Probabilities of Future Equity Returns

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)
von der KIT-Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von
M.Sc. Simon Walther

Referent:	Prof. Dr. Maxim Ulrich
Korreferentin:	Prof. Dr. Marliese Uhrig-Homburg
Tag der mündlichen Prüfung:	13. Mai 2020

Karlsruhe, im Mai 2020

Acknowledgements

I would like to express my deep gratitude to my advisor Professor Dr. Maxim Ulrich for his constant support, countless inspiring discussions, valuable advice and for always encouraging me to explore ideas. Many thanks also to Professor Dr. Marliese Uhrig-Homburg, Professor Dr. Melanie Schienle and Professor Dr. Melanie Volkamer for serving on my thesis committee.

I also thank my co-authors Jonas Rothfuss and Fabio Ferreira, whom I highly enjoyed working with. Special thanks goes to all my colleagues at the Chair of Financial Economics and Risk Management, in particular Dr. Elmar Jakobs, Dr. Stephan Florig, Lingjie Ni, Lukas Zimmer, Ralph Seehuber and Tobias Kuhlmann for many helpful discussions, a constant supply of coffee and encouragement and for making the past years such an enjoyable time.

I gratefully acknowledge financial support via a scholarship by the Konrad-Adenauer-Stiftung.

Above all, I would like to thank my wife Anna for her love, her steady support and for cheering me up when the mood was low. I thank my parents for their unlimited support throughout my life.

Contents

List of Figures

List of Tables

I Introduction

I.1 Motivation

With a total volume of 67 trillion US-$, public equity markets represent one of the most important investment opportunities in the world.[1] Despite their popularity, equity valuations fluctuate significantly over time, exposing investors to a considerable amount of risk. A striking example for these risks materialized about 10 years ago, when the Great Recession of 2008 wiped out 46% of equity wealth in a single year[2] and forced many states to bail out their tumbling financial institutions. Mostly unforeseen crash events like this show that missing knowledge about the fundamental laws behind price moves and changes in risks are a central issue for practitioners and academics alike, leading to the construction and use of approximate models, that may eventually proof wrong or even harmful (see, e.g., Colander et al., 2009; Martin, 2017). At the same time, Lewis (2011) describes in his famous book "The Big Short", how several investors anticipated the crash and traded on their insights, without sharing their views with the broad public. Although such reports render it imaginable to forecast crashs in the equity market at least to some extent, the attempt to create a general model for the robust prediction of such events appears presumptuous. Still, there is reason to believe that *someone* is able to identify a looming, but mostly overlooked, risk and is trading on her believes. If so, these views should be incorporated in equity prices and even more in prices of equity derivatives (Chakravarty et al., 2004; Cremers et al., 2019).

This thought opens a route to an alternative approach to quantifying risks in equity markets: the model-free measurement of expectations that are incorporated in observed market prices. The advantage of the approach is clear; if there is no model, there is also no chance of a too simplistic representation of complex relationships or over-fitting of

---

[1] According to data from World Bank (2020) for the end of 2018.

[2] See World Bank (2020).

spurious characteristics in the data. But how can one extract market expectations from observed prices only, without making strong assumptions? Breeden and Litzenberger (1978) showed that the risk-neutral density of equity returns can be recovered model-free from a panel of equity option prices, if markets are arbitrage-free. Unfortunately, a sufficient amount of option prices for the precise measurement of the risk-neutral density is often not observed and noise in the observed prices is relatively high. These characteristics necessitate the use of a smoothing and interpolation technique to obtain artificial option prices for a continuum of strike prices, that are in line with the observed market situation. Multiple of such techniques have been proposed (see, e.g., Jackwerth, 1999; Figlewski, 2008; OptionMetrics, 2016) and are heavily used in financial research today. Despite their workhorse status, no detailed assessment and comparison of the most popular techniques for recovering risk-neutral expectations has been performed so far. This thesis closes the gap in the literature, by uncovering that option-implied measures are sensitive to the employed smoothing and interpolation technique and providing guidance about which technique represents option markets most accurately.

While the model-free extraction of risk-neutral expectations has become a standard in the financial literature, a model-free estimation of physical expectations has long not been feasible. At the same time, the physical distribution of equity returns may be of even greater interest for risk and asset managers than the risk-neutral distribution, as it allows to quantify the probability of real-world events and extract measures like the value-at-risk or the expected shortfall. While studies like Aït-Sahalia and Lo (2000) estimate the unconditional physical distribution of equity returns model-free using a kernel density estimator, the same has not been possible for time-varying density estimates as the estimator requires a high amount of data, but only a single draw from each day's conditional physical density is observed. Moreover, in the presence of regime shifts and limited information, relying on past returns to predict the physical return density can lead to biases (Linn et al., 2018; Jackwerth and Menner, 2018). This problem can only be overcome by using *forward-looking* estimators of the physical return distribution, i.e., estimators that base their predictions on market expectations only, not on observations of the past. The most popular input to such estimators are again risk-neutral expectations. With all their theoretical appeal, the proposed forward-looking estimators in the literature come at a cost: They require strong assumptions about the pricing kernel to facilitate the translation of risk-neutral into physical expectations, thus losing their model-free character. In this thesis, we present a forward-looking estimator of the time-varying physical distribution of equity returns that is agnostic about the pricing kernel and relies only on minimal exogenous constraints. It can thus be considered model-free while still maintaining full flexibility.

The wedge between the risk-neutral and the physical density is the pricing kernel (Cochrane, 2001). Based on a firm theoretical foundation, the pricing kernel is a central building block for academic research. Due to its tight connection to investors' preferences, knowledge about the empirical properties of the pricing kernel directly helps to improve our understanding of financial markets. In a seminal study, Jackwerth (2000) recovers the pricing kernel empirically and describes local increases in the pricing kernel, a result that was at odds with most asset pricing models at the time.[3] With the advent of forward-looking estimators for the physical equity return distribution, a debate re-emerged about whether these local increases actually exist or had just been a result of biases in the backward-looking estimators for the physical return density (see, e.g., Linn et al., 2018; Cuesdeanu and Jackwerth, 2018a). Equipped with our model-free estimators of the risk-neutral and physical distribution of equity returns, we are able to recover model-free pricing kernels. Our pricing kernel estimates are both forward-looking and time-varying, which has not been possible before. We use our measurements to extract a number of insights about key properties of the pricing kernel.

I.2 Structure of the Thesis

In this thesis, we develop and present a set of tools that may proof helpful for both practitioners and academics. The structure of the thesis follows the logical grouping of these tools. Chapter II covers the precise measurement of the risk-neutral density. We present and examine our forward-looking estimator for the physical density of equity returns in chapter III. Chapter IV then analyses their ratio, the pricing kernel. In more detail, this thesis is structured as follows:

In chapter II, which is based on the working paper Walther and Ulrich (2019), we inspect the sensitivity of multiple option-implied measures to the method that is used in constructing the implied volatility surface. Different state-of-the-art and widely used approaches lead to economically surprisingly large differences in these measures and some methods lead to systematic biases, especially for out-of-the-money Put options. To over-

---

[3] Many classical asset pricing models imply that the pricing kernel decreases monotonically in equity returns, see for example Mehra and Prescott (1985); Bansal and Yaron (2004); Cochrane (2001).

come this problem we propose two new volatility surfaces, one for end-of-day and one for intraday applications. Our end-of-day method builds upon a one-dimensional kernel regression, while our intraday method employs a Bayesian approach to filter stable and accurate volatility smiles from tick-by-tick option trade observations. We assess the statistical accuracy of our proposed methods relative to existing state-of-the-art parametric, semi- and non-parametric volatility surfaces by means of leave-one-out cross-validation. Based on 14 years of end-of-day and intraday S&P 500 and Euro Stoxx 50 option data, we show that our methods represent option market information more accurately than existing approaches of the literature.

Having established accurate measurements for the risk-neutral distribution, we next turn to the estimation of the physical distribution of equity returns in chapter III, which is based on the working paper Walther et al. (2019). In this chapter, we present a forward-looking estimator for the time-varying physical return distribution with minimal prior restrictions of the shape of the distribution and no exogenous assumptions about the economy or preferences. Our estimator, which is based on a neural network, derives its forecasts from the option-implied moments of the most accurate volatility surface from chapter II and predicts the conditional mean and volatility of returns such that profitable trading strategies can be derived. In contrast to backward-looking estimators and alternative forward-looking parametric and non-parametric approaches, its distribution forecasts cannot be rejected in statistical tests and they feature lower prediction errors and higher conditional log likelihood values than the alternatives. By deliberately varying the input variables of our estimator, we uncover nonlinear relationships between physical and risk-neutral moments, which appear necessary to accurately capture changes in the physical distribution of equity returns.

Chapter IV then combines our estimators for the risk-neutral and the physical distribution from the previous chapters to obtain daily measures for the pricing kernel at the monthly time horizon. Despite their time-varying nature, our pricing kernels are non-parametric, forward-looking, agnostic about preferences, economic state variables or their dynamics and rely only on minimal technical constraints. Still, our realized pricing kernel estimates are clearly linked to economic state variables like the term spread, the credit spread or liquidity.[4] We decompose the expected variance of the log pricing kernel and find

---

[4] For example, Ang (2014, ch. 9) shows that the term spread reliably predicted economic activity in the past, Gilchrist and Zakrajšek (2012) provides similar evidence for credit spreads and Næs et al. (2011) identifies stock market liquidity as a "leading indicator of the real economy".

that jumps contribute a considerable portion to overall pricing kernel risk. Building on statistical tests, we identify a strong U-shape in the pricing kernel as a function of equity returns, which never ceases to exist in our sample. Overall, we find clear signs of time variation in the pricing kernel, both with respect to its shape and volatility.

Finally, we recapitulate the main results of this thesis in chapter V and briefly discuss ideas for future research.

II The Risk-Neutral Return Distribution

Many popular option-implied metrics such as the variance or skewness of the risk-neutral distribution (RND) and the variance risk premium are calculated based on an estimate of the option-implied volatility surface. We document here that the method for constructing the volatility surface affects these standard option-implied quantities. Our findings hold more generally for any quantity that is extracted from the aggregation of option prices along the strike range.

State-of-the-art methodologies such as the semi-parametric spline interpolation[1] of Figlewski (2008) or the three-dimensional kernel regression of OptionMetrics (2016) produce surprisingly large differences in standard option-implied quantities. In our sample for S&P 500 options (2004 - 2017), Bakshi et al. (2003) risk-neutral variance for medium-term maturities, computed with exactly the same procedure, but based on the volatility surface from the interpolation scheme of Figlewski (2008) or OptionMetrics (2016), differs in relative terms by more than 10% on average. The one-month-ahead variance risk premium in S&P 500 options varies across both volatility surfaces by a relative margin of on average 60%. Differences are even more troublesome for risk-neutral skewness, where we document relative differences in the order of 200% and more.

The key question to ask is which volatility surface represents market information most accurately? As information is extracted from option prices by means of deterministic manipulations of the observed portions of the volatility surface, it is natural that it is the most accurate volatility surface that also reprices options most accurately. We therefore perform a detailed empirical investigation to understand which volatility surface cap-

---

[1] The spline interpolation represents an implied volatility smile parametrically. However, there is no explicit parametric form of the risk-neutral density that the volatility smile implies. For this reason, we consider this methodology, and more broadly the class of parametric implied volatility models that do not allow to pin down the RND parametrically, as semi-parametric.

tures market information most accurately. Our test incorporates the semi-parametric spline interpolation (Figlewski, 2008), a three-dimensional non-parametric kernel regression (OptionMetrics, 2016), and the parametric Gram-Charlier expansion (Beber and Brandt, 2006). In addition, we also propose a one-dimensional non-parametric kernel regression method and a Bayesian filtering model for capturing the intraday movements of the implied volatility surface. We compare the statistical accuracy of these volatility surfaces by means of leave-one-out cross-validation root mean squared errors (RMSEs) and mean absolute errors (MAEs). Calculating the average integrated squared second derivative of the respective implied volatility smiles allows us to identify differences in their smoothness. Our tests expand across two dimensions: (i) options on the S&P 500 and on the Euro Stoxx 50 and (ii) with an end-of-day and intraday frequency. The time span of the analysis is 2004 to 2017 for US data and 2002 to 2017 for European data.

Our main findings are as follows: First, the one-dimensional kernel regression generates the most accurate end-of-day volatility surface for S&P 500 options by means of the lowest leave-one-out cross-validation RMSE and MAE. For the case of intraday transaction data, the Bayesian filtering model appears to represent the unobserved and by single trades only scarcely measured volatility surface best.

Second, for the end-of-day analysis of S&P500 options, the spline-based volatility surface turns out to be the second best, with a RMSE (MAE) that is on average 128% (91%) higher than the RMSE (MAE) of the one-dimensional kernel regression surface. The three-dimensional kernel regression produces a RMSE that is more than 5 times larger than the RMSE of the one-dimensional kernel regression and an 11 times higher MAE. The Gram-Charlier volatility surface produces the largest RMSEs (MAEs), on average over 9 (23) times larger than the RMSE (MAE) of the best performing volatility surface. The results for Euro Stoxx 50 options confirm our general findings, though the one-dimensional kernel regression and the spline interpolation appear to perform roughly at par here.

Third, state-of-the-art volatility surfaces turn out to be less accurate than the one-dimensional kernel regression volatility surface because these surfaces do not accurately capture market information in the thinly traded out-of-the-money regions of the volatility surface, which correspond to the tails of the RND. The volatility surface based on spline interpolation shows weakness in capturing the left tail of short-term options. The three-dimensional kernel regression shows severe shortcomings in capturing the left tail of short-, medium- and long-term options. The method of the Gram Charlier expansion partially captures the at-the-money region of volatility smiles well, but shows otherwise shortcomings in capturing the left and right tail of options across all maturities.

Fourth, for the intraday analysis we find that our proposed Bayesian volatility surface is the most accurate by a large margin. This finding holds for all considered error measures and assets, even if the alternative methods are equipped with a higher amount of trade observations as inputs. Among the non-Bayesian methods, the one-dimensional kernel regression performs best, if data is very scarce. However, the indication about the most accurate method among the non-Bayesian approaches depends on the number of trades that are used to construct the implied volatility surface and diverges between the error measures. While the one-dimensional kernel regression continues to lead in most set-ups with respect to the RMSE, the spline interpolation quickly shows lower MAEs than all other non-Bayesian methods as the number of utilized input observations grows. This finding can be rationalized by occasional over-fitting in the spline method, which occurs less often if more observations are available for fitting the spline.

Our analysis concludes that option-implied information can differ substantially across volatility surfaces. Despite being widely used in the literature, the three-dimensional kernel regression method of OptionMetrics (2016) underpredicts option-implied tail risk at the end-of-day data frequency which translates into systematic biases in risk-neutral skewness and variance. Instead, our one-dimensional kernel regression appears to produce the most accurate volatility surface for the end-of-day use-case. For intraday set-ups, our proposed Bayesian volatility surface appears to be the most accurate in all of our tests.

Our research study adds to the growing empirical finance literature that exploits option-implied information. By now, this literature is too vast to be reviewed here in detail. Hence, we cannot give credit to all studies and have to leave out important contributions. However, we discuss a selection of recent studies and focus on how these have constructed the option-implied volatility surface.

A large and diverse amount of research studies work with the volatility surface of OptionMetrics (2016): For example, Buss and Vilkov (2012) estimate option-implied correlations and CAPM-betas and find that higher option-implied betas go along with higher average returns. Chang et al. (2012) predict a stock's beta based on Bakshi et al. (2003) implied volatility and skewness estimates. Martin and Wagner (2019) extract a measure for predicting the risk premium for each stock from its associated implied volatility surface.

Christoffersen et al. (2017) propose a parametric option pricing model where illiquidity is a driver of the jump intensity of the underlying and can thus explain time variation in the implied volatility surface. Du and Kapadia (2012) compare the information content of the VIX and Bakshi et al. (2003) implied variance and construct a tail risk index from the difference between the two measures. Hofmann and Uhrig-Homburg (2018) use the wedge between implied volatility observations and respective OptionMetrics (2016) estimates to construct a limits of arbitrage measure.

Driessen et al. (2009) compare portfolios of single stock options with options on the index and find evidence for a substantial option-implied correlation risk premium. Bollerslev and Todorov (2011) introduce various risk measures for realized and expected continuous risk and jump risk under the physical and risk-neutral probability measure and find that a significant portion of the equity risk premium is compensation for jump risk. The option-implied volatility surface in these innovative studies is constructed based on end-of-day closing prices and with a version of the spline interpolation methodology that we use in this paper.

Martin (2017) shows that options contain information about the lower bound of the underlying's expected return. Schneider and Trojani (2015) construct tradable option-implied strategies for higher moments. These studies do not interpolate the option-implied volatility surface, but work with observed option-implied volatilities.

Wright (2016) adopts the multi-dimensional kernel regression of Aït-Sahalia and Lo (1998) to construct a monthly option-implied volatility surface of real interest rates. The author pools all end-of-day volatilities (roughly 25 per day and maturity) within one month to stabilize the procedure and obtain a sufficiently smooth volatility surface. Swanson (2016) highlights that pooling prevents that study from working at a higher frequency. Moreover, Swanson (2016) suggests to use the spline methodology to construct daily implied volatility surfaces. As evidence in Bliss and Panigirtzoglou (2002) suggests, the spline method works well for 10 or more option prices per day and maturity. In order to stay in a non-parametric framework while still obtaining smoothness, Jackwerth and Rubinstein (1996) and Jackwerth (2000) propose a method that directly fits the volatility surface to observed data by minimizing the squared fitting error while at the same time maximizing the smoothness of the surface. Our research study contributes to this discussion, as we formally compare both the statistical accuracy and smoothness of several state-of-the-art volatility surfaces across different frequencies and different currency zones. For our tests, we select two kernel regression approaches as representatives for the non-parametric class of volatility surfaces. These methods do not explicitly optimize for smoothness, thus enabling us to perform a fair evaluation, since neither the kernel regression of OptionMetrics (2016) nor the spline method of Figlewski (2008) or the Gram-Charlier expansion features an explicit smoothness treatment. As we document, especially the one-dimensional kernel regression is still well capable of producing smooth volatility surfaces.

A large body of the literature studies parametric representations of the implied volatility surface. An early example is Schönbucher (1999), who assumes a diffusion process for the evolution of implied volatility.[2] He derives restrictions on the parameters to ensure arbitrage-freeness and to prevent the potential model-implied emergence of bubbles in implied volatility as the maturity decreases. The latter leads to a constraint that can essentially serve as a model for the volatility smile at a given maturity. Based on a two-dimensional diffusion for the forward price of an asset and its volatility, Hagan et al. (2002) derive a closed-form solution for a parametric volatility smile in their $SABR$ model. The model is completely specified by four parameters, which essentially describe the level, skew and curvedness of the volatility smile. However, as Gatheral (2006) points out, the lack of a mean reversion component in the volatility diffusion makes it only applicable to short-maturity options. In his SVI model, Gatheral (2004) assumes a parametric function for the volatility smile that is similar to the representation of Schönbucher (1999), but features an additional parameter to locate the volatility smile across the strike range. By construction, the SVI model assumes that the implied volatility smile becomes approximately linear in the tails. This assumption is not always fulfilled in the data, which has triggered the development of generalized versions of the SVI model that allow concavity, for example Zhao and Hodges (2013) or Damghani and Kos (2013). Damghani (2015) further elaborates on this work in the context of the FX option market. His IVP model features an explicit treatment of the bid-ask spreads in option prices and allows to incorporate liquiditiy factors. By adding a maturity interpolation scheme, he is able to capture the whole volatility surface with a dramatic reduction in parameters that need to be estimated. Figlewski (2008) follows a different approach: He argues to use a 4th-order smoothing spline with one knot point at-the-money to model the volatility smile at each maturity separately. The higher amount of free parameters is accepted to reach higher

---

[2] Assuming a diffusion process for implied volatility is different from a diffusion process for the volatility of the underlying, as in Heston (1993). In the model of Schönbucher (1999), movements in implied volatility are correlated with movements in the underlying, though explicitly feature components that are independent of the underlying asset.

accuracy while still smoothing out noise. In our study, we use the method of Figlewski (2008) as representative of the class of parametric models of the implied volatility surface.

Survey and methodology papers that are closest to ours are Jackwerth (1999), Jondeau and Rockinger (2000), Bliss and Panigirtzoglou (2002) and Bahaludin and Abdullah (2017). Relative to these studies, our empirical assessment covers a much longer time span (14 years), US and EU equity option markets and distinguishes between characteristics of the option data sets.

II.1 Relationship between the RND and the Volatility Surface

At a given maturity, the implied volatility smile of a set of options on some underlying is linked tightly to the risk-neutral probability density function of the returns of that underlying, such that knowing one allows to infer the other. As Breeden and Litzenberger (1978) show, at time t, the τ-maturity RND $q_t(K,\tau)$ for a potential price in $t+\tau$ of K can be calculated as the second derivative of (European-style) Call prices $C_t(K,\tau)$ with respect to their strike:

$$q_t(K,\tau) = \left.\frac{\partial^2 C_t(x,\tau)}{\partial x^2}\right|_{x=K} = \lim_{\Delta K \to 0} \frac{C_t(K-\Delta K,\tau) - 2C_t(K,\tau) + C_t(K+\Delta K,\tau)}{(\Delta K)^2}. \tag{II.1}$$

Given a dense set of Call price observations along the strike range, the right-hand side of equation II.1 can be calculated. However in most cases, the traded grid of strikes is not dense enough for the approximation error of the limit in equation II.1 to become negligible. The solution is to enrich the set of observed option prices by interpolation. Such an interpolation can be performed on Call prices directly, however, as Jackwerth (1999) points out, interpolating implied volatility is preferable due to numerical stability.

Implied volatility is a transformation of the option price, that is obtained by inverting an option pricing model and setting the volatility parameter of that model such that it matches the observed option price perfectly. It is important to note that the pricing model does not need to be correct. It is merely used as a bijective mapping between the space of implied volatilities and the space of option prices. We use the model of Black and Scholes (1973) to map implied volatilities to observed option prices for options of European exercise style. For American-style options, the early exercise premium of these contracts needs to be taken into account.[3] Here, we recommend to distinguish between two cases: Many American-style options are written on Future contracts as the underlying or on underlyings that pay an approximately continuous dividend yield. For these options, the model of Barone-Adesi and Whaley (1987) provides a fast and precise approximation for the American option price and the early exercise premium. Options on underlyings that have discrete payoffs at specific points of time during the lifetime of the option constitute the second case. Options on dividend-paying single stocks belong to this category. Here, the binomial tree model of Cox et al. (1979) allows to consider the timing of the single payments and adjusts the early exercise premium appropriately.

The previous discussion pointed out a route to translate implied volatilities to the RND: First, implied volatilities are converted to European option prices via the Black and Scholes (1973) and then the second derivative of the price with respect to the option's strike is approximated numerically. At the same time, the reverse transformation is feasible, too. The price of an option is the discounted expected payoff under the risk-neutral measure (Cochrane, 2001, ch. 17):

$$C_t(K,\tau) = e^{-r_{f,t}\tau} E_t^{\mathcal{Q}}(\max(S_{t+\tau} - K, 0)) = e^{-r_{f,t}\tau} \int_0^\infty \max(x - K, 0)\, q_t(x,\tau)\, dx. \quad \text{(II.2)}$$

Here, $r_{f,t}$ denotes the risk-free rate in t, $S_{t+\tau}$ is the underlying price in $t+\tau$ and $E_t^{\mathcal{Q}}(\cdot)$ represents the expectation under the risk-neutral measure $\mathcal{Q}$, conditional on all information that is available at time t. Equation II.2 shows the pricing formula for a Call option, pricing a Put option is straight-forward by replacing the payoff with the respective counter-part for Puts.

Given estimates for the RND $q_t(x,\tau)$ at a dense grid of x, the integral in equation II.2 can be approximated numerically with high precision (Atkinson, 1989, ch. 5). A translation of the option price to implied volatility can then be made by inverting the Black and Scholes (1973) model.

---

[3] Since the option pricing model need not to be correct to map implied volatilities to observed option prices, one could also just use the Black and Scholes (1973) model for American-style options, too. However, several second-tier applications, like the calculation of Bakshi et al. (2003) model-free implied volatility requires a panel of *European-style* option prices, making it necessary to eliminate the early exercise premium from the observed American-style option prices.

II.2 Constructing a Volatility Surface

II.2.1 Estimation of Expected Dividend Yields

Many option contracts are written on underlyings that pay out dividends or coupons, or that experience storage costs before the expiration of the option contract. If such pre-expiration payoffs exist[4], they must be considered in the option's price, as such payoffs predictably affect the price of the underlying during the lifetime of the option.[5] For this study, both the S&P 500 and the Euro Stoxx 50 index, are price indices and are thus quoted ex-dividends. More generally, dividend payments need to be taken into account for all options on equity underlyings, since filtering out invalid option prices from our data set (see section II.3) and calculating implied volatilities requires an estimate of the expected dividend payment under the risk-neutral measure. For options with European exercise style, we determine the maturity-specific expected dividend yield via the put-call parity. For American exercise style options on single stocks, we need to take care of the discreteness of the dividend payment, its timing and how it affects the early exercise premium of the option. In the following paragraphs, we will provide a detailed description of our estimation procedure of risk-neutral dividend expectations based on observed option prices. These estimations are based on the full option data sample, before applying any filters.

European-style Options We make use of the put-call parity to obtain daily model-free estimates of the expected dividend yield for an option's underlying. More precisely, let C be the Call price and P be the Put price of 2 options with maturity τ on the same underlying with price S and forward price F. Both options have the same strike K. Let the risk-free rate be r and the dividend yield be q. Following Hull (2018), we can express the put-call parity in the following equation:

$$C - P = e^{-r\tau}(F - K) = e^{-q\tau}S - e^{-r\tau}K. \tag{II.3}$$

---

[4] Costs can be seen as negative payoffs.

[5] Some options, for example the fixed income options at the Eurex, are written on Futures as underlying. Here, no payoff adjustment is necessary, as any pre-expiration payoff is already reflected in the price of the Future. For these options, the "payoff yield" does not need to be estimated and is set to zero.

Solving this equation for the dividend yield q yields

$$q = \frac{1}{\tau}\left[ln(S) - ln(C - P + e^{-r\tau}K)\right]. \tag{II.4}$$

In order to estimate the dividend yield with this equation, one needs to find a Put and a Call price with the same strike and maturity at the same point of time. In our end-of-day data sets, settlement or bid/ask prices for a high amount of strikes and maturities for both Puts and Calls are available. As all end-of-day option prices are recorded at the same point of time within the day, the price of the underlying coincides for all price quotes. For each pair of options at a specific date, strike and maturity, we calculate the implied dividend yield using equation II.4 and take the median over the strike range to arrive at a single estimate of the dividend yield per date and maturity. If the estimate turns out to be negative, we replace it with 0.

The situation is more complex for intraday transaction data, since quoted trades do not occur simultaneously. For this reason, there will always be a time differential when matching a Put and a Call with the same strike and maturity for calculating the dividend yield via the put-call parity. However, the time differential might be large enough for the price or the risk-neutral volatility expectation of the underlying to change. In that case, the put-call parity does not hold any more, even if the expected dividend yield remains constant. When matching Put and Call prices in the intraday set-up, we therefore impose the constraints that the price quotes are from the same day and that the price of the underlying has changed less than 0.01%. If multiple pairs of a Put and a Call fulfill these constraints, we choose the ones with the smallest time differentials. We assume that the risk-neutral volatility expectations for such pairs of option contracts differ only negligibly. Again, we take the median over the expected dividend yield estimates of all Put-Call pairs for a certain date and maturity to arrive at a single dividend yield estimate.

American-style Options The estimation of option-implied dividend yields is more complex for American-style options, since the put-call parity does not hold for these options. However, the price of an American Call option C^A and an American Put option P^A with the same strike and maturity must fulfill the put-call inequality (Hull, 2018),

$$S - K - D \leq C^A - P^A \leq S - Ke^{-r\tau} = \hat{D}^{lb}, \tag{II.5}$$

where D denotes the present value of the risk-neutral dividend expectation.[6] The put-call inequality allows to extract a lower bound for the dividend yield:

$$D \geq S - K - C^A + P^A. \tag{II.6}$$

By expressing the American option prices C^A and P^A as European option prices plus a respective early exercise premium, $C^A = C + EEP^{call}$ and $P^A = P + EEP^{put}$, the lower bound can be restated as

$$\hat{D}^{lb} = S - Ke^{-r\tau} - C + P + (EEP^{put} - EEP^{call} - K(1 - e^{-r\tau})). \tag{II.7}$$

We know from the put-call parity for European-style options that $D = S - Ke^{-r\tau} - C + P$, so the error of the lower-bound dividend estimate is $\hat{D}^{lb} - D = EEP^{put} - EEP^{call} - K(1 - e^{-r\tau})$. It might well be that one makes a smaller estimation error by simply using American option prices in the put-call parity to estimate dividend yields. In that case, we can write

$$\hat{D}^{parity} = S - Ke^{-r\tau} - C^A + P^A = S - Ke^{-r\tau} - C + P + EEP^{put} - EEP^{call} \tag{II.8}$$

and the error between the put-call parity dividend estimate and the true risk-neutral dividend expectation is given by $\hat{D}^{parity} - D = EEP^{put} - EEP^{call}$.

In order to decide on a method for the expected dividend estimation for American-style options, we analyze the method-specific errors for a panel of artificial option contracts. Given the characteristics of such an artificial option contract, we use the binomial tree pricing model of Cox et al. (1979) to calculate the American Put and Call prices for a grid of strikes $K \in \{0.8S, 0.825S, 0.85S, ..., 1.175S, 1.2S\}$. We then use the lower-bound method and put-call parity to estimate the option-implied dividend for the Put-Call pair of each strike. Similar to the dividend yield estimation for European-style options, $\hat{D}^{lb}$ and $\hat{D}^{parity}$ are then obtained by taking the median expected dividend estimate over all strikes. Finally, we calculate a third estimate, $\hat{D}^{partial}$, which also uses the put-call parity, but aggregates only over option contracts with a strike that fulfills $1 \leq \frac{K}{S} \leq 1.1$. It is important to keep in mind that we know the true dividend expectation of our artificial options, since we set it as a parameter of the option price. We can hence calculate

---

[6] The following relationship holds between the dividend yield and the present value of expected dividends: $D = S(1 - e^{-q\tau})$

Table II.1: Option-Implied Dividend Expectations: Estimation Errors by Method

S	τ	r	σ	D	τ^D	lb	parity	partial
100	30	0.02	0.25	$0.01S$	0.5τ	-27.7	-11.3	1.1
100	30	0.02	0.25	0.0	-	-0.140	0.010	0.028
100	30	0.02	0.25	$0.1S$	0.5τ	-19.5	-17.8	-3.6
100	365	0.02	0.25	$0.01S$	0.5τ	-182.5	17.5	25.0
100	365	0.02	0.25	$0.04S$	0.5τ	-57.4	-7.4	-0.7
100	365	0.02	0.25	$0.1S$	0.5τ	-39.4	-19.4	-11.6
100	30	0.02	0.25	$0.01S$	0.1τ	-15.3	1.2	3.3
100	30	0.02	0.25	$0.01S$	0.9τ	-54.4	-38.0	-14.9
100	30	0.06	0.25	$0.01S$	0.1τ	-54.7	-5.4	9.6
100	30	0.02	0.5	$0.01S$	0.1τ	-24.1	-7.7	-1.7

The table shows estimation errors for expected dividends based on artificial American-style options. The first six columns show the parameters that characterize each artificial option contract. The estimation methods are the lower-bound ('lb') method, the put-call parity ('parity') method and the put-call parity method with aggregation only over options with a moneyness between 1 and 1.1 ('partial'). τ^D specifies the point of time during the lifetime of the option of the discrete dividend payment D. Errors are given in percent of the actual dividend payment, except for the case of no dividend payment, where we report the absolute deviation from the correct dividend estimate.

estimation errors for our dividend estimates of the true dividend expectation. Table II.1 shows the estimation error $\frac{\hat{D}^x-D}{D}, x \in \{lb, parity, partial\}$ for each method and a panel of artificial option contracts.

Not surprisingly, the lower-bound method always under-estimates the true dividend expectation. The extent of this under-estimation can be quite large: The method frequently only estimates half of the correct dividend expectations and one expected dividend estimate is even negative, causing an estimation error below -100%. The put-call parity and the partial put-call parity estimates are generally closer to the true expected dividend in our tests. With the partial put-call parity estimate showing a lower absolute error in 6 out of 10 test cases and an average absolute estimation error of 7.9% compared to 14.0% for the full put-call parity estimate, we recommend using this approach for estimating option-implied dividends for American-style options in practice.

At the point of time of an option trade, the actual payment date of a future dividend is generally unknown. The timing of the dividend can however affect the early exercise premium of an American-style option. It is therefore necessary to estimate the date of the dividend payment during the lifetime of the option contract. A dividend-paying company usually follows a certain dividend payout frequency. For example, if a stock is currently paying a dividend every half year, we consider it to follow a semi-annual payment regime. For dividend payment dates that have not been announced yet, we recommend the following payment-regime-dependent set of rules to construct expected dividend payment dates:

- **Annual Dividend**: Use the same month and date as the last observed dividend payment.
- **Semi-annual and Quarterly Dividend**: For each dividend payment in a year, compute the average day of the year of the dividend payment over the last 3 years. For each year in the future, adjust that day to match the weekday of the last dividend payment, such that the deviation from the respective average day of the year is minimal.
- **Monthly Dividend**: Use the day of the month of the last observed dividend payment.
- **Other Dividend Frequency**: For each dividend payment in a year, compute the average day of the year of the dividend payment over the last 3 years. Use that day as projected dividend payment day in the future.
- **Monthly, Annual and Other Frequencies**: If the projected payment date is on a weekend, use the closest trading day. If the projected payment date is a non-trading holiday, use the previous trading day.

II.2.2 One-Dimensional Kernel Regression with Tail Extrapolation

We now continue to propose a kernel regression specification for interpolating the implied volatility surface. Our methodology features a number of favorable characteristics. First, it is a non-parametric approach, which ensures flexibility. Second, it is essentially only one-dimensional, which makes it more robust and less data-intensive than existing multi-dimensional kernel regression approaches. Third, our approach ensures the volatility surface to be arbitrage-free, borrowing a technique from Fengler (2009). Fourth, conditional on having sufficient data, our approach takes special care of capturing market information that is hidden in the thinly traded tails.

