$$\theta^a - \overline{\theta} = \frac{\frac{\theta_3-\overline{\theta}}{1-\theta_3^2}\sigma_3^2}{2\frac{\sigma_1^2}{1-\theta_i^2} + \frac{\sigma_3^2}{1-\theta_3^2} + 2\frac{\sigma_{12}}{1-\theta_1\theta_1}} = \frac{\frac{\theta_3-\overline{\theta}}{1-\theta_3^2}\sigma_3^2}{\frac{\sigma_3^2}{1-\theta_3^2}} = \theta_3 - \overline{\theta} = 2\left(\frac{\theta_3-\theta_1}{3}\right)$$

Which is exactly what we found before. This particular counterexample is of type b) of the necessary conditions described above. It is clearly a very extreme case and the conditions underlying it (66% of the prices constituting the index perfectly negatively correlated and more persistent than the remaining 33% of the prices) or any conditions remotely resembling these ones are far from being found in price data. Sectoral prices tend to be moderately positively correlated and whenever a subsample is moderately negatively correlated, its persistence is far from dominating the persistence of the other sectoral prices constituting the index. Finally, we notice that the covariance matrix of the panel estimator would be singular in this case. We have not encountered problems with singularity or near singularity in the estimation.

We conclude that these types of counter-example (and for that matter other counter-examples that could be built using necessary conditions of type a)), although interesting in theory, are not relevant in our application.