RETURNS OR DIFFERENCES? METHODS FOR RISK FUNCTIONAL FORM SELECTION

SERGEI MOROZOV

Abstract. We describe and illustrate three categories methods that help decide on whether risk factors, underlying portfolio risk measurement framework such as VaR, should be represented as returns or differences (or some hybrid form). Methods in the first category rank alternative representations by their performance with respect to stationarity tests, in-sample or out-of-sample measures of goodness of fit, or by information-theoretic considerations. These methods are largely informal and must be handled with care, as the two representations are not nested, goodness-of-fit tests can be biased if their null distributions are themselves estimated, traditional unit root tests of stationarity are designed to address somewhat different question, while homoscedasticity tests tend to have power only against specific alternatives, are sensitive to non-normality and often inconclusive. Second category of methods revolves around some form of elasticity of volatility model which conveniently nests both return and difference representations. Depending on the time-series in question, and one's willingness to make distributional or prior assumptions, GMM, maximum likelihood or Bayesian estimation procedures may be called upon. Third category nests return and difference representations in the wider non-parametric class and includes methods to estimate the volatility as a smooth function of the level, allowing for a possibility of functional representation switch depending on the level. The methods are illustrated using daily interest rate swap data, for which we find that the return formulation is preferable when rates are below roughly 2.8%, as is prevalent in most recent four year sub-sample, while the difference representation works better for rates above this cutoff.

JEL classification: C11; C58

Keywords: Stationarity, goodness-of-fit, density forecast, unit root tests, elasticity of volatility, conditional heteroscedasticity, mutual information, stochastic volatility, GMM, maximum likelihood, Markov chain Monte Carlo, Gibbs sampler, sequential Monte Carlo, risk representation, non-parametric diffusion function

---

Date: May 31, 2014.

Version 2.06.

The author gratefully acknowledges Richard Barraclough, Victor Chernozhukov, Jamel Saadaoui, Louis Scott and participants of Third International Symposium in Computational Economics and Finance 2014 for helpful comments and suggestions that improved the paper. Parts of the paper were written when the author was at Morgan Stanley. Standard disclaimer applies.

✉ Sergei.Morozov@ey.com.

Contents

List of Figures

List of Tables

1. Introduction

Accurate portfolio risk measurement is predicated on striking the right balance between two sources of information that impinge on the risk forecast – current market conditions and historical experience. Indeed, a key assumption underlying risk measurement methodologies is that the future looks like the past.[1]

A technical basis for treating the future like the past is the concept of stationarity, meaning that the joint probability distribution of potential losses for a fixed portfolio does not change when shifted in time. Since the potential loss distribution of a portfolio is assembled from arbitrary exposures to risk factors, it follows that risk factors themselves should be expressed in stationary form, if the same set of factors is to be used for an arbitrary portfolio. Expressing time-series in a stationary form is a way of addressing frequent critique of risk estimates as based on the data that are no longer relevant to the current conditions. In stationary form, all data look statistically similar and retain their relevance.

At its core, the entire field of time-series econometrics can be viewed as a quest to suitably transform data in order to achieve stationarity. For instance, models for detrending, seasonal adjustment and even GARCH all revolve around separating predictable features from purely random noise that is distributed independently and identically over time. A direct consequence for the more precise risk measurement would be to apply transformation to stationarity to all the risk factors, either one at a time or, better still, jointly. Indeed, this is the approach taken in assorted academic and regulatory papers advocating the use of GARCH for VaR (see, e.g., Angelidis, Benos, and Degiannakis (2004) and references therein). Unfortunately, for large portfolios exposed literally to thousands of different risk factors, the method of fitting sophisticated time-series models to either the entire set of factors jointly or to individual series one at a time founders on the reefs of dimensionality, complexity and implementability.

It should be of no surprise, then, that, currently, only two forms of risk factor functional representation are widely used, relative (i.e. proportional) change ("returns") and absolute change ("differences"), and the choice between the two can have a nontrivial impact on the estimated value-at-risk. As the nature of the risk factor time series going into the overall risk measurement framework may evolve over time, the choice of functional form should be periodically reevaluated. Similarly, for a novel risk factor the choice should be established from scratch. This paper aims to lay out supporting statistical foundations to making intelligent choice about risk functional form. In doing so, it provides both description of several statistical methods behind the choice and guidance to the practical application. We illustrate various approaches using a dataset of medium- to long-term US interest rate swap rates since the turn of the century.

The statistical methods for the functional specification choice that we discuss fall into three broad categories. First category of methods takes as given that the only two options – returns or differences – are available. Methods in this category attempt to rank the two alternatives based on statistical criteria such as goodness-of-fit statistics, stationarity tests or measures linked to model complexity . We argue that such non-nested approaches must be applied with caution. Goodness-of-fit criteria could be a shaky basis for functional form comparison, for three reasons.

[1]Value-at-Risk (VaR) models, in use at most banks to satisfy market risk capital requirements and concerning with tail risks during "normal" times, identify the relevant past by selecting a relatively short recent time window, typically no longer than 4 years. Stress scenarios, on the other hand, may look further into the past for guidance on large shock magnitudes and their frequency of incidence.

First, goodness-of-fit is commonly a joint hypothesis of an appropriate functional form transformation and parametric distributional assumptions. Fitting parsimonious parametric forms to unconditional distributions of most financial time-series is likely to be overwhelmingly rejected no matter the transformation used, while fitting complex flexible models may lead to the in-sample non-rejection for both forms and failure out-of-sample. Second, many of these tests tend to have low power in finite samples, particularly if parametric assumptions are relaxed. Finally, many of these tests (such as Kolmogorov-Smirnov or Anderson-Darling) are biased when the null distribution is itself estimated instead of being pre-specified in advance.[2] Testing for stationarity must also be handled with similar care. Oftentimes, testing for stationarity is taken to mean testing for unit roots but that only relates to the first order stationarity, telling us whether a time-series has a stochastic trend and whether it needs to be further differenced in order to achieve stationarity. For risk measurement this is insufficient, especially as most daily financial time-series tend not to have unit roots over medium-term sample periods, whether measured in returns or differences, leading to inconclusive and inconsistent ranking of the two forms. Second- (and higher-) order stationarity is more important, as are issues of structural breaks and regime switches. A battery of heteroscedasticity tests, as exemplified by section 3.3, is a better starting point even when these tests have power only against specific alternatives. Even then, finding ways to neutralize non-constant variance may not be sufficient for the task at hand since other measures of variation may be important. Information-based quantities, touched upon in section 3.4, may provide additional food for thought, but are not easy to link directly to objects of interest.

The second category of methods centers around constant elasticity of variance (CEV) specification and includes some of its extensions. The main merit of this approach is that CEV nests both returns and differences, thus standard tests could be used to discriminate the two. Depending on the strengths of assumptions one is willing to make, there are several estimation and testing techniques available. At the one extreme is Generalized Method of Moments (GMM) approach that only uses limited information embodied in selected moment or orthogonality conditions. No assumption is made about the distributional form of residuals. If, on the other hand, we are willing to commit to a particular distribution of residuals, maximum likelihood estimation approach exploits this additional information to gain further statistical efficiency. At the other extreme, Bayesian methods allow incorporating prior information to restrict parameter space more tightly and guide inference toward more intuitively appealing regions in the parameter space. Aside from ability to incorporate prior information, Bayesian methods have some technical advantages as well. For example, Bayesian approach eschews difficult issues of multivariate optimization that plague maximum likelihood and GMM. Furthermore, posterior inference does not rely on any asymptotic approximations and so applies in finite sample. Hence, it will result in tighter confidence sets and more precise inference, assuming the model is correctly specified.[3] Finally, an attractive feature of Bayesian approach is its modular nature so that Bayesian CEV model can be, without difficulty, extended to a multivariate panel setting where multiple time-series are constrained to have common elasticity of variance. In the interest rate modeling practice, the assumption of constant elasticity of variance appears too restrictive. Patterns in elasticity estimators over subsamples and along the yield curve suggest that, instead, elasticity of volatility tends to decline as

[2] Modifications to Kolmogorov-Smirnov and Anderson-Darling test are available for select parametric families of null distributions.

[3] Indeed, Chan, Karolyi, Longstaff, and Sanders (1992) estimate scale of variance so imprecisely that the parameter is not significantly different from zero, while it should be strictly positive.

the rates rise. To accommodate this feature we propose a couple of variable elasticity models. Further extensions allow differential elasticities for upward and downward moves and introduce additional sources of randomness in volatility.

Third category nests the return and difference specifications within an even richer functional class of diffusion coefficient representations that is described non-parametrically. In other words, the object of estimation is now the conditional volatility *function* $\sigma(\cdot)$. Particular case of constant function specifies difference representation, while affine functions correspond to return representation. While a variety of intermediate shapes is possible, which makes it harder to form a definitive conclusion about functional form for the entire range of values that a time-series can take, the diffusion function representation facilitates local functional form choice, i.e., choice conditional on current levels of the series. In other words, if diffusion coefficient function appears more or less constant within some range, the difference specification is preferred within that range. Elsewhere, in regions of rapidly changing volatility, the return representation would likely work better. The additional flexibility of being able to switch between representations depending on the level of the series could therefore be useful. Combining methods in this category with variable elasticity models from the second category yields a fairly robust conclusion that the US interest rate swap rates should be modeled as returns below roughly 280 basis points and as differences above the cutoff.

The rest of the paper is organized as follows. Section 2 provides brief description of the interest rate swaps dataset that we use to illustrate our methods. Section 3 is dedicated to the critical analysis of non-nested approaches such as goodness-of-fit tests (section 3.1), unit root tests (section 3.2), heteroscedasticity tests (section 3.3) and mutual information (section 3.4). Section 4 lays out CEV model and three ways to estimate it, with sections 5 through 10 extending the model in various directions, including panels, variable elasticity settings and environments where volatility has its own sources of randomness. Section 11 describes non-parametric method of estimating the volatility function and tests of parametric models against non-parametric alternative. Finally, section 12 offers concluding remarks. Appendices A through D explain estimation details in more depth.

2. Sample Risk Factor Dataset

To put the issue into the practical perspective, we study risk form selection on a recent daily dataset of non-seasonally-adjusted US dollar interest rate swap rates over July 3, 2000 through July 26, 2012 (3,019 data points excluding holidays) from the Federal Reserve Bank of St. Louis database. To focus more clearly on the behavior of rates that is not tied to the jumps in the overnight Federal Funds policy target rate, we excluded all maturities less than one year. All 8 series are plotted in Figure 1 with shaded areas representing the last four and last one years of data. Visual inspection of plots of returns and differences in Figure 2 indicates that the difference representation looks more uniformly jagged throughout the entire sample, and thus appear to have a better chance of selection. Post mid-2007, the return time-series visually appear to have less time variation in volatility. Statistical analysis that follows will generally reinforce these initial observations.

3. Globally Non-nested Tests

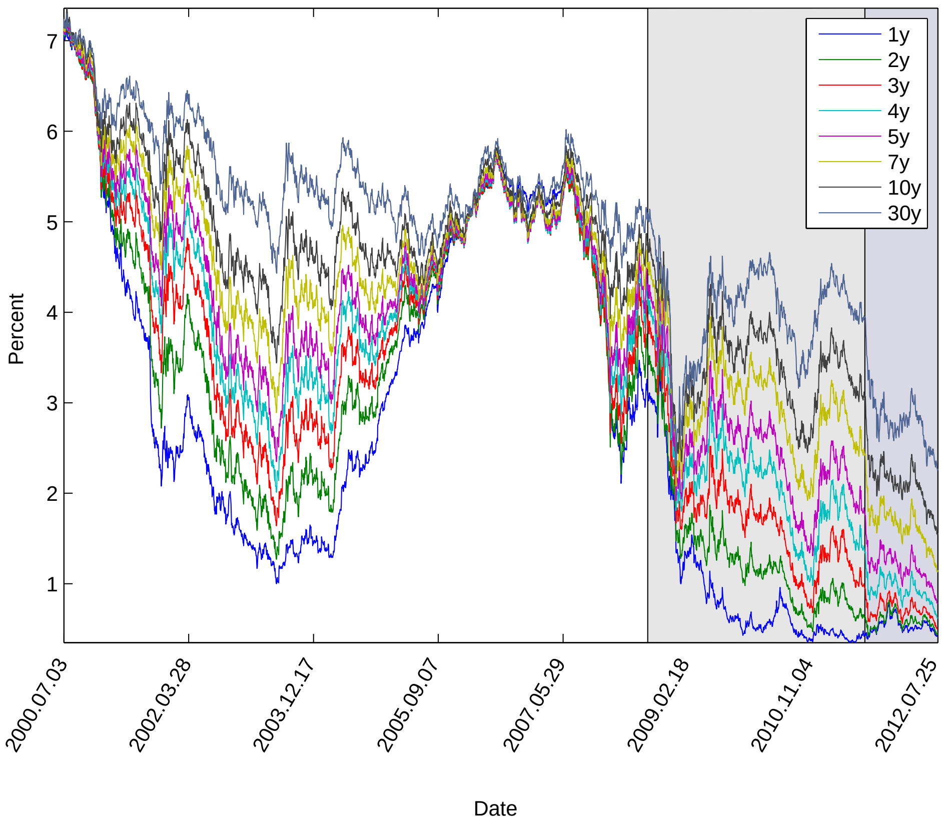

FIGURE 1. US dollar interest rate swap rates.

A beguiling approach to try to distinguish whether the interest rate dynamics should be modeled via absolute or proportional changes is to test separately how well each form captures certain desirable properties such as adequate model fit, satisfactory degree of stationarity, low model complexity or muted sensitivity to outside information. Relative to such properties, it is tempting to consider the two alternatives as non-nested in a sense that it is not possible to derive one representation by means of an exact set of parametric restrictions or as a result of a limiting process (Gouriéroux and Monfort, 1994). Moreover, it would seem attractive to treat them as globally non-nested, that is, without reference to any encompassing model (Pesaran, 1987). The methods in this section follow this approach with varying degrees of sophistication but tend not to lead to a clear verdict. Such failure is due to the lack of due care as well as confusion over purposes of model selection versus hypothesis testing as the literature on non-nested hypothesis testing showed quite compellingly (Pesaran and Weeks, 1999). We will revisit this point at the end of the section.

3.1. Goodness-of-Fit Measures. Measures of goodness-of-fit typically summarize the discrepancy between observed values and values implied by hypothesized statistical distribution or

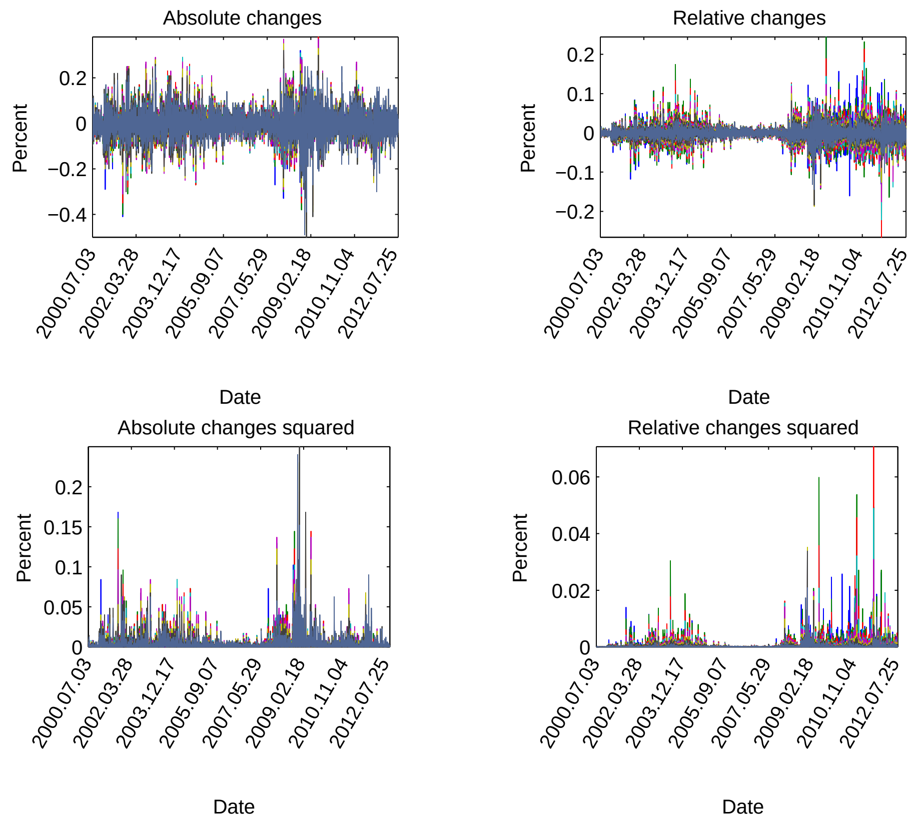

FIGURE 2. Changes in US dollar interest rate swap rates.

statistical model. To use these measures for ranking rival functional forms we have to make assumption about unconditional distribution that the data under the two alternative representation have to satisfy.

While the most common kind of distributional assumption is that of normality, financial time-series are notorious for their non-normality, so we are not even going to try normal distribution as the null.[4] Instead, we will be using location-scale Student t distribution. In this case, the hypothesis of difference representation is

$$\Delta r_t \sim t\left(\mu_d, \sigma_d^2, \nu_d\right), \tag{3.1}$$

while that of return representation is

$$\frac{\Delta r_t}{r_{t-1}} \sim t\left(\mu_r, \sigma_r^2, \nu_r\right). \tag{3.2}$$

Thus, the idea here is to produce goodness-of-fit measures and compare them across the two representations. Importantly, one has to select measures that are invariant with respect scale and location, which excludes a number of popular measures such as root mean square error. Below we explore the utility of few common ones.

[4] p-value of the null of normality for a one-year sample using either representation is on the order of 10^{-50}.

3.1.1. Pearson's χ^2 Statistics. The Pearson's goodness-of-fit statistics is one of the best known ways to evaluate the agreement between fitted and empirical distributions. It is given by

$$P(g) = \sum_{i=1}^{g} \frac{(n_i - \mathbb{E}n_i)^2}{\mathbb{E}n_i}, \tag{3.3}$$

where g is a number of cells, n_i is the number of observations in cell i, and $\mathbb{E}n_i$ is the expected number of observations (based on maximum likelihood estimates). For i.i.d. observations the asymptotic distribution of $P(g)$ under the null of correct distribution is χ^2_{g-k-1} where k is the number of independent parameters fitted. This test is in fact equivalent to an in-sample density forecast test (Diebold, Gunther, and Tay, 1998). The Pearson's statistics is therefore ″small″ when all of the observed counts (proportions) are close to the expected counts (proportions) and it is ″large″ when one or more of the expected proportions deviates notably from what is expected under the null hypothesis of the two distributions being the same. The choice of the number of cells used to evaluate the statistics is far from obvious. Palm and Vlaar (1997) propose $g = 50$ when the number of observations is 2,252, and suggest that the number of cells increase at the rate $N^{0.4}$. In consequence, we use $\texttt{floor}\left(2.28N^{0.4}\right)$ as our preferred cell size.

3.1.2. Information Criteria. Information criteria are grounded in the concept of entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality. Information criteria can be said to describe the tradeoff between bias and variance in model construction, or, loosely speaking, between precision and complexity of the model. Information criteria make direct use of the maximum likelihood estimates of the fitted distribution, thus they directly judge which distribution within a parametric family is more likely to have generated the data. Alternative information criteria differ in the way they penalize the number of parameters. Since the number of parameters in (3.1) and in (3.2) is the same, the choice among different information criteria does not matter. Reported results use the Akaike Information Criterion with correction for finite sample size (Brockwell and Davis, 1998).

3.1.3. Kolmogorov-Smirnov-Lilliefors and Anderson-Darling Tests. The Kolmogorov-Smirnov test (KS test) is a non-parametric test of equality of one-dimensional probability distributions used to compare a sample with a reference probability distribution. The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. The null distribution of this statistic assumes that the sample is drawn from the reference distribution.

The Kolmogorov-Smirnov test can be modified to serve as a goodness-of-fit test. However, if the reference distribution is estimated instead of being specified in advance, the p-values are incorrect. Correcting the Kolmogorov-Smirnov test for this bias when the parameters estimated are parameters of scale and location leads to the Lilliefors family of tests (Lilliefors, 1967). Adjustments are specific to a chosen scale-location family of distributions and have not been developed in the literature for the Student t distributions (with pre-specified degrees of freedom). It is also known that using the sample to modify the null hypothesis reduces the power of a test.

An alternative to the Kolmogorov-Smirnov-Lilliefors test is the Anderson-Darling statistics that weighs the squared distance between the empirical and the reference cumulative distribution functions by the inverse of its variance, emphasizing more the tails of the distributions than the

Kolmogorov-Smirnov distance. The Anderson-Darling test can also serve as a goodness-of-fit-test and needs similar bias adjustments to enable testing when the reference distribution is from a location-scale family with estimated location-scale parameters.

Since bias correction for both tests only works for location-scale families and depends on the shape of the reference distributions, we set the degrees of freedom parameter $\nu = 5$ (implying finite first four moments) and derived the critical values of Kolmogorov-Smirnov-Lilliefors and Anderson-Darling tests by Monte Carlo simulation. Application of these tests to the interest rate swap data is documented in section 3.1.6.

3.1.4. Histogram Binning. Parametric goodness-of-fit tests above describe the difficulty of coercing the sample into the straightjacket of a parametric form. Non-parametric methods are commonly used to unfetter the data from such constraints by substituting the *divine insight* of knowing the distributional form of the data generating process up to a finite number of unknown parameters with less forceful assumptions such as smoothness or local shape restrictions. An explicit advantage of non-parametric methods is the recognition that fitted models are inherently misspecified approximations. Increasing complexity or reducing the degree of smoothness of a fitted model tends to decrease the misspecification bias at the cost of inflating the estimation variance. Non-parametric methods make the complexity of the fitted model depend upon the sample in a way that balances this trade-off. Non-parametric models are typically indexed by a tuning parameter which controls the degree of complexity. One such parameter that originated in coding and information theory is the *stochastic complexity*. This metric is defined, relative to a collection of models, as the fewest number of bits in a probabilistic sense with which the data can be encoded using a code designed by help of the models in the class.[5] A simple class of models admitting a particular elegant formulation in terms of stochastic complexity is that of histograms. Histograms are "minimalist" in terms of the degree of smoothing as they have a finite number of "steps" and associated discontinuities. Compared to kernel methods, histograms have a lesser tendency to underestimate the peaks of the sampling density. They are also the oldest and most commonly used. Here, unknown parameters represent the number and locations of bins. Since with large enough number of bins histograms can fit almost anything, the problem of determining the number of parameters is precisely that of choosing the model complexity that is required to achieve a satisfactory fit.

Considering all possible partitions of the real line to support every conceivable kind of irregular histogram requires computationally heavy and intricate search procedures that destroy much of the benefit of bias reduction. For this reason, we limit our exploration of histogram binning to the two kinds of regular histograms – *regular* histograms with equal binwidth and *equi-depth* histograms with equal probability in each bin. Among many binwidth estimators for the regular histograms we have chosen an estimator due to Hall and Hannan (1988) based on stochastic complexity as the outcomes with other penalized likelihood estimators were similar. For the equi-depth histogram binning, we used the likelihood cross-validation approach, correcting for duplicate observations.[6] We complement these two with an irregular histogram estimator of Kumar, Heikkonen, Rissanen, and Kaski (2006) that represents a piecewise constant function

[5] Further examples of such tuning parameters include kernel bandwidth, modulator of orthogonal basis decomposition, wavelet shrinkage parameters, etc. See Wasserman (2006).

[6] Equi-depth histograms are naturally linked to k-nearest neighbor density estimators. Ignoring inexact divisibility of the sample size into an integer number of equally-sized bins, the question of the optimal number of bins can be reformulated as the problem of finding the optimal number of neighbors, k. This problem has been studied in the literature.

derived from an expansion within the Haar wavelet basis. This estimator is described in appendix A. In general, wavelet-based density estimators feature local adaptation that imparts an ability to remove noise without compromising sharp details of the original signal. This can be useful to neutralize boundary biases that tend to beleaguer the other two estimators. Unfortunately, sample variability of all three is an endemic concern.

The outcomes of the three binning procedures are discussed in section 3.1.6.

3.1.5. Probability Integral Transform of Density Forecast. An obvious limitation of the foregoing measures is their emphasis on the in-sample fit. This limitation may be an important consideration for risk measurement applications where the predictive ability of different models is a key concern. In order to attend to this requirement, we add an assessment of the out-of-sample performance to our list of measures. The assessment is based on the probability integral transform of the density forecast.

In plain language, the probability integral transform assigns a model-based probability value to an out-of-sample realization. In a perfect world, these p-values are uniformly distributed on the $[0, 1]$ interval and such distributional equivalence can be formally tested.

We put the above schematic into practice as follows. For each alternative functional form, we form an empirical density estimate based on a fixed window of data. The p-value of the first out-of-sample observation following the window can be estimated by its magnitude relative to the in-sample observations. The window is then rolled forward, and the next empirical density estimate is formed. The process is initiated using the earliest sub-sample of the sample under consideration and continues until the entire dataset is exhausted. In line with statistics described in previous sections, we use one year and four year sub-samples over the full sample as well as over the last four years. The distributional equivalence of generated predictive p-values to the standard uniform distribution is tested by means of the Pearson χ^2 statistics already described.

To expand the range of possibilities beyond absolute and proportional changes, we also test distributional equivalence for measures of change of the form $\frac{\Delta r_t}{r_{t-1}^{\gamma}}$ for all $\gamma \in [-1, 2]$ and characterize γ^* that gives rise to the highest p-value. The idea here is to judge closeness of the optimal γ to zero or one as an indication of preference for the difference specification corresponding to $\gamma = 0$ or the return specification corresponding to $\gamma = 1$. Since such procedure does not deliver confidence bounds about the optimal γ and does not constitute a nested test, its results have to be regarded somewhat informally. These results are discussed in section 3.1.6.

3.1.6. Empirical Results. Across the entire 12 year of data, goodness-of-fit measures sharply reject the Student t hypothesis for the unconditional distribution of both absolute and relative changes, indicating difficulty of fitting simple parametric models to financial time-series. In spite

---

Mack and Rosenblatt (1979) showed that in the univariate case the optimal choice involves balancing

$$\operatorname{bias}\left(\hat{f}_{kNN}(x)\right) \propto \frac{f''(x)}{f(x)} \frac{k^2}{T^2} + O\left(\left(\frac{k}{T}\right)^3\right)$$

with

$$\operatorname{var}\left(\hat{f}_{kNN}(x)\right) \propto \frac{\left(f(x)\right)^2}{k} + o\left(\frac{1}{k}\right)$$

for the unknown true density $f(x)$. Using standard reference densities with infinite support to minimize integrated MSE leads to divergent integrals and $k^{opt} \to \infty$. Using reference densities with finite support opens up the quandary of having to specify the support bounds and may produce perplexing results such as those of Orava (2012) where k^{opt} does not depend on any sample statistics and may exceed the sample size. To circumvent this predicament, we gave likelihood cross-validation a go in order to derive k^{opt} and, by implication, B^{opt}.

of the overwhelming rejection, these measures tend to favor the difference form. Similarly, the Akaike information criterion, while silent about any degree of absolute fit, also supports difference formulation for all maturities. In the 4-year sample, returns specification achieves acceptable and fit for maturities over three years. The fit is better than that of the differences for all maturities with the sole exception of the Pearson's χ^2 at 3-year maturity. The AIC reinforces the preference for returns except for 2-year and 30-year maturities. Similar conclusions can be made about the most recent year of data where returns fit well and better than differences at all maturities over three years. For these maturities, the AIC results are also in agreement. For shorter maturities there is a mild disagreement amongst different measures as well as deterioration of the absolute fit.

Histogram binning results are inconclusive, with disagreements between different methods as to the best choice at any given maturity in any of the three subsamples.

As regards predictive p-values, as shown separately in Table 5, results signal mild preference for return specification when using either four year or one year estimation window in the full sample, except for longer maturities and the short estimation window. Using last four years as a testing ground for predictive p-values using one year estimation window is also in favor of return formulation. Optimal γ leads to inconclusive inferences except for the short estimation sub-samples over the most recent four years.

Even with a disagreement across different measures, the widespread divergence of functional form conclusions between the full sample and more recent sub-samples is meaningful since the obvious distinction between the full sample and the more recent ones is that the full sample contains a larger proportion of higher interest rates. The theme of dependence of the functional form choice on the level of interest rates will be revisited several times in the sections that follow.

3.1.7. Critique. Goodness-of-fit measures provide tenuous basis for functional form comparison for a number of reasons. Fitting a parametric *unconditional* distribution to either proportional or absolute change representation amounts to assuming that the sample is drawn i.i.d. from the reference distribution. Emphatic rejections in tests relying on parametric assumptions for both representations provide no way to tell apart failure of parametric fit from the violations of the i.i.d. assumption. Strong rejection of both representations does not instill measurable confidence in the preference for whatever form happened to have a better, yet still failing, measure of fit. And over moderately long horizons sharp rejections of fit to a simple distribution are almost inevitable for most daily financial time-series. At the other extreme, fitting very complex unconditional distributions may result in excellent in-sample fit for both representations but lead to a failure out-of-sample. Fitting complex model will likely produce very different models for the two representation and will make non-nested model comparison even more dubious. Finally, even when the differences in fit are pronounced, the test statistics need to be carefully adjusted for bias when parameters of the reference distribution are themselves estimated, which is not always straightforward.

Although measures introduced in sections 3.1.4 and 3.1.5 are not susceptible to the choice of incorrect parametric distribution, it is of little consolation as their guidance is still inconclusive as both forms are rejected and best-fit elasticities are far away from zero or unity.

