Version of November 24, 2017

# Risk Neutral Densities: A Review

Stephen Figlewski*

* Professor of Finance
New York University Stern School of Business
44 West 4th Street
New York, NY 10012-1126

email: sfiglews@stern.nyu.edu
tel: 212-998-0712

I thank the NASDAQ Educational Foundation and the Instituto Tecnológico Autónomo de México for generous support during this project. I am grateful to the many people who have shared their thoughts and expertise on these issues with me over the years.

ABSTRACT

# Risk Neutral Densities: A Review

Trading in options with a wide range of exercise prices and a single maturity allows a researcher to extract the market's risk neutral probability density (RND) over the underlying price at expiration. The RND contains investors' beliefs about the true probabilities blended with their risk preferences, both of which are of great interest to academics and practitioners alike. With particular focus U.S. equity options, this article reviews the historical development of this powerful concept, practical details of fitting an RND to option market prices, and the many ways in which investigators have tried to distill true expectations and risk premia from observed RNDs. I touch on areas of active current research including the "pricing kernel puzzle" and the "volatility surface," and offer thoughts on what has been learned about RNDs so far and fruitful directions for future research.

Keywords: Risk neutral densities, option risk premia, implied volatility, option pricing

Black and Scholes' (1972, 1973) (BS) option pricing model introduced the idea that important pricing information, specifically the future volatility of the underlying stock, could be extracted from the market price of an option by inverting the valuation formula.  Implied volatility (IV) today plays a hugely important role in both academic finance and derivatives trading.  Researchers soon realized that not only volatility, but an entire probability distribution for the future stock price could be extracted from the options market.  Because the pricing equation is derived from an arbitrage relation between the option and its underlying stock in which all risk has been hedged away, the model and any parameters implied out from it are not affected by investors' risk preferences.  The implied returns distribution is therefore called the Risk Neutral Density, abbreviated hereafter as the RND.

In this paper, I review the major ideas and theoretical developments in this interesting and important field.  Complete coverage or anything like it would require far more space than is available.  Over the years, there have been a number of excellent reviews that together provide broad coverage of the existing research.  To limit the territory, I will restrict the discussion almost entirely to the U.S. equities market, which leaves out research on risk neutral densities for interest rates, currencies, and commodities, as well as equity markets in other countries.  This is meant to be a review of the key ideas, not the literature, although it will cover many specific articles.  I therefore apologize to the authors whose excellent work I do not mention explicitly here, as well as, especially, to any authors whose excellent work I may have misunderstood and misrepresented, hopefully few.

Risk neutral pricing was first discussed by Cox and Ross (1976).  Rubinstein (1976) and Brennan (1979) explored the joint conditions on a returns process and a "representative agent's" utility function such that risk neutral pricing obtains in discrete-time even without the support of an arbitrage trade. They proved that lognormal returns and constant relative risk aversion utility are necessary and sufficient conditions for option values to satisfy the BS equation; Normal returns and constant absolute risk aversion are necessary and sufficient to produce the so-called Normal model.[1] Subsequently, Stapleton and Subrahmanyam (1984) extended the proofs to the multivariate case.

These ideas were explored by a number of researchers in the 1970s, but the seminal articles most cited today are Breeden and Litzenberger (1978), who showed how to obtain the full RND from option prices, and Harrison and Kreps (1979), who proved that in a market with no profitable arbitrage trades, it is always possible to combine investors' objective probability estimates over future states of the world with their utility-based marginal valuations of $1 in each of those states. Risk neutral investors facing this "risk neutral" density would value all derivatives exactly as they are priced in a risk averse real world market. Moreover, this RND is unique if the derivatives market is complete.

Much research has focused on extracting, understanding, and making use of the potentially valuable information on investor expectations and risk preferences contained in an RND, with volatility attracting by far the most attention. The main challenge is to separate beliefs about the "true" or "empirical" or "statistical" distribution that reflects investors' best judgment about probabilities over future returns--commonly known as the "P" distribution--from the risk neutralized RND, or "Q" distribution, which modifies the P-distribution to incorporate risk preferences. The Q-distribution is what can be extracted from market option prices. To find

[1] Constant relative risk aversion is a feature of the power utility function, $U(W) = W^{1-a}/(1-a)$, where a is the coefficient of relative risk aversion, and the limit as a goes to 0 is log utility, $U = \log(W)$. Constant absolute risk aversion is a feature of the exponential utility function, $U(W) = e^{-aW}$.

the equivalent P-distribution, further assumptions are required on either the returns process or the representative agent's utility function.

The P- and Q-distributions are related to each other through the pricing kernel, which embeds the representative agent's risk preferences. Let $S_T$ denote the state of the world at date T (often assumed to be proportional to the level of the stock market portfolio) and $k(S_T)$ be the value in date 0 dollars of $1 to be received in future state $S_T$. For any asset V whose payoff at T is some function $H(S_T)$ of the state at that date, the time 0 value $V_0$ is given by

$$V_0 = e^{-rT} \int_{S_T} H(S_T) g(S_T)\, dS_T = e^{-rT} \int_{S_T} H(S_T) k(S_T) f(S_T)\, dS_T \qquad (1)$$

where g( . ) is the risk neutral Q-density and f( . ) is the empirical P-density. Letting V be an Arrow-Debreu state claim that pays $H(S_T)$ = $1 in state $S_T$ and nothing otherwise, we get the pointwise relationship between the P- and Q-densities as

$$g(S_T) = k(S_T) f(S_T), \quad \mathit{for\ all\ states\ } S_T \mathit{\ in\ the\ domain}^{2} \mathit{\ of\ } g \mathit{\ and\ } f \qquad (2)$$

If V is a T-period call option with strike price X, its date T payoff is $C(S_T)$ = Max (0, $S_T$ – X), and (1) becomes

$$C_0 = e^{-rT} \int_X^{\infty} (S_T - X) g(S_T)\, dS_T \qquad (3)$$

in which the integral is over the portion of the RND above X where the call has a positive payoff.

Taking the partial derivative with respect to X and rearranging gives

$$1 + e^{rT} \frac{\partial C}{\partial X}\bigg|_{S_T} = 1 - \int_X^{\infty} g(S_T)\, dS_T = G(S_T) \qquad (4)$$

[2] Harrison and Kreps' proof of equivalence between the P- and Q-densities requires that they share the same support, that is, that there are no states with zero probability under one density that have positive probability under the other.

where *G(S_T)* is the cumulative risk neutral probability distribution at $S_T$. Taking the partial derivative with respect to X a second time gives the RND as the forward value of the second partial derivative of the call value with respect to the strike price.[3]

$$e^{rT}\frac{\partial^2 C}{\partial X^2}\bigg|_{X=S_T} = g(S_T) \tag{5}$$

This is the key equation that allows extraction of the RND from a panel of option prices. In practice, the partial derivative in (5) is replaced with a discrete approximation using prices of options with a single maturity spanning a wide (and dense) range of strikes. This is readily accomplished for broad index options like those on the S&P 500 index, but it can be hard to apply for thinly traded options with only a small number of available strikes. The alternative to the nonparametric approach is to constrain the fitted RND to take a particular form, such as lognormal or some more flexible density. The next section reviews the important principle that the value of a derivative is connected to its underlying by arbitrage. In theory, this means options prices contain no information beyond what is in the spot price. But with limits to arbitrage real world RNDs can reflect both market expectations about the true returns distribution (the P-density) and also risk premia.

Section 3 briefly reviews the historical development of this idea, as option pricing models evolved from Black and Scholes (1973) to the current collection of multifactor jump-diffusion specifications. Section 4 covers practical details of extracting RNDs from option market prices and describes many different approaches that have been proposed.

RNDs contain information about true probabilities and about risk premia. Section 5 reviews attempts to deduce information about the true returns process, i.e., the P-density, and Section 6 looks at what has been learned about risk premia from the Q-density. But it is a well-

[3] It is easy to show that the same equation (5) also holds if H( ) refers to a put option.

known "anomaly" that empirical pricing kernels are badly behaved. If the S&P 500 index can be taken as a reasonable proxy for the whole U.S. wealth portfolio, higher stock prices correspond to greater wealth and a lower margin utility of $1. But RNDs extracted from index options produce pricing kernels with upward sloping portions in the middle, implying that risk aversion drops as wealth increases in the most likely range of the distribution of returns, strongly contradicting standard utility theory. Section 7 reviews hypotheses that attempt to explain the "pricing kernel puzzle," and the empirical evidence that contradicts most of them.

Section 8 then looks more closely at what is probably the most important application of RNDs: calculation and prediction of the VIX volatility index. The VIX appears to have become a systematic volatility factor in the options market, playing a similar role to that of the S&P index in the CAPM. Section 9 suggests two broad avenues for future research and concludes.

## 2. The Underlying Economics of Risk Neutral Valuation

Equation (1) applies to all derivatives. Before getting into estimation and interpretation of option RNDs, it is useful to think about a much simpler derivative, a forward or futures contract (we will treat them as being equivalent in this discussion). This will help clarify the intuition behind risk neutral valuation, how risk premia enter derivatives prices, and the conceptual difficulty with assuming a representative investor in modeling derivatives trading.

A forward contract locks in a price F at which an underlying asset will be purchased on future date T. The classical pricing model begins by envisioning a population of "speculators". Individual speculators may have different expectations over $S_T$, but the market as a whole is rational, so the mean expectation should be the objective (P-distribution) expected value of $S_T$. With only speculators in the market, the futures price at date $t < T$ should satisfy

$$F_t \ = \ E[\ S_T \mid \Phi_t\ ] \qquad (6)$$

where $\Phi_t$ is the market's information set at date t.

Hedgers constitute a separate class of market participants, and they are the reason for the futures market to exist. Hedgers do not forecast $S_T$, but simply take futures positions opposite to their exposure in the spot market. Farmers and others who worry about falling prices hedge by selling futures, which drives F below the objective expected value of $S_T$ into a pattern of "normal backwardation." The lower futures price causes some speculators to switch sides and buy the now-underpriced contracts, thus taking on the risk the hedgers are unloading. The price discount creates an expected profit to the speculators, while the hedgers have expected losses, which they consider the cost of insurance. Backwardation in the futures market is thus a risk premium.[4]

This classical futures pricing model illustrates two important points that also affect options. First, the risk premium in futures arises from the interplay between speculators (information traders) and hedgers. We cannot combine the two groups into a single "representative investor" without losing the whole purpose of the market, which is to transfer risk from one group to the other. A market made up of identical investors can agree on a fair price for a given derivative contract, but there would never be any trading, because identical investors won't take opposite sides of a zero-sum contract like a future or an option. Although the assumption is nearly universal in theoretical modeling, a representative agent model is not able to capture what actually happens in a real world derivatives market.

Second, the direction and size of the risk premium in a particular market will reflect factors related to the risk transfer, including price volatility, contract maturity, risk aversion of

[4] If hedging is mainly done by consumers and producers who worry about rising prices, they will go long in futures, which bids the futures price up above E[ $S_T \mid \Phi_t$ ]. The market will be in "contango," and the risk premium will be earned by speculators on the short side of the futures market.

hedgers and speculators (buyers and writers in options markets), the number of contracts traded, and the relative numbers and wealth of those who want to take long and short positions. Variations in any of these factors should cause market risk premia to differ in different markets and to fluctuate over time within a single market. For instance, variations in hedging activity in commodities with a seasonal pattern in supply (e.g., wheat) or demand (e.g., heating oil) should create time-varying risk premia that embed that pattern. The key point is that risk premia in derivatives markets can be expected to be dynamic and to reflect the economics of the underlying spot market.

