Technical Trading Rules

The Econometrics of Predictability

This version: May 7, 2014

May 7, 2014

UNIVERSITY OF OXFORD

Overview

- Technical Trading Rules
  - Filter Rules
  - Moving Average Oscillator
  - Trading Range Break Out
  - Channel Breakout
  - Moving Average Convergence/Divergence
  - Relative Strength Indicator
  - Stochastic Oscillator
  - Simple Momentum
  - On-Balance Volume
- Model Combination

Technical Trading

- Technical trading is one form or predictive modeling
- It is mostly a graphical, rather than statistical tool
- Constructs rules based on price movements
- Rules, while often used graphically, can usually be written down in mathematical expressions
- This can be used to formally allow for testing for technical trading rules
  - Testing the rules is going to be the basis of the assignments this term
  - Using appropriate methodology for evaluation will be important

Data

- Daily DJIA for 12 months
- Use high, low and close
- Compute the rules, but focus on the visualization of the rule
- Rule implementation
  - Red dot is sell
  - Green dot is buy

Filter Rules

Definition ($x\%$ Buy Filter Rule)

A $x\%$ filter rule buys when price has increased by $x\%$ from the previous low, and liquidates when the price has declined $x\%$ from the high measured since the position was opened.

Definition ($x\%$ Sell Filter Rule)

A $x\%$ filter rule sells when price has declined by $x\%$ from the previous high, and liquidates when the price has increased $x\%$ from the low measured since the position was opened.

- These are a momentum rule
- If using both rules with the same percentage, will always have an long or short position, since after a decline of $x\%$, a short is opened, and after a rise of $x\%$ a long is opened

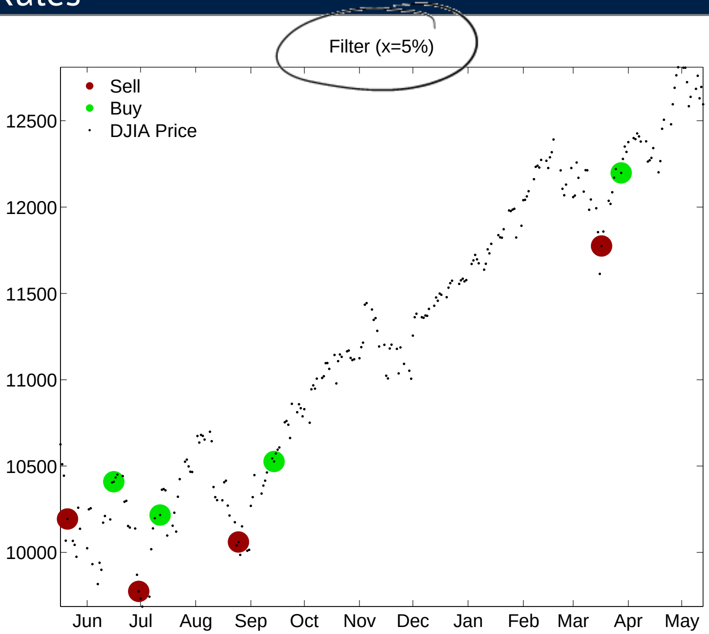
Filter Rules

- A modified rule allows for periods where there is no long or short

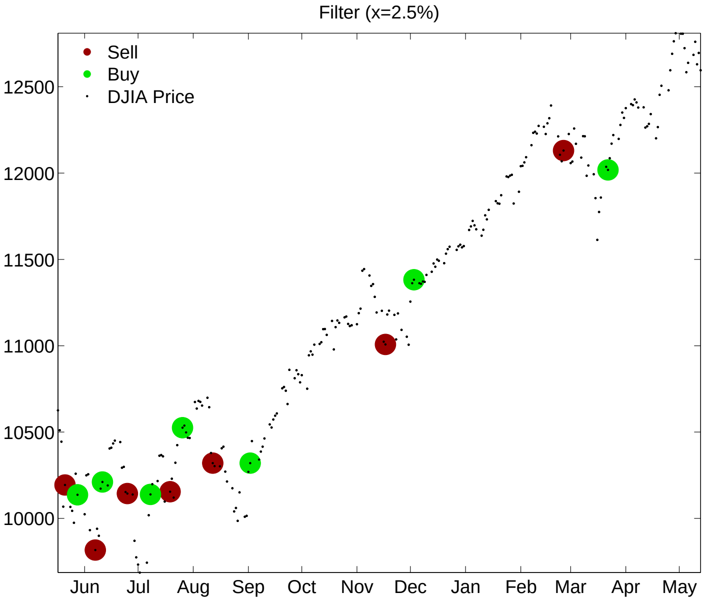
Definition ($x\%/y\%$ Buy Filter Rule)

A $x\%$ filter rule buys when price has moved up by $x\%$ from the previous low, and liquidates when the price has declined $y\%$ from the high measured since the position was opened.

- The sell rule is similarly defined, only using the relative low
- $y \leq x$, and $y = x$ then reduces to previous rules
- Do not have to use both long and short rules


Filter (x=5%)
Sell
Buy
DJIA Price
12500
12000
11500
11000
10500
10000
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May

Filter Rules



Moving-Average Oscillator

Definition (Moving-Average Oscillator)

The moving average oscillator requires two parameters, m and n, $n > m$,

$$MA_t = m^{-1} \sum_{i=t-m+1}^{t} P_i - n^{-1} \sum_{i=t-n+1}^{t} P_i$$

- This is obviously the difference between an m period MA and a n period MA
- Momentum rule
- It is used as an indicator to buy when positive or sell when negative
  - Usually used to initiate a trade when it first crosses, not simply based on sign

Moving-Average Oscillator

UNIVERSITY OF OXFORD

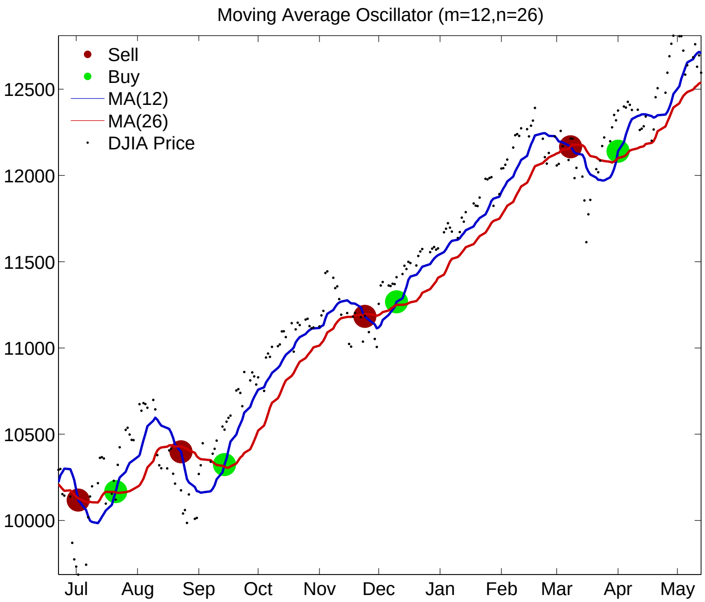
- MA_t is not enough to determine a buy rule, since the direction of the crossing matters
- Formally the buy and sell can be defined as the difference of MA_t

$$\begin{array}{ll} \text{Buy if} & \operatorname{sgn}(MA_t) - \operatorname{sgn}(MA_{t-1}) = 2 \\ \text{Sell if} & \operatorname{sgn}(MA_t) - \operatorname{sgn}(MA_{t-1}) = -2 \end{array}$$