3.2. Unit Root Tests. Unit root tests are considered a staple component of a suite of stationarity tests. But for risk measurement purposes they must be treated with caution. The reason for the caution is threefold. First, unit root tests are designed to capture first order non-stationarity. That is,

Swap Maturity	Statistic	Full sample		4-year sample		1-year sample	
		Diff	Ret	Diff	Ret	Diff	Ret
1y	Pearson χ^2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	AIC	11126.37	12049.20	3815.7159	3254.07	816.89	814.99
	Kolmogorov-Smirnov-Lilliefors	0.000010	0.000009	0.000020	0.000044	0.000271	0.000233
	Anderson-Darling	0.000100	0.000100	0.000100	0.002100	0.002700	0.006900
2y	Pearson χ^2	0.000000	0.000000	0.004207	0.004748	0.000000	0.955873
	AIC	10566.59	11438.23	3737.5613	3272.31	847.78	849.78
	Kolmogorov-Smirnov-Lilliefors	0.000065	0.000062	0.000067	0.000139	0.000549	0.000716
	Anderson-Darling	0.000600	0.000100	0.000500	0.019200	0.005000	0.017900
3y	Pearson χ^2	0.000000	0.000000	0.096554	0.000000	0.000000	0.066455
	AIC	10255.72	11152.53	3604.1453	3277.36	883.35	859.21
	Kolmogorov-Smirnov-Lilliefors	0.000189	0.000076	0.000197	0.000342	0.000415	0.000507
	Anderson-Darling	0.001700	0.000200	0.001600	0.066200	0.004000	0.022100
4y	Pearson χ^2	0.000000	0.000000	0.006312	0.410597	0.000000	0.255325
	AIC	10094.63	10893.88	3466.5580	3250.51	877.67	840.97
	Kolmogorov-Smirnov-Lilliefors	0.000307	0.000178	0.000574	0.001245	0.000513	0.021905
	Anderson-Darling	0.002900	0.000500	0.005300	0.266000	0.006800	0.090500
5y	Pearson χ^2	0.000000	0.000319	0.001079	0.157838	0.000000	0.675605
	AIC	10028.64	10726.47	3432.5816	3244.25	859.82	819.45
	Kolmogorov-Smirnov-Lilliefors	0.000339	0.000300	0.001343	0.001958	0.003545	0.070858
	Anderson-Darling	0.003700	0.000800	0.009500	0.236100	0.023900	0.324000
7y	Pearson χ^2	0.000000	0.000001	0.000009	0.000280	0.000004	0.010249
	AIC	9887.32	10549.19	3351.0533	3201.78	818.76	783.84
	Kolmogorov-Smirnov-Lilliefors	0.000296	0.000306	0.001427	0.001274	0.067961	0.103722
	Anderson-Darling	0.003800	0.001200	0.023200	0.119600	0.132000	0.201700
10y	Pearson χ^2	0.000000	0.019104	0.000000	0.021652	0.003218	0.032559
	AIC	9886.23	10520.87	3322.3569	3249.78	812.72	780.64
	Kolmogorov-Smirnov-Lilliefors	0.000258	0.000209	0.001819	0.001953	0.031502	0.151916
	Anderson-Darling	0.003100	0.001200	0.024300	0.121200	0.192800	0.203300
30y	Pearson χ^2	0.000000	0.000326	0.020735	0.137334	0.246047	0.361616
	AIC	9807.73	10573.16	3281.5881	3340.43	810.04	788.14
	Kolmogorov-Smirnov-Lilliefors	0.000250	0.000239	0.001165	0.009653	0.172425	0.353523
	Anderson-Darling	0.002600	0.000800	0.020300	0.081300	0.240100	0.309200

TABLE 1. P-values of Kolmogorov-Smirnov-Lilliefors, Anderson-Darling and Pearson's goodness-of-fit measures under Student t null, and the Akaike Information Criterion corrected for finite sample size.

the tests outcomes provide or deny support for whether the series (already in return or difference form) need to be differenced further. As a consequence, none of the second- or higher-order non-stationarity may be reflected in the test outcomes. Second reason is a technical one. Unit root test tend to have low power against many similar alternatives, especially in small samples. In other words, unit root tests provide a fairly dim lens to study the properties of the data and have hard time distinguishing the null and alternative hypotheses. Third, unit root tests provide no rigorous basis for comparison across different functional form specifications, because the two rival functional form hypotheses are not nested, and distributions of tests statistics may involve, for example, a choice of the number of lags, with the diametrically opposite ranking achieved depending on that number of lags.

Swap Maturity	Full sample				4-year window				1-year Window			
	Diff		Ret		Diff		Ret		Diff		Ret	
	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*
1y	8	-4173.33	11	-4118.90	9	-1490.76	9	-942.40	4	-218.21	5	-151.95
2y	18	-3499.33	12	-4144.51	9	-1264.21	10	-1013.97	5	-176.47	7	-199.70
3y	16	-3207.59	8	-4219.48	16	-1193.41	8	-1077.54	5	-234.69	5	-216.81
4y	15	-3269.55	12	-4226.24	15	-1131.10	13	-1045.31	5	-218.78	7	-196.46
5y	14	-3215.64	13	-3822.86	14	-1075.93	14	-941.95	5	-209.46	7	-180.98
7y	13	-3280.31	10	-3825.31	12	-998.44	10	-908.52	4	-158.73	9	-136.89
10y	14	-3271.28	8	-3823.79	14	-964.79	8	-932.14	13	-129.72	16	-102.39
30y	10	-3096.63	9	-4065.60	10	-875.36	9	-970.32	8	-132.80	6	-138.52

TABLE 2. Optimal number of bins and stochastic complexity of equi-width histogram.

Swap Maturity	Full sample		4-year window		1-year Window	
	Diff	Ret	Diff	Ret	Diff	Ret
1y	77	88	41	32	6	16
2y	70	169	48	54	5	21
3y	69	166	54	61	13	32
4y	75	106	46	73	16	31
5y	71	113	43	77	10	28
7y	78	130	10	67	9	27
10y	13	124	57	109	5	61
30y	10	126	16	98	7	42

TABLE 3. Optimal number of bins in equi-depth histogram.

Swap Maturity	Full sample				4-year window				1-year Window			
	Diff		Ret		Diff		Ret		Diff		Ret	
	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*	B^*	C^*
1y	2	-280973	2	-278845	5	-62818	12	-61587	19	-9152	15	-8981
2y	2	-278206	2	-279672	8	-63139	7	-65637	15	-9005	13	-8998
3y	2	-276422	2	-281704	8	-64198	8	-63622	9	-9510	9	-9222
4y	2	-277025	2	-280594	7	-67178	8	-62767	8	-10401	8	-9290
5y	2	-277456	2	-279889	7	-65063	8	-62183	8	-9985	8	-9275
7y	2	-277980	2	-280820	6	-63943	7	-62120	8	-9209	8	-10082
10y	2	-278486	2	-281558	6	-63707	7	-62517	8	-8642	7	-8833
30y	2	-279198	2	-281644	6	-65008	7	-63372	9	-8531	10	-8645

TABLE 4. Optimal number of retained wavelet coefficients and minimum description length.

These arguments should not be viewed as a denunciation of tests of first order stationarity. For example, it would still make sense to test the series for breaks and regime switches, with the functional form representation experiencing fewer of those being the preferred one. Unfortunately, the development of tests along these lines remains to be formally implemented. The rudimentary attempts at regime modeling involve judgments about the sample period selection, and trying to distinguish long swings of upward drift in the series (e.g. "rising rates regime") from similarly

Swap Maturity	4y subsamples of full sample			1y subsamples of full sample			1y subsamples of 4y sample		
	Diff	Ret	γ^*	Diff	Ret	γ^*	Diff	Ret	γ^*
1y	-468.854	-185.658	0.394	-53.456	-63.221	-0.681	-152.604	-63.696	1.999
2y	-297.041	-45.698	0.394	-37.271	-20.200	-0.860	-83.199	-26.340	2.000
3y	-224.632	-80.888	0.341	-25.102	-24.895	-0.391	-58.186	-25.722	2.000
4y	-98.589	-58.498	0.394	-10.718	-21.337	0.000	-39.993	-12.716	1.996
5y	-77.543	-56.962	0.400	-7.244	-18.513	-0.690	-14.225	-14.806	1.987
7y	-62.625	-72.247	0.403	-9.791	-27.027	-0.538	-7.273	-12.532	1.982
10y	-71.955	-90.385	0.397	-4.505	-14.770	-0.621	-12.840	-1.982	2.000
30y	-114.656	-79.027	0.394	-15.586	-33.904	-0.533	-9.160	-6.877	2.000

TABLE 5. Log p-values of Pearson's χ^2 test for the distributional equivalence of predictive p-values to the standard uniform distribution and best-fit elasticity of volatility.

persistent declines ("declining rates regime"). Still, the focus on volatility regimes has to dominate once the issues of mean trend breaks and mean regime switches are addressed.

Out of curiosity, we ran KPSS tests (Kwiatkowski, Phillips, Schmidt, and Shin, 1992) (with quadratic spectral kernel) on returns and differences of all interest rate swap series under study with results reported in Table 6. The maximum lag order (bandwidth) was derived from an automatic bandwidth selection procedure recommended in Hobijn, Franses, and Ooms (2004). We also used Monte Carlo simulations in order to expand the range of critical values tabulated in Kwiatkowski, Phillips, Schmidt, and Shin (1992). Expanded range is shown in appendix B.

Except for the shortest maturities over the entire 12 years of data, unit root is not a problem for either yield returns or yield differences. Proportional change representation tends to perform better with only exceptions being 30-year maturity over the full sample and up to 2-year maturities over 1-year sample while the lag choice does not seem to matter for our dataset. Even so, the difference specification is competitive for most maturities and subsamples and we cannot tell whether differences in test p-values are sufficiently significant to express confident preference in favor of returns.

3.3. Homoscedasticity Tests.

Testing for second order stationarity means to assess equality of the second moments over time. As second moments are not observed, they must be estimated by either linking them to observables in a time-series model specification or by partitioning the time domain into various subsets. These subsets can be either identified directly by defining a method to group observations over time, or indirectly through the values of another pre-defined explanatory variable.

The tests below give examples of all three methods and are not meant to be exhaustive.

3.3.1. Levene Equality of Variances Test.

Generic Levene F-test (Levene, 1960), in its original formulation or a more robust version due to Brown and Forsythe (1974), is used to assess equality of variances across different samples. Associating different subsamples of returns or difference series with different time windows leads to what we call *time heteroscedasticity* test. Associating instead with different values of the level series produces *level heteroscedasticity* test. The latter formulation is also known as Goldberg-Quandt parametric test (Goldfeld and Quandt, 1965). The test results for both formulations are shown in Table 7. In both cases, the two groups of

Swap Maturity	Form	Lags							Optimal
		1	2	3	4	5	6	10	Lags
Full sample									
	Diff	0.000004	0.000005	0.000006	0.000006	0.000007	0.000007	0.000021	6
1y	Ret	0.022855	0.022970	0.022140	0.021390	0.020469	0.020000	0.022658	6
	Diff	0.002513	0.002480	0.002353	0.002196	0.002005	0.001861	0.002067	6
2y	Ret	0.185120	0.177170	0.173270	0.171170	0.166110	0.160540	0.160360	6
	Diff	0.023341	0.022311	0.021280	0.020253	0.018923	0.017837	0.018358	6
3y	Ret	0.262780	0.248710	0.244560	0.245530	0.243920	0.238410	0.239120	6
	Diff	0.071480	0.068252	0.064374	0.060721	0.056459	0.052660	0.052349	6
4y	Ret	0.272020	0.260550	0.256170	0.254560	0.251100	0.244730	0.240280	6
	Diff	0.149600	0.141800	0.133340	0.126420	0.118470	0.111700	0.107100	6
5y	Ret	0.296190	0.281290	0.274890	0.270480	0.263990	0.255470	0.247270	6
	Diff	0.259610	0.248780	0.238810	0.230970	0.221950	0.214330	0.207140	6
7y	Ret	0.340640	0.320980	0.309010	0.300250	0.292040	0.283320	0.274930	6
	Diff	0.404290	0.391740	0.380620	0.370270	0.359130	0.350270	0.338520	6
10y	Ret	0.442310	0.424370	0.409900	0.397050	0.383700	0.370720	0.356280	6
	Diff	0.636610	0.618300	0.600110	0.587000	0.574840	0.566790	0.558070	6
30y	Ret	0.633820	0.612030	0.589770	0.574020	0.558080	0.543560	0.537250	6
4-year sample									
	Diff	0.106150	0.096609	0.093189	0.089152	0.089297	0.091301	0.100670	5
1y	Ret	0.559710	0.550930	0.536670	0.521470	0.505360	0.495610	0.512000	5
	Diff	0.310050	0.282250	0.278050	0.263840	0.238230	0.215930	0.198030	5
2y	Ret	0.784610	0.767440	0.764380	0.762490	0.754840	0.745630	0.738970	5
	Diff	0.433720	0.384360	0.380320	0.372620	0.349660	0.325590	0.301490	6
3y	Ret	0.658480	0.632390	0.631590	0.638660	0.640920	0.634400	0.627840	5
	Diff	0.481650	0.442810	0.441050	0.433380	0.409860	0.388650	0.366460	4
4y	Ret	0.572660	0.548970	0.547310	0.549080	0.545300	0.535730	0.518300	6
	Diff	0.491810	0.447770	0.449670	0.443930	0.421420	0.399820	0.380960	5
5y	Ret	0.544080	0.510180	0.508030	0.505200	0.495120	0.481760	0.462120	5
	Diff	0.454350	0.410860	0.413950	0.411870	0.394080	0.378740	0.362910	5
7y	Ret	0.509610	0.476650	0.468680	0.461060	0.444650	0.426490	0.404860	5
	Diff	0.398110	0.367960	0.370070	0.367690	0.351180	0.335910	0.318360	5
10y	Ret	0.484390	0.457910	0.447700	0.435960	0.413050	0.392250	0.371410	5
	Diff	0.228330	0.199130	0.194340	0.190380	0.181540	0.174000	0.173110	5
30y	Ret	0.354730	0.326360	0.304760	0.292370	0.277430	0.262240	0.260610	5
1-year sample									
	Diff	0.612150	0.636460	0.660250	0.674800	0.689680	0.700940	0.670450	6
1y	Ret	0.645210	0.646360	0.652970	0.657940	0.660850	0.662580	0.627940	6
	Diff	0.469600	0.501550	0.536470	0.592200	0.634880	0.657010	0.654730	6
2y	Ret	0.449260	0.482880	0.506310	0.564860	0.601410	0.621170	0.612360	6
	Diff	0.192960	0.220940	0.244280	0.287430	0.324370	0.342920	0.360930	6
3y	Ret	0.217210	0.251330	0.273840	0.315770	0.352810	0.369950	0.374680	6
	Diff	0.147490	0.169660	0.180340	0.198460	0.213110	0.221120	0.229250	6
4y	Ret	0.177080	0.197160	0.214990	0.237860	0.247980	0.251890	0.241120	6
	Diff	0.140960	0.154930	0.163260	0.170870	0.172320	0.170560	0.170570	6
5y	Ret	0.182340	0.194050	0.201480	0.207270	0.202750	0.196580	0.182530	6
	Diff	0.132870	0.139680	0.136490	0.126790	0.116860	0.108880	0.099978	6
7y	Ret	0.192050	0.196650	0.189540	0.175500	0.159900	0.146240	0.120090	6
	Diff	0.124180	0.128480	0.117460	0.097393	0.085696	0.076800	0.065579	6
10y	Ret	0.189430	0.193740	0.179270	0.153810	0.124340	0.099253	0.080132	6
	Diff	0.102830	0.097702	0.081707	0.062923	0.048584	0.040520	0.029107	6
30y	Ret	0.178200	0.173990	0.142680	0.105900	0.086024	0.066859	0.044128	6

TABLE 6. P-values of KPSS unit root test.

Swap Maturity	Form	Time Heteroscedasticity: Full sample F stat	Full sample p-value	4-year sample F stat	4-year sample p-value	1-year sample F stat	1-year sample p-value
1y	Diff	109.05	0.000000	82.98	0.000000	0.00	0.000000
	Ret	206.77	0.000000	1.18	0.278357	1.36	0.243887
2y	Diff	151.77	0.000000	144.66	0.000000	0.00	0.000000
	Ret	94.71	0.000000	1.30	0.253920	1.53	0.217077
3y	Diff	64.93	0.000000	229.74	0.000000	27.49	0.000000
	Ret	131.14	0.000000	3.28	0.070474	13.87	0.000242
4y	Diff	75.45	0.000000	101.27	0.000000	16.88	0.000054
	Ret	151.45	0.000000	0.08	0.227002	11.85	0.000674
5y	Diff	57.92	0.000000	88.75	0.000000	14.22	0.000203
	Ret	128.17	0.000000	0.09	0.237606	2.91	0.089530
7y	Diff	0.00	0.000000	32.58	0.000000	22.11	0.000004
	Ret	138.63	0.000000	0.24	0.374037	13.45	0.000299
10y	Diff	0.00	0.000000	51.15	0.000000	11.55	0.000786
	Ret	106.79	0.000000	1.09	0.297518	5.69	0.017791
30y	Diff	68.09	0.000000	53.40	0.000000	12.13	0.000586
	Ret	177.00	0.000000	7.05	0.008047	0.97	0.325649
		Level Heteroscedasticity					
1y	Diff	213.29	0.000000	162.69	0.000000	0.00	0.000000
	Ret	871.32	0.000000	2.29	0.130461	2.31	0.129630
2y	Diff	127.95	0.000000	144.58	0.000000	0.00	0.000000
	Ret	454.80	0.000000	1.94	0.163749	1.04	0.309264
3y	Diff	0.00	0.000000	237.59	0.000000	23.89	0.000002
	Ret	515.82	0.000000	0.66	0.417222	4.35	0.037948
4y	Diff	0.00	0.000000	70.80	0.000000	4.45	0.035987
	Ret	502.13	0.000000	13.36	0.000271	0.01	0.095879
5y	Diff	0.00	0.000000	100.17	0.000000	3.52	0.061665
	Ret	439.28	0.000000	7.65	0.005788	0.08	0.224863
7y	Diff	73.17	0.000000	39.92	0.000000	10.93	0.001087
	Ret	390.65	0.000000	27.01	0.000000	0.00	0.047020
10y	Diff	69.04	0.000000	0.00	0.000000	9.19	0.002688
	Ret	296.53	0.000000	35.43	0.000000	0.09	0.235839
30y	Diff	73.92	0.000000	0.00	0.000000	1.45	0.230323
	Ret	261.77	0.000000	43.39	0.000000	0.33	0.435341

Table 7. Levene-Brown-Forsythe and Goldfeld-Quandt test results.

observations are the first and the last third of the sample in order to improve the power of the test.

Homogeneity of variance between first and last four years of the full 12 year sample is unequivocally denied by the test, while homogeneity between observations corresponding to high interest rates and observations corresponding to low rates is rejected even more categorically. Returns tend to do worse by both methods. In the shorter subsamples, the evidence is less decisive with relative change formulation not rejected by the first test for most of maturities in the four year sample as well as by the second test for short maturities in the four year sample and most maturities in the one year sample. Even when the return specification is rejected, the rejection is less pronounced than that for the difference alternative.

3.3.2. Non-parametric Goldfeld-Quandt Test. Levene-Brown-Forsythe and parametric Goldfeld-Quandt tests, though reasonably robust, may still be sensitive to non-normality, especially under conditions where samples are collected from population distributions that are skewed (Nordstokke and Zumbo, 2010). They also rely on asymptotic large-sample approximations. The non-parametric Goldfeld-Quandt test (Goldfeld and Quandt, 1965) is an exact test robust to non-normality. In this test the absolute values are sorted in ascending order according to pre-defined explanatory "deflator" variable , and the test statistics is the number of observations that exceed all previous observations ("peak" observations)

Test results are shown in Table 8.[7] Low power of the test is evident in fairly inconsistent conclusions between the case when the explanatory variable is time and the case when the level of the series serves as a covariate, as well as in a number of ties. Using time as the sorting variable, the test favors the difference specification over the full sample and its four year subsample but is inconclusive in the one year sample. Using level as the sorting variable, the test generally favors the return specification, albeit with a number of exceptions.

The weakness of this test is the requirement to order absolute residuals according to a known explanatory variable. If the heteroscedasticity structure depends on the unknown or latent variable, the test offers little guidance. Another major shortcoming is that the variance must be a monotonic function of explanatory variable. For example, when faced with a quadratic function mapping a deflator to the variance, the test might improperly accept the null of homoscedasticity. We suspect that relatively high p-values in Table 8 can be explained by such non-monotone behavior. Lastly, the test only allows a single sorting covariate.

3.3.3. Information Matrix Test. Variance homogeneity tests in sections 3.3.1 and 3.3.2 point to the association of volatility with the level of the series. A regression context might be useful to provide further details. Information matrix test (Cameron and Trivedi, 1990) may be particularly illuminating since it could be decomposed into tests for conditional heteroscedasticity, conditional non-normal skewness and conditional non-normal kurtosis via three auxiliary regressions involving powers of residuals against potential drivers of heteroscedasticity (e.g., level of the series).[8]

We performed information matrix tests using lagged level as an explanatory factor and displayed the test statistics in Table 9. To conserve space, only test statistics are displayed.[9] Overall, it appears that conditional skewness is not an issue and conditional kurtosis is only moderately severe. On the other hand, conditional heteroscedasticity is very strong in the full sample and remains elevated in the shorter subsamples. Ranking the two alternatives according to the IM_2 statistics, we observe again that in the full sample the difference form is preferred, while for the 1-year sample it is returns that display lower heteroscedasticity. Over the 4-year sample, returns also dominate except for 30 year rates.

Comparing with the earlier heteroscedasticity tests we find complete consensus for 5- and 7-year maturities in both recent subsamples and for 30-year maturity in 1-year subsample. Notable disagreement among different tests arises for 2-year and 10-year maturities in the long

[7] Exact p-values were obtained from $P\,(Number\ of\ peaks \le k) = \frac{1}{T!}\sum_{i=1}^{k+1}|S_T^{(i)}|$ where $S_T^{(i)}$ are Stirling numbers of the first kind (Abramowitz and Stegun, 1965).

[8] Specifically, $IM = IM_2 + IM_3 + IM_4 = \sum_{i=2}^{4} TR^2_{un,i}$, where $R^2_{un,2}$ is uncentered R^2 from a regression of squared residuals against a full quadratic form in explanatory variable(s), $R^2_{un,3}$ comes from a regression of a cubic function of residuals against a full linear form in explanatory variables, and $R^2_{un,4}$ is from a regression of a quartic function of residuals against a constant term. With a single explanatory factor each IM_i is asymptotically χ^2_1.

[9] For comparison, χ^2_1 critical values are 3.841 at 5%, 6.635 at 1% and 10.827 at 0.1%.

Swap Maturity	Form	Time Heteroscedasticity					
		Full sample		4-year sample		1-year sample	
		Peaks	p-value	Peaks	p-value	Peaks	p-value
	Diff	4	0.1142	6	0.4743	5	0.4031
1y	Ret	13	0.0188	9	0.1102	4	0.4093
	Diff	3	0.0479	8	0.1967	4	0.4093
2y	Ret	13	0.0188	10	0.0565	4	0.4093
	Diff	5	0.2195	6	0.4743	4	0.4093
3y	Ret	15	0.0035	12	0.0115	5	0.4031
	Diff	6	0.3570	6	0.4743	3	0.2305
4y	Ret	12	0.0391	8	0.1967	5	0.4031
	Diff	9	0.2258	7	0.3202	3	0.2305
5y	Ret	11	0.0756	6	0.4743	3	0.2305
	Diff	14	0.0084	9	0.1102	4	0.4093
7y	Ret	15	0.0035	8	0.1967	4	0.4093
	Diff	12	0.0391	7	0.3202	4	0.4093
10y	Ret	12	0.0391	8	0.1967	5	0.4031
	Diff	10	0.1358	5	0.3600	4	0.4093
30y	Ret	13	0.0188	6	0.4743	4	0.4093
		Level Heteroscedasticity					
	Diff	22	0.0000	20	0.0000	4	0.4093
1y	Ret	9	0.2258	9	0.1102	3	0.2305
	Diff	8	0.3467	6	0.4743	3	0.2305
2y	Ret	4	0.1142	4	0.2103	2	0.0986
	Diff	15	0.0035	14	0.0017	11	0.0037
3y	Ret	8	0.3467	8	0.1967	8	0.0627
	Diff	12	0.0391	12	0.0115	9	0.0269
4y	Ret	9	0.2258	9	0.1102	9	0.0269
	Diff	14	0.0084	14	0.0017	11	0.0037
5y	Ret	9	0.2258	10	0.0565	9	0.0269
	Diff	10	0.1358	10	0.0565	9	0.0269
7y	Ret	9	0.2258	9	0.1102	8	0.0627
	Diff	13	0.0188	13	0.0046	9	0.0269
10y	Ret	10	0.1358	11	0.0266	9	0.0269
	Diff	6	0.3570	6	0.4743	10	0.0104
30y	Ret	5	0.2195	5	0.3600	8	0.0627

TABLE 8. Goldfeld-Quandt peaks test results.

sample, 10-year and 30-year maturities in the 4-year sample and at the short end of the curve in the short sample. Lack of perfect agreement across tests is likely driven by inability of these tests to detect non-monotone variation in volatility, especially as different tests fail detection in different ways. We will revisit non-monotone variation in volatility relative to the level of the series in the subsequent sections.

3.3.4. Engle's ARCH Test. Engle's Lagrange multiplier test (Engle, 1982) without covariates looks for autoregressive patterns in the two bottom panels of Figure 2. Table 10 shows the results when the assumed autocorrelation is of the first order. In the full sample, the null of no autoregressive conditional heteroscedasticity is soundly defeated albeit less so for returns at most maturities. In the four year sample, returns are again preferred and the null is no longer rejected for some

Swap Maturity	Form	Full sample				4-year sample				1-year sample			
		IM	$\mathbf{IM}_2$	$\mathbf{IM}_3$	$\mathbf{IM}_4$	IM	$\mathbf{IM}_2$	$\mathbf{IM}_3$	$\mathbf{IM}_4$	IM	$\mathbf{IM}_2$	$\mathbf{IM}_3$	$\mathbf{IM}_4$
	Diff	118.16	97.84	6.78	13.54	134.04	127.12	0.37	6.55	20.92	14.31	5.11	1.50
1y	Ret	196.07	177.81	6.97	11.29	11.48	2.41	3.90	5.18	22.12	13.25	5.12	3.75
	Diff	148.33	123.50	2.05	22.78	110.47	101.91	0.58	7.99	11.25	5.29	3.50	2.46
2y	Ret	179.83	166.39	2.99	10.45	7.33	0.80	0.82	5.71	5.21	1.98	1.45	1.78
	Diff	137.18	106.85	1.06	29.27	90.79	80.93	0.09	9.77	7.20	3.66	2.33	1.21
3y	Ret	163.41	157.43	0.03	5.96	5.41	1.53	0.19	3.69	4.34	1.00	2.14	1.20
	Diff	109.20	83.14	0.25	25.81	71.07	60.84	0.10	10.14	14.47	9.11	4.10	1.27
4y	Ret	179.19	171.21	0.24	7.74	6.82	1.91	0.30	4.62	9.51	4.14	3.92	1.44
	Diff	91.32	70.45	0.05	20.82	53.92	45.02	0.42	8.48	19.00	12.45	5.10	1.45
5y	Ret	180.19	169.91	0.88	9.40	7.96	1.88	0.71	5.36	11.47	5.36	4.33	1.77
	Diff	56.95	44.31	1.09	11.55	34.98	27.56	1.71	5.71	18.33	11.12	5.67	1.54
7y	Ret	164.44	157.10	2.42	4.92	6.66	2.42	1.43	2.81	9.16	3.21	4.21	1.73
	Diff	37.23	25.03	3.36	8.84	19.42	12.40	2.29	4.73	18.81	9.13	5.45	4.23
10y	Ret	147.72	140.47	3.60	3.64	8.85	4.78	1.68	2.39	9.61	2.24	2.79	4.58
	Diff	61.72	49.30	6.00	6.42	11.99	5.58	2.45	3.97	9.92	5.72	2.41	1.79
30y	Ret	226.77	215.30	5.56	5.92	31.86	25.12	2.51	4.23	5.15	2.20	1.55	1.40

TABLE 9. Information matrix tests and their Cameron-Trivedi decomposition.

maturities. In the one year sample, both representations tend not to reveal much in a way of autoregressive patterns in squared changes making the risk form selection ambiguous.

A particular weakness of ARCH model is the choice of the number of lags. Using model selection criteria such as AICc typically results in excessively large number of lags. For our dataset AICc-selected number of lags for different series varies between 30 and 54 using the full sample, 11 and 18 using the four year sample, and settles at 8 for all series using the one year sample. Although not shown to conserve space, test comparisons using different number of lags (either the same for all series or using optimal number for each series) often contradict the results of Table 10, rendering the choice of form erratic, arbitrary and not convincing. We do not recommend this test for these reasons.

3.4. Mutual Information.

Homoscedasticity tests in section 3.3 are, in effect, attempts to discern patterns in volatility of daily changes as a function of the preceding level of the series. Although evolving volatility is not directly observable, it could be estimated or, simpler yet, proxied by the squared daily changes. Correlation between the squared daily changes and the preceding levels can then be used as a simple indicator whether a linear pattern exists. The risk representation with minimal such correlation exhibits weaker relationship and would be preferred. The problem is that correlation would only capture linear comovements.

