Comments on "Prediction Using Several Macroeconomic Models" by Gianni Amisano and John Geweke

James Mitchell[*,†]

[*]Department of Economics, University of Leicester
[†]National Institute of Economic and Social Research, London

5 May 2012

Amisano and Geweke (AG)

- AG emphasise densities (move beyond RMSE loss) and combination
- AG's empirical conclusions
  1. Full Bayesian (FB) predictive densities beat 'plug-in' densities (which use the posterior mode)
  2. Pooling either using equal or optimised weights (but not BMA) is better than any individual FB density

Amisano and Geweke (AG)

- AG emphasise densities (move beyond RMSE loss) and combination
- AG's empirical conclusions

  1. Full Bayesian (FB) predictive densities beat 'plug-in' densities (which use the posterior mode)
  2. Pooling either using equal or optimised weights (but not BMA) is better than any individual FB density

     - Equal weights is better than real-time optimised weights
     - But suspect the equal weighted combination is still poorly calibrated on the basis of *pits* (absolute) density forecast evaluation tests

My comments

1. Things to do with forecast densities
2. Things to do with combination and the model space

Forecast Densities

- AG emphasise both densities and combination
- Difficulties with RMSE based evaluation increasingly recognised
  - 'Recent' work by Granger & Pesaran (2000, JoF), Gneiting (2011, JASA), Mitchell & Wallis (2011, JAE), Elliott *et al.* (2005, ReStud), Rossi & Sekhposyan (2012)...

Forecast Densities

- AG emphasise both densities and combination
- Difficulties with RMSE based evaluation increasingly recognised
  - 'Recent' work by Granger & Pesaran (2000, JoF), Gneiting (2011, JASA), Mitchell & Wallis (2011, JAE), Elliott *et al.* (2005, ReStud), Rossi & Sekhposyan (2012)...
- AG build on this push towards densities in macro using the log score and *pits* as evaluation tools
- Brings us back to the first of their empirical conclusions:
  - Full Bayesian (FB) predictive densities beat 'plug-in' densities (which use the posterior mode)
- But each of their (intrinsic) densities is Gaussian

Forecast Densities

- AG emphasise both densities and combination
- Difficulties with RMSE based evaluation increasingly recognised
  - 'Recent' work by Granger & Pesaran (2000, JoF), Gneiting (2011, JASA), Mitchell & Wallis (2011, JAE), Elliott *et al.* (2005, ReStud), Rossi & Sekhposyan (2012)...
- AG build on this push towards densities in macro using the log score and *pits* as evaluation tools
- Brings us back to the first of their empirical conclusions:
  - Full Bayesian (FB) predictive densities beat 'plug-in' densities (which use the posterior mode)
- But each of their (intrinsic) densities is Gaussian
  - Is FB *working* as it provides one means of introducing some much needed non-Gaussianity?

Statistical vs. economic loss

- AG use statistical evaluation tests which look at the *whole* density
  - But what about outliers and use of CRPS or median, rather than mean, log score?
- More generally, what's the *relevant* region of the forecast density?
  - i.e., what are the forecasts for?

Statistical vs. economic loss

- AG use statistical evaluation tests which look at the *whole* density
  - But what about outliers and use of CRPS or median, rather than mean, log score?
- More generally, what's the *relevant* region of the forecast density?
  - i.e., what are the forecasts for?
- Garratt, Mitchell & Vahey (2012) evaluate using a loss function based on the probability of deflation
  - Find that economic evaluation of a deflation event provides more discrimination between competing densities than statistical tests

Statistical vs. economic loss

- AG use statistical evaluation tests which look at the *whole* density
  - But what about outliers and use of CRPS or median, rather than mean, log score?
- More generally, what's the *relevant* region of the forecast density?
  - i.e., what are the forecasts for?
- Garratt, Mitchell & Vahey (2012) evaluate using a loss function based on the probability of deflation
  - Find that economic evaluation of a deflation event provides more discrimination between competing densities than statistical tests
  - See Diks, Panchenko and van Dijk (2011, JoE) on the statistical evaluation of tail events

Combining probabilistic forecasts

- AG is part of a programme of work on combining models/forecasts
- AG combine *several* models; others combine *many* models
- Models might all be individually misspecified
    - What use is the Bayes Factor between two misspecified models?

