MANCHESTER 1824

The University of Manchester

Economics Discussion Paper

EDP-0641

Does Spread Really Predict the Short Rate? Explaining Empirical Anomalies in the Expectations Theory

By

Erdenebat Bataa, Dong H. Kim, Denise R. Osborn

December 2006

Download paper from:

http://www.socialsciences.manchester.ac.uk/economics/research/discuss.htm

Economics
School of Social Sciences
The University of Manchester
Oxford Road
Manchester M13 9PL
United Kingdom

Does Spread Really Predict the Short Rate? Explaining Empirical Anomalies in the Expectations Theory

Erdenebat Bataa

Centre for Growth and Business Cycles Research, Economics, University of Manchester

Dong H. Kim

Dept. of Economics, Korea University

Denise R. Osborn*

Centre for Growth and Business Cycles Research, Economics, University of Manchester

December 2006

* Corresponding author. Denise Osborn, Centre for Growth and Business Cycles Research, Economics, School of Social Sciences, University of Manchester, Oxford Road, Manchester, UK, M13 9PL. Email: denise.osborn@manchester.ac.uk, Tel: +44(0)161-275-4861. Details of other authors: Erdenebat Bataa, Email: e.bataa@manchester.ac.uk, Tel: +44(0)161-275-4860, Dong Heon Kim, Email: dongkim@korea.ac.kr, Tel: +82-2-3290-2226.

Acknowledgements: The authors would like to thank Chris Orme, Simon Peters and Econometrics seminar participants at the University of Manchester for their useful comments. Erdenebat Bataa would also like to acknowledge financial assistance from the University of Manchester and the Open Society Institute.

Abstract

Empirical studies often find that the spread between longer and shorter rates does not have predictive power for future longer rates, violating the Expectations Theory (ET). Although the predictive power of the spread for future shorter rates is largely in accordance with the ET, especially when the forecast period is long, researchers often find this holds to varying degrees across samples (country-wise or time-wise). We show this pattern may be due to the powers of all tests depending on interest rates' maturities and their persistency in small samples. This paper also compares the powers of tests of the ET against the under/over-reaction and the time varying term premium alternatives across various maturity combinations, levels of persistency and sample sizes. Tests perform best and are comparable to each other at the shortest end of the term structure, but deteriorate as the distance between maturities of longer and shorter rates increase. However, this deterioration is of varying degrees for different tests and its speed diminishes as we depart from the shortest end. In general Lagrange multiplier and distance metric tests emerge as being the most powerful and least sensitive to interest rate maturities and their persistency.

JEL classification: G10; E43.

Keywords: expectations hypothesis; term structure of interest rates; vector autoregression

"*…The simple expectations theory, in combination with the hypothesis of rational expectations, has been rejected many times in careful econometric studies. But the theory seems to reappear perennially in policy discussions as if nothing had happened to it…We are reminded of the Tom and Jerry cartoons that precede feature films at movie theatres. The villain, Tom the cat, may be buried under a ton of boulders, blasted through a brick wall (leaving a cat shaped hole), or flattened by a steamroller. Yet seconds later he is up again plotting his evil deeds*" [1].

1. Introduction

Economists and investors believe that a better understanding of the relationship between interest rates of various maturities leads to better decision making. One of the most important theories of this relationship is the expectations theory (ET), according to which investing in a succession of short-term bonds gives the same expected return as investing in a long-term bond, when adjustment is made for the assumed constant term premium[2].

The empirical literature on this theory is huge, yet there is little sign that research interest in this topic has waned, as the theory is constantly being subjected to scrutiny using new datasets and new methodologies. One of the most puzzling results, reported as early as Macaulay (1938), is that two main implications of the theory lead to different conclusions, which Campbell and Shiller (1991) describe "*…the slope of the term structure almost always gives a forecast in the wrong direction for the short term change in the longer bond, but gives a forecast in the right direction for long term changes in short rates*". The former is typically statistically different from the ET forecast. Although Stambaugh (1988) notes that the regression used to test the first implication is very sensitive to measurement error in the long term interest rate, Campbell and Shiller (1991) show that these rejections are quite robust even if this is correctly accounted for using instrumental variables.

The second implication, relating to predictions of future short rates, is rejected in Campbell and Shiller (1991) only when the longer rate used to compute the spread is less than 36 months, so that the rejections in this case cluster at the short end of the term structure, Indeed this implication is tested much more frequently, leading to rejections at the short end of the term structure, e.g. in Shiller, *et al.* (1983), Mankiw and Summers (1984) and Evans and Lewis (1994). This pattern is also present in more recent studies by

[1] Shiller, Campbell and Schoenholtz (1983), pp 174-175.
[2] Another implication of the EH is that the forward interest rate must equal the expected spot rate. This implication is the subject of studies by Fama and Bliss (1987), Backus, Foresi, Mozumdar and Wu (2001), and Fama (2006), among others and will not discussed in the present paper.

Sarno, Thornton and Valente (2006) and Bataa, Kim and Osborn (2006), but is not universally accepted. For example, Longstaff (2000) does not reject the ET at the very short end using high frequency data while Taylor (1992) finds very strong evidence against it at the long end. Moreover, its performance seems to differ across countries and sample periods. Hardouvellis (1994) finds that, among the G7 countries, it is strongly rejected only for the US while Gerlach and Smets (1997) extend this conclusion using Eurocurrency rates in 17 countries. However, the evidence for Germany seems controversial. Jondeau and Ricart (1999) reject the implication for Germany and the US but not for France and the UK and Bekaert and Hodrick (2001) also reject the null in Germany and the US but not in the UK using the extended data of Gerlach and Smets (1997). In contrast, Cuthbertson, Hayes and Nitzsche (2000) and Boero and Torricelli (2002) use estimated German term structure data and provide supportive evidence for the theory. Country-specific studies such as Dahlquist and Jonnson (1995) for Sweden, Engsted (1996) for Denmark, Cuthbertson (1996) for the UK and Cuthbertson and Bredin (2001) for Ireland also support the theory.

Several alternatives have been proposed to explain these anomalies, which include time varying term premia, the overreaction hypothesis, monetary policy regime change and the finite sample properties of different tests. Shiller *et al*. (1983) conclude: "…*Variations in risk premiums are so large as to destroy any information in the term structure about future interest rates*". An unobserved time-varying term premium is modelled in various ways: using levels of interest rates, yield spreads, and unemployment rates (Shiller, 1979; Mankiw and Summers, 1984), using second moments of explanatory variables (Engle Lillien and Robins, 1987; Engle and Ng, 1993), employing panal data method (Harris, 2001) and as a difference between the actual and the theoretical spread derived under the ET (Carriero, Favero and Kaminska, 2006). Tzavalias and Wickens (1997) argue the two empirical implications are in accordance with the theory once a time-varying term premium that is correlated with the spread is allowed

However, Campbell and Shiller (1991) and Hardouvellis (1994) assert that markets overreact to monetary policy announcements, changing their expectations of future spot rates by more than is warranted, and this explains the contradictory test results on the two theory implications. Mankiw and Miron (1986) attribute the poor performance of the EH over certain periods to the monetary policy pursued by the US Fed, with the EH performing better in periods of monetary targeting than in periods of interest rate targeting (and even better before the foundation of the Fed). Supporting evidence on this conjecture is found in Kugler (1988), Hardouvellis (1988) and Simon (1990). Rudebusch (1995),

Roberds, Runkle and Whiteman (1996), Fuhrer (1996) and Balduzzi, Bertola and Foresi (1997) attempt to reconcile the ET with data, with partial success, by explicitly modelling Fed behaviour in the process governing the short term interest rate. More recently, Kozicki and Tinslay (2005) stress the importance of imperfect policy credibility of the Central Bank on the performance of the theory.

Finally, there is a strand of literature that suggests the tests themselves may lead to false rejections in finite samples. Early studies of this possibility consider either a single equation test, or employ a VAR as the data generating process (DGP) and test implications imposed by the ET on the VAR parameters using Campbell and Shiller's (1987) Wald test. Bekaert, Hodrick and Marshall (1997, 2001) document that the finite sample distributions of the test statistics under the null, including those from the single equation test, can be quite different from their asymptotic counterparts in the presence of highly persistent short rates and "peso problems". Using survey data, Froot (1989) finds that the rejections in the single equation test can be due to the rational expectations hypothesis, not necessarily due to the theory itself. Shea (1992) illustrates that the Wald test can lead to different conclusions depending on how one specifies the null, while Bekaert and Hodrick (2001) document its extreme size distortion and suggest an LM test. On the other hand, Bekaert, Wei and Xing (2006) and Sarno *et al.* (2006) include more macroeconomic and financial variables into the VAR as conditioning information and obtain more uniform rejections of the theory across the maturity spectrum.

The primary goal of this paper is to explore in more depth the finite sample properties of the tests, extending previous analyses by considering well specified alternative hypotheses against the null of the ET, and reconciling the above mentioned contradictory results across the maturity spectrum and/or samples (country-wise or time-wise). We start from Campbell and Shiller's (1991) comment on the previous literature: *"…Different studies use different econometric methods, test different implications of the expectations theory, and look at different interest rates"* to which Driffil, Psaradakis and Sola (1997) add "…*different studies also use data drawn from different places and periods of time*". Less importantly, studies use samples of different sizes. We take sample sizes typically used in the literature and ask: Suppose the ET is either true or false, but uniformly so across interest rate maturity pairs and/or periods. Would we get different results from the various econometric methods on testing the second implication, which produced most contradictory results, *ceteris paribus*? If the tests are sensitive to interest rate maturities and persistency, the (non)rejection pattern across the term structure maturity spectrum and/or different samples is not necessarily due to the theory itself.

Moreover, given the extensive empirical literature on the topic, there is a surprising lack of papers investigating the properties of various tests, proposed for the ET, in a unified framework. Not only is it of interest to know how tests perform when interest rates of different maturities and persistency are used in order to assess previous findings, but also to identify the most robust ones to be recommended for use in future research. Both asymptotic and wild-bootstrap finite sample versions of all tests are examined, following the recommendations of Horowitz and Savin (2000) and Horowitz (2001). As well as considering the conventional single equation test and more novel VAR based tests, we propose "new" *t* statistic forms of the implied regression slope and variance ratio tests, building on Campbell and Shiller (1987) and Bekaert and Hodrick (2001). The VAR based tests include a Wald (*W* hereafter) test proposed by Campbell and Shiller (1987), and a Lagrange Multiplier (*LM*) and Distance Metric (*DM*) tests of Bekaert and Hodrick (2001), the latter of which is equivalent to the Likelihood Ration (*LR*) test of Sargent (1979) under normality. Bekaert and Hodrick regression tests (2001) argue that the *LM* test is superior to the *DM* and *W* tests and is fast gaining popularity, being used in Bekaert, Wei and Xing (2006), Sarno *et al.* (2006) and Bataa *et al.* (2006), among others.

The paper is organized as follows. Section 2 explains the ET and conventional ways to test it, followed in Section 3 by the classic trinity of *LM*, *DM* and *W* tests as used for ET testing in a VAR framework. Section 4 then develops the regression and variance ratio test statistics. The main results of the paper are contained in Section 5, which details our Monte Carlo study. Section 6 concludes.

2. The ET and Conventional Test

Most modern asset pricing theories that admit no arbitrage opportunity deliver the following general relationship between long and short rates[3]:

$$R_{n,t} = \frac{1}{k}\sum_{i=0}^{k-1} E(R_{m,t+mi}\left|\Xi_t\right.) + \pi_{(n,m),t} \ , \qquad (1)$$

where $R_{n,t}$ and $R_{m,t}$ are long and short rates at time *t*, respectively, $E(R_{m,t+mi}\left|\Xi_t\right.)$ is the mathematical expectation of the short rates at $t+mi$, $i = 0, 1, 2,\ldots, k\text{-}1$, formed at time *t* conditional of the information set available to the market, Ξ_t. Here $k=n/m$ is the maturity

[3] See e.g. Shiller (1979), Kozicky and Tinsley (2001) and Bekaert and Hodrick (2001).

multiple defined for simplicity to be an integer, m is the maturity of a shorter rate and n is the maturity of a longer rate; $\pi_{(n,m),t}$ is a term premium and if it is constant (1) represents the ET and if it is zero then we have the simple or pure ET, PET. Following Shiller (1982) and Melino (2001) we ignore the constant term premium in the following discussion as this drops out of (1) if the data are demeaned.

The ET in (1) is rarely tested directly, probably due to most empirical results concluding the series are integrated, in which case conventional statistical theory is not appropriate. Rather, another implication of the ET is usually tested, which is based on the ability of the spread between long and short rates to predict future short rate changes,

$$\sum_{i=1}^{k-1}(1-\frac{i}{k})E(\Delta_m R_{m,t+mi}|\Xi_t) = S_{(n,m),t}\,, \qquad (2)$$

where $\Delta_m R_{m,t+m} = R_{m,t+m} - R_{m,t}$ and $S_{(n,m),t} = R_{n,t} - R_{m,t}$, which is obtained by subtracting $R_{m,t}$ from both sides of (1). Equation (2) implies the current spread predicts a cumulative change in the shorter term (m-period) interest rate over n periods. If one assumes rational expectations, so that

$$E(R_{m,t+mi}|\Xi_t) = R_{m,t+mi} + v_{t+mi}\,,$$

where v_{t+mi} has zero mean and is orthogonal to the information set available at time t, probably the most commonly tested equation of the ET is obtained, namely

$$\sum_{i=1}^{k-1}(1-\frac{i}{k})\Delta_m R_{m,t+mi} = \gamma + \beta S_{(n,m),t} + w_{(n,m),t} \qquad (3)$$

where $w_{(n,m),t}$ is a moving average process of order (n-m) and under the null hypothesis of the ET, β should be unity. [4]

However, there are several econometric difficulties with the conventional regression approach applied to (3). Firstly, it is inefficient as we lose n-m observations at the end of the sample period. For example, in studies of Campbell and Shiller (1991), Sarno *et al.* (2006) and Bataa *et al.* (2006) that use monthly data, n is as large as 120, i.e.

[4] Another implication of (1), that is less empirically supported and therefore called a "contrarian" test in Thornton (2006), is that the yield spread predicts the m-period change in the longer- term yield, which is tested (see e.g. Campbell and Shiller 1991) using $R_{n-m,t+m} - R_{n,t} = \gamma + \alpha\frac{m}{n-m}S_{(n,m),t} + v_t$; under the null α is unity.

the long rate maturity is 10 years. Secondly, using the realized returns as a proxy for expected returns is at best problematic. Elton (1999) strongly argues against such an approach, which implies if the test rejects the null it is impossible to distinguish if it is due to failure of the way rational expectations are handled or the ET itself. Even if rational expectations are correctly dealt with, the error term $w_{(n,m),t}$, is a MA(n-m), so standard errors have to be corrected, for example using the method described in Hansen and Hodrick (1980), or Newey and West (1987). But as Richardson and Stock (1991) and Hodrick (1992) illustrate, these adjustments do not work well when n-m is not small relative to the sample size. Thirdly, as discussed in Mankiw and Shapiro (1986) and Campbell, Lo and MacKinlay (1997) the regressor is serially correlated and correlated with lags of the dependent variable, and this can cause finite sample problems as well.

3. Testing the ET in a VAR framework

Recent work has focussed on testing the second implication of the ET in a VAR framework. In this section we outline this approach, first in terms of the relevant asymptotic distributions, before considering inference using empirical finite sample distributions.

3.1 Asymptotic distributions

Probably the biggest problem in the single equation framework is deriving the market expectation, $E(R_{m,t+mi}|\Xi_t)$. If we assume expectations are formed linearly and the information set available to the market, Ξ_t, can be proxied by some observable set I_t, $\Xi_t \supset I_t$, then the aforementioned problems may be avoided using a VAR framework. The idea is old and can be traced back to at least Sargent (1979). As in Campbell and Shiller (1991) we consider a stationary vector stochastic process for $\mathbf{y}_t = \left[\Delta R_{m,t}, S_{(n,m),t}\right]'$.[5] Assuming the process for $\mathbf{y}_t$ is represented by a demeaned VAR of order p with error covariance matrix $\mathbf{\Sigma} = E(\mathbf{u}_t\mathbf{u}_t')$,

[5] This specification can be interpreted as an assumption that interest rates are nonstationary, specifically $I(1)$, and hence, as demonstrated in Hall, Granger and Anderson (1992), both VAR variables are stationary according to the ET. However, even if interest rates are stationary, they are highly persistent and for the finite sample sizes typically used for analysis the reduction of this persistence by differencing is advantageous for VAR modelling.

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{y}_{t-i} + \mathbf{u}_t, \tag{4}$$

it can be written as a first order VAR in companion form such that $\mathbf{z}_t = \mathbf{\Phi}\mathbf{z}_{t-1} + \mathbf{v}_t$, where the companion matrix $\mathbf{\Phi}$ is of dimension $2p \times 2p$:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \cdots & \mathbf{\Phi}_{p-1} & \mathbf{\Phi}_p \\ \mathbf{I}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_2 & \mathbf{0} \end{bmatrix}$$

while $\mathbf{z}_t$ has $2p$ elements, $\mathbf{z}_t = \left[\mathbf{y}_t', \mathbf{y}_{t-1}', ..., \mathbf{y}_{t-p+1}'\right]'$, $\mathbf{v}_t$ is a $2p$ vector equal to $\left[\mathbf{u}_t', 0, 0, ..., 0\right]'$ which is uncorrelated over time. Thus $\mathbf{z}_t$ summarizes the whole history of $\mathbf{y}_t$.

Now define vectors $\mathbf{e}_i$, $i = 1, 2$; each of dimension $2p$, with unity in the i^{th} position and zeros everywhere else such that

$$\Delta R_{m,t} = \mathbf{e}_1' \mathbf{z}_t \text{ and } S_{(n,m),t} = \mathbf{e}_2' \mathbf{z}_t. \tag{5}$$

Using the ET in (2), the spread between long and short rates, which will be referred to as the theoretical spread henceforth, is[6]

$$S^*_{(n,m),t} \equiv \mathbf{e}_1' \mathbf{\Lambda} \mathbf{z}_t, \tag{6}$$

where $\mathbf{\Lambda} \equiv \mathbf{\Lambda}(\mathbf{\Phi}, n, m) \equiv \mathbf{\Phi}[\mathbf{I} - m/n(\mathbf{I} - \mathbf{\Phi}^n)(\mathbf{I} - \mathbf{\Phi}^m)^{-1}](\mathbf{I} - \mathbf{\Phi})^{-1}$. If the ET is true, the expected spread must be equal to the theoretical spread, and this equality holds when the nonlinear restrictions on the VAR parameters

$$\mathbf{e}_2' = \mathbf{e}_1' \mathbf{\Lambda} \tag{7}$$

are valid.

Applying a Wald test to the restrictions in (7), Campbell and Shiller (1987) and Shea (1992) find overwhelming evidence against the ET in contrast to Sargent (1979) and Melino (2001) whose LR test did not reject it. Recently, Bekaert and Hodrick (2001) suggest LM and DM tests using the GMM hypothesis testing framework of Newey and McFadden (1994) and find the LM test has better small sample properties than Wald and

[6] The derivation is provided in Appendix A.

DM tests, the latter of which is equivalent to the LR test in terms of size and power when joint normality holds. Since the Bekaert and Hodrick (2001) methodology is relatively new and general enough to accommodate other tests, it is summarised here with the suggested extensions of Bataa *et al.* (2006).

The Generalized Method of Moments (GMM) estimator of Hansen (1982) is used to estimate the VAR in (4). Defining $\dot{\mathbf{\Phi}} = [\mathbf{\Phi}_1,...,\mathbf{\Phi}_p]'$, the vector of nonlinear orthogonality conditions can be written as $E[\mathbf{g}(\mathbf{x}_t,\boldsymbol{\theta})] = \mathbf{0}$, where $\mathbf{x}_t \equiv (\mathbf{y}_t', \mathbf{z}_{t-1}')'$, $\boldsymbol{\theta} = vecr(\dot{\mathbf{\Phi}})$. Estimation uses the corresponding sample moment conditions for a sample of size *T*, namely $\mathbf{g}_T(\boldsymbol{\theta}) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(\mathbf{x}_t,\boldsymbol{\theta})$.

It proceeds by selecting $\boldsymbol{\theta}$ to minimize the GMM criterion function

$$J_T(\boldsymbol{\theta}) \equiv -\frac{1}{2}\mathbf{g}_T(\boldsymbol{\theta})'\hat{\mathbf{\Omega}}_T^{-1}\mathbf{g}_T(\boldsymbol{\theta}),$$

where, assuming the VAR of (4) is correctly specified with $\mathbf{u}_t$ uncorrelated, the weighting matrix, $\hat{\mathbf{\Omega}}_T^{-1}$, is a consistent estimate of the inverse of

$$\mathbf{\Omega} \equiv E[\mathbf{g}(\mathbf{x}_t,\boldsymbol{\theta})\mathbf{g}(\mathbf{x}_t,\boldsymbol{\theta})']. \tag{8}$$

If we denote the Jacobian matrix as $\mathbf{G} \equiv E\nabla_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{x},\boldsymbol{\theta}_0)$, where $\nabla_{\boldsymbol{\theta}}$ denotes derivative with respect to $\boldsymbol{\theta}$, then the GMM asymptotic distribution theory guarantees

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0},\mathbf{B}^{-1}), \tag{9}$$

where $\hat{\boldsymbol{\theta}}_T$ is the GMM estimator obtained using *T* observations, $\boldsymbol{\theta}_0$ is the corresponding true value, $\xrightarrow{d}$ denotes convergence in distribution and $\mathbf{B} \equiv \mathbf{G}'\mathbf{\Omega}^{-1}\mathbf{G}$.

The null hypothesis of (7) can be written as:

$$H_0 : \mathbf{a}(\boldsymbol{\theta}_0) \equiv \mathbf{e}_2' - \mathbf{e}_1'\mathbf{\Lambda} = \mathbf{0}, \tag{10}$$

where $\mathbf{a}(\boldsymbol{\theta}_0)$ is a $2p$ dimensional vector, with $\mathbf{A} \equiv \nabla_{\boldsymbol{\theta}}\mathbf{a}(\boldsymbol{\theta}_0)$. Notice that the null is composite, i.e. it does not fully specify the data generating process (DGP). For example, it restricts only $2p$ out of $4p$ VAR slope parameters and does not say anything about the

conditional covariance matrix of the VAR residuals. The Lagrangian for the constrained GMM maximization problem is

$$L(\boldsymbol{\theta},\boldsymbol{\gamma}) = -\frac{1}{2}\mathbf{g}_T(\boldsymbol{\theta})'\hat{\boldsymbol{\Omega}}_T^{-1}\mathbf{g}_T(\boldsymbol{\theta}) - \mathbf{a}(\boldsymbol{\theta})'\boldsymbol{\gamma};$$

where $\boldsymbol{\gamma}$ is a vector of Lagrange multipliers, and $\hat{\boldsymbol{\Omega}}_T$ is a consistent estimate of $\boldsymbol{\Omega}$ obtained from (8) using the sample mean in place of the expectation. The first order conditions for the solution of this problem are

$$\begin{bmatrix} -\sqrt{T}\mathbf{G}_T'\boldsymbol{\Omega}_T^{-1}\mathbf{g}_T(\bar{\boldsymbol{\theta}}_T) - \nabla_{\boldsymbol{\theta}}\mathbf{a}(\bar{\boldsymbol{\theta}}_T)\sqrt{T}\bar{\boldsymbol{\gamma}}_T \\ -\sqrt{T}\mathbf{a}(\bar{\boldsymbol{\theta}}_T) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and, under the null, the constrained GMM estimator satisfies $\bar{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$, so $\nabla_{\boldsymbol{\theta}}\mathbf{a}(\bar{\boldsymbol{\theta}}_T) \xrightarrow{p} \mathbf{A}$, where $\xrightarrow{p}$ denotes convergence in probability. Then Newey and McFadden (1994) show that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \bar{\boldsymbol{\theta}}_T) \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1}\mathbf{A}\mathbf{B}^{-1}). \tag{11}$$

The Wald statistic used to test (10) is based on deviations of the unconstrained estimates from values consistent under the null. On the other hand, the LM or score statistic is based on deviations of the constrained estimates from values solving the unconstrained problem. Finally, the DM statistic is based on the difference between the GMM objective functions at the constrained and unconstrained estimators using the same weighting matrix. Specifically,

$$W = T\mathbf{a}(\hat{\boldsymbol{\theta}}_T)'(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1}\mathbf{a}(\hat{\boldsymbol{\theta}}_T) \xrightarrow{d} \chi^2(2p) \tag{12}$$

$$LM = T\bar{\boldsymbol{\gamma}}_T'\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'\bar{\boldsymbol{\gamma}}_T \xrightarrow{d} \chi^2(2p) \tag{13}$$

$$DM = -2T(J(\bar{\boldsymbol{\theta}}_T) - J(\hat{\boldsymbol{\theta}}_T)) \xrightarrow{d} \chi^2(2p), \tag{14}$$

where p is the VAR lag length.

