Financial Model Calibration Using Consistency Hints

Yaser S. Abu-Mostafa

***Abstract*—We introduce a technique for forcing the calibration of a financial model to produce valid parameters. The technique is based on learning from hints. It converts simple curve fitting into genuine calibration, where broad conclusions can be inferred from parameter values. The technique augments the error function of curve fitting with consistency hint error functions based on the Kullback–Leibler distance. We introduce an efficient EM-type optimization algorithm tailored to this technique. We also introduce other consistency hints, and balance their weights using canonical errors. We calibrate the correlated multifactor Vasicek model of interest rates, and apply it successfully to Japanese Yen swaps market and U.S. Dollar yield market.**

***Index Terms*—Canonical error, computational finance, consistency hint, cross entropy, EM algorithm, financial engineering, interest rates, Kullback–Leibler distance, model calibration, multifactor models, optimization, overfitting, Vasicek model, volatility term structure.**

I. Introduction

THE calibration of a financial model is the process of tuning the model parameters to fit market data. Unlike the parameters of generic learning models such as neural networks, the parameters of financial models correspond to economic and financial quantities. For instance, they might correspond to the volatility of a given market, or to the steady-state interest rate. These semantic aspects of the parameters are often lost in the process of "curve fitting." We may end up with a good fit that nonetheless assigns improbable or contradictory values to the parameters. For instance, we may fit the prices of bonds very well, only to find that a volatility parameter in the formula is five times what it should be. Such an inconsistency needs to be avoided since the plausibility of the solution depends on the plausibility of the model it is based on.

In order to force the calibration process to conform with the characteristics of the model parameters, we will supplement it with consistency *hints* about these parameters. Hints [2], [3] are the auxiliary pieces of information appended to the data to help direct the learning process toward more plausible solutions. Consistency hints can have a dramatic impact on the calibration. A case in point is illustrated in Figs. 1 and 2. Fig. 1 shows the results of fitting market data with and without consistency hints. Both fits appear to be equally good, and the hints do not seem to make a difference. However, a huge difference is shown in Fig. 2. Using the parameter values from the two calibrations of Fig. 1, we computed the market volatility implied by these parameters. When the hints are used, the volatility is in almost perfect agreement with the historical value it is meant to predict. When the hints are not used, the volatility is completely off. This contrast could not have been detected by comparing the two fits of Fig. 1, on which the calibrations were based.

Hints were first introduced in the context of neural networks [1] to reduce overfitting, which results from having too many weights [12]. Such redundancy allows the learning algorithm to fit idiosyncrasies of the training data that have nothing to do with the function being learned. Inconsistency in calibration is a manifestation of overfitting, too. As we saw in Fig. 1, we can fit the same set of market data with different sets of parameters, some consistent and some not. This means that the parameters are redundant, and therefore susceptible to overfitting. Since hints must always be valid properties in the context they are used, they will steer the fit toward the more consistent solution.

The calibration of complex models is more prone to overfitting than that of simple models, since complex models have more parameters that can be exploited in the fit. Without techniques such as consistency hints, complex models may have to be avoided altogether because of this drawback. However, these complex models are needed to explain the behavior of financial markets more accurately. For instance, multi-factor interest rate models are more realistic in representing the behavior of interest rates than single-factor models. Consistency hints impose an increasingly tighter constraint on higher-order models, thus regulating the overfitting potential proportionately.

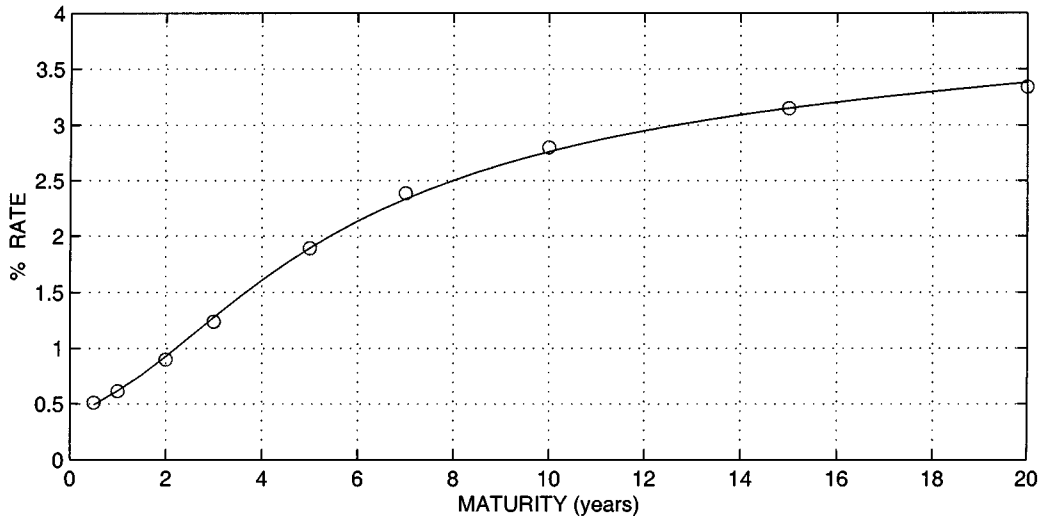
Depending on the application, the use of consistency hints may be crucial to the final results. Although the calibration is concerned with fitting market data, we are not just after a good fit, but also a *correct* fit. The fit may be only a means to infer other quantities, such as the volatility of Fig. 2. The fit may also be used to help a specific application, such as *relative-value trading*, which is based on whether the model prediction is higher or lower than the current market value. Even for two equally good fits like those of Fig. 1, this prediction can be different. For instance, the model prediction of the 15-year par rate is higher than the market value when hints are used, but it is about the same as the market value when hints are not used. If we are going to base a trade on the model prediction, we must have a reason to believe that one fit or the other is more credible, beyond just being a good-looking fit.

To describe how consistency hints are used in financial model calibration, we will consider a multi-factor interest rate model. Section II introduces this model and develops the basic framework for calibration. Section III defines consistency hints and derives the formulas that quantify the hint errors. Section IV

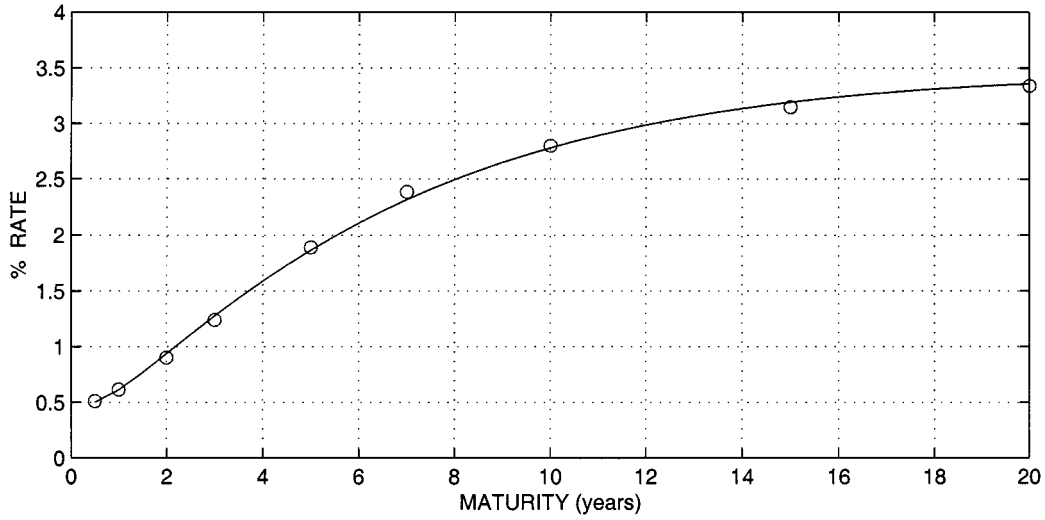
Manuscript received December 30, 2000; revised March 26, 2001. This work was supported by the Center for Neuromorphic Systems Engineering, an Engineering Research Center supported by the National Science Foundation under NSF Cooperative Agreement EEC 9402726.

The author is with the Learning Systems Group, Departments of Electrical Engineering, Computer Science, and Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125 USA.

Publisher Item Identifier S 1045-9227(01)05017-2.



(a)



(b)

Fig. 1. The results of calibrating a financial model to swaps market data, with and without consistency hints. The two fits are virtually indistinguishable. (a) Fitting swap par rates without hints. (b) Fitting swap par rates with hints.

discusses implementation issues and experimental results. Section V takes a look at calibration from a probabilistic point of view, and provides a more principled framework for our techniques, including the introduction of canonical errors. Finally, for self sufficiency, the Appendix provides brief mathematical derivations for the main functions of the interest rate model we use.

II. The Interest-Rate Model

Interest-rate models are among the more sophisticated financial models, and their calibration is quite challenging. We are going to use the Vasicek model for interest rates [14], [18] as a paradigm for employing consistency hints in the calibration of financial models. This concrete example will enable us to do a full derivation of the consistency hint equations and to illustrate the numerical results using real-life data. It is fairly straightforward to adapt our method to the calibration of other interest-rate models that have analytic solutions, as well as to analogous financial models that deal with other markets.

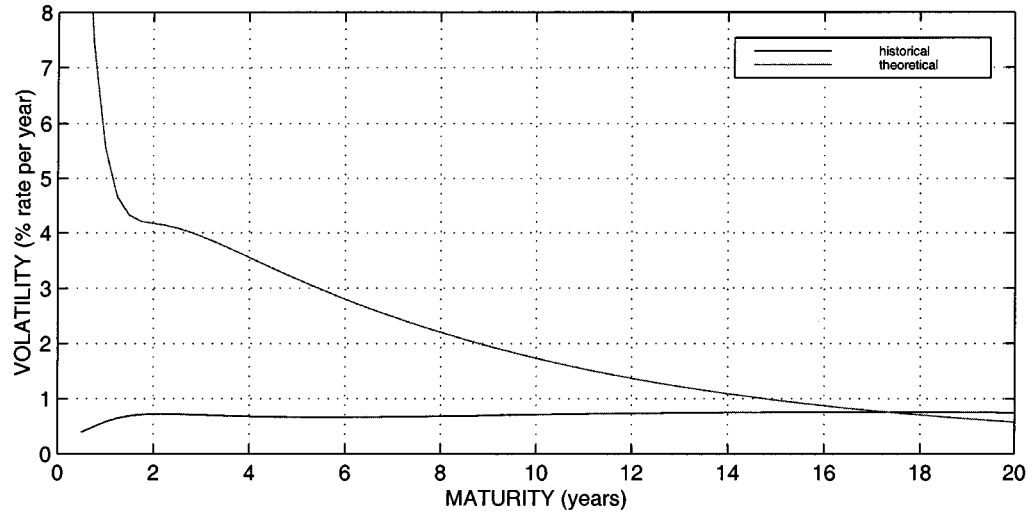
A. Vasicek Model

The premise of the Vasicek model is that the evolution of interest rates in time is driven by two forces. The first is a 'drift' toward a steady-state or equilibrium value of what the interest rate should be. The second is an injection of random movements into the interest rate as a result of the unpredictable economic environment. How these two forces interact is what defines a Vasicek model.