Combining probabilistic forecasts

- AG is part of a programme of work on combining models/forecasts
- AG combine *several* models; others combine *many* models
- Models might all be individually misspecified
  - What use is the Bayes Factor between two misspecified models?

Combining probabilistic forecasts

- AG is part of a programme of work on combining models/forecasts
- AG combine *several* models; others combine *many* models
- Models might all be individually misspecified
  - What use is the Bayes Factor between two misspecified models?
- Density combination (ensembling) a great way to produce more accurate/robust probabilistic forecasts
- Now used at central banks (in particular Norges Bank) when nowcasting & forecasting using a suite of models
- Probabilistic Forecasting Institute (ProFI) has been set up
  - to stimulate and coordinate research into new methods for probabilistic forecasting, evaluation and communication
  - to exchange ideas for operationalising methodologies

Equal vs. weighted combinations

- In AG equal weights appears to beat real-time optimised weights as the 3 models perform *pretty* similarly (in fact badly on basis of *pits*)
- When there's more **diversity** between the models than in AG it can pay to use real-time optimised weights

Equal vs. weighted combinations

- In AG equal weights appears to beat real-time optimised weights as the 3 models perform *pretty* similarly (in fact badly on basis of *pits*)
- When there's more **diversity** between the models than in AG it can pay to use real-time optimised weights
    - Hall & Mitchell (2005/7 OBES/IJF) combine BoE, NIESR and AR densities; Jore, Mitchell & Vahey (2010, JAE) combine break and no-break VARs; Bache et al. (2011, JEDC) combine many VARs with Norges Bank's DSGE

Equal vs. weighted combinations

- In AG equal weights appears to beat real-time optimised weights as the 3 models perform *pretty* similarly (in fact badly on basis of *pits*)
- When there's more **diversity** between the models than in AG it can pay to use real-time optimised weights
  - Hall & Mitchell (2005/7 OBES/IJF) combine BoE, NIESR and AR densities; Jore, Mitchell & Vahey (2010, JAE) combine break and no-break VARs; Bache et al. (2011, JEDC) combine many VARs with Norges Bank's DSGE
  - In a nowcasting application, Mazzi *et al.* (2010, Eurostat) find use of Occam's Window restores advantages of equal weighted density combinations, by eliminating *bad* models from the combination

Equal vs. weighted combinations

- In AG equal weights appears to beat real-time optimised weights as the 3 models perform *pretty* similarly (in fact badly on basis of *pits*)
- When there's more **diversity** between the models than in AG it can pay to use real-time optimised weights
    - Hall & Mitchell (2005/7 OBES/IJF) combine BoE, NIESR and AR densities; Jore, Mitchell & Vahey (2010, JAE) combine break and no-break VARs; Bache et al. (2011, JEDC) combine many VARs with Norges Bank's DSGE
    - In a nowcasting application, Mazzi *et al.* (2010, Eurostat) find use of Occam's Window restores advantages of equal weighted density combinations, by eliminating *bad* models from the combination

Equal vs. weighted combinations

- In AG equal weights appears to beat real-time optimised weights as the 3 models perform *pretty* similarly (in fact badly on basis of *pits*)
- When there's more **diversity** between the models than in AG it can pay to use real-time optimised weights
    - Hall & Mitchell (2005/7 OBES/IJF) combine BoE, NIESR and AR densities; Jore, Mitchell & Vahey (2010, JAE) combine break and no-break VARs; Bache et al. (2011, JEDC) combine many VARs with Norges Bank's DSGE
    - In a nowcasting application, Mazzi *et al.* (2010, Eurostat) find use of Occam's Window restores advantages of equal weighted density combinations, by eliminating *bad* models from the combination
- More general question is, how should we choose the model space? Statistically, economically...