Newey and McFadden (1994) show how to obtain values for the above trinity of statistics starting from any initial $\sqrt{T}$ consistent estimator $\tilde{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}_0$. We evaluate **B** and **A** at $\tilde{\boldsymbol{\theta}}_T$, and the constrained estimator is obtained from the Lagrangian first order conditions,

$$\begin{bmatrix} \tilde{\tilde{\boldsymbol{\theta}}}_T \\ \tilde{\tilde{\boldsymbol{\gamma}}}_T \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\theta}}_T \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{A} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{G}_T' \hat{\boldsymbol{\Omega}}_T^{-1} \mathbf{g}_T(\tilde{\boldsymbol{\theta}}_T) \\ -\mathbf{a}(\tilde{\boldsymbol{\theta}}_T) \end{bmatrix}, \tag{15}$$

using $\boldsymbol{\gamma} = \mathbf{0}$ as the initial consistent estimator of the Lagrangian multipliers. Newey and McFadden (2001) note $\mathbf{a}(\tilde{\tilde{\boldsymbol{\theta}}}_T)$ is not necessarily zero in finite samples and Bekaert and Hodrick (2001) suggest to further iterate on (15), i.e., $\tilde{\tilde{\boldsymbol{\theta}}}_T$ obtained from (15) is put back on the right hand side and iterated until $\mathbf{a}(\tilde{\tilde{\boldsymbol{\theta}}}_T) = \mathbf{0}$.[7]

3.2 Bootstrap inference

Bekaert and Hodrick (2001) also provide finite sample versions of their tests using a bootstrap procedure. According to Horowitz (2001), the bootstrap should only be used if the asymptotic distribution of a test statistic is pivotal, i.e. does not depend on unknown population parameters. Although the asymptotic distributions in (12) to (14) depend on the unknown lag length of the assumed VAR data generating process (DGP), Bekaert and Hodrick (2001) apparently rely on the consistency of the SIC to select that order. The VAR parameters, estimated subject to the constraint in (10), and a bootstrap of the corresponding residuals are used as the DGP to estimate the finite sample distributions of the test statistics.

Although Bekaert and Hodrick (2001) use an *iid* bootstrap or assume a GARCH model for the VAR residuals, we use a recursive design wild bootstrap that has been shown to deal better with general forms of volatility clustering.[8] For the estimated constrained VAR parameters $\overline{\boldsymbol{\Phi}}_1, ..., \overline{\boldsymbol{\Phi}}_p$ and corresponding residual vector $\overline{\mathbf{u}}_t$ for time period *t*, we generate a bootstrap sample as

$$\mathbf{y}_t^* = \sum_{i=1}^{p} \overline{\boldsymbol{\Phi}}_i \mathbf{y}_{t-i}^* + \mathbf{u}_t^*, \ \mathbf{u}_t^* = \omega_t \overline{\mathbf{u}}_t, \ t = 1, \ldots, T, \tag{16}$$

in which the scalar random variable ω_t follows the Rademacher distribution, taking the possible values of negative and positive unity with equal probabilities.[9] This choice is justified by recent Monte Carlo studies of Davidson and Flachaire (2001), Godfrey and Orme (2004) and Godfrey and Tremayne (2005). For each of a large number of data sets

[7] In our application the tolerance level for convergence is set at 10^{-8}.
[8] See Goncalves and Killian (2004) and discussions in Bataa *et al*. (2006).
[9] Bataa *et. al* (2006) follow Stine (1987) in randomizing the starting values. They split the observed data into $T - p + 1$ overlapping blocks of length p and one of these is selected randomly as the starting point.

generated in this way, we estimate (12) to (14) to derive the empirical finite sample distributions of the test statistics.

4. Mixed testing approach

Bekaert and Hodrick (2001) note that the inferential efficiency of the single equation method can be improved by considering implications of the VAR parameters for the slope coefficient of regression (3). Sarno *et al.* (2006) and Bekaert *et al.* (2006) extend this idea to the variance ratio of the theoretical and actual long rates. We term these a mixed approach. This section first describes how previous studies use this mixed approach make inference and then argues such inference is invalid because the null hypothesis does not fully specify the assumed DGP. The second subsection then develops the studentized versions that avoid this problem.

4.1 Slope coefficient and variance ratio tests

The basic idea of the mixed approach is to generate empirical distributions of the implied OLS slope coefficient for equation (3) and the variance ratio statistic under the null, using a large number of datasets generated from (16) with an *iid* bootstrap, which then allow computation of empirical *p*-values for the test statistics obtained from real data.

If the ET is true, the population slope coefficient from a regression of the actual spread on the theoretical spread must be unity. Therefore, from (5) and (6), the implied slope coefficient is

$$\beta(\boldsymbol{\theta}) = \frac{\mathbf{e}_1'\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{e}_2}{\mathbf{e}_2'\boldsymbol{\Psi}\mathbf{e}_2}, \tag{17}$$

namely the covariance between the dependent and independent variables in (3) divided by the variance of the dependent variable, where $vec(\boldsymbol{\Psi}) = (\mathbf{I} - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi})^{-1} vec(\boldsymbol{\Sigma})$. Similarly, the variance ratio of the theoretical and actual spreads can also expressed in terms of the VAR parameters as[10]

[10] Campbell and Shiller (1991) use the concept to evaluate the economic significance of the ET, as this ratio should be close to one if the ET is true, but Sarno *et al.* (2006) and Bekaert *et al.* (2006) provide the compact expression for the variance ratio of the theoretical and actual long rate rates, as their DGP includes interest rates in levels.

$$v(\boldsymbol{\theta}) = \frac{\mathbf{e}_1'\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'\mathbf{e}_1}{\mathbf{e}_2'\boldsymbol{\Psi}\mathbf{e}_2}. \tag{18}$$

However, it is important to note that the estimated DGP in (16) is only one of possibly many DGPs under the null. This implies applying a simple bootstrap procedure to either (17) or (18) to conduct inference is problematic, because the ET constitutes a composite null hypothesis in terms of the assumed DGP. For the ET, only 2*p* restrictions are implied on the 4*p* elements of the VAR coefficient matrices, so that there are potentially many DGPs that could generate the data under the null. Therefore, as discussed by Horowitz and Savin (2000), valid inference requires the use of size-corrected critical values. These set the supremum of the test's rejection probability over all admissable DGPs under H_0 to a pre-specified level α, rather than taking the critical value estimated as the α level quantile from a specific and arbitrary finite sample distribution. Estimating the appropriate size-corrected critical value will obviously be very expensive, if not impossible, due to the need to consider all DGPs admissible under the composite null hypothesis. Perhaps more importantly, this critical value may be infinite, leading to a test with no power if the number of possible DGPs under H_o is large (Dufour, 1997), or the power of the test may be the same as the size (Bahadur and Savage, 1956).

4.2 Studentized slope and variance ratio tests

As just argued, finite sample bootstrap critical values obtained using (17) and (18) are not valid for inference on the ET at the specified level of significance α. Indeed this may be the reason why Bekaert *et al.* (2006) and Sarno *et al.* (2006) find somewhat contradictory results from the *LM* test and the implied test statistics of (17) and (18). However, if the ET is true, the population values of these statistics must equal unity. That is, the ET in (10) implies the null hypotheses

$$\mathrm{H}_0\colon\ \beta(\boldsymbol{\theta}_0) = v(\boldsymbol{\theta}_0) = 1 \tag{19}$$

This, in turn, suggests bootstrapping simple *t* statistics, which are asymptotically pivotal, associated with the null that (17) and (18) are equal to unity, rather than bootstrapping the implied slope coefficient and variance ratio directly.[11] This leads to the new tests we propose.

First note that $\beta(\bar{\boldsymbol{\theta}}_T) = v(\bar{\boldsymbol{\theta}}_T) = 1$, then from (11) and a Taylor's series approximation, the asymptotic distributions are

[11] See Hall (1994) and Horowitz (2001) for the importance of bootstrapping asymptotically pivotal statistics.

$$\sqrt{T}(\beta(\hat{\boldsymbol{\theta}}_T)-1) \xrightarrow{d} N(0, \mathbf{H}'\mathbf{B}^{-1/2}(\mathbf{I}-\mathbf{M})\mathbf{B}^{-1/2}\mathbf{H}), \quad (20)$$

$$\sqrt{T}(v(\hat{\boldsymbol{\theta}}_T)-1) \xrightarrow{d} N(0, \mathbf{L}'\mathbf{B}^{-1/2}(\mathbf{I}-\mathbf{M})\mathbf{B}^{-1/2}\mathbf{L}), \quad (21)$$

where $\mathbf{H} \equiv \nabla_{\boldsymbol{\theta}}\beta(\bar{\boldsymbol{\theta}}_T)$ and $\mathbf{L} \equiv \nabla_{\boldsymbol{\theta}}v(\bar{\boldsymbol{\theta}}_T)$ are gradients that can be calculated using numerical derivatives.[12] These expressions indicate the problem with the straightforward use of finite sample inference applied to the slope coefficient and the variance ratio, because the variances depend on parameters that are not fully specified under the null hypothesis. However, the asymptotic distributions of the *t* statistics, which will be referred to as *t2* and *t3* respectively, obtained from (20) and (21) are standard normal and do not depend on the parameters of the specific DGP from the set satisfying the ET that generated the data under the null hypothesis.

In particular, the studentized test statistics we propose are obtained as

$$t2 = \sqrt{T}(\beta(\hat{\boldsymbol{\theta}}_T)-1)/(\mathbf{H}'\mathbf{B}^{-1/2}(\mathbf{I}-\mathbf{M})\mathbf{B}^{-1/2}\mathbf{H})^{1/2} \quad (22)$$

$$t3 = \sqrt{T}(v(\hat{\boldsymbol{\theta}}_T)-1)/(\mathbf{L}'\mathbf{B}^{-1/2}(\mathbf{I}-\mathbf{M})\mathbf{B}^{-1/2}\mathbf{L})^{1/2}. \quad (23)$$

Because of the asymptotic standard normal property of these statistics, the bootstrap provides valid higher order approximations to their finite sample distributions under the null (Horowitz 2001).

Section 5. Monte Carlo study

This section first sets out the methodology used in our Monte Carlo analysis and then compares the control parameters with those estimated from real data to ensure the empirical relevance of the exercise and finally discusses the results.

5.1 Methodology

To our knowledge only two studies to date have compared the finite sample properties of ET tests. Bekaert and Hodrick (2001) compare the *LM*, *DM* and *W* tests of (12)-(14), while Sarno *et al.* (2006) compare the *LM* test with its extended versions. The extensions are to

[12] See for example, Campbell *et al.* (1997, p540). We also used the distribution in (11) under the null, as in related literature of Hodrick (1992) and Bekaert and Hodrick (1992), but our specification was found to perform slightly better in our Monte Carlo Experiments.

include inflation in the VAR, while they also consider testing the ET using more than two interest rates. The former study finds the *LM* performs better than the other two statistics and the latter concludes the extensions increase the power of the *LM* test.

Our study differs from these in at least three important aspects. Firstly, both previous studies take the unrestricted VAR, estimated on the observed data, as the DGP under the alternative hypothesis. However, this assumes that the ET does not hold in the observed data. More importantly, even if this assumption is true, the procedure means that the alternative, against which the null is being tested, is unknown. As Melino (2001) argues, a good statistical methodology considers the alternative hypotheses that are the most plausible and constructs tests which are as sensitive as possible in detecting differences between the maintained hypothesis and these particular alternatives. Therefore, we explicitly consider the widely cited overreaction and time varying term premium hypotheses as alternatives to the null of the ET.

The second important difference is that we compare the finite sample properties of not only the classic trinity of tests in (12)-(14), but also the conventional single equation tests obtained from (3) that will be referred to as *t1*, and studentized implied regression and variance ratio tests, *t2* and *t3*, obtained in (22) and (23) respectively. The performance of the conventional test can be predicted, given its econometric problems discussed in Section 2 and results in Bekaert *et al.* (1997), but the latter two have not been considered in the previous literature. Finally, we explicitly analyse the effects of interest rate maturities and persistency on the powers of the tests.

In order to ensure that the DGP resembles the real world and various levels of interest rate persistency induced by different monetary policy regimes (Mankiw and Miron, 1986), we use observed term structure data from the US and the UK, the countries where most and least evidence against the ET has been reported, as the basis of our Monte Carlo study. To be explicit, we estimate a first order VAR for the mean with a multivariate GARCH (1,1) process of Engle and Kroner (1995) applied to the resulting residuals:

$$\mathbf{y}_t = \boldsymbol{\delta} + \mathbf{A}\mathbf{y}_{t-1} + \mathbf{e}_t ,$$

$$\mathbf{e}_t = \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\xi}_t ,$$

$$\boldsymbol{\Sigma}_t = \mathbf{D}'\mathbf{D} + \mathbf{F}'\mathbf{e}_{t-1}\mathbf{e}_{t-1}'\mathbf{F} + \boldsymbol{\Lambda}'\boldsymbol{\Sigma}_{t-1}\boldsymbol{\Lambda} ,$$

where $\mathbf{y}_t = \left[\Delta R_{m,t}, S_{(n,m),t}\right]'$ and $\boldsymbol{\xi}_t$ is the vector standard normal variable. The VAR and GARCH parameters are estimated using 1 and 3-month UK government Treasury Bill rates from January 1979 to May 2004 obtained from DataStream® and 1 and 2-month US zero coupon yield data from Jan 1952 to Dec 2003 from Sarno *et al.* (2006). This specific choice of data has no other intention except to generate DGPs that are empirically relevant. These are:

US: $\mathbf{A} = \begin{bmatrix} 0.069 & 1.035 \\ -0.09 & 0.363 \end{bmatrix}$, $\boldsymbol{\delta} = \begin{bmatrix} -0.0009 \\ 0.0002 \end{bmatrix}$,

$\mathbf{D} = \begin{bmatrix} 0.095 & 0.012 \\ 0.000 & 0.000 \end{bmatrix}$, $\mathbf{F} = \begin{bmatrix} 0.218 & 0.074 \\ 0.021 & 0.397 \end{bmatrix}$, $\mathbf{G} = \begin{bmatrix} 0.951 & -0.034 \\ 0.014 & 0.920 \end{bmatrix}$;

UK: $\mathbf{A} = \begin{bmatrix} -0.020 & 0.631 \\ 0.060 & 0.368 \end{bmatrix}$, $\boldsymbol{\delta} = \begin{bmatrix} -0.024 \\ 0.001 \end{bmatrix}$,

$\mathbf{D} = \begin{bmatrix} 0.041 & -0.019 \\ 0.000 & 0.000 \end{bmatrix}$, $\mathbf{F} = \begin{bmatrix} 0.229 & 0.062 \\ -0.171 & 0.203 \end{bmatrix}$, $\mathbf{G} = \begin{bmatrix} 0.958 & -0.037 \\ 0.166 & 0.966 \end{bmatrix}$;

We consider four DGPs,

DGP1: UK VAR+UK GARCH

DGP2: US VAR+US GARCH

DGP3: UK VAR+US GARCH, and

DGP4: US VAR+UK GARCH.

Using these processes we generate samples of four different sizes, with 150, 300, 600 and 1500 observations, plus 1000 observations that are discarded. Since most studies of the ET use monthly data, the three smaller sample sizes we employ reflect the lengths of actual data typically available to a researcher.

Asymptotic and bootstrapped versions of all tests are considered. Although asymptotic tests are easy to use, there is now considerable evidence that they can suffer from large size distortions in finite samples. In contrast, bootstrapped tests are computationally expensive but are designed to have the correct size. In our size-power study the critical values come from the $\chi(2)$ and $N(0,1)$ distributions for the asymptotic tests and from the null empirical distributions for the bootstrapped tests.

To implement the bootstrap, we apply the recursive procedure described in Section 3 to estimate the VAR parameters that are restricted to satisfy the ET null hypothesis of (10) and obtain the corresponding residuals. These parameters and residuals are then used in (16) as the wild bootstrap DGP. This DGP is used to generate 1000 datasets, for each of the four sample sizes, that are used to estimate the six empirical distributions for the LM, DM, W, *t1*, *t2* and *t3* statistics. For the first trinity, the critical values are simply the 95% quantiles of these distributions. For the various *t* statistics, they are the 2.5 and 97.5% quantiles from the empirical distributions, allowing for two tailed alternatives.

Following the recommendations of Shiller (1987) and Melino (2001), we concentrate on analysing the test powers against interesting alternative hypotheses. Two of the most prominent alternatives to the ET are the under/over-reaction hypothesis and the time varying term premium. We capture these alternatives by generating term structure data using

$$S_{(n,m),t} = \delta S^{*}_{(n,m),t} - \tau_t , \qquad (24)$$

where the theoretical spread $S^{*}_{(n,m),t}$ is defined in (6) and $\tau_t = c\mathbf{i}'\mathbf{z}_t$ where $\mathbf{i}$ is a unit vector. For simplicity, the term premium τ_t here depends on the sum of current $\Delta R_{m,t}$ and $S_{(n,m),t}$[13]. When c is zero, (24) corresponds to the ET if δ = 1, to the over-reaction hypothesis if $\delta > 1$ and to the under-reaction hypothesis if $\delta < 1$. If $\delta = 1$ and c is nonzero, we have the time varying term premium hypothesis.

The alternative given by (24) translates into the VAR parameter restrictions

$$\mathbf{e}_2' = \delta \mathbf{e}_1' \mathbf{\Lambda} - c\mathbf{i}' . \qquad (25)$$

We consider different values of the parameters δ and c, specifically $\delta \in [0.6, 1.4]$ and $c \in [-0.25, 0.25]$. For each pair (δ, c), and for a given sample size T, the VAR coefficients are estimated satisfying the restrictions of (25), not those of (10). These VAR coefficients restricted under the alternative hypothesis and the corresponding residuals are then used in (16) as DGP to generate 1000 datasets where the ET does not hold. For each of the datasets we obtain the six test statistics corresponding to the null in (10) and the proportions of the test statistics that are greater (and, for the two tailed tests, lower) than the relevant critical values provide the estimates of power. When δ = 1 and c = 0, this exercise yields an

[13] Of course one can estimate a VAR of order greater than 1 and let the current term premium depend only on past information, possibly also allowing the contributions of past $\Delta R_{m,t}$ and $S_{(n,m),t}$ to differ.

estimate of the actual size of a test and if this actual size is greater (smaller) than nominal 5% the test is over(under)sized.

To summarise, we first generate 1500 observations, after discarding initial 1000 observations, from each DGP starting from the same random number generator and use first 150, 300, 600 and 1500 observations as raw data in the size-power calculation for various n and m. This allows us to examine the effects of n and m that are embedded in the null $\mathbf{e}_2' = \mathbf{e}_1'\mathbf{\Lambda}(\mathbf{\Phi}, n, m)$ defined in (7) on the sizes and powers of various tests while keeping $\mathbf{\Phi}$ fixed. By using two VAR slope parameters estimated from the UK and US data we are able to examine the effect of $\mathbf{\Phi}$ while keeping n and m fixed. Finally we keep all of the above, specifically $\mathbf{\Phi}, n$ and m, fixed and examine the effects of dynamics that are not restricted by the ET by employing two different conditional volatility processes.

5.2 Corroboration

Before presenting our main results we briefly assess how relevant our choice of values for δ and c are in terms of observed series by estimating these parameters using US term structure data. The theoretical spread is calculated as in Bataa *et al.* (2006) and the actual spread is regressed on the theoretical spread to obtain an estimate of δ. A time varying term premium proxy is obtained as a difference between the actual and theoretical spreads, as in Carriero *et al.* (2006), and we regress this on the sum of current spread and the first difference of the shorter rate. Table 1 provides the resulting point estimates of δ and c, along with their standard errors, for all conventional maturity pairs between 1 month and 10 years for three different sample periods: whole sample (Jan 1952- Dec 2003), pre-1979-1982 monetary policy change (Jan 1952- Dec 1978) and post monetary policy change (Jan 1982-Dec 2003).

Most slope estimates are statistically significant, and ignoring two negative estimates of δ, range between 0.21 and 3.39, while estimates of c range between -0.22 and 0.75. In particular, for the three different sample periods 6.3%, 56.3% and 20.8% of the point estimates of δ are in the range $\delta \in$[0.6, 1.4] and 35.4%, 83.3% and 20.8% of the point estimates of c are in the range $c \in$[-0.25, 0.25]. Although the range of values we consider for δ and c is not fully representative of the data and therefore could be extended we had to restrict ourselves because of the computational burden. Consistent with Campbell and Shiller (1991) and Fuhrer (1996), most of the estimates of δ are greater than unity (85.42%, 87.50% and 91.67%) implying the actual spread is more volatile than