This classical theory of futures pricing, and the extension to options that I am suggesting, tacitly assume the underlying asset cannot be stored and carried over time, which is not true for foreign exchange, most financial instruments and many physical commodities. If storage is possible, an alternative way to lock in an $S_T$ value is simply to buy the underlying today and hold it until date T. To prevent profitable arbitrage, the return to buying spot, hedging with futures, and carrying the riskless position must equal the riskless interest rate, which leads to the well-known "cost of carry" pricing relationship:[5]

$$F = e^{rT} S_0 \tag{7}$$

Notice that the expected value of the future $S_T$ is missing and discounting is at the riskless rate: there is no risk premium. In a world of risk neutral investors, the expected future spot price is just today's spot price future-valued at the riskless rate, so (7) holds in a risk neutral world whether storage is possible or not. When a derivative's payoff can be replicated exactly by a portfolio of instruments available today (the underlying plus riskless borrowing, in this case), arbitrage forces the equilibrium futures price to be the same under risk aversion as in a world of

[5] For simplicity, we are abstracting from dividends and other cash payouts and assuming there are no physical or financial carrying costs on the underlying other than interest.

risk neutral investors.  This is the principle of risk neutral valuation:  If a riskless arbitrage can be set up with the derivative, it should be priced with no risk premium.  A corollary is that under risk neutral valuation, a derivative's price provides no information about either expected returns or risk premia on unhedged investments beyond what is in the spot price.

Option payoffs are nonlinear and much harder to replicate than futures.  Black and Scholes' conceptual breakthrough was to derive an arbitrage strategy that is riskless in a frictionless market and allows the option's payoff to be replicated by delta-hedging between the stock and cash.  Like equation (7), neither risk aversion nor expected stock returns enter the BS equation.  If the arbitrage were possible in practice, even highly risk averse investors would not require a risk premium in options prices (relative to the stock price).

Theoretical valuation models leave out many real-world impediments to arbitrage, including transactions costs, margin and financing requirements, and uncertainty over volatility and other model parameters. The continuous rebalancing needed in delta-hedging is especially expensive in terms of transactions costs and cannot be made riskless in practice.  When models derived for frictionless markets are "taken to the data" they often uncover significant discrepancies: real world limits to arbitrage allow mispricing (in terms of the model) to arise and to persist over time.

A contingent claim is a redundant security in an arbitrage-based theoretical model because all risk can be hedged away.  How closely market prices follow a given model will depend on how forcefully arbitrageurs pursue the trade.  If arbitrage is very easy, as it is in foreign exchange forwards where the underlying is just money, mispricing should be rare, small, and transitory. The harder and riskier the arbitrage trade is in practice, the larger and more persistent mispricing can be and the more unhedgeable risk potential arbitrageurs must bear.

Options markets fall somewhere between the pure arbitrage-free world of risk neutral valuation and a market where limits to arbitrage are so severe that no linkage at all exists connecting spot prices to those in the future and derivatives are priced like other nonredundant securities. It is limits to arbitrage that, contrary to theory, allow an RND extracted from options prices to reveal expectations about future returns (the P-distribution) and premia for risks that cannot be hedged away (the Q-distribution and the pricing kernel). Moreover, we should expect any substantial change in the cost or risk of the arbitrage trade to be reflected in option prices.

This line of reasoning was pursued in Canina and Figlewski (1993) (CF), which explored an important options market where delta-hedging was essentially impossible. During the sample period (1983-87), the S&P 100 (OEX) index was the most actively traded options contract in the world, but continuous rebalancing of a portfolio containing 100 different stocks was too onerous to be done. It was widely believed at the time that implied volatility was the best available prediction of future realized volatility because it represented the expectations of well-informed investors in an efficient market. But CF's statistical analysis showed the exact opposite: OEX options appeared to contain no information at all about future realized volatility—the limits to arbitrage in that market at that time were too large.[6]

Figlewski and Webb (1993) studied the informational effect of a different market constraint: short sales restrictions on the underlying stocks. Buying puts or writing calls can substitute for a short sale of the stock. Options allow investors who cannot or prefer not to short the stock to accomplish the same thing indirectly: they trade options with marketmakers, who

[6] The OEX was an important but very special case. The results in Canina and Figlewski (1993) confirmed that limits to arbitrage can be a crucial factor in how information enters option prices, but it always represented a polar case, not a general property of all options. The paper's main contribution to research on RNDs was to demonstrate that IV was not necessarily the best possible volatility forecast and to stimulate numerous subsequent studies of the question using different markets, time periods, and statistical techniques. The nearly universal conclusion from this research has been that, in general, implied volatility does contain useful information about future realized volatility, but it is not an informationally efficient forecast.

then hedge by shorting the stock.  Figlewski and Webb found that short interest in the stock increased when options were introduced, and subsequently the difference between put and call option IVs was positively related to the amount of short selling of the stock.  Options were not redundant instruments: The ability to trade puts and write calls helped to complete the market by reducing the impact of constraints on short sales of the stock, and the effect of the trading that that permitted was then reflected in options prices in the market.

In similar fashion, marketmakers who follow delta-hedging or other risk management strategies typically have to bear some unhedgeable risk, at least at first, so their price quotes can be expected to reflect their risk tolerance, trading capital, and inventory positions.   High convexity makes hedging harder and riskier, and Figlewski and Freund (1994) found that options with high gammas were priced higher relative to Black-Scholes than less convex contracts.  Cao and Han (2013) found a higher premium on idiosyncratic variance risk than on systematic risk and attributed the difference to greater difficulty in delta-hedging for marketmakers.  Bollen and Whaley (2004) found a strong empirical connection between public order flow and option implied volatilities.  IVs for stock index puts reflected daily variations in buying pressure, while IVs of calls on individual stocks varied significantly with call order flow.  In both cases, the initial impact dissipated over time as marketmakers gradually adjusted their risk exposure.  In the most direct evidence so far, Gârleanu, Pedersen, and Poteshman (2009) analyzed actual trading positions of options marketmakers and showed a close link between the unhedgeable risk in their options inventory and their option price quotes.  Moreover, prices of similar options are also affected, in proportion to their covariance with the unhedged risks.

## 3.  The Evolution of Risk Neutral Density Modeling

The CBOE launched exchange-traded options in 1973, the same year the BS model appeared in the Journal of Political Economy.  It was a happy conjunction of events for academic researchers, who had an exciting new theoretical model to test and a new market generating price data to test it on.  Expected future volatility was needed, but an implied value could be extracted from the market.    Unfortunately, empirical tests, starting with Black and Scholes themselves (1972), showed that volatility did not enter option market prices in the way it was modeled.

Options on the same stock with different strikes consistently exhibit quite different IVs, which violates the model.  (Yet, to this day, IVs are almost always computed using the BS equation.)   Out of the money (OTM) options had higher IVs than at the money (ATM) contracts with the same maturity, the well-known "smile" pattern.  Black (1976) offered one possible explanation: falling stock prices cause the underlying firm to become more leveraged, which in turn should increase the stock's volatility.  This explanation became known as the "leverage effect" and the strong empirical regularity that equity volatilities, both implied and realized, go up when stock prices fall has kept this title, even for cases where a leverage argument makes no sense, e.g. options on all-equity firms, commodities, and interest rates.  As Bakshi, Kapadia and Madan (2003) note, the leverage argument applies to individual stocks and implies that smiles for index options should be less skewed on average than those for the component stocks.  But this is strongly contradicted by the data. (Toft and Prucyk (1997), Dennis and Mayhew (2002)). Dennis and Mayhew found empirically that more leveraged firms actually had flatter volatility skews than average.

Attempts to explain the smile tend to fall into three (non-exclusive) classes.  The first is that volatility is not constant. The evidence pointed to stochastically time-varying variance, which causes Gaussian shocks to generate fat-tailed non-Gaussian RNDs.  Fat tails make OTM options worth more than in the BS model (using ATM IV) because there are more large returns than a lognormal distribution allows for. [7] But the early empirical options research was seriously limited by lack of data and especially by lack of computer power. The kinds of models with multiple stochastic factors that are routinely estimated today on vast data sets were impossible in the 1970s and 80s; the earliest work was generally limited to a handful of options and only a few years of price data.[8] Bhattacharya (1983), discussed below, was a notable exception. Research in the 1970s and early 1980s tried to flatten out the smile by using different returns processes, while still limiting risk to a single stochastic factor, the diffusion term dz.

Cox and Ross (1976) suggested the Constant Elasticity of Variance (CEV) model as one of several alternatives to the fixed volatility lognormal. The CEV changes the stochastic term from $\sigma S\ dz$ in the BS lognormal diffusion to $\sigma S^{\gamma}\ dz$. Volatility becomes stochastic, increasing or decreasing with S, depending on the elasticity parameter γ. However, empirical testing (Macbeth and Merville (1980), Emanuel and Macbeth (1982)) showed that while fitting an additional parameter necessarily improved in-sample performance, the fitted elasticities were quite different for different stocks and varied sharply over time.

Another way to get fatter tails is to modify a lognormal RND directly. This cuts the connection between the instantaneous dynamics (needed for hedging) and the density as of option expiration, but it allows a better match to empirical RNDs. Jarrow and Rudd (1982) proposed using a generalized Edgeworth expansion, a technique similar to Taylor series, that approximates a given density in terms of another more tractable one by matching its moments.

[7] Time variation in volatility is not enough to generate a smile pattern unless it is also stochastic. In a frictionless market, volatility that changes over time but whose instantaneous value at each future date is known at the outset still allows risk to be totally eliminated by delta-hedging. The BS IV in that case is the square root of the (known) integrated future variance from the present through option expiration.

[8] For example, Macbeth and Merville (1980) used one year of daily data on six stocks. In updating that study, Emanuel and MacBeth (1982) were pleased to be able to add a second year of data for the same six stocks.

The resulting option pricing formula is in the form of the BS equation plus three adjustment terms for the differences in variance, skewness and kurtosis between the distribution of the data and a lognormal with the same mean.

With no explicit connection between the RND and a returns process that produces it, the RND shape is simply fitted to a mathematically tractable approximating curve. Then risk neutral valuation is invoked to obtain an option pricing formula, despite the lack of any riskless arbitrage trade. Attempts to obtain better-performing RNDs by modifying the lognormal directly include Corrado and Su (1996), who suggested using a Gram-Charlier approximation and Madan and Milne (1994) who tried Hermite polynomials.

The Binomial model, developed into a valuation methodology by Cox, Ross and Rubinstein (1979), features fixed volatility at every node. Later, Rubinstein (1985, 1994), and others) explored extended Binomial structures to accommodate volatility that varies through the tree. Stochastic volatility was obtained in the Binomial, like in the CEV model, without adding another stochastic variable. The most successful model in which this is possible is Generalized Autoregressive Conditional Heteroskedasticity (GARCH), developed first by Engle (1982) and extended by Bollerslev (1987).

GARCH is set in discrete time, which unfortunately precludes eliminating risk by continuous hedging, although GARCH is readily applied to real world price data that are also only observed at discrete intervals. GARCH features a returns equation in which i.i.d. shocks are multiplied by a time-varying volatility parameter. A second volatility equation specifies period t+1 variance as a weighted combination of period t variance and period t's squared return shock. Thus the same shock drives both returns and volatility. Although each shock is Gaussian, a series of large ones that would bring an OTM option into the money will also produce higher return volatility, and a higher theoretical option price than with the current volatility. GARCH can generate volatility smiles, but there is a practical problem. Out of sample volatility forecasts under GARCH rapidly converge toward the long-run mean value, so GARCH option pricing models are generally unable to produce smiles as steep as are exhibited by longer maturity options (Jackwerth (2000)).

