CAHIER 0199

SEMI-PARAMETRIC WEAK INSTRUMENT REGRESSIONS WITH AN APPLICATION TO THE RISK-RETURN TRADE-OFF

Benoit PERRON

Université de Montréal

Centre de recherche
et développement en économique

C.P. 6128, Succursale Centre-ville
Montréal (Québec) H3C 3J7

Téléphone : (514) 343-6557
Télécopieur : (514) 343-5831
Adresse électronique : crde@crde.umontreal.ca
Site Web : http://www.crde.umontreal.ca/

CAHIER 0199

SEMI-PARAMETRIC WEAK INSTRUMENT REGRESSIONS WITH AN APPLICATION TO THE RISK-RETURN TRADE-OFF

Benoit PERRON[1]

[1] Centre de recherche et développement en économique (C.R.D.E.) and Département de sciences économiques, Université de Montréal

January 1999

This paper first circulated under the title "Risk Premia in Financial Data : Lessons from Weak Instrument Asymptotics". The author wishes to thank Peter Phillips, Oliver Linton, Donald Andrews, Hyungsik Moon and seminar participants at Yale University, University of Toronto, Université de Montréal, Concordia University, Rutgers University, and Florida International University for helpful comments and suggestions. The usual disclaimer necessarily applies. Financial assistance from the Alfred P. Sloan Foundation and the Social Sciences and Humanities Research Council (SSHRC) of Canada is gratefully acknowledged.

RÉSUMÉ

Des recherches récentes démontrent qu'une corrélation faible entre les instruments et les variables explicatives peut mener à de sérieux problèmes d'inférence dans les régressions avec variables instrumentales. Nous étendons l'analyse locale à zéro des modèles avec instruments faibles aux modèles avec des instruments et régresseurs estimés et avec de la dépendance dans les moments supérieurs. Ainsi, cet environnement devient applicable aux modèles linéaires avec des variables anticipatoires qui sont estimées de façon non paramétrique. Deux exemples de tels modèles sont la relation entre le risque et les rendements en finance et l'impact de l'incertitude de l'inflation sur l'activité économique réelle. Nos résultats démontrent que l'inférence basée sur les tests du multiple de Lagrange (LM) est plus robuste à la présence d'instruments faibles que l'inférence basée sur les tests de Wald. En utilisant des intervalles de confiance construits selon les tests de LM, nous concluons qu'il n'y a pas de prime de risque significative dans les rendements de l'indice S&P 500, les rendements excédentaires entre les Bons du Trésor de 6 mois et de 3 mois et les rendements du taux de change spot entre le yen japonais et le dollar américain.

Mots clés : variables instrumentales, instruments faibles, analyse locale à zéro, tests du multiple de Lagrange, tests de Wald, prime de risque, anticipations, modèles semi-paramétriques, noyaux, réseaux de neurones

ABSTRACT

Recent work shows that a low correlation between the instruments and the included variables leads to serious inference problems. We extend the local-to-zero analysis of models with weak instruments to models with estimated instruments and regressors and with higher-order dependence between instruments and disturbances. This makes this framework applicable to linear models with expectation variables that are estimated non-parametrically. Two examples of such models are the risk-return trade-off in finance and the impact of inflation uncertainty on real economic activity. Results show that inference based on Lagrange Multiplier (LM) tests is more robust to weak instruments than Wald-based inference. Using LM confidence intervals leads us to conclude that no statistically significant risk premium is present in returns on the S&P 500 index, excess holding yields between 6-month and 3-month Treasury bills, or in yen-dollar spot returns.

Key words : instrumental variables, weak instruments, local-to-zero analysis, LM tests, Wald tests, risk premium, expectations, semi-parametric models, kernels, neural networks

1. Introduction

Recently, the problem of weak correlation between instruments and regressors in instrumental variable regression has become a focal point of much research. Staiger and Stock (1997) developed an asymptotic theory for this type of problem using a local-to-zero framework. They show that standard asymptotics for IV estimators can be highly misleading when this correlation is low. Following the methodology of Staiger and Stock, Zivot, Startz, and Nelson (forthcoming) and Wang and Zivot (1998) show that usual testing procedures are unreliable in such situations. Earlier analyses of models under partial identification conditions was given in Phillips (1989) and Choi and Phillips (1992), and Dufour(1997).

This paper extends the weak instrument literature using the Staiger and Stock framework in two ways: first, we will analyze a restricted class of semi-parametric models in which both regressors and instruments are estimated, and second we will allow for higher-order dependence between the instruments and the disturbances. These extensions are meant to make the analysis applicable to the many theoretical models in finance and macroeconomics that suggest a linear relationship between a random variable and an expectation term of the general form,

$$y_t = \beta' x_t + \delta' Z_t + e_t \tag{1.1}$$

where y_t is a scalar, x_t is a vector of exogenous and predetermined variables, and Z_t is a vector of unobservable expectation variables. Of particular interest is the case where Z_t is a conditional variance term, and in this framework, interest centers on the parameter δ as it measures the response of y_t to increased risk.

One example of this type of problem includes the risk-return trade-off in finance where agents have to be compensated with higher expected returns for holding riskier assets. A model like (1.1) will hold as an approximation in this case if y_t is the return on an aggregate portfolio. This trade-off has been examined by several authors, including French, Schwert, and Stambaugh (1987) and Glosten, Jagannathan, and Runkle (1993). In this case, Z_t is the conditional variance of the asset, and x_t would generally include variables measuring the fundamental value of the asset. For example, if the asset is an exchange rate, potential fundamental variables include the interest rate differentials,

relative money stocks, relative outputs, and relative inflation rates. A second example of this model is in analyzing the effect of inflation uncertainty on real economic activity. Here, Z_t is the variance of the inflation rate conditional on past information, and y_t is some real aggregate variable such as real GDP or industrial production.

The estimation of these models has proven difficult because a proxy has to be constructed for the unobservable expectation term. A complete parametric approach would assume functional forms for the expectation processes of agents which can then be estimated along with (1.1) by, for example, maximum likelihood. A semi-parametric approach, which is of interest in this paper, leaves the functional form of the expectation terms unspecified but uses the linear structure in (1.1) to estimate the parameters of interest once estimates of the expectation terms are obtained.

When Z_t is a variance term, Engle, Lilien, and Robins (1987) have introduced the parametric AutoRegressive Conditional Heteroskedasticity-in-Mean (ARCH-M) model which postulates that $Z_t = \sigma_t^2$, the variance of the returns, follows an ARCH(p) model. A popular generalization is the Generalized ARCH-M (GARCH-M) model with σ_t^2 of the form:

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \ldots + \alpha_p e_{t-p}^2 + \gamma_1 \sigma_{t-1}^2 + \ldots + \gamma_q \sigma_{t-q}^2 \tag{1.2}$$

with (1.1) and (1.2) estimated jointly by maximum likelihood. Two problems surface when using such models. First, maximization of the likelihood function can be very difficult unless p and q are kept small. Second, estimates in the mean equation will be inconsistent if the variance equation is misspecified because the information matrix is not block diagonal. Given the lack of restrictions on the behavior of the conditional variance provided by economic theory, this seems quite problematic.

An alternative approach that is robust to specification was suggested by Pagan and Ullah (1988) and Pagan and Hong (1991). Their suggestion is to first replace Z_t by its realized values, say Y_t, estimating this quantity non-parametrically, and using a non-parametric estimate of Z_t as an instrument. This approach is itself problematic since it does not solve the necessity to keep the number of conditioning variables low. Moreover, a common problem when using such a semi-parametric approach is that the estimated conditional variance is poorly correlated with $\widehat{Y}_t$, the estimated realized values. This paper will focus on addressing this second problem. The first problem is addressed by using non-parametric estimators that are less susceptible to the so-called curse of dimensionality,

such as neural networks and a semi-parametric estimator suggested by Engle and Ng (1993).

The rest of the paper is divided as follows: section 2 presents the instrumental variable procedure described above in detail under the standard assumptions. In section 3, we present evidence on the presence of weak instruments in the risk-return trade-off. Next, in section 4, we develop asymptotic theory for the instrumental variable estimator described above under the weak instrument assumption. In section 5, results from a limited simulation experiment are presented to outline the difficulties involved in carrying out analysis in this type of models. Section 6 contains the results from applying the techniques developed in previous sections to three financial data sets, returns on the Standard and Poor's 500 index, excess holding yields on Treasury bills and yen-dollar spot returns. Finally, section 7 provides some concluding comments.

2. Semi-parametric models with conditional expectations

As discussed above, we consider linear models such as,

$$y_t = \beta' x_t + \delta' Z_t + e_t \tag{2.1}$$

where y_t is a scalar, x_t is a $k_1 \times 1$ vector of exogenous and predetermined variables, and Z_t is a $k_2 \times 1$ vector of unobservable expectation variables. One example of particular interest is where Z_t is a variance term of the form $E[Y_t|\mathcal{F}_t]$, with $Y_t = (\psi_t - E[\psi_t|\mathcal{F}_t])(\psi_t - E[\psi_t|\mathcal{F}_t])'$ and where $\mathcal{F}_t$ is the information set available to agents in the economy at the beginning of period t. In this framework, interest centers on the parameter δ as it measures the response of y_t to increased risk. Such models were first investigated along the lines followed here by Pagan and Ullah (1988).

The first step in tackling this problem is to replace the conditional expectation Z_t by the realized value Y_t. In the following, we assume that Y_t is not observable as is the case in the variance example since Y_t is itself a function of an expectation. Thus, an extra step is required in replacing Y_t by an estimate, $\widehat{Y}_t$. The model to be estimated is then:

$$\begin{aligned} y_t &= \beta' x_t + \delta' \widehat{Y}_t + e_t + \delta' \left(Y_t - \widehat{Y}_t \right) + \delta' (Z_t - Y_t) \\ &= \beta' x_t + \delta' \widehat{Y}_t + u_t \end{aligned}$$

In general, an ordinary least squares regression of y_t on x_t and $\widehat{Y}_t$ will lead to inconsistent estimates of β and δ due to the correlation between $\widehat{Y}_t$ and $(Z_t - Y_t)$. The solution suggested by Pagan (1984) and by Pagan and Ullah (1988) is to use an instrumental variable estimator with $\widehat{Z}_t$ used as instruments for $\widehat{Y}_t$. In fact, to obtain consistent estimates, any variable in $\mathcal{F}_t$ could be used as instrument. We could consider finding an optimal instrument as $E\left[\widehat{Y}_t|\mathcal{F}_t\right]$ which in general will be different from $\widehat{Z}_t$ because of the bias arising from the estimation of Y_t. The steps used to construct the estimator are illustrated in figure 1.

This problem will be semi-parametric when Y_t and Z_t are estimated non-parametrically. As in many semi-parametric models, despite the lower rate of convergence of the non-parametric estimators, the estimates of β and δ will converge at the usual $\sqrt{n}$ rate under certain conditions.

Define $\overline{Z}_t = (x_t, Z_t)$, $\overline{Y}_t = (x_t, Y_t)$, $\overline{Z} = \left(\overline{Z}_1, \ldots, \overline{Z}_n\right)'$, $\overline{Y} = \left(\overline{Y}_1, \ldots, \overline{Y}_n\right)'$ with $\widetilde{Z}$ and $\widetilde{Y}$ similarly defined but with $\widehat{Z}_t$ and $\widehat{Y}_t$ replacing Z_t and Y_t. Further let $\overline{u}_t = e_t + \delta'(Z_t - Y_t)$ and $\theta = (\beta, \delta)'$. Consider the IV estimator for this model:

$$\widehat{\theta} = \left(\widetilde{Z}'\widetilde{Y}\right)^{-1} \widetilde{Z}'y$$

A direct adaptation of the proof of Andrews (1994) gives the distribution of this estimator in the usual assumptions for the instruments.

Lemma 2.1. (*Andrews* (1994)) *Suppose the following:*

1. $\frac{1}{\sqrt{n}} \sum E_t\left(\widehat{Y}_t - Y_t\right) \xrightarrow{p} 0$
2. $\widehat{Z}_t \xrightarrow{p} Z_t, \quad Z_t < \infty \ \forall t$
3. θ_0 is the interior of $\Theta \subset R^{k+1}$
4. $\frac{1}{\sqrt{n}} \sum \overline{Z}_t \overline{u}_t \xrightarrow{d} N(0, S)$ where $S = \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n} \sum_{t=1}^{n} E\overline{u}_t \overline{u}_s \overline{Z}_s \overline{Z}_t^{'}$.
5. $M = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E_t \overline{Z}_t \overline{Y}_t^{'}$ is non-singular

Then, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N\left(0, M^{-1}S\left(M^{-1}\right)'\right)$.

The condition of most interest here is that we need $\widehat{Y}$ to be $n^{\frac{1}{2}}$-consistent for the asymptotic distribution of the IV estimator of $\widehat{\theta}$ not to be affected by replacing Y_t and Z_t by $\widehat{Y}_t$ and $\widehat{Z}_t$. This will generally not be the case when $\widehat{Y}$ is estimated non-parametrically. However, it will hold in the special case where Z_t is a variance term. Let $\tau_{1t} = E[\psi_t|\mathcal{F}_t]$ and $\widehat{\tau}_{1t}$ be an estimate of this quantity. Then,

$$\begin{aligned} E_t\left[\frac{1}{\sqrt{n}}\sum\left(\widehat{Y}-Y\right)\right] &= \frac{1}{\sqrt{n}}\sum E_t\left(\left[\psi_t-\widehat{\tau}_{1t}\right]^2-\left[\psi_t-\tau_{1t}\right]^2\right) \\ &= \frac{1}{\sqrt{n}}\sum\left(\widehat{\tau}_{1t}-\tau_{1t}\right)^2 \end{aligned}$$

which will be $o_p(1)$ if $\widehat{\tau}_{1t}$ is consistent for τ_{1t} at rate $n^{\frac{1}{4}}$. Conditions under which this holds can be found in Andrews (1995). The distribution is still affected by replacing Z_t by Y_t, however, as $\overline{u}_t = e_t + \delta'(Z_t - Y_t)$.

This estimator has been applied in Pagan and Ullah (1988), Pagan and Hong (1991), Bottazzi and Corradi (1991), and Sentana and Wadhwani (1991). Except for Pagan and Ullah, all these papers analyze the trade-off between financial returns and risk as postulated by mean-variance analysis. Pagan and Ullah look at the forward premium in the foreign exchange market and the real effects of inflation uncertainty.

3. Evidence of weak instruments

When using the above instrumental variable estimator, the quality of the instrument $\widehat{Z}_t$ will determine the quality of the asymptotic approximation described in lemma 2.1. There is a large amount of work in the simultaneous equation literature devoted to the importance of strong instruments for the finite-sample distribution to be well approximated by a normal distribution (one example is Nelson and Startz (1990)). Essentially, the non-singularity condition (Assumption 5) in the previous lemma is close to being violated.

Unfortunately, in our case of interest in which $Y_t = e_t^2$ and $Z_t = \sigma_t^2$, it will generally be the case that the correlation between the two estimates, $\widehat{e}_t^2$ and $\widehat{\sigma}_t^2$, is very low. Tables 1, 2, and 3 show the value of R^2 for the regression of $\widehat{e}_t^2$ on a constant and $\widehat{\sigma}_t^2$ for three financial data sets using three

non-parametric estimators with different conditioning variables and smoothing parameters. The estimates are constructed in three ways. First, we use a multivariate kernel estimate defined as:

$$\widehat{\tau}_{jt} = \frac{\sum_{i \neq t} y_i^j K\left(\frac{w_i - w_t}{b}\right)}{\sum_{i \neq t} K\left(\frac{w_i - w_t}{b}\right)}$$

as the estimate of the mean of y_t^j for $j = 1, 2$ with the kernel function $K(w)$ taken to be the multivariate standard normal and the bandwidth $b = c\widehat{s}n^{\frac{-1}{p+4}}$ where $\widehat{s}$ is the sample standard deviation of y_t, n is the sample size, p is the number of variables in the conditioning set and c is a constant taking three different values, 0.5, 1, or 2. The conditioning variables, w, are taken to be lagged values of the returns. We then define $\widehat{e}_t^2 = (y_t - \widehat{\tau}_{1t})^2$ and obtain an estimate of σ_t^2 as:

$$\widehat{\sigma}_t^2 = \widehat{\tau}_{2t} - (\widehat{\tau}_{1t})^2 .$$

A theoretical analysis of this non-parametric estimator of the conditional variance can be found in Masry and Tjostheim (1995).

The second estimation method used is neural networks. A good introduction to these methods is Kuan and White (1994). The advantage of this approach over the kernel is that it is not subject to the curse of dimensionality. The version we will adopt has one hidden layer with logistic and identity activation functions. the number of nodes will be allowed to equal 2, 4 and 8. The representation is:

$$\widehat{\tau}_j = \sum_{k=1}^{K} \widehat{\theta}_k \frac{1}{1 + e^{-w'\widehat{\beta}_j}}$$

for $j = 1, 2$.