Table 1. Empirical Estimates of δ and c

	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
	1			**2**			**3**			**4**			**6**			**9**			**12**			**24**			**60**		
2	0.42	0.42	0.25																								
	0.07	*0.11*	*0.09*																								
	-0.02	0.01	-0.10																								
	0.01	*0.02*	*0.03*																								
3	0.21	0.68	0.60																								
	0.05	*0.14*	*0.09*																								
	-0.04	0.07	-0.01																								
	0.02	*0.02*	*0.04*																								
4	0.30	0.25	1.21	0.44	1.58	1.09																					
	0.05	*0.08*	*0.13*	*0.06*	*0.13*	*0.09*																					
	-0.01	0.04	0.17	0.01	0.04	0.10																					
	0.02	*0.03*	*0.04*	*0.01*	*0.02*	*0.02*																					
6	0.63	0.31	1.38	1.75	1.61	1.26	2.03	1.39	0.96																		
	0.07	*0.08*	*0.12*	*0.08*	*0.11*	*0.14*	*0.14*	*0.12*	*0.27*																		
	0.11	0.12	0.28	0.13	0.11	0.29	0.10	0.06	0.21																		
	0.02	*0.04*	*0.04*	*0.02*	*0.02*	*0.04*	*0.01*	*0.02*	*0.03*																		
9	1.25	0.31	1.56				2.05	1.45	2.09																		
	0.07	*0.09*	*0.10*				*0.10*	*0.10*	*0.21*																		
	0.20	0.25	0.36				0.19	0.15	0.37																		
	0.02	*0.04*	*0.04*				*0.02*	*0.02*	*0.03*																		
12	1.33	0.45	1.69	1.55	1.50	1.01	1.89	1.43	2.68	1.89	1.57	2.71	1.64	1.74	2.27												
	0.07	*0.09*	*0.09*	*0.05*	*0.09*	*0.05*	*0.07*	*0.09*	*0.18*	*0.08*	*0.11*	*0.19*	*0.08*	*0.10*	*0.20*												
	0.22	0.28	0.40	0.21	0.21	0.21	0.23	0.18	0.45	0.19	0.19	0.37	0.09	0.12	0.20												
	0.02	*0.04*	*0.03*	*0.02*	*0.03*	*0.03*	*0.02*	*0.03*	*0.03*	*0.02*	*0.03*	*0.03*	*0.01*	*0.02*	*0.03*												
24	1.76	1.54	1.92	2.41	1.53	2.23	2.36	1.53	2.19	1.88	1.27	2.16	1.77	1.43	2.21				3.17	1.29	1.20						
	0.08	*0.08*	*0.08*	*0.14*	*0.06*	*0.09*	*0.18*	*0.07*	*0.10*	*0.05*	*0.05*	*0.11*	*0.04*	*0.05*	*0.08*				*0.22*	*0.04*	*0.19*						
	0.40	0.30	0.44	0.53	0.27	0.51	0.55	0.26	0.50	0.32	0.17	0.48	0.25	0.18	0.39				0.25	0.05	0.31						
	0.02	*0.03*	*0.03*	*0.02*	*0.03*	*0.03*	*0.02*	*0.03*	*0.03*	*0.02*	*0.03*	*0.03*	*0.02*	*0.02*	*0.02*				*0.02*	*0.02*	*0.03*						
36	1.92	1.46	1.54	2.85	1.39	2.02	1.90	1.36	2.37	1.98	1.36	2.59	1.85	1.35	2.74	1.91	1.22	3.17	2.01	1.14	3.39						
	0.07	*0.06*	*0.06*	*0.10*	*0.05*	*0.07*	*0.04*	*0.05*	*0.09*	*0.04*	*0.05*	*0.09*	*0.04*	*0.04*	*0.10*	*0.04*	*0.04*	*0.14*	*0.04*	*0.04*	*0.21*						
	0.43	0.29	0.35	0.56	0.24	0.49	0.38	0.23	0.54	0.39	0.22	0.55	0.33	0.19	0.52	0.30	0.10	0.53	0.27	0.04	0.51						
	0.02	*0.03*	*0.03*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*						
48	1.97	1.45	1.59	2.81	1.37	1.97	1.91	1.34	2.29	1.99	1.32	2.48	1.79	1.29	2.64				1.93	1.10	3.01	-0.66	1.03	3.13			
	0.06	*0.05*	*0.05*	*0.08*	*0.05*	*0.06*	*0.04*	*0.05*	*0.07*	*0.04*	*0.04*	*0.08*	*0.03*	*0.04*	*0.08*				*0.03*	*0.03*	*0.10*	*0.24*	*0.03*	*0.38*			
	0.45	0.28	0.37	0.58	0.24	0.48	0.41	0.22	0.54	0.42	0.21	0.55	0.35	0.17	0.54				0.33	0.03	0.53	0.39	-0.07	0.44			
	0.02	*0.03*	*0.03*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.02*	*0.02*				*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.03*			
60	1.97	1.45	1.56	2.70	1.37	1.96	1.94	1.33	2.33	2.00	1.32	2.28	1.92	1.24	2.44				-0.24	1.08	2.88						
	0.05	*0.05*	*0.05*	*0.07*	*0.04*	*0.06*	*0.04*	*0.04*	*0.07*	*0.04*	*0.04*	*0.06*	*0.03*	*0.03*	*0.06*				*0.19*	*0.03*	*0.08*						
	0.45	0.28	0.36	0.57	0.24	0.48	0.43	0.22	0.55	0.44	0.21	0.53	0.40	0.15	0.53				0.75	0.03	0.56						
	0.02	*0.03*	*0.03*	*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*				*0.02*	*0.02*	*0.02*						
120	1.80	1.41	1.42	1.68	1.34	1.61	1.85	1.32	1.75	1.93	1.29	1.62	1.80	1.23	1.53				1.65	1.14	1.33	1.58	1.15	1.21	1.98	1.78	0.72
	0.04	*0.04*	*0.04*	*0.03*	*0.03*	*0.04*	*0.03*	*0.03*	*0.04*	*0.03*	*0.03*	*0.03*	*0.03*	*0.03*	*0.03*				*0.02*	*0.03*	*0.02*	*0.02*	*0.03*	*0.02*	*0.03*	*0.06*	*0.01*
	0.42	0.27	0.30	0.38	0.23	0.39	0.43	0.22	0.43	0.45	0.20	0.38	0.41	0.16	0.34				0.34	0.09	0.24	0.28	0.06	0.15	0.21	0.15	-0.22
	0.02	*0.02*	*0.02*	*0.01*	*0.02*	*0.02*	*0.01*	*0.02*	*0.02*	*0.01*	*0.02*	*0.02*	*0.01*	*0.02*	*0.02*				*0.01*	*0.02*	*0.01*	*0.01*	*0.02*	*0.01*	*0.02*	*0.03*	*0.02*

Note: Empirical estimates of δ (first row) and c (second row) and their standard errors (*in italics*) are reported for various maturity pairs using data from *Sarno et al.* (2006). I, II, and III denotes sample periods, which are 1952:01-2003:12, 1952:01-1978:12 and 1982:01-2003:12 respectively.

the theoretical spread; 93.75%, 97.92% and 93.75% of the point estimates for *c* are positive, again consistent with Tzavalis and Wickens (1997). Given that the set of parameter values we consider is indeed empirically relevant we discuss our results next.

5.3 Results

Table 2 reports the empirical sizes for all tests from DGP1 and DGP2 for various maturity pairs at a nominal significance level of 5%.[14] For each DGP, the first panel reports size results for m=1 in common with a range of values for the longer maturity n and the lower part reports for n=180 in common with a range of values for the maturity m. In the final panel we keep k(=n/m) constant and increase both n and m. It appears that size does not depend on the DGP. It is, however, evident that the *W, LM* and *DM* trinity is oversized and this size distortion does not disappear even with a sample of 1500 observations. The size distortion is generally smaller for *LM* than for *W* and *DM* tests in small samples, confirming the conclusions of Bekaert and Hodrick (2001). However, the size differences among these three become largely indistinguishable as the sample size increases. In contrast, the implied regression and variance ratio tests, *t2* and *t3*, seem to have the least size distortion.

Unlike the other tests, the size of conventional test, *t1*, depends on the maturities of interest rates. For example, for a maturity pair 1&3 months, size is 6% with 300 observations, however this inflates into a staggering 53% when the maturity of the longer rate is 120, which implies a cross maturity spectrum comparison of ET performance becomes senseless unless one is willing to compensate for the number of observations "lost" in the estimation process. This "loss" of observations has two sources: one is physical loss of observations in trying to calculate the left hand side of equation (3) and the other is the loss of independent observations resulting from a high degree of MA correction.

The second important question is to analyse how powerful the tests are against empirically relevant local alternatives. Using the asymptotic distribution, Figure 2 illustrates size-power curves for the sample size of 300 observations from DGP2 where the tests have their highest powers compared to other DGP's.[15] The powers of the tests against the under/overreaction hypothesis are plotted in Panel A while those against the time varying term premium are shown in Panel B. While Table 2 showed no obvious pattern for

[14] Results from DGP3 and DGP4 were quantitatively very close and qualitatively the same. They are not reported to conserve space but available on request.
[15] Full results are available from the authors upon request.

Table 2. Empirical Size Results

T		150	300	600	1500	150	300	600	1500	150	300	600	1500	150	300	600	1500
	Panel A: DGP1																
									n								
		3				*9*				*24*				*120*			
LM		0.15	0.12	0.15	0.17	0.14	0.12	0.15	0.15	0.15	0.13	0.16	0.16	0.15	0.13	0.16	0.16
DM		0.18	0.13	0.16	0.17	0.17	0.17	0.17	0.16	0.18	0.17	0.19	0.17	0.18	0.17	0.19	0.18
W	*m=1*	0.18	0.13	0.16	0.17	0.18	0.17	0.18	0.16	0.19	0.17	0.18	0.18	0.19	0.17	0.18	0.18
t1		0.08	0.06	0.06	0.07	0.11	0.10	0.09	0.07	0.23	0.20	0.13	0.10	N.A.	0.53	0.31	0.16
t2		0.06	0.04	0.05	0.05	0.06	0.07	0.07	0.07	0.06	0.08	0.07	0.07	0.06	0.08	0.07	0.07
t3		0.06	0.04	0.05	0.06	0.07	0.08	0.07	0.06	0.08	0.08	0.07	0.07	0.09	0.09	0.08	0.07
									m								
		1				*6*				*36*				*60*			
LM		0.15	0.13	0.16	0.16	0.15	0.13	0.16	0.16	0.14	0.13	0.15	0.15	0.14	0.13	0.15	0.17
DM		0.18	0.17	0.19	0.18	0.18	0.17	0.19	0.17	0.18	0.17	0.17	0.16	0.17	0.15	0.16	0.18
W	*n=180*	0.19	0.17	0.18	0.18	0.19	0.17	0.18	0.18	0.18	0.17	0.16	0.17	0.19	0.17	0.16	0.17
t1		N.A.	N.A.	0.44	0.21	N.A.	N.A.	0.44	0.21	N.A.	0.69	0.38	0.20	N.A.	0.57	0.34	0.20
t2		0.06	0.08	0.07	0.07	0.06	0.08	0.07	0.07	0.07	0.06	0.06	0.07	0.06	0.06	0.05	0.06
t3		0.09	0.09	0.08	0.07	0.08	0.09	0.08	0.07	0.06	0.07	0.07	0.07	0.06	0.06	0.07	0.06
	Panel B: DGP2																
									n								
		3				*9*				*24*				*120*			
LM		0.16	0.15	0.14	0.16	0.15	0.15	0.14	0.16	0.14	0.16	0.14	0.17	0.15	0.15	0.14	0.16
DM		0.18	0.16	0.14	0.17	0.17	0.17	0.15	0.17	0.17	0.17	0.15	0.17	0.17	0.17	0.15	0.17
W	*m=1*	0.18	0.16	0.14	0.17	0.18	0.17	0.15	0.16	0.18	0.16	0.15	0.16	0.18	0.16	0.16	0.16
t1		0.09	0.06	0.05	0.08	0.14	0.09	0.10	0.08	0.22	0.13	0.11	0.08	N.A.	0.40	0.24	0.13
t2		0.06	0.05	0.04	0.07	0.05	0.06	0.04	0.06	0.05	0.06	0.04	0.06	0.05	0.06	0.04	0.06
t3		0.05	0.04	0.04	0.06	0.06	0.07	0.05	0.06	0.08	0.07	0.06	0.06	0.08	0.08	0.06	0.06
									m								
		1				*6*				*36*				*60*			
LM		0.15	0.15	0.14	0.16	0.15	0.15	0.14	0.17	0.15	0.15	0.15	0.17	0.15	0.14	0.15	0.16
DM		0.17	0.17	0.15	0.17	0.17	0.17	0.15	0.17	0.17	0.16	0.15	0.17	0.17	0.16	0.16	0.17
W	*n=180*	0.18	0.16	0.16	0.16	0.18	0.17	0.16	0.16	0.17	0.17	0.15	0.16	0.18	0.16	0.17	0.17
t1		N.A.	N.A.	0.35	0.17	N.A.	N.A.	0.33	0.16	N.A.	0.54	0.31	0.16	N.A.	0.47	0.28	0.15
t2		0.05	0.06	0.04	0.06	0.05	0.06	0.04	0.06	0.05	0.06	0.05	0.06	0.05	0.05	0.04	0.07
t3		0.08	0.08	0.06	0.06	0.08	0.08	0.06	0.06	0.08	0.07	0.05	0.06	0.07	0.07	0.05	0.06

Note: Table reports empirical sizes of the tests at the nominal significance level of 0.05**.** N.A. for the *t1* test indicates there are an insufficient number of observations available for the MA correction after losing *n-m* observations in the calculation of the test statistic.

the size distortion across various DGPs and in relation to location of the maturity pair in the term structure spectrum, except for *t1*, there are clear patterns for test powers. The three rows of the graphs in each panel are designed to show three different aspects of the effect of maturities, the first row showing the effects of increasing *n* for a given *m*, the second row of increasing *m* for a given *n*, and the third row increasing both *n* and *m* while keeping *k* (=*n*/*m*) constant. All the tests are considerably powerful at the shortest end of the maturity spectrum. However as one deviates from there the powers decrease, but with varying degrees for various tests. It can be seen that for large *n* and small *m* the plotted size-power curve of *t1* either becomes a horizontal line at zero as there is an insufficient number of observations for the MA correction after loosing *n-m* observations in the calculation process or straight lines running from the bottom left to the top right against the under/over-reaction hypothesis and from the bottom right to the top left of the graph

against the time varying term premium alternative. It is evident that the other size-power curves also depend on *n* and *m*, especially those of *t2* and *t3*. Although *t2* and *t3* tests have the closest empirical sizes to the nominal one, their powers can be smaller than size, when the null of the ET is tested against the over-reaction hypothesis or the time varying term premium that is negatively correlated with the sum of the spread and the change in shorter rate. In contrast, even though *LM, DM* and *W* tests are oversized, they have good powers against the over-reaction hypothesis and the time varying term premium.

In Figures 2-5 we explicitly compare the tests' powers across 4 DGPs and 4 sample sizes, using bootstrapped versions to avoid the problem that tests with higher size distortion may incorrectly appear more powerful. The reported are the same three row graphs designed to illustrate the effects on interest rate maturities on test powers. Panels A1-A4 (A1-A3 for DGP3 and DGP4) consider test powers against the alternative of over/under-reaction hypothesis and Panels B1-B4 (B1-B3 for DGP3 and DGP4) report those against time varying term premium.

In Panel As the powers are in general lowest for DGP1 (that is, based on UK data characteristics) and strongest in DGP2 (based on the US).[16] Indeed, Panel A1 of Figure 2 (UK mean and volatility) shows power to be low with a sample size of 150 observations, except at the very short end of the maturity spectrum. Except for the large sample size (T=1500), Figure 2 shows power to be greater against the under-reaction than the over-reaction hypothesis, whereas, with the exception of the small sample size of 150 observations, this is less marked in Figure 3. Conditional volatility of the UK DGP seems to have the strongest negative effect on power, as when we combine US conditional volatility with the UK mean (that is DGP3, Figure 4) test powers dramatically increase while doing the converse (that is DGP4, Figure 5) entails a dramatic decrease in test performances. From these Panels one can also see that the *t1* and *W* tests are most sensitive to increasing *n* for a given *m* and decreasing *m* for a given *n*, but the sensitivity of the latter diminishes much faster than that of the former as the sample size increases. The *W* test is also most powerful against the over-reaction hypothesis. In small samples the implied regression, *t2*, and variance ratio, *t3*, tests are most powerful against the overreaction hypothesis at the shortest end of the term structure but they lose power to the *LM* and *DM* tests as the sample size increases and/or interest rate maturities change in a pattern described above.

[16] We do not provide A4 panels that correspond to 1500 observations for DGP3 and DGP4 because of the computational cost.

In Panels B1-B4 (B1-B3 for DGP3 and DGP4) we report power of the tests of the ET against the time varying term premium alternative. All tests perform best at the shortest end, as before. Interestingly, the dependence of the test performances on the specific DGP is less pronounced than in Panel As. There is no clear evidence for *LM*, *DM* and *W* tests for such a dependence but comparing B1 in Figures 2 & 3 and similarly B2 in these Figures, the *t*-tests perform better in Figure 3, especially when $c < 0$. In this case there is no close competitor to *LM* and *DM* tests which perform much better than the *t2* and *t3* tests, the latter of which can even have power below size for negative *c*. The worst test is again *t1*, while *W* performs very comparably with the forerunners as the sample size increases.

Overall, the performances of all tests are at their best when both *m* and *n* are small and decrease as the distance between *m* and *n* increase. However, this decrease is of varying degrees for each of the tests and its speed diminishes as we depart from the shortest end of the maturity spectrum. For example, in DGP1-Panel B when *m* is kept constant at 1 and *n* is increased from 3 to 9 all the powers reduce dramatically, but the reductions for *LM*, *DM* and *W* are relatively small than those based on *t*-tests. However, as we further increase *n* to 120 the power reduction that follows is much smaller and arguably less important for *t*-tests. When maturity pairs are already further away from the shortest end the specific values for *n* and *m* appears not to matter as long as *k(=n/m)* is constant. The size-power curves for maturity pairs 24&48 and 60&120 are visually extremely similar from both panels and across DGPs for all sample sizes, except for *t1*. Another interesting observation is that in DGP3-Panel B the test powers are higher for the latter maturity pair than for the former for all sample sizes. Moreover, increasing both *m* and *n* can improve the test powers as long as this also reduces *k*, as if large *n* and *m* cancel out each other. For example all tests, except *t1* and *W,* are more powerful at the maturity pair 60&120 than at 1&24 in all sample sizes and all DGP's excluding and also DGP3. However these sensitivities to interest rate maturities and *k* are much less pronounced, except *t1*, for the large sample size of 1500 observations.

6. Conclusions

This paper provides extensive Monte Carlo evidence that addresses the empirical puzzle that the spread between longer and shorter rates predicts future movements in the shorter rate in accordance with the ET if the forecast horizon is long, but not otherwise. We also study if the persistency of interest rates can explain a general finding in this empirical literature that indicates the ET appears to hold in some samples (country-wise and time-wise) as opposed to others.

We do indeed find that in small samples, the powers of the tests previously employed in the literature, and new ones we propose, depend on the interest rates' maturities and their persistency of conditional volatility when the null hypothesis of the ET is tested against empirically relevant alternative hypotheses such as the under/over-reaction and the time varying term premium hypotheses. This suggests that the cross sample (country-wise or time-wise) and/or maturity spectrum comparison of the performance of the theory is, strictly speaking, impossible in samples typically used in the literature, as the powers of the tests are not the same. Since such dependence diminishes as the sample size increases, this suggests powerful tests might be obtained and their results can be compared across maturity spectrum and/or samples if one uses high frequency data such as daily data provided by the Bank of England and the US Federal Reserve.

Our secondary goal was to compare finite sample performances of the conventional regression test (*t1*), VAR based *LM*, *DM*, *W* tests considered in Bekaert and Hodrick (2001), and the implied regression (*t2*) and variance ratio (*t3*) tests, which are constructed using the asymptotic distribution of the restricted parameter estimator. Although forms of the latter two tests have been suggested in previous literature, we argue that the use of these leads to invalid inference, whereas we avoid this problem by proposing a *t* ratio specification for these tests. An extensive Monte Carlo analysis is used to achieve this goal, with tests conducted using both the asymptotic distributions of the statistics and the empirical finite sample distributions obtained using a wild bootstrap procedure.

The conventional test's actual size depends on the maturities of interest rates and it can go up to 56% in our simulations when the nominal significance level is 5%. *LM, DM* and *W* tests are oversized, the former being least size distorted, which is consistent with Bekaert and Hodrick (2001), while the implied regression and variance ratio tests have the closest actual sizes to the nominal level. However, the powers of the latter two can be less

than their size, when the null of the ET is tested against the under-reaction hypothesis or against a time varying term premium that is negatively correlated with sum of the spread and change in short rate.

In order to compare powers of the tests while keeping their sizes at the nominal level we use a wild bootstrap. The performances of all tests are at their best and comparable with each other against both alternative hypotheses when both m and n are small, but decrease as the distance between m and n increase. However, this decrease is of varying degrees for each of the tests and its speed diminishes as we depart from the shortest end. The worst performance comes from the conventional test ($t1$) while LM and DM tests are most consistent and powerful. However it is worth mentioning that all the conclusions rely on the assumption of known DGP and it is indeed possible that the implied regression and variance tests may outperform in situations of unknown lag order for the assumed VAR-DGP as their asymptotic distributions are pivotal compared to those of LM, DM and W.

Appendix A: Derivation of the ET restrictions on VAR parameters

Restrictions imposed by the ET on the companion-form VAR parameters are derived in this Appendix.

The spread predicted by the ET can be obtained from (2) as, $S_{(n,m),t} = \sum_{i=1}^{k-1}(1-\frac{i}{k})E(\Delta^m R_{m,t+im}|\Xi_t)$. This spread is not observable, but can be proxyed using the VAR. Assuming the process for $\mathbf{y}_t = \left[\Delta R_{m,t}, S_{(n,m),t}\right]'$ is represented by a demeaned VAR of order *p*, its companion form can be written as, $\mathbf{z}_t = \mathbf{\Phi}\mathbf{z}_{t-1} + \mathbf{u}_t$, and hence $E(\mathbf{z}_{t+j}|I_t) = \mathbf{\Phi}^j \mathbf{z}_t$, assuming $E(\mathbf{u}_{t+i}|I_t) = \mathbf{0}$, $i > 0$. Thus, using the notation of the text, $E(\Delta R_{m,t+j}|I_t) = \mathbf{e}_1' E(\mathbf{z}_{t+j}|I_t) = \mathbf{e}_1'\mathbf{\Phi}^j\mathbf{z}_t$ as $\Delta R_{m,t} = \mathbf{e}_1'\mathbf{z}_t$.

In what follows we use the following two simple results from matrix algebra:

$$\mathbf{I} + \mathbf{\Phi} + \mathbf{\Phi}^2 + ... + \mathbf{\Phi}^{n-m-1} = (\mathbf{I}-\mathbf{\Phi})^{-1} - \mathbf{\Phi}^{n-m}(\mathbf{I}-\mathbf{\Phi})^{-1} = (\mathbf{I}-\mathbf{\Phi}^{n-m})(\mathbf{I}-\mathbf{\Phi})^{-1}$$

which follows since $\mathbf{I} + \mathbf{\Phi} + \mathbf{\Phi}^2 + ... = (\mathbf{I}-\mathbf{\Phi})^{-1}$ and

$$\mathbf{I} + \mathbf{\Phi}^m + \mathbf{\Phi}^{2m} + ... + \mathbf{\Phi}^{n-m} = (\mathbf{I}-\mathbf{\Phi}^m)^{-1} - \mathbf{\Phi}^n(\mathbf{I}-\mathbf{\Phi}^m)^{-1} = (\mathbf{I}-\mathbf{\Phi}^n)(\mathbf{I}-\mathbf{\Phi}^m)^{-1}$$

since $\mathbf{I} + \mathbf{\Phi}^m + \mathbf{\Phi}^{2m} + ... = (\mathbf{I}-\mathbf{\Phi}^m)^{-1}$.

Now define a theoretical spread $S^*_{(n,m),t}$, which is related to the spread under the null hypothesis of the ET as , $S_{(n,m),t} = S^*_{(n,m),t} + w_t$, where w_t has mean zero and

$$\begin{aligned}
S^*_{(n,m),t} &\equiv \sum_{i=1}^{k-1}\left(1-\frac{i}{k}\right)E(\Delta^m R_{m,t+im}|I_t) = \sum_{i=1}^{k-1}\left(1-\frac{i}{k}\right)\sum_{j=0}^{m-1}E(\Delta R_{m,t+im-j}|I_t) \\
&= \sum_{i=1}^{k-1}\left(1-\frac{i}{k}\right)\{\mathbf{e}_1'[\mathbf{\Phi}^{im} + \mathbf{\Phi}^{im-1} + ... + \mathbf{\Phi}^{im-m+1}]\mathbf{z}_t\} \\
&= \frac{1}{k}\{\mathbf{e}_1'[(k-1)(\mathbf{\Phi}^m + \mathbf{\Phi}^{m-1} + ... + \mathbf{\Phi}) + (k-2)(\mathbf{\Phi}^{2m} + \mathbf{\Phi}^{2m-1} + ...\mathbf{\Phi}^{m+1}) + ... \\
&\quad + (\mathbf{\Phi}^{n-m} + \mathbf{\Phi}^{n-m-1} + ... + \mathbf{\Phi}^{n-2m+1})]\mathbf{z}_t\} \\
&= \frac{1}{k}\{\mathbf{e}_1'[\mathbf{\Phi}(\mathbf{I}-\mathbf{\Phi})^{-1} - \mathbf{\Phi}^{n-m+1}(\mathbf{I}-\mathbf{\Phi})^{-1} + \mathbf{\Phi}(\mathbf{I}-\mathbf{\Phi})^{-1} - \mathbf{\Phi}^{n-2m+1}(\mathbf{I}-\mathbf{\Phi})^{-1} + ... \\
&\quad + \mathbf{\Phi}(\mathbf{I}-\mathbf{\Phi})^{-1} - \mathbf{\Phi}^{m+1}(\mathbf{I}-\mathbf{\Phi})^{-1}]\mathbf{z}_t\} \\
&= \frac{1}{k}\{\mathbf{e}_1'\mathbf{\Phi}[(k-1)\mathbf{I} - (\mathbf{I}-\mathbf{\Phi}^m)^{-1} + \mathbf{\Phi}^n(\mathbf{I}-\mathbf{\Phi}^m)^{-1} + \mathbf{I}](\mathbf{I}-\mathbf{\Phi})^{-1}\mathbf{z}_t\} \\
&= \mathbf{e}_1'\mathbf{\Phi}[\mathbf{I} - m/n(\mathbf{I}-\mathbf{\Phi}^n)(\mathbf{I}-\mathbf{\Phi}^m)^{-1}](\mathbf{I}-\mathbf{\Phi})^{-1}\mathbf{z}_t = \mathbf{e}_1'\mathbf{\Lambda}\mathbf{z}_t.
\end{aligned}$$

The actual spread, on the other hand, is $S_{(n,m),t} = \mathbf{e}_2' \mathbf{z}_t$. Taking expectations in these expressions, it can be seen that the expected spread and the theoretical spread are equal when $\mathbf{e}_2' = \mathbf{e}_1' \mathbf{\Lambda}$.

Alternatively, the theoretical slope coefficient when $S_{(n,m),t}$ is regressed on $S^*_{(n,m),t}$ is equal to unity.