Aït-Sahalia and Lo (1998) propose to approximate unobserved option-implied volatilities by the following kernel regression set-up:

$$\hat{\sigma}(F_j, K_j, \tau_j) = \sum_{i=1}^{N} \frac{k(F_j - F_i)\, k(K_j - K_i)\, k(\tau_j - \tau_i)}{\sum_{i=1}^{N} k(F_j - F_i)\, k(K_j - K_i)\, k(\tau_j - \tau_i)}\, \sigma(F_i, K_i, \tau_i), \tag{II.9}$$

where N is the number of observed option prices that enter the kernel regression as inputs, $\sigma(.)$ is the observed implied volatility and $\hat{\sigma}$ is the interpolated implied volatility for a desired tuple of strike K_j, underlying forward price F_j and maturity τ_j. The kernel function $k(x)$ has a Gaussian shape with its own bandwidth parameter h_x for each dimension, i.e.,

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2h_x}\right)}. \tag{II.10}$$

We apply several refinements that distinguish our approach from the tested application in Aït-Sahalia and Lo (1998). First, we reduce the input dimension by one unit, as we combine the underlying forward price with the strike level. More precisely, we define the observed moneyness measure as $m_i = \frac{K_i}{F_i}$ and apply the kernel regression to $m_j - m_i$ instead of $F_j - F_i$ and $K_j - K_i$.[7] In most cases, this step can be regarded as a mere technical re-scaling of the strike axis, which does not affect the interpolation accuracy of the technique. In our end-of-day set-up, we only use the price observations of a single day to construct the respective day's implied volatility surface. For a given maturity, these price observations all have the same underlying forward price as they are observed at the same point of time. The denominator of the ratio $\frac{K_i}{F_i}$ is therefore constant. Similarly, in intraday set-ups, we construct the implied volatility surface based on a small number of subsequent option trade observations. In the vast majority of cases, the change in the underlying price during the time between these option trades is very small, such that the underlying forward price can be considered approximately constant. Here, the switch in the kernel regression to a moneyness axis can hence be seen as a re-scaling of the strike dimension, too.

The second refinement of the Aït-Sahalia and Lo (1998) approach is the removal of the time to maturity dimension, τ, from the kernel regression. This means that the kernel regression is purely used to approximate the option-implied volatility surface in the moneyness dimension. Mathematically, we arrive at the following kernel weighting

---

[7] Although this adjustment has already been proposed by Aït-Sahalia and Lo (1998), their subsequent empirical analysis works with K and F, separately.

for the moneyness dimension:

$$\hat{\sigma}(m_j, \tau) = \sum_{i=1}^{N_\tau} \frac{k(m_j - m_i)}{\sum_{i=1}^{N_\tau} k(m_j - m_i)} \, \sigma(m_i, \tau), \qquad m_i = \frac{K_i}{F_i}, \tag{II.11}$$

where N_τ denotes the size of the set of observed option prices with maturity τ.[8] We set the bandwidth parameter for the moneyness dimension by the following rule-of-thumb: We first compute the average distance between two neighboring strikes

$$\overline{\Delta K}(\tau) = \frac{1}{N_\tau - 1}(K^{max}(\tau) - K^{min}(\tau)), \tag{II.13}$$

where $K^{max}(\tau)$ ($K^{min}(\tau)$) represents the maximum (minimum) observed strike at maturity τ. We now set the moneyness bandwidth to $h_m(\tau) = 0.75 \, \overline{\Delta K}(\tau)$. While our choice of the coefficient of 0.75 may appear ad-hoc at first, our tests reveal that the results barely change if it is set at any value within a range of $[0.6, 1]$.

Third, far out-of-the money options are usually not observed, but very important ingredients for capturing the tails of the risk-neutral density. For the analysis at the end-of-day frequency, we use a linear extrapolation scheme to increase the number of 'observed' implied volatilities in both tails before applying the kernel smoother. The extrapolation works as follows: We start with a set of implied volatilities $\{\sigma_k\}_{k\in[1,N]}$ for some point of time. We split this set into subsets $\{\sigma_{i,\tau}\}_{i\in[1,N_\tau]}$ with all elements in a subset having the same time to maturity τ. We assume that the elements of $\{\sigma_{i,\tau}\}_{i\in[1,N_\tau]}$ are sorted ascendingly by their moneyness m_i.

For the left tail extrapolation, we use the first 5 implied volatilities of each subset. The number of 5 observations is a trade-off between stability on the one hand and including only the most extreme tail observations on the other hand. Unreported results show that our findings are robust to variations in that considered number of observations. On these 5 observations, we estimate the following linear regression using OLS:

$$\sigma_{i,\tau} = \alpha_\tau + \beta_\tau \, m_i + \epsilon_i, \quad \epsilon_i \sim i.i.d. \tag{II.14}$$

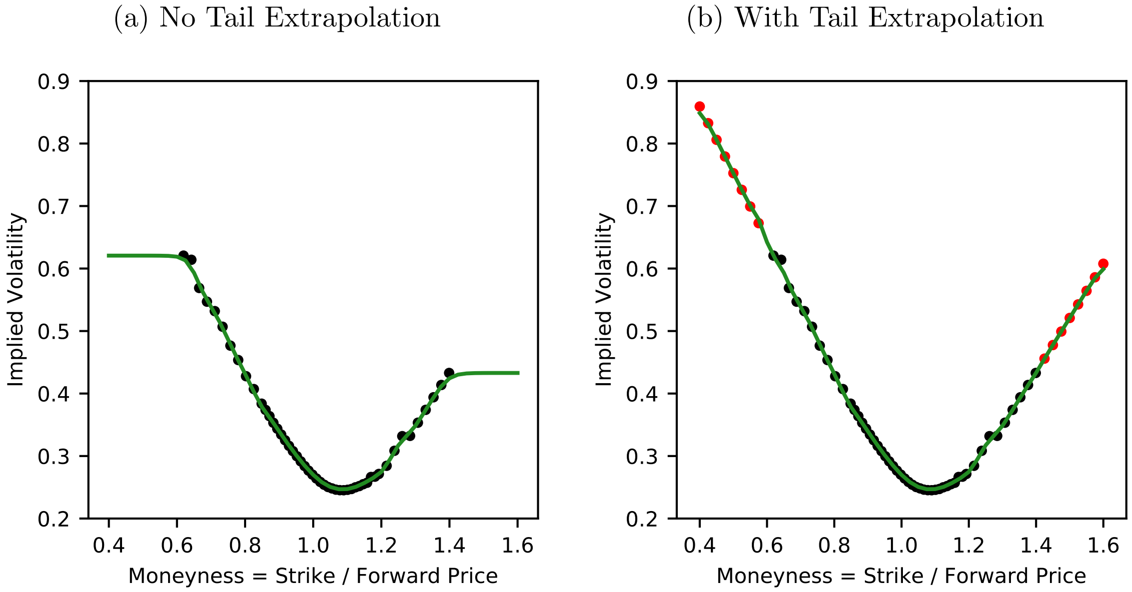
---

[8] If needed, we recommend a linear interpolation along the log-maturity axis between two observed maturities:

$$\hat{\sigma}(m_j, \tau_j) = \hat{\sigma}(m_j, \tau_l) + \frac{\ln \tau_j - \ln \tau_l}{\ln \tau_h - \ln \tau_l}(\hat{\sigma}(m_j, \tau_h) - \hat{\sigma}(m_j, \tau_l)), \tag{II.12}$$

where τ_h (τ_l) represents the next longer (shorter) observed time to maturity, relative to the targeted time to maturity τ_j.

Figure II.1: Effect of Tail Extrapolation on the Kernel Regression


(a) No Tail Extrapolation

(b) With Tail Extrapolation

This figure visualizes for a random day of the sample, May 22nd, 2012, the effect of the linear tail extrapolation on the option-implied volatility smile for Euro Stoxx 50 options with 24 days to maturity. The black dots mark the observed implied volatilities for that day and maturity, the red dots mark the additional artificial observations that are introduced by our tail extrapolation. The green line depicts the kernel regression of the implied volatility smile. Panel (a) shows the kernel regression without tail extrapolation, panel (b) includes the tail extrapolation.

Starting from the lowest observed moneyness, we proceed in steps of $\overline{\Delta K}(\tau)$ and calculate the extrapolated implied volatility as $\tilde{\sigma}_\tau(m) = \alpha_\tau + \beta_\tau\, m$ until reaching a moneyness of $m = 0.4$. We proceed similarly for the right tail, using the last 5 observations in $\{\sigma_{i,\tau}\}_{i\in[1,N_\tau]}$ and extrapolating from the largest observed moneyness in steps of $\overline{\Delta K}(\tau)$ until a final moneyness of 1.6. We use the union of observed implied volatilities and thus artificially created implied volatilities as inputs for the kernel regression. Figure II.1 presents the tail extrapolation visually for a sample day and maturity.

Our last refinement to the Aït-Sahalia and Lo (1998) methodology is to guarantee that the resulting option-implied volatility surface is consistent with an arbitrage-free asset market by applying the algorithm of Fengler (2009). This algorithm adjusts the implied volatility surface minimally, until no-arbitrage constraints in the interpolation are fulfilled. If the volatility surface from the kernel regression is already arbitrage-free, the algorithm does not lead to any changes.

Table II.2 displays the performance gains that we achieve with each adjustment of the Aït-Sahalia and Lo (1998) method. Clearly, dropping the maturity dimension from the kernel regression leads to the strongest improvement, lowering the RMSE by 77% for the

Table II.2: Cross-Validation Errors for Refinement Steps of the Kernel Regression

	S&P 500		Euro Stoxx 50	
Step	RMSE	MAE	RMSE	MAE
Initial	0.0479	0.0139	0.1964	0.0518
No maturity dimension	0.0109	0.0023	0.0156	0.0024
Tail Extrapolation	0.0091	0.0019	0.0101	0.0021
No-arbitrage Enforcement	0.0092	0.0019	0.0100	0.0019

This table shows the leave-one-out cross-validation root mean squared error (RMSE) and mean absolute error (MAE) for our refinements steps of the kernel regression method of Aït-Sahalia and Lo (1998). All numbers refer to end-of-day data. *Initial* represents the original method, applied day by day. We first remove the maturity dimension (*No maturity dimension*) and then add the linear tail extrapolation in implied volatility space (*Tail Extrapolation*). Finally, we additionally enforce no-arbitrage on the volatility surface in the *No-arbitrage Enforcement* step.

S&P 500 and 92% for the Euro Stoxx 50. Further adding the tail extrapolation leads to an additional 16% reduction in the RMSE of the S&P 500 implied volatility surface (minus 35% for the Euro Stoxx 50). Finally, enforcing no-arbitrage does generally not lead to a higher accuracy of the constructed implied volatility surface. However, as the error figures are not rising, we can enforce an arbitrage-free implied volatility surface without having to sacrifice accuracy.

II.2.3 A Bayesian Intraday Volatility Surface

In this section, we propose a novel methodology for constructing the implied volatility surface on a tick-by-tick basis. There are essentially two approaches to constructing an intraday volatility surface, that is, a volatility surface that updates throughout the trading day. One approach is to base the surface on bid-ask price quotes. These quotes are available at any point of time during the day for a broad set of strikes and maturities. The set-up is analogous to the construction of an end-of-day volatility surface, which is either build on settlement prices or bid-ask quotes, too. Obviously, the large amount of input data over a dense grid of strikes and maturities simplifies the construction of the volatility surface and may lead to more robust volatility surface estimates. However, as Vergote and Gutiérrez (2012) point out, bid-ask quotes for different options may be recorded at different points of time in the past, such that the prices may not be consistent across strikes and maturities. Also, high bid-ask spreads might introduce high errors into the measurement of implied volatilities, a problem that Birru and Figlewski (2012) emphasize.

We follow a second approach in constructing the intraday implied volatility surface by basing our estimate on tick-by-tick transaction data. The use of transaction data prevents the imprecise measurement of implied volatility and asynchronicity, that arise from bid-ask quotes, though at the downside of a much sparser set of observations. At any point of time, only a single observation, the recorded trade, is available as current information. Handling this kind of sparseness is key on the way to a high-quality high-frequency volatility surface. Using Bayesian inference, our method sequentially updates an internal estimate of the complete implied volatility smile for a certain maturity with each new incoming observation. This allows to closely track developments in the option market, while smoothing out noise and maintaining key features of the implied volatility smile, like slope or curvature. In the following paragraphs, we will provide a detailed exposition of our methodology.

We start by redefining the moneyness measure as

$$m = \frac{\ln \frac{K}{F}}{\sqrt{\tau}}. \tag{II.15}$$

The division by the square root of maturity neutralizes the moneyness measure from changes in the maturity of the option contract.[9] Over a historic interval, we collect the daily j-th end-of-day volatility smile, that is, the front maturity's volatility smile, the second-shortest maturity's volatility smile, etc. We label these smiles $\{\sigma_{t,m}\}_j$ for a specific date t and moneyness m, where j is iterating over the maturity number. In the following outline, we always consider options of the same expiration date at a time, thus we drop the index j for ease of notation.

For each date t, we collect the volatility smiles at a fix grid of moneynesses $\{m_i\}_{i\in[1,N]}$ for the past 2 years, amounting to 504 trading days. In order to obtain measurements for implied volatility at the grid points $\{m_i\}$, we first apply the methodology of section II.2.2

---

[9] To see this, consider the following example: Assume a 1-year option on an underlying with a yearly volatility of 25% and a log-normal return distribution. The probability that the underlying price exceeds the strike $K = e^{0.05}F$ at the end of the option's maturity is roughly 42%. Now assume the same situation for an option with a remaining maturity of one week. The probability of the underlying price exceeding $K = e^{0.05}F$ after one week is only 7.4%. By decreasing maturity, the same log moneyness $ln\,\frac{K}{F}$ is hence "drifting" into the tails. Implied volatilities in the tails have very different characteristics than implied volatilities near the at-the-money region, though. All else equal, dividing by the square root of the maturity erases this maturity-induced drift towards the tails and makes moneyness comparable across maturities.

to observed end-of-day data. We combine the volatility smile observations to a matrix,

$$V = \begin{bmatrix} \sigma_{t-504,m_1} & \sigma_{t-504,m_2} & \cdots & \sigma_{t-504,m_N} \\ \sigma_{t-503,m_1} & \sigma_{t-503,m_2} & \cdots & \sigma_{t-503,m_N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{t-1,m_1} & \sigma_{t-1,m_2} & \cdots & \sigma_{t-1,m_N} \end{bmatrix}.$$

Here, each row in V corresponds to an end-of-day volatility smile before t. Let $\tilde{V}$ be a matrix that consists of the demeaned columns of V, $\tilde{V}_{t-\Delta t,m_i} = V_{t-\Delta t,m_i} - \bar{V}_{m_i}$, $\bar{V}_{m_i} = \frac{1}{504}\sum_{t'=t-504}^{t-1} \sigma_{t',m_i}$. We apply a PCA on $\tilde{V}$ and obtain the eigenvectors EV_i, eigenvalues ev_i and principal components PC_i. We choose M such that the first M principal components (PCs) explain 95% of the variance of $\tilde{V}$. In general, it holds:

$$PC = \tilde{V}\; EV, \qquad \tilde{V} = PC\; EV^{-1},$$

where PC and EV are matrices with PC_i and EV_i as columns. Approximating V with the first M principal components of $\tilde{V}$ yields the approximating relationship

$$V \approx PC_M\; EV_M^{-1} + \bar{V}, \tag{II.16}$$

with PC_M being a $T \times M$ matrix of the first M columns of the matrix of PCs, EV_M^{-1} being a $M \times N$ matrix of the first M rows of the full EV^{-1} matrix and $\bar{V}$ a $T \times N$ matrix where the i-th column is set to $\bar{V}_{m_i}$. Note, that this converts the PCs back to implied volatility at the respective moneyness grid points $\{m_i\}$. However, we usually observe the implied volatility at a moneyness m that is not part of the regular moneyness grid, but which lies between 2 points of that grid, $m_l \leq m \leq m_h$. We assume that this implied volatility observation can be accurately represented by a linear interpolation of the implied volatilities of the neighbor grid points:

$$\begin{aligned} \sigma_{t,m} &= \frac{m-m_l}{m_h-m_l}\sigma_{t,m_h} + (1 - \frac{m-m_l}{m_h-m_l})\sigma_{t,m_l} = c\,\sigma_{t,m_h} + (1-c)\,\sigma_{t,m_l} \\ &\approx c\,([PC_M]_t\,[EV_M^{-1}]_h + \bar{V}_{m_h}) + (1-c)\,([PC_M]_t\,[EV_M^{-1}]_l + \bar{V}_{m_l}) \\ &= [PC_M]_t\,\left(c\,[EV_M^{-1}]_h + (1-c)\,[EV_M^{-1}]_l\right) + c\,\bar{V}_{m_h} + (1-c)\,\bar{V}_{m_l} \\ &= [PC_M]_t\,f(m) + g(m). \end{aligned} \tag{II.17}$$

$[EV_M^{-1}]_h$ and $[EV_M^{-1}]_l$ stand for the h-th and l-th column in EV_M^{-1}, $[PC_M]_t$ is the row in PC_M that corresponds to time t and $f(m)$ and $g(m)$ are functions of the moneyness of the implied volatility observation m. Hence, for each observation $\sigma_{t,m}$, we have a deterministic approximation based on the reduced set of principle components $[PC_M]_t$.

We now treat the first M principle components as latent state variables in the linear state-space system

$$\sigma_{t,m} = g(m) + [PC_M]_t\, f(m) \;+\; \epsilon_t, \qquad \epsilon_t \sim N(0, s^{2,\epsilon}), \tag{II.18}$$

$$[PC_M]_t = [PC_M]_{t-1} \;+\; \nu_t, \qquad \nu_t \sim N(0, \Sigma^{\nu}), \quad \Sigma^{\nu}_{ij} = 0,\, i \neq j, \tag{II.19}$$

which we estimate with a Kalman Filter (Kalman, 1960). We start in the morning of each day and, with a slight abuse of notation, iterate the time index t through the incoming intraday trade observations. Note, that the matrix V remains the same for all trades in a day, which also means that EV_M^{-1} remains the same throughout the day. The very first PC values, $[PC_M]_0$, are set to the end-of-day PCs of the previous day. At each point of time, the model maintains a prior estimate of the full implied volatility smile based on $f(m)$, $g(m)$ and the current values of PC_M. With each new trade observation, the Kalman Filter updates the state PC_M and hence the full internal representation of the implied volatility smile. In the model, we assume that a potential drift component in the state variables is negligibly small due to the short time interval between 2 observations in intraday set-ups, such that $E_{t-1}(PC_{i,t}) = PC_{i,t-1}$ with the conditional expectation at time $t-1$, $E_{t-1}(\cdot)$.

Due to the restricted nature of the model, the only free parameters are $s^{2,\epsilon}$, the diagonal elements of Σ^{ν} and the initial state variances $\Sigma^{\nu,0}$. Unfortunately, the precise estimation of the variances in a linear state-space system, in particular the state variances, is challenging (Welch and Bishop, 2001, p. 35). Luckily, our state variables, the principle components of V, are only unobserved intraday and we observe a time series of our states on an end-of-day basis for the 504-day time period before the current day t. We use this end-of-day time series to form a prior expectation about the parameters Σ^{ν}. Such a prior can be incorporated in the maximum likelihood estimation of the parameters $\theta = \{s^{2,\epsilon}, \Sigma^{\nu}, \Sigma^{\nu,0}\}$,

$$\begin{aligned} p(\theta|\{\sigma_{t,m}\}, [PC_M]_0) &= \frac{p(\{\sigma_{t,m}\}|\theta, [PC_M]_0)\, p(\theta)}{p(\sigma_{t,m})} \\ &\propto p(\{\sigma_{t,m}\}|\theta, [PC_M]_0)\, p(\theta). \end{aligned}$$

In many applications, nothing is known about the parameters θ, such that one remains agnostic about their values, essentially setting $p(\theta)$ to a uniform distribution and $p(\theta|\{\sigma_{t,m}\}, [PC_M]_0) \propto p(\{\sigma_{t,m}\}|\theta, [PC_M]_0)$. However, given information about the parameters to estimate, we can add this information to the estimation by choosing a more informative prior $p(\theta)$. Specifically, we set $p(\Sigma^\nu) = \phi(\Sigma^\nu; \bar{\Sigma}^\nu, S^\nu)$, where ϕ denotes the pdf of a normal distribution, $diag(\bar{\Sigma}^\nu) = \frac{1}{|\{\sigma_{t,m}\}|} diag(Cov(\Delta PC_M))$ being the scaled covariance of the changes in the end-of-day PC_M observations and $S^\nu = \frac{1}{|\{\sigma_{t,m}\}|}\frac{1}{504}(\mu_4(\Delta PC_M) - diag(\bar{\Sigma}^\nu)^2)$ being the variance of the variance estimate $diag(\bar{\Sigma}^\nu)$, using the centralized fourth moment of ΔPC_M, $\mu_4(\Delta PC_M)$.[10] Note that the estimators for the variance and variance of variance of the PC changes are scaled by the amount of observed trades $|\{\sigma_{t,m}\}|$ on the day under consideration. This is because the end-of-day variance of PC innovations consists of the sum of variances of the intraday PC innovations, which accumulate to the daily PC innovation.

We remain agnostic about the values of the remaining parameters $s^{2,\epsilon}$ and $\Sigma^{\nu,0}$, i.e., we set their respective priors to a uniform distribution. For each day and maturity, we fit the parameters θ to the observed set of intraday option trades. We then use these parameters to filter the intraday time series of PCs and thus the implied volatility smile for the respective maturity.

II.2.4 Alternative Volatility Surface Estimators

A broad range of methods for constructing the option-implied volatility surface has been introduced to the literature. After giving a quick overview on the most common parametric, semi-parametric and non-parametric approaches here, we will shortly present each of these methodologies in more detail in the following subsections.

One popular representative for a parametric approach is the Gram-Charlier expansion. It approximates an unknown density by starting with a Gaussian density and, similar to a Taylor expansion, iteratively adding higher-order components to reduce deviations between the true unknown density and its previous approximation. This parametric approach captures volatility, skewness and kurtosis of the unknown risk-neutral density

---

[10] Using the relationship $Var(X) = E((X-\mu)^2) = E(X^2) - (E(X))^2$, $E(X) = \mu$, the variance of the variance of a random variable X is $Var((X-\mu)^2) = E((X-\mu)^4) - (E((X-\mu)^2))^2 = \mu_4 - \sigma^4$, with the centralized fourth moment μ_4 and the variance σ^2 of X. Given an estimate s^2 for σ^2 based on n observed realizations of X, the central limit theorem holds (van der Vaart, 2000, p.16) and implies

$$\sqrt{n}(s^2 - \sigma^2) \xrightarrow{p} N(0, \mu_4 - \sigma^4).$$

and allows for closed-form prices of Calls and Puts (Backus et al., 2004), which can straight-forward be translated to implied volatilities (see section II.1).

Interpolating observed implied volatilities with polynomial splines is certainly the most popular semi-parametric approach for constructing the implied volatility surface. Several variations of the spline interpolation approach exist, mainly differing in the degree of the spline or in the dimension over which the interpolation runs (see, for example, Bliss and Panigirtzoglou (2002), Jiang and Tian (2005), or Chang et al. (2012)). We follow Figlewski (2008) and Fengler (2009) as their procedure nests many other approaches of the literature.

An often applied non-parametric approach is the kernel regression. This is the class of methods that our volatility surface estimator of section II.2.2 belongs to. Another prominent example of this class of methods, that is often used in financial research, is the kernel regression specification of OptionMetrics (2016). That data provider approximates the option-implied volatility surface with a three-dimensional kernel regression. The corresponding three bandwidth parameters are fixed in OptionMetrics (2016).

In the following expositions, let $\{O_i\}_{i\in[1,\ldots,N]}$, be a panel of option prices with N being the number of observed option prices that are used to construct the volatility surface. As all methodologies apply per point of time, we desist from a time index to save notation. In addition to the previously introduced underlying price S_i, strike K_i and maturity τ_i, each option price $O_i = O(K_i, \tau_i, I_i)$ is associated with a Black-Scholes option delta Δ_i and an indicator I_i, which is 1 for Call options and 0 for Put options. In this notation, O_i just groups all option prices, since for Calls $O(K_i, \tau_i, 1) = C(K_i, \tau_i)$ and for Puts $O(K_i, \tau_i, 0) = P(K_i, \tau_i)$.

II.2.4.1 Gram-Charlier Expansion

The Gram-Charlier expansion applies a fourth-order approximation of the risk-neutral density $q(x, \tau)$. More formally,

$$q(x,\tau) = \phi(x;\sigma_\tau^2) - \frac{\gamma_{1,\tau}}{3!}\frac{\partial^3\phi}{\partial x^3}(x;\sigma_\tau^2) + \frac{\gamma_{2,\tau}}{4!}\frac{\partial^4\phi}{\partial x^4}(x;\sigma_\tau^2) + H.O.T.(x^5),\ \forall x,\tau, \tag{II.20}$$

where $\phi(x;\sigma_\tau^2)$ is the Gaussian density function with volatility σ_τ, evaluated at the point x. $\gamma_{1,\tau}$ and $\gamma_{2,\tau}$ are maturity-specific free parameters that account for the degree of skewness and excess kurtosis in $q(x,\tau)$, and $H.O.T.$ stands for higher order error terms.

We follow Beber and Brandt (2006) and fit the free parameters $(\sigma_\tau, \gamma_{1,\tau}, \gamma_{2,\tau})$ by minimizing the squared pricing errors between observed option prices and Backus et al. (2004) implied option prices, i.e.

$$\min_{\sigma_\tau, \gamma_{1,\tau}, \gamma_{2,\tau}} \sum_{i=1}^{N} \Big(O(K_i, \tau_i, I_i) - \hat{O}(K_i, \tau_i, I_i; \sigma_\tau, \gamma_{1,\tau}, \gamma_{2\tau}) \Big)^2 \tag{II.21}$$

where the Backus et al. (2004) implied call price coincides with

$$\begin{aligned} \hat{C}(K, \tau; \sigma_\tau, \gamma_{1,\tau}, \gamma_{2,\tau}) =& e^{-r\tau}(F\phi(d) - K\phi(d - \sigma_\tau)) \\ &+ Fe^{-r\tau}\phi(d)\sigma_\tau \left[\frac{\gamma_{1,\tau}}{3!}(2\sigma_\tau - d) - \frac{\gamma_{2,\tau}}{4!}(1 - d^2 + 3d\sigma_\tau - 3\sigma_\tau^2) \right] \\ d =& \frac{ln(F/K) + \sigma_\tau^2/2}{\sigma_\tau}. \end{aligned}$$

and F is the value of a future contract on the underlying with maturity τ and $\phi(\cdot)$ denotes the pdf of a standard normal distribution. Put prices are obtained via the put-call parity. In line with Jondeau and Rockinger (2000), we fit the three free parameters separately for each observed maturity τ.

II.2.4.2 Smoothing Spline

We follow Figlewski (2008) and Fengler (2009) as their procedure nests many other approaches of the literature. Their variation of the smoothing spline algorithm consists of 4 steps: (i) data pre-processing, (ii) spline interpolation, (iii) tail extrapolation and (iv) data post-processing. We now sketch each step in more detail.

According to Figlewski (2008), a pre-processing step of observed at-the-money (ATM) implied volatilities is necessary due to a practical problem of the data: Many approaches construct the option-implied volatility surface with only out-of-the-money (OTM) options due to their higher liquidity and hence supposably higher informativeness about the current market environment. However, the switch from OTM Puts to OTM Calls at a moneyness of $m = 1$ results in many occasions in a small jump in implied volatility, as ATM Puts often imply a slightly different implied volatility than ATM Calls. This volatility jump may deteriorate the quality of the fitted spline.

The pre-processing step mitigates this problem by considering, separately for each maturity, a small region around the ATM level, which contains options that fulfill $0.98 \leq m_i = \frac{K_i}{F} \leq 1.02$. Let K_{high} be the highest and K_{low} be the lowest traded strike within that moneyness region. By linearly combining the observed Put- and Call-implied volatilities $\sigma_{Put}(K_i)$ and $\sigma_{Call}(K_i)$, we smooth the hard volatility jump in the region of ATM options and replace the observed implied volatilities with their smoothed counterparts (Figlewski, 2008):

$$\sigma^*(K_i) = w\, \sigma_{Put}(K_i) + (1-w)\, \sigma_{Call}(K_i), \tag{II.22}$$

$$w = \frac{K_{high} - K_i}{K_{high} - K_{low}}, \ \forall K_i \in [0.98\, F_i,\ 1.02\, F_i]. \tag{II.23}$$

In the main interpolation step, we now fit a spline to the observed and pre-processed option-implied volatilities for each maturity separately. Figlewski (2008) proposes to use a 4th-order polynomial with one knot at a moneyness of $m = 1$. Using a 4th-order polynomial guarantees a continuous 3rd derivative of the spline, which translates into a smooth RND without any edges. In addition, it is not too restrictive on the shape of the implied volatility smile, while still smoothing out market noise. If enough option observations to determine all parameters of the spline are available, we use this setting for interpolating the moneyness range. In settings where a sufficient amount of implied volatility observations for the parameter estimation is not available, we simply drop the knot point and reduce the polynomial's order, thus effectively reducing the number of parameters.

The third step concerns the approximation of the tails of the RND. Depending on the moneyness of the traded options, it might occur that these do not reach out far enough into the out-of-the-money regions of the volatility smile to accurately recover these tails. In these situations, an extrapolation of the RND and thus the volatility surface into the tails is necessary. For that task, Figlewski (2008) proposes to translate the initial volatility curve of each maturity to its corresponding RND estimate and fit a generalized extreme value density (GEV) to both tails of the RND. We then extrapolate the RND based on the fitted GEV tail and backwards calculate option prices based on that extrapolated density via numerical integration of the expected payoff (see section II.1).

Finally, the fourth step of the interpolation method follows the algorithm of Fengler (2009) to post-process the implied volatility surface in order to exclude arbitrage in the estimator.

II.2.4.3 Three-Dimensional Kernel Regression

For this approach, we follow the methodology of the OptionMetrics IvyDB US File and Data Reference Manual, Version 3.1, Rev. 8/19/2016. The implied volatility estimate for an option with a given delta Δ_j, a maturity τ_j and a Put/Call flag $I_j \in \{0, 1\}$ is given by

$$\hat{\sigma}(\Delta_j, \tau_j, I_j) = \sum_{i=1}^{N} \frac{\nu_i \; k(\Delta_j - \Delta_i, ln(\tau_j) - ln(\tau_i), I_j - I_i)}{\sum_{i=1}^{N} \nu_i \; k(\Delta_j - \Delta_i, ln(\tau_j) - ln(\tau_i), I_j - I_i)} \, \sigma(\Delta_i, \tau_i, I_i), \tag{II.24}$$

where the left hand side stands for the approximation of the unobserved implied volatility, while the right hand side is a weighted average of all N observed implied volatilities. Especially, the right hand side of equation II.24 contains options of all maturities, which differs from the other methodologies that we consider. The weight that each observed $\sigma(\Delta_i, \tau_i, I_i)$ obtains is determined by that option's vega ν_i and the proximity of the observation to the desired target point of the interpolation. This proximity is translated into a weight via the Gaussian kernel function

$$k(x, y, z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2h_1} + \frac{y^2}{2h_2} + \frac{z^2}{2h_3}\right)}. \tag{II.25}$$

The IvyDB US Reference Manual specifies the bandwidth parameters to coincide with

$$h_1 = 0.05, \; h_2 = 0.005, \; h_3 = 0.001$$

and limits the kernel regression to the delta interval of $[0.2, 0.8]$.

The volatility surface of OptionMetrics (2016) is not necessarily arbitrage-free. In order to compare apples with apples, we need all considered volatility surfaces to be arbitrage-free, so we again apply the post-processing step of Fengler (2009) on the original OptionMetrics (2016) volatility surface. All of our our results are qualitatively and quantitatively robust to whether or not one applies this post-processing step. In unreported results we have found that ensuring no-arbitrage improves the statistical accuracy of the OptionMetrics (2016) volatility surface.

II.3 Data

We use data for two of the most actively traded equity option contracts: options on the S&P 500 and options on the Euro Stoxx 50. The former is traded at the CBOE, while

Table II.3: Summary Statistics

	EOD		Trades	
	S&P 500	Euro Stoxx 50	S&P 500	Euro Stoxx 50
Total number of price observations	9,246,926	6,877,495	43,749,042	16,921,911
Avg number of maturities per day	17.21	16.58	16.96	12.65
Avg number of strikes per maturity	61.27	45.28	16.97	5.93
Total number of short-term ATM prices	771,000	73,000	24,627,422	5,866,507
Total number of short-term Left Tail prices	1,326,413	274,227	4,449,056	1,956,662
Total number of short-term Right Tail prices	293,823	198,558	1,330,209	903,600
Total number of medium-term ATM prices	1,186,655	293,028	7,348,889	3,148,740
Total number of medium-term Left Tail prices	3,536,880	1,218,912	3,996,122	3,016,860
Total number of medium-term Right Tail prices	1,048,097	982,266	1,628,093	1,586,504
Total number of long-term ATM prices	120,389	340,796	100,016	168,532
Total number of long-term Left Tail prices	656,001	1,768,878	179,571	170,934
Total number of long-term Right Tail prices	307,668	1,727,830	89,664	103,572

This table provides some aggregated summary statistics of our option data panel. 'EOD' stands for 'end-of-day' (data-rich environment), whereas the 'Trades' columns summarize the aggregated data for our intraday analysis (data-poor environment).

the latter is traded at the Eurex. The S&P 500 and the Euro Stoxx 50 do both stand for the Blue Chip stocks of their respective currency zone. Both option contracts are of European exercise style.

Historical data for S&P 500 options come from the CBOE Livevol data shop. We collect both, end-of-day bid-ask prices as well as intraday transaction prices. For the end-of-day data set, we use the mid price for our calculations, whereas we use actual trade prices for the intraday data set. The end-of-day data is from January 2004 to July 2017. The intraday data spans the period January 2004 to October 2017. The data comes with the matched bid and ask prices of the underlying at the point of time of the record.

We obtain end-of-day settlement prices and intraday transaction prices for Euro Stoxx 50 options and the underlying Euro Stoxx 50 index from the *Karlsruher Kapitalmarktdatenbank* (KKMDB), which is hosted at the Finance Institute of the Karlsruhe Institute of Technology. The KKMDB receives its data directly from the Eurex and Stoxx. The index data comes in 15 second intervals, while the intraday option data is time stamped at the point of time of the trade. We match the option trade with the index price directly at or prior to the trade's time stamp. The end-of-day data spans January 2002 to September 2017, whereas the intraday data covers the period January 2003 to September 2017. Table II.3 summarizes our option data sets in more detail.

For the risk-free rate, we use the US-Dollar and Euro OIS curves from Bloomberg, which come in discrete maturities between 1 day and 20 years and which we obtain at the daily frequency. We match each option record with the risk-free rate for its respective date. We interpolate the OIS rate linearly along the maturity dimension to match the maturity of the respective option.