- sgn is the signum function which returns $x/|x|$ for $x \neq 0$ and 0 for $x = 0$

Moving Average Oscillator

UNIVERSITY OF OXFORD



Trading Range Breakout/Support and Resistance

UNIVERSITY OF OXFORD

Definition (Trading Range Breakout)

The trading range break out is takes one parameter, m, and is defined

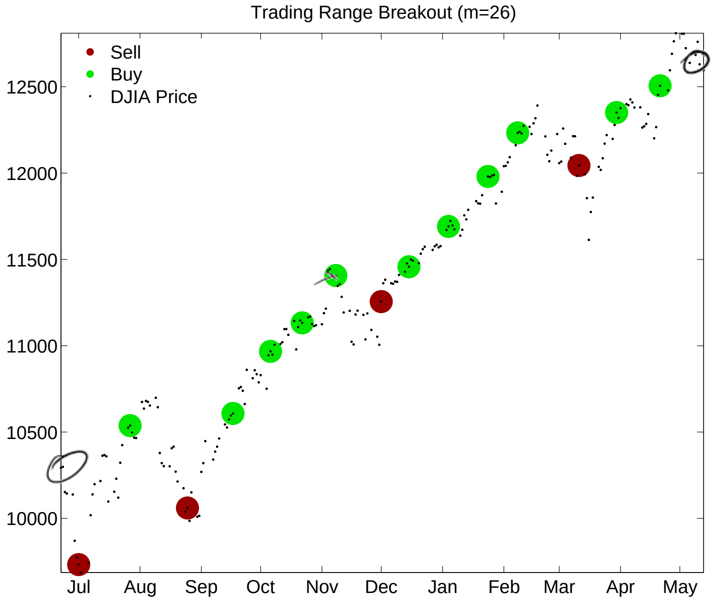
+1

$$TRB_t = \left(P_t > \max\left(\{P_i\}_{i=t-m}^{t-1}\right)\right) - \left(P_t < \min\left(\{P_i\}_{i=t-m}^{t-1}\right)\right)$$

−1

- Positive values (1) indicate that the price is above the m-period *moving maximum*, negative values -1 indicate that it is below the m-period *moving minimum*.
- Momentum rule
- Buy on positive signals, sell on negative signals
- If no signal, then takes the value 0

Trading Range Breakout



Channel Breakout

UNIVERSITY OF OXFORD

Definition (x% Channel Breakout)

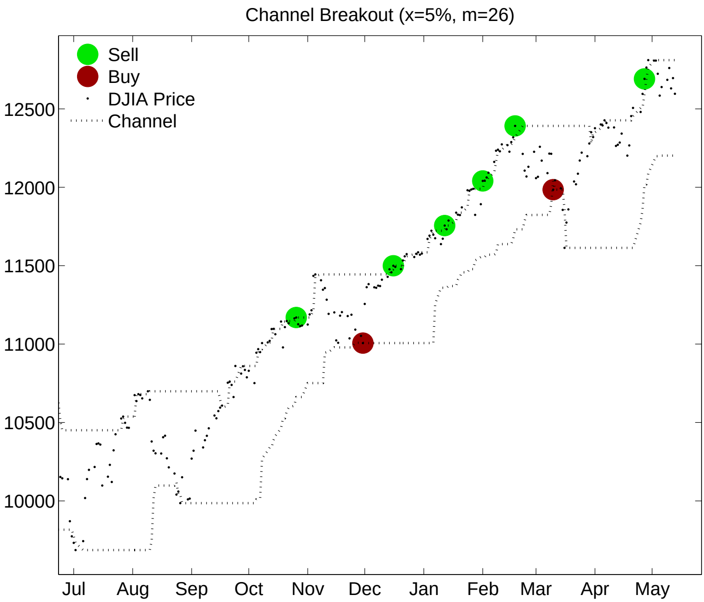
The $x\%$ channel breakout rule, using a m-day channel, is defined

$$\text{Buy if} \quad P_t > \max\left(\{P_i\}_{i=t-m}^{t-1}\right) \cap \frac{\max\left(\{P_i\}_{i=t-m}^{t-1}\right)}{\min\left(\{P_i\}_{i=t-m}^{t-1}\right)} < (1+x)$$

$$\text{Buy if} \quad P_t < \min\left(\{P_i\}_{i=t-m}^{t-1}\right) \cap \frac{\max\left(\{P_i\}_{i=t-m}^{t-1}\right)}{\min\left(\{P_i\}_{i=t-m}^{t-1}\right)} < (1+x)$$

- Momentum rule
- $x\%$ denotes the channel
- Modification of trading range breakout with second condition which may reduce sensitivity to volatility

Channel Range Breakout



Moving Average Convergence/Divergence (MACD)

UNIVERSITY OF OXFORD

Definition (Moving Average Convergence/Divergence (MACD))

The moving-average convergence/divergence indicator takes three parameters, m, n and d, and is defined

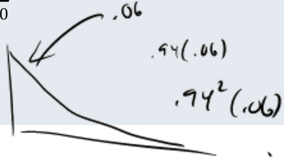
.67

.67(1/3)