Outstanding puzzles: selecting the model space

- AG combine a DFM, a DSGE and a (B)VAR density

Outstanding puzzles: selecting the model space

- AG combine a DFM, a DSGE and a (B)VAR density
  - All individually poor: intrinsic densities in each case are Gaussian
  - And the 3 models are pretty *similar*; they're all linear or linearised Gaussian
  - So with equal weights - if the world is non-linear, non-Gaussian - it is hard to see how this AG sparse linear combination is getting it any more than one of their individual models
  - None of them explicitly accommodate TVP, breaks, nonlinearities etc.
  - Indeed all models estimated (recursively) on data back to 1951q1

Outstanding puzzles: selecting the model space

- AG combine a DFM, a DSGE and a (B)VAR density
  - All individually poor: intrinsic densities in each case are Gaussian
  - And the 3 models are pretty *similar*; they're all linear or linearised Gaussian
  - So with equal weights - if the world is non-linear, non-Gaussian - it is hard to see how this AG sparse linear combination is getting it any more than one of their individual models
  - None of them explicitly accommodate TVP, breaks, nonlinearities etc.
  - Indeed all models estimated (recursively) on data back to 1951q1
- What use is it to know that model X is the most valued member of this combination?
  - Should we look for (one or many?) model(s) in the area of X?

Outstanding puzzles: selecting the model space

- AG combine a DFM, a DSGE and a (B)VAR density
    - All individually poor: intrinsic densities in each case are Gaussian
    - And the 3 models are pretty *similar*; they're all linear or linearised Gaussian
    - So with equal weights - if the world is non-linear, non-Gaussian - it is hard to see how this AG sparse linear combination is getting it any more than one of their individual models
    - None of them explicitly accommodate TVP, breaks, nonlinearities etc.
    - Indeed all models estimated (recursively) on data back to 1951q1
- What use is it to know that model X is the most valued member of this combination?
    - Should we look for (one or many?) model(s) in the area of X?

Outstanding puzzles: selecting the model space

- AG combine a DFM, a DSGE and a (B)VAR density
  - All individually poor: intrinsic densities in each case are Gaussian
  - And the 3 models are pretty *similar*; they're all linear or linearised Gaussian
  - So with equal weights - if the world is non-linear, non-Gaussian - it is hard to see how this AG sparse linear combination is getting it any more than one of their individual models
  - None of them explicitly accommodate TVP, breaks, nonlinearities etc.
  - Indeed all models estimated (recursively) on data back to 1951q1
- What use is it to know that model X is the most valued member of this combination?
  - Should we look for (one or many?) model(s) in the area of X?
- 'Dependence' between models
  - AG look simply at the correlation between the 3 models' log scores
  - X=DFM 'moves against the market' (negative correlation)
  - But dependence is lower in the tails $\Rightarrow$ nonlinear dependence, copula?

Combining several or many models?

- Strategies for selecting the model space (rather than how we combine)

Combining several or many models?

- Strategies for selecting the model space (rather than how we combine)

1. 'Ensemble' modelling
    - Combining many, many models rather than a small number as in AG
    - Some similarities with the meteorology literature

Combining several or many models?

- Strategies for selecting the model space (rather than how we combine)

1. 'Ensemble' modelling
    - Combining many, many models rather than a small number as in AG
    - Some similarities with the meteorology literature
2. Grand ensemble' (Garratt, Mitchell & Vahey, 2012)
    - Combine one *group* of models prior to combining with another *group*

Forecast diversity is important

- Recall Tobin's advice when picking financial assets:
  - 'don't put your eggs in one basket'

Forecast diversity is important

- Recall Tobin's advice when picking financial assets:
  - 'don't put your eggs in one basket'
- So why not combine many models? Still manageable computationally
- Linear Opinion Pool becomes more flexible as N increases: better ability to approximate non-Gaussian and non-linear DGPs
- There is then no need, as in AG, to select one DFM, one DSGE and one VAR

Forecast diversity is important

- Recall Tobin's advice when picking financial assets:
  - 'don't put your eggs in one basket'
- So why not combine many models? Still manageable computationally
- Linear Opinion Pool becomes more flexible as N increases: better ability to approximate non-Gaussian and non-linear DGPs
- There is then no need, as in AG, to select one DFM, one DSGE and one VAR
- What do AG's pooled densities look like?
  - Are they approximately Gaussian?
  - What feature of them accounts for their improvement over BMA?
    - Their shape (seems unlikely) or their location?