We apply a number of filtering steps to ensure that only valid option prices enter the construction of implied volatility surfaces. We only consider options with a price that fulfills basic no-arbitrage considerations. More precisely, we keep a Call option price C, if $C \geq Se^{-q\tau} - Ke^{-r\tau}$, and a Put option price P, if $P \geq Ke^{-r\tau} - Se^{-q\tau}$, with the current price of the underlying S, a strike of K, maturity of τ years, the risk-free rate r and the continuous dividend yield q. While this study focuses on options with European exercise style, we still want to point out where the application of the described procedures deviates for options of American exercise style.[11] In the filtering phase, an additional lower price bound is imposed on American exercise style options. Since these options can be exercised at any point of time, their value must always exceed the intrinsic value of the option contract. We would hence only keep an American Call option price C^A (American Put option price P^A), if, in addition to fulfilling the lower price bound constraint for European Calls (Puts), it fulfills $C^A \geq S - K$ ($P^A \geq K - S$). Independent of the exercise style, we remove options for which the calculation of implied volatility did not converge. In line with Bliss and Panigirtzoglou (2002), we only keep out-of-the-money options and recover prices for in-the-money options (if required) through put-call parity or, for American exercise style options, by repricing the option based on the out-of-the-money implied volatility. The latter is not done for the three-dimensional kernel regression of OptionMetrics (2016), as that method explicitly works with both, Call and Put prices, in a delta range of 0.2 to 0.8 and a vega of less than 0.5.

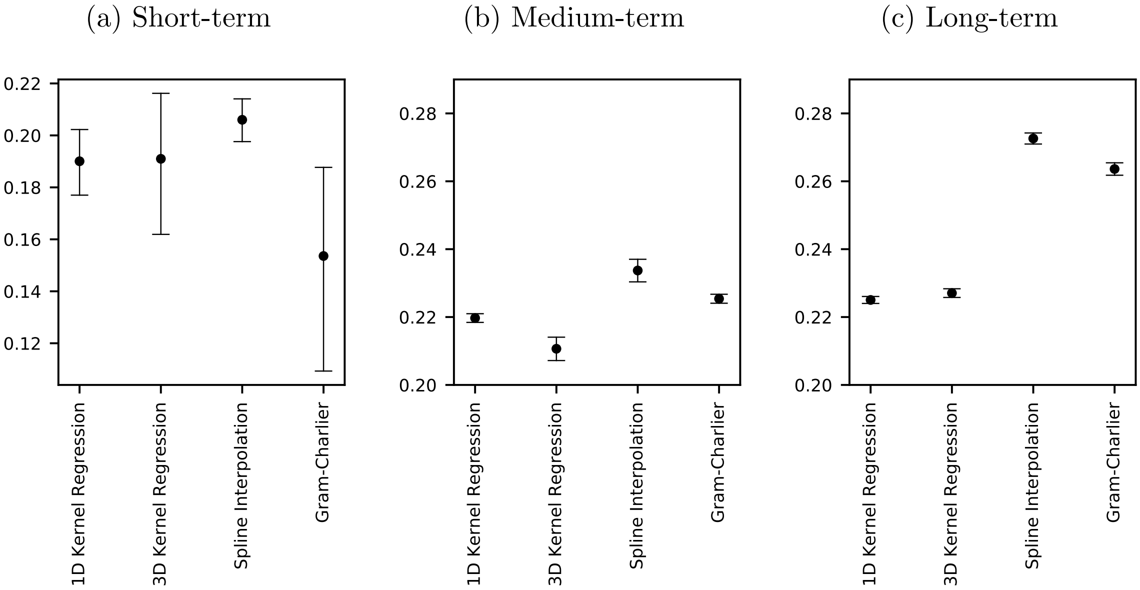
II.4 On the Uniqueness of Option-Implied Information

Option-implied risk and return measures, such as forward-looking risk-neutral variance, skewness or the variance risk premium do have a unique mathematical representation. Yet, the numerical implementation of such measures does often require the approximation of an integral of option prices across the whole strike spectrum, even though only a subset of these strikes are actually traded in the market. Conditional on a correct volatility

---

[11] For example, most options on single stock underlyings are of American exercise style.

Figure II.2: Unconditional Mean of Model-Free Implied Volatility by Methodology

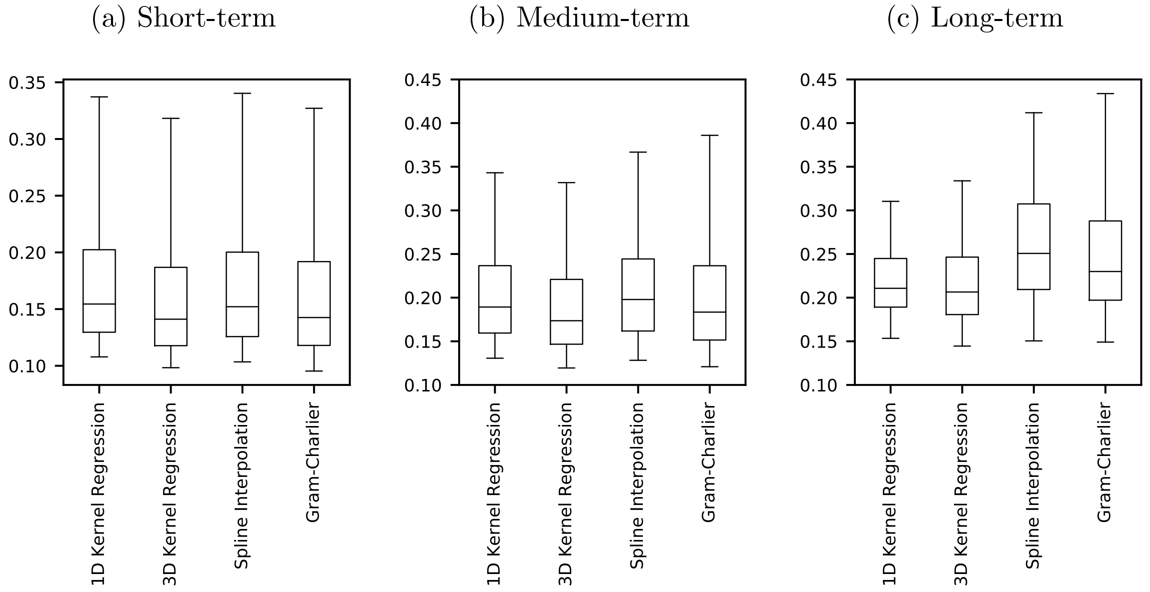


This figure presents the sample mean of the Bakshi et al. (2003) model-free implied volatility for S&P 500 options (black dots), by maturity and interpolation methodology, together with 95% confidence intervals. The short-term category refers to maturities below 30 days, the mid-term category groups maturities between 30 and 365 days, the long-term category summarizes maturities that are longer than 365 days.

surface, Jiang and Tian (2005) show how to implement such an integration numerically. But in real-life applications the problem is that different state-of-the-art approaches for constructing the volatility surface in high resolution across the strike dimension do result in different volatility surfaces. It is the goal of this chapter to assess the severity of these differences for a subset of methodologies and to perform a thorough statistical analysis to identify the methodology that comes closest to the unobserved true volatility surface.

Figure II.2 summarizes the sample mean of the Bakshi et al. (2003) implied volatilities for S&P500 options with a remaining maturity of (i) less than 30 days (short-term), (ii) 30 to 365 days (medium-term) and (iii) more than 365 days (long-term), respectively. Each sample mean is based on one of the four volatility surfaces that we have discussed above. We highlight two insights: First, different volatility surfaces imply different values for a standard risk measure like Bakshi et al. (2003) implied volatility. The average medium-term Bakshi et al. (2003) implied volatility is in our sample 21.0% when using the three-dimensional kernel regression volatility surface and 23.5% when working with the spline volatility surface. That 12% relative spread is statistically significant and obviously large from an economic point of view. Second, while the spline interpolation method and the kernel regression methods tend to produce Bakshi et al. (2003) volatilities that are close

Figure II.3: Distribution Characteristics of Model-Free Implied Volatility by Methodology



This figure presents boxplots for the estimates of the Bakshi et al. (2003) model-free implied volatility for S&P 500 options, by maturity and interpolation methodology. The short-term category refers to maturities below 30 days, the mid-term category groups maturities between 30 and 365 days, the long-term category summarizes maturities that are longer than 365 days. The rectangular boxes mark the quartiles of the distribution of implied model-free volatility, the horizontal lines within the boxes mark the median. The solid lines outside of the boxes expand to the 5% and 95% percentiles.

to each other for short-term options, the differences build up for maturities larger than 30 days. Noteworthy, the kernel regression methods produce average term structures of Bakshi et al. (2003) implied volatilities that are flatter than the term structures of the other two interpolation methodologies. This results in comparatively low average medium- and long-term estimates for Bakshi et al. (2003) implied volatility.

In contrast to the disperse, yet precisely estimated, sample means of the Bakshi et al. (2003) volatilities, the boxplots in figure II.3 describe the respective full distribution of each interpolation method's Bakshi et al. (2003) implied volatility. The distributions largely overlap for short- and medium-term maturities. For medium-term maturities and even more strongly for long-term maturities, the downward bias of the kernel regression methods' estimates becomes visible, which is even more pronounced in the right tail of the distribution.

Differences in model-free implied variance estimates induce different variance risk premium (VRP) estimates. In line with Drechsler and Yaron (2011), we define the variance risk premium (VRP) as the difference in expected variance under the risk-neutral measure

$\mathcal{Q}$ and the physical measure,

$$VRP_t = E_t^{\mathcal{Q}}[\text{Total Return Variation}(t, t+21)] - E_t[\text{Total Return Variation}(t, t+21)]$$

where $t+21$ stands for a one-month (or 21-trading-day) increment. We use Bakshi et al. (2003) model-free implied variance as the measure for the expected total return variation under $\mathcal{Q}$. For those days, during which the one-month maturity option is not traded, we linearly interpolate the implied variance measures for surrounding traded maturities along the maturity dimension to obtain a one-month-ahead expectation. We use an adjusted form of the model of Corsi (2009) to obtain total return variation expectations under the physical measure:

$$RV_t = r_t^2 = \left(\ln \frac{S_t}{S_{t-1}}\right)^2, \quad RV_t^{(w)} = \frac{1}{5}\sum_{i=t-4}^{t} r_i^2, \quad RV_t^{(m)} = \frac{1}{21}\sum_{i=t-20}^{t} r_i^2$$

$$\ln RV_{t+21}^{(m)} = \alpha + \beta_1 \ln RV_t + \beta_2 \ln RV_t^{(w)} + \beta_3 \ln RV_t^{(m)} + \epsilon_{t+21}. \tag{II.26}$$

The parameters α, β_1, β_2 and β_3 are estimated via OLS.

Table II.4 summarizes the average annualized VRP for S&P 500 and Euro Stoxx 50 options for different volatility surfaces, respectively. The average annualized VRP for S&P 500 (Euro Stoxx 50) options has been estimated to be between 1.5% and 2.4% (1.9% and 2.9%), depending on the volatility surface. In relative terms, the spline-based volatility surface results on average in a 60% higher S&P 500 VRP estimate relative to the same estimate for the three-dimensional kernel regression. The differences between the average VRP estimates are for nearly all pairs of methods statistically strongly significant.

Table II.4: Average Annualized Monthly Variance Risk Premium and Sharpe Ratio

	S&P 500		Euro Stoxx 50	
	Average VRP	Sharpe ratio	Average VRP	Sharpe Ratio
3D Kernel Regression	0.015 (0.0006)	0.44	0.019 (0.0006)	0.5
1D Kernel Regression	0.020 (0.0006)	0.54	0.029 (0.0007)	0.68
Spline Interpolation	0.024 (0.0008)	0.57	0.028 (0.0007)	0.65
Gram-Charlier Expansion	0.016 (0.0010)	0.26	0.024 (0.0007)	0.56

This table shows average annualized one-month variance risk premia and VRP Sharpe Ratios. Standard errors are given in parenthesis. The integration scheme and model for the physical variance expectations are the same for all estimates, such that the only difference is the method for constructing the implied volatility surface.

Table II.5: Average Bakshi et al. (2003) Skewness by Time Period and Method

(a) S&P 500

Period	Start Date	End Date	3D Kernel Regression	1D Kernel Regression	Spline Interpolation	Gram-Charlier Expansion
Pre-Crisis	2004-01-01	2007-12-31	-1.256 (0.005)	-2.347 (0.022)	-2.583 (0.042)	-1.332 (0.007)
Financial Crisis	2008-01-01	2009-03-31	-0.881 (0.005)	-1.307 (0.01)	-1.366 (0.035)	-1.210 (0.01)
Post-Crisis	2009-04-01	2017-07-21	-1.316 (0.002)	-2.791 (0.007)	-3.703 (0.025)	-1.492 (0.002)

(b) Euro Stoxx 50

Period	Start Date	End Date	3D Kernel Regression	1D Kernel Regression	Spline Interpolation	Gram-Charlier Expansion
Pre-Crisis	2002-01-01	2007-12-31	-0.826 (0.009)	-1.858 (0.009)	-1.612 (0.01)	-1.347 (0.015)
Financial Crisis	2008-01-01	2009-03-31	-0.708 (0.015)	-1.324 (0.010)	-1.335 (0.013)	-1.242 (0.01)
Between-Crisis	2009-04-01	2010-03-31	-0.657 (0.016)	-1.622 (0.011)	-1.721 (0.015)	-1.135 (0.009)
Euro Crisis	2010-04-01	2013-12-31	-0.699 (0.009)	-1.855 (0.01)	-1.921 (0.012)	-1.277 (0.007)
Post-Crisis	2014-01-01	2017-09-30	-0.805 (0.008)	-1.918 (0.009)	-1.826 (0.015)	-1.287 (0.007)

This table summarizes the sample average of the end-of-day Bakshi et al. (2003) model-free option-implied skewness, for different time periods and different interpolation schemes. All numbers in panel (a) are calculated for S&P 500 options, whereas numbers in panel (b) refer to Euro Stoxx 50 options. Only options with maturities between 15 and 91 days are considered. Standard errors are given in parenthesis.

Monthly Sharpe ratios for the VRP range from 0.26 to 0.57 for the S&P 500 and from 0.5 to 0.68 for the Euro Stoxx 50. We highlight that these economically large differences are a direct result of the choice of the inter- and extrapolation method that builds the basis for a volatility surface; the input data and the VRP calculation scheme is the same across all methods.

Table II.5 states the sample mean of model-free option-implied skewness for maturities between 15 and 91 days. Option-implied skewness is calculated as in Bakshi et al. (2003) and reported for different sub-samples.[12] We split the US data set into a pre-financial-crisis, a crisis and a post-financial-crisis period. We apply the same cuts for the European data set, but further split the post-financial-crisis period into a between-crisis-, an euro-crisis-, and a post-euro-crisis period due to the high impact of the European sovereign debt crisis for European stock markets. The risk-neutral skewness estimates are nearly all significantly different across different volatility surfaces. The estimates of the three-dimensional kernel regression are roughly only half the size when compared to our proposed one-dimensional kernel regression or the spline method. Further, the changes in the average risk-neutral skewness between one subperiod and the next also

---

[12] The results do not change qualitatively for different maturity intervals.

differ among the volatility surfaces, sometimes even disagreeing on the sign of the change. For example, risk-neutral skewness based on the spline interpolation become less negative on average after the Euro Crisis in Europe, while the estimates based on all other volatility surfaces become more negative.

In summary, information extracted from option markets is supposed to be unique. But our analysis documents that this information is sensitive with regard to the volatility surface that a researcher uses. We have shown that risk-neutral model-free estimates for variance and skewness differ by a large margin across different state-of-the-art methods. While we have used the Bakshi et al. (2003) moments for explanatory purpose, our findings hold more generally for any quantity that is extracted from the aggregation of option prices along the strike range.

II.5 Assessing the Accuracy of a Volatility Surface

We proceed with a discussion about potential evaluation schemes for the relative advantages and disadvantages of different state-of-the-art volatility surfaces. Our previous findings have documented that standard option-implied measures differ across volatility surfaces. As these measures are deterministic functions of the volatility surface, it is natural that the most accurate volatility surface does imply the most accurate option-implied risk measures.

At the same time, accuracy is only one of multiple practical considerations when constructing an implied volatility surface. It is, for example, well thinkable to accept a lower accuracy in favor of a more informative representation of the volatility surface. Especially parametric models can provide such representations. Depending on the model, single parameters can be interpreted directly and serve as measures that express the situation at the option market in lower dimension. For example, the parameters of the SVI model of Gatheral (2004) inform about the level of the implied volatility smile, its rotation and how wide it is. Parametrization also allows to easily share an implied volatility surface, as it can be fully reconstructed from the relatively few parameters. The model for the volatility smile allows to extrapolate beyond observed option prices in a manner that is consistent with the central part of the modeled volatility smile. No-arbitrage constraints can be incorporated into the construction of the implied volatility surface at the parameter estimation phase already (Damghani and Kos, 2013; Damghani, 2015). These arguments speak in favor of a parametric representation of the implied volatility surface.

However in this chapter, we explore how accurate different volatility surfaces capture option market information. The result of our analysis can then serve as an important input, next to the previously mentioned concerns, for the choice of a construction method in practice.

One might also be willing to deliberately sacrifice some accuracy in the representation of the observed implied volatilities in favor of a smoother volatility surface (Jackwerth, 2000; Jackwerth and Rubinstein, 1996). This is especially true, if the noise in the observations is expected to be large enough to not be fully canceled out by the interpolation and smoothing method. In that case, the constructed volatility surface would show spikes or bumps, which would decrease its smoothness. On the other hand, a very smooth volatility surface might plane out important features of the observations and thus introduce biases.[13] Therefore, we are going to implement a thorough investigation of the statistical accuracy of popular volatility surfaces and compare them with respect to their smoothness.

We follow a rich machine learning literature that assesses the accuracy of a model based on leave-one-out cross-validation (Geisser, 1993; Kohavi, 1995). The advantage of this cross-validation approach for our study is that every method that we use to construct an option-implied volatility surface is evaluated with data that was not used to construct the surface. This allows us to detect over- and under-fitting.

We evaluate the statistical quality of each volatility surface with end-of-day and intraday data for S&P 500 and Euro Stoxx 50 options. End-of-day data is characterized by a rich panel of option prices for different strikes and maturities, that all refer to the same point of time. We call this to be the *data-rich* environment. In contrast, intraday data is characterized by a limited amount of observed trade prices in a given time interval. We therefore call the intraday application to be a *data-poor* environment. The intraday set-up becomes increasingly data-poor as the considered time interval for pooling trade observations shrinks and thus the time resolution increases.[14]

The highest possible time resolution of a method is bound by the minimum amount

---

[13] The introduction of biases by very smooth volatility surfaces can easily be seen by the fact that the smoothest volatility surface is an uncurved plane.

[14] End-of-day data does not necessarily create a data-rich environment. For some options, only a handful of strikes are traded, thus effectively constituting a data-poor environment in end-of-day data. On the other hand, if bid-ask quotes are used on an intraday basis or the time interval for pooling intraday trade prices becomes large enough, there will likely be enough observations for distinct strikes to constitute a data-rich environment.

of option trades that a method requires for constructing the option-implied volatility surface. Technically, the Gram-Charlier expansion requires only 3 observed option prices for distinct strikes at the same maturity. Price observations at 3 different strikes are also the minimum of the kernel regression methods and the spline interpolation for capturing the key characteristics level, slope and curvature of the implied volatility smile. Opposed to these methods, the Bayesian intraday volatility surface from section II.2.3 can maintain a realistic estimate of the implied volatility smile after processing a single new price observation. However, the Bayesian method updates an existing estimate and does not construct the volatility smile from scratch based on new observations only, like the other methods do. For the Gram-Charlier, kernel regression and spline methods, we therefore undertake four independent intraday analyses, namely relying on 3, 4, 5 and 10 pooled trade observations at different strikes per maturity. In each case, we compare their performance among each other and with the Bayesian method that updates on every single trade.

For the data-rich end-of-day case, the cross-validation works as follows: Given price observations of a day, we in turn leave out a single observation and calculate the methodology-implied volatility estimate at the strike and maturity that was left out. We repeat this procedure for each price observation. The data-poor intraday set-up with 3, 4, 5 and 10 price observations per maturity is treated similarly: For each new transaction, we use the corresponding 3, 4, 5 or 10 preceding transactions with differing strikes for the same maturity in order to create an estimate for the next transaction's implied volatility. In the case of the Bayesian volatility surface, we use the prior volatility estimate at the strike and maturity of the next transaction. For all intraday volatility surfaces, this approach is basically assessing out-of-sample how well a volatility surface predicts future option prices.

Given a set of N evaluation samples with implied volatility observations $\sigma_i, i \in \{1, ..., N\}$, our primary evaluation measure is the root mean squared error (RMSE), that arises when comparing the respective method's implied volatility estimator $\hat{\sigma}_i$ with the left out observation σ_i:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sigma_i - \hat{\sigma}_i)^2}. \tag{II.27}$$

By squaring the errors, the RMSE penalizes large deviations of the constructed volatility surface from the observed implied volatilities more strongly than small deviations. However, as discussed above, one might be willing to accept occasional large errors in favor of a smoother volatility surface. This is especially true, if one expects such large errors to be due to outliers, that are not representative of the true unobserved volatility surface. For this reason, we also calculate the mean absolute error (MAE),

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left|\sigma_i - \hat{\sigma}_i\right|, \tag{II.28}$$

which is less responsive to such occasional large errors and is still comparably low, if these errors only occur seldom and the remaining fit is good. For both error measures, we exclude observations with an implied volatility above 10.[15]

We also compare the end-of-day volatility surfaces with respect to their smoothness directly. Following Jackwerth (2000), we measure smoothness as the sum of squared second derivatives of the implied volatility smile. More precisely, for each day t and maturity τ, we consider a moneyness range of $[1 - s\sqrt{\tau}, 1 + s\sqrt{\tau}]$, where s is an estimate for the unconditional annual volatility of the underlying. We discretize this range in steps of $\Delta = \frac{2s\sqrt{\tau}}{99}$ and construct the volatility smile $\{\hat{\sigma}_{j,t,\tau}\}_{j\in[0,100]}$ for the grid points of that day and maturity with each method. Our measure for the smoothness of the smile is then calculated as

$$S_{t,\tau} = \sum_{j=1}^{99}\left(\frac{\hat{\sigma}_{j-1,t,\tau} - 2\hat{\sigma}_{j,t,\tau} + \hat{\sigma}_{j+1,t,\tau}}{\Delta^2}\right). \tag{II.29}$$

Finally, we compute the mean of that smoothness measure across all days and maturities and take the square root for better readability. In that, we exclude the 1% of the smiles with the highest and the 1% with the lowest smoothness measure to mitigate the impact of the tails of the distribution of smoothness estimates on the average smoothness measure.

The MSE Performance Measure and Smoothness The following discussion shows that the MSE is indeed a valid measure for assessing the quality of a volatility surface construction methodology, especially considering the desire to obtain a smooth volatility surface that still captures all characteristics of the true unobserved volatility surface.[16] In the end-of-day set-up, all price recordings for a given day refer to the same point of time, but may be measured with noise. We can split the error of a construction methodology, $\sigma_i - \hat{\sigma}_i$, into a noise component ϵ_i^{noise} and a methodology-specific bias

---

[15] In rare cases, observed option prices translate into unreasonably high implied volatilities. Our threshold of 10 leads to the exclusion of 0.012% or less of the data points, depending on the data set.

[16] The measure that we print in our tables, the RMSE, is the square root of the MSE. Since taking the root does not affect the relative ordering of different methods, our arguments for the MSE apply similarly for the RMSE.

component ϵ_i^{bias}. The noise component is the difference between the implied volatility observation σ_i and the true unobserved implied volatility and is thus the same across all methodologies. All methodologies are not purely interpolating between observed option prices, but attempt to smooth the volatility smile in order to reduce the noise error. However, smoothing comes at the cost that a potential over-smoothing introduces a bias into the implied volatility estimator, which is independent of the noise component. Such a bias results in an error ϵ_i^{bias} between the implied volatility estimate $\hat{\sigma}_i$ and the true unobserved implied volatility. The methodology-specific residual can thus be rewritten as $\epsilon_i^{method} = b\,\epsilon_i^{noise} + \epsilon_i^{bias}$, where we would expect b to be positive but smaller than 1. If a method just replicates the exact implied volatility observations without any smoothing, its volatility surface would not be biased ($\forall i : \epsilon_i^{bias} = 0$) but exposed to the full noise error component ($b = 1$). Substituting these considerations into the MSE equation yields for the expected MSE

$$E\left[MSE\right] = E\left[(b\,\epsilon_i^{noise} + \epsilon_i^{bias})^2\right] = b^2 E\left[(\epsilon_i^{noise})^2\right] + E\left[(\epsilon_i^{bias})^2\right]. \tag{II.30}$$

The expected MSE consists of 2 components: The noise component and the bias component. If a method works particularly well in smoothing out price measurement noise, its b parameter is close to 0 and the expected MSE reduces. At the same time, the bias error directly increases the MSE if a method starts to over-smooth the implied volatility curve. Explicitly splitting the estimation error into its two components would be spurious, as the noise component is not observed directly. However, a perfect standardization methodology would have just the right smoothness to cancel out noise without introducing any biases. Both objectives lead to a lower MSE.

The situation becomes a bit more complex for the intraday set-up. Here, transaction prices are recorded at different points of time, leading to a situation, in which there is a small time gap between the option prices that are underlying the standardization and the option price that is used to evaluate the standardization. During this time, market expectations may have changed, which may result in a change in the true implied volatility surface. When calculating the mean squared error, the time difference introduces a third error component, ϵ_i^{time}, into the estimation error. We assume this error component to be independent from the noise and the bias components in the estimation error. For the intraday set-up, the expected MSE can hence be written as

$$E\left[MSE\right] = b^2 E\left[(\epsilon_i^{noise})^2\right] + E\left[(\epsilon_i^{bias})^2\right] + E\left[(\epsilon_i^{time})^2\right]. \tag{II.31}$$

Since all methods are evaluated on the same data set, $E\left[(\epsilon_i^{time})^2\right]$ is the same for all considered methodologies. While the time error component biases the measurement of the full method-specific MSE upwards, a comparison of standardization methodologies is still valid for the intraday set-up, as methodologies with better noise cancellation and no biases still possess a lower expected MSE.

II.6 The Accuracy of Volatility Surfaces

This section summarizes key findings of our empirical assessment of the statistical quality of different state-of-the-art volatility surfaces. We start with the end-of-day analysis for S&P 500 and Euro Stoxx 50 options. The section ends with the presentation of our findings for the intraday analysis.

II.6.1 End-of-day: Data-rich Environment

The aggregated RMSEs for the data-rich end-of-day environment are summarized in table II.6, the respective MAE error figures can be found in table II.7. For S&P 500 options we find that the volatility surface of the one-dimensional kernel regression produces the lowest RMSE (0.0092) and MAE (0.0019). The spline interpolation's volatility surface ranks second with an overall RMSE (MAE) that is 118% (91%) higher. The three-dimensional kernel regression ranks third place, with an RMSE (MAE) that is roughly 6 (12) times larger, relative to the one-dimensional kernel regression. The volatility surface of the Gram-Charlier expansion turns out to be the least accurate, producing a RMSE (MAE) that is more than 9 (23) times larger than the most accurate volatility surface.

Looking at the RMSE results for the Euro Stoxx 50 confirms that the volatility surface that is generated by the one-dimensional kernel regression is the most accurate one, followed by the spline-based volatility surface, whose overall RMSE is 37% higher. However, the ranking is reversed with respect to the MAE, where the spline method performs best and the kernel regression shows a 29% higher error. This reversal is the result of a downward bias of the one-dimensional kernel regression for deep out-of-the-money medium-term options with a moneyness below 0.5. If these options were to be excluded from the error calculation, the one-dimensional kernel regression and the spline method would produce roughly the same MAE for Euro Stoxx 50 data. Finally, the three-dimensional kernel regression and the Gram-Charlier expansion produce RMSEs (MAEs) that are 14 (25) and 12 (24) times higher, respectively.

Table II.6: End-of-day Cross-Validation RMSE

(a) S&P 500

Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**
all	578%	100% (0.0092)	228%	927%
short				
center	397%	100% (0.0082)	373%	362%
left	1036%	100% (0.0088)	513%	1643%
right	201%	101%	100% (0.0109)	221%
medium				
center	235%	131%	210%	100% (0.0084)
left	622%	100% (0.0102)	138%	1039%
right	192%	100% (0.0059)	110%	232%
long				
center	249%	100% (0.0022)	150%	164%
left	515%	116%	100% (0.0066)	1073%
right	185%	100% (0.0052)	120%	209%

(b) Euro Stoxx 50

Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**
all	1397%	100% (0.0100)	137%	1220%
short				
center	371%	100% (0.0129)	398%	155%
left	1340%	100% (0.0395)	181%	506%
right	3369%	113%	100% (0.0120)	5204%
medium				
center	125%	100% (0.0010)	460%	328%
left	2521%	219%	100% (0.0059)	1892%
right	4021%	152%	100% (0.0022)	4337%
long				
center	164%	109%	100% (0.0020)	226%
left	1302%	216%	100% (0.0015)	2268%
right	755%	107%	100% (0.0013)	1169%

This table summarizes the cross-validation RMSE for the data-rich (end-of-day) set-up overall and by region of the implied volatility surface. The *short* maturity section groups options with remaining maturities of less than 30 days, *medium* maturity refers to a remaining maturity of 30 to 365 days and *long* maturity represents options with a remaining maturity of more than 365 days. The moneyness $m = \frac{K}{F}$ axis is divided into a *left*, *center* and *right* section, with the cutting lines being located at 0.95 and 1.05. Error figures are given in percent of the respective best-performing method's RMSE for each section of the implied volatility surface. For each section's best performing method, we report its RMSE in parenthesis.

Table II.7: End-of-day Cross-Validation MAE

(a) S&P 500

Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**
all	1174%	100% (0.0019)	191%	2370%
short				
center	506%	100% (0.0013)	483%	655%
left	1844%	100% (0.0032)	379%	3535%
right	377%	100% (0.0037)	112%	375%
medium				
center	299%	100% (0.0005)	370%	561%
left	1487%	100% (0.0021)	121%	3275%
right	249%	100% (0.0018)	114%	432%
long				
center	252%	100% (0.0007)	195%	303%
left	706%	100% (0.0024)	128%	1402%
right	349%	100% (0.0014)	131%	490%

(b) Euro Stoxx 50

Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**
all	2536%	129%	100% (0.0015)	2435%
short				
center	309%	100% (0.006)	155%	104%
left	2618%	100% (0.0123)	104%	1184%
right	7037%	100% (0.0035)	122%	9905%
medium				
center	127%	100% (0.0005)	217%	422%
left	4284%	265%	100% (0.0014)	4570%
right	3262%	112%	100% (0.0012)	3419%
long				
center	165%	100% (0.0010)	106%	284%
left	1401%	119%	100% (0.0008)	2066%
right	640%	100% (0.0006)	116%	1388%

This table summarizes the cross-validation MAE for the data-rich (end-of-day) set-up overall and by region of the implied volatility surface. The *short* maturity section groups options with remaining maturities of less than 30 days, *medium* maturity refers to a remaining maturity of 30 to 365 days and *long* maturity represents options with a remaining maturity of more than 365 days. The moneyness $m = \frac{K}{F}$ axis is divided into a *left*, *center* and *right* section, with the cutting lines being located at 0.95 and 1.05. Error figures are given in percent of the respective best-performing method's MAE for each section of the implied volatility surface. For each section's best performing method, we report its MAE in parenthesis.

We now continue to report the RMSE for different regions of the option-implied volatility surface. We split the option-implied volatility surface along the maturity and moneyness dimension. Options with a maturity of less than 30 calendar days are considered short-term, options with a maturity between 30 and 365 calendar days are considered medium-term, while options with a maturity of more than one year are called long-term. Along the moneyness axis, we label a moneyness of 0.95 to 1.05 as 'center', whereas the 'left tail' ('right tail') is characterized by a moneyness of below (above) 0.95 (1.05). In combination, these splits yield nine different regions of the option-implied volatility surface, for which we calculate the RMSE and MAE of each method separately.

Looking at panels (a) and (b) of tables II.6 and II.7 highlights that the short-term and left tail regions of the volatility surface appear to be hardest to capture as all methods produce the highest errors here. For the S&P 500, the one-dimensional kernel regression performs best in most sections of the implied volatility surface with respect to the RMSE and in all sections with respect to the MAE. Spline interpolation tends to outperform the one-dimensional kernel regression for medium- and long-term options for the Euro Stoxx 50, though. This result is more pronounced in the RMSE figures than in the MAE figures. The three-dimensional kernel regression and the Gram-Charlier expansion both show severe difficulties in capturing the left and the right tail of the surface. The problem of the Gram-Charlier volatility surface is that the parametric risk-neutral density approximation turns out to be insufficient for capturing market information in the tails. We identify that the relative weakness of the three-dimensional kernel regression is that it only considers options with a delta of 0.2 to 0.8, which ignores market information about the tails.

We proceed by comparing the smoothness of the volatility surfaces. Table II.8 displays our smoothness measures, the sum of squared second derivatives of a volatility smile. A smooth volatility surface has low second derivatives along the strike dimension and thus a low smoothness measure. Conversely, a high smoothness measure is an indicator for more curved volatility smiles. The volatility surface of the three-dimensional kernel regression is the least smooth in our tests. On the other end, the one-dimensional kernel regression appears to be smoother than most of the alternatives. For the spline interpolation, there is an interesting divergence in the smoothness measure between its comparably rough S&P 500 volatility surface and its very smooth Euro Stoxx 50 surface. This goes hand in hand with the low errors of the spline interpolation in constructing the Euro Stoxx 50 surface. It seems like the spline interpolation is not able to cancel out all noise in the S&P

Table II.8: Smoothness of the Volatility Surface

Method	S&P 500	Euro Stoxx 50
3D Kernel Regression	1657 (115)	3315 (213)
1D Kernel Regression	600 (55)	303 (24)
Spline Interpolation	1402 (407)	196 (46)
Gram-Charlier Expansion	623 (46)	1096 (73)

This table shows the average smoothness of the constructed implied volatility surface for each method. For each day and maturity, we sum up the squared second derivative along the moneyness dimension, aggregate across all days and maturities by taking the mean and report the root of this figure. The second derivative is calculated for a moneyness interval between -3 and 3 times the unconditional standard deviation of the underlying, de-annualized to the respective maturity. Due to the existence of rare outliers, we remove the 1% largest and smallest single smoothness values from the calculation. Standard errors are given in parenthesis.

