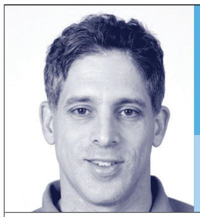

Kent Osband

FINFORMATICS

How to Measure Really Small Things

The orthodoxy has tendency to ignore drift which leaves opportunity for finformaticians the market over…

Traders in financial assets implicitly compare the trading price to the stream of dividends the assets stand to generate. Clearly, a key determinant of value—usually *the* key determinant—will be the long-term drift or rate of growth or rather, it's clear to everyone except orthodox finance theorists.You see, orthodox finance has been mesmerized by Black-Scholes—where the drift, exactly offset by risk-aversion, drops out of equilibrium option values—into believing that the drift never matters. But this assumes that everyone knows what the drift is. In reality, that hardly ever applies. It can't, both because the drift is too shrouded in noise to measure exactly and because the drift tends to change over time. In consequence, people with s uperior knowledge of the drift stand to profit from betting against the market consensus.

In previous *Finformatics* we worked the optimal way to update probability estimates of drift, assuming a Markov switching process between various constant values. Given a multitude of regimes i with drift μ_I and volatility σ, an objective or subjective likelihood p_i of being in regime i, instantaneous switching probabilities λ_{ij} for migrating from regime i to and other regime j, $\lambda_{ii} \equiv -\sum_{i \neq j} \lambda_{ij}$ defined as the negative of the instantaneous probability of switching out of regime i, and an observation $dx = \mu_? dt + \sigma dz$ for standardized Brownian noise dz and unknown drift $\mu_?$, the optimal Learning Equation is again:

$$dp_i = p_i \left(\frac{\mu_i - E}{\sigma} \right) dW + \langle \lambda_{ji} \rangle dt$$

where $\langle \cdot \rangle \equiv \sum_j \cdot p_j$ denotes the expectations operator over index j, $E \equiv \langle \mu_j \rangle$ denotes the expected drift, and $dW \equiv \dfrac{dx - Edt}{\sigma}$ is the best estimate of standardized Brownian motion given expectations.

As we recall, the Learning Equation is best interpreted as the sum of belief revision and expected regime-switching. The belief revision is the product of three effects: the current conviction p_i, the idiosyncrasy $\dfrac{\mu_i - E}{\sigma}$ of the current belief measured relative to observation noise, and the news or surprise value of the standardized residual dW. The expected regime-switching is just the difference between expected movements into regime i and movements out of regime i.

Obviously this is much more comprehensive than the orthodox single-state framework. But at first glance it's not nearly comprehensive enough. How do we incorporate drifts that may depend on other variables including time? And how do we tractably handle updating a possible continuum of regimes? That's the focus of this article.

Caveats

Let me pause to acknowledge that the Learning Equation applies only to really small things. By "really small" I mean infinitely divisible without any discrete jumps. We know it can't apply to bigger things, not always, because if dW is big enough relative to the idiosyncrasy and has the wrong sign, the corresponding p_i will drop below zero.

Interestingly, daily observations aren't always small enough. Consider for example a discrete approximation to the Learning Equation that posits various regimes with yearly Sharpe ratios ranging from -2 to $+2$ and updates using daily data. The daily Sharpe ratios μ/σ of the various regimes will range from approximately $-1/8$ to $+1/8$, for a maximal idiosyncrasy of 1/4, It follows that daily outliers over 4 standard deviations could generate negative probabilities, absent modification. I won't deal here with the modifications, simply assume that we're measuring finely enough with clean enough data not to worry about such things.

Another caveat is that we're assuming perfect measurement of dx, without noise or dullness in the measurement stick itself. Otherwise we couldn't measure finely enough to be absolutely sure about observation noise σ. In practice we are indeed not absolutely sure about σ ; we can't be because things like tick size bounds and bid/ask spreads get in the way. I'm not going to address that here either. We need to learn to walk before we fly. Besides, it's extremely important to appreciate that most of our problems reflect the problem of estimating drift on short intervals no matter how good our measuring stick is.