A more general approach to correlation is to determine how similar the joint distribution of levels and changes is to the factored distribution of the two variables by means of their Kullback-Leibler divergence, also known as the *mutual information*. For two random variables X and Y, this quantity is defined as:

$$MI(X,Y) := \iint_{\mathbb{X}\times\mathbb{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y), \tag{3.4}$$

where $\mathcal{H}(X)$ and $\mathcal{H}(Y)$ are the marginal entropies and $\mathcal{H}(X,Y)$ is the joint entropy. It can be interpreted as the reduction in uncertainty about X after observing Y. Since for correlated

Swap Maturity	Form	Full sample		4-year sample		1-year sample	
		Test stat	p-value	Test stat	p-value	Test stat	p-value
1y	Diff	75.90	0.000000	9.62	0.001925	0.01	0.926370
	Ret	38.70	0.000000	2.35	0.124972	0.08	0.783083
2y	Diff	75.04	0.000000	19.35	0.000011	0.70	0.403496
	Ret	48.90	0.000000	6.13	0.013289	7.33	0.006777
3y	Diff	66.77	0.000000	18.62	0.000016	2.02	0.155257
	Ret	49.31	0.000000	5.82	0.015812	3.46	0.062767
4y	Diff	58.21	0.000000	13.40	0.000252	3.36	0.066871
	Ret	58.40	0.000000	6.09	0.013562	3.02	0.082082
5y	Diff	49.58	0.000000	8.91	0.002838	1.19	0.276108
	Ret	36.91	0.000000	1.36	0.243151	0.30	0.581721
7y	Diff	44.03	0.000000	8.60	0.003354	0.15	0.697681
	Ret	35.37	0.000000	1.59	0.207762	0.07	0.793606
10y	Diff	38.07	0.000000	7.14	0.007524	0.11	0.742755
	Ret	31.42	0.000000	2.24	0.134467	0.54	0.464427
30y	Diff	70.34	0.000000	18.18	0.000020	0.00	0.992551
	Ret	93.56	0.000000	16.68	0.000044	0.01	0.927006

TABLE 10. Engle's ARCH LM test results.

Gaussian variables, $MI(X,Y) = -\log\left(1-\rho^2\right)/2$, one can convert $MI(X,Y)$ into a correlation measure that is equivalent to Gaussian (linear) correlation in terms of its information content,

$$\rho_{MI}\,(X,Y) = \sqrt{1-e^{-2MI(X,Y)}}. \tag{3.5}$$

An advantage of the above transformation is a more intuitive interpretation as it ranges from 0 to 1 as we go from independence to complete dependence. As a nonlinear measure, mutual information takes into account the whole dependence structure and captures both linear and nonlinear associations between random variables. Although mutual information is considered very powerful, its estimation is a long-standing difficult problem. As the joint and marginal distributions in the above definition may indeed be quite general, most approaches fall into the non-parametric class. These include histogram-based methods, splines, kernel density, nearest neighbor, polynomial density expansions, empirical processes and even wavelets (Beirlant, Dudewicz, Györfi, and van der Meulen, 1997; Walters-Williams and Li, 2009). Histogram methods are most commonly used. These methods bin observation into ranges, so that the densities in (3.4) are approximated by various piecewise constant functions. Unfortunately, the number of bins used, and the location of the bin edges can have a significant impact on the results. While some ways to sidestep density estimation have been suggested, such as methods based on *order statistics* (Vašíček, 1976; Learned-Miller, 2004), or by using so-called *entropic spanning graphs* (Hero, Ma, Michel, and Gorman, 2002), removal of bias at the boundary tends to be pestilent. Another approach is to try many different bin sizes and locations and to record the maximum mutual information attained, appropriately normalized as in the maximal information coefficient (MIC) of Reshef, Reshef, Funicane, Grossman, McVean, Turnbaugh, Lander, Mitzenmacher, and Sabeti (2011):

$$MIC\,(X,Y) = \max_{n_x, n_y,\; n_x n_y \le B} \frac{\max\limits_{G\in\mathbb{G}(n_x,n_y)} MI\,(X_{\mathbb{G}}, Y_{\mathbb{G}})}{\log\min(n_x,n_y)}, \tag{3.6}$$

Swap Maturity	Form	Full sample		4-year sample		1-year sample	
		ρ_{MI}	MIC	ρ_{MI}	MIC	ρ_{MI}	MIC
1y	Diff	0.7920	0.4934	0.6360	0.2592	0.4082	0.0912
	Ret	0.7466	0.4075	0.7531	0.4186	0.8509	0.6436
	γ^*	-0.3105		0.9377		1.5374	
2y	Diff	0.7506	0.4144	0.6381	0.2614	0.4585	0.1180
	Ret	0.6955	0.3305	0.6219	0.2445	0.8278	0.5781
	γ^*	-0.6002		0.7571		1.8879	
3y	Diff	0.7271	0.3761	0.5644	0.1918	0.4202	0.0971
	Ret	0.6723	0.3007	0.5973	0.2206	0.8327	0.5912
	γ^*	-0.6868		0.6704		1.7305	
4y	Diff	0.7024	0.3399	0.7128	0.3547	0.5114	0.1516
	Ret	0.6764	0.3057	0.6173	0.2398	0.8390	0.6087
	γ^*	0.0000		0.8496		1.5568	
5y	Diff	0.7506	0.4144	0.6487	0.2731	0.4816	0.1319
	Ret	0.6522	0.2770	0.6075	0.2302	0.8064	0.5254
	γ^*	-0.9867		0.9031		0.0000	
7y	Diff	0.7065	0.3458	0.5859	0.2103	0.5134	0.1529
	Ret	0.6880	0.3206	0.6033	0.2263	0.7483	0.4104
	γ^*	-0.9858		1.2227		1.5503	
10y	Diff	0.6693	0.2970	0.4189	0.0965	0.6055	0.2284
	Ret	0.7035	0.3416	0.6586	0.2844	0.7137	0.3560
	γ^*	-0.9443		1.3321		1.7890	
30y	Diff	0.5901	0.2140	0.6434	0.2672	0.4656	0.1222
	Ret	0.7453	0.4054	0.7163	0.3598	0.7279	0.3773
	γ^*	0.0000		1.7090		1.9788	

TABLE 11. Mutual information statistics against levels.

where $\mathbb{G}\left(n_x, n_y\right)$ is the set of two-dimensional grids of size $n_x \times n_y$ covering sample support $\widehat{\mathbb{X}} \times \widehat{\mathbb{Y}}$, $X_{\mathbb{G}}$ and $Y_{\mathbb{G}}$ are discretizations of X and Y onto this grid, while $B \propto T^{0.6}$ is an upper bound on the number of bins. It can be shown that the MIC lies between 0 and 1, where 0 represents no relationship between the variables and 1 represents a noise-free relationship of any form, not just linear.

Table 11 gives an example of ρ_{MI} and MIC statistics enacted on our dataset. The risk representation with lower ρ_{MI} or lower MIC is preferred as it indicates that level effects are captured already. Also included is the elasticity of volatility parameter γ that gives the lowest mutual information. According to the table, returns and differences are competitive in the full sample except at longer maturities, where differences win. Differences add additional tenors to their tally using 4-year sample, and completely dominate over the 1-year sample. These results are rather at odds with those of the earlier sections, although it should be pointed out that the MIC-minimal elasticity of volatility seem to be located significantly above unity. These findings are hard to interpret and we only conjecture that ρ_{MI} and MIC estimates are unduly influenced by boundary effects due to the limited resolution of the interest rate grid near its zero boundary and associated "discrete-like" moves in that region.

3.5. Discussion. Difficulties reconciling results of testing various hypotheses across the two rival specifications stem from misconstruing principles of model selection and of hypothesis testing. The model selection process aims to *choose* one of the models under consideration for a particular purpose with a specific utility function in mind (Pesaran and Weeks, 1999). It treats all models symmetrically and should end in a definitive outcome. Hypothesis testing, on the other hand, attaches a different status to the null and to the alternative hypotheses, treating them asymmetrically. By design, it looks for evidence of departure from the null hypothesis in the direction of alternative hypotheses. Rejection of the null hypothesis does not necessarily imply acceptance of any of the alternatives. Hypothesis testing is better suited to inferential problems where the empirical validity of a theoretical prediction is the main goal.

Because of the unbalanced treatment of available models in the hypothesis testing approach, the choice of the null hypothesis plays a pivotal role. In the case of non-nested models, particularly globally non-nested models, there is no natural null. Carrying out the analysis with different models treated as the null can lead to ambiguities, as we have just observed.[10]

Furthermore, the simple comparison approaches applied above are oblivious of the careful research work in the area of non-nested hypothesis testing. That literature, starting with the pioneering work of Cox (1961, 1962), identified three principal approaches.

(1) Modified log-likelihood ratio procedure. From statistical point of view, the usual log-likelihood ratio or Wald statistics are not centered at zero under the null when the hypotheses under consideration are not nested. For example, in the likelihood ratio test, the degrees of freedom of the chi-square statistics is equal to the reduction in the dimension of the parameter space after imposing the necessary set of zero restrictions. When neither of the two models nests the other model, the attendant parameter spaces are unrelated. Cox's 1961; 1962 contribution was to note that this problem can be overcome if a consistent estimate of expected log-likelihood ratio under the null can be obtained. Closed form expressions for the Cox test are only available for a comparatively small number of special cases. Simulation-based approximations and bootstrap-based procedures have been used to adjust the distribution of test statistics. These adjustments need to be tailored to specific features of non-nested tests and to achieve desirable size and power properties.

(2) Comprehensive model approach. This method relies upon a third comprehensive model, artificially constructed so that each of the non-nested models is a special case. The abundance of ways in which a comprehensive model can be constructed is intimidating with widely explored examples, such as the exponential mixture, subject to important limitations concerning disappearance of coefficients of the alternative model under the null, problems with identification and unfavorable consequences for the power of the test. Fortunately, a very natural comprehensive model exists and will be explored in the sequel.

(2) Encompassing procedure. This approach attempts to directly test ability of one model to explain one or more features of the rival model. The Wald and Score Encompassing

[10] Pesaran and Weeks (1999) point out that ambiguity is objectionable only if the primary objective is to *select* a specific model for forecasting or decision making, but not if the goal is to evaluate the comparative strengths and weaknesses of opposing explanations. Ability to assess strengths and weaknesses is an asset for statistical inference and model building. For example, a rejection of both models can point to developing a third model incorporating the desirable features of the original, as well as being theoretically meaningful.

> Tests (WET and SET) are typically constructed under the assumption that one of the rival models is correct, although in some cases this assumption had been relaxed. The implementation of these tests tends to face several difficult hurdles such as derivation of the binding function linking parameters of different models, general non-invertibility of the covariance matrices involved in the test, etc. As a recent example of this approach, Giacomini and Komunjer (2005) propose a method of comparing and combining conditional quantile forecasts based on out-of-sample framework and on the principle of encompassing. While promising, conclusions of their methods varied depending on the tail quantile.

More integrated approach to non-nested hypothesis testing and model selection requires a more formal definition. Non-nested hypothesis testing literature addressed this by means of a variety of "closeness" criteria for measuring the divergence of one distribution function with respect to another, similar to information measures in section 3.1. Given a measure of closeness, a model is strictly nested within another model if and only if the divergence of the second model with respect the first model is zero throughout the entire parameter space of the first model but is non-zero for some parameter configuration of the second model. Globally non-nested models obtain if both divergence of the second model with respect to the first and the divergence of the first model with respect to the second are non-zero for both parameter spaces. Partial non-nesting corresponds to the case when divergences in both directions are non-zero for some parameters in both spaces. Lastly, the two models are observationally equivalent if divergences in both directions are zero for all values of parameters. Using these definitions, it is clear that absolute and proportional specification are globally non-nested when formulated as in the preceding subsections. Closeness measures can also be defined from the perspective of the true model. Vuong's (1989) approach takes up this idea and proposes a probabilistic approach to model selection based on testing the hypothesis that the models under consideration are equally close to the true model against the hypothesis that one model is closer than another. Since in our case, both functional specifications can be nested so that they correspond to different value of the same parameter, differences in log-likelihood from the unrestricted estimate serve similar purpose of making a probabilistic decision of selecting the best (i.e. the closest) model. For this reason, we do not need to pursue Vuong's (1989) approach formally.

4. Constant Elasticity of Volatility Model

The analysis of previous section highlighted the need to attend to the seemingly non-nested character of the two rival hypotheses to gain firmer statistical footing. Fortunately, both functional specifications can be viewed as special cases of modeling of $\frac{\Delta r_t}{r_{t-1}^{\gamma}}$, where $\gamma = 0$ corresponds to modeling absolute change and $\gamma = 1$ corresponds to modeling proportional change. This brings the constant elasticity of volatility model to the forefront.

4.1. The Model.

We now assume that upon the time-series of the risk factor of interest follows discrete-time version of constant elasticity of volatility model

$$\Delta r_t = \mu + \sigma r_{t-1}^{\gamma} \epsilon_t, \tag{4.1}$$

where σ and γ denote the scale and elasticity of volatility.[11] The CEV model in (4.1) has been popular and useful in describing the dynamics of the short-term interest rate[12] (hence use of r_t notation) and in models of equity volatility (under risk-neutral measure).

Mean drift parameter μ will be shown to be quite small and could be safely omitted. The model also ignores the possibility of the mean reversion since accounting for it is notoriously difficult, at least in the interest rate context. Although monetary policy interventions by the central banks may induce occasional relatively large jumps in the policy rates, the above formulation attributes all interest rate moves to continuous shocks instead of jumps. We do so because jumps in the U.S. interest rates outside of the federal funds market are hard to separately identify, especially as the policy rates have been exceptionally low and stable throughout most of the sample period.[13] Our dataset may be partially immune to the problem due to the focus on rates with at least one year maturity.

4.2. Limited Information Approach: GMM. Generalized Method of Moments (GMM) is an econometric procedure for estimating the parameters of the model by choosing parameters so as to match select few moments of the model to those of the data as closely as possible, with a weighting matrix determining the relative importance of matching each moment. It is important to realize the generality of the GMM. Without providing additional details, saying "we estimated the parameters by GMM" is essentially the same as saying "we estimated the model on a computer". Indeed, a large variety of estimation procedures, including OLS, instrumental variables estimation, two-stage least squares can be phrased as GMM procedures.

For a parametric model described by a parameter vector θ, the art of the GMM consists in selecting moments to match and selecting a weighting matrix. The mechanics of the GMM is then to minimize a quadratic distance measure

$$\min_{\theta} J_T(\theta) = \min_{\theta} m(\theta)' W m(\theta), \tag{4.2}$$

where $m(\theta)$ is a vector of L moment conditions and W is a $L \times L$ positive definite weighting matrix. A key advantage of the GMM over other estimation procedures is the weakness of statistical assumptions needed.

In the CEV context, we specify the following six moments

$$m_t = \begin{bmatrix} r_t - r_{t-1} - \mu \\ (r_t - r_{t-1} - \mu)^2 - \sigma^2 r_{t-1}^{2\gamma} \end{bmatrix} \otimes \mathbf{z}_t \tag{4.3}$$

to match, where $\mathbf{z}_t = (1, r_t, r_t^2)$ is a vector of instruments. The first two imply matching of the first and second conditional moments. The other four are orthogonality conditions, forcing the residual terms to be uncorrelated with regression terms. Obviously one can extend the vector of instruments to any number of functions of the sample, and the fact that only a limited number is used is the reason why this approach is of limited information - information in higher

[11] In the literature, the model is routinely called *the constant elasticity of variance model*, as in Chan, Karolyi, Longstaff, and Sanders (1992) or Chan, Choy, and Lee (2007), since elasticity of variance is 2γ. We prefer volatility nomenclature as it links more naturally with the problem at hand.

[12] Nominal interest rates cannot follow unrestricted random walk over very long horizons since they are bounded above and below. However, the persistence pattern over (relatively) short horizons is better approximated in this way, in contrast to modeling levels of rates themselves as stationary variables.

[13] Additionally, as jumps are not observed, the estimated jump model would prevent using historical simulation in order to compute various risk measures. Jump model is likely to be more useful for pricing and best applied to instantaneous forward rates.

order cross-moments is simply ignored. The loss of information leads to loss of efficiency, which manifests as wider confidence bounds for the estimates. On the flip side, possibly misspecified assumptions are not necessary. For example, the assumption of normally distributed error terms is unnecessary. GMM-type estimators remain asymptotically unbiased under mild stationarity requirements, even in the presence of conditionally heteroscedastic error terms.

Since we are estimating three model parameters with six moment conditions, the model is overidentified. This is a deliberate choice so to enable testing over-identifying restrictions.

GMM estimation can be very finicky in the choice of inverse weighting matrix (also known as spectral density matrix) which can account for various forms of heteroscedasticity and/or serial correlation, choice of optimization algorithm, initial guesses[14] and choice of instruments. We have selected spectral density matrix estimate of White (1980) over alternative approaches based on its better performance in recovering true parameters using pilot artificial samples from (4.1) with normal and Student t residuals. We used the same method to tune other GMM parameters such as the number of lags in estimation of the spectral density matrix, number of convergence steps to tune weighting matrix, demeaning, etc.

GMM estimation results are shown in Table 12. Over the full sample, most estimates cluster around zero providing justification for the difference form. Elasticity estimates universally increase when the estimation window is confined to the last four years, with only 30-year rates remaining in difference space. Over the most recent one year, GMM estimates are again closer to returns but the uncertainty bounds surrounding these are far too loose for comfort. These findings prompt an idea that elasticity of volatility is a declining function of interest rate level.

Asymptotic χ^2 J-test for overidentifying restrictions (Gallant, 1987), shown in Table 13, suggests that over the full 12 years of data, the CEV model constraints are too restrictive, which is perhaps not too surprising as the large volume of data can overwhelm any simple model. At one-year window, the CEV model is not rejected in about half the cases. Not surprisingly, an examination of individual moment restrictions (not shown to save space) indicates that it is matching the second moments that is a problem that leads to model rejection. Subsample parameter instability seen in Table 12 is more informative regarding possible extensions of the model needed to improve fit.

4.3. Full Information Approach: Maximum Likelihood.

Full information approaches require complete parametric specification for the distribution of residuals ϵ_t. Figure 3 contrasts two common choices that we can implement – normal and Student t distribution. Student t-distribution has fatter tail, and so does CEV process with t-distributed residuals. The upper panel displays two simulated paths starting from the common value and with perfect rank correlation of moves. It makes clear that the size of the move under the Student t-distribution is larger than under normal distribution, conditional on rank of the residual realization. The lower panel display kernel-smoothed densities of the two paths. Again, t-distribution tends to produce fatter tails when it drives CEV -type process. This could be an important feature *if* the normal residual distribution does not produce tail fat enough.

[14] We based initial guess on a two-stage least squares approach. In the first stage $\hat{\mu} = \overline{\Delta r_t}$. Second stage uses regression $\log\left((\Delta r_r - \hat{\mu})^2\right) + 1.270362845 = \log\sigma^2 + \gamma\log\left(r_{t-1}^2\right) + \tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t = \log\left(\epsilon_t^2\right) + 1.270362845$. Constant -1.270362845 is intended to ensure zero mean of $\tilde{\epsilon}_t$ and is valid for normally distributed ϵ_t.

Swap Maturity	Parameter	Full sample Estimate	Full sample Bounds	4-year sample Estimate	4-year sample Bounds	1-year sample Estimate	1-year sample Bounds
1y	μ	-0.002	(-0.003,-0.001)	-0.000	(-0.002,0.001)	0.000	(-0.002,0.002)
	σ	0.031	(0.029,0.034)	0.037	(0.034,0.041)	0.019	(0.002,0.036)
	γ	0.122	(0.052,0.194)	0.890	(0.742,1.039)	0.187	(-1.207,1.581)
2y	μ	-0.002	(-0.004,-0.001)	-0.001	(-0.003,0.001)	-0.000	(-0.003,0.002)
	σ	0.041	(0.038,0.045)	0.037	(0.034,0.041)	0.027	(0.014,0.041)
	γ	0.088	(0.020,0.156)	0.912	(0.738,1.086)	0.458	(-0.537,1.454)
3y	μ	-0.003	(-0.005,-0.001)	-0.001	(-0.004,0.002)	0.000	(-0.003,0.003)
	σ	0.046	(0.041,0.050)	0.037	(0.033,0.042)	0.029	(0.012,0.045)
	γ	0.101	(0.027,0.174)	0.868	(0.683,1.052)	0.746	(-0.551,2.042)
4y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.002)	-0.000	(-0.004,0.003)
	σ	0.050	(0.044,0.056)	0.036	(0.031,0.042)	0.030	(0.021,0.039)
	γ	0.091	(0.013,0.170)	0.839	(0.640,1.036)	1.190	(-0.730,3.110)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.002)	-0.001	(-0.005,0.003)
	σ	0.053	(0.046,0.060)	0.034	(0.028,0.040)	0.026	(0.020,0.032)
	γ	0.084	(-0.005,0.173)	0.837	(0.635,1.039)	1.397	(-0.166,2.960)
7y	μ	-0.003	(-0.005,-0.000)	-0.002	(-0.006,0.002)	-0.002	(-0.007,0.003)
	σ	0.059	(0.049,0.069)	0.032	(0.025,0.038)	0.023	(0.010,0.035)
	γ	0.023	(-0.084,0.131)	0.800	(0.591,1.009)	1.204	(-0.004,2.412)
10y	μ	-0.002	(-0.004,0.000)	-0.002	(-0.006,0.002)	-0.003	(-0.009,0.004)
	σ	0.074	(0.059,0.089)	0.035	(0.025,0.045)	0.022	(0.006,0.038)
	γ	-0.105	(-0.235,0.024)	0.635	(0.389,0.882)	1.013	(0.024,2.002)
30y	μ	-0.002	(-0.004,0.000)	-0.003	(-0.007,0.002)	-0.002	(-0.009,0.005)
	σ	0.143	(0.107,0.178)	0.086	(0.049,0.123)	0.026	(-0.001,0.053)
	γ	-0.558	(-0.711,-0.405)	-0.154	(-0.482,0.173)	0.701	(-0.320,1.721)

TABLE 12. GMM estimation results.

Swap Maturity	Full sample J-test	Full sample p-value	4-year sample J-test	4-year sample p-value	1-year sample J-test	1-year sample p-value
1y	86.027	0.000	6.984	0.072	3.202	0.361
2y	149.542	0.000	10.971	0.012	2.791	0.425
3y	157.656	0.000	7.518	0.057	7.053	0.070
4y	133.993	0.000	5.104	0.164	8.083	0.044
5y	126.020	0.000	5.596	0.133	8.273	0.041
7y	92.017	0.000	5.856	0.119	9.839	0.020
10y	50.021	0.000	6.345	0.096	12.309	0.006
30y	12.033	0.007	11.124	0.011	13.311	0.004

TABLE 13. GMM J-test for overidentifying restrictions.

Under the normal assumption, the likelihood of one observation Δr_t is given by

$$p\left(\Delta r_t | r_{t-1}, \mu, \sigma, \gamma\right) = \frac{1}{\sigma r_{t-1}^{\gamma}\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2 r_{t-1}^{2\gamma}}\left(\Delta r_t - \mu\right)^2\right). \tag{4.4}$$

Under the Student t-distribution assumption, the likelihood is instead

$$p\left(\Delta r_t | r_{t-1}, \mu, \sigma, \gamma, \nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma r_{t-1}^{\gamma}\sqrt{\nu\pi}}\left(1 + \frac{1}{\nu}\left(\frac{\Delta r_t - \mu}{\sigma r_{t-1}^{\gamma}}\right)^2\right)^{-(\nu+1)/2}. \tag{4.5}$$

As $\nu \to \infty$, t_ν-distribution converges to normal and so does the CEV likelihood.

Swap Maturity	Parameter	Full sample		4-year sample		1-year sample	
		Estimate	95% Bounds	Estimate	95% Bounds	Estimate	95% Bounds
1y	μ	-0.001	(-0.002,0.000)	0.000	(-0.001,0.002)	0.000	(-0.002,0.003)
	σ	0.033	(0.032,0.034)	0.039	(0.037,0.041)	0.030	(0.022,0.041)
	γ	0.389	(0.352,0.425)	0.994	(0.910,1.077)	0.774	(0.302,1.245)
2y	μ	-0.001	(-0.003,0.001)	-0.000	(-0.002,0.002)	-0.000	(-0.003,0.003)
	σ	0.044	(0.042,0.047)	0.040	(0.038,0.041)	0.033	(0.024,0.045)
	γ	0.312	(0.267,0.358)	0.937	(0.850,1.024)	0.731	(0.172,1.290)
3y	μ	-0.002	(-0.004,0.001)	-0.001	(-0.003,0.002)	-0.000	(-0.003,0.003)
	σ	0.050	(0.047,0.053)	0.038	(0.036,0.040)	0.049	(0.039,0.062)
	γ	0.235	(0.181,0.288)	0.899	(0.808,0.991)	1.857	(1.229,2.485)
4y	μ	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)	-0.001	(-0.005,0.002)
	σ	0.056	(0.051,0.060)	0.035	(0.033,0.038)	0.037	(0.033,0.040)
	γ	0.161	(0.099,0.223)	0.893	(0.789,0.997)	2.277	(1.705,2.850)
5y	μ	-0.002	(-0.004,0.001)	-0.002	(-0.005,0.002)	-0.002	(-0.006,0.002)
	σ	0.061	(0.056,0.067)	0.033	(0.029,0.036)	0.025	(0.022,0.028)
	γ	0.100	(0.028,0.171)	0.922	(0.802,1.042)	2.220	(1.663,2.776)
7y	μ	-0.002	(-0.004,0.000)	-0.002	(-0.006,0.002)	-0.003	(-0.008,0.003)
	σ	0.072	(0.064,0.081)	0.030	(0.026,0.035)	0.017	(0.012,0.023)
	γ	-0.040	(-0.125,0.046)	0.889	(0.739,1.040)	1.967	(1.344,2.590)
10y	μ	-0.002	(-0.004,0.000)	-0.002	(-0.006,0.002)	-0.003	(-0.009,0.003)
	σ	0.087	(0.075,0.100)	0.033	(0.027,0.041)	0.013	(0.007,0.022)
	γ	-0.176	(-0.273,-0.079)	0.716	(0.529,0.902)	1.903	(1.176,2.630)
30y	μ	-0.002	(-0.004,0.000)	-0.003	(-0.007,0.002)	-0.003	(-0.010,0.005)
	σ	0.154	(0.130,0.184)	0.085	(0.062,0.116)	0.009	(0.003,0.023)
	γ	-0.602	(-0.713,-0.490)	-0.111	(-0.353,0.130)	1.892	(0.957,2.828)

TABLE 14. Maximum likelihood estimation results with Gaussian innovations.

The log-likelihood of the sample $\{r_t\}_{t=0}^T$ is then

$$\mathcal{L}\left(\mu,\sigma,\gamma,\nu\middle|\{r_t\}_{t=0}^T\right)=\sum_{t=1}^T \log p\left(\Delta r_t|r_{t-1},\mu,\sigma,\gamma,\nu\right), \tag{4.6}$$

with $\nu=\infty$ distinguishing model with normal innovations from that with Student t residuals.

The magnitudes and patterns of the maximum likelihood estimates, in Table 14 for the normal residuals and in Table 15 for the Student t case, are in line with the expectations set by the GMM approach. To reiterate, estimates of the elasticity of volatility are in the vicinity or below zero using the full sample and rise steeply if the earlier data, characterized by higher interest rates, are excluded, continuing to support an idea of elasticity declining with level. The steep rise effect is slightly stronger in the Gaussian case. With respect to the parameter ν that governs tail behavior of Student t distribution, we find that significant departures from normality are evident at the short end of the curve using the full sample. Elsewhere, indications of non-normality are more moderate, although the likelihood ratio test (not shown) strongly rejects the null of normal CEV. On a technical side, we note that log-likelihood function is better behaved than the GMM criterion, making for an easier effort in obtaining the estimates.

Variation of $\hat{\gamma}_{MLE}$ across different sub-samples is very notable. To formally assess its statistical significance, we applied the Nyblom stability test (Nyblom, 1989) that amounts to the Lagrange Multiplier test of constant parameters against the martingale difference alternative with constant

Swap Maturity	Parameter	Full sample		4-year sample		1-year sample	
		Estimate	95% Bounds	Estimate	95% Bounds	Estimate	95% Bounds
1y	μ	-0.001	(-0.002,0.000)	-0.001	(-0.002,0.000)	0.000	(-0.002,0.002)
	σ	0.021	(0.020,0.022)	0.030	(0.028,0.033)	0.019	(0.011,0.031)
	γ	0.299	(0.257,0.342)	0.950	(0.845,1.054)	0.379	(-0.360,1.117)
	ν	2.883	(2.542,3.224)	5.223	(3.640,6.805)	5.698	(2.189,9.207)
2y	μ	-0.002	(-0.003,0.000)	-0.001	(-0.003,0.001)	0.000	(-0.002,0.002)
	σ	0.030	(0.028,0.032)	0.031	(0.028,0.033)	0.022	(0.014,0.036)
	γ	0.304	(0.253,0.354)	0.960	(0.846,1.073)	0.624	(-0.243,1.492)
	ν	3.387	(2.926,3.849)	5.037	(3.514,6.560)	4.218	(2.005,6.432)
3y	μ	-0.002	(-0.004,-0.000)	-0.001	(-0.003,0.001)	0.000	(-0.003,0.003)
	σ	0.035	(0.032,0.038)	0.029	(0.027,0.031)	0.029	(0.020,0.041)
	γ	0.259	(0.200,0.318)	0.938	(0.822,1.054)	1.312	(0.441,2.183)
	ν	3.829	(3.260,4.398)	4.994	(3.483,6.506)	3.945	(1.937,5.952)
4y	μ	-0.002	(-0.004,-0.000)	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)
	σ	0.040	(0.036,0.043)	0.028	(0.025,0.030)	0.028	(0.024,0.032)
	γ	0.193	(0.124,0.261)	0.908	(0.780,1.036)	1.943	(1.170,2.716)
	ν	4.174	(3.513,4.835)	5.299	(3.644,6.954)	5.058	(2.113,8.002)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.001)	-0.002	(-0.005,0.002)
	σ	0.043	(0.039,0.048)	0.026	(0.023,0.030)	0.022	(0.018,0.026)
	γ	0.143	(0.065,0.222)	0.894	(0.749,1.039)	1.940	(1.233,2.646)
	ν	4.326	(3.622,5.030)	5.322	(3.619,7.024)	6.551	(1.826,11.276)
7y	μ	-0.003	(-0.005,-0.000)	-0.002	(-0.006,0.001)	-0.003	(-0.008,0.002)
	σ	0.052	(0.045,0.060)	0.026	(0.022,0.031)	0.017	(0.012,0.025)
	γ	0.010	(-0.085,0.105)	0.806	(0.629,0.983)	1.750	(1.015,2.485)
	ν	4.798	(3.950,5.646)	5.806	(3.807,7.805)	11.612	(-3.057,26.282)
10y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.006,0.002)	-0.003	(-0.009,0.003)
	σ	0.062	(0.052,0.073)	0.028	(0.022,0.036)	0.013	(0.007,0.024)
	γ	-0.119	(-0.231,-0.007)	0.651	(0.432,0.870)	1.699	(0.868,2.531)
	ν	4.826	(3.967,5.684)	5.188	(3.533,6.844)	12.198	(-5.594,29.991)
30y	μ	-0.002	(-0.004,-0.000)	-0.002	(-0.006,0.002)	-0.003	(-0.011,0.004)
	σ	0.109	(0.088,0.135)	0.061	(0.042,0.090)	0.011	(0.004,0.031)
	γ	-0.516	(-0.650,-0.382)	-0.062	(-0.352,0.229)	1.558	(0.514,2.603)
	ν	5.572	(4.473,6.672)	4.845	(3.407,6.284)	8.888	(0.065,17.712)

TABLE 15. Maximum likelihood estimation results with Student t innovations.

hazard of parameter change (Hansen, 1990).[15,16] To make a sharper contrast, location and scale parameters μ and σ were assumed to be constant under the alternative, even though there is some evidence that the volatility scale parameter may be unstable as well, see section 9. Three versions of the test were run, making different assumptions about which subset of $\{\gamma, \nu\}$ is potentially unstable. The results are shown in Table 16 with the p-values corrected for the finite sample size distortions under the null with the help of a Monte Carlo simulation. We conclude that elasticity of volatility is not stable in the full sample, but mostly stable in the shorter subsamples. As the later subsamples are associated with lower interest rates, this finding points to a potential usefulness of the variable elasticity of volatility. We also find that even though the estimates of

[15]The martingale difference alternative allows for substantial flexibility, such as random walk or single structural break alternatives (Nyblom, 1989). However, the test is not informative about the timing or type of the structural change.