500 data, which produces a more curved volatility surface and higher fitting errors. At the same time, a smooth volatility surface does not necessarily imply low errors, as can be seen from the S&P 500 results for the Gram-Charlier expansion. That method might produce a smooth volatility surface, though it does not accurately reflect the information in observed option prices as can be seen from its high error figures.

In short summary, the cross-validation of the data-rich environment recommends to use a volatility surface that was constructed based on the one-dimensional kernel regression. The spline interpolation is a good alternative for Euro Stoxx 50 data, though appears to not be able to cancel out all noise in the S&P 500 data. The volatility surface of the Gram-Charlier expansion should only be applied if one is interested in option-implied information from the at-the-money region and should be avoided when inferring conclusions about the tails. The three-dimensional kernel regression produces RMSEs that relative to the one-dimensional kernel regression are roughly 6 times larger for S&P 500 options and roughly 14 times larger for Euro Stoxx 50 options.

II.6.2 Intraday: Data-poor Environment

Here, we summarize key findings for the intraday analysis. For the non-Bayesian methods, we start with the highest time resolution (3 trades), followed by interpolations based on 4, 5 and 10 trades with distinct strikes. We compare the performance of these methods with the intraday Bayesian volatility surface that is updated on each trade. The Bayesian volatility surface requires the estimation of state and measurement variances within the Kalman filter model. However, these parameters are hard to pin down in practice (Mehra, 1970; Matisko and Havlena, 2012). We therefore investigate a second specification of the Bayesian volatility surface, where the state variances are not estimated at all, but for each day set fix to the historically observed variances of the principle components of the end-of-day volatility surfaces of the 2 years preceding that day. The initial state variance is then set to 5 times the historical state variance.[17] The only parameter that is left to estimate in this specification is the measurement variance. While this second specification is certainly not optimal, it can be estimated much faster and is much less prone to estimation errors. We call this second specification the *rule-of-thumb* Bayesian volatility surface, while the first specification of section II.2.3 is called the *optimal* Bayesian volatility surface.

Our comparison of the accuracy of the intraday methods can be found in tables II.9 and II.10. Our findings show that the Bayesian approach produces the most accurate intraday volatility surfaces across all assets, error measures and regions of the volatility surface. Overall, the best non-Bayesian approach shows a more than 7 (6) times higher RMSE than the best Bayesian specification for the S&P 500 (Euro Stoxx 50). The magnitude of the performance differential is similar with respect to the MAE. Especially the left-tail regions of the implied volatility surface are captured much more accurately by the Bayesian volatility surfaces, as can be seen by the highly increased errors of the non-Bayesian methods here. The Bayesian volatility surface appears to benefit from the high liquidity in left-tail options and manages to transmit this information to better describe less liquid sections of the implied volatility surface. Among the two Bayesian volatility surface specifications, the version with fixed state variances appears to be roughly at par with the optimized state variances. For the S&P 500, the *rule-of-thumb* version shows even lower errors, mainly due to the comparably bad performance of the *optimal* version in the right tail. The *optimal* version is leading in most regions of the implied volatility surface of both equity indices with respect to the MAE, while the picture is more disperse with respect to the RMSE. In most cases, the performance between the two versions of the Bayesian volatility surface is similar. Overall, we conclude that one can benefit from the higher speed and estimation stability of the *rule-of-thumb* Bayesian volatility surface, without having to sacrifice much accuracy.

Among the non-Bayesian methods, the one-dimensional kernel regression produces the most accurate overall volatility surface. The three-dimensional kernel regression is the

---

[17] Our tests show that variations in the initial state variance do not have a significant impact on the performance of the Bayesian volatility surface. We increase the initial state variance with respect to the overall state variance since it captures the over-night innovation in the states, between the previous day's end-of-day volatility surface and the volatility surface of the first trade of the current day.

Table II.9: Intraday Cross-Validation RMSE

(a) S&P 500

	Non-Bayesian				**Bayesian**	
Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
all	835%	722%	2747%	917%	129%	100% (0.0098)
short						
center	531%	379%	1659%	660%	100% (0.0098)	101%
left	1792%	1703%	5898%	1811%	100% (0.0085)	118%
right	316%	445%	834%	775%	103%	100% (0.0158)
medium						
center	483%	401%	1685%	660%	235%	100% (0.0079)
left	1463%	1324%	6241%	1608%	104%	100% (0.0066)
right	361%	399%	880%	703%	196%	100% (0.0127)
long						
center	347%	270%	813%	608%	100% (0.017)	102%
left	1645%	1267%	5818%	1757%	100% (0.0059)	117%
right	949%	729%	2361%	1663%	245%	100% (0.0087)

(b) Euro Stoxx 50

	Non-Bayesian				**Bayesian**	
Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
all	673%	606%	1757%	924%	100% (0.0081)	104%
short						
center	330%	228%	380%	439%	100% (0.0108)	108%
left	1282%	1253%	2446%	1371%	100% (0.0067)	109%
right	502%	547%	734%	906%	100% (0.0087)	110%
medium						
center	792%	628%	1861%	1433%	105%	100% (0.0049)
left	1340%	1237%	5235%	1757%	100% (0.0050)	103%
right	839%	788%	1666%	1600%	118%	100% (0.0058)
long						
center	1059%	669%	2952%	2063%	100% (0.0064)	127%
left	1628%	1238%	7637%	1886%	279%	100% (0.0052)
right	1398%	1107%	3042%	2616%	151%	100% (0.0062)

This table summarizes the out-of-sample RMSE for the data-poor (intraday) set-up overall and by region of the implied volatility surface. The *short* maturity section groups options with remaining maturities of less than 30 days, *medium* maturity refers to a remaining maturity of 30 to 365 days and *long* maturity represents options with a remaining maturity of more than 365 days. The moneyness $m = \frac{K}{F}$ axis is dividend into a *left*, *center* and *right* section, with the cutting lines being located at 0.95 and 1.05. Evaluations are performed based on 3 prior trades on different strikes and the same maturity for the non-Bayesian methods. Error figures are given in percent of the respective best-performing method's RMSE for each section of the implied volatility surface. For each section's best performing method, we report its RMSE in parenthesis.

Table II.10: Intraday Cross-Validation MAE

(a) S&P 500

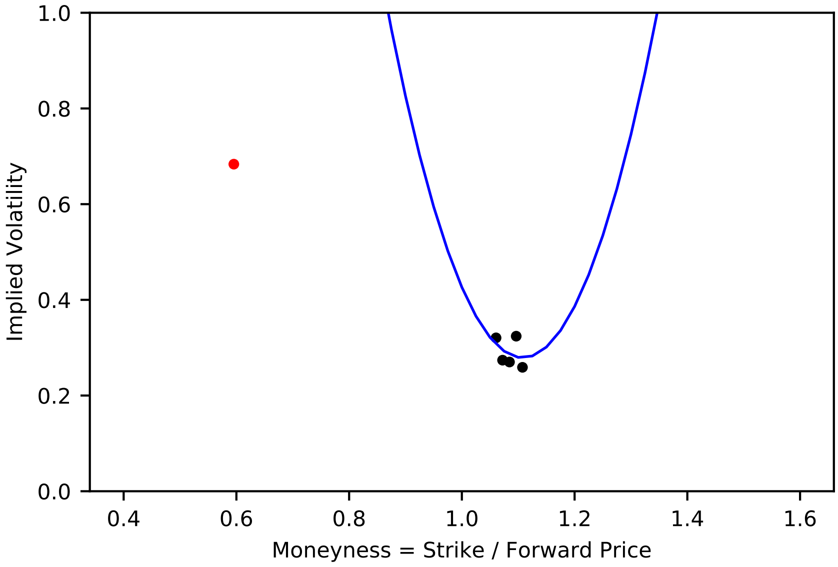
	Non-Bayesian				**Bayesian**	
Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
all	796%	649%	904%	892%	103%	100% (0.0048)
short						
center	482%	399%	465%	623%	100% (0.0048)	105%
left	2196%	1821%	2814%	2118%	100% (0.0045)	112%
right	289%	339%	470%	691%	100% (0.0088)	105%
medium						
center	469%	454%	487%	613%	143%	100% (0.0036)
left	1880%	1604%	2739%	2109%	100% (0.0027)	112%
right	317%	418%	526%	731%	149%	100% (0.0057)
long						
center	999%	870%	1277%	1813%	100% (0.0026)	109%
left	2086%	1457%	2297%	2241%	100% (0.0026)	120%
right	912%	907%	1501%	1932%	166%	100% (0.0035)

(b) Euro Stoxx 50

	Non-Bayesian				**Bayesian**	
Region	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
all	672%	627%	621%	860%	100% (0.0032)	115%
short						
center	299%	283%	182%	347%	100% (0.0043)	116%
left	1357%	1144%	887%	1279%	100% (0.0030)	122%
right	281%	342%	382%	579%	100% (0.0041)	122%
medium						
center	676%	672%	624%	1059%	100% (0.0023)	111%
left	1548%	1368%	1779%	1805%	100% (0.0022)	114%
right	547%	705%	849%	1319%	100% (0.0027)	111%
long						
center	970%	927%	1223%	2283%	100% (0.0026)	113%
left	1484%	1247%	3101%	2154%	104%	100% (0.0023)
right	1289%	1191%	1652%	2743%	100%	100% (0.0028)

This table summarizes the out-of-sample MAE for the data-poor (intraday) set-up overall and by region of the implied volatility surface. The *short* maturity section groups options with remaining maturities of less than 30 days, *medium* maturity refers to a remaining maturity of 30 to 365 days and *long* maturity represents options with a remaining maturity of more than 365 days. The moneyness $m = \frac{K}{F}$ axis is dividend into a *left*, *center* and *right* section, with the cutting lines being located at 0.95 and 1.05. Evaluations are performed based on 3 prior trades on different strikes and the same maturity for the non-Bayesian methods. Error figures are given in percent of the respective best-performing method's MAE for each section of the implied volatility surface. For each section's best performing method, we report its MAE in parenthesis.

Figure II.4: Example for Over-Fitting in Spline Interpolation



This figure visualizes the severity of the over-fitting in the spline interpolation methodology for a particular data point of the Euro Stoxx 50 sample, namely 30-day options on May 16th, 2012 (01:35pm). The black dots represent the 5 previous trade observations based on which the spline interpolation constructs the spline, which is drawn in blue. The red dot marks the next trade observation.

second most accurate surface, though the performance differential between the two is not as large as for the end-of-day set-up. In our sample, we find that the spline-based volatility surface suffers from frequent outliers in the data, which increases the RMSE enormously. This is not surprising, because the spline interpolation is prone to over-fitting when it is applied to very few observations. Figure II.4 visualizes the potential problem of over-fitting when constructing the spline-based volatility surface with 5 or less data points. As the MAE is less responsive to occasional large errors, which are a result of over-fitting, the spline interpolation performs better with respect to the MAE than with respect to the RMSE. It even shows a slightly lower MAE than the one-dimensional kernel regression for the Euro Stoxx 50 intraday set-up with 3 trades. The volatility surface of the Gram-Charlier expansion ranks always in the back, though with a much smaller error differential to the other non-Bayesian methods than in the end-of-day set-up. There does not appear to occur any over-fitting in the Gram-Charlier expansion, however in all of our tests, the general fit seems to be worse than for the kernel regression methods.

For the non-Bayesian methods, basing an implied volatility estimate on only 3 price observations is certainly extreme. We therefore report our main findings for building intraday volatility surfaces with 4, 5 and 10 price observations per maturity, respectively,

Table II.11: Intraday Cross-Validation Errors for Varying Amount of Trades Used

(a) S&P 500

	Traditional				**Bayesian**	
Interpolation Base	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
RMSE					129%	100% (0.0098)
3 trades	835%	722%	2747%	917%		
4 trades	766%	639%	1645%	846%		
5 trades	730%	578%	1135%	807%		
10 trades	640%	431%	514%	715%		
MAE					103%	100% (0.0048)
3 trades	796%	649%	904%	892%		
4 trades	707%	532%	527%	788%		
5 trades	655%	450%	388%	734%		
10 trades	514%	273%	238%	635%		

(b) Euro Stoxx 50

	Traditional				**Bayesian**	
Interpolation Base	**3D Kernel Regression**	**1D Kernel Regression**	**Spline Interpolation**	**Gram-Charlier Expansion**	**Optimal**	**Rule-of-thumb**
RMSE					100% (0.0081)	104%
3 trades	673%	606%	1757%	924%		
4 trades	646%	536%	847%	893%		
5 trades	627%	492%	567%	895%		
10 trades	636%	393%	321%	882%		
MAE					100% (0.0032)	115%
3 trades	672%	627%	621%	860%		
4 trades	634%	528%	396%	831%		
5 trades	604%	467%	333%	828%		
10 trades	574%	341%	291%	853%		

This table summarizes out-of-sample RMSE and MAE for different amount of trades used for the implied volatility surface. Error figures are given in relative terms towards the lowest respective error for each asset and error measure. This lowest error is given in paranthesis.

thus sacrificing time resolution in favor of increased precision of the estimated volatility surfaces. Table II.11 summarizes the results and restates the performance of the Bayesian methods for comparison. As one would expect, adding more observations as input for the interpolations, improves the statistical accuracy for all methods. Clearly, the spline-based volatility surface benefits most, since the higher amount of observations reduces the tendency of over-fitting. Its RMSE differential towards the best method is falling rapidly as the volatility surface is based on more price observations. With the decreasing occurrence of over-fitting, spline interpolation quickly passes the one-dimensional kernel regression with respect to the MAE and even out-performs the other non-Bayesian approaches with respect to the RMSE in Euro Stoxx 50 options, when using 10 trades for the volatility surface. Still, even with 10 trades at hand, no method can get close to the performance of the Bayesian volatility surfaces. We conclude that the Bayesian approach is generally superior to its alternatives for the intraday set-up. In set-ups where the Bayesian approach is not an option, such as data-poor end-of-day environments, the spline interpolation is a good alternative for the one-dimensional kernel regression, if an intermediate amount of price observations is used and if the occasional occurrence of an over-fitted volatility smile is acceptable.

II.7 Summary

This chapter makes three contributions to the literature. First, we are the first to document that option-implied information, such as well-known forward-looking measures for variance, skewness and the variance risk premium, are sensitive to the choice of the underlying volatility surface. These findings hold more generally for any quantity that is extracted from the aggregation of option prices along the strike range.

Second, we propose a novel methodology for estimating the implied volatility surface tick-by-tick, based on a Bayesian filtering model. For each maturity, the method maintains an internal estimate of the volatility smile, which is updated each time a new trade observation for that maturity becomes available. Crucially, the new observation causes an update of the whole implied volatility smile, such that the implied volatility estimates at different strikes than the observed trade's strike are affected, too.

Third, we implement a thorough statistical assessment of the accuracy of common parametric, semi-parametric and non-parametric approaches that the literature has entertained for constructing volatility surfaces at the end-of-day and intraday frequency. For the intraday frequency, this assessment is expanded to our proposed Bayesian volatility surface. The non-Bayesian methods under consideration are the Gram-Charlier expansion, the spline interpolation of Figlewski (2008), the three-dimensional kernel regression of OptionMetrics (2016) and the one-dimensional kernel regression with a linear tail extrapolation in the end-of-day setting. We have recorded the root mean squared error (RMSE) and mean absolute error (MAE) based on a leave-one-out cross-validation for each method and have compared the smoothness of the constructed end-of-day volatility surfaces directly. The test assets are S&P 500 and Euro Stoxx 50 options at the daily and intraday frequency over 14 years.

The result of our analysis is that the volatility surface that is constructed with the one-dimensional kernel regression is generally the most accurate one for end-of-day options on the S&P 500 and the Euro Stoxx 50. We recommend to use that volatility surface for extracting option-implied information at the daily frequency. For the intraday set-up, the proposed Bayesian volatility surface appears to be highly superior to the alternatives, even if these alternatives form their estimates of the volatility surface on an intermediate amount of trade observations. If the Bayesian volatility surface is not an option, for example for data-poor end-of-day set-ups, the one-dimensional kernel regression showed the highest accuracy again if only very few observations are available. However, if an intermediate amount of observations is available, spline interpolation might be a preferable alternative, despite the occasional occurrence of over-fitting. The parametric Gram-Charlier volatility surface is in many cases too restrictive to approximate the true risk-neutral distribution and thus the true volatility surface accurately. The three-dimensional kernel regression of OptionMetrics (2016) performs only slightly worse than the one-dimensional kernel regression in the intraday analysis, but shows severe shortcomings in capturing the volatility surface for out-of-the-money Puts at the end-of-day frequency.

III The Physical Return Distribution

$\mathcal{P}$, the conditional physical probability density function, is the key input to every application of Expected Utility Theory, which itself builds the foundation to any modern economic decision making problem.[1] For instance, any valuation problem is solved by finding the expected risk-adjusted present value of future payoffs, where the expectation is taken with regard to $\mathcal{P}$.[2] Analogous, the optimal solution to any consumption and investment problem depends crucially on $\mathcal{P}$.[3] The vast majority of the literature follows parametric or non-parametric backward-looking approaches to identify $\mathcal{P}$.[4] One serious limitation of backward-looking methodologies is their limited informativeness about forward-looking events in the presence of regime shifts or non-stationary data.[5] A strand of the literature has therefore relied on the seminal work of Ross (2015) and Hansen and Scheinkman (2009) to identify $\mathcal{P}$ with forward-looking option data and parametric assumptions on the representative investor's marginal utility.[6] Despite its theoretical appeal, recent empirical evidence for the S&P 500 concludes that the forward-looking $\mathcal{P}$ density of returns cannot be accurately recovered based on Ross (2015).[7] We suggest a third strategy for estimating $\mathcal{P}$, which, similar to the important contribution of Linn et al. (2018), is a pure econometric approach that relies only on minimal mathematical continuity and differentiability assumptions and which hence can be considered free of

---

[1] In this work, our notion of $\mathcal{P}$ does not explicitly differentiate between objective and subjective probabilities. Early seminal work that discusses both concepts are Ramsey (1931), de Finetti (1937), von Neumann and Morgenstern (1947) and Savage (1954).

[2] For early work of this seminal concept see Gordon (1962), Lucas Jr (1978), Mehra and Prescott (1985), Hansen and Singleton (1982), Hansen and Singleton (1983). Classical textbook treatments are Cochrane (2001) and Duffie (2001).

[3] See Merton (1969), Merton (1975) for seminal work in optimal consumption and portfolio planning.

[4] Noteworthy applications of parametric models on historical returns are Rosenberg and Engle (2002), Barone-Adesi et al. (2008), Barone-Adesi and Dall'O (2010). Important non-parametric kernel density estimators on historical returns are Jackwerth (2000), Jackwerth (2004) and Aït-Sahalia and Lo (1998).

[5] Thorough explanations of this concern can be found in Bliss and Panigirtzoglou (2004) and Linn et al. (2018) and Cuesdeanu and Jackwerth (2018a).

[6] See Schneider and Trojani (2019), Jensen et al. (2019), Jackwerth and Menner (2018), Borovička et al. (2016), Carr and Yu (2012) and Walden (2017) for recent generalizations of Ross (2015).

[7] See Jackwerth and Menner (2018) and Dillschneider and Maurer (2018).

parametric assumptions of the underlying economy.

Our econometric methodology follows the economic idea of Ross (2015) to back out $\mathcal{P}$ from inherently forward-looking option prices. Yet, instead of relying on simplifying assumptions on an investor's marginal utility, we do instead borrow from a well-developed machine learning literature, which applies Mixture Density Networks (MDN) to approximate any conditional probability density function to arbitrary precision (Bishop, 1994). This allows us to extract $\mathcal{P}$ from option data with only minimal continuity and differentiability requirements and to remain agnostic about the underlying structure of the economy and preferences. Our method can also be applied to determine which factors are truly useful in identifying $\mathcal{P}$. Our analysis with S&P 500 return and option data also documents that backward-looking information such as past S&P 500 returns, returns of the Fama and French (1993) value and size factor, momentum returns and their respective realized volatilities do not add noteworthy information about $\mathcal{P}$ that is not yet part of option-implied return moments.

Our option-implied MDN approach works as follows: We assume that the conditional probability of next day's S&P 500 return is a weighted sum of Gaussian distributions, which introduces the aforementioned continuity and differentiability conditions on $\mathcal{P}$. As conditioning variables we use the risk-free rate and two sets of factors. One set of factors captures forward-looking option-implied risk-neutral moments of the S&P 500; namely (i) the $SVIX^2$ from Martin (2017), (ii) the risk-neutral skewness and (iii) risk-neutral kurtosis from Bakshi et al. (2003). The second set of factors captures a subset of classical backward-looking return factors such as the return of the value, size and momentum factor and their respective 10-trading-day rolling window variances. Each mean and variance of the Gaussian mixture model as well as the weight that each Gaussian density obtains is allowed to be a function of the conditioning variables. It is key that we are agnostic about the type of function that these quantities follow. Instead, we approximate these functions by a feed-forward neural network with one hidden layer of neurons and let the neural network learn from the data, how to best approximate these functions and hence, $\mathcal{P}$. A detailed exposition of how to design well-specified MDNs and how to estimate them can be found in (Rothfuss et al., 2019a) and (Rothfuss et al., 2019b).

We confront our methodology with end-of-day S&P 500 return and CBOE option data, spanning the period January 2004 to July 2017, to answer five questions. First, do backward-looking factor returns add information about $\mathcal{P}$ that is not yet spanned by forward-looking option-implied information? In order to answer this question we compare the in-sample and out-of-sample log likelihood values of three MDNs which only differ with regard to the set of conditioning variables. One MDN conditions on forward-looking option data and the risk-free rate only. A second one conditions on backward-looking factor returns and their historical volatility. A third MDN conditions on both types of information, forward-looking option data including the risk-free rate and backward-looking factor returns and their historical volatilities. When comparing the respective log likelihood values, we follow Welch and Goyal (2008) and compare all likelihood values to a baseline case, which we specify to be a non-parametric kernel density estimator (Jackwerth, 2000; Aït-Sahalia and Lo, 1998) of the unconditional return distribution. Our analysis concludes that a MDN with only forward-looking option-implied information is sufficient to beat the non-parametric kernel density estimator by a margin of roughly 5%, both in- and out-of-sample. While adding backward-looking factor information further increases the in-sample log likelihood by a relative margin of 0.6%, the out-of-sample relative increase in the log likelihood is only 0.04%. We hence conclude that the backward-looking factor returns and their historical volatilities do not add noteworthy information about $\mathcal{P}$ and can hence be easily skipped when building a forward-looking $\mathcal{P}$ estimate.

The second question of this chapter is to test whether each of the four considered $\mathcal{P}$ estimates could be the data generating process for the realized time series of daily S&P 500 returns. We follow Jackwerth and Menner (2018) and apply a Berkowitz (2001) and Knüppel (2015) test. Based on out-of-sample data, we reject that the non-parametric kernel density estimator and the backward-looking MDN are proper characterizations of $\mathcal{P}$. We fail to reject the hypothesis for the forward-looking MDN. These tests also show a text book like pattern for over-fitting when working with backward-looking $\mathcal{P}$ estimates. All of the considered backward-looking densities cannot be rejected using in-sample data, yet, they do not generalize well to previously unseen out-of-sample data and are strongly rejected here.

Based on the first two research questions we conclude that the forward-looking $\mathcal{P}$ estimate that combines a MDN with forward-looking option data and the risk-free rate provides an accurate statistical description of daily S&P 500 returns. As a third research question we want to understand whether our forward-looking $\mathcal{P}$ estimate is useful from a financial economic point of view. To assess that we implement two dynamic trading strategies that rely on $\mathcal{P}$ as a signal for trading and compare the resulting Sharpe ratio to a static trading strategy that does not rely on $\mathcal{P}$. One strategy goes long (short) the S&P 500 on days where the forward-looking expected return is positive (negative). We compare the Sharpe ratio of this strategy to a buy-and-hold strategy. The second trading strategy shorts delta-neutral straddles and closes the position after one day. We increase the short position on days where the conditional forward-looking variance expectation is falling and we reduce it for days where the variance forecast is rising. We compare the outcome of the straddle timing strategy to a static short straddle strategy. With regard to the first trading strategy, we document an in-sample (out-of-sample) Sharpe ratio of 0.69 (0.74), relative to a 0.39 (0.65) Sharpe ratio of the static buy-and-hold strategy. We document an in-sample (out-of-sample) Sharpe ratio for the dynamic straddle strategy of 4.9 (5.4), relative to a 2.7 (3.2) Sharpe ratio for the static short straddle strategy. Based on the in-sample and out-of-sample Sharpe ratio results, we conclude that the forward-looking $\mathcal{P}$ density carries economically meaningful information about the time series properties of daily stock returns.

Our fourth research question aims to understand which of the forward-looking option-implied conditioning variables are especially important for the forward-looking $\mathcal{P}$ density. We perform an adjusted Patton and Timmermann (2010) test to assess the impact of a particular option-implied conditioning variable onto $\mathcal{P}$. Our findings highlight that all of the considered option-implied variables are informative about $\mathcal{P}$ at the 1% significance level. From all option-implied quantities, we find $SVIX^2$ to have the strongest effect on the ex-ante mean, variance, skewness and kurtosis of $\mathcal{P}$. Bakshi et al. (2003) option-implied skewness and kurtosis are only important for pinning down the forward-looking $\mathcal{P}$ skewness and kurtosis.

The adjusted Patton and Timmermann (2010) test does not reveal how the option-implied moments feed into $\mathcal{P}$. The fifth research question does therefore aim to learn from the feed-forward neural network. We follow Davison and Hinkley (1997) and construct confidence intervals for the predictive relationship that the neural-network-based MDN implies. With regard to VIX^2, we identify a close to linear positive predictive relationship to the next day's expected return, well in line with the economic model of Martin (2017). Also of interest is the observation of a positive relationship between $SVIX^2$ and the expected $\mathcal{P}$ variance. The identified relationship is linear for days where the annualized $SVIX$ is above 17%. As to the relationship between $SVIX^2$ and forward-looking $\mathcal{P}$ skewness we document that as option-implied volatility drops from an annualized value of 17.3% to 12.2%, next day's $\mathcal{P}$ skewness reduces from roughly 0 to -0.62.

Our work is closest related to the young, yet fast growing and influential, literature on estimating $\mathcal{P}$ from option data. Ross (2015) develops an economic technique to recover the physical return density from its risk-neutral counterpart. His key economic restriction is that the representative investor's ratio of marginal utility between two states is transition independent and thus constant over time. The empirical success of the Ross (2015) theorem is mixed. Audrino et al. (2015) show that a trading strategy with signals extracted from Ross (2015) recovered $\mathcal{P}$ moments outperforms trading signals from risk-neutral moments. On the other hand, Jackwerth and Menner (2018) apply a series of statistical tests on the Ross (2015) implied $\mathcal{P}$ density. The authors reject the hypothesis that realized S&P 500 returns are drawn from the Ross (2015) implied $\mathcal{P}$ density. Jackwerth and Menner (2018) identify that a key challenge to the empirical success of the Ross (2015) theorem is the difficulty to obtain the required transition state prices from option price data. Findings in Bakshi et al. (2018) for options on 30-year Treasury bond futures do also challenge the adequacy of the Ross (2015) required pricing kernel.

Jensen et al. (2019) generalize the work of Ross (2015). The authors replace the time-homogeneity restriction on the pricing kernel with a weaker time-separability constraint. The authors show that this extension improves the resulting accuracy of the implied $\mathcal{P}$ volatility forecast, yet it still does not pass a Berkowitz (2001) test. In contrast to these important contributions, we show how to transform option-implied information to $\mathcal{P}$ without relying on the economically important, yet empirically difficult to measure, concept of a pricing kernel. Our approach is only data-driven and uses a small-scale feed-forward neural network as part of a MDN to uncover $\mathcal{P}$ from a panel of option prices.

Our work is also related to the model-free $\mathcal{P}$ recoveries of Schneider and Trojani (2019) and Linn et al. (2018). Schneider and Trojani (2019) use economically motivated sign restrictions on tradable higher moment risk premiums to derive constraints on the physical conditional moments of returns. Their recovered $\mathcal{P}$ estimate is free of technical assumptions on the underlying economy and shown to predict S&P 500 returns. The model-free approach of Linn et al. (2018) is an innovative econometric method that estimates $\mathcal{P}$ using the forward-looking option-implied density and the inverse of the Radon-Nikodym derivative. The authors' approach relies on a finite order cubic B-spline to approximate the inverse of the Radon-Nikodym derivative with a set of time-varying option-implied densities and return realizations that are sampled from the corresponding $\mathcal{P}$ density. Their model-free estimate for $\mathcal{P}$ shares the same information set as our option-implied forward-looking density and similar to Linn et al. (2018), our approach of recovering $\mathcal{P}$ is a pure econometric approach. Different to the previous two contributions however, we do not identify $\mathcal{P}$ based on the conceptually important, though rather indirect route via a pricing kernel. Instead, we directly approximate the unknown $\mathcal{P}$ distribution by a small-scale MDN with a feed-forward neural network that conditions on option-implied information.

Our work also adds to the growing literature that uses machine learning and neural network techniques in finance applications. Early work in this field includes Hutchinson et al. (1994) and Yao et al. (2000) who use neural networks to price options on the S&P 500 and Nikkei 225 futures. More recently, Ludwig (2015) documents several advantages when using a neural network to interpolate the option-implied volatility surface. Dunis et al. (2011) and Zhao et al. (2018) show how to use neural networks for portfolio formation and trading strategies. The influential study of Gu et al. (2019) compares a range of machine learning techniques for time series and cross-sectional return predictions. The authors conclude that well-performing machine learning techniques, such as neural networks, benefit from their ability to capture important nonlinearities in the data. We add to this literature by showing that machine learning techniques are not only useful for return predictions, but they also help improve our understanding on deep financial economic questions, such as how to use the rich cross-section of option data, which is available in real time, to learn about the full conditional return density under $\mathcal{P}$.

The concept of using MDNs and neural networks to approximate conditional probability density functions has been primarily developed in the computer science literature. The seminal work in that field is Bishop (1994). Recently, there has been a new interest in the literature to further improve on that technique. Rothfuss et al. (2019a) develop a noise regularization scheme for machine learning tools like MDNs to prevent over-fitting in applications that have to rely on small training samples. In our implementation of the MDN we test for over-fitting using their regularization technique. Rothfuss et al. (2019b) use a controlled simulation study to assess how different non-parametric conditional density estimators perform when asked to learn a specific density function. The authors conclude that a MDN dominates the other considered non-parametric density estimators. On a technical level, we differ from these studies by letting the training algorithm choose the size of the neural network, which allows the data to determine the proper degree of complexity and which renders noise regularization unnecessary in our case. On an economic level, we differ from these studies by using this well developed tool from the computer science literature to address one of the most fundamental questions in financial economics, namely how to find a good forward-looking estimate for the conditional density of daily S&P 500 returns.

III.1 Model-Free Return Density

In this section, we present our conditional density estimator, that is able (but not restricted) to form physical density estimates based on forward-looking information alone. We consider this estimator to be model-free in the sense that we do not restrict in any way the shape of the distribution or the relationship between the distribution's shape and the inputs that it is conditioned on. Instead, the estimator learns these characteristics during the training phase from the data. The estimator, called Mixture Density Network (MDN), combines a neural network with the class of mixture density models and has been developed by Bishop (1994).

Let $r_t = ln \frac{S_t}{S_{t-1}}$ be the log return of an asset's price at time t and $X_{t-1} = (x_{1,t-1}, ..., x_{i,t-1}, ..., x_{N,t-1})^\intercal \in \mathcal{I} \subseteq \mathbb{R}^N$ be a set of N predictor variables, that we can observe in $t-1$. We start by expressing the conditional physical log return density $p(r_t|X_{t-1})$ as a mixture density model, more precisely, as a Gaussian mixture model

$$p(r_t|X_{t-1}) = \sum_{k=1}^{K} \alpha_k(X_{t-1})\,\mathcal{N}(r_t|\mu_k(X_{t-1}), \sigma_k^2(X_{t-1})). \tag{III.1}$$

Here, the conditional density is constructed as the weighted sum of K Gaussian densities with respective weights $\alpha_k(X_{t-1})$, means $\mu_k(X_{t-1})$ and variances $\sigma_k^2(X_{t-1})$. The weights, means and variances of the mixture model are unknown functions of the input variables X_{t-1}. In order to form a valid density, it must hold $\sum_{k=1}^{K} \alpha_k(X_{t-1}) = 1$. As Bishop (2006) emphasizes, this specification is flexible enough to approximate almost any distribution to arbitrary precision, provided the number of Gaussians K is large enough.

A MDN uses a feed-forward neural network to jointly estimate $\alpha_k(X_{t-1})$, $\mu_k(X_{t-1})$ and $\sigma_k^2(X_{t-1})$. For robustness, we restrict ourselves to neural networks with one hidden layer of neurons. To formalize that, we express the distribution's parameters as a function of the input variables that we condition on: $f : \mathcal{I} \rightarrow \mathbb{R}_+^K \times \mathbb{R}^K \times \mathbb{R}_+^K$, $f(X_{t-1}) = (\alpha_1(X_{t-1}), ..., \alpha_K(X_{t-1}), \mu_1(X_{t-1}), ..., \mu_K(X_{t-1}), \sigma_1^2(X_{t-1}), ..., \sigma_K^2(X_{t-1}))^\intercal$ and write the j-th element of f as

$$f_j(X_{t-1}) = \sum_{h=1}^{H} w_{2,j,h}\, \phi \left(\sum_{i=1}^{N} w_{1,i,h}\, x_{i,t-1} \right) + w_{2,j,bias}. \tag{III.2}$$

Here, we assume that the input vector X_{t-1} already contains a constant element. In essence, the neural network, which is characterized by equation III.2, can be seen as a weighted sum of homogeneous functionals of the input variables. The difference between the functionals only lies in the weights that are assigned to each input. Note that these first-level weights $w_{1,i,h}$ are shared among all elements of f, while the second-level weights $w_{2,j,h}$ and $w_{2,j,bias}$ are element-specific. The function $\phi(\cdot)$, called *activation function*, should be continuous, bounded and non-constant and we follow common practice in the literature in choosing the tangens hyperbolicus function, i.e., $\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$.

As Hornik (1991) showed, this neural network specification is flexible enough to approximate any continuous function $f_j(X_{t-1})$ to arbitrary precision, provided that the number of hidden neurons H is large enough. Put differently, the neural network's parameters $w_{1,i,h}$, $w_{2,j,h}$ and $w_{2,j,bias}$ define the function $f_j(X_{t-1})$, which is selected from the full space of continuous functions in the model training phase. Beside the number of hidden neurons H, which we let our training algorithm choose freely, we impose no restrictions on $f_j(X_{t-1})$ in any way. In combination with the flexible density definition in equation III.1, the MDN can approximate any conditional distribution $p(r_t|X_{t-1})$ without prior restrictions about this distribution (Bishop, 1994).