Complex Regimes

With these caveats, let's return to the original Markov-switching framework but relax our identification of each regime i with a fixed μ_i. Instead, let's allow each μ_i to be a function f_i of a vector Y of other variables. One of those other variables can be time, which allows for deterministic changes in drift. The variables can also include economic fundamentals like GDP growth or profit margins. There's no limit other than that, properly speaking, Y should not include the beliefs p about which regime applies.

Now if we review the statement of the Learning Equation, we see that holding beliefs and regime-switching probabilities fixed, none of the updates depend on anything other than the drift in a given regime and the expected drift over all regimes. Whether $\mu_i \equiv f_i(Y_i)$ denotes a very complex regime or simply a fixed state is totally irrelevant, provided we know its current value. That irrelevance is borne out in the derivation as well. So the answer about what needs to change in the Learning Equation is very simple: nothing!

That conclusion can be discomfiting. Consider for example two regimes that currently have the same drift. Or even betters, consider two regimes that always have the same drift and differ only in relevant switching propensities? How will the Learning Equation distinguish them? If can't, other than thru the slow changes of expected regime-switching. On reflection, however, we can't expect observation to distinguish regimes when the regimes produce identical observations. And the Learning Equation is always first and foremost geared to seeking truth from observation.

One potential application comes in weighing two different models of reality. Suppose that we're sure that only one of models A or B applies. Classical statistics would force us to derive potentially super-complex likelihood functions for each of the two models, take their ratio, and discard the worse one when the ratio gets high enough. The Learning Equation will simply update the probability p_A that model A is correct according to

$$dp_A = p_A(1-p_A)\left(\frac{\mu_A - \mu_B}{\sigma}\right)dW$$

Note that this applies regardless of the structure of model A or model B. What could be simpler? It's so simple that if we have non-normal or lumpy observations it's tempting to reformulate the decision problem so as to apply this updating rule as an approximation.

Distilling a Continuum

The main practical worry in applying the Learning Equation is that, even though each particular update is simple, the number of regimes that need updating may be prohibitively high. In principle, we may even need a continuum of regimes, and no one can count a continuum.

Fortunately, one kind of belief continuum turns out to be extremely easy to update. That's where beliefs stay normal. Consider first the special case where all switching probabilities λ are zero. If the initial beliefs are normal and a normal observation is recorded, it is well known that the updated beliefs are normal too. Using the word "precision" to describe the inverse variance, the new mean can be calculated as the precision-weighted average of the prior mean and the observation, while the precision of the new beliefs equals the sum of the precisions of prior and observation. Hence, if the prior mean and variance are M and V respectively, then the new mean drift and variance after an observation dx/dt with variance σ^2/dt will be

$$\begin{aligned}
E_{new} &= \frac{1/V\,E + dt/\sigma^2 \cdot dx/dt}{1/V + dt/\sigma^2} = \frac{E + Vdx/\sigma^2}{1 + Vdt/\sigma^2} \\
&\cong (E + Vdx/\sigma^2)(1 - Vdt/\sigma^2) \\
&\cong E + \frac{V}{\sigma}\left(\frac{dx - Edt}{\sigma}\right) = E + \frac{V}{\sigma}dW \\
V_{new} &= \frac{1}{1/V + dt/\sigma^2} = \frac{V\sigma^2}{Vdt + \sigma^2} = \frac{V}{1 + Vdt/\sigma^2} \\
&\cong V(1 - Vdt/\sigma^2) = V - \frac{V^2}{\sigma^2}dt
\end{aligned}$$

We can derive the same results more elegantly using the cumulant form of the Learning Equation, which we developed in previous *Finformatics*. Unfortunately, the derivation given there contains a mistake, so let me quickly redo the relevant parts. Given a characteristic function (Poisson transform) $\varphi(b) \equiv \langle e^{ib\mu} \rangle$ of a distribution, the cumulants are the coefficients C_n in the Taylor series expansion below of $\xi(b)$, the logarithm of the characteristic function:

