The Shape of the Risk Premium: Evidence from a Semiparametric GARCH Model*

Oliver Linton† London School of Economics

Benoit Perron ‡ Université de Montréal

September 22, 2000

Abstract

We examine the relationship between the risk premium on the S&P500 index total return and its conditional variance. We propose a new semiparametric model in which the conditional variance process is parametric, while the conditional mean is an arbitrary function of the conditional variance. For monthly S&P 500 excess returns, the relationship between the two moments that we uncover is nonlinear and nonmonotonic. Moreover, we find considerable persistence in the conditional variance as well as a leverage effect as documented by others.

KEYWORDS: ARCH; Asset Pricing; Backfitting; Fourier Series; Kernel; Risk Premium.

JEL CLASSIFICATION: C13, C14, G12

*We are grateful to Adrian Pagan, participants at the 1999 EC^2 conference in Madrid and at seminars at Montéal, Queen's, and UC, Santa Barbara, and two anonymous referees for comments and discussion.

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: lintono@lse.ac.uk.

‡Corresponding author. Département de sciences économiques and CRDE, Université de Montréal C.P. 6128, Succursale Centre-ville, Montréal (Québec) Canada, H3C 3J7. Email: benoit.perron@umontreal.ca. Financial aassistance under the Mathematics of Information Technology and Complex Systems (MITACS) network is gratefully acknowledged.

1 INTRODUCTION

Modern asset pricing theories such as Abel (1987, 1998), Cox, Ingersoll, and Ross (1985), Merton (1973), and Gennotte and Marsh (1993) imply restrictions on the time series properties of expected returns and conditional variances on market aggregates. These restrictions are generally quite complicated, depending on utility functions as well as on the process driving asset returns. However, in an influential paper Merton (1973) obtained very simple restrictions albeit under somewhat drastic assumptions; he showed in the context of a continuous time partial equilibrium model that

$$\mu_t = E[(r_{mt} - r_{ft})|I_{t-1}] = \gamma \text{var}[(r_{mt} - r_{ft})|I_{t-1}] = \gamma\sigma_t^2, \tag{1}$$

where r_{mt}, r_{ft} are the returns on the market portfolio and risk-free asset respectively, while I_{t-1} is the market wide information available at time $t-1$. The constant γ is the Arrow–Pratt measure of relative risk aversion.

The simplicity of the above restrictions and their apparent congruence with the original CAPM restrictions (see Sharpe (1964) and Lintner (1965)) has motivated a large number of empirical studies that test some variant of this restriction. A convenient statistical framework for examining the relationship between the quantities μ_t and σ_t^2 in financial discrete time series is the ARCH class of models, see the survey papers of Bollerslev, Chou, and Kroner (1992) and Bollerslev, Engle, and Nelson (1994) for references. Engle, Lilien, and Robins (1987) examined the relationship between government bonds of different maturities using the ARCH–M model in which the errors follow an ARCH(p) process and $\mu_t = \mu(\sigma_t^2)$ for some parametric function $\mu(\cdot)$. They examined $\mu = \gamma_0 + \gamma_1\sigma_t$ and $\mu = \gamma_0 + \gamma_1 \log(\sigma_t^2)$, finding that the latter specification provided the better fit. French, Schwert, and Stambaugh (1987) and Nelson (1991) also examine this relationship using GARCH models.

Gennotte and Marsh (1993) argue that the linear relationship (1) should be regarded as a very special case. They construct a general equilibrium model of asset returns and derive the equilibrium relationship

$$\mu_t = \gamma\sigma_t^2 + g(\sigma_t^2), \tag{2}$$

where the form of $g(\cdot)$ depends on preferences and on the parameters of the distribution of asset returns. If the representative agent has logarithmic utility, then $g \equiv 0$ and the simple restrictions of Merton pertain. In addition, Backus, Gregory, and Zin (1989) and Backus and Gregory (1993) provide simulation evidence that, $g(\cdot)$ and hence $\mu(\cdot)$ could be of arbitrary functional form in general equilibrium.

Pagan and Hong (1990) argue that the risk premium μ_t and the conditional variance σ_t^2 are highly nonlinear functions of the past whose form is not captured by standard parametric GARCH–M models. They estimate μ_t and σ_t^2 nonparametrically finding evidence of considerable nonlinearity.

They then estimated δ from the regression

$$r_{mt} - r_{ft} = \beta' x_t + \delta \sigma_t^2 + \eta_t, \tag{3}$$

by least squares and instrumental variables methods with σ_t^2 substituted by the nonparametric estimate, finding a negative but insignificant δ. Perron (1999) analyses this approach using weak instrument asymptotics and finds similar results.

There are a number of drawbacks with their approach. Firstly, the conditional moments are calculated using a restricted conditioning set - the information set used in defining μ_t, σ_t^2 contained only a finite number of lags, i.e., $I_{t-1} = \{y_{t-1}, \ldots, y_{t-p}\}$ for some fixed p and data series $y_t = r_{mt} - r_{ft}$. This greatly restricts the dynamics for the variance process. In particular, if the conditional variance is highly persistent, the non-parametric estimator of the conditional variance provides a poor approximation as confirmed by the simulation evidence reported in Perron (1998). Secondly, linearity of the relationship between μ_t and σ_t^2 is imposed, and this seems to be somewhat restrictive in view of earlier findings.

In this paper, we investigate the relationship between the risk premium and the conditional variance of excess returns on the CRSP value-weighted index. We consider a semiparametric specification that differs from previous treatments. In particular, we choose a parametric form for the variance dynamics (in our case EGARCH), while allowing the mean to be an unknown function of σ_t^2. This model takes account of the high level of persistence and leverage effect found in stock index return volatility, while at the same time allowing for an arbitrary functional form to describe the relationship between risk and return at the market level. We develop two estimation methods for this model: a Fourier series method and a method based on kernels. The kernel method is based on iterative one-dimensional smoothing and is similar in this respect to the backfitting method for estimating additive nonparametric regression, see Hastie and Tibshirani (1990). We also suggest a bootstrap algorithm for obtaining confidence intervals. Using these methods, we find evidence of a nonlinear relationship between the risk premium and the conditional variance.