References:

Backus, D., Foresi, S., Mozumdar, A., Wu, L., 2001. "Predictable Changes in Yields and Forward Rates", *Journal of Financial Economics* 59, 281-311,

Bahadir, R.R. and Savage, L.J. 1956. "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems", *Annals of Mathematical Statistics* 27, 1115-1122,

Balduzzi, P., Bertola, G., Foresi, S., 1997. "A Model of Target Changes and the Term Structure of Interest Rates", *Journal of Monetary Economics* 39, 223- 249,

Bataa, E., Kim, D.H. and Osborn, D.R. 2006. "Extectations Hypothesis Tests in the Presence of Model Uncertainty". Economics Discussion Paper Series, EDP 0611, University of Manchester,

Bekaert, G., & Hodrick, R.J., 1992. "Characterizing Predictable Components in Excess Returns on Equity and Foreign Exchange Markets", *Journal of Finance* 47, 467- 509,

_____, 2001. "Expectations Hypotheses Tests", *Journal of Finance* 56(4), 1357-1394,

Bekaert, G, Hodrick, R.J, Marshall, D,A., 1997. "On Biases in Tests of the Expectations Hypothesis of the Term Structure of Interest Rates", *Journal of Financial Economics* 44, 309-348,

_____, 2001. "Peso Problem Explanations for Term Structure Anomalies", *Journal of Monetary Economics* 48, 241-270,

Bekaert, G., Wei, M. and Xing, Y. 2006. "Uncovered Interest Rate Parity and the Term Structure", *Journal of International Money and Finance*, forthcoming,

Campbell, J.Y., Shiller, R.J., 1987. "Cointegration and Tests of Present Value Models". *Journal of Political Economy* 95, 1062-1088,

_____, 1991. "Yield Spreads and Interest Rate Movement: A Bird's Eye View". *Review of Economic Studies* 58, 495-514,

Campbell, J.Y, Lo, A.W, MacKinlay, A.C., 1997. The Econometrics of Financial Markets, Princeton University Press, New Jersey,

Carriero, A., Favero, C.A. and Kaminska, I. 2006. "Financial Factors, Macroeconomic Information and the Expectations Theory of the Term Structure of Interest Rates", *Journal of Econometrics* 131(1-2), 339- 358,

Cuthbertson, K., 1996. "The Expectations Hypothesis of the Term Structure: The UK Interbank Market", *Economic Journal*, 578- 592,

Cuthbertson, K., Hayes, S., Nitzsche, D., 1996. "Are German Money Market Rates Well Behaved?", *Journal of Economic Dynamics & Control* 24, 347- 360,

Cuthbertson, K., and Bredin, D., 2001. "Risk Premia and Long Rates in Ireland", *Journal of Forecasting* 20, 391- 403,

Dahlquist, M., Jonsson, G., 1995. "The Information in Swedish Short-maturity Forward Rates", *European Economic Review* 39, 1115-1131,

Davidson, J., Flachaire, E., 2001. "The Wild Bootstrap, Tamed at Last". Working Paper, Darp58, STICERD, LSE,

Dufour, J-M. 1997. "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models". *Econometrica* 65, 1365- 1387,

Elton, J.E. 1999. "Expected Return, Realized Return, and Asset Pricing Tests". *Journal of Finance* 54(4), 1199- 1220,

Engle, R.F., Lilien, D.M., Robins, R.P. 1987. "Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model", *Econometrica* 55(2), 391-407,

Engle, R.F., Ng, V.K., 1993. "Time Varying Volatility and the Dynamic Behaviour of the Term Structure", *Journal of Money, Credit and Banking* 25(3), 336- 349,

Engsted, T., 1996. "The Predictive Power of the Money Market Term Structure", *International Journal of Forecasting* 12, 289- 295,

Evans, M.D.D., Lewis, K,K. 1994. "Do Stationary Risk Premia Explain It All?", *Journal of Monetary Economics* 33, 285- 318,

Fama, E.F., 1984. "The Information in the Term Structure". *Journal of Financial Economics* 13, 509-528,

_____, 2006. "The Behavior of Interest Rates". *Review of Financial Studies* 19(2), 359- 379,

Fama, E.F., Bliss, R.R., 1987. "The Information in Long Maturity Forward Rates". *American Economic Review* 77(4), 680-692,

Froot, K.A., 1989. "New Hope for the Expectations Hypothesis of the Term Structure of Interest Rates", *Journal of Finance* 44, 283-305,

Fuhrer, J.C. 1996. "Monetary Policy Shifts and Long-Term Interest Rates", *Quarterly Journal of Economics* 111(4), 1183-1209,

Gerlach, S., Smets, F., 1997. "The Term Structure of Euro-rates: Some Evidence in Support of the Expectations Hypothesis", *Journal of International Money and Finance* 16(2), 305-321,

Godfrey, L.G. and Orme, C.D. 2004. "Controlling the Finite Sample Singificance levels of Heteroskedasticity-robust Tests of Several Linear Restrictions on Regression Coefficients", *Economics Letters* 82, 281-287,

Godfrey, L.G. and Tremayne, A.R. 2005. "The Wild Bootstrap and Heteroskedsticity- Robust Tests for Serial Correlation in Dynamic Regression Models", *Computational Statistics & Data Ananlysis* 49, 377- 395,

Goncalves, S., Kilian, L., 2004. "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form". *Journal of Econometrics* 123, 89- 120,

Hall, A. D., Anderson, H. M., Granger, C. W. J., 1992. "A Cointegration Analysis of Treasury Bill Yields", *Review of Economics and Statistics* 74, 116-126,

Hansen, L. P., 1982. "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica* 50, 1029-1054,

Hansen, L. P. and Hodrick, R.J. "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis", *Journal of Political Economy* 88(5), 829- 853,

Hardouvelis, G.A. 1988. "The Predictive Power of the Term Structure during Recent Monetary Regimes". *Journal of Finance* 43, 339-356,

_____1994. "The Term Structure Spread and Future Changes in Long and Short Rates in the G7 Countries". *Journal of Monetary Economics* 33, 255-283,

Harris, R.D.F. 2001. "The Expectations Hypothesis of the Term Structure and Time Varying Risk Premia: A Panel Data Approach", *Oxford Bulletin of Economics and Statistics* 63(2), 233- 245,

Hodrick, R. 1992. "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement", *Review of Financial Studies* 5, 357- 386,

Horowitz, J.L. 2001. "The Bootstrap", in Heckman, J.S. & Leamer, E. eds. *Handbook of Econometrics* 5, North-Holland, Amsterdam, The Netherlands, 3159- 3228,

Horowitz, J.L. & Savin, N.E. 2000. "Empirically Relevant Critical Values for Hypothesis Tests: A Bootstrap Approach", *Journal of Econometrics* 95, 375- 89,

Kozicki, S. and Tinsley, P.A. 2001. "Shifting Endpoints in the Term Structure of Interest Rates". *Journal of Monetary Economics* 47, 613- 652,

_____, 2005. "What Do You Expect? Imperfect Policy Credibility and Tests of the Expectations Hypothesis". *Journal of Monetary Economics* 52, 421- 447,

Kugler, P., 1988. "An Empirical Note on the Term Structure and Interest Rate Stabilization Policies", *Quarterly Journal of Economics* 103(4), 789- 792,

Longstaff, F. A. 2000. “The Term Structure of Very Short-term Rates: New Evidence for the Expectations Hypothesis”, *Journal of Financial Economics* 58, 397-415,

Macaulay, R.F. 1938. “Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856.” New York, National Bureau of Economic Research,

Mankiw, N.G., Shapiro, M.D., 1986. “Do We Reject Too Often? Small Sample Properties of Tests of Rational Expectations Models”. *Economics Letters* 20, 139-145,

Mankiw, N.G., Summers, L.H., 1984. “Do Long Term Interest Rates Overreact to Short- Term Interest Rates?”, *Brookings Papers on Economic Activity* 1, 223- 247,

Melino, A., 2001. “Estimation of a Rational Expectations Model of the Term Structure”. *Journal of Empirical Finance* 8, 639-688,

Newey, W. K., McFadden, D. L., 1994. “Large Sample Estimation and Hypothesis Testing”. *Handbook of Econometrics* 4. Elsevier Science. Amsterdam, The Netherlands, 2111-2245.

Newey, W. K. and K. D. West., 1987, “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix”. *Econometrica* 55, 703-708,

Richardson, M., & Stock, J.H., 1989. “Drawing Inferences from Statistics Based on Multiyear Asset Returns”, *Journal of Financial Economics* 25, 323- 348,

Roberds, W., Runkle, D., and Whiteman, C. H., 1996. "A Daily View of Yield Spreads and Short-term Interest Rate Movements", *Journal of Money, Credit and Banking* 28, 34-53,

Rudebusch, G.D., 1995. “Federal Reserve Interest Rate Targeting, Rational Expectations, and the Term Structure”, *Journal of Monetary Economics* 35, 245- 274,

Sargent., T. 1979. “A Note on Maximum Likelihood Estimation of the Rational Expectations Model of the Term Structure”, *Journal of Monetary Economics* 5, 133- 143,

Sarno, L., Thornton, D.L., and Valente, G., 2006. “New Evidence on the Expectations Hypothesis of the Term Structure of Bond Yields”. *Journal of Financial and Quantitative Analysis*, forthcoming,

Shea, G., 1992. “Benchmarking the Expectations Hypothesis of the Interest Rate Term Structure: An Analysis of Cointegrating Vectors”, *Journal of Business & Economic Statistics*, 10(3), 347-66,

Shiller, R.J., 1979. “The Volatility of Long Term Interest Rates and Expectations Models of the Term Structure”. *Journal of Political Economy* 87, 1190- 1219,

_____, 1982. “Alternative Tests of Rational Expectations Models: The Case of the Term Structure”, *Journal of Econometrics* 16, 71- 87,

Shiller, R. J., Campbell J. Y., Schoenholtz K. L., 1983. “Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates”, *Brookings Papers on Economic Activity* 1, 173-217,

Simon, D.P. 1990. “Expectations and the Treasury Bill-Federal Funds Rate over Recent Monetary Policy Regimes”, *Journal of Finance* 45, 467-477,

Stambaugh, R.F. 1988. “The Information in the Forward Rates: Implications for Models of the Term Structure”, *Journal of Financial Economics* 21, 41-70,

Taylor, M. P. 1992. “Modelling the Yield Curve”, *Economic Journal* 102, 524-537,

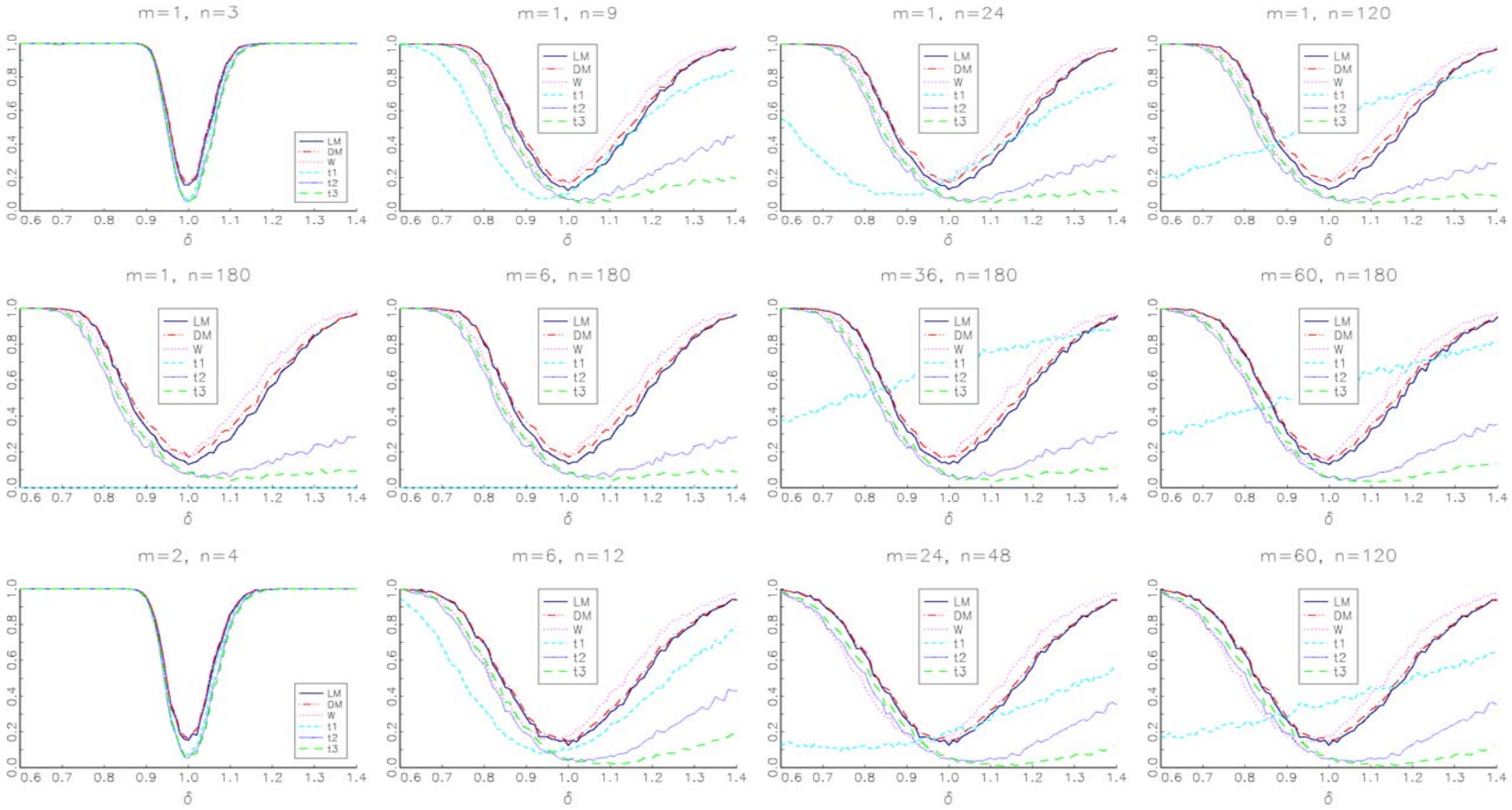
Thornton, D.L. 2006. “Tests of the Expectations Hypothesis: Resolving Campbell-Shiller Paradox”, *Journal of Money, Credit and Banking* 38(2), 511-542,

Tzavalis, E., Wickens, M., 1997. “Explaining the Failures of the Term Spread Models of the Rational Expectations Hypothesis of the Term Structure”, *Journal of Money, Credit and Banking* 29, 364-380,

Figure 1. Size and power: Asymptotic critical value (DGP2: US VAR(1) mean, US GARCH(1,1) residual)

T=300

Panel A: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis



Panel B: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

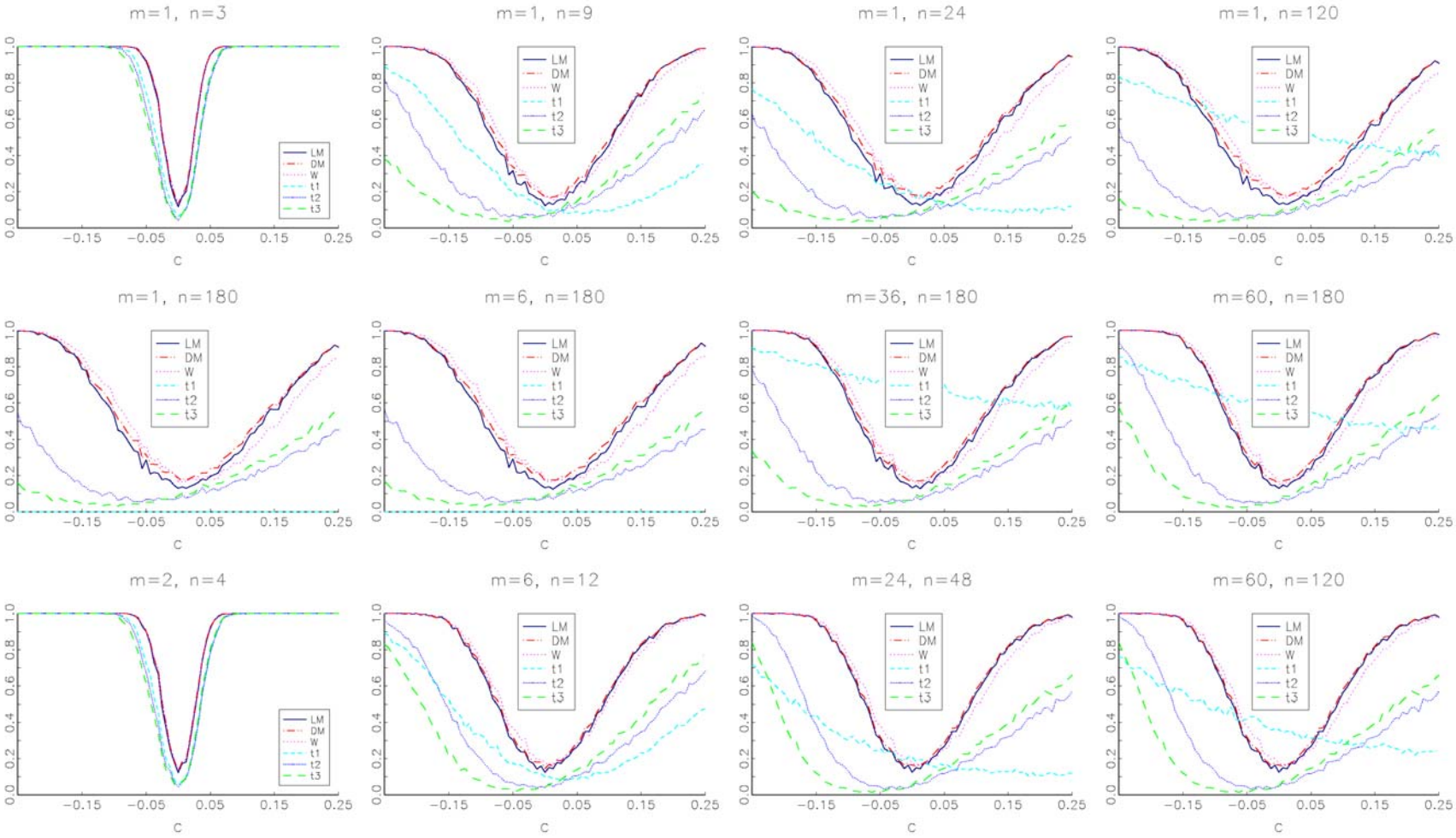
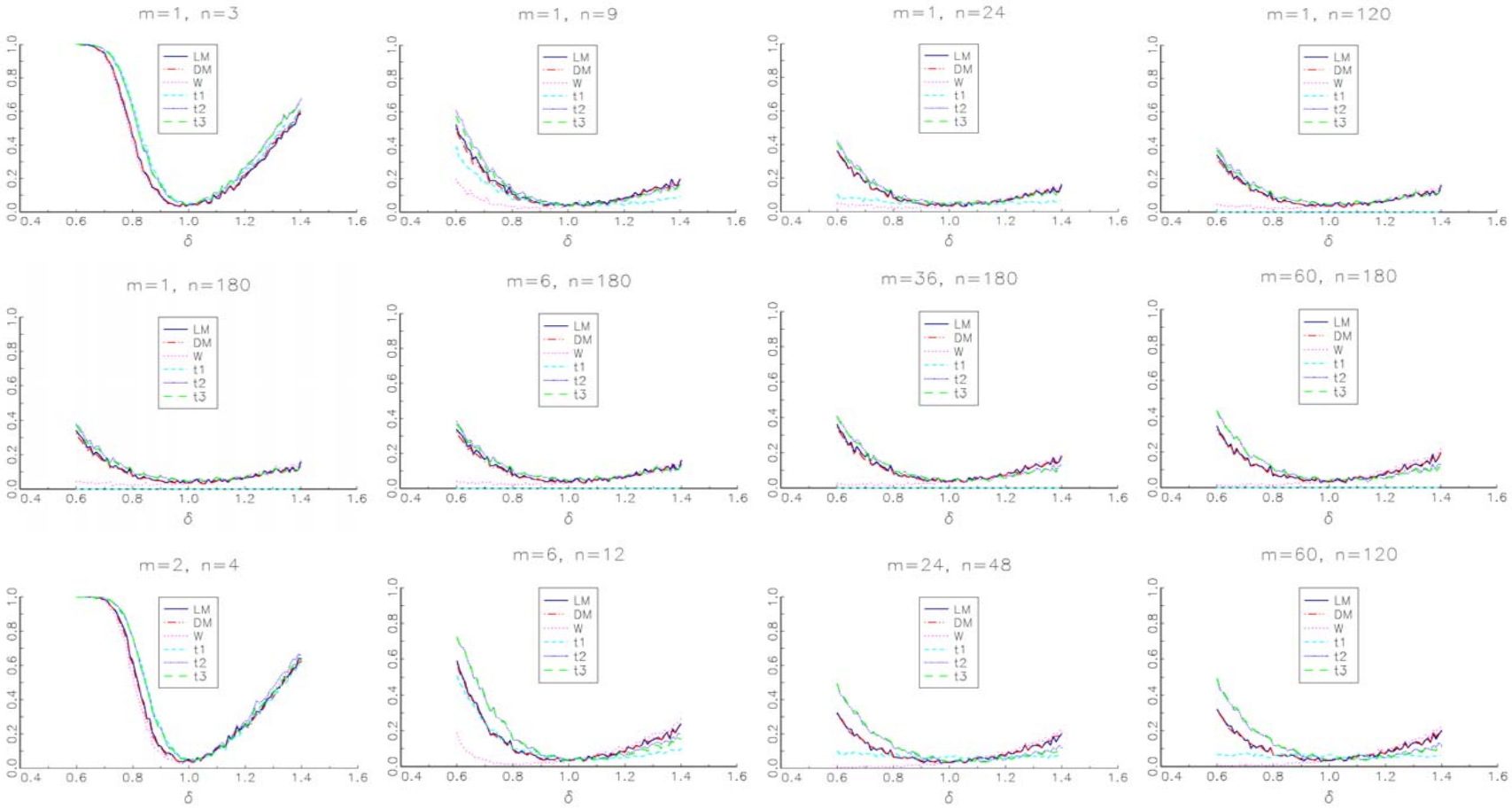


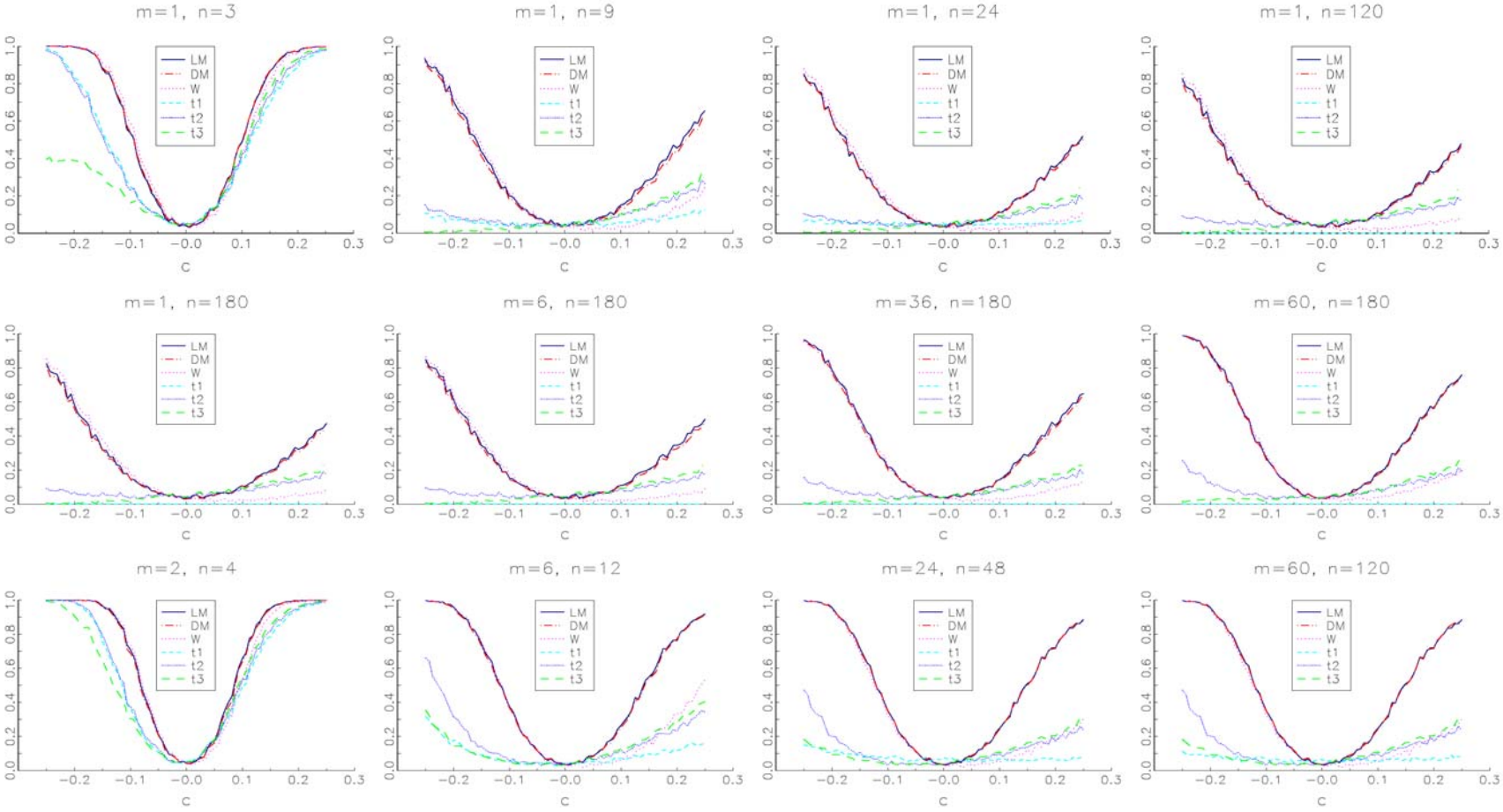
Figure 2. Size and Power: Bootstrap critical value (DGP1: UK VAR(1) mean, UK GARCH(1,1) residual)