III.1.1 Estimation

The parameters of our Gaussian mixture density from equation III.1 are subject to a number of natural constraints: The variances of the component densities $\sigma_k^2(X_{t-1})$ must be positive. Also, the weights $\alpha_k(X_{t-1})$ must be positive and add up to 1. Translating these restrictions into a set of constraints for the neural network's weights would be very challenging, if not impossible. Instead, we transform the weights and variances of the Gaussian components in equation III.1 to enforce fulfillment of the constraints without needing to constrain the neural network. More precisely, let $\tilde{\alpha}_k(X_{t-1})$ be the MDN output for a weight and $\tilde{\sigma}_k^2(X_{t-1})$ be the MDN output for the variance of a mixture component.

The actual weight and variance of that component are then set to

$$\sigma_k^2(X_{t-1}) = e^{\tilde{\sigma}_k^2(X_{t-1})}, \tag{III.3}$$

$$\alpha_k(X_{t-1}) = \frac{e^{\tilde{\alpha}_k(X_{t-1})}}{\sum_{i=1}^{K} e^{\tilde{\alpha}_i(X_{t-1})}}. \tag{III.4}$$

These transformations allow us to maintain an unconstrained output for the neural network while fulfilling the natural parameter constraints of a mixture density model.

We now collect a training data set $\{(r_t, X_{t-1})\}_{t\in[1,T]}$. As we are aiming for a full density estimator, we fit the neural network parameters $\theta = (w_{1,i,h}, w_{2,j,h}, w_{2,j,bias})$ by maximizing the conditional log likelihood of the observed data points:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{T} ln\, \hat{p}(r_t|X_{t-1};\theta). \tag{III.5}$$

There exists a large body of literature about the training of neural networks. Bishop (2006) and Géron (2017) provide a good introduction and overview of different training techniques. We use stochastic gradient descent with the adaptive learning rate method of Kingma and Ba (2015) to train our Mixture Density Network. In a nutshell, this local optimization technique repeatedly iterates over the complete training data set and adjusts the parameters step-by-step to draw closer to the optimal solution. The starting values for the weights of the MDN are set randomly. Within each iteration, the method starts in the beginning of the training data set, feeds the respective input data point into the neural network and computes the model-implied likelihood of the associated observed return. The optimizer now slightly adjusts the weights along their gradients of the likelihood function to increase the likelihood of that return in subsequent estimations. Afterwards, it proceeds to the next observation and repeats updating the weights until the end of the data set is reached. This process is repeated L times.

A central issue when training neural networks is determining the optimal degree of complexity of the model. Models with too limited capacity may not be a able to sufficiently capture the structure of the data, introducing a restriction bias. On the other hand, if a model is too expressive, it is prone to over-fit the training data, resulting in poor generalization. There are three hyper-parameters embedded in our approach that allow to adjust the complexity of the Mixture Density Network: the number of mixture components K, the number of hidden neurons H and the number of training iterations L.

Increasing any of these hyper-parameters makes the model more complex.

In order to pin down the hyper-parameters and find the best degree of complexity for our use case, we perform a cross-validation-based hyper-parameter grid search. First, we consider a set of discrete values for each hyper-parameter and construct a grid by forming all possible combinations of the respective hyper-parameter values. Next, we divide the training data set into 10 equally sized folds. For each hyper-parameter combination on the grid, we in turn leave out one of the folds and use the 9 remaining folds for training. We now train the MDN 5 times on this trimmed training data set with different starting values for the weight optimization. After each training, we record the average conditional log likelihood for the returns in the fold that was left out. We arrive at a performance figure for a given hyper-parameter combination by taking the average log likelihood across all folds and starting values. Finally, we choose the hyper-parameter combination with the highest performance, that is, the highest cross-validation conditional likelihood. If any of the selected hyper-parameters is located at one of the borders of the grid, i.e., if the lowest or highest value of the discrete value set of a hyper-parameter is picked, we expand that parameter's value set and repeat the entire grid search until no hyper-parameters at the borders of the grid are picked.

III.1.2 Alternative Physical Density Estimators

Here, we present several alternative density estimators that we use in our robustness tests. A large range of conditional density estimators are known to the literature. On a high level, one can distinguish between parametric (Shumway and Stoffer, 2017) and non-parametric methods (Li and Racine, 2007). Recently, a third class of modeling approaches emerged, which is based on machine learning techniques (Bishop, 2006; Ambrogioni et al., 2017). Our proposed MDN approach also belongs to this class. In essence, machine learning methods are parametric, but they are often so flexible that they can approximate a very large class of alternative parametric models without requiring exogenous information about which exact model to approximate. In a simulation study, Rothfuss et al. (2019b) show the superiority of these methods compared to non-parametric density estimators for a range of distributions that are relevant to financial applications.

In the following, we will give an overview of the methods that we employ in our analysis. To fix notation, let r_t stand for the log return of an asset's price at time t. Let $X_{t-1} = (x_{i,t-1})_{i\in\{1,\dots,N\}}$ stand for a vector of predictor variables that can be observed in $t-1$.

We are now interested in the conditional density $p(r_t|X_{t-1})$ of r_t.

III.1.2.1 Parametric Density

A first approach to specifying $p(r_t|X_{t-1})$ is to pose a parametric structure on the evolution of the asset price and thus return. The well-known class of ARMA-GARCH models (Bollerslev, 1986) is a natural candidate for this task, however, it is by construction backward-looking as it conditions on past return data. Engle and Patton (2007) and Sharma et al. (1996) investigate such time series models that are enriched with exogenous predictor variables. We follow their intuition, but shut down the backward-looking channel by dropping past return information from the model. More precisely, we assume that log returns are conditionally normally distributed, with mean and variance that are linear in the forward-looking input variables:

$$r_t = \alpha + \beta\, X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{III.6}$$

$$\sigma_t^2 = a + b\, X_{t-1} + \nu_t, \qquad \nu_t \sim i.i.d.,\ E(\nu_t) = 0. \tag{III.7}$$

Here, X_t is the vector of forward-looking variables and log returns are distributed according to $r_t \sim \mathcal{N}(\alpha + \beta\, X_{t-1}, a + b\, X_{t-1})$. We estimate the parameters in a two-pass estimation. In a first step, we regress the observed returns in the training set on X_{t-1} via OLS, thus obtaining initial estimates for α and β. We square the residuals and regress these squared residuals on X_{t-1} again, which provides us with OLS estimates for a and b. In the second estimation step, we use the previous parameter estimates as starting values for a joint maximum likelihood estimation.

III.1.2.2 Non-Parametric Density

Three kernel-based methods represent the non-parametric approach in our analysis. First, we use the kernel density estimator to estimate the unconditional joint distributions of $(r_t, X_{t-1})^\intercal$ and the unconditional distribution of X_{t-1} alone and form the ratio between these two estimates to obtain the conditional distribution of r_t. Putting more weight on the training data points that are closest to the current $(r_t, X_{t-1})^\intercal$ observation leads to neighborhood kernel density estimation. Finally, least-squared conditional density estimation relaxes the weight constraints that are imposed upon the estimator by the previous methods. We will present the three methods in more detail in the following paragraphs.

Conditional Kernel Density Estimation Given a set of training data points $\{y_t\}_{t\in[1,T]}$ the unconditional kernel density estimator for a potential point of interest y reads as follows (Rosenblatt, 1956; Parzen, 1962):

$$\hat{p}(y) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{y-y_t}{h}\right) \tag{III.8}$$

Kernel density estimation (KDE) can be understood as placing a simple kernel function $K(\cdot)$ into each data point y_t and forming an equally weighted mixture of the T densities. The difference between y and a training data point y_t is scaled by the bandwidth parameter h. In the case of a multivariate variable $y \in \mathbb{R}^J, J > 1$, the unconditional density can be estimated as the product of marginal kernel density estimates:

$$\hat{p}(y) = \prod_{j=1}^{J} \hat{p}(y_j) = \prod_{j=1}^{J} \frac{1}{Th_j}\sum_{t=1}^{T} K\left(\frac{y_j - y_{j,t}}{h_j}\right). \tag{III.9}$$

In that, y_j denotes the j-th dimension of the column vector y and h_j is the corresponding bandwidth for that dimension. One popular choice of $K(\cdot)$, that we use, too, is the Gaussian kernel

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{III.10}$$

Provided a continuous kernel function such as equation III.10, the estimated pdf in equation III.9 is continuous, too. Beyond the appropriate choice of $K(\cdot)$, a central challenge in kernel density estimation is the selection of the bandwidth parameter h, which controls the smoothness of the estimated pdf. We determine h by minimizing the Integrated Mean Squared Error (IMSE) using a cross-validation approach, as recommended by Li and Racine (2007).

The non-parametric KDE approach can be extended to the conditional case (Conditional KDE; CKDE), in which we are interested in. Here, unconditional KDE can be used to estimate both the joint density of returns and inputs by setting $y_t = (r_t, X_{t-1})^\intercal$ and the marginal density of the inputs by setting $y_t = X_{t-1}$. Then, the conditional density estimate of returns follows as the density ratio

$$\hat{p}(r_t|X_{t-1}) = \frac{\hat{p}(r_t, X_{t-1})}{\hat{p}(X_{t-1})}, \tag{III.11}$$

where both the enumerator and denominator are the sums of kernel functions as in equation III.9.

Neighborhood Conditional Kernel Density Estimation Similar to kernel density estimation, neighborhood conditional kernel density estimation (NKDE) employs standard kernel density estimation in a local ϵ-neighborhood around a point of interest y (Sugiyama and Takeuchi, 2010). The method uses kernels in the training data points as well, however, rather than using all past observations, NKDE only considers a local subset of the training samples $\{y_t\}_{t \in \mathcal{K}_{y,\epsilon}}$, where $\mathcal{K}_{y,\epsilon}$ contains all sample indices that fulfill $||y_t - y||_2 \leq \epsilon$.

In order to obtain a conditional density estimate, we again first build unconditional density estimates for $\hat{p}(r_t, X_{t-1})$ by setting $y_t = (r_t, X_{t-1})^\intercal$ and $\hat{p}(X_{t-1})$ by setting $y_t = X_{t-1}$. We then apply equation III.11. Again, we use a Gaussian kernel function and select both the bandwidth hyper-parameter h and the neighborhood hyper-parameter ϵ via cross-validation.

Least-Squares Conditional Density Estimation The Least-Squares Conditional Density Estimation (LSCDE) approach of Sugiyama and Takeuchi (2010) estimates the conditional density as a linear combination of kernel functions $K(\cdot)$

$$\hat{p}_\alpha(r_t|X_{t-1}) \propto \alpha^T K((r_t, X_{t-1})^\intercal) \tag{III.12}$$

Here, $K((r_t, X_{t-1})^\intercal) = (K_1((r_t, X_{t-1})^\intercal), ..., K_M((r_t, X_{t-1})^\intercal))^T$ is a vector of kernel functions. The main difference between LSCDE and the previous kernel methods is the direct estimation of the weights α via regression. Furthermore, the kernel functions are not necessarily bound to the past observations. In principle, any number of kernel functions, that are located anywhere in the domain of $(r_t, X_{t-1})^\intercal$ is possible. Practically, Sugiyama and Takeuchi (2010) advise picking randomly a number of past observations at which kernels are located. This number is typically much smaller than the amount of observations, thus making the estimation of α more robust. The parameters $\alpha \in \mathbb{R}^M$ are then obtained by minimizing the integrated squared error

$$J(\alpha) = \int \int (\hat{p}_\alpha(r_t|X_{t-1}) - p(r_t|X_{t-1}))^2 p(X_{t-1}) dX_{t-1} dr_t. \tag{III.13}$$

Sugiyama and Takeuchi (2010) derive the closed-form solution for α for the case of Gaussian kernels, which we follow. After having obtained $\alpha^* = \arg\min_\alpha J(\alpha)$, the conditional density of r_t can be computed as follows:

$$\hat{p}_\alpha(r_t|X_{t-1}) = \frac{(\alpha^*)^\intercal K((r_t, X_{t-1})^\intercal)}{\int (\alpha^*)^\intercal K((r_t, X_{t-1})^\intercal) dr_t} \tag{III.14}$$

The denominator in equation III.14 is traceable and can be computed analytically (Sugiyama and Takeuchi, 2010). Hence, neither numerical optimization nor numerical integration is needed for obtaining conditional density estimates with LSCDE. However, three hyper-parameters need to be determined: the bandwidth parameter of the Gaussian kernels, the number of kernel functions to use and a regularization parameter that can be used in the estimation of α^*. As before, we estimate these hyper-parameters via cross-validation.

III.1.2.3 Kernel Mixture Networks

Beside the Mixture Density Network, we consider one further neural network density estimator, the Kernel Mixture Network (KMN). While MDNs resemble a purely parametric conditional density model, the Kernel Mixture Network, combines both non-parametric and parametric elements (Ambrogioni et al., 2017). Similar to MDNs, a mixture density model of $\hat{p}(r_t|X_{t-1})$ is combined with a neural network, which takes the conditioning variable X_{t-1} as an input. However, the neural network only controls the weights of the mixture components while the centers and scales of the Gaussian components are fixed. Figuratively, one can imagine the neural network as choosing between a very large amount of pre-existing kernel functions to build up the final combined density function. Differing from the kernel density methods, multiple kernel functions with differing bandwidths but equal location are considered, giving the KMN an indirect way of controlling the bandwidths and locations of the components, too. Still, the focus on representing the weights as only output variable might result in a higher performance for the KMN. As for MDNs, we employ Gaussians as mixture components, wherein the bandwidths parameter directly coincides with the standard deviation.

Let M be the number of kernel centers μ_m and S the number of kernel scales σ_s. The KMN conditional density estimate reads as follows:

$$\hat{p}(r_t|X_{t-1}) = \sum_{m=1}^{M} \sum_{s=1}^{S} w_{m,s}(X_{t-1}) \mathcal{N}(r_t|\mu_m, \sigma_s^2) \tag{III.15}$$

Here, the neural network only models the weights $w_{m,s}(X_{t-1})$, which must add up to

1 in order to form a valid density estimator. Hence, as in equation III.4, the output nonlinearity of the neural network is chosen as a softmax function. Ambrogioni et al. (2017) propose to choose the kernel centers μ_m by sub-sampling the training data and recursively removing each point $(r_i, X_{i-1})^\intercal, i < t$ that is closer than a constant δ to any of its predecessor points. This can be seen as a naive form of clustering which depends on the ordering of the data set. Instead, we use the well-established clustering method k-means for selecting the kernel centers. Overall, the KMN model is more restrictive than the MDN as the locations and scales of the mixture components are fixed during inference and cannot be directly controlled by the neural network.

The KMN has a number of hyper-parameters, namely the number of hidden neurons in the neural network, the number of locations for the Gaussian components[8] and the number of training iterations. We select these hyper-parameters as described for the MDN in section III.1.1.

III.2 Data

Our forward-looking estimate of the physical return density heavily relies on risk-neutral information. We use the moments of the risk-neutral distribution, which can be inferred from option prices, as inputs for our estimation. To this end, we obtain end-of-day quotes for S&P 500 options from the Chicago Board Options Exchange (CBOE). The option data spans the period January 2004 until July 2017 and comes with matched underlying prices at the time of the option market's close. Bakshi et al. (2003) showed how the variance, skewness and kurtosis of the risk-neutral distribution can be backed out from option prices in a model-free way via aggregation over the strike range. Martin (2017) introduced the SVIX, a similarly constructed measure, that is closely related to the risk-neutral variance and that constitutes a lower bound for the expected return under reasonable assumptions. Due to its theoretical merits, we thus replace the risk-neutral variance estimate of Bakshi et al. (2003) with the SVIX in our physical density predictions.

A careful and precise estimation of the SVIX and the risk-neutral moments from the raw option data is crucial for our analysis. Several popular approaches for building the implied volatility surface and thus risk-neutral moment estimates have been compared in chapter II. Large differences in option-implied estimates due to the choice of the calculation

---

[8] We use k-means to determine the locations of the Gaussian components. At each location, we consider two components with a scale of 0.3 and 0.7. The scales are independent of the volatility of returns, as they are normalized to unit variance prior to estimation and model fitting.

Table III.1: Summary Statistics

			Percentiles				
	Mean	Volatility	10%	25%	50%	75%	90%
S&P 500 return	0.03	1.17	-1.11	-0.39	0.07	0.53	1.13
Risk-free rate	0.005	0.007	0.00	0.00	0.001	0.008	0.019
SVIX^2	394.30	494.35	134.12	167.20	238.33	413.58	720.12
QSkew	-2.46	0.94	-3.46	-2.93	-2.41	-1.94	-1.51
QKurt	24.37	16.62	8.59	13.16	20.85	30.27	43.11
Value return	0.003	0.64	-0.57	-0.28	-0.00	0.26	0.58
Size return	0.002	0.57	-0.67	-0.34	0.01	0.34	0.67
Momentum return	-0.002	0.96	-0.94	-0.38	0.06	0.43	0.89
S&P 500 risk	1.88	6.01	0.22	0.40	0.84	1.77	3.71
Value risk	0.41	0.92	0.04	0.07	0.14	0.29	0.77
Size risk	0.33	0.41	0.10	0.15	0.23	0.35	0.59
Momentum Risk	0.93	2.01	0.09	0.16	0.32	0.74	1.83
N	3303						

The table shows summary statistics for our full S&P 500 data set. All figures are in percent. We use log returns, the factor risk rows represent 10-trading-day rolling window variances of the respective log returns. The SVIX is by definition a measure for risk-neutral volatility. We use its squared version to make it consistent with our remaining risk measures, which are expressed in variance terms, and because SVIX^2 represents a lower bound for the expected equity risk premium, according to Martin (2017). QSkew (QKurt) denote the risk-neutral skewness and kurtosis measures. All quantities refer to the daily time-interval, beside the SVIX^2, which is annualized due to its definition.

method have been identified there. These differences are especially pronounced in the tail regions of the risk-neutral density, which are of high importance for the risk-neutral skewness and kurtosis measures that we use. Taking these findings into account, we use the one-dimensional kernel regression over the strike range for interpolating the implied volatility surface at observed maturities. After having obtained the risk-neutral measures for observed maturities by aggregating appropriately over the strike range, we linearly interpolate these along the maturity dimension to a fix maturity of 30 calendar days. In that, we exclude options with maturities of less than 7 days as the term structure of risk-neutral moments becomes increasingly nonlinear for very short-term options. At times, this filter removes all observations with maturities that are shorter than 30 calendar days, in which case we linearly extrapolate the moment estimates that relate to longer-term maturities.

In order to compare our results for purely forward-looking inputs with density predictions based on backward-looking data, we also obtain time series information. First, as we are interested in predicting the density of S&P 500 returns, we get daily S&P 500 closing prices from Bloomberg. Furthermore, we collect daily return time series for the size, value and momentum factor from Kenneth French's website and treat the S&P 500 as a measure for the market factor. To obtain a time-varying physical measure for factor risk, we compute the variance of factor returns over rolling 10-trading-day windows for each factor separately. We use Overnight Index Swap (OIS) rates as our measure for the risk-free rate, which we source from Bloomberg. We obtain all of these time series data for the same time frame as we have option data available.

We merge risk-neutral moment, return and physical risk time series that we gather and match them with subsequent daily S&P 500 returns. We drop days for which one of our inputs it not observed. Table III.1 contains summary statistics of our data set. We split this full data set into a training and a validation subset. The training set consists of the first 80% of observations, while the last 20% of observations form the validation set. The estimation that we laid out in section III.1.1, including the hyper-parameter selection, is only based on the training data set, such that the validation set is completely untouched before we evaluate a predictor's performance on it.

III.3 Forward-Looking Return Density

III.3.1 Purely Forward-Looking Information

We start our analysis by inspecting the predictive information content of risk-neutral measures with respect to the complete daily return distribution. In particular, we check whether restricting the predictor variables to purely forward-looking measures comes along with a reduction in prediction performance. We measure performance of a density predictor via the average log likelihood of the return observations $\{r_{t+1}\}_{t=1,\dots,T}$, given the density estimator $\hat{p}(\cdot)$ and the predictive variables $\{X_t\}_{t=1,\dots,T}$:

$$\frac{1}{T}\sum_{t=1}^{T}\ln\hat{p}(r_{t+1}|X_t) \tag{III.16}$$

Intuitively, a conditional density estimator will reduce the likelihood of some potential returns and increase the likelihood of other returns based on some observations of X_t. If X_t is informative about the future return and if the density estimator is well specified, the ex post realizations of the returns will be more likely to show up in the regions with increased likelihood, thus leading to an increase in the average log likelihood.

As the average log likelihood is a relative performance measure between competing model specifications, we compare the Mixture Density Network estimator to a benchmark density estimator. In the spirit of Welch and Goyal (2008), who use the unconditional mean of the in-sample data set as benchmark for in- and out-of-sample mean predictions, we use the smoothed empirical in-sample return density as our benchmark. For smoothing, we apply a Gaussian kernel density estimator and select the bandwidth parameter via cross-validation, as suggested by Härdle (1991). We estimate three MDNs, which differ in their input variables. The first MDN uses purely forward-looking information, i.e., the risk-free rate, the SVIX^2, risk-neutral skewness and kurtosis. The second MDN is based on purely backward-looking information, which is past S&P 500, value, size and momentum factor returns and variances. Finally, we consider a MDN specification that uses both types of information as input variables. In all cases, including the benchmark kernel density, we first construct the estimator based on the training data set. We then calculate the average log likelihoods of the fitted density estimators for the training (in-sample) and the validation (out-of-sample) data set, separately. As fitting a MDN employs a local optimization routine, the final conditional estimator may depend on the random starting values of the optimization. To address this concern, we fit each estimator 100 times with different starting values and calculate the average performance across all fits. We also calculate the standard deviation of the individual performance evaluations to get a sense for the stability of the estimator with respect to the starting values.

Panel a of table III.2 contains the result of our performance evaluation. Throughout our estimations, the MDN estimator showed a higher average log likelihood than the benchmark method, both in-sample and out-of-sample. In relative terms, the average log likelihood of the forward-looking estimator is 4.7% higher than the benchmark in-sample and 4.6% higher out-of-sample. The similar performance increase in the in- and out-of-sample data set is a first hint that the forward-looking estimator is generalizing well. The standard deviation of the log likelihood due to different starting values is low, indicating that the estimation method for the neural network is not easily trapped in a local optimum, but converges to a robust estimator. Using purely backward-looking information not only comes along with a lower log likelihood, but also only reaches a 0.9% performance increase against the unconditional benchmark in the out-of-sample data set, after a 4.0% increase in the in-sample data set. This decrease of the performance

Table III.2: Performance and Validity of MDN Conditional Density Estimates

(a) Average Log Likelihood

Model	In-Sample	Out-of-Sample
Benchmark	3.1774	3.4318
Forward-looking	3.3327 (0.0056)	3.5968 (0.0094)
Backward-looking	3.3156 (0.0056)	3.4704 (0.0096)
All information	3.3524 (0.0082)	3.5984 (0.0141)
N	2642	661

(b) Density Test p-Values

	Berkowitz (2001)		Knüppel (2015)	
Model	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
Benchmark	0.000	0.000	0.466	0.001
Forward-looking	0.066	0.81	0.7033	0.974
Backward-looking	0.767	0.000	0.681	0.000
All information	0.537	0.549	0.605	0.386
N	2642	661	2642	661

Panel a of this table shows the average log likelihood of S&P 500 return observations for different conditional density estimators over 100 estimation runs with different starting values in the optimization. We call the unconditional kernel density estimator on the in-sample data set *Benchmark*. We estimate three conditional MDN estimators, which differ in their input data. The *forward-looking* estimator is trained solely on risk-neutral information, the *backward-looking* estimator conditions on past returns and physical risk measures only. We also train an *all information* estimator, which uses both sets of input variables. Standard errors of the neural network log likelihood estimates due to random starting values of the neural network optimization are given in parenthesis in panel a. Panel b shows p-values of the null hypothesis that the density estimator correctly specifies the true return density.

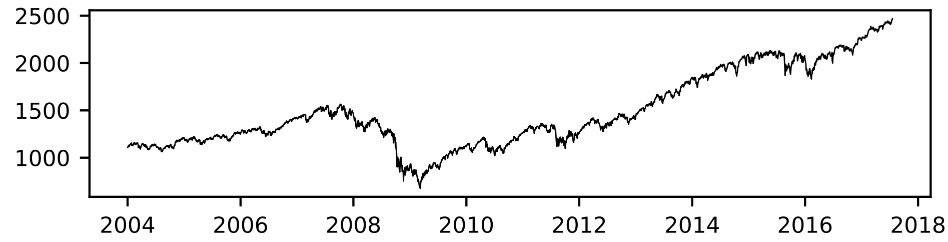
spread is a sign for over-fitting in the backward-looking estimator. Furthermore, adding backward-looking information to the forward-looking MDN does not appear to improve the estimator sustainably. The performance slightly increases in-sample, but does nearly not change out-of-sample, leading to a decrease in the performance spread against the bechmark between the in- and out-of-sample data sets. It thus seems like past return and risk information only introduces a slight tendency of over-fitting into the estimator. Our results show that a purely foreward-looking MDN estimator of the conditional return density improves on the unconditional return density and is most favorable with respect to performance and generalization. Figure III.1 shows time series charts of the S&P 500 price development along with the expected physical moments of our forward-looking density estimator.

We continue by asking whether the MDN estimator is consistent with the data from a statistical point of view. To this purpose, we apply the tests of Berkowitz (2001) and Knüppel (2015). The null hypothesis of these tests is that the observations of a random variable are drawn from a given density estimator. Panel b of table III.2 contains p-values for this null hypothesis for the in-sample and out-of-sample data set, separately. To obtain the p-values for the out-of-sample data set, we use the density estimator that was trained on the in-sample data set. In general, the findings of the two tests are strongly aligned. The tests reveal that the unconditional kernel density estimator appears to be inappropriate, as it is strongly rejected out-of-sample and even in-sample by the Berkowitz (2001) test. The forward-looking estimator shows high p-values and we cannot reject the null that this estimator describes the true conditional density of returns. For the backward-looking estimator, in-sample p-values are high, but the estimator is very strongly rejected out-of-sample by both tests. Again, this is a clear sign of the over-fitting behavior of the backward-looking estimator. The estimator that uses both backward- and forward-looking information is not rejected by either test. Beside the Berkowitz (2001) test of the in-sample data set, the p-values are considerably lower than for the forward-looking estimator, though.

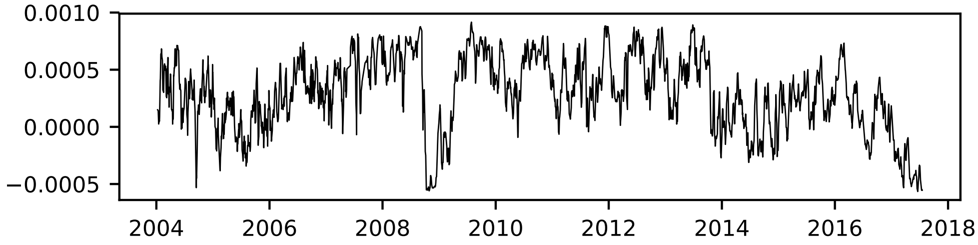
At first sight, it appears puzzling that the benchmark estimator is rejected in-sample by the Berkowitz (2001) test. The p-value of the forward-looking estimator is low, too. This finding can be explained by the structure of the Berkowitz (2001) test, which not only checks whether the distribution matches, but also if there is autocorrelation in the conditional probabilities. If we disable the autocorrelation check, the p-value of the benchmark rises to 0.111, the forward-looking estimator's p-value even increases to

Figure III.1: S&P 500 Price Development and Expected $\mathcal{P}$ Moments

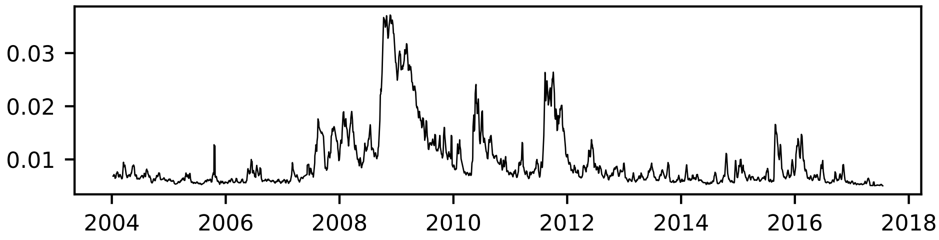
(a) S&P 500 Price



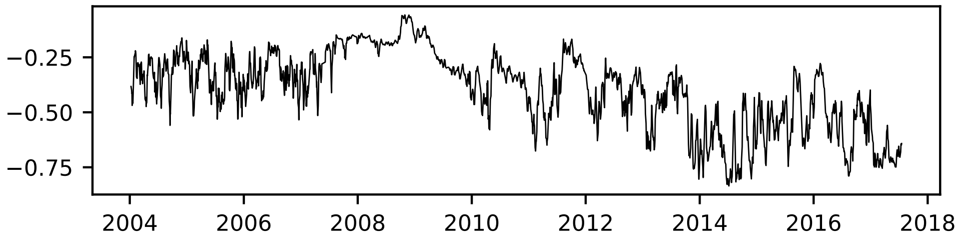
(b) Expected Mean



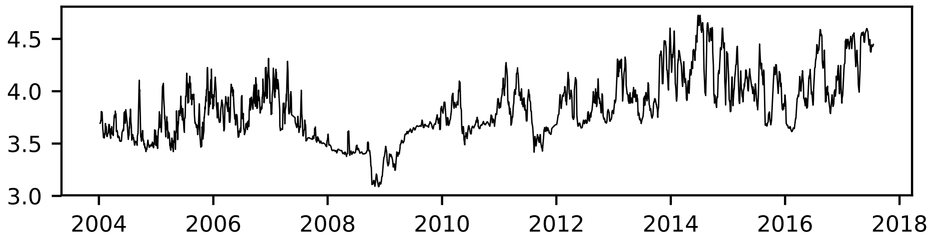
(c) Expected Standard Deviation



(d) Expected Skewness



(e) Expected Kurtosis



Panel (a) shows the price development of the S&P 500 between January 2004 and July 2017. Panels (b) to (e) show one-week rolling window averages of the MDN-implied forward-looking mean, standard deviation, skewness and kurtosis expectation.

0.805 in the in-sample data set. Autocorrelation in the conditional probabilities can be induced by autocorrelation in the returns, if the estimator does not account for that. As neither the benchmark nor the forward-looking estimator conditions on past return data, they cannot correct for return autocorrelation. In contrast, the backward-looking estimator, which makes use of past return data, shows a high in-sample p-value. The exploitation of return autocorrelation is dangerous though, as can be seen from the out-of-sample p-values. Here, the forward-looking estimator has a very high p-value, while the backward-looking estimator is strongly rejected. The reason for this observation is that the autocorrelation does not persist in the out-of-sample data set. Again, the result is due to the over-fitting behavior of the backward-looking estimator, which does not show up in the forward-looking estimator.

III.3.2 Trading on the Conditional Return Distribution

If the forward-looking MDN estimator is informative about the future return distribution, as our previous results suggest, it should be possible to derive profitable trading strategies based on the density forecasts. In this section, we present two simple trading schemes that exploit information about the conditional physical mean and volatility, that the MDN implicitly predict. Essentially, both trading strategies use the moment prediction as a signal to increase or decrease a market position.

For our first strategy, we obtain the daily conditional one-day-ahead mean expectation $\hat{m}_{1,t+1}$ as

$$\hat{m}_{1,t+1} = E[r_{t+1}|X_t] = \int_{-\infty}^{\infty} u\, \hat{p}(u|X_t)\, du. \tag{III.17}$$

Here, $\hat{p}(\cdot)$ denotes the MDN estimator and X_t is a vector of all forward-looking variables that we observe at time t. Since the MDN allows to evaluate $\hat{p}(r_{t+1}|X_t)$ at any r_{t+1}, we can numerically approximate the integral in equation III.17 with high precision (Atkinson, 1989, ch. 5). Each day, we invest a fraction of total wealth into the S&P 500 at its closing price and hold that position for one day. The fraction of wealth that we invest is defined as $w_t = \frac{\hat{m}_{1,t+1}}{\sigma(\hat{m}_{1,t+1})}$, where $\sigma(\hat{m}_{1,t+1})$ represents the standard deviation of the mean estimate. Note that the position size is determined by the expected mean itself. Any other denominator beside the standard deviation could also be used and would only scale the returns of our trading strategy. Our investment scheme can lead to negative position sizes, in which case we initiate a short position in the S&P 500. If the w_t exceeds 1, we assume that the position is leveraged accordingly. We compare the strategy returns to a

Table III.3: Mean Trading Strategy: Summary Statistics

	Full data set		Out-of-sample	
Strategy	Buy-and-Hold	MDN	Buy-and-hold	MDN
Mean return	8.48 (5.11)	17.67 (7.02)	8.43 (7.89)	9.94 (8.25)
Mean excess return	7.18 (5.11)	16.38 (7.03)	8.28 (7.9)	9.78 (8.26)
Excess return volatility	18.52	25.45	12.79	13.38
Sharpe Ratio	0.388	0.694	0.647	0.743
N	3303	3303	661	661

The table shows the annualized expected return, excess return, excess return volatility and Sharpe Ratio of the conditional mean trading strategy in the S&P 500, compared with the buy-and-hold strategy. The numbers in parenthesis are standard errors of the average return estimates. The full data set covers January 2004 to July 2017. The out-of-sample data set covers the last 20% of the full data set and thus starts in November 2014.

simple buy-and-hold strategy's returns.

Table III.3 gives an overview of this simple mean trading strategy. On average, the market timing strategy is invested with 80.7% of total wealth, with a standard deviation of 100%. Although the average fraction of wealth invested is lower than for the buy-and-hold strategy, the market timing strategy manages to realize an average yearly excess return of 16.4%, which is 9.2% higher than the average return of the buy-and-hold strategy. However, this higher average excess return comes along with an increased volatility of annualized 25.5%. Still, using the MDN mean forecast as a signal to time the market increases the Sharpe ratio of the buy-and-hold strategy of 0.388 by more than 78% to 0.694. The results for the full data set are confirmed by looking at the out-of-sample data set only. Although the market volatility and thus the potential for market timing has been significantly lower in the out-of-sample set than in the full data set, our market timing strategy still increases the Sharpe ratio of the buy-and-hold strategy from 0.647 to 0.743, an increase of nearly 15%. This increase results from a 1.5% higher annual excess return for the market timing strategy with a volatility increase of only 0.59%. In summary, timing the market based on the MDN's mean return forecasts appears to be a beneficial and robust trading strategy.