In the next section we discuss the specification of our model, while in Section 3 we describe how to obtain point and interval estimates. In Section 4, we present our empirical results. In section 5, we present the results of a small simulation experiment, while section 6 concludes.

2 A SEMIPARAMETRIC-MEAN EGARCH MODEL

We suppose that the realized risk premium y_t is generated as follows

$$y_t = \mu(\sigma_t^2) + \varepsilon_t \sigma_t, \ t = 1, 2, \ldots, T, \tag{4}$$

where ε_t is a martingale difference sequences with unit variance, while $\mu(\cdot)$ is a smooth function, but of unknown functional form. The restriction that $E[y_t|\mathcal{F}_{t-1}]$, where $\mathcal{F}_{t-1} = \{y_{t-j}\}_{j=1}^{\infty}$, only depends on the past through σ_t^2 is quite severe but is a consequence of asset pricing models such as for example Backus and Gregory (1993) and Gennotte and Marsh (1993). In any case, it is possible to generalize this formulation in a number of directions. It is straightforward to incorporate fixed explanatory variables, lagged σ_t^2, or lagged y_t either as linear regressors or inside the unknown function $\mu(\cdot)$. More complicated dynamics for ε_t, such as an ARMA(p, q) model, and a multivariate extension can also be accommodated.

We propose using a parametric function for the conditional variance so as to allow for rich dynamics in the volatility. To be specific we shall largely consider the Exponential GARCH model introduced by Nelson (1991):

$$h_t \equiv \log(\sigma_t^2) = a + \sum_{j=1}^{p} b_j \log(\sigma_{t-j}^2) + \sum_{k=1}^{q} \left[c_k \left(|\varepsilon_{t-k}| - E\left|\varepsilon_{t-k}\right|\right) + d_k \varepsilon_{t-k}\right]. \tag{5}$$

The presence of the lagged dependent variables h_{t-j} ensures very rich dynamics for the variance process itself; richness which cannot be achieved as yet by nonparametric models of the conditional variance. The above model also allows both the sign and the level of ε_{t-k} to affect σ_t^2 — good news and bad news can have different effects on volatility, hence allowing the possibility of the so-called leverage effect in stock returns. The parameter d controls the relative importance of the symmetric versus asymmetric effects. A number of economic arguments have been advanced to support this specification. For example, Black (1976) and Christie (1982) suggest that since downside risk to the owners of a company is limited by bankruptcy laws, owners have an incentive to adopt more risky investment when the value of the firm is low. Therefore, return volatility will be negatively correlated with returns. Evidence of a leverage effect in stock returns is widespread in the literature and can be found in Nelson (1991) for daily data and in Braun, Nelson, and Sunier (1991) for monthly data.

A number of authors, e.g., Nelson (1991), have found that standardized residuals from estimated GARCH models are leptokurtic relative to the normal, see also Engle and Gonzalez–Rivera (1991). We therefore assume that ε_t has a distribution within the exponential power family

$$f(\varepsilon) = \frac{\nu \exp\left(-\frac{1}{2}|\varepsilon/\lambda|^{\nu}\right)}{\lambda 2^{(1+1/\nu)}\Gamma(1/\nu)};\ \lambda = [2^{(-2/\nu)}\Gamma(1/\nu)/\Gamma(3/\nu)]^{1/2}, \tag{6}$$

where Γ is the gamma function. The GED family of errors includes the normal, uniform and Laplace as special cases. The distribution is symmetric about zero for all ν, and has finite second moments for $\nu > 1$. With this density, we obtain that $E\left|\varepsilon_t\right| = (\lambda 2^{1/\nu}\Gamma\left(2/\nu\right))/\Gamma\left(1/\nu\right)$ [see Hamilton (1994), p. 669].

We assume that the parameter values satisfy the requirements for stationarity given in Nelson (1991). Carrasco and Chen (1999) establish a general result about the dependence properties of

a general class of volatility models, which suggests that the process y_t is β-mixing under some conditions.

Newey and Steigerwald (1997) have recently shown how quasi-likelihood estimators in GARCH models based on distributions other than the normal are generally inconsistent. Therefore, we also investigate our EGARCH(p,q) specification for the variance combined with a normal error distribution.

The main difference between our model and previous treatments is that we do not restrict the functional form of $\mu(\cdot)$ d sulrul. This has a number of implications both for estimation and testing. In particular, a simple consistent estimator of $\mu(\cdot)$ is difficult to obtain and would appear to depend on first obtaining consistent estimates of the parameters of the variance process. On the other hand, to estimate these parameters we need to have a good estimate of $\mu(\cdot)$. In the next section we propose a solution to this problem.

3 ESTIMATION

3.1 Parametric Estimation

Estimation of the unknown parameters by maximum likelihood when $\mu(\cdot)$ is known apart from a finite number of parameters, say τ, is considered in Engle, Lilien, and Robins (1987) and Nelson (1991). In this case, let $\theta = (\phi, \tau)$, where $\phi = (a, b_1, \ldots, b_p, c_1, \ldots, c_q, d_1, ..., d_q, \nu)'$, while τ are the vector of unknown mean parameters. Then $\varepsilon_t(\theta)$ and $h_t(\theta)$ can be built up recursively given initial conditions, and the conditional log-likelihood function is

$$\ell_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = \sum_{t=1}^{T} \log(f(\varepsilon_t(\theta); \nu) - \frac{1}{2} \sum_{t=1}^{T} h_t(\theta), \tag{7}$$

The likelihood function can be maximized with respect to ϕ, τ using the BHHH algorithm, viz.

$$\theta^{[i+1]} = \theta^{[i]} - \lambda^{[i]} \left[\sum_{t=1}^{T} \dot{\ell}_{t\theta} \dot{\ell}'_{t\theta} \right]^{-1} \sum_{t=1}^{T} \dot{\ell}_{t\theta}, \tag{8}$$

where $\lambda^{[i]}$ is a variable step length chosen to maximize the log likelihood function in the given direction, and the score functions $\dot{\ell}_{t\theta}$ are evaluated at $\theta^{[i]}$. Although the likelihood function is not smooth in all parameters [because of the presence of the absolute value of ε], this derivative based method seems to work well in practice.