[16]Test formulation relies on analytic scores and the Hessian of the log-likelihood function. Details of the test derivation are relegated to appendix C.

Swap Maturity	Parameter	Full sample		4-year sample		1-year sample	
		Estimate	95% Bounds	Estimate	95% Bounds	Estimate	95% Bounds
1y	μ	-0.001	(-0.002,0.000)	-0.001	(-0.002,0.000)	0.000	(-0.002,0.002)
	σ	0.021	(0.020,0.022)	0.030	(0.028,0.033)	0.019	(0.011,0.031)
	γ	0.299	(0.257,0.342)	0.950	(0.845,1.054)	0.379	(-0.360,1.117)
	ν	2.883	(2.542,3.224)	5.223	(3.640,6.805)	5.698	(2.189,9.207)
2y	μ	-0.002	(-0.003,0.000)	-0.001	(-0.003,0.001)	0.000	(-0.002,0.002)
	σ	0.030	(0.028,0.032)	0.031	(0.028,0.033)	0.022	(0.014,0.036)
	γ	0.304	(0.253,0.354)	0.960	(0.846,1.073)	0.624	(-0.243,1.492)
	ν	3.387	(2.926,3.849)	5.037	(3.514,6.560)	4.218	(2.005,6.432)
3y	μ	-0.002	(-0.004,-0.000)	-0.001	(-0.003,0.001)	0.000	(-0.003,0.003)
	σ	0.035	(0.032,0.038)	0.029	(0.027,0.031)	0.029	(0.020,0.041)
	γ	0.259	(0.200,0.318)	0.938	(0.822,1.054)	1.312	(0.441,2.183)
	ν	3.829	(3.260,4.398)	4.994	(3.483,6.506)	3.945	(1.937,5.952)
4y	μ	-0.002	(-0.004,-0.000)	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)
	σ	0.040	(0.036,0.043)	0.028	(0.025,0.030)	0.028	(0.024,0.032)
	γ	0.193	(0.124,0.261)	0.908	(0.780,1.036)	1.943	(1.170,2.716)
	ν	4.174	(3.513,4.835)	5.299	(3.644,6.954)	5.058	(2.113,8.002)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.001)	-0.002	(-0.005,0.002)
	σ	0.043	(0.039,0.048)	0.026	(0.023,0.030)	0.022	(0.018,0.026)
	γ	0.143	(0.065,0.222)	0.894	(0.749,1.039)	1.940	(1.233,2.646)
	ν	4.326	(3.622,5.030)	5.322	(3.619,7.024)	6.551	(1.826,11.276)
7y	μ	-0.003	(-0.005,-0.000)	-0.002	(-0.006,0.001)	-0.003	(-0.008,0.002)
	σ	0.052	(0.045,0.060)	0.026	(0.022,0.031)	0.017	(0.012,0.025)
	γ	0.010	(-0.085,0.105)	0.806	(0.629,0.983)	1.750	(1.015,2.485)
	ν	4.798	(3.950,5.646)	5.806	(3.807,7.805)	11.612	(-3.057,26.282)
10y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.006,0.002)	-0.003	(-0.009,0.003)
	σ	0.062	(0.052,0.073)	0.028	(0.022,0.036)	0.013	(0.007,0.024)
	γ	-0.119	(-0.231,-0.007)	0.651	(0.432,0.870)	1.699	(0.868,2.531)
	ν	4.826	(3.967,5.684)	5.188	(3.533,6.844)	12.198	(-5.594,29.991)
30y	μ	-0.002	(-0.004,-0.000)	-0.002	(-0.006,0.002)	-0.003	(-0.011,0.004)
	σ	0.109	(0.088,0.135)	0.061	(0.042,0.090)	0.011	(0.004,0.031)
	γ	-0.516	(-0.650,-0.382)	-0.062	(-0.352,0.229)	1.558	(0.514,2.603)
	ν	5.572	(4.473,6.672)	4.845	(3.407,6.284)	8.888	(0.065,17.712)

TABLE 16. Nyblom stability test for CEV model with Student t errors.

the degrees of freedom parameters may appear drifting towards normality over time, the drift is largely a consequence of reduced sample size.

Notice that in both MLE models, the confidence bounds are tighter than those from the GMM. One can view the maximum likelihood as a limiting case of GMM: under MLE the distribution of errors is specified so that in a sense all of the moments are included.

Figure 3 continues comparison of MLE and GMM results via an error plot indicating two standard error bands around respective estimates of γ. It makes apparent a slight tendency of the GMM estimates of elasticity of volatility to fall below maximum likelihood estimates, but not statistically significantly in most cases. Vertical elongation of error bands, albeit visually obscured by unequal axes, confirms higher efficiency of likelihood-based methods, while the GMM offers a better robustness to deviations from Student t assumption. The cost of robustness is the GMM error bounds that are much too wide when only one year of data is used.

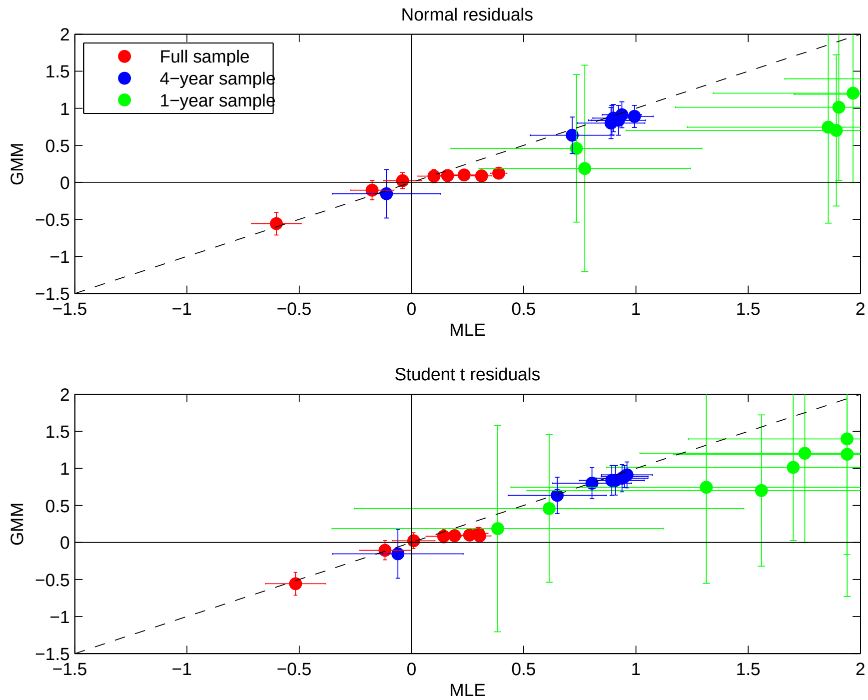

FIGURE 3. Comparison of GMM and maximum likelihood estimates.

4.4. Full Information Approach: Bayesian Estimation. Bayesian approach goes further than maximum likelihood – in addition to the information contained in the data sample, it utilizes prior information. Prior information could be either innate views, natural parameter restrictions (e.g., certain stability constraints) or information from earlier or related studies.

Bayesian approach also resolves a potential problem with maximum likelihood algorithm in that required multivariate optimization could be sometimes difficult, especially in the multi-modal case, and could lead one to point estimates outside the reasonable range.

Since we are interested in full information approaches, we again have to specify parametric form of the residual distribution. As before and out of the same considerations for robustness against fat tails, we specify either normal or Student t-distributed residuals.[17] The degrees of freedom parameter will be allowed to be unknown random parameter with its own prior.

For estimation and sampling convenience we adopt the scale mixture of normals representation for the Student-t family of distributions[18]

$$\epsilon_t \big| u_t \sim \mathcal{N}(0, u_t), \text{ where } \frac{1}{u_t} \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{4.7}$$

with $\mathcal{G}(df, S)$ denoting gamma distribution with degrees of freedom (shape) parameter df and scale parameter S.

[17] Chan, Choy, and Lee (2007) advocate the exponential power distribution family that encompasses normal and Laplace distributions.

[18] Scale mixture representation draws on the familiar genesis of t distribution as a ratio of normal and χ^2 distributions. Indeed, $\mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ is χ^2_ν.

4.4.1. Prior Distributions. Regarding the prior distributions of the model parameters $(\mu, \sigma, \gamma, \nu)$ we follow the literature in postulating that all components are independent.

Mean parameter μ is a priori normally distributed:

$$\mu \sim \mathcal{N}\left(m_{\mu}^{0}, \sigma_{0\mu}^{2}\right). \tag{4.8}$$

Since weakly informative priors obtain when $\sigma_{0\mu}$ is large, we set $\sigma_{0\mu} = 10$. We center the distribution at zero ($m_{\mu}^{0} = 0$) as the results in the previous sections indicate that $\hat{\mu} \approx 0$. Conditional posterior is also Gaussian:

$$\mu | \sigma, \gamma, u_t, \{r_t\}_{t=1}^{T} \sim \mathcal{N}\left(m_{\mu}^{1}, \sigma_{1\mu}^{2}\right), \tag{4.9}$$

where

$$\sigma_{1\mu}^{-2} = \sigma_{0\mu}^{-2} + \sigma^{-2} \sum_{t=2}^{T} r_{t-1}^{-2\gamma} \frac{1}{u_t}, \tag{4.10}$$

$$m_{\mu}^{1} = \sigma_{1\mu}^{-2} \left(\sigma_{0\mu}^{-2} m_{\mu}^{0} + \sigma^{-2} \sum_{t=2}^{T} r_{t-1}^{-2\gamma} \frac{\Delta r_t}{u_t} \right). \tag{4.11}$$

The scale of volatility is assumed to follow an inverse-gamma prior distribution:

$$\sigma^2 \sim \mathcal{IG}\left(df_0, S_0\right) \tag{4.12}$$

with degrees of freedom parameter df_0 and scale parameter S_0. This distribution is convenient since it is a conjugate prior for the univariate normal sampling model (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, 2013), so that the posterior distribution, conditional on μ, γ and u_t, is also inverse-gamma:

$$\sigma^2 \left| \left\{ \frac{r_t - \mu}{r_{t-1}^{\gamma} \sqrt{u_t}} \right\}_{t=2}^{T} \right. \sim \mathcal{IG}\left(df_1, S_1\right), \tag{4.13}$$

where

$$df_1 = df_0 + T - 1, \tag{4.14}$$

$$S_1 = \frac{df_0 S_0 + \sum_{t=2}^{T} \frac{1}{u_t} \left(\frac{\Delta r_t}{r_{t-1}^{\gamma}} - \overline{\frac{\Delta r_t}{r_{t-1}^{\gamma}}} \right)^2}{T - 1}. \tag{4.15}$$

To express ignorance about the scale of the process we have chosen vague prior for σ^2 by setting $df_0 = 0.2$ and $S_0 = 100{,}000$.

For the elasticity of variance parameter, we propose uniform distribution on [−1, 2].

$$\gamma \sim \mathcal{U}[-1, 2]. \tag{4.16}$$

This choice is motivated by desire to limit the range of possible values coupled with the lack of substantially precise prior information. Indeed, the results in Chan, Karolyi, Longstaff, and Sanders (1992) and Lubrano (2001) suggest that elasticity moderately in excess of unity is common for short term interest rates while using −1 as a lower bound is designed to conservatively bracket the range [0, 1] that is popular in the interest rate option pricing literature. The prior mean corresponds to the CIR square root model (Cox, Ingersoll, and Ross, 1985).

Lastly, we specify the prior for the degrees of freedom parameter as uniform on [1, 30]:

$$\nu \sim \mathcal{U}[1, 30]. \tag{4.17}$$

$\nu = 1$ corresponds to the Cauchy distribution with infinite mean, and we therefore assume that CEV residuals do not have tails fatter than Cauchy's. Similarly, it is well known that Student t distribution with ν of about 30 comes very close to normal distribution, and it would be very hard and perhaps pointless to try to distinguish values of ν above that range.

4.4.2. Objects of Interest. The main object of interest is the joint posterior distribution of the four unknown parameters, $p\left(\mu, \sigma, \gamma, \nu \middle| \{r_t\}_{t=1}^T\right)$, and the corresponding four marginal posterior distributions. Direct calculation of these is complicated, since we have to integrate our latent mixing parameters and because priors for γ and ν are not conjugate. Direct marginalizing γ or ν out of the joint posterior is also difficult, since it involves three-dimensional integration needed to infer the normalizing constant in

$$p\left(\mu, \sigma, \gamma, \nu \middle| \{r_t\}_{t=1}^T\right) \propto p\left(\{r_t\}_{t=1}^T \middle| \mu, \sigma, \gamma, \nu\right) p(\mu, \sigma, \gamma, \nu). \tag{4.18}$$

Instead, we turn to the popular Markov chain Monte Carlo (MCMC) class of algorithms (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, 2013; Robert and Casella, 2000). An appealing feature of MCMC is its modular nature as extended models can be developed by building additional blocks without destroying the foundation. This will be seen in the subsequent sections. Details of algorithm are relegated to appendix D.1.

4.4.3. Posterior Inference. Posterior medians and posterior 95% confidence bounds for different swap maturities and subsamples are shown in Table 17. The results are in excellent agreement with ML estimates in Table 15, particularly regarding the magnitudes and patterns of elasticity estimates over maturities and over different subsamples. These patterns continue suggesting downward-sloping elasticity as the rates go higher. Consistency with MLE illustrates general vagueness of our prior assumptions. The upper bounds of posterior confidence intervals for γ are, however, substantially smaller for the 1-year data window due to asymmetry of the marginal posterior distribution while the asymptotic MLE bounds are based on a normal distribution and thus are symmetric.

Indeed, a conflict between prior and likelihood information is evident in Figure 4 where posterior density of γ for 30-year swap rate estimated over 1-year subsample leans against the upper bound of the prior's support. Similar conflicts occur for all maturities over 4-year, whereas posterior densities for shorter maturities tend to gravitate toward the center of the prior's support. The model can be easily re-estimated by widening the support of the prior distribution. However, no such conflicts arise for longer samples, so we continue to trust our prior beliefs about plausible range for elasticity and leave the priors alone.

5. Multivariate Common Constant Elasticity of Volatility Model

5.1. The Model. It is often inconvenient to treat related risk factor time-series separately with regard to selection of functional form. Indeed, if the different series represent rates or spreads on similar instruments across the spectra of maturity or credit rating, assigning some to returns and others to differences makes explaining comovements more difficult, opens a door to the suspicion of gaming the system and leaves an impression of arbitrariness, especially if supporting evidence is vague.

Thus, it makes sense to constrain the elasticity of variance in the cross-section of related risk factors, while still allowing differential volatility scales and other parameters. In other words,

Swap Maturity	Parameter	Full sample		4-year sample		1-year sample	
		Posterior Median	Posterior 95% Bounds	Posterior Median	Posterior 95% Bounds	Posterior Median	Posterior 95% Bounds
1y	μ	-0.001	(-0.002,0)	-0.001	(-0.002,0)	0.000	(-0.002,0.002)
	σ	0.021	(0.02,0.023)	0.030	(0.028,0.033)	0.019	(0.012,0.031)
	γ	0.299	(0.257,0.342)	0.951	(0.85,1.056)	0.374	(-0.351,1.103)
	ν	2.897	(2.576,3.273)	5.388	(4.025,7.592)	6.479	(3.554,15.438)
2y	μ	-0.002	(-0.003,0)	-0.001	(-0.003,0.001)	0.000	(-0.002,0.002)
	σ	0.030	(0.028,0.032)	0.031	(0.029,0.033)	0.023	(0.014,0.036)
	γ	0.304	(0.253,0.354)	0.960	(0.849,1.073)	0.598	(-0.232,1.457)
	ν	3.404	(2.985,3.918)	5.191	(3.891,7.272)	4.579	(2.791,8.742)
3y	μ	-0.002	(-0.004,0)	-0.001	(-0.003,0.001)	0.000	(-0.003,0.003)
	σ	0.035	(0.032,0.038)	0.029	(0.027,0.032)	0.029	(0.021,0.038)
	γ	0.258	(0.2,0.319)	0.937	(0.823,1.054)	1.265	(0.414,1.925)
	ν	3.853	(3.341,4.501)	5.154	(3.849,7.146)	4.274	(2.655,8.015)
4y	μ	-0.002	(-0.004,0)	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)
	σ	0.040	(0.036,0.043)	0.028	(0.025,0.031)	0.028	(0.024,0.032)
	γ	0.194	(0.124,0.262)	0.910	(0.779,1.035)	1.715	(1.065,1.987)
	ν	4.217	(3.621,4.978)	5.480	(4.094,7.841)	5.641	(3.247,12.464)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.001)	-0.002	(-0.006,0.002)
	σ	0.044	(0.039,0.048)	0.026	(0.023,0.03)	0.023	(0.02,0.027)
	γ	0.143	(0.065,0.222)	0.896	(0.747,1.042)	1.750	(1.183,1.986)
	ν	4.367	(3.74,5.19)	5.530	(4.074,7.949)	7.880	(3.965,23.405)
7y	μ	-0.003	(-0.005,0)	-0.002	(-0.006,0.002)	-0.003	(-0.008,0.002)
	σ	0.052	(0.045,0.06)	0.026	(0.022,0.031)	0.018	(0.015,0.025)
	γ	0.012	(-0.085,0.104)	0.811	(0.63,0.987)	1.664	(1.014,1.981)
	ν	4.837	(4.079,5.834)	6.071	(4.372,8.972)	15.128	(5.704,28.941)
10y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.006,0.002)	-0.003	(-0.01,0.003)
	σ	0.062	(0.052,0.073)	0.028	(0.022,0.036)	0.015	(0.011,0.026)
	γ	-0.116	(-0.228,-0.007)	0.658	(0.443,0.872)	1.619	(0.868,1.979)
	ν	4.869	(4.095,5.89)	5.374	(3.958,7.737)	16.196	(5.913,29.232)
30y	μ	-0.002	(-0.004,0)	-0.002	(-0.006,0.002)	-0.003	(-0.011,0.004)
	σ	0.109	(0.088,0.136)	0.061	(0.041,0.089)	0.011	(0.007,0.028)
	γ	-0.517	(-0.651,-0.381)	-0.051	(-0.342,0.241)	1.544	(0.659,1.974)
	ν	5.655	(4.682,7.022)	5.000	(3.765,7.05)	12.526	(4.941,28.496)

TABLE 17. Bayesian estimation results for CEV-t model.

consider the following model for the cross-section of N related risk factor time-series:

$$\Delta r_{it} = \mu_i + \sigma_i r_{it}^{\gamma} \epsilon_{it}, \quad \epsilon_{it} \sim t_{\nu_i}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \tag{5.1}$$

5.2. Estimation. This model could be treated with GMM or maximum likelihood methods, but neither likelihood function nor GMM objective factorize exactly along different sets of series-specific parameters, which calls for multivariate optimization, which could be delicate to set up and ensure convergence to the global optimum. On the other hand, Bayesian approach of the previous section can be extended to the common elasticity of volatility setting in a straightforward fashion. [19] Therefore, we focus on Bayesian computation to facilitate this model.

[19]Bayesian approach is also natural for a multilevel hierarchical versions of (5.1) that incorporate random effects of maturity, e.g.:

$$\Delta r_{it} \sim \mathcal{N}\left(\mu_i, \exp\left(\log\sigma_i + H(z_i;\theta_{zi})\log r_{it}\right)\right), \ \mu_i \sim \mathcal{N}\left(\mu_0, \sigma_\mu^2\right), \ \log\sigma_i \sim \mathcal{N}\left(s_0 i, \sigma_s^2\right), \ \theta_{zi} \sim \mathcal{N}\left(\theta_{z0}, \sigma_\theta^2\right),$$

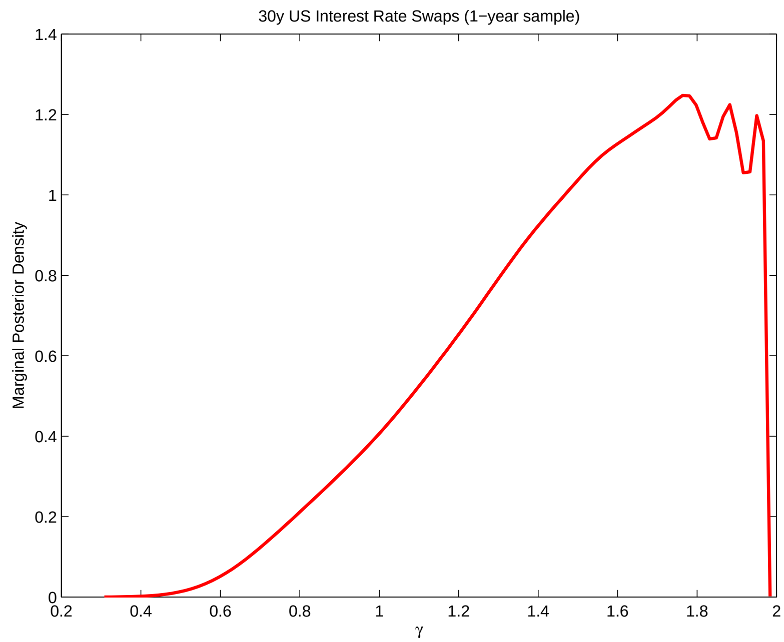

FIGURE 4. Posterior distribution of elasticity of volatility of US dollar 30-year interest rate swap rates over 1-year subsample in CEV-t model.

Again, we make an independent prior assumption for all μ_i, σ_i, γ, and ν_i:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \gamma, \boldsymbol{\nu}) = p(\gamma) \prod_{i=1}^{N} p(\mu_i)\, p(\sigma_i)\, p(\nu_i)\,, \tag{5.2}$$

where boldface variables denote vectors of parameters in the cross-section.

Component priors take the same convenient functional form as in the single time-series case, except for wider prior for ν_i:

$$\mu_i \sim \mathcal{N}\left(m_\mu^0, \sigma_{0\mu}^2\right), \tag{5.3}$$

$$\sigma_i^2 \sim \mathcal{IG}\left(df_0, S_{i0}\right), \tag{5.4}$$

$$\gamma \sim \mathcal{U}[-1, 2], \tag{5.5}$$

$$\nu_i \sim \mathcal{U}[1, 130]. \tag{5.6}$$

The MCMC algorithm for this model is similar and can be found in appendix D.2.

5.3. Posterior Inference. Table 18 gives posterior medians and posterior 95% confidence bounds for all parameters of common constant elasticity of volatility model, while Figure 5 compares posterior shapes for γ distributions. The results are very similar for all maturity-specific parameters except for the scale of volatility at the long end of the curve using the full sample and for 30-year maturity using 4-year sample. Common elasticity of volatility is estimated higher than the

where $H(\cdot)$ is any function that maps $\mathbb{R}$ onto a suitable finite interval to constrain values of γ to a plausible range and θ_{zi} are parameters describing the stochastic process of (transformed) elasticity of volatility. We do not pursue such models here in order to keep the presentation concise.

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	-0.001	(-0.002,0)	-0.001	(-0.002,0)	0.000	(-0.002,0.002)
	σ	0.023	(0.021,0.024)	0.029	(0.027,0.032)	0.038	(0.03,0.048)
	ν	2.812	(2.509,3.161)	5.298	(3.99,7.417)	6.046	(3.385,14.996)
2y	μ	-0.002	(-0.003,0)	-0.001	(-0.003,0.001)	0.000	(-0.002,0.003)
	σ	0.033	(0.031,0.035)	0.031	(0.029,0.033)	0.035	(0.028,0.043)
	ν	3.382	(2.973,3.898)	5.165	(3.86,7.234)	4.293	(2.663,8.057)
3y	μ	-0.002	(-0.004,0)	-0.001	(-0.003,0.001)	0.000	(-0.003,0.003)
	σ	0.037	(0.035,0.039)	0.030	(0.027,0.032)	0.031	(0.025,0.037)
	ν	3.871	(3.34,4.522)	5.137	(3.853,7.243)	4.362	(2.685,8.327)
4y	μ	-0.002	(-0.004,0)	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)
	σ	0.039	(0.037,0.041)	0.028	(0.026,0.03)	0.027	(0.023,0.031)
	ν	4.205	(3.622,4.961)	5.474	(4.087,7.787)	5.399	(3.132,11.897)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.001)	-0.002	(-0.006,0.002)
	σ	0.040	(0.038,0.042)	0.027	(0.024,0.029)	0.024	(0.021,0.028)
	ν	4.342	(3.719,5.144)	5.494	(4.074,7.948)	7.284	(3.792,20.653)
7y	μ	-0.003	(-0.005,0)	-0.002	(-0.006,0.002)	-0.003	(-0.008,0.002)
	σ	0.040	(0.038,0.042)	0.024	(0.022,0.027)	0.020	(0.017,0.024)
	ν	4.743	(4.011,5.712)	6.157	(4.422,9.243)	14.460	(5.705,28.759)
10y	μ	-0.003	(-0.005,0)	-0.002	(-0.006,0.002)	-0.004	(-0.01,0.003)
	σ	0.039	(0.036,0.041)	0.022	(0.02,0.024)	0.016	(0.013,0.021)
	ν	4.702	(3.98,5.668)	5.469	(4.022,7.904)	15.707	(5.719,29.054)
30y	μ	-0.002	(-0.004,0)	-0.002	(-0.006,0.002)	-0.003	(-0.011,0.004)
	σ	0.035	(0.033,0.037)	0.018	(0.016,0.02)	0.013	(0.009,0.017)
	ν	4.778	(4.031,5.755)	4.522	(3.463,6.162)	12.455	(4.933,28.516)
	γ	0.206	(0.183,0.229)	0.883	(0.835,0.931)	1.446	(1.15,1.734)

Table 18. Bayesian estimation results for common elasticity of volatility model.

average of individual elasticities in all cases, with little overlap of posterior distribution using full or 4-year samples. For each sample, the overlap of posterior distributions for ν_i is also substantial, with an exception of one- and two-year maturities using full sample. This suggests further reduction in the number of distinct model parameters.

All in all, the full sample is suggestive of difference representation of interest rate swap rate risk, while risk over 4-year and 1-year windows would be better represented in yield return space.

6. Asymmetric Elasticity of Volatility

Asymmetric patterns in volatility has been subject of an extensive literature (Engle and Patton, 2001). In the interest rate context, such studies has been motivated by the zero bound considerations. As a simple extension, we specify the following multivariate model that displays dependency of the elasticity of volatility on the direction of the move:

$$\Delta r_{it} = \mu_i + \sigma_i r_{it}^{\gamma(\Delta r_{it-1})} \epsilon_{it}, \quad \epsilon_{it} \sim t_{\nu_i}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \tag{6.1}$$

where $\gamma\left(\Delta r_{it}\right) = \begin{cases} \gamma_1, & \text{if } \Delta r_{it} \leq 0 \\ \gamma_2, & \text{if } \Delta r_{it} > 0. \end{cases}$ The model is estimated by Bayesian methods above, with the same priors. Priors for both downward and upward elasticity parameters, γ_1 and γ_2, are

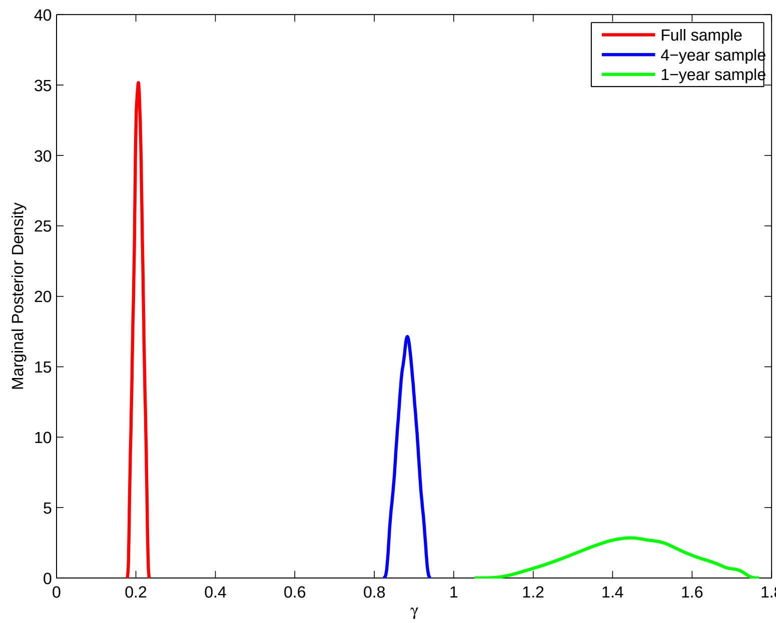

FIGURE 5. Posterior distributions of common elasticity of volatility of US dollar interest rate swap rates.

uniform on $[-1, 2]$. The posterior densities of the two elasticities are shown in Figure 6, while the posterior distributions for other parameters are omitted for brevity since they are similar to those in section 4. The posterior distributions are indeed distinct for all sub-samples. Of the three sub-samples, only in the shortest sub-sample the downward elasticity is such that the zero lower bound is unambiguously respected. In the two longer sub-samples, the magnitudes of elasticities switch ranking.