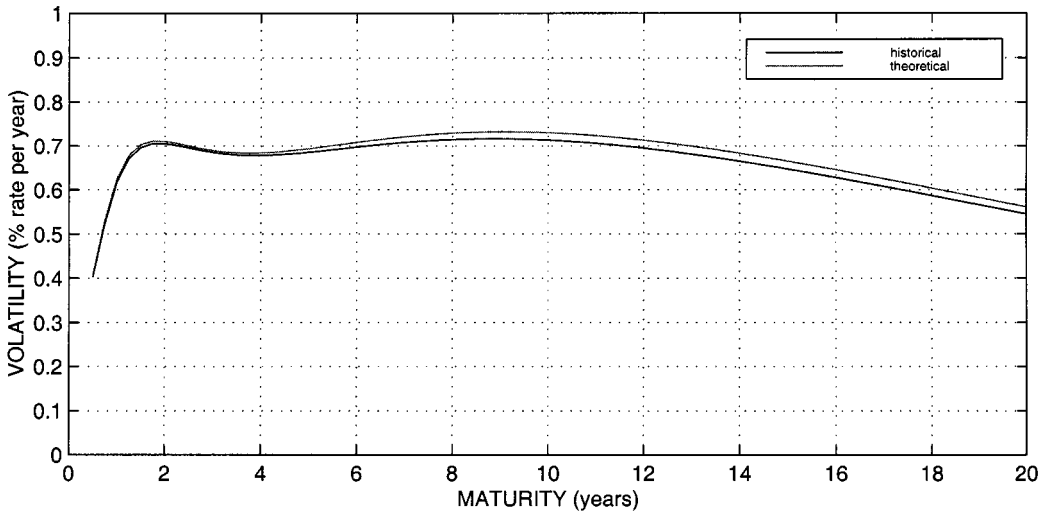
In its simplest form, the model uses a steady-state interest rate θ, a speed k of converging to that steady state, and a volatility or "randomness level" σ, to describe the instantaneous interest rate as a function x governed by the equation

$$dx = k(\theta - x)\,dt + \sigma\,dW$$

where dt is the infinitesimal increment in time, and dW is an infinitesimal stochastic variable (W is formally a Wiener process). The drift element is captured by the $k(\theta - x)\,dt$ portion of dx, and indeed this term pushes x toward θ. If $x > \theta$, this term



(a)



(b)

Fig. 2. The volatility term structure of forward rates (6 months to 20 years) corresponding to the fits of Fig. 1. (a) Historical versus theoretical volatility without hints. (b) Historical versus theoretical volatility with hints. In spite of those fits being almost identical, the theoretical volatility in (a) is in gross violation of the historical volatility it is supposed to predict, while in (b) they are in almost perfect agreement. Consistency hints are not used in (a), but used in (b).

is negative, hence x will drift downwards toward θ, while if $x < \theta$, this term is positive hence x will drift upwards, again toward θ. The value of k modulates the change dx that results in this drift, and hence determines the speed of converging to the steady state θ. The $\sigma\, dW$ portion of dx adds the random component to the interest rate. x accumulates the different $\sigma\, dW$'s that occur as time goes by, but this accumulated random component is subject to decaying as x drifts toward θ by virtue of the $k(\theta - x)\, dt$ term. Fig. 3 shows an evolution of the instantaneous interest rate under this model.

The focus of this paper is not the stochastic differential equation (SDE) itself, but the functions of interest rate that are derived from the SDE. The parameters of the SDE will appear in the expressions of these functions (see the Appendix), and when the functions are calibrated to market data, the values of the parameters are determined. The understanding of what these parameters signify and how they interact is important to appreciate how consistency hints come into play.

With this in mind, let us illustrate the more general form of the Vasicek model. This form is called the multifactor model because it asserts that the interest rate is not just a single x as in the above equation, but rather a superposition of several x's of analogous form. These x's are the "factors," and each of them follows the same basic equation. Thus,

$$dx_n = k_n(\theta_n - x_n)\, dt + \sigma_n\, dW_n$$

for $n = 1, \ldots, N$ where N is the number of factors. The interest rate r is given by the sum of these factors

$$r = \sum_{n=1}^{N} x_n$$

The philosophy behind having multiple factors stems from the observation that there are different time scales for the behavior of interest rates. Some aspects of this behavior are observed in a short time horizon (high-speed factors or large k_n),

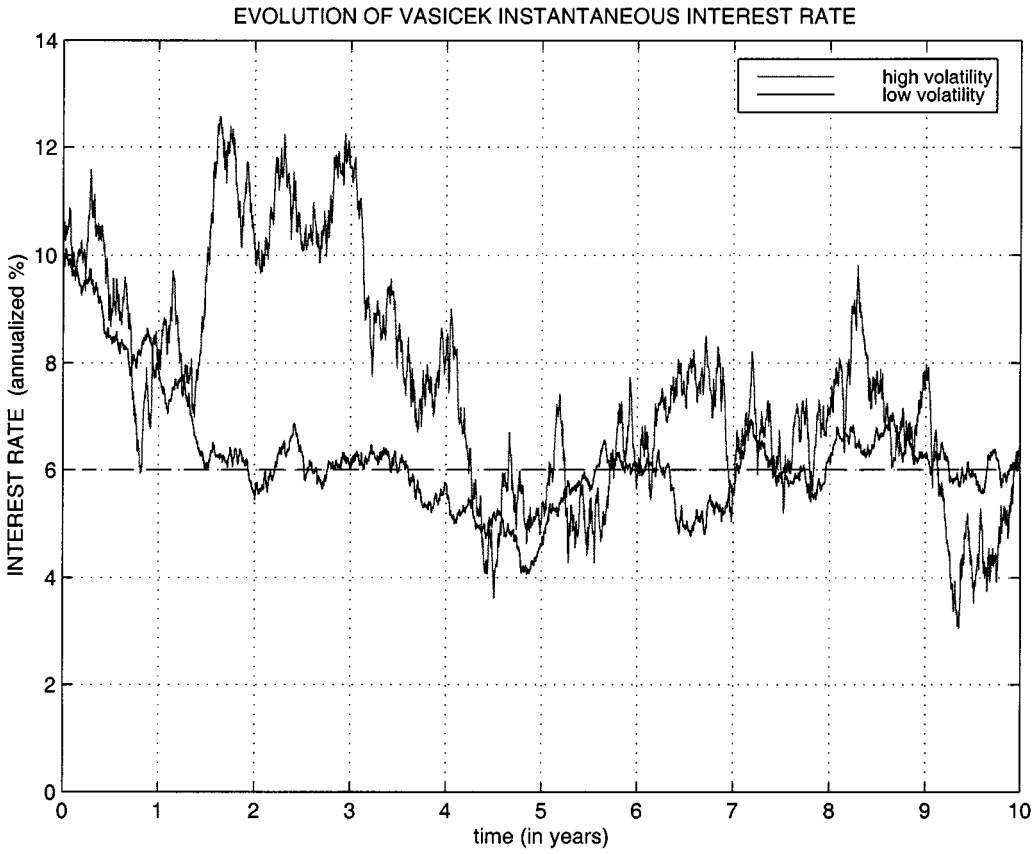


Fig. 3. Simulation of instantaneous interest rates under the Vasicek model. Two scenarios with different volatilities are presented for the same steady-state rate of 6%, and the same mean reversion speed.

and some aspects are observed in a longer horizon (low-speed factors or small k_n). Each factor has its own steady-state θ_n and its own volatility σ_n. The corresponding stochastic elements dW_n are not always independent, hence there are correlation coefficients ρ_{ij} between dW_i and dW_j as part of the model parameters. The model is sometimes referred to as a *correlated* multi-factor Vasicek.

It is obvious that the multi-factor model provides more flexibility for fitting the data by introducing more parameters that can be exploited in the calibration process. Therefore, a 3-factor Vasicek model is more powerful than a 2-factor Vasicek model. By the same token, the 3-factor Vasicek model will be more prone to overfitting, i.e., to fitting the idiosyncrasies of a particular data set at the expense of proper generalization to new data, because it has more resources for such a fit. This problem limits the number of factors that can be used in practice, even if more factors are needed to model real markets. Multifactor models need techniques like the ones we are introducing in this paper to be reliably calibrated. Consistency hints constrain the multitude of parameters in these models so as to keep overfitting in check. The constraining is based on legitimate rules that may be inadvertently violated if the calibration is done without the hints.

B. Calibration

We now address how the Vasicek model is used to fit market data, or, equivalently, how market data is used to calibrate the Vasicek model. Let $\mathbf{p}$ denote the vector of all the parameters in the Vasicek N-factor model. A market function related to interest rates, be it the price of a 30-year bond or the yield of three-month CD, will have a theoretical value based on the model that is function of $\mathbf{p}$, say $f(\mathbf{p})$. It will also have an actual value observed in the market, say $\tilde{f}$. If the model is correct, and the value of $\mathbf{p}$ is chosen properly, we would have

$$f(\mathbf{p}) = \tilde{f}$$

Since the model is not perfect, we have to settle for a $f(\mathbf{p})$ that comes closest to the above equation. For instance, we can pick the value of $\mathbf{p}$ that minimizes the error function

$$E_0 = (f(\mathbf{p}) - \tilde{f})^2$$

If we have several market functions $f_1, \ldots, f_M$, say the prices of bonds of different maturities, we can minimize

$$E_0 = E_0(\mathbf{p}) = \sum_{m=1}^{M} (f_m(\mathbf{p}) - \tilde{f}_m)^2$$

Variations of this error measure are of course possible. We will refer to this error as the *fit error*, as distinct from the *consistency error* to be introduced in Section III.

Calibrating the model to market data is the process of determining $\mathbf{p}$ that minimizes the error. It is no different from computing the weights of a neural network by minimizing the error between the network prediction and the actual data, except

that the "weights" here are parameters coming from a financial model.

The Appendix shows how $f_m(\mathbf{p})$ can be derived from the Vasicek model SDEs for different market functions. Once a formula for $f_m(\mathbf{p})$ is obtained, the calibration process can proceed without involving the SDEs themselves. In our experiments, we use two sets of market functions. The first set consists of par rates in the Japanese Yen swaps market, and the second set consists of the yield of the US Dollar for different maturities. The market values for the swaps and the yield can change from day to day, if not from moment to moment. Therefore, the calibration attempts to simultaneously fit quantities occurring at different times, e.g., at the daily close of the market. The same notation of $f_m(\mathbf{p})$ will still work in this case since the index $m = 1, \ldots, M$ can refer to the same type of function but at different times, or to different types of functions. As long as there is a model-based formula for each $f_m(\mathbf{p})$ used in the fit, no notational distinction is needed.

C. Discrete Time

If we calibrate the Vasicek model based on market data available at a discrete-time sequence $t[1] < t[2] < \cdots < t[l] < \cdots < t[L]$, e.g., at the daily close of the market, it is helpful to view the model through discrete-time difference equations that approximate the continuous-time SDE's (see the Appendix for more details). The index l of the discrete-time sequence is made explicit in these difference equations

$$\Delta x_n[l] = k_n(\theta_n - x_n[l])\Delta t[l] + \sigma_n w_n[l]\sqrt{\Delta t[l]}$$

for $n = 1, \ldots, N$ and $l = 1, \ldots, L-1$, where

$$\begin{aligned} \Delta t[l] &= t[l+1] - t[l] \qquad l = 1, \ldots, L-1 \\ \Delta x_n[l] &= x_n[l+1] - x_n[l] \qquad l = 1, \ldots, L-1 \quad \text{and} \\ & \qquad n = 1, \ldots, N. \end{aligned}$$

The stochastic elements $w_n[l]$ are normally distributed with zero mean and a covariance given by

$$\mathcal{E}(w_i[l]w_j[l]) = \rho_{ij}$$

for $i, j = 1, \ldots, N$ and $l = 1, \ldots, L-1$, with $\rho_{ii} = 1$. Each $w_n[l]$ is independent of all the others with different l. The instantaneous interest rate r is given by

$$r[l] = \sum_{n=1}^{N} x_n[l] \qquad l = 1, \ldots, L \quad \text{and} \quad n = 1, \ldots, N.$$

Numerical simulations of the Vasicek model, such as the one used to generate Fig. 3, are based on this discrete-time version.

The discrete model spells out the parameters $\mathbf{p}$ that go into the calibration process. $\mathbf{p}$ consists of long-term parameters or constants, and short-term parameters or state variables. The long-term parameters, denoted by $\mathbf{p}_{\mathcal{L}}$, are

$$\begin{aligned} \text{speeds of mean reversion:} \quad & k_n \qquad n = 1, \ldots, N \\ \text{steady-state means:} \quad & \theta_n \qquad n = 1, \ldots, N \\ \text{volatilities:} \quad & \sigma_n \qquad n = 1, \ldots, N \\ \text{correlations:} \quad & \rho_{ij} \qquad i, j = 1, \ldots, N. \end{aligned}$$

Long-term parameters are constant with regard to the time index l. Short-term state variables, denoted by $\mathbf{p}_{\mathcal{S}}$, depend on l

$$\text{state variables:} \quad x_n[l] \qquad n = 1, \ldots, N \quad \text{and} \quad l = 1, \ldots, L.$$

There is a total of $\frac{1}{2}(N^2 + 5N)$ long-term parameters[1] in an N-factor Vasicek model, and a total of NL state variables when we have market data at L discrete-time instances. Hence

$$\mathbf{p} = (\mathbf{p}_{\mathcal{L}}, \mathbf{p}_{\mathcal{S}}) \text{ has } \frac{N}{2}(N + 2L + 5) \text{ parameters.}$$

Once both $\mathbf{p}_{\mathcal{L}}$ and $\mathbf{p}_{\mathcal{S}}$ are determined through calibration, the values of the stochastic elements $w_n[l]$ can be solved for using the model difference equations. It is through $w_n[l]$ that consistency will be defined.

