CAHIER 0899

THE SHAPE OF THE RISK PREMIUM : EVIDENCE FROM A SEMIPARAMETRIC GARCH MODEL

Oliver LINTON and 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 0899

THE SHAPE OF THE RISK PREMIUM : EVIDENCE FROM A SEMIPARAMETRIC GARCH MODEL

Oliver LINTON[1] and Benoit PERRON[2]

[1] Cowles Foundation for Research in Economics, Yale University

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

October 1999

The authors are grateful to Adrian Pagan for comments. Research was supported by the National Science Foundation.

RÉSUMÉ

Nous étudions la relation entre la prime de risque sur l'indice S&P 500 et sa variance conditionnelle. Nous utilisons le modèle SMEGARCH - Semiparametric-Mean EGARCH - selon lequel la variance conditionnelle suit un processus EGARCH, alors que la moyenne est une fonction arbitraire de la variance conditionnelle. Pour les rendements excédentaires mensuels sur l'indice S&P 500, la relation que nous trouvons est non linéaire et non monotone. De plus, nous trouvons beaucoup de persistance dans la variance conditionnelle ainsi qu'un effet de levier, tel que documenté par plusieurs autres auteurs.

Mots clés : modèles ARCH, évaluation d'actifs, séries Fourier, noyau, prime de risque

ABSTRACT

We examine the relationship between the risk premium on the S&P 500 index return and its conditional variance. We use the SMEGARCH - Semiparametric-Mean EGARCH - model in which the conditional variance process is EGARCH 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. Moreover, the shape of these relationships seems to be relatively stable over time.

Key words : ARCH models, asset pricing, backfitting, Fourier series, kernel, risk premium

1 Introduction

Modern asset pricing theories such as Abel (1987, 1998), Cox, Ingersoll, and Ross (1985), Merton (1973), and Gennotte and Marsh (1988) 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

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

where r_{mt}, r_{ft} are the returns on the market portfolio and riskless 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.[1] 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 (1989) 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)

[1] The two well-established empirical regularities about financial and macroeconomic time series — serially dependent conditional second moments and unconditional leptokurtosis — can potentially be explained by this class of models – see the original papers by Engle (1982) and the survey papers of Bollerslev, Chou, and Kroner (1992) and Bollerslev, Engle, and Nelson (1994) for references.

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 $E[r_{mt} - r_{ft}|I_{t-1}]$ and $\text{var}[r_{mt} - r_{ft}|I_{t-1}]$ 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 OLS and IV methods, 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 this approach. Firstly, the conditional moments are calculated using a finite conditioning set. 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 will provide a poor approximation as reported by Perron (1998) using simulation. 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 S&P500 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 method 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 of estimating additive nonparametric regression. 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 of unknown functional form, but smooth. We shall also suppose that ε_t is *i.i.d.*, although this is not strictly necessary for some purposes. 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 (1988). 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 model the conditional variance using the Exponential GARCH model considered by Nelson (1991):

$$h_t \equiv \log(\sigma_t^2) = a + \sum_{j=1}^{p} b_j \log(\sigma_{t-j}^2) + \sum_{k=1}^{q} c_k' v_{t-k}, \tag{5}$$

where $c_k = (c_{1k}, c_{2k})'$ and $v_t = (v_{1t}, v_{2t})'$ is a vector of mean zero *i.i.d.* innovations:

$$\begin{aligned} v_{1t} &= \varepsilon_t; \\ v_{2t} &= |\varepsilon_t| - E\,|\varepsilon_t|, \end{aligned} \tag{6}$$

which are identical to those in Nelson (1991).

Specifying a time series model for the log of the conditional variance (EGARCH) has a number of advantages over specifying the model for the level of the conditional variance (GARCH). In GARCH models, it is necessary to impose inequality constraints on the parameters during estimation to ensure that the variance process remains non-negative. The constraints typically employed exclude cyclical behavior for the variance.[2] If an intercept is included, a lower bound is imposed on the conditional variance process. This lower bound is estimated from the entire sample and may distort some of the properties of the fitted model. In particular, the standardized residuals may appear to have too much mass near zero — see for example Whistler (1988).

GARCH models essentially specify the behavior of the square of the data. In this case a few large observations, such as those between 1929–1933, can dominate the sample; the estimated parameters are

[2] Nelson and Cao (1992) establish a much weaker set of inequality constraints than Bollerslev (1986).

determined primarily by these massive observations. In EGARCH models, one is essentially modelling the log of the square of the data, and hence large observations get substantially down-weighted.[3] A more general class of specifications is considered in Higgins and Bera (1992).

The usual choice of innovations in the GARCH model is $\varepsilon_{t-1}^2\sigma_{t-1}^2$, which are not *i.i.d.*; this results in a number of conceptual anomalies. Nelson (1990) shows that the conditional moments may explode even when the process is strictly stationary and ergodic — i.e., the usual unit root condition does not delineate stationarity from nonstationarity; see also Lumsdaine (1996) and Lee and Hansen (1994) on this point. In addition, the driftless IGARCH(1,1) process collapses to zero almost surely. With *i.i.d.* innovations as in Nelson (1991) these conundrums do not occur, i.e., strong stationarity and weak stationarity coincide.

This formulation 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 the so-called leverage effect. The parameters c_{1k} and c_{2k} control 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 for this hypothesis 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[4]

$$f(\varepsilon) = \frac{\nu \exp\left(-\frac{1}{2}|\varepsilon/\lambda|^{\nu}\right)}{\lambda 2^{(1+1/\nu)}\Gamma(1/\nu)}; \; \lambda = [2^{(-2/\nu)}\Gamma(1/\nu)/\Gamma(3/\nu)]^{1/2}, \tag{7}$$

where Γ is the gamma function. With this density, we obtain that $E\,|\varepsilon_t| = \frac{\lambda 2^{1/\nu}\Gamma(2/\nu)}{\Gamma(1/\nu)}$ (Hamilton, 1994, p. 669).

Under the full specification (4-7), $E[\exp(dv_t)] < \infty$ for any d provided $\nu > 1$ and $|\mu(x)| \leq c|x|$ for some c and all x. Therefore, the unconditional variance of y_t exists provided the autoregressive polynomial $b(L) = 1 + b_1L + \cdots + b_pL^p$ has roots outside the unit circle. In this case, σ_t^2 and y_t are both weakly and strongly stationary.

[3] This is essentially the motivation Schwert (1989) gives for modeling the standard deviation. However, in EGARCH there is the converse problem that observations close to zero receive considerable weight in the estimation.

[4] 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 $\nu > 1$.

The main difference between this model and previous treatments is that we do not restrict the functional form of $\mu(\cdot)$ *a priori*. 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 and inference

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 $\phi = (a, b_1, \ldots, b_p, c_1', \ldots, c_q', \nu)' \equiv (\theta', \nu)'$ and τ be the vector of unknown mean parameters. Then $\varepsilon_t(\phi, \tau)$ and $h_t(\phi, \tau)$ can be built up recursively given initial conditions, and the conditional log-likelihood function is

$$\ell_T(\phi, \tau) = \sum_{t=1}^{T} \ell_t(\phi, \tau) = \sum_{t=1}^{T} \log(f(\varepsilon_t(\phi, \tau); \nu) - \frac{1}{2} \sum_{t=1}^{T} h_t(\phi, \tau), \tag{8}$$

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

$$\begin{pmatrix} \phi^{[i+1]} \\ \tau^{[i+1]} \end{pmatrix} = \begin{pmatrix} \phi^{[i]} \\ \tau^{[i]} \end{pmatrix} - \lambda^{[i]} \left[\sum_{t=1}^{T} \dot{\ell}_{t\phi} \dot{\ell}_{t\phi}' \right]^{-1} \sum_{t=1}^{T} \dot{\ell}_{t\phi}, \tag{9}$$

where $\lambda^{[i]}$ is a variable step length chosen to maximize the log likelihood function in the given direction, and the score functions $\dot{\ell}_{t\phi}$ are evaluated at $\phi^{[i]}, \tau^{[i]}$.

We propose constructing estimates of ϕ and $\mu(\cdot)$ in the semiparametric model by analogous methods. 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. The basis we will use is the Fourier Flexible Form of Gallant (1981). We describe the estimation and the construction of confidence intervals for each method in turn.