3.2 Semiparametric Estimation

We propose several methods of constructing estimates of ϕ and $\mu(\cdot)$ in the semiparametric model. We estimate μ using two main approaches: the first one consists of treating the $T \times 1$ vector $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_T)'$ as unknown parameters and estimating them through a kernel smoothing method inside the optimization routine. The second approach is to parametrize $\mu(\cdot)$ in a flexible way using series expansion methods. The basis we will use is the Fourier Flexible Form of Gallant (1981). Estimation of ϕ is then achieved by concentrating the likelihood function. We describe the estimation and the construction of confidence intervals for each method in turn.

3.2.1 Kernel Estimation

The first method estimates μ by a smoothing procedure based on kernels [see Härdle (1990) and Härdle and Linton (1994) for a discussion of kernel nonparametric regression estimation]. Suppose that we could obtain some estimate of $\mu(.)$, then one could easily estimate the parameters of the variance and error distribution using maximum likelihood on the residuals. Unfortunately, in our time series model, the relevant information set is the entire infinite past, i.e., $\mu(.) = E[y_t|\mathcal{F}_{t-1}]$ depends on the entire past of the series. One could argue — as do Pagan and Hong (1990) — that consistent estimates of $E[y_t|\mathcal{F}_{t-1}]$ could be obtained using nonparametric regression with a truncated information set $\mathcal{F}_{t-1}^{P(T)} = \{y_{t-1}, \ldots, y_{t-P}\}$, where $P(T) \Rightarrow \infty$ at a very slow rate. This estimate could then be used to obtain consistent estimates of the parameters of h_t. This is not a particularly appealing procedure from a practical point of view because of the high dimension of the conditioning set. Silverman (1986) dramatically illustrates the curse of dimensionality by showing the effective sample size needed to achieve a certain precision.

In many other semiparametric problems one can use a semiparametric profile likelihood method as described in Powell (1994) in which the nonparametric function is estimated for each given parameter value and then the parameters are chosen to minimize some criterion function that would have been the likelihood if the functions were known rather than estimated. In general such estimators are root-n consistent and asymptotically normal. However, there is a cost to not knowing the function μ, i.e., the semiparametric information bound is generally lower than the information bound when μ is finitely parameterized.

In our time series model, we can't define the corresponding profiled quantity $\widehat{\mu}_\phi(\sigma_t^2)$ so easily, since σ_t^2 depends on lagged ε's, which in turn depend on lagged μ's. Therefore, we need to know the entire function $\mu(.)$ [or at least its values at n sample points] to construct $\widehat{\mu}_\phi(\sigma_t^2)$. This might at first glance appear to make the estimation procedure hopeless, but this is a false impression. The same sorts of issues arise in the estimation of additive nonparametric models and an enormous

literature has arisen that proposes estimation algorithms, and, more recently, distribution theory, see for example Breiman and Freedman (1985), Hastie and Tibshirani (1990), Opsomer and Ruppert (1997), and Mammen, Linton, and Nielsen (1998). We borrow from this literature and suggest an estimation procedure based on iterative updating of both the finite dimensional parameters and the function $\mu(.)$. Our procedure first requires picking starting values for $\underline{\mu}$ and ϕ. We then define a modified version of the Newton–Raphson algorithm to update our estimates of ϕ. We then update our estimates of $\underline{\mu}$ using kernel estimates based on the previous iterations filtered log variances. The main advantage of the procedure is that it relies on only one-dimensional smoothing operations at each step, so that the curse of dimensionality does not operate. The main disadvantage is that the procedure is time consuming and may not converge or may converge to local minima etc.

For convenience we describe our algorithm for the case $p = q = 1$. We smooth on the log of variance h_t instead of the variance itself. Since the logarithm is a monotonic transformation, the two approaches are equivalent. Our main algorithm is as follows.

Kernel Estimation Algorithm

1. Choose starting values for $\phi^{[1]}$ and $\{h_s^{[1]}\}_{s=1}^T$.

2. Given $\{h_t^{[r-1]}\}_{t=1}^T$, calculate
$$\mu_t^{[r]} = \frac{\sum_{s \neq t} K\left(\frac{h_t^{[r-1]} - h_s^{[r-1]}}{\delta}\right) y_s}{\sum_{s \neq t} K\left(\frac{h_t^{[r-1]} - h_s^{[r-1]}}{\delta}\right)} \tag{9}$$
for $t = 1, 2, \ldots, T$, where $\delta > 0$ is a small bandwidth parameter, while K is a bounded kernel satisfying $\int K(u)du = 1$.

3. Given initial values $h_0^{[r]}(\phi)$ and $\varepsilon_0^{[r]}(\phi)$, define recursively for any parameter value ϕ
$$\begin{aligned} h_t^{[r]} &= a + bh_{t-1}^{[r]} + c_1\left(|\varepsilon_{t-1}^{[r]}| - E|\varepsilon_{t-1}^{[r]}|\right) + d_1\varepsilon_{t-1}^{[r]}, \\ \varepsilon_t^{[r]} &= \frac{y_t - \mu_t^{[r]}}{\exp(h_t^{[r]})}, \end{aligned}$$
for $t = 1, 2, \ldots, T$. Then for any ϕ construct $\ell_t^{[r]}(\phi) = \ell_t(\phi; \underline{\mu}^{[r]})$, the period t contribution to the r^{th} pseudo likelihood function, where $\underline{\mu}^{[r]} = (\mu_1^{[r]}, \ldots, \mu_T^{[r]})'$.

4. Calculate
$$\phi^{[r+1]} = \phi^{[r]} - \lambda^{[r]}\left[\sum_{t=1}^T \dot{\ell}_{t\phi}^{[r]} \dot{\ell}_{t\phi}^{\prime[r]}\right]^{-1} \sum_{t=1}^T \dot{\ell}_{t\phi}^{[r]}, \tag{10}$$
where $\dot{\ell}_{t\phi}^{[r]}$ is the vector of partial derivatives of $\ell_t^{[r]}(\phi)$ with respect to ϕ evaluated at $\phi^{[r]}, \underline{\mu}^{[r]}$.

