Vox Metrica

The Econometric Wisdom of Crowds

Matriculation Number: 170015890

University
of
St Andrews

School of Economics and Finance
Msc Economics

This dissertation is submitted in part requirement for the Degree of MSc in Economics at the University of St Andrews, Scotland. I declare that this Dissertation is **11,872** words in length. I have read and fully understand the University Policy on Good Academic Practice. I hereby declare that the attached piece of work is my own. It is written in my own words and I have acknowledged all of the sources that I have drawn upon.

August 2018

Supervisor: Professor Roderick McCrorie

Abstract

Can the principles that drive the psychological phenomenon of the wisdom of crowds be harnessed to create a serviceable new class of econometric time series predictor? This paper answers that question by conducting a multidisciplinary review of extant literature concerning the wisdom of crowds and integrating econometric theory with that stemming from this literature to define the vox predictor – the new class of econometric estimator based on the principles that manifest the wisdom of crowds. We develop a theoretical time series discussion that identifies a process with an unobserved autoregressive component that creates a time variant conditional distribution on the observed variables as one where the vox predictor could outperform the benchmark of ordinary least squares, owing to the former's flexibility in estimating the conditional expectation function. We then test the vox predictor with a simulation, finding our theoretical suspicions are confirmed and the vox predictor can out-predict ordinary least squares in certain circumstances. This establishes a very narrow context in which the vox predictor is serviceable, though the fact that the main results of the simulation were preempted by our theoretical exposition carries the promise that appropriately designed vox predictors could be advantageous econometric predictors in a wide range of contexts.

Table of Contents

Page

LIST OF TABLES

LIST OF FIGURES

FIGURE **Page**

CHAPTER 1

INTRODUCTION

Explorations of collective wisdom have always presupposed the potentially wise collective to be human, but the principles that enable human crowds to arrive at sagacious solutions need not be so anthropologically restricted. Building on the work of Hogarth (1978), chapter two unveils operative mathematical principles that allow the average of a crowd's guesses to be often more accurate than the average guess, and how a violation of these principles can lead to the collective folly found in psychological and political literature. The second section of chapter two establishes how the literature advocates the use of means of select subsets of crowds in the absence of ideal conditions, relating an important characteristic that leads to select crowd success back to those that produce whole crowd success through an original contribution to the mathematical exploration of the wisdom of crowds: the bracketing theorem. The development of the multidisciplinary literature in chapter two lays bare the form that an econometric predictor based on the wisdom of crowds – the vox predictor – must take, and the next chapter brings that form into an econometric context.

Ordinary Least Squares [OLS] pervades applied econometric literature [2] [19] and therefore presents itself as a standard of comparison for a new econometric predictor, though its basic nature need not result in an unrigorous test if the comparison is applied to the right time series process. Chapter three recapitulates the standard theory regarding OLS to identify those processes in which it has unparalleled prediction performance as measured by mean squared error and then provides a formal definition for the vox predictor within the time series context. A slightly altered time series process with unobserved variables is proposed as one where the vox predictor has potential to out-predict OLS. The conditional distribution of the proposed process on the observed variables is derived in the third section of chapter three with the use of the conditional normality theorem proved in the second section. In the limit, the proposed process becomes a stochastic and homoscedastic unchanging linear function of the observed variables; a domain where OLS is an ideal predictor. However, in finite time, the time variant nature of the conditional distribution entertains the possibility that the vox predictor could best OLS. This motivates the predominantly limited sample simulations conducted and displayed in chapter four.

The first section of chapter four explains the parameters and procedure of the simulation whose code is displayed in appendix A.8. The second section tabulates the main body of simulation outputs comparing the mean of the mean squared prediction errors of the vox predictor and OLS over one thousand trials of the same calibration for the simulation. These results are discussed in the third section and are used as inspiration for supplementary simulation outputs that enhance the experiment's conclusions. The central contention of the theoretical investigation in chapter three is confirmed: the vox predictor can outperform OLS in limited samples; this serves as a springboard into contemplating the potentialities of the vox predictor and the possible extensions to this paper using the given code and theorems, all of which is covered in the same section three of chapter four.

Chapter five concludes the paper and our contribution to both the wisdom of crowds literature and econometric literature. These two fields are enhanced separately through the brackting theorem, the conditional normality theorem, and our analysis of a time series process with an unobserved autoregressive component. However, the unification of these fields through the introduction of the vox predictor is this paper's truly unique contribution.

CHAPTER 2

RELATED LITERATURE AND CROWD THEORY

2.1 Whole Crowds

In 1906, Francis Galton collated 787 guesses of the weight of an ox in a county fair competition. He did this in an attempt to demonstrate the collective mediocrity of common people, but to his astonishment found that the median estimate of the fair punters was correct to within 0.8 percent of the true weight of the ox, showing that the average of the crowd's guesses was far superior to the average guess. Galton published these results in Nature [7] and called this phenomena of collective intelligence 'vox populi,' which has since been popularised as 'the wisdom of crowds' by James Surowiecki (2004). In the eponymous book, Surowiecki details how crowds can come to optimal solutions when faced with multiple types of problems, but time series econometrics, and therefore this dissertation, deals with the category of problem Surowiecki calls a 'cognition problem.' Cognition problems are concerned with the processing of information to arrive at a conclusion about an objective state of affairs. Answering a question on 'Who Wants To Be A Millionaire?' is a cognition problem, as is determining the present value of a company's future free cash flow. Galton's crowd of bettors was also solving a cognition problem, but their outstanding collective result was not guaranteed only by virtue of their being a crowd. There are four conditions Surowiecki gives that ensure crowd wisdom, and whose lack thereof can just as easily turn individually smart members of a crowd into an unintelligent collective.[1] They are:

1. Diversity (of opinion): members of a crowd must have distinct perspectives or pieces of private information, even if this is constituted by an eccentric interpretation of the known facts.

2. Decentralization: The power to make decisions cannot be concentrated in one or a few members of the crowd, members are relatively disassociated from each other, and members can specialize in and rely on local knowledge.

3. Independence: people's judgments must not be determined or heavily influenced by those of others in the crowd.

4. Aggregation: there must be a mechanism to turn a set of private judgments into a collective one, such as taking the mean or median of the estimates.

The importance of independence and diversity of opinion can be seen in the psychological phenomenon of groupthink[2] [11]. Groupthink occurs when a collective makes erroneous decisions because pressure to conform to the group's consensus prevents individuals from checking the group's biases. The fact that there is a group consensus means there is not a great degree of diversity, and social pressure prevents those who do have diverse opinions from submitting them, meaning their contributions are no longer independent from those of the other members of the group. This prevents members of a group from contributing their private information and the fact everyone agrees augments the group's confidence, compounding rather than canceling errors in judgment. Janis (1972) documents how groupthink led to disastrous foreign policy evaluations and decisions, a domain where authority ultimately lies in the hands of a few politicians and all those responsible for a policy share the same networks. The fact that decisions were made within these tight-knit networks that included one's superiors ensured the lack of diversity and independence inherent in groupthink, and this underscores how a lack of decentralization can in turn lead to an absence of independence and diversity.

Decentralisation does not guarantee independence and diversity, however, as demonstrated by information cascades, which can occur in very decentralised stock markets.

[1] Indeed, Surowiecki's coining of the 'wisdom of crowds' was itself based on Charles Mackay's (1841) 'Extraordinary Popular Delusions and the Madness of Crowds.'

[2] The coining of this term is based on 'doublethink' from George Orwell's '1984,' which referred to the simultaneous holding of contradictory beliefs in order to conform to the novel's totalitarian society.

Information cascades happen when enough people start acting or making judgments in a particular way that subsequent actors or judges ignore their private information and go along with the established majority, which only compounds the effect by increasing the majority. Milgram et al. (1969) demonstrated how we are influenced by crowds by placing participants in their experiment on a busy New York street and having them look up at the sky. When only one or two participants did so, passers by hardly noticed, but when the number of participants was large, passers by frequently stopped and stared along with the crowd, even though there was nothing to see. The fact that passers by were willing to trust that a group of strangers was acting wisely enough that it merited imitation, despite having no private information confirming that it was, is at the heart of what causes information cascades. Ironically it is people's belief in the wisdom of crowds that can undermine it. We can see this in stock market bubbles where enough influential investors have (what may be perfectly valid) private information indicating some companies are undervalued and invest accordingly. Subsequent investors that have (again what may be perfectly valid) private information indicating otherwise trust the prevailing consensus on the reasonable assumption that the previous investors' collective private information and action make it more probable that some stocks are undervalued despite one's own private information. In this way, information indicating stocks are overvalued does not influence the stock price, and stocks are bought because everyone believes they wouldn't have been bought by others were there no good reason.

The principles discussed here are put into a more quantitative context with the expression below from Hogarth[3] (1978). Participants of a crowd who are solving a cognition problem with a numerical answer will submit estimates that are equal to the true answer plus an error.

$$x_i = \mu + \epsilon_i \tag{2.1}$$

Where x_i is a participant i's estimate, μ is the true value being estimated, and ϵ_i is i's error. Independence can naturally be interpreted as the errors being independently distributed, whereas diversity can be interpreted as the errors being mean zero, since a diverse crowd ought not be systematically biased. So long as the errors also have finite variance and satisfy the Lindeberg condition [12], then the Lindeberg central limit theorem guarantees that

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \overset{a}{\sim} N\left(0, \frac{S_n}{n^2}\right) \tag{2.2}$$

[3] This expression has been slightly simplified.

where S_n is the sum of the variances σ_i^2 of the $i = 1,2,\ldots,n$ errors. Since the variances are finite, we can state that $S_n \leq n \cdot \left(\sup_{i\in\{1,\ldots,n\}} \sigma_i^2 \right)$ with $\sup_{i\in\{1,\ldots,n\}} \sigma_i^2 < \infty$. The variance of the mean of the errors is therefore less than or equal to $n^{-1} \cdot \left(\sup_{i\in\{1,\ldots,n\}} \sigma_i^2 \right)$ and converges to zero as crowd size tends to infinity, which implies that the average of the participants' estimates converges in probability to the true value μ as crowd size grows to infinity. Independence, diversity, and aggregation by way of the mean are, along with regularity conditions, sufficient to produce crowds that are perfectly accurate in the limit, relegating decentralisation to a supporting condition for independence and diversity.[4] If estimates proceed sequentially and become dependent if enough errors randomly go one way, we can see how information cascades develop, with a bunch of (even unbiased) estimates randomly overestimating μ and subsequent estimates then becoming dependent and compounding those random errors. Groupthink can be similarly viewed, except that estimates start out biased and dependence exists from the beginning, and herein lies the value of decentralization, since in a decentralised crowd, information cascades require a relatively infrequent series of random errors.

Evidence of the principles here outlined can be found in Nofer and Hinz (2014) who examined the market returns attainable by following the average recommendation of an online stock community where members could submit their own stock recommendations. The authors found that following the community's recommendation yielded an annualised return 0.59 percentage points higher than that attainable from following the public recommendations of leading banks, though this result must be contextualised by the fact that banks may have ulterior motives to their public stock recommendations, such as maintaining relationships with clients or assisting their investments; moreover, there is a degree to which popular sentiment on stocks can be self-fulfilling, but the fact that only high market capitalisation stocks were considered limits the influence the community could have on their price, and Chen et al. (2011) found in their analysis – which showed that the proportion of negative comments in an online stock community predicts stock returns – that stocks favoured by the community did not experience returns reversals in the long run. Nofer and Hinz's more significant result comes from a random effects model of the determinants of the daily returns of the stock community's recommended stocks, which includes among its independent variables an indicator function for the presence

[4] Though the specialisation and local knowledge that comes with decentralization could be interpreted as reducing the errors of the variance.

of bank recommendations and an indicator variable for the introduction of a ranking system that allowed every user to see how proficient every other user had been in their recommendations. Given a general propensity to trust perceived experts [14] [4] [24], the activation of these indicators ought to decrease the online community's independence and thus the stocks they pick ought to exhibit lower returns, which is exactly what the authors found when they estimated the model.[5] Though they also found that measures of gender and age diversity were not statistically significant, this is the wrong kind of diversity and does not contradict the preceding, as demographic diversity does not entail diversity of opinion.

Prediction markets provide another context that evidentiates the wisdom of crowds, and have the added virtue that outcomes being predicted are unambiguously independent of people's predictions. Simmons et al. (2011) ran an experiment where NFL fans biased[6] in favour of the favourites proved to be worse than chance in picking the winners[7] of NFL games, but when the question was reframed to eliminate bias by asking fans to predict the point spread, which determines whether they estimate the favourites will beat the point spread or not, then the mean and median fan estimate tended to outperform chance. A psychological bias pervading a crowd constitutes a lack of diversity of opinion, and thus Simmons' results are consistent with our theoretical discussion. The strength of this evidence is limited by the fact that the groups being experimented on were always less than 100 in number, but further support can be seen in the fact that bookies often exploit fan bias by taking on unbalanced positions that effectively bet on the favourites losing [23]. Prediction markets extend beyond sports – see Ray (2006) for an overview – and the Iowa Electronic Markets [IEM] (the presidential stock market [IPSM] in particular) have received attention in the literature for their poll-beating electoral predictions, shown by figure 2.1. The market's prediction is garnered from the price of index contracts whose payout is proportional to the vote share a candidate gets; the price, when appropriately scaled, is the market's expectation of that candidate's vote share [3]. Wolfers and Zitzewitz (2004) further show that the consensus of professional forecasters cannot beat the IEM, so it seems that, unlike spread betting in sports, it is difficult to beat the market. We might expect this to be down to the quality of the crowd, but Forsythe et al. (1992) show that the disproportionately republican participants in

[5] Both coefficients were statistically significant at the 1% level at 0.003 and 0.001, respectively, which is considerable given the outcome variable is *daily* returns.

[6] Fans exhibited the assimilation-contrast effect [22] [18].

[7] In terms of beating the point spread.

IPSM were biased in their favour on average. The puzzle of how IPSM can be so accurate despite the bias of its participants is resolved by its anatomy: The market works by its issuing or purchasing a complete bundle of contracts that offer a certain payoff[8] for a price equal to that certain payoff, and then letting the participants trade all available contracts by double auction. It is marginal traders – traders who often place bids or offers at the top of their respective queues – rather than average traders that determine the prevailing price in a double auction process; that process in turn determines contracts' price and therefore the market's expectation, and Forsythe et al. document that these marginal traders did not exhibit the same biases as the average trader, allowing the market price to accurately predict outcomes. This raises the spectre of selecting crowds within crowds for the purpose of forecasting.