III. Consistency Hints

The calibration of a Vasicek model infers the values of the parameters $\mathbf{p}$ by minimizing the error between the model-based functions $f_m(\mathbf{p})$ and the market values $\tilde{f}_m$. As we have shown in Figs. 1(a) and 2(a), it is possible to attain a very small error between $f_m(\mathbf{p})$ and $\tilde{f}_m$, while creating a huge discrepancy between other model-based functions and their market values. It is conceivable that the problem is inherent, i.e., the model is not powerful enough to match all these quantities simultaneously. However, as we saw in Fig. 1(b) and Fig. 2(b), the Vasicek model has no such limitation. There is another "consistent" solution for the parameters $\mathbf{p}$ that achieves an equally good fit without the discrepancy. Indeed, the redundancy of the parameters $\mathbf{p}$ in the expression of $f_m(\mathbf{p})$ allows for several solutions, possibly infinitely many. Some of these solutions are consistent, and some are not. How do we make sure that the calibration process picks a consistent $\mathbf{p}$? To answer this, we first need to spell out exactly what it means for $\mathbf{p}$ to be consistent.

A. Consistency

The criterion for consistency cannot be based merely on the ability to fit many quantities simultaneously, for the issue would then be confused with the sheer power of the model. Instead, consistency would reconcile the role of $\mathbf{p}$ as generic parameters in a formula $f_m(\mathbf{p})$ used for fitting, with their role as meaningful quantities in the basic equations that gave rise to that formula. In doing so, it produces parameters that stand the best chance of fitting other functions that can be legitimately derived from the same set of basic equations.

Let us see how this applies to the Vasicek model. Consider the basic equation of the discrete-time version

$$\Delta x_n[l] = k_n(\theta_n - x_n[l])\Delta t[l] + \sigma_n w_n[l]\sqrt{\Delta t[l]}.$$

After the calibration is done, one can substitute the values of the fitted parameters in the above equation and solve for the "implied" $w_n[l]$, i.e., the particular realization of the stochastic elements $w_n[l]$ that must have occurred to generate this fit. However, there are basic assumptions about the statistics of $w_n[l]$ that were utilized in deriving the $f_m(\mathbf{p})$ functions used for the fit. If

[1]Counting ρ_{ij} for only $i > j$ since $\rho_{ii} = 1$ and $\rho_{ij} = \rho_{ji}$.

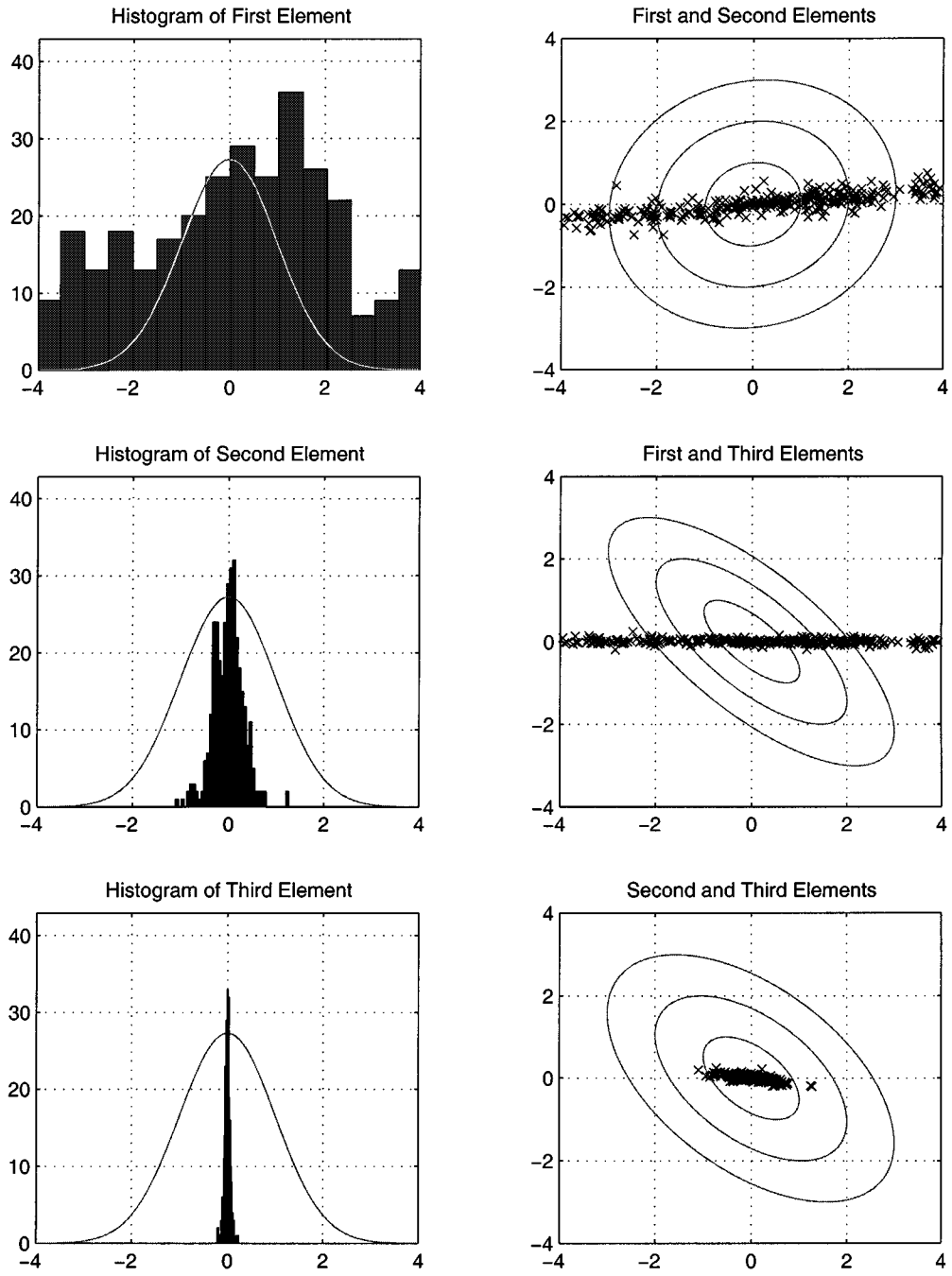


Fig. 4. Histograms and scatter diagrams of the implied stochastic elements from a calibration without consistency hints. The superimposed curves are the theoretical density and the σ, 2σ, and 3σ contours that the sample is supposed to follow, but grossly violates.

the implied $w_n[l]$ do not satisfy these assumptions, the fit is inconsistent with the model it is based on. This leads us to the following rule.

> Consistency Hint: The stochastic elements implied by the fit should obey the statistical assumptions of the model.

This rule enforces the desired property at the level of the building blocks of the model. The consistency of other "higher level" functions will follow suit, since they are derived from these building blocks. Indeed, the discrepancy of Fig. 2(a) can be traced back to a violation of the consistency hint. Fig. 4 shows the histograms and scatter diagrams of $w_n[l]$ without the hint. Also shown are the theoretical curves of where things should be according to the assumptions of the model. Fig. 4 corresponds to the fit of Fig. 1(a), and it is interesting to see how such a legitimate-looking fit has the hidden gross violation of statistics depicted in Fig. 4.

Fig. 5 shows that the histograms and scatter diagrams are far better behaved when the hint is used. These correspond to the fit of Fig. 1(b) and the volatility term structures of Fig. 2(b). As we argued, the higher-level functions in Fig. 2 inherit the consistency of the basic model.

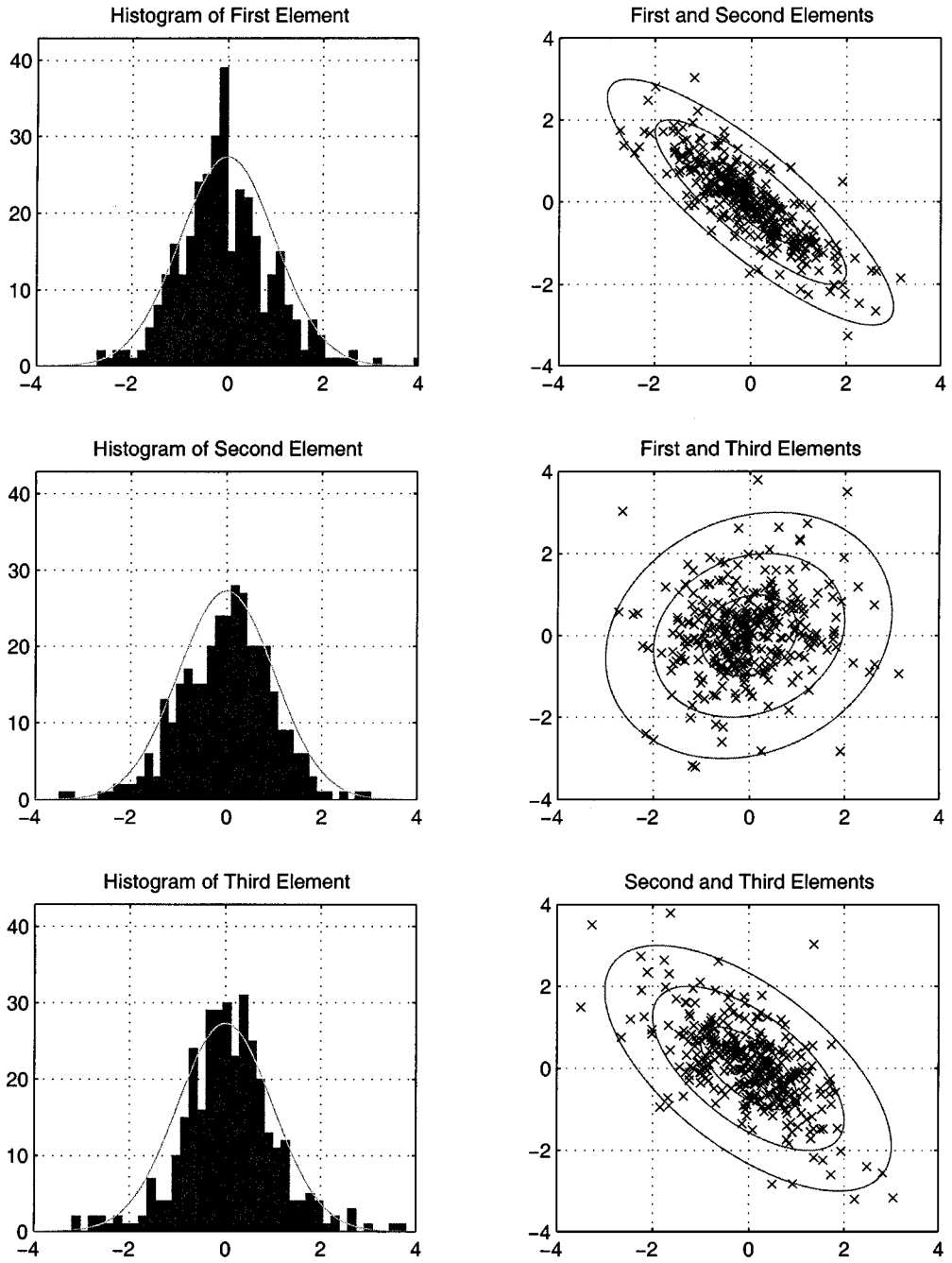


Fig. 5. Histograms and scatter diagrams of the implied stochastic elements from a calibration with consistency hints. The superimposed curves are the theoretical density and the σ, 2σ, and 3σ contours. Compared to Fig. 4, the theoretical distributions are largely followed.

B. Entropy Measure

To formalize the consistency hint, we need to quantify the agreement/disagreement between the distribution of the implied $w_n[l]$ and the distribution of the theoretical $w_n[l]$. One obvious way of doing this is by measuring the Kullback–Leibler distance $K(p||q)$ [9] between the two distributions. Given two probability density functions (pdf's) $p(u)$ and $q(u)$, $K(p||q)$ is defined by

$$K(p||q) = \int p(u) \log \frac{p(u)}{q(u)} du.$$

The Kullback-Leibler distance has the property that $K(p||q) \geq 0$ with equality if, and only if, $p = q$. It can serve as an "error function" to be minimized in order to match p to q.

