Title
Essays on Forecasting Financial Markets Using Decomposition, Constraints and Extreme Learning Machine

Permalink
https://escholarship.org/uc/item/3f38g17s

Author
Xi, Zhou

Publication Date
2013

Peer reviewed|Thesis/dissertation
Essays on Forecasting Financial Markets Using Decomposition, Constraints and
Extreme Learning Machines

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Zhou Xi

December 2013

Dissertation Committee:

Professor Tae-Hwy Lee, Co-Chairperson
Professor Gloria Gonzalez-Rivera, Co-Chairperson
Professor Marcelle Chauvet
Professor Aman Ullah
The Dissertation of Zhou Xi is approved:

______________________________
Committee Co-Chairperson

______________________________
Committee Co-Chairperson

University of California, Riverside
Acknowledgments

I would like to gratefully and sincerely thank Professor Tae-Hwy Lee for his guidance, understanding, patience, and most importantly, his encouragement during my graduate studies at UC Riverside. I would also like to thank my committee members, Professor Gloria Gonzalez-Rivera, Professor Marcelle Chauvet and Professor Aman Ullah for their help.
This dissertation is dedicated to myself, who suffered a lot in pursuing the Ph.D degree.
ABSTRACT OF THE DISSERTATION

Essays on Forecasting Financial Markets Using Decomposition, Constraints and Extreme Learning Machines

by

Zhou Xi

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, December 2013
Professor Tae-Hwy Lee, Co-Chairperson
Professor Gloria Gonzalez-Rivera, Co-Chairperson

Chapter 1 and 2 discuss how to use a decomposition model to make a density forecast of the financial return and how to improve this density forecast by imposing matching moment constraints. The density forecast model is based on a decomposition of financial returns into the absolute return and the sign of the return. We also use the maximum entropy principle for the out-of-sample density forecast subject to the constraint that matches the mean forecasts from the decomposition model and a simple regression model. In Chapter 1 (joint with Professor Tae-Hwy Lee), we show that when the mean forecast from the decomposition model deviates from that of the mean return, imposing the matching mean forecast constraint will tilt the density forecast of the decomposition model and improve over the density forecast of the original decomposition model. In Chapter 2 (joint with Professor Tae-Hwy Lee and Ru Zhang), we further improve the decomposition model by using dependent copula functions, and we show that the risk forecast produced by the decomposition density forecast model is superior to RiskMetrics in terms of giving higher coverage probability and lower predictive quantile loss in extreme events of large loss for monthly returns.
Chapter 3 and 4 (joint with Professor Tae-Hwy Lee and Ru Zhang) deal with the testing of nonlinearity of time series data by using artificial neural network (ANN). In Chapter 3, we find that the original Lee, White and Granger (LWG, 1993) test is sensitive to the randomly generated activation parameters since they consider a fairly small number (10 or 20) of random hidden unit activations. To solve this problem, we simply increase the number of randomized hidden unit activations to a very large number (e.g., 1000). We show that using many randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a real empirical data. This robustification is reliable and useful in practice, and can be achieved at no cost as increasing the number of random activations is almost costless given today's computer technology. In Chapter 4, we further consider different types of regularization of the dimensionality, such as principal component analysis (PCA), Lasso, Pretest, partial least squares (PLS), among others. We demonstrate that while these supervised regularization methods such as Lasso, Pretest, PLS, may be useful for forecasting, they may not be used for testing because the supervised regularization would create the post-sample inference or post-selection inference (PoSI) problem. Our Monte Carlo simulation shows that the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso.
# Contents

List of Figures \hspace{1cm} x

List of Tables \hspace{1cm} xi

1 Density Forecast Using Decomposition and Maximum Entropy 1
   1.1 Introduction ........................................... 1
   1.2 Density Forecast Models ............................... 5
      1.2.1 Benchmark Model ................................. 5
      1.2.2 Decomposition Model ............................ 7
   1.3 Decomposition Model with Moment Constraints ......... 10
   1.4 Density Forecast Evaluation .......................... 17
   1.5 Empirical Analysis ................................... 19
      1.5.1 Data .............................................. 19
      1.5.2 Results .......................................... 20
   1.6 Conclusions ........................................... 23

Reference 25

2 Risk Forecast of Financial Returns Using Decomposition and Maximum Entropy 38
   2.1 Introduction ......................................... 38
   2.2 Density Forecast Models ............................... 44
      2.2.1 Benchmark Models ................................. 44
      2.2.2 Decomposition Model using Copulas ............... 49
   2.3 Decomposition with Moment Constraints ............... 56
   2.4 Density Forecast Evaluation .......................... 60
   2.5 Risk Forecast ......................................... 62
   2.6 Empirical Results .................................... 67
      2.6.1 Data .............................................. 67
      2.6.2 Results .......................................... 68
   2.7 Conclusion ........................................... 73

Reference 75
### 3 Testing for Neglected Nonlinearity Using Artificial Neural Networks with Many Randomized Hidden Unit Activations

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>92</td>
</tr>
<tr>
<td>3.2 The ANN Test</td>
<td>94</td>
</tr>
<tr>
<td>3.3 Small $q$ vs. Large $q$ in Monte Carlo Size and Power</td>
<td>100</td>
</tr>
<tr>
<td>3.4 Small $q$ vs. Large $q$ in Sensitivity to Randomized Hidden Unit Activations</td>
<td>106</td>
</tr>
<tr>
<td>3.5 Small $q$ vs. Large $q$ in Applications</td>
<td>109</td>
</tr>
<tr>
<td>3.6 Conclusions</td>
<td>112</td>
</tr>
</tbody>
</table>

### 4 Testing for Neglected Nonlinearity Using Regularized Artificial Neural Networks

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>132</td>
</tr>
<tr>
<td>4.2 Testing for Neglected Nonlinearity Using ANN</td>
<td>138</td>
</tr>
<tr>
<td>4.3 Regularizing the ANN Test</td>
<td>142</td>
</tr>
<tr>
<td>4.3.1 Unsupervised Regularization of the ANN Test Using PCA</td>
<td>142</td>
</tr>
<tr>
<td>4.3.2 Supervised Regularization of the ANN Test Using Lasso</td>
<td>143</td>
</tr>
<tr>
<td>4.3.3 Supervised Regularization of the ANN Test Using PLS</td>
<td>144</td>
</tr>
<tr>
<td>4.3.4 Supervised Regularization of the ANN test Using Pretests</td>
<td>145</td>
</tr>
<tr>
<td>4.3.5 Supervised Regularization of the ANN test Using PCA-first-and-then-Lasso</td>
<td>145</td>
</tr>
<tr>
<td>4.3.6 The PoSI Problem</td>
<td>147</td>
</tr>
<tr>
<td>4.4 Monte Carlo</td>
<td>148</td>
</tr>
<tr>
<td>4.4.1 DGPs and Simulation Design</td>
<td>148</td>
</tr>
<tr>
<td>4.4.2 Results</td>
<td>150</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>153</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Adding a Moment Condition under Normality ........................................ 33  
1.2 Plots of \( I(\lambda_t) \) and \( I(\lambda^*_t) \) .................................................. 34  
1.3 Mean Predictions from Model 1 and Model 2, and the Actual Data .... 36  
1.4 Mean Predictions from the Decomposition Model and the Predictive Regression Model Using All of GW’s Predictors, and the Actual Data ... 37  
2.1 Estimated copula parameter \( \theta \) over time ........................................ 88  
2.2 Mean forecasts comparison for copula models of Model 2 and historical mean ................................................................. 89  
2.3 Risk forecast comparison ................................................................. 90  
3.1 Monte Carlo Distribution of \( T_n(q) \) under \( H_0 \) .......................... 128  
3.2 P-values of \( T_n(q) \) under \( H_0 \) with \( m = 100 \) Randomizations of q Hidden Units ................................................................. 129  
3.3 Empirical Applications with \( m = 100 \) Randomized Hidden Unit Activations130  
4.1 Distribution of \( T^{PCA}_n \) under \( H_0 \) ............................................. 166  
4.2 Distribution of \( T^{Lasso}_n \) under \( H_0 \) .......................................... 167  
4.3 Distribution of \( T^{PLS}_n \) under \( H_0 \) ........................................... 168  
4.4 Distribution of \( T^{Pretest}_n \) (PCA-first-and-then-Pretest) under \( H_0 \) .... 169  
4.5 Distribution of \( T^{PCA–Lasso}_n \) (PCA-first-and-then-Lasso) under \( H_0 \) ... 170
List of Tables

1.1 Average Out-of-sample Logarithmic Scores of the Benchmark models and the Decomposition model ............................................. 29
1.2 Average Out-of-sample Logarithmic Scores of Binary Prediction .......... 30
1.3 Average Out-of-sample Logarithmic Scores of Different Density Forecast Models ................................................................. 31
2.1 Test for Independence .................................................................. 80
2.2 Average Out-of-Sample log Scores of Normal Distribution Model ...... 80
2.3 Average Out-of-Sample log Scores of Decomposition Models .......... 81
2.4 Out-of-sample Value-at-Risk Forecast Comparison: \( Q_t(12) \) .......... 82
2.5 Out-of-sample Value-at-Risk Forecast Comparison: \( Q_t(3) \) .......... 83
2.6 Out-of-sample Value-at-Risk Forecast Comparison: \( Q_t(1) \) .......... 84
2.7 Out-of-sample Value-at-Risk Forecast Comparison: \( R_t(12) \) .......... 85
2.8 Out-of-sample Value-at-Risk Forecast Comparison: \( R_t(3) \) .......... 86
2.9 Out-of-sample Value-at-Risk Forecast Comparison: \( R_t(1) \) .......... 87
3.1 Monte Carlo: Size and Power of the ANN Test ............................... 117
3.2 P-values of \( T_n(q) \) with \( m = 100 \) Randomizations of \( q \) Hidden Unit Activations ................................................................. 120
3.3 Empirical Analysis: P-values, Bonferroni Bounds, and Rejection Frequencies ............................................................................. 124
4.1 Size and Power of LWG, Lasso, PLS, and Pretest (with \( q = 200 \)) .... 164
4.2 Size and Power of PCA-first-and-then-Lasso with \( k = 100, 50, 10, 5 \) .... 165
Chapter 1

Density Forecast Using Decomposition and Maximum Entropy

1.1 Introduction

This paper discusses how to use a decomposition model to make a density forecast of the financial return and how to improve this density forecast by imposing matching moment constraints. As is well known that financial returns are not predictable in conditional mean, it is not our interest to forecast the mean. Rather, our interest is to forecast the conditional density of financial returns. The density forecast model is based on a decomposition of a financial return series into a product of the absolute value and the sign of the return. This paper shows how best we can use the two components, both of which are quite easily predictable, in forecasting density. That does not necessarily imply we can predict the financial returns in the conditional mean.
using the decomposition. Our interest is not forecasting the mean returns, but forecasting
the conditional density of financial returns using the decomposition. For example,
in financial risk management, we are more interested in the density (especially in tails)
than the mean. This paper also shows how to improve the density forecast using the
decomposition. Our trick is to find the maximum entropy density that satisfies simple
moment conditions. However, although the decomposition model satisfies stylized
distribution properties and provides predictability of the two components, its mean forecast
constructed from the decomposition model may not match with the mean forecast
from a simple predictive regression. As the simple model such as zero mean, constant
mean, or linear mean tends to give less estimation error than the complicated decompo-
sition model at least for the mean forecast, we use the maximum entropy principle for
the out-of-sample density forecast subject to the matching moment constraint that the
mean forecasts from the decomposition model and the simple model are equal. Imposing
such a natural constraint in the moment condition can help the out-of-sample density forecasts. We show that when the mean forecast from the decomposition model devi-
ates from that of the simple linear predictive regression, imposing the matching mean
forecast constraint will tilt the density forecast of the decomposition model to produce
the maximum entropy density forecast. The tilted density forecast can improve over
the density forecast of the original decomposition model.

Traditionally econometric modelling has been focused on the conditional mo-
mements of variables of interest, particularly mean and variance. Modern research has
shifted from the moment based modelling to modelling of the conditional distribution.
Density forecast has recently become a popular topic. The common density forecast
models for financial returns assume a particular distribution such as Gaussian, Student
t, or some variants to capture fat-tails or skewness. However, these distributional assumptions of the financial returns, although widely used, are inconsistent with some empirical findings on time series properties. In particular, Granger and Ding (1995, 1996) and Rydén, Teräsvirta and Åsbrink (1998) provide several stylized facts about the financial returns. Let $r_t$ be the return on a financial asset at time $t$, $|r_t|$ denote the absolute value, and $\text{sign}(r_t) = 1 (r_t > 0) - 1 (r_t < 0)$. Here are three distributional properties (DP) that are particularly important for this paper.

**DP1:** $|r_t|$ and $\text{sign}(r_t)$ are independent.

**DP2:** $|r_t|$ has the same mean and standard deviation.

**DP3:** The marginal distribution for $|r_t|$ is exponential.

The common distributional assumption fails to capture any of these three properties since it takes the financial return as an indivisible entity, rather than as two components. The density forecast model based on the joint density of $|r_t|$ and $\text{sign}(r_t)$ can be obtained from modelling of marginal densities of the absolute return and the sign of the return and their copula, which can capture the above three DP properties.

In addition to the three distributional properties which enable the margin forecasts of $|r_t|$ and $\text{sign}(r_t)$ to be linked to their joint density forecast, the conditional means of the margins are not martingale difference. That is, the conditional means of $|r_t|$ and $\text{sign}(r_t)$ can be dynamically modelled unlike that of the returns $r_t$. Ding et al (1993) show that $|r_t|$ is easily predictable, while Korkie, Sivakumar, and Turtle (2002) and Christoffersen and Diebold (2006) show that $\text{sign}(r_t)$ is predictable as well. Let $I_t$ be the information set at time $t$. If the return sign indicator series $1 (r_{t+1} < 0)$ displays conditional mean serial dependence, namely, if $E [1 (r_{t+1} < 0) | I_t]$ is a nonconstant function of $I_t$, then the signs can be predicted. Further, let $\mu_{t+1} = E (r_{t+1} | I_t)$ be the
conditional mean and $\sigma_{t+1}^2 = E \left[ (r_{t+1} - \mu_{t+1})^2 | I_t \right]$ be the conditional variance. Then Christoffersen and Diebold (2006) note that

$$E \left[ 1(r_{t+1} < 0) | I_t \right] = \Pr(r_{t+1} < 0 | I_t) = \Pr \left( \frac{r_{t+1} - \mu_{t+1}}{\sigma_{t+1}} < \frac{-\mu_{t+1}}{\sigma_{t+1}} | I_t \right) = F \left( \frac{-\mu_{t+1}}{\sigma_{t+1}} \right),$$

which shows the sign is predictable if $\sigma_{t+1}$ is predictable and $\mu_{t+1}$ is not zero. Using a series expansion of the conditional distribution $F(\cdot)$, such as the Gram-Charlier expansion or the Edgeworth expansion, we can show that $1(r_{t+1} < 0)$ can be predictable because the conditional higher moments (skewness and kurtosis) are predictable. Because the absolute return and the sign of the stock market return are predictable, the margin density forecast models of $|r_t|$ and $\text{sign}(r_t)$ can be specified such that these serial dependence properties associated with the predictability are incorporated. It can eventually yield a more precise density forecast model for $r_{t+1}$.

Although the decomposition model satisfies the above three distribution properties, sometimes it cannot generate good moment predictions even for the first moment (mean). First, the mean of density forecast from the decomposition model may be far from zero which contradicts the fact that stock returns are close to zero in mean especially in high frequency. Second, in our empirical applications we find that when applying the decomposition model to the annualized monthly stock return, the predicted conditional mean fluctuates rather excessively around or even deviating from the historical mean (HM) predictions. But for annualized monthly return, the HM is quite smooth and stable around a constant near zero. If we fix this problem by imposing a moment condition, we may improve the decomposition model. Since our objective function is the logarithmic score in evaluating the density forecast, the improvement can be achieved by solving the constrained maximum entropy problem. Jaynes (1968)
provides the solution for a discrete density. A general solution for any distribution can be seen in Csiszár (1975). Robertson, Tallman, and Whiteman (2005) and Giacomini and Ragusa (2011) apply these solutions to macroeconomic data and obtain nice results. These results encourage us to consider imposing a smooth moment condition when the decomposition model is unable to match the moment forecast of a simple mean regression model. Indeed, we find that imposing a sensible moment (mean) condition to the decomposition model can improve the density forecast of the financial returns $r_{t+1}$.

The rest of the paper is structured as follows. Section 2 introduces the benchmark models with normal distribution, and the decomposition model with specifications for the absolute return and the sign of the return. In Section 3, we impose a moment condition constraint to the decomposition density forecast in section 2 and some analytical results are provided. Section 4 introduces score functions to evaluate the density forecast in the entire support. Section 5 includes empirical results which compares the benchmark density forecasts, the decomposition density forecast and the decomposition density forecasts with constraints. Finally Section 6 concludes.

1.2 Density Forecast Models

1.2.1 Benchmark Model

We consider a simple benchmark density forecast model for $r_{t+1} = \mu_{t+1} + \varepsilon_{t+1} = \mu_{t+1} + \sigma_{t+1}z_{t+1}$ with the conditional normal distribution $\varepsilon_{t+1}|I_t \sim N \left(0, \sigma^2_{t+1}\right)$, and $\varepsilon_{t+1} = \sigma_{t+1}z_{t+1}$, $z_{t+1}$ is i.i.d. $E(z_{t+1}|I_t) = 0$, $E(z^2_{t+1}|I_t) = 1$.

For the conditional mean specification, it is common to assume that it is equal to zero or equal to the historical mean. The weak form of efficient market hypothesis
(EMH) has been tested to be true in the US financial markets. According to the EMH, the best forecast of the conditional mean is zero. Many papers have demonstrated the validity of the EMH. Consider a linear predictive regression for the conditional mean

\[ r_{t+1} = \alpha + x_t'\beta + \varepsilon_{t+1}. \] (1.2)

If \( \alpha = 0 \) and \( \beta = 0 \), then \( \mu_{t+1} = 0 \) (zero mean, ZM). If \( \beta = 0 \), then \( \mu_{t+1} = \alpha \), which can be estimated by the historical mean (HM) at time \( t \), \( \bar{r}_t = \frac{1}{R} \sum_{s=t-R+1}^{t} r_s \), using the rolling window of \( R \) observations. Without restrictions on \( \alpha, \beta \), we have \( \mu_{t+1} = \alpha + x_t'\beta \) the linear predictive regression using a set of predictors in Goyal and Welch (2008).

Here \( x_t \) may be a single predictor or include many predictors. Once we have obtained fitted values \( \left( \hat{\alpha}_t, \hat{\beta}_t \right) \) at time \( t \), we set \( \hat{\mu}_{t+1} = \hat{\alpha}_t + x_t'\hat{\beta}_t \). Goyal and Welch (2008) find that none of their 17 predictors can make a better mean forecast than the historical mean. Their results further demonstrate that it is very difficult to use a linear mean model to outperform the historical mean specification. Campbell and Thompson (2008) add sign constraints to the coefficient of predictors which make their model essentially nonlinear in mean specification. And as a result, their model makes a better out-of-sample performance.

For the conditional variance, we assume a generalized autoregressive conditional heteroskedasticity (GARCH) model specification

\[ \sigma_{t+1}^2 = \gamma_0 + \gamma_1 \varepsilon_t^2 + \gamma_2 \sigma_t^2. \] (1.3)

In the benchmark density forecast model, \( z_{t+1} \) has a normal distribution and so does \( r_{t+1} \). Then benchmark density forecast is

\[ r_{t+1} | I_t \sim N \left( \hat{\mu}_{t+1}, \hat{\sigma}_{t+1}^2 \right) \] (1.4)
where \( \hat{\sigma}^2_{t+1} = \hat{\gamma}_0 + \hat{\gamma}_1 \epsilon_t^2 + \hat{\gamma}_2 \sigma_t^2 \) is the predicted conditional variance according to the GARCH (1,1) model using the fitted values \((\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)\) at time \(t\).

The benchmark model will be called Model 1, for which we can set \( \mu_{t+1} = 0 \) (Model 1-ZM), \( \mu_{t+1} = \alpha \) (Model 1-HM), or \( \mu_{t+1} = \alpha + x_t' \beta \) (Model 1-x, labelled with the variable name of \(x\), with \(x\) being a set of predictors from Goyal and Welch (2008)). For example, if \(x_t\) is the dividend price ratio (dp), it will be Model 1-dp. If all 13 predictors are used, it is Model 1-all.

### 1.2.2 Decomposition Model

Since the pioneering work by Granger and Ding (1995), the decomposition model has been studied in several papers. Korkie, Sivakumar, and Turtle (2002), Rydberg and Shephard (2003), and Anatolyev and Gospodinov (2010) are among more recent works. Write the stock return as

\[
  r_{t+1} - \mu_{t+1} = |r_{t+1} - \mu_{t+1}| \times \text{sign}(r_{t+1} - \mu_{t+1}) =: U_{t+1} V_{t+1}, \tag{1.5}
\]

where \(U_{t+1} = |r_{t+1} - \mu_{t+1}|\) and \(V_{t+1} = \text{sign}(r_{t+1} - \mu_{t+1})\). This is merely an identity unless further constraints and assumptions are made.

First, let us model the absolute return \(U_{t+1}\). Since it takes nonnegative values (like duration), we can apply a model similar to the autoregressive duration (ACD) model of Engle and Russell (1998). We assume

\[
  U_{t+1} = \psi_{t+1} e_{t+1}, \tag{1.6}
\]

where \(\psi_{t+1} = E(U_{t+1}|I_t)\) is the conditional mean of the absolute return \(u_{t+1}\) and \(e_{t+1}\) is an i.i.d. positive random variable with \(E(e_{t+1}|I_t) = 1\). For the density forecast, Engle and Russell (1998) consider exponential and Weibull distributions for \(e_{t+1}\). Other
papers consider different distributions of $e_{t+1}$. Grammig and Maurer (2000) consider the Burr distribution and Lunde (1999) proposes a generalized gamma distribution, both of which nest the exponential distribution as a special case. Based on the stylized facts DP2 and DP3 of the absolute returns, we consider only the exponential distribution. More complicated distributions may not behave well since they may deviate from the stylized facts, especially DP3. We find that, from the data, absolute return has a strictly decreasing density. But if we use more complicated distributions, they cannot guarantee this property. In our empirical experiments, we also find that the Weibull distribution gives a much worse result. For the conditional mean $\psi_{t+1}$ we take the following specification

$$
\psi_{t+1} = \delta_0 + \delta_1 |r_t - \mu_t| + \delta_2 \psi_t,
$$

which is an ACD-like model. While other (nonlinear) specifications such as a logarithmic model of Bauwens and Giot (2000) and a threshold model of Zhang, Russell and Tsay (2001) are possible, we find the above simple linear model is sufficient and a higher order specification is not necessary to make the density forecast more accurate. Once we estimated the parameters, we can obtain the forecasted conditional mean $\hat{\psi}_{t+1}$ and the forecasted density for $U_{t+1}$ will be an exponential distribution with mean equal to $\hat{\psi}_{t+1}$. Therefore,

$$
\int_{t+1}^U(u) = \frac{1}{\psi_{t+1}} \exp \left(-\frac{1}{\psi_{t+1}}u\right).
$$

Next, to model the sign of the return $V_{t+1}$, we assume a Bernoulli-type density for the binary event. If the sign of the actual stock return at $t + 1$ is positive, $v_{t+1} = 1$, ...
otherwise \( v_{t+1} = -1 \). Then the sign forecast density function will be:

\[
 f_{V_{t+1}}(v) = \begin{cases} 
 p_{t+1} & \text{if } v = 1 \\
 1 - p_{t+1} & \text{if } v = -1 
\end{cases}
\]

\[
 = \frac{v+1}{p_{t+1} (1 - p_{t+1})^{\frac{v+1}{2}}},
\]

where \( p_{t+1} := \Pr (v_{t+1} = 1|I_t) \). For the density forecast, we need to predict \( p_{t+1} \). The simplest way is to predict \( p_{t+1} \) from the historical percentage of positive returns, \( \hat{p}_{t+1} = \frac{1}{R} \sum_{s=t-R+1}^{t} 1 (r_s > \mu_s) \). This is a special case of the generalized linear model (GLM)

\[
 p_{t+1} = \mathbb{E} [1 (r_{t+1} > \mu_{t+1}) | I_t] = G (a + x_t^t b),
\]

where \( G(\cdot) \) is a link function. If \( G(\cdot) \) is the identity function, we have the ordinary least square estimator. \( G(\cdot) \) can be also the standard normal cumulative distribution function or the logistic function. In these two cases, we have the probit model or the logit model. Here we assume \( G(z) = \frac{e^z}{1 + e^z} \), the logit model. See, e.g., Lahiri and Liu (2012) for an extensive survey. The covariate \( x_t \) may be a single predictor or include many predictors. If we do not include \( x_t \) in the model \( (b = 0) \), we have the historical percentage of positive returns. Therefore, the GLM nests the historical percentage of positive returns. Once we get \( \hat{p}_{t+1} \), the density forecast of the sign of the return will be

\[
 f_{V_{t+1}}(v) = \frac{v+1}{\hat{p}_{t+1} (1 - \hat{p}_{t+1})^{\frac{v+1}{2}}},
\]

Finally, based on DP1, we assume that the absolute return \( U_{t+1} \) and the sign of the return \( V_{t+1} \) are independent. The joint density of \( (U_{t+1}, V_{t+1}) \) (conditional on \( I_t \)) using the independent copula is

\[
 f_{t+1}(u, v) = f_{t+1}^U(u) \times f_{t+1}^V(v).
\]

One can relax the assumption of independence between the absolute return and the sign of the return and use a copula function to model the dependence. However, from the stylized fact DP1, the absolute return and the sign of return are actually independent.
This is also supported by Anatolyev and Gospodinov (2010), who find that ignoring dependence performs better in out-of-sample prediction. They also find that the correlation between the absolute return and the sign of return is very weak. We follow them and DP1 to model the joint density of the absolute return and the sign of return using the independent copula.

The decomposition model will be called Model 2, for which we can set \( \mu_{t+1} = 0 \) (Model 2-ZM), \( \mu_{t+1} = \alpha \) (Model 2-HM), or \( \mu_{t+1} = \alpha + x_i \beta \) (Model 2-\( x \)), labelled with the variable name of \( x \), with \( x \) being a predictor from Goyal and Welch (2008). For example, if \( x \) is the book to market ratio (bm), it will be Model 2-bm.

### 1.3 Decomposition Model with Moment Constraints

A potential problem with the estimated joint density forecast model is that the mean prediction \( E(U_{t+1}V_{t+1}|I_t) \) from the the joint density forecast \( f_{t+1}(u,v) \) of the decomposition model (Model 2) may not be equal to the zero, which makes \( E(r_{t+1} - \mu_{t+1}|I_t) \neq 0 \). We can see, by assuming independence, that the mean of the joint density is equal to the product of the means of two marginal densities. Since \( f_{t+1}^U(u) \) is the exponential density forecast, its mean is \( \psi_{t+1} \) and is strictly positive; the mean of \( f_{t+1}^V(v) \) is \( (2\hat{p}_{t+1} - 1) \). Thus if and only if \( \hat{p}_{t+1} \) is 0.5, the mean of the joint density forecast is 0, the same as the benchmark mean forecast. If \( \hat{p}_{t+1} \) is not 0.5, the mean of the joint density forecast is not 0. Usually \( \hat{p}_{t+1} \) is greater than 0.5, meaning that we have more positive return months than negative return months, as stock prices tend to rise in the long run. This contradicts the fact that \( E(r_{t+1} - \mu_{t+1}|I_t) = 0 \) of the benchmark model (which must hold from the definition of \( \mu_{t+1} \)). With this in mind, we want to impose
the (conditional) moment condition that

\[ E (U_{t+1}V_{t+1}|I_t) = 0, \]  

(1.11)

into the decomposition model. This moment condition reflects the identity \( E (U_{t+1}V_{t+1}|I_t) = E (r_{t+1} - \mu_{t+1}|I_t) = 0. \)

To see the same issue from a different angle, recall that the decomposition in (2.5) is merely an identity. It holds for any number replacing \( \mu_{t+1} \) in (2.5), say, by zero, which will give another identity

\[ r_{t+1} = |r_{t+1}| \times \text{sign}(r_{t+1}) =: U_{t+1}V_{t+1}, \]  

(1.12)

where \( U_{t+1} = |r_{t+1}| \) and \( V_{t+1} = \text{sign}(r_{t+1}) \) are re-defined. This is equivalent to Model 2-ZM. We have the same problem that the mean prediction \( E (U_{t+1}V_{t+1}|I_t) \) from the joint density forecast \( f_{t+1}(u,v) \) of the decomposition model (Model 2-ZM) may not be equal to the mean prediction \( E (r_{t+1}|I_t) = \mu_{t+1} \) of the benchmark model (Model 1).

With this in mind, we want to impose the (conditional) moment condition that

\[ E (U_{t+1}V_{t+1}|I_t) = \mu_{t+1}, \]  

(1.13)

into the decomposition model. This moment condition reflects the identity \( E (U_{t+1}V_{t+1}|I_t) = E (r_{t+1}|I_t) = \mu_{t+1} \). In what follows we consider the decomposition (1.12) and the moment condition (1.13).

Since we are using the logarithmic score to evaluate the density forecast, imposing the moment constraint is equivalent to solve the following constrained maximization problem of the cross-entropy of the new density forecast \( h_{t+1}(u,v) \) with respect to the
original density forecast $f_{t+1}(u,v)$:

$$
\max_{h_{t+1}(u,v)} = - \int \int \log \frac{h_{t+1}(u,v)}{f_{t+1}(u,v)} h_{t+1}(u,v)dudv
$$

(1.14)

subject to

$$
\int \int m_t(u,v) h_{t+1}(u,v)dudv = 0,
$$

(1.15)

and

$$
\int \int h_{t+1}(u,v)dudv = 1,
$$

(1.16)

where $f_{t+1}(u,v)$ is the density forecast from the decomposition model and $h_{t+1}(u,v)$ is a new density forecast satisfying the moment constraint of (2.28). If the conditional mean forecast condition in (1.13) is imposed to the decomposition (1.12), we have

$$
m_t(u,v) = uv - \hat{\mu}_{t+1}
$$

(1.17)

and the expectation in (1.13) is evaluated using the new tilted density forecast $h_{t+1}(u,v)$.

Note that the moment constraint function $m_t(u,v)$ is denoted with the subscript $t$ as it is measurable with respect to the information $I_t$ at time $t$ (as $\hat{\mu}_{t+1}$ is $I_t$-measurable). The moment condition (2.28) with (2.30) will make the new joint density forecast $h_{t+1}(u,v)$ have the same mean forecast as the benchmark Model 1.

The maximization of (2.27) subject to (2.28) is well known in the maximum entropy literature. Jaynes (1957) was the first to consider this problem. Jaynes (1957, 1968) provides a solution for discrete density. A general solution for any distribution can be seen in Csiszár (1975). See also Maasoumi (1993), Zellner (1994), Golan, Judge, and Miller (1996), Ullah (1996), Bera and Bilias (2002), among others.

The solution to the maximization problem, if exists, is given by

$$
h_{t+1}(u,v) = f_{t+1}(u,v) \exp\{\eta_t^* + \lambda_t^* m_t(u,v)\},
$$

(1.18)
where

$$\lambda^*_t = \arg \min_{\lambda_t} I_t(\lambda_t),$$  \hspace{1cm} (1.19)$$

$$\eta^*_t = \log\{I_t(\lambda^*_t)\}^{-1},$$  \hspace{1cm} (1.20)$$

and

$$I_t(\lambda_t) = \int \int \exp\left[\lambda_t m_t(u, v)\right] f_{t+1}(u, v)\,du\,dv.$$

We construct a new density forecast by exponentially tilting through $\lambda^*_t$ and normalizing it through $\eta^*_t$. This derivation can also be found in recent econometric applications of the maximum entropy, e.g., Imbens, Spady and Johnson (1998), Robertson, Tallman and Whiteman (2005), Park and Bera (2006), Bera and Park (2008), Stengos and Wu (2010), and Giacomini and Ragusa (2011).

We can see the effect of imposing moment constraints through two examples. In the first example, the original density is $N(2, 1)$ and we impose the moment constraint that the mean is 0. In the second example, the original density is $N(0, 4)$ and we impose the moment constraint that the variance is 1. The plots of the results are shown in Figure 1. In Figure 1a, by imposing the mean constraint, we shifts the normal distribution without change its shape. In Figure 1b, imposing the variance constraint changes the shape of the normal distribution but does not shift it. In both cases, the distribution remains as the normal distribution. Therefore, in practice imposing moment constraint is not very useful if assuming normality. This is because imposing a moment constraint is only a suboptimal result and cannot be better than estimating the conditional mean and the conditional variance jointly.