The CEV, the Binomial and GARCH all allow volatility to change stochastically, without requiring a second stochastic factor. But the empirical evidence strongly indicated that a single stochastic factor was not adequate to capture real world option pricing. Stochastic volatility (SV) models began to be developed in the late 1980s by Hull and White (1987), Scott(1987), and Wiggins(1987). These articles were limited in that variance was stochastic, but uncorrelated with returns. Heston (1992) is the now-classic article in which a closed-form option pricing model is derived for an asset whose returns feature stochastic variance driven by a second, correlated, mean-reverting diffusion. Heston's derivation made use of a powerful technique, recently introduced to option pricing by Stein and Stein (1991), that allowed closed-form valuation equations to be obtained for a very much wider range of returns processes. After Heston, Fourier transforms, and a variety of related transform methods have become essential technology in modern derivatives modeling (see, e.g., Duffie, Pan, and Singleton (2000)).

In SV models, volatility shocks are correlated with shocks to the returns process, but imperfectly, so volatility becomes a second (and unhedgeable) source of risk. Correlation between return shocks and variance shocks is typically estimated to be around -0.7 for the broad U.S. stock market, but much lower for individual stock options. SV models produce more realistic-looking smiles than models with deterministic or less-flexible volatility processes, but they have trouble matching high market prices for deep OTM puts close to maturity, because the probability a diffusion could lead to a large enough price change to bring them into the money becomes infinitesimal. A model that allows prices for the underlying asset to make non-diffusive jumps seems to be required.

Allowing jumps produces more realistic fatter-tailed returns distributions even with constant diffusive volatility and jump parameters. Cox and Ross (1976) discussed returns processes with jumps, but Merton (1976) is credited with developing option pricing within a jump diffusion framework. He explicitly did not assume risk neutral pricing, and hence was not able to derive a general option pricing model independent of preferences. However, if jump risk is diversifiable, it should not be priced in equilibrium, in which case the BS equation, with suitably adjusted volatility, still holds. Hull and White (1987) develops this result further.

Bates (1996) tried to find a returns process with enough tail risk to explain market pricing for FX options on the Deutschemark. He found strong evidence for the existence of jumps. But jumps were not enough unless there was also a substantial jump risk premium. This article, although not on stock index options, is noteworthy because it was one of the first papers to focus on the need for risk premia in options prices even in the context of a model with a realistic jump-diffusion returns process.

Das and Sundaram (1999) considered whether models needed stochastic (diffusive) volatility or jumps to produce the observed volatility skew. They showed that near to maturity put prices exhibit so much implied tail risk that their steep smiles could only be explained by (downward) price jumps. But jump effects average out over longer horizons, so jumps cannot explain volatility smiles for longer maturity puts. They concluded that both stochastic diffusive volatility and stochastic jumps are required to explain the volatility smiles at both short and longer maturities. Bates (2000) looked at S&P 500 index options in the aftermath of the 1987 market crash, when the "smile" turned into a downward sloping "skew" for index options, and also concluded that both stochastic diffusive volatility and jumps must be included to capture the volatility process.

Stochastic volatility with jumps (SVJ) models are now standard, and apparently necessary to capture real world option pricing. But volatility itself is not observable, nor is it easy to disentangle jumps from diffusive price changes over discrete observation intervals, so putting jumps into an SV option model adds additional unobservable and unhedgeable latent factors. One must specify intensity--how often jumps occur--and also a distribution for the stochastic jump size. Down-jumps (market crashes) appear to have different characteristics from up-jumps, so two different jump processes may be allowed, with four new parameters to be fitted. Another issue is whether the occurrence of a jump should affect the returns equation, the volatility equation, or both. Is it plausible that there can be a jump in one without affecting the other? If both returns and volatility need to jump, how many more parameters are required to fully capture the relationship? And since jumps can't be perfectly hedged, investors might well require a premium, which must be modeled and estimated, on each new source of risk that affects return or volatility.

The development of highly structured models accelerated after 2000, with major advances in data availability, computer power, and econometric methodology. An important development was the integration of model estimation by fitting both the time series of stock returns and the cross section of option prices together. Pan (2002) is a prominent example. Previously, the normal approach was to use the returns time series to estimate the returns process and then to compute option prices from that process and compare them to market option prices.

Other econometricians began to exploit the availability of high frequency intraday data in computing realized volatility. A major problem in fitting SVJ models to daily data was that volatility had to be estimated using a reasonable number of past days, but volatility was moving stochastically within the sample. When it became possible to compute returns intraday over intervals as short as 5 minutes, one could reasonably assume that the volatility within the day was a constant, and also, that any large price change in such a short interval could be reliably classified as a jump. Andersen Fusari and Todorov (2015) and Bollerslev and Todorov (2011) are good examples of the value of computing realized volatility with high frequency data.

The introduction of the VIX volatility index, enhanced by the new nonparametric construction methodology (Chicago Board Options Exchange (2003)) provided a new index of overall S&P 500 volatility and added another dimension to the observable information about volatility expectations (and volatility risk premia). Futures on the VIX began trading in 2004 and VIX options started in 2006. Numerous exchange-traded funds (ETFs) tied to the returns on the S&P and to the VIX have also been launched. The existence of so many interconnected new markets has provided a wealth of new data, which at the time of this writing, is still largely unexplored in academic research. Up to this time, research on the behavior of risk neutral densities was largely an academic exercise, although there was some interest in the research departments of some central banks. With the explosion of trading activity surrounding the VIX, there has been a comparable expansion of research to understand and to predict the "volatility surface." The (Black-Scholes) IVs extracted from the prices of traded options are plotted in strike price – moneyness space, and the discrete IVs are connected, with appropriate smoothing, into a surface in three dimensions. Since success in modeling the movement in the volatility surface may translate into profitable option trading strategies, this area has attracted a great deal of research among practitioners. We will discuss some of this in Section 8.

## 4. Extracting a Risk Neutral Density

This section will review numerous methods that have been used to extract RNDs from options prices. One must recognize that an RND combines the effects of many disparate factors, including investors' beliefs about the true returns process, the market's aggregation of their preferences over future states of the world, the strength of trading constraints and limits to arbitrage, and rational and irrational beliefs about future prices. There is no reason to expect the resulting RND to be a member of any particular family of known distributions, or to remain constant over time. Although it is common to fit a known density to an RND, as we will now discuss, it should be kept in mind that this remains essentially a curve fitting exercise. Indeed, Orosi (2015) treats it as such and extracts plausible RNDs with no underlying theory beyond constraints that they satisfy the static no-arbitrage conditions.

We begin with a look at the nature of the input data.

### 4.1 Data issues

The starting point is a set of market prices for options written on a single underlying. A separate RND exists for each option maturity. Connecting RNDs across expirations creates an RND surface in three dimensions: exercise price, maturity and probability. This section only considers RNDs for a single maturity.

Most empirical work on U.S. options has used prices from either the Berkeley Option Data Base (BODB), available for the years 1976 through 1995, or OptionMetrics thereafter. The BODB contains every quote or trade registered for every option traded at the CBOE throughout the trading day, reported to the nearest second.  Intraday data allow careful treatment of bid-ask spreads, timing issues, and other frictions affecting actual trades, but because of the huge volume, academic research with BODB mostly used closing prices.  Bhattacharya (1983) did use the intraday data to check how consistently the options market obeyed both static no-arbitrage relationships like convexity and the BS model.  Violations were much more common for the BS model, but mispricing was nearly always smaller than transactions cost bounds.

The BODB ended when the CBOE stopped providing data in 1995.  OptionMetrics then became the primary data source for stock options research, with extensive coverage of all listed equity options (not just those on CBOE) beginning in January 1996.  OptionMetrics data, however, only covers closing prices.

An RND is a snapshot of the options market across the spectrum of strikes at a single point in time, so it is important to use option bid and ask quotes, that are continuously available. Recorded trade prices are too sparse across strikes and too spread out over the trading day, even for extremely liquid contracts like those on the S&P 500 index.  The possible objection that marketmakers' quotes may not reflect investors' expectations and risk preferences is unfounded, for the most part.  Marketmakers must post firm bids and offers on every option—spreads may be wide but they are real prices at which trades of reasonable size can be executed.  Quotes are not stale even for highly illiquid contracts, but are continually being updated as the stock price fluctuates, to keep the marketmaker from being "picked off" by astute investors.

A riskless interest rate and an adjustment for dividends and other cash payouts is also needed to extract an RND.  OptionMetrics provides data for dividend yields and the riskless rate, calculated as an interpolated value between LIBOR rates for the nearest maturities bracketing option expiration.[9] Future dividends through option expiration are estimated by extrapolating the current yield, with ex-dividend dates for non-index underlyings estimated by an algorithm.[10]

An RND applies to the option expiration date, but nearly all U.S. equity options are American, so it is necessary to convert option quotes into equivalent European option prices. The standard approach, used by OptionMetrics and others, is to use the Binomial model to find the American option's implied volatility after adjusting for early exercise, and then to use that IV in the Black-Scholes equation to compute the price of the equivalent European contract.[11]

An RND can be extracted from call prices or put prices or both together. Most of the value for deep in the money (ITM) contracts is intrinsic value. Optionality is the only part that embeds volatility but it is small relative to bid-ask spreads on ITM options, so it is common to eliminate them and to extract RNDs (and the VIX index) using only OTM calls and puts.[12,13] Lastly, when weekly data frequency is desired, many researchers pick Wednesdays to minimize the number of days with shocks from significant data releases and similar events.

[9] It is not clear what the most appropriate riskless rate to apply to long and short options position is. Many researchers have used U.S. Treasury bill rates. In principle, the risk free interest rate should be related to the cost of funding for options positions that require borrowing, and also to the rate that can be earned when a position generates a cash inflow at the beginning. There might be a substantial spread between these two rates. Fortunately, for short maturity options, which are the most actively traded, the interest rate does not affect option prices much except for those that are deep in the money.

[10] Some authors try to extract the expected future dividend yield from put-call parity. This seems misguided: firms try hard to maintain their dividend payout, so last period's dividend is generally a very good estimate of future dividends. Inverting put-call parity has the effect of bringing any pricing noise in the options market into the dividend calculation. See also footnote 13.

[11] See the OptionMetrics reference manual (2008) for full details.

[12] Equation (5) applies to calls or to puts. To combine the two in a single estimation, put prices are transformed into equivalent call prices using put-call parity.

[13] In the market, puts and calls with the same exercise price frequently have slightly different IVs, even though arbitrage should prevent it. IVs for puts typically appear higher than for calls, but this depends on the interest rate used in the calculation, so the size and even the direction of the difference is not unambiguous. The arbitrage trade to close the gap when put IV is above call IV would require buying the call, shorting the (relatively overpriced) put, selling the stock short, investing the proceeds at the riskless rate, and carrying the position to option expiration. Not surprisingly, the costs and other frictions associated with this trade are large enough to allow some IV differences to persist. Figlewski (2009) suggests combining both call and put IVs in the neighborhood of the current stock price and imposing a smooth transition from one to the other, but this is rarely done. For example, the VIX calculation makes no adjustment for the discontinuity.