5. Repeat until convergence. We define convergence in terms of the relative gradient and the change in the nonparametric estimate, i.e.,

$$\max\left\{\max_k\left|\frac{\sum_{t=1}^T \dot{\ell}_{t\phi_k}\cdot\phi_k}{\ell(\phi)}\right|, \frac{1}{T}\sum_{t=1}^T\left|\frac{\underline{\mu}^{[r+1]}-\underline{\mu}^{[r]}}{\underline{\mu}^{[r]}}\right|\right\}<\varepsilon, \tag{11}$$

for some small prespecified ε. Denote the resulting estimates by $\hat{\phi}$ and $\hat{\underline{\mu}}$. ■

We are unable to prove convergence of the above algorithm, although in practice it seems to work reasonably well and to give similar answers for a range of starting values. Note that convergence of the backfitting algorithm for separable nonparametric regression has only been shown in some special cases, specifically when the estimator is linear in the dependent variable. However, backfitting has been defined and widely used to estimate more general models than additive nonparametric regression [see Hastie and Tibshirani (1990)], and is widely believed to do a good job in such cases.

An alternative implementation is to iterate to convergence on the computation of ϕ in step 4 for each $\mu_t^{[r]}$, and then to update $\mu_t^{[r]}$ as in step 2 above. One can also modify the algorithm to substitute more recent values of μ in step ?

In practice, the estimated parameters of h_t appear to be quite robust to different parametric specifications of the mean equation. The filtered estimate of h_t based on $\mu_t^{[0]} = T^{-1}\sum_{s=1}^T y_s$ should be close to the true h_t and should provide good starting values. We also use the fitted values from am EGARCH-M model as starting values to check for robustness. As in the parametric case, additional iterations should improve the performance of the estimated parameters and function.

The stopping rule (11) was arrived at after some experimentation. It is desirable to ensure that the entire parameter vector (ϕ, μ) is convergent.

3.2.2 Fourier Series Estimation

The second approach is to parametrize the mean equation using a flexible functional form. By letting the number of terms grow with sample size and with a suitable choice of basis functions, this method can approximate arbitrary functions. This is an example of sieve estimation, but for a given sample size, it reduces to a parametric method with a finite number of parameters, and the estimation algorithm is just the BHHH given above.

The basis we will use is the flexible Fourier form of Gallant (1981) which adds sine and cosine terms to a quadratic function. Because it uses trigonometric terms, it is convenient for the data to lie in the $[0, 2\pi]$ interval. To do so, we recenter and rescale the estimates of h_t and define a new

variable

$$h_t^* = (h_t - \underline{h}) \frac{2\pi}{\left(\overline{h} - \underline{h}\right)},$$

where $\underline{h}$ and $\overline{h}$ are scalars such that $\underline{h}$ is less than $\min(h_t)$ and $\overline{h}$ is greater than $\max(h_t)$. Then the Fourier approximation is

$$\mu(h_t^*) = \gamma_0 + \gamma_1 h_t^* + \gamma_2 h_t^{*2} + \sum_{j=1}^{M} \psi_j \sin(j h_t^*) + \sum_{j=1}^{M} \varphi_j \cos(j h_t^*). \tag{12}$$

The number of terms to estimate is $p + 2q + 2M + 4$.

4 INFERENCE

There is a general theory of inference for maximum likelihood and quasi-maximum likelihood estimators in time series, see Wooldridge (1994) for a state of the art survey. Specifically, Bollerslev and Wooldridge (1992) showed, under high level conditions, that quasi-maximum likelihood estimators of parameters in an ARCH model can be consistent and asymptotically normal provided only that the mean and the variance equations are correctly specified. However, their theory was based on high-level conditions, which turned out to be rather difficult to verify even in the most simple cases. Papers that have derived the asymptotic theory for these models from primitive conditions are: Weiss (1986) for ARCH models, and Lumsdaine (1996) and Lee and Hansen (1994) for the GARCH(1,1) model. For other specifications in the GARCH class, the asymptotic theory that is used in practice is not known to be valid. Specifically, no-one has established the distribution theory for the EGARCH model of Nelson even in the special case with no mean effects and normal errors. However, there is much simulation evidence to support the normal approximation in the general class of models, and the results of Bollerslev and Wooldridge (1992) are widely believed to hold more generally, and are frequently used in practice.

Given the complicated structure of our semiparametric model it is not surprising that we cannot provide rigorous asymptotic theory for our estimators. However, if h_t were observed, a kernel estimate of $\mu(\cdot)$ as in (9) would be consistent and asymptotically normal under appropriate conditions, since the process h_t is weakly dependent. Therefore, the results of Robinson (1983) can be applied to establish consistency, provided $\delta(T) \to 0$ at an appropriate rate; this argument can be extended to the case where h_t is replaced by a consistent parametric estimate. Indeed, the asymptotic distribution of nonparametric estimates is usually independent of any preliminary parametric estimations [Powell (1994)]. We therefore expect $\hat{\mu}_t$ to be consistent. As regards $\hat{\phi}$, we expect it to be $\sqrt{T}$ consistent and to have a limiting normal distribution with the variance including some component arising from the estimation of μ.

We now turn to construction of standard errors for the parameter estimates and the risk premium. In the former case, we report analytical and bootstrap standard errors. The analytical standard errors are obtained by taking the outer product of the gradient with respect to the estimated parameters. For the kernel estimators, the estimated parameters are just ϕ, the parameters of the error distribution and the variance process, while for the series estimator we are estimating these parameters jointly with the pseudo parameters τ of the mean function. For the series estimator we therefore compute standard errors from the matrix $[\sum_{t=1}^{T} \dot{\ell}_{t\theta}\dot{\ell}'_{t\theta}(\widehat{\theta})]^{-1}$, while for the kernel estimators we compute them from the smaller matrix $[\sum_{t=1}^{T} \dot{\ell}_{t\phi}\dot{\ell}'_{t\phi}(\widehat{\phi},\widehat{\mu})]^{-1}$. The kernel standard errors asymptotically understate the true uncertainty associated with the parameter estimates, since they neglect the loss of efficiency associated with the non-parametric estimation of $\mu(\cdot)$.

