How can the yield curve predict an economic recession?

Ge Desheng

Peking University HSBC Business School

July 1, 2020

Outline

Contents

Introduction

Yield curve inversion

- long-term yield $<$ short-term yield
- the most efficient recession predictor in the last decades.

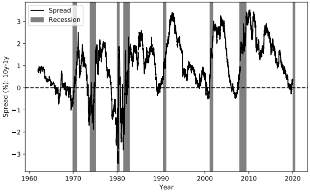


Figure: Term spread of (1y, 10y)

Related theories:

- Expectations hypothesis: Market participants' anticipation of decrease in short-term yields $\rightarrow$ lower long-term rates (Moneta, 2005).
- Tightened Monetary policy $\rightarrow$ flat curve (Estrella & Mishkin, 1997)
- Stable income preference. Buy long-term bond for pay-off in case of income reduction $\rightarrow$ Decline in long-term yields. (Harvey, 1988).
- No consensus on the channels yet.

This paper:

- Reconfirms and improves the predictive power of the yield curve.

Related literature

Continuous regression

- Predictor for output growth (Harvey, 1988, 1989; Stock & Watson, 1989) and inflation (Mishkin, 1990; Estrella, 2005).
- Decomposition of term spread. Term premia: Hamilton and Kim (2002); Short-term yield: Ang et al. (2006). Neither dominates.

Binary classification (this paper)

- Recession binary in Probit regression (Wright, 2006; Estrella & Trubin, 2006)
- Better than continuous variables (e.g., GDP growth rate). (Evgenidis & Siriopoulos, 2014)

Choice of yield rates

- [3m, 10y]: Chinn and Kucko (2015), Estrella and Hardouvelis (1991).
- [1y, 5y]: Mishkin (1990).

Research motivation

Is term spread the best indicator among all information from yield curve?

- Use machine learning algorithm to choose the best features from the whole yield curve.

Can we find "general" spread instead of the naïve one?

- Naïve inversion: long - short < 0 (widely used in literature)
- General form (linear combination): a$\times$long - b$\times$short$+$c $<$0.

Model performance evaluation under the machine learning framework.

Contents

Standard procedure of machine learning

Feature Selection

- Logistic regression with L1 regularization
- Rank of feature importance from random forest

Classifier Estimation

- Split the estimation sample: training vs. testing set.
- Logistic regression.
- Bayesian treatment and Laplace approximation.

Performance Evaluation

Confusion matrix

Prediction

Out-of-sample prediction

Logistic or Probit regression

- Inputs: Yield curve rates x
- Entry: $z_i = \sum_{j=1}^{n} w_j x_{ij} + w_0$.
- Activation function: $\phi(z)$. Probability of positive label.
- Output: 1 if $\phi(z) >$ threshold, otherwise 0.
- Choice of ϕ (Logistic or Probit)
- Linear classifier: $z = \phi^{-1}(\text{threshold})$.

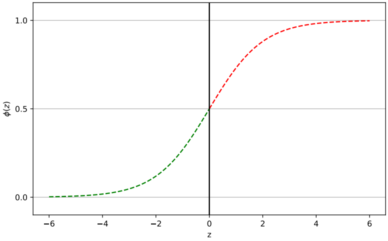


Figure: Sigmoid activation function

Given observations $\{(x_i, y_i)|i = 1, 2, \ldots, n\}$, the likelihood is given by

$$\mathcal{L}(w) = \prod_{i=1}^{n} \Pr(y = y_i|x_i, w) = \prod_{i=1}^{n} \phi(z_i)^{y_i}(1 - \phi(z_i))^{(1-y_i)}$$

Numerical solution of w: Newton-Raphson iteration

Objective: $\min E(w) = -\ln \mathcal{L}(w) + \frac{\gamma}{2}\|w\|$. (Cross-entropy + penalty).

- Diamond constraints, same as Least Absolute Shrinkage and Selection Operator (LASSO)
- γ: Hyper parameter; controls the strength of regularization.

- Conditional optimization of w
- Optimal solutions on axes (many zeros)
- Increase the strength until only two non-zero w_i, w_j
- Corresponding features i and j are selected.
- Use the selected features for estimation (e.g., logistic regression).