Note that $\int \log \frac{h_{t+1}(u, v)}{f_{t+1}(u, v)} h_{t+1}(u, v)\,du\,dv$ in the objective function of (2.27) is the Kullbeck-Leibler (1951) information criterion (KLIC) divergence measure between the
new density and the original density. If the (conditional) moment constraint is true, the
difference of expected (conditional) logarithmic scores between $h_{t+1}(u,v)$ and $f_{t+1}(u,v)$
is nonnegative. To be specific, if the conditional moment constraint $E[m_t(u,v)] = 0$ is true, then $E[\log h_{t+1}(u,v) - \log f_{t+1}(u,v)] = E(\eta_t)$ is nonnegative, since $\eta_t = KLIC(h,f)$ is nonnegative. See also Giacomini and Ragusa (2011). In this case, as the optimal $\eta^*_t$ is always positive, $I_t(\lambda^*_t)$ must be between 0 and 1.

To find $\lambda^*_t$, we need to take the integral of the original joint density times
an exponential function. This can be done in two ways. The first way is to use the
numerical integral. For the density that is not in the exponential family, the integral
above is difficult or impossible to be calculated analytically. Thus we have to use the
numerical integral. Since

$$I_t(\lambda_t) = \int \int \exp[\lambda_t m_t(u,v)] f_{t+1}(u,v) du dv = E_t(\exp[\lambda_t m_t(u,v)]) , \quad (1.22)$$

where the expectation is taken over the density forecast $f_{t+1}(u,v)$, we can first generate
$S$ random draws $\{u^s_t, v^s_t\}_{s=1}^S$ from $f_{t+1}(u,v)$, then calculate $I_t(\lambda_t) = \frac{1}{S} \sum_{s=1}^S \exp\{\lambda_t m(u^s_t, v^s_t)\}$. Once we solve for $\lambda^*_t$ by minimizing $I_t(\lambda_t)$, then we obtain $\eta^*_t = \log\{I_t(\lambda^*_t)\}^{-1}$. The problem with the numerical integration is that, as we discuss below, $I(\lambda_t)$ is flat for a wide range of $\lambda_t$, so that the numerical integration may not well behave and the algorithm
may stop before reaching $\lambda^*_t$. Therefore the nonnegativity of $\eta^*_t$ cannot be guaranteed.

The second way is to solve the maximization problem analytically. Solving
it analytically will give the most accurate results of $\lambda^*_t$ and $\eta^*_t$ and ensure that $\eta^*_t$ is
nonnegative. To illustrate, consider a simple case with $m_t(u,v) = uv - \hat{\mu}_{t+1} = uv$ (i.e.,
with $\hat{\mu}_{t+1} = 0$) in the moment condition (2.30). In this case, the analytical expression
of $I_t(\lambda_t)$ is obtained as follows. Let $a_t = \frac{1}{\psi_t}$.

$$I_t(\lambda_t) = \int \int f_{t+1}(u,v) \exp[\lambda_t m_t(u,v)] \, du \, dv$$

$$= \int \int_{v=-1}^{v=1} f_{t+1}(u,v) \exp[\lambda_t m_t(u,v)] \, du \, dv + \int \int f_{t+1}(u,v) \exp[\lambda_t m_t(u,v)] \, du \, dv$$

$$= \int_{0}^{\infty} a_t \exp(-a_t u)(1 - p_t) \exp(-\lambda_t u) \, du + \int_{0}^{\infty} a_t \exp(-a_t u)p_t \exp(\lambda_t u) \, du$$

$$= \frac{a_t(1 - p_t)}{a_t + \lambda_t} \int_{0}^{\infty} (a_t + \lambda_t) \exp[-(a_t + \lambda_t) u] \, du$$

$$+ \frac{a_t p_t}{a_t - \lambda_t} \int_{0}^{\infty} (a_t - \lambda_t) \exp[-(a_t - \lambda_t) u] \, du$$

$$= \frac{a_t(1 - p_t)}{a_t + \lambda_t} + \frac{a_t p_t}{a_t - \lambda_t}.$$  \hfill (1.23)

A plot of $I_t(\lambda_t)$ with $a_t = 8$ and $p_t = 0.55$ is given in Figure 2a. These values of $a_t$ and $p_t$ are estimated from actual annualized monthly equity premium series. A plot of $I_t(\lambda_t)$ with $a_t = 8$ and $p_t = 0.65$ is given in Figure 2c, with the values of $a_t$ and $p_t$ estimated from actual annualized monthly stock return series. We can see that for a wide range of $\lambda_t$, $I_t(\lambda_t)$ is near flat. $I_t(\lambda_t)$ appears to change little over the flat area. Therefore when using numerical integration $I_t(\lambda_t) = \frac{1}{S} \sum_{s=1}^{S} \exp[\lambda_t m_t(u_s^t, v_s^t)]$ to find the optimal value, it can easily stop somewhere in the flat area where $I_t(\lambda_t)$ may be above 1 (and thus $\eta^*$ may be less than 0).

To avoid this to happen, one can use the analytical solution for $\lambda_t^*$ from solving the first order condition (FOC)

$$\frac{dI_t(\lambda_t)}{d\lambda_t} = -\frac{1 - p_t}{(a_t + \lambda_t)^2} + \frac{p_t}{(a_t - \lambda_t)^2} = 0.$$  \hfill (1.24)

Solving for $\lambda_t$ and choosing the solution whose absolute value is less than $a$, we get:

$$\lambda_t^* = \frac{a_t \left(-1 + 2\sqrt{p_t(1 - p_t)}\right)}{2p_t - 1},$$  \hfill (1.25)

if $p_t \neq \frac{1}{2}$. Note that $I_t(0) = 1$ and in fact $I_t(\lambda_t) < 1$ for some $\lambda_t$. To look more closely at the bottom of Figure 2a and Figure 2c, these figures are magnified into Figure 2b.
and Figure 2d for a narrower domain of $-2 < \lambda_t < 2$ that includes the optimal value $\lambda_t^*$. It can be seen that $I_t(\lambda_t^*) < 1$ at the optimal value of $\lambda_t$.

Plug $\lambda_t^*$ back into $I_t(\lambda_t)$, we can find the minimized value of the integral

$$I_t(\lambda_t^*) = \frac{(1 - p_t)(2p_t - 1)}{2p_t - 2 + 2\sqrt{p_t(1 - p_t)}} + \frac{p_t(2p_t - 1)}{2p_t - 2\sqrt{p_t(1 - p_t)}}.$$  \hspace{1cm} (1.26)

Figure 2e plots $I_t(\lambda_t^*)$ against $p_t$. Note that the optimal value of $I_t(\lambda_t^*)$ is smaller than 1 for all values of $p_t$ (except 0.5), which means that $\eta^*$ is greater than 0 for all $p_t$. It is important to note that, the more $p_t$ deviates from 0.5, the more room we can have for improvement from imposing the moment condition. This is because $\eta^*$ can be substantially less than 1 when $p_t$ deviates from 0.5. It is interesting to see that $I_t(\lambda_t^*)$ does not depend on $a_t$ but depends only on $p_t$.\footnote{This is due to the particular moment condition $m_t(u,v) = uv - \hat{\mu}_{t+1}$ and $\hat{\mu}_{t+1} = 0$ (2.30) in deriving this. It is not true in general with a different moment condition. For example, if the moment condition with $\hat{\mu}_{t+1} \neq 0$, $I_t(\lambda_t^*)$ will depend on $a_t$ and $p_t$.}

The decomposition model (Model 2-ZM) with the moment constraint imposed will be called Model 3. For simplicity, we consider only the decomposition model Model 2-ZM to construct Model 3. We do not consider Model 2-HM (with $\mu_{t+1} = \alpha$) and Model 2-\(x\) ($\mu_{t+1} = \alpha + x_{t}^\prime \beta$), to extend them to Model 3 with imposing the moment (mean) condition. Model 3 incorporates the mean forecast through the moment condition $m_t(u,v) = uv - \hat{\mu}_{t+1}$ in (2.30). If $\hat{\mu}_{t+1} = 0$, it will be labelled as Model 3-ZM. If $\hat{\mu}_{t+1} = \bar{r}_t$, it will be called Model 3-HM. If $\hat{\mu}_{t+1} = \hat{\alpha}_t + x_{t}^\prime \hat{\beta}_t$ then Model 3-\(x\). Model 2-ZM is the original density forecast model $f_{t+1}(u,v)$, while Model 3 is the tilted density forecast model $h_{t+1}(u,v)$. A goal of this paper is to examine if imposing the moment constraint can improve Model 3 over Model 2-ZM in density forecast.
1.4 Density Forecast Evaluation

We can use a scoring rule to evaluate density forecasts. A scoring rule is a positive-oriented criterion to evaluate density forecasts. A larger expected score usually means that the associated density forecast is better. Formally speaking, a score function or scoring rule $S(f, y)$ for a single point takes value in the real line $\mathbb{R}$ or in the extended real line $\mathbb{R} = [-\infty, \infty]$, where $f$ is the density forecast and $y$ is the realized value. Let $E_h S(f, y) = \int S(f, y) h(y) dy$ be the expected value of $S(f, y)$ under the distribution $h$. A scoring rule is said to be proper if $E_h S(h, y) \geq E_h S(f, y)$ for all $f$ and $h$. If the equality holds only if $f = h$, then the score function is strictly proper.

Since for a proper scoring rule, the expected score of the true density is always greater than the expected score of any other density. This property encourages people to reveal his or her true belief. For example, one of the most popular scoring rules is the likelihood based scoring rule:

$$S^L(f, y) = \log f(y). \quad (1.27)$$

The difference of the expected scores $[E_h S(h, y) - E_h S(f, y)]$ is the KLIC divergence measure. It can be shown that the likelihood based score is strictly proper because

$$KLIC(h, f) = E_h [S^L(h, y) - S^L(f, y)] = E_h [\log h(y) - \log f(y)] \geq 0 \quad (1.28)$$
due to the Jensen’s inequality applied to the logarithmic function which is a concave function. See, e.g., Rao (1965), White (1994), and Ullah (1996).

To apply the logarithmic score to the decomposition model, let $y = (u, v)$ and

$$S^L(f_{t+1}, y_{t+1}) = S^L(f_{t+1}, (u_{t+1}, v_{t+1})) = \log f_{t+1}(u_{t+1}, v_{t+1}). \quad (1.29)$$

In fact, the logarithmic score of this joint density of $U$ and $V$ can be compared with that of the benchmark model. Since $r_t \equiv U_t \ast V_t$, in the benchmark model we implicitly
assume that the joint density forecast between $U$ and $V$, conditional on the information set $I_t$ is:

$$f_{t+1}^B(u, v) = \frac{1}{\sqrt{2\pi \hat{\sigma}_{t+1}}} \exp\left\{-\frac{(uv - \hat{\mu}_{t+1})^2}{2\hat{\sigma}_{t+1}^2}\right\},$$

(1.30)

where $u \geq 0$ and $v = 1$ or $-1$. One can show that the above function is a density since the integral is 1. Also, $S^L(f_{t+1}^B(r, r_{t+1}) = S^L(f_{t+1}^B(u, v), (u_{t+1}, v_{t+1}))$. So we are actually comparing two joint densities between $U$ and $V$.

An advantage of assuming independence in the decomposition model (based on the stylized fact, DP1) is that we can improve the density forecast by improve any of the two marginal density forecasts. From equation (2.6),

$$S^L(f_{t+1}, (u_{t+1}, v_{t+1})) = \log f_{t+1}^U (u_{t+1}) + \log f_{t+1}^V (v_{t+1}).$$

(1.31)

Once we obtain a better marginal density forecast for $u$, the logarithmic score of the joint density will be improved since the marginal density of $v$ remains unaffected. The opposite is also true. Therefore, we can always try different specifications for the absolute return and the sign of the return to make a better joint density forecast.

After calculating the score function at each time $t$, we can compare density forecasts by computing the average out-of-sample scores. That is

$$\bar{S}_P = \frac{1}{P} \sum_{t=T-P+1}^T S_t,$$

(1.32)

where $S_t$ is the logarithmic score $S^L(f_{t+1}, (u_{t+1}, v_{t+1}))$ or the conditional likelihood score $S^{CL}(f_{t+1}, (u_{t+1}, v_{t+1}))$ at time $t$. The density forecast with a higher value of $\bar{S}_P$ is considered to be the better density forecast.
1.5 Empirical Analysis

1.5.1 Data

We use the data in Goyal and Welch (2008). In addition to their original monthly data, we calculate the annualized monthly stock return as in Campbell and Thompson (2008). We also use equity premium. Since the difference between equity premium and stock return is the risk free rate which is relatively small and smooth compared to the equity premium, the equity premium has similar distribution properties (DP1, DP2, DP3) as those of the stock return. Therefore we can apply the decomposition model to equity premium as well. Denote by \( P_t \) the S&P500 index at month \( t \). The monthly simple one-month return from month \( t \) to month \( t+1 \) is defined as

\[
R_t(1) \equiv \frac{P_{t+1}}{P_t} - 1,
\]

and one-month excess return is 
\[ Q_t(1) \equiv R_t(1) - r_f \]

with \( r_f \) being the risk-free interest rate. Following Campbell, Lo and MacKinlay (1997, p. 10), we define the \( k \)-period return from month \( t \) to month \( t+k \) as

\[
R_t(k) \equiv \frac{P_{t+k}}{P_t} - 1 = \left( \frac{P_{t+k}}{P_{t+k-1}} \right) \times \cdots \times \left( \frac{P_{t+1}}{P_t} \right) - 1
= (1 + R_{t+k-1}(1)) \times \cdots \times (1 + R_t(1)) - 1
\]

and following Campbell and Thompson (2008) we define the \( k \)-period excess return as

\[
Q_t(k) \equiv (1 + R_{t+k-1}(1) - r_f) \times \cdots \times (1 + R_t(1) - r_f) - 1
= (Q_{t+k-1}(1) + 1) \times \cdots \times (Q_t(1) + 1) - 1
= \left[ \prod_{j=1}^{k} (Q_{t+k-j}(1) + 1) \right] - 1.
\]

We let \( r_{t+1} = R_t(k) \) or \( Q_t(k) \), and consider \( k = 1, 12 \) as reported in Campbell and Thompson (2008).
We use 13 predictors including dividend price ratio (dp), earning price ratio (ep), book to market ratio (bm), long term yield (lty), inflation (infl), dividend payout ratio (de), dividend yield (dy), stock variance (svar), default yield spread (dfy), long term rate of return (ltr), net equity expansion (ntis), treasure bill rates (tbl) and cross-sectional premium (csp). For a detailed definition and calculation of each of the predictors please see Goyal and Welch (2008).

We consider the data from May 1937 to December 2002 because this time period includes the most updated data for all 13 predictors of Goyal and Welch (2008). We divide the data equally into \( R \) in-sample observations and \( P \) pseudo out-of-sample observations. The models are estimated using rolling windows of the fixed size \( R \). That is, at each time \( t \) we use the data starting from \( t - R + 1 \) and ending at time \( t \) to estimate parameters of a model and then make one-month ahead forecast for the next month \( t + 1 \). For annualized monthly data, to avoid using future information, we only use data up to month \( t - 11 \) for estimation.

1.5.2 Results

In Table 1, it can be clearly seen from the results of annualized monthly returns data that the decomposition model easily beats the benchmark model. The average logarithmic score of the decomposition model improves upon that of the benchmark model by a huge margin. Also one can see that the predictive ability of Model 1-ZM and Model 2-ZM is worse for the annualized monthly stock return than those for the annualized monthly equity premium. Turning to the monthly returns without annualizing (with \( k = 1 \)), one can find that the difference of average logarithmic scores between the decomposition models and the benchmark models is much narrower. We can also see
that the decomposition model fails to beat the zero mean benchmark model for monthly equity premium.

Table 2 presents the results of using different predictors in sign prediction. If we do not include any predictor, the estimation is equal to the historical average of positive returns and the results are the same as those in Table 1. By looking at the first two columns of the table, we can see that it is difficult to predict the sign for annualized data since only one predictor can make a better density forecast for stock return and this number becomes 6 for equity premium. However, the sign prediction is much easier for monthly returns. For both equity premium and stock return, there are 8 out of 13 predictors that can generate higher average logarithmic scores. And if we use all those 13 predictors, they can still produce better forecasts. The reason why monthly returns can be predicted is due to the stronger time dependence of the conditional variance in high frequency data. Since the conditional variance is easier to predict, according to equation (2.1), the sign of the return is also easier to predict.

Table 3 reports the results under normal distribution (Model 1), the results when using conditional mean prediction for the decomposition (Model 2), and the results when using the conditional mean prediction as a moment constraint (Model 3). The findings are summarized as the following. First, by looking at the results of Model 1, we confirm the results of Goyal and Welch (2008). That is all of the predictors cannot help to make a better forecast if they are put into a linear regression. Second, by comparing the results of Model 1 and Model 2, it demonstrates again that the decomposed density forecast easily beats the normal density forecast. Therefore, we should use those predictors through a nonlinear model (e.g. the decomposition model). Third, we can see that the historical mean is always among the best predictions. So even
if we use the decomposition model and impose moment constraint, the historical mean prediction in mean is still difficult to beat when compare with other mean predictions. Finally, by examining the results of Model 3, we can see that moment constraints will always make the density forecast worse when applying to the equity premium; but for stock return all of them make the prediction better except the one with all 13 predictors.

To help understand these results, we plot the actual return, and theoretical mean of the decomposition model and the historical mean prediction in Figure 3. From Figure 3b, we can see for the annualized monthly stock returns, the decomposition model generates a very volatile mean forecast. We know that the mean of annualized data should not be volatile since annualizing smooths out the noise. Therefore the decomposition model fail to deliver a good mean prediction for stock return. The historical mean prediction, on the other hand, is smooth and fits the actual stock return well. Thus it is a solid mean constraint candidate and that is why after imposing this moment constraint, we actually improve a lot upon the original density forecast. Figure 3a tells us a different story. For equity premium, the theoretical mean of the decomposition model is already smooth and it is also close to the historical mean prediction which means there is not much space for improvement. Also due to the estimation error in the numerical integral, it is natural to expect that imposing the moment constraint will not generate a better forecast. Another reason for the better prediction in stock return is because stock return is more frequent to be positive. As discussed in Section 3 with Figure 2, the probability of positive returns that deviates further away from 0.5 will lead to higher possibility of improvement. Since equity premium is equal to the stock return minus the risk free rate, the stock return must be more frequently positive and therefore it is more likely for the moment constraint to be effective.
We also plot the actual return, and theoretical mean of the decomposition model and the mean prediction using all predictors in Figure 4. Since the mean prediction using all predictors is not smooth and it also makes opposite prediction in the 1990s, it generates the worst density forecast among all models even when applying to stock return. Thus, it is crucial to find good moment constraints. Adding a bad constraint can necessarily make the density forecast worse. The above arguments can also explain why using conditional mean prediction before the decomposition can make better predictions when using equity premium, and why using the conditional mean prediction as a moment constraint does a good job when applying to the stock return data.

1.6 Conclusions

In this paper, we use a decomposition model that uses the identity that financial return can be written as a product of the absolute return and the sign of the return. It may sound trivial as it is simply an identity. However, the insights from Granger and Ding (1995) and Ding, Granger, and Engle (1993) lead us to follow up a recent work by Anatolyev and Gospodinov (2010). We evaluate the decomposition model in terms of density forecast by using the logarithmic score and the conditional likelihood score. We show that the decomposition model can dominate the normal density forecast models in terms of log score functions in the entire density support. In particular, inspired by the recent econometrics literature on imposing some natural constraints or moment conditions that can help the out-of-sample forecasts, we consider the density forecast of the decomposition model with the maximum entropy principle for “out-of-sample” forecast subject to the “out-of-sample” mean forecast constraint. We
show that when the mean forecast from the decomposition model deviates from that of the simple linear predictive regression, as in the stock return case, imposing the mean forecast constraint will tilt the density forecast from the decomposition model towards the direction that satisfies the mean forecast from a simple model, and can improve the density forecast of the decomposition model. Of course, if a poor mean forecast constraint is used for the moment condition in maximizing the out-of-sample density forecast entropy, it can damage and make the density forecast worse.
References


Table 1.1 Average Out-of-sample Logarithmic Scores of the Benchmark models and the Decomposition model

<table>
<thead>
<tr>
<th></th>
<th>$Q_t(12)$</th>
<th>$R_t(12)$</th>
<th>$Q_t(1)$</th>
<th>$R_t(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1-ZM</td>
<td>-0.0399</td>
<td>-0.3840</td>
<td>1.6613</td>
<td>1.6544</td>
</tr>
<tr>
<td>Model 1-HM</td>
<td>-0.1333</td>
<td>-0.1309</td>
<td>1.6560</td>
<td>1.6595</td>
</tr>
<tr>
<td>Model 2-ZM</td>
<td>0.1763</td>
<td>0.0267</td>
<td>1.6611</td>
<td>1.6610</td>
</tr>
</tbody>
</table>

Notes: Data start from May 1937 and end at December 2002. There are total 788 monthly observations. The full sample period is divided equally for in-sample estimation and pseudo out-of-sample forecast evaluation. $R_t(12), Q_t(12)$ are annualized monthly equity premium and stock return as defined in (2.49) and (2.50). $R_t(1)$ and $Q_t(1)$ are monthly equity premium and stock return. Model 1-ZM is a benchmark density forecast model with normal distribution, zero mean $\mu_t = 0$, and GARCH. Model 1-HM is another benchmark density forecast model with normal distribution, constant mean (historical mean) $\mu_t = \alpha$, and GARCH. Model 2-ZM is the joint density forecast model using the decomposition with $\mu_t = 0$ (ZM) and with $p_t$ being the historical percentage of positive returns in sign specification.
Table 1.2 Average Out-of-sample Logarithmic Scores of Binary Prediction

<table>
<thead>
<tr>
<th>predictor</th>
<th>$Q_t(12)$</th>
<th>$R_t(12)$</th>
<th>$Q_t(1)$</th>
<th>$R_t(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0.1736</td>
<td>0.0267</td>
<td>1.6611</td>
<td>1.6610</td>
</tr>
<tr>
<td>dp</td>
<td>0.1587</td>
<td>-0.0209</td>
<td>1.6628</td>
<td>1.6621</td>
</tr>
<tr>
<td>dy</td>
<td>0.1466</td>
<td>-0.0458</td>
<td>1.6647</td>
<td>1.6626</td>
</tr>
<tr>
<td>ep</td>
<td>0.1784</td>
<td>0.0057</td>
<td>1.6633</td>
<td>1.6625</td>
</tr>
<tr>
<td>de</td>
<td>0.1857</td>
<td>0.0303</td>
<td>1.6627</td>
<td>1.6634</td>
</tr>
<tr>
<td>svar</td>
<td>0.1630</td>
<td>0.0198</td>
<td>1.6607</td>
<td>1.6611</td>
</tr>
<tr>
<td>dfy</td>
<td>0.1586</td>
<td>0.0207</td>
<td>1.6581</td>
<td>1.6576</td>
</tr>
<tr>
<td>infl</td>
<td>0.1772</td>
<td>0.0258</td>
<td>1.6632</td>
<td>1.6626</td>
</tr>
<tr>
<td>lty</td>
<td>0.1646</td>
<td>0.0155</td>
<td>1.6680</td>
<td>1.6645</td>
</tr>
<tr>
<td>ltr</td>
<td>0.1746</td>
<td>0.0267</td>
<td>1.6614</td>
<td>1.6617</td>
</tr>
<tr>
<td>tbl</td>
<td>0.1763</td>
<td>0.0176</td>
<td>1.6627</td>
<td>1.6598</td>
</tr>
<tr>
<td>bm</td>
<td>0.1795</td>
<td>0.0036</td>
<td>1.6591</td>
<td>1.6598</td>
</tr>
<tr>
<td>csp</td>
<td>0.0631</td>
<td>-0.0719</td>
<td>1.6585</td>
<td>1.6604</td>
</tr>
<tr>
<td>ntis</td>
<td>0.0830</td>
<td>-0.1157</td>
<td>1.6600</td>
<td>1.6604</td>
</tr>
<tr>
<td>all</td>
<td>-0.4393</td>
<td>-0.6203</td>
<td>1.6715</td>
<td>1.6736</td>
</tr>
</tbody>
</table>

Notes: Average Out-of-sample Logarithmic Scores when different predictors are used.

The probability of a positive return is specified as $\Pr(v_{t+1} = 1|I_t) = G(a + x_t'b)$ with $G(z) = \frac{e^z}{1+e^z}$ being the logistic function. $x$ can be a single predictor or include many predictors.
Table 1.3 Average Out-of-sample Logarithmic Scores of Different Density Forecast Models

Panel A. Annualized Monthly Return \((k = 12)\)

<table>
<thead>
<tr>
<th></th>
<th>Equity Premium (Q_t(k))</th>
<th>Stock Return (R_t(k))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>ZM</td>
<td>-0.0039</td>
<td>0.1763</td>
</tr>
<tr>
<td>HM</td>
<td>-0.1333</td>
<td>0.1612</td>
</tr>
<tr>
<td>dp</td>
<td>-0.2734</td>
<td>0.0860</td>
</tr>
<tr>
<td>dy</td>
<td>-0.2060</td>
<td>0.0807</td>
</tr>
<tr>
<td>ep</td>
<td>-0.2886</td>
<td>-0.0909</td>
</tr>
<tr>
<td>de</td>
<td>-0.3487</td>
<td>0.1967</td>
</tr>
<tr>
<td>svar</td>
<td>-0.5859</td>
<td>0.1570</td>
</tr>
<tr>
<td>dfy</td>
<td>-0.3755</td>
<td>-0.0574</td>
</tr>
<tr>
<td>infl</td>
<td>-0.1035</td>
<td>0.1849</td>
</tr>
<tr>
<td>lty</td>
<td>-0.4349</td>
<td>0.0642</td>
</tr>
<tr>
<td>ltr</td>
<td>-0.1292</td>
<td>0.1438</td>
</tr>
<tr>
<td>tbl</td>
<td>-0.7774</td>
<td>0.0613</td>
</tr>
<tr>
<td>bm</td>
<td>-0.4382</td>
<td>0.0774</td>
</tr>
<tr>
<td>csp</td>
<td>-0.2477</td>
<td>0.1814</td>
</tr>
<tr>
<td>ntis</td>
<td>-0.7253</td>
<td>0.0338</td>
</tr>
<tr>
<td>all</td>
<td>-2.1306</td>
<td>0.0366</td>
</tr>
</tbody>
</table>
Panel B. Monthly Return \((k = 1)\)

<table>
<thead>
<tr>
<th></th>
<th>Equity Premium (Q_t) ((k))</th>
<th>Stock Return (R_t) ((k))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>ZM</td>
<td>1.6613</td>
<td>1.6611</td>
</tr>
<tr>
<td>HM</td>
<td>1.6560</td>
<td>1.6620</td>
</tr>
<tr>
<td>dp</td>
<td>1.6539</td>
<td>1.6575</td>
</tr>
<tr>
<td>dy</td>
<td>1.6537</td>
<td>1.6572</td>
</tr>
<tr>
<td>ep</td>
<td>1.6510</td>
<td>1.6609</td>
</tr>
<tr>
<td>de</td>
<td>1.6504</td>
<td>1.6591</td>
</tr>
<tr>
<td>svar</td>
<td>1.6501</td>
<td>1.6628</td>
</tr>
<tr>
<td>dfy</td>
<td>1.6473</td>
<td>1.6566</td>
</tr>
<tr>
<td>infl</td>
<td>1.6574</td>
<td>1.6631</td>
</tr>
<tr>
<td>lty</td>
<td>1.6465</td>
<td>1.6537</td>
</tr>
<tr>
<td>ltr</td>
<td>1.6574</td>
<td>1.6637</td>
</tr>
<tr>
<td>tbl</td>
<td>1.6464</td>
<td>1.6628</td>
</tr>
<tr>
<td>bm</td>
<td>1.6454</td>
<td>1.6602</td>
</tr>
<tr>
<td>csp</td>
<td>1.6541</td>
<td>1.6583</td>
</tr>
<tr>
<td>ntis</td>
<td>1.6503</td>
<td>1.6593</td>
</tr>
<tr>
<td>all</td>
<td>1.5965</td>
<td>1.6524</td>
</tr>
</tbody>
</table>

Notes: The mean specification are \(\hat{\mu}_{t+1} = 0\) (ZM), \(\hat{\mu}_{t+1} = \bar{r}_t\) (HM), or \(\hat{\mu}_{t+1} = \hat{\alpha} + x'_t\hat{\beta}\) with different predictors. Model 1 is the benchmark model with normal distribution. Model 2 is the decomposition model. Model 3 is the decomposition model and with a moment constraint. More details on these models are explained in Sections 2 and 3. “all” denotes the models using all 13 predictors of Goyal and Welch (2008) for \(x\).
Figure 1.1 Adding a Moment Condition under Normality

(a) Mean constraint

(b) Variance constraint

Notes: In both panels, the solid line represents the original distribution and the dashed line represents the new distribution with the moment constraint. In Panel (a), the original distribution is $N(2,1)$ and the mean constraint of zero is imposed. In Panel (b), the original distribution is $N(0,4)$ and the variance constraint of 1 is imposed.
Figure 1.2 Plots of $I(\lambda_t)$ and $I(\lambda^*_t)$

(a) $I(\lambda_t)$ with $a = 8$ and $p_t = 0.55$

(b) $I(\lambda_t)$ with $a = 8$ and $p_t = 0.55$

(c) $I(\lambda_t)$ with $a = 8$ and $p_t = 0.65$

(d) $I(\lambda_t)$ with $a = 8$ and $p_t = 0.65$

(e) Optimal $I(\lambda_t^*)$ as a function of $p_t$

Notes: Panels (a,b,c,d) are plots of of $I(\lambda_t)$ against $\lambda_t$ for fixed values of $a$ and $p_t$. Panel (a,b) are plots of $I(\lambda_t)$ against $\lambda_t$ for fixed values of $a = 8$ and $p_t = 0.55$ which are the estimated values for the annualized monthly equity premium $Q_t(12)$. Panel (b) magnifies Panel (a) for $-2 < \lambda_t < 2$. Panel (c,d) are plots of $I(\lambda_t)$ against $\lambda_t$ for fixed
values of $a = 8$ and $p_t = 0.65$ which are the estimated values for the annualized monthly stock returns $R_t(12)$. Panel (d) magnifies Panel (c) for $-2 < \lambda_t < 2$. Panel (e) is a plot of the optimal $I(\lambda^*)$ against $p_t$. 
Figure 1.3 Mean Predictions from Model 1 and Model 2, and the Actual Data

(a) Annualized monthly equity premium

(b) Annualized monthly stock returns

(c) Monthly equity premium $Q_t (1)$

(d) Monthly stock returns $R_t (1)$

Notes: Presented are the two predicted values and the actual data for $Q_t (12)$, $R_t (12)$, $Q_t (1)$, $R_t (1)$. The green line represents the actual data for equity premium or stock return. The blue line is the mean prediction from the benchmark normal density forecast, Model 1-HM. The red line is the mean prediction from the decomposition model, Model 2-ZM, that is the mean of the joint density forecast $f_{t+1}(u, v)$, which is obtained, under the independence assumption between $u$ and $v$, from the product of mean predictions of the marginal densities $f_{t+1}^U(u)$ and $f_{t+1}^V(v)$.
Figure 1.4 Mean Predictions from the Decomposition Model and the Predictive Regression Model Using All of GW’s Predictors, and the Actual Data

(a) Annualized monthly equity premium  (b) Annualized monthly stock returns

Notes: Two predicted values versus the actual data for $Q_t(12)$, $R_t(12)$, $Q_t(1)$, $R_t(1)$. The green line represents the actual data for equity premium or stock return. The blue line is the mean prediction from using the linear predictive regression using all 13 predictors of Goyal and Welch (2008). The red line is the mean prediction from the decomposition model, that is the mean of the joint density forecast $f_{t+1}(u,v)$, which is obtained, under the independence assumption between $u$ and $v$, from the product of mean predictions of the marginal densities $f_{t+1}^U(u)$ and $f_{t+1}^V(v)$. 