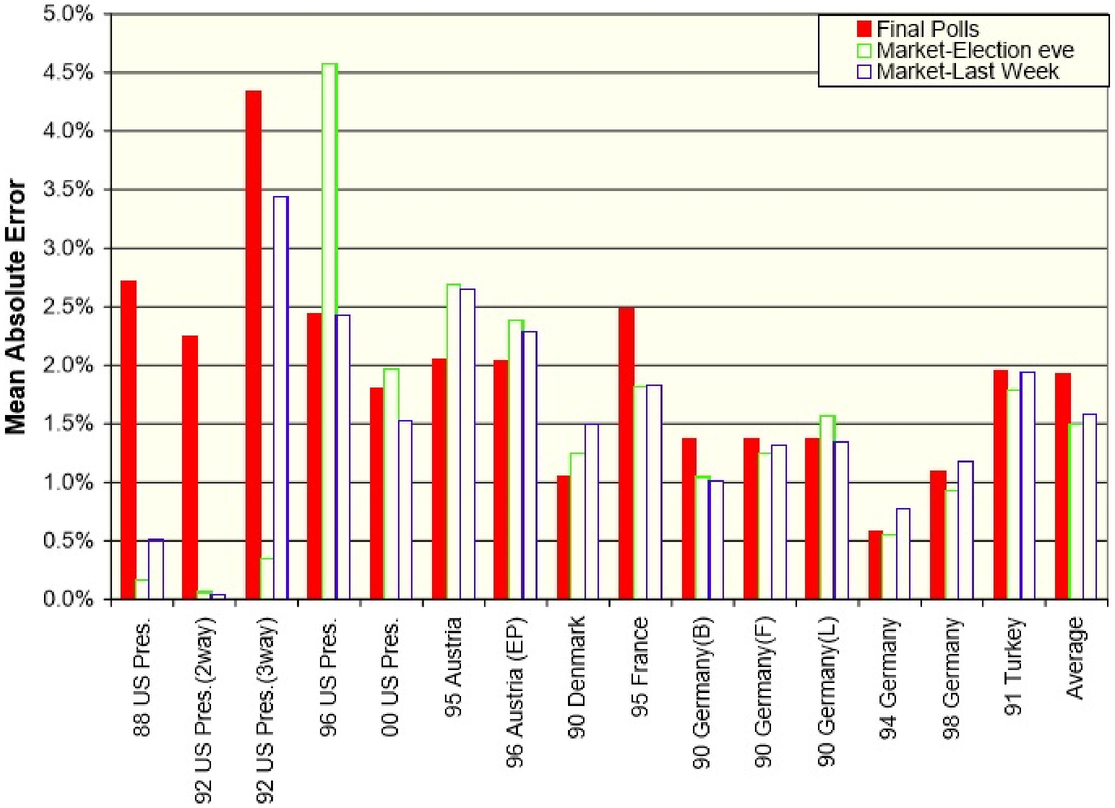

FIGURE 2.1. [3] This figure shows the average absolute errors of predictions of candidate vote shares by the Iowa market on election eve (Market-Election), a week before elections (Market-Last week), and by final polls, across a series of elections.

[8] So, for example, a bundle containing a index contract in each candidate running.

2.2 Select Crowds

Ghiselli (1964, Chapter 7) shows that the correlation between an outcome variable and the mean of a set of predictor variables is given by

$$\rho_{y\bar{x}} = \frac{k^{1/2}\bar{\rho}_{yx}}{\left[1+(k-1)\bar{\rho}_{x_i x_j}\right]^{1/2}} \tag{2.3}$$

where k is the number of predictor variables, $\bar{\rho}_{yx}$ is the average of the correlations between the outcome variable and each predictor variable, and $\bar{\rho}_{x_i x_j}$ is the average of the correlations between each predictor variable and every other predictor variable. Accepting $\rho_{y\bar{x}}$ as a measure for the accuracy of a crowd, Hogarth (1978) uses 2.3 to show that, when a crowd does not exhibit the ideal conditions explored by 2.1 and 2.2, the average of a select subset of a crowd may be more accurate than the whole crowd in that including the prediction of an extra crowd member to a preexisting select crowd estimate[9] only improves accuracy if

$$\rho_{yx_{k+1}} > \rho_{y\bar{x}} \cdot \left(\left[(k+1)+(k+1)k\bar{\rho}'_{x_i x_j}\right]^{1/2} - \left[k+(k-1)k\bar{\rho}_{x_i x_j}\right]^{1/2}\right) \tag{2.4}$$

where $\rho_{yx_{k+1}}$ is the correlation between the additional predictor and the outcome variable and $\bar{\rho}'_{x_i x_j}$ is the average correlation between predictors once the additional predictor is included. Taking the mean of a select crowd formed by the inclusion of members until 2.4 is no longer satisfied thus presents itself as an amelioration to a crowd lacking diversity and independence, and whose estimates are likely correlated. This conclusion is borne out by the work of Mannes et al. (2014), who investigate the performance of taking the average of the k best members – or 'judges' – of a crowd as identified by historical performance, for both real and simulated crowds, as shown by figures 2.2 and 2.3. The simulated crowds are defined by two characteristics: bracketing and dispersion (of expertise). Bracketing refers to the frequency with which any two judges' estimates fall on opposite sides of the truth, whereas dispersion is defined as the coefficient of variation[10] of the mean of judges' absolute error across all simulated judgments. Bracketing relates to diversity and independence inasmuch as a diverse and independent crowd will avoid estimates conglomerating on one side of the truth and is thus likely to have high bracketing. Indeed, a series of diverse and independent estimates drawn from symmetric distributions will exhibit maximal bracketing, since estimates will fall half-and-half on either side of the truth.

[9] i.e. the mean of the select crowd.

[10] This is just the ratio of the standard deviation to the mean.

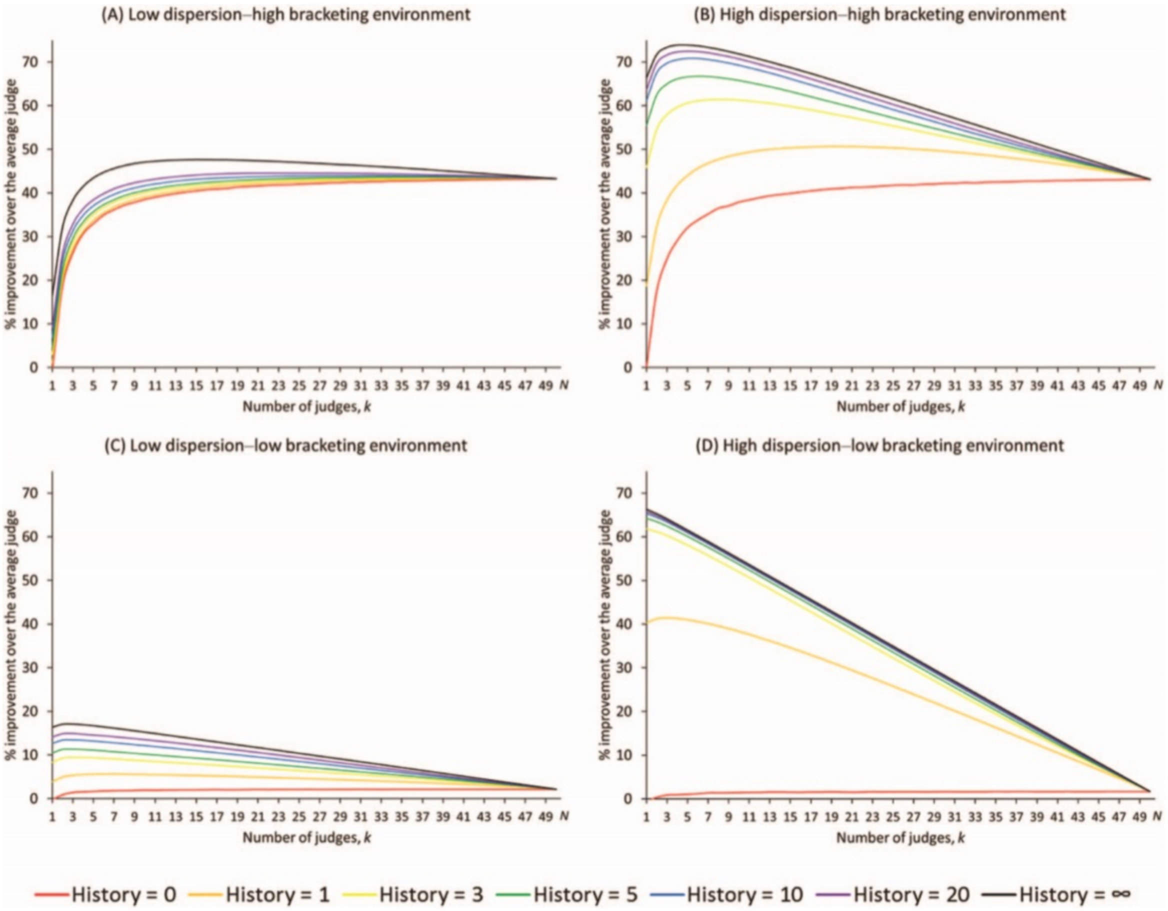

FIGURE 2.2. [14] This figure shows the percentage improvement (in average absolute error across simulated judgments) over the average judge (i.e. over the mean absolute error across all judgments and judges) of the strategy of taking the mean of the k best judges as identified by lowest average absolute error over a variable number of past judgments, indicated by the colour of the lines. This is for four simulated environments with high or low bracketing or dispersion as identified in the titles of the graphs.

Highly skewed distributions are necessary for bracketing not to follow from diversity and independence. The same cannot be said of dispersion inasmuch as a diverse and independent crowd of estimates with some specified symmetry or skew could have any level of dispersion. Perhaps due to this theoretical connection between whole crowds and bracketing, Mannes et al. posit that the error canceling benefits of crowd aggregation will manifest in higher bracketing environments, whereas crowds characterised by higher dispersion are best navigated by seeking a more select group. figure 2.2 bears this hypo-

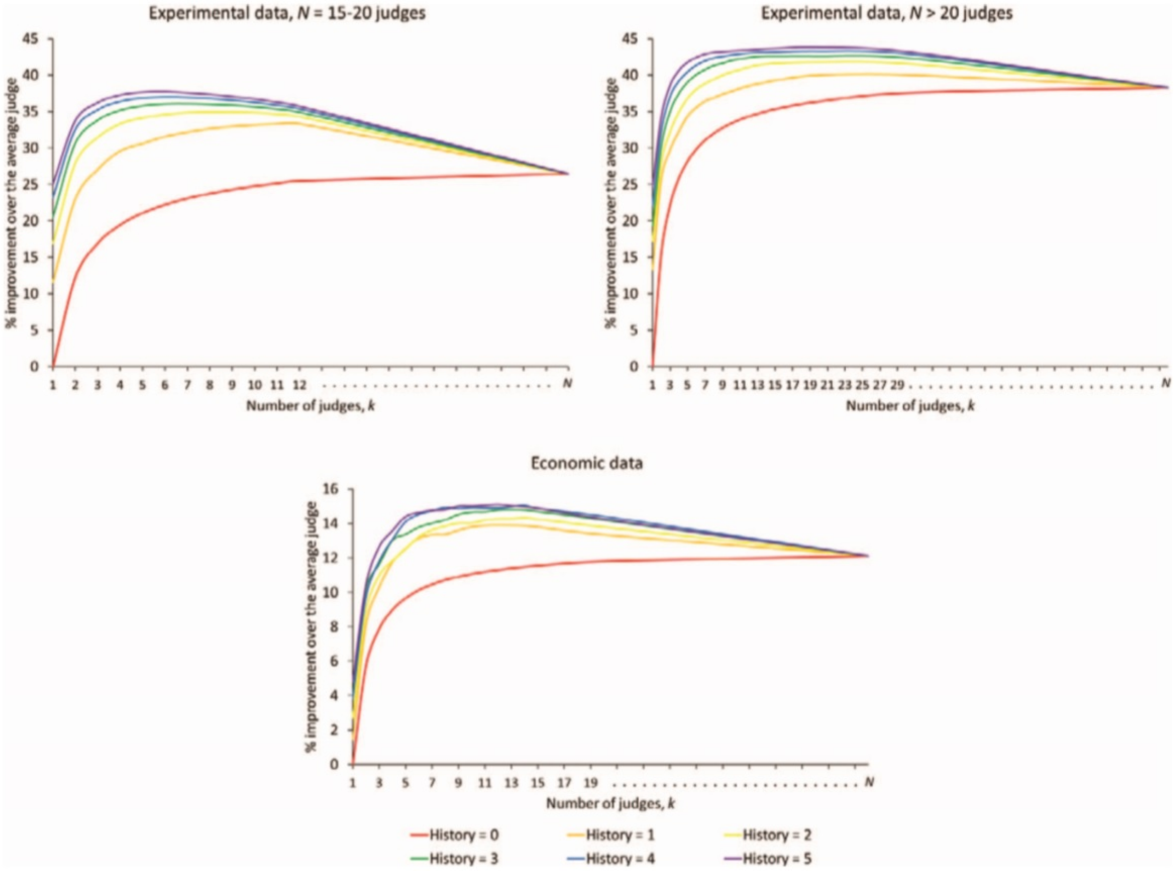

FIGURE 2.3. [14] This figure is similar to the above, but the judgment environments are produced from observed data. 'Experimental' data comes from experiments where participants make miscellaneous estimates, such as about distances displayed in pictures. 'Economic' data comes from published estimates of professional economists. Experimental data is subdivided into experiments with $N = 15$ to 20 or $N > 20$ participants, and the lines indicate the average (over the set of experiments in that category) improvement over the average judge of taking the mean of the k best judges as identified by indicated periods of history.

thesis out, with the optimal size of select crowds in the different simulated environments being negotiated between bracketing and dispersion. We see that a whole crowd strategy, represented on the graphs at $k = N$, demonstrates a significant improvement over the average judge in high bracketing environments, but barely so in low bracketing environments. It is possible to establish mathematical properties regarding the connection between bracketing and collective wisdom, and thereby formalise this discussion.

Consider therefore a series of i.i.d. estimates w_i of a true parameter ϕ drawn from a distribution $f(w)$. Then, we define the bracketing rate as

$$B = P(((w_i < \phi) \cap (w_s > \phi)) \cup ((w_i > \phi) \cap (w_s < \phi))) \tag{2.5}$$

for $i \neq s$, which follows from the aforestated definition. If f is continuous or does not include ϕ in its support, then we may state

$$B = 2F(\phi)(1 - F(\phi)) \tag{2.6}$$

Where $F(\phi)$ is the cumulative distribution at ϕ. If f is a discrete probability density (or mass) function [p.d.f.] with ϕ in the support, then 2.6 is modified to $B = 2(F(\phi) - f(\phi))(1 - F(\phi))$. Therefore, for most p.d.f.s, the bracketing rate is wholly determined by the cumulative distribution at ϕ, which in turn can be used in conjunction with some bounds of f's support to demarcate the supremum error of averaging a whole crowd's estimates when the crowd size tends to infinity. This is proved in one of this paper's original theorems – the bracketing theorem – but we first define two sets for $A, \Omega > 0$.

$$\mathfrak{D}^{A,\Omega,p} = \{f \mid f \text{ is a p.d.f. with support } \mathbb{S} \subseteq [\phi - A, \phi + \Omega] \text{ and cumulative density at } \phi \text{ of } F(\phi) = p\} \tag{2.7}$$

$$\mathfrak{D}_{-}^{A,\Omega,p} = \mathfrak{D}^{A,\Omega,p} \setminus \{f \mid f \text{ is a discrete p.d.f. with } \phi \in \mathbb{S}\} \tag{2.8}$$

Bracketing Theorem. As crowd size tends to infinity, the supremum absolute error of taking the mean of the whole crowd to estimate a true parameter ϕ when estimates of the crowd are i.i.d. draws from some distribution $f \in \mathfrak{D}^{A,\Omega,p}$ is delimited by the cumulative density $F(\phi) = p$ and the bounds $\phi - A$ and $\phi + \Omega$. Mathematically, the statement is

$$\sup_{f \in \mathfrak{D}^{A,\Omega,p}} \left| \frac{1}{n} \sum_{i=0}^{n} w_i - \phi \right| \xrightarrow{P} \max\{\Omega(1-p), Ap\} \text{ as } n \longrightarrow \infty$$

Proof. This theorem is proved through the use of lemmas 2.1-6...