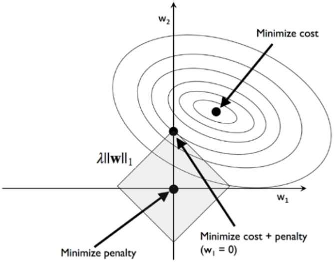


Figure: L1 regularization

Point estimation of w has the over-fitting problem. In Bayesian statistics, w is suppose to have a prior distribution $p(w)$. Denoted $p(w|y)$ as the posterior, then the predictive model is

$$p(C|X, y) = \int p(C|X, w)p(w|y)\,\mathrm{d}w$$

where C is the binary label, 1 or 0 (recession status)

Bayesian theorem

- $p(w|y) \propto p(w)p(y|w)$.
- Log form: $\ln p(w|y) = \ln p(w) + \ln(y|w) +$ Constant. (Bishop, 2006).
- Posterior = Prior + Likelihood.

Laplace approximation for $p(w|y)$, denoted as $q(w)$.

Purpose: get the conjugate posterior from a Gaussian prior $\mathcal{N}(w|m_0, S_0)$.

- Log-posterior: $\ln p(w|y) = \ln \mathcal{N}(w|m_0, S_0) + \ln \mathcal{L}(w) + \text{Constant}$
- $E(w)$ increases by a quadratic form. S_0 is positive definite.
- Maximum posterior (MAP) solution of w.Then optimize S_0
- Gaussian posterior: $q(w) = \mathcal{N}(w|w_{\text{map}}, \tilde{S})$

$$p(y=1|x) \approx \int \phi(w'x)q(w)\,\mathrm{d}w = \int \phi(\eta)p(\eta)\,\mathrm{d}\eta$$

$$p(\eta) = \int \delta(\eta - w'x)q(w)\,\mathrm{d}w$$

- $p(\eta)$: Marginal of the joint Dirac delta and Gaussian posterior.
- $p(\eta) \sim \mathcal{N}(w|\mu_\eta, \sigma_\eta^2)$, where $\mu_\eta = w'_{\text{map}}x$ and $\sigma_\eta^2 = x'\tilde{S}x$
- MacKay (1992) suggests using Probit function $\psi(k\eta)$ to approximate logistic sigmoid $\phi(\eta)$ around $\eta = 0$. $k = \sqrt{2\pi}/4$.
- $p(y=1|x) \approx \phi\left(\mu_\eta / \sqrt{1 + \pi\sigma_\eta^2/8}\right)$

Evaluation metrics

Cross-validation: Estimator selection

- Grid search of hyper parameters γ.
- Choose the estimator with best mean score.

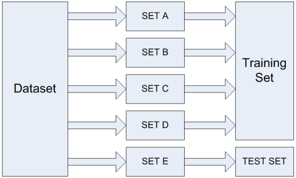


Figure: cross-validation

Confusion matrix: Performance of the estimation.

Table: Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

- Accuracy: $\frac{TP+TN}{TP+FN+FP+TN}$; Recall: $\frac{TP}{TP+FN}$
- Derivative indicator: Receiver Operating Characteristic (ROC) curve

Contents

Overview

- Selected Interest Rates (daily data), 11 series from US Federal Reserves website (H.15).
- **Short-term**: 1m, 3m, 6m, 1y, 2y, 3y.
- **Long-term**: 5y, 7y, 10y, 20y, 30y.
- Missing data and robustness check: Fitted via Nelson-Siegel-Svensson method (Nelson & Siegel, 1987; Svensson, 1994). Available at the FRB website (by Gürkaynak et al., 2007).

Period of availability

- 1962 - : 1y, 2y, 3y, 5y, 7y, 10y
- 1977 - : 30y
- 1982 - : 20y, 3m, 6m
- 2001 - : 1m

Business cycle

- Monthly data from NBER's website. 7 recession periods since 1962.
- Recession includes troughs but not the peak (Estrella & Trubin, 2006)
- Generated on a daily basis to match with yield curve.

Given recession period set Ω, the recession s_t is generated as

$$s_t = \chi_\Omega(t) = \begin{cases} 1, & t \in \Omega, \\ 0, & t \notin \Omega. \end{cases}$$

Recession binary y_t enters the model with a lag of m months: $y_t = s_{t+m}$.
Choice of the lag:

- Wright (2006) and Ang et al. (2006) use 18 months.
- Industry practitioners use 14 months. (This paper)
- Estrella and Trubin (2006) use 12 months.

Contents

Select features via logistic regression (LR) with L1 regularization.
Yield curve data are standardized (mean 0 and standard deviation 1).