Let $\mathbf{w}[l] = (w_1[l], w_2[l], \ldots, w_N[l])^T$ (column vector), and let $p(\mathbf{w})$ be the pdf of the implied $\mathbf{w}[l]$[2] and $q(\mathbf{w})$ be the pdf of the theoretical $\mathbf{w}[l]$. While q can be written explicitly as a Gaussian in terms of the model parameters, p is not explicitly known. It is only represented by a sample (the implied $\mathbf{w}[l]$; $l = 1, \ldots, L-1$ that p 'generated'). To evaluate $K(p||q)$, we can employ density estimation techniques [17] to get p, then

[2]Assuming the implied $\mathbf{w}[l]$ are identically distributed for different l, like their theoretical counterparts.

evaluate the integral. Alternatively, we can try to estimate the integral directly from the sample. We can rewrite[3]

$$K(p||q) = \int p(\mathbf{w}) \log \frac{1}{q(\mathbf{w})} d\mathbf{w} - \int p(\mathbf{w}) \log \frac{1}{p(\mathbf{w})} d\mathbf{w}.$$

The first term is the *cross entropy* between p and q, and the second term is the *entropy* of p. Since the form of p is unknown, we use the maximum-entropy principle [16] to estimate the second term. If P is the covariance matrix of p, the maximum-entropy value of $\int p(\mathbf{w}) \log(1/p(\mathbf{w}))\, d\mathbf{w}$ occurs when p is Gaussian. We evaluate this integral and further reduce the expression of $K(p||q)$ to[4]

$$\tfrac{1}{2}\left(\log(|Q|) + \int p(\mathbf{w})(\mathbf{w}^T Q^{-1}\mathbf{w})\, d\mathbf{w} - \log(|P|) - N\right)$$

where $|\cdot|$ denotes the determinant, $Q = [\rho_{ij}]$ is the covariance matrix of q (ρ_{ij} come from the Vasicek model), and N is the dimension of $\mathbf{w}$ (the number of Vasicek factors). To estimate the remaining integral, we use the sample average[5]

$$\frac{1}{L-1} \sum_{l=1}^{L-1} \mathbf{w}[l]^T Q^{-1} \mathbf{w}[l]$$

and to estimate $|P|$, we use the sample covariance matrix Σ of $\mathbf{w}[l]$; $l = 1, \ldots, L-1$. Hence, we get the entropy-based expression

$$\frac{1}{2}\left(\log(|Q|) - \log(|\Sigma|) - N + \frac{1}{L-1} \sum_{l=1}^{L-1} \mathbf{w}[l]^T Q^{-1} \mathbf{w}[l]\right)$$

as an estimate for $K(p||q)$ that can be completely determined from the model parameters. Dropping the $\frac{1}{2}$, we arrive at our first consistency hint error function

$$E_1 = \log(|Q|) - \log(|\Sigma|) - N + \frac{1}{L-1} \sum_{l=1}^{L-1} \mathbf{w}[l]^T Q^{-1} \mathbf{w}[l]$$

which becomes part of the overall objective function together with the fit error E_0. Notice that E_1 is an "optimistic" estimate, since the actual entropy of p may not be as big as the maximum-entropy estimate. Notice also that finite-sample variations may drive the value of E_1 slightly negative (Fig. 10(a)) in spite of $K(p||q)$ being strictly nonnegative.

C. Initial State

The error function E_1 quantifies the consistency of the stochastic elements $\mathbf{w}[l]$; $l = 1, \ldots, L-1$. In addition to $\mathbf{w}[l]$, there is another stochastic element in the Vasicek model, which is the initial state $\mathbf{x}[1] = (x_1[1], x_2[1], \ldots, x_N[1])^T$. The initial state is stochastic because it accumulates all the stochastic elements that happened from $t = -\infty$ until $t = t[1]$, the earliest time in which market data is available for calibration. To find the statistics of the initial state, we start from the integral equation for the continuous-time x_n in the Appendix

$$x_n(t_2) = x_n(t_1) e^{-k_n(t_2 - t_1)} + \theta_n(1 - e^{-k_n(t_2 - t_1)}) + \sigma_n e^{-k_n t_2} \int_{t_1}^{t_2} e^{k_n \tau}\, dW_n(\tau).$$

Substituting $t_1 = -\infty$ and $t_2 = t$ (the initial time), we get

$$x_n(t) = \theta_n + \sigma_n e^{-k_n t} \int_{-\infty}^{t} e^{k_n \tau}\, dW_n(\tau).$$

Therefore, the initial x_n are jointly Gaussian with mean

$$E(x_n) = \theta_n$$

and covariance

$$\mathcal{E}((x_i - \theta_i)(x_j - \theta_j)) = \frac{\sigma_i \sigma_j \rho_{ij}}{k_i + k_j}$$

by an argument similar to that in the Appendix.

$\mathbf{x}[1]$ together with $\mathbf{w}[l]$; $l = 1, \ldots, L-1$ determine all the state variables of the model by induction. Since $\mathbf{x}[1]$ is independent of $\mathbf{w}[l]$, consistency would also require that the implied $\mathbf{x}[1]$ be reconciled with the model statistics. Defining consistency for $\mathbf{x}[1]$ is more problematic than for $\mathbf{w}[l]$, since we have a single implied $\mathbf{x}[1]$ as opposed to $L-1$ implied elements in the case of $\mathbf{w}[l]$. One definition is based on maximizing the value of the pdf, which results in the hint error function

$$E_2 = \log(|S|) + (\mathbf{x}[1] - \Theta)^T S^{-1} (\mathbf{x}[1] - \Theta)$$

where $\Theta = (\theta_1, \ldots, \theta_N)^T$, and S is the covariance matrix $[\sigma_i \sigma_j \rho_{ij}/k_i + k_j]$. Another related definition drops the $\log(|S|)$ term from the expression of E_2. This version measures how far the initial state is from its expected value, in units of variance along each coordinate.

The three errors E_0, E_1, and E_2 are merged to create a single objective function $\hat{E}(E_0, E_1, E_2)$ to be minimized. $\hat{E}(E_0, E_1, E_2)$ can be a simple weighted sum of E_0, E_1 and E_2, as we used in the experiments of Section IV, or can be a more principled combination as discussed in Section V.

IV. Implementation

In this section, we address the practical aspects of calibration using consistency hints, and discuss experimental results for Japanese Yen swaps and US Dollar yield data.

A. The Algorithm

Let $t[1] < t[1] < \cdots < t[L]$ be the calibration window, i.e., the times when market data are available, and let $\tilde{f}_m$; $m = 1, \ldots, M$, be the market data. The calibration algorithm determines the values of the parameters $\mathbf{p}$ that optimize the objective function $\hat{E}(E_0, E_1, E_2)$. First, we describe how the algorithm evaluates $\hat{E}$ for a given $\mathbf{p}$, then we turn our attention to optimization.

$\mathbf{p}$ consists of long-term parameters $\mathbf{p}_{\mathcal{L}}$, namely the Vasicek constants $k_n, \theta_n, \sigma_n, \rho_{ij}$, and short-term parameters $\mathbf{p}_{\mathcal{S}}$, namely the state variables $x_n[l]$. Given $\mathbf{p}_{\mathcal{L}}$ and $\mathbf{p}_{\mathcal{S}}$, we can evaluate the market functions $f_m(\mathbf{p})$; $m = 1, \ldots, M$ using the formulas

[3]We use a simplified notation for the multiple integral.

[4]Throughout the paper, we use standard properties of Gaussian distributions [6], [8], [10], [11].

[5]An efficient estimator if $\mathbf{w}[l]$ are statistically independent for different l.

TABLE I

Specifications of the Data		
	JPY Swaps	USD Yield
Dates	11/27/96 - 3/20/98	1/1/84 - 12/31/88
Calibration Window size	343 trading days	1247 trading days
Maturities in years	0.5,1,2,3,5,7,10,15,20	0.25,0.5,1,2,3,5,7,10,30
Average Interest Rate	1.78 %	8.73 %
Market Conditions	crisis[6], non-trending	normal, trending

derived in the Appendix. Therefore, we can evaluate $E_0 = \sum_{m=1}^{M}(f_m(\mathbf{p})-\tilde{f}_m)^2$. To evaluate E_1, we need the implied stochastic elements $\mathbf{w}[l]; l = 1, \ldots, L-1$. We can solve for $w_n[l]; l = 1, \ldots, L-1, n = 1, \ldots, N$, in terms of $x_n[l]; l = 1, \ldots, L, n = 1, \ldots, N$, using the Vasicek difference equations. We get

$$w_n[l] = \frac{(x_n[l+1] - x_n[l]) - k_n(\theta_n - x_n[l])(t[l+1] - t[l])}{\sigma_n \sqrt{t[l+1] - t[l]}.}$$

To evaluate E_2, we use the initial state $x_n[1]; n = 1, \ldots, N$. Finally, E_0, E_1, and E_2 are substituted into the expression for $\hat{E}(E_0, E_1, E_2)$. We thus have evaluated $\hat{E}$ as a function of $\mathbf{p}$.

For optimization, since $\hat{E}$ is highly nonlinear in $\mathbf{p}$, an iterative method such as conjugate gradient [7] is employed. The gradient of $\hat{E}$ is needed for such a method, but a numerical gradient can be used. At every iteration, the gradient of $\hat{E}$ with regard to all parameters is evaluated. This creates a computational bottleneck, since a typical calibration may have more than 1000 parameters.

A closer look at the functional dependencies reveals that the errors and parameters can be organized into two categories, leading us to a more efficient, EM-type optimization [5]. The short-term parameters $\mathbf{p}_{\mathcal{S}}$ are handled separately from the long-term parameters $\mathbf{p}_{\mathcal{L}}$, and the fit error E_0 is handled differently from the hint errors E_1 and E_2. The algorithm works as follows.

Initialize $\mathbf{p}_{\mathcal{L}}$ to a fixed value, and initialize the corresponding $\mathbf{p}_{\mathcal{S}}$ by minimizing E_0. Repeat the following two steps:

1) Minimize $\hat{E}(E_0, E_1, E_2)$ with regard to $\mathbf{p}_{\mathcal{L}}$, while holding $\mathbf{p}_{\mathcal{S}}$ constant.
2) Minimize E_0 w.r.t. $\mathbf{p}_{\mathcal{S}}$, while holding $\mathbf{p}_{\mathcal{L}}$ constant.

In Step 1, the state variables are fixed, and the objective function $\hat{E}$ is minimized with regard to the long-term parameters (12 in total for the 3-factor Vasicek used in our experiments). Step 2, as well as the initialization step, minimize the fit error E_0 only. The function $f_m(\mathbf{p})$, which is the main ingredient of E_0, depends on the long-term parameters and only N state variables (those corresponding to time $t[l]$, when the data point $\tilde{f}_m$ is observed). Therefore, for fixed long-term parameters, each term in the sum $E_0 = \sum_{m=1}^{M}(f_m(\mathbf{p}) - \tilde{f}_m)^2$ can be minimized *separately* with regard to only N variables ($N = 3$ in our experiments). Notice that, while the total number of parameters grows with the size of the calibration window L, the number of parameters to be optimized at one time using this algorithm does not change, which allows the computation to scale well.

In spite of having no guarantee of convergence (since the two steps have different objective functions), the algorithm works well in practice. It usually reaches a good value of $\hat{E}$ in less than 20 iterations of steps 1 and 2.

Since the values of k_n, σ_n, ρ_{ij} are constrained by the model ($k_n > 0$, $\sigma_n \geq 0$, and $[\rho_{ij}]$ is positive definite), the optimization in question is a constrained type. However, the constraints can be enforced by defining k_n in terms of another variable κ_n as e^{κ_n} or $\kappa_n^2 + \epsilon$, by absorbing the sign of σ_n in ρ_{ij}, and by adding a penalty term if any eigenvalue of $[\rho_{ij}]$ becomes smaller than ϵ. Within few iterations, the solution usually steers clear of the penalty area.

B. Experimental Results

We ran the calibration algorithm with and without consistency hints on two sets of interest rate market data, the Japanese Yen swaps and the US Dollar yield. In both cases, we calibrated a 3-factor Vasicek model on daily market data, using the market close values for nine different maturities of swaps and yield. Table I[6] compares the two data sets.