Lemma 2.1. $\sup_{f \in \mathfrak{D}^{A,\Omega,p}} |E(w) - \phi| = \max\{\sup_{f \in \mathfrak{D}^{A,\Omega,p}} E(w) - \phi, |\inf_{f \in \mathfrak{D}^{A,\Omega,p}} E(w) - \phi|\}$

Proof. The least upper bound of $|E(w)-\phi|$ is achieved by making $E(w)-\phi$ either as large or as small as possible. The maximum of the absolute value of the supremum and infimum is therefore least upper bound of the absolute value of $E(w)-\phi^{11}\ldots\Box$

Lemma 2.2. $\sup\limits_{f\in\mathfrak{D}_{-}^{A,\Omega,p}} E(w)-\phi=\Omega(1-p)$

Proof. Consider first continuous p.d.f.s. Then $\forall f : f\in\mathfrak{D}_{-}^{A,\Omega,p}$ and f is a continuous p.d.f.

$$E(w)=\int\limits_{\phi-A}^{\phi+\Omega} wf(w)\mathrm{d}w=\int\limits_{\phi-A}^{\phi} wf(w)\mathrm{d}w+\int\limits_{\phi}^{\phi+\Omega} wf(w)\mathrm{d}w$$

$$<\int\limits_{\phi-A}^{\phi}\phi f(w)\mathrm{d}w+\int\limits_{\phi}^{\phi+\Omega}(\phi+\Omega)f(w)\mathrm{d}w=\phi p+(\phi+\Omega)(1-p)$$

$$\Longleftrightarrow E(w)-\phi<\Omega(1-p)$$

Similarly for discrete p.d.f.s, $\forall f : f\in\mathfrak{D}_{-}^{A,\Omega,p}$ and f is a discrete p.d.f.

$$E(w)=\sum_{i:w_i\in\mathbb{S}} w_i f(w_i)=\sum_{i:w_i\in[\phi-A,\phi]\cap\mathbb{S}} w_i f(w_i)+\sum_{i:w_i\in(\phi,\phi+\Omega]\cap\mathbb{S}} w_i f(w_i)$$

$$\leq\sum_{i:w_i\in[\phi-A,\phi]\cap\mathbb{S}}\phi f(w_i)+\sum_{i:w_i\in(\phi,\phi+\Omega]\cap\mathbb{S}}(\phi+\Omega)f(w_i)=\phi p+(\phi+\Omega)(1-p)$$

$$\Longleftrightarrow E(w)-\phi\leq\Omega(1-p)$$

Thus the least upper bound of $E(w)-\phi$ cannot be greater than $\Omega(1-p)$. Note that $E(w)-\phi=\Omega(1-p)$ for those sets $\mathfrak{D}_{-}^{A,\Omega,p}$ where $p=0$ and p.d.f.s with support $\mathbb{S}=\{\phi+\Omega\}$.

Now consider the series of discrete p.d.f.s $\{f_n\}_{n\in(0,A]}\subset\mathfrak{D}_{-}^{A,\Omega,p}$ defined by

$$f_n(w)=\begin{cases} p & w=\phi-n\\ 1-p & w=\phi+\Omega\\ 0 & \text{otherwise}\end{cases}$$

$\lim\limits_{n\to 0}E_n(w)-\phi=\lim\limits_{n\to 0}(\phi-n)p+(\phi+\Omega)(1-p)-\phi=\Omega(1-p)$, where $E_n(w)$ is the expectation of $w\sim f_n(w)$. This means the least upper bound of $E(w)-\phi$ cannot be less than $\Omega(1-p)$ since an element of $\mathfrak{D}_{-}^{A,\Omega,p}$ can always be found so as to produce a value greater than any

[11] We do not require absolute value bars around the supremum since, if the supremum is negative, then the infimum will necessarily be at least as great in absolute value.

number less than $\Omega(1-p)$. Given that the least upper bound cannot be either greater than or less than the upper bound $\Omega(1-p)$, it must be that $\sup\limits_{f\in\mathfrak{D}_{-}^{A,\Omega,p}} E(w)-\phi=\Omega(1-p)\ldots\square$

Notes on Proof of Lemma 2.2. Instead of using a series of discrete p.d.f.s f_n, the series $\{f_m\}_{m\in(0,A]}\subset\mathfrak{D}_{-}^{A,\Omega,p}$ defined by

$$f_m(w)=\begin{cases} p/m & \phi-m\le w<\phi \\ (1-p)/m & \phi+\Omega-m\le w\le\phi+\Omega \\ 0 & \text{otherwise}\end{cases}$$

could have been used instead to the same effect, so Lemma 2.2 is valid when considering only continuous p.d.f.s as well.

Lemma 2.3. $|\inf\limits_{f\in\mathfrak{D}_{-}^{A,\Omega,p}} E(w)-\phi|=Ap$

Proof. The proof is a mirror image of the proof of Lemma 2.2. Thus, $\forall f: f\in\mathfrak{D}_{-}^{A,\Omega,p}$ and f is a continuous p.d.f.

$$E(w)=\int\limits_{\phi-A}^{\phi} wf(w)\mathrm{d}w+\int\limits_{\phi}^{\phi+\Omega} wf(w)\mathrm{d}w>\int\limits_{\phi-A}^{\phi}(\phi-A)f(w)\mathrm{d}w+\int\limits_{\phi}^{\phi+\Omega}\phi\mathrm{d}w$$

$$=(\phi-A)p+\phi(1-p)\Longleftrightarrow E(w)-\phi>-Ap$$

And similarly for discrete p.d.f.s (with equality for those sets $\mathfrak{D}_{-}^{A,\Omega,p}$ where $p=1$ and p.d.f.s with support $\mathbb{S}=\{\phi-A\}$). Thus the greatest lower bound of $E(w)-\phi$ cannot be less than $-Ap$.

Now consider the series of discrete p.d.f.s $\{f_n\}_{n\in(0,\Omega]}\subset\mathfrak{D}_{-}^{A,\Omega,p}$ given by

$$f_n(w)=\begin{cases} p & w=\phi-A \\ 1-p & w=\phi+n \\ 0 & \text{otherwise}\end{cases}$$

$\lim\limits_{n\to0}E_n(w)-\phi=-Ap$ so the greatest lower bound cannot be greater than $-AP$. Note again that a series of continuous p.d.f.s could have been used instead. It follows from the preceding that $|\inf\limits_{f\in\mathfrak{D}_{-}^{A,\Omega,p}} E(w)-\phi|=Ap\ldots\square$

Lemma 2.4 $\sup\limits_{f\in\mathfrak{D}^{A,\Omega,p}\setminus\mathfrak{D}_{-}^{A,\Omega,p}} E(w)-\phi=\Omega(1-p)$

Proof. This proposition considers discrete p.d.f.s within $\mathfrak{D}^{A,\Omega,p}$ with supports that include ϕ. The set of these p.d.f.s is $\mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}$ and we proceed with the proof in a similar fashion as that in Lemma 2.2. Thus, $\forall f : f \in \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}$

$$E(w) = \sum_{i:w_i \in [\phi - A, \phi] \cap \mathbb{S}} w_i f(w_i) + \sum_{i:w_i \in (\phi, \phi+\Omega] \cap \mathbb{S}} w_i f(w_i) \leq \phi p + (\phi + \Omega)(1-p)$$
$$\iff E(w) - \phi \leq \Omega(1-p)$$

Now consider the following p.d.f. which is a member of $\mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}$

$$f(w) = \begin{cases} p & w = \phi \\ 1-p & w = \phi + \Omega \\ 0 & \text{otherwise} \end{cases}$$

For this p.d.f., $E(w) - \phi = \Omega(1-p)$ and therefore $\sup_{f \in \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}} E(w) - \phi = \Omega(1-p) \ldots \square$

Lemma 2.5 $|\inf_{f \in \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}} E(w) - \phi| = Ap$

Proof. This proof is similar to the preceding proofs, but with a slight asymmetry due to the fact that $F(\phi)$ includes $f(\phi)$. So, $\forall f : f \in \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}$

$$\begin{aligned} E(w) &= \sum_{i:w_i \in [\phi - A, \phi) \cap \mathbb{S}} w_i f(w_i) + \phi f(\phi) + \sum_{i:w_i \in (\phi, \phi+\Omega] \cap \mathbb{S}} w_i f(w_i) \\ &\geq \sum_{i:w_i \in [\phi - A, \phi) \cap \mathbb{S}} (\phi - A) f(w_i) + \phi f(\phi) + \sum_{i:w_i \in (\phi, \phi+\Omega] \cap \mathbb{S}} \phi f(w_i) \\ &= (\phi - A)(p - f(\phi)) + \phi f(\phi) + \phi(1-p) > (\phi - A)p + \phi(1-p) \end{aligned}$$
$$\iff E(w) - \phi > -Ap$$

Now consider the series of p.d.f.s $\{f_n\}_{n \in (0, \min\{1,\Omega\}]} \subset \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}$ defined by

$$f_n(w) = \begin{cases} p - n & w = \phi - A \\ n & w = \phi \\ 1-p & w = \phi + n \\ 0 & \text{otherwise} \end{cases}$$

$\lim_{n \to 0} E_n(w) - \phi = -Ap$ and therefore $|\inf_{f(w) \in \mathfrak{D}^{A,\Omega,p} \backslash \mathfrak{D}_{-}^{A,\Omega,p}} E(w) - \phi| = Ap \ldots \square$

Lemmas 2.1-5 jointly imply that $\sup_{f \in \mathfrak{D}^{A,\Omega,p}} |E(w) - \phi| = \max\{\Omega(1-p), Ap\}$

Lemma 2.6. $\sup_{f \in \mathfrak{D}^{A,\Omega,p}} \left| \frac{1}{n} \sum_{i=0}^{n} w_i - \phi \right| \xrightarrow{P} \sup_{f \in \mathfrak{D}^{A,\Omega,p}} |E(w) - \phi|$

Proof. Lemma 2.6 follows from the weak law of large numbers and continuous mapping theorem

Lemmas 2.1-6 therefore jointly imply the bracketing theorem... ■

For $f \in \mathfrak{D}_{-}^{A,\Omega,p}$, the bracketing rate B is wholly determined by p, and we can see from the bracketing theorem that at $B_{max} = \max_{p \in [0,1]} B$ where $p = 0.5$ the supremum error is minimised for $A = \Omega$. Moreover, a p.d.f. where $p = 0.5$ and the bracketing rate is maximised could also be symmetric, in which case $E(w) = \phi$ and the error is zero. Note that in this case the drawn crowd would be independent and diverse. Knowing the bracketing rate pins down the limiting properties of a whole crowd average less than diversity and independence jointly do, but its apparent role in select crowds ensure it is revisited later in this paper.

Practical experience with select crowds can be seen in Hill and Ready-Campbell (2011), who longed (shorted) the up-voted (down-voted) stocks of the best members of The Motley Fools' CAPS online community. They invested equal amounts of money in each 'best member,' and for each member split his or her allocated funds equally between his or her up or down voted stocks. Best members were identified by past performance, with the size of the select crowd of best members and the time frames for evaluation being set by a genetic algorithm with the objective of maximising the strategy's Sharpe ratio [21]. Hill and Ready-Campbell's strategy outperforms a whole crowd strategy of longing or shorting stocks in proportion to their up or down votes from the whole online community, which itself outperforms the S&P 500. Combined with Mannes et al.'s simulated and experimental results, this militates in favour of a vox predictor that makes use of a performance related subset of a crowd of estimates, but to understand precisely the cognition problem a vox predictor must contend with, and the form of its contention, we must frame the discussion in a time series context.

CHAPTER 3

THEORETICAL TIME SERIES DISCUSSION

3.1 OLS and the Vox Predictor

If a vox predictor is to be compared to OLS, a theoretical review of OLS's predictive power can help to place the vox predictor in the right context. Below is simple linear data generating process [DGP].

$$y_t = X_t'\Phi + \varepsilon_t \tag{3.1}$$

Where y_t is a scalar outcome variable, X_t is a vector of predictor variables, Φ is a vector of parameters, and ε_t is a scalar mean zero and variance σ^2 shock independent of both $\varepsilon_{j\neq t}$ and X_t. The Gauss-Markov theorem guarantees that OLS will be the best (most efficient) linear unbiased estimator of Φ. So long as $\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$ and $\mathrm{E}\left(||X_t\varepsilon_t||^2\right)$ are finite, the Lindeberg condition holds for $X_t\varepsilon_t$, and $\frac{\sigma^2}{T}\sum_{t=1}^{T}\mathrm{E}[X_tX_t'|X_{t-1}\varepsilon_{t-1},\ldots,X_1\varepsilon_1]$ converges in probability to a positive definite matrix, then OLS consistently estimates Φ, as proved in appendix A.1. For processes such as 3.1 therefore, OLS efficiently, unbiasedly, and consistently estimates the conditional expectation function, which itself minimises the mean squared error for whichever set of observations it is applied to, as proved in appendix A.2. The mathematical certainty that the conditional expectation minimises mean squared error sets this metric – or its sample analogue – up as a natural way to evaluate the proficiency of prediction methods, with the superior method being that which best approximates the conditional expectation function. This gives a clear cognition problem for a vox predictor: estimate the conditional expectation function.

Given the preceding chapter, we know that independently drawing a large number of diverse[1] estimates of a parameter and taking the mean would estimate any true parameter of a conditional expectation function under the aforestated regularity conditions. Though random number generators can produce independent estimates, the true parameter would have to be already known in order to specify a diverse distribution from which to draw the estimates. Predicting an outcome variable with a perfectly accurate model of the conditional expectation function is therefore untenable, but our discussion in section 2.2 suggests using a (past performance related) select subset of a crowd of estimates of the outcome variable, which could be formed with independent draws of parameters of the supposed conditional expectation function.[2] This leads to definition 3.1.

Definition 3.1. *Vox Predictor*. A vox predictor is any predictor of a time series y_t given by $\hat{y}_t^{\text{vox}} = \bar{g}_t(X_t, \hat{\Phi}_i)$, where each guess $g(X_t, \hat{\Phi}_i)$ is a function of the vector or scalar of observed variables X_t and the vector or scaler of parameters $\hat{\Phi}_i$ produced by independently drawing at least one parameter from some distribution. $\bar{g}_t(X_t, \hat{\Phi}_i)$ is the mean of some performance related subset of $\{g(X_t, \hat{\Phi}_i)\}_{i \in \{1,\ldots,n\}}$ at time t for $i = 1, \ldots, n$ estimates.

Our discussion above indicates a vox predictor (or indeed any other predictor) could not outperform OLS when the DGP is given by 3.1 and X_t is composed of observed variables. However, the case may be different when some of the variables are unobserved, as in 3.2.

$$y_t = X_t'\Phi + Z_t'\Psi + \varepsilon_t \tag{3.2}$$

Where Z_t now represents a vector of unobserved variables and Ψ is another vector of parameters. The conditional expectation of y_t given X_t is thus

$$\mathrm{E}[y_t \mid X_t] = X_t'\Phi + \mathrm{E}[Z_t' \mid X_t]\Psi \tag{3.3}$$

The conditional expectation can still be recovered by OLS with similar regularity conditions to those required for 3.1 so long as Z_t is linearly related to X_t, as proved in appendix A.3. Thus, if unobserved variables are to prevent OLS's superiority to the vox predictor in the mean squared error sense from being a forgone conclusion, they

[1] i.e. whose mean is equal to the true parameter.

[2] Though parameter estimates may be independently drawn and therefore uncorrelated, the predictions that they subsequently produce are liable to correlated, and therefore 2.4 is applicable.

cannot be linearly related to the observed variables. Moreover, a vox predictor using a proper subset of $\{g(X_t,\hat{\Phi}_i)\}_{i\in\{1,\ldots,n\}}$ may amount to a time varying approximation of the conditional expectation function, whereas OLS returns a time invariant function, so a DGP with unobserved variables that drives a time variant conditional expectation function when conditioned only on observed variables could create a niche in which a vox predictor can shine. We therefore consider 3.4.

$$y_t = \phi y_{t-1} + \psi z_t + \varepsilon_t \text{ with } z_t = \beta z_{t-1} + e_t \text{ and } y_0 = z_0 = 0 \tag{3.4}$$

Where all variables and parameters[3] are scalar, past values of y_t are observed, z_t is unobserved, $\varepsilon_t \sim i.i.d.N(0,\sigma_\varepsilon^2)$, and $e_t \sim i.i.d.N(0,\sigma_e^2)$. z_t is not linearly related to y_{t-1}, and its unobserved and autoregressive nature suggest that if $\psi > 0$, then its being above (below) its mean at time t may cause ϕy_{t-1} to underestimate (overestimate) y_t, and if $\beta > 0$, it is liable to do so again at time $t+1$. If $\beta < 0$ then an overestimation is liable to follow an underestimation and vice versa. In either case, $y_{t-1} - \phi y_{t-2}$ appears relevant to estimating y_t, and its significance could be captured by the vox estimator in a way that OLS cannot. The significance of $y_{t-1} - \phi y_{t-2}$ can be deduced by deriving the distribution of y_t conditional on y_{t-1} and y_{t-2}, but this derivation, presented in section 3.3, requires the conditional normality theorem below, which is originally proved.

