Time Series and Forecasting
Lecture 2
Nowcasting, Forecast Combination, Variance Forecasting

Bruce E. Hansen

THE UNIVERSITY of WISCONSIN MADISON

Summer School in Economics and Econometrics
University of Crete
July 23-27, 2012

Today's Schedule

- Review
- VARs
- Nowcasting
- Combination Forecasts
- Variance Forecasting

Review

- Optimal point forecast of y_{n+1} given information I_n is the conditional mean $E(y_{n+1}|I_n)$
- Linear model $E(y_{n+1}|I_n) \simeq \boldsymbol{\beta}'\mathbf{x}_n$ is an approximation
- Estimate linear projections by least-squares
- Model selection should focus on performance, not "truth"
  - Best forecast has smallest MSFE
  - Unknown, but MSFE can be estimated
  - CV is a good estimator of MSFE
- Good forecasts rely on selection of leading indicators

Vector Autoregresive Models

- $\mathbf{y}_t$ is an p vector
- x_t are other variables (including lags)
- Ideal point forecast $E\left(\mathbf{y}_{n+1} | I_n\right)$
- Linear approximation

$$E\left(\mathbf{y}_{n+1} | I_n\right) \simeq A_1 \mathbf{y}_t + A_2 \mathbf{y}_{t-1} + \cdots + A_k \mathbf{y}_{t-k+1} + Bx_t$$

- Vector Autoregression (VAR)

$$\mathbf{y}_{t+1} = A_1 \mathbf{y}_t + A_2 \mathbf{y}_{t-1} + \cdots + A_k \mathbf{y}_{t-k+1} + Bx_t + e_{t+1}$$

- Estimation: Least squares

$$\mathbf{y}_{t+1} = \widehat{A}_1 \mathbf{y}_t + \widehat{A}_2 \mathbf{y}_{t-1} + \cdots \widehat{+ A}_k \mathbf{y}_{t-k+1} + \widehat{B} x_t + e_{t+1}$$

- One-Step-Ahead Point forecast

$$\widehat{\mathbf{y}}_{n+1} = \widehat{A}_1 \mathbf{y}_n + \widehat{A}_2 \mathbf{y}_{n-1} + \cdots + \widehat{A}_k \mathbf{y}_{n-k+1} + \widehat{B} x_n$$

Vector Autoregresive versus Univariate Models

- Let $\mathbf{x}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, ..., x_t)$
- Then a VAR is a set of p regression models

$$\begin{aligned} y_{1t+1} &= \boldsymbol{\beta}_1' \mathbf{x}_t + e_{1t} \\ &\vdots \\ y_{pt+1} &= \boldsymbol{\beta}_p' \mathbf{x}_t + e_{pt} \end{aligned}$$

- All variables $\mathbf{x}_t$ enter symmetrically in each equation
- Sims (1980) argued that there is no a priori reason to include or exclude an individual variable from an individual equation.

Model Selection

- Do not view selection as identification of "truth"
- Rather, inclusion/exclusion is to improve finite sample performance
  - minimize MSFE
- Use selection methods, equation-by-equation

Example: VAR with 2 variables

$$\begin{aligned} y_{1t+1} &= \widehat{\beta}_{11} y_{1t} + \widehat{\beta}_{12} y_{1t-1} + \widehat{\beta}_{13} y_{2t} + \widehat{e}_{1t} \\ &\vdots \\ y_{2t+1} &= \widehat{\beta}_{21} y_{1t} + \widehat{\beta}_{22} y_{2t} + \widehat{\beta}_{23} y_{2t-1} + \widehat{e}_{2t} \end{aligned}$$

- Selection picks y_{1t}, y_{1t-1}, y_{2t} for equation for y_{1t+1}
- Selection picks y_{1t}, y_{2t}, y_{2t-1} for equation for y_{2t+1}
- The two equations have different variables

- Same as system

$$\mathbf{y}_{t+1} = A_1 \mathbf{y}_t + A_2 \mathbf{y}_{t-1} + e_{t+1}$$

with

$$\begin{aligned} A_1 &= \begin{bmatrix} \beta_{11} & \beta_{13} \\ \beta_{21} & \beta_{22} \end{bmatrix} \\ A_2 &= \begin{bmatrix} \beta_{12} & 0 \\ 0 & \beta_{23} \end{bmatrix} \end{aligned}$$

- The VAR system notation is still quite useful for many purposes (including multi-step forecasting)

Nowcasting

- Forecasting current, near recent, or near future economic activity
- For example, 2nd quarter GDP (April-June 2012)
  - So far, we have used information up through first quarter
  - We have a fair amount of information
  - Quite a lot about the 2nd quarter itself

General Framework

- Two time scales
  - y_t (GDP)
  - x_v (interest rates)
  - $I_{t,v}$: information in y_j for $j \leq t$ and x_j for $j \leq v$
  - e.g., GDP up to 2011:1, interest rates up to today
- Optimal forecast of y_{t+1} given $I_{t,v}$ is conditional mean

$$E\left(y_{t+1} \middle| I_{t,v}\right) = \mu_{t,v}$$

Standard Linear Approximation

- Approximate conditional mean as linear and Markov

$$
\begin{aligned}
E\left(y_{t+1} | I_{t,v}\right) &= \mu_{t,v} \\
&\approx \beta_0 + \beta_1 y_t + \cdots + \beta_k y_{t-k+1} \\
&\quad + \gamma_0 x_v + \gamma_1 x_{v-1} + \cdots + \gamma_p x_{v-p}
\end{aligned}
$$

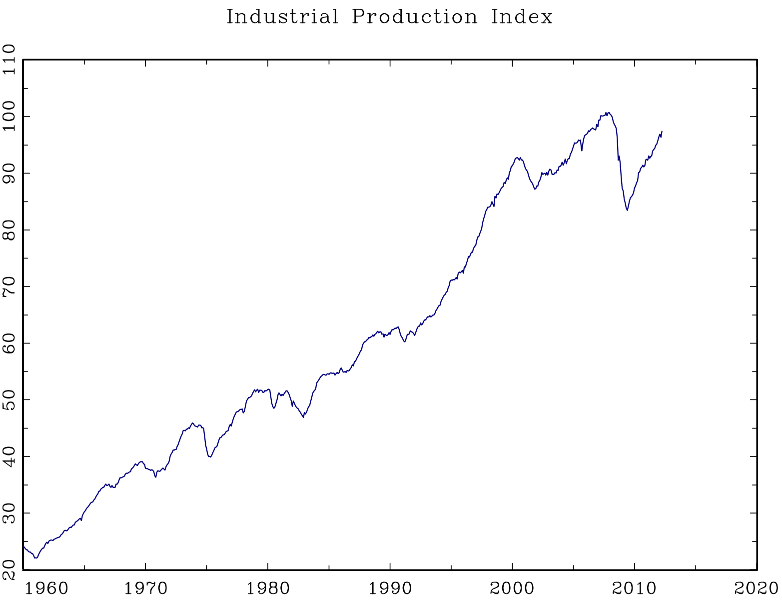
- Traditional solution (aggregate x_v to frequency t)
  - Sets $\gamma_j = 0$ for periods v before quarter t
  - Sets $\gamma_j = \gamma_k$ for periods j and k in common quarter t
  - Unreasonable restrictions
- Unrestricted approximation
  - Non-parsimonious
  - p may be very large

MIDAS

- Ghysels, Santa-Clara, and Valkanov
- Use parametric distributed-lag structure for coefficients γ_j
- Difficult to justify parametric restrictions