.67(1/3)²

$$\delta_t = (1-\lambda_m)\sum_{i=0}^{\infty}\lambda_m^i P_{t-i} - (1-\lambda_n)\sum_{i=0}^{\infty}\lambda_n^i P_{t-i}$$

$$S_t = (1-\lambda_d)\sum_{i=0}^{\infty}\lambda_d^i \delta_t$$

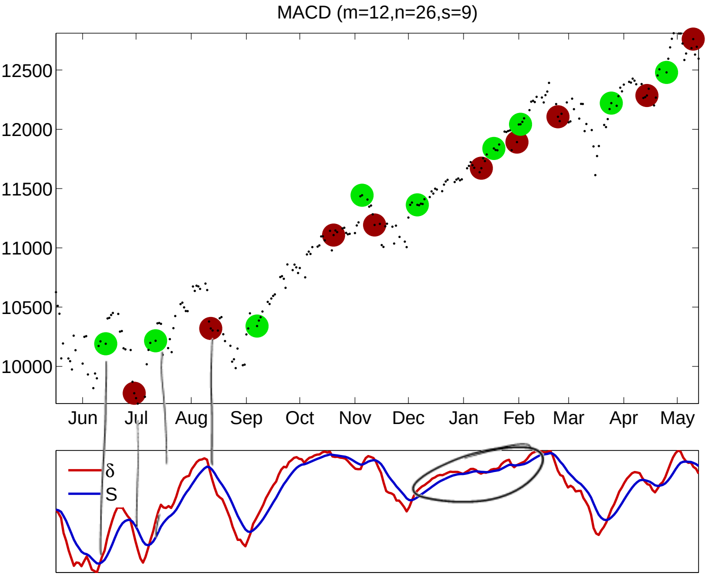


- Pronounced MAK-D
- $\lambda_m = 1 - \frac{2}{m+1}$, $\lambda_n = 1 - \frac{2}{n+1}$,$\lambda_d = 1 - \frac{2}{d+1}$
- S_t is the signal line
- Plot often has δ and S, and a histogram to indicate the difference $\delta_t - S_t$
- Difference is used to predict trends

$$\text{Buy if} \quad \operatorname{sgn}(\delta_t - S_t) - \operatorname{sgn}(\delta_{t-1} - S_{t-1}) = 2$$

$$\text{Sell if} \quad \operatorname{sgn}(\delta_t - S_t) - \operatorname{sgn}(\delta_{t-1} - S_{t-1}) = -2$$

Moving Average Convergence/Divergence



Relative Strength Indicator

UNIVERSITY OF OXFORD

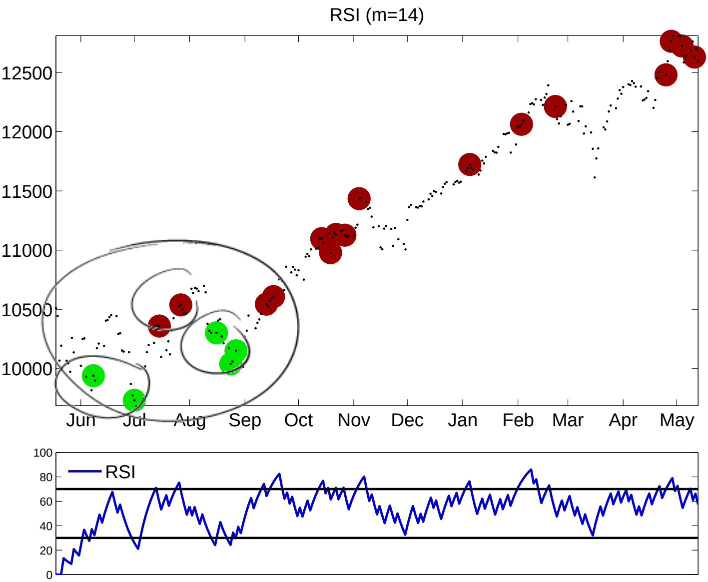
Definition (Relative Strength Indicator)

The relative strength indicator takes one parameter m and is defined as

$$RSI = 100 - \frac{100}{1 + \frac{\sum_{i=0}^{\infty} \lambda^i I_{[(P_{t-i} - P_{t-i-1})>0]}}{\sum_{i=0}^{\infty} \lambda^i I_{[(P_{t-i} - P_{t-i-1})<0]}}}, \quad \lambda = 1 - \frac{2}{m+1}$$

- The core of the indicator are two EWMAs
- Each EWMA is based on indicator variables or positive (top) or negative (bottom) returns
- If all positive, then indicator will equal 100, if all negative, indicator will equal 0
- EWMA can be replaced with MA
- Buy signals are indicated if RSI is *below* some threshold (e.g. 30), sell if *above* a different threshold (e.g. 70)
- RSI is a reversal rule

Relative Strength Indicator (Reversal)



Stochastic Oscillator

Definition (Stochastic Oscillator)

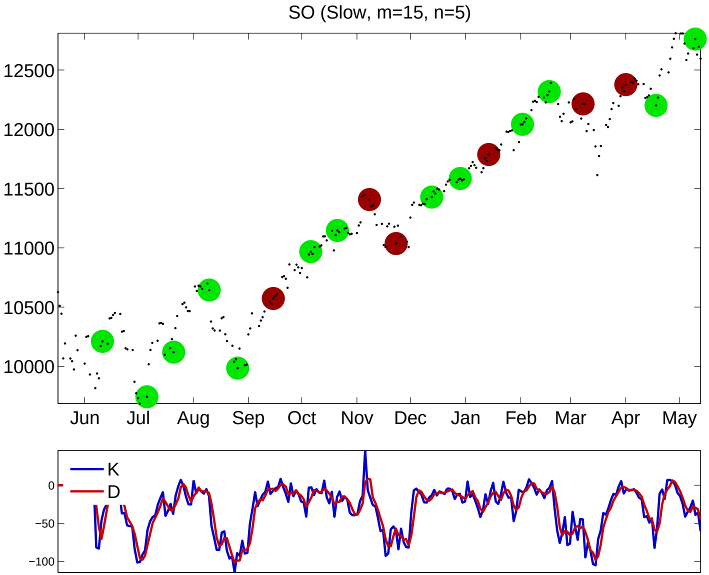
A stochastic oscillator takes two parameters m and n and is defined as

$$\%K_t = 100 \times \frac{P_t - \min\left(\{P_i\}_{i=t-m}^{t-1}\right)}{\max\left(\{P_i\}_{i=t-m}^{t-1}\right) - \min\left(\{P_i\}_{i=t-m}^{t-1}\right)}$$

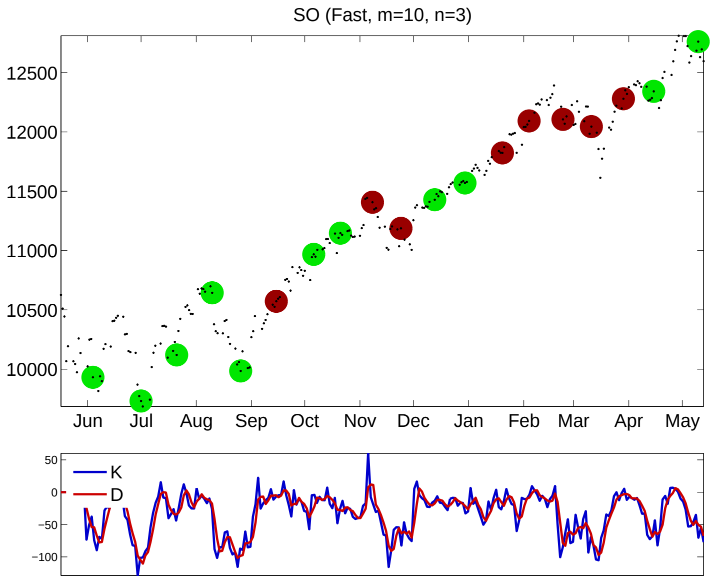
$$\%D_t = \frac{1}{n}\sum_{i=1}^{n} \%K_{t-i+1}$$