37
Chapter 2

Risk Forecast of Financial Returns Using Decomposition and Maximum Entropy

2.1 Introduction

This paper discusses two issues, first we consider how to use a decomposition model to make a density forecast of the financial return and how to improve this density forecast by imposing matching moment constraints; and then we consider the risk forecast and in particular Value-at-Risk (VaR) forecast produced from the density forecast. As is well known that financial returns are not predictable in conditional mean, it is not our interest to forecast the mean. Rather, our interest is to forecast the conditional density of financial returns as well as the risk forecasts. The density forecast model in this paper is based on a decomposition of a financial return series into a product of the absolute value and the sign of the return, and we consider both cases whether the two
parts are independent or not. This paper shows how best we can use the two components, both of which are relatively easily predictable, in forecasting density. That does not necessarily imply we can predict the financial returns in the conditional mean using the decomposition. Our interest is not forecasting the mean returns, but forecasting the conditional density of financial returns using the decomposition. However, we can show that a better density forecast model does produce a better risk forecast in term of VaR forecast. For example, in financial risk management, we are more interested in the density (especially in tails) than the mean, and it can be better predicted using our decomposition density forecast model. This paper also shows how to improve the density forecast using the decomposition through imposing some proper moment constraints. Our trick is to find the maximum entropy density that satisfies some simple moment conditions. The intuition is that although the decomposition model satisfies some of the stylized distribution properties and provides predictability of the two components, its mean forecast constructed from the decomposition model may not match with the mean forecast from a simple predictive regression. As the simple model such as zero mean, constant mean such as historical mean tends to give less estimation error than the complicated decomposition model at least for the mean forecast, we use the maximum entropy principle for the out-of-sample density forecast subject to the matching moment constraint that the mean forecasts from the decomposition model and the simple model are equal. We will explore whether imposing such a natural constraint in the moment condition can help the out-of-sample density forecasts. We show that when the mean forecast from the decomposition model deviates from that of the simple moment conditions, imposing the matching mean forecast constraint will tilt the density forecast of the decomposition model to produce the maximum entropy density forecast. The tilt-
Traditionally econometric modelling has been focused on the conditional moments of variables of interest, particularly mean and variance. Modern research has shifted from the moment based modelling to modelling of the conditional distribution. Density forecast has recently become a popular topic. The common density forecast models for financial returns assume a particular distribution such as Gaussian, Student $t$, or some variants to capture fat-tails or skewness. However, these distributional assumptions of the financial returns, although widely used, are inconsistent with some empirical findings on time series properties. In particular, Granger and Ding (1995, 1996) and Rydén, Teräsvirta and Åsbrink (1998) provide several stylized facts about the financial returns. Let $r_t$ be the return on a financial asset at time $t$, $|r_t|$ denote the absolute value, and $\text{sign}(r_t) = 1 (r_t > 0) - 1 (r_t < 0)$. The following three distributional properties (DP) are frequently studied by researchers.

**DP1:** $|r_t|$ and $\text{sign}(r_t)$ are independent.

**DP2:** $|r_t|$ has the same mean and standard deviation.

**DP3:** The marginal distribution for $|r_t|$ is exponential.

The common distributional assumption like normal or student-$t$ distribution fails to capture any of these three properties since it takes the financial return as an indivisible entity, rather than as two components. However, the density forecast model based on the joint density of $|r_t|$ and $\text{sign}(r_t)$ which can be obtained from modelling of marginal densities of the absolute return and the sign of the return as well as their copula, can capture the above three DP properties.

In addition to the three distributional properties which enable the margin
forecasts of $|r_t|$ and sign$(r_t)$ to be linked to their joint density forecast, the conditional means of the margins are not martingale difference. That is, the conditional means of $|r_t|$ and sign$(r_t)$ can be dynamically modelled unlike that of the returns $r_t$. Ding et al (1993) show that $|r_t|$ is easily predictable, while Korkie, Sivakumar, and Turtle (2002) and Christoffersen and Diebold (2006) show that sign$(r_t)$ is predictable as well. Let $I_t$ be the information set at time $t$. If the return sign indicator series $1(r_{t+1} < 0)$ displays conditional mean serial dependence, namely, if $E[1(r_{t+1} < 0)|I_t]$ is a nonconstant function of $I_t$, then the signs can be predicted. Further, let $\mu_{t+1} = E(r_{t+1}|I_t)$ be the conditional mean and $\sigma^2_{t+1} = E[(r_{t+1} - \mu_{t+1})^2|I_t]$ be the conditional variance. Then Christoffersen and Diebold (2006) note that

$$E[1(r_{t+1} < 0)|I_t] = \Pr(r_{t+1} < 0|I_t) = \Pr\left(\frac{r_{t+1} - \mu_{t+1}}{\sigma_{t+1}} < \frac{-\mu_{t+1}}{\sigma_{t+1}}|I_t\right) = F\left(\frac{-\mu_{t+1}}{\sigma_{t+1}}\right),$$

which shows the sign is predictable if $\sigma_{t+1}$ is predictable and $\mu_{t+1}$ is not zero. Using a series expansion of the conditional distribution $F(\cdot)$, such as the Gram-Charlier expansion or the Edgeworth expansion, we can show that $1(r_{t+1} < 0)$ can be predictable because the conditional higher moments (skewness and kurtosis) are predictable. Because the absolute return and the sign of the stock market return are predictable, the margin density forecast models of $|r_t|$ and sign$(r_t)$ can be specified such that these serial dependence properties associated with the predictability are incorporated. It can eventually yields a more precise density forecast model for $r_{t+1}$.

Although some studies shows that one of the stylized facts about the sign and absolute value of the return is that they are independent over time. However, as also indicated by empirical data, financial returns exhibits properties such as fat tail and negative skewness. The negative skewness observed from data indicates there is
possibility that the sign and absolute value of the return are astrally no independent with each other. As a result, we will consider both cases in our decomposition models in this study. We will first impose the assumption that the sign and absolute value of the return are independent to construct our decomposition model of the density forecast, and then we will incorporate their dependence through different copula functions to see whether release the independence assumption will improve the density forecast. Also, we will test for independency as well.

In addition, although the decomposition model satisfies the above three distribution properties, sometimes it cannot generate good moment predictions even for the first moment (mean). First, the mean of density forecast from the decomposition model may be far from zero which contradicts the fact that stock returns are close to zero in mean especially in high frequency. Second, in our empirical applications we find that when applying the decomposition model to the annualized monthly stock return, the predicted conditional mean fluctuates rather excessively around or even deviating from the historical mean (HM) predictions. But for annualized monthly return, the HM is quite smooth and stable around a constant near zero. If we fix this problem by imposing a moment condition to the density forecast model, we may improve the decomposition model. Since our objective function is the logarithmic score in evaluating the density forecast, the improvement can be achieved by solving the constrained maximum entropy problem. Jaynes (1968) provides the solution for a discrete density. A general solution for any distribution can be seen in Csiszár (1975). Robertson, Tallman, and Whiteman (2005) and Giacomini and Ragusa (2011) apply these solutions to macroeconomic data and obtain nice results. These results encourage us to consider imposing a smooth moment condition when the decomposition model is unable to match the moment forecast.
of a simple mean regression model. Indeed, we find that imposing a sensible moment (mean) condition to the decomposition model can improve the density forecast of the financial returns $r_{t+1}$.

One of the most important application to our density forecast models is its availability to make risk forecasts. Since a better density forecast will fit the out-of-sample data better, especially on the tail part, then we should expect it can produce better risk forecast as well, so that there is possibility to improve upon traditional risk forecast models using our improved density forecasts. To be more specific, we can apply density forecasts to risk forecasts such as the quantiles at different confidence level, or in particular Value-at-Risk at different levels, or risk spectrum which is the weighted average of return quantiles with weighs reflecting different risk aversion; we can also apply density forecasts to calculate the expected short fall, loss given default, unexpected loss, or tail Value-at-Risk. In this paper, we will focus on the forecasts of Value-at-Risk produced different density forecast models.

The rest of the paper is structured as follows. Section 2 introduces the benchmark models with normal distribution as well as the decomposition model which specifies the absolute return and the sign of the return, together with their copula functions. In Section 3, we impose a moment condition constraint to the decomposition density forecast in section 2 to see whether adding the constraint can further improve density forecasts. Section 4 introduces score functions to evaluate the density forecast. Section 5 introduces how to make risk forecast using the density forecast results and gives the comparison rule to evaluate risk forecasts. Section 5 includes empirical results which compares the benchmark density forecasts, the decomposition density forecast and the decomposition density forecasts with constraints, as well as the risk forecasts from all
models. Finally Section 6 concludes.

2.2 Density Forecast Models

2.2.1 Benchmark Models

The Normal Distribution Model

In our benchmark setting, we consider two models to forecast the density of the financial return, following Lee and Xi (2013). In the first model, we assume the return follows a normal distribution which is uniquely specified by its mean and variance. Let $\mu_{t+1}$ denote the conditional mean of $r_{t+1}$ based on the information set at time $T$, i.e.

$$
\mu_{t+1} = E(r_{t+1} | I_t),
$$

then the density forecast of $r_{t+1}$ is given by $r_{t+1} = \mu_{t+1} + \varepsilon_{t+1} = \mu_{t+1} + \sigma_{t+1} z_{t+1}$, where the error term is normally distributed conditional on $I_t$ with $\varepsilon_{t+1} | I_t \sim N(0, \sigma_{t+1}^2)$, and $\varepsilon_{t+1} = \sigma_{t+1} z_{t+1}$, where $z_{t+1} | I_t \sim \text{i.i.d. } N(0, 1)$.

For the conditional mean specification, it is common to assume that it is equal to zero or equal to the historical mean. According to the weak form of efficient market hypothesis (EMH), the best forecast of the conditional mean is zero. EMH has been tested to be true by many studies in the US financial markets. Consider a linear predictive regression for the conditional mean

$$
r_{t+1} = \alpha + x_t' \beta + \varepsilon_{t+1}.
$$

(2.2)

If $\alpha = 0$ and $\beta = 0$, then $\mu_{t+1} = 0$ (zero mean, ZM). In our first specification, we use $\hat{\mu}_{t+1} = 0$ as our out-of-sample forecast for the conditional mean of the return. If $\beta = 0$, then $\mu_{t+1} = \alpha$, where $\alpha$ can be estimated by the historical mean (HM) at time $t$, i.e.

$$
\hat{\mu}_{t+1} = \bar{r}_t = \frac{1}{R} \sum_{s=t-R+1}^{t} r_s.
$$

In our second specification, we use the historical mean by rolling window of $R$ observations to estimate the out-of-sample forecast for
the conditional mean of the return. In more general case, we can also use a set of covariates to forecast the conditional mean, where \( \mu_{t+1} = \alpha + x'_t \beta \) as in Goyal and Welch (2008). Goyal and Welch (2008) find that none of their 17 predictors can make a better mean forecast than the historical mean. Their results further demonstrate that it is very difficult to use a linear mean model to outperform the historical mean specification. Campbell and Thompson (2008) add sign constraints to the coefficient of predictors which make their model essentially nonlinear in mean specification. And as a result, their model makes a better out-of-sample performance. However, according to Lee and Xi (2013), the covariates based conditional mean performs worse than simpler model of ZM or HM in the density forecast. We also tried the covariates in this paper, and the results are also worse than ZM and HM. So in the following part we do not consider using covariates for the density forecast.

For the conditional variance specification in the normal distribution model, we assume a generalized autoregressive conditional heteroskedasticity (GARCH) model specification as in Lee and Xi (2013)

\[
\sigma^2_{t+1} = \gamma_0 + \gamma_1 \varepsilon^2_t + \gamma_2 \sigma^2_t.
\]

(2.3)

Therefore, the first benchmark density forecast is given by

\[
\hat{r}_{t+1} | I_t \sim N \left( \hat{\mu}_{t+1}, \hat{\sigma}^2_{t+1} \right)
\]

(2.4)

where \( \hat{\mu}_{t+1} \) is either estimated using zero mean or historical mean, and \( \hat{\sigma}^2_{t+1} = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} \varepsilon^2_t + \hat{\gamma}_{2,t} \sigma^2_t \) is the predicted conditional variance according to the GARCH (1,1) model using the fitted values \( (\hat{\gamma}_{0,t}, \hat{\gamma}_{1,t}, \hat{\gamma}_{2,t}) \) at time \( t \). This model will be refereed as Model 1 in the following sections, in particular, \( \hat{\mu}_{t+1} = 0 \) is labeled as M1-ZM, and \( \hat{\mu}_{t+1} = \bar{r}_t \) is labeled as M1-HM.
The Independent Decomposition Model

Since the pioneering work by Granger and Ding (1995), the decomposition model for financial returns has been studied in several papers, such as Korkie, Sivakumar, and Turtle (2002), Rydberg and Shephard (2003) and Anatolyev and Gospodinov (2010), among others. Lee and Xi (2013) applied the decomposition model for the density forecast for financial returns where the return can be decomposed as the sign part and the absolute value part. In their paper, they assume the sign and absolute value are independent with each other based on the stylized facts from Teräsvirta and Åsbrink (1998).

In this paper, we will generalize the results in Lee and Xi (2013) to incorporate the possible dependence properties between the sign part and the absolute value part. Therefore, we consider their density forecast model as our second benchmark model. The financial return can be decomposed as the product of its sign and absolute value as follows:

\[ r_{t+1} - \mu_{t+1} = |r_{t+1} - \mu_{t+1}| \times \text{sign}(r_{t+1} - \mu_{t+1}) =: U_{t+1}V_{t+1}, \quad (2.5) \]

where \( U_{t+1} = |r_{t+1} - \mu_{t+1}| \) and \( V_{t+1} = \text{sign}(r_{t+1} - \mu_{t+1}) \).

To get the density forecast of \( r_{t+1} \), several assumptions about the joint and marginal distribution of \( U_{t+1} \) and \( V_{t+1} \) are needed. First, according to the stylized facts of the financial returns, Lee and Xi (2013) imposed the independence assumption about the absolute return \( U_{t+1} \) and the sign of the return \( V_{t+1} \). Then the joint density of \( U_{t+1} \) and \( V_{t+1} \) can be written as:

\[ f_{U,V}^{UV}(u,v) = f_U(u) \times f_V(v). \quad (2.6) \]
To model the marginal distribution of the absolute return $U_{t+1}$, since it takes nonnegative values (like duration), a model similar to the autoregressive duration (ACD) model by Engle and Russell (1998) can be applied, that is,

$$U_{t+1} = \psi_{t+1} e_{t+1},$$

(2.7)

where $\psi_{t+1} = E(U_{t+1}|I_t)$ is the conditional mean of the absolute return $u_{t+1}$ and $e_{t+1}$ is an i.i.d. positive random variable with $E(e_{t+1}|I_t) = 1$.

To model $e_{t+1}$, Engle and Russell (1998) consider exponential and Weibull distributions, Grammig and Maurer (2000) consider the Burr distribution and Lunde (1999) proposes a generalized gamma distribution, both of the Burr and gamma distribution nest the exponential distribution. Based on the stylized facts DP2 and DP3 of the absolute returns, and results from Lee and Xi (2013), we consider only the exponential distribution as our benchmark and also throughout this paper. More complicated distributions may not behave well since they may deviate from the stylized facts, especially DP3. We can observe from the data that absolute return has a strictly decreasing density, yet if we use more complicated distributions, they cannot guarantee this property. In our empirical experiments, we also find that the Weibull distribution gives a much worse result.

The conditional mean $\psi_{t+1}$ is modeled using an ACD-like model:

$$\psi_{t+1} = \delta_0 + \delta_1 |r_t - \mu_t| + \delta_2 \psi_t,$$

(2.8)

While other (nonlinear) specifications such as a logarithmic model of Bauwens and Giot (2000) and a threshold model of Zhang, Russell and Tsay (2001) are possible, the above simple linear model is sufficient and a higher order specification is not necessary to make the density forecast more accurate.
The marginal density forecast of $U_{t+1}$ is then an exponential distribution with mean equal to the conditional mean forecast $\hat{\psi}_{t+1}$ from the above linear model. That is,

$$f_{U_t+1}(u) = \frac{1}{\hat{\psi}_{t+1}} \exp\left(-\frac{1}{\hat{\psi}_{t+1}} u\right).$$

(2.9)

In this decomposition model, we only consider the zero mean case, i.e. $\mu_{t+1} = 0$, to estimate $\hat{\psi}_{t+1}$. We also tried the historical mean case where $\mu_{t+1} = \bar{r}_t$, and as also mentioned by Lee and Xi (2013), the result using historical mean to decompose the density of return is not as good. So we will impose $\mu_{t+1} = 0$ in our decomposition model.

Next, the marginal density of the sign of the return $V_{t+1}$ can be modeled using a Bernoulli-type density since the event is binary. Let $v_{t+1} = 1$ when the sign of the actual stock return at $t+1$ is positive and otherwise $v_{t+1} = -1$. Then the sign forecast density function can be written as:

$$f_{V_t+1}(v) = \begin{cases} 
\hat{p}_{t+1} & \text{if } v = 1 \\
1 - \hat{p}_{t+1} & \text{if } v = -1 
\end{cases}$$

(2.10)

$$= \hat{p}_{t+1}^{\frac{v+1}{2}} (1 - \hat{p}_{t+1})^{\frac{1-v}{2}},$$

where $\hat{p}_{t+1} := \Pr (v_{t+1} = 1 | I_t)$.

To predict $p_{t+1}$, the simplest way is to use the historical percentage of positive returns, that is, $\hat{p}_{t+1} = \frac{1}{R} \sum_{s=t-R+1}^{t} \mathbf{1} (r_s > \mu_s)$. This is a special case of the generalized linear model (GLM) $p_{t+1} = E [\mathbf{1} (r_{t+1} > \mu_{t+1}) | I_t] = G (a + x_t'b)$, where $G(\cdot)$ is a link function. More complicated models can be used to estimate $p_{t+1}$. For instance, If $G(\cdot)$ is the identity function, we have the ordinary least square estimator. $G(\cdot)$ can be also the standard normal cumulative distribution function or the logistic function. In these two cases, we have the probit model or the logit model. However, these more
complicated models may not work better due to the parameter estimation uncertainty, as also showed in Lee and Xi (2013). So in this paper, we only consider the historical percentage estimator of $p_{t+1}$. Once we get $\hat{p}_{t+1}$, the density forecast of the sign of the return will be $f^V_{t+1}(v) = \hat{p}_{t+1} \left( 1 - \hat{p}_{t+1} \right)^{\frac{1-v}{2}}$.

This decomposition model assuming independence between $U_{t+1}$ and $V_{t+1}$ will be referred to as Model 2 independent copula (M2-I) in the following parts, since it is a special case of the decomposition model incorporating copula in the next section.

2.2.2 Decomposition Model using Copulas

Testing for Independence

Although stated as one of the stylized facts in some studies for the independence between the sign part and absolute value part of the financial return, a more flexible model which incorporates the possible dependence between them two could improve the forecast to some extent. To see the need to consider such flexible models, we will first implement a independency test.

If the absolute value of the return $U_{t+1}$ and the sign part $V_{t+1}$ are independent with each other, the joint density will be equal to the product of their marginal density. So in order to test the independency, the null hypothesis is:

\[ H_0^1 : f_{t+1}^{UV}(u,v) = f_{t+1}^U(u) \times f_{t+1}^V(v). \] (2.11)

Since the joint density function of $U_{t+1}$ and $V_{t+1}$ can be rewritten as the decomposition of their marginal density function and the conditional copula function, regardless of their dependency, as follows:

\[ f_{t+1}^{UV}(u,v) = f_{t+1}^U(u) \times f_{t+1}^V(v) \times c \left( F_{t+1}^U(u), F_{t+1}^V(v) \right) \] (2.12)
where \( c(F^U_{t+1}(u), F^V_{t+1}(v)) \) is defined the conditional copula density function such that \( c(w_1, w_2) = \frac{\partial^2 C(w_1, w_2)}{\partial w_1 \partial w_2} \), and \( C(\cdot) \) is denoted for the conditional copula function, and it is defined as a function obtained by the conditional joint cumulative distribution function of \( U_{t+1} \) and \( V_{t+1} \), \( F^U_{t+1}(u) \) and \( F^V_{t+1}(v) \), such that:

\[
F^U_{t+1}(u) = \Pr(U \leq u | I_t), F^V_{t+1}(v) = \Pr(V \leq v | I_t).
\]

Moreover, the conditional copula function can be rewritten as:

\[
C(w_1, w_2) = F^U_{t+1} \left( (F^U_{t+1})^{-1}(w_1), (F^V_{t+1})^{-1}(w_2) \right)
\]

where \( w_1 = F^U_{t+1}(u) \) and \( w_2 = F^V_{t+1}(v) \), while \( (F^U_{t+1})^{-1}(w_1) \) and \( (F^V_{t+1})^{-1}(w_2) \) denote the inverse function of the cumulative distribution function of \( U_{t+1} \) and \( V_{t+1} \).

From (2.11) and (2.12), to test the null hypothesis that the sign and absolute value of the return are independent, we just need to test that the conditional copula function is equal to 1. That is, the null hypothesis in (2.11) can be stated as the null hypothesis that the copula density is the independent copula,

\[
H^2_0 : c(F^U_{t+1}(u), F^V_{t+1}(v)) = 1.
\]

Following Lee and Yang (2013), Hong and Li (2005), we test for \( H^2_0 \) in (2.15) by estimating \( c(F^U_{t+1}(u), F^V_{t+1}(v)) \) using a nonparametric predictive copula density:

\[
\hat{c}_P(F^U_{t+1}(u), F^V_{t+1}(v)) = \frac{1}{P} \sum_{t-R}^{T-1} K_h(w_1, \hat{w}_{1,t+1})K_h(w_2, \hat{w}_{2,t+1}),
\]

where \( w_1 = F^U_{t+1}(u) \), \( w_2 = F^V_{t+1}(v) \), and with no confusion, we will skip the subscript in \( w_1 \) and \( w_2 \) and thereafter, while \( K_h(\cdot) \) is a kernel function and

\[
\hat{w}_1 = \frac{\hat{F}^U_{t+1}(u)}{R} \sum_{s=t-R+1}^{t} 1(u_s \leq u_{t+1}),
\]

\[
\hat{w}_2 = \frac{\hat{F}^V_{t+1}(v)}{R} \sum_{s=t-R+1}^{t} 1(v_s \leq v_{t+1}),
\]
are the out-of-sample marginal empirical distribution functions (EDF) that have been estimated using the rolling window samples of the most recent $R$ observations at each time $t = R, \ldots, T - 1$. To circumvent the boundary problem of the EDF, we follow the boundary-modified kernel used by Hong and Li (2005):

$$K_h(a, a') = \begin{cases} 
    h^{-1}k\left(\frac{a-a'}{h}\right)\int_{-1}^{1} k(u)du, & \text{if } a \in [0, h), \\
    h^{-1}k\left(\frac{a-a'}{h}\right), & \text{if } a \in [h, 1-h), \\
    h^{-1}k\left(\frac{a-a'}{h}\right)\int_{-1}^{1(1-a)/h} k(u)du, & \text{if } a \in (1-h, 1].
\end{cases} \quad (2.19)$$

where $k(\cdot)$ is a symmetric kernel function and $h$ is the bandwidth such that $h \to 0, nh \to \infty$ as $n \to \infty$. We will use the quadratic kernel function $k(z) = \frac{15}{16}(1 - u^2)^2 1(|z| \leq 1)$ in the empirical part, following Hong and Li (2005).

Then the test statistic for the null hypothesis $H_0^2$ in (2.15), based on a quadratic form, is given by:

$$\hat{M}_P = \int_0^1 \int_0^1 \left[ \hat{c}_P(w_1, w_2) - 1 \right]^2 dw_1 dw_2. \quad (2.20)$$

And following Lee and Yang (2013), $\hat{M}_P$ can be pivotalized by properly centered and scaled as

$$\hat{T}_P = \left[ Ph\hat{M}_P - A_h^0 \right] / V_0^{1/2}, \quad (2.21)$$

where $A_h^0$ is the nonstochastic centering factor and $V_0$ is the nonstochastic scale factor, with the form:

$$A_h^0 = \left[ (h^{-1} - 2) \int_{-1}^1 k^2(w_1)dw_1 + 2 \int_0^1 \int_{-1}^b k^2(w_1)dw_1 db \right]^2 - 1, \quad (2.22)$$

$$V_0 = 2 \left[ \int_{-1}^1 \left[ \int_{-1}^1 k(w_1 + w_2)k(w_2)dw_2 \right]^2 dw_1 \right]^2, \quad (2.23)$$

in which $k_b(\cdot) = k(\cdot) / \int_{-1}^b k(z)dz$. Hong and Li (2005) show, under some regularity conditions, $\hat{T}_P$ follows the standard normal distribution asymptotically as $P \to \infty$ under $H_0^2$ in (2.15). In the following empirical part, $\hat{T}_P$ and its asymptotic p-values
are reported. Larger $\hat{T}_P$ and smaller p-values are suggest rejection of $H^2_0$ in 2.15, which implies that the absolute value of the return $U_{t+1}$ and sign of the return $V_{t+1}$ are not independent.

**The Decomposition Model**

From our empirical results reported below, the independency between the sign and absolute value of the return is rejected, which indicates the need to incorporate the interdependency between them. So we will assume a more flexible density forecast decomposition model in this paper.

One way to incorporate dependence between $U_{t+1}$ and $V_{t+1}$ is to include copula in the joint density function $f^{UV}_{t+1}(u, v)$, as in equation 2.12. However, due to the binary property of the Bernoulli-Type distribution of the sign of the return $V_{t+1}$, such decomposition does not work for all types of copula functions. Instead, we follow Anatolyev and Gospodinov (2010) and consider the following representation of the joint density function decomposition.

$$f^{UV}_{t+1}(u, v) = f^U_{t+1}(u)\rho_{t+1}(1 - \rho_{t+1})^{1/2}$$  \hspace{1cm} (2.24)

where $\rho_{t+1} = \rho_{t+1}(F^U_{t+1}(u)) = 1 - \partial C(F^U_{t+1}(u), 1 - p_{t+1})/\partial w_1$. This decomposition can be proved following similar steps as in Anatolyev and Gospodinov (2010) with minor changes. Since the sign variable $v$ takes only the value of 1 and -1, and the absolute variable $u$ is continuous, thus

$$f^{UV}_{t+1}(u, v) = \frac{\partial F^{UV}_{t+1}(u, v)}{\partial u} - \frac{\partial F^{UV}_{t+1}(u, v - 2)}{\partial u}.$$
Plug in $F_{t+1}^{U,V}(u, v) = C(F_{t+1}^{U}(u), F_{t+1}^{V}(v))$, we get

$$f_{t+1}^{U,V}(u, v) = \frac{\partial C(F_{t+1}^{U}(u), F_{t+1}^{V}(v))}{\partial u} - \frac{\partial C(F_{t+1}^{U}(u), F_{t+1}^{V}(v - 2))}{\partial u}.$$

Since for the conditional copula density function, we always have $c(w_1, 1) \equiv w_1$, $c(w_1, 0) \equiv 0$, thus $\frac{\partial c(w_1, 1)}{\partial w_1} = 1$ and $\frac{\partial c(w_1, 0)}{\partial w_1} = 0$. Therefore, when $v = -1$, $F_{t+1}^{V}(v) = 1 - p_{t+1}$, $F_{t+1}^{V}(v - 2) = 0$, and

$$f_{t+1}^{UV}(u, v) = f_{t+1}^{U}(u)\left[\frac{\partial C(F_{t+1}^{U}(u), 1 - p_{t+1})}{\partial w_1} - 0\right].$$

and when $v = 1$, $F_{t+1}^{V}(v) = 1$, $F_{t+1}^{V}(v - 2) = 1 - p_{t+1}$, and

$$f_{t+1}^{UV}(u, v) = f_{t+1}^{U}(u)\left[1 - \frac{\partial C(F_{t+1}^{U}(u), 1 - p_{t+1})}{\partial w_1}\right].$$

Define $\rho_{t+1} = \rho_{t+1}(F_{t+1}^{U}(u)) = 1 - \frac{\partial C(F_{t+1}^{U}(u), 1 - p_{t+1})}{\partial w_1}$, then we can write the decomposition of $f_{t+1}^{UV}(u, v)$ in the form of the above equation 2.24.

In our empirical part, we will first consider the independent copula, which just simplifies to model 2 which impose the assumption that $U_{t+1}$ and $V_{t+1}$ are independent. Then we consider other three types of copula that incorporates dependence between $U_{t+1}$ and $V_{t+1}$: Frank copula, Clayton copula and Farlie-Gumbel-Morgenstern copula. Their conditional copula function, conditional copula density function and the $\rho$-function are given as follows.

(1) Independent copula

$$C^{\text{Indep}}(w_1, w_2) = w_1 w_2,$$

$$c^{\text{Indep}}(w_1, w_2) = 1,$$

$$\rho_{t+1}^{\text{Indep}}(w_1, p_{t+1}, \theta) = p_{t+1}.$$
(2) Frank copula

\[ C^{\text{Frank}}(w_1, w_2, \theta) = \frac{-1}{\theta} \log \left( 1 + \frac{(e^{-\theta w_1} - 1)(e^{-\theta w_2} - 1)}{(e^{-\theta} - 1)} \right), \theta \in (-\infty, +\infty) \setminus \{0\}, \]

\[ c^{\text{Frank}}(w_1, w_2, \theta) = \frac{\theta(1 - e^{-\theta})e^{-\theta(w_1+w_2)}}{[(1 - e^{-\theta}) - (1 - e^{-\theta w_1})(1 - e^{-\theta w_2})]^2}, \]

\[ \rho_{t+1}(w_1, p_{t+1}, \theta) = \left( 1 - \frac{1 - e^{-\theta(1-p_{t+1})}}{1 - e^{\theta p_{t+1}}} e^{\theta(1-w_1)} \right)^{-1}. \]

(3) Clayton copula

\[ C^{\text{Clayton}}(w_1, w_2, \theta) = \left( w_1^{-\theta} + w_2^{-\theta} - 1 \right)^{-1/\theta}, \theta \in [-1, +\infty) \setminus \{0\}, \]

\[ c^{\text{Clayton}}(w_1, w_2, \theta) = \frac{(1 + \theta)(w_1^{-\theta} + w_2^{-\theta} - 1)^{-\frac{1}{\theta}-2}}{(w_1 w_2)^{\theta+1}}, \]

\[ \rho_{t+1}(w_1, p_{t+1}, \theta) = 1 - \left( 1 + \frac{(1 - p_{t+1})^{-\theta} - 1}{w_1^{-\theta}} \right)^{-1/\theta-1}. \]

(4) Farlie-Gumbel-Morgenstern copula

\[ C^{\text{FGM}}(w_1, w_2, \theta) = w_1 w_2 (1 + \theta(1 - w_1)(1 - w_2)), \theta \in [-1, 1], \]

\[ c^{\text{FGM}}(w_1, w_2, \theta) = 1 + \theta - 2\theta w_1 - 2\theta w_2 + 4\theta w_1 w_2, \]

\[ \rho_{t+1}(w_1, p_{t+1}, \theta) = 1 - (1 - p_{t+1})(1 + \theta p_{t+1}(1 - 2w_1)). \]

where \( w_1 = F^U_{t+1}(u), w_2 = F^V_{t+1}(v) \) as before.

Notice that Frank copula and Farlie-Gumbel-Morgenstern copula are symmetric copula, while Clayton copula is asymmetric copula which shows lower tail dependence. For Frank and Farlie-Gumbel-Morgenstern copula, \( \theta < 0 \) implies negative dependence and \( \theta > 0 \) implies positive dependence. For Clayton copula, \( \theta \to +0 \) leads to independence between \( w_1 \) and \( w_2 \) and in this case \( \rho_{t+1} \to p_{t+1} \). For Frank copula, \( \theta \to 0 \) and \( \rho_{t+1} \to p_{t+1} \) implies independence and for Farlie-Gumbel-Morgenstern copula \( \theta = 0 \) and \( \rho_{t+1} = p_{t+1} \) implies independence.

In terms of estimation and density forecast process, one can jointly estimate all parameters, including the parameters in the copula function, as well as the parameters
of the joint density at the same time. However, since we have already obtained the marginal density forecasts of $U_{t+1}$ and $V_{t+1}$, we can take advantage of this and estimate the unknown parameters in the copula functions based on our marginal density forecast of $U_{t+1}$ and $V_{t+1}$. Shih and Louis (1995) show that this two step estimation is consistent but may not be efficient. Also see Chen, Fan and Tsyrennikov (2006) for a detailed discussion.