3.2 Conditional Normality Theorem

Conditional Normality Theorem. Let a and b be two independent normally distributed mean zero random variables with variances σ_a^2 and σ_b^2, respectively. If it is given that $a+b=x$, then a given x is normally distributed with mean $\sigma_a^2 x/(\sigma_a^2+\sigma_b^2)$ and variance $\sigma_a^2\sigma_b^2/(\sigma_a^2+\sigma_b^2)$.

Proof. Let f_a and f_b be the p.d.f.s of a and b, respectively. Given their independence, and the fact that the joint probability that a takes on some value and $a+b=x$ is merely the joint probability that a takes on that value and b takes on x minus that value, Bayes' formula tells us that $f(a\,|\,x) = f_a(a)f_b(x-a)/f_x(x)$. We proceed by first computing $f_x(x) = f(a+b=x)$, which is the sum of the point densities of all (mutually exclusive) realisations that lead a and b to sum to x; the set of these realisations is $\{(a, a-x)\}_{a\in\mathbb{R}}$ and thus we get

[3] The parameters are the greek letters and are non-zero.

$$
\begin{aligned}
f(a+b=x) &= \int_{-\infty}^{\infty} f_a(a) f_b(x-a)\,\mathrm{d}a \\
&= \int_{-\infty}^{\infty} \frac{e^{-\frac{a^2}{2\sigma_a^2}}}{\sqrt{2\pi\sigma_a^2}} \cdot \frac{e^{-\frac{(x-a)^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\mathrm{d}a \\
&= \frac{1}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{a^2}{\sigma_a^2}+\frac{(x-a)^2}{\sigma_b^2}\right)}\mathrm{d}a \\
&= \frac{1}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left(\sigma_b^2 a^2+\sigma_a^2(x-a)^2\right)}\mathrm{d}a \\
&= \frac{1}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left((\sigma_a^2+\sigma_b^2)a^2-2\sigma_a^2 ax+\sigma_a^2 x^2\right)}\mathrm{d}a \\
&= \frac{e^{-\frac{x^2}{2\sigma_b^2}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left((\sigma_a^2+\sigma_b^2)a^2-2\sigma_a^2 ax\right)}\mathrm{d}a \\
&= \frac{e^{-\frac{x^2}{2\sigma_b^2}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left(\left(a\sqrt{\sigma_a^2+\sigma_b^2}-\frac{\sigma_a^2 x}{\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2-\frac{\sigma_a^4x^2}{\sigma_a^2+\sigma_b^2}\right)}\mathrm{d}a \\
&= \frac{e^{-\frac{x^2}{2\sigma_b^2}}\cdot e^{\frac{\sigma_a^4x^2}{2\sigma_a^2\sigma_b^2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left(a\sqrt{\sigma_a^2+\sigma_b^2}-\frac{\sigma_a^2 x}{\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2}\mathrm{d}a \\
&= \frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_a^2\sigma_b^2}\left(a\sqrt{\sigma_a^2+\sigma_b^2}-\frac{\sigma_a^2 x}{\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2}\mathrm{d}a \\
&= \frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{a(\sigma_a^2+\sigma_b^2)-\sigma_a^2 x}{\sigma_a\sigma_b\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2}\mathrm{d}a
\end{aligned}
$$

We make the substitution $u = \dfrac{a(\sigma_a^2+\sigma_b^2)-\sigma_a^2 x}{\sigma_a\sigma_b\sqrt{\sigma_a^2+\sigma_b^2}}$

$$
\begin{aligned}
&= \frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sigma_a\sigma_b} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\cdot\frac{\sigma_a\sigma_b}{\sqrt{\sigma_a^2+\sigma_b^2}}\cdot\mathrm{d}u \\
&= \frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\mathrm{d}u = Q
\end{aligned}
$$

Having set the last line of the integration equal to Q, we proceed with the computation by squaring Q and converting into polar coordinates, which is a technique that can be found in such tomes as *Elementary Linear Algebra* [1] or chapter 14 of *Calculus* [25].

$$Q^2 = \left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\mathrm{d}u\right)\cdot\left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\mathrm{d}u\right)$$

$$= \left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\mathrm{d}u\right)\cdot\left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}v^2}\mathrm{d}v\right)$$

$$= \left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{2\pi\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2\cdot\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\mathrm{d}u\right)\cdot\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}v^2}\mathrm{d}v\right)$$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2+v^2)}\mathrm{d}u\mathrm{d}v$$

Now we make the substitutions $u = r\cos\theta$ and $v = r\sin\theta$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}((r\cos\theta)^2+(r\sin\theta)^2)}|\det(\mathbf{J})|\,\mathrm{d}r\mathrm{d}\theta$$

Where $\mathbf{J}$ is the Jacobian matrix which scales the area element, with $\det(\mathbf{J}) = \begin{vmatrix}\frac{\partial u}{\partial\theta} & \frac{\partial u}{\partial r}\\ \frac{\partial v}{\partial\theta} & \frac{\partial v}{\partial r}\end{vmatrix} = \begin{vmatrix}-r\sin\theta & \cos\theta\\ r\cos\theta & \sin\theta\end{vmatrix} = -r$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2(\cos^2\theta+\sin^2\theta)}r\,\mathrm{d}r\mathrm{d}\theta$$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2}r\,\mathrm{d}r\mathrm{d}\theta$$

We make the substitution $w = \dfrac{r^2}{2}$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)}\int_0^{2\pi}\int_0^{\infty} e^{-w}\mathrm{d}w\mathrm{d}\theta$$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)} \int_0^{2\pi} \left[-e^{-w}\right]_0^\infty \,\mathrm{d}\theta$$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{4\pi^2(\sigma_a^2+\sigma_b^2)} \int_0^{2\pi} \mathrm{d}\theta$$

$$= \frac{e^{-\frac{x^2}{\sigma_a^2+\sigma_b^2}}}{2\pi(\sigma_a^2+\sigma_b^2)}$$

Given that p.d.f.s output only positive values, the integral Q must be positive, and therefore we compute the positive square root of Q^2 to complete the computation.

$$Q = f_x(x) = \frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{\sqrt{2\pi(\sigma_a^2+\sigma_b^2)}}$$

We now use Bayes' formula to compute $f(a\,|\,x)$ and thereby conclude the proof

$$f(a\,|\,x) = \left(\frac{e^{-\frac{x^2}{2(\sigma_a^2+\sigma_b^2)}}}{\sqrt{2\pi(\sigma_a^2+\sigma_b^2)}} \right)^{-1} \cdot \frac{e^{-\frac{a^2}{2\sigma_a^2}}}{\sqrt{2\pi\sigma_a^2}} \cdot \frac{e^{-\frac{(x-a)^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{\frac{x^2}{2(\sigma_a^2+\sigma_b^2)} - \frac{a^2}{2\sigma_a^2} - \frac{(x-a)^2}{2\sigma_b^2}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{\frac{\sigma_a^2\sigma_b^2x^2 - \sigma_b^2(\sigma_a^2+\sigma_b^2)a^2 - \sigma_a^2(\sigma_a^2+\sigma_b^2)(x-a)^2}{2\sigma_a^2\sigma_b^2(\sigma_a^2+\sigma_b^2)}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{\frac{-\sigma_a^2a^2 - \sigma_b^2a^2 - 2\sigma_a^2\sigma_b^2a^2 + 2\sigma_a^4ax + 2\sigma_a^2\sigma_b^2ax - \sigma_a^4x}{2\sigma_a^2\sigma_b^2(\sigma_a^2+\sigma_b^2)}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{\frac{-(\sigma_a^2+\sigma_b^2)^2a^2 + 2\sigma_a^2(\sigma_a^2+\sigma_b^2)ax - \sigma_a^4x^2}{2\sigma_a^2\sigma_b^2(\sigma_a^2+\sigma_b^2)}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{-\frac{\left((\sigma_a^2+\sigma_b^2)a - \sigma_a^2x\right)^2}{2\sigma_a^2\sigma_b^2(\sigma_a^2+\sigma_b^2)}}$$

$$= \sqrt{\frac{\sigma_a^2+\sigma_b^2}{2\pi\sigma_a^2\sigma_b^2}}\, e^{-\frac{1}{2}\left(\frac{(\sigma_a^2+\sigma_b^2)a - \sigma_a^2x}{\sigma_a\sigma_b\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma_a^2\sigma_b^2/(\sigma_a^2+\sigma_b^2)}} e^{-\frac{1}{2}\left(\frac{(\sigma_a^2+\sigma_b^2)a-\sigma_a^2 x}{\sigma_a\sigma_b\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2} \\
&= \frac{1}{\sqrt{2\pi\sigma_a^2\sigma_b^2/(\sigma_a^2+\sigma_b^2)}} e^{-\frac{1}{2}\left(\frac{a-\sigma_a^2 x/(\sigma_a^2+\sigma_b^2)}{\sigma_a\sigma_b/\sqrt{\sigma_a^2+\sigma_b^2}}\right)^2} \\
&= \frac{1}{\sqrt{2\pi\sigma_{a|x}^2}} e^{-\frac{1}{2}\left(\frac{a-\mu_{a|x}}{\sigma_{a|x}}\right)^2}
\end{aligned}
$$

Where $\mu_{a|x} = \sigma_a^2 x/(\sigma_a^2+\sigma_b^2)$ and $\sigma_{a|x}^2 = \sigma_a^2\sigma_b^2/(\sigma_a^2+\sigma_b^2)$, which shows that $a\,|\,x$ is normally distributed with mean $\sigma_a^2 x/(\sigma_a^2+\sigma_b^2)$ and variance $\sigma_a^2\sigma_b^2/(\sigma_a^2+\sigma_b^2)$, as we set out to prove . . . ■

The conditional normality theorem uses mean zero random variables because this serves our purposes of deriving the distribution of $y_t\,|\,y_{t-1}, y_{t-2}$, but it can be extended to derive the distribution $a\,|\,x$ when a and b have non-zero means, as proved in appendix A.4.

3.3 Conditional Distribution of the DGP in 3.4

Here, we provide our own fully worked out construction of the conditional distribution, starting by breaking down the DGP in 3.4.

$$
\begin{aligned}
& y_{t-1} = \phi y_{t-2} + \psi z_{t-1} + \varepsilon_{t-1} \\
\Longrightarrow & \psi z_{t-1} = y_{t-1} - \phi y_{t-2} - \varepsilon_{t-1} \\
\Longrightarrow & \psi z_t = \beta y_{t-1} - \beta\phi y_{t-2} - \beta\varepsilon_{t-1} + \psi e_t \\
\Longrightarrow & y_t = \phi y_{t-1} + \beta y_{t-1} - \beta\phi y_{t-2} - \beta\varepsilon_{t-1} + \psi e_t + \varepsilon_t \\
\Longrightarrow & y_t = (\phi+\beta) y_{t-1} - \beta\phi y_{t-2} - \beta\varepsilon_{t-1} + \psi e_t + \varepsilon_t
\end{aligned} \tag{3.5}
$$

e_t and ε_t are independent of y_{t-1} and y_{t-2} but ε_{t-1} is not, which can be seen from

$$
y_{t-1} = \phi y_{t-2} + \psi z_{t-1} + \varepsilon_{t-1} \tag{3.6}
$$

$z_{t-1} = \sum_{i=0}^{t-2} \beta^i e_{t-1-i}$ is a sum of normal random variables each of which is independent of each other and of ε_{t-1}, making ψz_{t-1} and ε_{t-1} two independent normally distributed variables.[4] Therefore, if y_{t-1} and y_{t-2} are given, then the conditional normality theorem is applicable, since two independent normally distributed random variables sum to $y_{t-1} - \phi y_{t-2}$. We have only to determine the variance of these two random variables to determine the conditional distribution of ε_{t-1}. The variance of ψz_{t-1} is

$$\text{Var}(\psi z_{t-1}) = \psi^2 \sum_{i=0}^{t-2} \text{Var}(\beta^i e_{t-1-i}) = \psi^2 \sigma_e^2 \sum_{i=0}^{t-2} \beta^{2i} = \psi^2 \sigma_e^2 \frac{(1-\beta^{2(t-1)})}{1-\beta^2} \tag{3.7}$$

The variance of ε_{t-1} is simply σ_ε^2. Consequently, we can ascertain from the conditional normality theorem that

$$\text{E}[\varepsilon_{t-1} \,|\, y_{t-1}, y_{t-2}] = \frac{\sigma_\varepsilon^2 (y_{t-1} - \phi y_{t-2})}{\sigma_\varepsilon^2 + \psi^2 \sigma_e^2 \left(\frac{1-\beta^{2(t-1)}}{1-\beta^2}\right)} = \frac{\sigma_\varepsilon^2 (1-\beta^2)(y_{t-1} - \phi y_{t-2})}{\sigma_\varepsilon^2 (1-\beta^2) + \psi^2 \sigma_e^2 (1-\beta^{2(t-1)})} \tag{3.8}$$

$$\text{Var}[\varepsilon_{t-1} \,|\, y_{t-1}, y_{t-2}] = \frac{\psi^2 \sigma_\varepsilon^2 \sigma_e^2 \left(\frac{1-\beta^{2(t-1)}}{1-\beta^2}\right)}{\sigma_\varepsilon^2 + \psi^2 \sigma_e^2 \left(\frac{1-\beta^{2(t-1)}}{1-\beta^2}\right)} = \frac{\psi^2 \sigma_\varepsilon^2 \sigma_e^2 (1-\beta^{2(t-1)})}{\sigma_\varepsilon^2 (1-\beta^2) + \psi^2 \sigma_e^2 (1-\beta^{2(t-1)})} \tag{3.9}$$

This allows us to fully detail the distribution of y_t conditional on y_{t-1} and y_{t-2}. We first consider the conditional expectation of y_t.

$$\begin{aligned}
&\text{E}[y_t \,|\, y_{t-1}, y_{t-2}] \\
&= \text{E}[(\phi + \beta) y_{t-1} - \beta \phi y_{t-2} - \beta \varepsilon_{t-1} + \psi e_t + \varepsilon_t \,|\, y_{t-1}, y_{t-2}] \\
&= (\phi + \beta) y_{t-1} - \beta \phi y_{t-2} - \beta \text{E}[\varepsilon_{t-1} \,|\, y_{t-1}, y_{t-2}] \\
&= (\phi + \beta) y_{t-1} - \beta \phi y_{t-2} - \frac{\beta \sigma_\varepsilon^2 (1-\beta^2)(y_{t-1} - \phi y_{t-2})}{\sigma_\varepsilon^2 (1-\beta^2) + \psi^2 \sigma_e^2 (1-\beta^{2(t-1)})}
\end{aligned}$$

Let $D = \sigma_\varepsilon^2 (1-\beta^2) + \psi^2 \sigma_e^2 (1-\beta^{2(t-1)})$. Then we have

$$\frac{1}{D}\left((\phi + \beta) y_{t-1} D - \beta \left(\phi y_{t-2} D + \sigma_\varepsilon^2 (1-\beta^2)(y_{t-1} - \phi y_{t-2})\right)\right)$$

Let us consider first the rightmost bracketed term of the numerator

$$\begin{aligned}
&\phi y_{t-2} D + \sigma_\varepsilon^2 (1-\beta^2)(y_{t-1} - \phi y_{t-2}) \\
&= \phi \sigma_\varepsilon^2 (1-\beta^2) y_{t-2} + \phi \psi^2 \sigma_e^2 (1-\beta^{2(t-1)}) y_{t-2} + \sigma_\varepsilon^2 (1-\beta^2) y_{t-1} - \phi \sigma_\varepsilon^2 (1-\beta^2) y_{t-2} \\
&= \phi \psi^2 \sigma_e^2 (1-\beta^{2(t-1)}) y_{t-2} + \sigma_\varepsilon^2 (1-\beta^2) y_{t-1}
\end{aligned}$$

Considering the whole numerator and substituting in this expression yields

$$(\phi + \beta) y_{t-1} D - \beta \left(\phi \psi^2 \sigma_e^2 (1-\beta^{2(t-1)}) y_{t-2} + \sigma_\varepsilon^2 (1-\beta^2) y_{t-1}\right)$$

[4] since the sum of independent normally distributed random variables is itself a normally distributed random variable.