The third estimator was first proposed by Engle and Ng (1993). It provides more structure to the conditional variance and will approximate the conditional variance function much better than the kernel when the variance is persistent (see Perron [**?**] for simulation evidence). The estimator is implemented by first estimating the mean by a kernel estimate as above and then fitting a function for σ_t^2 as follows:

$$\sigma_t^2 = \omega + f_1(\widehat{e}_{t-1}) + \ldots + f_p(\widehat{e}_{t-p}) + \beta\sigma_{t-1}^2$$

where the $f_j(\cdot)$ are estimated as splines with knots using a Gaussian likelihood function. This allows for a flexible effect of recent information on the conditional variance while allowing for persistence. This framework includes most parametric models suggested in the literature such as the

entire GARCH class. The number of segments in the spline functions acts as a smoothing parameter and is allowed to take three values, 2, 4, and 8. The knots in the spline were selected using the order statistics such that each bin has roughly the same number of observation subject to the constraint of an equal number of bins in the positive and negative regions.

The first data set analyzed represents monthly excess returns on the Standard and Poor's 500 between January 1965 and December 1997 measured at the end of each month. The data is taken from CRSP, and the risk-free rate is the return on three-month Treasury bills. The second data set is made of quarterly excess holding yields on 6-month versus 3-month Treasury bills between 1959:1 and 1998:1. A similar, but shorter, data set has already been analyzed by Engle, Lilien, and Robins (1987) using their GARCH-M methodology and Pagan and Hong (1991) using the above instrumental variable estimator. Finally, the last data series is made of monthly returns on the yen-dollar spot rate obtained from *International Financial Statistics* between September 1978 and June 1998. The three data sets are plotted in figures 2-4.

A quick look at the tables reveals that of these three series, only the excess holding yield generally has R^2 higher than 0.1. The reason for this low correlation is that e_t^2 and σ_t^2 have very different volatility. Even if $E\left[e_t^2|\mathcal{F}_t\right] = \sigma_t^2$, financial returns are extremely volatile and therefore, the difference between e_t^2 and σ_t^2 can be quite large. This is true even if we did not have to estimate these two quantities; having to estimate them complicates matters further. We can illustrate by looking at the GARCH(1,1) model:

$$\begin{aligned} y_t &= \mu + \sigma_t \varepsilon_t = \mu + e_t \\ \sigma_t^2 &= \omega + \alpha e_{t-1}^2 + \beta \sigma_{t-1}^2. \end{aligned}$$

Andersen and Bollerslev (1997) show that the population R^2 in the regression

$$(y_t - \mu)^2 = \gamma_0 + \gamma_1 \widehat{\sigma}_t^2 + v_t$$

where $\widehat{\sigma}_t^2$ is the one-period ahead forecast obtained from the GARCH model is

$$R^2 = \frac{\alpha^2}{1 - \beta^2 - 2\alpha\beta}$$

which will in general be very small even though $E\left[(y_t-\mu)^2\,|\mathcal{F}_t\right]=\sigma_t^2$. Figure 5 plots the value of R^2 for different values of α and β. The value of R^2 is highly sensitive to the value of α. It is usual in the literature to find point estimates of GARCH(1,1) models in the neighborhood of $\alpha=0.05$ and $\beta=0.9$. The figure clearly shows that for such values, the correlation between e_t^2 and σ_t^2 will typically be quite low. The problem in this case is that σ_t^2 has very low variance relative to that of y_t^2; a low value of α means that σ_t^2 is nearly constant locally.

We can expect that tables 1 and 2 do not even provide an accurate picture of the problem of weak instruments. Using data sampled at higher frequency (e.g. daily) would result in even lower correlation. The lower frequency allows some averaging which reduces the variance of e_t^2.

4. Asymptotics with weak instruments

Staiger and Stock (1997) have recently shown, in the framework of a linear simultaneous equation system, that having instruments that are weakly correlated with the explanatory variables makes the usual asymptotic theory work poorly. Their assumed model is:

$$\begin{aligned} y &= Y\delta + X\beta + u \\ Y &= Z\Pi + X\Gamma + V \end{aligned}$$

where Y is the matrix of included endogenous variables that are to be replaced by instruments. Since in our case, it will always be true that the model is exactly identified (that is, there will be as many regressors as instruments since the instruments are estimates of the expected value of the regressors), we will concentrate on the case where Z is a $n\times k_2$ matrix.. The weak instrument assumption is imposed by assuming that:

$$\Pi=\frac{G}{\sqrt{n}} \tag{4.1}$$

for some fixed $k_2\times k_2$ matrix $G\neq 0$ This assumption implies that in the limit, Y and Z are uncorrelated.

It is possible to extend the analysis of weak instruments in Staiger and Stock (1997) to our case of interest. Because the correlation between estimated regressors $\widehat{Z}_t$ and estimated instruments $\widehat{Y}_t$ is

very low in the data, it might be plausible to assume:

$$Y = Z\Pi + X\Gamma + V \tag{4.2}$$

with

$$\Pi = \frac{G}{\sqrt{n}} \tag{4.3}$$

for some $G \neq 0$. Simple algebra leads to:

$$\begin{aligned} \widehat{Y} &= \widehat{Z}\Pi + \left(Z - \widehat{Z}\right)\Pi + \left(\widehat{Y} - Y\right) + X\Gamma + V \\ &= \widehat{Z}\Pi + X\Gamma + \zeta \end{aligned}$$

so that the correlation between $\widehat{Y}_t$ and $\widehat{Z}_t$ is also low. This differs from the Staiger and Stock framework in that both variables will be estimated, and we will also allow for higher-order dependence between the instruments and the disturbances.

There might be two reasons for a low correlation between the estimated instrument and explanatory variable. The first may be that the estimators used in constructing $\widehat{Z}_t$ and $\widehat{Y}_t$ are poor and will not approach their true value in small samples. On the other hand, the estimators may not be poor in any sense, but Y_t and its expected value may be weakly correlated. We saw one such example above in the GARCH(1,1) model.

We can give a different motivation for equations (4.2) and (4.3) in the case where the instrument is a variance term ($Z_t = \sigma_t^2$, $Y_t = e_t^2$) if we assume that the estimates of σ_t^2 and e_t^2 are obtained by the kernel method. In this case, a simple application of the results in Masry and Tjostheim (1995) leads to the joint distribution:

$$\sqrt{nb^p}\begin{pmatrix} \widehat{e}_t^2 - e_t^2 - B_1 \\ \widehat{\sigma}_t^2 - \sigma_t^2 - B_2 \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{pmatrix}\right)$$

where b is the bandwidth, p the number of conditioning variables, and B_1 and B_2 are bias terms. We can then assume that the covariance term is local to zero, say $J/\sqrt{n}$, and use the formula for the conditional distribution of Y_t to obtain:

$$\widehat{e}_t^2 - e_t^2 - B_{1t} = \Omega_2^{-1}\frac{J}{\sqrt{n}}\left(\widehat{\sigma}_t^2 - \sigma_t^2 - B_{2t}\right) + \nu_t$$

and since $e_t^2 = \sigma_t^2 + \upsilon_t$ with υ_t uncorrelated with the past, we can rewrite this as:

$$\begin{aligned}\widehat{e}_t^2 &= \Omega_2^{-1}\frac{J}{\sqrt{n}}\widehat{\sigma}_t^2 - \Omega_2^{-1}\frac{J}{\sqrt{n}}\sigma_t^2 - B_{2t} + \sigma_t^2 + \upsilon_t + B_{1t} + \nu_t \\ &= \Omega_2^{-1}\frac{J}{\sqrt{n}}\widehat{\sigma}_t^2 + \left(I - \Omega_2^{-1}\frac{J}{\sqrt{n}}\right)\sigma_t^2 + B_t + \zeta_t\end{aligned}$$

with $B_t = B_{1t} + B_{2t}$ so that the coefficient on $\widehat{\sigma}_t^2$ is local to zero.

Recall that the IV estimator of δ is:

$$\begin{aligned}\widehat{\delta} &= \left(\widehat{Z}' M_x \widehat{Y}\right)^{-1} \widehat{Z} M_X y \\ &= \delta + \left(\widehat{Z}' M_X \widehat{Y}\right)^{-1} \widehat{Z}' M_X u\end{aligned}$$

where $M_X = I - X\left(X'X\right)^{-1}X'$. In order to derive the distribution of $\widehat{\beta}$, we need to make an extra assumption on the reduced-form coefficients of X. We will also assume that they are local to zero:

$$\Gamma = \frac{H}{\sqrt{n}} \tag{4.4}$$

for some $k_1 \times k_2$ matrix $H \neq 0$. This assumption is made because if Γ were fixed, X and Y would be collinear in the limit and the moment matrices would be singular.

The distribution of the estimators is given in the following theorem. All proofs are relegated to the appendix.

Theorem 4.1. *Assume the same conditions as in lemma* 2.1, *but with the fifth condition replaced by* (4.2) *and* (4.3) . *Further, suppose the following hold:*

$$(n^{-1}X'X, n^{-1}X'Z, n^{-1}Z'M_XZ) \xrightarrow{p} (\textstyle\sum_{XX}, \sum_{XZ}, \sum_{ZZ})$$

$$\left(n^{-\frac{1}{2}}X'u, n^{-\frac{1}{2}}Z'M_Xu, n^{-\frac{1}{2}}X'V, n^{-\frac{1}{2}}Z'M_XV\right) \Rightarrow (\Psi_{Xu}, \Psi_{Zu}, \Psi_{Xv}, \Psi_{Zv}).$$

Define

$$\sigma_{Zu} = \lim_{n\to\infty} \tfrac{1}{n}\textstyle\sum_s \sum_t u_t u_s Z_s^{\perp} Z_t^{\perp\prime}$$

$$\sigma_{Zv} = \lim_{n\to\infty} \tfrac{1}{n}\textstyle\sum_s \sum_t Z_s^{\perp} V_s' V_t Z_t^{\perp\prime}$$

$\sigma_{Xu} = \lim_{n\to\infty} \frac{1}{n} \sum_s \sum_t u_t u_s X_s X_t^{'}$

$\sigma_{Xv} = \lim_{n\to\infty} \frac{1}{n} \sum_s \sum_t X_s V_s^{'} V_t X_t^{'}$

$\rho_Z = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} V_t^{'} Z_t^{\perp\prime} \sigma_{ZV}^{-\frac{1}{2}\prime} \sigma_{Zu}^{-\frac{1}{2}} Z_s^{\perp} u_s$

$\rho_x = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} V_t^{'} X_t \sigma_{Zv}^{-\frac{1}{2}\prime} \sigma_{Zu}^{-\frac{1}{2}} X_s u_s$

where $Z_t^{\perp}$ is the projection of Z_t onto X, *i.e.* it is the transpose of the t^{th} row of $Z^{\perp} = M_X Z$.

Then,

1. $\widehat{\delta} - \delta \xrightarrow{d} \Xi = \sigma_{Zv}^{-\frac{1}{2}} (\lambda + z_v)^{-1} \sigma_{Zu} z_u$ with $\lambda = \sigma_{Zv}^{-\frac{1}{2}} \sum_{ZZ} G$, where $z_u = z_v \rho_Z + (1 - \rho_Z \rho_Z')^{\frac{1}{2}} \xi$, and $(vec(z_v), \xi) \sim N\left(0, I_{k_2(k_2+1)}\right)$,
2. In addition, with (4.4), $\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \sum_{XX}^{-1} \left[\sigma_{Xu} x_u + \left(\sum_{XZ} G + \sum_{XX} H + \sigma_{Xv} x_v\right) \Xi\right]$, where $x_u = x_V \rho_x + \left(I_{k_1} - \rho_x \rho_x^{'}\right)^{\frac{1}{2}} \zeta$, and $(vec(x_v), \zeta) \sim N\left(0, I_{k_1(k_2+1)}\right)$.

Several aspects of this result can be pointed out, all the outcome of the poor identification of δ. First, the IV estimator of δ does not converge to the true value of the population parameter δ. Rather, it will converge to a random variable in the limit, as in Phillips (1989). Second, the limit distribution is the ratio of correlated normal random variables. This suggests that the distribution will, in some cases, have Cauchy-like behavior with thick tails and possibly bimodality. Moreover, the distribution depends on nuisance parameters λ, and ρ, making inference difficult. As $\lambda \to \infty$, Ξ will approach the usual normal distribution. The distribution of the coefficients on the exogenous variables x_t is contaminated by the poor identification of δ. Specifically, the usual standard errors will understate the true uncertainty as these are based on the first term of the limit distribution only. This will lead to over-rejection of the hypothesis $H_0 : \beta = \beta_0$.

The basic distribution theory described above is very closely related to that derived by Staiger and Stock. Some adjustments have to be made, however, because we do not assume that the instruments, Z_t, are independent of the error terms u_t and v_t; we only assume that they are uncorrelated. The adjustments allow for higher order dependence between Z_t on the one hand and u_t and v_t on the

other. In cases where there is no higher dependence between the instruments and the error terms, this distribution coincides with that derived by Staiger and Stock.

The assumptions on the properties of the data are given in terms of high-level conditions, a joint weak law of large numbers and a weak convergence result. This is done to make the conditions similar to those used by Staiger and Stock. Many sets of primitive conditions can lead to these two results. For example, sufficient conditions are that the vector (u_t, V_t) be a martingale difference sequence with respect to the filtration $\left\{(u_{t-j-1}, V_{t-j-1}, Z_{t-j}, X_{t-j}), \dot{j} \geq 0\right\}$ with uniform finite $(2+\eta)$ moments for some $\eta > 0$ and the vector (Z_t, X_t) be α-mixing with mixing numbers of size $-\frac{\kappa}{\kappa-1}$ and $(r+\kappa)$finite moments for some $r \geq 2$. Unfortunately, these conditions imply that in the variance case, $Z_t = \sigma_t^2$, we need σ_t^8 to be finite for all t. This is a very difficult requirement for financial data as there is evidence that many financial series do not even have four finite moments. For this reason, we will use highly aggregated data (for example monthly and quarterly data) for applications to financial data.

4.1. Inference

Recent work by Wang and Zivot (1998) and Zivot, Startz and Nelson (forthcoming) has shown how unreliable inference can be in the Staiger and Stock framework. In particular, they show that confidence intervals based on Wald statistics tend to be too narrow, thus leading to overrejection. Rather, these authors recommend the use of confidence intervals obtained from inverting LM statistics and the Anderson-Rubin statistic in the case where the model is overidentified.

Use of the asymptotic theory developed in the previous section is hampered by the presence of the nuisance parameters, λ and ρ, which cannot be consistently estimated. As Wang and Zivot (1998) have noticed, in the case of just-identified models as is the case here, if we use the restricted estimate of σ_{Zu}, test statistics will have a limiting χ^2 distribution. In over-identified models, these test statistics will be bounded from above by a $\chi^2(K)$ distribution where $K > k_2$ is the number of instruments. Thus LM statistics will be appropriate if our concern is to control the size of the test and construct asymptotically valid confidence intervals.

The LM confidence intervals can be obtained as the set of δ such that the LM test statistic does

not reject the null hypothesis. Zivot, Startz, and Nelson (forthcoming) have shown that inverting the LM statistic for δ involves solving a quadratic equation. The shape of the resulting confidence interval will vary: it could be a bounded set, the union of two unbounded intervals, or the entire real line. These are quite unusual in shape. The possibility that confidence intervals could be unbounded reflects the great uncertainty about the parameter of interest. Dufour (1997) has shown that a valid $(1-\alpha)$ confidence interval for a locally unidentified parameter will be unbounded with probability $(1-\alpha)$. Since Wald intervals are always bounded (being constructed by adding and subtracting two standard errors to the point estimate), they cannot provide valid inference in this type of model. Unfortunately, these Wald intervals are almost always used in practice.

In our case here, we need to adjust the LM statistic for the higher order dependence. This is done in the following proposition for our just-identified case:

Proposition 4.2. *Let* $g = \frac{1}{n}\widehat{Z}' M_X \left(y - \widehat{Y}\delta\right)$. *Then under the null hypothesis,* $H_0 : \delta = \delta_0$, $LM = ng'\sigma_{Zu}^{-1} g \xrightarrow{d} \chi^2(k_2)$.

Unfortunately, in this case, there is no easy way to write the inequality that defines the confidence intervals as a quadratic equation in δ. Confidence intervals must be computed numerically by defining a grid of δ and verifying for each point on the grid whether the LM statistic defined in the above proposition is less than the appropriate critical value from the $\chi^2(k_2)$ distribution. This method is easily implemented in the scalar case, but could hardly be carried out in high dimensions.

Another approach to obtaining confidence intervals, suggested by Staiger and Stock (1997), is to use the Anderson-Rubin statistic. It is usually defined as the F-statistic for the significance of δ^* in the regression

$$y - \widehat{Y}\delta_0 = X\beta^* + \widehat{Z}\delta^* + u^*$$

where $\beta^* = \beta + \Gamma(\delta - \delta_0)$, $\delta^* = \Pi(\delta - \delta_0)$, and $u^* = u + v(\delta - \delta_0)$. Since we have a case with heteroskedasticity, we need to use robust standard errors to compute the test statistic. It turns out that in the just-identified case, this statistic is identical to the above LM statistic. This fact is stated in the following proposition:

Proposition 4.3. *Let* $AR = n\widehat{\delta}^{*}\widehat{V}^{-1}\widehat{\delta}^{*}$ *where* $\widehat{V} = \sum_{ZZ}^{-1}\widehat{\sigma}_{Zu}\sum_{ZZ}^{-1}$. *Then, under the null hypothesis* $H_0 : \delta = \delta_0$, $AR = LM$.