Table: Feature selection

Period	Maturities
I) 2001/07/31-2018/10/31	1m, 3m, 6m, 1y, 2y, 3y, 5y, 7y, 10y, 20y, 30y
II) 1982/01/04-2018/10/31	3m, 6m, 1y, 2y, 3y, 5y, 7y, 10y, 20y, 30y
III) 1977/02/15-2018/10/31	1y, 2y, 3y, 5y, 7y, 10y, 30y
IV) 1962/01/02-2018/10/31	1y, 2y, 3y, 5y, 7y, 10y

- Ending date of estimation sample: 2019/12/31.
- Four data periods are used according to the data availability.
- 2018/10/31 is the 14-month lag of 2019/12/31.
- Selected features are in red.

Classifier optimization

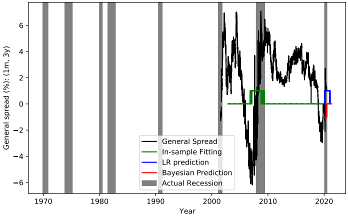
Estimation with selected pairs using LR + L1/L2 regularization.
Use grid search to select the model (best mean cross-validated score).

Table: Classifier estimation

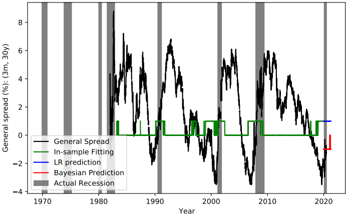
Pair	**Period**	**Classifier**(z)
[1m, 3y]	I) 2001/07/31 - 2018/10/31	$3.83r_{1m} - 2.91r_{3y} - 0.96$
[3m, 30y]	II) 1982/01/04 - 2018/10/31	$1.82r_{3m} - 2.13r_{30y} + 4.02$
[1y, 30y]	III) 1977/02/15 - 2018/10/31	$1.38r_{1y} - 1.5r_{30y} + 1.84$
[1y, 7y]	IV) 1962/01/02 - 2018/10/31	$1.82r_{1y} - 1.76r_{7y} + 0.3$

- Results are the same for L1/L2 regularization.
- Linear classifiers: $\phi(w_s r_s + w_l r_l + w_0) = \text{threshold}$, (w_s, w_l): coefficients for short-term and long-term yield rate; w_0: bias term.
- For example, $\phi(3.83r_{1m} - 2.91r_{3y} - 0.96) \geq 0.5$ implies a recession in 14 months if we use the default threshold.

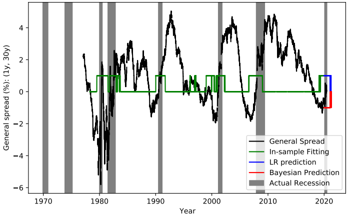
Graphical illustrations



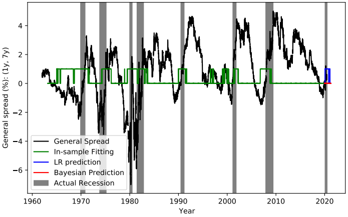
(a) $2.91r_{3y} - 3.83r_{1m} + 0.96$



(b) $2.13r_{30y} - 1.82r_{3m} - 4.02$



(c) $1.5r_{30y} - 1.38r_{1y} - 1.84$



(d) $1.76r_{7y} - 1.82r_{1y} - 0.3$

Performance Evaluation I

Confusion matrix

Display of the results in categories.

Table: Confusion matrix of empirical results

Pair	TP	FN	FP	TN	ACC	REC
[1m, 3y]	343	30	280	3664	93%	92%
[3m, 30y]	634	77	1579	6922	82%	89%
[1y, 30y]	998	170	1895	7362	80%	85%
[1y, 7y]	1389	335	3059	9412	76%	81%

- Threshold 0.5
- TP/FN/FP/TN are calculated on a daily basis
- Data imbalance: more none-recession observations.

Receiver Operating Characteristic

Relationship between True Positive Rate and False Positive Rate.

- Ideal model: $(0, 0) \rightarrow (0, 1) \rightarrow (1, 1)$.
- Random guess: $(0, 0) \rightarrow (1, 1)$.