The goal of these experiments is to assess *how consistency hints affect calibration*, rather than to evaluate the calibration method itself, the Vasicek model, or the optimization algorithm. Figs. 1, 2, 4, and 5 in the previous sections show the results of the JPY swaps experiment. We now present additional results from the USD yield experiment.

Fig. 6 shows the time evolution of the three state variables of the Vasicek model when the USD yield calibration uses consistency hints. Also shown is the theoretical range within which these variables should (and do) evolve. In contrast, Fig. 7 shows the case without the hints. The state variables are in gross violation of the range they should lie within.

Fig. 8 shows the time evolution of the instantaneous rate for the USD yield, with and without the hints. In spite of the two calibrated models being quite different, the instantaneous rates are similar, since they affect the value of the yield and we are using the same yield data in both cases. The situation is analogous to Figs. 1 and 2, where just looking at the two fits would not reveal the fundamental differences between the underlying models, but these differences result in vastly different volatility term structures.

Finally, we show the impact of enforcing consistency hints on the quality of the fit. It is conceivable that the hints may significantly constrain the fitting of the data, and a much worse

[6]The onset of the Asian crisis in 1997 inflated the short-term rates by what was called the "Japan premium."

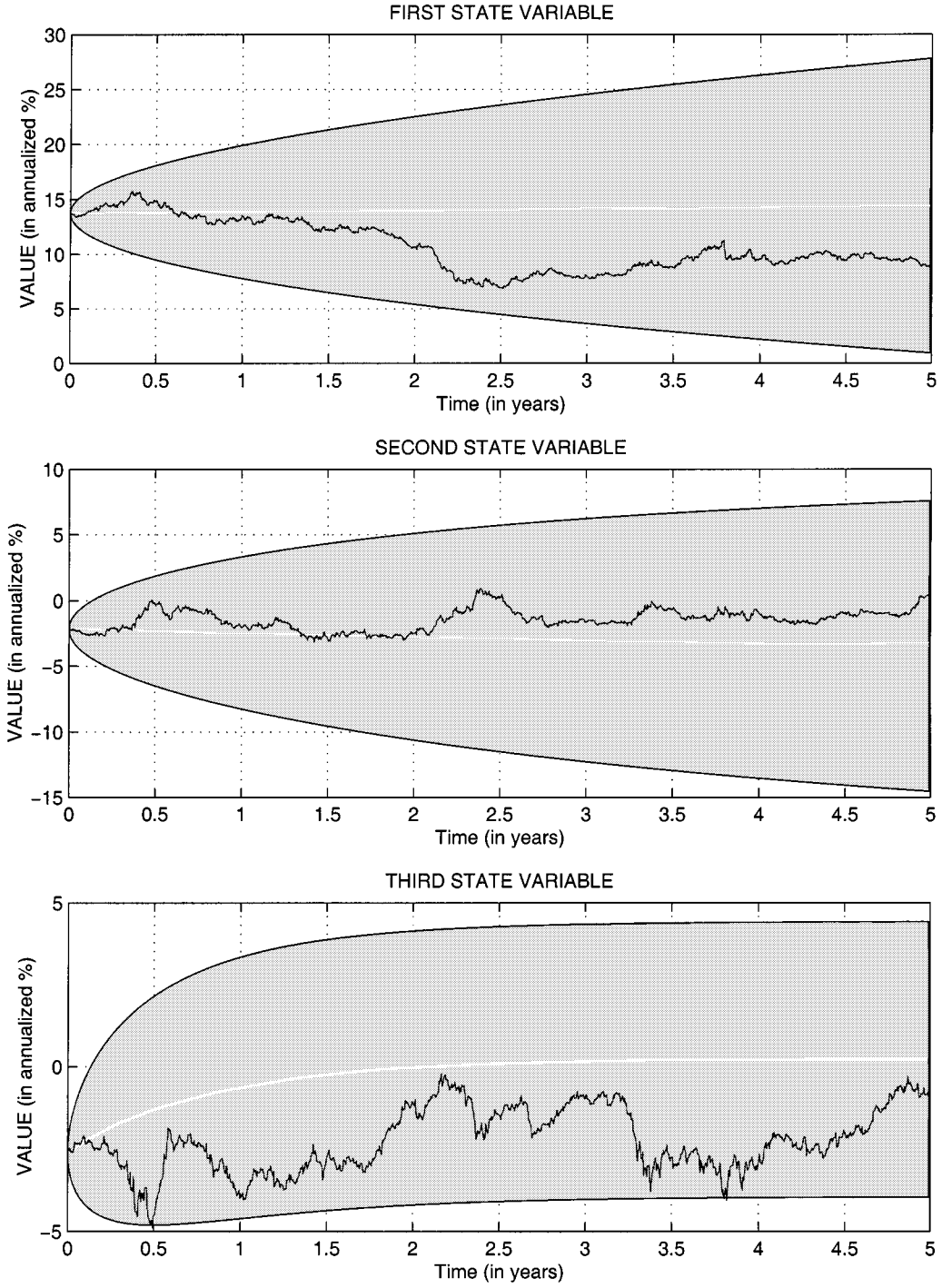


Fig. 6. Time evolution of the state variables in a 3-factor Vasicek model calibrated to USD yield data with consistency hints. The 'bubbles' show the 3σ range within which the evolution should take place.

fit error would result. However, as we see in Fig. 9, the impact is negligible in this case.

V. Statistical Interpretation

In this section, we put calibration in a statistical framework. This will provide a more principled way of making certain choices that would otherwise be made in a heuristic way. In particular,

1) it will provide a rationale for the relative weight between the fit error and the hint errors in the objective function;
2) it will enable us to bring other consistency hints, as well as a prior condition, into the picture;
3) it will provide a methodology for standardizing the different error measures, i.e., converting them to the same "units."

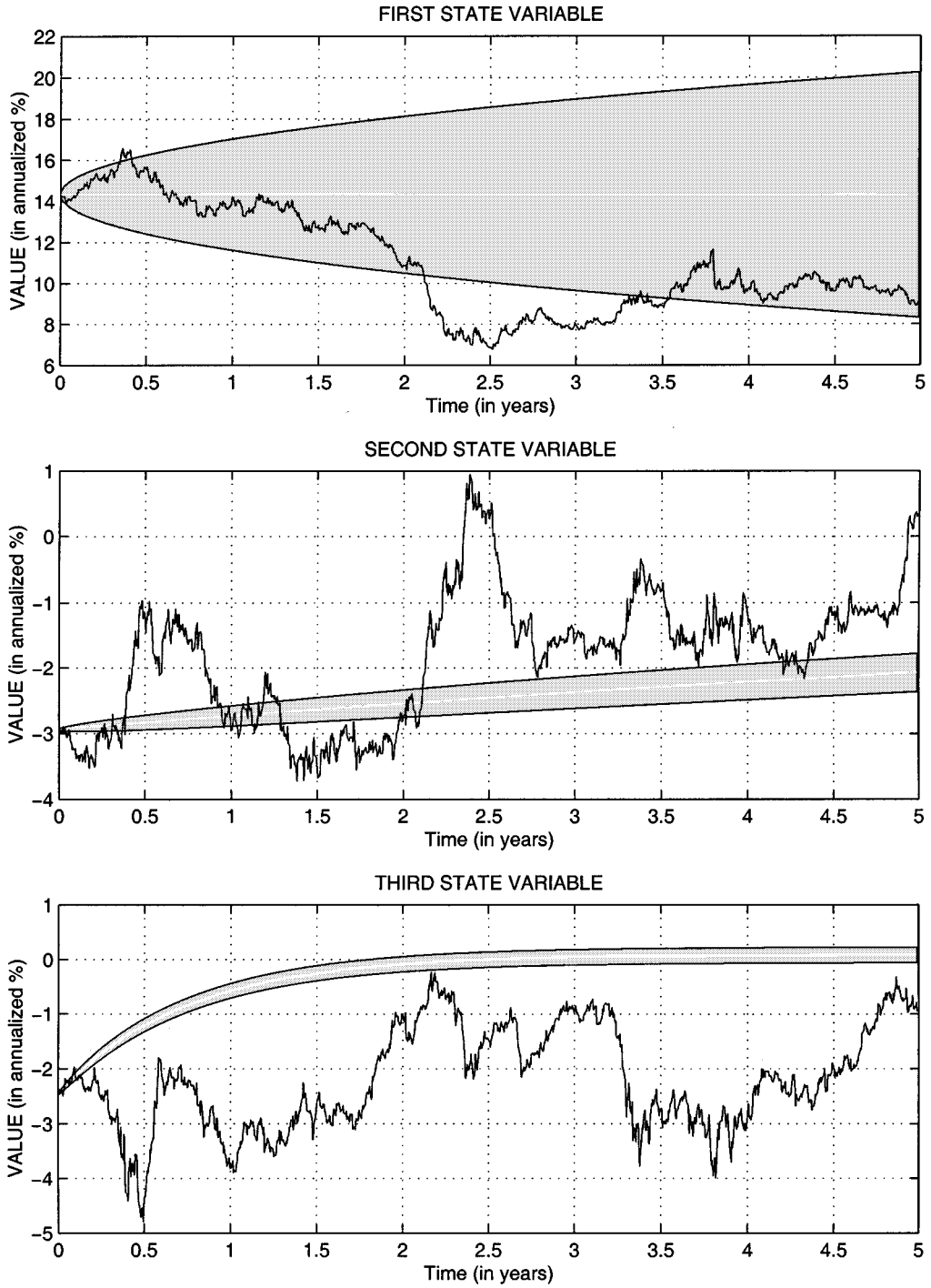


Fig. 7. Time evolution of the state variables in a 3-factor Vasicek model calibrated to USD yield data without consistency hints. The "bubbles" show the 3σ range within which the evolution should have taken place, but did not.

A. Probabilistic Setting

The premise of calibration is that the Vasicek model would be valid if the parameters (long-term $\mathbf{p}_{\mathcal{L}}$, and short-term $\mathbf{p}_{\mathcal{S}}$) were properly chosen. Validity of the model means that the pdf for generating $\mathbf{p}_{\mathcal{S}}$ has the form specified by the model, with $\mathbf{p}_{\mathcal{L}} = k_n, \theta_n, \sigma_n, \rho_{ij}$ determining the parameters of this pdf. The state variables $x_n[l]$; $n = 1, \ldots, N$, $l = 1, \ldots, L$, which are the short-term parameters $\mathbf{p}_{\mathcal{S}}$, are *generated* by the pdf. We obtain a simpler version of the pdf if we represent $\mathbf{p}_{\mathcal{S}}$ by the initial state $\mathbf{x}[1]$ and the stochastic elements $\mathbf{w}[l]$; $l = 1, \ldots, L-1$. This pdf[7] is given by

$$q(\mathbf{p}_{\mathcal{S}}) = q(\mathbf{x}[1], \{\mathbf{w}[l]\}) = \frac{1}{\sqrt{(2\pi)^N |S|}} e^{-(\mathbf{x}[1]-\Theta)^T S^{-1} (\mathbf{x}[1]-\Theta)/2}$$

[7]We use q to denote the joint pdf, and also to denote its marginal components as in Section III.