Example: GDP Nowcasting

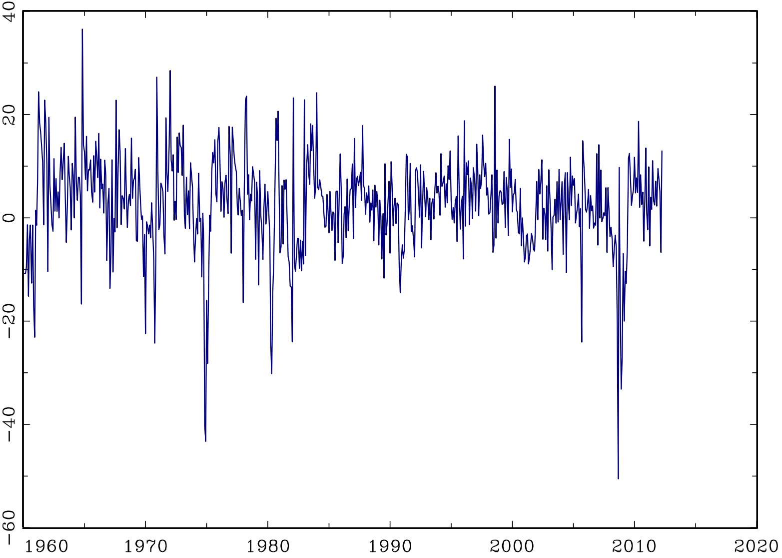
- Suppose we are interested in forecasting 2012 2nd quarter GDP growth
  - Economic activity for April, May and June
- For April, May and June, we have considerable information
  - Interest rates
  - unemployment rates
  - Industrial Production
  - Housing starts
  - Building Permits
  - Inflation

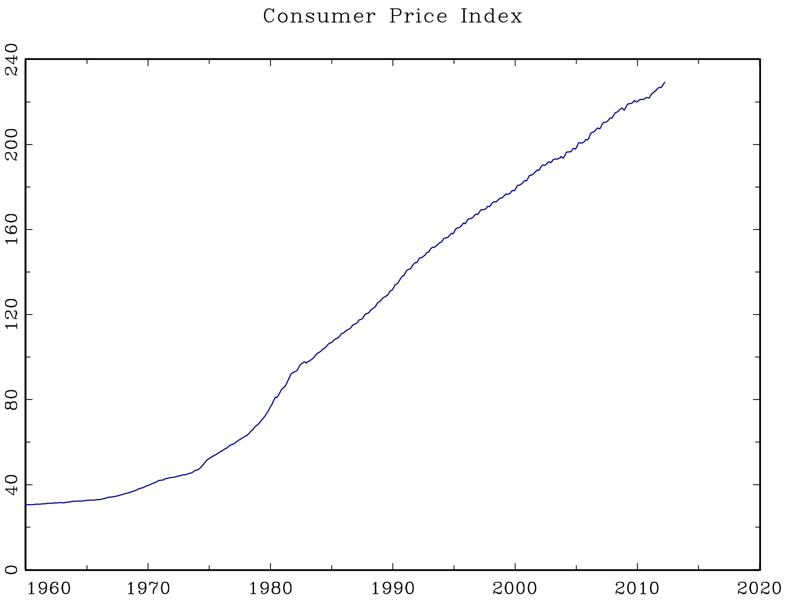

Industrial Production Index
110
100
90
80
70
60
50
40
30
20
1960
1970
1980
1990
2000
2010
2020