Parametric modeling and nonparametric estimation are the two main paths to go from option prices to an RND. The first imposes structure a priori either on the returns process or on the date T density without specifying price dynamics before that, but since such assumptions are never exactly satisfied in practice, parametric models find many options "mispriced" at current market prices. By contrast, nonparametric extraction automatically matches market pricing. This is important for marketmakers who need to trade at market prices, although it rules out using the RND to find trading opportunities in mispriced options. But fitting an RND nonparametrically when option prices are too noisy or too sparse does not work well and may not be possible at all. Stocks with relatively limited option trading may list only a handful of strikes, making it impossible to extract a reasonable RND by taking numerical second differences, while parametric specification allows a researcher to impose a plausible shape that can be fitted with a small number of parameters.[14]

Jackwerth (2004) argues that the extraction method makes little difference for the center of the distribution where there are many option prices, so differences among methods will occur in the tails where there are few observations available to identify the best model. Empirical testing will turn heavily on the particular tail observations or lack of them that happen to be present in a given data sample. Parametric models automatically impose a structure for the tails, even without sample data, while nonparametric methods cannot go beyond the range of traded option strikes and have to find some other way to fill out the tails of their RNDs.

[14] A common problem that can occur with any underlying arises following a large change in the price level. Option exchanges set exercise prices distributed around the current asset price. If it rises sharply, say, new contracts with higher strikes will be introduced, but they may not extend very far into the upper RND tail. Meanwhile, low strike options, once listed, will continue to be quoted until they expire. The opposite occurs following a strong downturn. In either case, the distribution of available exercise prices may become quite asymmetrical, thus providing relatively little information about the RND tail on one side.

Good in-sample fit is therefore relatively weak evidence that one has the true model. Moreover, as Bates (1996) points out, citing an argument raised in Berkowitz (2001), with our limited understanding of what drives volatility smiles and RNDs, the best estimate of tomorrow's surface may well be today's, so even out of sample tests may not reveal which of several models is really the best if they only look one day ahead.  Other model properties, like stability of parameter estimates over longer periods, may be more important in practice.

### 4.2  Parametric RNDs

Much early research tried to find a density from a known family that best fit empirical RNDs.  In Black-Scholes and the Binomial, the RND comes from a fully specified returns process, but in most cases the RND was fitted by approximation, separately from a model of short run dynamics.  Generalizing the RND while remaining within the Gaussian framework could be done in two ways: composite distributions based on the normal/lognormal, and mixture models.  The first category includes Gram-Charlier approximation (Jarrow and Rudd (1982), Corrado and Xu (1996, 1997), Rompolis and Tzavalis (2007)), Edgeworth densities (Rubinstein (1998)), and Hermite polynomials (Madan and Milne (1994), Abken et al. (1996), Xiu (2014)).

Several articles in the 1990s used mixture models to explore market expectations about macro events with strongly binary (or trinary) outcomes. Melick and Thomas (1997) fitted a mixture of three lognormals to explain RNDs from crude oil options prior to the 1990-91 Gulf War, to reflect three possibilities: no conflict, a major conflict, or no resolution before option expiration; Soderlind and Svensson (1997) reviewed  the literature to that point across many markets and explored a normal mixture approach; Gemmill and Saflekos (2000) used two lognormals in trying to tease out investors' expectations about the outcomes of British elections;

Bahra (1997) speaks favorably about mixture models in discussing RND research at the Bank of England.

Gaussian distributions have trouble fitting the tails of the RND, and approximations like Gram-Charlier can sometimes produce negative probabilities in tail regions, so over the years researchers have branched out to try other densities, of which there are many. The CEV has already been mentioned. A non-exhaustive list includes: the normal inverse Gaussian, a strong contender with good behavior in the tails and four parameters that can be calibrated to the empirical moments in the data (Eriksson, Ghysels and Wang (2009)); the generalized beta distribution, which is defined to lie between 0 and 1 can be fitted to the cumulative RND distribution (Aparicio and Hodges (1998)); Edgeworth expansion of the Binomial adapted to the implied tree methodology described below; the generalized extreme value distribution, which has desirable tail properties (Markose and Alentorn (2011)); RNDs produced by an underlying Lévy process that makes only jumps (Carr, Geman, Madan, and Yor (2003)), and others.

### 4.3 Nonparametric RNDs

The most common way to calculate an RND nonparametrically is numerical approximation to equation (5). The middle portion of the RND is extracted from the options market data and then some way must be found to deal with the tails. To begin, one smooths and fills in the option price curve as a function of the strike prices. But even for actively traded single-stock options, the number of available strikes is relatively small. To produce a reasonably smooth RND, the data is smoothed and filled in by interpolation to create approximate values on a much denser set of strikes. Shimko (1992) was the first to suggest transforming the option prices into Black-Scholes implied volatilities that are much closer in size than options prices. The BS equation is only used to transform prices into a more convenient space, something like taking logarithms. The inverse function is then applied to convert back from IVs to option prices after interpolation. There is no implication that BS is the market's pricing model. Smoothing and filling in an IV curve has frequently been done using a cubic spline, but because the RND calculation entails taking a second derivative and the second derivative of a cubic equation is continuous but not differentiable, the resulting density will often have large (artificial) spikes. This is easily avoided by using a 4th or higher degree spline.

Let $\{X_1, X_2, ..., X_N\}$ represent a dense set of strike prices selected to subdivide the range between the lowest and highest traded strike into small intervals, and ordered from lowest to highest. IVs for the traded contracts are computed from their market prices, and IVs for the new intermediate strikes are filled in by interpolation. Then, each interpolated $IV(X_i)$ is translated back into a call price $C(X_i,T)$ using the BS equation and the RND is estimated by numerically implementing eq. (5). Applying equation (4) to estimate the total probability in the left tail up to $X_2$, one approximates $\frac{\partial C}{\partial X}$ at $X_2$ and computes $G(X_2) \cong e^{rT} \frac{C_3 - C_1}{X_3 - X_1} + 1$, where G(.) is the approximated cumulative RND and $C_i$ is a shorthand expression for $C(X_i,T)$. The probability in the right tail from $X_{N-1}$ to infinity is approximated by,

$$1 - G(X_{N-1}) \cong 1 - \left( e^{rT} \frac{C_N - C_{N-2}}{X_N - X_{N-2}} + 1 \right) = -e^{rT} \frac{C_N - C_{N-2}}{X_N - X_{N-2}} .$$

The approximate density $g(X_n)$ at any $1 < n < N$ is given by[15]:

$$g(X_n) \approx e^{rT} \frac{C_{n+1} - 2C_n + C_{n-1}}{(\Delta X)^2} \tag{8}$$

[15] We assume here that the strike prices in the dense set are equally spaced such that $X_n - X_{n-1} = \Delta X$, for all $n$. Otherwise, the calculation of numerical derivatives needs to be adjusted appropriately.

The RND tails outside the range of traded exercise prices must be filled in in some other way. Shimko (1993) and Bliss and Panigirtzoglou (2002, 2004) simply extended the IVs at the lowest and highest available strikes into the unobserved tails, but this forces the tails to be lognormal, while the evidence is that they should be fatter.

Figlewski (2009) proposed fitting Generalized Extreme Value (GEV) distributions to the missing tails, citing the Fisher-Tippett Theorem which proves that the remote (right) tail of any plausible choice for a returns density will converge to the form of a GEV tail. The GEV is characterized by three parameters, location, scale and tail shape. Birru and Figlewski (2012) simplified this approach by substituting the Generalized Pareto distribution (GPD), shown in eq. (9), for the GEV.

$$F(S_T \mid S_T \geq c) \quad = \quad \begin{matrix} 1-\left(1 \ + \ \xi\left(\dfrac{S_T - c}{\sigma}\right)\right)^{-1/\xi} & \quad if \ \xi \neq 0 \\ 1-\exp\left(-\left(\dfrac{S_T - c}{\sigma}\right)\right) & \quad if \ \xi = 0 \end{matrix} \tag{9}$$

The GPD gives the distribution of values in the right tail of a GEV, so it has the same tail shape $\xi$. The location parameter c is the left endpoint, which is set equal to $X_{N-1}$ where the GPD tail will be attached to the empirical RND. If $\xi > 0$, the tail is fatter than the normal; $\xi < 0$ is a "thin" tail that does not extend to infinity; and $\xi = 0$ is a tail like the Normal distribution. Note that this describes only the right tail. To get the left tail using the same technique, one must transform the problem, fitting on $–S_T$.[16]

[16] Depending on the problem, it may also be necessary to translate the axis by adding a large constant to $–S_T$, so that the tail to be fitted lies fully in the positive quadrant.

### 4.4 Implied Trees

"Tree" models, starting with the Binomial Model of Cox, Ross, and Rubinstein (1979) (CRR) are extremely useful in practical option valuation, with much greater flexibility than the BS model. The plain vanilla CRR model produces a binomial distribution as the RND. Extending the model to trinomial branching allows much richer dynamics, with transition probabilities that can handle stochastic volatility and jumps. Rubinstein (1994) described how to imply out an entire binomial tree from a set of option market prices. The implied tree forces local Q-dynamics to reproduce the observed market RND as of option expiration, which links a returns generating process, and a procedure for hedging, to the RND. Similar ideas were explored by Derman and Kani (1994) in a binomial, and Dupire (1994), in a trinomial model with nonstochastic time-varying volatility. Jackwerth and Rubinstein (1996) used intraday data and applied a variety of adjustments and corrections to optimize the implied binomial tree procedure suggested in Rubinstein (1994).

Such models have been called "local volatility" models because volatility is allowed to change over time, albeit nonrandomly. However, the practical value of the approach was called into serious question by Dumas, Fleming and Whaley (1998) (DFW). An implied tree embeds an entire deterministic structure of volatilities covering every node. What DFW demonstrated is that this deterministic structure was generally so different one week later from what was implied on day t, that the model was outperformed by a simple quadratic function of strike price and maturity, both in terms of matching options market prices and also in hedging effectiveness. One clear implication was that nonstochastic volatility of stock returns is not a viable assumption for pricing real world options.

### 4.5 Other Methods

Aït-Sahalia and Lo (1998, 2000) developed a different nonparametric technique using kernel regression to estimate an option pricing surface that best fits market call prices conditional on the values of S, X, T, r, and d (stock price, strike price, maturity, interest rate, and dividend yield) such that the surface and its derivatives are "smooth" within a local region determined by the bandwidth parameter. The RND is extracted from the fitted call price surface. However, the approach is very data intensive and it becomes feasible only if data from multiple days are combined. Aït-Sahalia and Lo (1998) took option prices from the full year 1993 and assumed that a standardized RND function was constant over that year. This procedure assumes that all risk premia were constant conditional on (S, X, T, r, d). Computed smiles were downward sloping and became steeper for shorter maturities, but there is only one smile for the whole year at any given maturity.

Pinning down a nonparametric RND using the principle of maximum entropy was suggested by Stutzer (1996) and Buchen and Kelly (1996). The idea is that one wishes to find a density that is consistent with the observed data and any constraints the investigator places on the problem, e.g., no-arbitrage, but imposes the least constraint on the properties of the unknown information. This approach has esthetic appeal to modelers, but given the dependence on historical returns information, that is handled somewhat differently than in other models, maximum entropy has not been found to be superior empirically to more familiar approaches.

### 4.6 Ross Recovery

Ross (2016) presented important new ideas in this area, which caused quite a splash among RND researchers when the working paper first appeared. The approach seemed to accomplish something that was widely believed to be impossible: separating the P-density and the pricing kernel from the observed Q-density without making any restrictive assumptions about either the representative agent's utility function or the empirical returns process. In a discrete-time discrete-state Markov model, Ross applied Perron–Frobenius Theory to argue that given an observed Q-density, a unique P-density can be computed if the pricing kernel over future states does not depend on the current state. A full analysis of this argument, and the counterarguments offered by Borovička, Hansen, and Scheinkman (2016) and others, is well beyond the scope of this article. The underlying problem, however, is that what Ross assumed in developing his model is actually much stronger than it appears. Borovička et al. show that if the pricing kernel dynamics includes a martingale component, it is that component that Perron–Frobenius Theory extracts, and it need not correspond with investors' beliefs about the P-density. In practical terms, if risk premia are affected by random shocks, from events in the macroeconomy for example, the pricing kernel does depend on the current state and recovery fails. Although empirical testing of Ross's theory is still in the early stages, Jackwerth and Menner (2017) found that the P-density for the S&P 500 extracted from the model is a poor predictor of future returns and the pricing kernels it leads to are highly implausible in many cases.