3.1 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]. In many other semiparametric problems one can use a profile likelihood method as in Klein and Spady (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 includes the profiled nonparametric estimator. For example, consider the following i.i.d. problem, which is similar to ours. We observe $\{y_i, x_i\}_{i=1}^n$, where

$$y_i = \mu(\sigma_i^2) + \varepsilon_i \sigma_i \tag{10}$$

$$\sigma_i^2 = g(\theta, x_i), \quad i = 1, \ldots, n,$$

where $g(\theta, \cdot)$ is a known function and ε_i are i.i.d. with mean zero and variance one. For each θ, we can write

$$\widehat{\mu}_{\theta i}(s) = \frac{\sum_{j=1, j\neq i}^n K\left(\frac{s-g(\theta,x_j)}{\delta}\right) y_j}{\sum_{j=1, j\neq i}^n K\left(\frac{s-g(\theta,x_j)}{\delta}\right)}, \tag{11}$$

where $\delta(T)$ is a bandwidth parameter such that $\delta(T) \to 0$ as $T \to \infty$, while K is a bounded kernel satisfying $\int K(u)du = 1$, for example the normal density, and define the semiparametric profile likelihood

$$\widehat{\ell}(\theta) = -\frac{1}{2}\sum_{i=1}^n \log g(\theta, x_i) - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - \widehat{\mu}_{\theta i}(g(\theta, x_i)))^2}{g(\theta, x_i)}.$$

We can take the minimizer of $\widehat{\ell}(\theta)$, $\widehat{\theta}$, as an estimate of θ. This problem has not been explicitly treated before in the literature, but is similar to many others. We expect that $\widehat{\theta}$ is 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 the case of our time series model, we can't define the corresponding 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 an entire vector of $\mu's$ to construct $\widehat{\mu}_\phi(\sigma_t^2)$.[5] We are instead led to iterative updating of both the finite dimensional parameters and the function μ, i.e., our profiled likelihood is the limit of a sequence of operations. Our

[5] An alternative method that works for (10) is to compute the residuals directly $u_i = y_i - \widehat{E}(y_i|x_i)$, where $\widehat{E}(y_i|x_i)$ is some nonparametric estimate of $E(y_i|x_i)$, and then to estimate the conditional variance by maximum likelihood or least squares based on these residuals. Unfortunately, in the time series model, the relevant information set here is the entire infinite past, so it is hard to make use of the conditional moments $E[y_t|\mathcal{F}_{t-1}]$ directly. 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

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 $\phi \equiv (\theta', \nu)'$. We then update our estimates of $\underline{\mu}$ using kernel estimates based on the previous iterations filtered log variances. Our method has some similarity to the 'backfitting' methodology for estimating additive nonparametric regression, see Hastie and Tibshirani (1990), in that it amounts to iterative one-dimensional smoothing operations. In our case, we combine parametric and nonparametric estimation in each iteration.

For convenience we describe the algorithm for the case $p = 1$ and $q = 1$. We shall also 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.

3.1.1 Simultaneous Estimation of ϕ and $\mu(\cdot)$

We use the following 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

$$\mu_t^{[r]} = \frac{\sum_{s \neq t} K\left(\frac{h_t^{[r-1]} - h_s^{[r-1]}}{\delta}\right) y_s}{\sum_{s \neq t} K\left(\frac{h_t^{[r-1]} - h_s^{[r-1]}}{\delta}\right)}, \ t = 1, 2, \ldots, T. \tag{12}$$

   where $\delta(T)$ is a bandwidth parameter such that $\delta(T) \to 0$ as $T \to \infty$, 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

$$h_t^{[r]} = a + bh_{t-1}^{[r]} + c_1 \varepsilon_{t-1}^{[r]} + c_2 \left(\left| \varepsilon_{t-1}^{[r]} \right| - E \left| \varepsilon_{t-1}^{[r]} \right| \right),$$

$$\varepsilon_t^{[r]} = \frac{y_t - \mu_t^{[r]}}{\exp(h_t^{[r]})}, \ t = 1, 2, \ldots, T,$$

   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 $\underline{\mu}^{[r]} = (\mu_1^{[r]}, \ldots, \mu_T^{[r]})'$.

---

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.

4. Calculate

$$\phi^{[r+1]} = \phi^{[r]} - \lambda^{[r]} \left[\sum_{t=1}^{T} \dot{\ell}_{t\phi}^{[r]} \dot{\ell}_{t\phi}^{\prime[r]} \right]^{-1} \sum_{t=1}^{T} \dot{\ell}_{t\phi}^{[r]}, \tag{13}$$

where $\dot{\ell}_{t\phi}^{[r]}$ is the vector of partial derivatives of $\ell_t^{[r]}(\phi)$ with respect to ϕ evaluated at $\phi^{[r]}$.

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

$$\max \left\{ \max_k \left| \frac{\sum_{t=1}^{T} \dot{\ell}_{t\phi_k} \cdot \phi_k}{\ell(\phi)} \right|, \frac{1}{T} \sum_{t=1}^{T} \left| \frac{\underline{\mu}^{[r+1]} - \underline{\mu}^{[r]}}{\underline{\mu}^{[r]}} \right| \right\} < \varepsilon, \tag{14}$$

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

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.[6] An alternative implementation is to iterate to convergence on the computation of ϕ in (4) for each $\mu_t^{[r]}$, and then to update $\mu_t^{[r]}$ as in step (2) above. The stopping rule (14) was arrived at after some experimentation. It is desirable to ensure that the entire parameter vector (ϕ, μ) is convergent. In practice, this methods seems to work quite well.

No asymptotic theory is available for $\widehat{\phi}$; however, for comparison, no theory is available for maximum likelihood estimation of parametric EGARCH models without mean effects — see Nelson (1991). The only cases where a solid asymptotic theory that relies on primitive conditions for members of the GARCH class of models exists are ARCH models, see Weiss (1986), and the GARCH(1,1) model, see Lumsdaine (1996) and Lee and Hansen (1994). For any other specification in the GARCH class, the asymptotic theory that is used in practice is not known to be valid.

If h_t were observed, a kernel estimate of $\mu(\cdot)$ as in (12) would be consistent and asymptotically normal under appropriate conditions;[7] 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. In practice, the parameters of h_t appear to be quite robust to different parametric specifications of the mean equation. The filtered estimate of

[6] Note that convergence of the backfitting algorithm has only been shown in special cases. However, even in the absence of convergence proofs, as in the application to generalized additive models, the method seems to work well.

[7] The process h_t is weakly dependent — it is an ARMA process. Therefore, the results of Robinson (1983) can be applied to establish consistency, provided $\delta(T) \Rightarrow 0$ at an appropriate rate.

h_t based on, for example, $\mu_t^{[0]} = T^{-1}\sum_{s=1}^{T} y_s$ should be close to the true h_t and should provide good starting values. As in the parametric case, additional iterations should improve the performance of the estimated parameters and function. 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.

Choice of bandwidth is a nontrivial problem here, since it may be 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 simple, second best, approach to this, see below.

3.2 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.

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

$$\mu(h_t^*) = \gamma_0 + \gamma_1 h_t^* + \gamma_2 h_t^{*2} + \sum_{j=1}^{M} \psi_j \sin(jh_t^*) + \sum_{j=1}^{M} \varphi_j \cos(jh_t^*). \tag{15}$$

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

3.3 Constructing Standard Errors

The remaining problem is the 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:

$$se_\phi = \mathrm{sqrt}\left(\mathrm{diag}\left(\left[\sum_{t=1}^{T} \dot{\ell}_{t\phi}\dot{\ell}_{t\phi}'\right]^{-1}\right)\right),$$

for both the kernel and series estimators. These standard errors should understate the true uncertainty associated with the parameter estimates since they neglect the loss of efficiency associated with the non-parametric estimation of $\mu(\cdot)$ in the case of the kernel and the sieve nature of the estimation of $\mu(\cdot)$ in the case of the series. However, these standard errors are quite easy to compute.

The alternative method for constructing standard errors is the bootstrap (see Härdle (1990)). We give an algorithm for calculating such confidence intervals for $p = q = 1$ in the case of the kernel.

3.3.1 Bootstrap Standard Errors

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 $\widehat{\hat{\varepsilon}}_t = \hat{\varepsilon}_t - \overline{\hat{\varepsilon}}$, where $\overline{\hat{\varepsilon}} = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t$.

2. Draw a random sample $\{\varepsilon_t^*\}_{t=1}^{T}$ from the empirical distribution of $\widehat{\hat{\varepsilon}}_t$.

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

$$h_t^* = \hat{a} + \hat{b} h_{t-1}^* + \hat{c}_{11} \varepsilon_{t-1}^* + \hat{c}_{12} \left[\left| \varepsilon_{t-1}^* \right| - E \left| \varepsilon_{t-1}^* \right| \right].$$

   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 oversmoothes the data should be chosen [see Härdle (1990) for a discussion of this], 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 $\mu(\cdot)$.

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):

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

where $f(h_t)$ 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 $\widehat{\gamma}$ as the estimated mean parameters and H_t be the vector of slopes, i.e., $\partial\mu/\partial\gamma|_{\widehat{\gamma}}$. For instance, for the Fourier series

$$H_t = \begin{pmatrix} 1 \\ h_t^* \\ h_t^{*2} \\ \cos(h_t^*) \\ -\sin(h_t^*) \\ \vdots \\ M\cos(h_t^*) \\ -M\sin(h_t^*) \end{pmatrix}. \tag{16}$$

Then,

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

where $\text{var}(\widehat{\gamma})$ is the appropriate submatrix of the covariance matrix of $\widehat{\phi}$ obtained by either the outer product of gradient or the bootstrap as described above.

4 Empirical Results

We examine the monthly risk premium on the excess returns on the S&P500 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); it is plotted in figure 1. In Table 1 below we report sample cumulants for the raw data (multiplied by 100) over various subperiods: FP (1926–1997), I (1926–1945), II (1946–1973), and III (1973–1997).

*** Table 1 Here ***

The table reveals large differences in cumulants, in particular between the first period and the rest of the sample. This sub-period has much higher variance, positive skewness, and extremely fat tails relative to the rest of the sample. The second subperiod appears very calm with lowest variance and lowest fourth-order cumulant.

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. To simplify computations, the bandwidth was selected according to Silverman's rule of thumb:

$$\delta = 1.06\sigma(h_t) T^{-\frac{1}{5}},$$

where $\sigma(h_t)$ is the standard deviation of h_t, and updated at each iteration to reflect the new estimates of h_t. 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. 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. 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 c_{11} is strongly negative. There is some disagreement over the leverage effect of ε_{t-2}, but this parameter is much less precisely estimated. 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 effects, with the exception of the intercept, are also significant.