The above propositions thus give us two equivalent ways to construct asymptotically valid confidence intervals. The two methods are exactly the same as long as the same estimate of σ_{Zu} is used to construct either LM or AR. The performance of these intervals in a small sample situation will be analyzed in the simulation experiment in the next section.

5. Simulation Results

In this section, the behavior of the procedures described above will be analyzed through a small simulation experiment. Important issues to be analyzed include the choice of smoothing parameters, the appropriateness of the various confidence intervals, and the distribution of the resulting estimators.

Consider the GARCH-M$(1,1)$ DGP:

$$\begin{aligned} y_t &= \gamma + \delta\sigma_t^2 + e_t = \gamma + \delta\sigma_t^2 + \sigma_t\varepsilon_t \\ \sigma_t^2 &= \omega + \alpha e_{t-1}^2 + \beta\sigma_{t-1}^2 \\ \varepsilon_t &\sim i.i.d.N\left(0,1\right) \end{aligned}$$

In terms of the above notation, we have $v_t = e_t^2 - \sigma_t^2$, $u_t = e_t - \delta v_t$, $Y_t = e_t^2$, and $Z_t = \sigma_t^2$.

The various parameters are set to the estimates obtained from an identical GARCH-M(1,1) model for the S&P 500 data which are $\gamma = -0.0094$, $(\omega, \alpha, \beta) = (1.44 \times 10^{-4}, 0.0659, 0.8546)$, and $\delta = 6.6764$. These point estimates are similar to those usually obtained in this context, for example by Glosten, Jagannathan, and Runkle (1993), and will lead to a rather persistent σ_t^2 and to a weak instrument. Throughout, samples of 450 are drawn, with the first 50 observations deleted to remove the effect of the initial condition (taken as the mean of the unconditional distribution). The length of the sample nearly matches that of the S&P data.

One disadvantage of the current setup is that the correlation between $\widehat{\sigma}_t^2$ and $\widehat{e}_t^2$ cannot be controlled. We can control the correlation between the unobservable variables, but due to estimation, the correlation between observable variables will be different in general.

The values of the nuisance parameters in this setup can be obtained in terms of the moments of the conditional variance process as:

$$\begin{aligned}
\sigma_{Zv} &= (\kappa_4 - 1)\left[E\left(\sigma_t^8\right) - 2E\left(\sigma_t^2\right)E\left(\sigma_t^6\right) + E\left(\sigma_t^2\right)^2 E\left(\sigma_t^4\right)\right] \\
\sigma_{Zu} &= \delta^2\sigma_{Zv} + E\left(\sigma_t^6\right) - 2E\left(\sigma_t^2\right)E\left(\sigma_t^4\right) + E\left(\sigma_t^2\right)^3 - 2\delta\kappa_3\left[E\left(\sigma_t^7\right) - 2E\left(\sigma_t^5\right)E\left(\sigma_t^5\right) + E\left(\sigma_t^3\right)E\left(\sigma_t^2\right)^2\right] \\
\sigma_{uv} &= \kappa_3 E\left(\sigma_t^3\right) - \delta\sigma_{Zv} \\
\lambda &= \frac{\sqrt{n}\left[E\left(\sigma_t^4\right) - E\left(\sigma_t^2\right)^2\right]}{\sigma_{Zv}^{\frac{1}{2}}}
\end{aligned}$$

where $\kappa_j = E\left(\varepsilon_t^j\right)$ is the j^{th} moment of ε_t.

The values of the first 4 even moments of σ_t^2 are derived recursively in Bollerslev (1986) as a function of ω, α, and β and the moments of ε_t. This allows for the easy computation of the nuisance parameters. For the values given above, these parameters are $\lambda = 2.145$, $\rho = -0.472$, $\sigma_{Zv} = 3.404e-12$, and $\sigma_{Zu} = 6.819e-10$. The population R^2 between e_t^2 and σ_t^2 is 2.77%.

Figure 6 shows a plot of the asymptotic distribution using the above estimates of the nuisance parameters and that of the normal distribution obtained under the usual asymptotic theory, namely

$$\sqrt{n}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} N\left(0, \sum\nolimits_{ZY}^{-1}\sigma_{Zu}\left(\sum\nolimits_{ZY}^{-1}\right)'\right)$$

where $\sum_{ZY} = \lim_{n\to\infty}\frac{1}{n}Z'M_XY$. All quantities are normalized as t-ratios; this makes the usual normal theory above the standard normal. The figure is drawn with 100,000 draws taken from each distribution. The weak instrument approximation is slightly skewed, but its main feature is the much fatter tails than those of the standard normal distribution. The mass points at -10 and 10 represent the mass that lies outside of the $[-10, 10]$ interval.

Figure 7 shows the same picture for $n = 5000$. Since the weak instrument approximation approaches the standard normal as $n \to \infty$ in this case because $\lambda \to \infty$, we see that both the skewness and the excess kurtosis are much reduced for this sample size.

Figure 8 shows the distribution of the infeasible IV estimator using the actual values of σ_t^2 and e_t^2 generated; this estimator is infeasible since these values are unobservable in practice. This experiment was repeated 20,000 times. The asymptotic approximation captures most of the features

of the finite-sample distribution of the IV estimator. It matches the two tails well but overestimates the mass in the middle of the distribution. The usual normal approximation does not capture the tail behavior at all and does not do better in the area around the peak of the distribution.

The results of all simulation experiments are summarized in tables 4-10. The first column of each table shows the median of the IV estimator (rather than the mean because of the heavy tails of the distributions). The next two columns indicate the coverage rate of the appropriate 95% confidence intervals. The last column contains the mean R^2 of a regression of $\widehat{e}_t^2$ on a constant and $\widehat{\sigma}_t^2$.

The first line of these tables reports results of the infeasible estimator discussed above. The IV estimator appears slightly biased upward as expected given the skewness observed in figure 8. The Wald confidence interval has a coverage rate that is much lower than its nominal level, while the LM interval has coverage rate that is only slightly too low. The under-coverage of the Wald-based confidence intervals is expected given the theoretical results that these should have zero coverage asymptotically and the heavy tails of the distribution in figure 8. A researcher using these intervals would over-reject the null hypothesis $H_0 : \delta = \delta_0$ when it is true.

For the remaining experiments, estimates of e_t^2 and σ_t^2 are necessary. As before, these are obtained in three ways. The first one is a kernel-based estimator with a multivariate Gaussian kernel and with bandwidth selected according to the rule $b_k = c\widehat{s}_k n^{-\frac{1}{p+4}}$, where p is the number of conditioning variables (taken to be lagged values of y_t), $\widehat{s}_k$ is the sample standard deviation of the k^{th} conditioning variable, and c is a constant. Three values of c were used: 0.5, 1, and 2. These are the same choices as those used to obtain the values presented in table 1 above. The second estimator is based on artificial neural networks with one hidden layer and logistic and identity activation functions. The number of nodes are set at 2, 4, and 8 as was done in the construction of table 2. Finally, the last estimator is the Engle-Ng estimator used in the construction of table 3 with 2, 4, and 8 bins. Each experiment was repeated 1000 times.

The need to estimate σ_t^2 and e_t^2 changes the result quite dramatically relative to the infeasible estimator. The results using the kernel estimates are presented in tables 4-6 and figures 9-11. In all cases, the estimator of δ is strongly biased downward, but this bias goes down as the bandwidth increases. In general, a small bandwidth is preferable in semi-parametric estimation as it leads to less

biased but more variable non-parametric estimates that get averaged in the second step. However, additional smoothing is appropriate in this case because we need to keep the conditioning set small despite the high persistence. This finding is consistent with the results in Perron (1998). Surprisingly, the Wald intervals have in general better coverage than their LM counterparts. However, the coverage rate of LM intervals improves substantially as the bandwidth increases.

The figures explain this phenomenon. The kernel estimator is not a very good estimator of the conditional variance in this case as it does not capture persistence well. Hence, the finite sample distribution of $\widehat{\delta}$ is nowhere near the one obtained by using the infeasible IV estimator. However, this situation improves with a larger bandwidth, and this explains why the results approach those obtained using the infeasible estimator as the bandwidth increases. Nevertheless, all distributions are heavily skewed to the left and are not well summarized by any asymptotic approximation.

The results using the neural networks are presented in tables 7-10 and figures 12-14. The distribution of the estimator of δ is well-centered with $p = 1$. With more than one lagged value in the conditioning set, however, the estimator is biased downward. The two sets of confidence intervals have coverage rate that is too low, but the LM intervals perform much better. In fact, with $p = 1$, the coverage rate of the LM intervals is almost correct. the The weak instrument approximation does not provide a very good approximation to the finite-sample distribution of the estimator due to the bias, but of course neither does the usual normal theory.

Finally, the results for the Engle-Ng estimator are presented in table 7 and figure 15 for $p = 1$. The results provided by this method are excellent. The bias in the estimation of the risk parameter is small (but slightly negative). Once again, the LM-based confidence intervals perform better with a coverage rate that is close to their nominal level of 95%. The asymptotic approximation provided by the weak instrument theory is excellent. Moreover, there is only slight sensitivity to the choice of the smoothing parameter. The distributions with 2, 4 or 8 bins are essentially indistinguishable in the figure.

It would thus appear that only the Engle-Ng procedure provides a good approximation to the conditional variance as it leads to an IV estimator with a distribution that is close to that of the infeasible IV estimator. The other two (as well as other non-parametric estimators such as nearest

neighbors or local polynomials) face the disadvantage that they must be made conditional on a small information set. In theory, neural networks do not suffer from the curse of dimensionality and could be estimated conditional on a much larger number of lagged values. In practice, this is difficult as the optimization becomes more problematic, and the performance does not seem to improve remarkably.

Thus, the results above suggest extreme caution when using estimated instruments and explanatory variables in instrumental variable regression. It however appears that inference can be done robustly by using LM confidence intervals and using the semi-parametric estimator of the conditional variance. Once the conditional variance is estimated, the approximation provided by the weak instrument distribution is much superior to that provided by the usual normal approximation.

6. Empirical results

In this section, we will analyze our three financial data sets to seek evidence of a risk-return trade-off. To reiterate, the series are monthly returns on the S&P 500 index, quarterly excess holding yield between 6-month and 3-month Treasury bills and monthly returns on the yen-dollar spot rate.

For each series, we postulate a model of the form

$$y_t = \beta' x_t + \delta \sigma_t^2 + e_t$$

with x_t being a vector of explanatory variables specific to each series and $\sigma_t^2 = E\left[\left\{y_t - E\left[y_t|\mathcal{F}_{t-1}\right]\right\}^2 |\mathcal{F}_{t-1}\right]$ where $\mathcal{F}_{t-1}$ is the information set used by the agents in forming the corresponding expectation.

For both series, the conditional variance was estimated using each of the three methods discussed above: kernel, neural networks, and the Engle-Ng estimator described above. We only report the results using he kernel estimate with a bandwidth constant of 2 since this value reduced the bias in the estimation and provided confidence intervals with better coverage in the simulation experiment, neural networks with 4 nodes, and the Engle-Ng estimator with 4 bins. Results for the other choices are available upon request.

The LM confidence intervals were computed by numerically inverting the LM statistic. A grid of 1000 equi-spaced points between -100 and 100 was used for this purpose. For this reason, the infinite or very large confidence intervals got truncated at these two endpoints.

6.1. Stock returns

The first series represents monthly excess returns on the S&P 500 index between January 1965 and December 1997. The data is taken from CRSP, and the risk-free rate is the 3-month Treasury bill rate. The trade-off between risk and return has been extensively studied for similar series with conflicting results. For example, French, Schwert, and Stambaugh (1987) find a positive relation between returns and the conditional variance, while Glosten, Jagannathan, and Runkle (1993) find a negative relationship using a modified GARCH-M methodology. This conflicting evidence is not surprising in light of the results obtained by Backus, Gregory, and Zin (1989) and Backus and Gregory (1993). Using a general equilibrium setting, they provide simulation evidence that the relationship between expected returns and the variance of returns can go in either direction, depending on specification.

The estimation results are presented in table 11. In addition to the point estimates and the robust t-statistics, we present Wald-based and LM-based 95% confidence intervals for the coefficient on the risk variable, δ, as well as the partial R^2 between $\widehat{\sigma}_t^2$ and $\widehat{e}_t^2$. The results are unambiguous on the presence of a relationship between the excess returns and the conditional variance. In all cases but one, there is no significant effect of risk on returns. The only exception is the kernel estimator with 3 lags which shows a significant positive relationship using the Wald inference. However, the main feature of the results is the much wider confidence intervals obtained using the LM principle. Wald confidence intervals dramatically understate the uncertainty of the estimated parameters.

6.2. Excess holding yield

Following Engle, Lilien, and Robins (1987), the excess holding yield between 6-month Treasury bill and 3-month Treasury bill is defined as:

$$y_t = \frac{(1+R_t)^2}{(1+r_{t+1})} - (1+r_t)$$

where R_t and r_t are the yield on the 6-month and 3-month T-bill between t and $t+1$ respectively. Quarterly data between the first quarter of 1959 and the first quarter of 1998 is used. A similar (and shorter) series has been studied by Engle, Lilien, and Robins (1987) and by Pagan and Hong (1991). The first paper applied the ARCH-M methodology, while the second one used the above

semi-parametric instrumental variable estimator. A plot of the data is provided in figure 3.

The results of the estimation are presented in table 12 for lag lengths between 1 and 3. The variables included in the vector of exogenous and predetermined variables x_t include a constant, and the interest spread $R_t - r_t$. All point estimates are positive with the exception of the kernel with one lag. Three cases show a significant relationship using Wald inference. In these three cases. the LM intervals are very wide and reverse the conclusion. In fact the LM intervals are much wider than their Wald counterparts in all cases.

Also note that the interest spread is significant at the 5% level using standard testing procedures in all cases but one. This is to be expected given the second part of theorem 4.1 as the usual standard errors understate the level of uncertainty associated with the estimators of the coefficients of the exogenous variables.

6.3. Yen-dollar exchange rate

The other data series considered consists of monthly returns on the yen-dollar spot exchange rate between September 1978 and June 1998. This series is plotted in figure 4. The returns are assumed to depend on the differential between Japanese and U.S. interest rates as postulated by the uncovered parity condition, as well as their own lag values. The interest rate used is the 3-month LIBOR offer rate. The data was obtained from the IFS CD-Rom.

The results from the estimation are presented in table 13. Once again, few confidence intervals show a statistically significant risk premium term. The only exceptions are the neural network with 2 lags which shows a significantly negative relation and 3 cases where the LM intervals take the unusual disjoint shape. The Wald confidence intervals are already wide, but the LM intervals are even wider. The partial R^2 between estimated squared innovations and the estimated conditional variance is much lower than for the excess holding yield as was documented in tables 1-3.

Note that the coefficient on the interest rate differential seems quite precisely estimated between -3 and -4 for all specifications and is significantly *negative* using standard testing procedures. Many studies using the uncovered interest rate parity condition find such a significantly negative coefficient on the interest differential (see Froot and Thaler (1990) for a survey of the literature). The inclusion

of the variance term does not change the results much, neither does the inclusion of monthly dummies. This is also true for the GARCH-M specification. In this latter case however, the risk premium term is significantly positive. However, this significance has to be taken with care given the second result of theorem 4.1.

7. Conclusion

This paper follows several others in showing that inference using instrumental variables is greatly affected by a low correlation between the instruments and the explanatory variables. It extends the current literature to linear semi-parametric models with non-parametrically estimated regressors and instruments and to cases with higher-order dependence. The analysis shows that the limit theory in this case is similar to that currently available in the literature.

Simulation evidence reveals that the additional step of estimating both the regressor and the instrument may lead to a large loss in the quality of asymptotic approximations. Using a semi-parametric estimator proposed by Engle and Ng (1993) and carrying out inference using Lagrange Multiplier procedures allows for inference that is more robust than the alternatives considered here.

Empirical application to three financial series suggests that conclusions may hinge on the use of appropriate confidence intervals. Using the appropriate LM confidence intervals and the semi-parametric estimator of the conditional variance leads us to conclude that none of the series considered includes a statistically significant risk premium. This differs in many cases from inference based on the usual Wald confidence intervals and on a parametric GARCH-M model. However, because of the wide confidence bands, the results are also consistent with the presence of large risk premia. The data is simply not informative enough to precisely estimate the relationship between risk and returns.

Further work on this problem is clearly warranted. In particular, other estimators such as maximum likelihood are likely to face similar problems as the IV estimator analyzed here. Moreover, Bayesian methods might be helpful in this case as a prior distribution on the reduced form coefficients is intuitive. Finally, the development of data-based selection procedures for the smoothing parameters appears important given the sensitivity of the results to this choice.