Growth Rate

$$x_t = \ln IP_t - \ln IP_{t-1}$$

Industrial Production Index Growth Rate

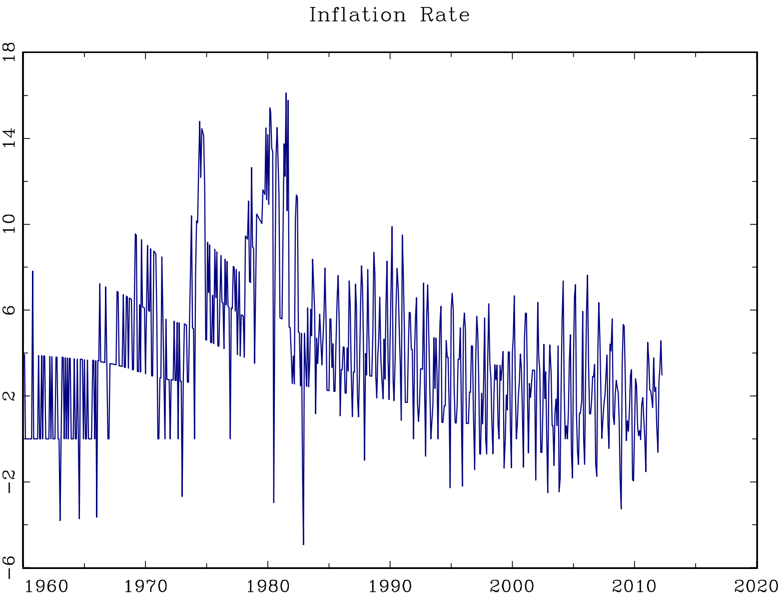




Consumer Price Index
0
40
80
120
160
200
240
1960
1970
1980
1990
2000
2010
2020

One Month Inflation Rate

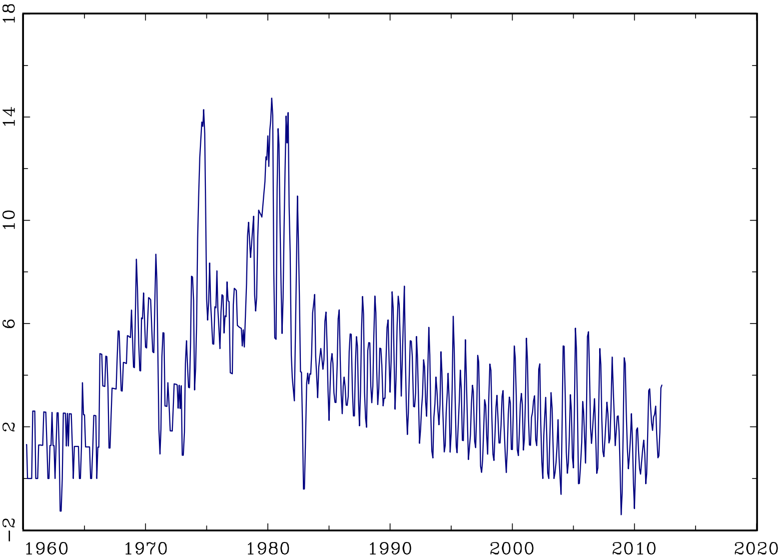
$$INF_t = \ln CPI_t - \ln CPI_{t-1}$$


Inflation Rate
18
14
10
6
2
−2
−6
1960
1970
1980
1990
2000
2010
2020

Three Month Inflation Rate

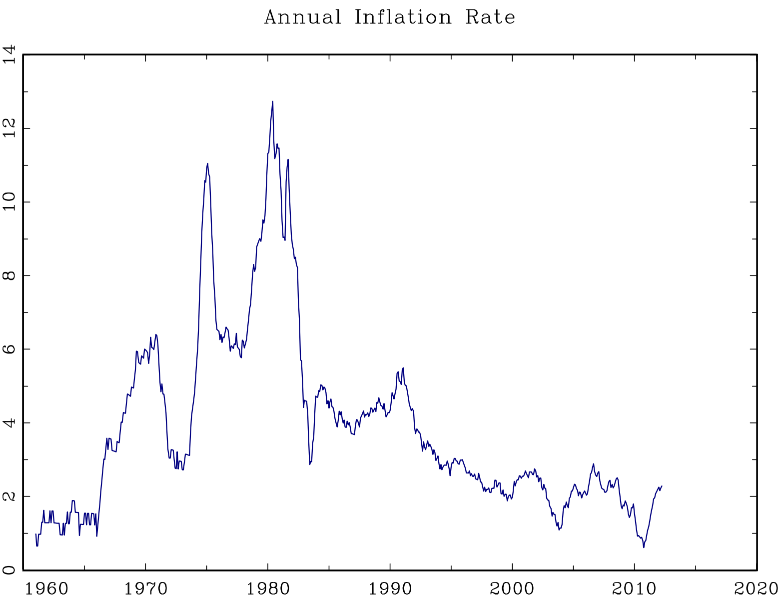
$$INF_t = \ln CPI_t - \ln CPI_{t-3}$$

3−Month Inflation Rate



One Year Inflation Rate

$$INF_t = \ln CPI_t - \ln CPI_{t-12}$$


Annual Inflation Rate
0
2
4
6
8
10
12
14
1960
1970
1980
1990
2000
2010
2020

Nowcasting Regression

- GDP growth as a linear function of
  - Previous 2 quarters GDP growth
  - Contemporaneous 3 months of
    - Term Spread (10 year over 3 month)
    - Default Spread (BAA over AAA yield)
    - Industrial Production
    - Building Permits
    - Housing Starts
  - (Or whatever is available at time of forecast)

Notation

- $t = year$
- $q = quarter$, $q = 1, 2, 3, 4$
- $m = month$ in quarter, $m = 1, 2, 3$
- $GDP_{t,q}$= GDP in year t, quarter q
  - Convention: $GDP_{t,0} = GDP_{t-1,4}$
- $IP_{t,q,m} = IP$ in year t, quarter q, month m

Example Models

- Monthly Data through First Month of Forecast Quarter

$$GDP_{t,q} = \beta_1 GDP_{t,q-1} + \beta_2 GDP_{t,q-2} + \beta_3 IP_{t,q,1} + \beta_4 IP_{t,q-1,3} + \cdots$$

- Monthly Data through Second Month of Forecast Quarter

$$GDP_{t,q} = \beta_1 GDP_{t,q-1} + \beta_2 GDP_{t,q-2} + \beta_3 IP_{t,q,2} + \beta_4 IP_{t,q,1} + \cdots$$

- Regressor Construction from Monthly Variables
  - Divide into "first", "second" and "third" months of quarters
  - Now you have 3 quarterly observations for each variable

Nowcasting Estimates

- Based on data through April (first month of forecast quarter)
- Selected variables:
  - $\Delta \log(GDP_t)$ (one lag)
  - IP_1, IP_3, IP_2 (first, previous third, and previous second months)
  - HS_1, HS_3 (first and previous third months)

	$\hat{\beta}$	$s(\hat{\beta})$
Intercept	0.32	(0.62)
$\Delta \log(GDP_t)$	-0.07	(0.06)
Industrial Production$_1$	0.17	(0.02)
Industrial Production$_3$	0.07	(0.02)
Industrial Production$_2$	0.12	(0.03)
Housing Starts$_1$	4.00	(1.14)
Housing Starts$_3$	-2.64	(1.14)

Nowcasting Point Forecast

- 2nd Quarter GDP Growth: 2.93
- Fitted model: $CV = 5.339$
  - Note that yesterday's best fitting model had $CV = 10.28$
  - Point forecast changes from 1.53 to 2.93
  - Adding contemporaneous IP very useful

Flexibility

- As each piece of information becomes available, that variable can be added to regression
- Sequence of nowcast estimates, updated with new information

Recommendation

- Make use of higher frequency information
- Be creative and flexible
- Handling high-dimensional p is similar to many other high-dimensional problems
  - Model selection, combination, shrinkgae
- Requires frequent re-estimation of distinct forecasting models as new information arises
  - Requires significant empirical care and attention to detail

Combination Forecasts

Diversity of Forecasts

- Model choice is critical
  - Classic approach: Selection
  - Modern approach: Combination
- Issues:
  - How to select from a wide set of models/forecasts?
    - Model selection criteria
  - How to combine a wide set of models/forecasts?
    - Weight selection criteria

Foundation

- The ideal point forecast minimizes the MSFE
- The goal of a good combination forecast is to minimize the MSFE

Forecast Selection

- M forecasts: $\mathbf{f} = \{f(1), f(2), ..., f(M)\}$
- Selection picks $\hat{m}$ to determine the forecast $f = f(\hat{m})$
- M weights: $\mathbf{w} = \{w(1), w(2), ..., w(M)\}$
- A combination forecast is the weighted average

$$\begin{aligned} f(\mathbf{w}) &= \sum_{m=1}^{M} w(m) f(m) \\ &= \mathbf{w}'\mathbf{f} \end{aligned}$$

- Combination generalizes selection

Possible restrictions on the weight vector

- $\sum_{m=1}^{M} w(m) = 1$
  - Unbiasedness
  - Typically improves performance
- $w(m) \geq 0$
  - nonnegativity
  - regularization
  - Often critical for good performance
- $w(m) \in \{0, 1\}$
  - Equivalent to forecast selection
  - $f(\mathbf{w}) = f(m)$
  - Selection is a special case of combination
  - Strong restriction

OOS Forecast Combination

- Sequence of true out-of-sample forecasts $\mathbf{f}_t$ for y_{t+1}
- Combination forecast is $f(\mathbf{w}) = \mathbf{w}'\mathbf{f}$
- OOS empirical MSFE

$$\hat{\sigma}^2(\mathbf{w}) = \frac{1}{P} \sum_{t=n-P}^{n} \left(y_{t+1} - \mathbf{w}'\mathbf{f}_t\right)^2$$

- PLS selected the model with the smallest OOS MSFE
- Granger-Ramanathan combination: select $\mathbf{w}$ to minimize the OOS MSFE
- Minimization over $\mathbf{w}$ is equivalent to the least-squares regression of y_t on the forecasts

$$y_{t+1} = \mathbf{w}'\mathbf{f}_t + \varepsilon_{t+1}$$

Granger-Ramanathan (1984)

- Unrestricted least-squares

$$\hat{\mathbf{w}} = \left(\sum_{t=n-P}^{n} \mathbf{f}_t \mathbf{f}_t' \right)^{-1} \sum_{t=n-P}^{n} \mathbf{f}_t y_{t+1}$$

- This can produce weights far outside $[0, 1]$ and don't sum to one
- Granger-Ramanathan's intuition was that this flexibility is good
    - But they provided no theory to support conjecture
- Unrestricted weights are not regularized
    - This results in poor sampling performance

Alternative Representation

- Take $y_{t+1} = \mathbf{w}'\mathbf{f}_t + \varepsilon_{t+1}$, subtract y_{t+1} from each side

$$0 = \mathbf{w}'\mathbf{f}_t - y_{t+1} + \varepsilon_{t+1}$$

- Impose restriction that weights to sum to one.

$$0 = \mathbf{w}'\left(\mathbf{f}_t - y_{t+1}\right) + \varepsilon_{t+1}$$

- Define $\mathbf{e}_{t+1} = \mathbf{w}'\left(\mathbf{f}_t - y_{t+1}\right)$, the (negative) forecast errors. Then

$$0 = \mathbf{w}'\mathbf{e}_{t+1} + \varepsilon_{t+1}$$

- This is the regression of 0 on the forecast errors
- But it is still better to also impose non-negativity $w(m) \geq 0$