We proceed with our second trading strategy, which makes use of the implicit volatility forecast of the MDN. Bakshi and Kapadia (2003) describe a negative variance risk premium, that can be earned by shorting options. We therefore base our volatility trading strategy on shorting at-the-money (ATM) straddles. A straddle consists of a Call and a Put option with the same strike and maturity. On expiration, the position earns a negative return, if the price of the underlying barely changed. On the other hand, if the price of the underlying moved strongly, the position will earn a positive return. On average, ATM straddles have significantly negative returns, which is consistent with a negative variance risk premium. However, it might be possible to time straddle returns if a robust volatility forecast is available.

To this purpose, we calculate the conditional one-day-ahead second moment $\hat{m}_{2,t+1}$ on a daily basis as

$$\hat{m}_{2,t+1} = E[r_{t+1}^2|X_t] = \int_{-\infty}^{\infty} u^2\,\hat{p}(u|X_t)\,du. \tag{III.18}$$

For each day, we select the strike that is closest to the S&P 500 price and the option series with the shortest maturity larger than 6 days. Thus, the combined position's delta exposure is approximately 0. At the selected strike and maturity, we initiate a straddle at the CBOE's reported end-of-day mid price. After one day, we close the position at its end-of-day mid price. Unconditionally, we would short the straddle position to earn the variance risk premium (Bakshi and Kapadia, 2003). However, if we predict the S&P 500's second moment to rise, we reduce the short position, or even go long the straddle. On the other hand, if we predict a lower second moment for the next day than for today, we increase the short position. More precisely, the position weight is defined as

$$\begin{aligned} \Delta\hat{m}_{2,t+1} &= \hat{m}_{2,t+1} - \hat{m}_{2,t} \\ w_t &= -1 + \frac{\Delta\hat{m}_{2,t+1}}{\sigma(\Delta\hat{m}_{2,t})}, \end{aligned}$$

where $\sigma(\Delta\hat{m}_{2,t})$ is the standard deviation of the daily change in the second moment of the $\mathcal{P}$ density. If our second moment is informative about ex post variation of returns, we expect this straddle timing to be beneficial for two reasons. First, a higher second moment increases the probability for larger returns in the underlying. Since a straddle essentially represents a bet on variation and a large return realization increases the payoff, higher second moments of the return distribution lead to higher expected straddle payoffs and thus increase the straddle's price. Second, as Bakshi and Kapadia (2003) show, higher return variation comes along with an increase in the variance risk premium, which would translate into an increased straddle price. Together, straddle prices should rise if variation increases. For this reason, we reduce the short position if the MDN predicts an increase in the second moment of the return distribution and vice versa.

Table III.4: Variation Trading Strategy: Summary Statistics

	Full data set		Out-of-sample	
Strategy	Fix Short	MDN	Fix Short	MDN
Mean return	523.86 (53.9)	1392.02 (79.03)	753.23 (144.12)	1464.58 (168.16)
Mean excess return	525.15 (53.9)	1393.29 (79.02)	753.61 (144.13)	1464.97 (168.17)
Excess return volatility	195.01	285.92	233.24	272.15
Sharpe Ratio	2.69	4.87	3.23	5.38
N	3299	3299	660	660

The table shows the annualized expected return, excess return, excess return volatility and Sharpe Ratio of the conditional variation trading strategy in S&P 500 straddles, compared with the fix short strategy. The numbers in parenthesis are standard errors of the average return estimates. The full data set covers January 2004 to July 2017. The out-of-sample data set covers the last 20% of the full data set and thus starts in November 2014.

Table III.4 compares the returns of the fix always-short strategy and our straddle timing strategy. As is known in the literature, average returns to shorting options in general, and straddles in particular, are sizable. Fix shorting of straddles earns an average log excess return of 525% per year.[9] This high average return is put into perspective by the standard deviation of 195%, leading to a Sharpe ratio of 2.69. By using the MDN estimator's variation forecast, the Sharpe ratio can be increased by more than 80% to 4.87. Out-of-sample, our variation forecast increases the Sharpe ratio of 3.23 by 66% to 5.38.

Taken together, our results show that the MDN estimator can be used to time the market. This market timing becomes possible due to the predictive capabilities of our forward-looking estimator for the mean and volatility of S&P 500 returns. In both cases, the results for the out-of-sample set strongly support the findings for the full data set. This confirms our previous findings that the MDN is not over-fitting the training data and generates a robust forward-looking density predictor.

III.3.3 Determinants of Index Return Distributions

In the previous sections, we've established the Mixture Density Network as a flexible and robust predictor for the S&P 500 return distribution. The MDN bases its predictions on moments of the risk-neutral distribution. As a result, we can express the conditional

---

[9] We calculate returns with respect to the initial price of the straddle, which is why they appear so large. In the literature, returns are sometimes also calculated with respect to the underlying price, which technically makes average returns appear lower. The true return depends on the degree of leverage that an investor is willing to accept.

return distribution, represented by its conditional moments, as a function of the risk-neutral moments. We further ask whether all of these risk-neutral moments matter for the density prediction. To this end, we apply an adjusted form of the test of Patton and Timmermann (2010), to allow statistical inference about the impact of a predictor variable on the conditional return distribution.

We start our analysis by estimating the MDN on the complete data set. Let x_i be the i-th predictor variable of the MDN, in our case the i-th risk-neutral moment, and $X' = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)^\intercal$ be a vector of the remaining input variables. As it is not essential for this analysis, we dropped the time index for ease of notation. We obtain the 10% (90%) percentile of the observed value range of x_i, x_i^{low} (x_i^{high}), and discretize the interval $[x_i^{low}, x_i^{high}]$ into 100 data points with equal step size $\{x_i^{(n)}\}_{n\in[1,100]}$. For each discretized data point, we calculate the MDN-implied conditional l-th centralized and normalized moment for $l \in [1, 2, 3, 4]$, i.e., mean, variance, skewness and kurtosis:

$$\hat{m}_{1,i}^{(n)} = E(r^1|(x_i^{(n)}, X')), \tag{III.19}$$

$$\hat{m}_{2,i}^{(n)} = E((r - \hat{m}_{1,i}^{(n)})^2|(x_i^{(n)}, X')), \tag{III.20}$$

$$\hat{m}_{3,i}^{(n)} = \frac{E((r - \hat{m}_{1,i}^{(n)})^3|(x_i^{(n)}, X'))}{\left(\hat{m}_{2,i}^{(n)}\right)^{1.5}}, \tag{III.21}$$

$$\hat{m}_{4,i}^{(n)} = \frac{E((r - \hat{m}_{1,i}^{(n)})^4|(x_i^{(n)}, X'))}{\left(\hat{m}_{2,i}^{(n)}\right)^{2}} \tag{III.22}$$

In that, we set the elements of the vector X' to their unconditional medians. Following Patton and Timmermann (2010), a test for the impact of x_i on $\hat{m}_{l,i}$ can now be built on this empirical representation of the functional relationship between the i-th predictor variable and the l-th conditional return moment. We identify the indices u and d that satisfy $\hat{m}_{l,i}^{(u)} = \max_n \hat{m}_{l,i}^{(n)}$ and $\hat{m}_{l,i}^{(d)} = \min_n \hat{m}_{l,i}^{(n)}$, that is, the indices on the discretized range of $\{x_i\}$ for which we observed the lowest and the highest conditional moment estimate. If x_i does not predict the l-th conditional moment, this lowest and highest conditional moment estimate should be equal, which leads us to the null hypothesis

$$H_0: \quad \hat{\Delta}_{l,i} = \hat{m}_{l,i}^{(u)} - \hat{m}_{l,i}^{(d)} = 0. \tag{III.23}$$

We can quantify the likelihood of this null hypothesis via bootstrapping. We apply the stationary bootstrap of Politis and Romano (1994) and resample the training data

$B = 1000$ times. For each resampled data set, we refit the MDN and obtain the bootstrapped moment samples along the discretized range of x_i, $\hat{m}_{l,i}^{(n,b)}$, $b \in \{1, ..., B\}$. For each bootstrap, we calculate the test statistic $\hat{\Delta}_{l,i}^{(b)} = \hat{m}_{l,i}^{(u,b)} - \hat{m}_{l,i}^{(d,b)}$ and eventually estimate the p-value for H_0 as

$$J_{l,i}^{(b)} = \hat{\Delta}_{l,i}^{(b)} - \hat{\Delta}_{l,i}$$

$$\hat{p}_{l,i}^{0} = \frac{1}{B} \sum_{b=1}^{B} 1_{\left[J_{l,i}^{(b)} > \hat{\Delta}_{l,i}\right]},$$

where $1_{\left[J_{l,i}^{(b)} > \hat{\Delta}_{l,i}\right]}$ is an indicator function that is 1 if the condition in brackets is fulfilled and 0 otherwise.

Table III.5 shows our estimated p-values for H_0 for each input and the first four moments of the forward-looking return density. All of our forward-looking input variables are informative about the conditional return density at the 5% significance level. In particular, SVIX^2 helps predict the mean, standard deviation and higher moments of the return distribution. For this input variable, the null hypothesis of no influence is very strongly rejected for all $\mathcal{P}$ moments. Risk-neutral skewness and kurtosis are especially relevant for predicting their physical counterparts. The evidence for an impact of these variables on the conditional mean and volatility is comparably weak. Finally, the risk-free rate only helps to pin down volatility expectations, which is significant at the 5% level. Surprisingly, we cannot reject the null that the risk-free rate is uninformative about the S&P 500 return. Our tests reveal that especially option-implied variables are robust predictors of the entire return distribution. However, these tests do not tell us about the shape of the predicting relationships, which is in principle unconstrained. We will now take a detailed

Table III.5: Impact of Predicting Variable on the Conditional Return Density

Variable	Mean	Variance	Skewness	Kurtosis
Risk-free rate	0.416	0.037**	0.286	0.288
SVIX^2	0.000***	0.000***	0.000***	0.004***
QSkew	0.231	0.117	0.000***	0.000***
QKurt	0.052*	0.165	0.007***	0.001***

The table shows p-values for the null hypothesis that the predicting variable in the first column is not informative about the moment in the column label of columns 2 to 5. We use stationary bootstrapping with 1000 iterations and the test of Patton and Timmermann (2010) to obtain these estimates. We add a '*' if the null hypothesis is rejected at the 10% level, '**' if it is rejected at the 5% level and '***' for rejection at the 1% confidence level.

look at these relationships, as estimated by the MDN.

In order to qualitatively and quantitatively inspect the impact of a forward-looking variable on the return density prediction, we make use of our empirical representations of the conditional moments along the input range, $\hat{m}_{l,i}^{(n)}$, and their respective bootstrap samples $\hat{m}_{l,i}^{(n,b)}$. We follow Davison and Hinkley (1997) in constructing 90% confidence intervals for the predicting relationships that the MDN identifies. For that, we first calculate the differences $\delta_{l,i}^{(n,b)} = \hat{m}_{l,i}^{(n,b)} - \hat{m}_{l,i}^{(n)}$. Let $\delta_{l,i}^{(n,0.05)}$ ($\delta_{l,i}^{(n,0.95)}$) be the empirical 5% (95%) percentile of $\delta_{l,i}^{(n,b)}$ for each $x_i^{(n)}$ and predicted return moment l. The bootstrapped confidence interval for $\hat{m}_{l,i}^{(n)}$ is then $[\hat{m}_{l,i}^{(n)} - \delta_{l,i}^{(n,0.95)}, \hat{m}_{l,i}^{(n)} - \delta_{l,i}^{(n,0.05)}]$.

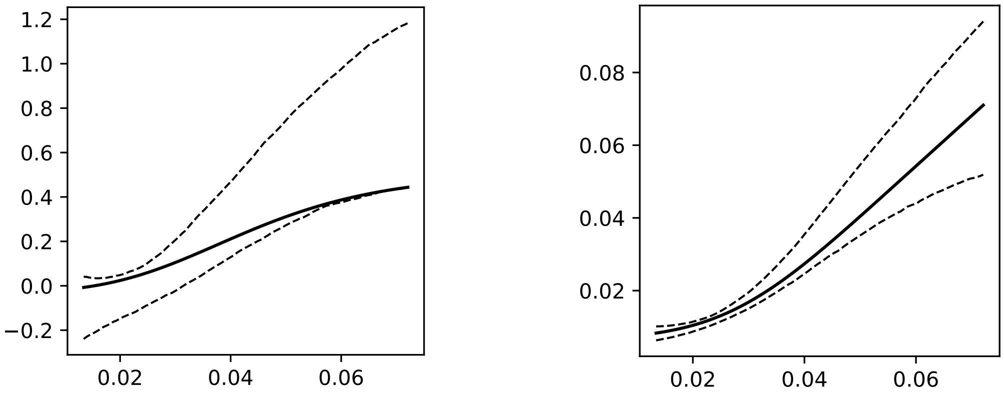
Figure III.2 plots $\hat{m}_{l,i}^{(n)}$ for the input-moment relationships that we identified as significant at the 0.1% confidence level in our previous tests. We start our inspection with the only forward-looking variable that predicts S&P 500 returns at that level of significance in our set-up, the SVIX^2. Panel a shows how the annualized conditional day-ahead return expectation changes, as the SVIX^2 changes. According to the MDN, the expected return rises with an increase in SVIX^2 and a linear relationship is well possible. This is in line with Martin (2017), who shows that the SVIX^2 constitutes a lower bound for the equity risk premium under certain conditions. For high values of SVIX^2 though, it appears like the lower bound is not tight, at least on the daily horizon. For example, an annualized SVIX^2 of 0.07 leads to an annualized day-ahead mean expectation well above 40% if all other variables are at their unconditional medians. It is important to note that we use the SVIX^2 at the 30-day horizon to form expectations about the day-ahead return. If there exists a downward-sloping term structure in the SVIX, the lower bound relationship might still be tight. Calculating backwards, an annualized day-ahead return expectation of 40% would imply a value of the SVIX at the one-day horizon of 63.2%, if the lower bound is tight.[10]

In essence, the SVIX^2 is a measure for risk-neutral variance expectations. Several studies (Chernov, 2007; Busch et al., 2011; Bekaert and Hoerova, 2014) document a strong relationship between risk-neutral variance expectations and subsequent variance realizations. Panel b of figure III.2 shows the MDN-implied variance forecast as a function of the SVIX^2. Note that the MDN is not constrained to a linear link between risk-neutral and physical variance, but instead retrieves the relationship from the data. Nevertheless, the MDN also models an approximately linear relationship, but only if the SVIX

---

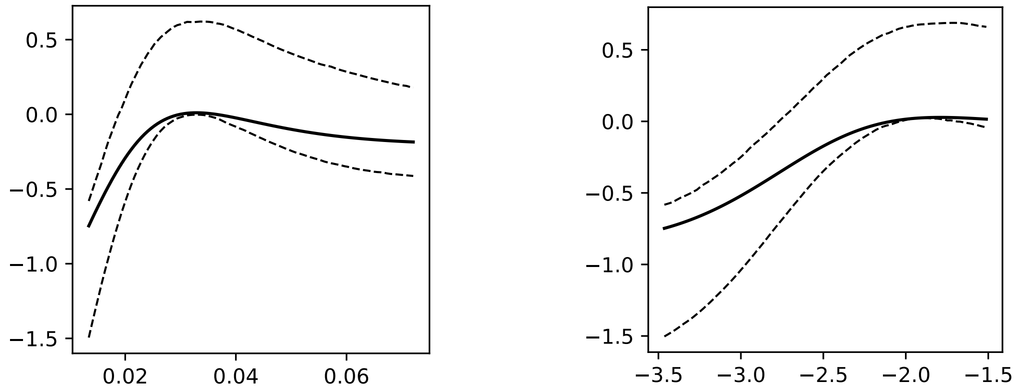
[10] The value results from the fact that SVIX^2 represents the lower bound for the equity risk premium, according to Martin (2017), and $0.632 \approx \sqrt{0.4}$.

Figure III.2: Conditional S&P 500 Moment by Risk-Neutral Moment



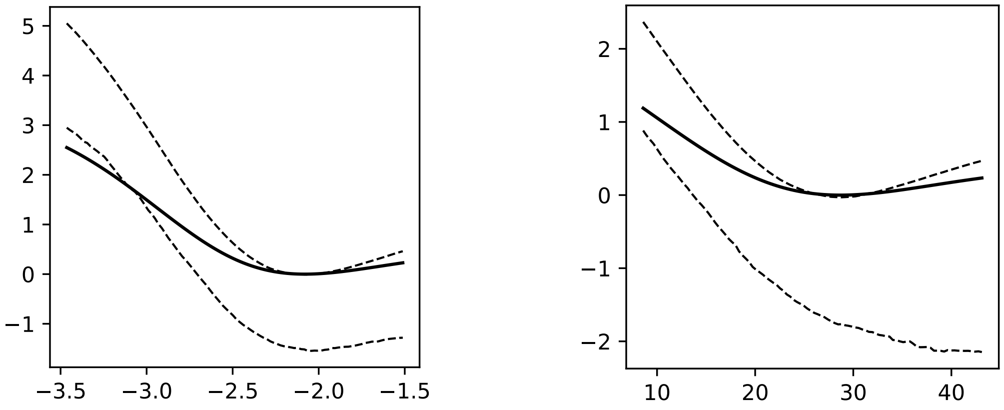
(a) Conditional Mean by SVIX^2

(b) Conditional Variance by SVIX^2



(c) Conditional Skewness by SVIX^2

(d) Conditional Skewness by QSkew



(e) Conditional Excess Kurtosis by QSkew

(f) Conditional Excess Kurtosis by QKurt

This figure shows selected predicting relationships between a forward-looking input variable and a conditional physical moment. Each panel's title indicates what it shows: A title of X *by* Y would mean that the MDN moment forecast of X is plotted on the vertical axis, and conditioned on Y, which is plotted on the horizontal axis. All other input variables beside Y are set to their unconditional median. The conditional mean and variance forecasts are annualized. Bootstrapped Davison and Hinkley (1997) 90% confidence bands are added in light lines.

is above 17% ($\text{SVIX}^2 > 0.03$). For lower values of the SVIX though, there appears to exist a nonlinearity, which causes higher physical variance expectations compared to an overall linear relationship. At the same time, the 90% confidence bands become very narrow at this point, indicating that the estimator is relatively certain about the shape of the relationship for low SVIX values. The relationship allows the following interpretation: An increase of the SVIX from low to intermediate levels is primarily driven by an increase in the variance risk premium, as physical variance expectations are not rising strongly. Further increases in the SVIX are then also driven by increases in physical variance expectations.

We now turn to the impact of option-implied variables on physical expectations of higher moments of the return density. Panels c and d of figure III.2 show, how skewness expectations change with the SVIX^2 and risk-neutral skewness, panels e and f illustrate the response of conditional kurtosis to changes in risk-neutral skewness and kurtosis. The MDN estimates strongly nonlinear responses of conditional higher moments to changes in the forward-looking inputs. For example, an increase in SVIX^2 from 0.015 (corresponds to a SVIX of 12.2%) to 0.03 (corresponds to a SVIX of 17.3%) leads to a corresponding increase in physical skewness from -0.62 to nearly 0. However, further increases in the SVIX^2 have nearly no significant impact on conditional skewness expectations. This finding is consistent with a notion that risk hides in the left tail in times of low market volatility. In times of higher risk-neutral volatility expectations, the daily return distribution is not significantly left skewed. A similar finding exists for the impact of risk-neutral skewness. Here, an increase in risk-neutral skewness also increases physical skewness expectations, but only up to a risk-neutral skewness of about -2. Similarly, for low values of risk-neutral skewness, conditional kurtosis is high and the return distribution becomes less heavy-tailed as it becomes less left skewed under $\mathcal{Q}$. Again, the effect disappears at a risk-neutral skewness of about -2 and expected return kurtosis stays nearly constant for larger risk-neutral skewness values. The MDN identifies a surprising relationship between risk-neutral and physical return kurtosis. As risk-neutral kurtosis increases, physical kurtosis drops until it reaches the normal distribution's kurtosis and stays roughly constant beyond this point.

In summary, the MDN identifies a range of nonlinearities in the relationship of risk-neutral and physical moments. The forward-looking return distribution appears to be closer to the normal distribution in crisis times, when risk-neutral variance is high, and more strongly left skewed in calm times. At the same time, strong priced-in fears about sudden crashes, as signaled by low risk-neutral skewness expectations, appear justified by a more heavy-tailed and left skewed physical return distribution. Our tests show that the MDN identifies relationships between risk-neutral and physical moments that are statistically robust and economically relevant.

III.4 Robustness

III.4.1 Conditional Density Estimation Method

In this chapter, we present a forward-looking $\mathcal{P}$ density estimator that builds on a neural network. However, the forward-looking feature is introduced by basing our forecasts purely on risk-neutral information, which is independent of the use of a Mixture Density Network as conditional density estimator. We therefore inspect in this section whether more traditional density estimators can also be used to form valid $\mathcal{P}$ densities, and whether the MDN does indeed perform better than these alternatives. To this purpose, we perform an out-of-sample horse race between the MDN and popular parametric, non-parametric and machine learning techniques. We compare these methods with respect to their average conditional log likelihood, as defined in equation III.16, as well as their root mean squared prediction error for the mean and volatility:

$$\text{Mean RMSE} = \sqrt{\frac{1}{T-t^*} \sum_{t=t^*+1}^{T} (r_t - E(r_t|X_{t-1}))^2}, \tag{III.24}$$

$$\sigma \text{ RMSE} = \sqrt{\frac{1}{T-t^*} \sum_{t=t^*+1}^{T} (|r_t - E(r_t|X_{t-1})| - \sigma(r_t|X_{t-1}))^2}, \tag{III.25}$$

where $E(r_t|X_{t-1})$ denotes the conditional mean expectation and $\sigma(r_t|X_{t-1})$ the conditional volatility expectation of a density estimator. All methods are trained on a training data set $\{(r_t, X_{t-1})\}_{t\in[1,t^*]}$, such that we obtain out-of-sample performance figures based on the validation data set $\{(r_t, X_{t-1})\}_{t\in[t^*+1,T]}$. As before, the validation data set consists of the last 20% of our total available data. Finally, we test whether a conditional density estimator can be rejected statistically by applying the distribution tests of Berkowitz (2001) and Knüppel (2015).

We consider six alternative density estimators in our performance evaluations. We provide a detailed description of these methods and how we perform estimation based on them in section III.1.2. Here, we quickly recapitulate the six alternatives and fix abbreviations.

Table III.6: Out-of-sample Performance of Conditional Density Estimators

Model	Avg. log $\mathcal{L}$	Mean RMSE	σ RMSE	Berkowitz (2001)	Knüppel (2015)
Benchmark	3.4318	0.8049	0.708	0.000	0.001
Linear	2.7125	0.8789	1.9254	0.000	0.000
CKDE	3.2231	0.8051	0.9914	0.000	0.000
NKDE	2.9564	1.2522	0.8617	0.000	0.000
LSCDE	3.0777	0.8041	1.2172	0.000	0.000
MDN	3.5968	0.8045	0.5284	0.81	0.974
KMN	3.5854	0.8042	0.5582	0.767	0.778

This table compares the out-of-sample performance for a range of forward-looking conditional density estimators for the S&P 500. The alternative conditional estimators are: Linear Conditional Gaussian (Linear), Conditional Kernel Density Estimation (CKDE), Neighborhood Kernel Density Estimation (NKDE), Least-Squares Conditional Density Estimation (LSCDE), Mixture Density Network (MDN) and Kernel Mixture Network (KMN). The second column shows the average log likelihood of observed returns for each estimator. The third and the fourth column report the root mean squared errors for the mean and standard deviation prediction. For the methods that involve an optimization (LSCDE, MDN, KMN), we report the error of the average estimator of 100 estimation runs with different starting values.

We use a Gaussian kernel density estimator as unconditional benchmark (Benchmark). We further estimate a parametric model, where we assume that returns are conditionally normally distributed and conditional mean and variance are linear in the forward-looking predictor variables (Linear). We inspect three different non-parametric estimators. The Conditional Kernel Density Estimator (CKDE) is closely related to our unconditional benchmark. It first forms an estimate of the joint distribution of r_t and X_{t-1} and then obtains a conditional density estimate by dividing by the marginal distribution of X_{t-1}. Related to this approach, the Neighborhood Conditional Density Estimator (NKDE) differs by only considering past observations whose inputs are close to X_{t-1} when forming the density estimate. The least-squares conditional density estimator (LSCDE) turns the kernel density estimation into a regression task by reducing the number of kernels, fixing their positions and only determining their weight for a given estimate. Finally, we consider another neural network approach, the Kernel Mixture Network (KMN).

Table III.6 shows the results of our horse race. The two neural network approaches show the highest average log likelihoods and lowest volatility prediction errors. As is well known, returns are nearly not predictable at the daily horizon, but still the 0.8045 (0.8042) RMSE of the MDN (KMN) mean prediction corresponds to an out-of-sample R^2 of 0.1% (0.17%). We pick the MDN over the KMN in our main analysis due to its slightly higher average conditional log likelihood and lower volatility forecasting error. The linear model shows the lowest out-of-sample performance of all analyzed methods. Its unfavorable performance compared to the unconditional benchmark is a sign for overfitting in this set-up. At the same time, the non-parametric methods also underperform the unconditional benchmark. One reason might lie in the size of our training set. Non-parametric methods typically require a large amount of data to build a powerful estimator. Our training data set contains 2642 data points, which might not be enough for these approaches. Finally, no density estimator except for the neural network estimators passes the Berkowitz (2001) and Knüppel (2015) tests. Put differently, we can reject the null hypothesis that returns are sampled from the respective conditional densities with almost certainty. Only the MDN and the KMN build conditional densities for which we cannot reject this null hypothesis.

It seems like the neural network approaches in general, and the MDN in particular, represent a bridge between the parametric and non-parametric world. They are comparable to non-parametric approaches in their flexibility, but their parametric structure enables us to form expressive density estimators based on relatively small training data sets. Our results suggest that neural network approaches are superior to standard alternatives and in our tests, they were the only approaches that produced a valid forward-looking density estimator.

III.4.2 International Evidence

It is thinkable that unknown characteristics in the relationship of the S&P 500 and its associated options work in favor of the MDN. For this reason, we repeat the analysis of section III.4.1 for the Euro Stoxx 50. The Euro Stoxx 50 represents the leading equity index for the euro area and highly liquid options on the index are traded at the Eurex.

Table III.7 contains the results of our performance evaluation for the Euro Stoxx 50. Throughout all methods and performance figures, it appears like the density of Euro Stoxx 50 returns is harder to predict than the density of the S&P 500. However, in our validation data set, daily S&P 500 returns had an average volatility of 0.8%, while the volatility of daily Euro Stoxx 50 returns was at 1.26%. Relative to the benchmark method, the MDN performs very similar as for the S&P 500: The average conditional log likelihood is 3.2% higher than the benchmark, compared to 4.8% for the S&P 500. The Mean RMSE grows by 0.1%, compared to a decrease of 0.05% in the S&P 500. Finally,

Table III.7: Euro Stoxx 50 Out-of-sample Performance of Conditional Density Estimators

Model	Avg. log $\mathcal{L}$	Mean RMSE	σ RMSE	Berkowitz (2001)	Knüppel (2015)
Benchmark	2.9957	1.2615	1.0686	0.000	0.067
Linear	2.7309	1.2657	1.6516	0.000	0.000
CKDE	2.8856	1.2611	1.2321	0.000	0.000
NKDE	2.6432	1.7439	1.2161	0.000	0.000
LSCDE	2.7906	1.2609	1.4234	0.000	0.000
MDN	3.0924	1.2627	0.8297	0.434	0.966
KMN	3.0883	1.2608	0.8635	0.888	0.901

This table compares the out-of-sample performance for a range of forward-looking conditional density estimators for the Euro Stoxx 50. The alternative conditional estimators are: Linear Conditional Gaussian (Linear), Conditional Kernel Density Estimation (CKDE), Neighborhood Kernel Density Estimation (NKDE), Least-Squares Conditional Density Estimation (LSCDE), Mixture Density Network (MDN) and Kernel Mixture Network (KMN). The second column shows the average log likelihood of observed returns for each estimator. The third and the fourth column report the root mean squared errors for the mean and standard deviation prediction. For the methods that involve an optimization (LSCDE, MDN, KMN), we report the error of the average estimator of 100 estimation runs with different starting values.

the RMSE of the volatility forecast is 22.4% below the benchmark, compared to a decrease of 25.4% in the S&P 500. In relative terms, the ordering of the different methods with respect to their performance is very similar to the S&P 500 application. The results of the Berkowitz (2001) and Knüppel (2015) tests also draw the same picture as for the S&P 500: The densities of all methods except for the neural network approaches are rejected at the 0.1% confidence level by at least one test.

Overall, the results of the Euro Stoxx 50 exercise confirm our previous findings. Note that our analysis does not state that the mechanics between forward-looking variables and returns are the same for the S&P 500 and the Euro Stoxx 50. It rather shows that the MDN is flexible enough to capture general and stable relationships between these variables in both markets.

III.4.3 Over-Fitting

A central concern when working with neural network approaches is over-fitting. Neural networks are highly flexible with respect to the approximated functional relationship in the training data. It can therefore happen that the trained model traces input-output characteristics that existed in the training data only due to randomness. The generalization capability of such a model would be severely reduced.

Table III.8: Out-of-sample MDN Performance with and without Noise Regularization

Noise Regularization	Yes	No
Avg. log $\mathcal{L}$	3.5936 (0.0094)	3.5968 (0.0094)
Mean RMSE	0.8045 (0.0012)	0.8045 (0.0012)
Mean MAE	0.5566 (0.0017)	0.5566 (0.0017)
σ RMSE	0.5327 (0.0069)	0.5284 (0.0075)
σ MAE	0.4271 (0.0094)	0.4209 (0.0102)

The table shows the effect of noise regularization on the out-of-sample performance of the MDN estimator. *Avg. log* $\mathcal{L}$ stands for the average log likelihood of the density prediction. *Mean RMSE* and *Mean MAE* are the root mean squared error (RMSE) and mean absolute error (MAE) of the mean forecast that is implied by the density prediction. σ *RMSE* and σ *MAE* are the respective root mean squared error and mean absolute error of the implied volatility forecast.

Over-fitting in neural networks is mainly driven by the size of the network and expresses itself in a good in-sample, but bad out-of-sample performance. Our estimation methodology is designed to address the over-fitting issue.[11] Still, ensuring that over-fitting does not appear in our forward-looking density estimator is of first order importance. Rothfuss et al. (2019b) propose noise regularization to prevent over-fitting in Mixture Density Networks. In particular, they propose to add small random noise terms to the input and return data during the training phase. Intuitively, the noise slightly blurs the training data, thus making it impossible for the neural network to identify the small random pseudo-relationships in the training data that lead to over-fitting. At the same time, fundamental relationships between the inputs and the returns are unaffected as they still hold in expectation. More formally, Rothfuss et al. (2019b) show that noise regularization introduces a curvature penalty term into the objective function of the estimator and thus introduces a tendency to smooth the conditional density estimate.

We re-estimate our forward-looking density estimator with noise regularization to check whether this additional shield against over-fitting improves the out-of-sample performance. We treat the size of the noise that is added to the training data as a hyper-parameter, which is determined in the hyper-parameter search phase of our training algorithm. Table III.8 compares the out-of-sample performance of the MDN estimator with and without noise regularization. None of the performance figures changes notably.

---

[11] We determine the network size during the hyper-parameter search. In that, different parts of the training data set are treated as validation set, that is not used for model fitting. Over-fitting increases the error in these validation sets. The algorithm chooses the network size such that the error in the validation sets is minimized, thus counteracting over-fitting.

Independent of the use of noise regularization, the variation in the performance figures due to different starting values is low, indicating stable convergence of the estimation. Together with the observation from table III.2 that the performance increase of the MDN against the benchmark is nearly the same in the in-sample and the out-of-sample data set, we conclude that our conditional density estimator does not over-fit.

III.5 Summary

In this chapter, we presented a forecasting method for the full physical distribution of returns, that is based on Mixture Density Networks. Our estimator is forward-looking, as it is purely based on option-implied, risk-neutral expectations about the future, which are measured in a model-free way. The MDN places very little constraints on the statistical return distribution and is agnostic about the stochastic discount factor. Our approach can therefore be considered as model-free.

We showed that the forward-looking estimator generalizes predicting relationships better than a backward-looking estimator, which is based on past return information. It also outperforms a number of alternative parametric and non-parametric forward-looking return distribution estimators. While the out-of-sample conditional densities of the backward-looking estimator and the forward-looking alternatives are rejected in our statistical tests, this is not the case for the conditional forward-looking return distribution of the MDN. It is therefore well possible that the MDN correctly describes the true conditional physical return distribution. The MDN uncovers significant nonlinear relationships between risk-neutral and physical moments. Our results indicate that an exploitation of these nonlinearities, in combination with the adaptive complexity of our estimation approach, is key in forecasting the return distribution.

The inclusion of other types of forward-looking information like analyst forecasts (Ulrich et al., 2019) or text measures (Engle et al., 2019) may improve the performance of the estimator even further. Furthermore, while we restrict our analysis to the inspection of equity index returns, the method can easily be expanded to other asset classes with associated options.

IV The Pricing Kernel

The pricing kernel is at the heart of modern asset pricing theory. Knowledge about its empirical characteristics and time series evolution is therefore of central importance in the modeling of financial markets. By definition, the pricing kernel arises as the ratio of the risk-neutral and the physical density. Consequently, given estimates for these densities, it is possible to invert out the implied pricing kernel and inspect its empirical properties.

Jackwerth (2000) was the first to use this relationship to extract the pricing kernel, quickly followed by Aït-Sahalia and Lo (2000) and Rosenberg and Engle (2002). Following different approaches for estimating the $\mathcal{P}$ and $\mathcal{Q}$ densities of equity returns, all three studies identify local increases in the pricing kernel. This finding was at odds with the implications of most asset pricing models at that time, which generally assume a monotonically decreasing pricing kernel in equity returns, and was thus labeled the *pricing kernel puzzle*. Although the pricing kernel puzzle has been confirmed in many studies since, the vast majority of these studies use backward-looking estimators for the $\mathcal{P}$ density to compute the pricing kernel. At the same time, too restrictive models for the $\mathcal{P}$ density can produce the pricing kernel puzzle for technical reasons (Sichert, 2020). Consequently, Linn et al. (2018) proclaim the use of forward-looking estimators for the $\mathcal{P}$ density[1] and find monotonically decreasing pricing kernels, thus reviving a debate about whether the pricing kernel puzzle actually exists. We add to this debate by presenting new evidence about key characteristics of the pricing kernel in a forward-looking and model-free set-up.

In this chapter, we estimate the daily pricing kernel in a non-parametric manner, using forward-looking information only. To the best of our knowledge, this has not been done before. The advantage of our approach lies in the minimal amount of necessary assumptions, making our pricing kernel estimates largely free of exogenous constraints.