The risk premium estimated using the kernel method is graphed in figure 2 as a function of h_t. Confidence intervals at the 95% level constructed using the pointwise kernel confidence intervals are also provided. The figures clearly reveal 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. The evolution of the estimated risk premium and conditional variance are presented in figures 3 and 4. These figures clearly reveal the episodes of high volatility: the Great Depression, World War II, the first oil shock, and the period around the crash of 1987.

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. We thus redraw the two estimates over the narrower range where most of the data lie 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 8 significant autocorrelation coefficients at the 5% level among the first 100 lags in the levels and 9 significant autocorrelations in the squares. Moreover, apart from a little skewness, the density of the standardized residuals is close to that of the generalized error distribution estimated by the data.

4.1 Subsample estimation

In order to see how robust our estimates are, we re-estimated the model over three sub-periods: 1926-1945, 1946-1973, and 1964-1997 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. Figures 11-13 show the estimated risk premium using the same scale as in figures 2 and 5. Because the last two subsamples were characterized by lower volatility than the beginning of the sample, the estimated log-volatility is concentrated towards the left of the graphs for those two periods. However, we clearly see the same overall non-monotonic shape of the risk premium for those values of the log variance as with the full sample. The first subsample stands out as having higher volatility, but the estimated risk premium has a similar shape.

5 Simulation

In order to appreciate the performance of our kernel procedure in estimating the risk premium in financial data, we generated data using the model we estimated. We thus generated data according to

$$h_t = -0.175 + 0.973 h_{t-1} - 0.234\varepsilon_{t-1} + 0.013\left[|\varepsilon_{t-1}| - E|\varepsilon_{t-1}|\right] + 0.193\varepsilon_{t-2} + 0.246\left[|\varepsilon_{t-1}| - E|\varepsilon_{t-1}|\right] \quad (17)$$

with ε_t drawn from a generalized error distribution with $\nu = 1.578$. The errors were generated using the algorithm of Tadikamalla (1980) based on rejection. For each value of h_t generated, we then associate to it the value of the estimated risk premium with a value of the log variance that is closest to h_t. The sample size was the same as ours, 864, and we repeated this experiment 100 times. The estimation was carried out in the same way as the empirical analysis above with starting values for the risk premium at the mean of the generated data and using the Gaussian kernel. The algorithm was started at the true value of the parameters. The results from the experiment are presented in table 4.

*** TABLE 4 HERE ***

The results are encouraging: despite the small number of replications, all the sample means are within one standard deviation of the true value with the exception of $\widehat{\nu}$ which appears to be quite imprecisely estimated. The estimated risk premium is underestimated and slightly more volatile than the true risk premium used to generate the data.

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 (1988). In particular, the risk premium is nonmonotonic. This result appears to be quite robust to the estimation method and the tuning parameters selected. Moreover, the shape of the risk-return relation seems to be robust over time. These results suggest that previous parametric GARCH–M models such as Engle, Lilien, and Robins (1987) are misleading about the underlying relationship between risk and return at the market level.

References

[1] Abel, A. B. (1987), "Stock Prices under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model," *Journal of Monetary Economics*, 22, 375-393.

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

[3] 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.

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

[5] Black, F. (1976), "Studies in Stock Price Volatility Changes," Proceedings of the 1976 Business and Economic Statistics Section, American Statistical Association.

[6] Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

[7] Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992), "ARCH Modelling in Finance," *Journal of Econometrics*, 52, 5-59.

[8] Bollerslev, T, R. F. Engle, and D. B. Nelson (1994), "ARCH Models" in Engle, R. F. and D. L. McFadden eds., *Handbook of Econometrics, volume IV*, Elsevier Science, 2959-3038.

[9] Braun, P. A., D. B. Nelson, and A. M. Sunier (1995), "Good News, Bad News, Volatility and Betas," J*ournal of Finance,* 50, 1575-1604.

[10] Christie, A. A. (1982), "The Stochastic Behavior of Common Stock Variances," *Journal of Financial Economics*, 10, 407-432.

[11] Cox, J., J. Ingersoll, and S. Ross (1985), "An Intertemporal General Equilibrium Model of Asset Prices," *Econometrica*, 53, 363-384.

[12] Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica*, 50, 987–1008.

[13] Engle, R. F., D. M. Lilien, and R. P. Robins (1987), "Estimating Time Varying Risk Premia in the Term Structure: The ARCH–M Model,"*Econometrica*, Vol. 55, 391-407.

[14] Engle, R. F., and G. Gonzalez–Rivera (1991), "Semiparametric ARCH Models," *Journal of Business and Economic Statistics*, 9, 345-359.

[15] French, K. R., G. W. Schwert, and R. B. Stambaugh (1987), "Expected Stock Returns and Volatility," *Journal of Financial Economics*, 19, 3-29.

[16] Gallant, A.R. (1981). On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form. Journal of Econometrics 15, 211-245.

[17] Gennotte, G., and T. Marsh (1988), "Valuations in Economic Uncertainty and Risk Premiums on Capital Assets," *European Economic Review*, 37, 1021-1041.

[18] Hamilton, James D. (1994), *Time Series Analysis*, Princeton University Press.

[19] Härdle, W. (1990), *Applied Nonparametric Regression.* Cambridge University Press.

[20] Härdle, W. and O. Linton (1994), "Applied Nonparametric Methods" in Engle, R. F. and D. L. McFadden, *Handbook of Econometrics, volume IV*, Elsevier Science, 2295-2339.

[21] Hastie, T.J. and R.J. Tibshirani (1990). Generalized Additive Models. Chapman and Hall.

[22] Higgins, M.L., and A.K. Bera (1992). A class of nonlinear ARCH models. *International Economic Review* 33, 137-158.

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

[24] Lintner, J. (1965), "The Valuation of Risky Assets and the Selection of Risky Investment in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47, 13-37.

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

[26] 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," *Econometrica*, 64, 575-596.

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

[28] Nelson, D. B. (1990), "Stationarity and Persistence in the GARCH(1,1) Model," *Econometric Theory*, 6, 318-334.

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

[30] Nelson, D. B. and C. Q. Cao (1991), "Inequality Constraints in the Univariate GARCH Model", *Journal of Business and Economic Statistics*, 10, 229-235.

[31] Pagan, A. R., and Y. S. Hong (1990), "Non-parametric Estimation and the Risk Premium." In W. A. Barnett, J. Powell, and G. 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, 51-75.

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

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

[34] Robinson, P. M. (1983), "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis,* 4, 185-207.

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

[36] Schwert, G. W. (1989), "Why Does Stock Market Volatility Change Over Time", *Journal of Finance*, 44, 1115-1153.

[37] Sharpe, W. (1964), "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance,* 19, 567-575.

[38] Tadikamalla, P. R. (1980), "Random Sampling from the Exponential Power Distribution", *Journal of the American Statistical Association*, 75, 683-686.

[39] Weiss, A. (1986). Asymptotic Theory for ARCH models: Estimation and Testing. Econometric Theory 2, 107-131.

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

Tables and Figures

Table 1

Raw Data by Sub Period

	FP	I	II	III
μ	0.357	0.409	0.346	0.326
σ^2	32.205	72.783	14.079	19.793
κ_3	80.413	342.114	-18.602	-28.469
κ_4	10174.626	28550.860	16.069	988.330

Table 2

Estimates		
	Kernel	Fourier
a	-0.167 (0.062) (0.099)	-0.451 (0.092) (0.134)
b_1	0.973 (0.010) (0.016)	0.927 (0.015) (0.022)
c_{11}	-0.234 (0.053) (0.065)	-0.306 (0.047) (0.054)
c_{12}	0.013 (0.083) (0.063)	-0.096 (0.068) (0.070)
c_{21}	0.193 (0.046) (0.090)	0.212 (0.045) (0.059)
c_{22}	0.246 (0.089) (0.087)	0.367 (0.079) (0.077)
ν	1.578 (0.102) (0.133)	1.590 (0.110) (0.142)
γ_0	-	0.122 (0.080) (0.150)
γ_1	-	-0.295 (0.040) (0.131)
γ_2	-	0.067 (0.008) (0.027)
γ_3	-	0.147 (0.036) (0.067)
γ_4	-	-0.143 (0.018) (0.060)
δ	0.207	-
$\ell =$	1427.02	1445.63
Linearity test $H_0 : \gamma_i = 0, i > 1$ $(p-value)$	-	16.744 (0.0008)

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

Table 3

Sub-period estimates			
	1926-1945	1946-1973	1974-1997
a	-0.181 (0.115)	-3.120 (2.145)	-0.531 (0.503)
b_1	0.968 (0.021)	0.524 (0.329)	0.916 (0.079)
c_{11}	-0.305 (0.127)	-0.187 (0.080)	-0.353 (0.103)
c_{12}	0.006 (0.178)	-0.061 (0.135)	0.065 (0.185)
c_{21}	0.168 (0.105)	-0.081 (0.144)	0.298 (0.107)
c_{22}	0.220 (0.217)	0.122 (0.146)	0.128 (0.208)
ν	1.306 (0.176)	1.979 (0.266)	1.686 (0.208)
δ	0.321	0.118	0.165
$\ell =$	305.62	625.95	502.70

Table 4

Simulation results			
	True	Mean estimate	Standard deviation
a	-0.167	-0.205	0.107
b_1	0.973	0.965	0.017
c_{11}	-0.234	-0.200	0.056
c_{12}	0.013	0.036	0.076
c_{21}	0.193	0.158	0.054
c_{22}	0.246	0.180	0.085
ν	1.578	1.998	0.234
δ	0.207	0.163	0.035
mean(μ)	0.004	0.001	0.002
std(μ)	0.009	0.013	0.003