REFERENCES

Andersen, Torben G. and Tim Bollerslev, "Answering the Critics: Yes, ARCH Models Do Provide Good Volatility Forecasts", NBER Working Paper 6023, 1997.

Andrews, Donald K., "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity", *Econometrica*, 62, 1994, 43-72.

Andrews, Donald K., "Examples of MINPIN Estimators: Supplement to Asymptotics for Semiparametric Models via Stochastic Equicontinuity", Mimeographed, 1992.

Andrews, Donald K., "Nonparametric Kernel Estimation for Semiparametric Models", *Econometric Theory*, 11, 1995, 560-596.

Backus, David K., Allan W. Gregory, and Stanley E. Zin, "Risk Premiums in the Term Structure: Evidence from Artificial Economies", *Journal of Monetary Economics*, 24, 1989, 371-399.

Backus, D. K., and A. W. Gregory (1993), "Theoretical Relations Between Risk Premiums and Conditional Variances," *Journal of Business and Economic Statistics*, 11, 177-185.

Bollerslev, Tim, "Generalized Conditional Heteroskedasticity", *Journal of Econometrics*, 31, 1986, 307-327.

Bottazzi, Laura and Valentina Corradi, "Analysing the Risk Premium in the Italian Stock Market: ARCH-M Models versus Non-parametric Models", *Applied Economics*, 23, 1991, 535-542.

Choi, In and Peter C. B. Phillips, "Asymptotic and Finite Sample Distribution Theory for the IV Estimators and Tests in Partially Identified Structural Relations", *Journal of Econometrics*, 51, 1992, 113-150.

Dufour, Jean-Marie, "Some Impossibility Theorems in Econometrics with Applications to Instrumental Variables, Dynamic Models, and Cointegration", *Econometrica*, 65, 1997, 1365-1387.

Engle, Robert F., and Victor K. Ng, "Measuring and Testing the Impact of News on Volatility", *Journal of Finance*, 48, 1993, 1749-1778.

Engle, Robert F., Lilien, David M., and Russell P. Robins, "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model", *Econometrica*, 55, 1987, 391-407.

French, Kenneth, G. William Schwert, and Robert F. Stambaugh, "Expected Stock Returns and Volatility", *Journal of Financial Economics*, 19, 1987, 3-29.

Froot, Kenneth A. and Richard H. Thaler, "Anomalies: Foreign Exchange", *Journal of Economic Perspectives*, 3, 1990, 179-192.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle, "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *Journal of Finance*, 48, 1993, 1779-1801.

Hall, Alastair R., Glenn D. Rudebusch, and David W. Wilcox, "Judging Instrument Relevance in Instrumental Variables Estimation", *International Economic Review*, 37, 1996, 283-298.

Kuan, Chang-Ming and Halbert White, "Artificial Networks: An Econometric Perspective", *Econometric Reviews*, 13, 1994, 1-91.

Masry, Elias and Dag Tjostheim, "Nonparametric Estimation and Identification of Nonlinear Time Series: Strong Convergence and Asymptotic Normality", *Econometric Theory*, 11, 1995, 258-289.

Nelson, Charles R., Startz, Richard, "The Distribution of the Instrumental Variables Estimator and its t-ratio when the Instrument is a Poor One", *Journal of Business*, 63, 1990, S125-S140.

Pagan, Adrian, "Econometric Issues in the Analysis of Regressions with Generated Regressors", *International Economic Review*, 25, 1984, 221-247.

Pagan, Adrian R. and Y. S. Hong, "Nonparametric Estimation and the Risk Premium" in Barnett, William A., James Powell, and George E. Tauchen eds., *Nonparametric and Semiparametric*

Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, Cambridge: Cambridge University Press, 1991, 51-75.

Pagan, Adrian and Aman Ullah. "The Econometric Analysis of Models with Risk Terms", *Journal of Applied Econometrics*, 3, 1988, 87-105.

Perron, Benoit, "A Monte Carlo Comparison of Non-parametric Estimators of the Conditional Variance", Mimeo, 1998.

Phillips, Peter C. B., "Partially Identified Econometric Models", *Econometric Theory*, 5, 1989, 181-240.

Sentana, Enrique and Sushil Wadhwani, "Semi-parametric Estimation and the Predictability of Stock Market Returns: Some Lessons from Japan", *Review of Economics Studies*, 58, 1991, 547-563.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments", *Econometrica*, 65, 1997, 557-586.

Stock, James and Jonathan Wright, "GMM with Weak Identification", Mimeographed, 1997.

Wang, Jiahui and Eric Zivot, "Inference on a Structural Parameter in Instrumental Regression with Weak Instruments", *Econometrica*, 66, 1998, 1389-1404.

Zivot, Eric, Richard Startz and Charles R. Nelson "Valid Confidence Intervals and Inference in the Presence of Weak Instruments", *International Economic Review*, forthcoming.

8. Appendix

1. Proofs

1.1. Preliminary results

Before proving the various theorems, we will collect the required preliminaries in the following lemma.

Lemma 1.1. *Suppose the conditions of theorem* (4.1) *hold. Then, the following hold:*

1. $\frac{1}{\sqrt{n}}\left(\widehat{Z}'M_X\widehat{Y}\right) = \frac{1}{\sqrt{n}}\left(Z'M_XY\right) + o_p\left(1\right)$
2. $\frac{1}{\sqrt{n}}\left[\widehat{Z}'M_X\left(Z-Y\right)\delta\right] = \frac{1}{\sqrt{n}}\left[Z'M_X\left(Z-Y\right)\delta\right] + o_p\left(1\right)$
3. $\frac{1}{\sqrt{n}}\left(\widehat{Z}'M_Xe\right) = \frac{1}{\sqrt{n}}\left(Z'M_Xe\right) + o_p\left(1\right)$
4. $\frac{1}{n}\left(\widehat{Z}'M_X\widehat{Z}\right) = \frac{1}{n}\left(Z'M_XZ\right) + o_p\left(1\right)$
5. $\frac{1}{\sqrt{n}}X'\widehat{Y} = \frac{1}{\sqrt{n}}X'Y + o_p\left(1\right)$
6. $\frac{1}{\sqrt{n}}\left[\widehat{Z}'M_X\left(Y-\widehat{Y}\right)\delta\right] \xrightarrow{p} 0$
7. $\frac{1}{\sqrt{n}}\left(\widehat{Z}'M_Xu\right) = \Psi_{Zu} + o_p\left(1\right).$

Proof. To prove the first result, note that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\widehat{Z}'M_X\widehat{Y} &= \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_X\left(\widehat{Y}-Y\right) + \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_XY \\
&\quad + \frac{1}{\sqrt{n}}Z'M_X\left(Y-\widehat{Y}\right) + \frac{1}{\sqrt{n}}Z'M_XY \\
&= \frac{1}{n}\left(\widehat{Z}-Z\right)'\left[\sqrt{n}M_X\left(Y-\widehat{Y}\right)\right] + \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_XY \\
&\quad + \frac{1}{n}Z'\left[\sqrt{n}M_X\left(\widehat{Y}-Y\right)\right] + \frac{1}{\sqrt{n}}Z'M_XY \\
&= \frac{1}{\sqrt{n}}Z'MY + E_t\left[\left(\widehat{Z}-Z\right)'\left[\sqrt{n}M_X\left(Y-\widehat{Y}\right)\right]\right] + \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_X\left(Y-Z\right) \\
&\quad + \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_X\left[Z-E\left(Z\right)\right] + \frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_XE\left(Z\right)
\end{aligned}$$

$$
\begin{aligned}
&+E_t\left[\left(Z-E\left(Z\right)\right)'\left[\sqrt{n}M_X\left(\widehat{Y}-Y\right)\right]\right]+E_t\left[E\left(Z\right)\left[\sqrt{n}M_X\left(\widehat{Y}-Y\right)\right]\right]+o_p\left(1\right)\\
= \;& \frac{1}{\sqrt{n}}Z'MY+A_1+A_2+A_3+A_4+A_5+A_6+o_p\left(1\right)
\end{aligned}
$$

We will next bound each of the A_i, $i=1,\ldots,6$. Let $|A|$ be the matrix norm of A. First,

$$
\begin{aligned}
|A_1| &= \left|E_t\left[\left(\widehat{Z}-Z\right)'\left[\sqrt{n}M_X\left(Y-\widehat{Y}\right)\right]\right]\right|\\
&\leq \left|\widehat{Z}-Z\right|\left|E_t\left[\sqrt{n}M_X\left(Y-\widehat{Y}\right)\right]\right|\\
&= o_p\left(1\right)
\end{aligned}
$$

by assumptions 1 and 2 where the second line follows from the fact that both $\widehat{Y}$ and Y are measurable with respect to $\mathcal{F}_t$ and the triangle inequality. Next,

$$
\begin{aligned}
|A_2| &= \left|\frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_X\left(Y-Z\right)\right|\\
&\leq \left|\widehat{Z}-Z\right|\left|\frac{1}{\sqrt{n}}M_X\left(Y-Z\right)\right|\\
&= o_p\left(1\right)
\end{aligned}
$$

by assumption 2 and since the quantity inside the second absolute value will be $O_p\left(1\right)$. The third term is:

$$
\begin{aligned}
|A_3| &= \left|\frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_X\left[Z-E\left(Z\right)\right]\right|\\
&\leq \left|\widehat{Z}-Z\right|\left|\frac{1}{\sqrt{n}}M_X\left[Z-E\left(Z\right)\right]\right|\\
&= o_p\left(1\right)
\end{aligned}
$$

again by assumption 2 and the term inside the second absolute value being $O_p\left(1\right)$. The fourth term can be bounded as:

$$
\begin{aligned}
|A_4| &= \left|\frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'M_XE\left(Z\right)\right|\\
&\leq \left|\frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)\right|\left|M_XE\left(Z\right)\right|\\
&= o_p\left(1\right)
\end{aligned}
$$

as $\widehat{Z} \xrightarrow{p} Z$ and $|E(Z)| < \infty$ with probability one since $Z_t < \infty$ for all t. The fifth term is:

$$\begin{aligned} |A_5| &= \left| E_t \left[(Z - E(Z))' \left[\sqrt{n} M_X \left(\widehat{Y} - Y \right) \right] \right] \right| \\ &\leq |Z - E(Z)| \left| E_t \left[\sqrt{n} M_X \left(\widehat{Y} - Y \right) \right] \right| \\ &= O_p(1) \cdot o_p(1) \\ &= o_p(1) \end{aligned}$$

by assumption 1. Finally, the sixth term can be bounded as:

$$\begin{aligned} |A_6| &= \left| E_t \left[E(Z) \left[\sqrt{n} M_X \left(\widehat{Y} - Y \right) \right] \right] \right| \\ &\leq |E(Z)| \left| E_t \left[\sqrt{n} M_X \left(\widehat{Y} - Y \right) \right] \right| \\ &= o_p(1) \end{aligned}$$

by assumption 1 and the fact that $|E(Z)| < \infty$ with probability one. Thus,

$$\frac{1}{\sqrt{n}} \widehat{Z}' M_X \widehat{Y} = \frac{1}{\sqrt{n}} Z' M_X Y + o_p(1)$$

as required.

The second result is obtained as:

$$\begin{aligned} \frac{1}{\sqrt{n}} \left[\widehat{Z}' M_X (Z - Y) \delta \right] &= \frac{1}{\sqrt{n}} \left[\left(\widehat{Z} - Z \right)' M_X (Z - Y) \delta \right] + \frac{1}{\sqrt{n}} \left[Z' M_X (Z - Y) \delta \right] \\ &= \frac{1}{\sqrt{n}} \left[Z' M_X (Z - Y) \delta \right] + o_p(1) \end{aligned}$$

where the last line follows from:

$$\begin{aligned} \left| \frac{1}{\sqrt{n}} \left(\widehat{Z} - Z \right)' M_X (Z - Y) \delta \right| &\leq \left| \widehat{Z} - Z \right| \left| \frac{1}{\sqrt{n}} M_X (Z - Y) \delta \right| \\ &= o_p(1) \cdot O_p(1) \\ &= o_p(1) \end{aligned}$$

The third result follows from:

$$\frac{1}{\sqrt{n}} \widehat{Z}' M_X e = \frac{1}{\sqrt{n}} \left[\left(\widehat{Z} - Z \right) \right]' M_X e + \frac{1}{\sqrt{n}} Z' M_X e$$

and noting that the first term can be bounded by:

$$\begin{aligned}\left|\frac{1}{\sqrt{n}}\left[\left(\widehat{Z}-Z\right)\right]' M_X e\right| &\leq \left|\widehat{Z}-Z\right|\left|\frac{1}{\sqrt{n}}M_X e\right| \\ &= o_p(1)\cdot O_p(1) \\ &= o_p(1)\end{aligned}$$

by assumptions 2 and 4.

The fourth result is proved by rewriting the left hand side as:

$$\begin{aligned}\frac{1}{n}\widehat{Z}' M_X \widehat{Z} &= \frac{1}{n}\left(\widehat{Z}-Z\right)' M_X \widehat{Z} + \frac{1}{n}Z' M_X \widehat{Z} \\ &= \frac{1}{n}Z' M_X Z + \frac{1}{n}\left(\widehat{Z}-Z\right)' M_X \left(\widehat{Z}-Z\right) + \frac{1}{n}\left(\widehat{Z}-Z\right)' M_X Z + \frac{1}{n}Z' M_X \left(\widehat{Z}-Z\right) \\ &= \frac{1}{n}Z' M_X Z + B_1 + B_2 + B_2^{'}\end{aligned}$$

where B_j, $j = 1, 2$, is each bounded in turn by an $o_p(1)$ term. For B_1, we obtain:

$$\begin{aligned}|B_1| &= \left|\frac{1}{n}\left(\widehat{Z}-Z\right)' M_X \left(\widehat{Z}-Z\right)\right| \\ &\leq \left|\frac{1}{\sqrt{n}}\left(\widehat{Z}-Z\right)'\iota\right| |M_X| \left|\frac{1}{\sqrt{n}}\iota'\left(\widehat{Z}-Z\right)\right| \\ &= o_p(1)\end{aligned}$$

by assumption 2. The second term is bounded as:

$$\begin{aligned}|B_2| &= \left|\frac{1}{n}\left(\widehat{Z}-Z\right)' M_X Z\right| \\ &\leq \left|\widehat{Z}-Z\right|\left|\frac{1}{n}M_X Z\right| \\ &= o_p(1)\cdot O_p(1) \\ &= o_p(1)\end{aligned}$$

by assumption 2. The fourth result follows.

The fifth result is obtained as:

$$\begin{aligned}\frac{1}{\sqrt{n}}X'\widehat{Y} &= \frac{1}{\sqrt{n}}X'Y + \frac{1}{\sqrt{n}}X'\left(\widehat{Y}-Y\right) \\ &= \frac{1}{\sqrt{n}}X'Y + o_p(1)\end{aligned}$$

by assumption 2.

The sixth result is obtained from the decomposition:

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}}\left[\widehat{Z}'M_X\left(Y-\widehat{Y}\right)\delta\right]\right| &= \left|\frac{1}{\sqrt{n}}\left[\left(\widehat{Z}-Z\right)'M_X\left(Y-\widehat{Y}\right)\delta\right]+\frac{1}{\sqrt{n}}\left[Z'M_X\left(Y-\widehat{Y}\right)\delta\right]\right| \\
&\leq \left|\frac{1}{\sqrt{n}}\left[\left(\widehat{Z}-Z\right)'\left(Y-\widehat{Y}\right)\delta\right]\right| \\
&\quad +\left|\frac{1}{\sqrt{n}}\left[Z-E\left(Z\right)\right]'M_X\left[\left(Y-\widehat{Y}\right)\right]\delta\right| \\
&\quad +\left|\frac{1}{n}E\left(Z\right)'M_X\left[\sqrt{n}\left(Y-\widehat{Y}\right)\right]\delta\right| \\
&\leq \left|\widehat{Z}-Z\right|\left|\frac{1}{\sqrt{n}}\iota'\left(Y-\widehat{Y}\right)\delta\right| \\
&\quad \left|Z-E\left(Z\right)\right|\left|\frac{1}{\sqrt{n}}\iota'\left(Y-\widehat{Y}\right)\delta\right| \\
&\quad +\left|E\left(Z\right)\right|\left|\frac{1}{\sqrt{n}}\iota'\left(Y-\widehat{Y}\right)\delta\right| \\
&= o_p\left(1\right)
\end{aligned}$$

where the last line follows from $\frac{1}{\sqrt{n}}\iota'\left(Y-\widehat{Y}\right)\xrightarrow{p}E_t\left[\sqrt{n}\iota'\left(Y-\widehat{Y}\right)\right]\xrightarrow{p}0$.