- Trading rules are based on intersections of the lines *and* the direction of of the intersection
- If $\%K_{t-1} < \%D_{t-1}$ and $\%K_t > \%D_t$, then a buy signal is indicated
- If $\%K_{t-1} > \%D_{t-1}$ and $\%K_t < \%D_t$, then a sell signal is indicated
- Often implemented using *fast* and *slow* periods, with feedback between the two

Stochastic Oscillator



Stochastic Oscillator



Bollinger Band

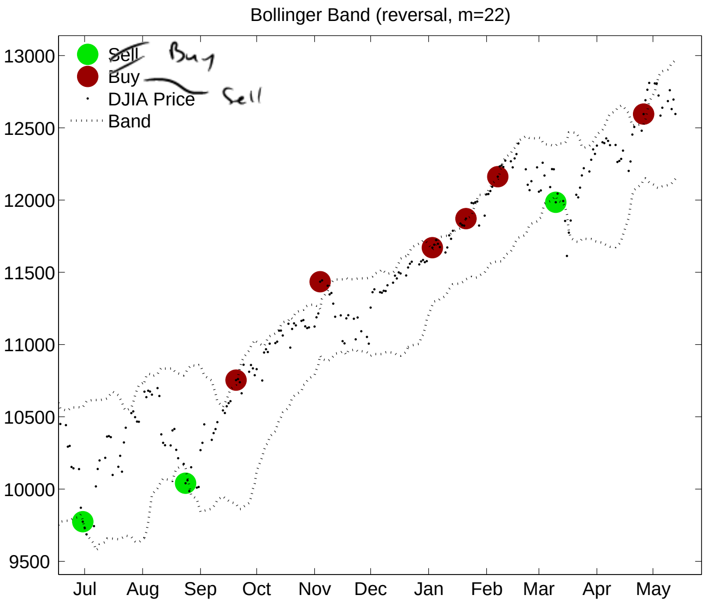
Definition (Bollinger Bands)

Bollinger bands plot the m-day moving average and the MA plus/minus 2 times the m-day moving standard deviation, where the moving averages are defined

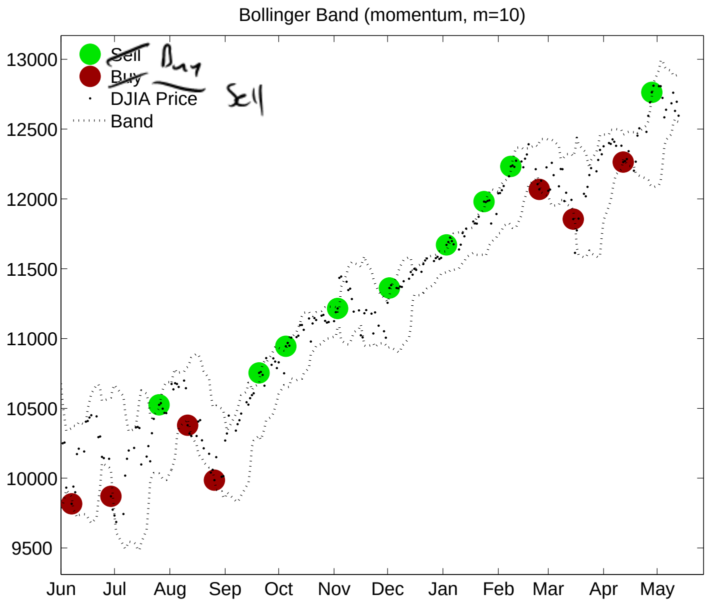
$$MA_t = m^{-1}\sum_{i=1}^{m} P_{t-i+1}, \sigma_t = \sqrt{m^{-1}\sum_{i=1}^{m}\left(\frac{(P_{t-i+1} - P_{t-i})}{P_{t-i}}\right)^2}$$

- Rules can be based on prices leaving the bands, and possibly then crossing of the moving average
- For example, buy when price hit bottom (reversal) and then sell when it hits the MA
- Alternatively buy when it hits the top (strong upward trend)

Bollinger Band



Bollinger Band



A Simple Momentum Rule

- Momentum is a common strategy
- Can construct a momentum rule as

$$S_t = \begin{cases} 1 & \text{if } P_t > P_{t-d} \\ 0 & \text{if } P_t \leq P_{t-d} \end{cases}$$

- Technically (trivial) moving average rule with d-day delay filter

UNIVERSITY OF OXFORD

Definition (On-Balance Volume)

On-Balance Volume (OBV) plots the difference between moving averages of signed daily volume, defined

$$OBV_t = \sum_{s=1}^{t} VOL_s D_s$$

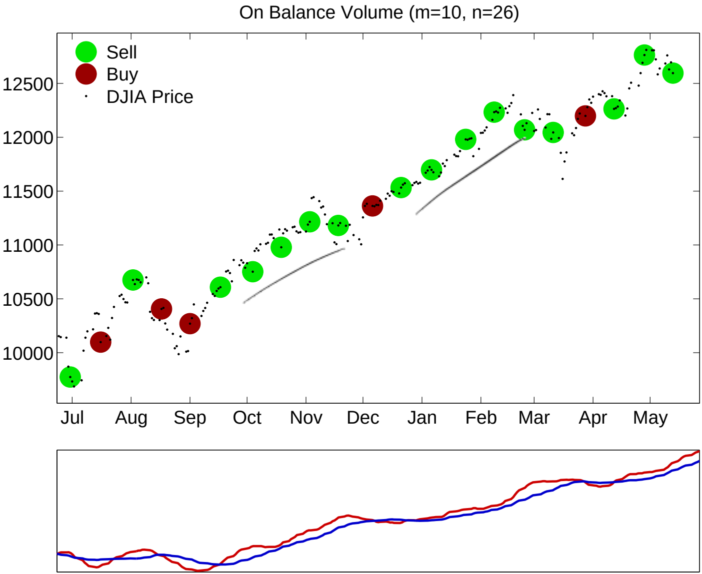
where VOL_s is the volume in period s, D_s is a dummy which is 1 if $P_t > P_{t-1}$ and -1 otherwise, and the trading signal is

$$S_t = \begin{cases} 1 & MA_{m,t}^{OBV} > MA_{n,t}^{OBV} \\ 0 & MA_{m,t}^{OBV} \leq MA_{n,t} \end{cases}$$

where $MA_{q,t}^{OBV} = q^{-1} \sum_{i=1}^{q} OBV_{t-i-1}$, $q = m, n$, $m < n$.