7. Variable Elasticity of Volatility Model

7.1. The Model. Changes in elasticity estimators over sub-samples dominated by low rates as well as non-parametric evidence in a later section 11 suggests that modeling elasticity of volatility as a constant throughout the entire time-series range is perhaps too restrictive. A common pattern emerging from non-parametric analysis of credit spread series is that elasticity tends to rise with spread level, but remains range-bound. Some kind of sigmoid link function would be able to capture this shape. In particular, we posit the following representation of the variable elasticity of volatility:[20]

$$\gamma(r) = \frac{3}{\pi} \operatorname{atan}(\beta_0 + \beta_1 r) + \frac{1}{2}. \tag{7.1}$$

[20]While the use of arc-tangent function may seem esoteric, this link function has appeared in the generalized linear models (Hardin and Hilbe, 2012; Ramalho, Ramalho, and Murteira, 2010) under the name of "cauchit" as it is related to the CDF of the Cauchy distribution. Our initial experience with the more standard, and less heavy-tailed, logistic alternative has shown it to be prone to numerical overflows for this particular model and dataset.

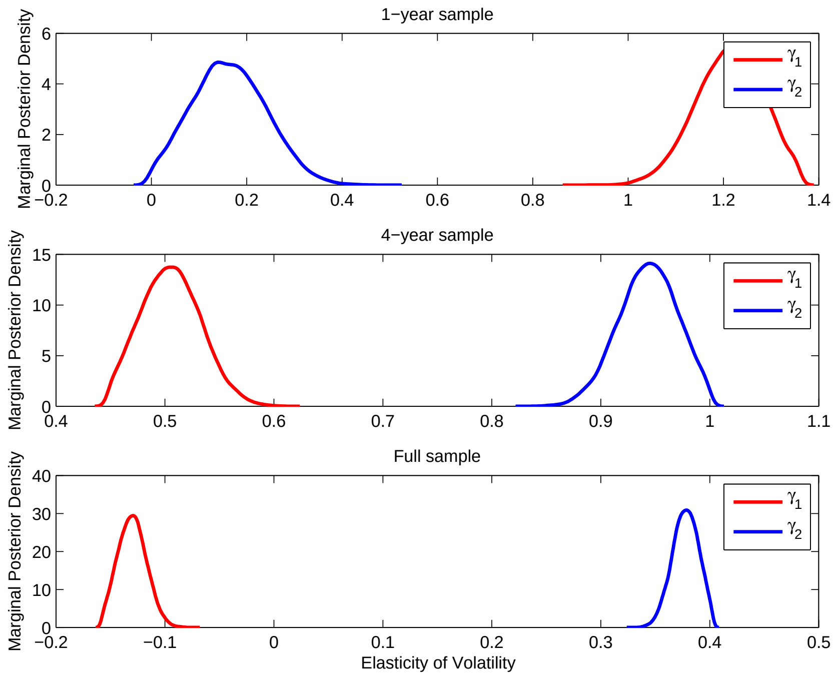

FIGURE 6. Posterior distributions of downward and upward elasticities of volatility of US dollar interest rate swap rates.

Scaling factor $3/\pi$ is there to ensure that $\gamma \in (-1, 2)$. With (7.1), the statistical model for a single time-series of interest becomes

$$\Delta r_t = \mu + \sigma r_{t-1}^{\gamma(r)} \epsilon_t. \tag{7.2}$$

We continue assuming Student t distribution for the error term ϵ_t.

The model is estimated by Bayesian approach. Prior assumptions regarding drift, volatility and degrees of freedom parameters remain the same as in 5.2. Additionally, we specify independent vague Gaussian priors for β_0 and β_1:

$$\beta_0 \sim \mathcal{N}\left(m_{\beta_0}, \sigma^2_{\beta_0}\right), \tag{7.3}$$

$$\beta_1 \sim \mathcal{N}\left(m_{\beta_1}, \sigma^2_{\beta_1}\right), \tag{7.4}$$

with $m_{\beta_0} = m_{\beta_1} = 0$ and $\sigma_{\beta_0} = \sigma_{\beta_1} = 100$.

7.2. Posterior Inference. Estimation results are in Table 19. Posterior distributions of new parameters β_0 and β_1 fall into three categories. The first category of estimates contain posterior distributions for short rates (up to 4-year maturity) using the full sample of data. This category is characterized by the entirely negative posterior for β_0 whereas the posterior for β_1 lies wholly in the positive territory. The signs of β_0 and β_1 are reversed in the second category which contains longer maturities over the full sample and all maturities over the four-year sample. The third category comprises all estimates using one-year sample with posterior distributions for both parameters having significant mass on both sides of zero, indicating that variable elasticity of

Swap Maturity	Parameter	Full sample		4-year sample		1-year sample	
		Posterior Median	Posterior 95% Bounds	Posterior Median	Posterior 95% Bounds	Posterior Median	Posterior 95% Bounds
1y	μ	-0.001	(-0.002, -0.000)	-0.001	(-0.002, 0.000)	0.000	(-0.002, 0.002)
	σ	0.005	(0.005, 0.005)	0.033	(0.029, 0.037)	0.038	(0.015, 0.048)
	β_0	-28.586	(-37.503,-21.072)	0.828	(0.524, 1.211)	-4.956	(-9.747, -0.509)
	β_1	28.640	(21.174, 37.664)	-0.204	(-0.389, -0.030)	12.728	(0.129, 24.378)
	ν	2.006	(2.000, 2.030)	5.271	(3.961, 7.376)	10.185	(4.367, 90.383)
2y	μ	-0.003	(-0.004, -0.001)	-0.001	(-0.003, 0.001)	0.000	(-0.002, 0.003)
	σ	0.006	(0.006, 0.006)	0.031	(0.029, 0.034)	0.034	(0.011, 0.047)
	β_0	-28.291	(-37.464,-20.658)	0.734	(0.403, 1.122)	-1.225	(-8.210, 11.741)
	β_1	28.427	(20.875, 37.540)	-0.116	(-0.280, 0.056)	4.838	(-9.653, 20.100)
	ν	2.005	(2.000, 2.027)	5.038	(3.802, 7.027)	4.587	(2.733, 9.125)
3y	μ	-0.002	(-0.003, -0.000)	-0.001	(-0.003, 0.001)	0.000	(-0.003, 0.003)
	σ	0.005	(0.005, 0.006)	0.029	(0.027, 0.032)	0.035	(0.025, 0.041)
	β_0	-26.335	(-35.197,-18.965)	0.694	(0.344, 1.096)	3.579	(-6.995, 20.666)
	β_1	26.426	(18.921, 35.417)	-0.087	(-0.232, 0.059)	4.779	(-10.466, 21.435)
	ν	2.005	(2.000, 2.027)	5.065	(3.819, 7.077)	4.367	(2.688, 8.407)
4y	μ	-0.003	(-0.005, -0.002)	-0.002	(-0.004, 0.001)	-0.001	(-0.004, 0.002)
	σ	0.005	(0.004, 0.005)	0.027	(0.025, 0.030)	0.028	(0.024, 0.033)
	β_0	-25.135	(-34.437,-17.399)	0.780	(0.353, 1.281)	6.205	(-6.754, 23.463)
	β_1	25.212	(17.520, 34.446)	-0.111	(-0.261, 0.031)	6.190	(-8.361, 22.902)
	ν	2.005	(2.000, 2.027)	5.437	(4.075, 7.695)	5.764	(3.297, 13.638)
5y	μ	-0.003	(-0.005, -0.001)	-0.002	(-0.005, 0.001)	-0.002	(-0.005, 0.002)
	σ	0.026	(0.024, 0.029)	0.024	(0.021, 0.028)	0.023	(0.020, 0.026)
	β_0	0.988	(0.798, 1.178)	0.968	(0.394, 1.657)	4.934	(-10.376, 22.873)
	β_1	-0.210	(-0.239, -0.180)	-0.148	(-0.323, 0.008)	6.999	(-5.580, 23.310)
	ν	5.579	(4.597, 6.870)	5.522	(4.090, 7.908)	8.725	(4.115, 64.043)
7y	μ	-0.003	(-0.005, -0.001)	-0.002	(-0.006, 0.001)	-0.003	(-0.008, 0.003)
	σ	0.024	(0.021, 0.028)	0.022	(0.018, 0.029)	0.017	(0.015, 0.020)
	β_0	0.925	(0.688, 1.166)	0.984	(0.140, 1.959)	3.933	(-11.765, 22.577)
	β_1	-0.185	(-0.218, -0.154)	-0.138	(-0.351, 0.038)	6.897	(-3.652, 23.477)
	ν	5.777	(4.774, 7.173)	6.092	(4.415, 9.126)	46.685	(7.705,125.361)
10y	μ	-0.003	(-0.005, -0.001)	-0.002	(-0.006, 0.002)	-0.003	(-0.009, 0.003)
	σ	0.024	(0.020, 0.031)	0.020	(0.014, 0.033)	0.012	(0.011, 0.015)
	β_0	0.779	(0.444, 1.078)	1.068	(-0.052, 2.470)	3.704	(-12.551, 22.046)
	β_1	-0.158	(-0.194, -0.121)	-0.158	(-0.430, 0.033)	6.995	(-2.901, 22.637)
	ν	5.308	(4.427, 6.496)	5.448	(4.017, 7.889)	54.923	(8.331,125.985)
30y	μ	-0.002	(-0.004, -0.000)	-0.003	(-0.007, 0.001)	-0.003	(-0.010, 0.005)
	σ	0.047	(0.028, 0.120)	0.015	(0.011, 0.032)	0.008	(0.007, 0.010)
	β_0	0.073	(-2.082, 0.622)	2.197	(0.497, 4.001)	3.331	(-14.918, 22.016)
	β_1	-0.124	(-0.162, 0.004)	-0.411	(-0.752, -0.141)	7.377	(-2.185, 22.801)
	ν	5.730	(4.745, 7.125)	5.164	(3.877, 7.333)	25.683	(5.929,120.255)

TABLE 19. Bayesian estimation results for variable elasticity of volatility model.

volatility is not necessary over such a short estimation window. Over that window, large values of the Student t shape parameter ν for the long end of the curve also confirm that non-normal fat tails are probably not required.

Typical shapes of elasticity of volatility are shown in the upper panels of Figure 7 where we selected representative examples from the first two categories. Namely, we used posterior medians estimated over the full sample for 1-year and 5-year interest rates swaps respectively.

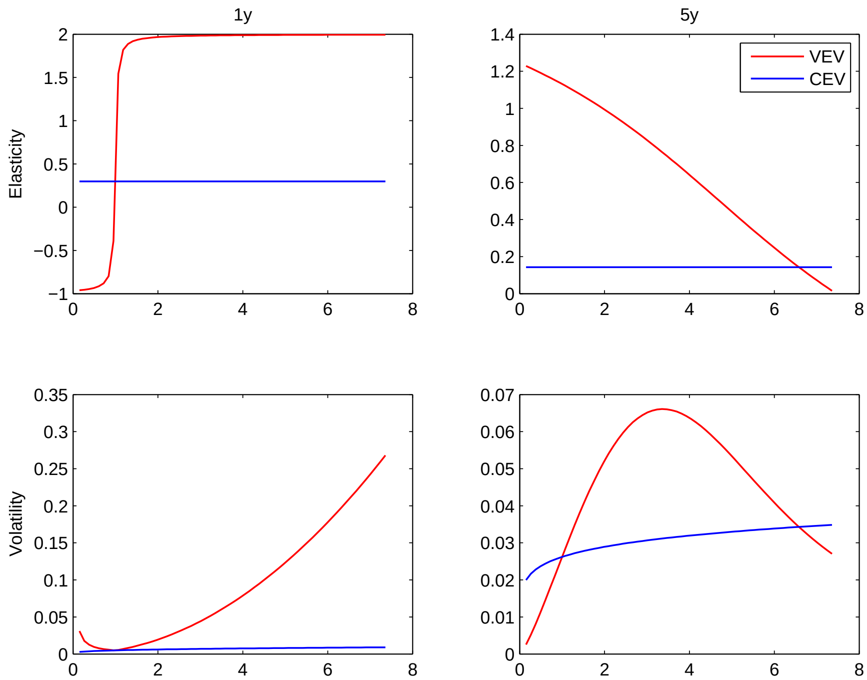

FIGURE 7. Typical estimated elasticity and volatility profiles.

The lower panels translate estimated elasticities into implied profiles of volatility function, using median σ to scale it into the right units. The volatility profile is non-monotone in both cases since elasticity eventually becomes negative for either very low rates or for very high rates. This may be an undesirable feature of our parameterization of the sigmoid elasticity function with forced negative asymptote. This feature is particularly pernicious for the first category where elasticity turns negative for very low rates since this may magnify the possibility of negative rates. A more flexible specification may do better by making sigmoid bounds free parameters. Alternatively, we can specify piecewise volatility function with different profiles for low and high rates. A version of the latter idea is pursued in section 8. For the second category of posterior distributions, negative elasticity at high rates is much less of a problem. The profile in the bottom right panel offers some support for shifting the risk form from returns to differences at about 3% interest rate level. The switch from returns to differences is in line with the evidence of section 3.3.

7.3. Model Comparison. Comparison of two models in the Bayesian framework is typically done via the posterior odds ratio. When the prior odds ratio is set to 1, as is done in much of the Bayesian literature in the absence of strong preferences for a particular model and also in the present paper, the log posterior odds ratio is given by the log Bayes factor, that is the difference of marginal log-likelihoods of the two competing models.

$$\log \frac{p(M_0|y)}{p(M_1|y)} = \log \frac{p(y|M_0)}{p(y|M_1)} + \log \frac{\pi(M_0)}{\pi(M_1)} = \log \frac{p(y|M_0)}{p(y|M_1)}. \tag{7.5}$$

Swap Maturity	Full sample	4-year sample	1-year sample
1y	1244.410	2.903	3.284
2y	1244.184	4.270	6.054
3y	1162.967	3.959	5.768
4y	1113.342	3.824	5.560
5y	-88.976	3.687	6.230
7y	-48.814	4.481	7.037
10y	-20.805	4.269	6.679
30y	0.423	1.059	7.156

Table 20. Log Bayes factors for constant elasticity of volatility model (M_0) against variable elasticity alternative (M_1).

To calculate the marginal likelihood, we follow Chib (1995) by using Bayes' theorem

$$\log p(y|M_k) = \log p(y|\theta, M_k) + \log p(\theta|M_k) - \log p(\theta|y, M_k), \tag{7.6}$$

where M_k denotes k^{th} model, $k \in \{0, 1\}$, whereas θ is a high density point in the parameter space, such as the posterior median. The first term on the right hand side of (7.6) is the standard log-likelihood, the second term is the log prior density, while the third term involves the posterior density of parameters. Since only the sample from the posterior density is available, we make use of a multivariate kernel density estimate to approximate the last term, following Kim, Shephard, and Chib (1998).

Table 20 presents log Bayes factors for comparison of Bayesian constant elasticity model with Student t residuals from section 4 against the sigmoid variable elasticity model currently discussed. To interpret log Bayes factors, we use Jeffreys's (1961) scale. Accordingly, the evidence in the full sample is decisive in favor of CEV model for maturities up to 4-year, decisive in favor of variable elasticity model for 5-year to 10-year maturities and barely worth mentioning in favor of CEV for 30-year maturities. In the four year sample, the evidence in favor of CEV is substantial for 1-year maturity, strong for 3-5 year maturities, barely worth mentioning for 30-year maturity and very strong in the remaining cases. In the one year sample, the evidence in favor of CEV is strong for 1-year maturity and decisive for all others. We can therefore conclude that sigmoid variable elasticity model is only a successful extension for longer maturities in full sample and is not needed in shorter subsamples. If variable elasticity is important for shorter maturities or shorter subsamples, we are advised to find better ways to account for the potential decline of elasticity of volatility with the level of interest rates.

8. Hybrid Piecewise Volatility Profiles

8.1. The Model.

Kainth, Kwiatkowski, and Muirden (2010), inspired by Rebonato, Mahal, Joshi, Buchholz, and Nyholm (2005), posit the following piecewise CEV-type process:

$$\Delta r_t = \mu + \sigma H\left(r_{t-1}; r_L, r_U, \lambda, \gamma\right) \epsilon_t, \tag{8.1}$$

where $H\left(r_{t-1}; r_L, r_U, \lambda, \gamma\right) = \left(\frac{r_{t-1}}{r_L}\right)^{\gamma} \mathbb{1}_{r_{t-1}<r_L} + \mathbb{1}_{r_L \le r_{t-1} < r_U} + \left(1 + \lambda \cdot (r_{t-1} - r_U)\right) \mathbb{1}_{r_U \le r_{t-1}}$. We add to (8.1) an assumption concerning error terms, $\epsilon_t \sim t_\nu$. A hypothetical volatility profile implied by (8.1) is shown in Figure 8.

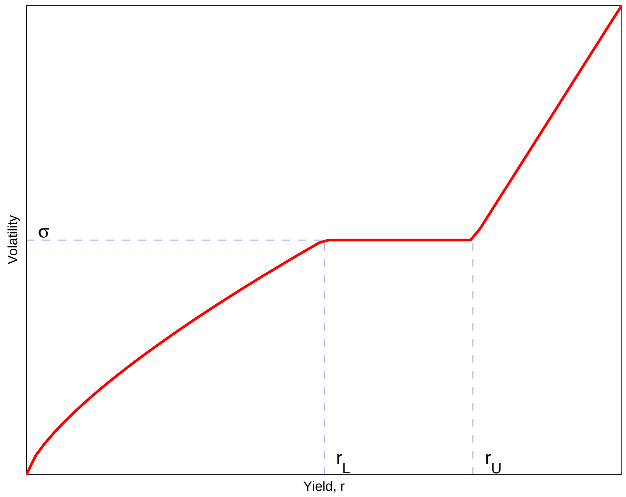

FIGURE 8. A hypothetical hybrid piecewise volatility profile.

Upper cutoff, r_U is identified if $\lambda \neq 0$, and larger λ makes identification easier. To ensure $r_U > r_L$, we set $r_U = r_L + \exp(\delta)$. In the absence of a priori knowledge about new parameters, we specify vague independent normal priors for $\log(r_L)$, δ and λ. Other priors are as before.

To conserve space and to help identification of cutoff parameters,[21] we constrained parameters that control the shape of the volatility profiles, namely r_L, r_U, γ and λ to be the same across the entire term structure. Bayesian posterior inference is presented in Table 21.

8.2. Posterior Inference. Using full sample, the distance between estimated r_L and r_U is very small, suggesting that the middle regime is not necessary. The slope of the relationship between volatility and swap rate in the third regime is negative which is also illustrated in posterior median volatility profile in Figure 9. This finding is in agreement with the variable elasticity of volatility results in Table 19 where posterior for the slope parameter β_1 lies unambiguously below zero. Both models therefore agree that for the full sample, the swap rates are best modeled in returns below the cutoff of about 280 basis points and as simple differences above it. For more recent subsamples, swap rates at most maturities do not reach that cutoff level. As a result sub-sample-specific cutoffs fall significantly so that most of the sample is located above the cutoff.[22] Using 4-years sample, piecewise model suggests a proportional law specification while sigmoid-type model of section 7 is somewhat less conclusive if slope parameter β_1 is maturity-specific since for some maturities and for sufficiently high rates during the last 4-years the elasticity drops close to zero. Imposing the same β_1 for all terms results in negative β_1 (not shown), which overall suggests percentage change representation. Finally, we conclude that the swap rate risks in the

[21]Cutoff parameters r_L and, especially, r_U are difficult to identify because of scarcity of data corresponding to the right-most section of the profile and because of small estimated values of λ for the full sample. If these parameters are left unconstrained across maturities, the estimates are close for maturities with wide span of rates, in agreement with Rebonato and Podugin (2010).

[22]Extremely wide posterior distribution of elasticity γ in the first regime is also a consequence of this since hardly any observation falls into that regime.

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	-0.001	(-0.002,0)	-0.001	(-0.002,0)	0.000	(-0.002,0.002)
	σ	0.047	(0.044,0.049)	0.002	(0.001,0.004)	0.002	(0.001,0.004)
	ν	3.708	(3.233,4.296)	5.389	(4.063,7.534)	6.634	(3.558,18.151)
2y	μ	-0.001	(-0.003,0)	-0.001	(-0.003,0.001)	0.000	(-0.002,0.003)
	σ	0.062	(0.059,0.066)	0.002	(0.001,0.004)	0.002	(0.001,0.003)
	ν	4.839	(4.083,5.823)	5.231	(3.938,7.322)	4.540	(2.763,8.949)
3y	μ	-0.002	(-0.004,0)	-0.001	(-0.003,0.001)	0.000	(-0.003,0.003)
	σ	0.066	(0.062,0.069)	0.002	(0.001,0.004)	0.002	(0.001,0.003)
	ν	5.347	(4.481,6.531)	5.151	(3.858,7.267)	4.198	(2.631,7.707)
4y	μ	-0.002	(-0.004,-0.001)	-0.002	(-0.004,0.001)	-0.001	(-0.004,0.002)
	σ	0.067	(0.064,0.07)	0.002	(0.001,0.004)	0.001	(0.001,0.003)
	ν	5.483	(4.564,6.702)	5.485	(4.061,7.777)	4.914	(2.945,9.664)
5y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.005,0.001)	-0.002	(-0.006,0.002)
	σ	0.069	(0.065,0.072)	0.002	(0.001,0.003)	0.001	(0.001,0.003)
	ν	5.441	(4.535,6.649)	5.555	(4.071,8.042)	6.200	(3.426,16.543)
7y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.006,0.002)	-0.003	(-0.009,0.002)
	σ	0.069	(0.065,0.072)	0.002	(0.001,0.003)	0.001	(0.001,0.003)
	ν	5.553	(4.607,6.846)	6.231	(4.437,9.432)	15.628	(5.222,113.934)
10y	μ	-0.003	(-0.005,-0.001)	-0.002	(-0.006,0.002)	-0.004	(-0.01,0.002)
	σ	0.068	(0.065,0.072)	0.001	(0.001,0.003)	0.001	(0.001,0.003)
	ν	5.142	(4.303,6.245)	5.462	(4.012,7.903)	27.290	(5.604,122.556)
30y	μ	-0.002	(-0.004,-0.001)	-0.002	(-0.006,0.002)	-0.004	(-0.011,0.004)
	σ	0.066	(0.063,0.07)	0.001	(0.001,0.002)	0.001	(0.001,0.003)
	ν	5.762	(4.755,7.155)	4.433	(3.38,6.021)	14.513	(4.931,110.844)
	r_L	2.867	(2.688,2.976)	0.000	(0,0.044)	0.001	(0,0.156)
	r_U	2.900	(2.774,3.158)	0.002	(0,0.063)	0.015	(0,0.194)
	λ	-0.124	(-0.133,-0.116)	14.205	(6.925,26.772)	17.651	(7.951,30.363)
	γ	0.725	(0.685,0.766)	0.505	(-0.924,1.909)	0.507	(-0.925,1.926)

TABLE 21. Bayesian estimation results for hybrid piecewise elasticity of volatility model.

most recent year are best represented as percentage changes in rates. Since the degrees of freedom parameters indicate only a moderate departure from normality, the added benefit of rates following a proportional law is a virtual guarantee that rates will remain positive.

8.3. Model Comparison. We applied Bayesian model selection methodology, described in section 7.3, to discriminate constant elasticity model of section 4.4 (M_0) against the current piecewise model (M_1). The log Bayes factors are shown in Table 22. Here, we find decisive evidence in favor of piecewise model for all maturities in the long sample except 30-year maturity. This resolves the curious issue we observed with the first group volatility profiles described in section 7.2 as driven by bounds on sigmoid elasticity function that are at odds with the data. For shorter subsamples, variable elasticity of volatility, as before, is unnecessary.

9. Constant Elasticity of Volatility with Stochastic Volatility Scale

The analysis thus far has not considered the possibility that volatility has its own source of uncertainty unrelated to the level of the series. Indeed, Andersen and Lund (1997), Ball and

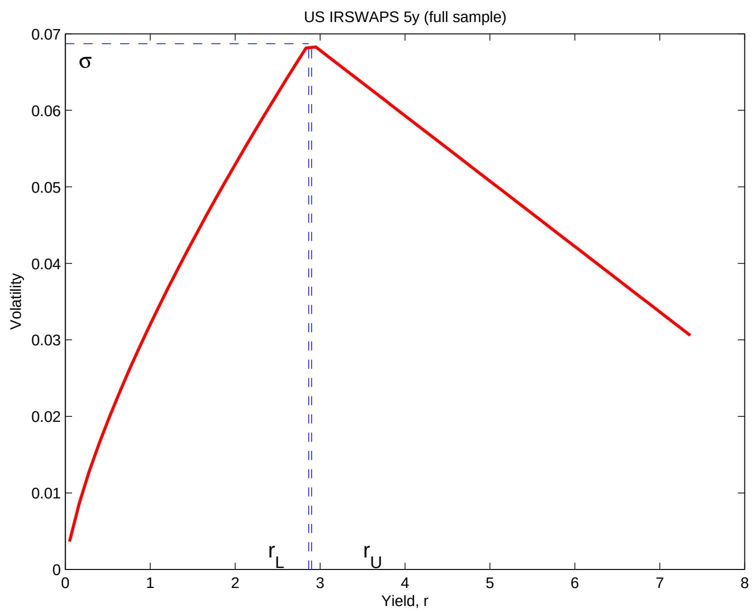

FIGURE 9. Hybrid piecewise volatility profile for 5-year swap maturity using full sample.

Swap Maturity	Full sample	4-year sample	1-year sample
1y	-100.883	13.532	15.079
2y	-135.886	11.547	13.972
3y	-127.043	12.224	13.021
4y	-102.355	13.460	16.148
5y	-76.828	13.762	16.995
7y	-34.559	15.048	16.478
10y	-6.642	16.884	14.656
30y	5.826	37.500	13.474

TABLE 22. Log Bayes factors for constant elasticity of volatility model (M_0) against hybrid piecewise alternative (M_1).

Torous (1999) and Trolle and Schwartz (2009) all find evidence of stochastic volatility in the interest rates.

To accommodate the stochastic volatility, we follow Andersen and Lund (1997) and Eraker (2001) and specify the following set of independent univariate models,[23]

$$\begin{aligned} \Delta r_{it} &= \mu_{ri} + \sigma_{it} r_{it}^{\gamma_i} \epsilon_{it}, & \epsilon_{it} &\sim t_{\nu_i}, \\ \log \sigma_{it} &= \mu_{\sigma i} + \rho_i \left(\log \sigma_{it-1} - \mu_{\sigma i}\right) + \sigma_{\eta i} \eta_{it}, & \eta_{it} &\sim \mathcal{N}\left(0, 1\right), \end{aligned} \tag{9.1}$$

[23]The model (9.1) also bears similarity to SABR framework of Hagan, Kumar, Lesniewski, and Woodward (2002) except that our formulation is in discrete time, is in the actuarial (physical) measure, does not have unit root in log-volatility and applies to the whole of the term structure. Relaxing unit root assumption is important as could be confirmed by improvement in DIC (not shown).

as well as two multivariate versions that constrain time evolution of volatility across different maturities in slightly different ways:

$$\begin{aligned} \Delta r_{it} &= \mu_{ri} + \bar{\sigma}_i \sigma_t r_{it}^{\gamma_i} \epsilon_{it}, & \epsilon_{it} &\sim \mathcal{N}(0,1), \\ \log \sigma_t &= \mu_\sigma + \rho(\log \sigma_{t-1} - \mu_\sigma) + \sigma_\eta \eta_t, & \eta_t &\sim \mathcal{N}(0,1), \end{aligned} \tag{9.2}$$

$$\begin{aligned} \Delta r_{it} &= \mu_{ri} + \bar{\sigma}_i \sigma_t r_{it}^{\gamma} \epsilon_{it}, & \epsilon_{it} &\sim \mathcal{N}(0,1), \\ \log \sigma_t &= \mu_\sigma + \rho(\log \sigma_{t-1} - \mu_\sigma) + \sigma_\eta \eta_t, & \eta_t &\sim \mathcal{N}(0,1). \end{aligned} \tag{9.3}$$

To ensure identification, we set $\bar{\sigma}_1 = 1$ for both panel models. Both multivariate models dispense with flexibility to capture fat tails in both state and measurement equations since, as the univariate model results below indicate, the evidence of fat tails is weak, consistently with improvements in fit due to combining stochastic volatility and CEV reported in Eraker (2001).

Priors for the new stochastic volatility parameters μ_σ, ρ and σ_η are similar to those frequently used in stochastic volatility literature (e.g., Yu (2005)) and are only slightly informative:

$$\mu_\sigma \sim \mathcal{N}(0,100), \qquad \rho \sim \mathcal{U}(-1,1), \qquad \sigma_\eta^2 \sim \mathcal{IG}(2.5, 0.025), \tag{9.4}$$

independently of each other and other parameters. Error terms ϵ_t and η_τ are uncorrelated at all leads and lags. This model belongs to the class of non-linear non-Gaussian state-space time-series models since the volatility dynamics are not observed. Some technical details related to the model estimation can be found in appendix D.3. Estimation results are reported in tables 23, 24 and 25.