Figure 1. Monthly returns on S&P 500
1926-1997

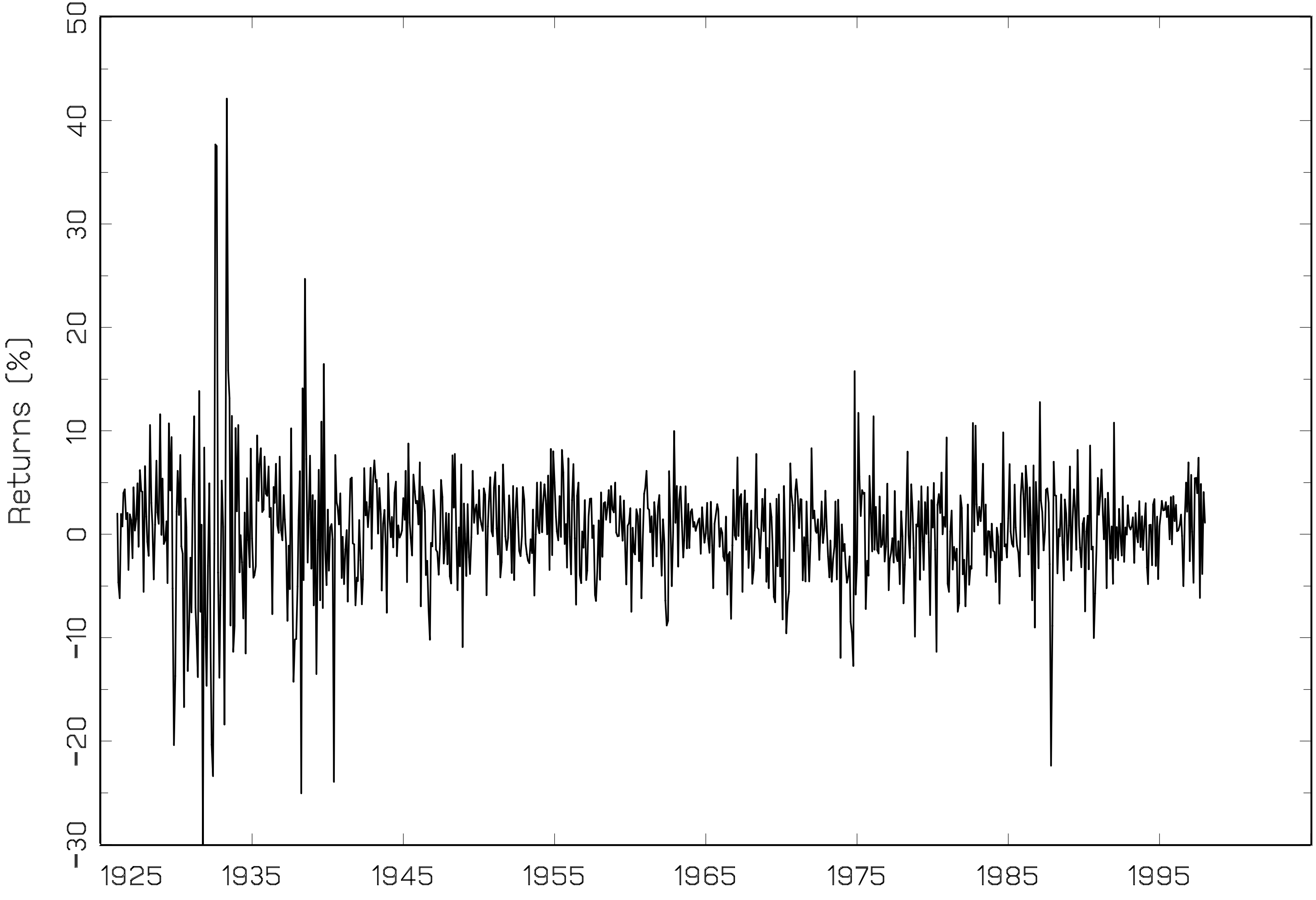

Figure 2. Smooth kernel estimate of risk premium
Full sample

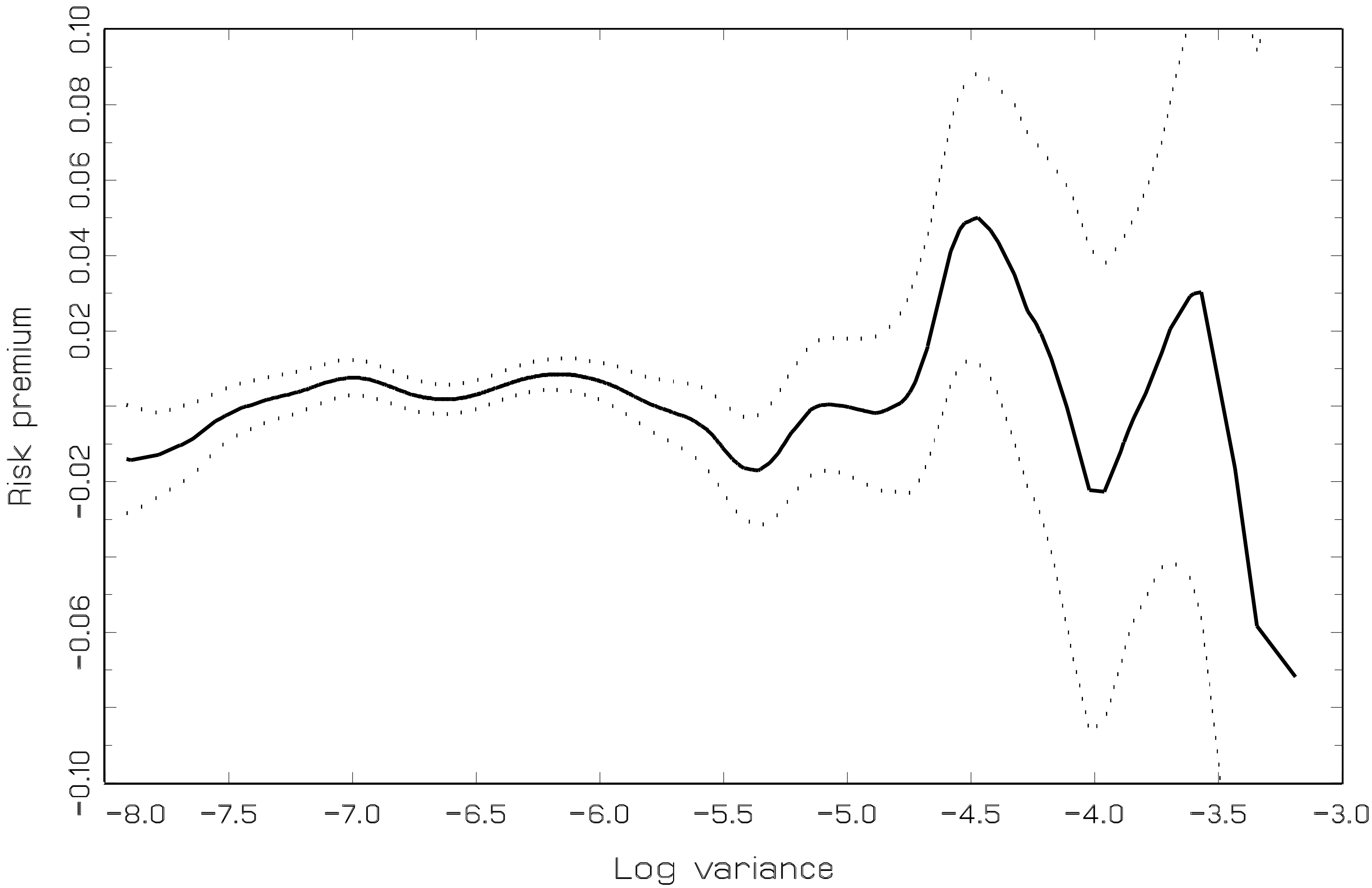

figure 3. Estimated risk premium
Kernel