In our framework, we need to estimate the parameter $\theta$ in the copula function which also goes into $\rho_{t+1}(F_{t+1}^U(u))$, that is, we can rewrite the joint density function as

$$f_{t+1}^{UV}(u,v,\theta) = f_{t+1}^U(u)(\rho_{t+1}(F_{t+1}^U(u),\theta))^{\frac{1}{2}} (1 - \rho_{t+1}(F_{t+1}^U(u),\theta))^{\frac{1}{2}}.$$  

We want to obtain the maximum likelihood estimator (MLE) of the parameter in the copula functions, which is given by:

$$\hat{\theta} = \arg \max_\theta \sum_{t=R+1}^T \log[f_{t+1}^U(u)(\rho_{t+1}(F_{t+1}^U(u),\theta))^{\frac{1}{2}} (1 - \rho_{t+1}(F_{t+1}^U(u),\theta))^{\frac{1}{2}}]$$

$$= \arg \max_\theta \sum_{t=R+1}^T \left[ \log(f_{t+1}^U(u)) + \frac{v+1}{2} \log(\rho_{t+1}(F_{t+1}^U(u),\theta)) + \frac{1-v}{2} \log(1 - \rho_{t+1}(F_{t+1}^U(u),\theta)) \right].$$

And since $f_{t+1}^U(u)$ does not depend on $\theta$, we can remove the first term in the likelihood function when we do the maximization. In our two-step estimation procedure, we first obtain the marginal density of $U_{t+1}$, and then we can have $\hat{F}_{t+1}^U(u)$ and substitute it into the above expression and get the MLE of $\theta$ by

$$\hat{\theta} = \arg \max_\theta \sum_{t=R+1}^T \left[ \frac{v+1}{2} \log(\rho_{t+1}(\hat{F}_{t+1}^U(u),\theta)) + \frac{1-v}{2} \log(1 - \rho_{t+1}(\hat{F}_{t+1}^U(u),\theta)) \right].$$ (2.25)

In the following parts, we will refer the decomposition model using different copula functions as model 2. In particular, we refer these models as model 2 independent copula (M2-I) which simplifies to the second benchmark model in the previous section, model 2 Frank copula (M2-F), model 2 Clayton copula (M2-C) and model 2 Farlie-Gumbel-Morgenstern copula (M2-FGM).
2.3 Decomposition with Moment Constraints

A possible problem with the joint density forecast model using decomposition is that the mean prediction $E(U_{t+1}V_{t+1}|I_t)$ from the joint density forecast function $f_{t+1}^{UV}(u,v)$ of the decomposition model (Model 2) may not be equal to zero, which contradicts the fact that we imposed $\mu_{t+1} = 0$ in estimation of $\psi_{t+1}$ in 2.8, also this will deviate from model 1 where either zero mean or historical mean is used. In other words, the estimated decomposition model is possible to make $E(r_{t+1} - \mu_{t+1}|I_t) \neq 0$, which is inconsistent to our decomposition model assumption, as mentioned in Lee and Xi (2013). For independent copula case, if and only if $\hat{p}_{t+1}$ is 0.5, then the mean of the joint density forecast is 0, and otherwise, we always have $E(r_{t+1} - \mu_{t+1}|I_t) \neq 0$.

From empirical results, $\hat{p}_{t+1}$ is greater than 0.5, since usually we have more positive return months than negative return months, as stock prices tend to rise in the long run. With this in mind, we want to impose the (conditional) moment conditions to the decomposition model that

$$E(U_{t+1}V_{t+1}|I_t) = \mu_{t+1},$$

(2.26)

into the decomposition model of the previous part. Since in our first benchmark model, we allow two cases for $\mu_{t+1}$, zero mean and historical mean. We will use both values as our constraints in this section.

Since in this paper, we will use the logarithmic score to evaluate and compare different density forecast models, we can see the issue from a different angle that imposing the moment constraint is equivalent to solve the following constrained maximization problem of the cross-entropy of the new density forecast $h_{t+1}^{UV}(u,v)$ with respect to the
original density forecast \( f_{t+1}^{UV}(u, v) \):

\[
\max_{h_{t+1}^{UV}(u, v)} = - \iint \log \frac{h_{t+1}^{UV}(u, v)}{f_{t+1}^{UV}(u, v)} h_{t+1}^{UV}(u, v) \, du \, dv
\]  
(2.27)

subject to

\[
\iint m_t(u, v) h_{t+1}^{UV}(u, v) \, du \, dv = 0, \tag{2.28}
\]

and

\[
\iint h_{t+1}^{UV}(u, v) \, du \, dv = 1, \tag{2.29}
\]

where \( f_{t+1}^{UV}(u, v) \) is the density forecast from the decomposition model and \( h_{t+1}^{UV}(u, v) \) is a new density forecast satisfying the moment constraint of (2.28). In particular, with the conditional mean forecast condition in (2.26) imposed to the decomposition model, we have

\[
m_t(u, v) = uv - \hat{\mu}_{t+1} \tag{2.30}
\]

and the expectation in (2.26) is evaluated using the new tilted density forecast \( h_{t+1}^{UV}(u, v) \).

Note that the moment constraint function \( m_t(u, v) \) is denoted with the subscript \( t \) as it is measurable with respect to the information \( I_t \) at time \( t \) (as \( \hat{\mu}_{t+1} \) is \( I_t \)-measurable). The moment condition (2.28) with (2.30) will make the new joint density forecast \( h_{t+1}(u, v) \) have the same mean forecast as the benchmark Model 1.

The maximization of (2.27) subject to the moment constraint (2.28) is well known in the maximum entropy literature. Jaynes (1957) was the pioneer to consider this problem and together with Jaynes (1968), the two studies provide a solution for discrete density. A general solution for any type of density can be found in Csiszár (1975), also see Maasoumi (1993), Zellner (1994), Golan, Judge, and Miller (1996), Ullah (1996), Bera and Bilias (2002), among others.

The solution to the above maximization problem, if exists, is given by

\[
h_{t+1}^{UV}(u, v) = f_{t+1}^{UV}(u, v) \exp \{ \eta_t^* + \lambda_t^* m_t(u, v) \}, \tag{2.31}
\]
where

$$\lambda_t^* = \arg \min_{\lambda_t} I_t(\lambda_t),$$  \hspace{1cm} (2.32)

$$\eta_t^* = \log \{I_t(\lambda_t^*)\}^{-1},$$  \hspace{1cm} (2.33)

and

$$I_t(\lambda_t) = \int \int \exp [\lambda_t m_t(u, v)] f_{t+1}^{UV}(u, v) du dv.$$  \hspace{1cm} (2.34)

That is, to get the solution $h_{t+1}^{UV}(u, v)$, we constructed a new density forecast by exponentially tilting through $\lambda_t^*$ and normalizing it through $\eta_t^*$. This derivation can also be found in recent econometric applications of the maximum entropy, as in Imbens, Spady and Johnson (1998), Robertson, Tallman and Whiteman (2005), Park and Bera (2006), Bera and Park (2008), Stengos and Wu (2010), and Giacominini and Ragusa (2011), among others. Note that $\int \log \frac{h_{t+1}^{UV}(u, v)}{h_{t+1}^{UV}(u, v)} h_{t+1}^{UV}(u, v) du dv$ in the objective function of (2.27) is the Kullbeck-Leibler (1951) information criterion (KLIC) divergence measure between the new density and the original density.

To find $\lambda_t^*$, we need first to find the function of $I_t(\lambda_t)$, which the integral of the joint density function of $U_{t+1}$ and $V_{t+1}$ times an exponential function of the moment constraint. Following Lee and Xi (2013), we will use the numerical integral in the empirical part, since the analytical solution to $I_t(\lambda_t)$ does not have an explicit expression under the historical mean constraint as well as for some copula functions. To implement the numerical integral, we follow the following procedure: since by 2.34, we have

$$I_t(\lambda_t) = E_t \left( \exp [\lambda_t m_t(u, v)] \right),$$  \hspace{1cm} (2.35)

where the expectation is taken over the joint density forecast $f_{t+1}^{UV}(u, v)$. Then to get the numerical integral, we just need to generate $S$ random draws $\{u_t^s, v_t^s\}_{s=1}^S$ from $f_{t+1}^{UV}(u, v)$,
and then calculate $I_t(\lambda_t) = \frac{1}{S} \sum_{s=1}^{S} \exp\{\lambda_t \lambda(u^s_t, v^s_t)\}$. Then $\lambda^*_t$ can be solved by minimizing $I_t(\lambda_t)$, and then $\eta^*_t$ is obtained by $\eta^*_t = \log\{I_t(\lambda^*_t)\}^{-1}$.

To generate the random draws of $\{u^*_t, v^*_t\}_{s=1}^{S}$ from $f_{t+1}^{UV}(u, v)$ under independent copula, we just need to generate $U_{t+1}$ and $V_{t+1}$ separately according to their marginal density functions since the joint density $f_{t+1}^{UV}(u, v)$ is just equal to the product of the two marginal density functions under independency. For the random draws under dependency, an easy way to generate $\{u^*_t, v^*_t\}_{s=1}^{S}$ from $f_{t+1}^{UV}(u, v)$ is to first generate $U_{t+1}$ based on its exponential marginal density function $f_{t+1}^{U}(u) = a \exp(-au)$, where $a = \frac{1}{\psi_{t+1}}$, and then generate $V_{t+1}$ based on the conditional density of $f^{V|U}(v|u)$, and since $V_{t+1}$ is binary, the conditional density should be Bernoulli-type. From 2.24, the conditional density of $V_{t+1}$ conditioning on $U_{t+1}$ is given by:

$$f_{t+1}^{V|U}(v|u) = \frac{f_{t+1}^{UV}(u, v)}{f_{t+1}^{U}(u)} = \rho^{\frac{1+v}{2}}(1 - \rho)^{\frac{1-v}{2}}, v = -1, 1$$

The decomposition model (Model 2) with the moment constraint imposed will be called Model 3. In particular, we will call model 3 imposing zero mean moment condition where $m_t(u, v) = uv - \hat{\mu}_{t+1}$ and $\hat{\mu}_{t+1} = 0$ as M3-ZM and imposing the historical mean condition where $\hat{\mu}_{t+1} = \bar{r}_t$ as M3-HM. And for each copula function, we will denote the model using the name of copula such as M3-ZM-I denote model 3 with zero mean constraint and independent copula. And the label for other copula functions will following similarity. The link between model 2 and 3 is that, Model 2 in this paper is the original density forecast model $f_{t+1}(u, v)$, while Model 3 is the tilted density forecast model $h_{t+1}(u, v)$. A goal of this paper is to examine if imposing the moment constraint can improve Model 3 over Model 2-ZM in density forecast with different copula functions.
2.4 Density Forecast Evaluation

To evaluate density forecasts from different models, we will apply a scoring rule. A scoring rule is a positive-oriented criterion to evaluate density forecasts. A larger expected score usually means that the associated density forecast is better. Formally speaking, a score function or scoring rule $S(f, y)$ for a single point takes value in the real line $\mathbb{R}$ or in the extended real line $\overline{\mathbb{R}} = [-\infty, \infty]$, where $f$ is the density forecast and $y$ is the realized value. Let $E_h S(f, y) = \int S(f, y) h(y) dy$ be the expected value of $S(f, y)$ under the distribution function $h(\cdot)$. A scoring rule is said to be proper if $E_h S(h, y) \geq E_h S(f, y)$ for all distribution functions $f(\cdot)$ and $h(\cdot)$. If the equality holds only if $f(\cdot) = h(\cdot)$, then the score function is strictly proper.

Since for a proper scoring rule, the expected score of the true density is always greater than the expected score of any other density. This property encourages people to reveal his or her true belief. For example, one of the most popular scoring rules is the likelihood based scoring rule:

$$S^L(f, y) = \log f(y).$$ (2.36)

The difference of the expected scores $[E_h S(h, y) - E_h S(f, y)]$ is the KLIC divergence measure. It can be shown that the likelihood based score is strictly proper because

$$KLIC(h, f) = E_h [S^L(h, y) - S^L(f, y)] = E_h [\log h(y) - \log f(y)] \geq 0$$ (2.37)

due to the Jensen’s inequality applied to the logarithmic function which is a concave function. See, e.g., Rao (1965), White (1994), and Ullah (1996).

To apply the logarithmic score to the decomposition model, let $y = (u, v)$ and

$$S^L(f_{t+1}, y_{t+1}) = S^L(f_{t+1}, (u_{t+1}, v_{t+1})) = \log f_{t+1}(u_{t+1}, v_{t+1}).$$ (2.38)
In fact, the logarithmic score of this joint density of $U$ and $V$ can be compared with that of the benchmark model. Since $r_{t+1} \equiv U_{t+1}V_{t+1}$, in the benchmark model we implicitly assume that the joint density forecast between $U_{t+1}$ and $V_{t+1}$, conditional on the information set $I_t$ is:

$$f_{t+1}^B(u,v) = \frac{1}{\sqrt{2\pi \hat{\sigma}_{t+1}}} \exp\{-\frac{(uv - \hat{\mu}_{t+1})^2}{2\hat{\sigma}_{t+1}^2}\}, \quad (2.39)$$

where $u \geq 0$ and $v = 1$ or $-1$. One can show that the above function is a valid density function. Also, $S^L(f_{t+1}^B(r), r_{t+1}) \equiv S^L(f_{t+1}^B(u,v), (u_{t+1}, v_{t+1}))$. So we are actually comparing two joint densities between $U_{t+1}$ and $V_{t+1}$.

An advantage of assuming independence in the decomposition model (based on the stylized fact, DP1) is that we can improve the density forecast by improve any of the two marginal density forecasts. From equation (2.6),

$$S^L(f_{t+1}, (u_{t+1}, v_{t+1})) = \log f_U^{t+1} (u_{t+1}) + \log f_V^{t+1} (v_{t+1}). \quad (2.40)$$

Once we obtain a better marginal density forecast for $u$, the logarithmic score of the joint density will be improved since the marginal density of $v$ remains unaffected. The opposite is also true. Therefore, we can always try different specifications for the absolute return and the sign of the return to make a better joint density forecast.

After calculating the score function at each time $t$, we can compare density forecasts by computing the average out-of-sample scores. That is

$$\bar{S}_P = \frac{1}{P} \sum_{t=T-P+1}^{T} S_t, \quad (2.41)$$

where $S_t$ is the logarithmic score $S^L(f_{t+1}, (u_{t+1}, v_{t+1}))$ or the conditional likelihood score $S^{CL}(f_{t+1}, (u_{t+1}, v_{t+1}))$ at time $t$. The density forecast with a higher value of $\bar{S}_P$ is considered to be the better density forecast.
2.5 Risk Forecast

One of the most important applications to our density forecast models is its availability to make risk forecasts. Since a better density forecast will fit the out-of-sample data better, then we should expect better risk forecast as well, so that there is possibility to improve upon traditional risk forecast models using our improved density forecasts. To be more specific, we can apply density forecasts to risk forecasts such as the quantiles at different confidence levels, or in particular Value-at-Risk (VaR) at different levels, or risk spectrum which is the weighted average of return quantiles with weights reflecting different risk aversion; we can also apply density forecasts to calculate the expected shortfall, loss given default, unexpected loss, or tail Value-at-Risk. In this paper, we will focus on the forecasts of Value-at-Risk produced by different density forecast models.

In this section, we set the Riskmetrics model of J.P. Morgan (1995) as our first benchmark. The Riskmetrics model forecasts quantile, denoted by \( \hat{\lambda}_{t+1} \), as follows:

\[
\hat{\lambda}_{t+1}(\alpha) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} \Phi^{-1}(\alpha),
\]

(2.42)

where \( \Phi^{-1}(\cdot) \) is the inverse of the standard normal cumulative distribution function so that \( \Phi(0.01) = -2.33, \Phi(0.05) = -1.645 \) and \( \Phi(0.1) = -1.28 \). While \( \hat{\mu}_{t+1} = \bar{r}_t = \frac{1}{t} \sum_{j=1}^{t} r_j \) is the historical mean with recursive window, and \( \hat{\sigma}_{t+1} \) is estimate recursively by the exponentially weighted moving average (EWMA)

\[
\hat{\sigma}_{t+1}^2 = 0.94\hat{\sigma}_{t}^2 + 0.06(r_t - \hat{\mu}_{t+1})^2.
\]

(2.43)

Another benchmark is the VaR forecast from normal distribution density forecast (M1). Since this density forecast is uniquely determined by its mean forecast
\( \hat{\mu}_{t+1} \) and standard deviation forecast \( \hat{\sigma}_{t+1} \), the VaR forecast is then given by \( \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} \Phi^{-1}(\alpha) \).

For the decomposition models M2 and M3, it is difficult to derive the VaR analytically or most of time no explicit solutions, so we use numerical methods to forecast VaR from the density forecast. Since we know the joint density forecast \( f_{t+1}^{UV}(u, v) \) in M2 and the joint density forecast under moment constraint \( h_{t+1}^{UV}(u, v) \) in M3, we can again generate \( S \) random draws of \( \{u^*_s, v^*_s\}_{s=1}^S \) from \( f_{t+1}^{UV}(u, v) \) or \( h_{t+1}^{UV}(u, v) \) accordingly, and calculate the return by \( r_t = u_tv_t \) and find the \( \alpha \) percentage level of VaR by taking the \( \alpha S \)'s lowest return from our random draw as the forecasted VaR for a given \( \alpha \).

To generate the random draws of \( \{u^*_s, v^*_s\}_{s=1}^S \) from \( f_{t+1}^{UV}(u, v) \), we can just follow the same procedure in the last section. However, to generate the random draws of \( \{u^*_s, v^*_s\}_{s=1}^S \) from \( h_{t+1}^{UV}(u, v) \) is less straightforward. Still we want to follow similar steps, that we first generate \( U_{t+1} \) based on its marginal density \( h_{t+1}^U(u) \), and then generate \( V_{t+1} \) based on the conditional density of \( h^V|U(v|u) \), and since \( V_{t+1} \) is binary, the conditional density should also be Bernoulli-type. Substitute 2.24 and 2.30 to 2.31, the joint density of \( U_{t+1} \) and \( V_{t+1} \) under moment constraint can be written as:

\[
\begin{align*}
    h_{t+1}^{UV}(u, v) &= f_{t+1}^U(u) \frac{1}{\sqrt{2\pi}} (1 - \rho_{t+1})^{1-u} \exp[\eta^* + \lambda^*(uv - \hat{\mu}_{t+1})].
\end{align*}
\]

Since the sign part is binary where \( v \) can only take value of 1 or -1, the marginal distribution of \( u \) can then be written as:

\[
\begin{align*}
    h_{t+1}^U(u) &= h_{t+1}^{UV}(u, -1) + h_{t+1}^{UV}(u, 1) \\
    &= f_{t+1}^U(u)(1 - \rho_{t+1}) \exp(\eta^* - \lambda^*\hat{\mu}_{t+1} + \lambda^*u) + f_{t+1}^U(u)\rho_{t+1} \exp(\eta^* - \lambda^*\hat{\mu}_{t+1} + \lambda^*u) \\
    &= \exp(\eta^* - \lambda^*\hat{\mu}_{t+1})f_{t+1}^U(u)[(1 - \rho_{t+1}) \exp(-\lambda^*u) + \rho_{t+1} \exp(\lambda^*u)].
\end{align*}
\]
then the conditional density of $V_{t+1}$ given realizations of $U_{t+1}$ is given by:

$$h_{t+1}(v|u) = \frac{h_{t+1}(u,v)}{h_{t+1}(u)} = \frac{\rho \frac{1}{\psi_{t+1}} (1 - \rho)^{\frac{1}{\psi_{t+1}}} \exp(\lambda u)}{(1 - \rho) \exp(-\lambda u) + \rho \exp(-\lambda u)}$$

However, it is not always easy to generate $u$’s from the $h_{t+1}(u)$ function since it is not a common density function especially when the $\rho_{t+1}$ function is complicated for some copulas. To solve this problem, we consider two possible methods.

The straight method to generate $u$ from $h_{t+1}(u)$ is to use PIT (probability integral transformation). Since we assume that $u$ follows an exponential distribution with mean equal to $\frac{1}{\psi_{t+1}}$, and again let $a = \frac{1}{\psi_{t+1}}$, we have $h_{t+1}^U(u) = a \exp(-au)$. Substituting this expression into the marginal distribution $h_{t+1}^U(u)$, we get

$$h_{t+1}^U(u) = \exp(\eta^* - \lambda^* \tilde{\mu}_{t+1}) a \exp(-au)[(1 - \rho_{t+1}) \exp(-\lambda^* u) + \rho_{t+1} \exp(\lambda^* u)],$$

and the cumulative density function (CDF) is then given by:

$$H_{t+1}^U(u) = \int h_{t+1}^U(u) du = \exp(\eta^* - \lambda^* \tilde{\mu}_{t+1}) a \frac{1 - \rho_{t+1}}{a + \lambda}(1 - \exp(-(a + \lambda)u)) + \frac{\rho_{t+1}}{a - \lambda}(1 - \exp(-(a - \lambda)u))$$

Once we get this analytical expression of the CDF of $U_{t+1}$, we can first generate random numbers from uniform distribution, and solve for $u$ from the inverse of its CDF function evaluated at the realizations of the uniform distribution. However, we can see from the above equation that $H_{t+1}^U(u)$ is a highly nonlinear function in $u$, and the only way to solve the inverse function of the CDF is through numerical methods, which could be very inefficient and inaccurate and thus affect the random draws of $u$, so we do not consider using PIT in the empirical part.
As an alternative method, we will use the Metropolis-Hasting Algorithm to generate $u$’s from $h_{t+1}^U(u)$ directly. This algorithm has been studied and used by statisticians. For a more detailed discussion, one can see Chib and Greenberg (1995) and Casella and Berger (1994, page 254-255). We explain the sketch of this algorithm as the follows:

Let $Y \sim f_Y(y)$ and $X \sim f_X(x)$, and $f_Y$ and $f_X$ have common support. If it is easy to generate $X$ from $f_X(x)$, then the following three step algorithm can generate $Y$ from $f_Y(y)$:

Step 1: Generate $X \sim f_X(x)$. Set $Z_0 = X$.

For $i = 1, 2, \ldots$:

Step 2: Generate $U_i \sim \text{Uniform}(0, 1)$, $X_i \sim f_X$, and calculate

$$q_i = \min \left\{ \frac{f_Y(Y_i)}{f_X(X_i)} \cdot \frac{f_X(Z_{i-1})}{f_Y(Z_{i-1})} \cdot 1 \right\}.$$  

Step 3: Set

$$Z_i = \begin{cases} 
X_i & \text{if } U_i \leq q_i \\
Z_{i-1} & \text{if } U_i > q_i 
\end{cases}.$$  

Then, as $i \to \infty$, $Z_i$ will converge to $Y$ in distribution.

Then in order to generate $u \sim h_{t+1}^U(u)$, since it’s easy to generate $u \sim f_{t+1}^U(u)$, we will set $u \sim f_{t+1}^U(u)$ as the $X$ variable above and $u \sim h_{t+1}^U(u)$ to be the $Y$ variable, and then $u \sim h_{t+1}^U(u)$ can be generated applying the Metropolis-Hasting Algorithm. Then we can get the forecast of the VaR according to the numerical method given above, and then evaluate and compare the VaR forecast from different density forecast models in terms of the predictive quantile loss and the empirical coverage probability.
To compare the productivity of VaR from different density forecasts, we use the "check function" of Koenker and Bassett (1978) and Bao, Lee and Saltoğlu (2006). The expected loss of quantile $l_t(\alpha)$ for a given level $\alpha$ has the form:

$$L(\alpha) = E[\alpha - 1(r_t < l_t(\alpha))][r_t - l_t(\alpha)].$$

(2.45)

where $l_t(\alpha)$ is the true VaR. Bertail et al. (2004) and Komunjer (2004) showed that the check function can be regarded as a predictive quasi-likelihood, therefore the expected loss $L(\alpha)$ can provide a measure of the lack-of-fit of a quantile model. Once we obtained the out-of-sample VaR forecasts $\hat{l}_t(\alpha)$’s, we can plug them into the above expression and evaluate the out-of-sample expected check function as:

$$\hat{L}(\alpha) = \frac{1}{F_\alpha} \sum_{t=R+1}^{T} [\alpha - 1(r_t < \hat{l}_t(\alpha))][r_t - \hat{l}_t(\alpha)].$$

(2.46)

Then a model which gives the VaR forecast $\hat{l}_t(\alpha)$’s with the minimum loss value of $L_F(\alpha)$ is considered as the best model.

As an alternative evaluation of risk forecast, when the CDF of $r_t$ is continuous in a neighborhood of $l_t(\alpha)$, $l_t(\alpha)$ minimizes $L(\alpha)$ and makes a condition for the correct conditional coverage probability

$$\alpha = E[1(r_t < l_t(\alpha))|I_{t-1}],$$

(2.47)

so $\alpha - 1(r_t < q_t(\alpha))$ is a martingale difference sequence and we can use this property to form a conditional moment test for the quantile models and also evaluate VaR forecasts. After replacing $l_t(\alpha)$ with $\hat{l}_t(\alpha)$, we can compute the empirical conditional coverage probability as:

$$\hat{\alpha}_F = \frac{1}{F} \sum_{t=R+1}^{T} 1(r_t < \hat{l}_t(\alpha)),$$

(2.48)

where the model which gives $\hat{\alpha}_F$ closest to the nominal value $\alpha$ is the preferred model.
2.6 Empirical Results

2.6.1 Data

In our empirical studies, we use the same data set as in Goyal and Welch (2008). In addition to their original monthly return, we calculate the annualized monthly stock return as in Campbell and Thompson (2008). We consider the density forecast of both stock return and equity premium. Since the difference between equity premium and stock return is the risk free rate which is relatively small and smooth compared to the equity premium, the equity premium should have similar distribution properties as those of the stock return discussed above. Therefore we can apply the decomposition model incorporating dependence as well as imposing constraints to equity premium as well.

Denote by \( P_t \) the S&P500 index at month \( t \). The monthly simple one-month return from month \( t \) to month \( t + 1 \) is defined as \( R_t(1) \equiv P_{t+1}/P_t - 1 \), and one-month excess return is denoted as \( Q_t(1) \equiv R_t(1) - r_f^t \) with \( r_f^t \) being the risk-free interest rate. Following Campbell, Lo and MacKinlay (1997, page 10), we define the \( k \)-period aggregated return from month \( t \) to month \( t + k \) as

\[
R_t(k) = \left( \frac{P_{t+k}}{P_t} - 1 \right) = \left( \frac{P_{t+k}}{P_{t+k-1}} \right) \times \cdots \times \left( \frac{P_{t+1}}{P_t} \right) - 1
\]

\[
= (1 + R_{t+k-1}(1)) \times \cdots \times (1 + R_t(1)) - 1
\]

\[
= \left[ \prod_{j=1}^{k} (R_{t+k-j}(1) + 1) \right] - 1. \quad (2.49)
\]
and following Campbell and Thompson (2008) we define the $k$-period aggregated excess return as

$$Q_t(k) \equiv (1 + R_{t+k-1}(1) - r_{t+k-1}^f) \times \cdots \times (1 + R_t(1) - r_t^f) - 1$$

$$= (Q_{t+k-1}(1) + 1) \times \cdots \times (Q_t(1) + 1) - 1$$

$$= \left[ \prod_{j=1}^{k} (Q_{t+k-j}(1) + 1) \right] - 1.$$

We allow $r_{t+1} = R_t(k)$ or $Q_t(k)$ when making the density and risk forecast and consider $k = 1, 3, 12$ as denoted in Campbell and Thompson (2008) which stands for the monthly, quarterly aggregated and annually aggregate stock return or equity premium.

We consider the data from May 1937 to December 2002, since although we did not report the result of using the covariates, we did used them and compared with the results in Lee and Xi (2013) which considers the same sample period. We divide the whole sample equally into $R$ in-sample observations and $P$ pseudo out-of-sample observations, where $R = P$. The models are estimated using rolling windows of the fixed size $R$. That is, at each time $t$ we use the data starting from $t - R + 1$ and ending at time $t$ to estimate parameters of a model and then make one-period ahead forecast for the next period $t + 1$. For annualized or quarterly aggregated monthly data, to avoid using future information, we only use data up to month $t - 11$ or $t - 2$ for estimation.

### 2.6.2 Results

Table 1 shows the test statistic and its $p$ value for test of independency between $U_{t+1}$ and $V_{t+1}$. From the table we can see that the null hypothesis of independence is clearly rejected for both stock return and equity premium and for all frequency. The test statistics are huge numbers from standard normal distribution and the $p$-values are 0 even up to four decimals. Many papers have already cited that the stock returns exhibit
negative skewness, which is an evidence of dependence between the absolute return and the sign of the return. Our test results are formal confirmation of this dependence and are in accordance with the literature while rejects the stylized facts DP1.

Figure 1 plots the estimated copula parameter $\theta$ for Frank, Clayton and FGM copula over time, and we can see that from the figures all the three copula parameters are away from 0, which again indicates that the absolute return and the sign of the return are not independent, since when the parameter converges to zero, all the three copula converges to independent copula. Moreover, from there figures we can also tell whether there is positive or negative dependence between the absolute return and the sign of the return. For stock returns, the dependence keeps to be positive since $\theta$ remains positive for all time period. Yet for equity premium, it seems there is a sign change around the 1980s, which shows there may be a structural break for the dependence between the absolute return and the sign of the return.

Table 2 and 3 shows the average log score of different density forecast models for the stock return and equity premium. By examining the numbers in these two tables, we can evaluate and compare density forecasts from M1, M2 and M3 to see whether decomposition model improves from normal distribution models, whether releasing the assumption of the independency between sign and absolute value helps as well as whether imposing moment conditions improves even more for the density forecast. First, compare the log score for M1 and M2 independent copula, we can observe that the decomposition model with independent copula improve a lot upon the benchmark normal distribution model, especially for annually and quarterly aggregated data, for both the stock return and equity premium. Moreover, using the other three copula functions which incorporate dependence between the sign and absolute value further
improves substantially upon the independent copula, for both stock return and equity
premium and all the aggregation frequencies. For example, the average log score jump
from M1 of −0.0399 to M2 of 0.1763 for independent copula, and then to M2 of 0.6434
for the Frank copula model.

In addition, we would like to see the effect of imposing moment condition on
the density forecast. Compare M3-ZM and M3-HM with independent copula of M2,
we can see that for stock return, M3 will improve upon M2 the constraint imposed
is “proper”, that is, if the data actually satisfies the moment condition. For stock
return, M3 always improves M2, yet for equity premium, only for quarterly aggregated
data, M3 can improve upon M2. When considering the dependence between sign and
absolute value, the results for M3 with the other three copula functions show that
adding moment constraints in general improves the density forecast since we chose
the “proper” constraint, and average log scores all goes up for one or both moment
constraints except for 3 months stock return. The ZM constraint is preferred for equity
premium for monthly and quarterly aggregated data and while the HM constraint is
better for equity premium and stock return for monthly and annually aggregated data.

To see the intuition of imposing the “proper” constraint, since the annually aggregated
return is aggregated and accumulated over 12 months, the aggregate return will deviate
more from zero for annualized month returns, and thus for 12 month aggregated returns,
historical mean should be a better constraint than zero mean. However, for monthly or
quarterly aggregated return, the mean does not aggregate too much from zero so the
ZM may be a better choice. Also similar reason for equity premium, which is equal to
the stock return minus the risk free rate, its mean should be closer to 0. That is why
the ZM constraint works better for risk premium for shorter return period while the
HM mean constraint is preferred for longer return period.

The magnitude of the improvement we can have from M2 to M3 depends on how far away the mean forecast from M2 deviates from the mean constraint we impose in M3. To see this point as well as the intuition to impose the “proper” constraint, we can look at Figure 2, which plots the mean forecasts from different decomposition models incorporating dependence against the historical mean forecast from real data. Figure 2-d shows that for annually aggregated stock return the decomposition models make mean forecasts which deviate from historical mean, but in figure 2-a we can see for annually aggregated equity premium the mean forecasts of decomposition models are close to historical mean. That is why in Table 3 the average log scores increase more for annually aggregated stock return than those for annually aggregated equity premium when we impose the historical mean constraint.

Now let us turn to out-of-sample risk forecast using VaR. Tables 4 through 9 report the predictive quantile loss and the empirical coverage probability for different aggregation frequencies of stock return and equity premium. In order to see more decimal numbers, we have multiplied 100 to both the predictive quantile loss as well as the empirical coverage probability value, so that all the numbers in these tables are in percentage. From these tables we can see the decomposition model of density forecast produces better VaR forecasts than Riskmetrics and normal distribution model for 1% quantile since we can always find at least one, and most cases more than one decomposition models among M2 and M3 using different copula functions which will produce a smaller loss. However, if we choose the 5% and 10% quantile VaR to forecast, the Riskmetrics model is hard to beat in terms of expected loss, but it may not always give a more accurate empirical coverage probability. The intuition behind why decomposition
model works for 1% quantile is due to the marginal density assumption for the absolute return $U_{t+1}$. Since we assume an exponential distribution for $U_{t+1}$ which has heavier tails than the normal distribution, it will make a more extreme and thus conservative VaR forecast, and thus is consistent with the truth that stock returns does have large drops in cases of financial crisis.