## 5. Extracting and Using Information about Future Returns from a Risk Neutral Density

Much academic RND research tries to model and understand where RNDs come from. But another line of research focuses mainly on extracting information from them. Research on risk premia is covered in the next section. This section reviews attempts to extract information about objective probabilities—the P-distribution—such as future realized volatility. But as just discussed, the nearly universal conclusion remains that it can't be done without imposing additional assumptions. Christoffersen, Jacobs and Chang (2013) provide an excellent and comprehensive review of the literature on extracting and forecasting with information from

RNDs, including multi-page tables that summarize the findings of a great number of articles that I am not able to discuss here. Bates (2003) gives an insightful "retrospection" on the empirical evidence up to that point.

### 5.1 Predicting the Risk Neutral Density Itself

Some researchers simply construct RNDs and present them as is, without attempting to decompose them further. Information is in the shape of the density itself, and especially in how it changes from period to period. Researchers at the Bank of England conducted some of the earliest work on RNDs, see Bahra (1997) and Clews et al. (2000) to be used for policy evaluation. More recently, the Minneapolis Federal Reserve began publishing RNDs semimonthly for a wide variety of markets (see Federal Reserve Bank of Minneapolis (2014)). They describe the curves, but do not attempt to break them down into an implied P-density and a pricing kernel. In fact, Feldman et al. (2015) argue that the RND is more relevant than the P-density for understanding households' behavior.

Such comparisons, and the use of RNDs in general, became more systematized after Bakshi, Kapadia and Madan (2003) presented relatively simple formulas for the moments of a risk neutral density in terms of the underlying option prices. The derivation built on Bakshi and Madan (2000)'s proof that any bounded payoff function can be spanned by a continuum of OTM calls and puts, This made risk neutral variance, skewness, and kurtosis values readily accessible without any need to deal with messy details of smoothing and interpolating IVs and figuring out how to handle the tails. Note, however, that the latter problem did not go away with the new methodology.

### 5.2 Predicting Excess Returns

Practitioners are always interested in trading strategies that produce excess returns, and many have been proposed and investigated. Academics are always interested in finding better pricing models, and the evidence that a new model is better is that it can produce excess returns from trades executed at current market prices. A vast literature has developed exploring returns on options strategies, many of which involve risk neutral densities and implied parameters. There is no way to review these ideas in detail here. However, it is worth observing that this research is generally portrayed as uncovering and exploiting "mispricing" in the market.

To interpret these results as mispricing puts no weight on risk premia: "excess" returns might simply be fair compensation for bearing hard-to-hedge risks that are not included in the pricing model. Or the mispricing may reflect the effects of market constraints, like short sale restrictions and financing issues. The implication of calling deviations from theoretical values mispricing is that option investors are wrong in the way they price options. If so, they can be expected to get smarter over time, and excess returns observed in past data should disappear in the future. Alternatively, perhaps investors are just earning fair premia as compensation for bearing risk. In that case, there may be excess return but there is no real excess profit to the representative investor. Or maybe the return is due to market constraints, and the trade cannot really be done by an investor who faces the same trading constraints as everyone else.

In short, the presence of excess returns means either the options market is using the wrong model and is informationally inefficient, or the mispricing is not really there when all risks and trading constraints are properly considered. A final caveat concerns an issue that is very hard to assess in practice: Given that a true mispricing exists, how large a trade would be possible: how big is the mispricing in terms of dollars of profit?

With those cautions about interpreting reported excess returns to strategies based on RNDs, we turn to forecasting volatilities.

### 5.3 Predicting Future Realized Volatility

From the beginning, extracting implied volatility from options prices elicited by far the most research interest on risk neutral densities. Practitioners simply wanted the right number to put into the BS formula, while academic research was interested in whether the market's expectation for future realized volatility was informationally efficient. The complete and surprising failure of IV from OEX options in the standard rationality test regression, as reported in Canina and Figlewski (1993), demonstrated that the informational efficiency of IV as a forecast of future volatility was open to question.

The basic rationality test is a simple regression of realizations on forecasts:

$$RV_t = a + b\ FV_t + u_t \qquad (10)$$

where $RV_t$ is realized future volatility from the present through option expiration, $FV_t$ , is a forecast of $RV_t$ as of date t, either IV or historical volatility calculated from past returns, and $u_t$ is the regression residual. If $FV_t = E[\ RV_t\ ]$ for some forecast, then a = 0, b = 1.0 and the variance of $u_t$ should be small. CF found $\hat{b}$ to be insignificantly different from 0 and $\hat{a}$ to be significantly positive for IV as the forecast. This test has been run repeatedly since then on IVs from other options on equities and indexes (for example, Fleming (1998), Christensen and Prabhala (1998)), and on a great many other underliers. As explained in footnote 6, because of the formidable barriers to arbitrage with OEX options in the mid-1980s, the CF results were a polar case. By far the most common finding when the rationality test is run is that $\hat{a}$ is positive, $\hat{b}$ is significantly greater than zero but often around 0.7 to 0.9, significantly less than 1.0, and when historical volatility is run with IV in the same ("encompassing") regression, it generally does not get a significant coefficient. Poon and Granger (2003) reviewed the extensive literature on this issue.

It should be noted that to the extent IV contains a risk premium—the strong evidence for this is reviewed in the next section—it is not the market's best (P-density) prediction of future volatility. IV is a Q-density volatility forecast and it should not pass the rationality test. When IV outperforms historical volatility but fails the rationality test, it is incorrect to conclude, as many papers do, that although it is not an efficient forecast, IV is the best available prediction of future volatility. The statistical evidence from regression (10) shows that the minimum variance estimate is not $IV_t$ but $\widehat{RV}_t = \hat{a} + \hat{b}\, IV_t$ where hats refer to regression estimates.

## 6. Extracting and Using Information about Risk Premia from a Risk Neutral Density

Option valuation models that exclude risk premia do not fit actual data well. Riskless arbitrage is not really possible, given transactions costs and insufficient hedging instruments to cover all of the risk factors, so options prices can and should include risk premia. This section describes the efforts to extract and to understand the risk premia embedded in option RNDs.

Empirically, delta-hedging works pretty well even with Black-Scholes deltas. Although not all option price risk is eliminated from a hedged position, theoretically, the directional risk that remains is largely idiosyncratic and should not be priced in equilibrium. The hedge should earn the riskless interest rate, and any consistent excess return, positive or negative, might be attributable to risk premia. Delta hedging reduces directional risk but does not affect volatility risk. Efforts to isolate a variance risk premium (VRP) frequently do so in the form of the excess return on a delta hedged position (Bakshi, Cao and Chen (1997), Dumas, Fleming and Whaley

(1998), Bakshi and Kapadia (2003a,b)). Research on the second moment of returns focuses on volatility in the BS framework, but includes the volatility of volatility in SV models, and also jump risk size and intensity in SVJ models.

Not all volatility risk is the same. Rather, idiosyncratic volatility is priced much differently from systematic volatility, which shows up in much steeper volatility smiles for index options than for options on individual stocks (Toft and Prucyk (1997), Dennis, Mayhew and Stivers (2006)). Recent modeling has uncovered strong evidence of at least one additional risk factor in the volatility process that appears to be particularly connected to the left tail of the returns distribution, and moves quite independently from the underlying returns process. Empirically, this tail risk factor explains a large fraction of the measured VRP.

Before proceeding further, it should be noted that different authors describe the VRP differently. One may declare the volatility risk premium is strongly negative (e.g., Bakshi and Kapadia (2003a), Carr and Wu (2009)) while another looking at essentially the same data says it is positive (e.g., Bollerslev, Tauchen and Zhou (2009)). They are generally saying the same thing, but using different measures. Investors dislike exposure to volatility, so they pay higher than BS prices for options that partially hedge volatility risk: The VRP in terms of option values is positive. Buying an option at a high price lowers its expected rate of return, so in terms of the expected return on an option held long by itself or in a delta hedge, the risk premium is negative. Finally, when a risk averse investor pays a higher option price than a risk neutral investor would, the risk premium measured in terms of IV is positive. Nevertheless, virtually all research on VRPs concludes that volatility risk is priced in the market and it shows up in higher option prices, higher option IVs, and lower expected returns on long option positions.

### 6.1 Risk Aversion

While the major interest in implying information out from options has always focused on the volatility, several papers in the late 1990s and early 2000s tried to use the RND from the stock market to explore and measure the representative investor's risk aversion. The efforts were econometrically interesting, but largely unsatisfying in terms of the implied utility functions.

Following the stock market crash of 1987 there was a striking change in the typical RND from S&P 500 options. Before 1987, the volatility smile was relatively symmetric between negative and positive returns, but afterwards it changed to the largely monotonic downward skew that has been its general shape since then. This presented serious challenges for efficient markets explanations—investors just seemed to be paying too much for OTM puts, given statistical estimates of the P-density and plausible levels of risk aversion. Rubinstein (1994) called the phenomenon "crash-o-phobia". Bates (1996, 2000) and Jackwerth (2000) both found it hard to come up with utility functions that were consistent with strongly left-skewed RNDs from the S&P index after the crash.

A frequently-cited work on implied risk aversion is Bliss and Panigirtzoglou (2004). They extracted RNDs for the S&P 500 and FTSE 100 indexes using the method of Bliss and Panigirtzoglou (2002) and constrained the representative agent's utility function to be either constant relative risk aversion (CRRA) or constant absolute risk aversion (CARA), which allowed them to compute an implied P-density. The risk aversion estimate that gave the best fit between the subjective P-density and realized returns across the full S&P sample was a fairly reasonable 3.53. However, there was strong period to period variability and a very steep term structure effect, with nearby maturities (1-2 weeks) implying much higher risk aversion than 6-week contracts. It is not clear how much regularity was forced onto the problem by the assumed utility function, and the estimation technique that imposed lognormality on the remote tails.

Bliss and Panigirtzoglou did not explore the behavior of the pricing kernel, but Jackwerth's (2000) results revealed a significant problem with it: In the region of market returns near the middle of the distribution, implied risk aversion begins to increase with increasing wealth (higher stock prices) instead of decreasing. This strange result has proven to be remarkably robust, and has come to be called the pricing kernel puzzle. Aït-Sahalia and Lo (2000) also found the pricing kernel for the S&P index to be very badly behaved using their nonparametric kernel smoothing RND technique described above. Discussion of this important subject is deferred to the next section.

### 6.2 The Variance Risk Premium

Work in this area is sometimes done on implied and realized volatility and sometimes on variance. Volatility is required in an option model, but variance is somewhat easier to deal with statistically. There are no important differences in the conclusions to be drawn from the two approaches and in describing the research, I will use them interchangeably.

A variance risk premium in option pricing can arise because risk control through delta hedging is hampered by transactions costs and discrete rebalancing. The risk in this case occurs even when volatility is nonstochastic, and the premium shows up as a (negative) excess return on a delta-hedged options position. In an SV model variance risk also arises because variance itself is stochastic. In that case, the VRP provides compensation for the uncertainty over future variance, and the size of the premium should be related to the volatility of volatility. This risk premium is often measured in terms of the difference between IV and expected realized volatility, that is, volatility (or variance) under the Q-distribution minus volatility under the P-distribution (e.g., Bollerslev et al. (2009), Bekaert et al. (2010), Zhou (2010)). If jumps are allowed, each of the unknown parameters that determine jump size and jump frequency may also carry a risk premium.  Differences in exactly what is tested and how make it harder to integrate results from different studies into a unified theory.  Pan (2002) performed a combined estimation on the underlying stock returns and options prices to fit the Bates (2000) jump-diffusion model, after demonstrating that the statistical evidence ruled out constant diffusive volatility and stochastic volatility without jumps.  She then found that more than a third of the variance risk premium could be attributed to jump risk.  Egloff, Leippold, and Wu (2010) provide a comprehensive review of the literature, except for the most recent articles.