Finally, the last result is obtained by rewriting the left hand side as:

$$\frac{1}{\sqrt{n}}\widehat{Z}'M_Xu=\frac{1}{\sqrt{n}}\widehat{Z}'M_Xe+\frac{1}{\sqrt{n}}\widehat{Z}'M_X\left(Y-\widehat{Y}\right)\delta+\frac{1}{\sqrt{n}}\widehat{Z}'M_X\left(Z-Y\right)\delta$$

and using results 2, 3, and 5 of the lemma. ■

1.2. Proof of theorem 4.1

The instrumental variable estimator of δ is

$$\begin{aligned}
\widehat{\delta}-\delta &= \left(\widehat{Z}'M_X\widehat{Y}\right)^{-1}\widehat{Z}'M_Xu \\
&= \left(Z'M_XY\right)^{-1}Z'M_Xu+o_p\left(1\right)
\end{aligned}$$

by results 1 and 7 of the lemma. To derive the asymptotic distribution, we can handle the inverse term as:

$$\frac{1}{\sqrt{n}}\left(Z'M_XY\right) = \frac{1}{\sqrt{n}}\left[Z'M_X\left(Z\Pi+V\right)\right]$$

$$
\begin{aligned}
&= \frac{1}{n} Z' M_X Z G + \frac{1}{\sqrt{n}} Z' M_X V \\
&\xrightarrow{d} \sum\nolimits_{zz} G + \Psi_{zv} \\
&= \sigma_{Zv}^{\frac{1}{2}} \left(\sigma_{Zv}^{-\frac{1}{2}} \sum\nolimits_{zz} G + z_v \right) \\
&= \sigma_{Zv}^{\frac{1}{2}} \left(\lambda + z_v \right)
\end{aligned}
$$

while $\frac{1}{\sqrt{n}} \left(Z' M_X u \right) \xrightarrow{d} \Psi_{zu} = \sigma_{Zu}^{\frac{1}{2}} z_u$ by assumption. Putting these pieces together gives us the desired result for the distribution of $\widehat{\delta}$:

$$
\widehat{\delta} - \delta \xrightarrow{d} \Xi.
$$

To derive the distribution of $\widehat{\beta}$, note that:

$$
\begin{aligned}
\widehat{\beta} &= \left(X'X \right)^{-1} X' \left(y - \widehat{Y} \widehat{\delta} \right) \\
&= \beta + \left(X'X \right)^{-1} X' \widehat{Y} \left(\delta - \widehat{\delta} \right) + \left(X'X \right)^{-1} X' u
\end{aligned}
$$

so that

$$
\begin{aligned}
\sqrt{n} \left(\widehat{\beta} - \beta \right) &= \left(\frac{X'X}{n} \right)^{-1} \left(\frac{X'\widehat{Y}}{\sqrt{n}} \right) \left(\delta - \widehat{\delta} \right) + \left(\frac{X'X}{n} \right)^{-1} \frac{X'u}{\sqrt{n}} \\
&= \left(\frac{X'X}{n} \right)^{-1} \left(\frac{X'Y}{\sqrt{n}} \right) \left(\delta - \widehat{\delta} \right) + \left(\frac{X'X}{n} \right)^{-1} \frac{X'u}{\sqrt{n}} + o_p(1) \\
&\xrightarrow{d} - \sum\nolimits_{XX}^{-1} \left(\sum\nolimits_{XZ} G + \sum\nolimits_{XX} H + \Psi_{Xv} \right) \Xi + \sum\nolimits_{XX}^{-1} \psi_{Xu}
\end{aligned}
$$

where the term in parentheses is derived from:

$$
\begin{aligned}
\frac{1}{\sqrt{n}} X'Y &= \frac{1}{\sqrt{n}} X' \left(Z\Pi + X\Gamma + V \right) \\
&= \frac{1}{n} X' Z G + \frac{1}{n} X' X H + \frac{1}{\sqrt{n}} X' V \\
&\xrightarrow{d} \sum\nolimits_{XZ} G + \sum\nolimits_{XX} H + \Psi_{Xv}
\end{aligned}
$$

by assumption. ■

1.3. Proof of Proposition 4.2

By result 7 of the lemma, $\sqrt{n} g \xrightarrow{d} \Psi_{Zu} \overset{d}{=} N\left(0, \sigma_{Zu}\right)$ under the null hypothesis. Standard arguments show the desired result, $n g' \sigma_{Zu}^{-1} g \xrightarrow{d} \chi^2\left(k_2\right)$. ■

1.4. Proof of Proposition 4.3

The estimator of δ^* is defined as:

$$\begin{aligned}
\widehat{\delta^*} &= \left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\widehat{Z}'M_X\left(y-\widehat{Y}\delta_0\right) \\
&= \left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\widehat{Z}'M_X\left(X\beta^*+\widehat{Z}\delta^*+u+v\left(\delta-\delta_0\right)\right) \\
&= \delta^*+\left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\widehat{Z}'M_Xu+\left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\widehat{Z}'M_Xv\left(\delta-\delta_0\right)
\end{aligned}$$

so that

$$\begin{aligned}
\sqrt{n}\left(\widehat{\delta^*}-\delta^*\right) &= \left(\frac{\widehat{Z}'M_X\widehat{Z}}{n}\right)^{-1}\frac{\widehat{Z}'M_Xu}{\sqrt{n}}+\left(\frac{\widehat{Z}'M_X\widehat{Z}}{n}\right)^{-1}\frac{\widehat{Z}'M_Xv\left(\delta-\delta_0\right)}{\sqrt{n}} \\
&= \left(\frac{\widehat{Z}'M_X\widehat{Z}}{n}\right)^{-1}\frac{\widehat{Z}'M_Xu}{\sqrt{n}}
\end{aligned}$$

under the null hypothesis. By results 4 and 7 of the lemma, $\sqrt{n}\left(\widehat{\delta^*}-\delta^*\right)\rightarrow N\left(0,\sum_{ZZ}^{-1}\sigma_{Zu}\sum_{ZZ}^{-1}\right)$.
The robust AR statistic is:

$$\begin{aligned}
AR &= n\left(y-\widehat{Y}\delta_0\right)'M_X\widehat{Z}\left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\left[\left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\sigma_{Zu}\left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\right]^{-1} \\
&\quad \left(\widehat{Z}'M_X\widehat{Z}\right)^{-1}\widehat{Z}'M_X\left(y-\widehat{Y}\delta\right) \\
&= n\left(y-\widehat{Y}\delta_0\right)'M_X\widehat{Z}\sigma_{Zu}^{-1}\widehat{Z}'M_X\left(y-\widehat{Y}\delta\right) \\
&= LM
\end{aligned}$$

after simplification with σ_{Zu} estimated under the null. ■

Table 1. R^2 from regression of $\widehat{e}_t^2$ on $\widehat{\sigma}_t^2$

Kernel-based estimates

Bandwidth is $b = c\widehat{s}T^{-\frac{1}{p+4}}$

		$c = 0.5$	1	2
S&P 500 returns	$p = 1$	0.001	0.000	0.061
1965:1-1997:12	2	0.000	0.001	0.085
	3	0.012	0.041	0.077
Yen-dollar returns	$p = 1$	0.022	0.004	0.026
1978:10-1998:6	2	0.000	0004	0.003
	3	0.011	0.004	0.005
Excess holding yield	$p = 1$	0.016	0.004	0.170
1959:1-1998:2	2	0.002	0.034	0.010
	3	0.008	0.017	0.046

Table 2. R^2 from regression of $\widehat{e}_t^2$ on $\widehat{\sigma}_t^2$

Neural network estimates

	number of nodes =	2	4	8
S&P 500 returns	$p = 1$	0.041	0.040	0.039
1965:1-1997:12	2	0.034	0.036	0.036
	3	0.020	0.020	0.020
Yen-dollar returns	$p = 1$	0.010	0.019	0.025
1978:10-1998:6	2	0.229	0.261	0.278
	3	0.234	0.075	0.075
Excess holding yield	$p = 1$	0.006	0.006	0.006
1959:1-1998:2	2	0.032	0.031	0.032
	3	0.002	0.002	0.002

Table 3. R^2 from regression of $\widehat{e}_t^2$ on $\widehat{\sigma}_t^2$

Engle-Ng estimates

	number of bins =	2	4	8
S&P 500 returns	$p = 1$	0.138	0.139	0.138
1965:1-1997:12	2	0.107	0.107	0.151
	3	0.098	0.088	0.073
Yen-dollar returns	$p = 1$	0.015	0.022	0.031
1978:10-1998:6	2	0.003	0.001	0.001
	3	0.004	0.001	0.000
Excess holding yield	$p = 1$	0.124	0.125	0.114
1959:1-1998:2	2	0.350	0.338	0.259
	3	0.152	0.142	0.148

Table 4. Simulation results

GARCH parameters from S&P 500 data

Kernel-based estimate of the conditional variance

$p = 1$

Bandwidth	Median	Coverage rate of 95% CI		First-stage
constant	IV estimator	Wald	LM	R^2 (%)
Actual	7.865	79.1	94.0	2.19
0.5	0.427	92.2	81.2	5.29
1	1.230	90.5	92.2	3.12
2	2.214	81.4	93.0	2.09

Table 5. Simulation results

GARCH parameters from S&P 500 data

Kernel-based estimate of the conditional variance

$p = 2$

Bandwidth constant	Median IV estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
0.5	-0.186	99.1	53.6	24.63
1	0.586	96.9	77.9	11.07
2	1.873	92.4	91.7	5.37

Table 6. Simulation results

GARCH parameters from S&P 500 data

Kernel-based estimate of the conditional variance

$p = 3$

Bandwidth constant	Median IV estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
0.5	-0.665	92.9	74.9	51.50
1	0.334	98.9	65.7	22.61
2	2.144	97.2	89.2	9.75

Table 7. Simulation results

GARCH parameters from S&P 500 data

Neural network estimate of the conditional variance

$p = 1$

Number of nodes	Median IV estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
2	6.449	51.2	95.5	0.35
4	6.938	51.4	93.8	0.36
8	6.258	53.7	93.5	0.33

Table 8. Simulation results

GARCH parameters from S&P 500 data

Neural network estimate of the conditional variance

$p = 2$

Number of nodes	Median IV estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
2	1.219	60.9	79.6	0.57
4	1.031	61.1	80.6	0.52
8	1.556	62.4	81.6	0.40

Table 9. Simulation results

GARCH parameters from S&P 500 data

Neural network estimate of the conditional variance

$p = 3$

Number of nodes	Median IV Estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
2	1.061	56.5	82.6	0.59
4	0.746	56.8	85.1	0.46
8	1.404	53.4	86.5	0.34

Table 10. Simulation results

GARCH parameters from S&P 500 data

Engle-Ng estimate of the conditional variance

$p = 1$

Number of bins	Median IV estimator	Coverage rate of 95% CI		First-stage R^2 (%)
		Wald	LM	
Actual	7.865	79.1	94.0	2.19
2	6.213	82.4	94.5	2.26
4	5.713	86.2	95.5	2.63
8	6.039	86.7	95.5	2.69

Table 11. Estimation results

S&P 500 returns, 1965-1997, robust t-statistics in parentheses

Estimator		$p=1$	$p=2$	$p=3$
Kernel ($c=2$)	constant ($\times 10^{-2}$)	0.25 (0.63)	−0.14 (−0.39)	−1.15 (−2.29)
	$\widehat{e}_t^2$	−0.40 (−0.20)	1.62 (0.99)	6.92 (2.63)
	Wald 95% CI	$[-4.4, 3.6]$	$[-1.6, 4.8]$	$[1.8, 12.1]$
	LM 95% CI	$[-100, -14.6] \cup [-7.6, 100]$	$[-100, 100]$	$[-100, 100]$
	1st stage R^2	0.061	0.085	0.077
Neural network (4 nodes)	constant ($\times 10^{-2}$)	0.27 (0.39)	0.09 (0.15)	0.10 (0.12)
	$\widehat{e}_t^2$	−0.45 (−0.14)	0.38 (0.14)	0.42 (0.10)
	Wald 95% CI	$[-6.9, 6.0]$	$[-5.1, 5.9]$	$[-7.8, 8.6]$
	LM 95% CI	$[-14.6, 6.0]$	$[-55.0, 7.0]$	$[-100, 100]$
	1st stage R^2	0.040	0.036	0.020
Engle-Ng (4 bins)	constant ($\times 10^{-2}$)	0.17 (0.60)	0.16 (0.49)	−0.04 (−0.12)
	$\widehat{e}_t^2$	−0.01 (−0.01)	0.05 (0.04)	1.10 (0.64)
	Wald 95% CI	$[-2.5, 2.5]$	$[-2.8, 2.9]$	$[-2.2, 4.4]$
	LM 95% CI	$[-100, 100]$	$[-100, 100]$	$[-39, 6.4]$
	1st stage R^2	0.139	0.107	0.088
GARCH-M(1,1)	constant ($\times 10^{-2}$)	−0.94 (−0.84)		
	$\widehat{e}_{t-1}^2$	6.68 (1.04)		
	Wald 95% CI	$[-6.2, 19.54]$		

Table 12. Estimation results

Excess holding yield, 1959:1-1998:2, robust t-statistics in parentheses

Estimator		$p=1$	$p=2$	$p=3$
Kernel ($c=2$)	constant ($\times 10^{-3}$)	0.45 (1.92)	−1.07 (−0.65)	−0.23 (−0.37)
	spread	1.50 (6.31)	1.98 (2.17)	1.79 (4.23)
	$\widehat{e}_t^2$	−24.56 (1.22)	119.25 (0.85)	34.63 (1.01)
	Wald 95% CI	$[-38.9, -10.2]$	$[-155.1, 393.6]$	$[-32.6, 101.8]$
	LM 95% CI	$[-100, 100]$	$[-100, 100]$	$[-49.6, 99.8]$
	1st stage R^2	0.167	0.008	0.041
Neural network (4 nodes)	constant ($\times 10^{-3}$)	−6.79 (−5.67)	−6.18 (−6.77)	−2.48 (−0.61)
	spread	0.46 (2.50)	0.57 (3.72)	1.20 (1.82)
	$\widehat{e}_t^2$	21.21 (6.11)	19.40 (7.50)	8.11 (0.65)
	Wald 95% CI	$[14.4, 28.0]$	$[14.3, 24.5]$	$[-16.2, 32.4]$
	LM 95% CI	$[-100, 100]$	$[-100, 100]$	$[-100, 17.8] \cup [23.8, 100]$
	1st stage R^2	0.042	0.090	0.020
Engle-Ng (4 bins)	constant ($\times 10^{-3}$)	0.01 (0.03)	0.05 (0.25)	−0.02 (−0.07)
	spread	1.66 (6.81)	1.70 (7.21)	1.68 (6.19)
	$\widehat{e}_t^2$	9.27 (1.18)	6.81 (0.59)	8.34 (0.73)
	Wald 95% CI	$[-6.1, 24.6]$	$[-15.6, 29.3]$	$[-14.2, 30.8]$
	LM 95% CI	$[-100, 100]$	$[-100, 100]$	$[-100, 53.0]$
	1st stage R^2	0.125	0.338	0.142
GARCH-M(1,1)	constant ($\times 10^{-3}$)	0.03 (0.29)		
	spread	1.30 (4.74)		
	$\widehat{e}_{t-1}^2$	51.43 (2.78)		
	Wald 95% CI	$[14.4, 88.4]$		

Table 13. Estimation results

Yen-dollar returns, robust t-statistics in parentheses

Estimator		$p=1$	$p=2$	$p=3$
Kernel ($c=2$)	constant	-0.02 (-2.22)	-0.03 (-1.99)	-0.03 (-1.98)
	differential	-4.19 (-4.40)	-4.51 (-3.77)	-4.37 (-3.67)
	$\widehat{e}_t^2$	6.95 (0.69)	19.96 (1.19)	15.54 (1.07)
	Wald 95% CI	$[-12.7, 26.6]$	$[-12.9, 52.8]$	$[-13, 44.1]$
	LM 95% CI	$[-20.6, 89.8]$	$[-9.2, 100]$	$[-8, 100]$
	1st stage R^2	0.026	0.008	0.012
Neural network (4 bins)	constant	-0.06 (-1.61)	0.02 (1.09)	0.01 (0.57)
	differential	2.84 (0.48)	-4.09 (-4.40)	-4.46 (-4.04)
	$\widehat{e}_t^2$	55.00 (1.27)	-0.28 (-1.97)	-1.00 (-1.29)
	Wald 95% CI	$[-29.8, 139.8]$	$[-0.6, 0]$	$[-2.5, 0.5]$
	LM 95% CI	$[-100, -69] \cup [18.8, 100]$	$[-0.6, -0.2]$	$[-3.8, 0]$
	1st stage R^2	0.008	0.321	0.115
Engle-Ng (4 bins)	constant	0.02 (0.60)	0.17 (0.34)	0.00 (0.04)
	differential	-3.05 (-1.80)	-1.38 (-0.11)	-3.48 (-3.17)
	$\widehat{e}_t^2$	-33.26 (-1.04)	-218.11 (-0.38)	-14.95 (-0.34)
	Wald 95% CI	$[-96.3, 29.7]$	$[-1354.9, 918.7]$	$[-102.1, 72.19]$
	LM 95% CI	$[-100, -9] \cup [33.6, 100]$	$[-100, -28.4] \cup [29.8, 100]$	$[-100, 100]$
	1st stage R^2	0.022	0.001	0.001
GARCH-M(1,1)	constant	-0.01 (-6.27)		
	differential	-3.37 (-69.61)		
	$\widehat{e}_t^2$	1.95 (25139.67)		
	Wald 95% CI	$[1.95, 1.95]$		