- Most rules make use of price signals
- OBV mixes volume information with indicator variable

On-Balance Volume



Additional Filters

- Many ways rules can be modified
- MAs and EWMAs can be swapped
- Can use a d-day delay filter to stagger execution of trade from signal
- Can use $b\%$-band with some filters to reduce frequency of execution
  - Requires the price price (or fast signal) to be $b\%$ above the band (or slow signal)
  - Relevant for most rules
  - Examples
    - Moving-Average Oscillator: Requires fast MA to be larger than $1 + b$ times slow for a buy signal, and smaller than $1 - b$ for a sell signal
    - Trading Range Breakout/Channel Breakout: Use $1 + b$ times max and $1 - b$ times min
- Can use k-day holding period, so that positions are held for k-days and other signal are ignored

From Technical Indicators to Trading Rules

UNIVERSITY OF OXFORD

- Most technical rules are interpreted as buy, neutral or sell – 1, 0 or -1
- Essentially applies a step function to the trading signal
- Can use a other continuous, monotonic increasing functions, although not clear which ones
- One options is to run a regression

$$r_{t+1} = \beta_0 + \beta_1 S_t + \epsilon_t$$

- S_t is a signal is computed using information up-to and including t
  - Can be discrete or continuous
- Maps to an expected return, which can then be used in Sharpe-optimization

Combining Multiple Technical Indicators

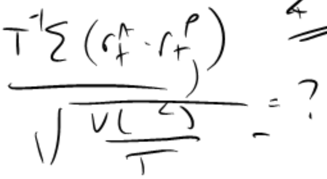
- Technical trading rules can be combined
- Not obvious how to combine when discrete
- Method 1: Majority vote
  - Count number of rules with signs 1, 0 or -1
- Method 2: Aggregation
  - Compute sum of indicators divided by number of indicators

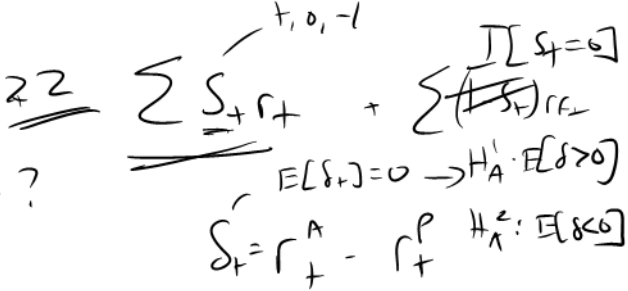
$$\tilde{S}_t = \frac{\sum_{i=1}^{k} S_{k,t}}{k}$$

    and go long/short $\tilde{S}_t$
  - Bound by 100% long and 100% short

Evaluating the Rules

- Obvious strategy it to look at returns, conditional on signal
- Important to have a benchmark model
  - Often buy and hold, or some other much less dynamic strategy
- Obvious test is t-statistic of difference in mean return between the active strategy and the benchmark
- Can also examine predictability for other aspects of distribution
  - Volatility
  - Large declines





Brock, Lakonishok and LeBaron

- One of the first systematically test trading rules
- Focused on two rules:
  - Moving Average Oscillator
  - Trading Range Breakout
- (Controversially) documented evidence of excess returns to technical trading rules
- Returns were large enough to cover transaction costs

$$\frac{1}{16\,000}$$

Moving Average Oscillator

- Moving Average Oscillators implemented for
  - $m = 1, n = 50$
  - $m = 1, n = 150$
  - $m = 5, n = 150$
  - $m = 1, n = 200$
  - $m = 2, n = 200$
- Use both the standard rule and one with a 1%-band filter
- Standard is implemented by taking the position and holding for 10 days, ignoring all other signals
- $b\%$-band version:
  - Requires an exceedence by 1% of the slow MA, but no crossing

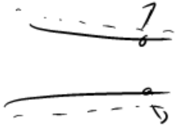
$$\text{Buy if} \quad \left(\frac{MA_t}{n^{-1} \sum_{i=t-n+1}^{t} P_i} \right) > \frac{b}{100}, \ \text{Sell if} \quad \left(\frac{MA_t}{n^{-1} \sum_{i=t-n+1}^{t} P_i} \right) < -\frac{b}{100}$$

  - If $b > 0$ then some days may have no signal
  - If $b = 0$ then all days are buys or sells

Trading Range Breakout

- Trading range breakout is implemented for
  - $m = 50$
  - $m = 100$
  - $m = 150$
- Implemented using the standard and with a 1% band
- b% band version is

$$TRB_t = \left(P_t > \left(1 + \frac{b}{100}\right) \max\left(\{P_i\}_{i=t-m}^{t-1}\right)\right) - \left(P_t < \left(1 - \frac{b}{100}\right) \min\left(\{P_i\}_{i=t-m}^{t-1}\right)\right)$$



Empirical Application

UNIVERSITY OF OXFORD

- A total of 26 rules are created
  - MAO: 5 $(m, n)\times$ 2 (Fixed or Variable Window) $\times 2$ $(b = 0, .01)$
  - TRB: 3 (m) $\times$ 2 $(b = 0, .01)$
- DJIA from 1897 until 1986
- Main result is that there appears to be predictability using these rules
- Strongest results were for the fixed windows MAO with $m = 1$, $n = 200$ and $b = .01$
- TRB with $m = 150$ and $b = .01$ also had a strong result
- Report
  - Number of buy and sell signals
  - Mean return during buy and sell signals
  - Probability of positive return for buy and sell signals
  - Mean return of a portfolio which both buys and sells