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*
- Aastveit *et al.* (2011, Norges Bank) find a 'grand ensemble' of VAR, leading indicator and factor models is effective when nowcasting

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*
- Aastveit *et al.* (2011, Norges Bank) find a 'grand ensemble' of VAR, leading indicator and factor models is effective when nowcasting
- Isn't the practical trick in portfolio management to find the assets that allow one to spread idiosyncratic risk?

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*
- Aastveit *et al.* (2011, Norges Bank) find a 'grand ensemble' of VAR, leading indicator and factor models is effective when nowcasting
- Isn't the practical trick in portfolio management to find the assets that allow one to spread idiosyncratic risk?
- ... similarly when combining models, more attention should be paid to how the model space is selected to accommodate forecast diversity

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*
- Aastveit *et al.* (2011, Norges Bank) find a 'grand ensemble' of VAR, leading indicator and factor models is effective when nowcasting
- Isn't the practical trick in portfolio management to find the assets that allow one to spread idiosyncratic risk?
- ... similarly when combining models, more attention should be paid to how the model space is selected to accommodate forecast diversity
  - Especially in the face of structural instabilities

The research agenda

- Unclear if AG's combinations/pools pass the *pits* tests; if not, perhaps we should question their model space and/or its size?
- Jore *et al.* (2010, JAE) find pooling many $N >> 3$ (simple) misspecified VAR models does, in fact, deliver well-calibrated densities on basis of *pits*
- Bache *et al.* (2011, JEDC) find an ensemble of a DSGE and many VARs is again effective in terms of *pits*
- Aastveit *et al.* (2011, Norges Bank) find a 'grand ensemble' of VAR, leading indicator and factor models is effective when nowcasting
- Isn't the practical trick in portfolio management to find the assets that allow one to spread idiosyncratic risk?
- ... similarly when combining models, more attention should be paid to how the model space is selected to accommodate forecast diversity
    - Especially in the face of structural instabilities
    - A known and leading cause of forecast failure

Modelling with structural instabilities

- Macro data are characterised by instabilities in, at least, both the mean and variance
- How should we accommodate these instabilities?

  1. In the individual/component models; and/or

Modelling with structural instabilities

- Macro data are characterised by instabilities in, at least, both the mean and variance
- How should we accommodate these instabilities?

  1. In the individual/component models; and/or
  2. When combining

Structural instabilities in the component models

Can think of using *robust* extensions to AG's 3 models

Structural instabilities in the component models

Can think of using *robust* extensions to AG's 3 models

1. Time Varying Parameter DFMs
2. TVP BVAR with stochastic volatility; Clark (2011, JBES)
3. DSGEs with time-variation; e.g. Justiniano and Primiceri (2008, AER)
    - Flexible stochastic trends; e.g. Canova (2011, QE), not simply deterministic as in Smets & Wouters (2007, AER)

Or accommodating structural instabilities when combining

- An ensemble of VARs (Jore and Garratt *et al.*) or DSGEs (Bache *et al.* 2010) estimated over different estimation windows
  - Crude but effective means of robustifying individually misspecified models to breaks in the conditional mean and importantly the variance

Or accommodating structural instabilities when combining

- An ensemble of VARs (Jore and Garratt *et al.*) or DSGEs (Bache *et al.* 2010) estimated over different estimation windows
  - Crude but effective means of robustifying individually misspecified models to breaks in the conditional mean and importantly the variance
- What about explicitly time-varying weights?
  - AG find optimised weights vary across pre, great and post Moderation 'regimes'
  - 3 models differ most in the probabilities they assign to tail events
- Why compute weights unconditionally (albeit recursively) over the whole evaluation period?