The second method of obtaining standard errors is through the bootstrap. There are now many methods for time series models including some that make very weak assumptions regarding the dependence structure, like block bootstrap and sieve bootstrap. In practice, however, their performance depends a lot on the implementation and the model structure. We instead prefer a bootstrap procedure that uses some of our model structure. We give an algorithm for calculating such confidence intervals for $p = q = 1$ in the case of the kernel procedure. We use the 'wild bootstrap' [see Härdle (1990, p 247)] because we do not wish to rule out higher order conditional heterogeneity, as this is relevant for the sampling variability of our estimators.

Wild Bootstrap Algorithm

1. Given estimates $\underline{\mu}$, $\hat{\omega}$, $h_t(\hat{\omega},\underline{\mu})$, and $\hat{\varepsilon}_t = \varepsilon_t(\hat{\omega},\underline{\mu})$, calculate the recentered residuals $\hat{\varepsilon}_t^c = (\hat{\varepsilon}_t - T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t)$.

2. Let z_t be a random variable with $E(z_t^j) = 0$ for $j = 1, 3$ and $E(z_t^j) = 1$ for $j = 2$. Draw a random sample $\{z_1, \ldots, z_T\}$ from this distribution and let $\varepsilon_t^* = \hat{\varepsilon}_t^c \cdot z_t$. The variable ε_t^* will satisfy $E(\varepsilon_t^*) = 0$, $E(\varepsilon_t^{*2}) = \widehat{\varepsilon}_t^{c2}$, $E(\varepsilon_t^{*3}) = 0$, and $E(\varepsilon_t^{*4}) = \widehat{\varepsilon}_t^{c4}$.

3. Given starting values h_0^* and ε_0^*, and $\{h_t\}_{t=1}^T$, define recursively

$$h_t^* = \hat{a} + \hat{b}h_{t-1}^* + \widehat{c}_1\left[|\varepsilon_{t-1}^*| - E|\varepsilon_{t-1}^*|\right] + \widehat{d}_1\varepsilon_{t-1}^*.$$

and

$$y_t^* = \mu(h_t^*; \{h_s\}_{s=1}^T) + \varepsilon_t^*\sigma_t^*,$$

with the appropriate choice of $\mu(\cdot)$. In the case of the kernel estimator, some auxiliary bandwidth parameter $\tilde{\delta}$ that oversmooths the data should be chosen, where

$$\mu(x; \{h_s\}_{s=1}^T, \delta) = \frac{\sum_{s\neq t} K\left(\frac{x-h_s}{\delta}\right) y_s}{\sum_{s\neq t} K\left(\frac{x-h_s}{\delta}\right)}.$$

4. Given $\{y_t^*\}_{t=1}^T$ calculate parameter estimates $\hat{\phi}^*$ using the above quasi-Newton procedure.

5. Repeat steps 2–4 m times. The standard errors are estimated from the sample standard deviation of the bootstrap parameter estimates $\hat{\phi}^*$. ■

This method of obtaining standard errors is time-consuming for large datasets since it relies on simulation. However, it should reflect fully the loss of precision associated with estimating $\mu(\cdot)$. We impose a condition of symmetry on the errors for simplicity. This allows for easy construction of the random variable z_t as a variable with discrete support points at -1 and 1 with equal probability. We also did not impose the restriction $E(\varepsilon_t^{*2}) = 1$ because this would have forced $E(z_t^2) = 1/\hat{\varepsilon}_t^{c2}$, which is numerically unstable and generates paths with very large outliers.

The second problem, the construction of confidence intervals for $\widehat{\mu}$ can be approached in two ways: we can think of standard errors that are conditional on a value of h_t [and therefore allows us to look at the issue of the shape of the risk premium], and those that are conditional on all observables and thus allow us to run real-time experiments, and would be of interest to a decision maker. The second type is more difficult to construct as h_t depends on the infinite past, hence these standard errors have to be built up recursively.

On the other hand, computing standard errors conditional on the value of h_t is rather simple. For the kernel method, the variance of $\widehat{\mu}_t$ is given by Härdle (1990):

$$\frac{1}{n\delta}\frac{\sigma_t^2 \int k\left(u\right)^2 du}{f\left(h_t\right)},$$

where $f\left(h_t\right)$ is the ergodic density of h_t evaluated at h_t. This quantity can be estimated by replacing σ_t^2 and $f(h_t)$ by estimates $\widehat{\sigma}_t^2$ and $\widehat{f}(\widehat{h}_t)$ respectively.

For the series approximation, we define $\widehat{\tau}$ as the estimated mean parameters and H_t be the vector of slopes, i.e., $\partial\mu/\partial\tau|_{\widehat{\tau}}$. For instance, for the Fourier series

$$H_t = (1, h_t^*, h_t^{*2}, \cos(h_t^*), -\sin(h_t^*), \ldots, M\cos(h_t^*), -M\sin(h_t^*))'. \tag{13}$$

Then,

$$\text{var}\left[\mu\left(h_t\right)|h_t\right] = H_t' \text{var}\left(\widehat{\tau}\right) H_t,$$

where $\text{var}(\widehat{\tau})$ is the appropriate submatrix of the covariance matrix of $\widehat{\theta}$ obtained by the bootstrap method as described above.

Finally, choice of bandwidth is a nontrivial problem here. It is necessary to undersmooth our estimate of $\mu(\cdot)$ to obtain good estimates of ϕ as has been pointed out by Robinson (1988) for example. We adopt a cross-validation approach in which we estimate the risk premium for a grid of δ and choose the value that maximizes the (leave-one-out) likelihood function for the current values

of ϕ. However, to obtain a reasonable choice of bandwidth, it was necessary to remove the outliers when doing this and we removed 5% at each end of the data.

5 NUMERICAL RESULTS

5.1 EMPIRICAL RESULTS

5.1.1 Data

We examine the monthly risk premium on the excess returns on the CRSP value-weighted index — the total monthly return on the index minus the monthly returns on T-bills— over the period January 1926 to December 1997. The data is obtained from the Center for Research on Security Prices (CRSP), which includes NYSE and AMEX and is perhaps the best readily available proxy for 'the market'.[1] The data are plotted in figure 1. In Table 1 below we report sample moments for the raw data over the whole sample and two subsamples, each containing half of the data: I (1926–1961) and II (1962–1997).