As described above, numerous empirical studies in the last twenty years have found that volatility under the risk neutral density is higher than empirical volatility.  Until recently, examining this question in detail was seriously hindered by the fact that empirical volatility is not observable, so any test is a joint test of the option pricing model and the procedure for estimating the expected value of future integrated variance in a model that might contain multiple latent stochastic variance and jump factors.  This problem has been somewhat mitigated by the use of intraday tick data.  By measuring returns at very short intervals, e.g. 5 minutes (Bollerslev, Tauchen, and Zhou (2009), Bollerslev and Todorov (2011)), one can compute an "instantaneous" realized volatility for each sample day from a reasonable number of observations.  This also allows identification of jumps, which are presumably much larger than diffusive price moves over such a short period. The final piece that needs to be incorporated is overnight variance.  Obviously, overnight returns are an important component of total variation over an option's life.  Overnight variance has not always been included in studies that compute integrated RV with intraday data, but it should be.

Disentangling the premia on multiple potential risk factors by deconstructing option smiles is challenging.  Some researchers have made use of prices for variance swap contracts which, in principle, represent direct measurement of the market premium on future integrated variance. (Carr and Wu (2009)). Unfortunately, variance and volatility swaps are only traded over the counter, so market data has not been easy to acquire. Carr and Wu used a replication strategy to construct a synthetic variance swap price series from traded option prices.

Recent studies have largely adopted models that include both diffusive stochastic volatility and jumps. Trying to tie up the loose ends uncovered in the early 2000's, they may include more than one diffusion with different long run means and speeds of mean reversion, for example, a slow moving one and a faster one. In earlier articles, to incorporate the possibility of a crash, an independent jump process was introduced into the specification with a negative coefficient. The jump specification was either one-sided like the exponential, or one that could accommodate a left skew, typically the lognormal. More elaborate current models may generalize to allow both down and up jumps with different characteristics and more sophisticated interactions among the risk factors, such as a positive dependence of jump frequency on the level of diffusive volatility (e.g., Andersen, Fusari, and Todorov (2015), Bollerslev, Todorov and Xu (2015)).

The overall conclusions about the VRP from the recent collection of highly developed models estimated with extraordinary econometric sophistication are that there is extremely strong evidence that option prices include a significant risk premium for exposure to the level of volatility and that this produces excess returns. The VRP varies widely over time and across stocks. It seems to be driven by more than one factor, with the second one being closely connected to the left tail of the RND, but largely independent of the overall level of volatility or to the right RND tail (e.g., Christoffersen, Heston, and Jacobs (2009), Bollerslev and Todorov (2011), Carr and Wu (2009)).

### 6.4 Index vs. Single Stock Variance Risk Premiums

As noted by Rubinstein (1994) and Bates (2000), volatility smiles are different for indexes than for their constituent stocks, with individual stocks continuing to show fairly symmetrical smile-like patterns while index options developed much steeper monotonic skews after 1987. Bakshi and Kapadia (2003b) concluded that the VRP was significantly negative for both (in terms of returns on a delta-hedged long option position), but much smaller for individual stocks. Dennis and Mayhew (2002) found the same. This raises the question whether systematic (index) volatility risk is priced in the market, while idiosyncratic volatility diversifies and is priced less or not at all.

Cao and Han (2013) analyzed the correlation between idiosyncratic variance and returns on delta hedged single stock options in the cross section. Controlling for the overall level of IV, options on stocks in the highest quintile earned significantly lower returns than those in the lowest quintile, supporting the idea that idiosyncratic volatility is priced in the options market. Cao and Han suggest that dealers find options on stocks that have high idiosyncratic volatility riskier to hedge with index options and more expensive to hedge with the underlying stock, so they charge higher premiums.

## 7. The Pricing Kernel Puzzle

As shown in (2), the pricing kernel $k(S_T)$ gives today's value of a \$1 payoff in future state $S_T$, relative to the probability of being in that state of the world. Payoffs in future states in which the investor is relatively poor should command higher prices today than payoffs in wealthy states. When the stock market is a proxy for overall wealth, low $S_T$ corresponds to a higher utility value on the payoff. In this case, the pricing kernel should be monotonically decreasing in $S_T$. But almost invariably, it isn't.

Figure 1 shows a typical example. The solid curve is the Q-density for the S&P 500 on Feb. 21, 2014, extracted from February maturity options on Feb. 3, 2014. The dotted curve is an estimate of the P-density for that date. It is lognormal with volatility set equal to a weighted combination of historical volatility of log returns and the VIX, where weights were those that would have produced the most accurate forecasts of 18-day RV in the past. The mean was set to the current riskless rate plus an assumed risk premium of 5% (see the Appendix to Figlewski and Malik (2014) for full details). The x-axis is in standard deviations of log returns.

FIGURE 1 GOES ABOUT HERE

The P-density is normal and the Q-density is left-skewed. The limited number of listed calls with high strike prices causes the RND to be truncated at the right end at 1.82σ while the left tail extends much further.[17] The dash-dotted curve is the pricing kernel, which shows the strong and highly counterintuitive but ubiquitous upward sloping middle portion, where risk aversion appears to increase with wealth.

There have been many efforts to explain the pricing kernel puzzle, none altogether successful. Empirical studies consistently find kernels with upward-sloping portions, although the description of the shape varies. Many focus on a middle portion, either the concave part or the convex part, and attempt to explain that, but do not extend the range of the x-axis to show negative slopes at both ends. The shape of the kernel in Figure 1 is the norm for the S&P 500, so an explanation that leads to a uniformly concave or convex kernel cannot be the complete

[17] The kernel was computed using only the RND from the prices of traded options. It does not include the tails extended with the GPD density.

concave explanation.  A full exposition of this subject is beyond the scope of this paper. Cuesdano and Jackwerth (2017) provide an excellent and comprehensive review.

Efforts to explain the puzzle within the standard modeling framework by looking for a combination of a data generating (P-density) process and a utility function defined over terminal wealth that could produce a pricing kernel like this all failed, including Bates (1996, 2000, 2008), Aït-Sahalia and Lo (2000), Pan (2002), Rosenberg and Engle (2002), Carr, Geman, Madan, and Yor (2002), Wu (2006), Ziegler (2007) and others.

Hens and Reichlin (2013) offer three potential theoretical explanations: investors are risk-loving, investors are bad forecasters of the P-density, or the options market is incomplete and investors care about other things in addition to terminal wealth.  They show that in theory, any of these can produce a generally downward-sloping kernel with upward sloping portion(s), but they do not attempt to apply the model empirically.

Presumably, we may reject generally risk-loving option market participants as a plausible description of real world markets.  Forecasting the market's subjective beliefs about the P-distribution is, of course, subject to considerable error.  Linn, Shive and Shumway (2017) argue that a new forward-looking nonparametric estimator can make the puzzle go away.  But while this is intriguing, it can only resolve the puzzle if the new P-density estimator is actually the way options traders have been forming their returns expectations all along, as they generated the option prices that have puzzled everyone else.

The third suggestion is that the options market is incomplete and that there are other dimensions than returns that investors care about and build into option prices.  Christoffersen, Heston and Jacobs (2013) show that if investors care about both returns and volatility risk, a U-shaped kernel can be generated.  Chabi-Yo (2012) is also able to generate a variety of shapes in a model where both expected return and volatility enter the utility function. It is not clear whether priced volatility risk is sufficient to produce the full shape as shown in Figure 1, rather than one that is consistently convex or concave throughout the full range of potential outcomes.

A different kind of explanation invokes heterogeneous beliefs. The representative agent assumption eliminates any effect of heterogeneity, but as pointed out above, such models cannot explain trading in zero-sum contracts. Ziegler (2007) explores how differences in risk aversion can affect the kernel, but finds that is not enough. Differences in expectations can generate pricing kernels with humps but he concludes that the degree of pessimism among the bears required to match market pricing of OTM puts was implausibly large. Bates (2008) develops a theoretical model with crash-o-phobic investors who buy insurance from regular investors through the options market and shows that a pricing kernel with bumps is possible (though his example does not produce upward sloping portions like those in Figure 1).

Bakshi, Madan and Panyototov (2010) take a different angle, starting with an assumption that the pricing kernel is U-shaped and seeing if this can help explain other aspects of options pricing, such as negative returns on OTM calls when the index is rising. Shefrin (2008) and Barone-Adesi, Mancini, and Shefrin (2013) offer a potential explanation rooted in behavioral finance. In their model, rational investors trade with overconfident ones. The missing factor their analysis tries to incorporate is "sentiment." The (2013) paper attempts to extract a confidence measure from the kernel and finds that it correlates with other confidence measures from the literature.

At this point, it is fair to say that the "pricing kernel puzzle" remains a puzzle, despite the continued efforts to decode it. What has been learned so far is that the nonintuitive shape is a robust feature of options markets, at least when variants of the standard econometric and extraction techniques are applied. Models that incorporate additional risk factors like volatility risk are capable of reproducing some of the features of empirical kernels but not to the point of being considered a full explanation. This area could use further investigation, and more imaginative efforts to bring behavioral factors into modeling heterogeneous expectations and risk preferences.

## 8. The Volatility Surface

Although researchers have had to go far beyond Black-Scholes in trying to construct option pricing models that can match the properties of observed stock returns and call and put prices, practitioners have largely stuck to "Practitioner Black-Scholes" (PBS), which uses the Black-Scholes equation but with a different implied volatility extracted from each individual option's current market price. This, of course, forces the model to accept the current price as the correct one. The objective of PBS is to obtain the option's delta for hedging purposes, and deltas are generally quite similar across models for options that are not too far out of the money. But BS delta assumes IV remains constant, while IVs change stochastically and sometimes by quite large amounts from one day to the next. Proper hedging also needs to adjust for vega risk, and that requires a prediction of next period's volatility.

The first effort to handle this problem came from the local volatility models, discussed in Section 4.4. An implied tree, for example, builds changing variance into each node so that the density at expiration matches the current RND. But as Dumas et al. (1998) demonstrated, IVs in the market did not follow the dynamics embedded in the fitted trees. Note that it is the dynamics of implied volatility not true volatility that must be modeled, because (leaving out financing cost) the change in the value of a delta hedge is $(\Delta C - \delta \Delta S)$, where $\delta$ is the option's Black-Scholes delta and the change in the call price $\Delta C$ depends on both $\Delta S$ and the change in IV.

Here I must mention an issue that I have personally always found bothersome: We know that Black-Scholes is not the way the market prices options. Yet we routinely extract BS IV and try to model its dynamics. What is the scientific basis for taking an incorrect model, extracting a set of fudge factors that artificially set its outputs for a single date equal to the outputs from the true but unknown model (i.e., market option prices), and then trying to apply elaborate statistical procedures to capture and predict the behavior of the wrong model's fudge factors?