Moving Average Oscillator, Variable Length

Period	Test	N(Buy)	N(Sell)	Buy	Sell	Buy > 0	Sell > 0	Buy-Sell
1897–1986	(1, 50, 0)	14240	10531	0.00047	−0.00027	0.5387	0.4972	0.00075
				(2.68473)	(−3.54645)			(5.39746)
	(1, 50, 0.01)	11671	8114	0.00062	−0.00032	0.5428	0.4942	0.00094
				(3.73161)	(−3.56230)			(6.04189)
	(1, 150, 0)	14866	9806	0.00040	−0.00022	0.5373	0.4962	0.00062
				(2.04927)	(−3.01836)			(4.39500)
	(1, 150, 0.01)	13556	8534	0.00042	−0.00027	0.5402	0.4943	0.00070
				(2.20929)	(−3.28154)			(4.68162)
	(5, 150, 0)	14858	9814	0.00037	−0.00017	0.5368	0.4970	0.00053
				(1.74706)	(−2.61793)			(3.78784)
	(5, 150, 0.01)	13491	8523	0.00040	−0.00021	0.5382	0.4942	0.00061
				(1.97876)	(−2.78835)			(4.05457)
	(1, 200, 0)	15182	9440	0.00039	−0.00024	0.5358	0.4962	0.00062
				(1.93865)	(−3.12526)			(4.40125)
	(1, 200, 0.01)	14105	8450	0.00040	−0.00030	0.5384	0.4924	0.00070
				(2.01907)	(−3.48278)			(4.73045)
	(2, 200, 0)	15194	9428	0.00038	−0.00023	0.5351	0.4971	0.00060
				(1.87057)	(−3.03587)			(4.26535)
	(2, 200, 0.01)	14090	8442	0.00038	−0.00024	0.5368	0.4949	0.00062
				(1.81771)	(−3.03843)			(4.16935)

Moving Average Oscillator, Fixed Length

Test	N(Buy)	N(Sell)	Buy	Sell	Buy > 0	Sell > 0	Buy-Sell
(1, 50, 0)	340	344	0.0029	−0.0044	0.5882	0.4622	0.0072
			(0.5796)	(−3.0021)			(2.6955)
(1, 50, 0.01)	313	316	0.0052	−0.0046	0.6230	0.4589	0.0098
			(1.6809)	(−3.0096)			(3.5168)
(1, 150, 0)	157	188	0.0066	−0.0013	0.5987	0.5691	0.0079
			(1.7090)	(−1.1127)			(2.0789)
(1, 150, 0.01)	170	161	0.0071	−0.0039	0.6529	0.5528	0.0110
			(1.9321)	(−1.9759)			(2.8534)
(5, 150, 0)	133	140	0.0074	−0.0006	0.6241	0.5786	0.0080
			(1.8397)	(−0.7466)			(1.8875)
(5, 150, 0.01)	127	125	0.0062	−0.0033	0.6614	0.5520	0.0095
			(1.4151)	(−1.5536)			(2.1518)
(1, 200, 0)	114	156	0.0050	−0.0019	0.6228	0.5513	0.0069
			(0.9862)	(−1.2316)			(1.5913)
(1, 200, 0.01)	130	127	0.0058	−0.0077	0.6385	0.4724	0.0135
			(1.2855)	(−2.9452)			(3.0740)
(2, 200, 0)	109	140	0.0050	−0.0035	0.6330	0.5500	0.0086
			(0.9690)	(−1.7164)			(1.9092)
(2, 200, 0.01)	117	116	0.0018	−0.0088	0.5556	0.4397	0.0106
			(0.0377)	(−3.1449)			(2.3069)

Trading Range Breakout

UNIVERSITY OF OXFORD

Test	N(Buy)	N(Sell)	Buy	Sell	Buy > 0	Sell > 0	Buy-Sell
(1, 50, 0)	722	415	0.0050	0.0000	0.5803	0.5422	0.0049
			(2.1931)	(−0.9020)			(2.2801)
(1, 50, 0.01)	248	252	0.0082	−0.0008	0.6290	0.5397	0.0090
			(2.7853)	(−1.0937)			(2.8812)
(1, 150, 0)	512	214	0.0046	−0.0030	0.5762	0.4953	0.0076
			(1.7221)	(−1.8814)			(2.6723)
(1, 150, 0.01)	159	142	0.0086	−0.0035	0.6478	0.4789	0.0120
			(2.4023)	(−1.7015)			(2.9728)
(1, 200, 0)	466	182	0.0043	−0.0023	0.5794	0.5000	0.0067
			(1.4959)	(−1.4912)			(2.1732)
(1, 200, 0.01)	146	124	0.0072	−0.0047	0.6164	0.4677	0.0119
			(1.8551)	(−1.9795)			(2.7846)
Average			0.0063	−0.0024			0.0087

The Standard Forecasting Model

- Standard forecasts are also popular for predicting economic variables
- Generically expressed

$$y_{t+1} = \beta_0 + \mathbf{x}_t \boldsymbol{\beta} + \epsilon_{t+1}$$

- $\mathbf{x}_t$ is a 1 by k vector of predictors ($k = 1$ is common)
- Includes both exogenous regressors such as the term or default premium and also autoregressive models
- Forecasts are $\hat{y}_{t+1|t}$

The forecast combination problem

UNIVERSITY OF OXFORD

- Two level of aggregation in the combination problem

1. Summarize individual forecasters' private information in point forecasts $\hat{y}_{t+h,i|t}$
   - Highlights that "inputs" are not the usual explanatory variables, but forecasts
2. Aggregate individual forecasts into consensus measure $C\left(\mathbf{y}_{t+h|t}, \mathbf{w}_{t+h|t}\right)$

- Obvious competitor is the "super-model" or "kitchen-sink" – a model built using all information in each forecasters information set
- Aggregation should increase the bias in the forecast relative to SM but may reduce the variance
- Similar to other model selection procedures in this regard

$$Y_{t+1} = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$= \left[\hat{\beta}_0 - \beta_0 + \hat{\beta}_1 X_t - \beta_1 X_t\right] + \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$(\hat{\beta}_1 - \beta_1) X_t$$

Why not use the "Super Model"

UNIVERSITY OF OXFORD

- Could consider pooling information sets

$$\mathcal{F}_t^c = \cup_{i=1}^n \mathcal{F}_{t,i}$$

- Would contain all information available to all forecasters
- Could construct consensus directly $C\left(\mathcal{F}_t^c; \boldsymbol{\theta}_{t+h|t}\right)$
- Some reasons why this may not work
  - Some information in individuals information sets may be qualitative, and so expensive to quantitatively share
  - Combined information sets may have a very high dimension, so that finding the best super model may be hard
    - Potential for lots of estimation error
- Classic bias-variance trade-off is main reason to consider forecasts combinations over a super model
  - Higher bias, lower variance

Linear Combination under MSE Loss

- Models can be combined in many ways for virtually any loss function
- Most standard problem is for MSE loss using only linear combinations
- I will suppress time subscripts when it is clear that it is $t+h|t$
- Linear combination problem is

$$\min_{\mathbf{w}} \mathrm{E}\left[e^2\right] = \mathrm{E}\left[\left(y_{t+h} - \mathbf{w}'\hat{\mathbf{y}}\right)^2\right]$$