Figure 1. Instrumental variable estimator

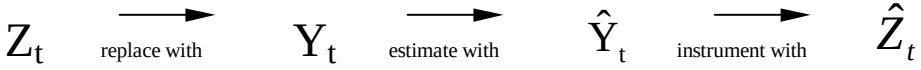

Fig. 2. Returns on S&P 500 index
1965–1997

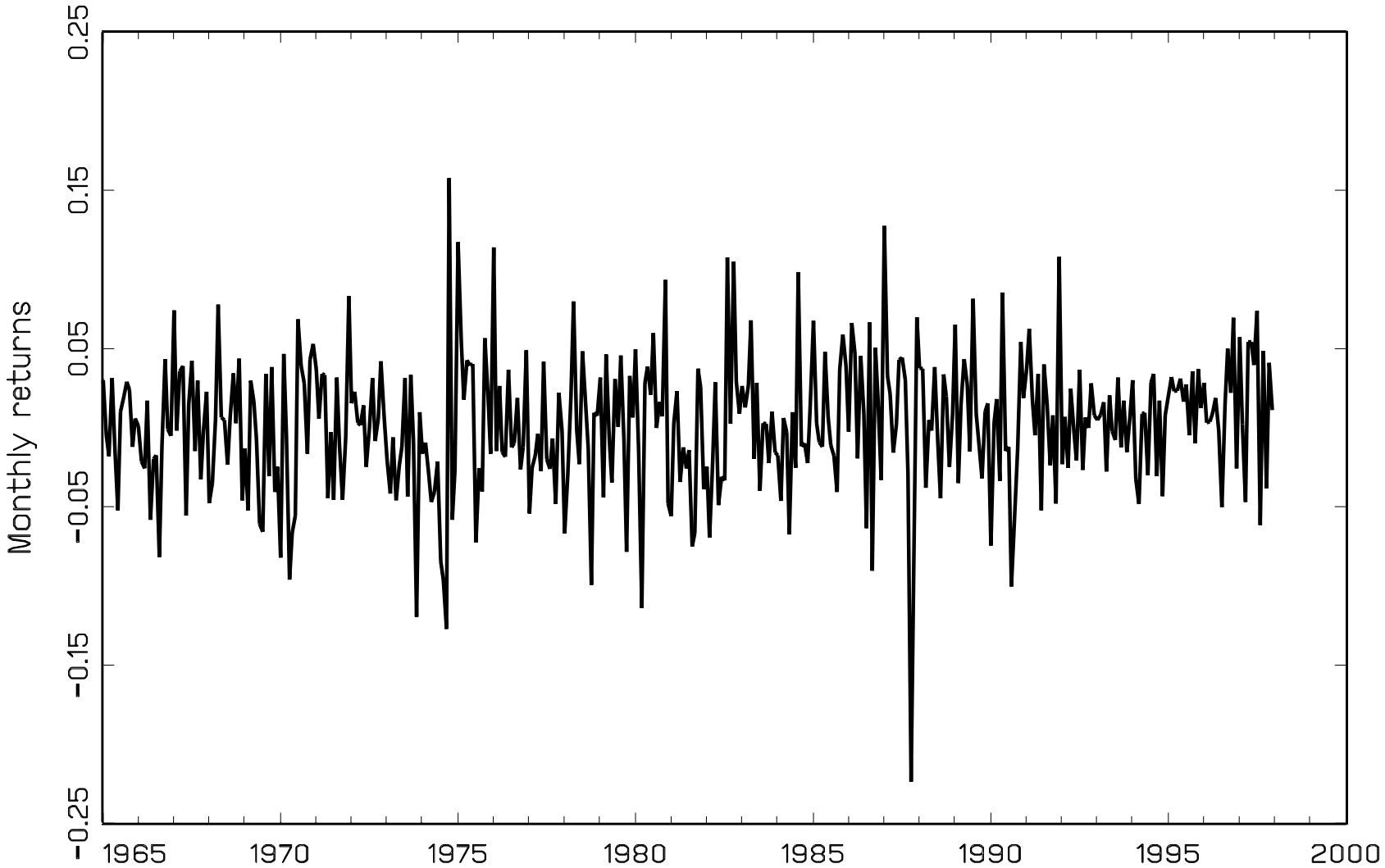

Fig. 3. Excess holding yield
1959:1–1998:2

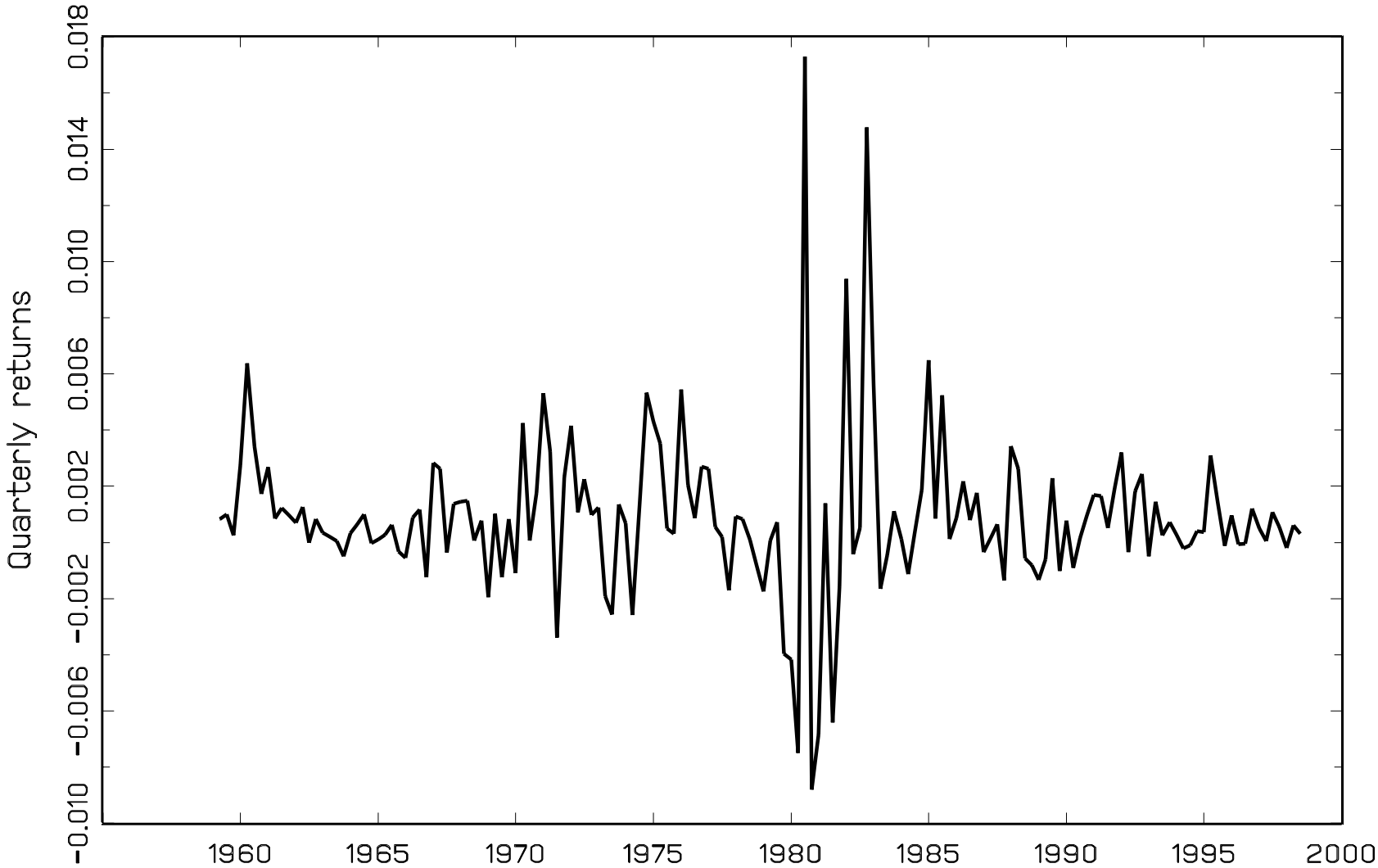

Fig. 4. Yen-dollar spot returns
1978-1998

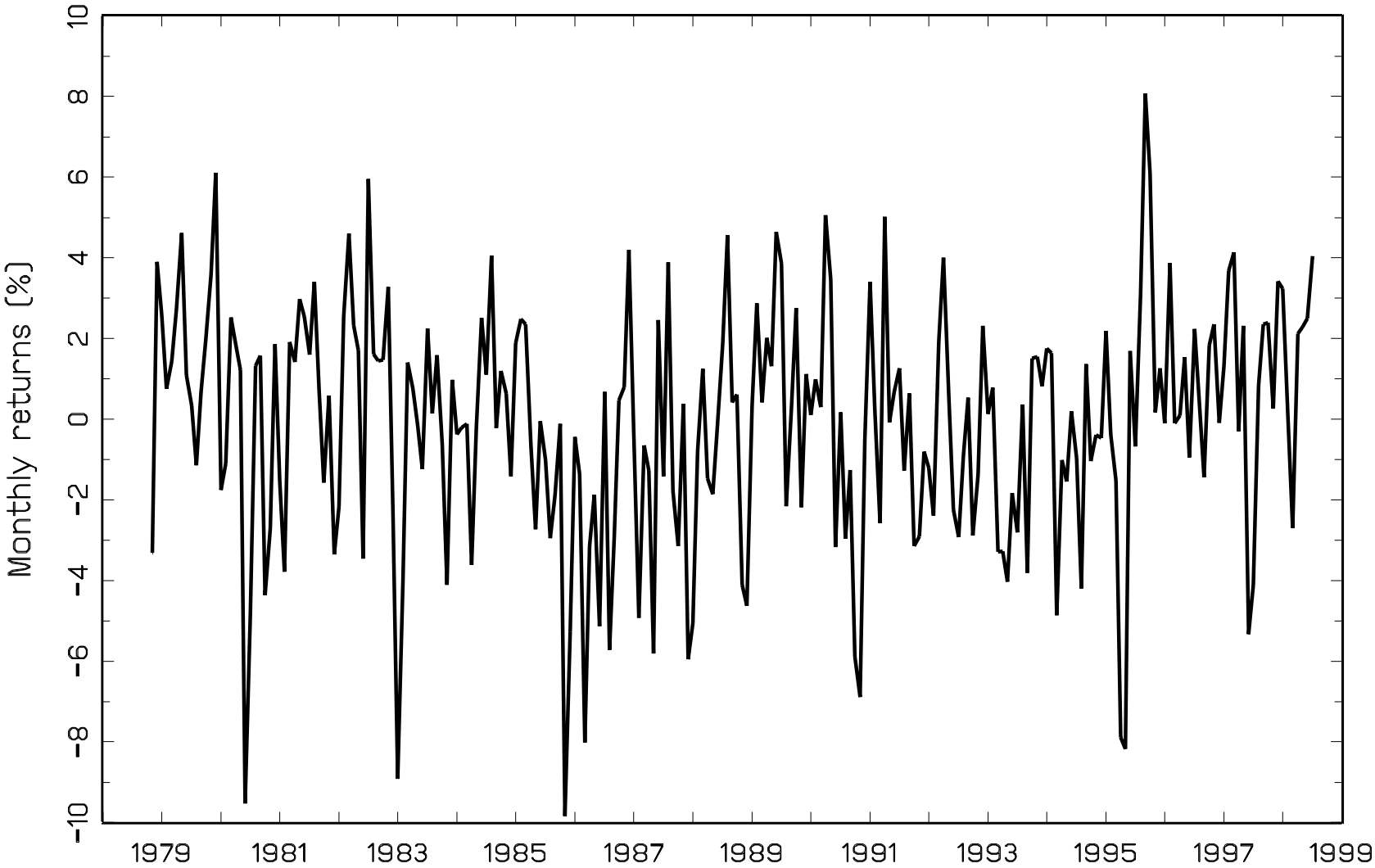

Fig. 5. Theoretical R-squared in GARCH(1,1) model

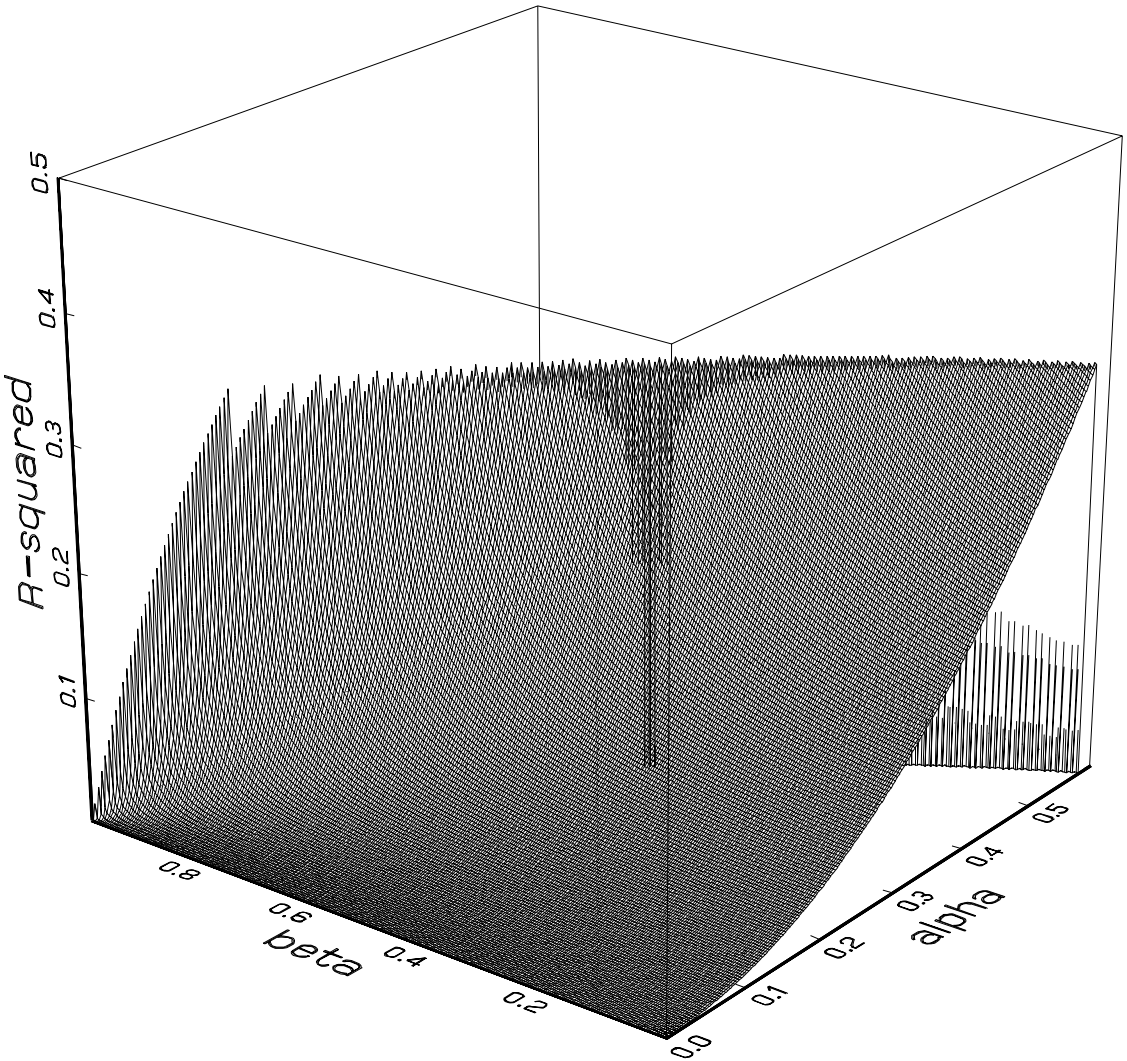

Fig 6. Distribution implied by GARCH(1,1) model