---

[1] A forward-looking $\mathcal{P}$ density is constructed based on expectations only, instead of observations about past events. In particular, all types of GARCH models are backward-looking. Mostly (including in this chapter), option data is used as input for forward-looking $\mathcal{P}$ densities. We discuss differences between backward-looking and forward-looking $\mathcal{P}$ density estimators in detail in chapter III.

In particular, we remain agnostic about preferences or the structure of the pricing kernel, adopting an approach that ”lets the data speak freely”.

Chapter III presented several forward-looking estimators of the $\mathcal{P}$ density, most of which are model-free in the sense that they can approximate any continuous probability density function solely by learning from the presented samples. Recovering the pricing kernels, that these methods imply, allows us to apply further tests to compare the estimators. We test statistically whether the recovered pricing kernels are unbiased discount factors for S&P 500 returns. Two neural network methods and a non-parametric conditional kernel density estimator for the $\mathcal{P}$ density pass this test at reasonable variance levels of the pricing kernel.

The three methods imply largely identical shapes for the recovered (log) pricing kernel, especially with respect to its central part. All methods identify a strong U-shape and a hump close to the zero log return. In order to be able to make reliable statements about the shape of the pricing kernel, we apply the test of Patton and Timmermann (2010) and directly test for shape characteristics. Tests of the pricing kernel puzzle in the literature usually either assume a parametric structure of the pricing kernel or only indirectly check for non-monotonicity. It appears like we are the first to apply non-parametric shape tests to estimates of the pricing kernel directly, especially to forward-looking pricing kernels. Our tests strongly confirm the existence of both a U-shape and a central hump in the unconditional log pricing kernel during our sample period between 2004 and 2017.

But do these shapes always exist? We construct daily measures for the magnitude of the U- and the hump shape and observe the U-shape reduce prior to the financial crisis and during its first months. After the bankruptcy of Lehman brothers, the U-shape strengthened again until 2010, since when it has become weaker. Despite these moves in the magnitude of the U-shape, it is always present at least to some extent. We can connect moves in the magnitude of the U-shape to the variance risk premium and the difference between risk-neutral and physical skewness. The situation is different for the hump shape. While it is clearly present in the non-crisis sample and statistically confirmed by our tests there, it vanishes during the financial crisis. In regressions of the magnitude of the hump on risk-neutral and physical return moments, we find that changes in risk-neutral moments can explain some of the variation in the hump's magnitude, but fluctuations in associated risk premia cannot. Our results indicate that the pricing kernel was flatter during the financial crisis than during calm times, as both the U- and the hump shape were less pronounced during the crisis.

Next, we combine our daily measurement of the pricing kernel for a spectrum of S&P 500 returns with our $\mathcal{P}$ density for S&P 500 returns and compute a daily estimate of the expected variance of the log pricing kernel. Expected variance was low in the beginning of the financial crisis. A low variance estimate implies that the pricing kernel is expected to experience only relatively small shocks, which is again a sign for a rather flat pricing kernel during that time. After a short spike following the Lehman collapse, variance only sustainably increased after the financial crisis and has experienced a general downward trend since. Applying the technique of Bollerslev and Todorov (2011), we split log pricing kernel variance into components due to continuous shocks and due to jumps. Overall, variation due to jumps accounts for 22% of pricing kernel variance, but the share of this component rises up to 60% at times, indicating that jumps are an important source of risk in the pricing kernel.

Finally, we use the tight theoretical connection between the pricing kernel and economic state variables to obtain a time series of realized economic shocks. We find that the distribution of these shocks is clearly left skewed and we can link spikes in the shock time series to important historic events. At the same time, economic shocks are strongly linked to the term spread, the credit spread, the Amihud (2002) illiquidity measure and the policy uncertainty index of Baker et al. (2016) in linear regressions. A negative economic shock is here connected to a decrease in the term spread and to an increase in the other three measures. Our time series of economic shocks can also well be explained by realized market returns and variance. Here, an one standard deviation upward shock in variance leads to a 0.62 standard deviation downward economic shock, while an one standard deviation upward return shock increases the economic shock by 0.35 standard deviations. Hence, variance appears to play a more important role for economic shocks than market returns do.

Our study adds to an active literature about the existence, the magnitude and the time series characteristics of the pricing kernel puzzle. After the puzzle had been identified by Jackwerth (2000), Aït-Sahalia and Lo (2000) and Rosenberg and Engle (2002), doubts about its existence re-emerged in Barone-Adesi et al. (2008), who find a monotonically downward sloping pricing kernel based on the asymmetric GARCH model of Glosten et al. (1993), fitted to both historic returns for the $\mathcal{P}$ density and option prices for the $\mathcal{Q}$ density. More formally, Barone-Adesi et al. (2012) cannot reject the hypothesis of monotonically decreasing pricing kernels in their statistical tests for most maturities, although releasing the parametric structure for the $\mathcal{Q}$ density and recovering it non-parametrically from option prices instead. However, their relinquishment of smoothing the implied volatility surface likely results in very noisy pricing kernel estimates, thus lowering the power of their applied tests.

In opposition to these findings, a number of studies confirm the pricing kernel puzzle empirically. Chabi-Yo et al. (2008) reproduce the findings of Jackwerth (2000) for the S&P 500 and in a sophisticated testing set-up Beare and Schmidt (2016) reject monotonically decreasing pricing kernels for that equity index. Grith et al. (2013) and Golubev et al. (2014) add international evidence by identifying the pricing kernel puzzle in data on the German DAX. At the same time, Chaudhuri and Schroder (2015) find generally decreasing pricing kernels for single stocks, which can be rationalized by the lower correlation between single stock returns and overall wealth. Cuesdeanu (2016) examines time variation in the pricing kernel puzzle for the S&P 500 and finds a w-shaped pricing kernel, with a hump around the zero return region during calm periods. Especially that central hump in the pricing kernel is examined by Sichert (2020), who confirms it conditionally, but who shows that at least its magnitude depends on the model that is used to construct the $\mathcal{P}$ density. His approach, a regime-swichting GARCH specification, significantly reduces humps, however, Sichert (2020) notes that the model might still not reflect nonlinearities accurately enough and its structure prevents it from capturing changes in conditional skewness. Note, that the conditional $\mathcal{P}$ density estimators in our study do not suffer from these drawbacks.

Among others, the study of Sichert (2020) shows the strong dependency of estimated pricing kernels on the backward-looking model specifications that are used in all of the previously mentioned studies to estimate the $\mathcal{P}$ density. For this reason, Linn et al. (2018) and Cuesdeanu and Jackwerth (2018a) make a strong case for using forward-looking $\mathcal{P}$ densities for analyses about the pricing kernel. The idea is not entirely new, an early example is Bliss and Panigirtzoglou (2004), but the technique has gained significantly more popularity after Ross (2015) proposed to recover the $\mathcal{P}$ density from option data alone. Evidence about the pricing kernel puzzle in forward-looking set-ups is mixed. Linn et al. (2018) remain agnostic about the shape of the pricing kernel and estimate a monotonically decreasing pricing kernel, thus concluding that the pricing kernel puzzle can be attributed to mis-specified backward-looking $\mathcal{P}$ density estimators. Opposingly, Audrino et al. (2015) and Cuesdeanu and Jackwerth (2018a) find U-shaped pricing kernels based on forward-looking $\mathcal{P}$ density estimators. The latter study explains the conflicting results with Linn et al. (2018) with their lack of scaling the $\mathcal{P}$ density and their specific choice of the objective function in fitting the pricing kernel estimates to the data. Jackwerth and Menner (2018) perform a detailed assessment of different Ross (2015) recovery schemes and reject decreasing pricing kernels in their statistical tests. All of these studies fit a fix forward-looking pricing kernel estimate to the data, which cannot capture time variation by design. Since we do not fit the pricing kernel, but estimate the $\mathcal{P}$ density directly, our approach allows to construct a daily estimate for the conditional pricing kernel.

Due to the high amount of studies that touch upon the pricing kernel puzzle, this short literature review is by no means complete and we certainly had to leave out important contributions. A very recommendable and comprehensive survey of the related literature can be found in Cuesdeanu and Jackwerth (2018b). Especially, their survey also discusses theoretical approaches for resolving the pricing kernel puzzle.

IV.1 Construction

Cochrane (2001, ch. 3.2) defines the risk-neutral probability $q(s)$ of a contingent future state s as a weighted version of the physical probability of that state $p(s)$, where the weight is determined by the pricing kernel $M(s)$:

$$q(s) = \frac{M(s)}{E(M(s))}\, p(s). \tag{IV.1}$$

In line with Jackwerth (2000), we use the cum dividend log return of the S&P 500 index $r_{t,\tau} = ln\, \frac{S_{t+\tau}+D_{t,\tau}}{S_t}$ as a measure for the state s, where S_t denotes the price of the S&P 500 in t and $D_{t,\tau}$ aggregates any dividend payments between t and $t+\tau$. Consequently, given estimates for the risk-neutral (physical) probabilities at time t, $q_t(r_{t,\tau})$ $(p_t(r_{t,\tau}))$, and the gross risk-free rate $R_{f,t,\tau} = \frac{1}{E(M_t(r_{t,\tau}))}$, one can invert equation IV.1 to

$$M_t(r_{t,\tau}) = \frac{q_t(r_{t,\tau})}{R_{f,t,\tau}\, p_t(r_{t,\tau})}. \tag{IV.2}$$

Chapters II and III are concerned with the precise estimation of the probabilities $q_t(r_{t,\tau})$ and $p_t(r_{t,\tau})$. We will build on the findings of these chapters to obtain our empirical estimate for the pricing kernel at time t for the monthly horizon.

For each of the methods described in chapter III, we train a physical density estimator for the cum-dividend log return $r_{t,\tau}$, with the horizon τ set to one month. For the risk-neutral probabilities, we obtain the volatility smile for the 30-day maturity from the volatility surface of the one-dimensional kernel regression and transform it to the risk-neutral density $q_t(K,\tau)$ as described in II.1. Before we can use these risk-neutral density estimates in equation IV.2, we need to measure the density with respect to cum-dividend log returns. Since the strike of an option contract is quoted in terms of the forward price of the underlying, it does not include dividends that are paid during the lifetime of the option. At the same time, option prices are informative about the risk-neutral expected dividend payment during the lifetime of the option, $E_t^{\mathbb{Q}}(D_{t,\tau})$, as described in section II.2.1. Setting $r_{t,\tau} = ln \frac{K+E_t^{\mathbb{Q}}(D_{t,\tau})}{S_t}$, by the law of integration by substitution (Rohde et al., 2012, ch. 3a), we can write for any two strikes K_a and K_b and their corresponding cum-dividend log returns $r_{t,\tau,a}$ and $r_{t,\tau,b}$

$$\int_{K_a}^{K_b} q_t(K,\tau)dK = \int_{r_{t,\tau,a}}^{r_{t,\tau,b}} q_t(S_t\, e^{r_{t,\tau}} - E_t^{\mathbb{Q}}(D_{t,\tau}),\tau)\, S_t\, e^{r_{t,\tau}} dr_{t,\tau} = \int_{r_{t,\tau,a}}^{r_{t,\tau,b}} q_t(r_{t,\tau})dr_{t,\tau}, \tag{IV.3}$$

which yields for the risk-neutral density (RND) with respect to cum-dividend log returns $q_t(r_{t,\tau}) = q_t(K,\tau)\, S_t\, e^{r_{t,\tau}}$.

Our thus obtained RND estimate has two unfavorable characteristic: First, the RND can become spiky, especially in the rarely traded tails of the distribution. This is a result of high sensitivity to noise in the numerical approximation of the second derivative of the Call price function. We therefore smooth the RND estimate to cancel out such approximation errors. We provide a detailed description of our smoothing methodology in section IV.1.1. Second, our RND estimates do not expand into the very deep tails of the distribution as options are not traded at such deep out-of-the-money strikes. We therefore follow Figlewski (2008) and fit the tails of a Generalized Extreme Value distribution to extrapolate the RND into the deep tails. Our procedure is described in detail in section IV.1.2.

IV.1.1 Smoothing the Risk-Neutral Density Estimate

A noisy RND estimate is characterized by the existence of multiple local modes and local minima, which can give the RND a "spiky" shape. While such a RND estimate does not hurt any of the formal criteria for a probability density function and is in line with no-arbitrage constraints, intuition speaks in favor of a smooth RND. This notion is strengthened when considering the sensitivity of a numerical second derivative approximation to noise in the Call price measurement (Iott et al., 1985).[2] For this reason, we further smooth the RND to cancel out approximation errors. In order to do so, we need to define a criterion that describes what a *smooth* RND actually is. Here, we need to balance the smoothness objective with the desire to stay as close as possible to the data and introduce as little restricting exogenous constraints as possible. We therefore consider those RNDs as smooth that fulfill the following rule:

A smooth density is unimodal.

Enforcing the property of a single mode on the RND effectively bans all spikes from the estimate. At the same time, saddle points in the RND estimate are accepted under this constraint.

We follow a local averaging approach to transform a raw RND estimate into a version that fulfills the smoothness constraint. The method smooths increasingly strongly until the constraint is fulfilled. More precisely, let $q_t(r_{t,\tau})$ stand for the original RND estimate with respect to cum-dividend log returns. Let $\mathbb{M}_{t,\tau} = \{r_i^{mode}\}_{t,\tau,i\in\mathbb{N}}$ be the set of log returns, where the RND has a mode. Modes are identified by the following set of rules:

$$q_t(r_{t,\tau,l}) + 0.0001\, q_t^{max} < q_t(r_i^{mode}) > q_t(r_{t,\tau,r}) + 0.0001\, q_t^{max} \tag{IV.4}$$

$$q_t^{max} = \max_{r_{t,\tau}} q_t(r_{t,\tau}) \tag{IV.5}$$

$$r_{t,\tau,l} = \begin{cases} \arg \min\limits_{r_{t,\tau} < r_i^{mode}} q_t(r_{t,\tau}) & \text{if } i = 1, \\ \arg \min\limits_{r_{t,\tau} \in (r_{i-1}^{mode}, r_i^{mode})} q_t(r_{t,\tau}) & \text{otherwise} \end{cases} \tag{IV.6}$$

$$r_{t,\tau,r} = \begin{cases} \arg \min\limits_{r_{t,\tau} > r_i^{mode}} q_t(r_{t,\tau}) & \text{if } i = |\mathbb{M}|, \\ \arg \min\limits_{r_{t,\tau} \in (r_i^{mode}, r_{i+1}^{mode})} q_t(r_{t,\tau}) & \text{otherwise} \end{cases} \tag{IV.7}$$

Here, inequality IV.4 states the identifying criterion for a mode, which is in essence a local maximum. Inequality IV.4 is a bit more restrictive than a plain local maximum criterion in that it requires the mode to stand out from the surrounding RND by a certain small amount of $0.0001\, q_t^{max}$. If a mode stands out less than that amount, we consider the RND as approximately constant in that region. Especially in the tails, where the RND takes very small values, it might well happen that very small local maxima emerge due

---

[2] While the kernel regression, that is used for constructing the volatility surface, attempts to smooth out noise in the Call price function, small error components may persist.

to computational inaccuracies and erasing these local maxima might require a very large amount of smoothing, which may result in over-smoothing. The minimum size condition for the local maximum in inequality IV.4 prevents that from happening and thus stabilizes the procedure and facilitates faster convergence.

The actual smoothing is done by applying a kernel regression with local bandwidth parameter to the RND estimate. The bandwidth parameter is iteratively increased, thus smoothing more and more strongly, until the smoothness constraint is fulfilled, i.e., $|\mathbb{M}| = 1$. We start by discretizing the RND estimate $q_t(r_{t,\tau})$ into J points of equal return step size $\{(q_j, r_j)\}_{j \leq J}$, with $r_j - r_{j-1} = \Delta r, j > 1$ and $q_j = q(r_j)$. As all procedures apply per point of time and maturity, we drop the time and maturity indices here in favor of a slim notation. We set $\Delta r = 0.005$, which is fine enough to capture all movements in the RND. Each point of the discretized RND gets assigned an individual bandwidth parameter b_j that is initially set to 0.01. We then apply algorithm 1.

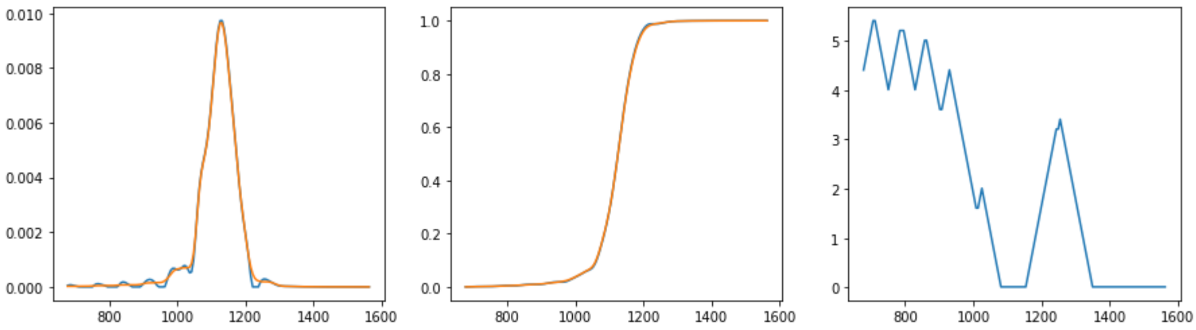
Algorithm 1: RND Smoothing

```
Input      : Raw RND estimate $q_j$
Parameters: $b_j \leftarrow 0.01$, $\Delta b \leftarrow 0.6$, $\theta \leftarrow 0.2$
Returns    : Smoothed RND estimate $\bar{q}_j$

$\bar{q}_j \leftarrow q_j$
$modeIndices \leftarrow$ findIndexOfModes($\bar{q}_j$)
while $size\ of\ modeIndices > 1$ do
    for $i \in modeIndices[1:]$ do
        $k \leftarrow \arg\min_j q_j, \quad modeIndices[i-1] \leq j \leq modeIndices[i]$
        $\tilde{b}_{i,j} \leftarrow b_j, \quad \forall j$
        $\tilde{b}_{i,k} \leftarrow \tilde{b}_{i,k} + \Delta b$
        $lb \leftarrow \tilde{b}_{i,k} - \theta$; $o \leftarrow 1$
        while $lb > 0.01$ do
            $\tilde{b}_{i,k+o} \leftarrow lb$; $\tilde{b}_{i,k-o} \leftarrow lb$
            $lb \leftarrow lb - \theta$; $o \leftarrow o + 1$
        end
    end
    $b_j \leftarrow \max_i \tilde{b}_{i,j}$
    $\bar{q}_j \leftarrow \frac{\sum_{l=1}^{J} \phi\left(\frac{j-l}{b_j}\right) q_l}{\sum_{l=1}^{J} \phi\left(\frac{j-l}{b_j}\right)}, \qquad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$
    $modeIndices \leftarrow$ findIndexOfModes($\bar{q}_j$)
end
```

Algorithm 1 identifies the local minimum between two modes of the current distribution estimate and increases the local bandwidth here. Qualitatively, a higher local bandwidth

Figure IV.1: Example for the Effect of RND Smoothing



The charts show the effect of smoothing on the 30-day maturity RND estimate of January 5th, 2004. The left plot shows the RND, the center plot shows the risk-neutral CDF and the right plot shows the local bandwidth parameter that is used in the smoothing procedure. The horizontal axes indicate the strike at which the RND is measured. In the left (center) plot, the original RND (CDF) estimate is drawn in blue, the smoothed version is drawn in orange. At a given strike, the original RND estimate is smoothed more strongly, the higher the bandwidth is at this strike. A local bandwidth of 0 indicates no smoothing at the respective strikes.

comes along with an increased smoothing in that area. In order to reach a smooth transition between highly and little smoothed areas of the density, the bandwidth is also increased in the region around the minimum, though less and less strongly the further away a point is from the minimum. Figure IV.1 shows a sample result of the smoothing procedure, including the local bandwidths that were used in smoothing. The local bandwidths are close to 0 in the central region of the RND, indicating that this region is effectively not smoothed at all. As there are no local minima in this region in the original RND estimate, it is already sufficiently smoothly measured according to our smoothness criterion. The bandwidth is only increased in the tails of the distribution, where the existence of local modes and minima suggests approximation errors. These approximation errors are then smoothed out due to increased local bandwidths.

Two parameters affect the smoothing behavior. The bandwidth incremental Δb and the bandwidth decay θ. Δb mainly impacts the convergence speed of smoothing. The higher Δb, the faster is the algorithm in reaching bandwidth parameters that allow a high degree of smoothing. However, if Δb is set too high, a single iteration might move the estimate from a multimodal, noisy RND to an over-smoothed RND. We set Δb to 0.6 to prevent over-smoothing to occur and the smoothing procedure to stop as soon as there's a single mode. To bring this value into perspective, keep in mind that the bandwidth parameter can be seen as the standard deviation of the Gaussian kernel function that is used as weighting function in the kernel regression. Since we form a local average of q_j based on the index j, neighboring raw RND observations have a distance of 1 ($(j+1)-j=1$) from each other. Increasing the standard deviation of the kernel by 0.6 in each observation is hence increasing the weight of a distant observation in the local average fairly slowly.

The bandwidth decay parameter θ must be chosen more carefully. If it is too large, the bandwidth decreases very strongly, resulting in a direct neighborhood of unsmoothed and smoothed sections of the RND. A sudden transition between such sections can introduce discontinuities and thus new noise, which may even make it impossible for the algorithm to produce unimodal RNDs. On the other hand, if θ is very small, the bandwidth is increased in sections of the RND that actually don't need any smoothing. We select θ by minimizing the squared pricing error of the smoothed RNDs in implied volatility space. To that end, we translate the smoothed RNDs to option prices and corresponding implied volatilities as described in section II.1. Denote the thus obtained implied volatility estimates as $\sigma(\bar{q}_t(K,\tau;\theta))$, where $\bar{q}_t(K,\tau;\theta)$ denotes the smoothed RND based on θ. θ is then selected in a grid search as the value that minimizes the squared implied volatility pricing error against all observed implied volatilities $\sigma_t(K,\tau)$:

$$\theta^* = \arg\min_{\theta} \sum (\sigma_t(K,\tau) - \sigma(\bar{q}_t(K,\tau;\theta)))^2 \tag{IV.8}$$

In the remaining exposition of this chapter, we replace the original RND estimate with its smoothed version, i.e., $q_t(r_{t,\tau}) = \bar{q}_t(r_{t,\tau};\theta^*)$.

IV.1.2 Tail Extrapolation

In its deep tail regions, the RND takes very small values, especially during times of low volatility. These low values are often too small to be accurately captured by the numerical approximation of the second derivative of the Call price function and thus rounded to 0. However, due to its definition (eq. IV.1), the RND must not become zero for any return with a non-zero physical probability. For all physical density estimators that we considered in chapter III, the physical probability is greater than 0 for all contingent future returns.

We solve this issue by following a procedure described by Figlewski (2008), who proposes to fit the tail of a Generalized Extreme Value (GEV) distribution to the empirical tails of the RND and to subsequently replace the empirical tails with the fitted GEV tails. The theoretical foundation of this approach lies in the Fisher-Tippett theorem, which states that the maximum sample, that was drawn from any unknown distribution, converges in distribution to a Gumbel, a Fréchet, or a Weibull distribution (McNeil et al., 2015). These three distributions are special versions of the GEV distribution. The GEV distribution hence plays a similar role for the tail of a distribution as the normal distribution plays for its mean estimate in the central limit theorem.

A GEV distribution is characterized by three parameters: its location $\mu \in \mathbb{R}$, its scale $\sigma > 0$ and a shape parameter $\xi \in \mathbb{R}$. Its probability density function is given by

$$p^{GEV}(x;\mu,\sigma,\xi) = \frac{1}{\sigma}(t(x))^{\xi+1}e^{-t(x)}, \quad t(x) = \begin{cases} \left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ e^{-\frac{x-\mu}{\sigma}}, & \text{if } \xi = 0 \end{cases} \tag{IV.9}$$

Figlewski (2008) proposes to determine the three parameters by matching 2 RND values and the value of the risk-neutral CDF at a given percentile. However, this approach is sensitive to noise in the RND estimates at the selected percentile. We therefore fit the parameters of the GEV distribution such that the squared error between the empirical RND estimate and the GEV tail is minimized. For a day t and maturity τ, let C^r be the empirical 97.5% percentile of the RND $q(r)$ (the time and maturity indices are dropped for brevity). For the right tail, the parameters of the respective GEV distribution are then determined by solving the optimization problem

$$\{\mu^*,\sigma^*,\xi^*\}_{right} = arg \min_{\{\mu,\sigma,\xi\}} \int_{C^r}^{\infty} (q(r) - p^{GEV}(r;\mu,\sigma,\xi))^2 dr. \tag{IV.10}$$

Since the Fisher-Tippett theorem applies to the maximum of a distribution, we need to flip the RND to approximate the left tail with a GEV distribution. Let $q'(r) := q(-r)$ and C^l be the empirical 2.5% percentile of the RND $q(r)$. The GEV parameters for the left tail can then be obtained by solving

$$\{\mu^*,\sigma^*,\xi^*\}_{left} = arg \min_{\{\mu,\sigma,\xi\}} \int_{-C^l}^{\infty} (q'(r) - p^{GEV}(r;\mu,\sigma,\xi))^2 dr. \tag{IV.11}$$

We approximate the integrals in equations IV.10 and IV.11 by discretely sampling the empirical RND in steps of $\Delta r = 0.005$ and computing the GEV tail at the same locations. We then replace the integral by a sum and dr with Δr. Having found the optimal parameters for the GEV tail, we replace the empirical RND tail for returns below the 2.5% percentile (above the 97.5% percentile) with the respective left (right) GEV tail.

IV.2 Data

Our analysis starts with option data, which we use to construct the end-of-day volatility surface. Based on that volatility surface, we compute both empirical estimates for the RND and the risk-neutral measures that are needed as input for the forward-looking $\mathcal{P}$ density. Realized shocks to the pricing kernel are then analyzed in the light of popular risk factors and further economic measures.

We collect end-of-day bid/ask quotes for S&P 500 options from the Chicago Board Options Exchange (CBOE). The data set includes matched underlying bid/ask prices, which were sampled at the closing time of the option market. Our option data covers the period between January 2004 and July 2017. We further collect daily Overnight Index Swap (OIS) rates from Bloomberg for the same time period, which we use as our measure for the risk-free rate. The OIS data comes in fix maturities between 1 day and 20 years and we interpolate the rates linearly to the respective option maturities. We apply the one-dimensional kernel regression methodology of section II.2.2 to construct the daily implied volatility surface based on mid prices for options and the underlying. Using the volatility smile of the 30-day maturity, we compute a daily risk-neutral density estimate with the procedures of section IV.1.

For each observed option maturity, we compute an estimate for the SVIX^2 of Martin (2017) and the model-free risk-neutral skewness and kurtosis of Bakshi et al. (2003). We linearly interpolate along the maturity dimension to obtain measures at a fix maturity of 30 calendar days. We use this calculation scheme instead of estimating the measures directly from the 30-day volatility smile to stick with conventions in the literature and industry (Martin, 2017; CBOE, 2019). We collect daily prices of the S&P 500 total return index, S_t, from Bloomberg and compute log returns $r_t = ln\, S_{t+\tau} - ln\, S_t$, $\tau = 30$ calendar days. If $S_{t+\tau}$ is not observed (e.g. because day $t + \tau$ is on a weekend), we use the most recent observed price. We fit the density estimators of sections III.1.1 and III.1.2 to obtain daily estimates for the forward-looking $\mathcal{P}$ density of cum-dividend S&P 500 log returns, namely, the unconditional kernel density estimator as benchmark (*Benchmark*), the conditional Gaussian density estimator where the mean and variance are linear in the input variables (*Linear*), the conditional kernel density estimator (*CKDE*), the neighborhood conditional kernel density estimator (*NKDE*), the least-squares conditional density estimator (*LSCDE*) and the two neural network estimators based on a kernel mixture network (*KMN*) and a mixture density network (*MDN*).

Table IV.1: Performance of Monthly Conditional Density Estimators

Model	Avg. log $\mathcal{L}$	Mean RMSE	σ RMSE	Berkowitz (2001)	Knüppel (2015)
Benchmark	1.8514	4.4463	3.5679	0.973	0.890
Linear	1.9140	4.4452	3.1255	0.908	0.361
CKDE	2.0256	4.4130	3.0158	0.764	0.832
NKDE	1.7584	4.8432	4.0280	0.000	0.037
LSCDE	1.6777	4.8485	4.1029	0.000	0.083
MDN	2.0431	4.3505	3.0373	0.639	0.767
KMN	2.0600	4.2525	3.0130	0.228	0.330

This table compares the performance for a range of forward-looking conditional density estimators for the monthly cum-dividend return density of the S&P 500 in our data sample. The methods are the same as in chapter III. The second column shows the average conditional log likelihood of realized monthly returns. The third and the fourth column report the root mean squared errors for the mean and standard deviation prediction in percent. Columns five and six show p-values for the tests of Berkowitz (2001) and Knüppel (2015), based on non-overlapping monthly return realizations.

Table IV.1 compares key performance measures for the monthly $\mathcal{P}$ return density predictions. The results broadly agree with the findings for daily returns in chapter III. Generally, the neural network methods outperform the alternatives in terms of average log likelihood and the root mean squared error of the mean and volatility prediction. At the same time, the *CKDE* method is nearly at par with the neural network methods for monthly return density predictions. This behavior was not observed for daily return densities in chapter III. The tests of Berkowitz (2001) and Knüppel (2015) require independent return draws from the proclaimed distributions, which necessitates the use of non-overlapping return realizations in the computation of their test statistics. The reduced number of independent monthly return observations makes the tests only weakly informative. At the 1% significance level, we can only reject the conditional densities of the *NKDE* and the *LSCDE* methods based on the test of Berkowitz (2001) and we cannot reject any conditional densities based on the test of Knüppel (2015). We will therefore compute pricing kernels for all methods and apply additional tests on these in our main analysis.

Finally, we collect returns for the US market, size, value and momentum risk factors from Kenneth French's website. The investor sentiment index (*Sent*) of Baker and Wurgler (2006) is sampled from the website of Jeffrey Wurgler and the daily news-based economic policy uncertainty index (*EPU*) of Baker et al. (2016) is collected from the website of

Table IV.2: Summary Statistics

			Percentiles				
	Mean	Volatility	10%	25%	50%	75%	90%
r_f^{3m}	1.203	1.666	0.02	0.06	0.19	1.86	4.55
$Term$	1.934	1.065	0.21	1.42	2.04	2.67	3.31
$Credit$	2.663	0.836	1.7	2.12	2.62	3.04	3.33
TED	0.465	0.462	0.19	0.22	0.32	0.48	0.97
$Sent$	-0.028	0.351	-0.58	-0.19	0.01	0.19	0.37
EPU	102.84	68.5	36.36	54.75	85.51	132.69	191.85
$ILLIQ$	6.04	6.9	0.61	1.63	3.9	8.05	13.62

The table shows summary statistics for measures that we use in our analysis. r_f^{3m} is the yield of 3-month treasuries, $Term$ is the yield spread between 10-year and 3-month treasuries, $Credit$ denotes the spread between the Moody's seasoned Baa yield and 10-year treasury yields and TED stands for the difference between 3-month Libor rates and the yield of 3-month US treasuries. $Sent$ denotes the investor sentiment index of Baker and Wurgler (2006), EPU is the news-based economic policy uncertainty index of Baker et al. (2016). $ILLIQ$ is the Amihud (2002) illiquidity measure. Since the latter is very small, we multiple it with 10^{15}. All yields and spreads are given in annualized percentage values. The data set spans the period from January 2004 to July 2017.

Scott Baker, Nick Bloom and Steven Davis. We also collect several interest rate related data sets (our labels are given in parenthesis) from the Fred database of the Federal Reserve Bank of St. Louis, namely, the yield on 3-month treasury bonds (r_f^{3m}), the spread between 10-year and 3-month treasury yields ($Term$), the spread between the Moody's Seasoned Baa bond yields and 10-year treasury yields ($Credit$) and the spread between the 3-month Libor and 3-month treasury yields (TED). Daily S&P 500 volume figures for the calculation of the Amihud (2002) illiquidity measure ($ILLIQ$) are sampled from Bloomberg. All of these time series cover at least the time period of our option data. The investor sentiment index of Baker and Wurgler (2006) is only available in end-of-month resolution. We interpolate its values linearly between two observations to obtain daily estimates.

Many of the data sets that we use in this study have been presented in previous chapters already. We therefore only provide summary statistics for the new data sets in table IV.2 and refer to tables II.3 and III.1 for information on the remaining data sets.

IV.3 Empirical Pricing Kernels

We start our evaluation of forward-looking pricing kernels by checking whether alternative forward-looking $\mathcal{P}$ densities produce valid unconditional pricing kernels in the first place. Since a valid $\mathcal{P}$ density estimate should produce valid pricing kernel estimates, this analysis can be seen as a further evaluation step of the forward-looking density prediction methods that were discussed in chapter III. Before we then turn to the inspection of time variation in pricing kernels in the next section, we formally describe the shape of the average pricing kernel, including statistically testing key shape characteristics. All pricing kernel estimates in this and the subsequent sections refer to the monthly maturity.

IV.3.1 Comparison

The Euler equation $E[M_t(r_{t,\tau})\, exp(r_{t,\tau})] = 1$ is among the core concepts of financial research. Under the absence of arbitrage opportunities, the equation must hold for the true pricing kernel (Cochrane, 2001), such that any pricing kernel estimator that systematically hurts the constraint can be rejected. This thought gives rise to a simple test for a candidate pricing kernel estimator: Given return realizations $\mathbf{r}_{t,\tau}$, $t \in [1, ..., T]$, we compute the *realized* pricing kernel $\mathbf{M}_{t,\tau} = M_t(\mathbf{r}_{t,\tau})$ as the pricing kernel in t at the realized return in $t + \tau$. We can now compute the average discounted realized return $\overline{\mathbf{M}_\tau \, \mathbf{R}_\tau} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{M}_{t,\tau} \, \mathbf{R}_{t,\tau}$, $\mathbf{R}_{t,\tau} = e^{\mathbf{r}_{t,\tau}}$ and apply a t-test to test for the null hypothesis

$$H_0: \quad E(\overline{\mathbf{M}_\tau \, \mathbf{R}_\tau}) - 1 = 0. \tag{IV.12}$$

The t-test of the Euler equation can help identify biased pricing kernel estimators. However, it is not very effective if the volatility of the realized pricing kernel $\mathbf{M}_{t,\tau}$ is very large.[3] At the same time, by rearranging the Euler equation, we know for the equity risk premium (Cochrane, 2001, ch.1)

$$E(R_{t,\tau}) - R_{f,t,\tau} = -R_{f,t,\tau} \rho \sigma_M \sigma_R, \tag{IV.13}$$

where σ_M is the unconditional volatility of the pricing kernel, σ_R is the unconditional volatility of the market return and ρ is the correlation between the two. $R_{f,t,\tau}$, $E(R_{t,\tau}) - R_{f,t,\tau}$ and σ_R can be estimated at high precision from market data. A very volatile pricing

---

[3] A large volatility of $\mathbf{M}_{t,\tau}$ translates in a large volatility of $\mathbf{M}_{t,\tau} \, \mathbf{R}_{t,\tau}$ (assuming no strong negative correlation) and thus a large standard error for the $\overline{\mathbf{M}_\tau \, \mathbf{R}_\tau}$ estimate.