Expanding D and canceling terms gives us

$$\phi\sigma_e^2(1-\beta^2)y_{t-1}+(\phi+\beta)\psi^2\sigma_e^2(1-\beta^{2(t-1)})y_{t-1}-\beta\phi\psi^2\sigma_e^2(1-\beta^{2(t-1)})y_{t-2}$$
$$=\phi\sigma_\varepsilon^2(1-\beta^2)y_{t-1}+\phi\psi^2\sigma_e^2(1-\beta^{2(t-1)})y_{t-1}+\beta\left[\psi^2\sigma_e^2(1-\beta^{2(t-1)})y_{t-1}-\phi\psi^2\sigma_e^2(1-\beta^{2(t-1)})y_{t-2}\right]$$
$$=\left[\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2(1-\beta^{2(t-1)})\right]\phi y_{t-1}+\beta\psi^2\sigma_e^2(1-\beta^{2(t-1)})(y_{t-1}-\phi y_{t-2})$$
$$=D\phi y_{t-1}+\beta\psi^2\sigma_e^2(1-\beta^{2(t-1)})(y_{t-1}-\phi y_{t-2})$$

Reintroducing the denominator by dividing through by D implies 3.10

$$\text{(3.10)}\qquad \mathrm{E}[y_t\,|\,y_{t-1},y_{t-2}]=\phi y_{t-1}+\frac{\beta\psi^2\sigma_e^2(1-\beta^{2(t-1)})(y_{t-1}-\phi y_{t-2})}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2(1-\beta^{2(t-1)})}$$

The above expression makes use of the formula for a sum of a geometric series, since we turned $\sum_{i=0}^{t-2}\beta^{2i}$ into $\frac{1-\beta^{2(t-1)}}{1-\beta^2}$ in our expression for the variance of z_{t-1}, but this formula fails to produce an intelligible expression for $|\beta|=1$. Therefore we also present an alternative expression in terms of $S(t)=\sum_{i=0}^{t-2}\beta^{2i}$ below.

$$\text{(3.11)}\qquad \mathrm{E}[y_t\,|\,y_{t-1},y_{t-2}]=\phi y_{t-1}+\frac{\beta\psi^2\sigma_e^2S(t)(y_{t-1}-\phi y_{t-2})}{\sigma_\varepsilon^2+\psi^2\sigma_e^2S(t)}$$

These expressions show that the conditional expectation is ϕy_{t-1} plus a *time varying* coefficient multiplied by $y_{t-1}-\phi y_{t-2}$. The coefficient is positive if β is positive, and negative if β is negative, showing that a prediction of y_t using ϕy_{t-1} ought to be augmented in the same direction as the last prediction's error if β is positive, and in the contrary direction if β is negative, confirming our suspicions in section 3.1. We turn now to the conditional variance of y_t, which is the sum of the conditional variances of $\beta\varepsilon_{t-1}$, ψe_t and ε_t, since these are independent. Thus

$$\text{(3.12)}\qquad \begin{aligned}\mathrm{Var}[y_t\,|\,y_{t-1},y_{t-2}]&=\beta^2\mathrm{Var}[\varepsilon_t\,|y_{t-1},y_{t-2}]+\psi^2\sigma_e^2+\sigma_\varepsilon^2\\&=\frac{\beta^2\psi^2\sigma_\varepsilon^2\sigma_e^2(1-\beta^{2(t-1)})}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2(1-\beta^{2(t-1)})}+\psi^2\sigma_e^2+\sigma_\varepsilon^2\end{aligned}$$

Again, we provide an alternative expression for the variance in terms of $S(t)$

$$\text{(3.13)}\qquad \mathrm{Var}[y_t\,|\,y_{t-1},y_{t-2}]=\frac{\beta^2\psi^2\sigma_\varepsilon^2\sigma_e^2S(t)}{\sigma_\varepsilon^2+\psi^2\sigma_e^2S(t)}+\psi^2\sigma_e^2+\sigma_\varepsilon^2$$

We can see from these expressions that the conditional variance is also split into the sum of time varying and time invariant components. If we call the conditional expectation $\mu(t)$ and the conditional variance $\sigma^2(t)$, then the fact that $\varepsilon_{t-1}\,|\,y_{t-1}, y_{t-2}$, e_t, and ε_t are normally distributed[5] allows us to state

$$y_t\,|\,y_{t-1}, y_{t-2} \sim N\left(\mu(t), \sigma^2(t)\right) \tag{3.14}$$

With our expressions for the conditional expectation and variance, we can explore how the conditional distribution of y_t develops over time. If we divide the numerator and the denominator of the time varying part of 3.10 by $\psi^2\sigma_e^2(1-\beta^{2(t-1)})$ then we obtain

$$\mu(t) = \phi y_{t-1} + \frac{\beta(y_{t-1}-\phi y_{t-2})}{\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^{2(t-1)})}+1} \tag{3.15}$$

If $|\beta| < 1$ [6] then $\psi^2\sigma_e^2(1-\beta^{2(t-1)})$ starts smaller than $\psi^2\sigma_\varepsilon^2$ and tends to $\psi^2\sigma_e^2$ as $t \longrightarrow \infty$. Thus, $\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^{2(t-1)})}+1$ diminishes as t grows and the coefficient on $y_{t-1}-\phi y_{t-2}$ gets larger in absolute value until it 'reaches' $\frac{\beta\psi^2\sigma_e^2}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2}$ in the limit. If $|\beta| \geq 1$ then we can examine the dynamics of the conditional expectation of y_t by expressing it in terms of $S(t)$.

$$\mu(t) = \phi y_{t-1} + \frac{\beta(y_{t-1}-\phi y_{t-2})}{\frac{\sigma_\varepsilon^2}{\psi^2\sigma_e^2 S(t)}+1} \tag{3.16}$$

$S(t)$ diverges to ∞ as $t \longrightarrow \infty$ and therefore $\frac{\sigma_\varepsilon^2}{\psi^2\sigma_e^2 S(t)} \longrightarrow 0$ meaning that the coefficient on $y_t - \phi y_{t-1}$ again grows in magnitude until it 'reaches' β at infinity.

Repeating this analysis for the conditional variance and dividing 3.12 through by $\psi^2\sigma_e^2(1-\beta^{2(t-1)})$, we see that

$$\sigma^2(t) = \frac{\beta^2\sigma_\varepsilon^2}{\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^{2(t-1)})}+1} + \psi^2\sigma_e^2 + \sigma_\varepsilon^2 \tag{3.17}$$

For $|\beta| < 1$ we see again that $\psi^2\sigma_e^2(1-\beta^{2(t-1)})$ starts smaller than $\psi^2\sigma_\varepsilon^2$ and tends to $\psi^2\sigma_e^2$ as $t \longrightarrow \infty$, meaning that $\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^{2(t-1)})}+1$ decreases and the time varying part of

[5] Note that $e_t = e_t\,|\,y_{t-1}, y_{t-2}$ and $\varepsilon_t = \varepsilon_t\,|\,y_{t-1}, y_{t-2}$.

[6] which is to say that z_t is stable.

the conditional variance increases until it 'reaches' $\frac{\beta^2\psi^2\sigma_\varepsilon^2\sigma_e^2}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2}$ at infinity. To consider the development of the conditional variance when $|\beta| \geq 1$ we re-express it in terms of $S(t)$

$$\sigma^2(t) = \frac{\beta^2\sigma_\varepsilon^2}{\frac{\sigma_\varepsilon^2}{\psi^2\sigma_e^2 S(t)}+1} + \psi^2\sigma_e^2 + \sigma_\varepsilon^2 \tag{3.18}$$

This shows, by the exact same reasoning as for the conditional expectation, that the time varying part of the conditional variance grows until it settles down to $\beta^2\sigma_\varepsilon^2$. These results jointly show that the asymptotic distribution of y_t conditional on y_{t-1} and y_{t-2} is

$$y_t \,|\, y_{t-1}, y_{t-2} \overset{a}{\sim} \begin{cases} N(\mu_{A,S}, \sigma_{A,S}^2) & |\beta| < 1 \\ N(\mu_{A,NS}, \sigma_{A,NS}^2) & |\beta| \geq 1 \end{cases}$$

where

$$\begin{aligned} \mu_{A,S} &= \phi y_{t-1} + \frac{\beta\psi^2\sigma_e^2(y_{t-1}-\phi y_{t-2})}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2} \\ \sigma_{A,S}^2 &= \frac{\beta^2\psi^2\sigma_\varepsilon^2\sigma_e^2}{\sigma_\varepsilon^2(1-\beta^2)+\psi^2\sigma_e^2} + \psi^2\sigma_e^2 + \sigma_\varepsilon^2 \\ \mu_{A,NS} &= \phi y_{t-1} + \beta(y_{t-1}-\phi y_{t-2}) \\ \sigma_{A,NS}^2 &= (1+\beta^2)\sigma_\varepsilon^2 + \psi^2\sigma_e^2 \end{aligned} \tag{3.19}$$

The asymptotic conditional distribution therefore settles down to a linear function of y_{t-1} and y_{t-2} plus a Gaussian random shock. If the process has been running 'forever,' OLS regression on y_{t-1} and y_{t-2} would consistently, efficiently, and unbiasedly estimate the conditional expectation, since the process becomes a manifestation of 3.1 in the limit. Forever is a very long time, however, and when considering finite time, the time dependent nature of the conditional expectation, along with the heteroscedasticity generated by the time varying variance, means that OLS regression will neither consistently nor efficiently nor unbiasedly estimate the conditional expectation of the process beyond the sample period, and this leaves open the possibility that the vox predictor could better capture the *time varying* augmentation of ϕy_{t-1}, as opposed to OLS's time invariant estimation of the augmentation. The process' heteroscedasticity raises the potentiality that Generalised Least Squares [GLS] might be a more appropriate measure to contend with the vox predictor. However, being an exponential process, $\beta^{2(t-1)}$ ought to drive relatively rapid convergence of the conditional variance, as ought $S(t)$. Therefore, it is unclear that the heteroscedasticity correcting advantage of GLS will outweigh the extra inaccuracy introduced by GLS's need to estimate a variance-covariance matrix. Moreover, the fact that GLS would be estimating a time varying conditional expectation function obfuscates its superiority over OLS, since GLS's advantage with processes exhibiting heteroscedasticity *and* a time varying conditional expectation function has not been established. The ambiguous theoretical advantage of either OLS or GLS makes the more established estimator better for comparison, as this paper will then be able to tie into more literature. Actually making this comparison is the task of the simulation in chapter 4, which generates data according to 3.4 and constructs a range of vox predictors that are tested against OLS.

CHAPTER 4

SIMULATION

4.1 Simulation Setup

4.1.1 Parameter List and Explanations

ϕ. This is the parameter ϕ from 3.4.

ψ. This is the parameter ψ from 3.4.

β. This is the parameter β from 3.4.

$\sigma_{\mathbf{e}}^{\mathbf{2}}$. This is the variance of e_t from 3.4.

$\sigma_{\varepsilon}^{\mathbf{2}}$. This is the variance of ε_t from 3.4.

TD. This is the size of the set of total data generated from the simulation. That data which does not form part of the sample is the test data. Sample data temporally precedes test data.

S. This is the size of the set of sample data.

CS. This is the size of the set of total guesses of the conditional expectation function.

μ_{vox}. This is one of the parameters that determines the mean of the distribution from which the component parameters of guesses are drawn.

$\sigma^{\mathbf{2}}_{\text{vox}}$. This is the variance of the distribution from which the component parameters of the guesses are drawn.

SS. This is the size of the set of select estimates whose average determines the vox predictor's prediction.

H. This is the total periods of history over which each guess of the conditional expectation is evaluated.

4.1.2 Platform and Data

The simulation is written in MATLAB, with the code (which is a contribution of this paper) accessible in appendix A.8. It generates data according to 3.4, producing the set of observations $\{y_t\}_{t\in\{1,\ldots,\mathbf{TD}\}}$ which is then split into sample and test data, with the sample data being used to run an OLS regression of y_t on y_{t-1} and y_{t-2} and the test data being used to evaluate the prediction performance of the vox predictor and OLS. For each calibration, the simulation is run one thousand times to produce the output described in subsection 4.1.5, with each calibration being a variation of parameters from the baseline calibrations described in subsection 4.1.4.

4.1.3 Vox Predictor

The simulation tests the performance of a variety of vox predictors, each with the same structure but with different parameters or set in a different context. Following on from definition 3.1, guesses take the form $g(X_t, \hat{\Phi}_i) = \hat{\phi} y_{t-1} + \hat{A}_i(y_{t-1} - \hat{\phi} y_{t-2})$, where $\hat{A}_i$ is an $i.i.d.N(\hat{A} + \mu_{\text{vox}}, \sigma^2_{\text{vox}})$ variable and $\hat{\phi}$ and $\hat{A}$ are recovered by solving the simultaneous equations $\hat{b}_1 = \hat{\phi} + \hat{A}$ and $\hat{b}_2 = \hat{\phi}\hat{A}$.[1] $\hat{b}_1$ and $\hat{b}_2$ are the OLS coefficients of a regression of y_t on y_{t-1} and y_{t-2}, $\hat{\phi}$ is the OLS estimate of ϕ recovered from the simultaneous equations, and $\hat{A}$ is effectively the OLS estimate for the time varying coefficient in 3.10, whose true value is henceforth denoted as A. The guesses are then

[1] This system usually has two solutions; we pick the solution where $\hat{\phi}$ is closest to the order one autoregressive [AR(1)] estimate of ϕ. If the solutions are imaginary, we use the AR(1) estimate of ϕ, given it produces estimates that are very close to the real AR(2) estimates.

ranked just before the vox predictor predicts y_t according to the lowest absolute sum $\sum_{h=1}^{\mathbf{H}} |y_{t-h} - (\hat{\phi} y_{t-1-h} + \hat{A}_i(y_{t-1-h} - \hat{\phi} y_{t-2-h}))|$. The best **SS** guesses are selected and their mean taken to produce the vox predictor's prediction of y_t: $\hat{y}_t^{\text{vox}} = \hat{\phi} y_{t-1} + \bar{A}_t(y_{t-1} - \hat{\phi} y_{t-2})$, where $\bar{A}_t$ is the mean of the draws for A from the selected guesses at time t. Note that this follows from the linearity of g. The time varying nature of $\bar{A}_t$ matches the time varying nature of the true coefficient, in contrast to $\hat{A}$, which does not, and it is this flexibility which gives the vox predictor the potential to beat OLS.