This argument is further proved by Figure 3. In Figure 3, we plot the VaR forecasts over time for 1%, 5% and 10% confidence level from four models: RiskMetrics, M1-ZM, M3-ZM-I, M3-ZM-C, since we want to compare the VaR results predicted from our decomposition density forecast model with the benchmark model of RiskMetrics and normal distribution model, and for both independent and dependent copula, the results from M2 and M3 does not vary much in terms of figure, so we just select the two from M3-ZM as an representative, and other decomposition models will follow similarity. From these figures we can see decomposition models do generate more conservative VaR forecast: their VaR forecasts are larger in absolute value than those from M1 and the RiskMetrics model for all confidence level and on the other hand, the RiskMetrics model produces a small absolute value of VaR forecast. Therefore, in terms of 1% confidence level VaR, where it captures the ability to cover loss from extreme negative return events from possible crisis, the decomposition model always works better than RiskMetrics since its conservative VaR forecast can efficiently account for those extreme events which is likely to happen in the financial merket. But when the confidence level increases and the level of loss becomes less extreme, thus a conservative VaR forecast may not be appropriate since it costs more fund on reserve for such less extreme events and that is why the decomposition model is not as good as RiskMetrics at 5% and 10% confidence level.
From these tables we can also see that decomposition models improve more for higher frequency aggregation of return. For 1 and 3 months aggregated returns, decomposition models can always beat the benchmark model in terms of both predictive quantile loss and the empirical coverage probability. But for 12 months aggregated return, the difference is not that large. This is because when we aggregate the return, extreme negative returns become less rare since it is averaged out by other normal or positive returns and there will be no significant loss if one look at the overall return. Thus the decomposition model lose its advantage of making a larger and more conservative VaR forecast. However, since the most common use of VaR is to give instructions for the reserve fund for possible extreme events like financial crisis on the daily or monthly return basis, our decomposition model is more useful compared to RiskMetrics in terms of its higher coverage probability and lower predictive quantile loss in the event of large loss happens.

2.7 Conclusion

This paper explores three questions: first, we consider whether decomposition model which specifies the sign and absolute value of the return for density forecast improves upon traditional normal distribution model, and in addition, whether releasing the assumption that the sign and absolute return are independent through incorporating copula functions to their join density improves upon the decomposition model imposing independent assumption. Second, we examine whether imposing moment conditions to decomposition model for density forecast can further improves the forecasts. And third, we apply the density forecast results to produce VaR forecasts and compare the results to traditional methods such as RiskMetrics and normal model to see if a better density
forecast will lead to a better risk forecast as well.

Our conclusion is as follows. On one hand, we find that decomposition model which incorporate the stylizes facts is indeed better than normal model which implies none of these facts. Moreover, test results show that the sign and absolute return are indeed dependent with each other, so that decomposition model which incorporates dependency through copula functions works better than imposing independency. On the other hand, we find that imposing moment constraint to decomposition density forecast models will further improve density forecasts as long as such constraint is proper. Furthermore, the decomposition density forecast model will produce better risk forecasts in terms of VaR forecasts compared to methods such as RiskMetrics at lower significance level such as 1%. And since the most common use of VaR is to give instructions for the reserve fund for possible extreme events like financial crisis on the daily or monthly return basis, our decomposition model is more useful compared to RiskMetrics in terms of its higher coverage probability and lower predictive quantile loss in the event of large loss happens.
References


Hong, Y., and H. Li (2005), “Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates”. *Review of


78

Table 2.1: Test for Independence

<table>
<thead>
<tr>
<th></th>
<th>$Q_t(12)$</th>
<th>144.7009</th>
<th>(0.0000)</th>
<th>$R_t(12)$</th>
<th>210.8172</th>
<th>(0.0000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_t(3)$</td>
<td>158.6323</td>
<td>(0.0000)</td>
<td>$R_t(3)$</td>
<td>213.3336</td>
<td>(0.0000)</td>
</tr>
<tr>
<td></td>
<td>$Q_t(1)$</td>
<td>168.6566</td>
<td>(0.0000)</td>
<td>$R_t(1)$</td>
<td>168.9395</td>
<td>(0.0000)</td>
</tr>
</tbody>
</table>

Notes: Reported are the Hong and Li (2005) test statistics for the null hypothesis of independence of $|r_t|$ and $\text{sign}(r_t)$. The asymptotic $p$-values are in brackets.

Table 2.2: Average Out-of-Sample log Scores of Normal Distribution Model

<table>
<thead>
<tr>
<th></th>
<th>M1-ZM</th>
<th>M1-HM</th>
<th>M1-ZM</th>
<th>M1-HM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_t(12)$</td>
<td>-0.0399</td>
<td>-0.1333</td>
<td>$R_t(12)$</td>
</tr>
<tr>
<td></td>
<td>$Q_t(3)$</td>
<td>0.8976</td>
<td>0.8507</td>
<td>$R_t(3)$</td>
</tr>
<tr>
<td></td>
<td>$Q_t(1)$</td>
<td>1.6613</td>
<td>1.6560</td>
<td>$R_t(1)$</td>
</tr>
</tbody>
</table>
Table 2.3: Average Out-of-Sample log Scores of Decomposition Models

<table>
<thead>
<tr>
<th></th>
<th>M2</th>
<th>M3-ZM</th>
<th>M3-HM</th>
<th>M2</th>
<th>M3-ZM</th>
<th>M3-HM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_t(12)$</td>
<td></td>
<td></td>
<td>$R_t(12)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>0.1763</td>
<td>0.1582</td>
<td>0.0620</td>
<td>0.0267</td>
<td>0.1661</td>
<td>0.2679</td>
</tr>
<tr>
<td>Frank</td>
<td>0.6434</td>
<td>0.6479</td>
<td>0.6580</td>
<td>0.5209</td>
<td>0.4918</td>
<td>0.5998</td>
</tr>
<tr>
<td>Clayton</td>
<td>0.6493</td>
<td>0.6535</td>
<td>0.6635</td>
<td>0.5200</td>
<td>0.4880</td>
<td>0.5975</td>
</tr>
<tr>
<td>FGM</td>
<td>0.6436</td>
<td>0.6481</td>
<td>0.6583</td>
<td>0.5232</td>
<td>0.4949</td>
<td>0.6019</td>
</tr>
<tr>
<td></td>
<td>$Q_t(3)$</td>
<td></td>
<td></td>
<td>$R_t(3)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>1.0764</td>
<td>1.3573</td>
<td>1.3516</td>
<td>1.0757</td>
<td>1.3258</td>
<td>1.3471</td>
</tr>
<tr>
<td>Frank</td>
<td>1.3557</td>
<td>1.3591</td>
<td>1.3513</td>
<td>1.3515</td>
<td>1.3322</td>
<td>1.3409</td>
</tr>
<tr>
<td>Clayton</td>
<td>1.3543</td>
<td>1.3584</td>
<td>1.3504</td>
<td>1.3529</td>
<td>1.3351</td>
<td>1.3464</td>
</tr>
<tr>
<td>FGM</td>
<td>1.3558</td>
<td>1.3591</td>
<td>1.3513</td>
<td>1.3514</td>
<td>1.3321</td>
<td>1.3409</td>
</tr>
<tr>
<td></td>
<td>$Q_t(1)$</td>
<td></td>
<td></td>
<td>$R_t(1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>1.6611</td>
<td>1.6609</td>
<td>1.6566</td>
<td>1.6610</td>
<td>1.6791</td>
<td>1.6844</td>
</tr>
<tr>
<td>Frank</td>
<td>1.7189</td>
<td>1.7203</td>
<td>1.7155</td>
<td>1.7093</td>
<td>1.7031</td>
<td>1.7095</td>
</tr>
<tr>
<td>Clayton</td>
<td>1.7178</td>
<td>1.7196</td>
<td>1.7148</td>
<td>1.7061</td>
<td>1.6973</td>
<td>1.7037</td>
</tr>
<tr>
<td>FGM</td>
<td>1.7189</td>
<td>1.7203</td>
<td>1.7155</td>
<td>1.7093</td>
<td>1.7031</td>
<td>1.7094</td>
</tr>
</tbody>
</table>
Table 2.4: Out-of-sample Value-at-Risk Forecast Comparison: $Q_t(12)$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha=1%$</th>
<th>$\alpha=5%$</th>
<th>$\alpha=10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$</td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>1.6993</td>
<td>0.7614</td>
<td>1.4529</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.9800</td>
<td>7.1066</td>
<td>2.2822</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.8533</td>
<td>4.8223</td>
<td>2.3689</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.8600</td>
<td>4.8223</td>
<td>2.3842</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.8538</td>
<td>4.8223</td>
<td>2.3828</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.8677</td>
<td>4.8223</td>
<td>2.3915</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.8341</td>
<td>4.5685</td>
<td>2.3181</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.7738</td>
<td>3.2995</td>
<td>2.3209</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.7737</td>
<td>3.2995</td>
<td>2.3132</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.7741</td>
<td>3.2995</td>
<td>2.3228</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.9191</td>
<td>5.3299</td>
<td>2.4344</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>1.0845</td>
<td>5.3299</td>
<td>2.6055</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>1.0920</td>
<td>5.3299</td>
<td>2.6050</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>1.0802</td>
<td>5.3299</td>
<td>2.6037</td>
</tr>
</tbody>
</table>
Table 2.5: Out-of-sample Value-at-Risk Forecast Comparison: $Q_t(3)$

<table>
<thead>
<tr>
<th></th>
<th>$\alpha=1%$</th>
<th>$\alpha=5%$</th>
<th>$\alpha=10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$</td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>0.9429</td>
<td>1.5228</td>
<td>0.9083</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.4382</td>
<td>4.0609</td>
<td>1.0980</td>
</tr>
<tr>
<td>M1-HM</td>
<td>0.4902</td>
<td>5.3299</td>
<td>1.1799</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.2687</td>
<td>1.2690</td>
<td>1.0651</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.2692</td>
<td>1.5228</td>
<td>1.0787</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.2687</td>
<td>1.5228</td>
<td>1.0763</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.2693</td>
<td>1.5228</td>
<td>1.0785</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.2798</td>
<td>1.2690</td>
<td>1.0480</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.2716</td>
<td>0.5076</td>
<td>1.0123</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.2709</td>
<td>0.5076</td>
<td>1.0138</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.2716</td>
<td>0.5076</td>
<td>1.0122</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.2780</td>
<td>1.5228</td>
<td>1.0655</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>0.2754</td>
<td>1.2690</td>
<td>1.0420</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>0.2761</td>
<td>1.2690</td>
<td>1.0470</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>0.2754</td>
<td>1.2690</td>
<td>1.0421</td>
</tr>
</tbody>
</table>
### Table 2.6: Out-of-sample Value-at-Risk Forecast Comparison: $Q_t(1)$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha=1%$</th>
<th>$\alpha=5%$</th>
<th>$\alpha=10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$</td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>0.5946</td>
<td>1.7766</td>
<td>0.5625</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.1980</td>
<td>2.7919</td>
<td>0.5667</td>
</tr>
<tr>
<td>M1-HM</td>
<td>0.2038</td>
<td>2.7919</td>
<td>0.5730</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.1816</td>
<td>1.2690</td>
<td>0.5641</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.1820</td>
<td>1.2690</td>
<td>0.5657</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.1818</td>
<td>1.2690</td>
<td>0.5645</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.1820</td>
<td>1.2690</td>
<td>0.5657</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.1805</td>
<td>1.2690</td>
<td>0.5689</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.1738</td>
<td>0.7614</td>
<td>0.5609</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.1745</td>
<td>0.7614</td>
<td>0.5616</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.1738</td>
<td>0.7614</td>
<td>0.5609</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.1804</td>
<td>1.2690</td>
<td>0.5707</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>0.1735</td>
<td>0.7614</td>
<td>0.5665</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>0.1747</td>
<td>0.7614</td>
<td>0.5668</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>0.1735</td>
<td>0.7614</td>
<td>0.5664</td>
</tr>
<tr>
<td></td>
<td>$\alpha=1%$</td>
<td>$\alpha=5%$</td>
<td>$\alpha=10%$</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>$\hat{L}_P(\alpha)$, $\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$, $\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$, $\hat{\alpha}_P$</td>
<td></td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>1.9249, 0.7614</td>
<td>1.5610, 3.0457</td>
<td>2.7621, 11.1675</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.7569, 3.5533</td>
<td>2.2512, 5.5838</td>
<td>3.6133, 8.1218</td>
</tr>
<tr>
<td>M1-HM</td>
<td>1.0250, 6.5990</td>
<td>2.4313, 12.1827</td>
<td>3.6106, 17.7665</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.8224, 2.5381</td>
<td>2.3257, 5.8376</td>
<td>3.5448, 10.6599</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.7662, 2.7919</td>
<td>2.2326, 9.1371</td>
<td>3.4912, 14.2132</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.8034, 2.5381</td>
<td>2.2729, 7.6142</td>
<td>3.5058, 12.6904</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.7561, 2.7919</td>
<td>2.2223, 9.8985</td>
<td>3.4834, 14.4670</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.8097, 2.2843</td>
<td>2.3826, 4.5685</td>
<td>3.6385, 7.6142</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.8567, 1.0152</td>
<td>2.7615, 3.2995</td>
<td>4.2386, 4.8223</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.8522, 1.0152</td>
<td>2.6398, 3.2995</td>
<td>3.9758, 5.3299</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.8601, 1.0152</td>
<td>2.8176, 3.0457</td>
<td>4.3456, 4.5685</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.8738, 2.7919</td>
<td>2.4549, 6.5990</td>
<td>3.7134, 11.6751</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>1.1466, 4.3147</td>
<td>2.7558, 7.3604</td>
<td>3.9759, 11.4213</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>1.1298, 4.5685</td>
<td>2.6781, 8.6294</td>
<td>3.8988, 11.9289</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>1.1520, 4.3147</td>
<td>2.7868, 7.1066</td>
<td>4.0065, 10.9137</td>
</tr>
<tr>
<td></td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{\alpha}_P$</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>0.9783</td>
<td>1.5228</td>
<td>0.9194</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.4083</td>
<td>3.0457</td>
<td>1.0531</td>
</tr>
<tr>
<td>M1-HM</td>
<td>0.4783</td>
<td>4.5685</td>
<td>1.1458</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.2671</td>
<td>1.2690</td>
<td>1.0303</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.2753</td>
<td>1.7766</td>
<td>1.0621</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.2645</td>
<td>1.5228</td>
<td>1.0540</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.2763</td>
<td>1.7766</td>
<td>1.0639</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.2735</td>
<td>1.0152</td>
<td>1.0155</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.2899</td>
<td>0.2538</td>
<td>1.0030</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.2917</td>
<td>0.2538</td>
<td>1.0065</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.2878</td>
<td>0.2538</td>
<td>1.0038</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.2708</td>
<td>1.0152</td>
<td>1.0469</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>0.2797</td>
<td>1.2690</td>
<td>1.0614</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>0.2779</td>
<td>1.2690</td>
<td>1.0699</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>0.2799</td>
<td>1.2690</td>
<td>1.0612</td>
</tr>
</tbody>
</table>
Table 2.9: Out-of-sample Value-at-Risk Forecast Comparison: $R_t(1)$

<table>
<thead>
<tr>
<th></th>
<th>$\alpha$=1%</th>
<th></th>
<th>$\alpha$=5%</th>
<th></th>
<th>$\alpha$=10%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
<td>$\hat{L}_P(\alpha)$</td>
<td>$\hat{\alpha}_P$</td>
</tr>
<tr>
<td>RiskMetrics</td>
<td>0.5997</td>
<td>1.5228</td>
<td>0.5568</td>
<td>6.0914</td>
<td>0.8808</td>
<td>9.8985</td>
</tr>
<tr>
<td>M1-ZM</td>
<td>0.1858</td>
<td>2.5381</td>
<td>0.5605</td>
<td>5.0761</td>
<td>0.8872</td>
<td>8.6294</td>
</tr>
<tr>
<td>M1-HM</td>
<td>0.1973</td>
<td>3.0457</td>
<td>0.5649</td>
<td>6.3452</td>
<td>0.8913</td>
<td>11.4213</td>
</tr>
<tr>
<td>M2-I</td>
<td>0.1770</td>
<td>1.0152</td>
<td>0.5567</td>
<td>5.3299</td>
<td>0.8774</td>
<td>10.6599</td>
</tr>
<tr>
<td>M2-F</td>
<td>0.1764</td>
<td>1.2690</td>
<td>0.5602</td>
<td>5.8376</td>
<td>0.8815</td>
<td>12.4365</td>
</tr>
<tr>
<td>M2-C</td>
<td>0.1773</td>
<td>1.0152</td>
<td>0.5558</td>
<td>5.8376</td>
<td>0.8788</td>
<td>10.4061</td>
</tr>
<tr>
<td>M2-FGM</td>
<td>0.1764</td>
<td>1.2690</td>
<td>0.5598</td>
<td>5.8376</td>
<td>0.8810</td>
<td>12.4365</td>
</tr>
<tr>
<td>M3-ZM-I</td>
<td>0.1778</td>
<td>0.7614</td>
<td>0.5628</td>
<td>4.5685</td>
<td>0.8796</td>
<td>9.6447</td>
</tr>
<tr>
<td>M3-ZM-F</td>
<td>0.1705</td>
<td>0.5076</td>
<td>0.5585</td>
<td>3.2995</td>
<td>0.8725</td>
<td>7.6142</td>
</tr>
<tr>
<td>M3-ZM-C</td>
<td>0.1704</td>
<td>0.5076</td>
<td>0.5598</td>
<td>3.2995</td>
<td>0.8737</td>
<td>7.6142</td>
</tr>
<tr>
<td>M3-ZM-FGM</td>
<td>0.1705</td>
<td>0.5076</td>
<td>0.5586</td>
<td>3.2995</td>
<td>0.8731</td>
<td>7.3604</td>
</tr>
<tr>
<td>M3-HM-I</td>
<td>0.1765</td>
<td>0.7614</td>
<td>0.5605</td>
<td>5.3299</td>
<td>0.8816</td>
<td>11.1675</td>
</tr>
<tr>
<td>M3-HM-F</td>
<td>0.1758</td>
<td>1.0152</td>
<td>0.5557</td>
<td>4.8223</td>
<td>0.8680</td>
<td>9.1371</td>
</tr>
<tr>
<td>M3-HM-C</td>
<td>0.1775</td>
<td>1.0152</td>
<td>0.5600</td>
<td>5.3299</td>
<td>0.8693</td>
<td>9.6447</td>
</tr>
<tr>
<td>M3-HM-FGM</td>
<td>0.1758</td>
<td>1.0152</td>
<td>0.5561</td>
<td>4.8223</td>
<td>0.8687</td>
<td>9.1371</td>
</tr>
</tbody>
</table>
Figure 2.1 Estimated copula parameter $\theta$ over time

(a) $Q_t(12)$  
(b) $Q_t(3)$  
(c) $Q_t(1)$  
(d) $R_t(12)$  
(e) $R_t(3)$  
(f) $R_t(1)$

Note: $\theta$ for Frank copula is denoted using the blue solid line, and $\theta$ for Clayton copula is denoted using red dashed line and $\theta$ for FGM copula is denoted by black dashed line.
Figure 2.2 Mean forecasts comparison for copula models of Model 2 and historical mean

(a) $Q_t(12)$  
(b) $Q_t(3)$  
(c) $Q_t(1)$  
(d) $R_t(12)$  
(e) $R_t(3)$  
(f) $R_t(1)$

Note: Mean forecast from Frank copula is denoted using the blue solid line, Mean forecast from Clayton copula is denoted using red dashed line and Mean forecast from is denoted by black dashed line. Green line is the historical mean of real data.
Figure 2.3. Risk forecast comparison

(a) $\alpha = 0.01$

(b) $\alpha = 0.05$

(c) $\alpha = 0.1$

Note: Black dashed line denotes the VaR forecast by RiskMetrics, red dashed line
denotes VaR forecasts by M1-ZM, green line denotes VaR forecasts by independent copula from M3-ZM density forecast, and blue line denotes VaR forecast using M3-ZM and Clayton copula density forecast.
Chapter 3

Testing for Neglected Nonlinearity Using Artificial Neural Networks with Many Randomized Hidden Unit Activations

3.1 Introduction

This paper revisits the test of White (1989) and Lee, White and Granger (L-WG, 1993), for neglected nonlinearity in conditional mean using the feedforward single layer artificial neural network (ANN). The advantage to use ANN model to test nonlinearity is that the ANN model inherits the flexibility as a universal approximator of
unknown functional form. The ANN test is designed to use the predictive ability of
the ANN hidden layer activations, which may be neglected in linear models. Because
the estimation of the ANN model is often difficult and the activation parameters in the
hidden layer are not identified under the null hypothesis of linearity, LWG suggested to
activate the ANN hidden units based on the randomly generated neural network activa-
tion parameters. LWG considered only a small set of random activation parameters
(limited by the computing power two decades ago). Nevertheless, their Monte Carlo
experiment demonstrated the excellent performance of the ANN test in size and pow-
er. The ANN test has been cited in numerous papers as a benchmark method in the
literature on testing neglected nonlinearity.

However, in this paper, we note that the size and power of any Monte Carlo
experiments are the empirical average frequencies of rejecting the null hypothesis, when
the null hypothesis is true (size) or when the null is not true (power), over many Monte
Carlo replications of the data generating process (DGP). Unlike in a Monte Carlo study
where the data are replicated multiple times, an empirical study has only one realized
sample. When the ANN test is applied to one realized sample, its performance is
largely affected by the randomly generated activation parameters. Applying the test to
a particular real data amounts to one single Monte Carlo replication. In this paper we
show that a small set of random activation parameters will make the performance of
the ANN test quite random. This was not noticed in LWG (1993) and any other papers
that have studied the ANN test, perhaps because most of these studies compare the
performance in Monte Carlo where the performance is measured in average rejection
over many replications. We will show that, when a real data is tested by the ANN
test, a small number of random activations makes the ANN test quite unstable and
sensitive to the random activations. Interestingly, however, we will also show that increasing the number of the randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a single real data set. This robustification is important and useful in practice, which can be achieved at no cost as increasing the number of random activations is almost costless given the computer technology available today.

The rest of the paper is organized as follows. Section 2 reviews the ANN test with randomized hidden unit activations. In Section 3, we examine the ANN test with Monte Carlo to confirm the LWG’s results on the excellent size and power of the randomly activated ANN test. In Section 4, for each simulated series, we point out a problem of the randomized ANN test when the number of randomized activations is small and then show that this problem can be easily resolved by simply increasing it to a very large number of randomized activations. In Section 5 we repeat what we have done in Section 4 using actual economic data. Section 6 concludes.

3.2 The ANN Test

The linear-augmented single hidden-layer feedforward ANN model has the following architecture:

\[ y_t = f(x_t, \theta) + \varepsilon_t := x_t'\alpha + \sum_{j=1}^{q} \beta_j \psi(x_t'\gamma_j) + \varepsilon_t, \]

where \( t = 1, \ldots, n, \) \( x_t = (x_{1t}, \ldots, x_{kt})' \), \( \theta = (\alpha', \beta', \gamma_1', \ldots, \gamma_q')' \), \( \alpha = (\alpha_1, \ldots, \alpha_k)' \), \( \beta = (\beta_1, \ldots, \beta_q)' \), and \( \gamma_j = (\gamma_{j1}, \ldots, \gamma_{jk})' \) for \( j = 1, \ldots, q \). and \( \psi(\cdot) \) is an activation function.\(^1\)

An example of the activation function is the logistic function \( \psi(z) = (1 + \exp(z))^{-1} \). \( \alpha \) is a column vector of connection strength from the input layer to the output layer; \( \gamma_j \) is

\(^1\)\( 'a := b' \) means that \( a \) is defined by \( b \), while \( 'a =: b' \) means that \( b \) is defined by \( a \).
a conformable column vector of connection strength from the input layer to the hidden units, \( j = 1, \ldots, q \); \( \beta_j \) is a (scalar) connection strength from the hidden unit \( j \) to the output unit, \( j = 1, \ldots, q \); and \( \psi \) is a squashing function (e.g., the logistic squasher) or a radial basis function. Input units \( \mathbf{x} \) send signals to intermediate hidden units, then each of hidden unit produces an activation \( \psi \) that then sends signals toward the output unit. The integer \( q \) denotes the number of hidden units added to the affine (linear) network.

Hornik, Stinchcombe and White (1989, 1990) show that neural network model in (3.1) is a nonlinear flexible functional form being capable of representing arbitrarily accurate approximations to any mappings. White (1990) and White and Wooldridge (1991) show that these approximations are learnable (i.e., consistently estimable) by proper control of the growth of network complexity \( q \) as network experience accumulates (i.e., the sample size \( n \) grows). While they give theoretical results for controlling the growth rate of \( q \) as a function of \( n \), the proper rate depends critically on the dependence properties of \( (y_t, \mathbf{x}_t')' \) which makes a choice of the growth rate for network complexity not immediately obvious. As a referee pointed out it would be desirable if we could give some guidance on the adaptive choice of \( q \). Unfortunately, to date there is no unified theory on this rate. See Chen (2007, p. 5575) for more discussion. While White (1990, p. 538) gave some guidelines on the choice of \( q \) for different \( n \) and different dependence properties, he recommended to use cross-validation to choose \( q \) in estimating ANN models in practice. This paper, however, deals with testing for neglected nonlinearity in a linear model without having to estimate the nonlinear ANN model. The main purpose of this paper is about the choice of \( q \) in using the ANN model for testing whether \( \beta_j \)'s are all zero. As we do this with the randomization of \( \gamma_j \)'s and then we take a small number of their principal components, a very large \( q \), even larger than \( n \),
may be used and may be more desirable as examined in Sections 4 and 5.

To test whether the process \( y_t \) is linear in mean conditional on \( x_t \), we consider the following null and alternative hypotheses

\[
H_0 : \Pr[E(y_t | x_t) = x_t' \alpha^*] = 1 \quad \text{for some } \alpha^* \in \mathbb{R}^k
\]

\[
H_1 : \Pr[E(y_t | x_t) = x_t' \alpha] < 1 \quad \text{for all } \alpha \in \mathbb{R}^k
\]

When the null hypothesis is rejected, a linear model is said to suffer from neglected nonlinearity. White (1989) and LWG (1993) developed a test for neglected nonlinearity likely to have power against a range of alternatives based on ANN models. See also Teräsvirta et al (1993) and Teräsvirta (1996) on the neural network test and its comparison with other specification tests. The neural network test is based on the activations of ‘phantom’ hidden units \( \psi(x_t' \gamma_j), j = 1, \ldots, q \). That is,

\[
H_0 : E[\psi(x_t' \gamma_j) \varepsilon_t] = 0, \quad j = 1, \ldots, q,
\] (3.2)

or

\[
E(\Psi_t \varepsilon_t) = 0,
\] (3.3)

where \( \Psi_t := (\psi(x_t' \gamma_1), \ldots, \psi(x_t' \gamma_q))' \) is a phantom hidden unit activation vector and \( \varepsilon_t \) is the error term from the two layer affine network \( y_t = x_t' \alpha + \varepsilon_t \) (with \( q = 0 \)). Evidence of correlation of \( \varepsilon_t \) with \( \Psi_t \) is evidence against the null hypothesis that \( y_t \) is linear in conditional mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations \( \Psi_t \) would permit an improvement in network performance. Thus the tests are based on the sample correlation of affine network errors with phantom hidden unit activations,

\[
n^{-1} \sum_{t=1}^{n} \Psi_t \hat{\varepsilon}_t = n^{-1} \sum_{t=1}^{n} \Psi_t(y_t - x_t' \hat{\alpha}),
\] (3.4)
where \( \hat{\alpha} \) is least squares estimator of \( \alpha \). Under suitable regularity conditions it follows from a central limit theorem that 
\[
\left( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\epsilon}_t \right) \overset{d}{\to} N(0, W) \quad \text{as} \quad n \to \infty,
\]
and if one has a consistent estimator for its asymptotic covariance matrix, say \( \hat{W}_n \), then an asymptotic chi-squared statistic can be formed as
\[
\left( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\epsilon}_t \right)' \hat{W}_n^{-1} \left( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\epsilon}_t \right) \overset{d}{\to} \chi^2_q.
\] (3.5)

It is well known that the ANN models are generally hard to estimate and suffer from possibly large estimation errors which can adversely affect their ability as a general approximator. To alleviate the estimation errors of the ANN, it is useful to note that, for given values of \( \gamma_j \)'s, the ANN is linear in \( x \) and the activation functions \( \Psi \) and therefore \( (\alpha', \beta') \) can be estimated from the linear regression once \( (\gamma_1, \ldots, \gamma_q) \) have been given. The LWG’s (1993) approach is to use a set of randomly generated \( (\gamma_1, \ldots, \gamma_q) \). The additional hidden unit activation functions \( \Psi_t(\gamma_1, \ldots, \gamma_q) \) are hidden (or phantom) because they do not exist under the null hypothesis. The \( (\gamma_1, \ldots, \gamma_q) \) are randomly generated in testing because they are nuisance parameters not identified under the null hypothesis.

This approach is shown to have excellent size and power properties from Monte Carlo simulation and has been used in many subsequent nonlinear testing papers as a benchmark method in comparison. However, it is not noted in the literature that the LWG’s excellent performance even with a small number \( (q = 10, 20) \) of the randomized phantom activations is in terms of the Monte Carlo size and power. The good size and power in Monte Carlo experiments are the average frequencies of rejecting the null hypothesis over multiple replications of the data generating process (DGP). The averaging in Monte Carlo smooths out the randomness of the test result in each replication. However, in an empirical application, unlike in a Monte Carlo study, multiple realizations
of the data are not possible or available. In this case, the ANN test is sensitive to the randomly generated activation parameters and its performance is generally unstable. When applying to real data, this randomness problem resulted from using different sets of randomized activation parameters $(\gamma_1, \ldots, \gamma_q)$ may lead to inconsistent conclusions.

One solution is the use of Bonferroni bounds of the p-values of the test statistics that are computed from $m$ randomizations of the activation parameters $\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^m$, as suggested in LWG (1993). However, the Bonferroni bounds still exhibit dependence on the randomized activations when $q$ is small (as shown later in Table 3 of Section 5).

Another solution is to integrate the test statistic over the nuisance parameter space of $(\gamma_1, \ldots, \gamma_q)$. However, this approach requires bootstrap or simulation to obtain the null distribution of the integrated statistic (more on this in Section 3.4).

In this paper, we show a much simpler solution. That is to increase the number of randomized hidden unit activations to a (very) large number (e.g., 1000). We show that ‘many’ randomly generated activation parameters can robustify the performance of the ANN test when it is applied to a real empirical data. It also makes the Bonferroni bounds tighter (as shown in Section 3.5). We will demonstrate this in the remaining sections of the paper in Monte Carlo and in empirical applications. While this proposal may sound trivial, no previous papers have noted this problem. It is partly because all studies were able to show the excellent performance via Monte Carlo simulations with a small $q$ and also because it was difficult to compute the singular value decomposition of a $q \times q$ matrix for a large $q$ (to compute the principal components). It was 1989 when LWG (1993) conducted their Monte Carlo on an IBM 286 PC. The set of randomly selected parameters $(\gamma_1, \ldots, \gamma_q)$ should be large enough so that it can be dense and make the ANN an universal approximator. A large set of $\gamma$’s (i.e., large $q$) enables $\sum_{j=1}^q \beta_j \psi (x' \gamma_j)$
to capture the maximal nonlinear structure. We will show that the proposal of increasing $q$ in fact provides a practically useful, powerful, and cheap solution to the randomness of random activations. The robustification is stable and reliable, and thus enables the ANN test to be employed in autopilot in its applications.

A large number $q$ of random activation parameters $(\gamma_1, \ldots, \gamma_q)$ will make the activation functions $\psi(x_t' \gamma_j)$ collinear with each other over $j = 1, \ldots, q$ and with $x_t$. Thus LWG (1993) conducted a test on $q^* < q$ principal components of $\Psi_t$ not collinear with $x_t$, denoted $\Psi_t^*$. The key to the success with the large number of randomized network activations is the regularization of the network performance by principal components for dimensionality reduction. The ANN test takes two steps, randomization and regularization.

Then LWG employed the asymptotically equivalent test statistic (under conditional homoskedasticity) which avoids explicit computation of $\hat{W}_n$

$$T_n (q, q^* \mid \gamma_1, \ldots, \gamma_q) := nR^2 \xrightarrow{d} \chi^2_{q^*} ,$$

(3.6)

where $R^2$ is uncentered squared multiple correlation from a standard linear regression of $\hat{\varepsilon}_t$ on $\Psi_t^*$ and $x_t$. This test is to determine whether or not there exists some advantage to be gained by adding hidden units to the affine network. In this paper, while we consider two values of $q$ (small and very large), we fix $q^* = 3$ to simplify our presentation. Different values of $q^*$ do not affect the conclusions of this paper. Therefore the test statistic will be henceforth denoted as $T_n (q, 3 \mid \gamma_1, \ldots, \gamma_q) =: T_n (q \mid \gamma_1, \ldots, \gamma_q)$ or simply $T_n (q)$.