Constrained Granger-Ramanathan

The constrained GR weights solve the problem

$$\min_{\mathbf{w}} \mathbf{w}'\mathbf{A}\mathbf{w}$$

subject to

$$\sum_{m=1}^{M} w(m) = 1$$

$$0 \leq w(m) \leq 1$$

where

$$\mathbf{A} = \sum_t \mathbf{e}_{t+1}\mathbf{e}'_{t+1}$$

is the $M \times M$ matrix of forecast error empirical variances/covariances

Quadratic Programming (QP)

- The weights lie on the unit simplex
- The constrained GR weights minimize a quadratic over the unit simplex
- QP algorithms easily solve this problem
  - Gauss (qprog)
  - Matlab (quadprog)
  - R (quadprog)
- Solution solution typical
  - Many forecasts will receive zero weight

Bates-Granger (1969)

- Assume $\mathbf{A} = \sum_t \mathbf{e}_{t+1}\mathbf{e}_{t+1}'$ is diagonal.
- Then the regression with the coefficients constrained to sum to one

$$0 = \mathbf{w}'\mathbf{e}_{t+1} + \varepsilon_{t+1}$$

  has solution

$$w(m) = \frac{\hat{\sigma}^{-2}(m)}{\sum_{j=1}^{M} \hat{\sigma}^{-2}(j)}$$

- This are the Bates-Granger weights.
- In many cases, they are close to equality, since OOS forecast variances can be quite similar

Bayesian Model Averaging (BMA)

- Put priors on individual models, and priors on the probability that model m is the true model
- Compute posterior probabilites $w(m)$ that m is the true model
- Forecast combination using $w(m)$
- Advantages
  - Conceptually simple
  - no theoretical analysis required
  - applies in broad contexts
- Disadvantages
  - Not designed to minimize forecast risk
  - Similar to BIC: asymptotically picks “true” finite models
  - does not distinguish between 1-step and multi-step forecast horizons

BMA Approximation

- BIC weights

$$w(m) \propto \exp\left(-\frac{BIC(m)}{2}\right)$$

- Simple approximation to full BMA method
- Smoothed version of BIC selection
- Works better than BIC selection in simulations

AIC Weights

- Smooted AIC

$$w(m) \propto \exp\left(-\frac{AIC(m)}{2}\right)$$

- Proposed by Buckland, Burnhamm and Augustin (1997)
- Not theoretically motivated, but works better than AIC selection in simulations

Comments

- Combination methods typically work better (lower MSFE) than comparable selection methods
- BIC and BMA not optimal for MSFE
- Granger-Ramanathan has similar senstive as PLS to choice of P
- Bates-Granger and weighted AIC have no theoretical grounding

Forecast Combination

$$\begin{aligned}
\widehat{y}_{n+1}(\mathbf{w}) &= \sum_{m=1}^{M} w(m)\widehat{y}_{n+1}(m) \\
&= \sum_{m=1}^{M} w(m)\mathbf{x}_n(m)'\widehat{\beta}(m) \\
&= \mathbf{x}_n'\widehat{\beta}(\mathbf{w})
\end{aligned}$$

where

$$\widehat{\beta}(\mathbf{w}) = \sum_{m=1}^{M} w(m)\widehat{\beta}(m)$$

- In linear models, the combination forecast is the same as the forecast based on the weighted average of the parameter estimates across the different models
- Computationally, it is easiest to calculate the M individual forecast $\hat{y}_{n+1}(m)$, then take the weighted average to obtain $\hat{y}_{n+1}(\mathbf{w})$

Combination Residuals

$$
\begin{aligned}
\widehat{e}_{t+1}(\mathbf{w}) &= y_{t+1} - \mathbf{x}_t' \widehat{\beta}(\mathbf{w}) \\
&= \sum_{m=1}^{M} w(m) \left(y_{t+1} - \mathbf{x}_t' \widehat{\beta}(m) \right) \\
&= \sum_{m=1}^{M} w(m) \widehat{e}_{t+1}(m)
\end{aligned}
$$

- In linear models, the residual from the combination model is the same as the weighted average of the model residuals.

Residual variance

$$\begin{aligned}\hat{\sigma}^2(\mathbf{w}) &= \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=1}^{M} w(m)\widehat{e}_{t+1}(m)\right)^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}'\widehat{\mathbf{e}}_{t+1}\right)^2 \\ &= \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w}\end{aligned}$$

where

$$\widehat{\mathbf{S}} = \frac{1}{n}\sum_{t=1}^{n}\widehat{\mathbf{e}}_{t+1}\widehat{\mathbf{e}}_{t+1}'$$

- The residual variance is a quadratic function of the covariance matrix of the M model residuals.

Point Forecast and MSFE

- Given $\widehat{y}_{n+1}(\mathbf{w})$ the forecast error is

$$
\begin{aligned}
y_{n+1} - \widehat{y}_{n+1}(\mathbf{w}) &= \mathbf{x}_n'\boldsymbol{\beta} + e_{t+1} - \mathbf{x}_n'\widehat{\boldsymbol{\beta}}(\mathbf{w}) \\
&= e_{n+1} - \mathbf{x}_n'\left(\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)
\end{aligned}
$$

- The mean-squared-forecast-error (MSFE) is

$$
\begin{aligned}
MSFE(\mathbf{w}) &= E\left(e_{n+1} - \mathbf{x}_n'\left(\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right)^2 \\
&\simeq \sigma^2 + E\left(\left(\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)' Q \left(\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \boldsymbol{\beta}\right)\right)
\end{aligned}
$$

- Minimizing MSFE is the same as minimizing the MSE of the coefficient estimate

Fitted values from Combination Forecast

$$\widehat{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M} w(m)\mathbf{x}_t'\widehat{\beta}(m)$$

and

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \sum_{m=1}^{M} w(m)\mathbf{X}(m)\widehat{\boldsymbol{\beta}}(m) \\
&= \sum_{m=1}^{M} w(m)\mathbf{X}(m)\left(\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\mathbf{X}(m)'\mathbf{y} \\
&= \sum_{m=1}^{M} w(m)\mathbf{P}(m)\mathbf{y} \\
&= \mathbf{P}(\mathbf{w})\mathbf{y}
\end{aligned}$$

where

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M} w(m)\mathbf{P}(m)$$

Fitted values from Combination Forecast (con't)

$$\widehat{\mu} = \mathbf{P}(\mathbf{w})\mathbf{y}$$

$$\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M} w(m)\mathbf{P}(m)$$

- In-sample fitted values are a linear operator on the dependent variable
- The operator $\mathbf{P}(\mathbf{w})$ is not a projection matrix
- It is a weighted average of projection matrices

Residual Fit

$$
\begin{aligned}
\widehat{\sigma}(\mathbf{w})^2 &= \frac{1}{n}\sum_{t=0}^{n-1}\widehat{e}_{t+1}(\mathbf{w})^2 \\
&= \frac{1}{n}\sum_{t=0}^{n-1}e_{t+1}^2 + \frac{1}{n}\sum_{t=0}^{n-1}\left(\mathbf{x}_t'\left(\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}\right)\right)^2 \\
&\quad -\frac{2}{n}\sum_{t=0}^{n-1}e_{t+1}\mathbf{x}_t'\left(\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}\right)
\end{aligned}
$$

- First two terms are estimates of

$$
MSFE(\mathbf{w}) = E\left(e_{n+1} - \mathbf{x}_n'\left(\widehat{\boldsymbol{\beta}}(\mathbf{w})-\boldsymbol{\beta}\right)\right)^2
$$