The vox predictor does not incorporate random draws of ϕ both to simplify the simulation and analysis into the one dimensional framework that has been explored in the related literature, and because trying to capture a time invariant parameter with a time variant estimator could introduce unnecessary variance and exacerbate the prediction effort. The normal distribution was chosen as that from which to draw estimates because by shifting the parameters of the normal distribution and dispersing or concentrating draws, the bracketing rate and dispersion can be traded off against each other (even as the true answer drifts higher). Moreover, Galton's crowd of ox guessers had an approximately normal distribution of guesses [7], so using a normal distribution ties back into the literature. Finally, the ranking procedure is taken from Mannes et al. (2014), who used it to great effect in their simulations.

4.1.4 Baseline Calibrations

The simulation is run with multiple variations stemming from two baselines, which are based around switching between a stable and unstable autoregressive process z_t, since this causes a significant shift in the asymptotic conditional distribution of y_t as proved in section 3.3. Choosing parameters that give vox predictors the advantage within these two scenarios makes for an informative simulation, since if a vox predictor cannot outperform OLS with the advantage, then the simulation disconfirms its utility; otherwise, it identifies a niche for future investigation. To this end, using positive parameters for β is amenable to the ranking system described in 4.1.3, since it ensures that high (low) values of y_t will likely be followed by high (low) values, and so the selection of large (small) parameter guesses will indeed produce a select crowd. The value of β also determines the speed with which the conditional distribution of y_t converges to its asymptotic limits. For the stable baseline, The closer β is to 1, the slower $\beta^{2(t-1)}$ will converge to 0, and so the slower the conditional distribution of y_t converges to $N(\mu_{A,S}, \sigma^2_{A,S})$. For the unstable baseline, $S(t)$ grows slowest when $\beta = 1$, similarly ensuring a slow convergence of the conditional distribution. Since OLS cannot capture the out of sample shift in the conditional expectation, ensuring a long transition to the asymptotic distribution will favor the vox predictor, and so we choose $\beta = 0.99$ and $\beta = 1$ for the stable and unstable baselines, respectively. By setting $\sigma_\varepsilon^2 = 1$ for both baselines, we ensure that y_t's own autoregressive component is driven by the standard normal distribution, which is the simplest normal distribution to interpret and allows us to then pick an appropriate σ_e^2 in relation to the standard normal distribution. Motivated values for both ψ and σ_e^2 can be determined given β and σ_ε^2 by maximising the difference between A in the asymptotic conditional distribution and A in the conditional distribution for y_3,[2] thereby ensuring that the time varying conditional distribution varies as much as it can. For the stable case, this difference, proved in appendix A.5, is

$$\frac{\beta^5\psi^2\sigma_e^2\sigma_\varepsilon^2}{\sigma_\varepsilon^4(1-\beta^2)+(2-\beta^4)\psi^2\sigma_e^2\sigma_\varepsilon^2+(1+\beta^2)\psi^4\sigma_e^4} \tag{4.1}$$

Appendix A.6 proves that the values of ψ and σ_e^2 that maximize this difference are

$$\psi = 0.46 \text{ and } \sigma_e^2 = 0.22 \tag{4.2}$$

We therefore choose these values for the stable baseline. The unstable baseline difference is given by 4.3 as proved in appendix A.7.

$$\frac{\beta\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\psi^2\sigma_e^2 S(3)} \tag{4.3}$$

It appears that either $\psi = 0$ or $\sigma_e^2 = 0$ would maximise 4.3, but this is chimerical, since it is clear from our discussion regarding 3.16 that if either of these values were zero, A would never converge to β, making the premise of our calculation of the difference in appendix A.7 invalid. Indeed, if either of these parameters were zero, A would be constant at zero, and 3.4 would collapse to an AR(1) process. There is, in fact, no definable maximum, since the smaller ψ and σ_e^2 are, the smaller does $\psi^2\sigma_e^2 S(t)$ in 3.16 become, and so the smaller does A start on its journey towards β, but when either parameter actually becomes zero, the journey never begins. Therefore, we could set ψ and σ_e^2 to values that are as small as possible, but this would reduce 3.4 to an uninteresting AR(1) process for all intents and purposes. Given that the value of σ_e^2 in the stationary baseline is already

[2] Which is the first dependent variable data point in the sample data.

almost a fifth of σ_{ε}^2 and the keeping of most parameters constant can better identify the effects of varying parameters, we keep σ_e^2 at 0.22. However, ψ is reduced from its more significant value to 0.1, which maintains 3.4 as an interesting process but gives a vox predictor the advantage. The last parameter of the DGP to determine is ϕ. Since we are conditioning on past values of y_t, the parameter ϕ is appended to observed variables and does not have the same impact on the variability of the conditional expectation function or conditional variance as do the parameters that drive the unobserved variables. Therefore ϕ is chosen for the sake of computability and simplicity, being assigned $\phi = 0.5$ in both baselines, which ensures that y_t's own autoregressive component in the DGP is mean reverting.

Turning to the vox predictor itself, for both baselines we take from Mannes et al. (2014) who set $\mathbf{CS} = 50$ in their simulations and recommend as a general rule $\mathbf{SS} = 5$. We set the standard deviation σ_{vox} to the values of 4.1 and 4.3 resultant from our calibrations for the stable and unstable baselines, respectively, which returns the values $\sigma_{\text{vox}}^2 = 0.37$ and $\sigma_{\text{vox}}^2 = 0.99$, again respectively. This is done to ensure a high bracketing rate, which, combined with setting $\mu_{\text{vox}} = 0$ for both baselines, might produce an improvement over OLS, since Mannes et al.'s simulations showed that select crowds improve over the best member of a crowd in the presence of bracketing. Thus, even if the OLS is the best predictor among all the guesses, the crowd's being drawn such that it is composed of nigh-indistinguishable guesses from OLS in conjunction with more dispersed estimates can still lead to improvement. Finally, the fact that A is a moving target suggests a short memory might be optimal, and this is confirmed in untabulated simulation results. We therefore set $\mathbf{H} = 1$ for both baselines.

A limited sample size within a large set of data will grant the vox predictor its niche. A large sample allows OLS to estimate using data from the DGP once it has 'reached' its asymptotic distribution, making it essentially unbeatable. Conversely, a small sample to estimate a large test set leads to greater uncaptured variation of the conditional expectation. we therefore set $\mathbf{S} = 52$ and $\mathbf{TD} = 1502$. This concludes our motivation for the parameters of the two baselines, whose values are summarised in tables 4.1 and 4.2.

4.1.5 Explanation of Simulation Output and Significance Tests

Section 3.1 presents the sample analogue of the mean squared error as a theoretically apt measure of the predictive prowess of the vox predictor and OLS, so we turn to the

TABLE 4.1. Parameter Calibration for the Stable Baseline

$\phi = 0.5$	$\psi = 0.46$	$\beta = 0.99$	$\sigma_e^2 = 0.22$
$\sigma_\varepsilon^2 = 1$	**TD** = 1502	**S** = 52	**CS** = 50
$\mu_{\text{vox}} = 0$	$\sigma_{\text{vox}}^2 = 0.37$	**SS** = 5	**H** = 1

TABLE 4.2. Parameter Calibration for the Unstable Baseline

$\phi = 0.5$	$\psi = 0.1$	$\beta = 1$	$\sigma_e^2 = 0.22$
$\sigma_\varepsilon^2 = 1$	**TD** = 1502	**S** = 52	**CS** = 50
$\mu_{\text{vox}} = 0$	$\sigma_{\text{vox}}^2 = 0.99$	**SS** = 5	**H** = 1

mean squared out of sample prediction error as a measure of performance – that is, we use the errors produced by predicting the data set $\{y_t\}_{t\in\{\mathbf{S}+1,\mathbf{S}+2,...\mathbf{TD}\}}$ using the vox predictor and an OLS regression on the sample data, respectively. Running the simulation only once would produce two mean squared prediction errors whose difference has uncontextualised significance. Running the simulation a thousand times produces one thousand identically and independently distributed realisations of the difference of the mean squared prediction error between the vox predictor and OLS. Thus, the central limit theorem can be applied to test if the mean of these one thousand differences is significantly different from zero with the use of the t statistic - the mean difference in mean squared prediction error divided by the standard error of this difference. In the output tables, we therefore display the mean mean squared prediction error of the vox predictor [MMSPEV] over the one thousand trials, the mean mean squared prediction error of OLS [MMSPEOLS], the mean of their difference [MD], and the t statistic produced by different calibrations of the parameters.

Our theoretical discussion emphasises the importance of sample size – since this is related to the number of time periods since the DGP started – in making OLS a beatable predictor, so varying **S** in our simulations will be our first and foremost result which can confirm or disconfirm our theoretical suspicions. The multiplicative nature of varying each parameter from two baselines makes a judicious choice of tabulated result imperative, so we choose to only additionally vary σ_{vox}^2 and **CS**, as well as running the stable baseline with only β changed to 1 and the unstable baseline with only β changed to 0.99. We vary σ_{vox}^2 since this parameter is the most significant in varying dispersion and bracketing of the vox predictor's guesses and its being varied can therefore be related to Mannes et al's results. We vary **CS** because Mannes et al.'s use of this crowd size is neither empirically nor theoretically justified, thus ensuring that new ground is covered in this simulation. We run the baselines with only β changed so as to isolate the effect of this significant distribution altering change. Untabulated experience with the simulation and our theoretical justifications of all other parameters leads to their exclusion.

4.2 Main Results

Results of the simulation are tabulated below, with all parameters being as in the stated baseline except for the one being varied. The first two tables tabulate the results of the baseline calibrations. Figures are rounded to two decimal places, which may produce minor disparity between the two mean mean squared prediction errors and the mean difference. Note that *negative* values of MD indicate an improvement of the vox predictor over OLS. One, two, or three stars appended to MD signify statistical significance at the 10%, 5%, and 1% level, respectively, for a two tailed t test.

TABLE 4.3. Simulation Results of the Stable Baseline

MMSPEV	MMSPEOLS	MD	t statistic
1.71	1.69	0.01	1.41

TABLE 4.4. Simulation Results of the Unstable Baseline

MMSPEV	MMSPEOLS	MD	t statistic
8.80	8.40	0.40	0.69

TABLE 4.5. Setting $\beta = 1$ Within the Stable Baseline

β	MMSPEV	MMSPEOLS	MD	t statistic
1	7.29	7.49	–0.20	–1.01

TABLE 4.6. Setting $\beta = 0.99$ Within the Unstable Baseline

β	MMSPEV	MMSPEOLS	MD	t statistic
0.99	2.00	1.72	0.28***	9.96

TABLE 4.7. Varying **S** Within the Stable Baseline

S	MMSPEV	MMSPEOLS	MD	t statistic
12	3.07	3.97	–0.90***	–10.60
22	2.28	2.55	–0.27***	–7.37
32	1.97	2.09	–0.13***	–4.62
42	1.83	1.86	–0.03*	–1.79
72	1.59	1.56	0.030***	4.61
102	1.49	1.46	0.03***	7.50
202	1.41	1.37	0.04***	13.95
302	1.39	1.35	0.04***	17.49
402	1.37	1.33	0.04***	19.56
502	1.36	1.32	0.04***	19.83
1002	1.35	1.32	0.04***	18.92

TABLE 4.8. Varying **S** Within the Unstable Baseline

S	MMSPEV	MMSPEOLS	MD	t statistic
12	37.45	46.23	–0.88***	–3.89
22	18.56	19.38	–0.82	–0.73
32	13.30	13.04	0.26	0.33
42	11.73	10.48	1.25	1.47
72	6.46	5.23	1.23***	3.38
102	4.91	4.52	0.39	1.61
202	2.35	1.96	0.39***	8.69
302	2.017	1.72	0.30***	8.49
402	2.09	1.63	0.46***	2.97
502	1.78	1.49	0.29***	12.69
1002	1.61	1.35	0.26***	21.63

TABLE 4.9. Varying σ^2_{vox} Within the Stable Baseline

σ^2_{vox}	MMSPEV	MMSPEOLS	MD	t statistic
0.01	1.72	1.76	–0.04***	–5.97
0.05	1.70	1.74	–0.04***	–5.16
0.1	1.69	1.73	–0.03***	–4.57
0.2	1.73	1.76	–0.03***	–3.08
0.3	1.74	1.74	0.01	0.79
0.4	1.74	1.71	0.03**	2.26
0.5	1.78	1.71	0.07***	4.55
1	2.12	1.76	0.37***	12.34
2	3.00	1.77	1.23***	16.32
5	10.05	1.72	8.34***	19.19
10	38.60	1.73	36.87***	22.02

TABLE 4.10. Varying σ^2_{vox} Within the Unstable Baseline

σ^2_{vox}	MMSPEV	MMSPEOLS	MD	t statistic
0.01	8.00	8.67	–0.66***	–4.66
0.05	6.67	7.10	–0.04***	–3.82
0.1	7.43	7.93	–0.50***	–3.43
0.2	6.51	6.69	–0.18*	–1.65
0.3	7.62	8.24	–0.63***	–3.32
0.4	7.58	7.83	–0.24	–1.44
0.5	8.99	9.29	–0.29	–0.71
1	8.25	7.22	1.03**	2.29
2	15.10	7.79	7.30***	4.05
5	35.35	6.51	27.84***	8.48
10	108.54	6.68	101.85***	12.45

TABLE 4.11. Varying **CS** Within the Stable Baseline

CS	MMSPEV	MMSPEOLS	MD	t statistic
10	1.74	1.74	0	0.08
20	1.72	1.72	0	0.08
30	1.69	1.68	0.01	1.22
40	1.74	1.75	–0.01	–0.88
70	1.72	1.70	0.016	1.33
100	1.76	1.75	0.01	0.58
200	1.74	1.72	0.016	1.47
300	1.74	1.74	–0	–0.10
400	1.74	1.75	–0.01	–0.55
500	1.72	1.70	0.02	1.43
1000	1.76	1.78	–0.02	–1.00

Note: '–0' refers to the fact that MD is a small *negative* number rounded to zero

TABLE 4.12. Varying **CS** Within the Unstable Baseline

CS	MMSPEV	MMSPEOLS	MD	t statistic
10	7.66	6.73	0.93**	2.12
20	8.57	8.31	0.26	0.60
30	9.72	7.99	1.74***	2.63
40	9.18	9.62	1.55**	2.50
70	8.56	7.79	0.77**	2.23
100	9.12	8.61	0.51	0.94
200	7.77	7.34	0.44	1.16
300	8.27	7.30	0.97	1.49
400	8.36	8.15	0.21	0.44
500	8.39	7.62	0.77*	1.94
1000	8.96	7.90	1.06*	1.94

4.3 Discussion, Extensions, and Supplementary Results

Varying **S** in both baselines (tables 4.7 and 4.8) confirms our theoretical suspicions that limited sample OLS is beaten by the more flexible vox predictor, though OLS improves in larger samples where the conditional distribution approaches its limit already in the sample and slows its pace in the test data. This is consistent with the vox predictor's ability to beat OLS in both baselines with a reduced σ^2_{vox} (tables 4.9 and 4.10), since Mannes et al.'s simulations show that bracketing leads select crowds to a significant improvement over the best member of a crowd, whereas dispersion militates in favour of the best member. The performance of both is expressed as an improvement over the average judge, however. Hence, if the average judge can be improved via a reduction in dispersion without sacrificing bracketing, then a select crowd strategy ought to become more accurate in absolute terms and relative to a potential best (or almost best) member strategy - read OLS. But if A has already almost reached its asymptotic limit, then reducing σ^2_{vox} will achieve this, as A will not move far from it's OLS estimate and even a tight distribution of guesses around the OLS estimate will bracket the true value of A throughout the test data, whilst also reducing dispersion and improving the average guess. Reducing σ^2_{vox} as sample size increases ought therefore to improve the performance of the vox predictor against OLS up to a point, which is what table 4.13 illustrates.