*** Table 1 Here ***

There is strong evidence of leptokurtosis and weaker evidence of skewness in the full sample and in the sub-sample. The table reveals some differences in moments across subsamples. In particular, the first sub-period has much higher variance, positive skewness, and fatter tails than the rest of the sample. The standard deviation is approximately ten times the size of the mean, and this appears to support the widely held view that it is fundamentally difficult to estimate any mean effect in the presence of such large volatility [making the association that global mean corresponds to signal and global standard deviation corresponds to noise+signal]. However, from the nonparametric point of view this evidence is not by itself convincing since the global moments are one end of the smoothing spectrum where bandwidth is infinite; the other end of the smoothing spectrum is where bandwidth is zero and corresponds to the point mean being equal to the observation itself and the point standard deviation being the same quantity. To illustrate this point we computed a running mean and running standard deviation with 7 observations and equal weighting. The results are in figure 2 and show the time-varying nature of the mean and volatility. At this frequency, the mean and standard deviation are much closer in magnitude.[2] Note also that this approach to estimating volatility provides similar estimates to those obtained from the dynamic models that we propose. Compare figures 2 with

[1] We also conducted an analysis on the S&P500 series and obtained similar results.

[2] Of course, since this method is using future information we are not addressing the fundamental issue of predictability here.

figure 4 - the shapes are quite similar. Estimated volatility is high around well-known events: the Depression years, World War II, the oil shock and the 1987 crash in both cases.

5.1.2 Estimation

We first discuss some model selection choices that had to be made. For the series estimator, values of the tuning parameters of up to 3 were considered with the models selected by the Akaike criterion (AIC) which maximizes $2\ln L(\omega) - 2k$ where k is the number of parameters in the model and the Bayesian criterion (BIC) which maximizes $2\ln L(\omega) - k\ln T$. Both criteria gave similar results: in both cases, $p = 1$ and $q = 2$ are selected, but the AIC chooses $M = 2$, while BIC chooses $M = 1$. We report the results for $M = 1$.

Because both approaches selected the same values of p and q, we chose these values when estimating the model using the kernel approach. Results for other choices of p and q are available from the authors upon request. It is difficult to compare the fit of the model estimated with the kernel for various values of p and q as the models are then non-nested. As explained above, the bandwidth was selected by cross-validation at each iteration within the algorithm, However, it was necessary to delete outliers when choosing the bandwidth, otherwise the bandwidth exploded to infinity. The bandwidth has the form:

$$\delta = k\sigma(h_t)\, T^{-\frac{1}{5}},$$

where $\sigma(h_t)$ is the standard deviation of h_t, updated at each iteration to reflect the new estimates of h_t with k allowed to vary between 0.5 and 2.5 in increments of 0.1. We set the values of $\underline{h}$ and $\overline{h}$ at -10 and -2 respectively based on the results from the kernel estimation which does not impose such restrictions. We also check to ensure that there is no value of h_t outside of these values in the course of optimization.

We now turn to the estimation results. The results from the estimation using the two methods considered here and their associated standard errors ($se_\phi(\phi)$) are presented in table 2.

*** Table 2 Here ***

Our parameter estimates appear quite robust to the method chosen to do the estimation. They are also consistent with many other studies in the area such as Nelson (1991), Glosten, Jagannathan, and Runkle (1993) or Bollerslev, Engle, and Nelson (1994). In particular, the estimate of b_1, which measures the degree of persistence, is high (above 0.9) and the estimate of the leverage effect d_1 is strongly negative. There is also agreement over the leverage effect of ε_{t-2}, but this parameter is positive. Finally, the estimated value of ν is around 1.6 which is again consistent with previous findings. The distribution we find has fatter tails than the normal which is a special case with $\nu = 2$.

Note that the bootstrap standard errors tend to be larger by up to 50% than the analytic standard errors.

The last row of table 2 provides results of a likelihood ratio test for the significance of the coefficients on the nonlinear terms in the Fourier series. The results clearly show that linearity is strongly rejected, even at a level of significance of 1%. The individual parameters are all significant.

The risk premium estimated using the kernel method is graphed in figure 3 as a function of h_t. Confidence intervals at the 95% level constructed using the pointwise kernel confidence intervals are also provided. The figure clearly reveals a non-monotonic relation between h_t and y_t. This is consistent with the findings of Backus and Gregory (1993) that in an artificial economy, the risk premium may have virtually any shape. Although the estimated risk premium is not significantly different from a constant at this level for some part of its range, the evidence is stronger in the middle range $h_t \in [-7.5, -5.5]$, which is where most of the data lie [see figure 4 for the plot of the marginal density]. The evolution of the estimated risk premium and conditional variance are presented in figure 4. The episodes of high volatility revealed by this figure coincide closely with those obtained by a simple running average as done in figure 2.

Figure 5 provides the shape of the risk premium estimated using the Fourier series. The graph also includes the analytical 95% confidence intervals conditional on h_t. Again, the estimated shape is nonlinear.

The two smoothing methods both have advantages and disadvantages. The kernel estimate appears rather wiggly in the end points where there is not much data. The Fourier series method on the other hand is very smooth and gives the appearance of being precisely estimated. However, there is a pronounced upward slope at the high end, which seems at odds with the kernel method finding. This end-trend is quite symptomatic of these polynomial-based methods; we view it with some skepticism. Notice also the difference in the standard errors for the two methods. The Fourier series method has a confidence band whose width is almost the same throughout the shown range, while the confidence band for the kernel is very wide at the end points, which reflects the relative paucity of the data in this region [see Figure 7 which shows a kernel density estimate of h_t]. Thus the Fourier series confidence band gives the appearance of being very precisely estimated in a region where we really don't have any data. This is because it is a global fitting method that draws its estimates from all the data. We thus redraw the two estimates on the same plot in figure 6. The methods agree quite closely on this subrange - there is a hump shape, which is first concave and then convex.

Finally, we provide some diagnostics on the standardized residuals $\hat{\varepsilon}_t = (y_t - \hat{\mu}_t)/\widehat{\sigma}_t$.We just report the results for the kernel, but similar results have been obtained for the series approach. The plots of the autocorrelogram of both the residuals and their squares indicates that they are close to white

noise: there are 4 significant autocorrelation coefficients at the 5% level among the first 100 lags in the levels and 9 significant autocorrelations in the squares.