9.1. Univariate Constant Elasticity of Stochastic of Volatility (CEV-SV) Model. If there are no cross-maturity constraints, the stochastic volatility is highly persistent, in line with much of the stochastic volatility literature. Estimates of elasticity of volatility are substantially similar across maturities, with wider posterior confidence intervals, and are only slightly higher than the ones reported in section 4, with all the same patterns across different sub-samples or maturities. Thus, incorporating stochastic volatility does not overturn our earlier conclusions. Across-the-board upward shift in the posterior distribution of the degrees of freedom parameter is also expected based on results of Eraker (2001). Since the Student t distribution is itself a scale mixture of normal distributions, having two different mixtures does not seem necessary.

9.2. Multivariate CEV Model with Common Stochastic Volatility. Next, we constrain stochastic elasticity process to be the same across the curve while letting other parameter vary freely. The most notable change we find is that the estimated persistence of volatility becomes much lower as the common stochastic volatility has to soak up some of the non-parallel moves of the yield curve. Posterior distributions of the elasticity of volatility are a bit wider than those of models without stochastic volatility, but they are centered nearly spot on the same values. Compared to the univariate models with stochastic volatility, elasticity estimates move back lower.

9.3. Multivariate Model with Common Elasticity and Stochastic Volatility. Finally, if we make all sources of time-variation across different maturities the same, and only the mean and scale parameter to vary in the cross-section, the results remain similar with respect to all model parameters except elasticities. As far as elasticity parameters are concerned, the full sample estimate of common γ resembles those from the short end of the curve, while the estimates from shorter sub-samples are more akin to those at the long end. The magnitude pattern across sub-samples remains consistent with earlier results.

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	-0.000	(-0.001, 0.001)	-0.001	(-0.002, 0.000)	0.000	(-0.002, 0.002)
	ν	6.311	(5.094, 8.187)	16.713	(6.636, 92.984)	49.651	(7.169, 99.189)
	γ	0.353	(0.152, 0.595)	0.901	(0.693, 1.092)	0.507	(-0.439, 1.458)
	μ_h	-3.810	(-4.099, -3.560)	-3.469	(-3.629, -3.330)	-3.825	(-4.464, -3.192)
	ρ	0.985	(0.974, 0.993)	0.897	(0.755, 0.954)	0.650	(-0.283, 0.913)
	σ_η	0.087	(0.073, 0.104)	0.153	(0.107, 0.239)	0.252	(0.141, 0.419)
2y	μ	-0.001	(-0.002, 0.000)	-0.001	(-0.003, 0.000)	-0.000	(-0.003, 0.002)
	ν	8.691	(6.563, 12.535)	18.537	(6.769, 93.830)	50.198	(6.854, 99.334)
	γ	0.365	(0.147, 0.616)	0.935	(0.733, 1.122)	0.874	(-0.236, 1.830)
	μ_h	-3.488	(-3.810, -3.209)	-3.416	(-3.546, -3.305)	-3.561	(-4.192, -3.015)
	ρ	0.986	(0.976, 0.994)	0.874	(0.654, 0.944)	0.693	(0.284, 0.904)
	σ_η	0.077	(0.065, 0.094)	0.164	(0.111, 0.283)	0.287	(0.164, 0.455)
3y	μ	-0.001	(-0.003, 0.000)	-0.002	(-0.004, 0.000)	-0.000	(-0.003, 0.002)
	ν	9.825	(7.214, 14.870)	16.067	(6.671, 91.496)	47.701	(5.697, 99.455)
	γ	0.364	(0.127, 0.654)	0.938	(0.718, 1.149)	1.469	(0.444, 1.971)
	μ_h	-3.416	(-3.799, -3.092)	-3.484	(-3.632, -3.345)	-3.392	(-3.785, -3.124)
	ρ	0.986	(0.977, 0.994)	0.894	(0.769, 0.953)	0.683	(0.242, 0.901)
	σ_η	0.075	(0.063, 0.090)	0.155	(0.107, 0.244)	0.292	(0.164, 0.465)
4y	μ	-0.002	(-0.004, 0.000)	-0.003	(-0.005, 0.000)	-0.001	(-0.004, 0.002)
	ν	11.901	(8.259, 21.238)	29.109	(8.382, 96.952)	46.093	(5.885, 99.195)
	γ	0.326	(0.065, 0.628)	0.902	(0.647, 1.171)	1.713	(1.002, 1.986)
	μ_h	-3.337	(-3.744, -2.991)	-3.519	(-3.715, -3.329)	-3.518	(-3.685, -3.367)
	ρ	0.984	(0.974, 0.992)	0.915	(0.824, 0.962)	0.544	(-0.170, 0.866)
	σ_η	0.076	(0.063, 0.091)	0.141	(0.101, 0.207)	0.268	(0.154, 0.418)
5y	μ	-0.002	(-0.004, -0.000)	-0.003	(-0.006, -0.000)	-0.002	(-0.006, 0.002)
	ν	14.386	(9.193, 33.172)	32.572	(8.926, 97.744)	45.582	(6.826, 99.050)
	γ	0.295	(0.013, 0.621)	0.884	(0.592, 1.174)	1.740	(1.113, 1.989)
	μ_h	-3.265	(-3.723, -2.888)	-3.564	(-3.819, -3.318)	-3.721	(-3.862, -3.570)
	ρ	0.982	(0.971, 0.991)	0.917	(0.825, 0.962)	0.214	(-0.610, 0.770)
	σ_η	0.078	(0.066, 0.095)	0.141	(0.101, 0.205)	0.240	(0.142, 0.379)
7y	μ	-0.002	(-0.004, -0.000)	-0.004	(-0.007, -0.000)	-0.003	(-0.008, 0.002)
	ν	22.656	(11.899, 82.931)	49.456	(12.451, 99.345)	55.230	(11.118, 99.546)
	γ	0.151	(-0.158, 0.472)	0.775	(0.379, 1.181)	1.618	(0.941, 1.976)
	μ_h	-3.074	(-3.547, -2.635)	-3.554	(-3.966, -3.162)	-3.972	(-4.188, -3.630)
	ρ	0.979	(0.966, 0.989)	0.936	(0.872, 0.971)	-0.046	(-0.716, 0.665)
	σ_η	0.078	(0.065, 0.095)	0.124	(0.093, 0.172)	0.201	(0.125, 0.323)
10y	μ	-0.003	(-0.004, -0.001)	-0.004	(-0.008, -0.000)	-0.004	(-0.010, 0.002)
	ν	26.780	(12.582, 92.099)	54.791	(14.564, 99.487)	56.355	(10.729, 99.942)
	γ	0.001	(-0.326, 0.399)	0.609	(0.069, 1.177)	1.542	(0.739, 1.973)
	μ_h	-2.876	(-3.461, -2.383)	-3.456	(-4.117, -2.846)	-4.172	(-4.518, -3.572)
	ρ	0.976	(0.962, 0.987)	0.953	(0.911, 0.981)	-0.066	(-0.694, 0.716)
	σ_η	0.081	(0.068, 0.099)	0.115	(0.089, 0.153)	0.204	(0.128, 0.329)
30y	μ	-0.002	(-0.004, -0.001)	-0.003	(-0.007, 0.001)	-0.004	(-0.011, 0.003)
	ν	37.290	(14.765, 96.539)	57.428	(15.238, 99.564)	53.915	(9.250, 99.428)
	γ	-0.401	(-0.736, -0.045)	-0.030	(-0.734, 0.727)	1.437	(0.357, 1.964)
	μ_h	-2.323	(-2.892, -1.787)	-2.757	(-3.758, -1.819)	-4.348	(-4.904, -3.242)
	ρ	0.968	(0.949, 0.981)	0.959	(0.921, 0.984)	0.128	(-0.629, 0.834)
	σ_η	0.084	(0.070, 0.104)	0.112	(0.086, 0.151)	0.211	(0.128, 0.344)

TABLE 23. Bayesian estimation results for individual constant elasticity of stochastic of volatility (CEV-SV) models.

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	-0.000	(-0.001, 0.000)	-0.002	(-0.003, -0.001)	0.001	(-0.001, 0.003)
	$\bar{\sigma}$	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	γ	0.254	(0.210, 0.298)	0.730	(0.628, 0.835)	1.871	(1.495, 1.995)
2y	μ	-0.001	(-0.002, -0.000)	-0.003	(-0.004, -0.002)	0.000	(-0.002, 0.002)
	$\bar{\sigma}$	1.175	(1.114, 1.241)	0.933	(0.863, 1.008)	0.517	(0.344, 0.801)
	γ	0.281	(0.230, 0.332)	0.636	(0.519, 0.752)	0.917	(0.218, 1.648)
3y	μ	-0.002	(-0.003, -0.001)	-0.004	(-0.005, -0.002)	-0.001	(-0.003, 0.001)
	$\bar{\sigma}$	1.261	(1.181, 1.346)	0.898	(0.822, 0.981)	0.492	(0.364, 0.684)
	γ	0.238	(0.180, 0.298)	0.619	(0.504, 0.741)	1.041	(0.307, 1.755)
4y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.006, -0.003)	-0.002	(-0.004, 0.000)
	$\bar{\sigma}$	1.374	(1.266, 1.490)	0.883	(0.789, 0.986)	0.448	(0.373, 0.581)
	γ	0.173	(0.102, 0.243)	0.597	(0.465, 0.726)	1.317	(0.602, 1.905)
5y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.007, -0.003)	-0.003	(-0.006, 0.000)
	$\bar{\sigma}$	1.534	(1.390, 1.698)	0.921	(0.804, 1.063)	0.409	(0.334, 0.536)
	γ	0.110	(0.026, 0.190)	0.514	(0.362, 0.658)	1.162	(0.445, 1.805)
7y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.007, -0.003)	-0.004	(-0.007, 0.000)
	$\bar{\sigma}$	1.899	(1.667, 2.170)	1.051	(0.867, 1.275)	0.428	(0.289, 0.650)
	γ	-0.048	(-0.150, 0.052)	0.358	(0.180, 0.538)	0.823	(0.104, 1.550)
10y	μ	-0.003	(-0.004, -0.001)	-0.006	(-0.008, -0.003)	-0.006	(-0.010, -0.001)
	$\bar{\sigma}$	2.425	(2.057, 2.863)	1.266	(0.984, 1.646)	0.378	(0.224, 0.690)
	γ	-0.190	(-0.310, -0.073)	0.213	(-0.009, 0.427)	0.947	(0.177, 1.650)
30y	μ	-0.002	(-0.003, -0.001)	-0.006	(-0.008, -0.003)	-0.005	(-0.011, 0.000)
	$\bar{\sigma}$	4.687	(3.738, 5.897)	2.602	(1.787, 3.805)	0.365	(0.183, 0.883)
	γ	-0.574	(-0.726, -0.424)	-0.292	(-0.574, -0.006)	0.952	(0.110, 1.642)
	μ_σ	-3.775	(-3.828, -3.722)	-3.517	(-3.596, -3.433)	-2.941	(-3.198, -2.773)
	ρ	0.244	(0.203, 0.284)	0.226	(0.153, 0.300)	0.164	(0.005, 0.318)
	σ_η	0.715	(0.694, 0.735)	0.632	(0.600, 0.667)	0.524	(0.468, 0.586)

Table 24. Bayesian estimation results for multivariate CEV model with common stochastic volatility.

9.4. Level-dependent or Stochastic Volatility? Figure 10 presents two comparisons of the estimated level-dependent and stochastic volatilities from univariate set of models (9.1), over the full sample, surrounded by fanning out posterior distributions to highlight the estimation uncertainty. The top plot corresponds to the short end of the yield curve, while the bottom plot corresponds to the long end. Both plots are shown in logs, so that the constant shift between the two fan charts in each plot can be used as the measure of relative strength for the two sources, while variability of the two over time can be used to indicate, on a forward-looking basis, the strength of the ability to evaluate risk. The relative strength is not identified in univariate models, so only the second distinction between the two sources can be assessed. The fanchart plot makes it apparent that CEV feature is mainly responsible for slowly evolving component of the rates volatility while the stochastic volatility tends to describe faster moving dynamic forces. The distinction becomes more pronounced at the intermediate maturities (not shown) and maxes out at the long end. The distinction becomes even more stark in the multivariate versions because of the diminished persistence of the stochastic volatility process (not shown).

The two sources of uncertainty can be combined to produce posterior predictive fancharts, similar to Cogley, Morozov, and Sargent (2005).

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	-0.001	(-0.001, 0.000)	-0.002	(-0.003, -0.001)	0.001	(-0.001, 0.003)
	$\bar{\sigma}$	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
2y	μ	-0.001	(-0.002, -0.000)	-0.003	(-0.005, -0.002)	0.000	(-0.002, 0.002)
	$\bar{\sigma}$	1.183	(1.139, 1.230)	0.952	(0.887, 1.022)	0.892	(0.782, 1.014)
3y	μ	-0.002	(-0.003, -0.001)	-0.004	(-0.006, -0.003)	-0.001	(-0.003, 0.001)
	$\bar{\sigma}$	1.196	(1.149, 1.246)	0.899	(0.827, 0.978)	0.743	(0.639, 0.862)
4y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.006, -0.003)	-0.002	(-0.004, 0.000)
	$\bar{\sigma}$	1.186	(1.136, 1.238)	0.859	(0.780, 0.944)	0.603	(0.496, 0.745)
5y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.007, -0.003)	-0.003	(-0.006, 0.000)
	$\bar{\sigma}$	1.202	(1.150, 1.257)	0.829	(0.746, 0.923)	0.522	(0.410, 0.682)
7y	μ	-0.002	(-0.004, -0.001)	-0.005	(-0.007, -0.003)	-0.004	(-0.007, 0.000)
	$\bar{\sigma}$	1.165	(1.112, 1.221)	0.788	(0.698, 0.888)	0.425	(0.309, 0.609)
10y	μ	-0.003	(-0.004, -0.002)	-0.006	(-0.008, -0.003)	-0.005	(-0.010, -0.001)
	$\bar{\sigma}$	1.166	(1.110, 1.226)	0.768	(0.673, 0.876)	0.356	(0.244, 0.547)
30y	μ	-0.002	(-0.003, -0.001)	-0.005	(-0.008, -0.003)	-0.005	(-0.011, 0.001)
	$\bar{\sigma}$	1.173	(1.113, 1.235)	0.768	(0.667, 0.885)	0.304	(0.194, 0.510)
	γ	0.364	(0.327, 0.401)	0.680	(0.604, 0.754)	1.412	(1.099, 1.688)
	μ_σ	-3.848	(-3.899, -3.796)	-3.531	(-3.605, -3.455)	-3.232	(-3.456, -3.026)
	ρ	0.269	(0.229, 0.309)	0.214	(0.141, 0.287)	0.157	(-0.001, 0.311)
	σ_η	0.723	(0.703, 0.745)	0.631	(0.599, 0.666)	0.521	(0.465, 0.583)

Table 25. Bayesian estimation results for common elasticity of stochastic volatility models.

10. Stochastic Elasticity of Volatility

Having explored possibilities of variable elasticity of volatility and having found rich diversity of shapes across sub-samples, perhaps we should allow the elasticity of volatility to have its own stochastic driver, in a manner akin to the stochastic volatility model explored in section 9.

$$\begin{aligned} \Delta r_t &= \mu_r + \sigma r_{t-1}^{\gamma_t} \epsilon_t, & \epsilon_t &\sim \mathcal{N}(0,1), \\ \gamma_t &= \frac{3}{2}\frac{e^{2z_t}-1}{e^{2z_t}+1} + \frac{1}{2}, \\ z_t &= \mu_z + \rho\,(z_{t-1}-\mu_z) + \sigma_\eta \eta_t, & \eta_t &\sim \mathcal{N}(0,1), \end{aligned} \tag{10.1}$$

where for simplicity and to relieve computational burden we model each tenor separately and assume Gaussian residual distributions. The second equation in the system is designed to limit the range of elasticities between -1 and 2. Models like this are potentially useful to appreciate the range of uncertainty associated with elasticity of volatility and to gauge direction of recent trends.

The outcomes of the MCMC estimation of posterior distributions are summarized in table 26 for each tenor and each sub-sample. The table is complemented with an example in Figure 11. In this and similar plots for other maturities and sample periods, the elasticity of volatility is highly persistent but is by no means close to constant as it vacillates widely within the posterior range reported in Table 17. Thus, the results are not particularly illuminating in explaining the driving forces behind the elasticity dynamics and extended examination of these may be warranted in the future research. For example, if we were to employ Student t residuals, the extent of elasticity

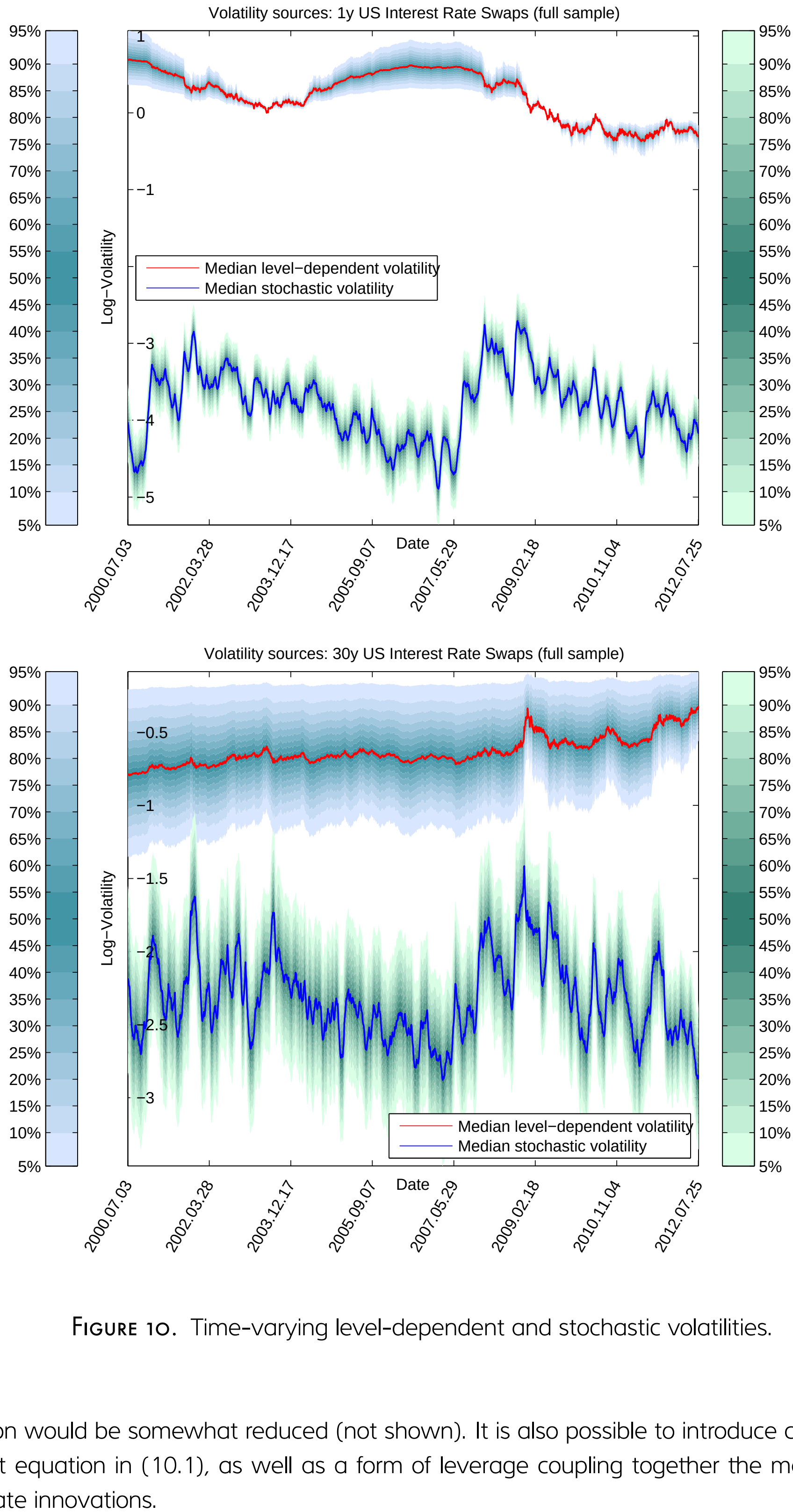

Figure 10. Time-varying level-dependent and stochastic volatilities.

variation would be somewhat reduced (not shown). It is also possible to introduce covariates in the last equation in (10.1), as well as a form of leverage coupling together the measurement and state innovations.

Swap Maturity	Parameter	Full sample Posterior Median	Full sample Posterior 95% Bounds	4-year sample Posterior Median	4-year sample Posterior 95% Bounds	1-year sample Posterior Median	1-year sample Posterior 95% Bounds
1y	μ	0.000	(-0.001, 0.001)	-0.000	(-0.001, 0.001)	0.001	(-0.002, 0.003)
	σ	0.050	(0.046, 0.054)	0.056	(0.051, 0.061)	0.063	(0.057, 0.070)
	μ_z	-0.132	(-1.039, 1.097)	1.010	(-0.979, 3.415)	2.580	(-1.178, 7.670)
	ρ	0.995	(0.989, 0.999)	0.989	(0.969, 0.999)	0.979	(0.798, 0.999)
	σ_η	0.094	(0.079, 0.112)	0.193	(0.136, 0.282)	0.315	(0.161, 0.789)
2y	μ	-0.000	(-0.002, 0.001)	-0.001	(-0.003, 0.001)	0.000	(-0.002, 0.003)
	σ	0.045	(0.042, 0.049)	0.048	(0.043, 0.052)	0.060	(0.055, 0.067)
	μ_z	-0.028	(-0.294, 0.282)	0.417	(-1.224, 2.586)	3.077	(-0.089, 8.966)
	ρ	0.988	(0.978, 0.996)	0.987	(0.959, 0.999)	0.953	(0.710, 0.998)
	σ_η	0.073	(0.062, 0.086)	0.220	(0.156, 0.321)	0.530	(0.268, 1.332)
3y	μ	-0.001	(-0.003, 0.001)	-0.002	(-0.004, 0.001)	-0.000	(-0.003, 0.003)
	σ	0.049	(0.046, 0.053)	0.050	(0.046, 0.054)	0.054	(0.049, 0.059)
	μ_z	-0.053	(-0.580, 0.852)	0.441	(-2.090, 4.182)	3.701	(-2.423, 11.487)
	ρ	0.994	(0.984, 0.999)	0.994	(0.979, 1.000)	0.967	(0.533, 0.999)
	σ_η	0.063	(0.055, 0.074)	0.144	(0.108, 0.198)	0.617	(0.301, 1.805)
4y	μ	-0.002	(-0.004, 0.000)	-0.003	(-0.006, -0.000)	-0.001	(-0.005, 0.003)
	σ	0.046	(0.042, 0.050)	0.045	(0.041, 0.049)	0.046	(0.042, 0.050)
	μ_z	-0.128	(-0.239, 0.023)	0.221	(-0.646, 2.203)	7.749	(1.941, 24.533)
	ρ	0.975	(0.955, 0.991)	0.986	(0.937, 0.999)	-0.126	(-0.804, 0.938)
	σ_η	0.064	(0.055, 0.076)	0.117	(0.091, 0.158)	0.203	(0.127, 0.564)
5y	μ	-0.002	(-0.004, -0.000)	-0.004	(-0.008, -0.001)	-0.003	(-0.009, 0.003)
	σ	0.050	(0.045, 0.055)	0.046	(0.042, 0.052)	0.049	(0.044, 0.055)
	μ_z	-0.228	(-0.320, -0.138)	-0.121	(-0.305, 0.038)	-0.957	(-5.073, 3.564)
	ρ	0.959	(0.938, 0.974)	0.912	(0.811, 0.966)	0.961	(0.528, 0.998)
	σ_η	0.067	(0.057, 0.079)	0.137	(0.100, 0.195)	0.499	(0.228, 1.663)
7y	μ	-0.003	(-0.005, -0.001)	-0.005	(-0.009, -0.001)	-0.005	(-0.011, 0.001)
	σ	0.068	(0.059, 0.079)	0.068	(0.057, 0.080)	0.070	(0.060, 0.079)
	μ_z	-0.445	(-0.580, -0.336)	-0.543	(-1.454, -0.190)	-1.379	(-4.370, -0.127)
	ρ	0.963	(0.940, 0.981)	0.975	(0.922, 0.998)	0.920	(-0.633, 0.998)
	σ_η	0.065	(0.056, 0.078)	0.104	(0.081, 0.137)	0.300	(0.160, 0.804)
10y	μ	-0.003	(-0.005, -0.001)	-0.005	(-0.008, -0.001)	-0.005	(-0.012, 0.001)
	σ	0.097	(0.080, 0.117)	0.094	(0.076, 0.112)	0.089	(0.073, 0.106)
	μ_z	-0.686	(-0.896, -0.528)	-0.782	(-1.826, -0.318)	-1.395	(-4.020, 0.208)
	ρ	0.970	(0.949, 0.986)	0.979	(0.941, 0.998)	0.965	(-0.336, 0.998)
	σ_η	0.070	(0.059, 0.084)	0.102	(0.081, 0.133)	0.223	(0.132, 0.523)
30y	μ	-0.002	(-0.004, -0.000)	-0.003	(-0.007, 0.001)	-0.005	(-0.013, 0.002)
	σ	0.150	(0.127, 0.172)	0.141	(0.115, 0.163)	0.132	(0.106, 0.153)
	μ_z	-1.103	(-1.370, -0.898)	-1.136	(-2.073, -0.689)	-1.653	(-5.074, 0.453)
	ρ	0.972	(0.951, 0.986)	0.978	(0.941, 0.997)	0.975	(0.784, 0.999)
	σ_η	0.086	(0.070, 0.108)	0.116	(0.088, 0.158)	0.210	(0.127, 0.408)

TABLE 26. Bayesian estimation results stochastic elasticity of volatility models.

Models with latent state variables, such as those described in this and the previous sections, are intricate and time-consuming to analyze. Outside of a limited number of special cases, advancing beyond a small collection of risk factors into a wider range of sources influencing risk measurement for a large portfolio is not currently practical with non-linear or non-Gaussian state space models. Therefore, these models are only recommended for the in-depth study of the most fundamentally important factors. Arguably, USD swap rates are among these.

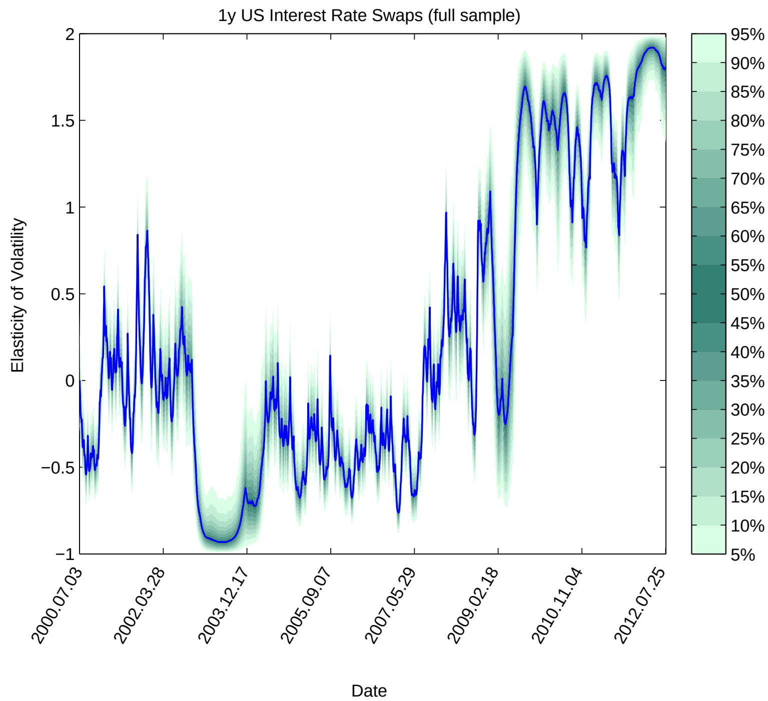

FIGURE 11. Distribution of stochastic elasticity of volatility profile for 1-year swap maturity.

11. Non-parametric Volatility Function

11.1. Local Volatility Profiles. More general encompassing alternative is to start with a general continuous time diffusion specification

$$dr_t = \mu(r_t) + \sigma(r_t)\, dW_t. \tag{11.1}$$

Several authors developed the idea of letting the diffusion coefficient of the instantaneous rate process to be modeled by the data themselves through an approach based on the non-parametric estimation of the conditional density of the series. The clear advantage of non-parametric specification is its flexibility.

Among several competing estimators, estimator of Stanton (1997) is lauded as simple and reliable by a comparative study of Renò, Roma, and Schaefer (2006).[24] It is given by

$$\hat{\sigma}^2(r) = \frac{\sum_{t=1}^{T-1} (r_{t+1} - r_t)^2 \mathcal{K}\left(\frac{r-r_t}{h}\right)}{\sum_{t=1}^{T-1} \mathcal{K}\left(\frac{r-r_t}{h}\right)}, \tag{11.2}$$

where $\mathcal{K}(x)$ is a kernel function that depends on the smoothing parameter (bandwidth) h. Gaussian kernel $\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$ is a convenient and common choice. The choice of

[24]Fan and Zhang (2003) examine higher order approximations and find that a reduction in asymptotic biases must be paid for with nearly exponentially escalating asymptotic variances. On the other hand, they argue that using local linear regression of squared changes on levels is a better alternative to local constant regression embedded inside estimator (11.2) in terms of boundary biases.

the smoothing parameter h regulates the tradeoff between bias and efficiency. For the estimator (11.2), Stanton (1997) recommends

$$h = 4\widehat{\text{std}}\,(r_t)\,T^{-\frac{1}{5}}. \tag{11.3}$$

Estimator (11.2) is intuitive. Indeed, one can recognize it as the average of squared differences weighted by a kernel function. With uniform kernel, the estimator amounts to the sample variance localized to the neighborhood of the point of interest. As one moves the spread value of interest through its range, the neighborhood moves as well and the estimate continually updated. With Gaussian kernel, diminishing consideration is given to the moves that occurred further away from the point of interest. Using Gaussian kernel allows selection of kernel bandwidth without worry about it being smaller than the minimal gap between observed rate which would pose a problem for kernels with finite support. An alternative interpretation of (11.2) is via local constant kernel regression of squared moves on the pre-move levels.[25] This interpretation can be used to replace recommended bandwidth (11.3) with more general bandwidth obtained by least squares cross-validation. Least squares cross-validation using our interest rate dataset, not shown, suggests that Stanton rule tends to oversmooth the volatility profile, but cross-validated bandwith results in peculiarly wiggly shapes that are harder to interpret.