T=150

Panel A1: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

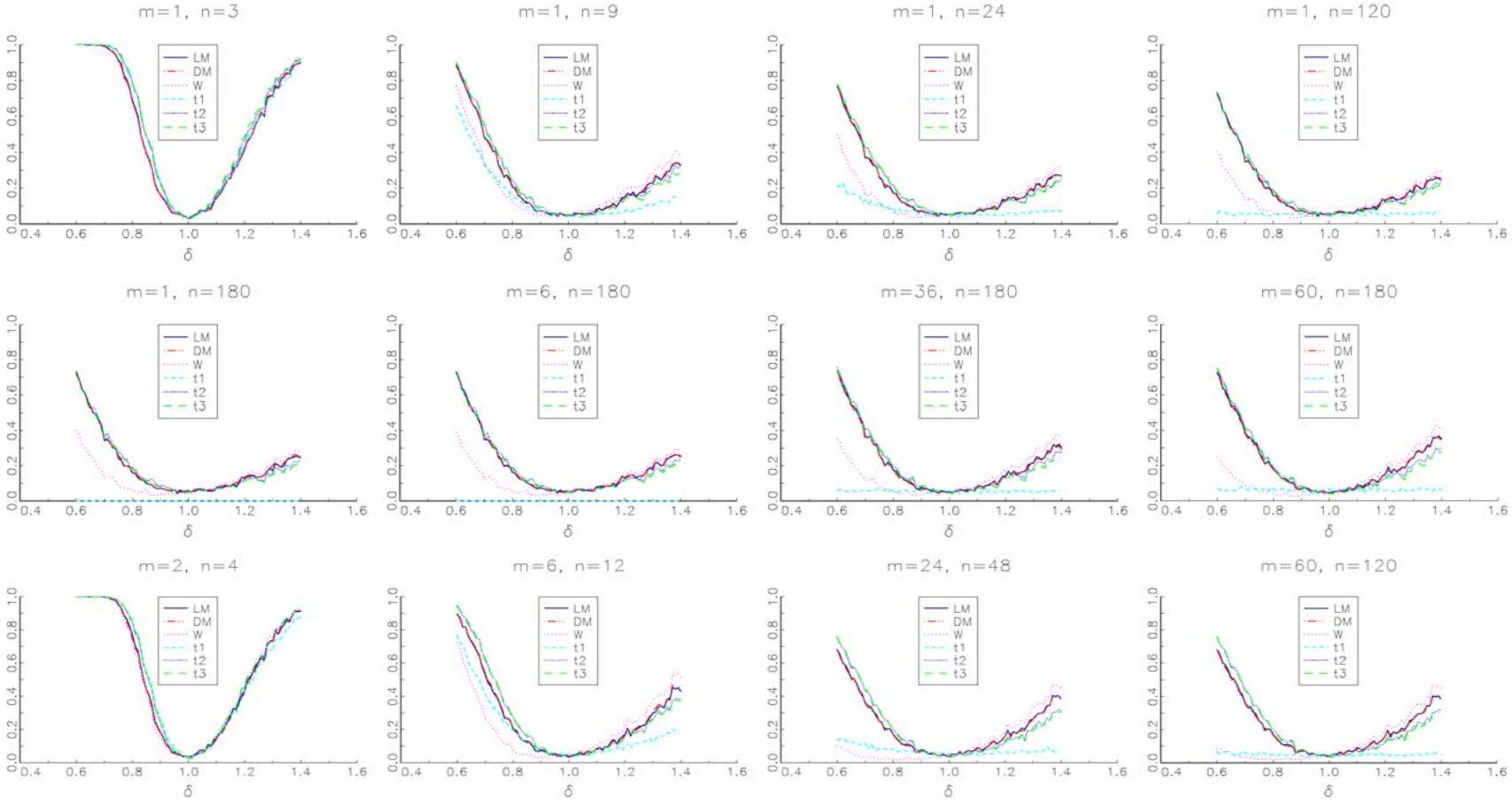


Panel B1: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

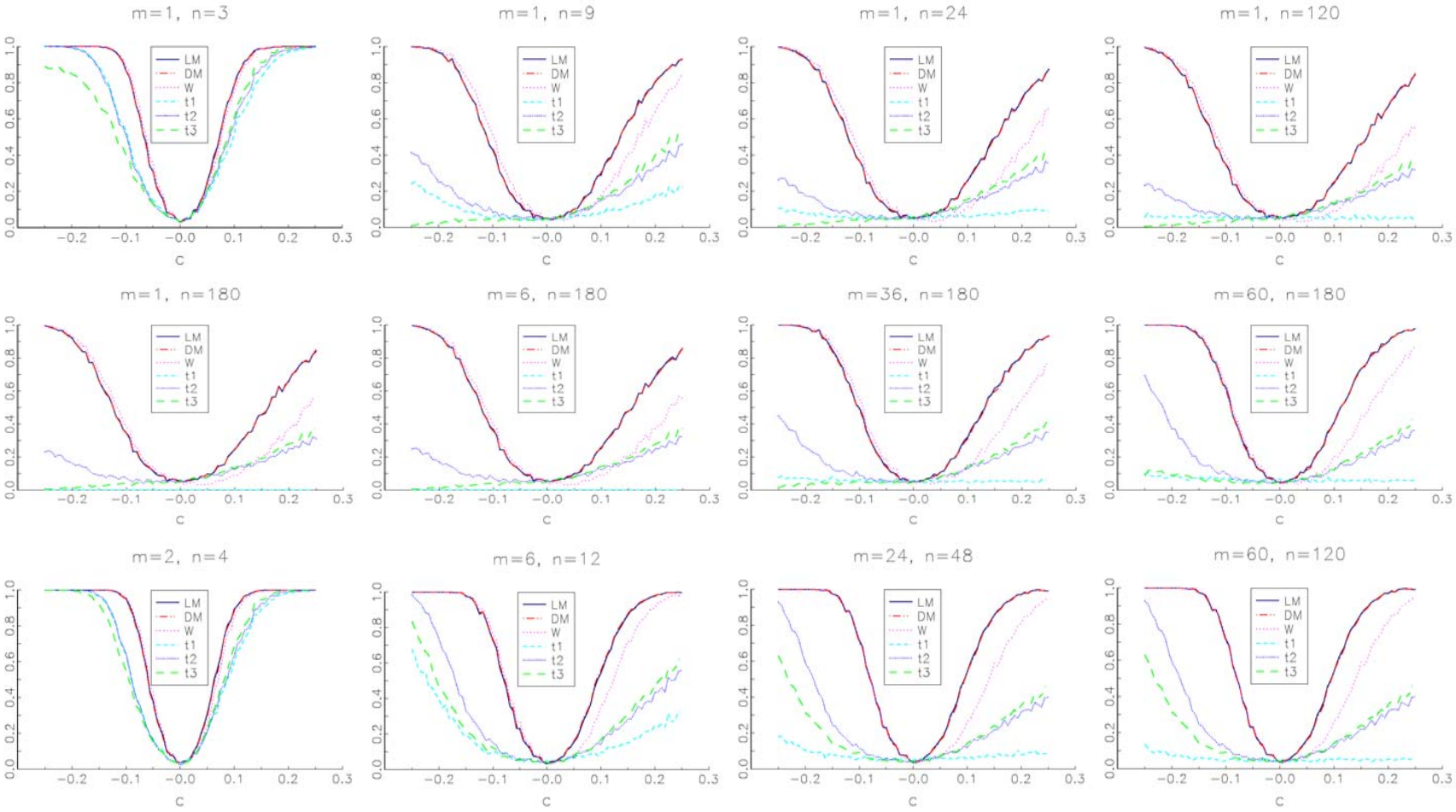


T=300

Panel A2: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

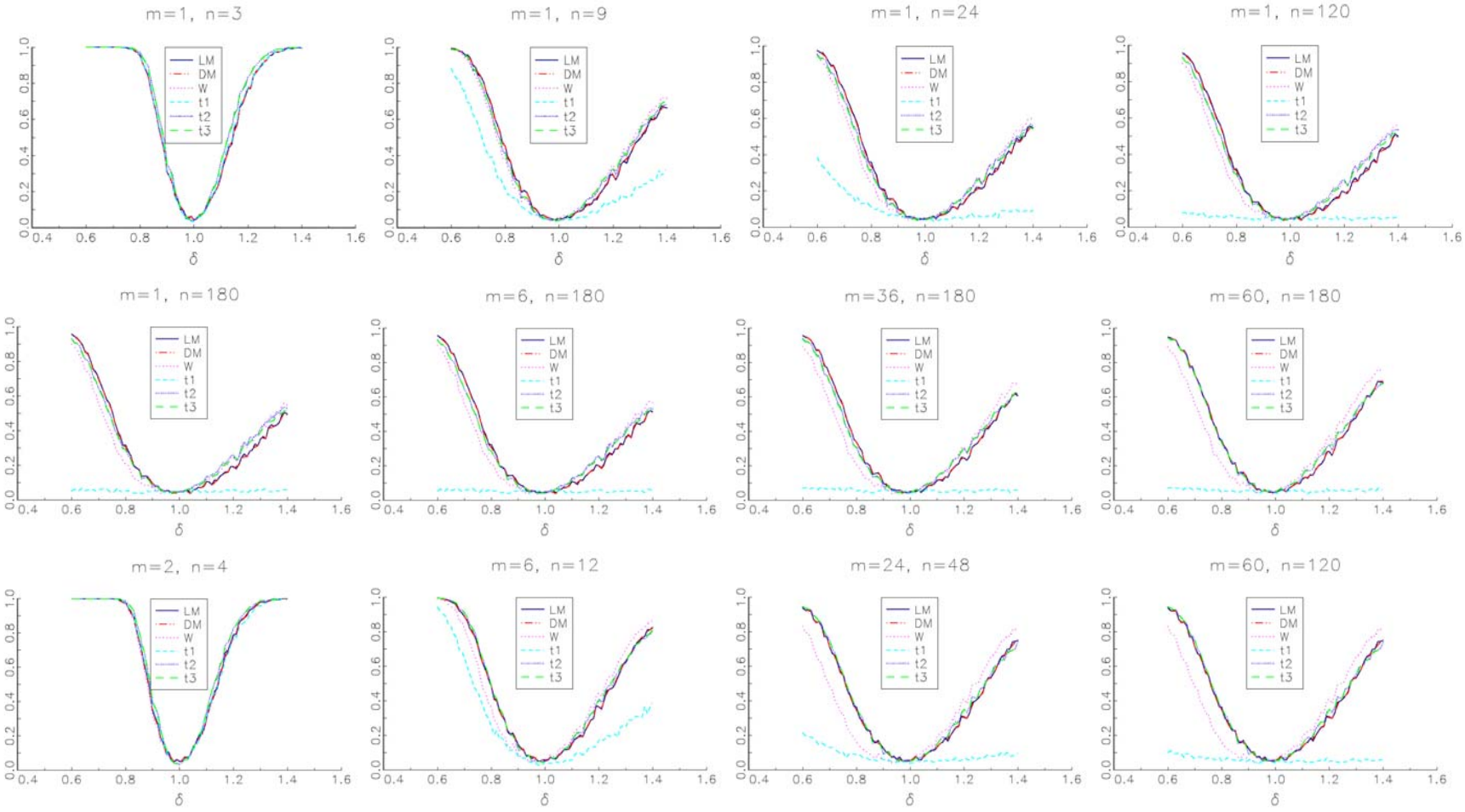


Panel B2: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

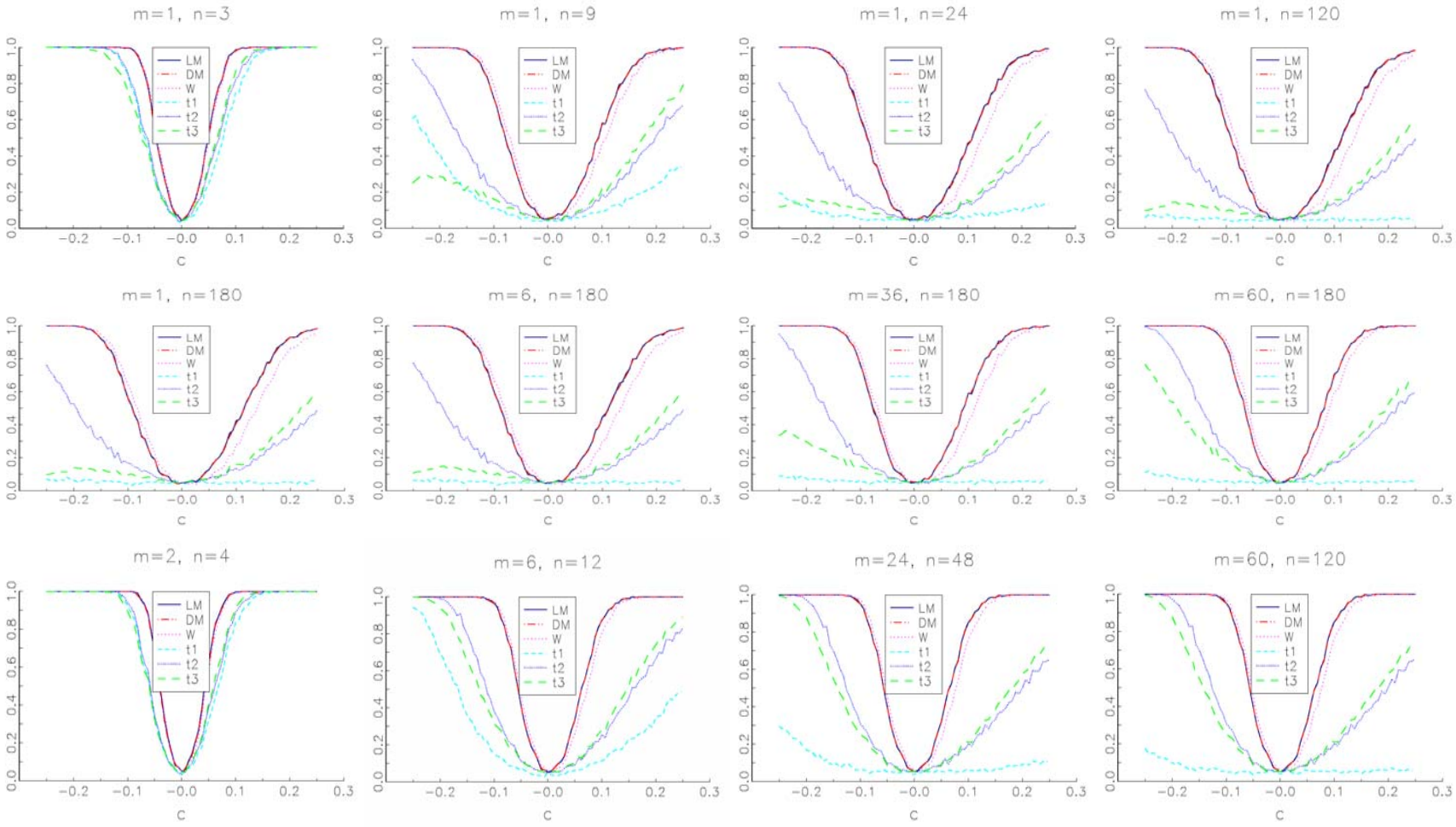


T=600

Panel A3: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

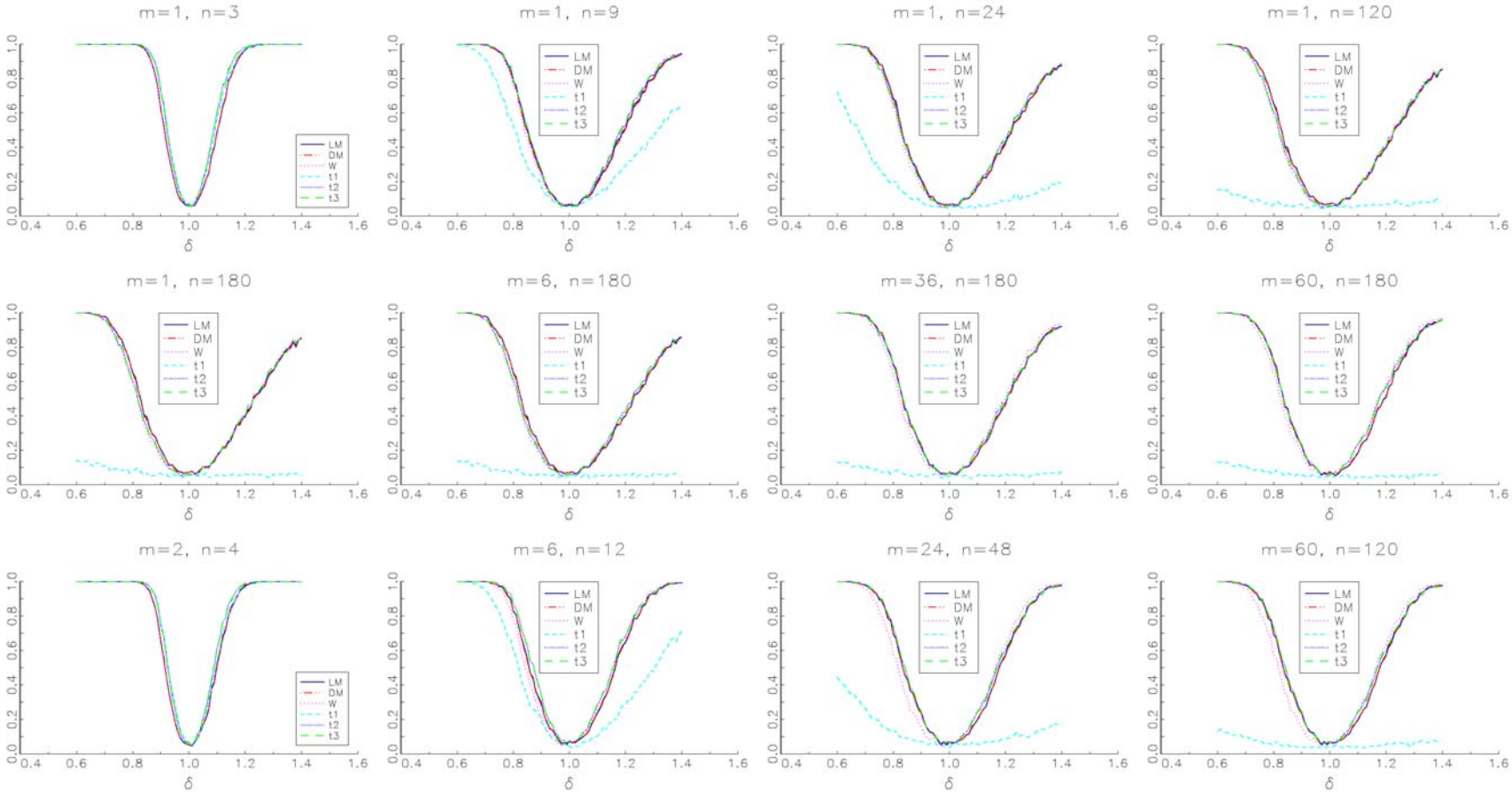


Panel B3: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis



T=1500

Panel A4: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis



Panel B4: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

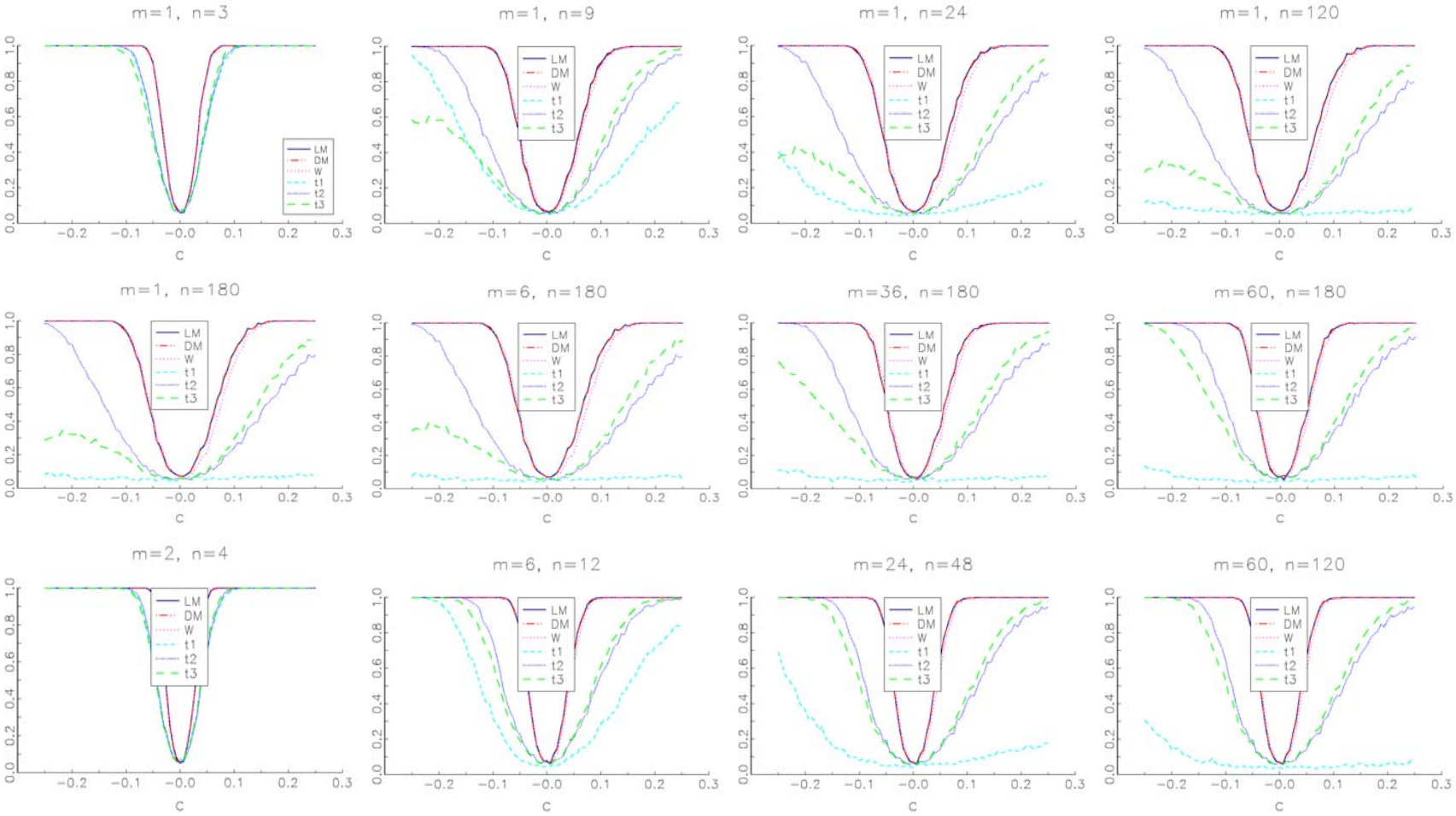
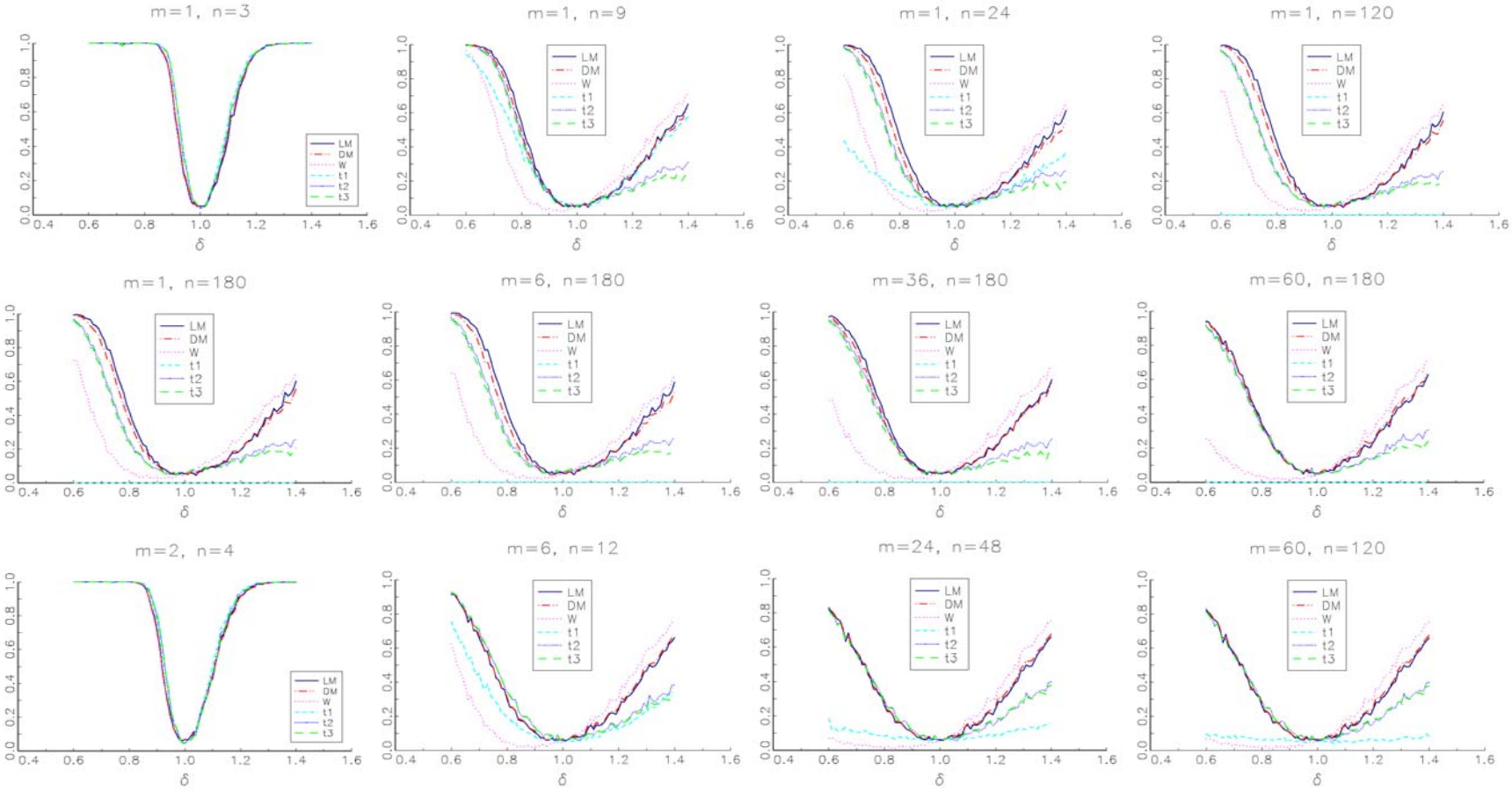


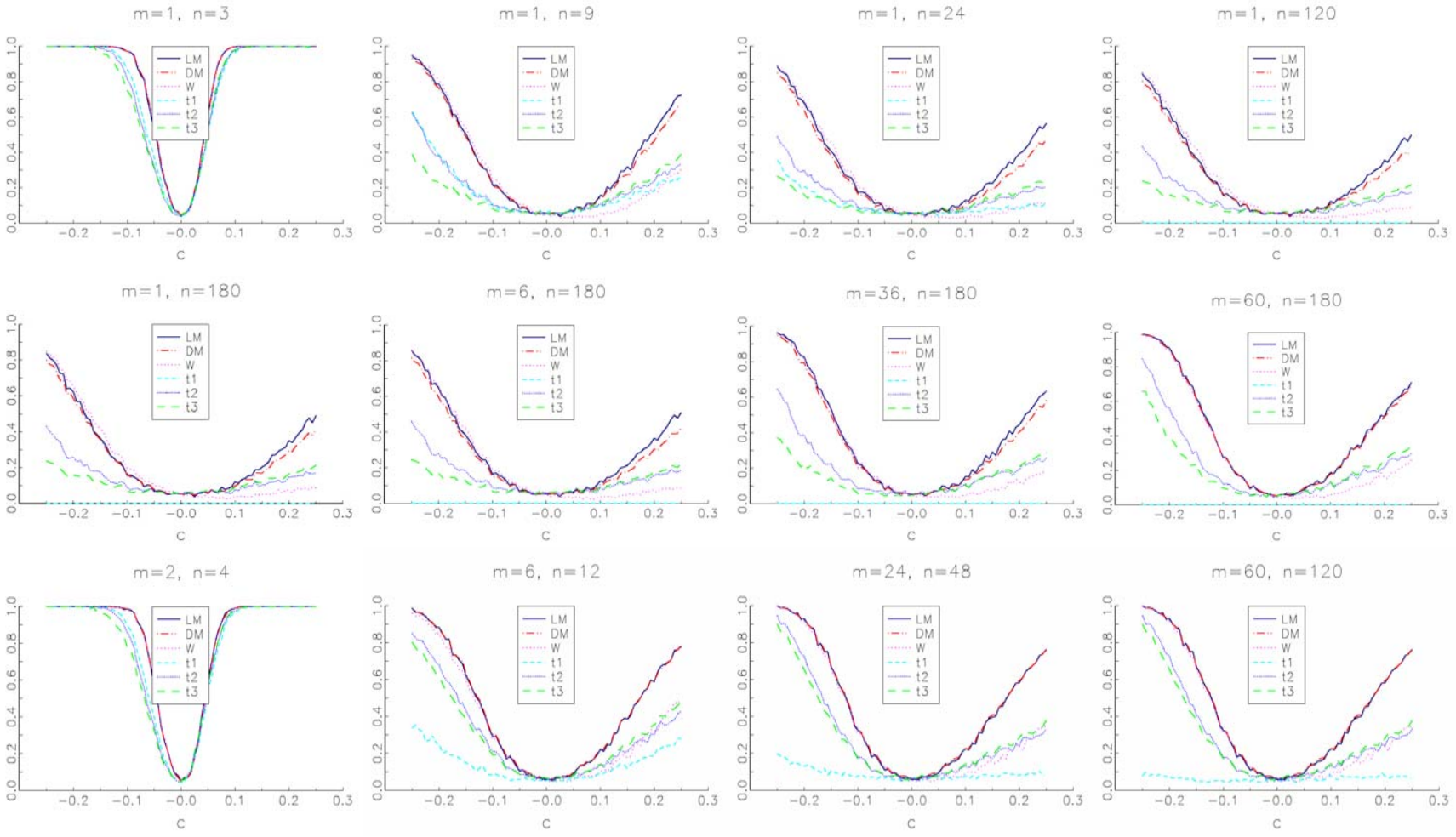
Figure 3. Size and Power: Bootstrap critical value (DGP2: US VAR(1) mean, US GARCH(1,1) residual)