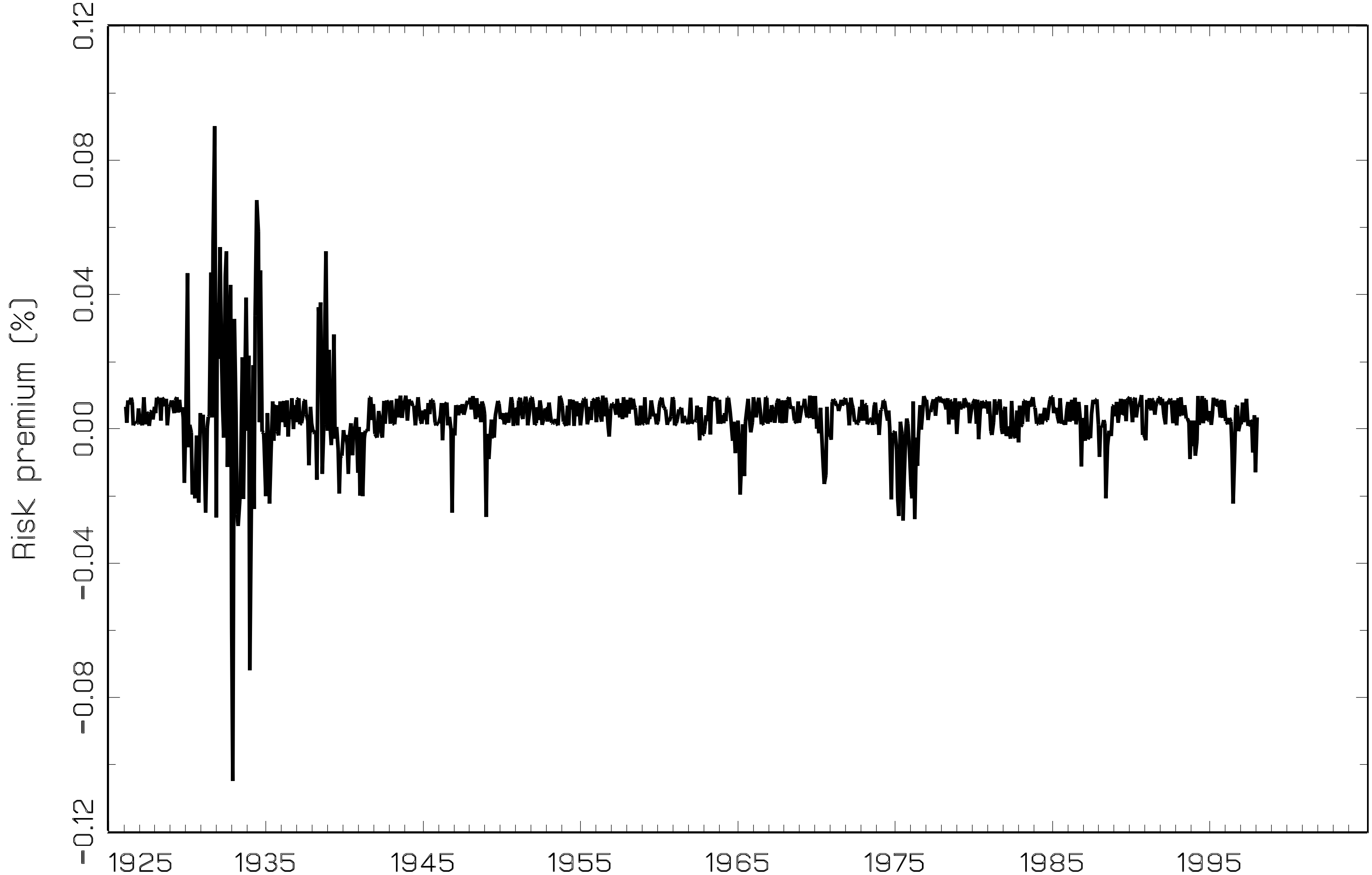

Figure 4. Estimated conditional variance
Kernel

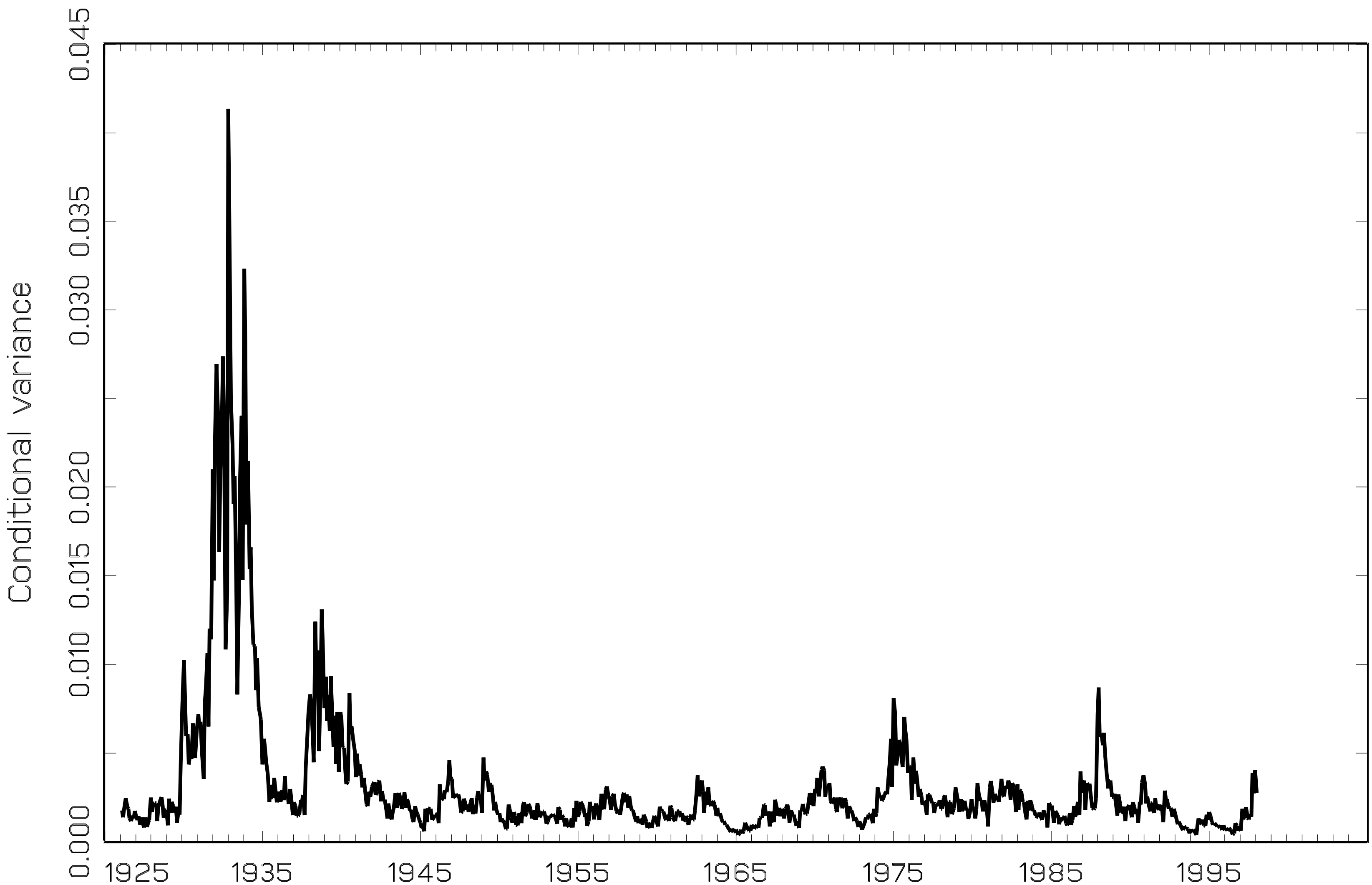

Figure 5. Fourier series estimate of risk premium

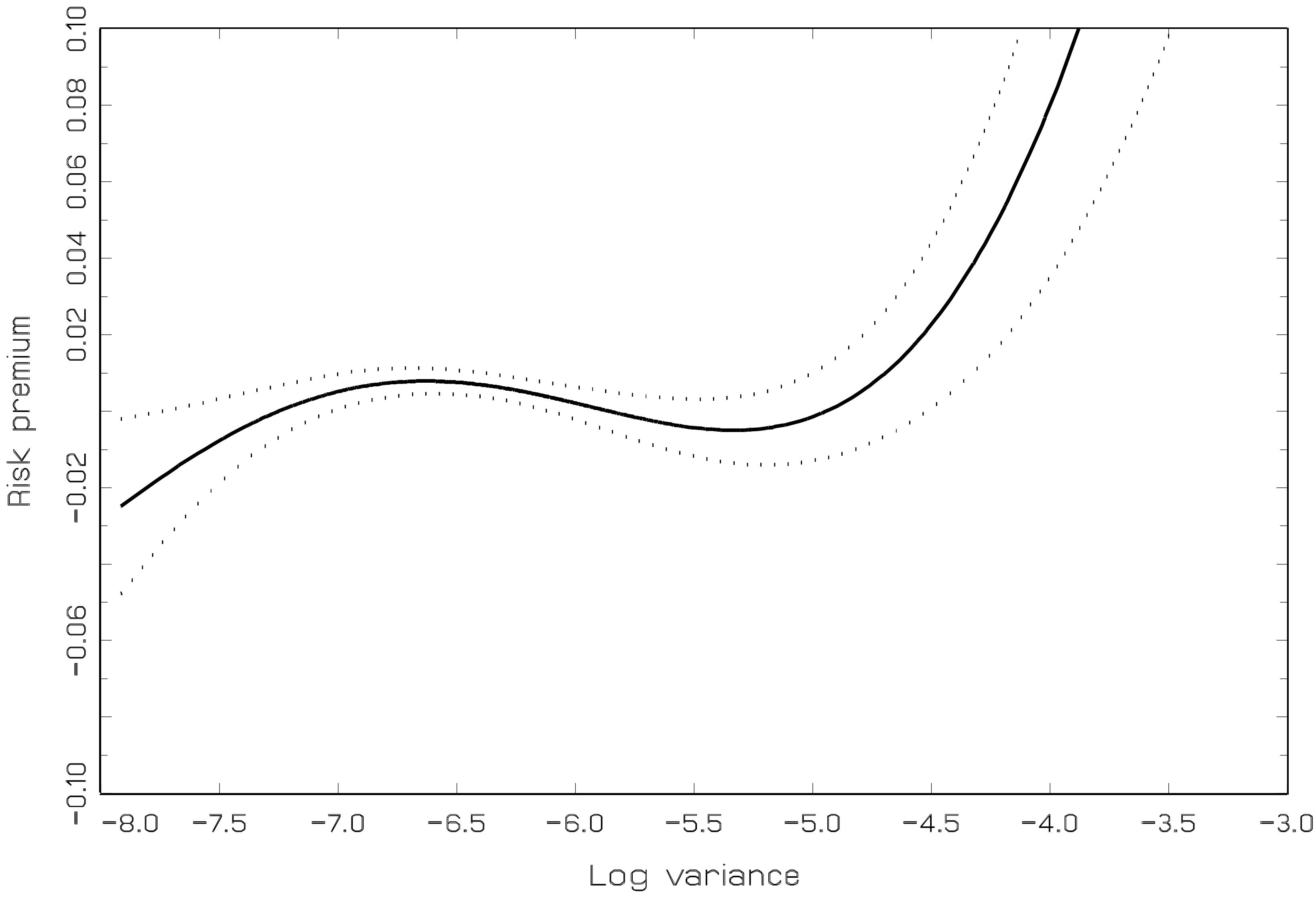

Figure 6. Kernel and Fourier series estimate of risk premium