Third term is

$$\begin{aligned}\sum_{t=0}^{n-1} e_{t+1}\mathbf{x}_t'\left(\widehat{\beta}(\mathbf{w})-\beta\right) &= \sum_{m=1}^{M} w(m)\sum_{t=0}^{n-1} e_{t+1}\mathbf{x}_t'\left(\widehat{\beta}(m)-\beta\right) \\ &= \sum_{m=1}^{M} w(m)\mathbf{e}'\mathbf{P}(m)\mathbf{e} \\ &= \mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}\end{aligned}$$

where

$$\mathbf{P}(m)=\mathbf{X}(m)\left(\mathbf{X}(m)'\mathbf{X}(m)\right)^{-1}\mathbf{X}(m)'$$

and

$$\mathbf{P}(\mathbf{w})=\sum_{m=1}^{M} w(m)\mathbf{P}(m)$$

Residual Variance as Biased estimate of MSFE

$$E\left(\widehat{\sigma}(\mathbf{w})^2\right) \simeq MSFE_n(\mathbf{w}) - \frac{2}{n}B(\mathbf{w})$$

where

$$\begin{aligned} B(\mathbf{w}) &= E\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\ &= \sum_{m=1}^{M} w(m)E\left(\mathbf{e}'\mathbf{P}(m)\mathbf{e}\right) \\ &= \sum_{m=1}^{M} w(m)B(m) \end{aligned}$$

Unbiased estimate of MSFE

$$C_n(\mathbf{w}) = \widehat{\sigma}(\mathbf{w})^2 + \frac{2}{n}B(\mathbf{w})$$

Bias Term

$$B(\mathbf{w}) = \sum_{m=1}^{M} w(m) B(m)$$

$$B(m) = \operatorname{tr}\left(Q(m)^{-1}\Omega(m)\right)$$

In homoskedastic case

$$B(m) = \sigma^2 k(m)$$

$$B(\mathbf{w}) = \sigma^2 \sum_{m=1}^{M} w(m) k(m)$$

a weighted average of the number of coefficients in each estimator.

Mallows Averaging Criterion

$$C_n(\mathbf{w}) = \widehat{\sigma}^2(\mathbf{w}) + \frac{2}{n}\widetilde{\sigma}^2 \sum_{m=1}^{M} w(m)k(m)$$

with $\widetilde{\sigma}^2$ an estimate from a "large" model

$$\widetilde{\sigma}^2 = \frac{1}{n-K}\sum_{t=0}^{n-1} \widehat{e}_{t+1}(K)^2$$

Hansen (2007, Econometrica) Mallows Model Averaging (MMA)

Mallows Weight Selection

Write

$$\sum_{m=1}^{M} w(m)k(m) = \mathbf{w}'\mathbf{K}$$

where $\mathbf{K} = (k(1), ..., k(M))'$. This is linear in $\mathbf{w}$
We showed earlier that $\widehat{\sigma}^2(\mathbf{w}) = \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w}$ is quadratic.
Linear/Quadratic criterion

$$C_n(\mathbf{w}) = \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w} + \frac{2}{n}\widetilde{\sigma}^2\mathbf{w}'\mathbf{K}$$

Forecast Model Averaging (FMA)

- Hansen (Journal of Econometrics, 2008)

$$C_n(\mathbf{w}) = \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w} + \frac{2}{n}\widetilde{\sigma}^2\mathbf{w}'\mathbf{K}$$

- Combination weights found by constrained minimization of $C_n(\mathbf{w})$

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\mathbf{w}'\widehat{\mathbf{S}}\mathbf{w} + \frac{2}{n}\widetilde{\sigma}^2\mathbf{w}'\mathbf{K}\right]$$

subject to

$$\sum_{m=1}^{M} w(m) = 1$$

$$0 \leq w(m) \leq 1$$

- Solution by Quadratic Programming (QP)

Theory of Optimal Weights

- $MSFE_n(\mathbf{w})$ is the MSFE using weights $\mathbf{w}$
- $\inf_{\mathbf{w}} MSFE_n(\mathbf{w})$ is the (infeasible) best MSFE, where the inf is over all feasible weights
- Let $\widehat{\mathbf{w}}$ be the selected weights
- Let $MSFE_n(\widehat{\mathbf{w}})$ denote the MSFE using the selected weighted average
- We say that weight selection is asymptotically optimal if

$$\frac{MSFE_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}} MSFE_n(\mathbf{w})} \xrightarrow{p} 1$$

Theory of Optimal Weights

- Hansen (2007, Econometrica)
- Mallows weight selection is asymptotically optimal under homoskedasticity
- No optimality proof yet for dependent data

Comparison of Granger-Ramanathan and FMA

- Both are solved by Quadratic Programming (QP)
- Both typically yield corner solutions – many forecasts will receive zero weight
- GR uses empirical (OOS) forecast errors, FMA uses sample residuals
- GR uses no penalty, FMA uses "average # of parameters" penalty
- FMA is an estimate of MSFE for homoskedastic one-step forecasts, GR has no optimality

Robust Mallows

$$C_n(\mathbf{w}) = \widehat{\sigma}^2(\mathbf{w}) + \frac{2}{n}\sum_{m=1}^{M} w(m)\,\mathrm{tr}\left(Q(m)^{-1}\Omega(m)\right)$$

$$\begin{aligned} Q(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t(m)'\right) \\ \Omega(m) &= E\left(\mathbf{x}_t(m)\mathbf{x}_t'(m)e_{t+1}^2\right) \end{aligned}$$

Sample estimate

$$\begin{aligned} C_n^*(\mathbf{w}) &= \widehat{\sigma}^2(\mathbf{w}) + \frac{2}{n}\sum_{m=1}^{M} w(m)\,\mathrm{tr}\left(\widehat{Q}(m)^{-1}\widehat{\Omega}(m)\right) \\ &= \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w} + \frac{2}{n}\mathbf{w}'\mathbf{B} \end{aligned}$$

where

$$\mathbf{B} = \left(\ \mathrm{tr}\left(\widehat{Q}(1)^{-1}\widehat{\Omega}(1)\right),\ \ \mathrm{tr}\left(\widehat{Q}(2)^{-1}\widehat{\Omega}(2)\right),\ \ \vdots\ \ \mathrm{tr}\left(\widehat{Q}(K)^{-1}\widehat{\Omega}(K)\right)\ \right)'$$

is vector of correction terms from robust Mallows selection.

Cross-Validation

- Leave-one-out estimator

$$\widehat{\beta}_{-t}(\mathbf{w}) = \sum_{m=1}^{M} w(m)\widehat{\beta}_{-t}(m)$$
$$= \sum_{m=1}^{M} w(m)\left(\sum_{j\neq t}\mathbf{x}_j(m)\mathbf{x}_j(m)'\right)^{-1}\left(\sum_{j\neq t}\mathbf{x}_j(m)y_{j+1}\right)$$

- Leave-one-out prediction residual

$$\widetilde{e}_{t+1}(m) = y_{t+1} - \sum_{m=1}^{M} w(m)\widehat{\beta}_{-t}(\mathbf{w})'\mathbf{x}_t(m)$$
$$= \sum_{m=1}^{M} w(m)\widetilde{e}_{t+1}(m)$$

where the second equality holds since the weights sum to one.