T=150

Panel A1: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

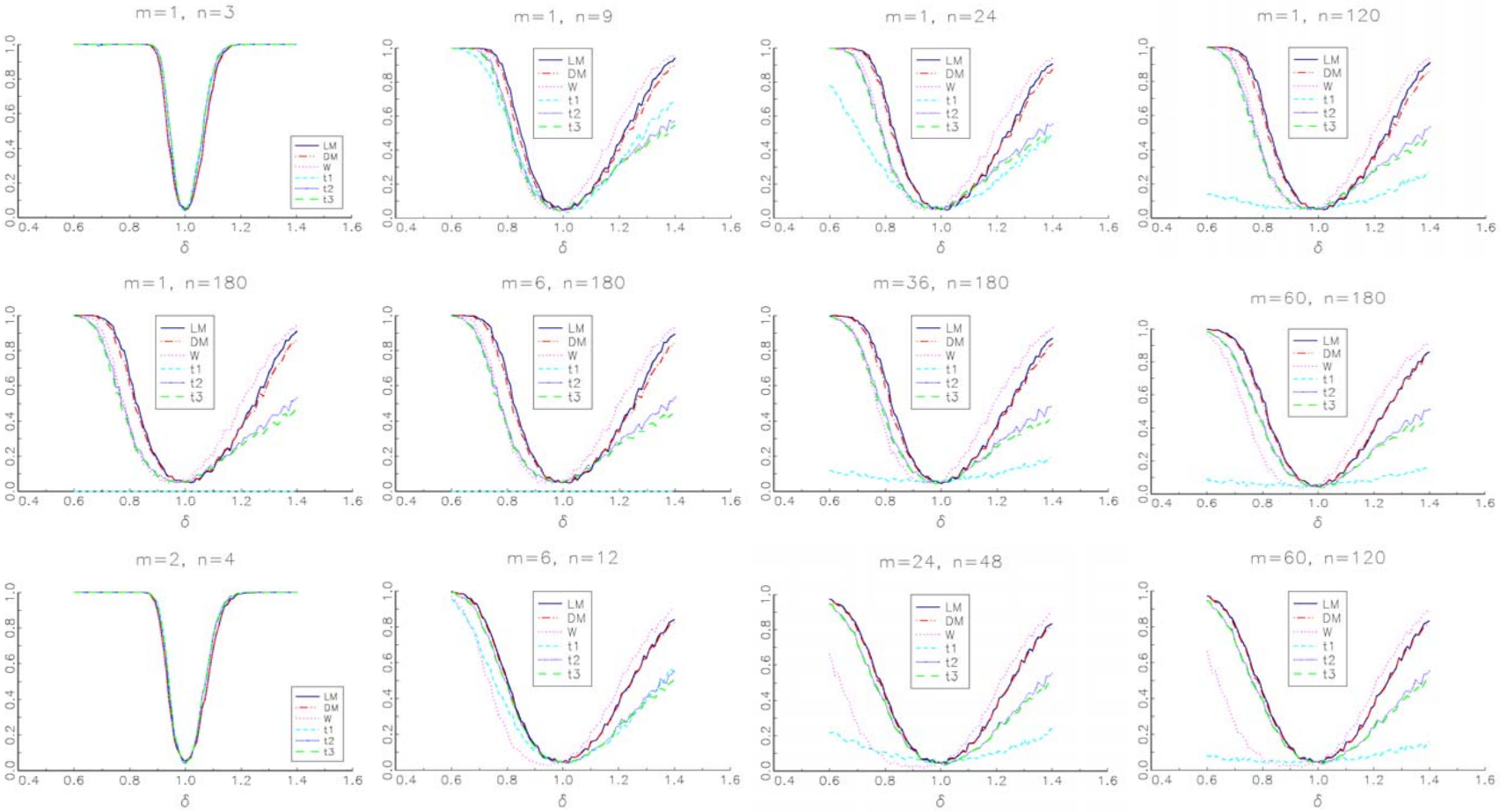


Panel B1: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

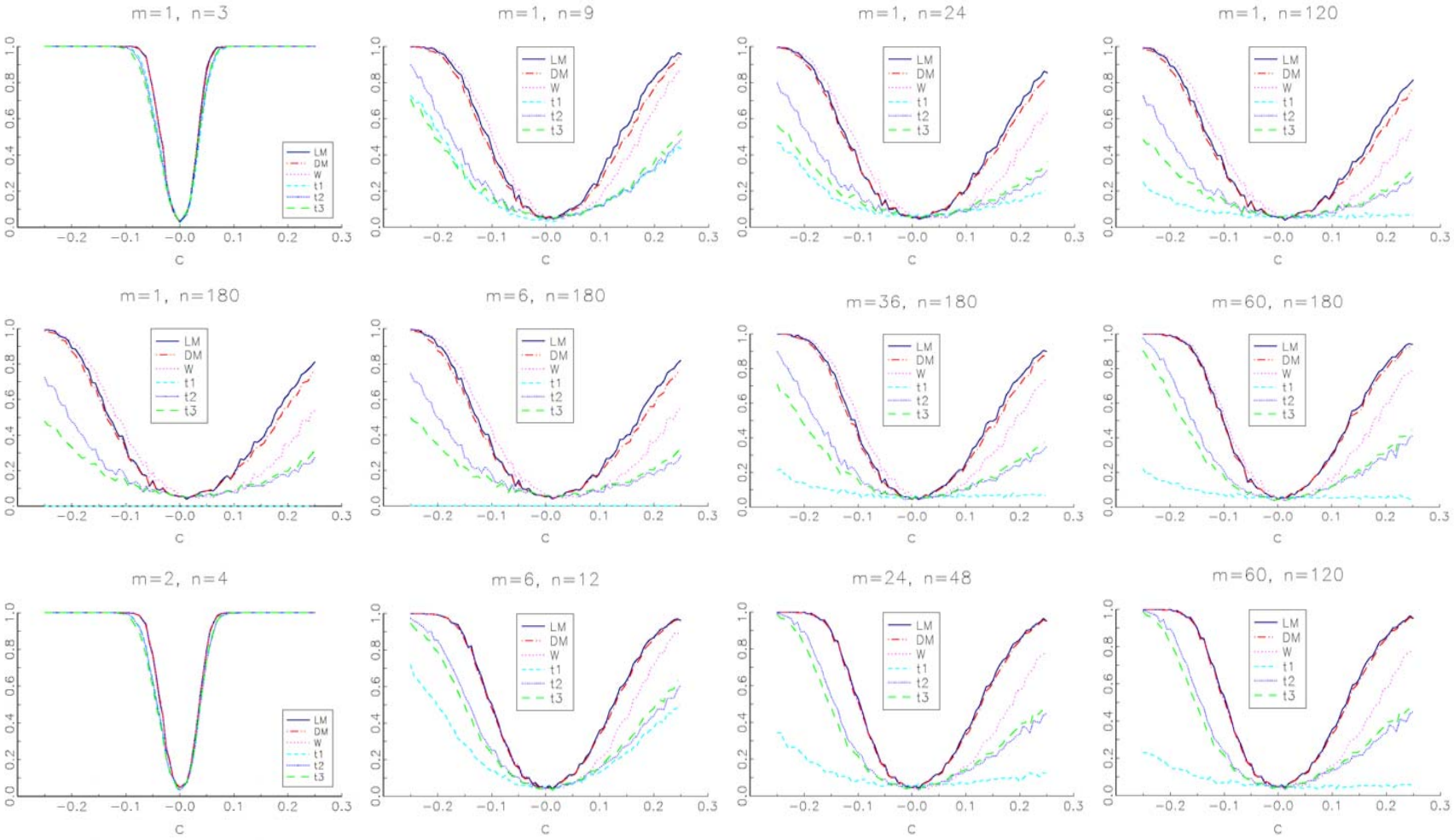


T=300

Panel A2: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

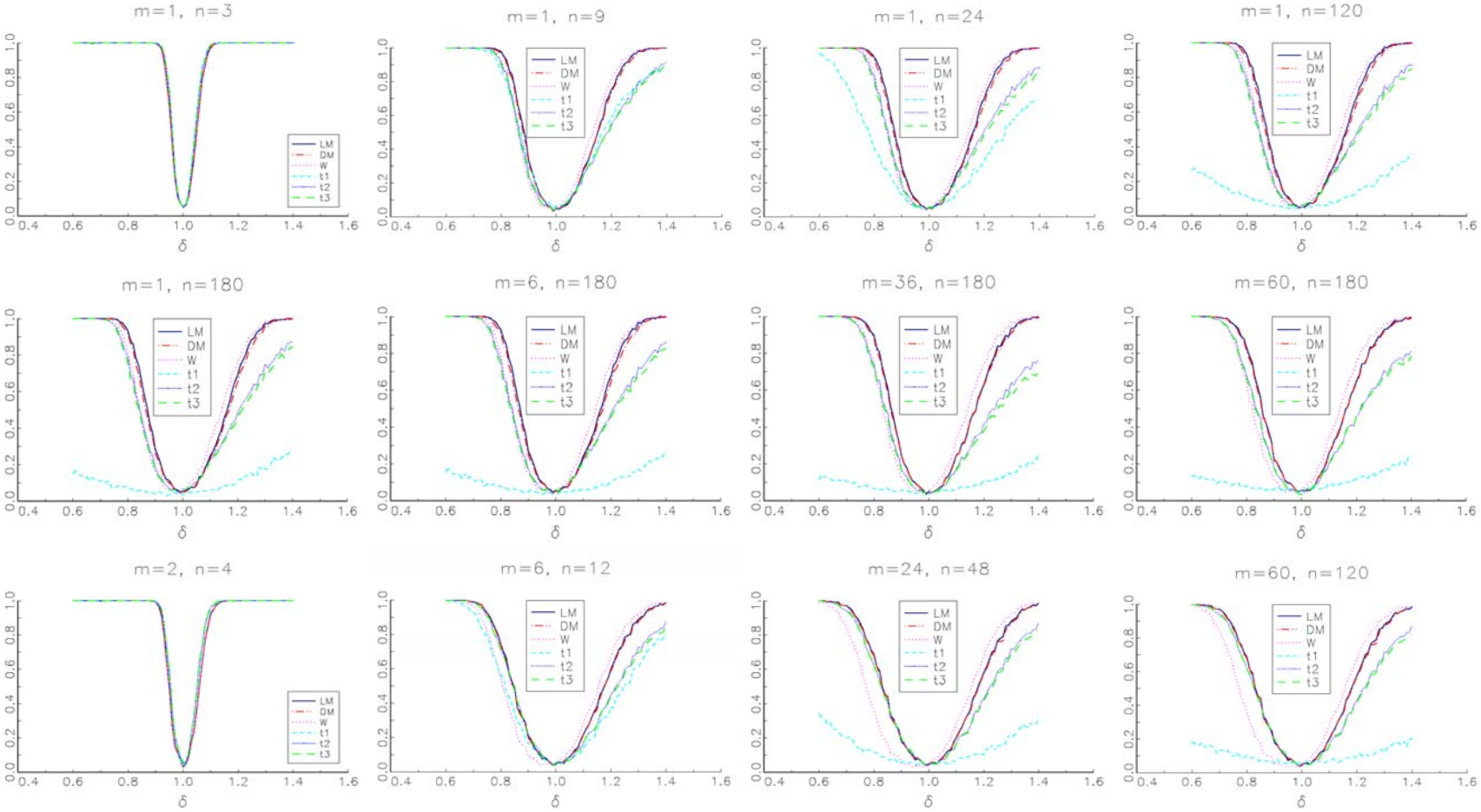


Panel B2: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

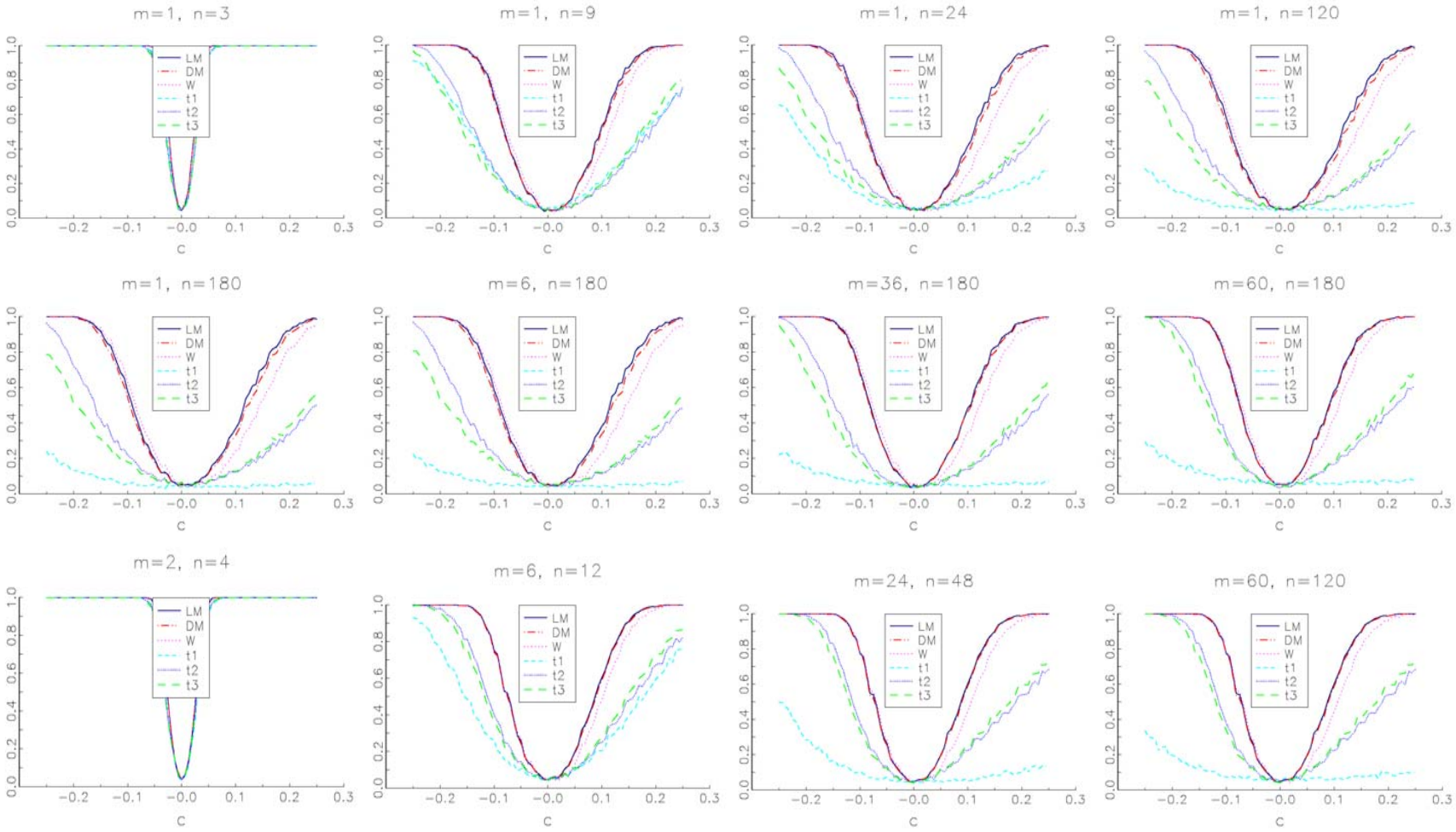


T=600

Panel A3: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

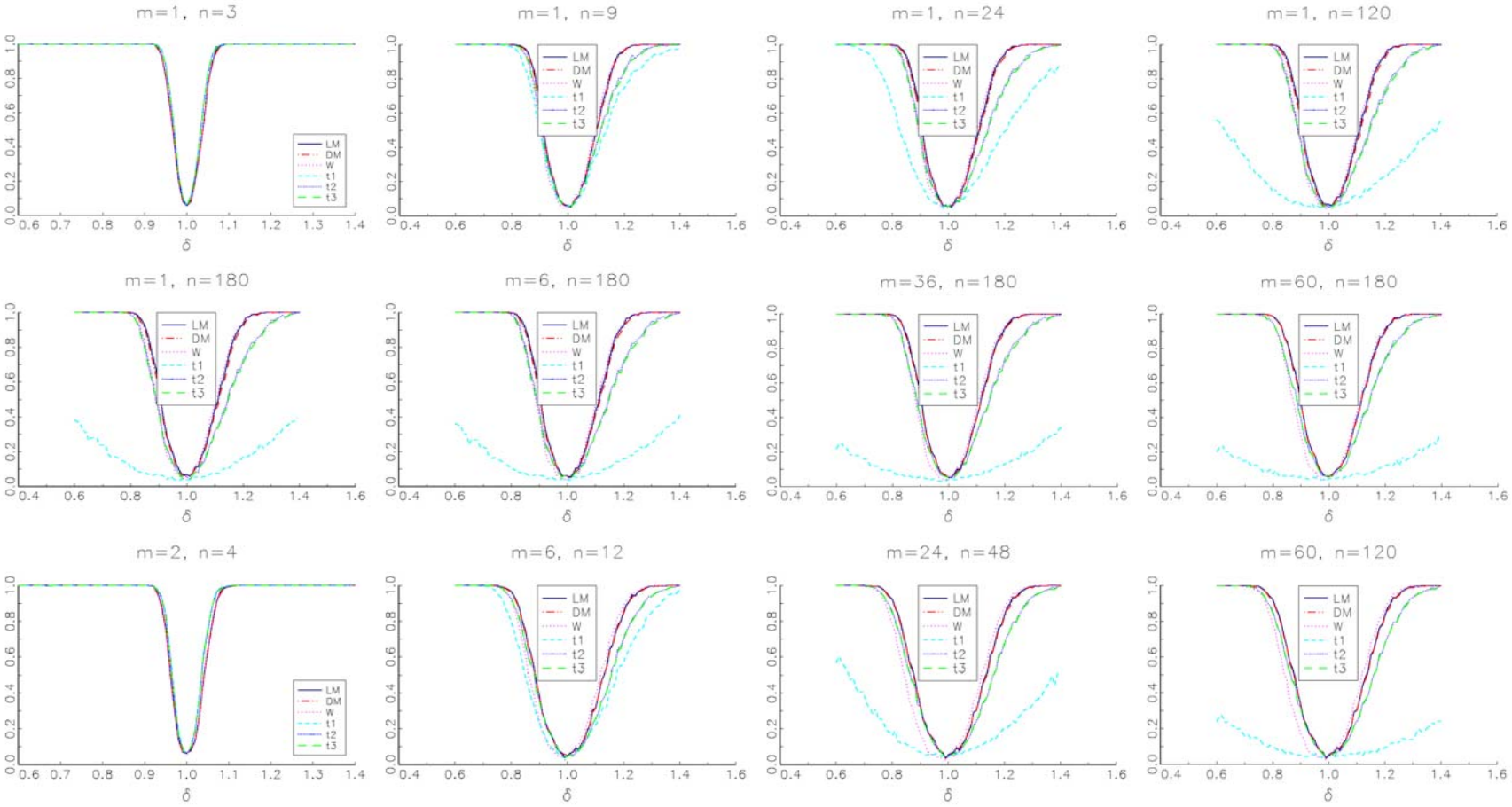


Panel B3: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis



T=1500

Panel A4: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis



Panel B4: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

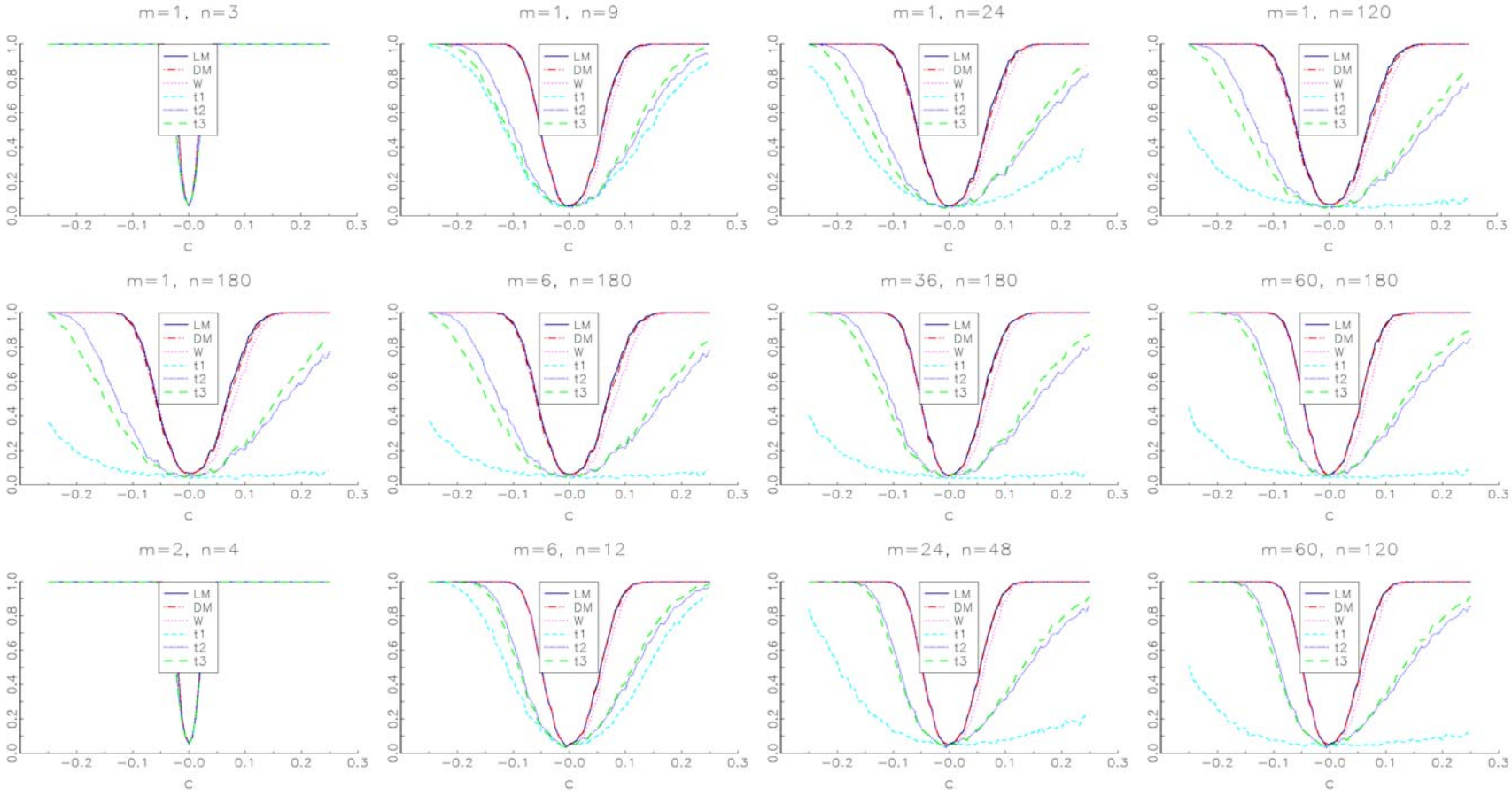
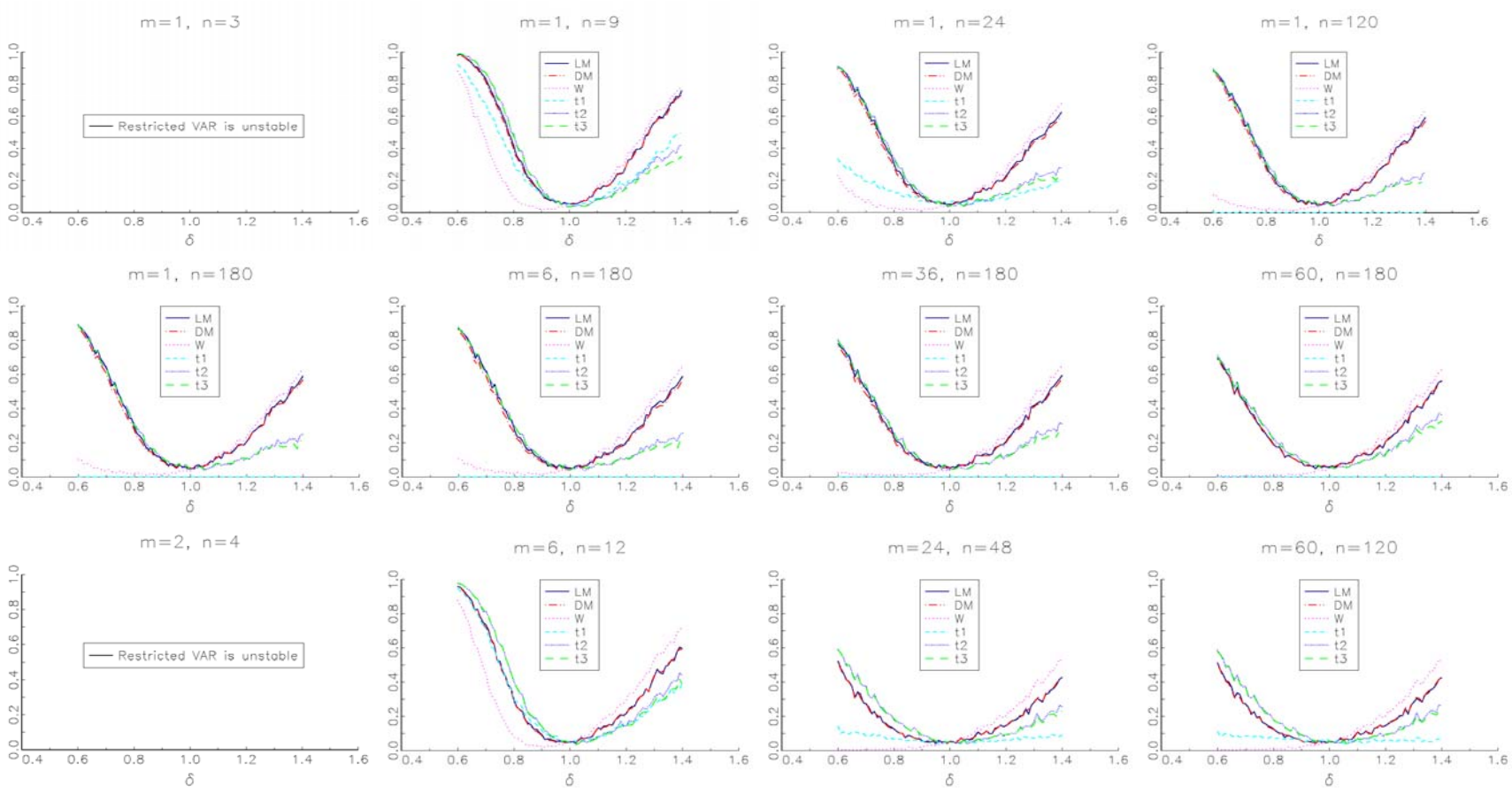


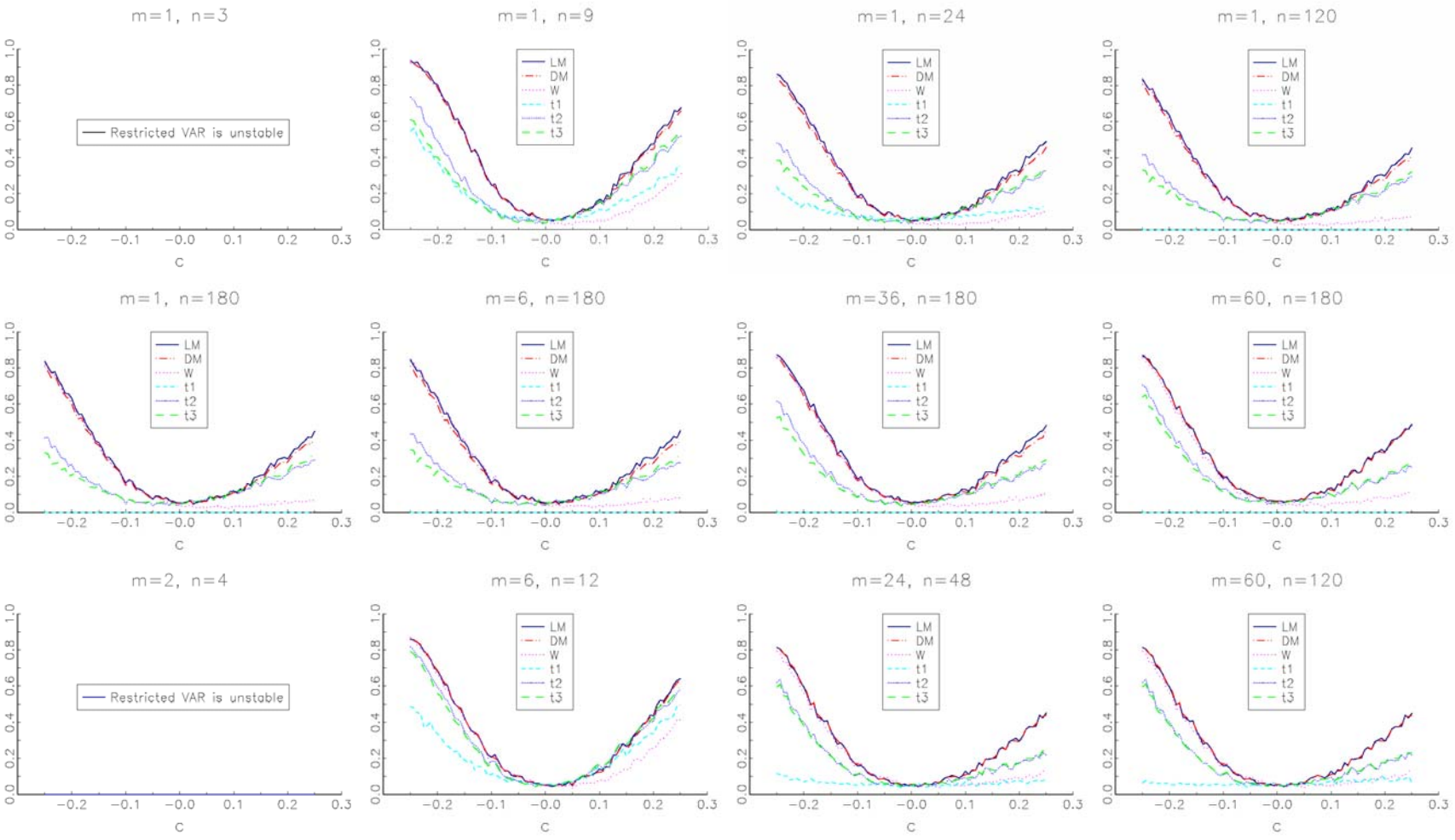
Figure 4. Size and Power: Bootstrap critical value (DGP3: UK VAR(1) mean, US GARCH(1,1) residual)