We implemented Gaussian kernel estimator (11.2) on our interest rate swap example. Univariate time-series results fall roughly in the three sub-groups. Short maturities (one to three year) are exemplified by profiles in the upper left panel of Figure 12 with upward sloping volatility shape based on four year sample and flat profiles based on either full sample or the most recent year. For medium maturities, shown in the upper right panel, volatility profile based on the most recent year is no longer flat but is rising gently. At the long end, profiles based on full sample and four year history exhibit volatility that declines with level, while the profile based on the most recent year slopes steeply upward, as can be seen in the lower left panel. The lower right panel of the figure pools all maturities together and estimates volatility profile assuming the same relationship across all maturities.[26] In this setting, profiles based on full sample and four year history exhibit rising and falling edges, while the profile based on the most recent year rises up.

The volatility profiles in Figure 12 help interpret earlier parametric results, particularly with regard to parameter instability across different subsamples as well as poor fit of some parametric models. For example, fitting a linear relationship to full sample profiles (red lines) over the entire range of interest rates is likely to result in virtually flat volatility profile with an attendant conclusion that the series should be represented as differences. On the other hand, limiting ourselves to the four year history, we find almost linearly increasing profiles for all maturities less than 30-year, indicating preference for the return specification for all maturities except 30-year.

11.2. Local Directional Volatility Profiles. In the spirit of section 6, it is possible to explore non-parametrically whether upward and downward elasticity profiles are different. This could be done by constraining non-parametric volatility estimator to subset of only upward and only downward moves. Doing so results in the volatility profiles shown in Figure 13. These profiles are consistent with the idea that downward moves are more likely to be represented as relative changes, since these profile tend to flatten at higher rate levels. Combining all maturities and using the longest

[25]More elaborate local linear kernel regression may result in negative fitted variance when projected outside the range of observed data.

[26]Pooling across maturities improves the bandwidth selection. It could be used as a preliminary step for an individual maturity volatility profile estimation, as in Sam and Jiang (2009).

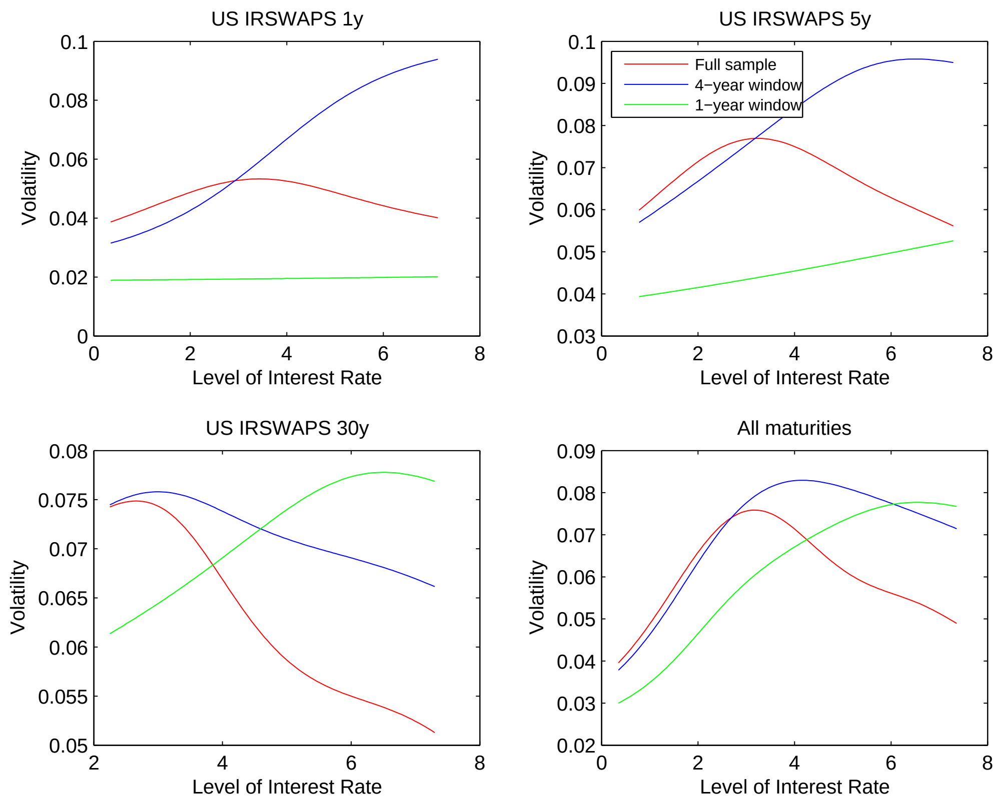

FIGURE 12. Non-parametric level-dependent volatility profiles for US dollar interest rate swap rates.

sample, the location of the switch is roughly 20 basis points higher for moves up than for moves down.

11.3. Estimation Uncertainty. Non-parametric estimators such as (11.2) are subject to estimation uncertainty that can be quite substantial in small samples. One way to create confidence envelopes about non-parametric volatility profiles is to resample the set of residuals $\hat{\epsilon}_t = \frac{\Delta r_t - \hat{\mu}}{\hat{\sigma}_h(r_{t-1})}$. In this method, bootstrap sample $\{\tilde{\epsilon}_t\}_{t=1}^{T}$ repeatedly feeds into synthetic observations $\tilde{r}_t$ and a new estimated volatility profile. A typical example is shown in Figure 14 for 30-year swap rates. It shows that although confidence bands are loose, the general shape remains consistent. The bands could be used to assess informally whether a parametric profile is within the confidence set of non-parametric estimator.

11.4. Testing Parametric Models against Non-parametric Alternative. A more formal comparison of parametric and non-parametric results can be made with the help of Generalized Likelihood Ratio test (Fan, Zhang, and Zhang, 2001; Fan and Zhang, 2003).[27]

[27] Further specifications tests of diffusion models can be found in Gallant and Tauchen (1996), Aït-Sahalia (1996), Corradi and Swanson (2005), as well as Hong and Li (2005).

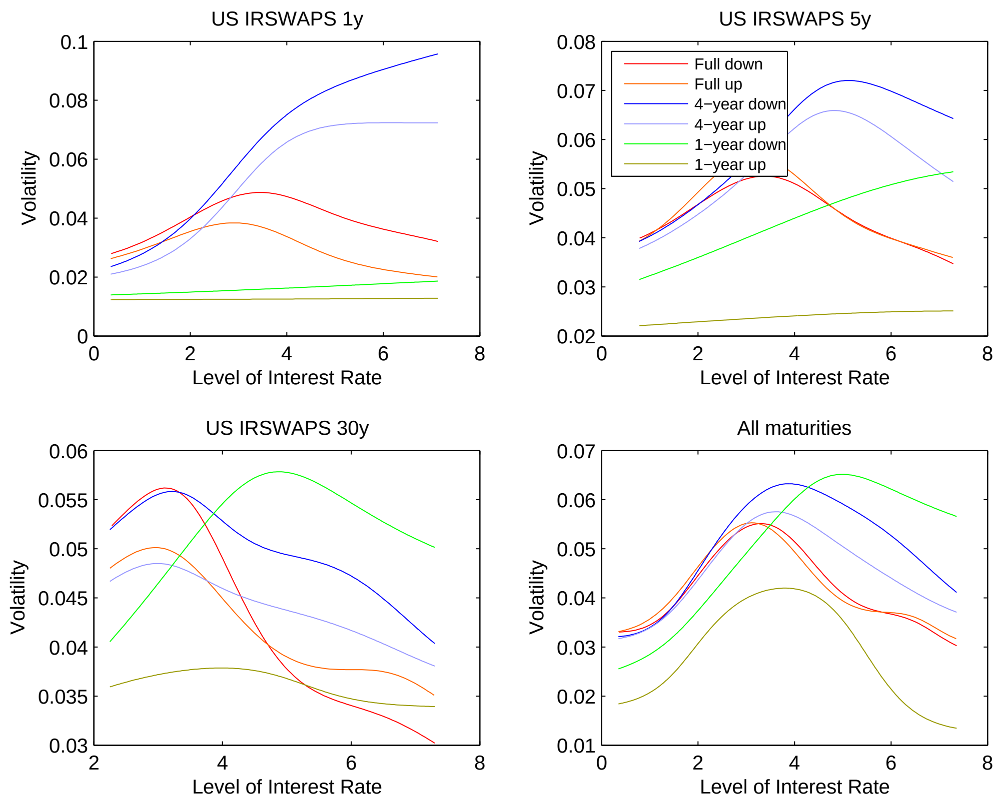

FIGURE 13. Level-dependent directional volatility in US dollar interest rate swap rates.

The test statistic is given by

$$g(h) = \frac{T-1}{2} \log \frac{RSS_0}{RSS_1(h)}, \tag{11.4}$$

where RSS_0 represents the residual sum of squares under a parametric null hypothesis

$$\log\left(\Delta r_t - \widehat{\mu}\right)^2 = \log\left(\widehat{\sigma}^2(r_{t-1})\right) + \log\left(\epsilon_t^2\right), \tag{11.5}$$

while $RSS_1(h)$ is the residual sum of squares under a non-parametric alternative

$$\log\left(\Delta r_t - \widehat{\mu}\right)^2 = \log\left(\widehat{\sigma}_h^2(r_{t-1})\right) + \log\left(\epsilon_t^2\right). \tag{11.6}$$

The test statistic can be viewed as a measure of distance between parametric and non-parametric profiles. Finite sample distributions of g for the trio of parametric elasticity models developed in preceding sections are not known. Instead, we can approximate the empirical null distributions of the GLR test statistics using a regression bootstrap method. Given fixed parameter estimates, this method repeatedly feeds standardized residuals into the model in order to generate artificial interest rate time-series. Each such time-series gives rise to simulated values for RSS_0, RSS_1 and, ultimately, g.

Results of the test are shown in Table 27. We can see some visible improvements in p-values for sigmoid variable elasticity version for the longest and shortest sample periods for certain maturities, but, overall, we cannot conclude that our attempts to model volatility parametrically

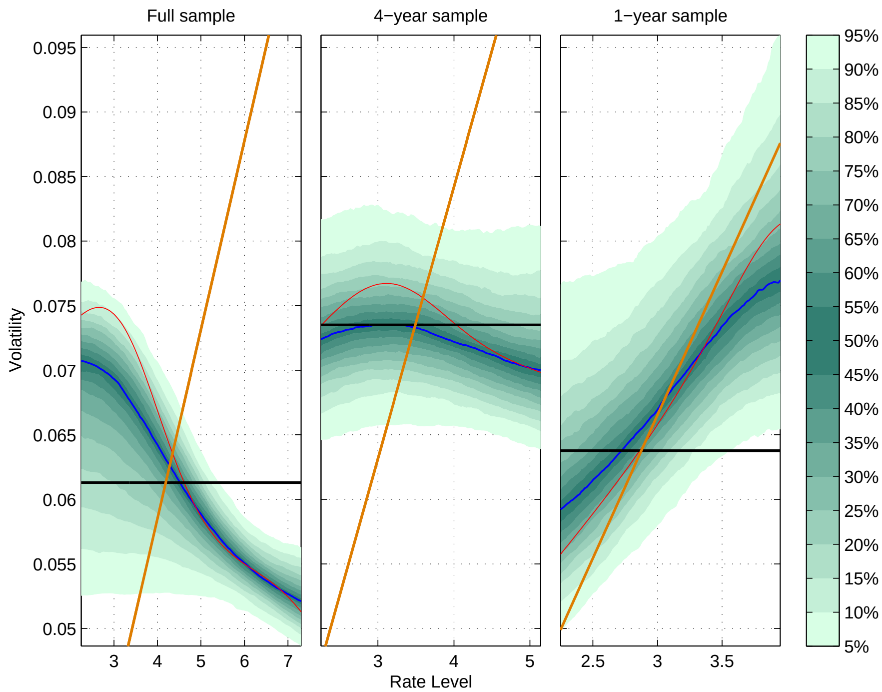

FIGURE 14. Estimation uncertainty of Stanton estimator of level-dependent volatility in 30-year US dollar interest rate swap rates. Stanton estimator is given by the red curve. Parametric alternatives are shown in black (differences) and brown (returns).

Swap Maturity	Constant Elasticity Full sample	Constant Elasticity 4-year sample	Constant Elasticity 1-year sample	Sigmoid Elasticity Full sample	Sigmoid Elasticity 4-year sample	Sigmoid Elasticity 1-year sample	Piecewise Elasticity Full sample	Piecewise Elasticity 4-year sample	Piecewise Elasticity 1-year sample
1y	0.000	0.000	0.000	0.011	0.000	0.170	0.000	0.000	0.000
2y	0.000	0.000	0.000	0.092	0.000	0.000	0.000	0.000	0.000
3y	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.000
4y	0.000	0.000	0.000	0.113	0.000	0.000	0.000	0.000	0.000
5y	0.000	0.000	0.000	0.309	0.000	0.001	0.000	0.000	0.000
7y	0.000	0.000	0.002	0.258	0.001	0.283	0.000	0.000	0.002
10y	0.000	0.000	0.001	0.409	0.000	0.315	0.000	0.000	0.018
30y	0.011	0.000	0.000	0.000	0.000	0.215	0.000	0.000	0.003

TABLE 27. P-values of Generalized Likelihood Ratio test of various parametric models against non-parametric diffusion estimator.

were particularly successful. This is probably because parametric models to do not capture fully downward sloping volatility at high levels.

11.5. Implications. Estimated diffusion coefficient function opens up the possibility for switching the risk form depending on the contemporaneous level, either between returns and differences or among a continuum of alternative fractional elasticities. Bottom right panel of Figure 12 is particularly exhortatory as it suggests a cutoff level of about 280 basis points below which the interest rates are best treated according to a proportional change law.

While the guidance that non-parametric methods provide is instructive, it should be emphasized that the cloud of uncertainty surrounding the diffusion shapes could be uncomfortably large unless the sample size is substantial. This is because there is no free lunch in statistics and not having to make a strong parametric assumption must be paid for by the loss of statistical efficiency, embodied in the wide confidence intervals. Point estimates such as those in Figure 12 ignore estimation uncertainty and, hence, should be taken with a grain of salt. Nonetheless, when taken as but one element of the battery of statistical methods aimed at settling the functional form choice, the non-parametric techniques could be a dependable final nail in the coffin of a dubious risk form proposal, such as the difference representation when the rates are low.

12. Concluding Remarks

Driven by the need to provide statistical basis to the functional form selection, we have described a number of techniques to help make that choice. The paper illustrated three kinds of approaches. In the first category are largely informal and often inconclusive non-encompassing methods that rank alternative functional forms on the basis of the stationarity tests, goodness-of-fit measures and stochastic complexity. In the second category of methods, we assume parametric elasticity of variance representation of the data generating process. Here, we attain formal statistical ground to select between the two alternatives that are now nested in a more general specification. Within this class we further distinguish three estimation approaches that hinge upon willingness to make additional distributional assumptions about model residuals, dependence structure across multiple related time-series, or ability to put informative priors on parameters of interest. In essence, the GMM, maximum likelihood and Bayesian methods allow us to refine bias-variance trade-off within the constant elasticity class of models. Obvious cost to the parametric methods such as assuming constant elasticity of variance structure is a nagging concern about misspecification. Tight standard errors are of no use if maintained assumptions are at odds with the true data generating process. Making elasticity variable is perfectly feasible except for the very wide gamut of alternative choices. Third and final category of methods is designed to mitigate the misspecification concern by appealing to non-parametric approaches and allows us to estimate the diffusion coefficient without assuming its functional form but making instead more general assumptions such as dimensionality, smoothness and existence of moments. Treating diffusion coefficient as an unknown function of the level of the series makes it possible to condition the current choice of functional form representation on the most recent values of the time series, switching between the two as conditions change. The cost is a loss of statistical efficiency. Thus, all three categories are best used together allowing one to move along the bias-variance trade-off frontier and make an educated functional form choice. This course of action was illustrated on an interest rate dataset where we were able to achieve a fairly robust conclusion that all series in this dataset should be treated in the return space for the short-to-medium-term risk measurement practice, at least until normalization of the interest rates takes them back above 2.8%.

Our final recommendation is to start with non-parametric profiles. Notwithstanding significant estimation uncertainty, these profiles can inform selection of a pertinent parametric model, especially if varying elasticity is in order. Armed with the estimated elasticity profiles, an analysis can proceed to formulating an appropriate variable elasticity model in order to refine and buttress initial impressions, selecting between the two alternatives as befits the relevant span of data and a purpose of selection. With this in mind, the choice of the risk functional specification will have meandering souls despair in the dark desert of ignorance no longer.

Appendix A. Histogram Binning with Thresholded Haar Wavelets

The Haar wavelet density estimator for a density $\hat{f}(x)$ based on a sample $\{x_t\}_{t=1}^{T}$ is given by the following truncated series expansion:

$$\hat{f}(x) = \sum_{k\in\mathbb{Z}} \hat{c}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1-1}\sum_{k\in\mathbb{Z}} \hat{d}_{j,k}\psi_{j,k}(x), \tag{A.1}$$

with Haar basis functions

$$\phi_{j_0,k}(x) = \begin{cases} 2^{j_0/2} & \text{if } \frac{1}{j_0}\log_2 k \le x < \frac{1}{j_0}\log_2(k+1) \\ 0 & \text{otherwise,} \end{cases}$$
$$\psi_{j,k}(x) = \begin{cases} 2^{j/2} & \text{if } \frac{1}{j}\log_2 k \le x < \frac{1}{j}\log_2\left(k+\frac{1}{2}\right) \\ -2^{j/2} & \text{if } \frac{1}{j}\log_2\left(k+\frac{1}{2}\right) \le x < \frac{1}{j}\log_2(k+1) \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

Since these basis functions form an orthonormal basis in the space of square-integrable functions, the coefficients can be estimated by

$$\hat{c}_{j_0,k} = \frac{1}{T}\sum_{t=1}^{T}\phi_{j_0,k}(x_t),$$
$$\hat{d}_{j,k} = \frac{1}{T}\sum_{t=1}^{T}\psi_{j,k}(x_t). \tag{A.3}$$

Functions $\phi_{j_0,k}$ are known as the *Haar scaling functions*, while functions $\psi_{j,k}$ are known as *Haar wavelets*. From the frequency domain perspective, the first sum in (A.1) describes low-frequency component of the signal x_t, while the second sum contains high-frequency detail. In view of that, coefficients $\hat{c}_{j_0,k}$ are known as the approximation coefficients and coefficients $\hat{d}_{j,k}$ as the detail coefficients. j_0 is termed the *primary resolution level* determining the scale of the largest effects that can be affected by smoothing inherent in the procedure. It is typically set to a value where a single scaling function $\phi_{j_0,k}$ has the range covering the whole of the observed data. j_1, is the finest resolution level, chosen to be the smallest j for which the $\phi_{jk}(x)$ each cover at most one of the distinct data values. Such j_1 is also known as the *minimum inter-point distance* choice. $\hat{c}_{j_0,k}$ and $\hat{d}_{j,k}$ are known as the empirical wavelet coefficients. Efficient computation of these uses the Daubechies-Lagarias algorithm (Daubechies and Lagarias, 1992) and the discrete wavelet transform (Burrus, Gopinath, and Guo, 1997) instead of the direct use of (A.3), which would be slow. In (A.3), only finitely many coefficients are non-zero.

Having computed all the empirical wavelet coefficients, the next stage is to threshold (or de-noise) them in order to compress the large amount of superfluous details caused by the minimum interpoint-distance method forcing each distinct data value into its own histogram bin. A number of ways have been devised in the literature in order to do so, including level-dependent thresholding rules in Donoho, Johnstone, Kerkyacharian, and Picard (1996) minimizing Stein's unbiased estimators of risk, a cross-validation approach in Nason (1994), a minimum description length principle in Kumar, Heikkonen, Rissanen, and Kaski (2006), wavelet coefficient significance tests in Herrick, Nason, and Silverman (2001), etc. We have chosen the minimum description length principle of Kumar, Heikkonen, Rissanen, and Kaski (2006) as a demonstration. The main idea of this approach is to model retained wavelet coefficients by an equal width histogram at each resolution level of the wavelet transform and the coefficients to be discarded by a single equal bin width histogram. Minimization of the total code length gives the optimal way of dividing the coefficients into those representing informative content and those representing extraneous noise. The main steps are summarized as follows.

(1) Obtain the set of all wavelet coefficients through the wavelet transform up to the highest possible level r.
(2) Recursively on resolution levels $i = 1, \dots, r$ from finest to crudest, fit an m-bin histogram to the coefficients of a given level i with the equal bin width determined by the full range of coefficients across all levels. With this histogram, select a tentative collection of bins S_i with the number of chose bins m_i. The bins in collection S_i contain a total of k_i tentatively chosen coefficients distributed into bins of S_i with counts $n_{i,(j)}$ for $j = 1, \dots, m_i$. If the chosen coefficients are retained, the residual coefficients are defined to equal zero at retained indices as well as for retained indices at all prior levels.
(3) Fit a histogram with M bins to the residual coefficients across all resolution levels.
(4) Find the optimal S_i and M by minimizing the total code length:

$$\min_{S_i,M}\Big\{\log_2\left(n_{i,(1)},\dots,n_{i,(m_i)},n_i-k_i\right)!+\log_2\left(n_i,m_i+1\right)!+\log_2\left(\nu_1,\dots,\nu_M\right)!$$
$$+\log_2\left(M,n-\tilde{k}_{i-1}-k_i\right)!+k_i\log_2\left(\frac{MR}{mR_i}\right)+\tilde{k}_{i-1}\log_2\left(\frac{M}{R_i}\right)$$
$$-(n-1)\log_2 M+2\log_2\log_2 M+(n+1)\log_2 R_i+2\log_2\log_2 R_i\Big\},$$

where n_i is the total number of coefficients at i^{th} resolution level, ν_j is the number of coefficients falling into the j^{th} bin of the M-bin histogram fitted to the residual string of coefficients, R is the range of all wavelet coefficients, R_i are the level-specific ranges of the coefficients, $\tilde{k}_{i-1} = \sum_{j=1}^{i-1}\hat{k}_j$ denotes the number of retained coefficients in the so far optimized sets S_j, $j < i$ and $(n_1, n_2, \dots, n_k)! = \frac{(n_1+n_2+\cdots+n_k)!}{n_1!n_2!\dots n_k!}$ is the multinomial coefficient. For the first level, this sum is zero.

The number and location of the optimally chosen bins is indicated by the retained coefficients, with coefficients from finer levels corresponding to narrower bins.

Appendix B. Extended Critical Values of KPSS test

In order to extend the range of critical values of the KPSS test we performed a Monte Carlo simulation using 2 billion replications of test statistics under the null using artificial samples of size 1,000. The results are tabulated in Table 28.

Appendix C. Nyblom Stability Test

Significance Level	Critical Value	
	No trend	Linear trend
0.000001	2.5146	0.6541
0.00001	2.0567	0.5423
0.0001	1.6121	0.4314
0.001	1.1718	0.3228
0.0025	1.0002	0.2804
0.005	0.8719	0.2487
0.01	0.7455	0.2175
0.025	0.5820	0.1773
0.05	0.4623	0.1477
0.1	0.3479	0.1190
0.2	0.2415	0.0914
0.3	0.1845	0.0755
0.4	0.1467	0.0643
0.5	0.1189	0.0555
0.6	0.0969	0.0480
0.7	0.0786	0.0413
0.8	0.0622	0.0348
0.9	0.0460	0.0278
0.95	0.0365	0.0234
0.975	0.0303	0.0202
0.99	0.0248	0.0172
0.995	0.0218	0.0155
0.9975	0.0194	0.0142
0.999	0.0170	0.0127
0.9999	0.0129	0.0101
0.99999	0.0104	0.0084
0.999999	0.0088	0.0072

TABLE 28. Extended (right-sided) critical values of KPSS test.

Let θ be the vector of model parameters partitioned as $\begin{pmatrix}\theta^1\\ \theta^2\end{pmatrix}$. The second group of parameters, θ^2 is treated as constant. The martingale formulation underlying the Nyblom test is

$$\begin{aligned}\mathbb{E}\left(\theta_t^1-\theta_{t-1}^1 \middle| \mathcal{I}_{t-1}\right) &= 0\\ \mathbb{E}\left(\left(\theta_t^1-\theta_{t-1}^1\right)\left(\theta_t^1-\theta_{t-1}^1\right)' \middle| \mathcal{I}_{t-1}\right) &= \delta^2 V_t\end{aligned}$$

for some adapted sequence of information sets $\mathcal{I}_t$ and a known array V_t. If V_t is constant, the case of constant instability hazard obtains.

The null and alternative hypotheses of the Nyblom test are

$$H_0: \quad \delta^2 = 0 \qquad H_1: \delta^2 > 0$$

For a single observation, define the score vector $\mathcal{S}_\tau^{\theta^1} = \frac{\partial}{\partial\theta^1}\log\mathcal{L}_\tau\left(\theta^1,\theta^2\right)$ and the Hessian of log-likelihood $\mathcal{H}_\tau^{\theta^1\theta^1} = \frac{\partial^2}{\partial\theta^1\partial\theta^{1\prime}}\log\mathcal{L}_\tau\left(\theta^1,\theta^2\right)$. For the entire sample, $\mathcal{S}^{\theta^1} = \sum_\tau \mathcal{S}_\tau^{\theta^1}$ and $\mathcal{H}^{\theta^1\theta^1} = \sum_\tau \mathcal{H}_\tau^{\theta^1\theta^1}$

Under constant instability hazard, the test statistics takes form

$$L = \frac{1}{T}\mathrm{tr}\left(V\sum_{t=1}^{T-1}\left(\sum_{\tau=1}^{t}\hat{\mathcal{S}}_{\tau}^{\theta^1}\right)\left(\sum_{\tau=1}^{t}\hat{\mathcal{S}}_{\tau}^{\theta^1}\right)'\right).$$

If we further assume $V = -\left(\hat{\mathcal{H}}^{\theta^1\theta^1}\right)^{-1}$, the asymptotic distribution of L is invariant to nuisance parameters.

Specializing the above to the CEV model with Student t residuals relies on the following analytic results.

$$\mathcal{S}_t^{\gamma} = \frac{\nu\left((\Delta r_t - \mu)^2 - \sigma^2 r_t^{2\gamma}\right)\log(r_t)}{(\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}},$$

$$\mathcal{H}_t^{\gamma\gamma} = -\frac{2\nu(\nu+1)\sigma^2(\Delta r_t - \mu)^2 r^{2\gamma}\log^2(r_t)}{\left((\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}\right)^2},$$

$$\begin{aligned}\mathcal{S}_t^{\nu} = &-\frac{1}{2}\left(\log\left(1 + \frac{r_t^{-2\gamma}(\Delta r_t - \mu)^2}{\nu\sigma^2}\right) + \psi\left(0, \frac{\nu}{2}\right) - \psi\left(0, \frac{1+\nu}{2}\right)\right)\\ &+ \frac{(\Delta r_t - \mu)^2 - \sigma^2 r_t^{2\gamma}}{2\left((\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}\right)},\end{aligned}$$

$$\mathcal{H}_t^{\gamma\nu} = \frac{(\Delta r_t - \mu)^2\left((\Delta r_t - \mu)^2 - \sigma^2 r_t^{2\gamma}\right)\log(r_t)}{\left((\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}\right)^2},$$

$$\mathcal{H}_t^{\nu\nu} = \frac{2\left((\Delta r_t - \mu)^4 + \nu\sigma^4 r_t^{4\gamma}\right) + \nu\left((\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}\right)\left(\psi\left(1, \frac{1+\nu}{2}\right) - \psi\left(1, \frac{\nu}{2}\right)\right)}{4\nu\left((\Delta r_t - \mu)^2 + \nu\sigma^2 r_t^{2\gamma}\right)^2},$$

where $\psi(k, \cdot)$ is the polygamma function.

Appendix D. Markov Chain Monte Carlo

Markov chain Monte Carlo is a general framework of specifying and simulating an irreducible aperiodic Markov chain with an ergodic transition kernel such that the complete posterior is the chain's invariant distribution. Running the simulation long enough then guarantees that the distribution of current draws is close enough to the target distribution. One particular celebrated special case is the so called Gibbs sampler (Gelfand and Smith, 1990) where the parameters are split into multiple blocks, parameters in one block are sampled conditional on the values of parameters in remaining blocks and the Markov chain is constructed by cycling through the drawing procedure for each conditional distribution. The sampler can be justified by the Clifford-Hammersley theorem (Robert and Casella, 2000). That the stationary distribution of the resulting chain is indeed the joint distribution of interest makes it possible to sample from joint posterior without knowledge of either the joint density or the marginal densities of the blocks. Having a sample from the joint posterior enables sampling from marginal posterior by simply ignoring draws of remaining parameters.

D.1. MCMC for CEV Model. In the case of the CEV model, the Gibbs sampler iterates through the following four steps:

(1) sample μ from $p\left(\mu \middle| \sigma, \gamma, \nu, \{r_t\}_{t=1}^{T}\right) = \mathcal{N}\left(m_{\mu}^{1}, \sigma_{1\mu}^{2}\right)$;

(2) sample σ from $p\left(\sigma^2 \middle| \mu, \gamma, \nu, \{r_t\}_{t=1}^{T}\right) = p\left(\sigma^2 \middle| \left\{\frac{r_t-\mu}{r_{t-1}^{\gamma}\sqrt{u_t}}\right\}_{t=2}^{T}\right) = \mathcal{IG}\left(df_1, S_1\right)$;

(3) sample γ from $p\left(\gamma \middle| \mu, \sigma, \nu, \{r_t\}_{t=1}^{T}\right)$;

(4) sample ν from $p\left(\nu \middle| \mu, \sigma, \gamma, \{r_t\}_{t=1}^{T}\right)$.