- $CV_n(\mathbf{w}) = \frac{1}{n}\sum_{t=0}^{n-1} \widetilde{e}_{t+1}(\mathbf{w})^2$ is an estimate of $MSFE_n(m)$
- Cross-validation (CV) criterion for regression combination/averaging

Cross-validation criterion for combination forecasts

$$
\begin{aligned}
CV_n(\mathbf{w}) &= \frac{1}{n}\sum_{t=1}^{n} \widetilde{e}_{t+1}(\mathbf{w})^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=1}^{M} w(m)\widetilde{e}_{t+1}(m)\right)^2 \\
&= \sum_{m=1}^{M}\sum_{\ell=1}^{M} w(m)w(\ell)\frac{1}{n}\sum_{t=1}^{n}\widetilde{e}_{t+1}(m)\widetilde{e}_{t+1}(\ell) \\
&= \mathbf{w}'\widetilde{\mathbf{S}}\mathbf{w}
\end{aligned}
$$

where

$$
\widetilde{\mathbf{S}} = \frac{1}{n}\widetilde{e}'\widetilde{e}
$$

is covariance matrix of leave-1-out residuals.

Cross-validation Weights

Combination weights found by constrained minimization of $CV_n(\mathbf{w})$

$$\min_{\mathbf{w}} CV_n(\mathbf{w}) = \mathbf{w}'\widetilde{\mathbf{S}}\mathbf{w}$$

subject to

$$\sum_{m=1}^{M} w(m) = 1$$

$$0 \leq w(m) \leq 1$$

Cross-validation for combination forecasts (theory)

- **Theorem**: $ECV_n(\mathbf{w}) \simeq C_n(\mathbf{w})$
- For heteroskedastic forecasts, CV is a valid estimate of the one-step MSFE from a combination forecast
- Hansen and Racine (Journal of Econometrica, 2012) show that the CV weights are asymptotically optimal for cross-section data under heteroskedasticity
- No optimality theory for dependent data

Computation (R)

- Min $(\frac{1}{2}\mathbf{w}'\widetilde{\mathbf{S}}\mathbf{w} + d'\mathbf{w})$ subject to $A'\mathbf{w} \geq b$
- Need *quadprog* package
  - Install under *packages*
  - `library(quadprog)`
- `QP <- solve.QP(D,d,A,b,b)`
- `w <- QP$solution`
- `w <- as.matrix(w)`
- help(solve.QP) for documentation
- $D = \widetilde{\mathbf{S}} = (e'e)/n$ where e is $n \times M$ matrix of leave-one-out residuals

Summary: Forecast Combination Methods

- Granger-Ramanathan (GR), forecast model averaging (FMA) and cross-validation (CV) all pick weight vectors by quadratic minimization
- GR only needs actual forecasts, the method can be unknown or a black box
- CV can be computed for a wide variety of estimation methods
  - optimality theory for linear estimation
- FMA limited to homoskedastic one-step-ahead models
- Smoothed AIC (SAIC) and BMA have no forecast optimality, and are designed for homoskedastic one-step-ahead forecasts.

Example: AR models for GDP Growth

- Fit AR(1) and AR(2) only
- Leave-one-out residuals $\tilde{e}_{1t}$ and $\tilde{e}_{2t}$
- Covariance matrix
$$\widetilde{\mathbf{S}} = \begin{bmatrix} 10.72 & 10.44 \\ 10.44 & 10.52 \end{bmatrix}$$
- The best-fitting single model is AR(2)
- The best combination is $\mathbf{w} = (.22, .78)'$
- $CV = 10.50$

Example: AR models for GDP Growth

- Fit AR(0) through AR(12)
- AR(0) is constant only
- Models with positive weight are AR(0), AR(1), AR(2)
- $\mathbf{w} = (.06, .16, .78)'$

$$\widetilde{\mathbf{S}} = \begin{bmatrix} 12.0 & 10.6 & 10.4 \\ 10.6 & 10.7 & 10.4 \\ 10.4 & 10.5 & 10.5 \end{bmatrix}$$

- $CV = 10.50$ (essentially unchanged)

Example: Leading Indicator Forecasts

- Fit AR(1), AR(2) with leading indicators
- Models with positive weight

	w
AR(1), Spread, Housing	0.13
AR(1), Spread, High-Yield, Housing	0.16
AR(1), Spread, High-Yield, Housing, Building	0.52
AR(2)	0.18
AR(2), Spread	0.01

- $CV = 9.81$

Example: Nowcasting

- Models with positive weight are
  - $w = .17$ on $\Delta \log(GDP_t)$, IP_1, IP_3, IP_2, HS_1,
  - $w = .83$ on $\Delta \log(GDP_t)$, IP_1, IP_3, IP_2, HS_1, HS_3
- $CV = 5.335$
- Point Forecast$=$ 2.91
- Essentially same as selected model

Summary: Forecast Combination by CV

- M forecasts $\widehat{f}_{n+1}(m)$ from n observations
- For each estimate m
  - Define the leave-one-out prediction error

$$\begin{aligned}\widetilde{e}_{t+1}(m) &= y_{t+1} - \widehat{\boldsymbol{\beta}}'_{(-t)}(m)\mathbf{x}_t(m) \\ &= \frac{\widehat{e}_{t+1}(m)}{1 - h_{tt}(m)}\end{aligned}$$

  - Store the $n \times 1$ vector $\widetilde{\mathbf{e}}(m)$
- Construct the $M \times M$ matrix

$$\widetilde{\mathbf{S}} = \frac{1}{n}\widetilde{e}'\widetilde{e}$$

- Find the $M \times 1$ weight vector $\mathbf{w}$ which minimizes $\mathbf{w}'\widetilde{\mathbf{S}}\mathbf{w}$
  - Use quadratic programming (quadprog) to find solution
- The combination forecast is $\widehat{f}_{n+1} = \sum_{m=1}^{M} w(m)\widehat{f}_{n+1}(m)$

Forecast Combination Criticisms

- There has been considerable skepticism about formal forecast combination method in the forecast literature
- Many researchers have found that equal weighting: $(w_m = 1/M)$ works as well as formal methods
- However, the formal methods which investigated are
  - Bates-Granger simple weights
    - Not expected by theory to work well
  - Unconstrained Granger-Ramanathan
    - Without imposing $[0, 1]$ weights, work terribly!
- Furthermore, most investigations examine pseudo out-of-sample performance
  - Identical to comparing models by PLS criterion
  - This is NOT an investigation of performance
  - Just a ranking by PLS

Another Example - 10-Year Bond Rate

- Estimated AR(1) through AR(24) models
- CV Selection picked AR(2)
- CV weight Selection: Models with positive weight
  - AR(0): $w = 0.04$
  - AR(1): $w = 0.04$
  - AR(2): $w = 0.47$
  - AR(6): $w = 0.23$
  - AR(22): $w = 0.22$
- MInimizing $CV = 0.0761$ (slightly lower than 0.0768 from AR(2))
- Point forecast 1.96 (same as from AR(2))

Variance Forecasting

Variance Forecasts

- Forecast uncertainty
  - Point forecasts insufficient!
- $\sigma_{t+1}^2 = \mathrm{var}\left(y_{t+1} | I_t\right)$
- In the model $y_{t+1} = \boldsymbol{\beta}' \mathbf{x}_t + e_{t+1}$
  - $\sigma_{t+1}^2 = \mathrm{var}\left(e_{t+1} | I_n\right) = E\left(e_{t+1}^2 | I_t\right)$