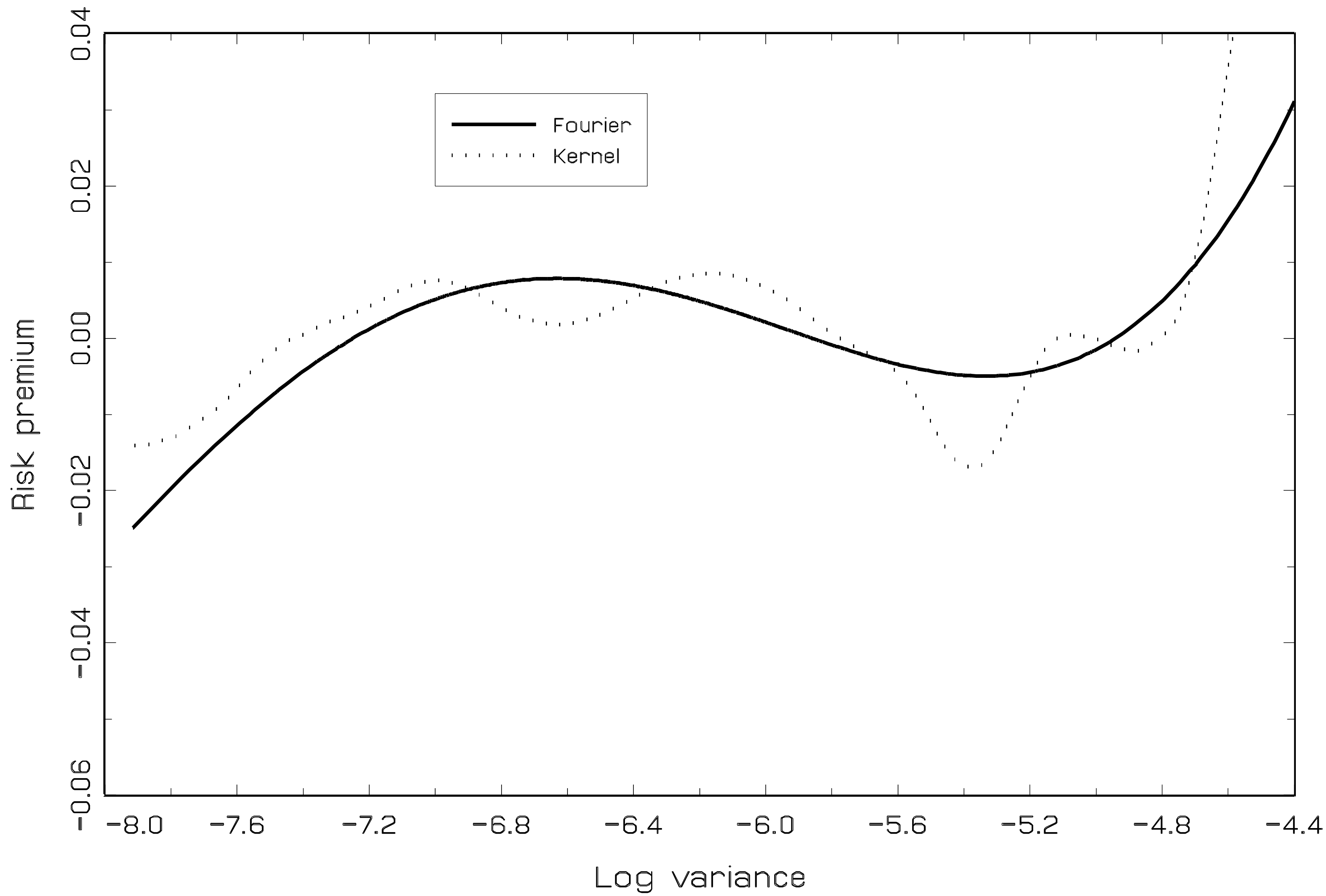

Figure 7. Standardized residuals
Kernel

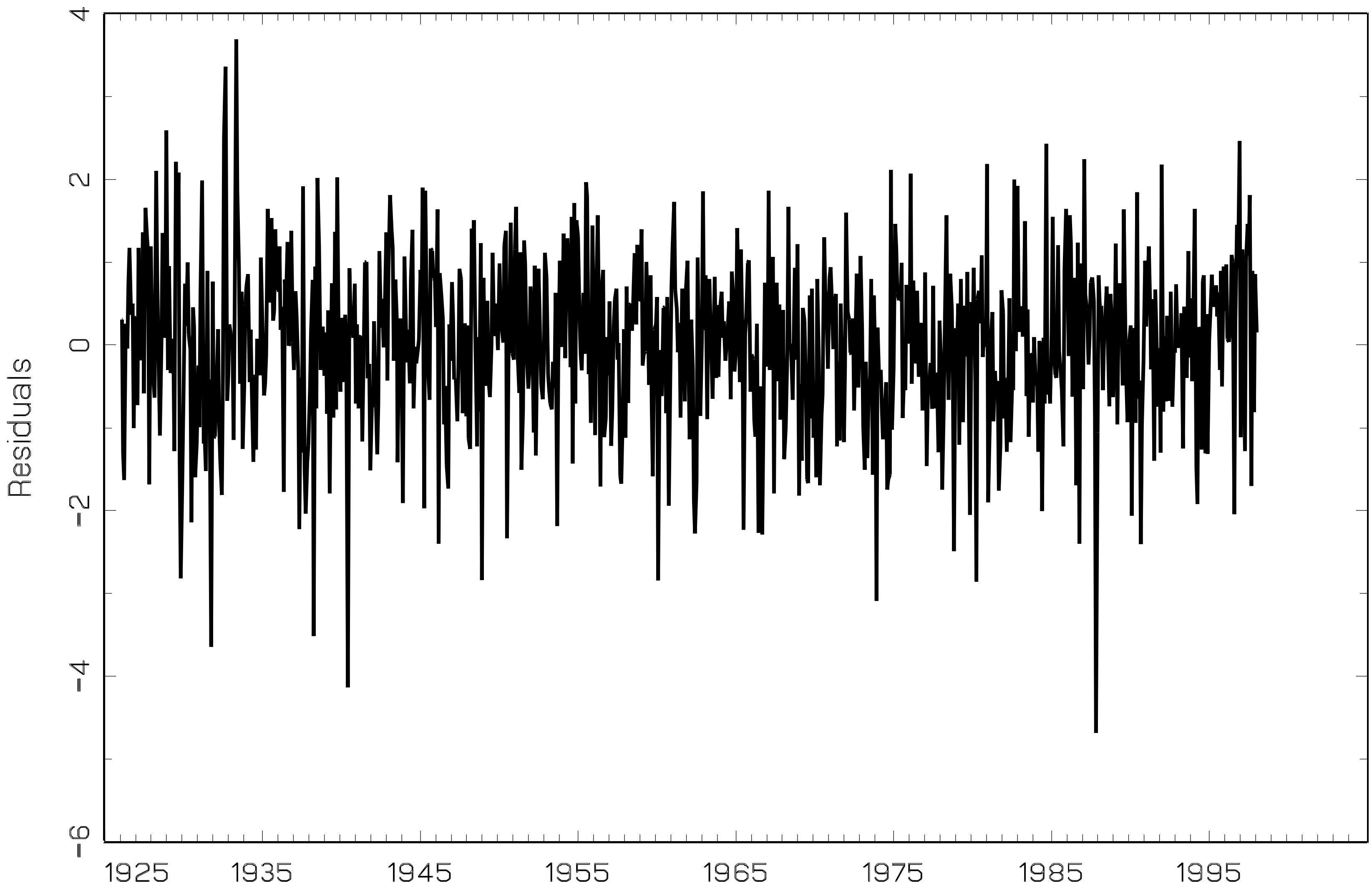

Figure 8. Autocorrelogram of standardized residuals
Kernel

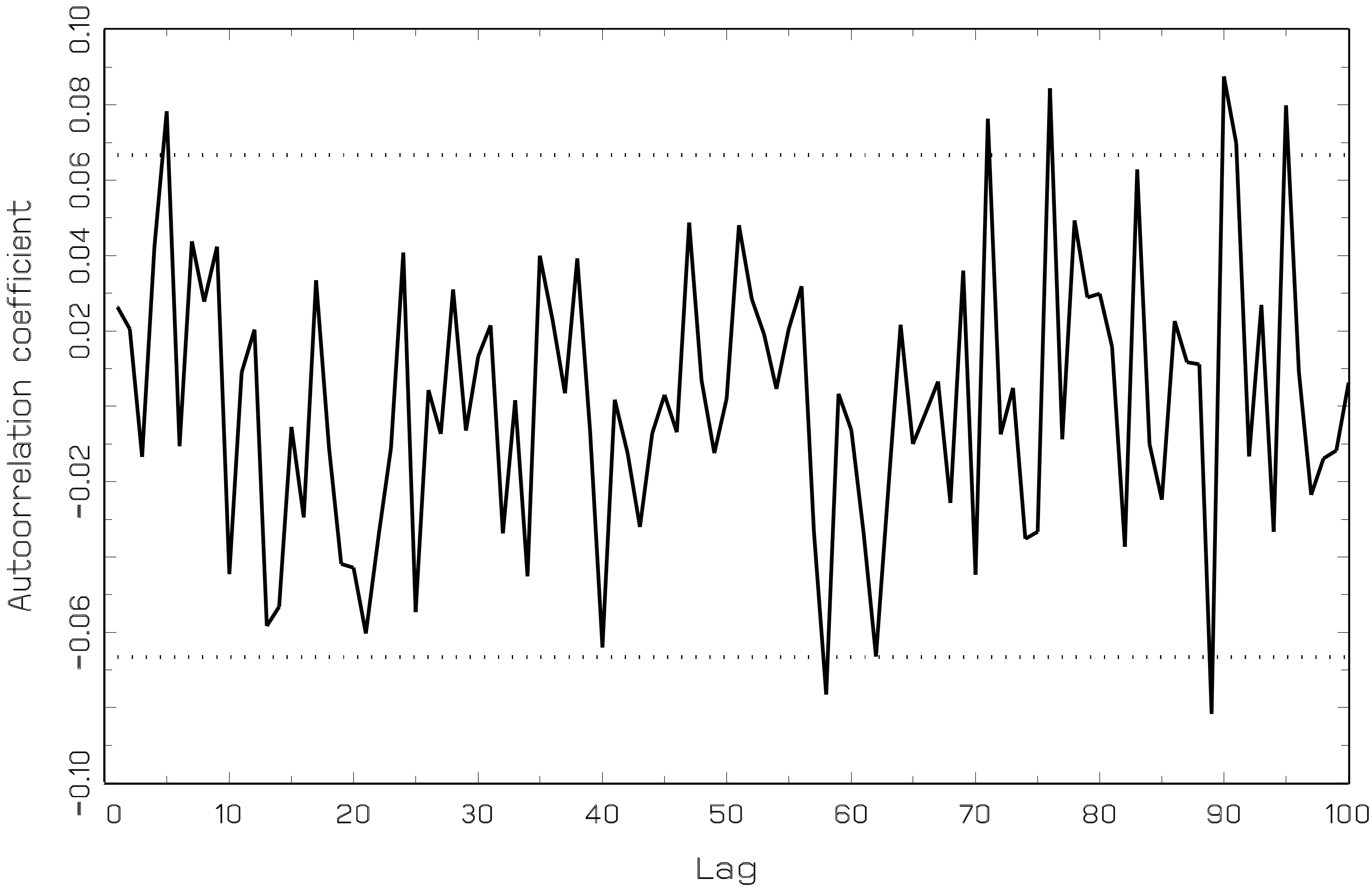