Steps:

- Threshold [0.01, 0.02, ..., 1]
- Optimization models
- $TPR = \frac{TP}{TP+FN}$
- $FPR = \frac{FP}{FP+TN}$

Smaller sample gives larger ROC AUC (Area Under Curve of ROC)

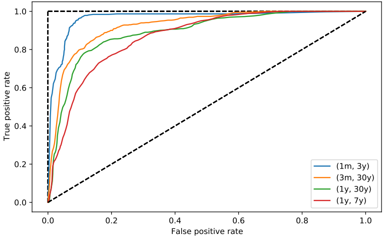


Figure: Receiver Operating Characteristic

Objective function for imbalanced classification: G-mean.

- Geometric mean of sensitivity (TPR) and specificity (1-FPR)
- Calculated from ROC.

Table: Threshold optimization

Pair	Threshold	G-mean	Sensitivity	Specificity
[1m, 3y]	0.28	0.93	0.96	0.90
[3m, 30y]	0.58	0.86	0.87	0.85
[1y, 30y]	0.57	0.83	0.83	0.84
[1y, 7y]	0.56	0.79	0.77	0.80

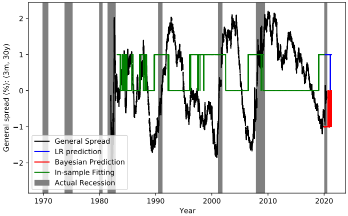
- Classifier: $\phi(z) =$ threshold. Same coefficients and bias term
- For example, $\phi(1.38r_{1y} - 1.5r_{30y} + 1.84) \geq 0.57 \rightarrow$ a recession in 14 months

Contents

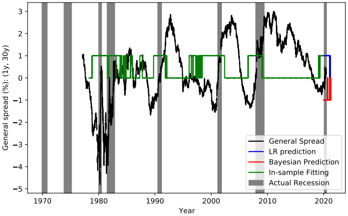
An alternative recession definiton: the occurrence within the next m months.

$$y_t = s_t \text{ or } s_{t+1} \text{ or } \cdots \text{ or } s_{t+m}$$

- $m = 14$.
- The same survival pairs: [3m, 30y] and [1y, 30y]



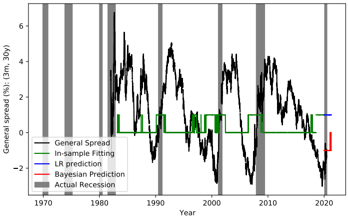
(a) (3m, 30y)



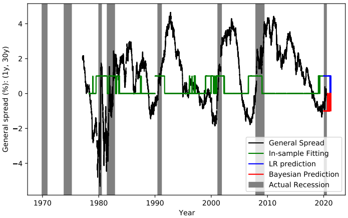
(b) (1y, 30y)

Figure: Estimation results from alternative recession definition

- Change the lag of recession to 12 months (Estrella & Trubin, 2006).
- Test if the long-short pairs are robust.
- The same survival pairs: [3m, 30y] and [1y, 30y].



(a) (3m, 30y)



(b) (1y, 30y)

Figure: Estimation results from alternative recession lag (12 months)

- More tests of lag.

Is the selection from L1 regularization reliable?
Use random forest (RF) - an ensemble of decision trees to measure feature importance.

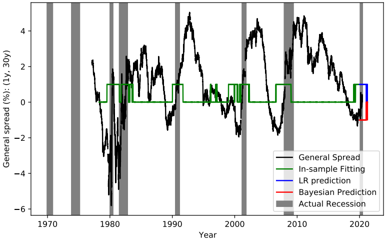
Table: Feature selection comparison

Period	**LR+L1**	**RF**
2001/07/31-2018/10/31	1m, 3y	1m, 3m
1982/01/04-2018/10/31	3m, 30y	3m, 20y
1977/02/15-2018/10/31	**1y, 30y**	**1y, 30y**
1962/01/02-2018/10/31	1y, 7y	1y, 2y

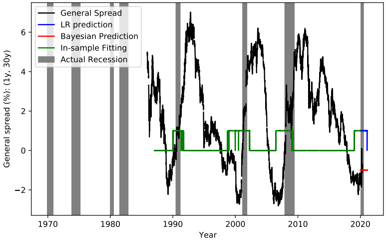
- RF gives the rank (feature importance) of yield curve rates. The two most important series are listed.
- (1y, 30y) is a robust long-short pair.

Robustness: Alternative yield curve series

- Gürkaynak et al. (2007) fit yield curve to a parametric form and monthly update the results in FRB website.
- Use (1y, 30y) as test pair.