$$\xi(b) \equiv \ln \varphi(b) = \sum_{n=0}^{\infty} \frac{(ib)^n}{n!} C_n$$

Since the characteristic function has differential

$$d\varphi(b) = \int e^{ib\mu} dp(\mu) = \int \langle \lambda_{\nu\mu} \rangle_\nu e^{ib\mu} d\mu + \langle e^{ib\mu}(\mu - E) \rangle \frac{dW}{\sigma}$$

its logarithm by Ito's rule has differential of:

$$\begin{aligned}
d\xi(b) &= \frac{d\varphi(b)}{\varphi(b)} - \frac{1}{2}\frac{\langle e^{ib\mu}(\mu - E)\rangle^2}{\varphi^2(b)\sigma^2}dt \\
&= \left[\frac{\int \langle \lambda_{\nu\mu}\rangle_\nu e^{ib\mu} d\mu}{\langle e^{ib\mu}\rangle} - \frac{1}{2}\left(\frac{\langle e^{ib\mu}(\mu - E)\rangle}{\langle e^{ib\mu}\rangle \sigma}\right)^2\right] dt + \frac{\langle e^{ib\mu}(\mu - E)\rangle}{\langle e^{ib\mu}\rangle}\frac{dW}{\sigma} \\
&\equiv \left[\frac{\int \langle \lambda_{\nu\mu}\rangle_\nu e^{ib\mu} d\mu}{\langle e^{ib\mu}\rangle} - \frac{Q^2(b)}{2\sigma^2}\right] dt + Q(b)\frac{dW}{\sigma}
\end{aligned}$$

Since $\dfrac{d\langle e^{ib\mu}\rangle}{db} = \langle i\mu e^{ib\mu}\rangle$, we can rewrite Q as

$$\begin{aligned}
Q(b) &\equiv \frac{\langle e^{ib\mu}(\mu - E)\rangle}{\langle e^{ib\mu}\rangle} = -i\frac{d\ln\langle e^{ib\mu}\rangle}{db} - E \\
&= -i\xi'(b) - E = \sum_{n=0}^{\infty}\frac{(ib)^n}{n!}C_{n+1} - C_1 = \sum_{n=1}^{\infty}\frac{(ib)^n}{n!}C_{n+1}
\end{aligned}$$

By equating each term in the Taylor expansion of $d\xi(b)$ we see that the volatility of C_n equals C_{n+1}/σ for all n, which in most cases implies a never-ending nontrivial chain. But with a normal distribution of beliefs the chain

>

simplifies enormously, because all normal cumulants of order 3 or higher vanish. In that case $Q = ibV$, and if all the λ's vanish too the updating is completely described by:

$$
\begin{aligned}
dC_1 &\equiv dE = \frac{V}{\sigma} dW \\
dC_2 &\equiv dV \equiv -\frac{V^2}{\sigma^2} dt \\
dC_3 &= dC_4 = \cdots = 0
\end{aligned}
$$

In other words, the distribution stays normal with the mean and variance updates above. Granted, this demonstration is more circuitous than a normal distribution requires, but it helps explain why normal distributions are so special. No other distribution of beliefs lends itself to such simple updating using the Learning Equation.

Non-zero λ could easily destroy normality. But suppose regime transitions are specially designed so as to preserve normality. In that case, $\frac{\int \langle \lambda_{\nu\mu} \rangle_\nu e^{ib\mu} d\mu}{\langle e^{ib\mu} \rangle}$ must be quadratic in b, implying

$$
\begin{aligned}
dE &= Fdt + \frac{V}{\sigma} dW \\
dV &\equiv \left(G^2 - \frac{V^2}{\sigma^2} \right) dt
\end{aligned}
$$

for some scalars F and G^2. That is the essence of the Kalman filter, which is the world's best-known model of regime-switching.

If desired we could allow F and G^2 to vary over time or with other variables Y. However, the case of constant G^2 and zero F is particularly interesting because it generates a filter even more basic than Kalman. Note that dV will be positive for $V < G\sigma$ and negative for $V > G\sigma$. Hence updating is driven toward an equilibrium in which $V = G\sigma$ and

$$
dE = GdW = \frac{G}{\sigma}(dx - Edt) = \frac{V}{\sigma^2}(dx - Edt)
$$

That's just a simple exponentially weighted average (EWA) with updating coefficient V/σ^2. Usually we simply interpret that updating coefficient as the inverse of the effective duration, and the EWA itself as just a cheap recursive substitute for simple averaging. But in our current framework we can regard the EWA as an optimal estimate under the conditions specified above, for V an equilibrium level of uncertainty.