In Section 3.3, we conduct a Monte Carlo to show the ANN test has good size and power even with a small $q = 20$. The size and power from Monte Carlo do not tell the problem discussed above from using a small $q$. To see the problem,
we conduct a different Monte Carlo, in Section 3.4. Only one realization (to mimic an empirical study) of \( \{y_t\}_{t=1}^{n=200} \) which is linear in mean is generated, for which the ANN statistic \( T_n(q | \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}) \) and its p-value \( P_i \) are computed from \( m \) different randomly generated activation parameters \( \left( \gamma_1^{(i)}, \ldots, \gamma_q^{(i)} \right)_{i=1}^m \). We show that the ANN statistic with a small number \((q = 20)\) of randomized phantom activations exhibits large variation over \( i = 1, \ldots, m \), while it becomes stable with a very large number \((q = 1000)\) of randomized phantom activations. Hence, we can improve and robustify the ANN test by simply increasing \( q \) (say, from 20 to 1000). Section 3.5 demonstrates this with the five US monthly economics time series. In practice, we suggest to choose \( q \) as large as possible provided the computational ability permits. This is because a larger \( q \) will stabilize the p-values. Since we take the principle components of the activation functions, we can allow \( q \) to be even larger than the number of observations \( n \). In our simulations and empirical experiments, for a moderately large data (with \( n \) around 200), choosing \( q \) to be 1000 leads to good results.

### 3.3 Small \( q \) vs. Large \( q \) in Monte Carlo Size and Power

The purpose of this section is to confirm the result of LWG (1993) that Monte Carlo studies will show excellent performance of the ANN test in terms of size and power, computed from 1000 Monte Carlo replications. To generate data we use the following DGPs, all of which have been used in the related literature. Two blocks of DGP are considered in this section: the first block has DGPs using the univariate autoregressive time series of \( y_t \) with one lagged endogenous input \( y_{t-1} \); the second block includes cross-sectional networks with two exogenous inputs \( x_{1t} \) and \( x_{2t} \) which follow a bivariate normal distribution. To see the sensitivity of the test statistic under
conditional heteroskedasticity, we also consider ARCH(1) and GARCH(1,1) processes for AR in Block 1. All DGPs below fulfil the conditions for the investigated testing procedures. For those regularity conditions and moment conditions, see White (1994, Chapter 9) for the ANN tests. All the error terms $\varepsilon_t$ below are i.i.d. $N(0, 4)$. $1(\cdot)$ is an indicator function which takes one if its argument is true and zero otherwise. The index $t = 1, \ldots, n$ with $n = 200$ being the sample size.

**Block 1 (Time-series data generating processes)**

1. Autoregressive (AR)

   $$y_t = 0.6y_{t-1} + \varepsilon_t$$

2. Threshold autoregressive (TAR)

   $$y_t = \begin{cases} 
   0.9y_{t-1} + \varepsilon_t & \text{if } |y_{t-1}| \leq 1 \\
   -0.3y_{t-1} + \varepsilon_t & \text{otherwise}
   \end{cases}$$

3. Sign autoregressive (SGN)

   $$y_t = \text{sgn}(y_{t-1}) + \varepsilon_t$$

   where $\text{sgn}(y_{t-1}) = 1(y_{t-1} > 0) - 1(y_{t-1} < 0)$.

4. Nonlinear autoregressive (NAR)

   $$y_t = \frac{0.7|y_{t-1}|}{|y_{t-1}| + 2} + \varepsilon_t$$

5. Markov regime-switching (MRS)

   $$y_t = \begin{cases} 
   0.6y_{t-1} + \varepsilon_t & \text{if } S_t = 0 \\
   -0.5y_{t-1} + \varepsilon_t & \text{if } S_t = 1
   \end{cases}$$
where $S_t$ follows a two-state Markov chain with transition probabilities $\Pr(S_t = 1 | S_{t-1} = 0) = \Pr(S_t = 0 | S_{t-1} = 1) = 0.3$.

**Block 2 (Cross-sectional data generating processes):**

Assume $x_{1t}, x_{2t}$ follow a bivariate normal distribution of $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ with $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$, and $\rho = 0$ or $0.7$. We have the following three cases:

1. **Linear**

   $$y_t = 1 + x_{1t} + x_{2t} + \varepsilon_t$$

2. **Interaction**

   $$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{1t}x_{2t} + \varepsilon_t$$

3. **Squared**

   $$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{2t}^2 + \varepsilon_t$$

In the simulations of the ANN test, LWG chose $q$ equal to 10 or 20 and $q^*$ equal to 2 or 3 in different DGPs, and the sample size of 50, 100, and 200. Moreover, they dropped the first largest principle component of $\psi(x_t'\gamma_j)$ to avoid the multicollinearity problem. In our paper, for the simulation results, we tried both the case with dropping the first principle component and without dropping the first principle component, the results were similar. So we keep the original LWG method to drop the first principal component for the LWG test in this paper. The information set is $x_t = y_{t-1}$ for Block 1 and $x_t = (x_{t1} \ x_{t2})'$ for Block 2.

In practice, we need to generate $\gamma$’s carefully so that $x_t'\gamma_j$ is within a suitable range. If $\gamma$’s are chosen to be too small then activation functions $\psi$’s are approximately
linear in $x$, and we want to avoid this situation since they can not capture much non-linearity; if $\gamma$’s are too large the activation functions $\psi$’s will take values close to 0 or 1 (their minimum or maximum values), and we want to avoid this situation as well. The logistic squasher $\psi(x'\gamma_j) = [1 + \exp(-x'\gamma_j)]^{-1}$ is used with $\gamma_j$ being generated from the uniform distribution on $[-2,2]$ and $y_t, x_t$ being rescaled onto $[0, 1]$.

Bierens (1990) suggested an alternative randomization method for obtaining a $\chi^2$ limiting distribution. Following theorem 4 in Bierens (1990) and applying to our context, suppose $\gamma_0$ is a point in the $q$-dimension $\Gamma$ space. Let $\hat{\gamma} = \arg \max_{\gamma \in \Gamma} \hat{T}(\gamma)$, where $\hat{T}(\gamma)$ is a consistent estimator of the statistic in equation (4.13). For some real numbers $\lambda > 0$ and $\rho \in (0, 1)$, let $\tilde{\gamma} = \gamma_0$ if $\hat{T}(\hat{\gamma}) - \hat{T}(\gamma_0) \leq \lambda n^\rho$, otherwise $\tilde{\gamma} = \hat{\gamma}$. Then under $H_0$, $\hat{T}(\tilde{\gamma})$ has a $\chi^2$ distribution. However, this result has some drawbacks. Firstly, the choice of $\tilde{\gamma}$ may be sensitive to the real numbers $\lambda$ and $\rho$. Secondly and more importantly, the choice of $\tilde{\gamma}$ depends on a $q$-dimensional maximization problem. If we choose $q$ to be too small, say 3, then the activation functions may not perform well as a universal approximator. If we choose $q$ to be moderately large like 10, then it will be very difficult to find the global maximum. Although theorem 5 in Bierens (1990) is more practical, it still requires the chosen sequence to be dense in the $\Gamma$ space and the required number of $\gamma$’s in the chosen sequence will explode exponentially as $q$ increases. This motivates us to use the principle components of the activation functions rather than the activation functions themselves in our statistics, and just simply generating a large number of $\gamma$’s randomly from uniform distribution.

In generating $\gamma_j$ randomly from the uniform distribution on $[-2,2]$, we did it in two different ways in our Monte Carlo experiment, namely by newly generating them for each replication or by fixing one same set of randomly generated $\gamma_j$ for all replications.
To compare the test results using randomized and fixed hidden units across replications, we report the Monte Carlo results using two methods to generate the $\gamma$’s in Panels A and B of Table 1.

Table 1A reports the size and power for the ANN test with $q = 20$ and $q = 1000$ using uniformly randomized generated hidden units across replications. The numbers in the tables are the rejection frequencies under the null hypothesis at 5% and 10% levels. It is seen that both $T_n(20)$ and $T_n(1000)$ have good size. The power for both are similar. Hence, $T_n(q)$ with small $q$ and large $q$ behaves equally well in size and power.

Figure 1 shows the Monte Carlo distribution of the test statistic $T_n(q)$ from the 1000 Monte Carlo replications with the sample size $n = 200$. The three figures in the left panel are for $T_n(20)$, and the three figures in the right panel are for $T_n(1000)$. The solid line shows the asymptotic distribution, $\chi^2_3$. All three DGPs in Figure 1 are linear in mean. Figure 1 confirms the size result of Table 1, showing that both $T_n(20)$ and $T_n(1000)$, despite the very different numbers of phantom activations, have the finite sample distributions very close to the asymptotic $\chi^2_3$ distribution. These findings hold for all three DGPs under the null – AR, Linear($\rho = 0$), and Linear($\rho = 0.7$), that are linear in mean.

Table 1B repeats Table 1A using fixed hidden unit activations. In Table 1B, we generate $\gamma_j$ from $U[-2, 2]$ and fix it across all 1000 replications. The results are similar to those in Table 1A – the size and power of $T_n(20)$ and $T_n(1000)$ are equally good. From Tables 1A, 1B we see that both randomly generated and fixed $\gamma$’s provide good size. For power, when $\gamma$’s are fixed, we see increasing power as we increase $q$ from 20 to 1000 for Block 1. But for Block 2, the performance is similar. In general, fixed $\gamma$’s can not beat randomly generated $\gamma$’s in terms of power.
We also examine the possible effect of the conditional heteroskedasticity on the test. The AR in Block 1 is modified to have conditionally heteroskedastic errors as follows:

\begin{align*}
\text{AR-ARCH} : & \quad y_t = 0.6y_{t-1} + \varepsilon_t, \quad h_t^2 = E(\varepsilon_t^2|y_{t-1}) = 0.9 + 0.1\varepsilon_{t-1}^2 \quad (3.7) \\
\text{AR-GARCH} : & \quad y_t = 0.6y_{t-1} + \varepsilon_t, \quad h_t^2 = E(\varepsilon_t^2|y_{t-1}) = 0.1 + 0.1\varepsilon_{t-1}^2 + 0.8h_{t-1}^2 \quad (3.8)
\end{align*}

In the cases when the errors are conditionally heteroskedastic, the test statistic in (3.6) is not valid. We use the test statistic in equation (4.13) with \( \Psi_t \) replaced by \( \Psi^*_t \) and a corresponding consistent covariance matrix used. The test statistic has a valid asymptotic distribution of \( \chi^2_q \). Table 1C reports the size of the test statistic, which is very close to the nominal size. The good size and good power of the randomized ANN tests under conditional homoskedasticity presented in Table 1A and Table 1B are not affected under conditional heteroskedasticity when the heteroskedasticity-robust statistics are employed as shown in Table 1C.

Table 1 and Figure 1 are in line with the known results in the literature showing outstanding properties of the ANN test even using a very small number of randomized hidden activations. These results do not show any difference in \( T_n(20) \) and \( T_n(1000) \), and thus they do not reveal some hidden problem of using a small number of randomized hidden activations.

In the next two sections, we show apparent difference in \( T_n(20) \) and \( T_n(1000) \). The main finding is that the ANN test with a small \( q \), say \( T_n(20) \), is not reliable to use in practice as it exhibits substantial variation to the random activations, while the ANN test with a large \( q \), say \( T_n(1000) \), is quite robust to the randomized activations as the large number of random activation is more dense in the nonlinear function space and thus reduces the variation of the statistic substantially.
To demonstrate the advantage of increasing $q$, we first conduct a Monte Carlo experiment again, in Section 3.4, but with only 5 replications for each DGP (rather than taking average over 1000 replications). We next apply $T_n(20)$ and $T_n(1000)$ to five monthly economic time series in Section 5 to show the advantage of $T_n(1000)$ over $T_n(20)$.

### 3.4 Small $q$ vs. Large $q$ in Sensitivity to Randomized Hidden Unit Activations

The simulation results reported in LWG (1993) and also in the previous section, show that the LWG has proper size and good power. However there is a hidden problem of the ANN test with small $q$. That is when $q$ is small, the statistic and the corresponding p-value are sensitive to the randomized hidden unit activations.

Consider a sample \( \{ y_t \}_{t=1}^{n=200} \) for which the ANN statistic $T_n \left( q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)} \right)$ and its p-value $P_i$ are computed from $m$ different randomly generated activation parameters $\left( \gamma_1^{(i)}, \ldots, \gamma_q^{(i)} \right)_{i=1}^{m}$. Even if we use one same sample, it is possible that we sometimes get a small statistic and fail to reject the null for some $i$, while other times we get a statistic large enough to reject the null for other $i$. Thus we may draw contradictory conclusions because of this sensitivity. As a result, the ANN test with small $q$ can not be applied to empirical data and we need a solution to this problem.

We can deal with this problem in the following three ways. One approach is Teräsvirta, Lin and Granger (1993), who use a Taylor series expansion of the ANN function $f(x_t, \theta)$ in (3.1) to write it into a parametric nonlinear approximation, and compare the estimated model with a linear model by the Wald test or LR test. The
second approach is to generate \((\gamma_1, \ldots, \gamma_q)\) randomly from their parameter space \(\Gamma\) and integrate the statistic over \(\Gamma\) with a certain weight function \(\phi(\gamma_1, \ldots, \gamma_q)\). This is to take a weighted average ANN statistic over the nuisance parameter space. The asymptotic theory has been established. But implementing this will require either the tabulation of the asymptotic distribution via simulation as it involves the integration of the Gaussian process or the use of bootstrap. Bierens (1982), Bierens (1990), Bierens and Ploberger (1997), and Härdle and Mammen (1993) take the statistics integrated over the nuisance parameter space. Corradi and Swanson (2002) use this method to test for nonlinear Granger-causality in out of sample. Alternative to taking the average of the statistic over nuisance parameter space \(\Gamma\), Rossi and Inoue (2012) take the maximum of the statistic over \(\Gamma\) and Hansen and Timmermann (2011) take the minimum p-value over \(\Gamma\). Their methods are in essence the same because of the one-to-one mapping between the statistic and the p-value. The asymptotic distributions of these statistics are integrals of Brownian motion. To obtain the correct critical value we need to either use bootstrap or follow the conditional p-value approach of Hansen (1996). Both methods are not easy to use so we turn to seek a simple and practical solution to the nuisance parameter problem.

This paper considers an obvious approach, the third approach, which is to increase \(q\) to a very large number. To compare how the ANN test works for small \(q\) and large \(q\), we simulate a sample \(\{y_t\}_{t=1}^{n=200}\) using DGP “Linear” in Block 2 with \(x_1\) and \(x_2\) following a bivariate normal distribution with correlation \(\rho = 0.7\). Then we generate \(m = 100\) different randomly generated activation parameters \(\left(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)_{i=1}^{m=100}\), with which the ANN test statistic \(T_n\left(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}\right)\) and its p-value \(P_i\) are computed. We plot the histogram of the p-values and statistics with \(q = 20\) or 1000 in Figure 3.
When \( q = 20 \), the p-values range from 0.0806 to 0.6719 for \( i = 1, \ldots, m = 100 \) (Figure 2a). We observe three of the 100 p-values are less than 0.10, which means in these three cases we incorrectly reject the null hypothesis at 10% level. When we increase \( q \) to 1000 the p-values range from 0.2784 to 0.4567, all above the 10% level (Figure 2b). From these experiments we conjecture that if \( q \) is large enough, the p-value will be concentrated to a small area or even converge to a point. The sample variances of the p-values are 0.0255 and 0.0013 for \( q = 20 \), 1000 respectively. We also plot histograms of the \( m \) test statistics \( \left\{ T_n(q \mid \gamma_1^{(i)}, \ldots, \gamma_q^{(i)}) \right\}_{i=1}^{m=100} \) with \( q = 20 \) (Figure 2c) and \( q = 1000 \) (Figure 2d). Since there is one-to-one mapping between the test statistic and the p-value, we shall see the similar pattern in the test statistic when \( q \) increases.

Table 2 reports the range and standard deviation (SD) of the p-values of \( T_n(q) \) for \( m = 100 \) randomized hidden unit activations. For each DGP, we report the results for 5 replications. For each replication, we conduct testing with \( m \) randomized hidden unit activations. Comparing the range and SD of the p-values for \( q = 20 \) and \( q = 1000 \), we find that when \( q \) increases the range of p-value gets tighter and SD gets smaller, which makes the test outcome more stable over the \( m \) randomizations of \( \gamma_j \)'s. When the DGP has an ARCH error tighter range and smaller SD are also found across all 5 replications as \( q \) increases from 20 to 1000. The results for AR-GARCH is not reported here since it is similar to the AR-ARCH case. Hence, increasing \( q \) makes the randomized ANN test more stable as well even under conditional heteroskedasticity.

Both Figure 2 and Table 2 show that increasing \( q \) is a good solution to the problem caused by randomizing the activation parameters. While the ANN statistic with a small number (\( q = 20 \)) of randomized phantom activations exhibits large variation over \( i = 1, \ldots, m \), it becomes stable with \( q = 1000 \). We can robustify the ANN test and
reduce its sensitivity to the randomization of γ’s by simply increasing q.

3.5 Small q vs. Large q in Applications

In this section, we compare $T_n(20)$ and $T_n(1000)$ for the same five monthly US economic time series used in LWG (1993) with updated time period from 1990:1 to 2011:12 with $n = 264$. The five series are US/Japan exchange rate (EX); US three-month T-bill interest rate (INT); US M2 money stock (M2); US personal income (PI), and US unemployment rate (UNE). We have made the same transformation as in LWG (p. 287), by taking logarithms and/or the first differencing, to ensure stationarity.

For each $\{y_t\}_{t=1}^{n=264}$ of these five series, we fit a linear AR(1) model under $H_0$, so that the ANN has one input $x_t = y_{t-1}$. The ANN statistic $T_n(q | \gamma_1^{(i)}, \ldots, \gamma_q^{(i)})$ and its p-value $P_i$ are computed from $m$ randomly generated activation parameters $(\gamma_1^{(i)}, \ldots, \gamma_q^{(i)})_{i=1}^m$. Table 3 reports the p-values $\{P_i\}$ with $i = 1, \ldots, m = 20$. Table 3 also reports the Hochberg’s (1988) Bonferroni bound $HB(m)$ and the Simple Bonferroni bound $SB(m)$, both to be defined below, computed using the first $m$ p-values (with $m = 5, 20$). Figure 3 presents the histograms of the p-values $\{P_i\}_{i=1}^{m=100}$.

For exchange rate and unemployment rate data, both the $T_n(20)$ and $T_n(1000)$ give consistent results among 20 times of tests. So with both $q = 20, 1000$, the null hypothesis of linearity is not rejected for exchange rate in all 20 p-values, but it is clearly rejected for unemployment rate by the ANN test using all $m = 20$ randomized hidden unit activations. However, for personal income PI, using $T_n(20)$ will give 2 times of failure of rejection in 20 randomized neural network activations, while using $T_n(1000)$ test, we reject the linearity using all 20 randomizations. For the M2 series, using $T_n(20)$ and $T_n(1000)$ will give us contradicting conclusions, as $T_n(20)$ rejects the
null hypothesis 8 times out of 20 and $T_n(1000)$ rejects linearity in all 20 statistics. In this case, using $T_n(1000)$ yields more reliable result. For the interest rate INT, both $T_n(20)$ and $T_n(1000)$ give some uncertainty in the results in the sense that there are 3 or 4 times of failure of rejection out of the total 20 randomized activations. To examine this case further, we further increase $q$. The results (not shown in the table) show that when $q$ increased to 2000, we can get 19 times of rejection out of 20. For INT, if $q = 1000$, some p-values are greater than 10% and some are even greater than 20%. But if $q = 2000$, all p-values are below 10% except one that is only slightly above it.

Table 3 reports the p-values for $T_n(20)$ and $T_n(1000)$ with $m = 20$ different randomly generated hidden unit activation parameters $(\gamma^{(i)}_1, \ldots, \gamma^{(i)}_q)_{i=1}^m$. A low p-value suggests a rejection of the null hypothesis of linearity in conditional mean. Since the tests may not give consistent results over the different randomized activations, we use Bonferroni bounds on the p-value as a reference value. Let $\{P_1, \ldots, P_m\}$ be the p-values of $m$ different randomized activations, and let $\{P_{(1)}, \ldots, P_{(m)}\}$ denote the ordered p-values from the smallest to the largest. Then the Bonferroni inequality leads to rejection of the null hypothesis at level $\alpha$ if $P_{(1)} \leq \frac{\alpha}{m}$, so we call $SB(m) := mP_{(1)}$ the Simple Bonferroni bound. One disadvantage of the Simple Bonferroni bound is that it is too conservative when $m$ is large. Hochberg (1988) modified the rejection rule to reject the null hypothesis if there exists an $i$ such that $P_{(i)} \leq \frac{\alpha}{m - i + 1}$, $i = 1, \ldots, m$. We call $HB(m) := \min_{i=1,\ldots,m}(m - i + 1)P_{(i)}$ the Hochberg Bonferroni bound. In Table 3, reported are $SB(m)$ and $HB(m)$ with $m = 5, 20$.

A disadvantage of the Simple Bonferroni bound is that it could be larger than 1, especially when $m$ is large. The Simple Bonferroni bound is more sensitive to $q$ than the Hochberg Bonferroni bound. Comparing Bonferroni bounds over $q = 20, 1000$, the
Hochberg Bonferroni bounds for $m = 5$ and $m = 20$ are close for $T_n(1000)$, but the difference between the two bounds $HB(5)$ and $HB(20)$ is larger for $T_n(20)$. Hence, increasing the number of the randomized hidden activations not only makes the ANN test more robust but also the Bonferroni bounds tighter. From the formula $HB(m) := \min_{i=1,\ldots,m}(m-i+1)P_i$, it is easy to see that, when $q$ is large, the Hochberg Bonferroni bound tend to be the maximum p-value $HB(m) \approx P(m)$ since the p-values tend to be concentrated to a small region as discussed in the previous section (Table 2 and Figure 2). However, when $q$ is small, the Hochberg Bonferroni bound may give inconsistent conclusion according to different values of $m$. For instance, for the money stock M2 series, for $T_n(20)$, we do not reject linearity when $m = 5$ at 10\% level, yet we reject linearity when $m = 20$ at 10\% level. And in this case, we can reject linearity using $T_n(1000)$ with both $m = 5$ and $m = 20$. Thus the Hochberg Bonferroni bound is preferred to the Simple Bonferroni bound. Moreover, if we use $T_n(1000)$ instead $T_n(20)$, we can take a smaller value of $m$ and get reliable conclusion.

The reported numbers in the last part of Table 3 are the rejection frequency in these $m = 20$ p-values that are less than 0.10 (at 10\% level), $REJ = \frac{1}{m} \sum_{i=1}^{m=20} 1(P_i \leq 0.10)$. To compare the rejection frequency using different approaches, we compare the rejection frequency using the Bonferroni approach to the False Discovery Rate (FDR) of Storey (2003) and Benjamini and Hochberg (1995). The results are reported in the last two rows of Table 3. $REJ-B$ is the rejection frequency using the Bonferroni approach, where $REJ-B = \sum_{i=1}^{m=20} 1(P_i \leq 0.10 \frac{i}{20})$. $REJ-FDR$ is the rejection frequency using FDR, where $REJ-FDR = \sum_{i=1}^{m=20} 1(P_i \leq 0.10 \frac{i}{20})$. Note that, using the Bonferroni approach we get fewer times of rejection for interest rate for both $q = 20$ and $q = 1000$, while for individual test, we can reject most of the time. This problem can be solved if we use
FDR. The results show that FDR can improve the power of the test for all the series. Storey (2003) pointed out that the positive False Discovery Rate (pFDR) could improve the power of FDR when the number of tests is large. In our study we find the rejection frequency of pFDR depends heavily on the choice of tuning parameter. As we get good power for our data with FDR, we do not report the results with pFDR here.

In addition to Table 3 for which 20 p-values (with \( m = 20 \)) are used, we also experiment this with \( m = 100 \) random draws of the hidden unit activations and 100 p-values are presented in Figure 3. For all five economic time series, the p-values tend to get concentrated at a narrow region or even converge to a single value when \( q = 1000 \) compared with \( q = 20 \). For M2 data, the p-values of \( T_n(20) \) range widely from 0 to 1 and close to 1 for around 40 times among the 100 p-values, while all the 100 p-values of \( T_n(1000) \) are near zero. For personal income, when \( q = 1000 \), we can get rejection among all 100 times of tests while when \( q = 20 \), we cannot reject for around 10 times of tests. For interest rate INT as \( REJ \) becomes \( \frac{19}{20} \) when we experiment it with \( q = 2000 \) (not shown). These results clearly indicate that choosing a large \( q = 1000 \) can give more stable conclusion compared with choosing a small \( q = 20 \).

### 3.6 Conclusions

In this paper, we revisit the ANN-based test statistics for neglected nonlinearity in conditional mean. The ANN test has a set of nuisance parameters that are not identified under the null hypothesis. As the nuisance parameters are identified only under the alternative, the alternative ANN model can be estimated to form a Wald-type test statistic. However, the estimation of the ANN models are known to be difficult and the estimated models are often contaminated by large estimation errors. To avoid
the estimation of the ANN models, LWG (1993) suggested a noble test in a Lagrange multiplier (LM) test framework for which the ANN model under the alternative hypothesis needs not be estimated. As suggested in LWG (1993), in constructing an LM test, the unidentified nuisance parameters under the null hypothesis can be randomly generated from their parameter space. LWG show excellent performance of the ANN test when a small number of hidden activations is based on the randomly generated nuisance parameters.

It has not been noted in the literature that the ANN test is sensitive to the number of the randomized activations. We demonstrate this sensitivity problem and propose a simple solution. We examine how the performance of the ANN test can be improved by simply increasing the number of randomized hidden unit activations. This paper shows that the benefit of increasing it is substantial. This robustification is reliable and does not require either the use of Bonferroni bounds or the integration of the test statistic over the nuisance parameter. We provide a practically useful insight to make the ANN test reliably applicable in applied work. As increasing the number of random activations is almost costless, the ANN test based on “many” randomized hidden unit neural network activations can be easily included in a diagnostics toolbox for applied research.
References


Table 3.1 Monte Carlo: Size and Power of the ANN Test

Panel A. Using Randomized Hidden Unit Activations

<table>
<thead>
<tr>
<th></th>
<th>$q = 20$</th>
<th></th>
<th>$q = 1000$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>AR</td>
<td>0.037</td>
<td>0.085</td>
<td>0.041</td>
<td>0.089</td>
</tr>
<tr>
<td>TAR</td>
<td>0.263</td>
<td>0.391</td>
<td>0.268</td>
<td>0.374</td>
</tr>
<tr>
<td>SGN</td>
<td>0.812</td>
<td>0.901</td>
<td>0.835</td>
<td>0.910</td>
</tr>
<tr>
<td>NAR</td>
<td>0.079</td>
<td>0.162</td>
<td>0.099</td>
<td>0.178</td>
</tr>
<tr>
<td>MRS</td>
<td>0.179</td>
<td>0.273</td>
<td>0.197</td>
<td>0.284</td>
</tr>
<tr>
<td>Linear($\rho = 0$)</td>
<td>0.047</td>
<td>0.105</td>
<td>0.049</td>
<td>0.097</td>
</tr>
<tr>
<td>Linear($\rho = 0.7$)</td>
<td>0.045</td>
<td>0.097</td>
<td>0.058</td>
<td>0.106</td>
</tr>
<tr>
<td>Interaction($\rho = 0$)</td>
<td>0.112</td>
<td>0.183</td>
<td>0.082</td>
<td>0.141</td>
</tr>
<tr>
<td>Interaction($\rho = 0.7$)</td>
<td>0.244</td>
<td>0.252</td>
<td>0.261</td>
<td>0.369</td>
</tr>
<tr>
<td>Squared($\rho = 0$)</td>
<td>0.191</td>
<td>0.297</td>
<td>0.186</td>
<td>0.272</td>
</tr>
<tr>
<td>Squared($\rho = 0.7$)</td>
<td>0.346</td>
<td>0.375</td>
<td>0.238</td>
<td>0.352</td>
</tr>
</tbody>
</table>
Panel B. Using Fixed Hidden Unit Activations

<table>
<thead>
<tr>
<th></th>
<th>$q = 20$</th>
<th>$q = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5% 10%</td>
<td>5% 10%</td>
</tr>
<tr>
<td>AR</td>
<td>0.055 0.095</td>
<td>0.047 0.103</td>
</tr>
<tr>
<td>TAR</td>
<td>0.197 0.285</td>
<td>0.318 0.402</td>
</tr>
<tr>
<td>SGN</td>
<td>0.838 0.846</td>
<td>0.908 0.915</td>
</tr>
<tr>
<td>NAR</td>
<td>0.081 0.115</td>
<td>0.151 0.184</td>
</tr>
<tr>
<td>MRS</td>
<td>0.134 0.166</td>
<td>0.201 0.254</td>
</tr>
<tr>
<td>Linear($\rho = 0$)</td>
<td>0.055 0.115</td>
<td>0.043 0.100</td>
</tr>
<tr>
<td>Linear($\rho = 0.7$)</td>
<td>0.040 0.101</td>
<td>0.035 0.093</td>
</tr>
<tr>
<td>Interaction($\rho = 0$)</td>
<td>0.131 0.215</td>
<td>0.068 0.132</td>
</tr>
<tr>
<td>Interaction($\rho = 0.7$)</td>
<td>0.190 0.280</td>
<td>0.213 0.334</td>
</tr>
<tr>
<td>Squared($\rho = 0$)</td>
<td>0.130 0.230</td>
<td>0.160 0.265</td>
</tr>
<tr>
<td>Squared($\rho = 0.7$)</td>
<td>0.193 0.284</td>
<td>0.245 0.367</td>
</tr>
</tbody>
</table>

Panel C. Size with Conditional Heteroskedasticity

<table>
<thead>
<tr>
<th></th>
<th>$q = 20$</th>
<th>$q = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5% 10%</td>
<td>5% 10%</td>
</tr>
<tr>
<td>AR-ARCH</td>
<td>0.048 0.112</td>
<td>0.058 0.107</td>
</tr>
<tr>
<td>AR-GARCH</td>
<td>0.058 0.122</td>
<td>0.055 0.113</td>
</tr>
<tr>
<td>AR</td>
<td>0.051 0.100</td>
<td>0.049 0.116</td>
</tr>
</tbody>
</table>

Notes: Sample size is $n = 200$. Reported values are the rejection frequencies of the $T_n(q)$ tests out of the total 1000 Monte Carlo replications, at 5% and 10% levels. In Panel A and Panel C, the hidden
unit activations $\gamma$’s are randomly generated for each replication. In Panel B, the hidden unit activations are fixed to be one random draw from $U[-2, 2]$ for all replications. The ANN test statistic in (3.6) is used in Panel A and Panel B, while the heteroskedasticity robust statistic of the form in (4.13) with the principal components is used in Panel C.
Table 3.2 P-values of $T_n(q)$ with $m = 100$ Randomizations of $q$ Hidden Unit Activations

Panel A. Block 1

<table>
<thead>
<tr>
<th></th>
<th>Range $q=20$</th>
<th>Range $q=1000$</th>
<th>SD $q=20$</th>
<th>SD $q=1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AR</strong></td>
<td>0.5540 0.0391</td>
<td>0.1315 0.0063</td>
<td>0.4582 0.0814</td>
<td>0.1030 0.0139</td>
</tr>
<tr>
<td></td>
<td>0.2468 0.0382</td>
<td>0.0686 0.0081</td>
<td>0.6711 0.0898</td>
<td>0.1435 0.0182</td>
</tr>
<tr>
<td></td>
<td>0.4973 0.1188</td>
<td>0.1353 0.0262</td>
<td>0.4582 0.0814</td>
<td>0.1030 0.0139</td>
</tr>
<tr>
<td><strong>TAR</strong></td>
<td>0.7319 0.0448</td>
<td>0.1691 0.0076</td>
<td>0.8275 0.2084</td>
<td>0.2567 0.0469</td>
</tr>
<tr>
<td></td>
<td>0.4369 0.0291</td>
<td>0.0943 0.0061</td>
<td>0.5807 0.1165</td>
<td>0.1389 0.0237</td>
</tr>
<tr>
<td></td>
<td>0.0485 0.0006</td>
<td>0.0092 0.0001</td>
<td>0.0004 0.0000</td>
<td>0.0000 0.0000</td>
</tr>
<tr>
<td><strong>SGN</strong></td>
<td>0.0010 0.0002</td>
<td>0.0002 0.0000</td>
<td>0.1525 0.0186</td>
<td>0.0262 0.0043</td>
</tr>
<tr>
<td></td>
<td>0.0280 0.0017</td>
<td>0.0062 0.0003</td>
<td>0.0004 0.0000</td>
<td>0.0000 0.0000</td>
</tr>
<tr>
<td></td>
<td>0.1791 0.0199</td>
<td>0.0361 0.0037</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 (Continued).