(a) Raw data



(b) Smoothed data

Figure: Predictive power of (1y, 30y) with raw and smoothed data

- No significant difference between the results of two data sets.

Contents

Conclusion

Conclusion of this paper:

- Machine learning confirms the predictive power of long-short yield spread on recession.
- (1y, 30y) is the robust pair and the general spread form is $1.5r_{30} - 1.38r_1 - 1.84 \leq \phi^{-1}(0.43)$.
- The indicator performs well in out-of-sample prediction. (accuracy=90%, lag=14 months)
- The proposed model predicts there would be a recession in 2020 even without the COVID-19 pandemic.

Further improvements:

- Activation function and threshold (e.g., softmax in the output layer of Neural Network).
- Properties of recession probability (e.g., discontinuity at turning point?).
- Can yield curve predict the duration of recession?

Ang, A., Piazzesi, M., & Wei, M. (2006). What does the yield curve tell us about gdp growth? *Journal of econometrics*, *131*(1-2), 359–403.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Chinn, M., & Kucko, K. (2015). The predictive power of the yield curve across countries and time. *International Finance*, *18*(2), 129–156.

Estrella, A. (2005). Why does the yield curve predict output and inflation? *The Economic Journal*, *115*(505), 722–744.

Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The journal of Finance*, *46*(2), 555–576.

Estrella, A., & Mishkin, F. S. (1997). The predictive power of the term structure of interest rates in europe and the united states: Implications for the european central bank. *European economic review*, *41*(7), 1375–1401.

Estrella, A., Rodrigues, A. P., & Schich, S. (2003). How stable is the predictive power of the yield curve? evidence from germany and the united states. *The Review of Economics and Statistics*, *85*(3), 629–644.

Estrella, A., & Trubin, M. (2006). The yield curve as a leading indicator: Some practical issues. *Current issues in Economics and Finance*, *12*(5).

Reference II

Evgenidis, A., & Siriopoulos, C. (2014). Does the yield spread retain its forecasting ability during the 2007 recession? a comparative analysis. *Applied Economics Letters*, *21*(12), 817–822.

Gürkaynak, R. S., Sack, B., & Wright, J. H. (2007). The us treasury yield curve: 1961 to the present. *Journal of monetary Economics*, *54*(8), 2291–2304.

Hamilton, J. D., & Kim, D. H. (2002). A re-examination of the predictability of the yield spread for real economic activity. *Journal of Money, Credit, and Banking*, *34*(2), 340–360.

Harvey, C. R. (1988). The real term structure and consumption growth. *Journal of Financial Economics*, *22*(2), 305–333.

Harvey, C. R. (1989). Forecasts of economic growth from the bond and stock markets. *Financial Analysts Journal*, *45*(5), 38–45.

MacKay, D. J. (1992). The evidence framework applied to classification networks. *Neural computation*, *4*(5), 720–736.

Mishkin, F. S. (1990). The information in the longer maturity term structure about future inflation. *The Quarterly Journal of Economics*, *105*(3), 815–828.

Moneta, F. (2005). Does the yield spread predict recessions in the euro area? *International Finance*, *8*(2), 263–301.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of business*, 473–489.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, *4*, 351–394.

Svensson, L. E. (1994). *Estimating and interpreting forward interest rates: Sweden 1992-1994* (tech. rep.). National bureau of economic research.

Wright, J. H. (2006). The yield curve and predicting recessions (march 2006). *FEDs Working Paper, No. 2006-7*.

Discussions

What is the implications of the results in this paper?

- The general spread can be used for prediction
- It supplements the naïve inversion as a tool for monitoring the macroeconomy (yield curve control?).

what is the main contribution of this paper?

- Let the data speak: select the pair for prediction.
- Threshold optimization: Distinguished from existing literature.