S&P 500 data
n = 400

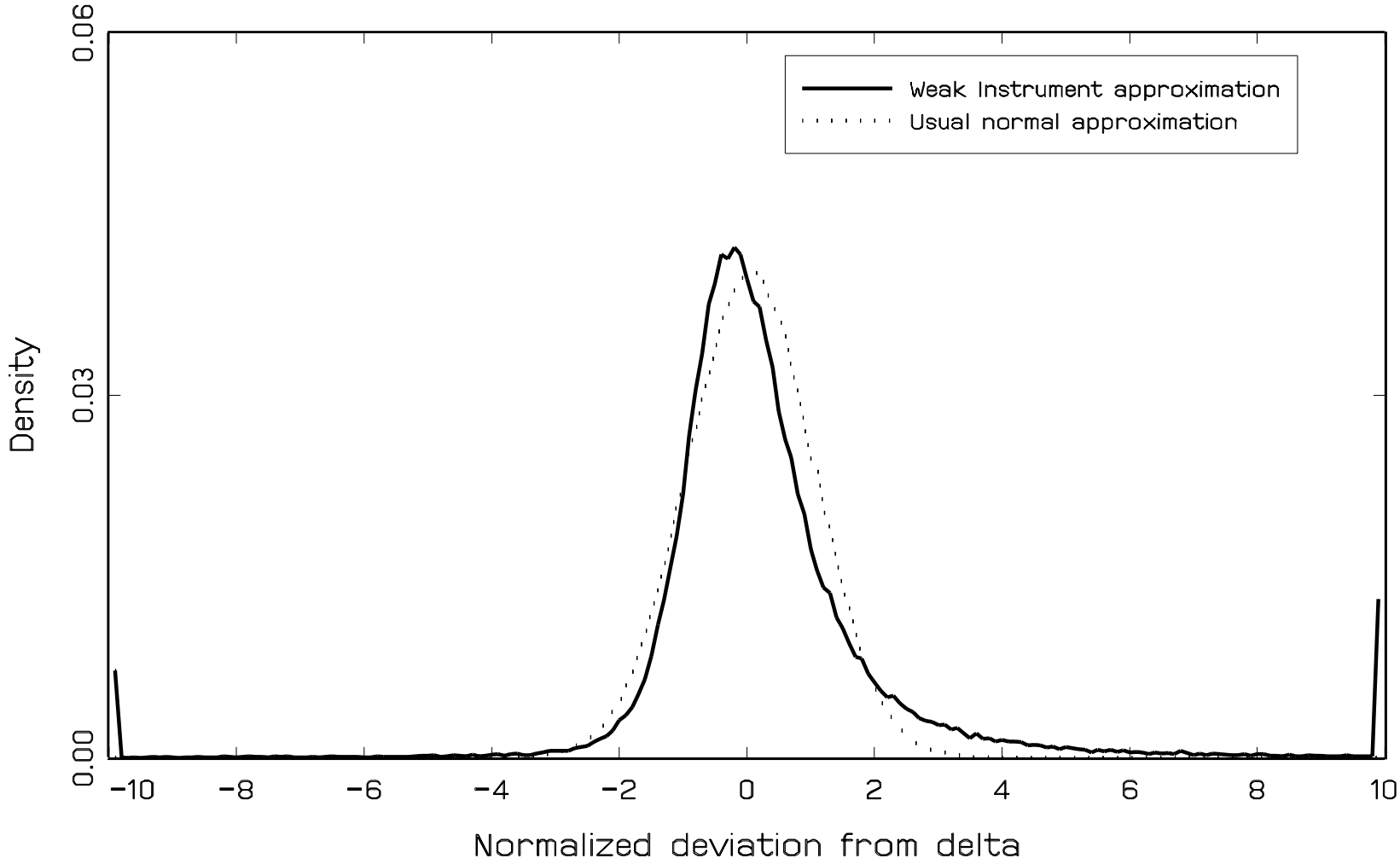

Fig 7. Distribution implied by GARCH(1,1) model
S&P 500 data
n = 5000

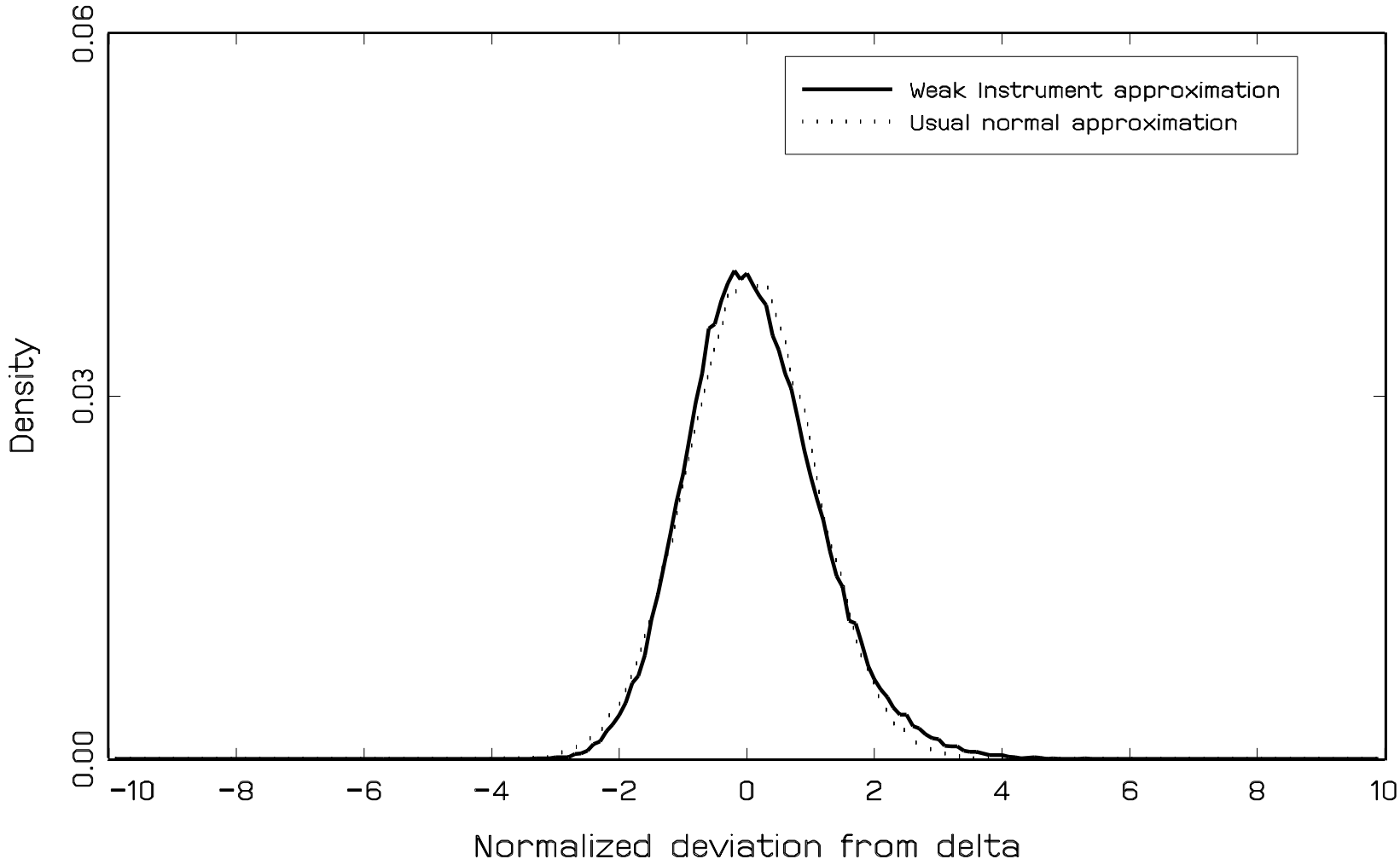

Fig. 8. Distribution of IV estimator
Actual values
n = 400

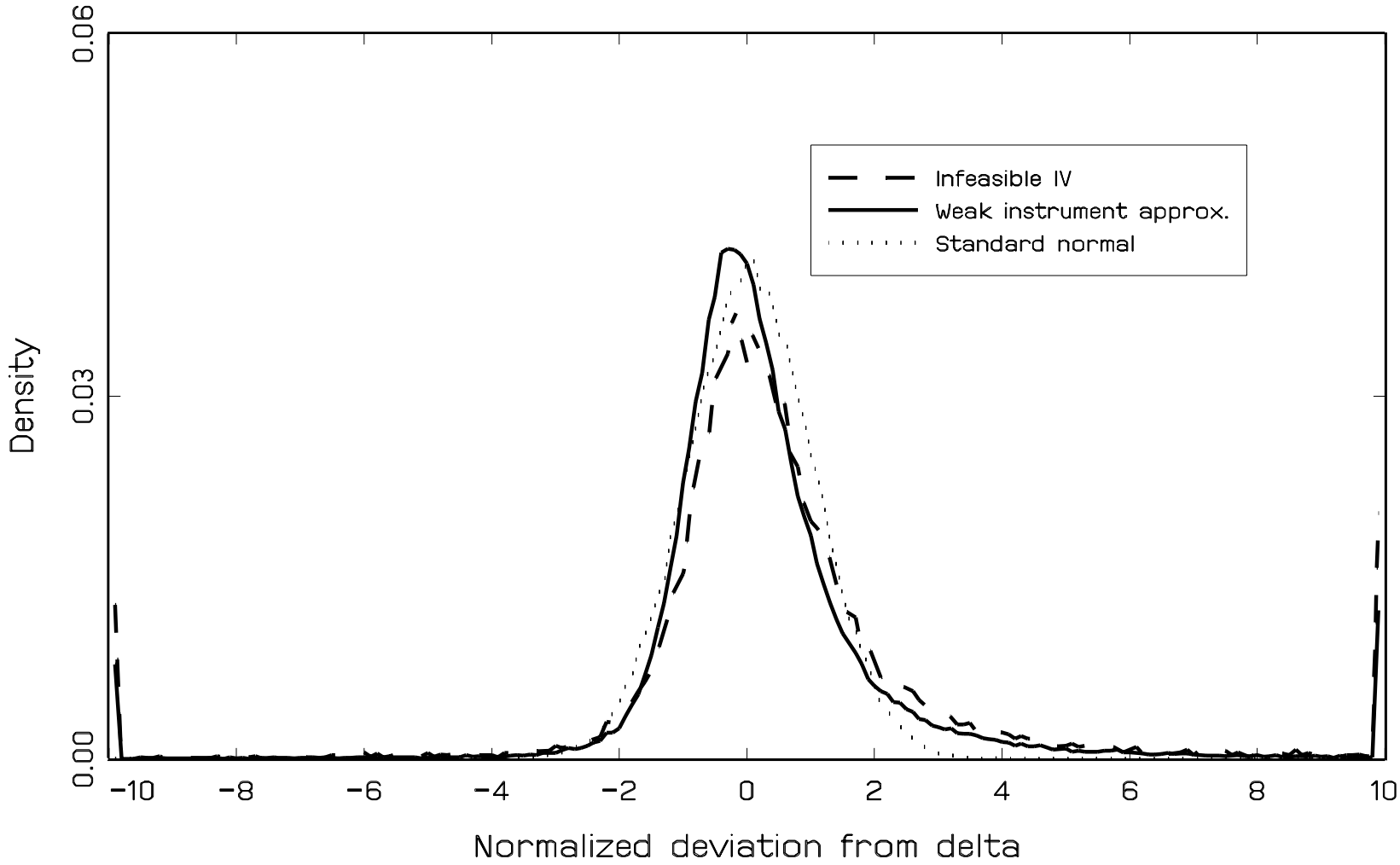

Fig. 9. Distribution of kernel-based IV estimator
n = 400, p = 1

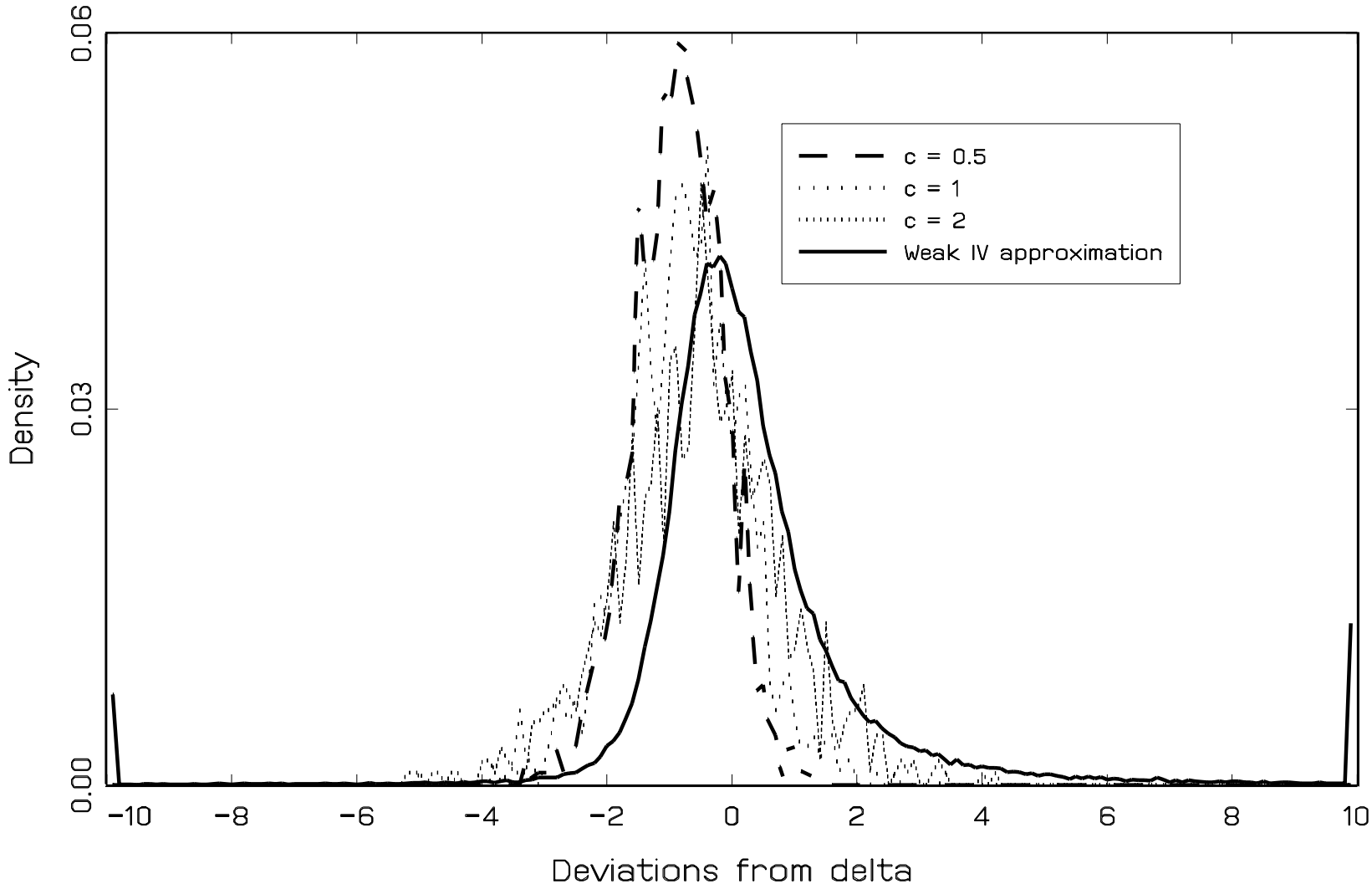

Fig. 10. Distribution of kernel-based IV estimator
n = 400, p = 2

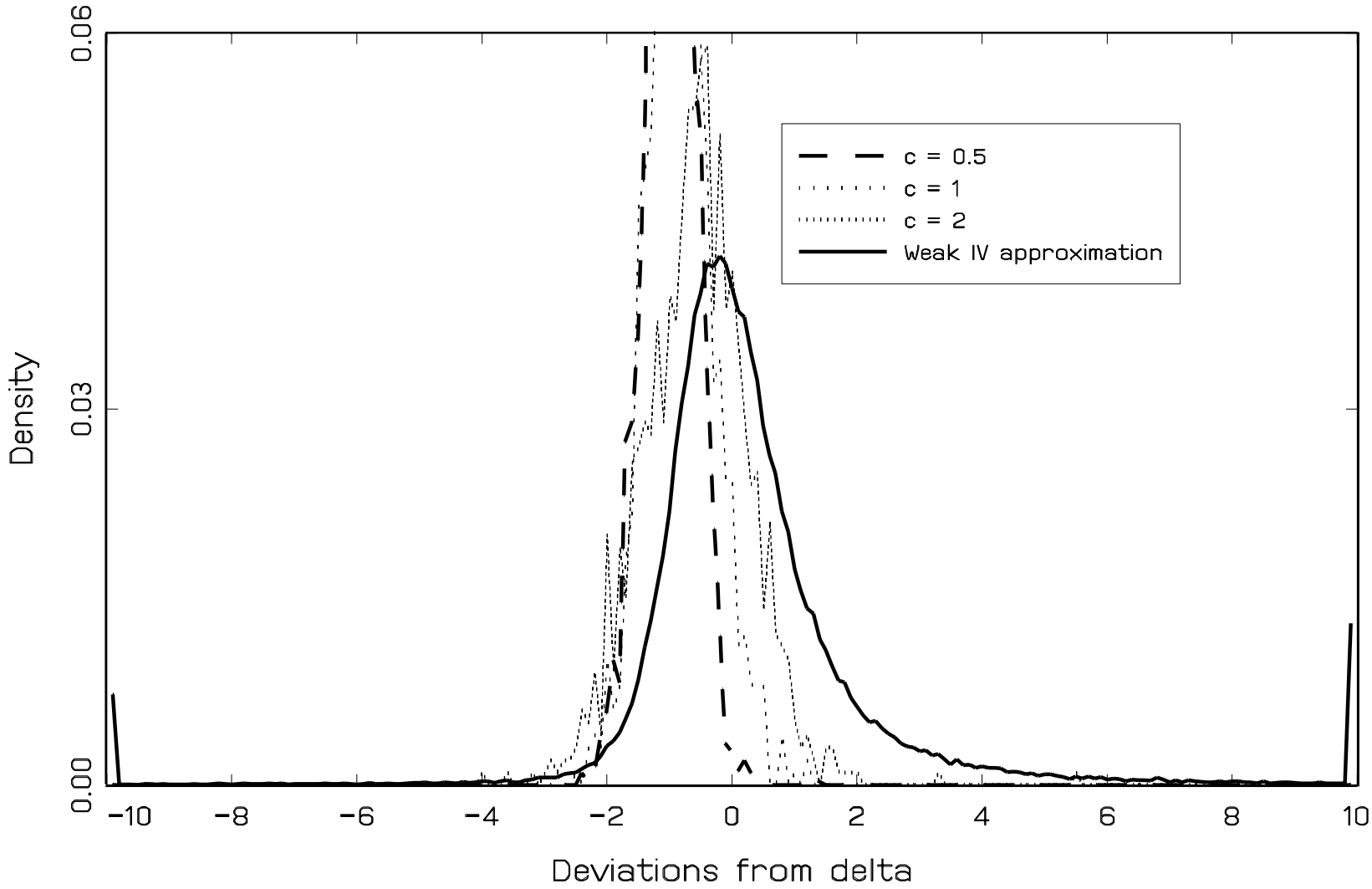