Panel A. Block 1

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th>SD</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(q=20)</td>
<td>(q=1000)</td>
<td>(q=20)</td>
<td>(q=1000)</td>
<td></td>
</tr>
<tr>
<td>NAR</td>
<td>0.8330</td>
<td>0.2214</td>
<td>0.2415</td>
<td>0.0424</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4501</td>
<td>0.0294</td>
<td>0.0812</td>
<td>0.0058</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8446</td>
<td>0.0595</td>
<td>0.1806</td>
<td>0.0121</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5092</td>
<td>0.1478</td>
<td>0.1435</td>
<td>0.0270</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0869</td>
<td>0.0218</td>
<td>0.0198</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td>MRS</td>
<td>0.8610</td>
<td>0.2738</td>
<td>0.2090</td>
<td>0.0557</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9992</td>
<td>0.1667</td>
<td>0.3300</td>
<td>0.0298</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2210</td>
<td>0.0337</td>
<td>0.0529</td>
<td>0.0072</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.3194</td>
<td>0.0247</td>
<td>0.0557</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.7001</td>
<td>0.2378</td>
<td>0.2062</td>
<td>0.0447</td>
<td></td>
</tr>
<tr>
<td>AR-ARCH</td>
<td>0.2754</td>
<td>0.1164</td>
<td>0.0784</td>
<td>0.0219</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5257</td>
<td>0.0696</td>
<td>0.1998</td>
<td>0.0155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6207</td>
<td>0.0188</td>
<td>0.1695</td>
<td>0.0034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.3866</td>
<td>0.0560</td>
<td>0.1281</td>
<td>0.0121</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1530</td>
<td>0.0046</td>
<td>0.0424</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 (Continued).

Panel B. Block 2

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th></th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q=20$</td>
<td>$q=1000$</td>
<td>$q=20$</td>
<td>$q=1000$</td>
</tr>
<tr>
<td>Linear $(\rho = 0)$</td>
<td>0.3846</td>
<td>0.0855</td>
<td>0.0781</td>
<td>0.0149</td>
</tr>
<tr>
<td></td>
<td>0.9793</td>
<td>0.5644</td>
<td>0.2943</td>
<td>0.1345</td>
</tr>
<tr>
<td></td>
<td>0.9389</td>
<td>0.1879</td>
<td>0.2807</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>0.7879</td>
<td>0.1093</td>
<td>0.2289</td>
<td>0.0212</td>
</tr>
<tr>
<td></td>
<td>0.6020</td>
<td>0.0755</td>
<td>0.1425</td>
<td>0.0179</td>
</tr>
<tr>
<td>Linear $(\rho = 0.7)$</td>
<td>0.2470</td>
<td>0.0373</td>
<td>0.0490</td>
<td>0.0065</td>
</tr>
<tr>
<td></td>
<td>0.4526</td>
<td>0.0120</td>
<td>0.0616</td>
<td>0.0027</td>
</tr>
<tr>
<td></td>
<td>0.9684</td>
<td>0.3642</td>
<td>0.3033</td>
<td>0.0819</td>
</tr>
<tr>
<td></td>
<td>0.3680</td>
<td>0.0322</td>
<td>0.0500</td>
<td>0.0067</td>
</tr>
<tr>
<td></td>
<td>0.5981</td>
<td>0.1580</td>
<td>0.1640</td>
<td>0.0361</td>
</tr>
<tr>
<td>Intersection $(\rho = 0)$</td>
<td>0.5982</td>
<td>0.1626</td>
<td>0.1646</td>
<td>0.0361</td>
</tr>
<tr>
<td></td>
<td>0.8872</td>
<td>0.2460</td>
<td>0.2114</td>
<td>0.0510</td>
</tr>
<tr>
<td></td>
<td>0.9198</td>
<td>0.1641</td>
<td>0.2285</td>
<td>0.0284</td>
</tr>
<tr>
<td></td>
<td>0.4399</td>
<td>0.0914</td>
<td>0.0995</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>0.8509</td>
<td>0.1895</td>
<td>0.2234</td>
<td>0.0332</td>
</tr>
</tbody>
</table>
Table 3.2 (Continued).

Panel B. Block 2

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th></th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>q=20</td>
<td>q=1000</td>
<td>q=20</td>
<td>q=1000</td>
</tr>
<tr>
<td>Intersection</td>
<td>0.5417</td>
<td>0.0558</td>
<td>0.0819</td>
<td>0.0101</td>
</tr>
<tr>
<td>(ρ = 0.7)</td>
<td>0.4461</td>
<td>0.0606</td>
<td>0.0943</td>
<td>0.0116</td>
</tr>
<tr>
<td></td>
<td>0.9064</td>
<td>0.4364</td>
<td>0.2627</td>
<td>0.0889</td>
</tr>
<tr>
<td></td>
<td>0.4144</td>
<td>0.1015</td>
<td>0.0911</td>
<td>0.0213</td>
</tr>
<tr>
<td></td>
<td>0.1813</td>
<td>0.0205</td>
<td>0.0308</td>
<td>0.0042</td>
</tr>
<tr>
<td>Squared</td>
<td>0.5354</td>
<td>0.0481</td>
<td>0.1253</td>
<td>0.0095</td>
</tr>
<tr>
<td>(ρ = 0)</td>
<td>0.9339</td>
<td>0.4809</td>
<td>0.2538</td>
<td>0.1140</td>
</tr>
<tr>
<td></td>
<td>0.8627</td>
<td>0.4690</td>
<td>0.2661</td>
<td>0.1000</td>
</tr>
<tr>
<td></td>
<td>0.7223</td>
<td>0.1195</td>
<td>0.1503</td>
<td>0.0259</td>
</tr>
<tr>
<td></td>
<td>0.8681</td>
<td>0.0881</td>
<td>0.2040</td>
<td>0.0196</td>
</tr>
<tr>
<td>Squared</td>
<td>0.3645</td>
<td>0.0391</td>
<td>0.0632</td>
<td>0.0094</td>
</tr>
<tr>
<td>(ρ = 0.7)</td>
<td>0.6779</td>
<td>0.1191</td>
<td>0.1435</td>
<td>0.0249</td>
</tr>
<tr>
<td></td>
<td>0.9652</td>
<td>0.3963</td>
<td>0.2997</td>
<td>0.0933</td>
</tr>
<tr>
<td></td>
<td>0.3984</td>
<td>0.0475</td>
<td>0.0592</td>
<td>0.0088</td>
</tr>
<tr>
<td></td>
<td>0.3805</td>
<td>0.0556</td>
<td>0.0663</td>
<td>0.0110</td>
</tr>
</tbody>
</table>

Note: Sample size n = 200. The p-values of $T_n(q)$ are computed for simulated data of each DGP from five replications. The statistic $T_n(q)$ is computed with $m = 100$ random draws of $\{\gamma_i^{(i)}\}_{i=1,\ldots,m}$ from $U[-2, 2]$. The table reports the range and standard deviation (SD) of the $m$ p-values in each of 5 replications with $q = 20, 1000$. 
<table>
<thead>
<tr>
<th>$q$ =</th>
<th>$i = 1$</th>
<th>0.1192</th>
<th>0.7126</th>
<th>0.0023</th>
<th>0.0156</th>
<th>0.3034</th>
<th>0.0034</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i = 2$</td>
<td>0.1752</td>
<td>0.7164</td>
<td>0.8501</td>
<td>0.0096</td>
<td>0.0576</td>
<td>0.0029</td>
</tr>
<tr>
<td></td>
<td>$i = 3$</td>
<td>0.4740</td>
<td>0.7926</td>
<td>0.0665</td>
<td>0.0032</td>
<td>0.0229</td>
<td>0.0033</td>
</tr>
<tr>
<td></td>
<td>$i = 4$</td>
<td>0.8045</td>
<td>0.7726</td>
<td>0.1198</td>
<td>0.0752</td>
<td>1.0000</td>
<td>0.0108</td>
</tr>
<tr>
<td></td>
<td>$i = 5$</td>
<td>0.1565</td>
<td>0.7589</td>
<td>0.0064</td>
<td>0.0024</td>
<td>1.0000</td>
<td>0.0041</td>
</tr>
<tr>
<td></td>
<td>$i = 6$</td>
<td>0.4497</td>
<td>0.8244</td>
<td>0.0034</td>
<td>0.2006</td>
<td>1.0000</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>$i = 7$</td>
<td>0.5505</td>
<td>0.7688</td>
<td>0.0125</td>
<td>0.0595</td>
<td>0.2030</td>
<td>0.0030</td>
</tr>
<tr>
<td></td>
<td>$i = 8$</td>
<td>0.5022</td>
<td>0.7575</td>
<td>0.0608</td>
<td>0.0535</td>
<td>0.0049</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td>$i = 9$</td>
<td>0.4750</td>
<td>0.7676</td>
<td>0.0258</td>
<td>0.1487</td>
<td>0.0049</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>$i = 10$</td>
<td>0.4628</td>
<td>0.7587</td>
<td>0.0407</td>
<td>0.0115</td>
<td>1.0000</td>
<td>0.0019</td>
</tr>
<tr>
<td></td>
<td>$i = 11$</td>
<td>0.4800</td>
<td>0.7097</td>
<td>0.0121</td>
<td>0.0013</td>
<td>1.0000</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>$i = 12$</td>
<td>0.2813</td>
<td>0.8057</td>
<td>0.1246</td>
<td>0.0146</td>
<td>0.3971</td>
<td>0.0115</td>
</tr>
<tr>
<td></td>
<td>$i = 13$</td>
<td>0.4717</td>
<td>0.6834</td>
<td>0.0003</td>
<td>0.1007</td>
<td>1.0000</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>$i = 14$</td>
<td>0.4730</td>
<td>0.6988</td>
<td>0.0217</td>
<td>0.0537</td>
<td>0.0495</td>
<td>0.0025</td>
</tr>
<tr>
<td></td>
<td>$i = 15$</td>
<td>0.5196</td>
<td>0.7630</td>
<td>0.0090</td>
<td>0.0015</td>
<td>0.8678</td>
<td>0.0029</td>
</tr>
<tr>
<td>$q$ =</td>
<td>EX</td>
<td>INT</td>
<td>M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1000</td>
<td>20</td>
<td>1000</td>
<td>20</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>$i = 16$</td>
<td>0.1573</td>
<td>0.7742</td>
<td>0.4018</td>
<td>0.0332</td>
<td>0.0033</td>
<td>0.0181</td>
<td></td>
</tr>
<tr>
<td>$i = 17$</td>
<td>0.5109</td>
<td>0.8010</td>
<td>0.0044</td>
<td>0.0191</td>
<td>0.0351</td>
<td>0.0067</td>
<td></td>
</tr>
<tr>
<td>$i = 18$</td>
<td>0.5241</td>
<td>0.7831</td>
<td>0.0102</td>
<td>0.0955</td>
<td>0.9987</td>
<td>0.0037</td>
<td></td>
</tr>
<tr>
<td>$i = 19$</td>
<td>0.4386</td>
<td>0.7584</td>
<td>0.0017</td>
<td>0.0333</td>
<td>0.0378</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>$i = 20$</td>
<td>0.4380</td>
<td>0.7807</td>
<td>0.0247</td>
<td>0.0354</td>
<td>0.1197</td>
<td>0.0042</td>
<td></td>
</tr>
<tr>
<td>$HB(5)$</td>
<td>0.5256</td>
<td>0.7926</td>
<td>0.0115</td>
<td>0.0120</td>
<td>0.1145</td>
<td>0.0082</td>
<td></td>
</tr>
<tr>
<td>$HB(20)$</td>
<td>0.8045</td>
<td>0.8244</td>
<td>0.0060</td>
<td>0.0260</td>
<td>0.0660</td>
<td>0.0181</td>
<td></td>
</tr>
<tr>
<td>$SB(5)$</td>
<td>0.5960</td>
<td>3.5630</td>
<td>0.0115</td>
<td>0.0120</td>
<td>0.1145</td>
<td>0.0145</td>
<td></td>
</tr>
<tr>
<td>$SB(20)$</td>
<td>2.3840</td>
<td>13.6680</td>
<td>0.0060</td>
<td>0.0260</td>
<td>0.0660</td>
<td>0.0380</td>
<td></td>
</tr>
<tr>
<td>$REJ$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{16}{20}$</td>
<td>$\frac{17}{20}$</td>
<td>$\frac{8}{20}$</td>
<td>$\frac{20}{20}$</td>
<td></td>
</tr>
<tr>
<td>$REJ-B$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{5}{20}$</td>
<td>$\frac{4}{20}$</td>
<td>$\frac{3}{20}$</td>
<td>$\frac{14}{20}$</td>
<td></td>
</tr>
<tr>
<td>$REJ-FDR$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{0}{20}$</td>
<td>$\frac{16}{20}$</td>
<td>$\frac{16}{20}$</td>
<td>$\frac{3}{20}$</td>
<td>$\frac{20}{20}$</td>
<td></td>
</tr>
<tr>
<td>$q =$</td>
<td>PI</td>
<td>UNE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 1000</td>
<td>20 1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 1$</td>
<td>0.0000 0.0000</td>
<td>0.0011 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 2$</td>
<td>0.0000 0.0000</td>
<td>0.0003 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 3$</td>
<td>0.0000 0.0000</td>
<td>0.0024 0.0005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 4$</td>
<td>0.0361 0.0000</td>
<td>0.0005 0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 5$</td>
<td>0.0000 0.0000</td>
<td>0.0003 0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 6$</td>
<td>0.0000 0.0000</td>
<td>0.0005 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 7$</td>
<td>0.0000 0.0000</td>
<td>0.0004 0.0004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 8$</td>
<td>0.0017 0.0000</td>
<td>0.0004 0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 9$</td>
<td>0.0000 0.0000</td>
<td>0.0006 0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 10$</td>
<td>0.0000 0.0000</td>
<td>0.0059 0.0007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 11$</td>
<td>0.0000 0.0000</td>
<td>0.0002 0.0005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 12$</td>
<td>0.0000 0.0000</td>
<td>0.0003 0.0007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 13$</td>
<td>0.0000 0.0000</td>
<td>0.0013 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 14$</td>
<td>0.0000 0.0000</td>
<td>0.0003 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 15$</td>
<td>0.0000 0.0000</td>
<td>0.0003 0.0007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$q$</td>
<td>PI $20$</td>
<td>PI $1000$</td>
<td>UNE $20$</td>
<td>UNE $1000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 16$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0003$</td>
<td>$0.0006$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 17$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0003$</td>
<td>$0.0006$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 18$</td>
<td>$0.3395$</td>
<td>$0.0000$</td>
<td>$0.0005$</td>
<td>$0.0007$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 19$</td>
<td>$0.2982$</td>
<td>$0.0000$</td>
<td>$0.0004$</td>
<td>$0.0006$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i = 20$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0002$</td>
<td>$0.0007$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$HB(5)$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0012$</td>
<td>$0.0008$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$HB(20)$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0030$</td>
<td>$0.0008$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SB(5)$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0015$</td>
<td>$0.0023$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SB(20)$</td>
<td>$0.0000$</td>
<td>$0.0000$</td>
<td>$0.0040$</td>
<td>$0.0089$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$REJ$</td>
<td>$\frac{18}{20}$</td>
<td>$\frac{20}{20}$</td>
<td>$\frac{20}{20}$</td>
<td>$\frac{20}{20}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$REJ-B$</td>
<td>$\frac{17}{20}$</td>
<td>$\frac{20}{20}$</td>
<td>$\frac{19}{20}$</td>
<td>$\frac{20}{20}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$REJ-FDR$</td>
<td>$\frac{18}{20}$</td>
<td>$\frac{20}{20}$</td>
<td>$\frac{20}{20}$</td>
<td>$\frac{20}{20}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: EX: US/Japan exchange rate. INT: US three-month T-bill interest rate. M2: US M2 money stock. PI: US personal income. UNE: US unemployment rate. We use AR(1) as a model under the null hypothesis in each case. The 20 rows ($i = 1, \ldots, 20$) show the 20 sets of p-values $\{P_i\}_{i=1}^m$ of the $\text{ANN}(q)$ test statistics with $q = 20$ or 1000. $HB(m) = \min_{i=1, \ldots, m}(m - i + 1) \times P_i$ is the Hochberg’s Bonferroni bound computed the first $m$ p-values $\{P_i\}_{i=1}^m$ with $m = 5$ or 20. $HB(5)$ in the Hochberg Bonferroni bound computed using the first 5 p-values $\{P_i\}_{i=1}^5$. $SB(m) = mP_{(1)}$ is the Simple Bonferroni bound computed the first $m$ p-values. $P_{(i)}$ is the $i$th smallest (ordered from the smallest to the largest) p-value among the $m$ p-values. The reported numbers in the last three rows are the rejection frequency in these $m = 20$ p-values that are less than 0.10 (at 10% level), $REJ = \frac{1}{m} \sum_{i=1}^{m=20} 1(P_i \leq 0.10)$, $REJ-B = \sum_{i=1}^{m=20} 1(P_i \leq \frac{0.10}{20})$, and $REJ-FDR = \sum_{i=1}^{m=20} 1(P_i \leq \frac{0.10}{20})$. 

127
Figure 3.1 Monte Carlo Distribution of $T_n(q)$ under $H_0$

(a) $q = 20$. DGP: AR

(b) $q = 1000$. DGP: AR

(c) $q = 20$. DGP: Linear($\rho = 0$)

(d) $q = 1000$. DGP: Linear($\rho = 0$)

(e) $q = 20$. DGP: Linear($\rho = 0.7$)

(f) $q = 1000$. DGP: Linear($\rho = 0.7$)

Note: The histograms are the Monte Carlo distribution of the test statistic $T_n(q)$ from the 1000 Monte Carlo replications with the sample size $n = 200$. The three figures in the left panel are for $T_n(20)$, and the three figures in the right panel are for $T_n(1000)$. The solid line is the $\chi^2_3$ density.
Figure 3.2 P-values of $T_n(q)$ under $H_0$ with $m = 100$ Randomizations of $q$ Hidden Units

(a) $q = 20$. P-values

(b) $q = 1000$. P-values

(c) $q = 20$. Test Statistics

(d) $q = 1000$. Test Statistics

Note: Sample size $n = 200$. The p-values and the test statistics $T_n(q)$ are computed for a simulated data from one replication of DGP, “Linear” with $\rho = 0.7$. For the same data, the statistic $T_n(q)$ is computed with $m = 100$ random draws of $\{\gamma_i^{(i)}\}_{i=1,...,m}$ from $U[-2, 2]$. The figures are frequency histograms of the $m$ p-values and the $m$ statistics. The top panels report the p-values and the bottom panels report the statistics. The left panels are for $q = 20$ and the right panels are for $q = 1000$. 

129
Figure 3.3 Empirical Applications with $m = 100$ Randomized Hidden Unit Activations

(a) P-values of EX with $q = 20$

(b) P-values of EX with $q = 1000$

(c) P-values of INT with $q = 20$

(d) P-values of INT with $q = 1000$

(e) P-values of M2 with $q = 20$

(f) P-values of M2 with $q = 1000$
Figure 3.3 (Continued).

(g) P-values of PI with $q = 20$

(h) P-values of PI with $q = 1000$

(i) P-values of UNE with $q = 20$

(j) P-values of UNE with $q = 1000$
Chapter 4

Testing for Neglected Nonlinearity Using Regularized Artificial Neural Networks

4.1 Introduction

In this paper we explore the issues in testing for functional forms, especially for neglected nonlinearity in parametric linear models. Many papers have appeared in the recent literature which deal with the issues of how to carry out various specification tests in parametric regression models. To construct the tests, various methods are used to estimate the alternative models. For example, Fan and Li (1996), Li and Wang (1998), Zheng (1996), and Bradley and McClelland (1996) use local constant kernel regression; Hjellvik, Yao and Tjøstheim (1998) and Tjøstheim (1999) use local polynomial kernel regression; Cai, Fan and Yao (2000) and Matsuda (1999) use nonparametric functional coefficient models; Poggi and Portier (1997) use a functional autoregressive model;
Granger and Teräsvirta (1993), Teräsvirta (1996), and Corradi and Swanson (2002) use
neural network models; Eubank and Spiegelman (1990) use spline regression; Hong and
White (1995) use series regression; Stengos and Sun (1998) use wavelet methods; and
Hamilton (2001) uses a parametric flexible regression model.

There are also many papers which compare different approaches in testing for
linearity. For example, Lee, White, and Granger (1993), Teräsvirta, Lin and Granger
(1993), Teräsvirta (1996), and Lee (2001) examine the neural network test and many
test and compare it with various tests including the neural network test. Blake and
Kapetanios (1999, 2000) extend the neural network test using a radial basis function
for the neural network activation function instead of using the typical logistic function

Lee and Ullah (2001, 2003) examine the
Härdle and Mammen (1993), and Aït-Sahalia, Bickel, and Stoker (2001). Fan and Li
Whang (2000) generalizes the Kolmogorov-Smirnov and Cramer-von Mises tests to the
regression framework and compare them with the tests of Härdle and Mammen (1993)
based on nonparametric estimates of conditional mean and variances and compare them
with a number of tests such as the bispectrum test and the BDS test.

This paper further investigates the artificial neural network (ANN) test. The
ANN test is a conditional moment test whose null hypothesis consists of conditional

\footnote{For radial basis functions, see (e.g.) Campbell, Lo, and MacKinlay (1997, p. 517).}
moment conditions that hold if the linear model is correctly specified for the conditional mean. The ANN test differs from other tests by the choice of the ‘test function’ that is chosen to be the ANN’s hidden layer activations. It can be checked for their correlation with the residuals from the linear regression model. The advantage to use an ANN model to test nonlinearity is that the ANN model inherits the flexibility as a universal approximator of unknown functional form. Hornick, Stinchcombe and White (1989) show that neural network is a nonlinear flexible functional form being capable of approximating any Borel measurable function to any desired level of accuracy provided sufficiently many hidden units are available.

We consider an augmented single hidden layer feedforward neural network model in which network output $y_t$ is determined given input $x_t$ as

$$y_t = x_t'\alpha + \sum_{j=1}^{q} \beta_j \Psi (x_t' \gamma_j) + u_t,$$

where $t = 1, \ldots, T$, $x_t = (x_{1,t}, \ldots, x_{N,t})'$, $\theta = (\alpha', \beta_1', \gamma_{1}', \ldots, \gamma_{q}')'$, $\alpha = (\alpha_1, \ldots, \alpha_N)'$, $\beta = (\beta_1, \ldots, \beta_q)'$, and $\gamma_j = (\gamma_{j,1}, \ldots, \gamma_{j,N})'$ for $j = 1, \ldots, q$, and $\Psi(\cdot)$ is an activation function. An example of the activation function is the logistic function $\Psi(z) = (1 + \exp(z))^{-1}$.

$\alpha$ is a conformable column vector of connection strength from the input layer to the output layer; $\gamma_j$ is a conformable column vector of connection strength from the input layer to the hidden units, $j = 1, \ldots, q$; $\beta_j$ is a (scalar) connection strength from the hidden unit $j$ to the output unit, $j = 1, \ldots, q$; and $\Psi$ is a squashing function (e.g., the logistic squasher) or a radial basis function. Input units $x$ send signals to intermediate hidden units, then each of hidden unit produces an activation $\Psi$ that then sends signals toward the output unit. The integer $q$ denotes the number of hidden units added to the affine (linear) network. When $q = 0$, we have a two layer affine network $y_t = x_t'\alpha + u_t$.

It is well known that the ANN models are generally hard to estimate and
suffer from possibly large estimation errors which can adversely affect their ability as a universal approximator. To alleviate the estimation errors of an ANN model, it is useful to note that, for given values of $\gamma_j$'s, the ANN is linear in $x$ and the activation function $\Psi$ and therefore $(\alpha', \beta')$ can be estimated from linear regression once $(\gamma_1', \ldots, \gamma_q')$ are estimated or given. As suggested in Lee, White and Granger (LWG 1993), a set of $\gamma$'s can be randomly generated. In this paper, we will generate a large set of $\gamma$'s such that $\sum_{j=1}^{q} \beta_j \Psi(x_t \gamma_j)$ can capture the maximal nonlinear structure. The LWG statistic is designed to detect neglected nonlinearity in the linear model by checking for correlation between the residual from a linear model and the additional hidden activation functions with randomly generated $\gamma$'s. The additional hidden activation functions are hidden (or phantom) because they do not exist under the null hypothesis. The $\gamma$'s are randomly generated in testing because they are not identified under the null hypothesis. The set of randomly selected $\gamma$'s should be large enough so that it can be dense and make the ANN a universal approximator.

While the architecture of the ANN model makes a universal approximator, it involves a very large number of parameters. Kock and Teräsvirta (2011) consider regularizing the complexity of an ANN model and demonstrate that the regularization of the large dimension is crucial in using ANN models for out-of-sample forecasting. This motivates us to consider regularizing the ANN for testing for neglected nonlinearity. In fact, LWG (1993) uses a (unsupervised) regularization method, namely the principal component analysis, for the randomly activated test functions. Kock and Teräsvirta (2011) consider two (supervised) regularization approaches. They insightfully notice that the supervised regularizations will result in the size distortion in inference, and they use these approaches only for forecasting.
One supervised regularization approach considered by Kock and Teräsvirta (2011) to select a small $q^*$ from a large $q$ number of $\gamma$’s is the simple-to-general algorithm, e.g., the QuickNet algorithm of White (2006), that adds one $\gamma$ and one activation function at a time to the ANN. The QuickNet expands starting from 0 activation to $q^*$ activations until the additional hidden unit activation is not found to improve the network capability. The second supervised regularization approach considered by Kock and Teräsvirta (2011) is the general-to-simple approach. This approach, from a variable-selection perspective, reduces the number of activations from an initial large number $q$ (say, 1000) to a smaller number $q^*$ by penalizing the complexity of the ANN model. The penalized regression methods include the smoothly clipped absolute deviation penalty (SCAD) (Fan and Li 2001), adaptive Lasso (Zou 2006), adaptive elastic net (Zou and Zhang 2009), the bridge estimator (Huang, Horowitz and Ma 2008), among others. In the case where $q$ is larger than the degrees of freedom, the marginal bridge estimator (Huang, Horowitz and Ma 2008) or the sure independence screening (SIS) (Fan and Lv 2008) may be used to reduce $q$ below the degrees of freedom and then apply these estimation methods.

The third approach is to follow LWG (1993) to compute the $q^*$ principal components of the $q$ additional hidden activation functions. Since the activation functions using randomly generated $\gamma$’s may be collinear with each other and with $\mathbf{x}_t$, LWG used principal components of the $q$ additional hidden activation functions. Unlike the above two supervised approaches, the principal components are not supervised for the output $y$.

The purpose of this paper is to examine the effect of various regularization on the ANN test for neglected nonlinearity when the ANN is activated based on a
large number of random activation parameters. We learn two points. First, when we consider the Lasso, the partial least square (PLS) method, the Pretest method, and a method combining Lasso with principle components, these supervised regularization methods bring size-distortion and the ANN test suffers from the post-sample inference or post-selection inference (PoSI) problem.\(^2\) Secondly, when we use the principle component analysis (PCA) as used in LWG (1993), this unsupervised regularization of the dimension reduction does not bring the PoSI problem, works really well for a large \(q\), and the asymptotic \(\chi^2(q^*)\) distribution does well in approximating the finite sample distribution of the ANN test statistic. To sum, while the supervised regularizations are useful in forecasting as studied by Bai and Ng (2008), Bair, Hastie, Paul and Tibshirani (2006), Inoue and Kilian (2008), Huang and Lee (2010), Hillebrand, Huang, Lee and Li (2011), Kock and Teräsvirta (2011), and Kock (2011), this paper shows that regularization should not be supervised in inference. Our Monte Carlo simulation shows that the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso. This paper also demonstrates that the use of unsupervised regularization by principal components does not lead to the PoSI problem.

The plan of the paper is as follows. In Section 2 we review the ANN test. Section 3 introduces various regularizations in two types, unsupervised and supervised. Section 4 presents the simulation results which demonstrate the PoSI problem of supervised methods. Section 5 concludes.

4.2 Testing for Neglected Nonlinearity Using ANN

Consider $Z_t = (y_t, \mathbf{x}_t')'$, where $y_t$ is a scalar and $\mathbf{x}_t$ may contain a constant and lagged values of $y_t$. Consider the regression model

$$y_t = m(\mathbf{x}_t) + \varepsilon_t, \quad (4.2)$$

where $m(x_t) \equiv E(y_t|x_t)$ is the true but unknown regression function and $\varepsilon_t$ is the error term such that $E(\varepsilon_t|x_t) = 0$ by construction. To test for a parametric model $g(\mathbf{x}_t, \theta)$ we consider

$$H_0 : m(\mathbf{x}_t) = g(\mathbf{x}_t, \theta^*) \text{ almost everywhere (a.e.) for some } \theta^*, \quad (4.3)$$

$$H_1 : m(\mathbf{x}_t) \neq g(\mathbf{x}_t, \theta) \text{ on a set with positive measure for all } \theta. \quad (4.4)$$

In particular, if we are to test for neglected nonlinearity in the regression models, set $g(\mathbf{x}_t, \theta) = \mathbf{x}_t'\alpha$, $\alpha \subset \theta$. Then under $H_0$, the process $\{y_t\}$ is linear in mean conditional on $\mathbf{x}_t$, i.e.,

$$H_0 : m(\mathbf{x}_t) = \mathbf{x}_t'\alpha^* \text{ a.e. for some } \alpha^*. \quad (4.5)$$

The alternative of interest is the negation of the null hypothesis, that is,

$$H_1 : m(\mathbf{x}_t) \neq \mathbf{x}_t'\alpha \text{ on a set with positive measure for all } \alpha. \quad (4.6)$$

When the alternative is true, a linear model is said to suffer from “neglected nonlinearity” (Lee, White, and Granger 1993).

If a linear model is capable of an exact representation of the unknown function $m(\mathbf{x}_t)$, then there exists a vector $\alpha^*$ such that (4.5) holds, which implies

$$E(\varepsilon_t^*|\mathbf{x}_t) = 0 \text{ a.e.,} \quad (4.7)$$
where \( \varepsilon^*_t = y_t - \mathbf{x}_t'\alpha^* \). By the law of iterated expectations \( \varepsilon^*_t \) is uncorrelated with any measurable functions of \( \mathbf{x}_t \), say \( h(\mathbf{x}_t) \). That is,

\[
E [h(\mathbf{x}_t)\varepsilon^*_t] = 0. \tag{4.8}
\]

Depending on how we choose the ‘test function’ \( h(\cdot) \), various specification tests may be obtained. The specification tests based on these moment conditions, so called the conditional moment tests, have been studied by Newey (1985), Tauchen (1985), White (1987, 1994), Bierens (1982, 1990), LWG (1993), Bierens and Ploberger (1997) and Stinchcombe and White (1998), among others. The ANN test exploits (4.8) with the test function \( h(\cdot) \) being chosen as the neural network hidden unit activation functions.

LWG (1993) considered the test of “linearity in conditional mean” using the ANN model. To test whether the process \( y_t \) is linear in mean conditional on \( \mathbf{x}_t \), they used the following null and alternative hypothesis:

\[
H_0 : \Pr \left[ E(y_t|\mathbf{x}_t) = \mathbf{x}_t'\alpha^* \right] = 1 \quad {\text{for some}} \ \alpha^* \\
H_1 : \Pr \left[ E(y_t|\mathbf{x}_t) = \mathbf{x}_t'\alpha \right] < 1 \quad {\text{for all}} \ \alpha
\]

The procedure to construct the LWG test statistic is as follows. Under the null hypothesis that \( y_t \) is linear in conditional mean, we first estimate a linear model of \( y_t \) on \( \mathbf{x}_t \), then if any nonlinearity is neglected in the OLS regression, it will be captured by the residual term \( \hat{u}_t \). Since the ANN model inherits the flexibility as a universal approximator of unknown functional form, we can apply an ANN function to approximate any possible types of nonlinearity in the residual term \( \hat{u}_t \).