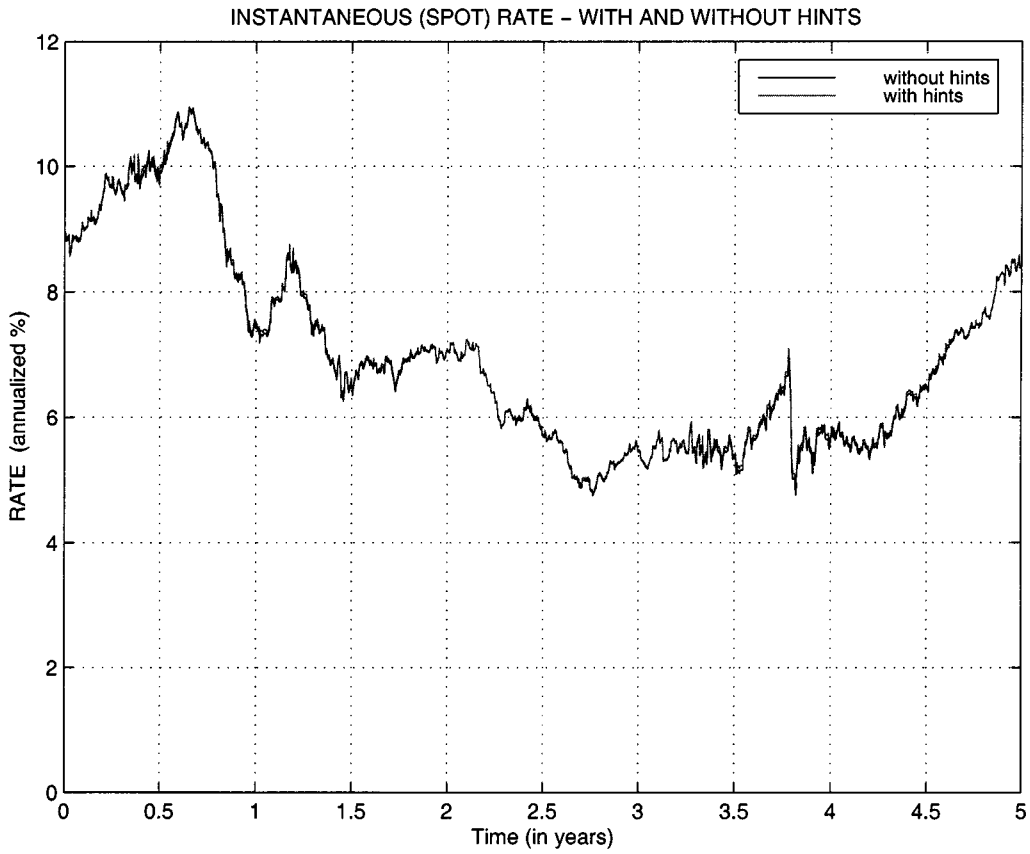


Fig. 8. In spite of the sharp contrast between Figs. 6 and 7, the instantaneous rates with or without consistency hints are virtually identical. The profound difference between the underlying models cannot be detected just by looking at these rates.

$$\times \prod_{l=1}^{L-1} \frac{1}{\sqrt{(2\pi)^N |Q|}} e^{-\mathbf{w}[l]^T Q^{-1} \mathbf{w}[l]/2}$$

where Q, S, and Θ are defined as in Section III.

Ideally, the correct values of the parameters would make every model function $f_m(\mathbf{p})$ identical to the market value $\tilde{f}_m$. In reality, however, the model will not perfectly match the data. Therefore, we must allow for some "noise" that separates $f_m(\mathbf{p})$ from $\tilde{f}_m$. We will view the data $\{\tilde{f}_m\}$ as well as the parameters $\mathbf{p}_\mathcal{L}$ and $\mathbf{p}_\mathcal{S}$ as random variables. Under this probabilistic scenario, some prior distribution generates $\mathbf{p}_\mathcal{L}$, which in turn specify the parameters of q, q generates $\mathbf{p}_\mathcal{S}$, then $\mathbf{p}_\mathcal{L}$ and $\mathbf{p}_\mathcal{S}$ determine $f_m(\mathbf{p})$, and $f_m(\mathbf{p})$ specify the parameters for generating $\tilde{f}_m$. The question becomes: Given the data, what is the *probability*[8] that the parameter values are correct? Applying Bayes rule, we get

$$\begin{aligned} \mathrm{P}(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S} | \{\tilde{f}_m\}) &= \frac{\mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})\, \mathrm{P}(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})}{\mathrm{P}(\{\tilde{f}_m\})} \\ &\propto \mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})\, \mathrm{P}(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S}) \\ &\quad \text{(fixed data } \{\tilde{f}_m\}) \\ &= \mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})\, \mathrm{P}(\mathbf{p}_\mathcal{S} | \mathbf{p}_\mathcal{L})\, \mathrm{P}(\mathbf{p}_\mathcal{L}). \end{aligned}$$

The most probable parameter values are the ones that maximize the product of these three probabilities. If we work with $-\log$(probability) instead of the probability itself, we will be minimizing the sum

$$\begin{aligned} &(-\log\, \mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})) \\ &\quad + (-\log\, \mathrm{P}(\mathbf{p}_\mathcal{S} | \mathbf{p}_\mathcal{L})) + (-\log\, \mathrm{P}(\mathbf{p}_\mathcal{L})). \end{aligned}$$

The three terms have a direct interpretation as

$$\begin{aligned} \text{fit error:} \quad & -\log\, \mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S}) \\ \text{consistency error:} \quad & -\log\, \mathrm{P}(\mathbf{p}_\mathcal{S} | \mathbf{p}_\mathcal{L}) \\ \text{prior error:} \quad & -\log\, \mathrm{P}(\mathbf{p}_\mathcal{L}). \end{aligned}$$

We will discuss these terms one at a time. The above sum provides the proper way of combining them once they are computed.

B. Fit Error

Given $\mathbf{p}_\mathcal{L}$, $\mathbf{p}_\mathcal{S}$. the model is fully specified. Therefore, we can calculate the functions $f_m(\mathbf{p})$ corresponding to the market data $\tilde{f}_m$. The fit probability $\mathrm{P}(\{\tilde{f}_m\} | \mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{S})$ would then penalize the "noise" that separates $\tilde{f}_m$ from the ideal $f_m(\mathbf{p})$. For example, if we assume that the noise is an additive zero-mean i.i.d. Gaussian, the fit error will be proportional to

$$\sum_{m=1}^{M} (f_m(\mathbf{p}) - \tilde{f}_m)^2$$

[8]or the probability *density*.

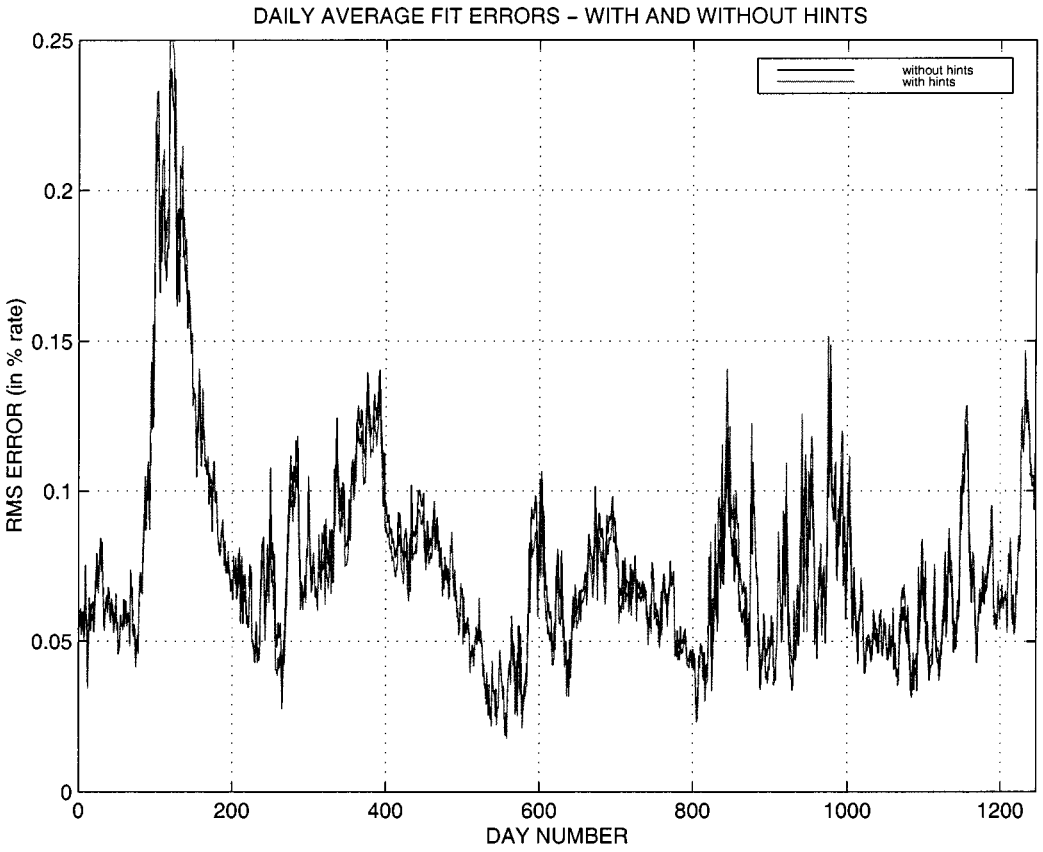


Fig. 9. The daily root mean square error in fitting the USD yield data, with and without consistency hints. With the hints constraining the fit, there is only a negligible increase in the fit error.

which is the expression for E_0 in Section II. The constant of proportionality is inversely related to the variance of the Gaussian. Thus, the relative weight between the fit error and the hint errors can be derived from assumptions about the noise level.

C. Consistency Error

The long-term parameters $\mathbf{p}_{\mathcal{L}}$ affect the consistency error directly by modifying S, Q, and Θ in the expression of q, and indirectly when we solve for the implied $\mathbf{w}[l]$ by substituting the state variables into the difference equations. The consistency error $-\log \mathrm{P}(\mathbf{p}_{\mathcal{S}}|\mathbf{p}_{\mathcal{L}})$ fixes $\mathbf{p}_{\mathcal{L}}$ in the expression of q, and evaluates $-\log(q(\mathbf{p}_{\mathcal{S}}))$. Substituting the expression of q, this reduces to the initial-state error E_2 plus the cross entropy part of E_1. Therefore, even without imposing hints *per se*, the Bayesian equation almost recreates the errors E_1 and E_2 of Section III.

Hints come into play because of overfitting. In order to optimize the objective function, we pursue many combinations of $\mathbf{p}_{\mathcal{L}}$ and $\mathbf{p}_{\mathcal{S}}$, based on a finite set of data. In doing so, we may introduce anomalies in the solution that would be very rare if we considered only one combination of the parameters. To avoid such anomalies, the search needs to be regularized or constrained. Hints provide constraints based on the properties of the model. As such, they do not exclude good solutions.

For instance, the entropy part of the hint error E_1 pulls $\mathbf{w}[l]$; $l = 1, \ldots, L-1$ away from the solution $\mathbf{w}[l] = \mathbf{0}$. This solution is the single most "probable" solution for $\mathbf{w}[l]$, since q assumes its maximum value there. The solution is nonetheless undesirable, since a typical solution for $\mathbf{w}[l]$ would have a variety of values that reflect the Gaussian distribution (the goodness of fit [10] seen in Fig. 5, but not in Fig. 4). If we generate a single solution, it is likely to be of the typical variety. However, if we actively seek a high-probability solution, we will get one, and it may be atypical. The contrast between "probable" and "typical" comes up in many contexts, most notably in information theory [4].

We will introduce other hint errors that also constrain the solution in a meaningful way. In deriving E_1, we made certain assumptions that we can exploit now to create the new hints. For instance, the Kullback-Leibler distance $K(p||q)$ should have been based on the full joint q, a situation we avoided because it would have rendered the entire $\mathbf{p}_{\mathcal{L}}$, $\mathbf{p}_{\mathcal{S}}$ a one-point sample, with no hope of creating a meaningful estimate. Working with the marginal q solved this problem, but left certain properties of the joint q untested. One such property is that $\mathbf{w}_l$ should be statistically independent for different l. We will create a correlation error function that penalizes statistical dependence. Also, the entropy part of E_1 was based on a Gaussian assumption about p, and we will create hint errors that penalize violations of this assumption. Finally, the entropy estimate was not sensitive to the mean of the distribution p, and we will create a bias error that penalizes p if it has a nonzero mean. Here are the details.

1) Bias: The form of q asserts that $\mathbf{w}[l]; l = 1, \ldots, L-1$, have zero mean. If so, $\mathbf{b}^T\mathbf{w}[l]$ must also have zero mean for any constant $\mathbf{b}$. Let[9]

$$\mu = \frac{1}{L-1}\sum_{l=1}^{L-1}\mathbf{w}[l]$$
$$\Sigma = \frac{1}{L-1}\sum_{l=1}^{L-1}(\mathbf{w}[l]-\mu)(\mathbf{w}[l]-\mu)^T.$$

Based on μ and Σ, we can define the bias error function

$$E_3 = \max_{\mathbf{b}}\frac{\mathbf{b}^T\mu}{\sqrt{\mathbf{b}^T\Sigma\mathbf{b}}}$$

which measures the normalized bias of $\mathbf{w}_l$ along the worst-case projection. The expression can be reduced to

$$E_3 = \sqrt{\mu^T\Sigma^{-1}\mu}$$

which is a simple function of the implied $\mathbf{w}[l]$.