TABLE 4.13. The Effect of Reduced σ^2_{ε} in Larger Sample

Baseline	**S**	σ^2_{vox}	MMSPEV	MMSPEOLS	MD	t statistic
Stable	102	0.0001	1.49	1.49	–0*	–1.69
Unstable	102	0.0001	3.76	3.80	–0.04	–1.51

Note: '–0' refers to the fact that MD is a small *negative* number rounded to 0

A greatly reduced σ^2_{vox} within a sample size of 102 turns statistically significant results in favour of OLS (as shown in tables 4.7 and 4.8) into results in favour of the vox predictor, as shown by the t statistics. It could be contended that with $\sigma^2_{\text{vox}} = 0.0001$, the vox predictor is tantamount to OLS. However, if our theory were not applicable, we would expect even a small σ^2_{vox} to only add unnecessary variance and exacerbate the mean squared prediction error of the vox predictor vis-á-vis OLS, at best leading neither OLS nor the vox predictor to be consistently better than the other. Yet, repetitions of the calibrations in table 4.13 give consistently negative t statistics as shown in table 4.14.

TABLE 4.14. Repeated Trials of Calibration in Table 4.13

Baseline	S	σ^2_{vox}	Trial No.	MMSPEV	MMSPEOLS	MD	t statistic
Stable	102	0.0001	1	1.48	1.48	–0	–1.59
			2	1.46	1.46	–0	–1.22
			3	1.46	1.46	–0*	–1.73
			4	1.47	1.47	–0*	–1.96
			5	1.46	1.46	–0	–1.40
Unstable	102	0.0001	1	3.55	3.55	–0	–1.42
			2	3.73	3.75	–0.02	–1.00
			3	3.68	3.72	–0.04	–1.41
			4	4.50	4.52	–0.02	–1.00
			5	3.49	3.49	–0	–0.99

Note: again '–0' refers to the fact that MD is a small *negative* number rounded to 0

The results concerning **CS** proffer a greater puzzle, since table 4.11 implies crowd size is largely unrelated to the vox predictor's performance within the stable baseline, whereas table 4.12 implies mid range **CS** of 100 to 400 could be better than more extreme values within an unstable baseline. The advantage of larger **CS** is that A has more potential to be precisely tracked, since whatever its value at t, there will likely be more guesses closer to the true value. The disadvantage is that a large random error will lead to the selection of only largely erroneous guesses in the next period, since five guesses can be more readily found in the direction of the error with a large **CS**, and better guesses will be excluded. This dichotomy appears to produce a complex geometry of optimal **CS** in the parameter space, which is a furtive area for further study, especially vis-á-vis the effects of switching between a stable or unstable series z_t. The importance of this switch is further emphasised by tables 4.5 and 4.6, which show a significant jump in the level of mean mean squared prediction error going from $\beta = 0.99$ to $\beta = 1$ – an unsurprising fact given the unobserved component of the DGP goes from having limited to having unlimitedly increasing variance.

The direct practical implications of our main results are limited inasmuch as it is rare to use a sample of around 100 or less to predict a remaining 1400 data points in DGPs with unobserved autoregressive variables, since there is no reason not to update sample data as time goes on. If vox predictors are significantly better at predicting data towards the end of the test set, but worse at the beginning, this would undermine the practicability of vox predictors for the DGP we have investigated. Consequently, we repeat the analysis with **TD** being only 10 greater than **S** in table 4.15, running five trials of each calibration in answer to the greater variance that mean squared prediction errors based on only 10 data points introduce to our results.

TABLE 4.15. Prediction of Outcomes Immediately Subsequent Small Samples

Baseline	**S**	**TD**	σ^2_{vox}	Trial No.	MMSPEV	MMSPEOLS	MD	t statistic
Stable	52	62	0.0001	1	1.38	1.39	–0.01***	–3.35
				2	1.37	1.38	–0.01***	–3.04
				3	1.36	1.37	–0.01***	–3.67
				4	1.33	1.35	–0.01***	–3.41
				5	1.37	1.38	–0.01***	–3.11
	102	112	0.0001	1	1.33	1.33	–0	–0.77
				2	1.31	1.31	–0*	–1.78
				3	1.32	1.32	–0	–1.43
				4	1.29	1.29	–0	–0.98
				5	1.35	1.36	–0	–0.89
Unstable	52	62	0.0001	1	1.34	1.35	–0.01***	–3.43
				2	1.39	1.40	–0.01***	–3.81
				3	1.34	1.35	–0.01***	–3.27
				4	1.33	1.34	–0**	–2.11
				5	1.35	1.36	–0.01*	–1.65
	102	112	0.0001	1	1.31	1.31	–0*	–1.91
				2	1.34	1.34	–0	–1.40
				3	1.34	1.34	–0	–1.34
				4	1.33	1.33	–0	–0.29
				5	1.32	1.33	–0	–1.42

Note: '–0' still refers to the fact that MD is a small *negative* number rounded to 0

The consistent, though very small and not always significant, outperformance of OLS by the vox predictors in table 4.15 indicate their ability to adapt to the time varying conditional expectation better than OLS even for those y_t immediately following the OLS sample, though the breadth of this ability depends on our results' robustness to less ideal parameters driving the DGP; this along with a general investigation of the effects of varying the parameters of the simulation is an extension in itself. An investigation of our results' robustness can also take the form of changing the metric of evaluation. In particular, using correlation instead of mean squared error would allow the results to be tied back into Hogarth's (1978) work presented in 2.4. Nonetheless, with further enhancement to increase the significance of the outperformance – such as the use of a genetic algorithm to determine parameters, which has already proven successful in the wisdom of crowds context in the work of Hill and Ready-Campbell (2014) – vox predictors may find fruitful application in predicting young economic time series, perhaps those related to new countries or countries having undergone recent dramatic structural change. Moving beyond our particular analysis of the DGP in 3.4, the toolkit of the conditional normality theorem may prove useful in dissecting the DGP in novel ways, such as by repeating the analysis with nonzero initial values, non-normal distributions of the errors, or discovering the distribution of y_t given more of its lags. A re-examination of the utility of the vox predictor in these lights may widen the field of its potential application, especially if non-convergent conditional distributions are discovered. Section 3.3 mentioned an alternative comparison of the vox predictor against GLS. This and other comparisons incorporating yet more econometric estimators, such as maximum likelihood, might more definitively determine the utility of the vox predictor. Study could also be conducted in relation to alternative DGPs with non-convergent time varying conditional distributions – perhaps a time series generated by the product of autoregressive processes as opposed to their linear combination. The form of the vox predictor need not be taken as final, either. Instead of having guesses be distinguished by distinct draws of parameters within a presupposed function, a select crowd of guesses of different functions incorporating disparate variables may prove effective, especially if the time series process being predicted is unknown. The restrictive nature of studying only the process in 3.4 makes the fact that we found theoretically consistent results the more important feature of this paper, since intelligibility bodes well for future work and advocates for the vox predictor to be considered along with other econometric predictors more advanced than OLS.

Chapter 5

Conclusion

This paper has brought together aspects of psychology, statistics, politics, and economics, and condensed them into an exploratory literature review that develops pre-existing mathematical insights into the wisdom of crowds and contributes its own in the form of the bracketing theorem. Theory regarding ordinary least squares is then exposed to transpose these mathematical insights into the time series context, begetting a new class of estimator – the vox predictor – whose aim is to approximate the conditional expectation function of a process. We proffer the time series process with an unobserved autoregressive variable in 3.4 as one where the vox predictor may outperform OLS and uncover its conditional distribution through the use of this paper's other theorem: the conditional normality theorem. The conditional distribution has two asymptotic limits depending on the stability or otherwise of the unobserved autoregressive variable, and the time-variant but convergent nature of the conditional distribution motivates a simulation that generates observations according to 3.4 and predominantly tests the limited sample prediction performance of a series of vox predictors against OLS, since it is identified that it is limited samples where the yet time variant conditional expectation function requires the flexibility in estimation that the vox predictor can offer. Within two baselines of the data generation process, one with stable and one with unstable z_t, the simulation varies parameters of the vox predictor and the size of sample and test data sets; its results support theoretical suspicions that the vox predictor can out-predict OLS in limited samples, but OLS reasserts itself in larger samples where the DGP settles down to a stochastic and homoscedastic unchanging linear function of the observed variables. Difficulty is encountered in interpreting simulation output resultant from varying one of the parameters that constitutes the vox predictor – **CS** or ‘crowd size’ – since this implies an inconsistent geometry of optimal values across the two baselines, but this lends itself to one fruitful extension among many. The MATLAB code provided in appendix A.8 allows for further work to carry out more extensive mapping of the effects of changing the parameters of the simulation, and this paper’s decomposition of 3.4 may prove useful in analysis of the vox predictor as applied to the process 3.4 conditioned on more variables, or as applied to other time series processes altogether. The specificity of the process studied and our control of its parameters means this paper must not be taken as presenting a finalised new procedure; instead the significance of this paper lies more in the fact that it shows the principles behind the wisdom of crowds can be harnessed to produce an econometric predictor that beats OLS in limited samples even for a time series that quickly converges to OLS’s ideal process. Testing the axioms of the wisdom of crowds in a time series context proved successful and consistent with extant theory. It may do so again. This paper ought therefore be taken as outlining the first steps towards collective wisdom in the world of metrics.

BIBLIOGRAPHY

[1] **Andrilli, Stephen and David Hecker**, *Elementary Linear Algebra*, 4th ed., London: Elsevier, 2010.

[2] **Angrist, Joshua D. and Jörn-Seffen Pischke**, *Mostly Harmless Econometrics*, Princeton: Princeton University Press, 2008.

[3] **Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz**, "Results from a Dozen Years of Election Futures Markets Research," in Charles R. Plott and Vernon L. Smith, eds., *Handbook of Experimental Economics Results, Volume 1.*, Amsterdam: Elsevier, 2008, chapter 80, pp. 742–751.

[4] **Budescu, David V., Adrian K. Rantilla, Hsiu-Ting Yu, and Tzur M. Karelitz**, "The effects of asymmetry among advisors on the aggregation of their opinions," *Organizational Behavior and Human Decision Processes*, 2003, *90* (1), 178–194.

[5] **Chen, Hailiang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang**, "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media," *The Review of Financial Studies*, 2014, *27* (5), 1367–1403.

[6] **Forsythe, Robert, George R. Neumann, Forrest Nelson, and Jack Wright**, "Anatomy of an Experimental Political Stock Market," *American Economic Review*, 1992, *82* (5), 1142–1161.

[7] **Galton, Francis**, "Vox Populi," *Nature*, 1907, *75* (1949), 450–451.

[8] **Ghiselli, Edwin E.**, *Theory of psychological measurement*, New York: McGraw-Hill, 1964.

[9] **Hill, Shawndra and Noah Ready-Campbell**, "Expert Stock Picker: The Wisdom of (Experts in) Crowds," *Journal of Electronic Commerce*, 2011, *15* (3), 73–102.

[10] **Hogarth, Robin M.**, "A Note on Aggregating Opinions," *Organizational Behavior and Human Performance*, 1978, *21*, 40–46.

[11] **Janis, Irving L.**, *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*, University of Michigan: Houghton Mifflin Company, 1972.

[12] **Lindeberg, Jarl W.**, "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 1922, *15* (1), 211–225.

[13] **Mackay, Charles**, *Extraordinary Popular Delusions and the Madness of Crowds*, London: Richard Bently, 1841.

[14] **Mannes, Albert E., Richard P. Larrick, and Jack B. Soll**, "The Wisdom of Select Crowds," *Journal of Personality and Social Psychology*, 2014, *107* (2), 276–299.

[15] **Milgram, Stanley, Leonard Bickman, and Lawrence Berkowitz**, "NOTE ON THE DRAWING POWER OF CROWDS OF DIFFERENT SIZES," *Journal of Personality and Social Psychology*, 1969, *13* (2), 79–82.

[16] **Nofer, Michael and Oliver Hinz**, "Are crowds on the internet wiser than experts? The case of a stock prediction community," *Journal of Business Economics*, 2014, *84* (3), 303–338.

[17] **Orwell, George**, *1984*, London: Secker and Warburg, 1949.

[18] **Parducci, Allen and Louise M. Marshall**, "Assimilation vs. Contrast in the Anchoring of Perceptual Judgments of Weight," *Journal of Experimental Psychology*, 1962, *63* (5), 426–437.

[19] **Pischke, Jörn-Steffen and Joshua D. Angrist**, *Mastering Metrics*, Princeton: Princeton University Press, 2014.

[20] **Ray, Russ**, "Prediction Markets and the Financial "Wisdom of Crowds"," *Journal of Behavioral Finance*, 2006, *7* (1), 2–4.

[21] **Sharpe, William F.**, "The Sharpe Ratio," *The Journal of Portfolio Management*, 1994, *21* (1), 49–58.

[22] **Sherif, Muzafer and Carl I. Hovland**, *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*, New Haven: Yale University Press, 1961.

[23] **Simmons, Joseph P., Leif D. Nelson, Jeff Galak, and Shane Frederick**, "Intuitive Biases in Choice vs. Estimation: Implications for the Wisdom of Crowds," *Journal of Consumer Research*, 2011, *38*, 1–15.

[24] **Soll, Jack B. and Albert E. Mannes**, "The effects of asymmetry among advisors on the aggregation of their opinions," *Judgmental aggregation strategies depend on whether the self is involved*, 2011, *27*, 81–102.

[25] **Strang, Gilbert**, *Calculus*, 2nd ed., Wellesley: Wellesley-Cambridge Press, 1991.

[26] **Surowiecki, James**, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, New York: Doubleday, 2004.

[27] **Wolfers, Justin and Eric Zitzewitz**, "Prediction Markets," *Journal of Economic Perspectives*, 2004, *18* (2), 107–126.

APPENDIX A

Appendix

A.1 Proof of the consistency of OLS

The OLS estimator is

$$\hat{\Phi}_{OLS} = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t y_t = \Phi + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t \varepsilon_t$$

ε_t's independence and the finititude of the second moment guarantee that $X_t\varepsilon_t$ is a martingale difference. Given the stated regularity conditions, the martingale central limit theorem applies and ensures that $\frac{1}{T}\sum_{t=1}^{T} X_t\varepsilon_t \xrightarrow{p} 0$, implying $\hat{\Phi}_{OLS} \xrightarrow{p} \Phi$.

A.2 Proof that the Conditional Expectation Minimises the Expected Squared Error

$$\begin{aligned}
\mathrm{E}\left[(y_t - f(X_t))^2\right] &= \mathrm{E}\left[\mathrm{E}\left[(y_t - f(X_t))^2 \,|\, X_t\right]\right] = \mathrm{E}\left[\mathrm{E}\left[(y_t - \mathrm{E}[y_t|X_t] + \mathrm{E}[y_t|X_t] - f(x_t))^2\right]\right] \\
&= \mathrm{E}[\mathrm{E}[(y_t - \mathrm{E}[y_t|X_t])^2 + 2(y_t - \mathrm{E}[y_t|X_t])(\mathrm{E}[y_t|X_t] - f(X_t)) + \\
&\quad (\mathrm{E}[y_t|X_t] - f(X_t))^2|X_t]] \\
&= \mathrm{E}[(y_t - \mathrm{E}[y_t|X_t])^2] + 2\mathrm{E}[\mathrm{E}[y_t\mathrm{E}[y_t|X_t] - y_t f(X_t) - \mathrm{E}[y_t|X_t]^2 + \\
&\quad \mathrm{E}[y_t|X_t]f(X_t)|X_t]] + \mathrm{E}[(\mathrm{E}[y_t|X_t] - f(X_t))^2] \\
&= \mathrm{E}[(y_t - \mathrm{E}[y_t|X_t])^2] + 2\mathrm{E}[\mathrm{E}[y_t|X_t]^2 - \mathrm{E}[y_t|X_t]^2 + \mathrm{E}[y_t|X_t]f(X_t) - \\
&\quad \mathrm{E}[y_t|X_t]f(X_t)] + \mathrm{E}[(\mathrm{E}[y_t|X_t] - f(X_t))^2]
\end{aligned}$$

$$= \mathrm{E}[(y_t - \mathrm{E}[y_t \,|\, X_t])^2] + \mathrm{E}[(\mathrm{E}[y_t \,|\, X_t] - f(X_t))^2]$$

This expression is minimised when $f(X_t) = \mathrm{E}[y_t \,|\, X_t]$ since the rightmost non-negative term is reduced to zero. Therefore, the conditional expectation function is the function of X_t that minimizes expected squared error.