Or accommodating structural instabilities when combining

- An ensemble of VARs (Jore and Garratt *et al.*) or DSGEs (Bache *et al.* 2010) estimated over different estimation windows
  - Crude but effective means of robustifying individually misspecified models to breaks in the conditional mean and importantly the variance
- What about explicitly time-varying weights?
  - AG find optimised weights vary across pre, great and post Moderation 'regimes'
  - 3 models differ most in the probabilities they assign to tail events
- Why compute weights unconditionally (albeit recursively) over the whole evaluation period?
- AG's result suggests use of conditonal, nonlinear, time-varying... weights
  - Estimate combo weights separately over pre, great or post Moderation data (condition)

Or accommodating structural instabilities when combining

- An ensemble of VARs (Jore and Garratt *et al.*) or DSGEs (Bache *et al.* 2010) estimated over different estimation windows
  - Crude but effective means of robustifying individually misspecified models to breaks in the conditional mean and importantly the variance
- What about explicitly time-varying weights?
  - AG find optimised weights vary across pre, great and post Moderation 'regimes'
  - 3 models differ most in the probabilities they assign to tail events
- Why compute weights unconditionally (albeit recursively) over the whole evaluation period?
- AG's result suggests use of conditonal, nonlinear, time-varying... weights
  - Estimate combo weights separately over pre, great or post Moderation data (condition)
  - Waggoner & Zha (2012) Markov-switching weights

Or accommodating structural instabilities when combining

- An ensemble of VARs (Jore and Garratt *et al.*) or DSGEs (Bache *et al.* 2010) estimated over different estimation windows
  - Crude but effective means of robustifying individually misspecified models to breaks in the conditional mean and importantly the variance
- What about explicitly time-varying weights?
  - AG find optimised weights vary across pre, great and post Moderation 'regimes'
  - 3 models differ most in the probabilities they assign to tail events
- Why compute weights unconditionally (albeit recursively) over the whole evaluation period?
- AG's result suggests use of conditonal, nonlinear, time-varying... weights
  - Estimate combo weights separately over pre, great or post Moderation data (condition)
  - Waggoner & Zha (2012) Markov-switching weights
  - Let the weights vary by region; e.g. DSGE good in middle of density, some other model better in the tails

Only linear opinion pools (LOP)...

- AG invoke McConway's marginalisation result to motivate LOP
- But what about log pools?
  - Kascha & Ravazzolo (2010, JoF) and Wallis (2011, AFE)
- Log pool is externally Bayesian when the weights sum to unity; Genest (1984, Annals of Statistics)
- LOP vs. LogOP depends on which way round you do the KLIC minimisation
  - The combined density is that density KLIC closest to the N individual density forecasts
- Nonlinear (copula) pools model the dependence between the component densities; Garratt, Mitchell & Vahey (2012)
  - Found COP beats optimised LOP in simulations

Multivariate versus variable-specific evaluation

- AG use multivariate log score (but they could be clearer on this)
- Captures dependence across variables, j
  - But is this basically linear, given intrinsic normality assumption?
- Why use univariate not multivariate *pits*? Guess calibration will only be worse if we evaluate the joint density directly
- What about tuning the combination weights to reflect the variable of interest?
- More generally, can think of tailoring the optimisation to reflect your (economic?) loss function across the vector $\mathbf{Y}_t$

Minor Suggestions

- Does the DSGE do better at longer forecast horizons?
  - Will complicate *pits* tests due to overlap
  - Sensitivity to estimation window (plausibility of a single common determinstic trend is contingent on sample period)
- Focus on specific regions of the density of economic interest
- Relationship of your moments-based *pits* test with that of Malte Knüppel's similar sounding test?
- Test if differences in log scores are statistically significant using Amisano/Giacomini test?
  - But does this mean you need rolling estimation for asymptotics?
- No need to ignore data revisions
  - Could add in a component model to handle revisions predictability