Mixtures of Normal Beliefs

Now that we've seen how to model normal beliefs simply, let's try to model mixtures of normal beliefs. Let $k = 1,..,K$ denote various normal components, with current means $E_1,..,E_K$, constant conditional variances $V_1,..,V_K$, and mixing probabilities $\theta_1,..,\theta_K$. Let us assume in addition that Markov transitions occur between components without blemishing the normality of either component, and let λ_{jk} denote the transition from component j to component k. Let $p_k(\mu)$ denotes the conditional density at μ given component k, and $p(\mu)$ the unconditional density, so that $p(\mu) = \sum_{k=1}^{K} p_k(\mu)$. The Learning Equation tells us to calculate updates as

$$
dp_k(\mu) = p_k(\mu) \left(\frac{\mu - E}{\sigma} \right) dW
$$

from which it follows that

$$
\begin{aligned}
dE_k &= \frac{\int_k \mu dp_k(\mu) d\mu}{\int_k p_k(\mu) d\mu} = \frac{\int_k \mu(\mu - E) p_k(\mu) d\mu}{\sigma \theta_k} dW = \frac{V_k}{\sigma} dW \\
d\theta_k &= \int_k dp_k(\mu) d\mu = \frac{dW}{\sigma} \int_k (\mu - E) p_k(\mu) d\mu \\
&= \frac{dW}{\sigma} \int_k (\mu - E_k + E_k - E) p_k(\mu) d\mu \\
&= \frac{dW}{\sigma} \int_k (\mu - E_k) p_k(\mu) d\mu + \frac{dW}{\sigma} \int_k (E_k - E) p_k(\mu) d\mu \\
&= 0 + \frac{dW}{\sigma}(E_k - E)\theta_k = \theta_k \frac{E_k - E}{\sigma} dW
\end{aligned}
$$

In other words, we can update the mixed normal distribution of beliefs by separately updating the conditional means of each component and the mixing weights. To confirm that the aggregate mean moves as required, note that

$$
\begin{aligned}
dE &= d\left(\sum_k \theta_k E_k \right) = \sum_k \theta_k dE_k + \sum_k E_k \, d\theta_k \\
&= \sum_k \theta_k \frac{V_k}{\sigma} dW + \sum_k E_k \theta_k \frac{E_k - E}{\sigma} dW \\
&= \frac{dW}{\sigma} (\langle V_k \rangle + Var(E_k)) = \frac{Var_{agg}}{\sigma} dW
\end{aligned}
$$

Again, it is quite easy for the regime transitions to muddy the normality of each component. To avoid this, let us simply assume that probability mass gets transferred only from one mixing component to another, without altering the conditional weights within any given component. Then if λ_{jk} denotes the instantaneous probability of shifting from component j to component k and λ_{kk} the negative of the probability of shirting out of k, then the updating equation for conditional means changes to

$$
dE_k = \langle \lambda_{kj} \rangle dt + \frac{V_k}{\sigma} dW
$$

while the updating equation for $d\theta_k$ remains the same.

Thus we have a very tractable way to update mixed normal distributions of beliefs, provided we restrict the transition matrix which is empirically difficult to identify anyway. Combine that with flexible interpretation of the various μ_i and their determinants and we obtain a tractable system for estimating virtually any diffusion. Basically it throws a bunch of measuring sticks at the diffusion, updates the mean estimate for each, and also updates the confidence in each measuring stick.

Moreover, while each update depends on every other, the requirements for information-sharing are incredibly small. A central coordinator should compute the current consensus mean E and the deviation dW of the latest observation from the consensus. The rest of the updating can be done locally, without any more detail from other regimes on estimation results or methodology, It's hard to imagine a more ridiculously simple updating system, unless it's simply ridiculous.