Figure of 30y and 1y

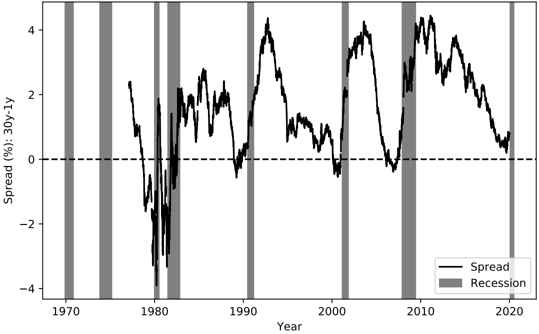
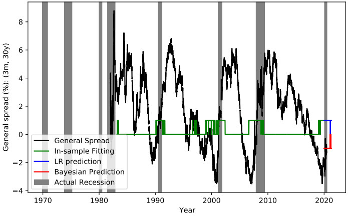


Figure: Term spread of (1y, 30y)

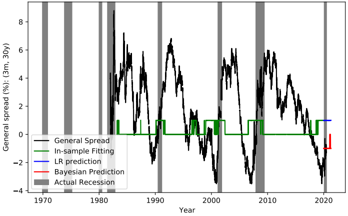
Table: Recessions since 1945

Peak	Trough	Rec.	Exp.	Cycle[P]	Cycle[T]
Nov. 1948	Oct. 1949	11	37	45	48
July 1953	May. 1954	10	45	56	55
Aug. 1957	Apr. 1958	8	39	49	47
Apr. 1960	Feb. 1961	10	24	32	34
Dec. 1969	Nov. 1970	11	106	116	117
Nov. 1973	Mar. 1975	16	36	47	52
Jan. 1980	July 1980	6	58	74	64
July 1981	Nov. 1982	16	12	18	28
July 1990	Mar. 1991	8	92	108	100
Mar. 2001	Nov. 2001	8	120	128	128
Dec. 2007	June 2009	18	73	81	91
1945-2009 (11 cycles)		11	58	69	69

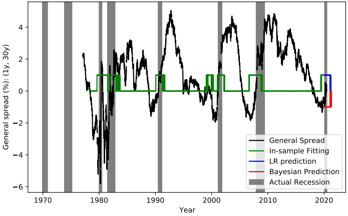
Optimal thresholds vs. default threshold



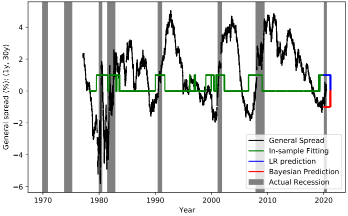
(a) (3m, 30y), threshold = 0.57



(b) (3m, 30y), threshold = 0.5



(c) (1y, 30y), threshold = 0.56



(d) (1y, 30y), threshold = 0.5

Test of other lags

Change the lag from one year to two years.

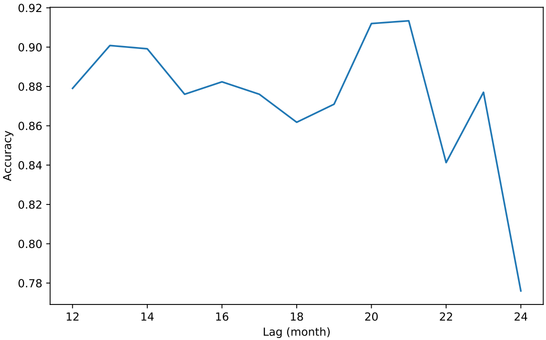
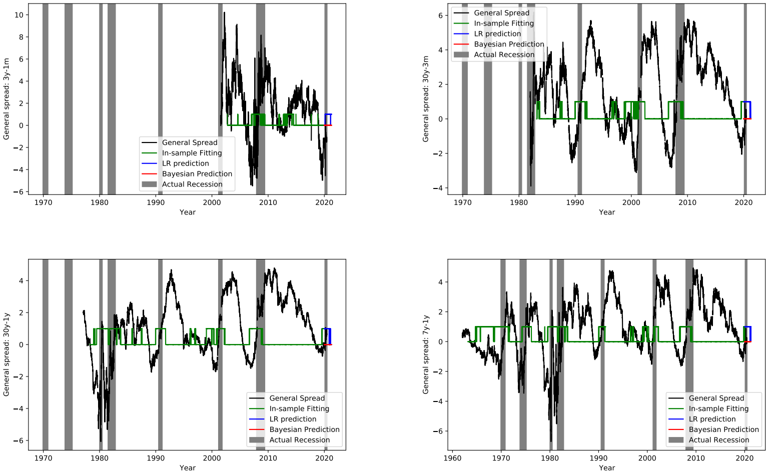


Figure: Out-of-sample prediction with different lags

- Tested pairs: [1y, 30y].
- Highest accuracy is about 0.91.

Results from naïve spread



Bayesian prediction fails!