The kernel of the conditional posterior for γ takes the following non-standard form:

$$p\left(\gamma \middle| \mu, \sigma, \nu, \{r_t\}_{t=1}^{T}\right) \propto \prod_{t=2}^{T} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma r_{t-1}^{\gamma}\sqrt{\nu\pi}} \left(1 + \frac{1}{\nu}\left(\frac{\Delta r_t - \mu}{\sigma r_{t-1}^{\gamma}}\right)^2\right)^{-(\nu+1)/2}. \tag{D.1}$$

Similarly, the kernel of conditional posterior for ν is also non-standard[28]:

$$p\left(\nu \middle| \mu, \sigma, \gamma, \{r_t\}_{t=1}^{T}\right) \propto \prod_{t=2}^{T} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sigma r_{t-1}^{\gamma}\sqrt{\nu\pi}} \left(1 + \frac{1}{\nu}\left(\frac{\Delta r_t - \mu}{\sigma r_{t-1}^{\gamma}}\right)^2\right)^{-(\nu+1)/2}. \tag{D.2}$$

Following Meyer and Yu (2000) and Chan, Choy, and Lee (2007) we make use of a freely available Bayesian software, JAGS, (Plummer, 2011), that provides an easy and efficient implementation of the Gibbs sampler with adaptive enhancements to deal with non-standard conditional posterior kernels as above (Gilks, Best, and Tan, 1995). Within the warm-up period, the algorithm progressively builds an almost-envelope for the target density, and afterward applies Metropolis-Hasting algorithm to deliver samples from non-standard conditional posteriors. This is designed to ensure good acceptance rate and fast exploration of the conditional posterior distribution.

In theory, output from an MCMC sampler converges to the target posterior distribution in the limit as the number of iterations tends to infinity. In practice, all MCMC runs are finite. By convention, the post-adaptation MCMC output is split into two parts: an initial "burn-in", which is discarded, and the remainder in which convergence is assumed to be achieved. Samples from the second part are used to create approximate summary statistics for the target distribution since any measurable function of a stationary and ergodic sequence is itself stationary and ergodic, so that the ergodic law of large number applies. We set the length of adaptation phase to 5,000 cycles, burn-in to 5,000 cycles and then generated 250,000 samples from the approximate posterior. These samples were further "thinned-out" by preserving every 25th sample in order to reduce the autocorrelation among subsequent draws. The remaining final sample of 10,000 parameter draws easily passed Heidelberger and Welch test (Heidelberger and Welch, 1983) as well as all other convergence diagnostics listed in Gamerman (1997) and accessible through CODA package in R.

[28]Right-hand side expressions in (D.1) and (D.2) are identical as both posteriors are proportional to the likelihood function. What separates the two is that the former is a function of γ, and the latter is a function of ν. Both equations also omit different normalization constants.

D.2. MCMC for Multivariate CEV Model. The Gibbs sampler algorithm for the multivariate version proceeds similarly:

(1) sample μ_i from

$$p\left(\mu_i \middle| \mu_{-i}, \sigma, \gamma, \nu, \{r_{it}\}_{i=1,t=1}^{N,T}\right) = p\left(\mu_i \middle| \sigma_i, \gamma, \nu_i, \{r_{it}\}_{t=1}^{T}\right) = \mathcal{N}\left(m_{\mu_i}^1, \sigma_{1\mu_i}^2\right)$$

for each $i = 1, \ldots, N$;

(2) sample σ_i from

$$p\left(\sigma_i^2 \middle| \mu, \sigma_{-i}, \gamma, \nu, \{r_{it}\}_{i=1,t=1}^{N,T}\right) = p\left(\sigma_i^2 \middle| \mu_i, \gamma, \nu_i, \{r_{it}\}_{t=1}^{T}\right) = \mathcal{IG}\left(df_1, S_{1i}\right)$$

for each $i = 1, \ldots, N$;

(3) sample γ from

$$\begin{aligned} & p\left(\gamma \middle| \mu, \sigma, \nu, \{r_{it}\}_{i=1,t=1}^{N,T}\right) \\ & \propto \prod_{i=1}^{N}\prod_{t=2}^{T} \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\Gamma\left(\frac{\nu_i}{2}\right)\sigma_i r_{it-1}^{\gamma}\sqrt{\nu_i\pi}} \left(1 + \frac{1}{\nu_i}\left(\frac{\Delta r_{it} - \mu_i}{\sigma_i r_{it-1}^{\gamma}}\right)^2\right)^{-(\nu_i+1)/2}; \end{aligned} \tag{D.3}$$

(1) sample ν_i from

$$\begin{aligned} & p\left(\nu_i \middle| \mu, \sigma, \gamma, \{r_{it}\}_{i=1,t=1}^{N,T}\right) = p\left(\nu_i \middle| \mu_i, \sigma_i, \gamma, \{r_{it}\}_{t=1}^{T}\right) \\ & \propto \prod_{t=2}^{T} \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\Gamma\left(\frac{\nu_i}{2}\right)\sigma_i r_{it-1}^{\gamma}\sqrt{\nu_i\pi}} \left(1 + \frac{1}{\nu_i}\left(\frac{\Delta r_{it} - \mu_i}{\sigma_i r_{it-1}^{\gamma}}\right)^2\right)^{-(\nu_i+1)/2} \end{aligned} \tag{D.4}$$

for each $i = 1, \ldots, N$,

where μ_{-i} is a shorthand for $\{\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_N\}$, and similarly for σ_{-i} and ν_{-i} symbols.

D.3. MCMC for Constant Elasticity of Stochastic Volatility and Stochastic Elasticity of Volatility Models. Models of sections 9 and 10 are more demanding in terms of techniques required for MCMC sampling from the joint posterior distribution as these belong to the class of nonlinear state-space models. In model (9.1), the time-varying volatility scale σ_t is the latent state variable. In model (10.1), the role of the latent state variable is taken by the evolving elasticity of volatility γ_t. In both cases, the entire joint posterior distribution of all latent states is needed, which is an object whose dimension grows with the sample size.

The generic MCMC approach, as implemented in JAGS, would introduce the following sequence of sampling steps into the algorithm:

- Sample S_t from $p\left(S_t \middle| S_{-t}, \theta\right)$ for all $t = 1, \ldots, T$,

where latent state S_t is either σ_t or γ_t, depending on the model, and θ encompasses all of the static model parameters. Such approach is inefficient and has proven to be very slow. This is due intermingling issues of high dependence between parameters and the latent process that makes standard Gibbs sampling strategies suboptimal (Roberts and Sahu, 1997; Papaspiliopoulos, Roberts, and Sköld, 2007), the difficulty of designing efficient simulation schemes for sampling from $p\left(S_{0:T} \middle| \theta\right)$ and having to deal with non-standard conditional distributions.

Optimizing sampling efficiency of nonlinear state-space model requires tailored algorithms. Fortunately, such algorithms exist. Two classes of state of the art methods are particularly appealing. The first category encompasses *sequential Monte Carlo* (SMC) (Doucet, de Freitas, and Gordon, 2001) family of methods and includes particle MCMC (Andrieu, Doucet, and Holenstein, 2010) and SMC^2 (Chopin, Jacob, and Papaspiliopoulos, 2013). In the context of state-space models, SMC-based methods sequentially approximate posterior densities of the state vector by a set of weighted random samples, called particles, as the dimension of the state grows with time. These are practical for a wide range of state-space models, including nonlinear and non-Gaussian models, and they fit well to recent, highly parallel, computer architectures. The second category is known as Hamiltonian Monte Carlo (HMC) (Liu, 2002; Neal, 2011; Hoffman and Gelman, 2014) that avoids correlation among successive draws by taking a series of steps informed by first-order gradient information. We used both approaches in order to gain comfort in features of the joint posterior distributions of parameters and unobserved volatility dynamics.

D.3.1. Particle MCMC and SMC^2. The posterior distribution of a state-space model with measurements $\mathbf{X} = X_{1:T} = \{X_t\}_{t=1}^{T}$, latent state vector $\mathbf{S} = S_{1:T} = \{S_t\}_{t=0}^{T}$ and static parameters θ can be factored as

$$p\left(\mathbf{S},\theta|\mathbf{X}\right) = p\left(\theta|\mathbf{X}\right) p\left(\mathbf{S}|\theta,\mathbf{X}\right). \tag{D.5}$$

Obtaining the first factor constitutes *parameter estimation*, while obtaining the second factor, conditional on the value of θ drawn from the first, constitutes *state estimation*.

The state estimation task can be handled by the *particle filter*. While a large assortment of variants is available, the most basic *bootstrap particle filter* works as follows:

(1) Initialization:
Draw P_S random samples (particles) $S_0^j \sim p\left(S_0|\theta\right)$ for $j = 1, \dots, P_S$ with uniform weights $w_0^j = 1/P_S$.

(2) Observation time iteration $t = 1, \dots, T$:

(a) Propagation step:
Each particle is advanced to the next observation time with $S_t^j \sim p\left(S_t|S_{t-1}^{a(j,t-1)},\theta\right)$ where $a\left(j,t-1\right)$ is the index of the particle's *ancestor* at the previous time.

(b) Weighting step:
Each particle at time t is weighted with the likelihood of the new observation $w_t^j = p\left(X_t|S_t^j,\theta\right)$. These weights are not normalized.

(c) Resampling step:
The set of weighted particles is transformed into a set with uniform weights by resampling particles with replacement, where the probability of each particle being drawn is proportional to its weight w_t^j. Particles with high weight tend to be replicated, while particles with low weight may be eliminated. Ancestor indices for the next time propagation step are determined in this step. At the end of this step, the population of particles represents a properly weighted sample from $p\left(S_t|\theta\right)$ in the sense that the weights are unbiased estimated of the Radon-Nikodym derivative between the target and proposal distribution.

(3) Smoothing:
The previous step provides samples from the *filtering* distribution $p\left(S_t|X_{1:t},\theta\right)$ at each $t = 1, \dots, T$. To obtain samples from the *smoothing* distribution $p\left(S_t|X_{1:T},\theta\right)$, each

terminal particle must be traced back through its ancestry. This requires storing entire histories of all particles.

A by-product of the particle filter output is an unbiased estimator of the likelihood increments

$$p\left(X_t \middle| X_{1:t-1}, \theta\right) = \frac{1}{P_S} \sum_{j=1}^{P_S} w_t^j, \tag{D.6}$$

and of the marginal likelihood

$$p\left(X_{1:t} \middle| \theta\right) = p\left(X_1 \middle| \theta\right) \prod_{\tau=2}^{t} p\left(X_\tau \middle| X_{1:\tau-1}, \theta\right), \tag{D.7}$$

for each $t = 1, \ldots, T$. Marginal likelihood of the entire observed sample, $p\left(X_{1:T} \middle| \theta\right)$, can be used as the model evidence in model comparison. In addition, tracing a single particle back through its ancestry provides an unbiased state sample $\tilde{S}_{0:T}$ from the state conditional posterior distribution $p\left(S_{0:T} \middle| X_{1:T}, \theta\right)$. The progressive exploration of the sequence of posterior distributions $p\left(S_{1:t}, \theta | X_{1:t-1}\right)$ is a key attraction of SMC methods as it allows the efficient reuse of samples across different times in contrast with MCMC methods which would typically have to be rerun for each time horizon.[29]

The parameter estimation task can also be handled by a variety of techniques. SMC^2 approach is to replace the MCMC over parameters with SMC over parameters and works similarly to SMC over state variables. It is initialized by drawing P_θ uniformly weighted particles from the prior distribution $p(\theta)$. It proceeds sequentially over observation times with a series of propagation, weighting and resampling steps, along with an additional *rejuvenation* step. An addition of the rejuvenation step is necessary because parameters do not change in time and, consequently, the propagation step is unable to reconstitute the number of unique values that may be depleted during the resampling step. Since the purpose of the rejuvenation is to diversify values of θ-particles while preserving their distribution, it is sufficient to take a single marginal Metropolis-Hastings step. In this step, a new value θ' is proposed for each θ-particle θ_t^j from some proposal distribution $q\left(\theta_t' \middle| \theta_t^j\right)$. The move is accepted with probability

$$\min\left(1, \frac{p\left(X_{1:T} \middle| \theta'\right) p\left(\theta'\right) q\left(\theta \middle| \theta'\right)}{p\left(X_{1:T} \middle| \theta\right) p\left(\theta\right) q\left(\theta' \middle| \theta\right)}\right), \tag{D.8}$$

where the marginal likelihood $p\left(X_{1:T} \middle| \theta'\right)$ is itself estimated by the particle filter (hence the SMC^2 appellation). If the move is accepted, the new θ value is the new draw. Otherwise, it remains at its previous value.

The SMC^2 algorithm has the highest degree of parallelism compared to SMC methods not relying on particle filtering for the parameter estimation task, such as the particle marginal Metropolis-Hastings (PMMH) algorithm (Andrieu, Doucet, and Holenstein, 2010). This is because both parameter and state particles can be manipulated simultaneously, except that resampling step needs to be synchronous.

[29] SMC^2 is a sequential but not an on-line algorithm since the computational load to maintain constant Monte Carlo error increases with time iterations (Chopin, Jacob, and Papaspiliopoulos, 2013). The difficulty of finding a genuinely generic on-line algorithm that would provide constant Monte Carlo error at a constant computational cost seems to be related to the fact that the target density is of increasing dimension.

The LibBi software library by Murray (2013) implements above SMC algorithms while supporting several hardware architectures and high-performance computing technologies, including multicore CPUs, MPI clusters and CUDA GPU programming on NVIDIA GPUs.

D.3.2. Hamiltonian Monte Carlo and the No-U-Turn Sampler. Hamiltonian Monte Carlo, also known as hybrid Monte Carlo, is based on a clever scheme that explores parameter space using the Hamiltonian dynamics of a fictitious physical system. In this system, the parameter vector of interest represents the position of a particle whose potential energy is given by the negative (unnormalized) probability. For each model variable, an auxiliary "momentum" variable is introduced, typically drawn independently from the standard Gaussian distribution. HMC alternates simple updates for these momentum variables with Metropolis updates in which a new parameter position is proposed using the end point of simulated trajectory under Hamiltonian dynamics propelled by previously drawn random initial kinetic energy. The benefit of the HMC proposal is that, even for a non-physical system, the resulting moves follow the dynamics of the target distribution more closely by taking advantage gradient information. Doing so is useful to speeding up exploration of the target distribution. The differential equations of Hamiltonian dynamics apply only to continuous variables and must be discretized for computer implementation. Simple leapfrog algorithm, also known as Störmer-Verlet integrator (Leimkuhler and Reich, 2004), is typically used. Standard HMC simulates the trajectory for a fixed number of discrete steps of a fixed step size.

Increased efficiency of HMC comes with two drawbacks. First, HMC requires the gradient of the log-posterior which could be tedious or even impossible to derive in closed form. Although automatic differentiation (Griewank and Walther, 2008) can be used to remove a user from the task, a performance penalty remains. Second, the optimal step size and the number of steps in the simulated trajectory are problem-specific and may require costly pilot runs to tune.

Hoffman and Gelman (2014) introduced an adaptive version of HMC, called the No-U-Turn Sampler (NUTS) that eliminates the need to tune the number of steps. NUTS uses a recursive algorithm that builds a set of candidate points dispersed over the target distribution and uses a geometric criterion that stops a trajectory when it begins to double back and retrace its steps. Once the trajectory is stopped, NUTS uses slice sampling to select a state along the trajectory as the next proposal. In addition, Hoffman and Gelman (2014) provide a method for adapting the global step sizes for each parameter to optimize a target Metropolis-Hastings rejection rate. The method is based on a modification of the primal-dual algorithm of Nesterov (2009) for stochastic optimization with vanishing adaptation. Altogether, NUTS can be run with no hand tuning to the shape of high-dimensional target distributions at all. The algorithm has been implemented as part of the new open-source Bayesian inference package called Stan (Stan Development Team, 2013). Stan includes modeling language similar to JAGS that allows writing models in familiar notation that could be transformed into efficient C++ code and then compiled into an executable. The results reported in sections 9 and 10 were obtained with Stan using the NUTS algorithm and were checked against posterior samples generated using LibBi.

References

Abramowitz, M., and I. A. Stegun (eds.) (1965): *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Dover Books on Mathematics. Dover Publications.

Aït-Sahalia, Y. (1996): "Testing Continuous-Time Models of the Spot Interest Rate," *The Review of Financial Studies*, 9(2), 385–426.

Andersen, T. G., and J. Lund (1997): "Estimating continuous-time stochastic volatility models of the short-term interest rate," *Journal of Econometrics*, 77(2), 343–377.

Andrieu, C., A. Doucet, and R. Holenstein (2010): "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72(3), 269–342.

Angelidis, T., A. Benos, and S. Degiannakis (2004): "The use of GARCH models in VaR estimation," *Statistical Methodology*, 1(1-2), 105–124.

Ball, C. A., and W. N. Torous (1999): "The Stochastic Volatility of Short-Term Interest Rates: Some International Evidence," *The Journal of Finance*, 54(6), 2339–2359.

Beirlant, J., E. J. Dudewicz, L. Györfi, and E. C. van der Meulen (1997): "Nonparametric Entropy Estimation: An Overview," *Journal of Mathematical Statistics Sciences*, 6(1), 17–39.

Brockwell, P. J., and R. A. Davis (1998): *Time Series: Theory and Methods*, Springer Series in Statistics. Springer, 2 edn.

Brown, M. B., and A. B. Forsythe (1974): "Robust Tests for the Equality of Variances," *Journal of the American Statistical Association*, 69(346), 364–367.

Burrus, C. S., R. A. Gopinath, and H. Guo (1997): *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, 1 edn.

Cameron, A. C., and P. K. Trivedi (1990): "The Information Matrix Test and Its Implied Alternative Hypotheses," Instutite of Governmental Affairs Papers 372, UC Davis, Davis, CA.

Chan, J. S. K., S. T. B. Choy, and A. B. W. Lee (2007): "Bayesian analysis of constant elasticity of variance models," *Applied Stochastic Models in Business and Industry*, 23(1), 83–96.

Chan, K. C., A. Karolyi, F. Longstaff, and A. Sanders (1992): "An Empirical Comparison of Alternative Models of the Short-term Interest Rate," *The Journal of Finance*, 47(3), 1209–1227.

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90(432), 1313–1321.

Chopin, N., P. Jacob, and O. Papaspiliopoulos (2013): "SMC^2: an efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 75(3), 397–426.

Cogley, T., S. Morozov, and T. J. Sargent (2005): "Bayesian fan charts for the UK inflation: Forecasting and sources of uncertainty in an evolving monetary system," *Journal of Economic Dynamics and Control*, 29(11), 1893–1925.

Corradi, V., and N. R. Swanson (2005): "Bootstrap specification tests for diffusion processes," *Journal of Econometrics*, 124(1), 117–148.

Cox, D. R. (1961): "Tests of Separate Families of Hypothesis," in *Proceedings of the Fourth Berkeley Symposium on Mathematical and Probability*, vol. 1: Contributions to the Theory of Statistics, pp. 105–123. University of California Press.

——— (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 24(2), 406–424.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

Daubechies, I., and J. C. Lagarias (1992): "Two-Scale Difference Equations II. Local Regularity, Infinite Products of Matrices and Fractals," *SIAM Journal on Mathematical Analysis*, 23(4), 1031–1079.

DIEBOLD, F. X., T. A. GUNTHER, AND A. S. TAY (1998): "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 31(4), 863–883.

DONOHO, D. L., I. M. JOHNSTONE, G. KERKYACHARIAN, AND D. PICARD (1996): "Density Estimation by Wavelet Thresholding," *The Annals of Statistics*, 24(2), 508–539.

DOUCET, A., N. DE FREITAS, AND N. GORDON (eds.) (2001): *Sequential Monte Carlo in Practice*, Information Science and Statistics. Springer-Verlag, New York.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1008.

ENGLE, R. F., AND A. J. PATTON (2001): "What good is a volatility model?," *Quantitative Finance*, 1(2), 237 – 245.

ERAKER, B. (2001): "MCMC Analysis of Diffusion Models With Application to Finance," *Journal of Business & Economic Statistics*, 19(2), 177–191.

FAN, J., AND C. ZHANG (2003): "A Reexamination of Diffusion Estimators With Applications to Financial Model Validation," *Journal of American Statistical Association*, 98(461), 118–134.

FAN, J., C. ZHANG, AND J. ZHANG (2001): "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29(1), 153–193.

GALLANT, A. R. (1987): *Nonlinear Statistical Models*, Wiley Series in Probability and Mathematical Statistics. Wiley.

GALLANT, A. R., AND G. TAUCHEN (1996): "Which Moments to Match?," *Econometric Theory*, 12(4), 657–681.

GAMERMAN, D. (1997): *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*, Texts in Statistical Science. Chapman & Hall, London.

GELFAND, A. E., AND A. F. M. SMITH (1990): "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85(410), 398–409.

GELMAN, A., J. B. CARLIN, H. S. STERN, D. B. DUNSON, A. VEHTARI, AND D. B. RUBIN (2013): *Bayesian Data Analysis*, no. 106 in Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, London, 3 edn.

GIACOMINI, R., AND I. KOMUNJER (2005): "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, 23, 416–431.

GILKS, W. R., N. G. BEST, AND K. K. C. TAN (1995): "Adaptive Rejection Metropolis Sampling within Gibbs Sampling," *Applied Statistics*, 44(4), 455–472.

GOLDFELD, S. M., AND R. E. QUANDT (1965): "Some Tests for Homoscedasticity," *Journal of the American Statistical Association*, 60(310), 539–547.

GOURIÉROUX, C., AND A. MONFORT (1994): "Testing Non-nested Hypotheses," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, chap. 44, pp. 2585–2640. Elsevier North-Holland, Oxford.

GRIEWANK, A., AND A. WALTHER (2008): *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, vol. 19 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, USA, 2 edn.

HAGAN, P. S., D. KUMAR, A. S. LESNIEWSKI, AND D. E. WOODWARD (2002): "Managing Smile Risk," *Wilmott Magazine*, pp. 84–108.

HALL, P., AND E. J. HANNAN (1988): "On stochastic complexity and nonparametric density estimation," *Biometrika*, 75(4), 704–714.

Hansen, B. E. (1990): "Lagrange Multiplier Tests for Parameter Stability in Non-Linear Models," Mimeo, University of Rochester.

Hardin, J. W., and J. M. Hilbe (2012): *Generalized Linear Models and Extensions*. Stata Press, College Station, TX, USA, 3 edn.

Heidelberger, P., and P. Welch (1983): "Simulation run length control in the presence of an initial transient," *Operations Research*, 31, 1109–1144.

Hero, A. O., B. Ma, O. J. Michel, and J. Gorman (2002): "Applications of Entropic Spanning Graphs," *IEEE Signal Processing Magazine*, 19(5), 85–95.

Herrick, D. R. M., G. P. Nason, and B. W. Silverman (2001): "Some New Methods for Wavelet Density Estimation," *Sankhyā, Series A*, 63(3), 394–411.

Hobijn, B., P. H. Franses, and M. Ooms (2004): "Generalizations of the KPSS-test for stationarity," *Statistica Neerlandica*, 58(4), 483–502.

Hoffman, M. D., and A. Gelman (2014): "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, forthcoming.

Hong, Y., and H. Li (2005): "Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates," *The Review of Financial Studies*, 18(1), 37–84.

Jeffreys, H. (1961): *Theory of Probability*. Oxford University Press, Oxford.

Kainth, D., J. Kwiatkowski, and D. Muirden (2010): "Modelling the CRM for the Correlation Trading Portfolio," Royal Bank of Scotland presentation.

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *The Review of Economic Studies*, 65(3), 361–393.

Kumar, V., J. Heikkonen, J. Rissanen, and K. Kaski (2006): "Minimum Description Length Denoising With Histogram Models," *IEEE Transactions on Signal Processing*, 54(8), 2922–2928.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root?," *Journal of Econometrics*, 54(1-3), 159–178.

Learned-Miller, E. G. (2004): "Hyperspacings and the Estimation of Information Theoretic Quantities," Technical Report 04-104, University of Massachusetts, Amherst.

Leimkuhler, B., and S. Reich (2004): *Simulating Hamiltonian Dynamics*, vol. 14 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, UK.

Levene, H. (1960): "Robust tests for equality of variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. by I. Olkin, pp. 278–292. Stanford University Press.

Lilliefors, H. W. (1967): "On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62(318), 399–402.

Liu, J. S. (2002): *Monte Carlo Stategies in Scientific Computing*, Springer Series in Statistics. Springer-Verlag, New York.

Lubrano, M. (2001): "Bayesian Model Choice for the Short Term US Interest Rate," Working paper, GREQAM-CNRS.

Mack, Y. P., and M. Rosenblatt (1979): "Multivariate k-nearest neighbor density estimates," *Journal of Multivariate Analysis*, 9(1), 1–15.

MEYER, R., AND J. YU (2000): "BUGS for a Bayesian analysis of stochastic volatility models," *Econometrics Journal*, 3, 198–215.

MURRAY, L. M. (2013): "Bayesian State-Space Modelling on High-Performance Hardware Using LibBi," Working Paper, CSIRO Mathematics, Informatics and Statistics.

NASON, G. P. (1994): "Wavelet Regression by Cross-Validation," Department of Statistics Technical Report 447, Stanford University, Stanford, CA.

NEAL, R. M. (2011): "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, Handbooks of Modern Statistical Methods, chap. 5, pp. 116–162. Chapman & Hall/CRC Press.

NESTEROV, Y. (2009): "Primal-dual subgradient methods for convex problems," *Mathematic Programming*, 120(1), 221–259.

NORDSTOKKE, D. W., AND B. D. ZUMBO (2010): "A New Nonparametric Levene Test for Equal Variances," *Psicólogica*, 31, 401–430.

NYBLOM, J. (1989): "Testing for Constancy of Parameters over Time," *Journal of the American Statistical Association*, 84(405), 223–230.

ORAVA, J. (2012): "k-Nearest Neighbour Kernel Density Estimation, The Choice of Optimal k," *Tatra Mountains Mathematical Publications*, 50(1), 39–50.

PALM, F. C., AND P. J. G. VLAAR (1997): "Simple Diagnostics Procedures for Modelling Financial Time Series," *Allgemeines Statistisches Archiv*, 81(1), 85–101.

PAPASPILIOPOULOS, O., G. O. ROBERTS, AND M. SKÖLD (2007): "A general framework for the parametrization of hierarchical models," *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 22(1), 59–73.

PESARAN, M. H. (1987): "Global and Partial Nonnested Hypotheses and Asymptotic Local Power," *Econometric Theory*, 3(1), 69–97.

PESARAN, M. H., AND M. WEEKS (1999): "Non-nested Hypothesis Testing: An Overview," Cambridge Working Papers in Economics 9918, University of Cambridge.

PLUMMER, M. (2011): *JAGS Version 3.1.0 User Manual*, International Agency for Research on Cancer.

RAMALHO, E. A., J. J. S. RAMALHO, AND J. M. R. MURTEIRA (2010): "Estimating and Testing Empirical Strategies for Fractional Regression Models," *Journal of Economic Surveys*, 25(2), 19–68.

REBONATO, R., S. MAHAL, M. JOSHI, L.-D. BUCHHOLZ, AND K. NYHOLM (2005): "Evolving yield curves in the real-world measures: a semi-parametric approach," *Journal of Risk*, 7(3), 29–61.

REBONATO, R., AND A. PODUGIN (2010): "Dependence of Magnitude of Rate Moves on Rate Levels," Discussion paper, Quantitative Client Research, The Royal Bank of Scotland.

RENÒ, R., A. ROMA, AND S. SCHAEFER (2006): "A Comparison of Alternative Non-parametric Estimators of the Short Rate Diffusion Coefficient," *Economic Notes*, 35(3), 227–252.

RESHEF, D. N., Y. A. RESHEF, H. K. FUNICANE, S. R. GROSSMAN, G. MCVEAN, P. J. TURNBAUGH, E. S. LANDER, M. MITZENMACHER, AND P. C. SABETI (2011): "Detecting Novel Associations in Large Data Sets," *Science*, 334, 1518–1524.

ROBERT, C. P., AND G. CASELLA (2000): *Monte Carlo Statistical Methods*, Springer Series in Statistics. Springer-Verlag, New York.

ROBERTS, G. O., AND S. K. SAHU (1997): "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), 291–317.

Sam, A. G., and G. J. Jiang (2009): "Nonparametric Estimation of the Short Rate Diffusion Process from a Panel of Yields," *Journal of Financial and Quantative Analysis*, 44(5), 1197–1230.

Stan Development Team (2013): "Stan: A C++ Library for Probability and Sampling, Version 2.1," http://mc-stan.org/.

Stanton, R. (1997): "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *The Journal of Finance*, 52(5), 1973–2002.

Trolle, A. B., and E. S. Schwartz (2009): "A General Stochastic Volatility Model for the Pricing of Interest Rate Derivatives," *The Review of Financial Studies*, 22(5), 2007–2057.

Vašíček, O. A. (1976): "A Test for Normality Based on Sample Entropy," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 38(1), 54–59.

Vuong, Q. H. (1989): "Likelihood Ratio Tests of Separate Families of Hypotheses in Time Series Analysis," *Econometrica*, 57(2), 307–333.

Walters-Williams, J., and Y. Li (2009): "Estimation of Mutual Information: A Survey," in *Rough Sets and Knowledge Technology: 4th International Conference, RSKT 2009*, ed. by R. Goebel, J. Siekman, and W. Wahlster, no. 5589 in Lecture Notes in Artificial Intelligence, pp. 389–396. Springer-Verlag.

Wasserman, L. (2006): *All of Nonparametric Statistics*, Springer Texts in Statistics. Springer.

White, H. (1980): "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48(4), 817–838.

Yu, J. (2005): "On leverage in a stochastic volatility model," *Journal of Econometrics*, 127(2), 165–178.

TEU