Figure 9. Autocorrelogram of squared standardized residuals
Kernel

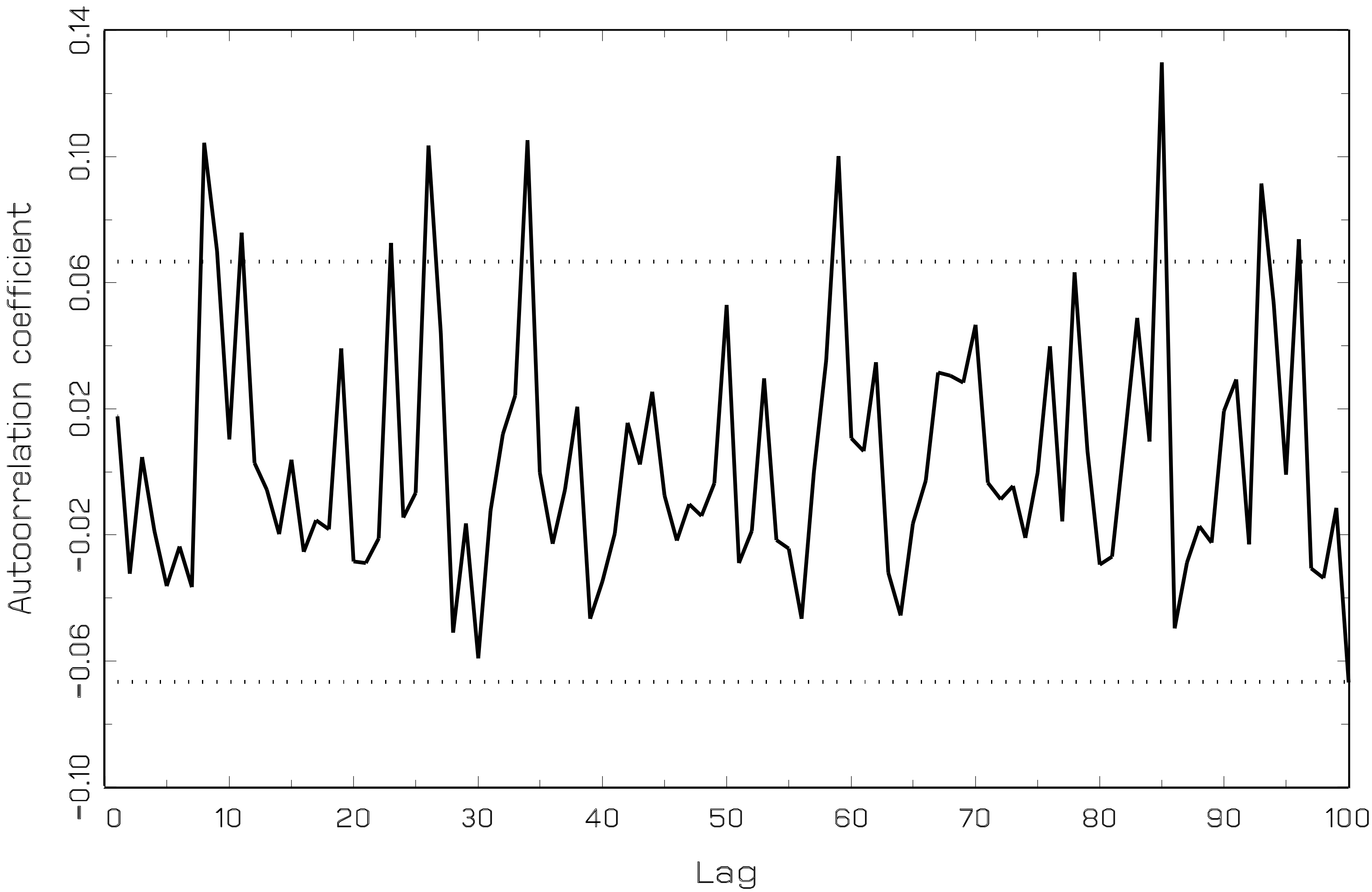

Figure 10. Density of standardized residuals vs. GED

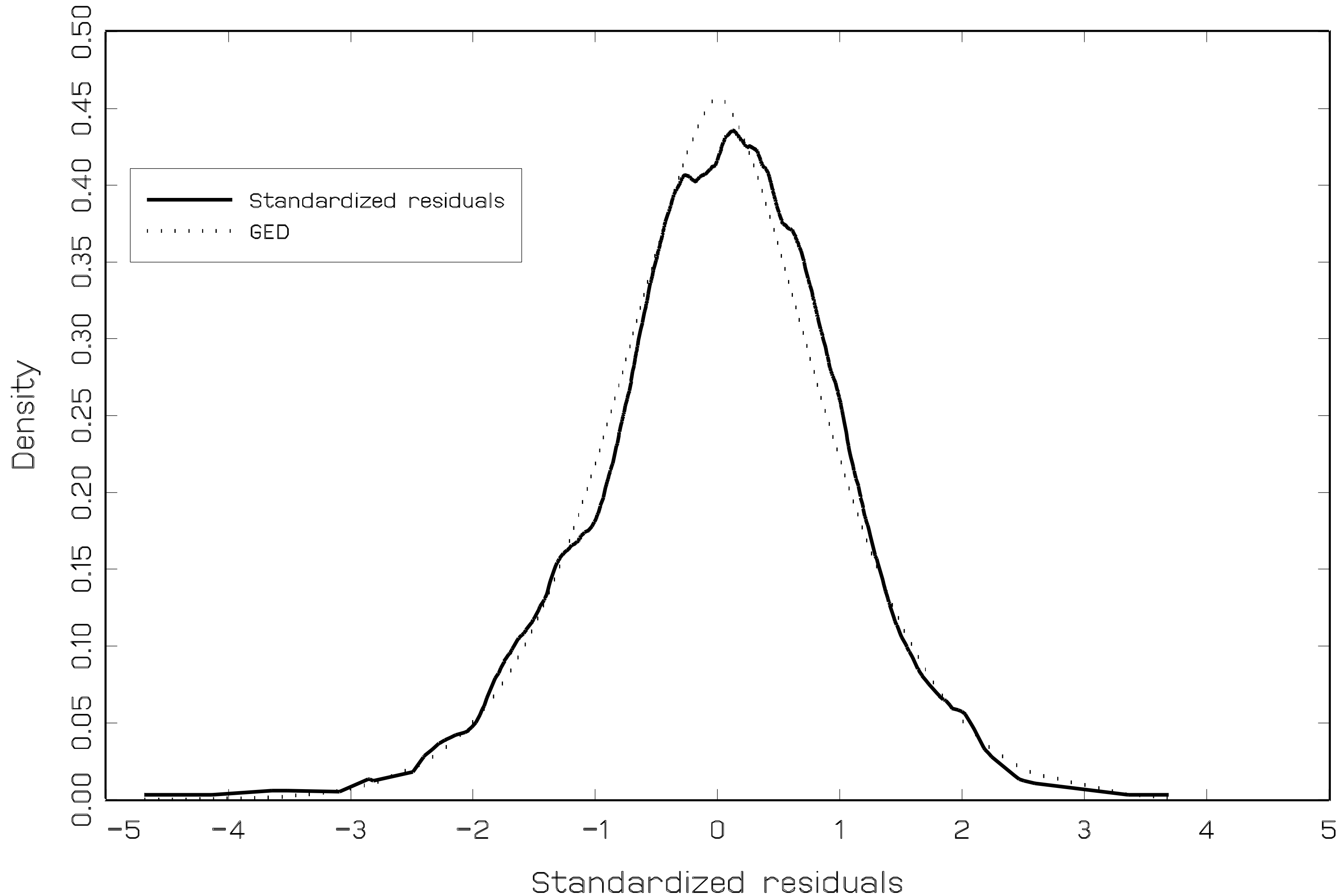

Figure 11. Smooth kernel estimate of risk premium