2) Correlation: q asserts that $\mathbf{w}[l]$ is uncorrelated with $\mathbf{w}[l+1]$, among other things. If so, $\mathbf{b}^T\mathbf{w}[l]$ must also be uncorrelated with $\mathbf{b}^T\mathbf{w}[l+1]$ for any constant $\mathbf{b}$. Let

$$C = \frac{1}{L-2}\sum_{l=1}^{L-2}(\mathbf{w}[l]-\mu)(\mathbf{w}[l+1]-\mu)^T.$$

Based on C and Σ, we can define the correlation error function

$$E_4 = \max_{\mathbf{b}}\left|\frac{\mathbf{b}^TC\mathbf{b}}{\mathbf{b}^T\Sigma\mathbf{b}}\right|$$

which measures the normalized covariance, again along the worst-case projection. The expression can be reduced to the maximum absolute eigenvalue of $(A+A^T)/2$, where

$$A = D^{-(1/2)}U^TCUD^{-(1/2)}$$

with D and U being the eigenvalue matrix and eigenvector matrix of Σ ($U^TU = I$ and $U^T\Sigma U = D$).

3) Gaussianity: q asserts that $\mathbf{w}[l]$ are normally distributed. If so, the higher order moments around the mean should be related to the variance accordingly. For instance, the third moment that measures *skewness* should be zero, and the fourth moment that measures *kurtosis*[10] should be three times the square of the variance[4]. One can define error functions E_5 and E_6 based on deviations from these values.

Together with E_1 and E_2, the new error measures E_3, E_4, E_5, and E_6 capture many aspects of the pdf q. The list is by no means exhaustive. It is inevitable for a finite sample realization of a pdf to have anomalies along some dimension. What we have done here was to develop consistency hints that penalize a few obvious anomalies that may arise with overfitting.

[9]For an unbiased version of Σ, a normalizing factor of $1/(L-2)$ instead of $1/(L-1)$ would be used.

[10]Kurtosis quantifies "fatness of the tail," which is among the more vulnerable aspects of the Gaussian assumption in models like the Vasicek.

D. Prior Error

$\mathrm{P}(\mathbf{p}_\mathcal{L})$ assigns a prior probability to the long-term parameters $\mathbf{p}_\mathcal{L} = k_n, \theta_n, \sigma_n, \rho_{ij}$. There are reasons for preferring one set of parameters over the other in the absence of any data. Some of the reasons are the following.

1) Hard constraints arising from the model assumptions such as $k_n > 0$, $\sigma_n \geq 0$, and $[\rho_{ij}]$ being positive definite.
2) Economic considerations such as plausible values for the equilibrium interest rate $\sum_{n=1}^{N}\theta_n$.
3) Moving window calibration that allows long-term parameters to change slowly from one window to the next. In this case, the solution for $\mathbf{p}_\mathcal{L}$ in the old window becomes the center of a concentrated prior distribution for the new window.

E. Canonical Errors

The consistency error functions that we derived have different scales. Some are based on pdfs, others on measures such as entropy, and others on various heuristics. Even the premise of an error function can vary. For instance, the bias error could have been based on a fixed projection instead of the worst-case projection. Therefore, the values of these error functions, in the absolute, do not mean much. In order to combine the errors in a meaningful way, we would like to convert them to a uniform scale. This can be done using probability as a common ground.

Let $E(\mathbf{p})$ be an error function. We only require that $E(\mathbf{p})$ be truly an *error* function, i.e., one for which larger values of E correspond to worse values of $\mathbf{p}$. If $\mathbf{p}$ is stochastic, E becomes a random variable. In this case, we define the canonical version $\mathbf{E}$ of E as follows:

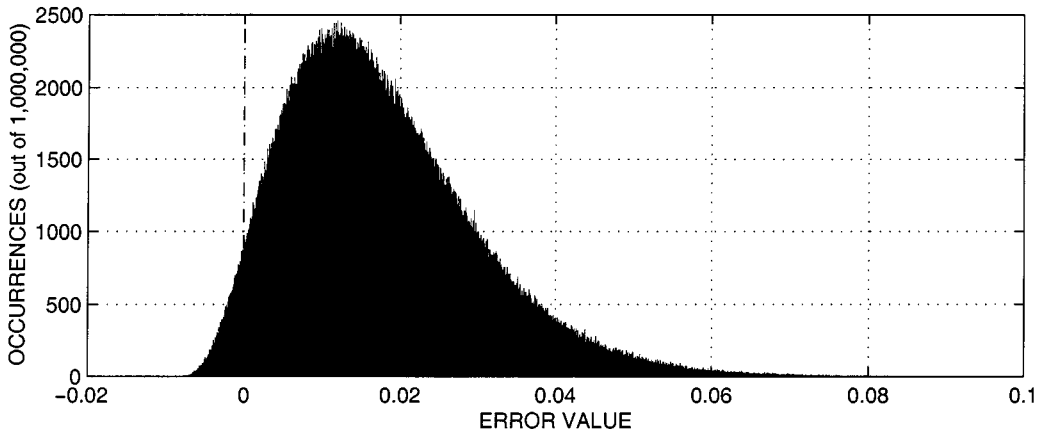
$$\mathbf{E}(\mathbf{p}) = -\log(\Pr(E \geq E(\mathbf{p}))).$$

In other words, the value of $\mathbf{E}$ for a given $\mathbf{p}$ is based on the total probability of all sets of parameters for which the value of E is no better than $E(\mathbf{p})$. One can view this as a natural grouping of the parameters induced by E.

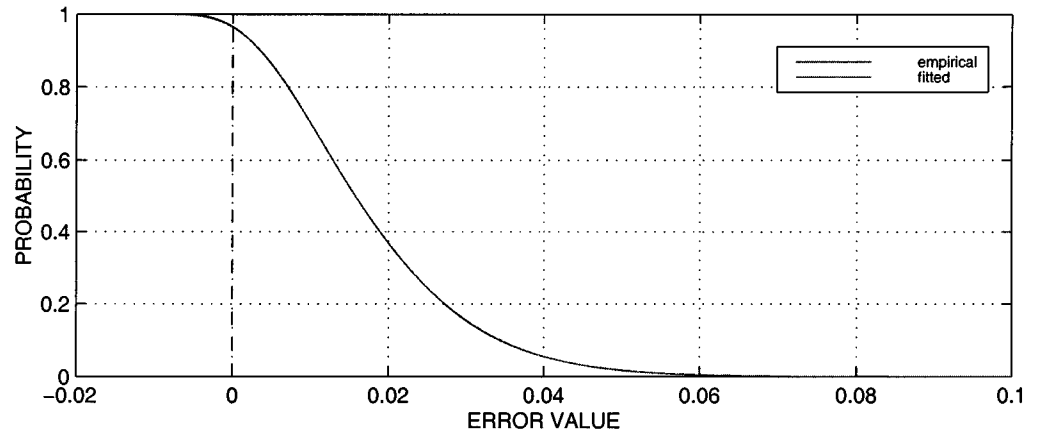
The definition implies that $\mathbf{E}(\mathbf{p})$ is actually $\mathbf{E}(E(\mathbf{p}))$. In some cases, it is possible to find an analytic formula for $\mathbf{E}(E)$. In other cases, $\mathbf{E}(E)$ can be evaluated based on numerical integration. If all else fails, it is possible to estimate $\mathbf{E}(E)$ using Monte Carlo simulations. To do this, generate the long-term parameters $\mathbf{p}_\mathcal{L}$ according to the prior (or fix them at a typical value), and generate $\mathbf{p}_\mathcal{S}$ according to q, then compute E and histogram it. $\mathbf{E}(E)$ can now be estimated from the histogram through curve fitting. The accuracy of the fit is more important for smaller values of E since the real tradeoff between different errors does not take place until they are relatively small. Fortunately, that's where more points fall in the histogram, allowing for a better fit.

In general, $\mathbf{E}$ will be different for different N (number of Vasicek factors), and will also vary with the calibration window size, sometimes in a predictable way. Fig. 10 illustrates the Monte Carlo procedure for the consistency error function E_1. We use the number of factors and the calibration window size of the JPY swaps experiment.

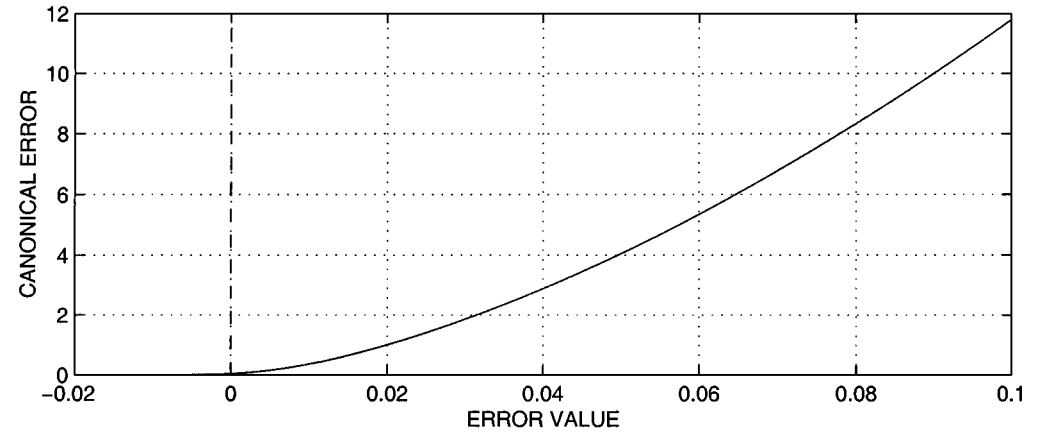
Regardless of the range of values for E, the canonical $\mathbf{E}$ will be greater than or equal to zero, with equality when E achieves



(a)



(b)



(c)

Fig. 10. Generating the canonical error version of E_1 for the 3-factor Vasicek. (a) Histogram of consistency hint error E1. A Monte Carlo simulation uses the model to generate a histogram of the values of E_1. (b) Implied probability that E1 exceeds a certain level. The histogram is used to infer the probability that E_1 exceeds a certain level, and an analytic formula is fit to that probability. (c) Canonical error function for E1. Taking $-\log$ of the formula, we get the value of the canonical error for any value of E_1.

its minimum possible value. The value of $\mathbf{E}$ has a uniform interpretation. For instance, $\mathbf{E} = 1$ always corresponds to a probability of e^{-1} or $\approx 0.37\%$.

If we have a number of statistically independent errors, their $\mathbf{E}$s can be combined by simple addition. Even with errors that are not quite statistically independent, our experience is that adding the canonical errors still works in practice.[11] This allows us to mix all types of error measures in the same objective function.

[11]Alternatively, one could define a joint version of $\mathbf{E}$ when the errors are not statistically independent. The Monte Carlo estimate in this case requires far more simulations.

VI. Conclusion

Calibration of financial models must conform to the assumptions of these models. If calibration is based only on fitting the data, it is liable to violate these assumptions. To guarantee that this does not happen, consistency hints are introduced as constraints on the calibration process. The Kullback-Leibler distance quantifies the main constraint. To balance the hint error functions, canonical errors are introduced. Consistency hints can be implemented with an efficient optimization algorithm. They are successfully applied to calibrating the correlated multifactor Vasicek model of interest rates in the JPY swaps market and the USD yield market.

Appendix I

In this Appendix, we provide the definitions and derivations of the correlated multi-factor Vasicek model. The reader may wish to get a more detailed account of interest-rate models [14], SDEs [15], and Ito calculus [13].