As mentioned above, we also investigated using the normal distribution for the innovation density. The results were very similar, and the shape of the risk premium was almost identical to figure 3. We therefore decided to not report these results, but they are available from the authors upon request.

5.1.3 Subsample Estimation

In order to see how robust our estimates are, we re-estimated the model over two sub-samples: 1926-1961 and 1962-1997 using the kernel method. The results are presented in table 3 below (with analytical standard errors in parentheses).

*** Table 3 Here ***

The results show quite a bit of instability in the point estimates. Nevertheless, the shape of the risk premium is relatively stable over time. Figure 9 shows the estimated risk premium using the same scale as in the other figures. Because the last subsample is characterized by lower volatility than the beginning of the sample, the estimated log-volatility is concentrated towards the left of the graph for that period. The risk premium we estimate in the second period is much flatter than that of the first period, though the point estimate suggests a similar non-monotonic shape as for the full sample and the first subsample.

5.2 MONTE CARLO

In order to appreciate the performance of our kernel procedure in estimating the risk premium in financial data, we carried out four simulation experiments. Each experiment is repeated 500 times on samples of size 500. To make the experiments as realistic as possible, the parameters if each experiment are set to values estimated from our dataset.

The first simulation experiment involves generating a risk premium from a linear model. We thus estimated an EGARCH-M model with GED errors from the data and used it to generate 500 samples. We then applied our non-parametric procedure to these simulated samples. The results are presented in the upper left panel of figure 10. The solid line represents the true risk premium which is linear. The line with the long dashes is the mean estimated function at each point on our equispaced grid. The short dashes represent the limits of a 95% confidence band. The method appears to do quite well as the mean estimate deviates from the true function marginally for all values of the conditional variance. The confidence interval is relatively narrow, although it widens dramatically for large volatility as there is less data

The second experiment used the model estimated by the Fourier series and GED errors to generate the data presented in the previous section. The results for this experiment are in the upper right panel of figure 10. The kernel procedure unveils the nonlinear mean function well except for small conditional variance.

The third experiment is a GARCH-M model with normal errors and linear mean. This experiment is designed to check the robustness of our results to misspecification in the conditional variance process and the innovation density (the parametric components of our model). The results are in the lower bottom panel of the figure. The kernel procedure discovers the linear mean very well. However, the confidence bands are very wide reflecting the additional uncertainty caused by misspecification.

Finally, the last experiment consists of a GARCH model with normal errors and mean function estimated with Fourier series. Once again, the mean function is well estimated where most data lies, but the confidence bands are once again very large due to misspecification.

Overall, these results suggest that our kernel procedure performs well in uncovering possible nonlinearities in the data. Yet, if the model were truly linear, the procedure would not mislead us. It is thus a useful tool for looking at the shape of the risk premium.

6 CONCLUSIONS

We have found a highly nonlinear relationship between the first two moments of index returns as suggested by Backus and Gregory (1993) and Gennotte and Marsh (1993). In particular, the risk premium appears to be non-monotonic and indeed camel-humped. This result appears to be quite robust to the estimation method and the tuning parameters selected. However, the estimated risk premiums are subject to quite a bit of variability and are not uniformly significantly different from a constant at the 95% level. This must temper our interpretations somewhat, although restricting attention to the range of volatility which occurs most often we get much stronger significance.

References

[1] Abel, A. B. (1987), "Stock Prices under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model," Mrxuqdo ri Prqhwdu| Hfrqrplfv, 22, 375-393.

[2] Abel, A. B. (1998), "Risk Premia and Term Premia in General Equilibrium", NBER Working Paper 6883.

[17] Engle, R. F., and Gonzalez–Rivera, G. (1991), "Semiparametric ARCH Models," Mrxuqdo ri Exvlqhvv dqg Hfrqrplf Vwdwlvwlfv, 9, 345-359.

[18] French, K. R., Schwert, G. W., and Stambaugh, R. B. (1987), "Expected Stock Returns and Volatility," Mrxuqdo ri Ilqdqfldo Hfrqrplfv, 19, 3-29.

[19] Gallant, A.R. (1981), "On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form", Mrxuqdo ri Hfrqrphwulfv/ 15, 211-245.

[20] Gennotte, G., and Marsh, T. (1993), "Valuations in Economic Uncertainty and Risk Premiums on Capital Assets," Hxurshdq Hfrqrplf Uhylhz, 37, 1021-1041.

[21] Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993), "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", Mrxuqdo ri Ilqdqfh, 48, 1779-1801.

[22] Hamilton, James D. (1994), Wlph Vhulhv Dqdo|vlv, Princeton University Press.

[23] Härdle, W. (1990), Dssolhg Qrqsdudphwulf Uhjuhvvlrq1 Cambridge University Press.

[24] Härdle, W. and Linton, O. (1994), "Applied Nonparametric Methods" in Engle, R. F. and D. L. McFadden, Kdqgerrn ri Hfrqrphwulfv/ yroxph LY, Elsevier Science, 2295-2339.

[25] Hastie, T. and R. Tibshirani (1990). Jhqhudol}hg Dgglwlyh Prghov. Chapman and Hall, London.

[26] Higgins, M.L., and Bera, A. K. (1992), "A Class of Nonlinear ARCH Models", Lqwhuqdwlrqdo Hfrqrplf Uhylhz 33, 137-158.

[27] Klein, R.W., and Spady, R. H. (1994): "An Efficient Semiparametric Estimator for Binary Choice Models." Hfrqrphwulfd 61, 387-421.

[28] Lintner, J. (1965), "The Valuation of Risky Assets and the Selection of Risky Investment in Stock Portfolios and Capital Budgets," Uhylhz ri Hfrqrplfv dqg Vwdwlvwlfv, 47, 13-37.

[29] Lee, S. and Hansen, B. (1994), "Asymptotic Theory for the GARCH(1,1) Quasi-Maximum Likelihood Estimator", Hfrqrphwulf Wkhru|, 10, 29-52.