A.3 Proof that OLS Consistently Estimates the Conditional Expectation Function with Linearly Related Unobserved Variables

Suppose that Z_t is related to X_t according to

$$Z_t = BX_t + e_t$$

Where B is a conformable matrix of coefficients and e_t is a vector of independent mean zero shocks with variance-covariance matrix Σ. An OLS regression of y_t from 3.2 on X_t results in

$$\begin{aligned}\hat{\Phi}_{OLS} &= \Phi + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t Z_t' \Psi + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t \varepsilon_t \\ &= \Phi + B'\Psi + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t e_t' + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X_t \varepsilon_t\end{aligned}$$

As with A.1, the finititude of $E(||X_t e_t'||^2)$, $E(||X_t \varepsilon_t||^2)$, and $\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$ ensure consistency along with the Lindeberg condition and convergence in probability of $\frac{\sigma^2}{T}\sum_{t=1}^{T} \mathrm{E}[X_t X_t' \,|\, X_{t-1}\varepsilon_{t-1}, \ldots, X_1\varepsilon_1]$ and $\frac{\mathrm{tr}(\Sigma)}{T}\sum_{t=1}^{T} \mathrm{E}[X_t X_t' \,|\, X_{t-1}e_{t-1}', \ldots, X_1 e_1']$ to positive definite matrices; these conditions ensure via the martingale central limit theorem that all terms except $\Phi + B'\Psi$ converge in probability to zero. The OLS stand in for the conditional expectation function therefore becomes

$$\hat{\mathrm{E}}_{y|X} = X_t'\Phi + X_t'B'\Psi$$

But $X_t'B'$ is precisely $\mathrm{E}[Z_t' \,|\, X_t]$ and therefore we see from 3.3 that OLS consistently estimates the conditional expectation function.

A.4 Conditional Normality Theorem for Non-Zero Mean Variables

Let a and b be independently normally distributed random variables with means μ_a and μ_b and variances σ_a^2 and σ_b^2, respectively. It follows that $a = a^* + \mu_a$ and $b = b^* + \mu_b$, where $a^* \sim N(0, \sigma_a^2)$ and $b^* \sim N(0, \sigma_b^2)$. if it is given that $a + b = x$, then it follows that $a^* + b^* = x - \mu_a - \mu_b = y$. The conditional normality theorem is therefore applicable, yielding $a^* | x \sim N(\sigma_a^2 y/(\sigma_a^2 + \sigma_b^2), \sigma_a^2 \sigma_b^2/(\sigma_a^2 + \sigma_b^2))$. Using the fact that $a = a^* + \mu_a$ we conclude that $a | x$ is normally distributed with mean $\sigma_a^2 y/(\sigma_a^2 + \sigma_b^2) + \mu_a$ and variance $\sigma_a^2 \sigma_b^2/(\sigma_a^2 + \sigma_b^2)$.

A.5 Total Change in Time Varying Coefficient of the Conditional Expectation – Stable Baseline

From 3.15 A at $t = 3$ is given by

$$\frac{\beta}{\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^4)} + 1} = \frac{\beta}{\frac{\sigma_\varepsilon^2(1-\beta^2)}{\psi^2\sigma_e^2(1-\beta^2)(1+\beta^2)} + 1} = \frac{\beta\psi^2\sigma_e^2(1+\beta^2)}{\sigma_\varepsilon^2 + \psi^2\sigma_e^2(1+\beta^2)}$$

Therefore, using 3.19, the difference between A in the limit and A at $t = 3$ is

$$\frac{\beta\psi^2\sigma_e^2}{\sigma_\varepsilon^2(1-\beta^2) + \psi^2\sigma_e^2} - \frac{\beta\psi^2\sigma_e^2(1+\beta^2)}{\sigma_\varepsilon^2 + \psi^2\sigma_e^2(1+\beta^2)} =$$
$$\frac{\beta\psi^2\sigma_e^2(\sigma_\varepsilon^2 + \psi^2\sigma_e^2(1+\beta^2)) - \beta\psi^2\sigma_e^2(1+\beta^2)(\sigma_\varepsilon^2(1-\beta^2) + \psi^2\sigma_e^2)}{(\sigma_\varepsilon^2(1-\beta^2) + \psi^2\sigma_e^2)(\sigma_\varepsilon^2 + \psi^2\sigma_e^2(1+\beta^2))} =$$
$$\frac{\beta\psi^2\sigma_e^2\sigma_\varepsilon^2 - \beta\psi^2\sigma_e^2\sigma_\varepsilon^2(1-\beta^2)(1+\beta^2)}{\sigma_\varepsilon^4(1-\beta^2) + \psi^2\sigma_e^2\sigma_\varepsilon^2(1+\beta^2)(1-\beta^2) + \psi^2\sigma_e^2\sigma_\varepsilon^2\psi^4\sigma_e^4(1+\beta^2)} =$$
$$\frac{\beta^5\psi^2\sigma_e^2\sigma_\varepsilon^2}{\sigma_\varepsilon^4(1-\beta^2) + (2-\beta^4)\psi^2\sigma_e^2\sigma_\varepsilon^2 + (1+\beta^2)\psi^4\sigma_e^4}$$

A.6 Maximisers of 4.1

We use the quotient rule to determine the numerator of the derivative of 4.1 with respect to ψ and then set it equal to 0, which returns us one first order condition, since the denominator is non-zero given our stipulations of σ_ε^2 and β. Thus, we obtain the numerator

$[\sigma_\varepsilon^4(1-\beta^2)+(2-\beta^4)\psi^2\sigma_e^2\sigma_\varepsilon^2+(1+\beta^2)\psi^4\sigma_e^4]\cdot 2\beta^5\psi\sigma_e^2\sigma_\varepsilon^2-\beta^5\psi^2\sigma_e^2\sigma_\varepsilon^2\cdot[2(2-\beta^4)\psi\sigma_e^2\sigma_\varepsilon^2+$
$4(1+\beta^2)\psi^3\sigma_e^4]=$
$2\beta^5\psi\sigma_e^2\sigma_\varepsilon^2[\sigma_\varepsilon^4(1-\beta^2)+(2-\beta^4)\psi^2\sigma_e^2\sigma_\varepsilon^2+(1+\beta^2)\psi^4\sigma_e^2-(2-\beta^4)\psi^2\sigma_e^2\sigma_\varepsilon^2-2(1+\beta^2)\psi^4\sigma_e^4]=$
$2\beta^5\psi\sigma_e^2\sigma_\varepsilon^2[\sigma_\varepsilon^4(1-\beta^2)-(1+\beta^2)\psi^4\sigma_e^4=0\Longrightarrow$
$\psi^4=\dfrac{\sigma_\varepsilon^4(1-\beta^2)}{\sigma_e^2(1+\beta^2)}$ or $\psi=0$

Notice that ψ and σ_e are interchangeable in 4.1, and thus the first order conditions also yield

$$\sigma_e^4=\frac{\sigma_\varepsilon^4(1-\beta^2)}{\psi^2(1+\beta^2)} \text{ or } \sigma_e=0$$

To determine the values that return a maximum, we notice first that our stipulations of β and σ_ε^2 guarantee that 4.1 is always greater than or equal to zero. Moreover, if either $\psi=0$ or $\sigma_e=0$ then 4.1 is zero, and therefore any point with the zero roots can be excluded as a maximum, since it will return the global minimum. Let us now consider the point formed by taking the positive roots. We get

$$\begin{aligned}\psi^4\sigma_e^2=\psi^4\sqrt{\frac{\sigma_\varepsilon^4(1-\beta^2)}{\psi^2(1+\beta^2)}}&=\frac{\sigma_\varepsilon^4(1-\beta^2)}{1+\beta^2}\\ \Longrightarrow\psi^3&=\sqrt{\frac{\sigma_\varepsilon^4(1-\beta^2)}{1+\beta^2}}\\ \Longrightarrow\psi&=\left(\frac{\sigma_\varepsilon^4(1-\beta^2)}{1+\beta^2}\right)^{1/6}\\ \Longrightarrow\sigma_e^2&=\left(\frac{\sigma_\varepsilon^4(1-\beta^2)}{1+\beta^2}\right)^{1/3}\end{aligned}$$

This is the only stationary point in the positive ψ-σ_e quadrant, and dividing the numerator and denominator of 4.1 by $\psi^2\sigma_e^2$ reveals that as either ψ or σ_e^2 tends to infinity, 4.1 tends to zero. Elsewhere, 4.1 is positive, so this stationary point must be a local maximum. 4.1 is also symmetric in both ψ and σ_e due to the indices appended to them, and this implies that any combination of positive and negative roots returns the same value. Therefore, we have found the global maxima and we choose positive values (a given for the standard deviation). Inputting our stipulated β and σ_ε^2 gives

$$\psi=0.4645448596 \text{ and } \sigma_e^2=0.2158019266$$

A.7 Total Change in Time Varying Coefficient of the Conditional Expectation – Unstable Baseline

From 3.11 and 3.19, The difference is

$$\beta - \frac{\beta\psi^2\sigma_e^2 S(3)}{\sigma_\varepsilon^2 + \psi^2\sigma_e^2 S(3)} = \frac{\beta(\sigma_\varepsilon^2 + \psi^2\sigma_e^2 S(3)) - \beta\psi^2\sigma_e^2 S(3)}{\sigma_\varepsilon^2 + \psi^2\sigma_e^2 S(3)} = \frac{\beta\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \psi^2\sigma_e^2 S(3)}$$

A.8 MATLAB Code

The MATLAB code for the simulation is presented on the next page.

```
% First I set the parameters of the simulation.
NumbTrials = 1000 ;

final = zeros(NumbTrials, 2) ;
for t = 1:NumbTrials

phi = 0.5;
psi = 0.1;
beta = 0.99;
sigma2_e = 0.22;
sigma2_varepsilon = 1;
Tdata = 62;
sample = 52 ;
crowdsize = 50 ;
mu_vox = 0 ;
sigma2_vox = 0.0001 ;
selectsize = 5 ;
history = 1 ;

% Now I generate the Series z_t

a = 0;
Z = zeros(Tdata,1);
shocks_e = normrnd(0,sigma2_e,[Tdata,1]);
for i = 1:Tdata
    a = shocks_e(i) + beta*a ;
    Z(i) = a ;
end
clearvars a i

% Now I generate the integrated z_t process

b = 0 ;
Zcum = zeros(Tdata,1);
for m = 1:Tdata
    b = Z(m) + phi*b ;
    Zcum(m) = b ;
end
clearvars b m

% Here I generate the part of the process driven by the varepsilon
 shocks

c = 0 ;
shocks_varepsilon = normrnd(0,sigma2_varepsilon,[Tdata,1]) ;
AR = zeros(Tdata,1) ;
for s = 1:Tdata
    c = shocks_varepsilon(s) + phi*c ;
    AR(s) = c ;
end
clearvars c s
```

```
% The Sum of AR and Zcum give us the data series, which we first split
 into
% our estimation data, which yields our OLS estimate

Y = AR + Zcum ;
Yesdep = Y(3:sample) ;
Yesdes = [Y(2:(sample - 1)) Y(1:(sample - 2))] ;

beta = (inv(Yesdes'*Yesdes))*Yesdes'*Yesdep ;
if (beta(1))^2 + 4*beta(2) >= 0
phihat1 = 0.5*(beta(1) + sqrt((beta(1))^2 + 4*beta(2)));
else
phihat1 =
 inv(Y(1:sample-1)'*Y(1:sample-1))*Y(1:sample-1)'*Y(2:sample) ;
end
A1 = beta(1) - phihat1 ;
phihat2 = 0.5*(beta(1) - sqrt((beta(1))^2 + 4*beta(2)));
A2 = beta(1) - phihat2 ;

% Now I turn to the estimator with step 1

Yexp = [Y((sample-history):(Tdata-2)) (Y((sample - history):(Tdata -
 2)) - phihat1*Y((sample-history-1):(Tdata-3)))] ;
Afull = [phihat1*ones(1,crowdsize); A1*ones(1,crowdsize) +
 normrnd(mu_vox,sigma2_vox,[1,crowdsize])] ;

% Step 2

Yhatvox = Yexp*Afull ;

% Step 3

Ymatdep = zeros(Tdata-sample-1+history,crowdsize) ;
for i = 1:Tdata-sample-1+history
    a = Y(sample - history + i)*ones(1,crowdsize) ;
    Ymatdep(i,:) = a ;
end
clearvars i a

e_hatvox = Yhatvox - Ymatdep ;

e_voxhis = zeros(Tdata-sample,crowdsize) ;
for u = 1:Tdata-sample
    a = abs(e_hatvox(u:u+history-1,:));
    e_voxhis(u,:) = sum(a);
end
clearvars u a

% Step 4

e_sorted = zeros(Tdata-sample,selectsize) ;
for i = 1:Tdata-sample
    a = e_voxhis(i,:) ;
    [~,idx] = sort(a) ;
    b = e_hatvox(i,idx) ;
    e_sorted(i,:) = b(1:selectsize) ;
end
clearvars a i b

% Step 5

Ymatcor = zeros(Tdata-sample,selectsize) ;
for i = 1:Tdata - sample
    Ymatcor(i,:) = Y(sample - 2 + i)*ones(1,selectsize) ;
end
clearvars i

select = e_sorted + Ymatdep(1:Tdata-sample,1:selectsize) -
 phihat1*Ymatcor;

% Step 6

Ymatcordiv = zeros(Tdata-sample, selectsize) ;
for i = 1:Tdata-sample
    Ymatcordiv(i,:) = (Y(sample - 2 + i) - phihat1*Y(sample - 3 +
 i))*ones(1,selectsize) ;
end
clearvars i

select2 = select./Ymatcordiv ;

% Step 7

vox_metrica = mean(select2') ;

% Step 8 and 9. I introduce OLS and calculate the out of sample
 prediction
% errors.

Yhatfinal = diag([phihat1*Y(sample:(Tdata-1)) (Y(sample:(Tdata-1))-
phihat1*Y((sample-1):(Tdata-2)))]*[ones(1,Tdata-sample) ;
 vox_metrica]) ;

e_finalvox = Y((sample+1):Tdata) - Yhatfinal ;
e_OLS = Y((sample+1):Tdata) - [Y(sample:(Tdata-1)) Y((sample-1):
(Tdata-2))]*beta ;

metric = [(1/(Tdata-sample))*e_finalvox'*e_finalvox (1/(Tdata-
sample))*e_OLS'*e_OLS] ;

clearvars -except metric final NumbTrials t

final(t,:) = metric ;
end

% Now I turn to tabulatable results.

Dif = final(:,1) - final(:,2) ;
```

```
Stat = mean(Dif) ;
SE = std(Dif)/sqrt(NumbTrials) ;
t_stat = Stat/SE ;

if abs(t_stat) > 2.576
    disp('1 percent')
elseif abs(t_stat) > 1.96
    disp('5 percent')
elseif abs(t_stat) > 1.65
    disp('10 percent')
else
    disp('not significant')
end

A = [mean(final(:,1)) mean(final(:,2)) Stat t_stat];

round(A*100)/100
```

```
1 percent

ans =

    1.3500    1.3500         0   -2.6900
```

Published with MATLAB® R2018a