The Vasicek N-factor model for interest rates is given by the following set of SDEs:

$$dx_n = k_n(\theta_n - x_n)\,dt + \sigma_n\,dW_n \qquad n = 1, \ldots, N$$

where $k_n > 0$, $\sigma_n \geq 0$, and θ_n are constants, and $W_n(t)$ are Wiener processes whose covariances are given by

$$\mathcal{E}(dW_i dW_j) = \rho_{ij}\,dt \quad \text{where} \quad \rho_{ii} = 1 \qquad i, j = 1, \ldots, N.$$

The instantaneous interest rate r is given by

$$r(t) = \sum_{n=1}^{N} x_n(t).$$

A. The Discount Function

The discount function $D(t, t+T)$ computes the value, at the present time t, of "a future dollar" at time $t+T$

$$D(t, t+T) = \mathcal{E}(e^{-\int_t^{t+T} r(\tau)\,d\tau})$$

which can also be interpreted as the price of a unit bond of maturity T. The following expression solves for $D(t, t+T)$ under the Vasicek model

$$\begin{aligned} D(t,t+T) = \prod_{i=1}^{N} \exp\Bigg(& -\frac{x_i(t)}{k_i}(1 - e^{-k_i T}) \\ & - \theta_i \left(T + \frac{e^{-k_i T} - 1}{k_i} \right) \\ & + \sum_{j=1}^{N} \left(\frac{\sigma_i \sigma_j \rho_{ij}}{2 k_i k_j} \left(\frac{(1 - e^{-T(k_i + k_j)})}{k_i + k_j} \right.\right. \\ & \left.\left. - \frac{(1 - e^{-T k_i})}{k_i} - \frac{(1 - e^{-T k_j})}{k_j} + T \right) \right) \Bigg). \end{aligned}$$

To prove this, we use the fact that $r(\tau) = \sum_{n=1}^{N} x_n(\tau)$, and integrate the SDEs to obtain

$$\begin{aligned} r(\tau) = \sum_{n=1}^{N} \Big(& x_n(t) e^{-k_n(\tau - t)} + \theta_n (1 - e^{-k_n(\tau - t)}) \\ & + \sigma_n e^{-k_n \tau} \int_t^{\tau} e^{k_n s}\,dW_n(s) \Big) \end{aligned}$$

for $\tau \geq t$. Therefore

$$\begin{aligned} \int_t^{t+T} r(\tau)\,d\tau = \sum_{n=1}^{N} \Bigg(& \frac{x_n(t)}{k_n}(1 - e^{-k_n T}) \\ & + \theta_n \left(T + \frac{e^{-k_n T} - 1}{k_n} \right) \\ & - \frac{\sigma_n}{k_n} \int_t^{t+T} (e^{k_n(s - t - T)} - 1)\,dW_n(s) \Bigg) \end{aligned}$$

where the last term resulted from integration by parts. Let us call this last term β

$$\beta = \sum_{n=1}^{N} \frac{\sigma_n}{k_n} \int_t^{t+T} (e^{k_n(s - t - T)} - 1)\,dW_n(s)$$

β is a zero-mean Gaussian with variance

$$\begin{aligned} \text{var}(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{N} & \frac{\sigma_i \sigma_j \rho_{ij}}{k_i k_j} \int_t^{t+T} \\ & \cdot (e^{k_i(s - t - T)} - 1)(e^{k_j(s - t - T)} - 1)\,ds \end{aligned}$$

since $E(dW_i(s)\,dW_j(s)) = \rho_{ij}\,ds$, and $E(dW_i(s_1)\,dW_j(s_2)) = 0$ for $s_1 \neq s_2$ by the properties of the Wiener processes. In terms of β, since $D(t, t+T) = \mathcal{E}(\exp(-\int_t^{t+T} r(\tau)\,d\tau))$, we can write

$$\begin{aligned} D(t,t+T) = \Bigg(\prod_{n=1}^{N} \exp\Bigg(& -\frac{x_n(t)}{k_n}(1 - e^{-k_n T}) \\ & - \theta_n \left(T + \frac{e^{-k_n T} - 1}{k_n} \right) \Bigg) \Bigg) \mathcal{E}(e^{-\beta}) \end{aligned}$$

but $\mathcal{E}(e^{-\beta}) = e^{1/2 \text{var}(\beta)}$ for a zero-mean Gaussian β[4]. Substituting

$$\begin{aligned} D(t,t+T) = \prod_{i=1}^{N} \exp\Bigg(& -\frac{x_i(t)}{k_i}(1 - e^{-k_i T}) \\ & - \theta_i \left(T + \frac{e^{-k_i T} - 1}{k_i} \right) \\ & + \sum_{j=1}^{N} \left(\frac{\sigma_i \sigma_j \rho_{ij}}{2 k_i k_j} \int_t^{t+T} (e^{k_i(s - t - T)} - 1) \right. \\ & \left. \cdot (e^{k_j(s - t - T)} - 1)\,ds \right) \Bigg). \end{aligned}$$

Carrying out the integration results in the required expression.

B. Other Market Functions

Many model-based market functions follow from the discount function. For example, the yield function $Y(t,t+T)$ estimates the interest rate between times t and $t+T$, expected at time t

$$Y(t,t+T)=\frac{-\log D(t,t+T)}{T}.$$

The forward rate function $F(t,t+T)$ is the instantaneous rate at time $t+T$ expected at time t

$$F(t,t+T)=\frac{-\partial}{\partial T}\log D(t,t+T).$$

The swap par rate is the fixed interest rate that can be evenly exchanged for a floating rate. It assumes that we are receiving at times $t+\Delta t, t+2\Delta t,\ldots,t+T$ the return on one dollar invested Δt earlier at the prevailing interest rate at the time of investment. In return, we must pay out at the same times $t+\Delta t$, $t+2\Delta t,\ldots,t+T$ constant payments of $R\Delta t$ each, which can be thought of as simple interest on one dollar invested Δt earlier at rate R. The par rate is the value of R that would make these two cash flows equitable. It is denoted by $R(t,t+T,\Delta t)$, and is given by

$$R(t,t+T,\Delta t)=\frac{1-D(t,t+T)}{\Delta t\sum_{i=1}^{T/\Delta t}D(t,t+i\Delta t)}.$$

For all of these functions, we can obtain a Vasicek formula by substituting the formula for $D(t,t+T)$. For instance

$$F(t,t+T)=\sum_{i=1}^{N}(x_i(t)e^{-k_iT}-\theta_i(1-e^{-k_iT}))+\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\sigma_i\sigma_j\rho_{ij}}{2k_ik_j}(e^{-Tk_i}-1)(e^{-Tk_j}-1)\right).$$

The only state variables appearing in these formulas are the "current states," i.e., the N state variables $x_n(t)$ at the present time t (the time when the quantities are measured). This fact simplifies the logistics of fitting market functions to market data.

The final market function used in this paper is the volatility term structure (VTS) of the forward rate. Given the Vasicek formula for $F(t,t+T)$, if we hold T constant, we can write

$$dF=\sum_{n=1}^{N}e^{-k_nT}\,dx_n.$$

Substituting from the Vasicek SDEs, the stochastic part of dF is given by $\sum_{n=1}^{N}e^{-k_nT}\sigma_n\,dW_n$. Therefore, the variance of dF is given by

$$\mathrm{var}(dF)=\sum_{i=1}^{N}\sum_{j=1}^{N}e^{-(k_i+k_j)T}\rho_{ij}\sigma_i\sigma_j\,dt.$$

The VTS is defined by $V(t,t+T)=\sqrt{\mathrm{var}(dF)/dt}$. Therefore,

$$V(t,t+T)=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}e^{-(k_i+k_j)T}\rho_{ij}\sigma_i\sigma_j}$$

which is constant with regard to t and does not depend on state variables. In Fig. 2, the theoretical VTS was computed by this formula, while the historical VTS was based on the sample standard deviation of changes in $F(t,t+T)$ from day to day.

C. Discrete-Time Approximation

To derive a discrete-time version of the Vasicek model, we consider one step in time from t to $t+\Delta t$, and integrate the SDEs to get

$$x_n(t+\Delta t)=x_n(t)e^{-k_n\Delta t}+\theta_n(1-e^{-k_n\Delta t})+\sigma_ne^{-k_n(t+\Delta t)}\int_t^{t+\Delta t}e^{k_ns}\,dW_n(s).$$

Rewriting $x_n(t+\Delta t)-x_n(t)$ as Δx_n and rearranging, we get

$$\Delta x_n=x_n(t)(e^{-k_n\Delta t}-1)+\theta_n(1-e^{-k_n\Delta t})+\sigma_n\int_t^{t+\Delta t}e^{k_n(s-t-\Delta t)}\,dW_n(s)$$

which can be rewritten as

$$\Delta x_n=\left(\frac{1-e^{-k_n\Delta t}}{\Delta t}\right)(\theta_n-x_n(t))\Delta t+\sigma_nw_n(t)\sqrt{\Delta t}.$$

The last term follows from the properties of Wiener processes, with $w_n(t)$ being jointly Gaussian with zero mean and a covariance given by

$$\mathcal{E}(w_i(t)w_j(t))=\left(\frac{1-e^{-(k_i+k_j)\Delta t}}{(k_i+k_j)\Delta t}\right)\rho_{ij}\qquad i,j=1,\ldots,N.$$

Furthermore, $w_n(t)$ are independent for different times with nonoverlapping Δt. This expression for Δx_n is the exact difference equation for the Vasicek model. If $k_n\Delta t\ll 1$, we can approximate it by a difference equation similar to the SDE.

$$\Delta x_n=k_n(\theta_n-x_n)\Delta t+\sigma_nw_n\sqrt{\Delta t}.$$

where $x_n=x_n(t)$, $w_n=w_n(t)$, and $\mathcal{E}(w_iw_j)=\rho_{ij}$. When discrete time is used, we adopt the usual notation of bracketed index arguments. Thus, time will be denoted by $t[l]$, and the corresponding x_n and w_n will be $x_n[l]$ and $w_n[l]$.

ACKNOWLEDGMENT

The author would like to acknowledge Dr. M. Magdon-Ismail for his useful hints and to thank the members of Caltech's Learning Systems Group for helpful discussions.

REFERENCES

[1] Y. Abu-Mostafa, "Learning from hints in neural networks," *J. Complexity*, vol. 5, pp. 192–198, June 1990.

[2] ——, "Machines that learn from hints," *Sci. Amer.*, vol. 272, no. 4, pp. 64–69, Apr. 1995.
[3] ——, "Hints," *Neural Comput.*, vol. 7, pp. 639–671, July 1995.
[4] T. Cover and J. Thomas, *Elements of Information Theory*: Wiley, 1991.
[5] A. Dempster *et al.*, "Maximum likelihood for incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B39, pp. 1–38, Jan. 1977.
[6] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1968, vol. 1.
[7] R. Fletcher and C. Reeves, "Function minimization by conjugate gradients," *Comput. J.*, vol. 7, pp. 149–154.
[8] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[9] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
[10] R. Larsen and M. Marx, *An Introduction to Mathematical Statistics and Its Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[11] R. Lupton, *Statistics in Theory and Practice*. Princeton, NJ: Princeton Univ. Press, 1993.
[12] J. Moody, "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems," in *Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, and R. Lippmann, Eds. San Mateo, CA: Morgan Kaufmann, 1992, vol. 4, pp. 847–854.
[13] S. Neftci, *An Introduction to the Mathematics of Financial Derivatives*. New York: Academic, 1996.
[14] R. Rebonato, *Interest-Rate Option Models*. New York: Wiley, 1996.
[15] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*. New York: Wiley, 1980.
[16] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. 26, pp. 26–37, Jan. 1980.
[17] B. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman and Hall, 1993.
[18] O. Vasicek, "An equilibrium characterization of term structure," *J. Financial Economics*, vol. 5, pp. 177–188, Nov. 1977.
[19] P. Wilmott *et al.*, *The Mathematics of Financial Derivatives*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

Yaser S. Abu-Mostafa received the B.Sc. degree from Cairo University, Cairo, Egypt, in 1979, the M.S.E.E. degree from the Georgia Institute of Technology, Atlanta, in 1981, and the Ph.D. degree from California Institute of Technology (Caltech), Pasadena, in 1983.

He is Professor of Electrical Engineering and Computer Science, and Head of the Learning Systems Group, at the California Institute of Technology. His research interests include machine learning, computational finance, and neural networks, and he has more than 80 technical publications including two articles in *Scientific American*.

Dr. Abu-Mostafa was a founding member of the IEEE Neural Networks Council, and the founding program chairman of the Neural Information Processing Systems (NIPS) Conference. He chaired a number of national and international conferences, most recently the international conference on *Computational Finance* (CF'99), and has served on the boards of ten journals. Since 1988, he has been a technical consultant for a number of financial firms, including *Citibank* for nine years. Among his awards are the 1996 Richard P. Feynman Prize for excellence in teaching, and the 1998 Kuwait State Award in Applied Science. He was awarded the Clauser Prize for the most original doctoral thesis.