- Requires information about first 2 moments of he joint distribution of the realization y_{t+h} and the time-t forecasts $\hat{\mathbf{y}}$

$$\left[\begin{array}{c} y_{t+h|t} \\ \hat{\mathbf{y}} \end{array}\right] \sim F\left(\left[\begin{array}{c} \mu_y \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}} \end{array}\right], \left[\begin{array}{cc} \sigma_{yy} & \boldsymbol{\Sigma}'_{y\hat{\mathbf{y}}} \\ \boldsymbol{\Sigma}_{y\hat{\mathbf{y}}} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} \end{array}\right]\right)$$

Linear Combination under MSE Loss

UNIVERSITY OF OXFORD

- The first order condition for this problem is

$$\frac{\partial \mathrm{E}\left[e^2\right]}{\partial \mathbf{w}} = -\mu_y \boldsymbol{\mu}_{\hat{\mathbf{y}}} + \boldsymbol{\mu}_{\hat{\mathbf{y}}}\boldsymbol{\mu}'_{\hat{\mathbf{y}}}\mathbf{w} + \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\mathbf{w} - \Sigma_{y\hat{\mathbf{y}}} = \mathbf{0}$$

- The solution to this problem is

$(X'X)^{-1}$

$$\mathbf{w}^\star = \left(\boldsymbol{\mu}_{\hat{\mathbf{y}}}\boldsymbol{\mu}'_{\hat{\mathbf{y}}} + \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\right)^{-1}\left(\Sigma_{y\hat{\mathbf{y}}} + \mu_y \boldsymbol{\mu}_{\hat{\mathbf{y}}}\right)$$

- Similar to the solution to the OLS problem, only with extra terms since the forecasts may not have the same conditional mean

Linear Combination under MSE Loss

- Can remove the conditional mean if the combination is allowed to include a constant, w_c

$$\begin{aligned} w_c &= \mu_y - \mathbf{w}^\star \boldsymbol{\mu}_{\hat{\mathbf{y}}} \\ \mathbf{w}^\star &= \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1} \boldsymbol{\Sigma}_{y\hat{\mathbf{y}}} \end{aligned}$$

- These are identical to the OLS where w_c is the intercept and $\mathbf{w}^*$ are the slope coefficients
- The role of w_c is the correct for any biases so that the squared bias term in the MSE is 0

$$\mathrm{MSE}[e] = \mathrm{B}[e]^2 + \mathrm{V}[e]$$

Understanding the Diversification Gains

- Simple setup

$$e_1 \sim F_1\left(0, \sigma_1^2\right), \; e_2 \sim F_2\left(0, \sigma_2^2\right), \; \mathrm{Corr}\left[e_1, e_2\right] = \rho, \; \mathrm{Cov}\left[e_1 e_2\right] = \sigma_{12}$$

- Assume $\sigma_2^2 \leq \sigma_1^2$
- Assume weights sum to 1 so that $w_1 = 1 - w_2$ (Will suppress the subscript and simply write w)
- Forecast error is then

$$y - w\hat{y}_1 - (1 - w)\,\hat{y}_2$$

- Error is given by

$$e^c = we_1 + (1 - w)\,e_2$$

- Forecast has mean 0 and variance

$$w^2\sigma_1^2 + (1 - w)^2\,\sigma_2^2 + 2w\,(1 - w)\,\sigma_{12}$$

Understanding the Diversification Gains

UNIVERSITY OF OXFORD

- The optimal w can be solved by minimizing this expression, and is

$$w^\star = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \; 1 - w^\star = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

- Intuition is that the weight on a model is higher the
  - Larger the variance of the other model
  - Lower the correlation between the models
- 1 weight will be larger than 1 if $\rho \geq \frac{\sigma_2}{\sigma_1}$
- Weights will be equal if $\sigma_1 = \sigma_2$ for any value of correlation
  - Intuitively this must be the case since model 1 and 2 are indistinguishable from a MSE point-of-view
  - When will "optimal" combinations out-perform equally weighted combinations? Any time $\sigma_1 \neq \sigma_2$
- If $\rho = 1$ then only select model with lowest variance (mathematical formulation is not well posed in this case)

Constrained weights

- The previous optimal weight derivation did not impose any restrictions on the weights
- In general some of the weights will be negative, and some will exceed 1
- Many combinations are implemented in a relative, constrained scheme

$$\min_{\mathbf{w}} \mathrm{E}\left[e^2\right] = \mathrm{E}\left[\left(y_{t+h} - \mathbf{w}'\hat{\mathbf{y}}\right)^2\right] \text{ subject to } \mathbf{w}'\boldsymbol{\iota} = 1$$

- The intercept is omitted (although this isn't strictly necessary)
- If the biases are all 0, then the solution is dual to the usual portfolio minimization problem, and is given by

$$\mathbf{w}^{\star} = \frac{\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\boldsymbol{\iota}}$$

- This solution is the same as the Global Minimum Variance Portfolio

Combinations as Hedge against Structural Breaks

- One often cited advantage of combinations is (partial) robustness to structural breaks
- Best case is if two positively correlated variables have shifts in opposite directions
- Combinations have been found to be more stable than individual forecasts
  - This is mostly true for static combinations
  - Dynamic combinations can be unstable since some models may produce large errors from time-to-time

Weight Estimation

- All discussion has focused on "optimal" weights, which requires information on the mean and covariance of both y_{t+h} and $\hat{\mathbf{y}}_{t+h|t}$
  - This is clearly highly unrealistic
- In practice weights must be estimated, which introduces extra estimation error
- Theoretically, there should be no need to combine models when all forecasting models are generated by the econometrician (e.g. when using $\mathcal{F}^c$)
- In practice, this does not appear to be the case
  - High dimensional search space for "true" model
  - Structural instability
  - Parameter estimation error
  - Correlation among predictors

Clemen (1989): "Using a combination of forecasts amounts to an admission that the forecaster is unable to build a properly specified model"

Weight Estimation

UNIVERSITY OF OXFORD

- Whether a combination is needed is closely related to forecast encompassing tests
- Model averaging can be thought of a method to avoid the risk of model selection
  - Usually important to consider models with a wide range of features and many different model selection methods
- Has been consistently documented that *prescreening* models to remove the worst performing is important before combining
- One method is to use the SIC to remove the worst models
  - Rank models by SIC, and then keep the x% best
- Estimated weights are usually computed in a 3rd step in the usual procedure
  - R: Regression
  - P: Prediction
  - S: Combination estimation
  - $T = P + R + S$
- Many schemes have been examined