10-Year Bond Rate

- Prediction Residuals
- Squares

Figure: Leave-One-Out Prediction Residuals

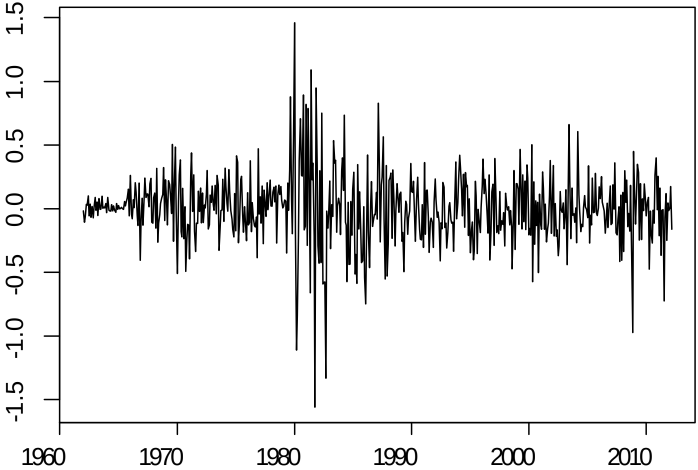
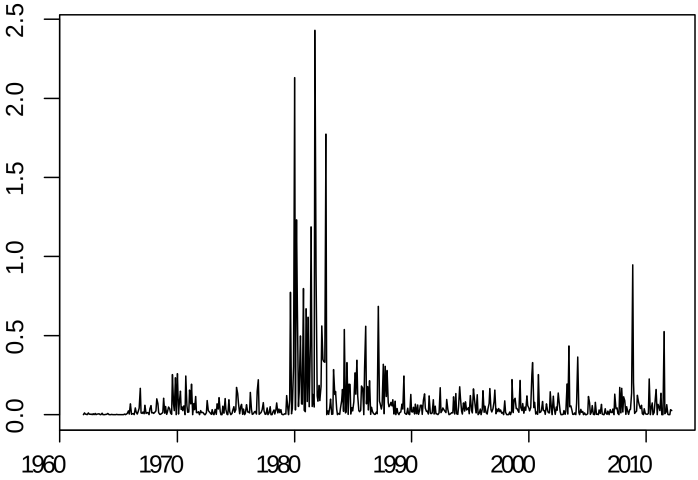


Figure: Squared Prediction Residuals



Variance Forecast Methods

- Constant Variance $\sigma_t^2 = \sigma^2$
  - Uncertainty not state-dependent
- GARCH
  - Common in financial data
  - Estimated by MLE
- Regression Approach
  - $\sigma_t^2 = E\left(e_{t+1}^2 | I_n\right) \approx \boldsymbol{\alpha}'\mathbf{x}_t$

2-Step Variance Estimation

- Start with residuals $\widehat{e}_{t+1}$
  - Better choice: leave-one-out residuals $\widetilde{e}_{t+1}$
- Estimate variance model (constant, ARCH, or regression)
- Obtain $\widehat{\sigma}_n^2$ from fitted model

Which Residuals?

- Least-squares residual variance biased toward zero
  - Forecast variance biased towards zero
- Leave-one-out residual variance estimates out-of-sample MSFE
  - This is appropriate

Joint Estimation: Mean and Variance

- Alternative to two-step estimation
  - I prefer 2-step as the regression coefficients preserve their projection interpretation
  - When the model is an approximation, the coefficient change their meaning under joint estimation

Constant Variance Model

- $\sigma_t^2 = \sigma^2$
- $\widehat{\sigma}_n^2 = \widehat{\sigma}^2 = \dfrac{1}{n-1} \sum_{t=1}^{n-1} \widetilde{e}_{t+1}^2$

Regression Variance Model

- $\sigma_t^2 \approx \boldsymbol{\alpha}' \mathbf{x}_t$
- $e_{t+1}^2 = \boldsymbol{\alpha}' \mathbf{x}_t + \eta_t$
- $\widehat{\boldsymbol{\alpha}} = \left(\sum_{t=1}^{n-1} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^{n-1} \mathbf{x}_t \widetilde{e}_{t+1}^2 \right)$
- $\widehat{\sigma}_n^2 = \widehat{\boldsymbol{\alpha}}' \mathbf{x}_n$
    - Easy, but not constrained to $(0, \infty)$

GARCH Models

- $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_t^2$
- Conditional variance of e_{t+1}
- Specifies conditional variance as function of recent squared innovations
- Large innovations (in magnitude) raise conditional variance
- Lagged variance smooths σ_t^2
- Non-negativity constraints: $\omega > 0$, $\beta \geq 0$, $\alpha > 0$

GARCH with Regressors

- $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_t^2 + \gamma x_t$
- $x_t > 0$ useful to constrain regressor to be positive

Gaussian Quasi-Likelihood

- Assume normality to construct quasi-likelihood
- Let $\theta = (\omega, \beta, \alpha)$. The density of e_{t+1} is

$$f_t(\theta) = \frac{1}{(2\pi\sigma_t^2)^{1/2}} \exp\left(-\frac{e_{t+1}^2}{\sigma_t^2}\right)$$

$$\log f_t(\theta) = \frac{1}{2}\left(\log(2\pi) + \log\left(\sigma_t^2\right) - \frac{e_{t+1}^2}{\sigma_t^2}\right)$$

- Negative log-likelihood

$$\mathcal{L}(\theta) = \sum_{t=0}^{n-1} \log f_t(\theta)$$

- Simple to calculate $\mathcal{L}(\theta)$ numerically
  - First calculate σ_t^2 given θ

Gaussian QMLE

- QMLE $\widehat{\theta}$ minimizes $\mathcal{L}(\theta)$
  - Easy using BFGS or other gradient method
  - Constrained optimization can be used to impose non-negative parameters
- Can write $\mathcal{L}(\theta)$ as a procedure and numerically minimize
  - For each θ
    - ★ Calculate σ_t^2 by recursion $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_t^2$ given σ_0^2
    - ★ Useful to trim $\sigma_t^2 >> 0$
    - ★ If $\sigma_t^2 \leq \sigma_0^2/100$ then set $\sigma_t^2 = \sigma_0^2/100$
    - ★ Calculate $\log f_t(\theta)$ and $\mathcal{L}(\theta)$

Computation (R)

- Use *tseries* package
  - Install under *packages*
  - `library(tseries)`
- `x.arch <- garch(e,order=c(1,1))`
- `x.arch <- garch(e,order=c(1,1),control=garch.control(start=st))`
  - st=starting values
- `archc=coef(x.arch)`
- `sd=predict(x.arch)`
- `like=logLik(x.arch)`
- `help(garch)`

Distribution Theory

- $\sqrt{n}\left(\widehat{\theta}-\theta\right) \rightarrow_d N((0, V)$
- $V = H^{-1}\Omega H^{-1}$
- $H = E\dfrac{\partial^2}{\partial\theta\partial\theta'}\log f_t(\theta)$
- $\Omega = E\dfrac{\partial}{\partial\theta}\log f_t(\theta)\dfrac{\partial}{\partial\theta}\log f_t(\theta)'$

Standard Errors

- $\widehat{H} = \frac{1}{n} \sum_{t=0}^{n-1} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_t(\widehat{\theta}) = \frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta'} \mathcal{L}(\widehat{\theta})$
- $\widehat{\Omega} = \frac{1}{n} \sum_{t=0}^{n-1} \frac{\partial}{\partial\theta} \log f_t(\widehat{\theta}) \frac{\partial}{\partial\theta} \log f_t(\widehat{\theta})'$
- Both can be calculated numerically
- $\widehat{V} = \widehat{H}^{-1} \widehat{\Omega} \widehat{H}^{-1}$
- Standard errors are square roots of diagonal elements of $n^{-1} \widehat{V}$

Model Selection

- Model with 2 ARCH lags and 2 regressors

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha_1 e_t^2 + \alpha_2 e_{t-1}^2 + \gamma_1 x_{1t} + \gamma_2 x_{2t}$$