[30] Lumsdaine, R. L. (1996), "Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models," Hfrqr0 phwulfd, 64, 575-596.

[31] Mammen, E, O. Linton, and J.P. Nielsen (1999): "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions." Wkh Dqqdov ri Vwdwlvwlfv 27, 1443-1490.

[32] Merton, R. C. (1973), "An Intertemporal Capital Asset Pricing Model," Hfrqrphwulfd, 41, 867-887.

[33] Nelson, D. B. (1990), "Stationarity and Persistence in the GARCH(1,1) Model," Hfrqrphwulf Wkhru|, 6, 318-334.

[34] Nelson, D. B. (1991), "Conditional Heteroscedasticity in Asset Returns: A New Approach," Hfrqrphwulfd, Vol. 59, 347-370.

[35] Nelson, D. B. and Cao, C. Q. (1991), "Inequality Constraints in the Univariate GARCH Model", Mrxuqdo ri Exvlqhvv dqg Hfrqrplf Vwdwlvwlfv, 10, 229-235.

[36] Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. Dqq1 Vwdwlvw. **25**, 186 - 211.

[37] Steigerwald, D. G. and W. K. Newey (1997), "Asymptotic Bias for Quasi-Maximum Likelihood Estimators in Conditional Heteroskedasticity Models", Hfrqrphwulfd, 65, ????

[38] Pagan, A. R., and Hong, Y. S. (1990), "Non-parametric Estimation and the Risk Premium." In W. A. Barnett, J. Powell, and G. Tauchen (eds.), Qrqsdudphwulf dqg Vhplsdudphwulf Phwkrgv lq Hfrqrphwulfv dqg Vwdwlvwlfv= Surfhhglqjv ri wkh Iliwk Lqwhuqdwlrqdo V|psrvlxp lq Hfrqrplf Wkhru| dqg Hfrqrphwulfv, Cambridge: Cambridge University Press, 51-75.

[39] Perron, B. (1999), "Semi-parametric Weak Instrument Regressions with an Application to the Risk-Return Trade-off", CRDE working paper 0199, Université de Montréal.

[40] Perron, B. (1998), "A Monte Carlo Comparison of Non-parametric Estimators of the Conditional Variance", Mimeo.

[41] Powell, J. (1994), "Estimation of Semiparametric Models", in Engle, R. F. and D. L. McFadden, Kdqgerrn ri Hfrqrphwulfv/ yro xph LY, Elsevier Science, 2443-2521.

[42] Robinson, P. M. (1983), "Nonparametric Estimators for Time Series," Mrxuqdo ri Wlph Vhulhv Dqdo|vlv/ 4, 185-207.

[43] Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," Hfrqrphwulfd, 56, 931-954.

[44] Schwert, G. W. (1989), "Why Does Stock Market Volatility Change Over Time", Mrxuqdo ri Ilqdqfh, 44, 1115-1153.

[45] Sharpe, W. (1964), "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," Mrxuqdo ri Ilqdqfh/ 19, 567-575.

[46] Tadikamalla, P. R. (1980), "Random Sampling from the Exponential Power Distribution", Mrxu0 qdo ri wkh Dphulfdq Vwdwlvwlfdo Dvvrfldwlrq, 75, 683-686.

[47] Weiss, A. (1986). Asymptotic Theory for ARCH models: Estimation and Testing. Hfrqrphwulf Wkhru|/ 2, 107-131.

[48] Whistler, D. (1988), "Semiparametric ARCH Estimation of Intra-Daily Exchange Rate Volatility," unpublished manuscript, London School of Economics.

[49] Wooldridge, J.M. (1994): "Estimation and Inference for Dependent Processes," in Engle, R. F. and D. L. McFadden, Kdqgerrn ri Hfrqrphwulfv/ yro xph LY, Elsevier Science, 2659-2738.

Tables and Figures

Table 1. Raw Data by Sub Period

	Full sample	I	II
Mean (×100)	0.3276	0.4674	0.1877
Variance (×100)	0.3034	0.4174	0.1897
Skewness	4.264e-5	1.1037e-4	-3.455e-5
Excess Kurtosis	8.1385	7.6340	2.3158

Table 2. Full sample estimates

	Kernel	Fourier	EGARCH-M
a	-0.151 (0.054) (0.097)	-0.374 (0.075) (0.140)	-0.155 (0.105)
b_1	0.976 (0.009) (0.015)	0.940 (0.012) (0.022)	0.975 (0.017)
c_1	-0.029 (0.083) (0.092)	-0.140 (0.062) (0.089)	0.012 (0.700)
c_2	0.278 (0.090) (0.094)	0.401 (0.072) (0.091)	0.235 (0.754)
d_1	-0.308 (0.055) (0.062)	-0.361 (0.042) (0.067)	-0.298 (0.096)
d_2	0.247 (0.047) (0.062)	0.261 (0.042) (0.065)	0.247 (0.142)
ν	1.547 (0.093) (0.134)	1.657 (0.099) (0.160)	1.561 (0.164)
γ_0	-	0.236 (0.236) (1.139)	0.014 (0.081)
γ_1	-	-0.378 (0.205) (1.021)	0.001 (0.012)
γ_2	-	0.080 (0.034) (0.175)	-
γ_3	-	0.146 (0.029) (0.125)	-
γ_4	-	-0.167 (0.081) (0.361)	-
Bandwidth constant	1.30	-	-
$\ell =$	1454.76	1466.16	1448.02
Linearity test $H_0 : \gamma_i = 0, i > 1$ $(p-value)$	-	36.28 (0.000)	-

Note: The numbers in parentheses are analytical and bootstrap standard errors respectively.

Table 3. Sub-period estimates

	1926-1961	1962-1997
a	-0.098 (0.061)	-0.476 (0.282)
b_1	0.984 (0.010)	0.926 (0.044)
c_1	0.044 (0.114)	-0.133 (0.135)
c_2	0.180 (0.124)	0.378 (0.147)
d_1	-0.211 (0.081)	-0.449 (0.078)
d_2	0.148 (0.065)	0.375 (0.082)
ν	1.543 (0.147)	1.762 (0.183)
Bandwidth constant	1.50	2.50
$\ell =$	684.71	775.47