First subsample

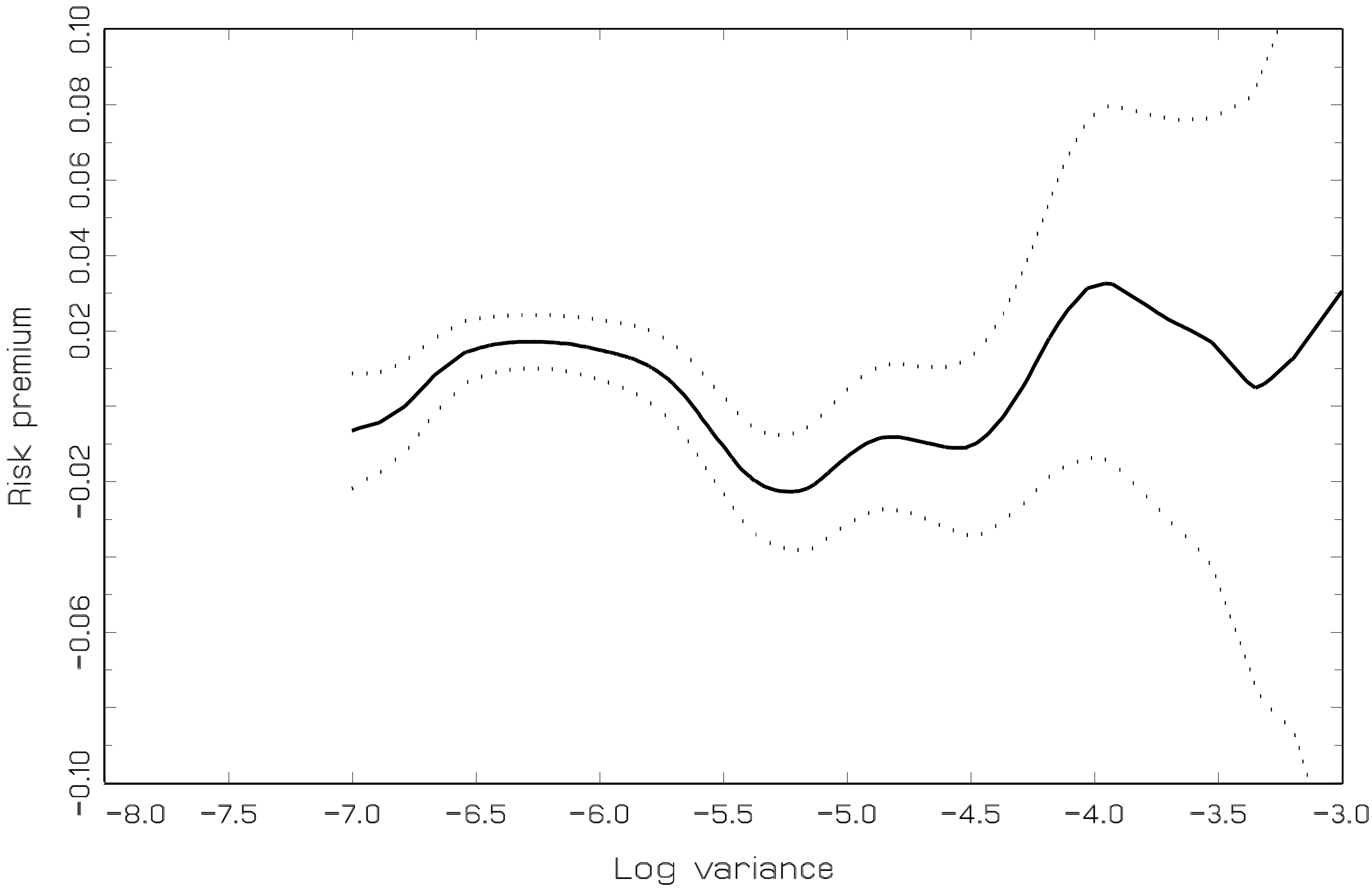

Figure 12. Smooth kernel estimate of risk premium
Second subsample

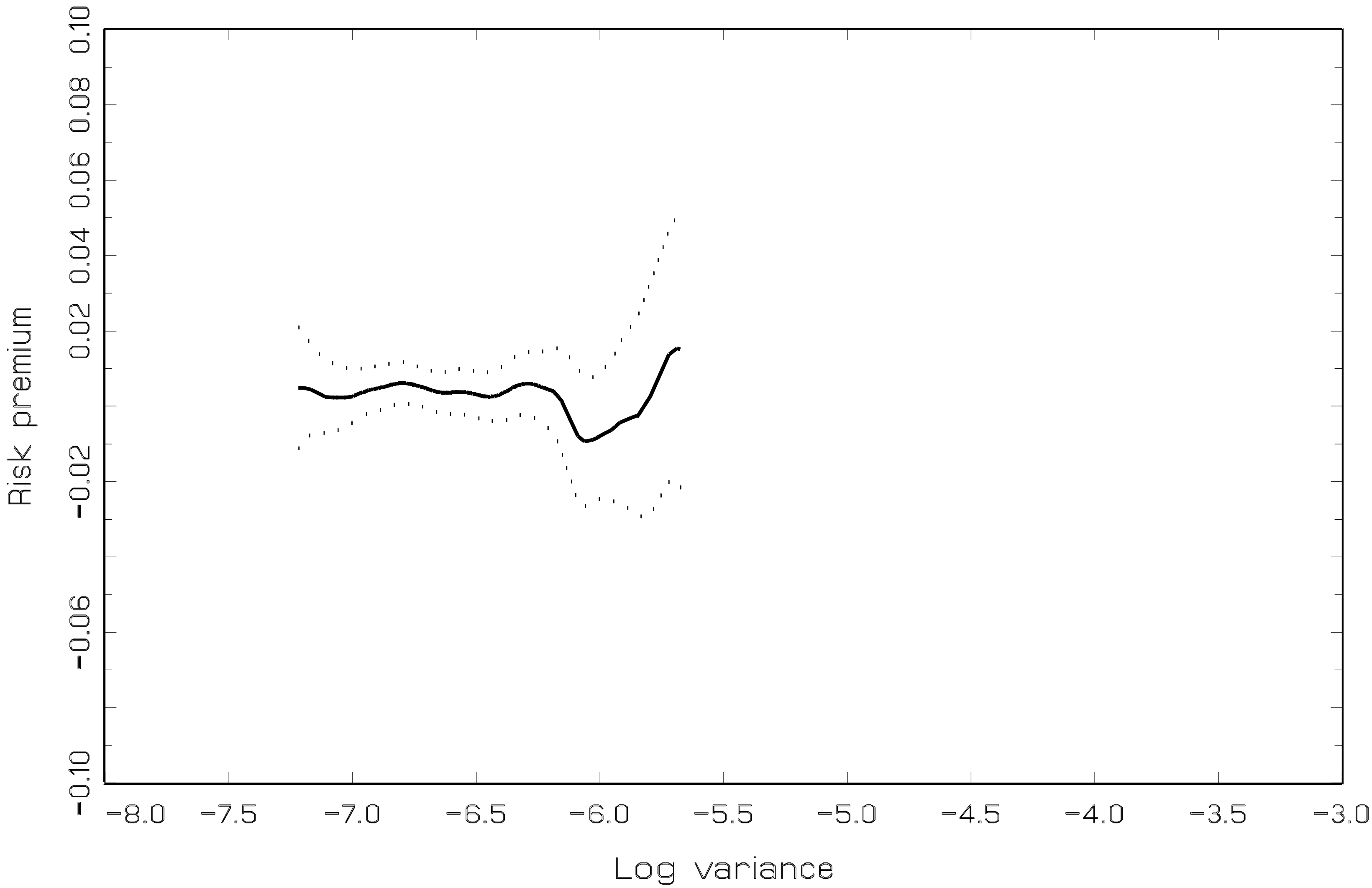

Figure 13. Smooth kernel estimate of risk premium
Third subsample