- How many lags? How many regressors?
- Presence of lagged σ_{t-1}^2 complicates issues
  - β not identified when $\alpha_1 = \alpha_2 = \gamma_1 = \gamma_2 = 0$
  - This means conventional tests and information criterion are not correct when the process is close to constant variance
  - We typically ignore this complication
- Since estimation is nonlinear MLE much of model selection & combination literature is not relevant
  - AIC and TIC are appropriate
  - Unfortunately, not easy to compute with standard packages

AIC and TIC for GARCH models

If model m has parameter vector $\theta(m)$ with $k(m)$ elements

- $AIC(m) = 2\mathcal{L}(\widehat{\theta}(m)) + 2k(m)$
- $TIC(m) = 2\mathcal{L}(\widehat{\theta}(m)) + 2\,\mathrm{tr}\left(\widehat{H}(m)^{-1}\widehat{\Omega}(m)\right)$
- Not standard output

Variance Forecast from GARCH model

- $\sigma^2_{n+1} = \omega + \beta\sigma^2_n + \alpha_1 e^2_n$
- $\widehat{\sigma}^2_{n+1} = \widehat{\omega} + \widehat{\beta}\widehat{\sigma}^2_n + \widehat{\alpha}_1 \widetilde{e}^2_n$
- $\widehat{\sigma}^2_{n+1}$ is estimated conditional variance of y_{n+1}
- Standard deviation $\sqrt{\widehat{\sigma}^2_{n+1}}$

Example: 10-Year Bond Rate

GARCH(1,1)

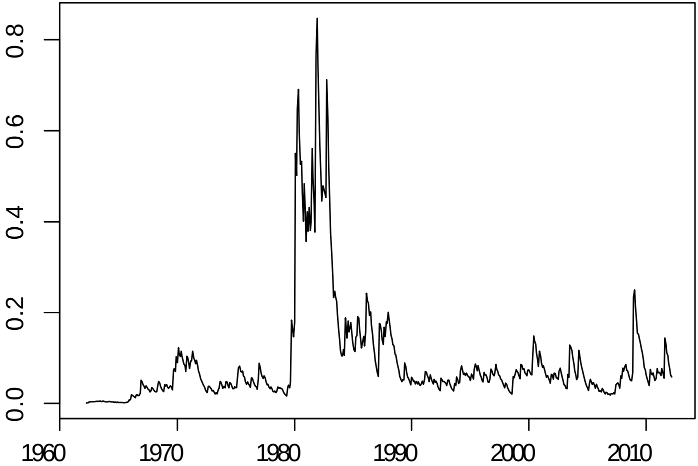
$$\sigma_t^2 = \omega + \alpha e_t^2 + \beta \sigma_{t-1}^2$$

	Estimate	*s.e.*
ω	0.0001	0.0001
α	0.200	0.041
β	0.835	0.025

Variance Forecast

- Conditional variance
  - $\widehat{\sigma}^2_{n+1} = 0.054$
  - $\widehat{\sigma}_{n+1} = 0.23$
- Unconditional
  - $\widehat{\sigma}^2 = 0.076$
  - $\widehat{\sigma} = 0.28$
- The conditional variance at present is similar, but somewhat smaller than the unconditional

Figure: Estimated Variance



Example: GDP Growth

Figure: GDP: Leave-One-Out Prediction Residuals

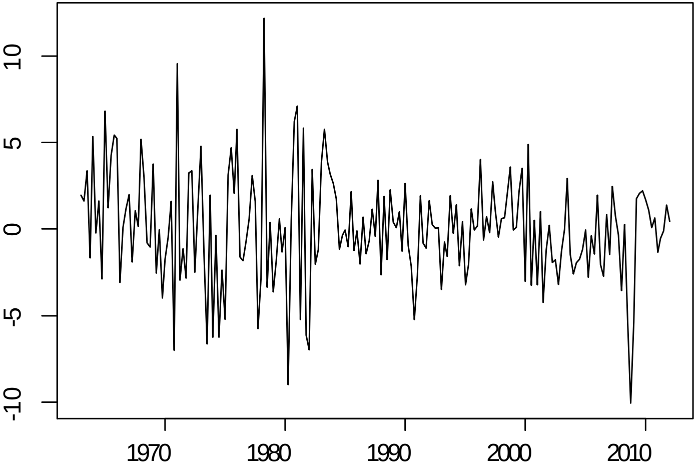
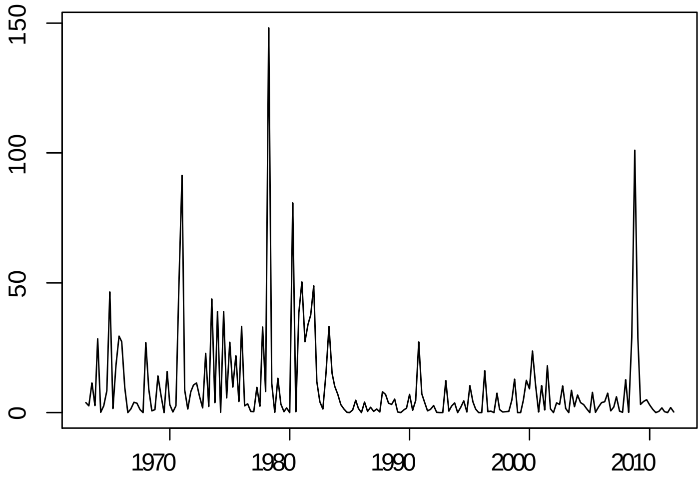


Figure: GDP: Squared Prediction Residuals



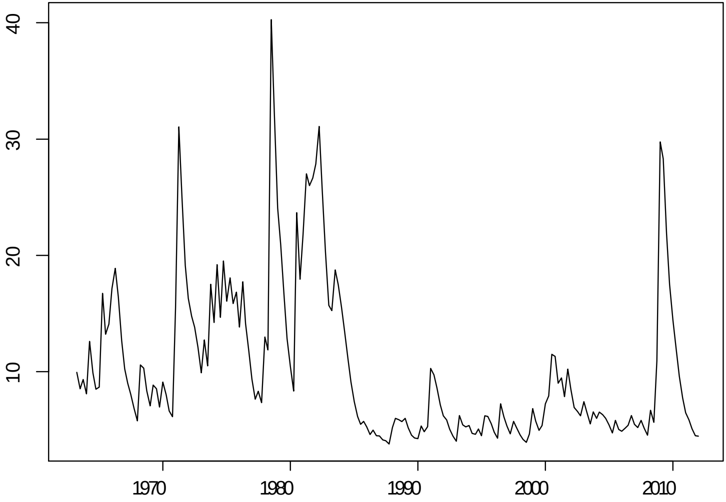
GARCH(1)

$$\sigma_t^2 = \omega + \alpha e_t^2 + \beta \sigma_{t-1}^2$$

	Estimate	*s.e.*
ω	0.81	0.46
α	0.21	0.06
β	0.72	0.06

- Conditional variance
  - $\widehat{\sigma}_{n+1}^2 = 4.1$
  - $\widehat{\sigma}_{n+1} = 2.0$
- Unconditional
  - $\widehat{\sigma}^2 = 9.8$
  - $\widehat{\sigma} = 3.1$

Figure: GDP: Estimated Variance



Assignment 2

- Take your regression models from yesterday
- Calculate forecast weights by cross-validation (CV).
- Use these weights to make a one-step point forecast for July 2012.
- Take the leave-one-out prediction residuals. Estimate a GARCH(1,1) model for the residuals. Calculate a one-step forecast standard deviation from the GARCH model, and compare with the unconditional standard deviation.