Fig. 11. Distribution of kernel-based IV estimator
n = 400, p = 3

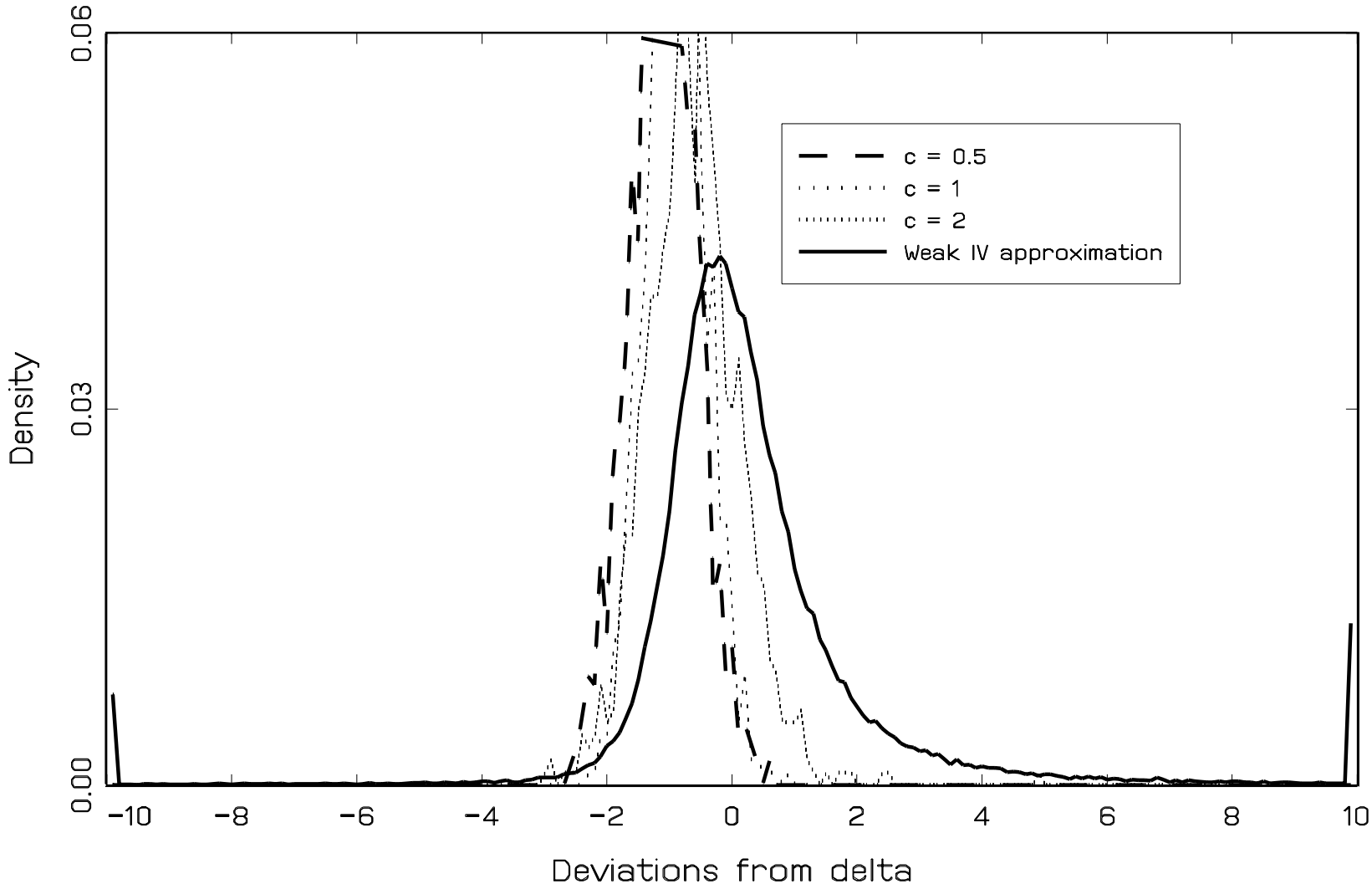

Fig. 12. Distribution of ANN-based IV estimator
n = 400, p = 1

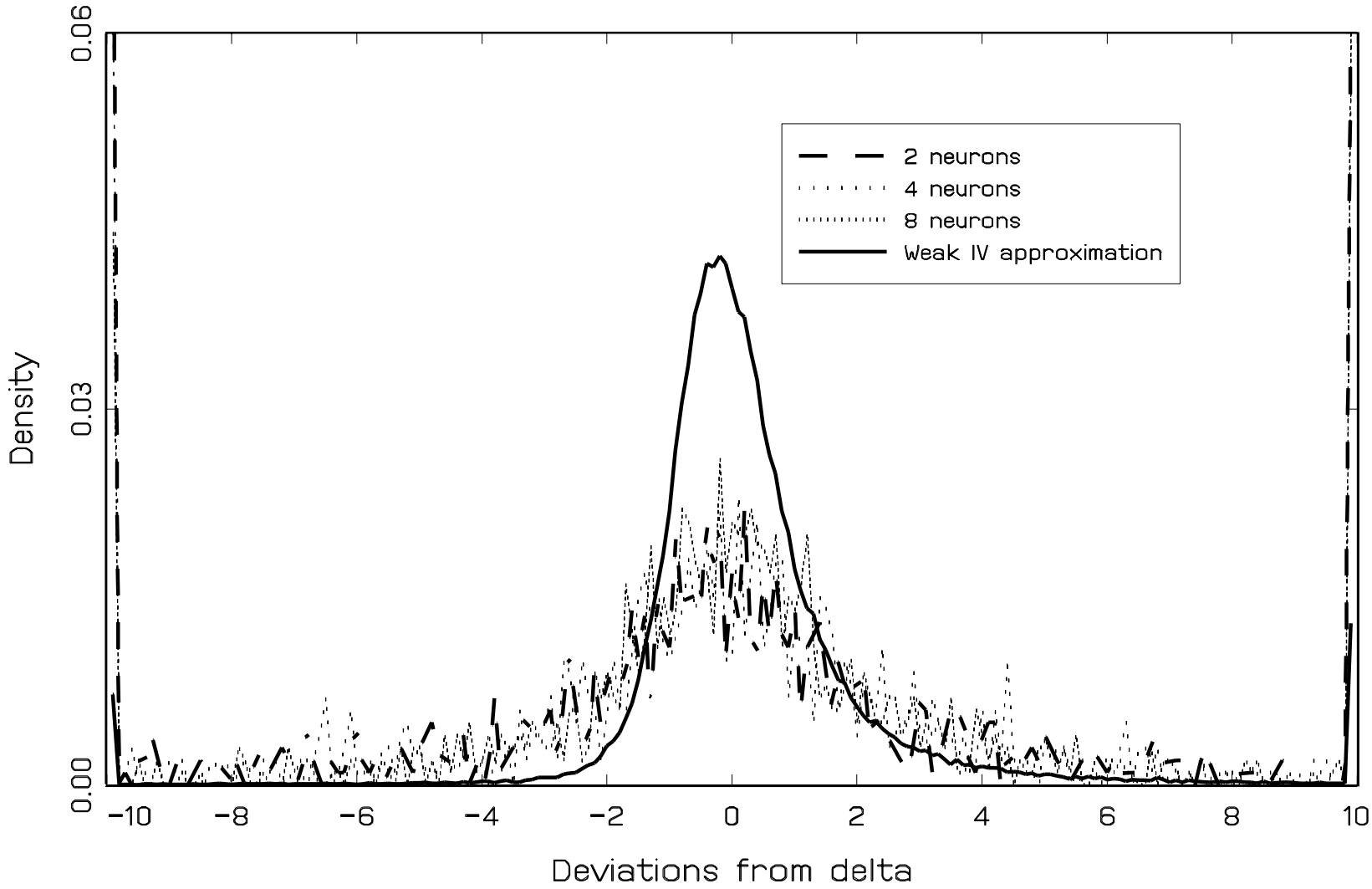

Fig. 13. Distribution of ANN-based IV estimator
n = 400, p = 2

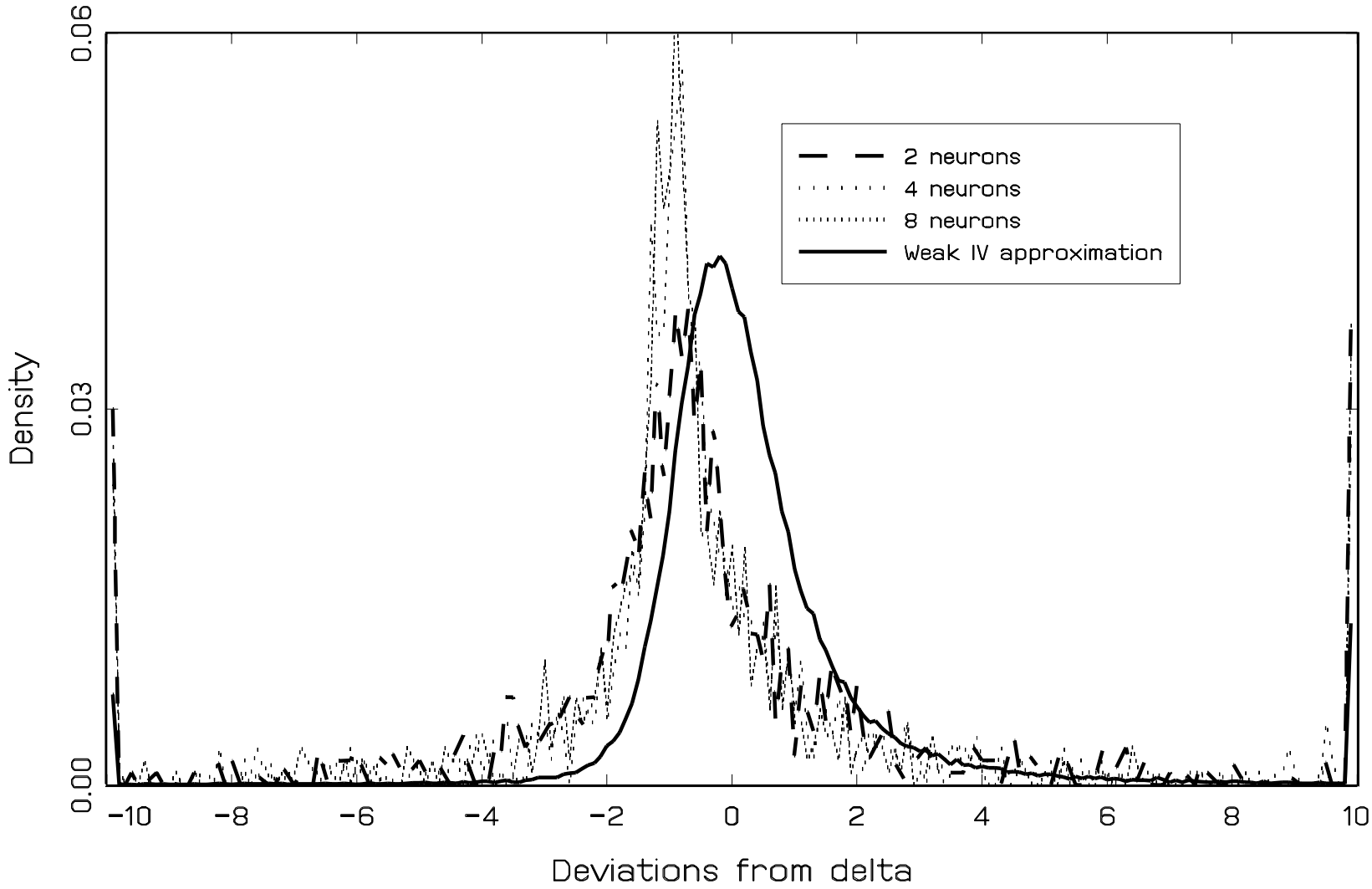

Fig. 14. Distribution of ANN-based IV estimator
n = 400, p = 3

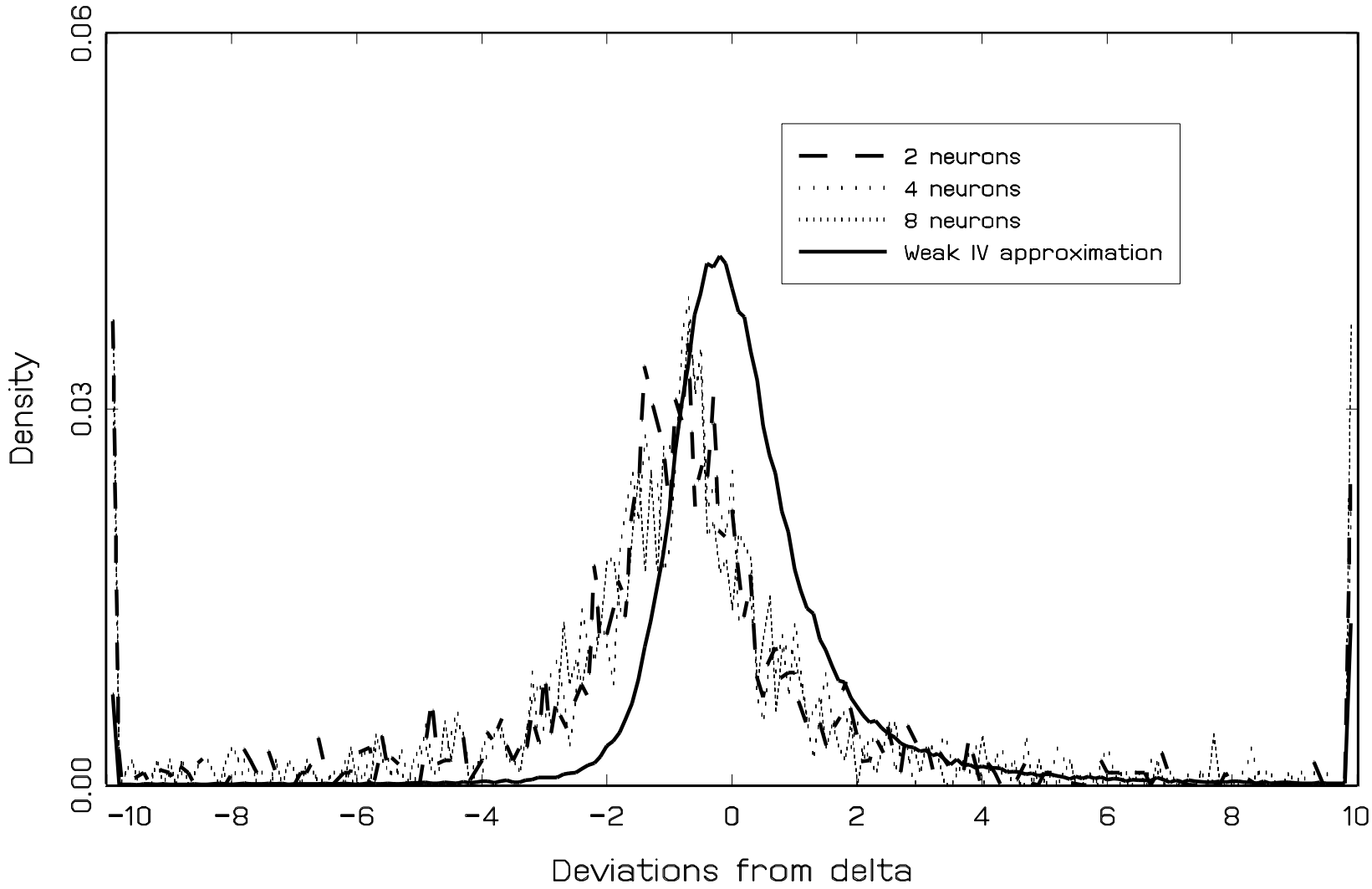

Fig. 15. Distribution of Engle–Ng IV estimator
n = 400, p = 1