Casting that objection aside, as practitioners and academics alike all do, there has been an enormous amount of research on modeling and forecasting Black-Scholes IVs. Equity options exhibit strong negative correlation between equity return and IV, so a change in S affects call value both from the direct effect and also because IV moves. A simple adjustment to the hedge ratio can help to offset this problem. When IV can change, the change in call value is given by $(\Delta C - \left(\delta + \frac{\partial C}{\partial IV}\frac{\partial IV}{\partial S}\right)\Delta S)$, where the new term is the BS call vega times $\left(\frac{\partial IV}{\partial S}\right)$, the change in IV due to the change in S. This can be taken directly from the slope of the volatility smile/skew, but that slope can change substantially from day to day (see, for example, Christoffersen, Heston and Jacobs (2009)), which limits the ability to eliminate vega risk this way.

SV and SVJ models specify returns processes that cover the full range of possible model outcomes, but hedge design can be quite tricky in these models, since there are typically multiple latent factors whose current values must be estimated. It is not clear that more detailed volatility processes in a theoretical model translate into better hedging performance in practice.

Parametric, semi-parametric, and nonparametric approaches have been explored to predict the behavior of the volatility surface without a fully specified underlying model. An early example is Dumas, Fleming and Whaley's (1998) "naive" quadratic model, which outperformed highly structured local volatility models.  Cont, da Fonseca, and Durrleman (2001) modeled the implied volatility surface as a mean-reverting diffusion, with three factors corresponding to changes in level, tilt, and curvature, respectively, as suggested by the principal components analysis presented in Cont and da Fonseca (2002).  Gatheral's (2006) book covers this area in detail from a sophisticated practitioner's perspective.

Currently, approaches that may be loosely related to underlying finance principles but are largely nonparametric are being actively explored.  In such models, the volatility surface is represented by IVs at discrete moneyness and maturity points and models for their stochastic evolution are developed, subject to no arbitrage constraints.  Carr and Wu (2016) fit a diffusion process to each IV.  Andersen, Fusari and Todorov (2013) combine a highly structured model for the returns process with an ad hoc four factor structure for the volatility surface.

Perhaps the most ambitious of these models is Israelov and Kelly (2017).  They try to model not just the expected IV change at each grid point, but its full distribution including stochastic shocks and model error.  Their approach tries to break each option's return into a delta part, coming from the change in the index, and an IV part.  To start with, they convert each option's market price into its equivalent Black-Scholes IV and interpolate across options onto a fixed moneyness-maturity grid.  This is done for every date in the sample, which yields a historical sample of standardized IV surfaces.  Then, for each grid point on date t a vector autoregression (VAR) is estimated from past returns at that point, with time-varying innovation volatility governed by a GARCH process. Two common factors, the S&P 500 and the VIX volatility index are used in the VAR, plus a small number of IV dynamics factors estimated from past data by principal components.  When projected values and risks for the whole surface have all been computed, they apply interpolation in the reverse direction, from the standardized grid points to compute the probability distributions for the  IVs and prices of the actual traded contracts.

There are many things to like in this paper, but one that seems especially fruitful is separating modeling of the dynamics of the IVs from a formally specified returns process, after acknowledging that none of our current models can really capture the behavior of the risk neutralization process, so a less ambitious approach, such as nonparametric principal components may be less likely to lead us astray through over-modeling and over-fitting.

## 9.  Conclusions and Future Directions for Research

This article has reviewed much excellent research on risk neutral densities.  The modeling and empirical work in this interesting field have now reached a very high level of sophistication.  Yet we are still far from fully understanding what RNDs mean and how they behave.  There is a lot of room for further investigation.

In the now-standard paradigm, a rational and well-informed representative agent solves (or behaves as if he has solved) a multifactor SVJ model to compute option values and hedge characteristics so he can maximize expected utility, defined over terminal wealth.  This line of attack has reached the point of seriously diminishing returns.  Adding another kind of jump or another kind of interaction among jumps, diffusions, and returns is unlikely to explain much more of the so far unexplained behavior of options pricing.  If we restrict our modeling to this framework, we will continue to look for our missing explanatory factors under the familiar street light rather than in the dark and as yet largely unexplored bushes, where new ideas are more likely to be found.  Let me briefly mention two major directions where further understanding is needed and are worth exploring.

First, as has been mentioned earlier, a derivatives market is a zero-sum game.  A world populated by identical representative agents can price options but would never support futures or options trading, so by making this assumption we effectively rule out what most real world market participants are actually doing.  Hedgers trade derivatives because they are more exposed or more risk averse than the average investor and they want to offload risk onto others.  They do not expect their trading at current market prices will produce returns equal to the riskless rate plus a fair risk premium; they expect low or negative returns on their derivatives trades as the cost of insurance.  Some information traders may enter the market passively, just to harvest risk premia, but most do so because they also do not believe their expected return is the riskless rate plus a fair risk premium; they expect to beat that rate, because they (think they) have better information than others about true values.

For the most part, heterogeneity has been swept under the rug in empirical work, partly because there is very little direct evidence on investor expectations that would allow differentiating between, say, "bulls" and "bears".  By contrast, is it easy to fall back on theoretical aggregation results that show that under quite general conditions, heterogeneous investors can be folded into a representative agent with appropriately averaged beliefs and risk aversion.  A few researchers have tried to focus on the effect of differences within the investor population, such as Ziegler (2007), Shefrin (2008), Bates (2008), and Barone-Adesi, Mancini, and Shefrin (2013)), but these articles only represent first steps in this direction.

The introduction of many new derivative markets all related to the same underlying S&P 500 index offers large and largely unexplored opportunities to go beyond the representative agent paradigm. Multiple markets now exist in both the future index level and also its volatility. The former include futures and options on the spot index, options on the index futures, and exchange-traded funds offering simple and leveraged exposure to the index, both long and short. Exposure to implied volatility (simple and leveraged, long and short) is traded in the form of futures, options, futures options, and a panoply of long and short ETFs on the VIX and other volatility indexes. The Chicago Board Options Exchange (2009, 2011) now trades contracts on implied correlation and implied skew of the S&P. It certainly seems plausible that investors who take long positions in an S&P ETF with double leverage have different expectations than those with double-leveraged positions in an ETF with short exposure to that index. An initial exploration of that specific hypothesis in Figlewski and Malik (2014) found significant differences in their RNDs and some interesting preliminary results, but just scratched the surface of a huge new area ripe for exploration.

A second promising direction for RND research is more direct modeling of the risk neutralization process. Several of the recent articles discussed above found that the left tail of the RND, or of the volatility smile, behaves in ways that seem largely independent of returns, and the model that fits the rest of the RND. The clear suggestion is that risk neutralization is affected by factors other than those associated with the returns process. These might be called sentiment. Israelov and Kelly (2017) take a step in the direction of treating returns and risk neutralization as separate processes, but they only model the latter nonparametrically with principal components. More careful attention to risk neutralization could offer a way to bring "behavioral finance" into formal valuation models, and to test hybrid models on market data.

The bottom line, then, must be: This is a great area with many potential directions for future research. The real world is providing vast amounts of new data for the project, and the creation of so many new derivative markets, all of which must be connected through (not-entirely-riskless) arbitrage allows high-resolution analysis of the expectations and risk preferences of real world investors.

# References

Aït-Sahalia Y. and A.W. Lo. (1998). "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices." Journal of Finance 53, 499-547.

Aït-Sahalia Y. and A.W. Lo. (2000). "Nonparametric Risk Management and Implied Risk Aversion." Journal of Econometrics 94, 9-51.

Andersen, Torben G, Fusari, Nicola, & Todorov, Viktor. 2015. The risk premia embedded in index options. Journal of Financial Economics, 117(3), 558–584.

Bakshi, G., Cao, C., and Chen, Z. (1997). Empirical performance of alternative option pricing models. The Journal of Finance, 52(5):2003–2049.

Bakshi, Gurdip, and Dilip Madan, 2000, "Spanning and derivative-security valuation," Journal of Financial Economics, 55(2), 205-238.

Bakshi, G. and Kapadia, N. (2003a). Delta-hedged gains and the negative market volatility risk premium. Review of Financial Studies, 16(2):527–566.

Bakshi, G. and Kapadia, N. (2003b). Volatility Risk Premiums Embedded in Individual Equity Options: Some New Insights." Journal of Derivatives 11 (1), 45-54.

Bakshi, Gurdip, Kapadia, Nikunj, & Madan, Dilip. 2003. Stock return characteristics, skew laws, and the differential pricing of individual equity options. Review of Financial Studies, 16(1), 101–143.

Bakshi, G., Madan, D. and G. Panayotov, 2010, Returns of claims on the upside and the viability of U-shaped pricing kernels, Journal of Financial Economics 97, 130-154.

Banz, R.W., Miller, M.H., 1978. Prices for state-contingent claims: some estimates and applications. Journal of Business 51, 653–672.

Barone-Adesi, G., Engle, R. F., and Mancini, L. (2008) A GARCH option pricing model with filtered historical simulation, Review of Financial Studies 21, 1223–1258.

Barone-Adesi, G., Mancini, L. and H. Shefrin, 2012, Behavioral finance and the pricing kernel puzzle: estimating sentiment, risk aversion and time preference, working paper, Swiss Finance Institute.

Bates, D. (1996). "Jumps and Stochastic Volatility: Exchange Rate Process Implicit in Deutsche Mark Options." Review of Financial Studies 9, 69-107.

Bates, D.S., 2000. Post-'87 Crash fears in the S&P 500 futures option market. Journal of Econometrics 94, 181–238.

Bates, David S., 2003, “Empirical option pricing:  a retrospection,” Journal of Econometrics, 116, 387-404.

Benzoni, L., Collin-Dufresne, P., and Goldstein, R. S. (2011) Explaining asset pricing puzzles associated with the 1987 market crash, Journal of Financial Economics 101, 552–573.

Berkowitz, Jeremy, 2001, "Testing Density Forecasts, With Applications to Risk Management." Journal of Business & Economic Statistics 19, 465–474.

Bhattacharya, M. (1983). "Transactions data tests of efficiency of the Chicago Board Options Exchange." Journal of Financial Economics 12, (2), 161-185.

Birru, J. and S. Figlewski (2012).  "Anatomy of a Meltdown:  The Risk Neutral Density for the S&P 500 in the Fall of 2008."  Journal of Financial Markets 15, 151-180.

Black, F., (1976). "Studies of stock price volatility changes." Proceedings of the 1976 Meetings of the American Statistical Association, 177–181.

Black, Fischer and Myron S. Scholes (1972) The valuation of option contracts and a test of market efficiency, Journal of Finance, 27 (2), 399–418.

Black, F., and M. Scholes, 1973, 'The Pricing of Options and Corporate Liabilities," Journal of Political Economy. 81. 637-659.

Bliss, R. and N. Panigirtzoglou (2002).  "Testing the Stability of Implied Probability Density Functions."  Journal of Banking and Finance 26, 381-422.

Bliss, R. and N. Panigirtzoglou (2004).  "Option Implied Risk Aversion Estimates."  Journal of Finance 59, 407-446.

Bollerslev, T., 1987. A conditional heteroskedastic time series model for speculative prices and rates of return. Review of Economics and Statistics 69, 542–547.

Bollerslev, Tim, Tauchen, George, & Zhou, Hao. 2009. Expected Stock Returns and Variance Risk Premia.  Review of Financial Studies, 22(11), 4463–4492.

Bollerslev, Tim, & Todorov, Viktor. 2011. Tails, fears, and risk premia. The Journal of Finance, 66(6), 2165–2211.

Borovička, J., Hansen, L., Scheinkman, J., 2016. Misspecified recovery. Journal of Finance 71, 2493–2544.

Breeden, Douglas and Robert Litzenberger  (1978).  "Prices of State-Contingent Claims Implicit in Option Prices."  Journal of Business 51, 621-652.