T=150

Panel A1: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

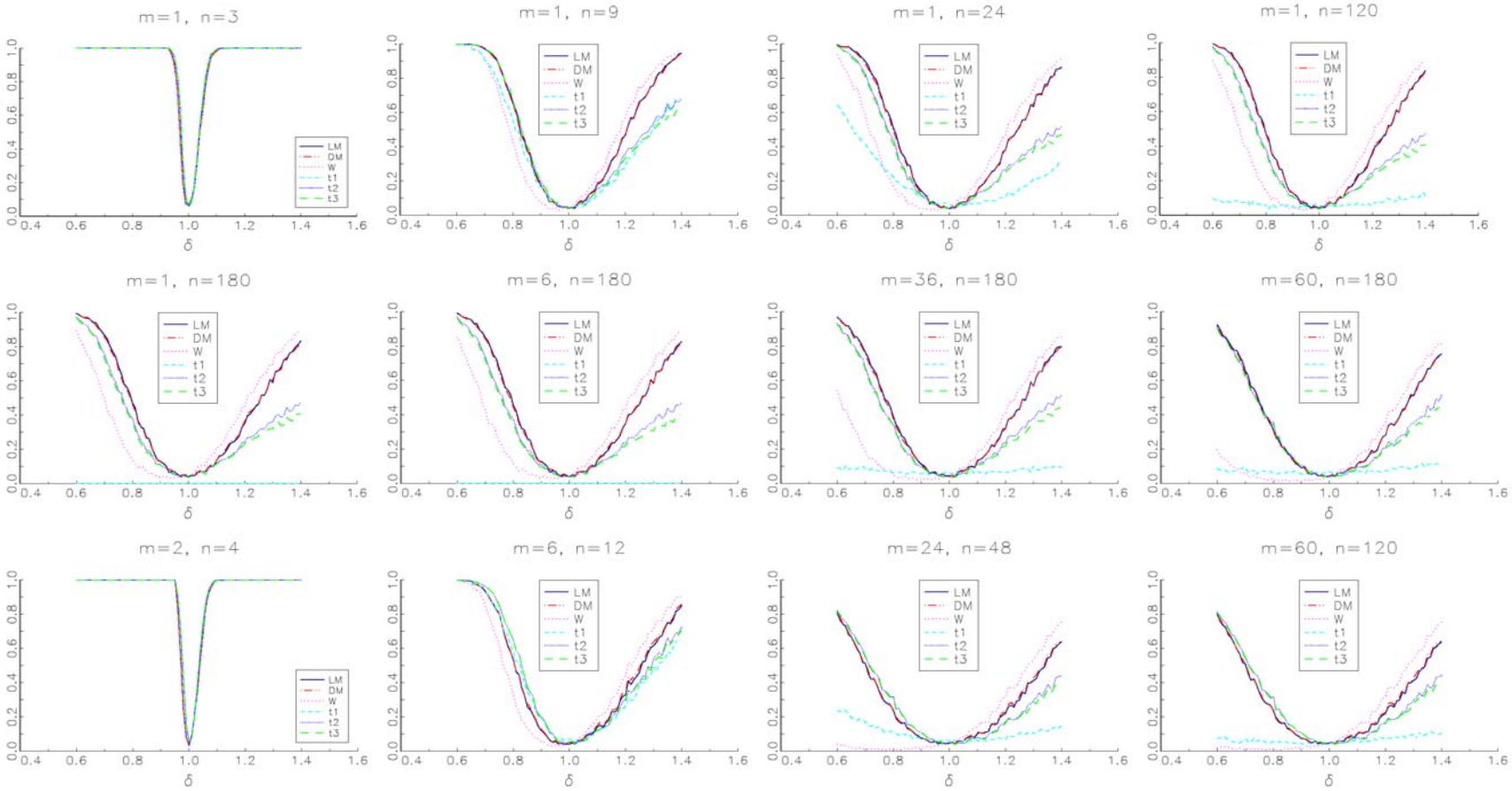


Panel B1: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

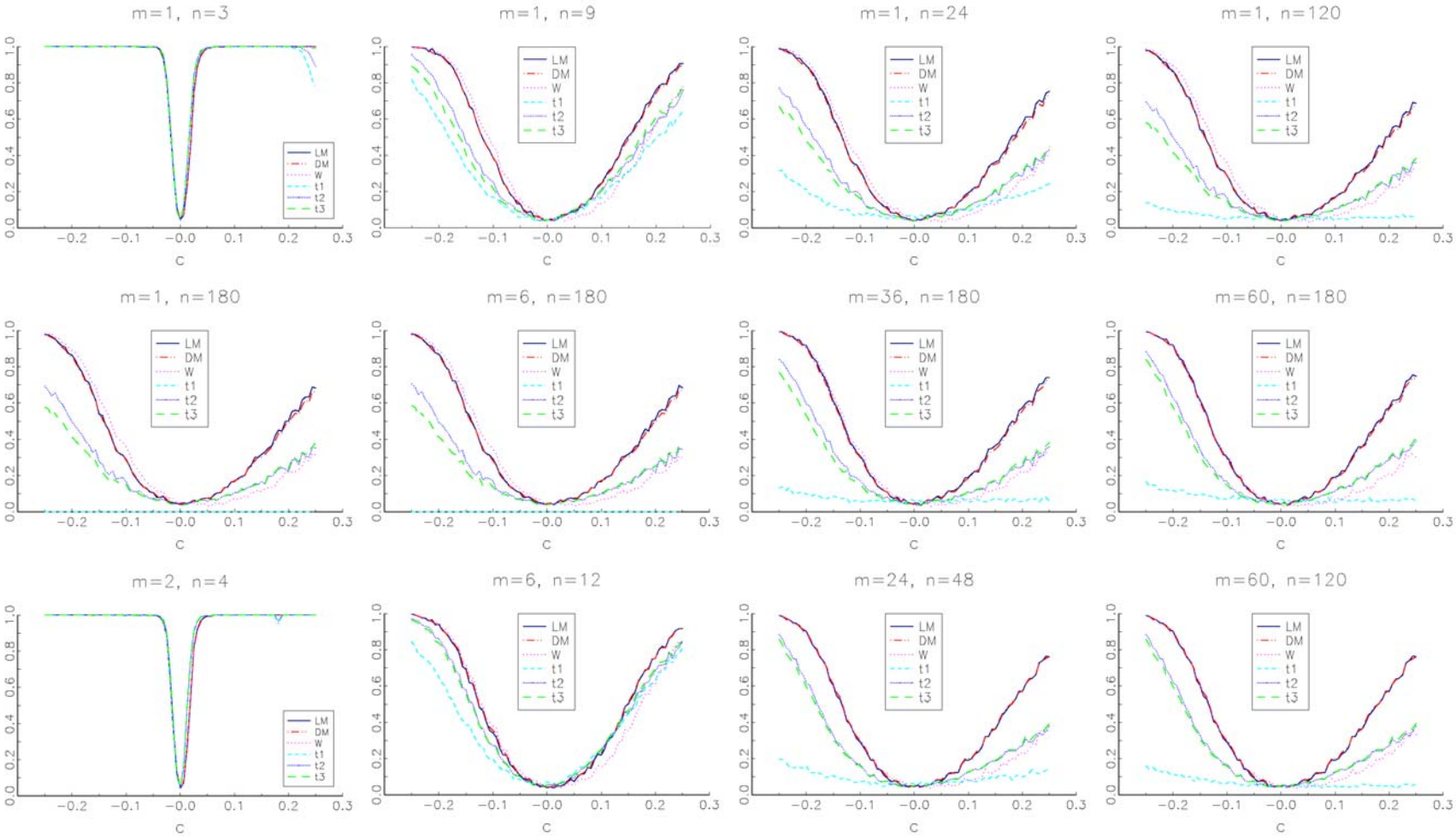


T=300

Panel A2: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

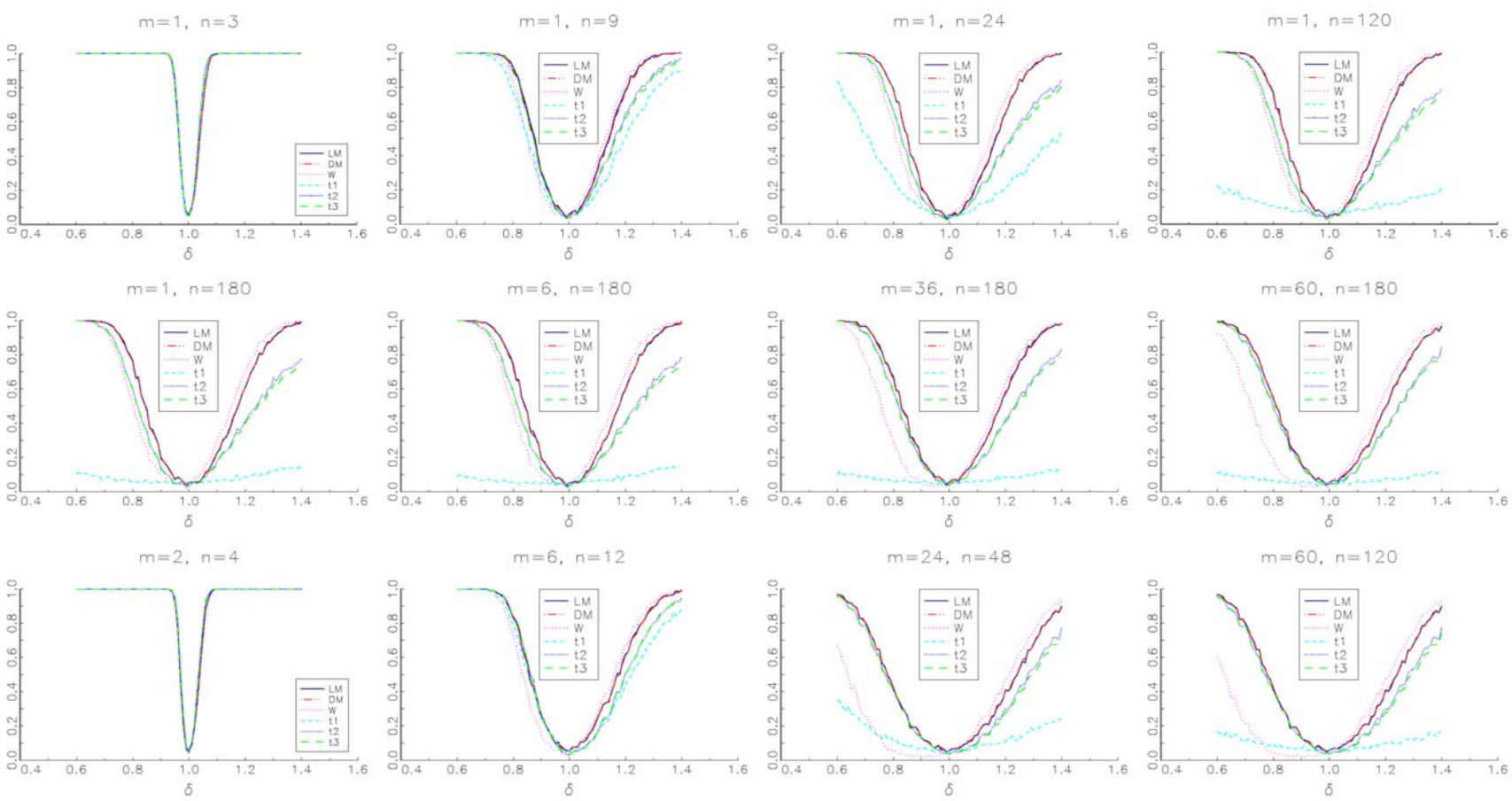


Panel B2: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis



T=600

Panel A3: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis



Panel B3: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

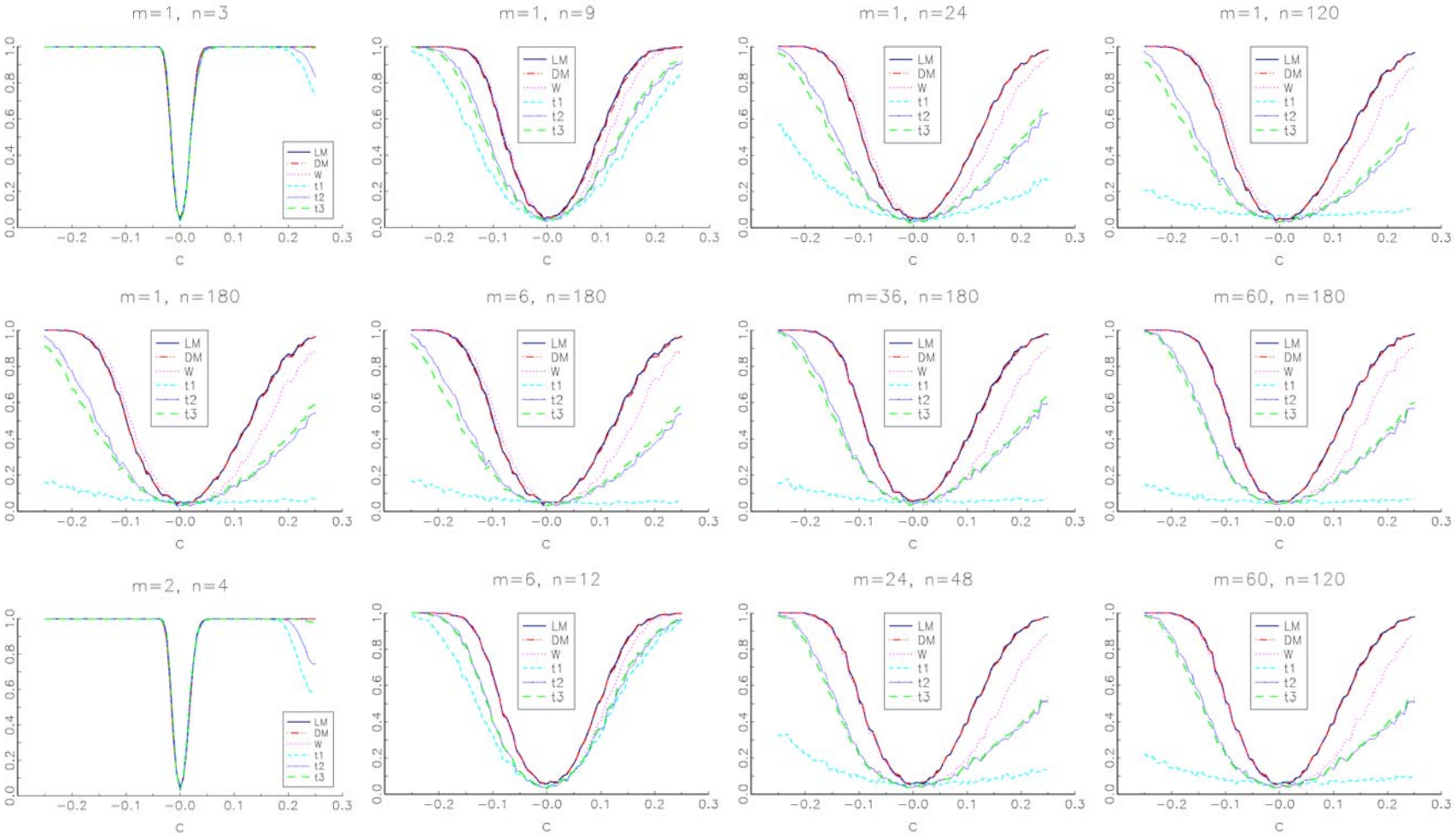
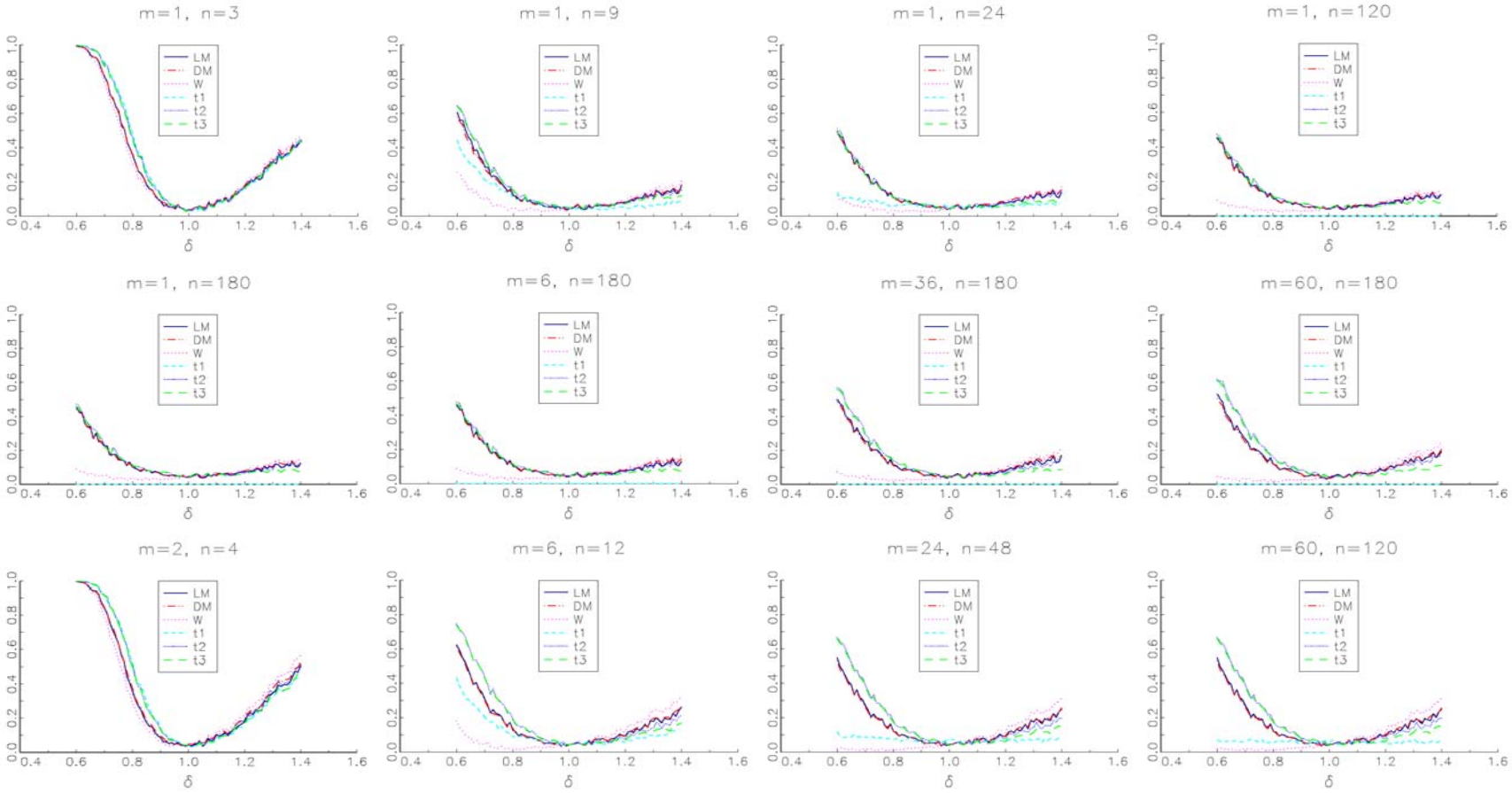


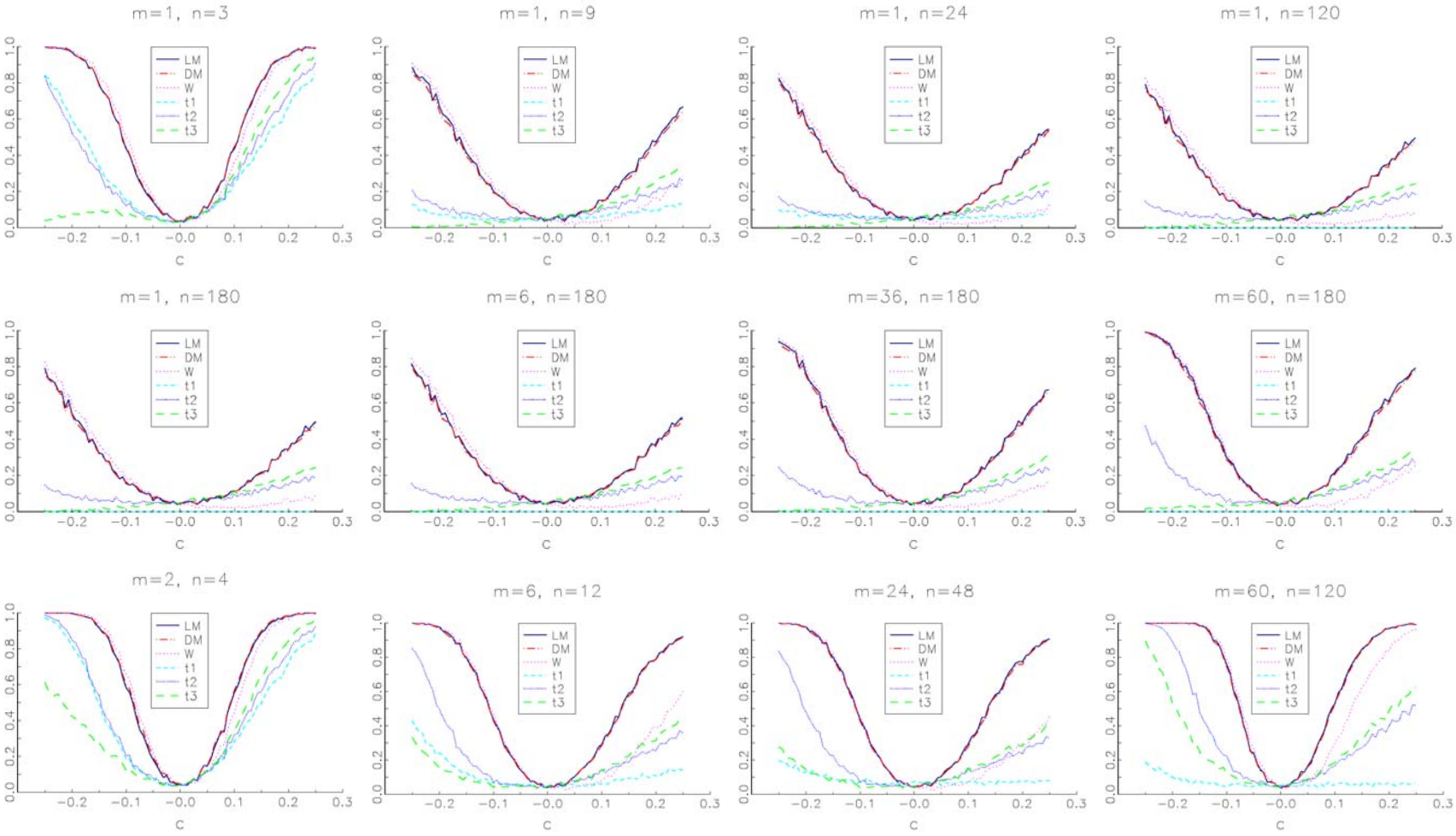