Weight Estimation

- Standard least squares with an intercept

$$y_{t+h} = w_0 + \mathbf{w}'\hat{\mathbf{y}}_{t+h|t} + \epsilon_{t+h}$$

- Least squares without an intercept

$$y_{t+h} = \mathbf{w}'\hat{\mathbf{y}}_{t+h|t} + \epsilon_{t+h}$$

- Linearly constrained least squares

$$y_{t+h} - \hat{y}_{t+h,n|t} = \sum_{i=1}^{n-1} w_i \left(\hat{y}_{t+h,i|t} - \hat{y}_{t+h,n|t}\right) + \epsilon_{t+h}$$

  - This is just a constrained regression where $\sum w_i = 1$ has been implemented where $w_n = 1 - \sum_{i=1}^{n-1} w_i$
  - Imposing this constraint is thought to help when the forecast is persistent

$$e^c_{t+h|t} = -w_0 + \left(1 - \mathbf{w}'\boldsymbol{\iota}\right) y_{t+h} + \mathbf{w}'\mathbf{e}_{t+h|t}$$

  - $\mathbf{e}_{t+h|t}$ are the forecasting errors from the n models
  - Only matters if the forecasts may be biased

Weight Estimation

- Constrained least squares

$$y_{t+h} = \mathbf{w}'\hat{\mathbf{y}}_{t+h|t} + \epsilon_{t+h} \text{ subject to } \mathbf{w}'\boldsymbol{\iota}=1,\ w_i \geq 0$$

  - This is not a standard regression, but can be easily solved using quadratic programming (MATLAB `quadprog`)
- Forecast combination where the covariance of the forecast errors is assumed to be diagonal
  - Produces weights which are all between 0 and 1
  - Weight on forecast i is

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}}$$

  - May be far from optimal if ρ is large
  - Protects against estimator error in the covariance

Weight Estimation

UNIVERSITY OF OXFORD

- Median
  - Can use the median rather than the mean to aggregate
  - Robust to outliers
  - Still suffers from not having any reduction in parameter variance in the actual forecast
- Rank based schemes
  - Weights are inversely proportional to model's rank

$$w_i = \frac{\mathcal{R}^{-1}_{t+h,i|t}}{\sum_{j=1}^{n} \mathcal{R}^{-1}_{t+h,j|t}}$$

  - Highest weight to best model, ratio of weights depends only on relative ranks
  - Places relatively high weight on top model
- Probability of being the best model-based weights
  - Count the proportion that model i outperforms the other models

$$\begin{aligned} p_{t+h,i|t} &= T^{-1} \sum_{t=1}^{T} \cap_{j=1, j\neq i}^{n} I\left[L\left(e_{t+h,i|t}\right) < L\left(e_{t+h,j|t}\right)\right] \\ y^{c}_{t+h|t} &= \sum_{i=1}^{n} p_{t+h,i|t}\hat{y}_{t+h,i|t} \end{aligned}$$

Weight Estimation

- Time-varying weights
  - These are ultimately based off of multivariate ARCH-type models
  - Most common is EWMA of past forecast errors outer-products
  - Often enforced that covariances are 0 so that combinations have only non-negative weights
  - Can be implemented using rolling-window based schemes as well, both with and without a 0 correlation assumption
  - Time-varying weights are thought to perform poorly when the DGP is stable since they place higher weight on models than a non-time varying scheme and so lead to more parameter estimation error

Broad Recommendations

- Simple combinations are difficult to beat
  - $1/n$ often outperforms estimated weights
  - Constant usually beat dynamic
  - Constrained outperform unconstrained (when using estimated weights)
- Not combining and using the best fitting performs worse than combinations – often substantially
- Trimming bad models prior to combining improves results
- Clustering similar models (those with the highest correlation of their errors) *prior* to combining leads to better performance, especially when estimating weights
  - Intuition: Equally weighted portfolio of models with high correlation, weight estimation using a much smaller set with lower correlations
- Shrinkage improves weights when estimated
- If using dynamic weights, shrink towards static weights

Equal Weighting

- Equal weighting is hard to beat when the variance of the forecast errors are similar
- If the variance are highly heterogeneous, varying the weights is important
  - If for nothing else than to down-weight the high variance forecasts
- Equally weighted combinations are thought to work well when models are unstable
  - Instability makes finding "optimal" weights very challenging
- Trimmed equally-weighted combinations appear to perform better than equally weighted, at least if there are some very poor models
  - May be important to trim both "good" and "bad" models (in-sample performance)
    - Good models are over-fit
    - Bad models are badly mis-specified

Shrinkage Methods

- Linear combination

$$\hat{y}^{c}_{t+h|t} = \mathbf{w}'\hat{\mathbf{y}}_{t+h|t}$$

  Standard least squares estimates of combination weights are very noisy
- Often found that "shrinking" the weights toward a *prior* improves performance
- Standard prior is that $w_i = \frac{1}{n}$
- However, do not want to be *dogmatic* and so use a distribution for the weights
- Generally for an arbitrary *prior weight* $\mathbf{w}_0$,

$$\mathbf{w}|\tau^2 \sim N(\mathbf{w}_0, \mathbf{\Omega})$$

- $\mathbf{\Omega}$ is a correlation matrix and τ^2 is a parameter which controls the amount of shrinkage

Shrinkage Methods

- Leads to a weighted average of the prior and data

$$\bar{\mathbf{w}} = \left(\boldsymbol{\Omega} + \hat{\mathbf{y}}'\hat{\mathbf{y}}\right)^{-1}\left(\boldsymbol{\Omega}\mathbf{w}_0 + \hat{\mathbf{y}}'\hat{\mathbf{y}}\hat{\mathbf{w}}\right)$$

- $\hat{\mathbf{w}}$ is the usual least squares estimator of the optimal combination weight
- If $\boldsymbol{\Omega}$ is very large compared to $\mathbf{y}'\mathbf{y} = \sum_{t=1}^{T}\mathbf{y}_{t+h|t}\mathbf{y}'_{t+h|t}$ then $\bar{\mathbf{w}} \approx \mathbf{w}_0$
- On the other hand, if $\mathbf{y}'\mathbf{y}$ dominates, then $\bar{\mathbf{w}} \approx \hat{\mathbf{w}}$
- Other implementation use a g-prior, which is scalar

$$\bar{\mathbf{w}} = \left(g\hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{y}}'\hat{\mathbf{y}}\right)^{-1}\left(g\hat{\mathbf{y}}'\hat{\mathbf{y}}\mathbf{w}_0 + \hat{\mathbf{y}}'\hat{\mathbf{y}}\hat{\mathbf{w}}\right)$$

- Large values of $g \geq 0$ least to large amounts of shrinkage
- 0 corresponds to OLS

$$\bar{\mathbf{w}} = \mathbf{w}_0 + \frac{\hat{\mathbf{w}} - \mathbf{w}_0}{1+g}$$