The neural network test is based on a test function \( h(\mathbf{x}_t) \) chosen as the activations of “phantom” hidden units \( \psi(\mathbf{x}_t'\gamma_j), \; j = 1, \ldots, q \), where \( \gamma_j \) are randomly generated column vectors independent of \( \mathbf{x}_t \). \( \gamma_j \)'s are not identified under the null

\[
E \left[ \psi (x_t' \gamma_j) \varepsilon_t^* \right] = 0 \quad j = 1, \ldots, q, \tag{4.9}
\]

under \( H_0 \), so that

\[
E (\Psi_t \varepsilon_t^*) = 0, \tag{4.10}
\]

where

\[
\Psi_t = (\psi (x_t' \gamma_1), \ldots, \psi (x_t' \gamma_q))'
\]

is a phantom hidden unit activation vector. Evidence of correlation of \( \varepsilon_t^* \) with \( \Psi_t \) is evidence against the null hypothesis that \( y_t \) is linear in mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations \( \psi (x_t' \gamma_j) \) would permit an improvement in network performance. Thus the tests are based on sample correlation of affine network errors with phantom hidden unit activations,

\[
n^{-1} \sum_{t=1}^{n} \Psi_t \hat{\varepsilon}_t = n^{-1} \sum_{t=1}^{n} \Psi_t (y_t - x_t' \hat{\alpha}), \tag{4.12}
\]

where \( \hat{\varepsilon}_t = y_t - x_t' \hat{\alpha} \) are estimated by OLS. Under suitable regularity conditions it follows from the central limit theorem that \( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\varepsilon}_t \overset{d}{\rightarrow} N (0, W^*) \) as \( n \rightarrow \infty \), and if one has a consistent estimator for its asymptotic covariance matrix, say \( \hat{W}_n \), then an asymptotic chi-square statistic can be formed as

\[
\left( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\varepsilon}_t \right)' \hat{W}_n^{-1} \left( n^{-1/2} \sum_{t=1}^{n} \Psi_t \hat{\varepsilon}_t \right) \overset{d}{\rightarrow} \chi^2 (q). \tag{4.13}
\]

Construct the following auxiliary regression:

\[
\hat{u}_t = x_t' \alpha + \sum_{j=1}^{q} \beta_j \psi (x_t' \gamma_j) + v_t,
\]

where \( t = 1, \ldots, T, \ x_t = (x_{1,t}, \ldots, x_{N,t})', \ \theta = (\alpha', \beta', \gamma_1', \ldots, \gamma_q')', \ \alpha = (\alpha_1, \ldots, \alpha_N)', \ \beta = (\beta_1, \ldots, \beta_q)', \) and \( \gamma_j = (\gamma_{j,1}, \ldots, \gamma_{j,N})' \) for \( j = 1, \ldots, q, \) and \( \psi (\cdot) \) is an activation
function. LWG chose the logistic function $\psi(z) = (1 + \exp(z))^{-1}$ as the activation function. If there is nonlinearity remained in the residual, we expect the goodness of fit for the auxiliary regression is high. However, one problem to estimate the auxiliary regression is that, when $q$ is large, there may exist multicollinearity between $\psi(x_t'\gamma_j)$ and $x_t$ and among $\psi(x_t'\gamma_j)$ themselves. LWG suggested to choose $q^*$ principle components of $q$ activation functions $\psi(x_t'\gamma_j)$, with $q^* < q$, and then use these $q^*$ principle components to run the auxiliary regression. Under the null hypothesis that the sequence $y_t$ is linear conditional on $x_t$, the goodness of fit in the auxiliary regression will be low. LWG (1993) constructed an LM-type test statistic which has an asymptotic $\chi^2(q^*)$ distribution under the null hypothesis. In their simulations, LWG chose $q$ equal to 10 or 20 and $q^*$ equal to 2 or 3 in different data generating processes (DGP), and the sample size 50, 100, or 200. Moreover, they dropped the first principle component of $\Psi_t$ to avoid the multicollinearity problem. In this paper, we have tried the ANN test both with and without dropping the first principle component, the results do not change much. Thus we keep the original LWG method with dropping the first principal component for the ANN test in this paper.

In practice, we need to generate $\gamma$’s carefully so that $\psi(x_t'\gamma_j)$ is within a suitable range. If $\gamma$’s are chosen to be too small then activation functions $\psi(x_t'\gamma_j)$ are approximately linear in $x$. We want to avoid this situation since they can not capture much nonlinearity. If $\gamma$’s are too large the activation functions $\psi(x_t'\gamma_j)$ take values close to $\pm 1$ (their maximum or minimum values), and we want to avoid this situation as well. In our study, for different $x$’s we generate $\gamma$’s from uniform distributions with different supports so that the activation functions are neither too small or too large.
4.3 Regularizing the ANN Test

As discussed above, LWG (1993) regularized the large number of the network activation functions using principle components in order to avoid possible collinearity problem. The \( q^* < q \) principle components are used out of \( q \) activations. We note that the principle components make its variance largest, yet may not necessarily be the ones that best explain the residuals from the OLS regression, \( \hat{u}_t \). In other words, these principle components are not “supervised” for \( y_t \) and thus for \( \hat{u}_t \). The regularization may be supervised so that the activations that are uncorrelated with \( \hat{u}_t \) can be dropped and the activations that are correlated with \( \hat{u}_t \) can be selected to increase the power of the test. Such regularization methods include the Lasso method, the PLS method, the Pretest method, the PCA-first-and-then-Lasso method, and etc. We first review the PCA method in the next subsection, and then other regularization methods in the following subsections.

4.3.1 Unsupervised Regularization of the ANN Test Using PCA

LWG (1993) found that the elements of \( \Psi_t \) in (4.11) tend to be collinear with \( x_t \) and with themselves and computation of \( \hat{W}_n \) can be tedious. Thus they conducted a test on \( q^* < q \) principal components of \( \Psi_t \) not collinear with \( x_t \), denoted \( \Psi_t^* \), and employ the equivalent test statistic (under conditional homoskedasticity) that avoids explicit computation of \( \hat{W}_n \), denoted \( T_{n}^{\text{PCA}} \)

\[
T_{n}^{\text{PCA}} \equiv nR^2 \overset{d}{\to} \chi^2 (q^*) ,
\]  

(4.14)

where \( R^2 \) is uncentered squared multiple correlation from a standard linear regression of \( \hat{\varepsilon}_t \) on \( \Psi_t^* \) and \( x_t \). This test is to determine whether or not there exists some advantage
to be gained by adding hidden units to the affine network.

It should be noted that the asymptotic equivalence of (4.13) and (4.14) holds under the conditional homoskedasticity, \( E(\varepsilon_t^*|x_t) = \sigma^2 \). Under the presence of conditional heteroskedasticity such as ARCH, \( T_n^{\text{PCA}} \) will not be \( \chi^2(q^*) \) distributed. To resolve the problem in that case, we can either use (4.13) with \( \hat{W}_n \) being estimated robust to the conditional heteroskedasticity (White 1980, Andrews 1991), or use (4.13) with the empirical null distribution of the statistic computed by a bootstrap procedure that is robust to the conditional heteroskedasticity (Wu 1986, Liu 1988).

4.3.2 Supervised Regularization of the ANN Test Using Lasso

The Lasso method is a shrinkage method which can be used as a selector of the activation functions for the ANN test. We use a penalized regression for the auxiliary model where the coefficients of \( \Psi(x_t'\gamma_j) \) are shrunken to zero if it is smaller than a particular value. The Lasso problem can be written as

\[
\hat{\beta}^{\text{Lasso}} = \arg \min_{\beta} \left\{ \sum_{t=1}^{T} \left( \hat{u}_t - x_t'\alpha - \sum_{j=1}^{q} \beta_j \Psi(x_t'\gamma_j) \right)^2 + \lambda \sum_{j=1}^{q} |\beta_j| \right\}
\]

The Lasso method uses the \( L^1 \)-penalty term \( |\beta_j| \), and it has the sparsity property such that some of the \( \beta_j \)'s that are small will be shrunken to zero, yet it does not have oracle property according to Fan and Li (2007) in the sense that it will give biased estimates of \( \beta_j \) even when sample size is large. The Lasso method is easier to implement than some other methods that has the oracle property. Since the activation functions are selected according to its explanation power to \( \hat{u}_t \), the Lasso is a supervised regularization. The tuning parameter \( \lambda \) determines the number of activation functions selected. To get the test statistic using the Lasso method, we will do the auxiliary regression of \( \hat{u}_t \) on the \( q^* \) selected activation functions \( \Psi^* \) (denoting \( q^* \)-vector of Lasso-selected activations), and
get $T_{nLasso}^L = nR_{Lasso}^2$. We choose $\lambda$ such that $q^* = 3$. In Section 4.4, we will examine if it has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from $q$ to $q^*$.

4.3.3 Supervised Regularization of the ANN Test Using PLS

Like PCA, the PLS method constructs variables using linear combinations of activation functions. Yet like Lasso, it is supervised using information about $\hat{u}_t$. The algorithm of the PLS method used in this test is described as follows:

1. Standardize each $\Psi (x'_t^j)$ to zero mean and unit variance. Set $\tilde{u}^{(0)}_t = \bar{u}_t$, $\Psi (x'_t^j)^{(0)} = \Psi (x'_t^j)$, for $j = 1, \ldots, q$, where $\iota = (1, \ldots, 1)'$.

2. For $m = 1, \ldots, q$,

   (a) Construct the linear combination, $z_m = \sum_{j=1}^q \omega_m \Psi (x'_t^j)^{(m-1)}$, where the weight is equal to the covariance between $\Psi (x'_t^j)^{(m-1)}$ and $\hat{u}_t$: $\omega_m = \text{cov} (\Psi (x'_t^j)^{(m-1)}, \hat{u}_t)$.

   (b) Regress $\hat{u}_t$ on $z_m$, and get the coefficient: $\hat{\theta}_m = \text{cov}(z_m, \hat{u}_t)/\text{var}(z_m)$.

   (c) Update $\tilde{u}^{(m)}_t$ by $\tilde{u}^{(m)}_t = \tilde{u}^{(m-1)}_t + \hat{\theta}_m z_m$.

   (d) Update $\Psi (x'_t^j)^{(m)}$ by orthogonalizing each $\Psi (x'_t^j)^{(m-1)}$ with respect to $z_m$: $\Psi (x'_t^j)^{(m)} = \Psi (x'_t^j)^{(m-1)} - \left[ \text{cov}(\Psi (x'_t^j)^{(m-1)}, z_m)/\text{var}(z_m) \right] z_m$, $j = 1, \ldots, q$.

3. The fitted value of residual terms by PLS is given by $\tilde{u}^{(m)}_t$ and the selected linear combinations of activation functions are given by $z_m$.

In this test, we select the first $q^*$ largest $z_m$, and then do auxiliary regression of $\hat{u}_t$ on $z_m$ to get the test statistic $T_{nPLS}^L = nR_{PLS}^2$. In Section 4.4, we will examine if it
has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from $q$ to $q^*$.

### 4.3.4 Supervised Regularization of the ANN test Using Pretests

The PCA shrinkage includes all the information of the activation vector $\Psi_t$, including those that are irrelevant to explain the residuals from the linear regression. We may consider to make further shrinkage from the principle components. In this section, we consider the Pretest method on the principle components, as implemented by Inoue and Kilian (2008). We first get $k = 20$ principle components from the $q$ activation vector $\Psi_t$, and then regress the residual from the OLS regression on these $k$ principle components. Then we choose $q^* = 3$ principle components corresponding to the coefficients with the highest absolute $t$-values. Then the test statistic for this Pretest method is equal to $T_{n,\text{Pretest}} = nR_{\text{Pretest}}^2$. Similarly, we will examine if it has the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from $q$ to $q^*$, in Section 4.4.

The Pretest method described here is essentially the “PCA-first-and-then-Pretest”. In the next subsection, we will consider the “PCA-first-and-then-Lasso”.

### 4.3.5 Supervised Regularization of the ANN test Using PCA-first-and-then-Lasso

Instead of using Pretest to supervise the original ANN test, we also use the Lasso method to supervise the principle components. In this subsection, we combine the PCA and the Lasso method. We first get a relatively larger number of $k$ (e.g. 100, 50, 10 or 5) principle components from the $q$-vector $\Psi_t$ of activation functions, and then use
the Lasso method to shrink them except for the $q^* = 3$ principle components. In this way, we can select the principle components that best fits the residuals from the OLS regression and increase the power of the test. We then do the auxiliary regression using the selected $q^*$ principle components and get the test statistic $T_{n}^{\text{PCA-lasso}} = nR_{\text{PCA-lasso}}^2$.

In Section 4.4, we will examine if the ANN test using this method of “PCA-first-and-then-Lasso” can still follow the asymptotic $\chi^2(q^*)$ distribution or if it is subjected to the PoSI problem due to the supervision in regularizing the dimension from $q$ to $q^*$.

4.3.6 The PoSI Problem

Regularized methods of estimation have been developed intensively in the past 20 years. Examples includes the Bridge estimator of Frank and Friedman (1993), the least absolute selection and shrinkage (Lasso) estimator of Tibshirani (1996), the least angle regression (LARS) of Efron, Hastie, Johnston, Tibshirani (2004), the smoothly clipped absolute deviation (SCAD) estimator of Fan and Li (2001), and the traditional hard-thresholding Pretest methods. It is tempting to use these supervised regularization in reducing the large number of randomized ANN activations. However, as noted in Leeb and Pötscher (2003, 2005, 2006, 2008), Pötscher and Leeb (2009), Berk et al (2011), and others, subset-searches like the Lasso shrinkage method suffer from the Post Sample Inference (PoSI) problem. See also Hoover (2012) on a related issue of size distortion resulted from model-search. In Section 4.4, we show that PLS, Pretest, PCA-first-and-then-Lasso will cause the PoSI problem that the distribution under the null hypothesis is different from the $\chi^2(q^*)$ distribution. cf. Leeb and Pötscher (2008).

To illustrate the PoSI problem, we take the Lasso supervision as an example. When using the Lasso method to select the activation functions, we are actually making
selection between the following two models:

\[ M_0 : Y = X_0 \beta_0 + v_1 \]

versus

\[ M_1 : Y = X_0 \beta_0 + X_1 \beta_1 + v_2, \]

where \( Y \) is the residual term \( \hat{u}_t \), \( \beta_0 \) and \( \beta_1 \) are vectors of parameters, \( X_0 \) and \( X_1 \) are partitions of the activation function \( \Psi_t \) and \( v_1, v_2 \) are the error terms. If the Lasso method shrinks \( \beta_1 \) to 0, then we use model \( M_0 \) to test the null hypothesis \( H_0 : \beta_0 = 0 \), and we denote the corresponding LM test statistic by \( T_{n,M_0} \); if the Lasso method does not shrink \( \beta_1 \) to 0, we pick up model \( M_1 \) and obtain the test statistic \( T_{n,M_1} \). Let \( M \) be the model selected, therefore the test statistic accounting for model selection is:

\[ T = T_{n,M_0} \times 1(M=M_0) + T_{n,M_1} \times 1(M=M_1), \]

where \( 1(\cdot) \) is the indicator function.

If \( M_0 \) is the true model, we know \( T_{n,M_0} \) follows a \( \chi^2(q_0) \) distribution with \( q_0 \) equal to \( \dim \beta_0 \); on the other hand, if \( M_1 \) is the true model, \( T_{n,M_1} \) has a \( \chi^2(q_1) \) distribution with \( q_1 \) equal to \( \dim \beta_0 + \dim \beta_1 \). In both cases, we know the exact distribution and can find the critical value. However, since we randomly draw \( \gamma_j \)'s and randomly activate \( \psi(x_t' \gamma_j), j = 1, \ldots, q \), many elements in the activation vector \( \Psi_t \) can be highly collinear and as a result the Lasso method may not distinguish the two models. Hence, even if \( M_0 \) is the true model the Lasso supervision may include some incorrect activation functions, and the distribution of the test statistic can be a mixture of two \( \chi^2 \) distributions with different degrees of freedom. To make things worse, as every time we randomly generate different sets of \( \Psi_t \), we cannot compute the probability of choosing \( M_0 \) or \( M_1 \) as the true model. This means that we cannot obtain the exact distribution.
of the test statistic and the usual $\chi^2_q^*$ critical value is invalid. This will be shown via simulation in the next section. As will be shown, the PoSI problem is especially severe with PLS and Pretest while it seems relatively mild or even negligible with Lasso.

4.4 Monte Carlo

4.4.1 DGPs and Simulation Design

To generate data we use the following DGPs, all of which have been used in the related literature. There are two blocks. All the error terms $\epsilon_t$ below are i.i.d. $N(0, 2^2)$. Two blocks of DGP are considered. The first block has DGP’s using the univariate series of $y_t$, and the second block introduces two external variables $x_{1t}$ and $x_{2t}$ which follow a bivariate normal distribution. All DGPs below fulfil the conditions for the investigated testing procedures. For those regularity conditions and moment conditions, see White (1994, Chapter 9) for the ANN tests.

Block 1 (Time-series data generating processes)

1. Autoregressive (AR)

$$y_t = 0.6y_{t-1} + \epsilon_t$$

2. Threshold autoregressive (TAR)

$$y_t = \begin{cases} 
0.9y_{t-1} + \epsilon_t & \text{if } |y_{t-1}| \leq 1 \\
-0.3y_{t-1} + \epsilon_t & \text{otherwise}
\end{cases}$$

3. Sign autoregressive (SGN)

$$y_t = \text{sgn}(y_{t-1}) + \epsilon_t$$
where

$$\text{sgn}(y_{t-1}) = \begin{cases} 
1 & \text{if } y_{t-1} > 0 \\
0 & \text{if } y_{t-1} = 0 \\
-1 & \text{otherwise}
\end{cases}$$

4. Nonlinear autoregressive (NAR)

$$y_t = \frac{0.7|y_{t-1}|}{|y_{t-1}| + 2} + \varepsilon_t$$

5. Markov regime-switching (MRS)

$$y_t = \begin{cases} 
0.6y_{t-1} + \varepsilon_t & \text{if } S_t = 0 \\
-0.5y_{t-1} + \varepsilon_t & \text{if } S_t = 1
\end{cases}$$

where $S_t$ follows a two-state Markov chain with transition probabilities $\Pr(S_t = 1|S_{t-1} = 0) = \Pr(S_t = 0|S_{t-1} = 1) = 0.3$.

**Block 2 (Cross-sectional data generating processes):**

This block includes DGPs similar to those in Zheng (1996). Assume $x_{1t}, x_{2t}$ follow a bivariate normal distribution of $N(0, 0, 1, 1, \rho)$ where the correlation $\rho = 0$ or 0.7. We have the following three cases:

1. Linear

$$y_t = 1 + x_{1t} + x_{2t} + \varepsilon_t$$

2. Cross-Product

$$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{1t}x_{2t} + \varepsilon_t$$

3. Squared

$$y_t = 1 + x_{1t} + x_{2t} + 0.2x_{2t}^2 + \varepsilon_t$$
For the simulations, the information set are $x_t = y_{t-1}$ for Block 1 and $x_t = (x_{t1} x_{t2})'$ for Block 2. The logistic squasher $\psi = [1 + \exp(-x'\gamma)]^{-1}$ is used with $\gamma$ being generated randomly from a uniform distribution on an interval depending on the data range. The number of additional hidden units to the affine network $q = 200$ is used. We set $q^* = 3$ for all regularization methods for simplicity.

4.4.2 Results

Tables 1 and 2 report the size and power for ANN test with $q = 200$ using various regularization methods (PCA, Lasso, PLS, Pretest, “PCA-first-and-then-Lasso”). The numbers in the tables are the rejection frequencies of the null hypothesis at 5% and 10% levels. The sample size $n$ is equal to 200. We use 1000 Monte Carlo replications. As demonstrated in LWG (1993) and Lee, Xi and Zhang (2012), the ANN test with PCA, that is an unsupervised regularization, exhibits good size under null hypothesis from observing the rows for AR, Linear ($\rho = 0$), Linear ($\rho = 0.7$). It also exhibits good power against a variety of nonlinear structures. In Figure 1 we plot the histograms of the test statistic under the null hypothesis. The solid line is the probability density function of $\chi_3^2$ distribution. In all three cases of AR and Linear, the finite sample distribution (histogram) of the test statistic is very close to its asymptotic $\chi_3^2$ distribution, which means the unsupervised ANN test with PCA has good size not only in 5% and 10% levels but also across the entire distribution. This demonstrates that use of unsupervised regularization for the ANN test does not lead to the PoSI problem.

To contrast, it seems that the use of supervised regularization for the ANN test do lead to the PoSI problem to some different extent depending on different method. Looking at the size in Table 1, we may see only slight over-rejections at 10% level for
Linear ($\rho = 0.7$). While the power of the supervised ANN test using Lasso are quite similar to those of the unsupervised ANN test with PCA in Block 1, it is higher in Block 2. Because Table 1 presents only the 5% and 10% quantiles in the right tail (i.e., 95% and 90% quantiles) of the null distribution of the statistic, the results of the tables do not show the difference between PCA and Lasso. However, comparing Figure 1 and Figure 2 for the entire distribution can tell some apparent difference especially in the left tail and to some lesser degree in the middle of the null distribution (but not in the right tail as shown in the tables). From Figure 2, we can see that the Lasso method suffers from the PoSI problem in the sense that the distributions of the test statistic diverge from the theoretical asymptotic $\chi^2_3$ distribution. This can be more clearly seen in the AR case in Block 1. But for the cross sectional cases in Block 2, the histograms of the test statistics are still close to the $\chi^2_3$ distribution, although they are not as good as the ones in Figure 1. Hence, it seems that the PoSI problem is relative mild or even negligible with Lasso.

For the size of the supervised ANN test using PLS, we observe from Table 1 typical over-rejections at 5% and 10% levels in all three linear cases. This clearly shows that the PoSI problem is severe for the PLS supervision, which leads to power much higher than those of the unsupervised ANN test with PCA. In Figure 3, we can see the histograms of test statistics shift out of the $\chi^2_3$ distribution, which again implies the PoSI problem.

For the PCA-first-and-then-Pretest method (in short, the Pretest method), the PoSI problem is most obvious. Table 1 shows the test results for $k = 20$, we can

---

3 At 5% level, since the $p$-value is Bernoulli distributed with success probability of 0.05, the standard error of the $p$-value from the 1000 Monte Carlo replication is $\sqrt{(0.05 \times 0.95)/1000} \approx 0.0069$. The 95% confidence interval is $0.05 \pm 1.96 \times 0.0069 = (0.0365, 0.0635)$. At 10% level, the standard error of the $p$-value is $\sqrt{(0.1 \times 0.9)/1000} = 0.0095$, and the 95% confidence interval is $0.10 \pm 1.96 \times 0.0095 = (0.0814, 0.1186)$. 4151
see that even the size under 5% and 10% is close to 1. We also tried different values of 
$k$, and the results are similar, so we do not report them in the table. Figure 4 shows
the distribution of test statistic for Pretest method with $k = 20$, which diverge heavily
from the $\chi^2_3$ distribution.

Finally, to show how different degrees of supervised regularization lead to d-
different degrees of PoSI problem, we experiment the supervised ANN test using the
PCA-first-and-then-Lasso with different values of $k$, the number of the principal com-
ponents selected by PCA in the first step of the method. The PCA-first-and-then-Lasso
method has two steps. The first step is to compute principal components of the $q = 200$
randomly activated neural network hidden units. Among them we select the first $k$
principal components. Then in the second step we select $q^* = 3$ of the $k$ principal com-
ponents. We consider $k = 3, 5, 10, 20, 50, 100, 200$. When $k = q = 200$, this method is
the same as Lasso (as presented in Figure 2), for which there is no role of the first-step in
the PCA-first-and-then-Lasso as no principal components are used. When $k = q^* = 3$,
this method is the same as PCA (as presented in Figure 1), for which there is no role
of the second-step in the PCA-first-and-then-Lasso as no Lasso is used. If $k$ is very
small, for example $k = 5$ (as presented in Figure 5a,b), this method is similar to the
unsupervised ANN test with PCA. In the other extreme, if $k$ is very large, say $k = 100$
(as presented in Figure 5e,f), then the LASSO will play a very important role but PCA
will have little effect on the test. Table 2 shows the size and power of this method using
different values of $k = 5, 10, 20, 100$. Let us first look at the size. The test behaves
reasonably good when $k$ is equal to 5 because when $k$ is small, this test is close to the
unsupervised ANN test with PCA and therefore suffers little from the PoSI problem.
But when $k$ increases to 50 and 100, we can see the over-rejection from the PoSI problem.
becomes more severe. The PoSI problem can be found in Figure 5, where we draw the histograms of test statistics for different $k$. For $k = 5$, the histograms are very close to the $\chi^2_3$ distribution. But as $k$ increases to 50 and 100, the histograms gradually shift to the right which indicates over-rejection.

When it comes to the power, the supervised ANN test using the PCA-first-and-then-Lasso method does very badly especially when $k$ is large. Table 2 shows that the power for $k = 50$ and $k = 100$ are substantially lower than the power for $k = 5$ in all cases except for MRS. When comparing with the unsupervised ANN test with PCA, this test shows inferior power in most cases. The reason for this lowered power is ascribe to how the Lasso works. In the LWG test, we choose the second to the fourth principle components which account for a large fraction of the variance of $\Psi_t$, so that they contain a lot of information and therefore can help detect the nonlinearity. But the Lasso will keep principle components with larger coefficients in the regression. Hence those principle components with large coefficients but maybe with less information can be kept; those ones with small coefficients but maybe with more information are dropped. That may be why the PCA-first-and-then-Lasso method performs poorly in power. When we increase $k$, it is more likely that the Lasso may pick up unimportant principle components and will reduce the power even more. On the other hand, if we set $k = q^*$ the Lasso to PCA test is essentially the LWG’s original ANN test, and this explains the increasing power when $k$ is very small.

4.5 Conclusions

In this paper, we applied the ANN model to test neglected nonlinearity in conditional mean. The ANN test uses the residuals from a linear model and check for their
correlation with the ANN’s hidden unit activation functions. We generated a large num-
ber of activation functions based on the randomly drawn activation parameters. The
large number of the activation functions is necessary to get good approximation of an
unknown nonlinear functional form. Then in order to avoid the collinearity problem, we
apply different regularization methods to select a moderate number of activation func-
tions. One regularization method suggested by LWG (1993) is the principle component
analysis (PCA), which is unsupervised. In this paper, we consider four supervised regu-
larization methods to select a subset of many activation functions. We show that the use
of supervised regularization such as Lasso, PLS, Pretest would lead to the post-selection
inference (PoSI) problem, while the PCA does not lead to such problem.

A way of avoiding the PoSI problem is to conduct the simultaneous inference for
all possible submodels under consideration which will make the resulting post-selection
inference valid but conservative, by using a Bonferroni-type bound as used by LWG
(1993) for PCA. As Leeb and Pötscher (2008) noted, finding the distribution of post-
selection estimates is hard and perhaps impossible. Pötscher and Leeb (2009) show
that the distribution of regularized estimators by Lasso, SCAD, and Pretest are highly
non-normal (non chi-squared in our testing set-up of this paper). Nevertheless, a valid
post-selection inference is possible via simultaneous inference as studied by Berk, Brown,
Buja, Zhang and Zhao (2011). Whether/how the simultaneous inference may be applied
for Lasso, Pretest, PLS requires further research.

We note that the PoSI “problem” is not necessarily a problem. Knowing the
PoSI problem could provide valuable information. The question is what for. The answer
is that the PoSI problem can be a measure of the possible gain by supervision, and
therefore it will be useful information for forecasting. The over-rejection in inference
due to the PoSI problem of the various supervised regularization methods shows that the null distribution of the test statistic based on the regularized (selected) randomized ANN activations can be shifted towards the right tail, especially when the Pretest method is in use. While it is a serious problem in inference, it may be a valuable information for forecasting. The degree of the PoSI problem can be translated into a measure of supervision in the regularization, i.e., a measure of the information contents for the forecast target from the variables (predictors) selected through the supervision. However, the results from Table 2 for the PCA-first-and-then-Lasso method indicates that this may not be a straightforward matter because it is shown that more supervision does not necessarily increase the power of the ANN test. It remains to be studied that it might be possible that the more supervised regularization can lead to poor forecasting performance of the ANN model. Hence, it will be interesting to examine whether the different degrees of the PoSI problem among the different regularization methods may be carried over to different degrees of improvement in forecasting ability of the ANN model. We leave this in our research agenda.
References


157


CREATES Research Paper 27.


Table 2.1 Size and Power of LWG, Lasso, PLS, and Pretest (with $q = 200$)

<table>
<thead>
<tr>
<th></th>
<th>PCA</th>
<th>Lasso</th>
<th>PLS</th>
<th>Pretest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>AR</td>
<td>0.047</td>
<td>0.102</td>
<td>0.054</td>
<td>0.098</td>
</tr>
<tr>
<td>TAR</td>
<td>0.243</td>
<td>0.373</td>
<td>0.248</td>
<td>0.354</td>
</tr>
<tr>
<td>SGN</td>
<td>0.841</td>
<td>0.914</td>
<td>0.735</td>
<td>0.829</td>
</tr>
<tr>
<td>NAR</td>
<td>0.104</td>
<td>0.183</td>
<td>0.086</td>
<td>0.238</td>
</tr>
<tr>
<td>MRS</td>
<td>0.167</td>
<td>0.259</td>
<td>0.164</td>
<td>0.344</td>
</tr>
<tr>
<td>Linear ($\rho = 0$)</td>
<td>0.043</td>
<td>0.088</td>
<td>0.052</td>
<td>0.112</td>
</tr>
<tr>
<td>Linear ($\rho = 0.7$)</td>
<td>0.043</td>
<td>0.091</td>
<td>0.057</td>
<td>0.129</td>
</tr>
<tr>
<td>Cross Product ($\rho = 0$)</td>
<td>0.075</td>
<td>0.126</td>
<td>0.216</td>
<td>0.364</td>
</tr>
<tr>
<td>Cross Product ($\rho = 0.7$)</td>
<td>0.240</td>
<td>0.362</td>
<td>0.320</td>
<td>0.456</td>
</tr>
<tr>
<td>Squared ($\rho = 0$)</td>
<td>0.178</td>
<td>0.277</td>
<td>0.219</td>
<td>0.303</td>
</tr>
<tr>
<td>Squared ($\rho = 0.7$)</td>
<td>0.220</td>
<td>0.341</td>
<td>0.267</td>
<td>0.384</td>
</tr>
</tbody>
</table>

Notes: Sample Size $n = 200$. $q = 200$. “Pretest” denotes “PCA-first-and-then-Pretest”. $k = 20$ is used for the Pretest method.
Table 2.2 Size and Power of PCA-first-and-then-Lasso with $k = 100, 50, 10, 5$

<table>
<thead>
<tr>
<th></th>
<th>$k = 100$</th>
<th>$k = 50$</th>
<th>$k = 10$</th>
<th>$k = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>AR</td>
<td>0.085</td>
<td>0.158</td>
<td>0.078</td>
<td>0.142</td>
</tr>
<tr>
<td>TAR</td>
<td>0.126</td>
<td>0.204</td>
<td>0.125</td>
<td>0.206</td>
</tr>
<tr>
<td>SGN</td>
<td>0.204</td>
<td>0.262</td>
<td>0.226</td>
<td>0.287</td>
</tr>
<tr>
<td>NAR</td>
<td>0.089</td>
<td>0.161</td>
<td>0.096</td>
<td>0.165</td>
</tr>
<tr>
<td>MRS</td>
<td>0.190</td>
<td>0.267</td>
<td>0.186</td>
<td>0.280</td>
</tr>
<tr>
<td>Linear ($\rho = 0$)</td>
<td>0.096</td>
<td>0.183</td>
<td>0.067</td>
<td>0.121</td>
</tr>
<tr>
<td>Linear ($\rho = 0.7$)</td>
<td>0.097</td>
<td>0.178</td>
<td>0.065</td>
<td>0.117</td>
</tr>
<tr>
<td>Cross Product ($\rho = 0$)</td>
<td>0.109</td>
<td>0.183</td>
<td>0.096</td>
<td>0.154</td>
</tr>
<tr>
<td>Cross Product ($\rho = 0.7$)</td>
<td>0.114</td>
<td>0.199</td>
<td>0.100</td>
<td>0.172</td>
</tr>
<tr>
<td>Squared ($\rho = 0$)</td>
<td>0.108</td>
<td>0.187</td>
<td>0.078</td>
<td>0.168</td>
</tr>
<tr>
<td>Squared ($\rho = 0.7$)</td>
<td>0.110</td>
<td>0.196</td>
<td>0.082</td>
<td>0.139</td>
</tr>
</tbody>
</table>

Notes: Sample Size is $n = 200. \ q = 200.$
Figure 4.1 Distribution of $T_n^{PCA}$ under $H_0$

(a) DGP: AR compared to $\chi^2_3$

(b) DGP: Linear($\rho = 0$) compared to $\chi^2_3$

(c) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$
Figure 4.2 Distribution of $T_{n}^{\text{Lasso}}$ under $H_0$

(a) DGP: AR compared to $\chi_3^2$

(b) DGP: Linear ($\rho = 0$) compared to $\chi_3^2$

(c) DGP: Linear ($\rho = 0.7$) compared to $\chi_3^2$
Figure 4.3 Distribution of $T_n^{\text{PLS}}$ under $H_0$

(a) DGP: AR compared to $\chi^2_3$

(b) DGP: Linear($\rho = 0$) compared to $\chi^2_3$

(c) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$
Figure 4.4 Distribution of $T_{n}^{\text{Pretest}}$ (PCA-first-and-then-Pretest) under $H_0$

(a) DGP: AR compared to $\chi^2_3$

(b) DGP: Linear($\rho = 0$) compared to $\chi^2_3$

(c) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$
Figure 4.5 Distribution of $T_{n}^{\text{PCA-Lasso}}$ (PCA-first-and-then-Lasso) under $H_0$

(a) DGP: AR compared to $\chi^2_3$

(b) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$

(c) DGP: AR compared to $\chi^2_3$

(d) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$

(e) DGP: AR compared to $\chi^2_3$

(f) DGP: Linear($\rho = 0.7$) compared to $\chi^2_3$