Figure 5. Size and Power: Bootstrap critical value (DGP4: US VAR(1) mean, UK GARCH(1,1) residual)

T=150

Panel A1: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

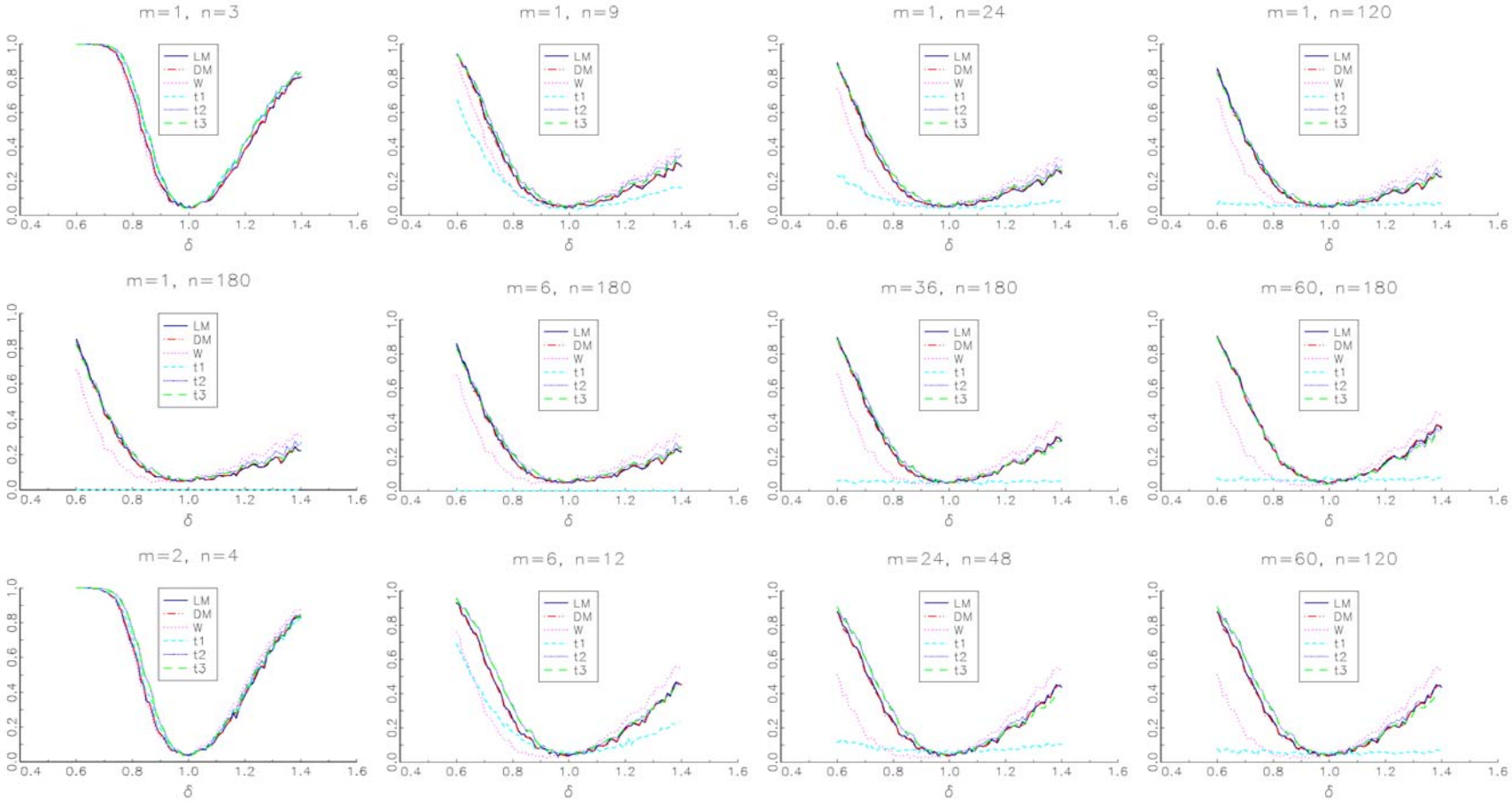


Panel B1: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis

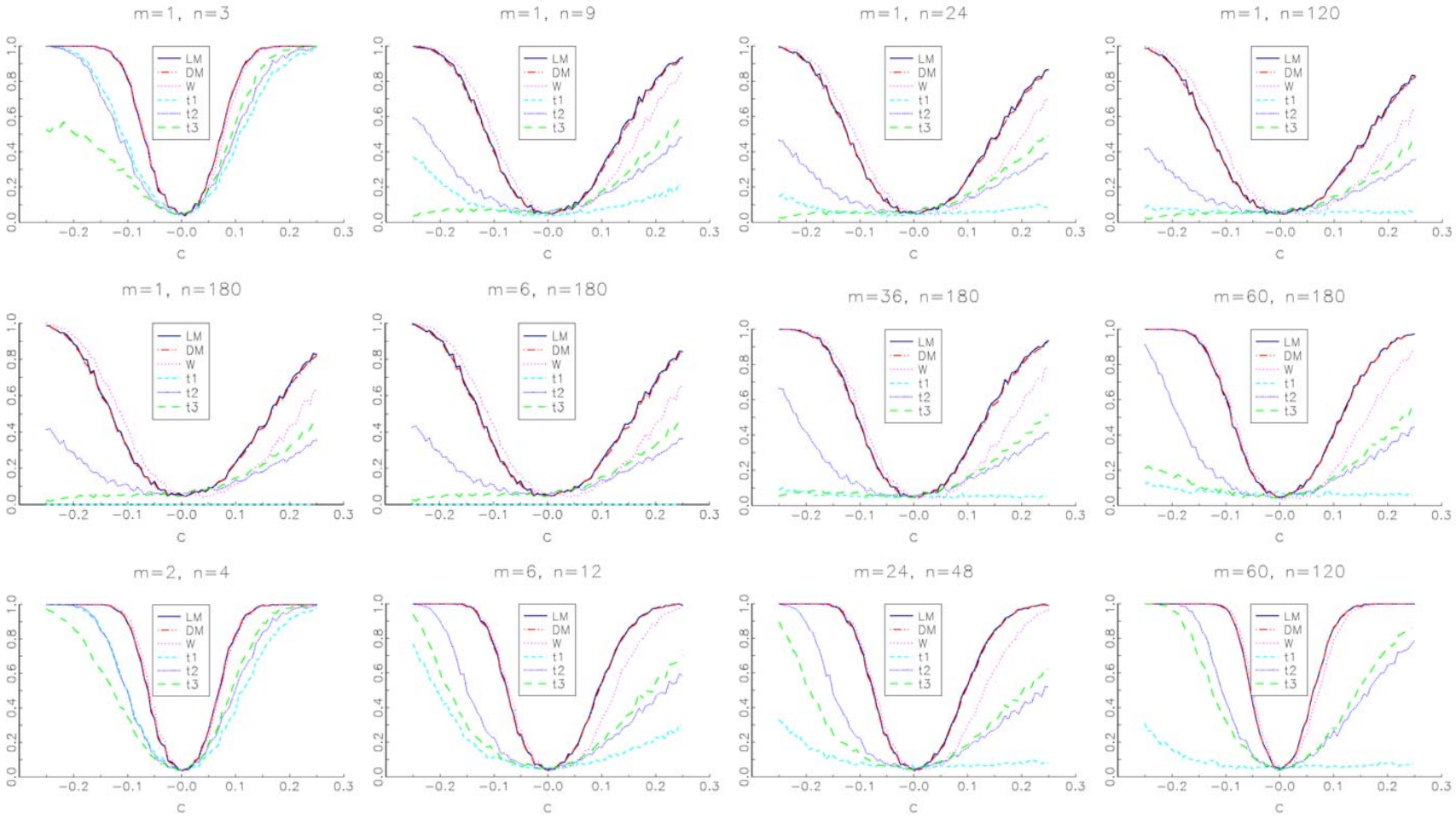


T=300

Panel A2: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis

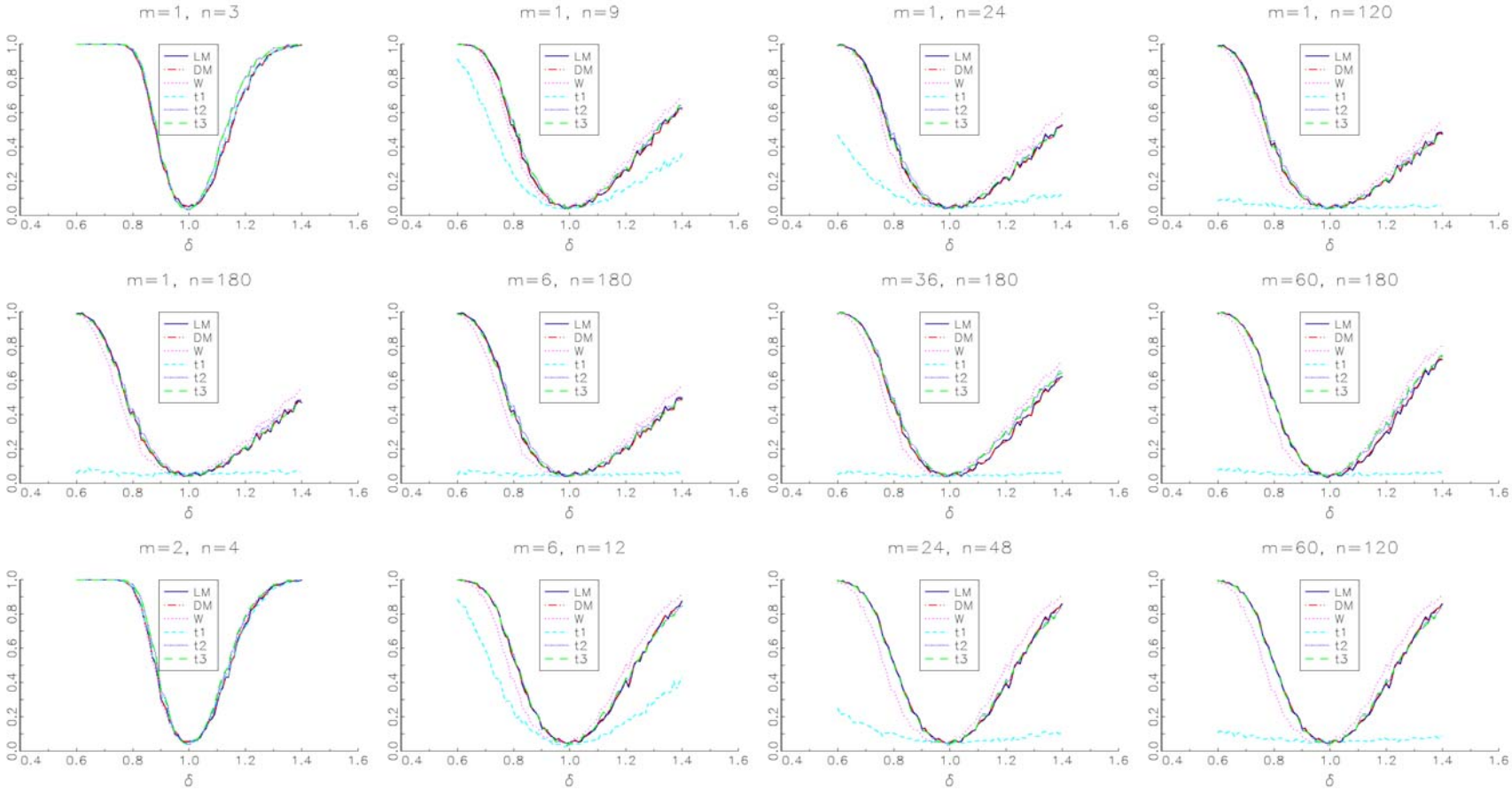


Panel B2: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis



T=600

Panel A3: H_O: ET (δ=1) , H_A: Over (δ>1)/Under (δ<1) Reaction Hypothesis



Panel B3: H_O: ET (c=0) , H_A: Time Varying Term Premium (c≠0) Hypothesis