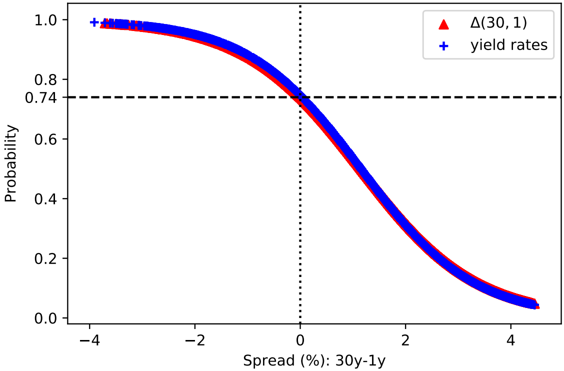


Figure: Naïve spread (red) and general form

- Output of LR with L2
- Red line is the model of Estrella et al. (2003).

Objective

Minimize the cost function: $E(w) = -\ln \mathcal{L}(w)$

Numerical optimization (Newton-Raphson iteration)

- Initialization of w.
- $w_{new} = w_{old} - H^{-1}(w)\nabla E(w)$.
- $H(w)$: Hessian matrix of $E(w)$.
- Better than gradient descent: $w_{new} = w_{old} - \eta \nabla E(w)$.

Regularization

- L1: $E(w) = -\ln \mathcal{L}(w) + \frac{\gamma}{2}\|w\|$. Sparse solution, feature shrinkage.
- L2: $E(w) = -\ln \mathcal{L}(w) + \frac{\gamma}{2}\|w\|^2$. Predictive model estimation.
- γ: Hyper parameter; controls the strength of regularization.

We aim to find the proper $q(w)$ to get the predictive distribution of output label C. The Laplace approximation is defined as

$$q(w) = \frac{|A|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(w - w_0)' A (w - w_0)\right\} = \mathcal{N}(w|w_0, A^{-1}) \quad (1)$$

It is actually a centered Gaussian at $w = w_0$ where A is the precision (A^{-1} is variance matrix). w_0 can be solved by numerical method, such as iterated reweighted least squares. Notice that the precision is the negative of Hessian matrix of $q(w)$.

$$A = -\nabla\nabla \ln q(w) \quad (2)$$

The posterior $p(w|y)$ is approximated by $q(w)$. $\tilde{S} = -\nabla\nabla \ln p(w|y)$. According to equation , $\tilde{S}$ is the sum of two terms. The first term of $\tilde{S}$ is S_0^{-1}. The second term is the 2nd-derivative of log-likelihood (Bishop, 2006, p. 218).

The log-likelihood itself is a sum of $y_i \ln(\phi(z_i)) + (1 - y_i) \ln(1 - \phi(z_i))$. For the sake of brevity, we define

$$\pi_i = \ln(\phi(z_i)) + (1 - y_i) \ln(1 - \phi(z_i)) \tag{3}$$

Notice that $1 - \phi(z_i) = \phi(-z_i)$, where $z_i = w' x_i$ and x_i is the i-th observation features (vector form). We have

$$\frac{\partial \pi_i}{\partial w} = y_i \frac{\phi'(z_i)}{\phi(z_i)} x_i - (1 - y_i) \frac{\phi'(z_i)}{\phi(-z_i)} x_i$$

The first order derivative of π_i can be written as

$$\begin{aligned} \frac{\partial \pi_i}{\partial w} &= y_i \phi(-z_i) x_i - (1 - y_i) \phi(z_i) x_i \\ &= y_i (1 - \phi(z_i)) x_i - (1 - y_i) \phi(z_i) x_i \\ &= (y_i - \phi(z_i)) x_i \end{aligned} \tag{4}$$

Then, the second order derivative of π_i with respect to w is

$$\begin{aligned}\frac{\partial^2 \pi_i}{\partial w^2} &= -\phi'(z_i)x_i x_i' = -\phi(z_i)(1-\phi(z_i))x_i x_i' \\ \frac{\partial^2 \ln \mathcal{L}(\mathbf{w})}{\partial w^2} &= -\sum_{i=1}^{n}\phi(z_i)(1-\phi(z_i))x_i x_i'\end{aligned} \tag{5}$$

Hence, the precision of Gaussian approximation posterior is

$$\tilde{S} = S_0^{-1} + \sum_{i=1}^{n}\phi(z_i)(1-\phi(z_i))x_i x_i' \tag{6}$$