Canina, Linda and Stephen Figlewski, 1993, “The Informational Content of Implied Volatility,” Review of Financial Studies, 6 (3), 659-81.

Carr, P., Geman, H., Madan, D., Yor, M. (2002). The fine structure of asset returns: an empirical investigation. Journal of Business 75, 305–332.

Chicago Board Options Exchange. (2003). VIX CBOE Volatility Index. http://www.cboe.com/micro/vix/vixwhite.pdf.

Chicago Board Options Exchange. (2009). CBOE S&P 500® Implied Correlation Index

Chicago Board Options Exchange., 2011. The CBOE skew index - SKEW. Technical Report.

Chabi-Yo, F. (2012). Pricing Kernels with Stochastic Skewness and Volatility Risk." Management Science 58, 624–640.

Christensen, Bent J., and Nagpurnanand Prabhala, 1998, “The Relation Between Implied and Realized Volatility,” Journal of Financial Economics, 50(2), 125-150.

Christoffersen, P., Heston, S., Jacobs, K.. (2009). "The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well." Management Science 55, 1914– 1932.

Christoffersen, P., S. Heston, and K. Jacobs. 2013. Capturing option anomalies with a variance-dependent pricing kernel. Review of Financial Studies 26:1962–2006.

Christoffersen, Peter, Kris Jacobs, and Bo-Young Chang, 2013, Forecasting with option-implied information, Handbook of Economic Forecasting 2, 581–656.

Coval, J. D. and Shumway, T. (2001). Expected option returns. The Journal of Finance, 56(3):983– 1009.

Corrado, C., and T. Su. 1996. “Skewness and Kurtosis in S&P 500 Index Returns Implied by Option Prices.” Journal of Financial Research, vol. 19, no. 2:175-192.

Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. Journal of Financial Economics 3, (1-2) 145-166.

Cox, J. C.; Ross, S. A.; Rubinstein, M. (1979). "Option pricing: A simplified approach". Journal of Financial Economics. 7 (3): 229.

Cuesdano, H., and J. Jackwerth, 2017, The pricing kernel puzzle: survey and outlook, working paper, University of Konstanz.

Dennis, P, & Mayhew, S. (2002). Risk-Neutral Skewness: Evidence from Stock Options. Journal of Financial and Quantitative Analysis, 37, 471–493.

Dennis, P., S. Mayhew and C. Stivers, 2006, Stock returns, implied volatility innovations, and the asymmetric volatility phenomenon, Journal of Financial and Quantitative Analysis 41, 381-406.

Derman, Emmanual, and Iraj Kani, 1994, “Riding on a Smile,” RISK, 7, 32-39.

Driessen, Joost, Pascal J. Maenhout, and Grigorty Vilkov, 2009, The price of correlation risk: evidence from equity options, Journal of Finance 64, 1377-1406.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. Econometrica, 68(6):1343–1376.

Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions: Empirical tests. The Journal of Finance, 53(6):2059–2106.

Dupire, B., 1994. Pricing with a smile. RISK 7, 18–20.

Emanuel, David C., and James D. MacBeth, 1982, Further results on the constant elasticity of variance call option pricing model, Journal of Financial and Quantitative Analysis 17 (4), 533-554

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

Engle, Robert, and Stephen Figlewski, 2015, Modeling the dynamics of correlations among implied volatilities, Review of Finance 19, 991-1018.

Feldman, Ron; Heinecke, Ken; Kocherlakota, Narayana; Schulhofer-Wohl, Sam; and Thomas Tallarini. 2015. “Market-Based Probabilities: A Tool For Policymakers,” Federal Reserve Bank of Minneapolis working paper.

Figlewski, Stephen. 2009. Estimating the implied risk neutral density. In:Bollerslev,T., Russell, J.R.,Watson,M.(Eds.),Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle.

Figlewski s and Freund S. 1994. The Pricing of Convexity Risk and Time Decay in Options Markets. Journal of Banking and Finance 18: 73-91.

Figlewski, S and Webb, G. 1993. Options, Short Sales, and Market Completeness. Journal of Finance 48 (2):761-777.

Fleming, J., 1998. The quality of market volatility forecasts implied by S&P 100 index option prices. Journal of Empirical Finance 5, 317–345.

Garleanu, N., Pedersen, L. H., Poteshman, A. M., 2009. Demand-based option pricing. Review of Financial Studies 22 (10), 4259-4299.

Gemmill, Gordon, and Apostolos Saflekos, 2000, “How useful are implied distributions? evidence from stock index options,” Journal of Derivatives, 7(3), 83-98.

Harrison, J. M. and D.M. Kreps (1979). "Martingales and arbitrage in multiperiod securities markets." Journal of Economic Theory 20, 381-408.

Hens, T. and C. Reichlin (2013). "Three Solutions to the Pricing Kernel Puzzle." Review of Finance,17, 1029-64.

Heston, S. “A Closed Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options.” Review of Financial Studies, 6 (2) (1993), pp. 327-343.

Hull, J., White, A., 1987. The pricing of options on assets with stochastic volatility. Journal of Finance 42, 281–300.

Israelov and Kelly (2017). Forecasting the Distribution of Option Returns. University of Chicago Working Paper.

Jackwerth, J.C. (1999). "Implied Binomial Trees: A Literature Review." Journal of Derivatives 7, 66-82.

Jackwerth, J.C. (2000). "Recovering Risk Aversion from Option Prices and Realized Returns." Review of Financial Studies 13, 433-451.

Jackwerth, J.C. (2004). Option-Implied Risk-Neutral Distributions and Risk Aversion. Charlotteville: Research Foundation of AIMR.

Jackwerth, J.C. and Mark Rubinstein (1996). "Recovering Probability Distributions from Option Prices." Journal of Finance 51, 1611-1631.

Jackwerth, J. and Menner, M (2017)._"Does the Ross Recovery theorem work empirically?" Working Paper. University of Konstanz.

MacBeth, J., and L. Merville (1980). "Tests of the Black-Scholes and Cox Call Option Valuation Models." Journal of Finance , Vol. 35, pp. 285-301.

Madan, D. and Milne, F. (1994) "Contingent claims valued and hedged by pricing and investing in a basis." Mathematical Finance, Vol. 4, No. 3, 223-245

Melick W.R. and C.P. Thomas (1997). "Recovering an Asset's Implied PDF from Option Prices: An Application to Crude Oil during the Gulf Crisis." Journal of Financial and Quantitative Analysis 32, 91-115.

Merton, R.C., 1976. Option pricing when underlying stock returns are discontinuous. Journal of Financial Economics 3, 125–144.

OptionMetrics (2008). Ivy DB file and Data Reference Manual Version 2.5.10. New York: OptionMetrics LLC.

Poon, S.-H., Granger, C., (2003). Forecasting volatility in financial markets: a review. Journal of Economic Literature 41, 478–539.

Rompolis, L.E. Tzavalis. (2007) “Retrieving Risk Neutral Densities based on Risk Neutral Moments through a Gram–Charlier Series Expansion.” Mathematical and Computer Modelling, Vol. 46, No. 1–2, pp. 225–234.

Rosenberg, J. and R. Engle, (2002). Empirical pricing kernels, Journal of Financial Economics 64, 341-372.

Ross, S., 2015, The recovery theorem, Journal of Finance 70, 615-648.

Rubinstein, M. (1985) "Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978." Journal of Finance 40, 455-480.

Rubinstein, M. (1994). "Implied Binomial Trees." Journal of Finance 49, 771-818.

Scott, L.O., 1987. Option pricing when the variance changes randomly: theory, estimation, and an application. Journal of Financial and Quantitative Analysis 22, 419–438.

Shefrin, H., 2008, A Behavioral Approach to Asset Pricing. Elsevier Academic Press, Boston, Second edition.

Shimko, David. 1993. "The Bounds of Probability." RISK 6, 33-37.

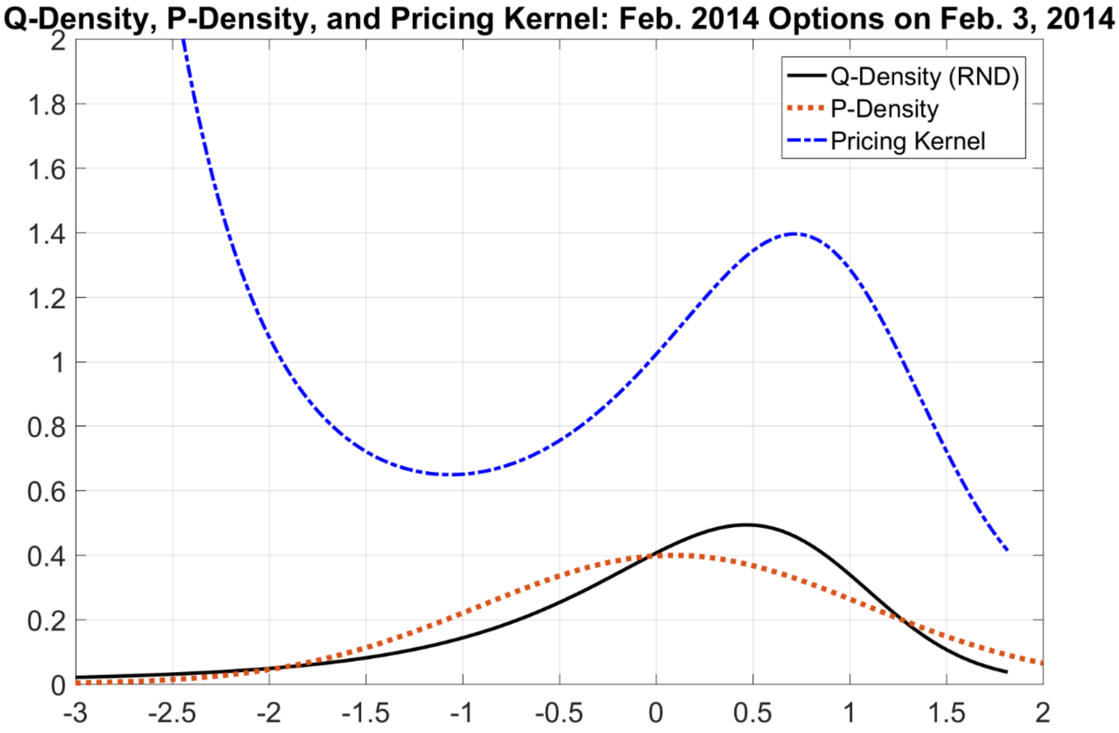
Stapleton, R. C., & Subrahmanyam, M. G. (1984). "The Valuation of Multivariate Contingent Claims in Discrete Time Models." Journal of Finance, 39, 207–228.

Stein, E.M., Stein, J.C., 1991. Stock price distributions with stochastic volatility: an analytic approach. Review of Financial Studies 4, 727–752.

Wiggins, J.B., 1987. Option values under stochastic volatility: theory and empirical estimates. Journal of Financial Economics 19, 351–377.

Xiu, D., 2014. Hermite polynomial based expansion of European option prices. Journal of Econometrics 179, 158–177.

# Figure 1



This Figure plots the risk neutral density (the Q-density) extracted from options on the SPDR Exchange-Traded Fund, frequently used as a tradable version of the S&P 500 index. The observation date was February 3, 2014 for options expiring on February 21, 2014 (solid line). The dotted line is an estimate of the P-density for Feb. 21. It is lognormal with annualized mean equal to the Feb. 3 riskless rate plus 5% risk premium and volatility equal to a weighted combination of the VIX index and 3 month historical volatility, where the weights were those which produced the most accurate forecasts of future realized volatility over an 18-day horizon in historical data up to that date. The pricing kernel is the dot-dash line, computed pointwise as the ratio of the Q-density to the P-density.