Table IV.3: Comparison of Pricing Kernels

$\mathcal{P}$ Density	p-value Euler eq.	σ_M	HJ bound
Benchmark	0.004	249.89%	6.289
Linear	0.223	4896.26%	67.6
CKDE	0.914	64.46%	2.232
NKDE	0.000	288.53%	6.672
LSCDE	0.141	32438.66%	68.878
KMN	0.076	43.16%	1.553
MDN	0.068	41.29%	1.483

The table shows key characteristics of different forward-looking pricing kernel estimators. *Benchmark* is an unconditional kernel density estimator for returns, *Linear* makes mean and variance of a Gaussian linearly dependent on the forward-looking variables. *CKDE*, *NKDE* and *LSCDE* are conditional non-parametric kernel density estimators. The *KMN* and *MDN* methods use neural networks to predict the $\mathcal{P}$ density. The second column shows the p-value for the t-test of equation IV.12, the third column shows the unconditional volatility estimate for the realized pricing kernel at the monthly maturity and the fourth column shows the Hansen and Jagannathan (1991) bound of the unconditional pricing kernel estimate.

kernel estimate (high σ_M) would therefore imply a low absolute correlation between the market return and the pricing kernel and thus shocks to aggregate wealth (low $|\rho|$). Intuitively, a very low correlation between these two quantities appears unrealistic. We therefore use the volatility of the pricing kernel realizations σ_M as a further measure for comparing our pricing kernel estimators.

Finally, Hansen and Jagannathan (1991) showed that equation IV.13 gives rise to an upper bound for the absolute Sharpe ratio of any tradable portfolio:

$$\frac{\sigma_M}{E(M_t(r_{t,\tau}))} \geq \frac{|E(R_{t,\tau}) - R_{f,t,\tau}|}{\sigma_R}. \tag{IV.14}$$

With a slight abuse of notation, we let $R_{t,\tau}$ stand for the gross return of any asset in this equation and σ_R for its corresponding volatility. The condition directly follows from the fact that $R_{f,t,\tau} = \frac{1}{E(M_t(r_{t,\tau}))}$ and that the absolute correlation between an asset's return and the pricing kernel must be smaller or equal to 1, $|\rho| \leq 1$. It is important to note that a portfolio with the Sharpe ratio at the Hansen-Jagannathan (HJ) bound is indeed tradable in a complete market. HJ bounds that are much higher than typically observed Sharpe ratios are therefore puzzling as they would imply that highly compensated future states cannot be traded.

Table IV.3 compares the alternative forward-looking pricing kernel estimators. At the

1% significance level, the t-test of the Euler equation based on realized market returns rejects the null hypothesis for the unconditional *Benchmark* method and the *NKDE* method. These methods appear to produce biased pricing kernels. While not being rejected in the test of the Euler equation, the *Linear* and *LSCDE* methods show extremely high estimates for the volatility of the realized pricing kernel and subsequently very high HJ bounds. Given the observed average risk-free rate, equity risk premium and return volatility in our data sample and following equation IV.13, these figures would translate into an unconditional correlation between the pricing kernel and the market return of -0.003 for the *Linear* method and -0.0004 for the *LSCDE* method. These correlation estimates appear unrealistically close to 0. The findings go hand in hand with the comparably low prediction performance of the *Linear* and the *LSCDE* method for the monthly $\mathcal{P}$ density of returns (see table IV.1). We therefore conclude that these two methods are not well suited for building realistic estimates of the pricing kernel.

The remaining three methods, the non-parametric *CKDE* and the two neural network estimators *MDN* and *KMN* appear to produce valid pricing kernel estimates. They cannot be rejected in a test of the Euler equation at the 5% significance level, their pricing kernel volatility estimates are roughly aligned and their estimates for the HJ bound appear realistic for the monthly horizon. The three methods also showed the highest prediction performance for the monthly $\mathcal{P}$ density of S&P 500 returns in table IV.1, with the neural network methods slightly ahead of the *CKDE* method. We will therefore use all three methods in our inspection of the shape of the unconditional pricing kernel in the next subsection.

IV.3.2 The Shape of the Unconditional Pricing Kernel

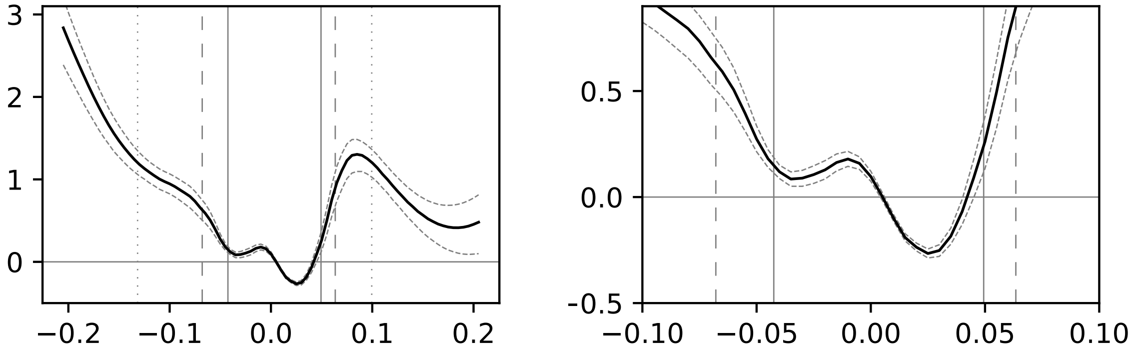
The shape of the unconditional forward-looking pricing kernel is subject of an on-going debate in the literature, which we sketched in the introduction of this chapter. We add to this debate here by applying statistical tests to check for the existence of a U-shape and a central hump in average pricing kernels based on forward-looking data. To that end, we first compute the average log pricing kernel as a function of the market return

$$\overline{m}(r_\tau) = \frac{1}{T}\sum_{t=1}^{T} m_t(r_{t,\tau}), \quad m_t(r_{t,\tau}) = ln\, M_t(r_{t,\tau}). \tag{IV.15}$$

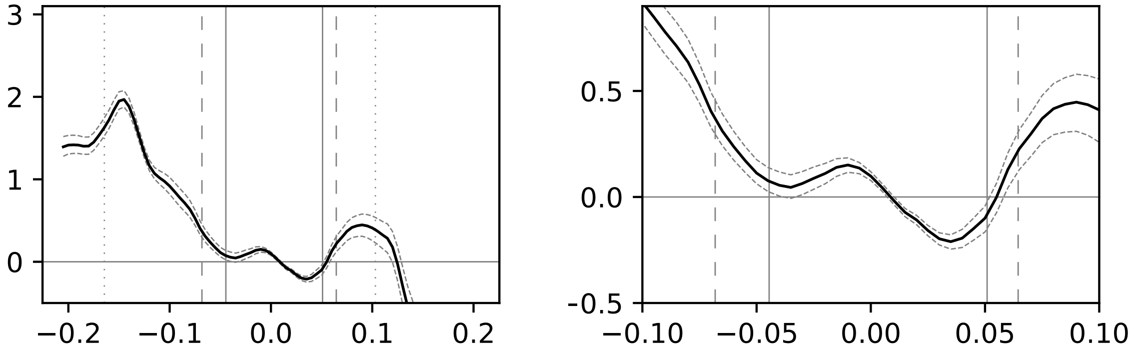
Figure IV.2 shows the average log pricing kernels for the *CKDE*, the *KMN* and the *MDN*

Figure IV.2: Average Empirical Log Pricing Kernels

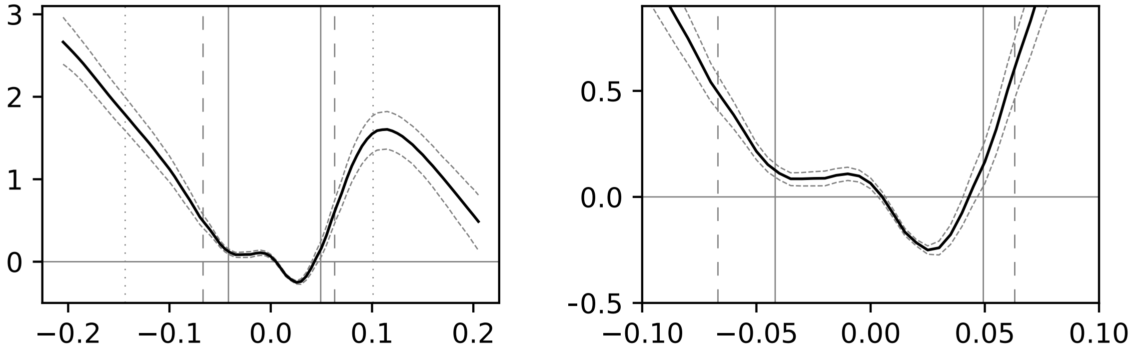
(a) CKDE, full return spectrum (b) CKDE, central returns



(c) KMN, full return spectrum (d) KMN, central returns



(e) MDN, full return spectrum (f) MDN, central returns



Each row shows the average forward-looking log pricing kernels of another method over log returns. All pricing kernels refer to the monthly maturity. The panels on the left show the average log pricing kernel for the full spectrum of log returns, the panels on the right zoom in on the important central part of the log return spectrum. Bootstrapped 90% confidence intervals are drawn in dashed lines. The vertical lines show percentiles of the average return distribution: the 10% (90%) percentiles are drawn in solid lines, the 5% (95%) percentiles are dashed and the 1% (99%) percentiles are dotted.

methods. For all methods, the central part of the average pricing kernels, between the 5% and 95% percentiles of the return distribution, looks very much alike. The pricing kernels follow an overall U-shape, with a clear hump at small negative returns. Judging from the plots, it is not evident whether the local increase at the hump is robust or whether a locally stable pricing kernel constitutes the hump. Our shape tests will answer this question shortly.

The fact that most of our log pricing kernel estimates are positive may seem puzzling at first, since the pricing kernel should discount a certain payoff in expectation. However, our charts do not show how likely a certain return realization is. In fact, the regions of the return spectrum where the log pricing kernel is negative are highly likely to realize. Consequently, our estimated pricing kernels discount strong enough to offset the equity risk premium, such that we cannot reject that the Euler equation holds.[4]

For large positive and negative returns, the average log pricing kernel estimates diverge for the *CKDE* and *MDN* method on the one hand and the *KMN* method on the other hand. However, as we will show in section IV.4.1, the *KMN* method does not produce reliable pricing kernel estimates in crisis times, which are the times when most tail events realize. For this reason, we discard the *KMN* estimates in our description of the shape of the pricing kernel in the tails of the return distribution. For these regions of the return distribution, the *CKDE* and the *MDN* imply very similar shapes. As expected, the pricing kernel is increasing as one moves deeper into the left tail. For the right tail however, the pricing kernel estimates turn again after forming the central U-shape and eventually decrease for large positive returns. Our findings are consistent with investors who dislike losses and variance, thus forming the central U-shape. The decreasing pricing kernel in the right tail can be reconciled if variance aversion gets at some point outweighed by the utility increase due to large positive returns.

We will now turn to our statistical tests for the U-shape and the local increase at the central hump. Similar to the shape tests in chapter III, we will apply the testing framework of Patton and Timmermann (2010). Their test was originally developed to test for monotonically increasing average returns in a portfolio sorting exercise, but the technique can easily be adjusted to monotonic decreases or zig-zag shapes by simply flipping signs. We start by defining what we consider a U-shape: Let r_{min} be the return where the average log pricing kernel reaches its global minimum. Moreover, let $r_l = -2\sigma_r$ and $r_r = 2\sigma_r$, with

---

[4] The average monthly cum-dividend return for the S&P 500 has been 0.75% in our sample.

Table IV.4: Shape Tests for Unconditional Pricing Kernels

$\mathcal{P}$ Density	U-shape	hump-shape
CKDE	0.000	0.000
KMN	0.000	0.000
MDN	0.000	0.010

The table shows p-values for the H_0 hypotheses of equations IV.16 and IV.17, which are obtained by bootstrapping the test statistic 10000 times using the stationary bootstrap of Politis and Romano (1994) with an average block size of 22.

the unconditional log return volatility σ_r. A U-shape is given if $\overline{m}(r_l) > \overline{m}(r_{min}) < \overline{m}(r_r)$ holds. For our test, this gives rise to the null hypothesis

$$H_0^U : \quad \exists\, \Delta_i \leq 0, \quad i \in \{1, 2\},\ \Delta_1 = \overline{m}(r_l) - \overline{m}(r_{min}),\ \Delta_2 = \overline{m}(r_r) - \overline{m}(r_{min}). \quad \text{(IV.16)}$$

If this null hypothesis is rejected, we must conclude that the U-shape in the log pricing kernel estimates is robust within the return interval $[-2\sigma_r, 2\sigma_r]$. For the central hump, we proceed similarly. Let r_{lmax} be the return where the average log pricing kernel reaches its local maximum in the hump and r_{lmin} be the return where the average log pricing kernel has its local minimum left of r_{lmax}. A hump is given, if we can reject the null hypothesis

$$H_0^{hump} : \quad \exists\, \Delta_i \leq 0, \quad i \in \{1, 2\},\ \Delta_1 = \overline{m}(r_{lmax}) - \overline{m}(r_{min}),\ \Delta_2 = \overline{m}(r_{lmax}) - \overline{m}(r_{lmin}). \quad \text{(IV.17)}$$

Following Patton and Timmermann (2010), the test statistic is given by $min\, \Delta_i$ and we obtain the distribution of the test statistic under the null hypothesis based on 10000 bootstrapped samples of $\overline{m}(r_\tau)$. We use the stationary bootstrap of Politis and Romano (1994) with an average block size of 22 days to account for time dependencies, which exist due to the overlapping horizons of our daily pricing kernel estimates.

The bootstrapped p-values for H_0^U and H_0^{hump} can be found in table IV.4. Clearly, H_0^U is rejected for all methods. We can thus conclude that a U-shape exists in the central region of the return spectrum. A similar observation can be made for the hump shape. H_0^{hump} is rejected at any confidence level for the *CKDE* and the *KMN* estimators and at the 1% confidence level for the *MDN* estimator. Thereby, our tests explicitly require a local increase in the pricing kernel to form the hump. A hump-shaped disturbance of the overall U-shape may also be formed by a saddle point in the pricing kernel. However, this possibility was explicitly rejected in our tests and we conclude that the average log pricing kernel has a local increase for slightly negative returns, thus forming a hump, and a larger increase for medium positive returns, thus resembling a U on a larger scale.

IV.4 Time Series Characteristics

A key advantage of our approach is the possibility to obtain daily pricing kernel estimates. We can thus assess the evolution of conditional pricing kernels. In a first step, we analyze how the shape of the pricing kernel changes over time. In that, we will put special attention to the differences between crisis times and calm times. Afterwards, we will inspect the conditional variance of the pricing kernel. Subsequently splitting our variance estimates into a component for continuous and a component for jump risk completes this section.

IV.4.1 Crisis and Calm Times Pricing Kernels

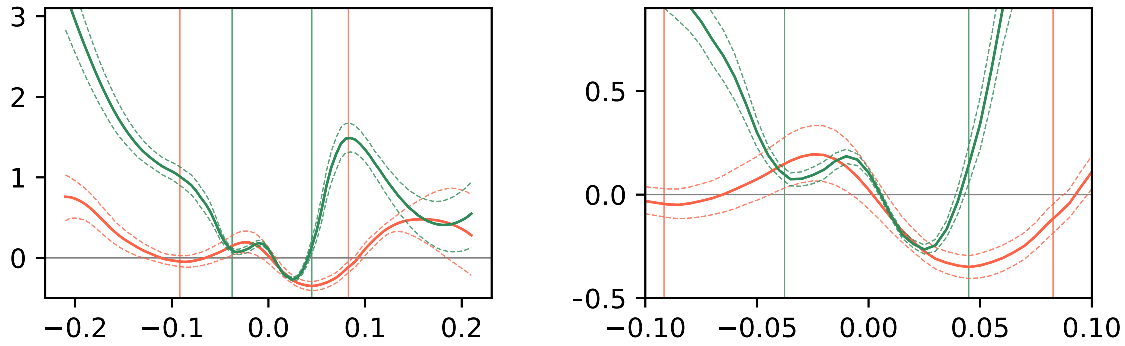
We start our evaluation of conditional pricing kernels by repeating the analysis of section IV.3 for crisis periods and calm periods, separately. The NBER reports a single recession period during our data sample, the financial crisis between December 2007 and June 2009. We use this time period as our crisis sample and the remaining data as our calm sample. Table IV.5 shows the p-values for the null hypothesis that the Euler equation holds for each sample. While we cannot reject the null hypothesis for the *CKDE* and *MDN* pricing kernels with confidence, the *KMN* pricing kernels appear to not fulfill the Euler equation in crisis times at the 1% significance level. For this reason, we discard this method from further analysis of time variation.

Table IV.5: Testing the Euler Equation for Sub-Samples

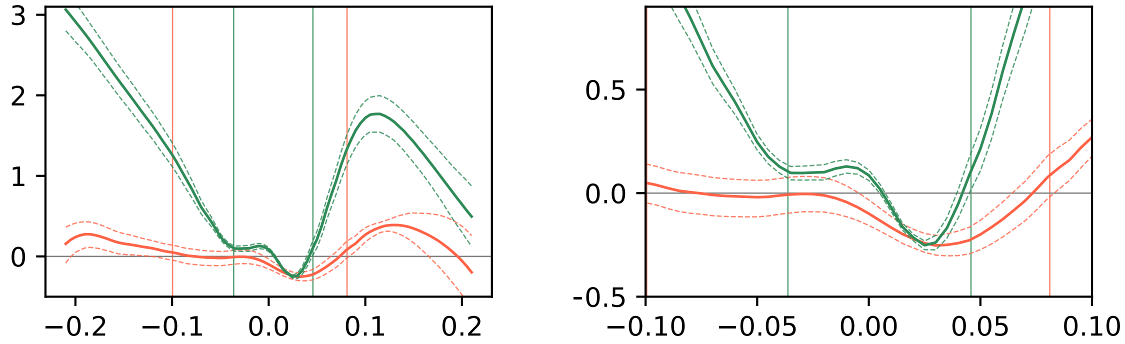
$\mathcal{P}$ Density	Crisis Sample	Calm Sample
CKDE	0.675	0.911
KMN	0.006	0.321
MDN	0.159	0.14

The table shows p-values for the null hypothesis that the Euler equation is fulfilled by the pricing kernel estimates of some methods. *CKDE* is the conditional non-parametric kernel density estimator. The *KMN* and *MDN* methods use neural networks to predict the $\mathcal{P}$ density. The crisis sample spans the NBER recession period between 12/2007 and 06/2009. The calm sample combines the periods from 01/2004 - 11/2007 and 07/2009 - 07/2017.

Figure IV.3: Average Log Pricing Kernels by Economic Regime



(a) CKDE, full return spectrum (b) CKDE, central returns



(c) MDN, full return spectrum (d) MDN, central returns

Each row shows the average month-ahead empirical log pricing kernels of another method over log returns. The panels on the left show the average log pricing kernel for the full spectrum of log returns, the panels on the right zoom in on the central part of the log return spectrum. The average log pricing kernel during crisis times is plotted in red, the calm times average log pricing kernel is plotted in green. The vertical lines indicate the 10% (90%) percentiles of the respective average return distribution. The percentiles for the crisis sample are in red, for the non-crisis sample in green.

A visual comparison of the average log pricing kernel during crisis times and calm times can be found in figure IV.3. We compute these period-specific average pricing kernels by applying equation IV.15 for the crisis sample and calm sample, separately. Comparing the estimates yields some interesting insights. First, the calm period pricing kernels look very similar to the shapes of the unconditional pricing kernels. This may be a result of the dominance of the calm period in our full data sample. At first sight, crisis times pricing kernels look very different from calm times pricing kernels. However, return volatility is much higher during crisis times, as indicated by the vertical lines, which depict the 10% and 90% percentiles of the conditional return distribution. The vertical lines cut the respective pricing kernels at similar "points" in their evolution. For example, the 10% percentile cuts the *CKDE* pricing kernel close to its left local minimum. This is

Table IV.6: Shape Tests for Conditional Pricing Kernels

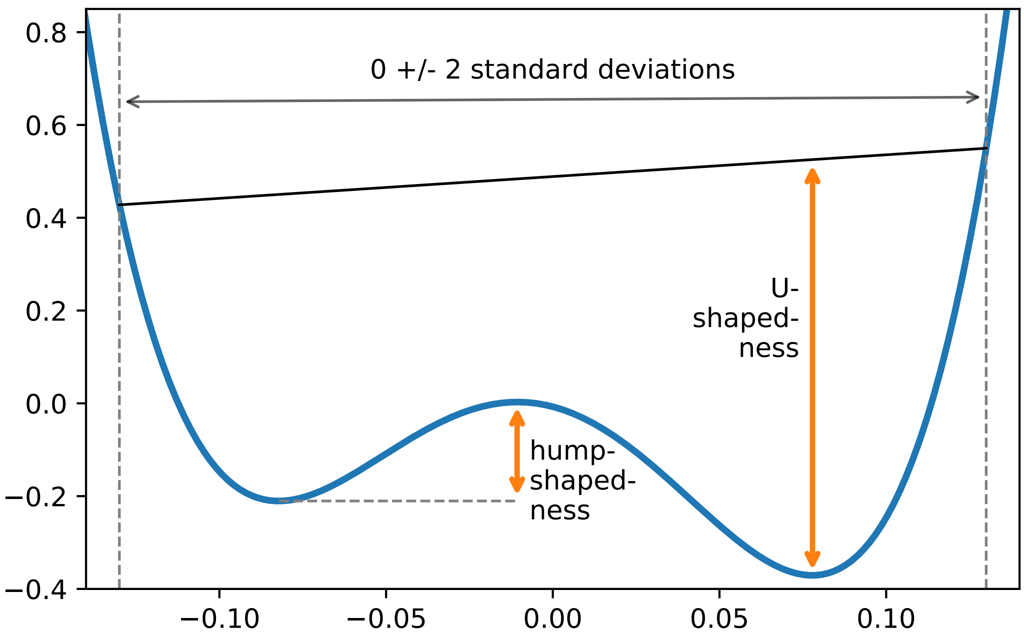
	Crisis Sample		Calm Sample	
$\mathcal{P}$ Density	U-shape	hump shape	U-shape	hump shape
CKDE	0.000	0.000	0.000	0.000
MDN	0.000	0.182	0.000	0.001

The table shows p-values for the H_0 hypotheses of equations IV.16 and IV.17, which were obtained by bootstrapping the test statistic 10000 times using the stationary bootstrap of Politis and Romano (1994) with an average block size of 22. The crisis sample spans the period between 12/2007 and 06/2009. The calm sample combines the periods from 01/2004 - 11/2007 and 07/2009 - 07/2017.

the same for the crisis times and the calm times pricing kernel. On the other side of the distribution, the 90% percentile cuts the *MDN* pricing kernel right after it turned positive, again the same for the crisis times and calm times pricing kernel. It seems like crises times pricing kernels resemble horizontally stretched versions of calm times pricing kernels. Such an observation is consistent with the existence of one or several unpriced components in equity index returns and a higher volatility of these components during crisis times. In that case, the return density would widen during a crisis while the pricing kernel remains unchanged, thus stretching the shape of the pricing kernel. Still, crisis times pricing kernels are different even from their stretched calm times counterparts. They generally appear to be slightly lower, which is consistent with stronger discounting during crisis times. Also, the magnitude of the U- and hump shapes may be different, which we analyze in detail in the next subsection.

The p-values of our shape tests for the U- and the hump shape in crisis and calm times log pricing kernels can be found in table IV.6. The null hypothesis of no U-shape is clearly rejected by our tests for both pricing kernel estimators. This finding holds for both crisis times and calm times. Note that we use the period-specific volatility of the underlying for identifying r_l and r_r, thus neutralizing potential effects of horizontal stretching. For the hump shape, the results are less aligned between methods. While we can reject the null hypothesis of no hump shape at all times for the *CKDE* estimator, we cannot confirm this finding with the *MDN* estimator for crisis times. Here, a hump is identified for calm times at the 0.1% confidence level, but we cannot conclude the existence of a hump with a local increase in the pricing kernel during crisis times. This finding reinforces a question that asks about the time variation in the magnitude of the U- and hump shapes.

Figure IV.4: Illustration of the Shape Measures for the Pricing Kernel



The figure illustrates how the U- and hump shape measures for the pricing kernel are defined. The blue line indicates a log pricing kernel estimate over log returns on the horizontal axis. The orange lines indicate the shape measures. In the chart, we depict the case of the majority of observations, where the hump is to the left of the global minimum of the pricing kernel.

IV.4.2 The Shape of Conditional Pricing Kernels

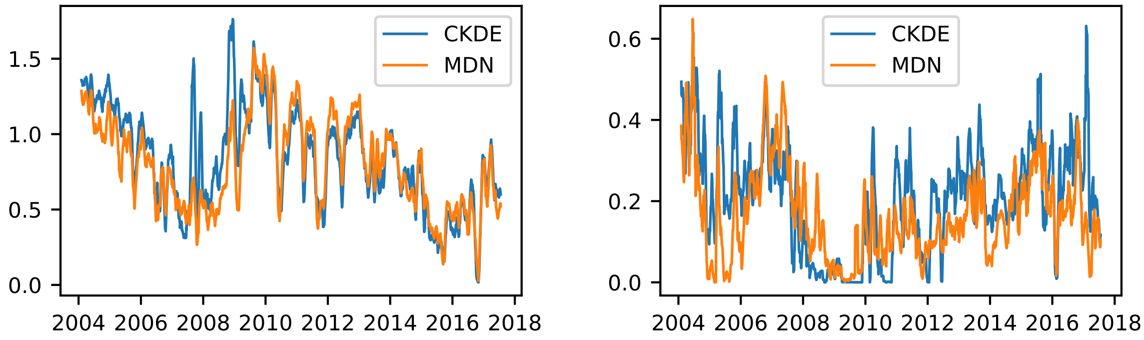
We now continue with the investigation of day-to-day time variation in the shape of the pricing kernel. We define two measures for the magnitude of the U-shape and the hump. For the U-shape, the measure is defined as follows:

$$\mathcal{M}_t^U = \left(m_t(r_l) + \frac{r_{min} - r_l}{r_r - r_l}(m_t(r_r) - m_t(r_l)) \right) - m_t(r_{min}). \tag{IV.18}$$

Here, we determine r_{min}, r_l and r_r for each day separately according to the day's pricing kernel estimate and its respective expected $\mathcal{P}$ variance. By using the daily conditional variance estimate, we neutralize any effect that may arise from varying volatility of unpriced components in equity returns. For the hump shape, our measure is

$$\mathcal{M}_t^{hump} = m_t(r_{lmax}) - m_t(r_{lmin}). \tag{IV.19}$$

Figure IV.5: Time Series Evolution of Pricing Kernel Shape Measures


(a) Magnitude of U-shape

(b) Magnitude of Hump

The plots show the time series of $\mathcal{M}_t^U$ (left plot) and $\mathcal{M}_t^{hump}$ (right plot). The blue line depicts pricing kernels that are induced by CKDE $\mathcal{P}$ density, while the orange line uses the MDN $\mathcal{P}$ density. For the sake of a smoother exposition, the plots show 22-trading-day moving averages of the respective daily measures.

When inspecting pricing kernels of single days, it occurs regularly that the local minimum of the log pricing kernel at slightly negative returns is below the local minimum at slightly positive returns. This feature of the data leads to a switch in the location of the hump, from the left of the overall pricing kernel's minimum to its right. In order to still be able to capture the hump as a deviation from the overall U-shape, we therefore need to slightly re-define r_{lmax} and r_{lmin} in these situations. We still let r_{lmax} be the return where the hump reaches its maximum. If $r_{lmax} < r_{min}$, we let $r_{lmin} < r_{lmax}$ be the return where the log pricing kernel has its minimum before continuing the overall U-shape. This is the same situation as in the unconditional case. If however $r_{lmax} > r_{min}$, the local minimum that constitutes the hump switches sides to the right of the hump's top and $r_{lmin} > r_{lmax}$ is again the return where the log pricing kernel has its minimum before continuing the U-shape for even larger returns. Figure IV.4 gives a visual impression of our two measures. Effectively, they measure the vertical magnitude of the respective shapes.

Time series charts of these magnitudes are shown in figure IV.5. In order to smooth out noise, we apply a 22-trading-day moving average to all estimates. For both measures, the time series of shape magnitudes as implied by the *CKDE* and the *MDN* methods strongly co-move. With 0.87, correlation between the two methods' shape measures is higher for the U-shape than for the hump shape, where we detect a correlation of 0.66. During our entire data sample, the U-shape never vanishes enduringly.[5] However,

---

[5] Keep in mind that M_t^U would become 0 for a monotonically declining convex pricing kernel, as $r_{min} = r_r$ in that case.

it appears to decrease in calm times and increase in and after crisis times. If the U-shape reflects investor's aversion to variance (Cuesdeanu and Jackwerth, 2018b), these movements suggest that investors become increasingly averse to variance changes during crises, but their aversion reduces again after some time passed in calm periods. For the hump shape, the situation is different. It appears to vanish in the late financial crisis and in its aftermath. During calm times, it appears to become increasingly pronounced. Generally, the hump shape appears to manifest more clearly if the U-shape reduces. The correlation between the two measures is -0.34 according to the *MDN* and -0.21 according to the *CKDE*.

We will now describe time variation in M_t^U and M_t^{hump} a bit more formally, by regressing the measures on expected moments. More precisely, we estimate the following regressions via OLS:

$$\mathcal{M}_t^{\{U,hump\}} = \alpha + \beta\,\mu_t + e_t^{\{U,hump\}}; \quad e_t^{\{U,hump\}} \sim i.i.d., E(e_t^{\{U,hump\}}) = 0. \tag{IV.20}$$

Here, μ_t is a vector of mean, variance and skewness estimates under $\mathcal{P}$ or $\mathcal{Q}$ at time t for the same (monthly) time horizon that $\mathcal{M}_t^{\{U,hump\}}$ spans. We calculate these measures from our estimates of the $\mathcal{P}$ and $\mathcal{Q}$ density via numerical integration. Prior to the regression, we normalize all variables to a mean of zero and unit standard deviation. We run multiple regressions that differ in the set of moments that we use as exogenous information. The β coefficients of our regressions can be found in table IV.7. For the sake of a slim representation, we only report the coefficient estimates for the *MDN* estimator here.

The regression results confirm our previous intuition. Expected returns have a weakly significant influence on the magnitude of the U-shape, which tends to become more pronounced if expected returns are higher. However, the effect is not very strong. At the same time, the variance risk premium, the difference in variance expectations under $\mathcal{Q}$ and under $\mathcal{P}$, is a strong determinant of the U-shape, which can be seen from the very significant and nearly equal coefficients of opposite sign in regression (4) of panel (a). A one standard deviation increase in the variance risk premium comes along with a nearly 2.5 standard deviation increase in the magnitude of the U-shape. The finding confirms our previous notion that the U-shape is primarily an expression of variance aversion. Finally, expected $\mathcal{Q}$ skewness appears to be an important driver of the U-shape, too. Here, more negative $\mathcal{Q}$ skewness reduces the U-shape. With constant $\mathcal{P}$ skewness, $\mathcal{Q}$

Table IV.7: Explaining Shape Changes in the Pricing Kernel

(a) U-shape

	$\mu^{1,\mathcal{P}}$	$\mu^{1,\mathcal{Q}}$	$\mu^{2,\mathcal{P}}$	$\mu^{2,\mathcal{Q}}$	$\mu^{3,\mathcal{P}}$	$\mu^{3,\mathcal{Q}}$	Adj. R2
(1)	0.176* (2.32)						0.031
(2)	0.17* (2.21)	-0.013 (-0.15)					0.03
(3)			-0.029 (-0.51)				0.001
(4)			-2.467*** (-8.66)	2.475*** (8.77)			0.181
(5)					-0.225*** (-3.32)		0.05
(6)					-0.147* (-2.49)	0.378*** (7.28)	0.187

(b) Hump Shape

	$\mu^{1,\mathcal{P}}$	$\mu^{1,\mathcal{Q}}$	$\mu^{2,\mathcal{P}}$	$\mu^{2,\mathcal{Q}}$	$\mu^{3,\mathcal{P}}$	$\mu^{3,\mathcal{Q}}$	Adj. R2
(1)	0.037 (0.82)						0.001
(2)	-0.019 (-0.42)	-0.136*** (-3.78)					0.016
(3)			-0.219*** (-8.18)				0.048
(4)			0.36 (1.92)	-0.588** (-2.88)			0.058
(5)					-0.031 (-0.74)		0.001
(6)					-0.091** (-2.61)	-0.288*** (-7.29)	0.08

The tables show the β coefficients of the regressions from equation IV.20. $\mu^{1,\{\mathcal{P},\mathcal{Q}\}}$, $\mu^{2,\{\mathcal{P},\mathcal{Q}\}}$ and $\mu^{3,\{\mathcal{P},\mathcal{Q}\}}$ represent the expected mean, variance and skewness under the respective measure. All variables are de-meaned and normalized to a standard deviation of 1. Newey-West adjusted t-statistics with 22 lags are given in parenthesis. '*' marks parameter estimates at the 5% significance level, '**' marks the 1% significance level and '***' marks the 0.1% significance level.

skewness becomes more negative if investors fear large downward returns more strongly. This may be an expression of rising risk aversion against losses relative to risk aversion against variance, thus flattening out the U. At the same time, an increased likelihood of negative returns, as expressed by a more negative $\mathcal{P}$ skewness, comes along with a strengthening U-shape. In any case, the low absolute size of the coefficients implies that the effect of skewness on the magnitude of the U-shape is weaker than the effect of the variance risk premium.

For the hump shape, the regression results differ in a few important aspects. Instead of the mean $\mathcal{P}$ return, the risk-free rate as first moment of the $\mathcal{Q}$ density affects the hump. Very significantly, a rise in the risk-free rate reduces the hump. The effect is not very strong, though. Much more strongly, $\mathcal{P}$ and $\mathcal{Q}$ variance expectations appear to affect the hump shape, with a higher variance expectation reducing the hump. Importantly, it is not the variance risk premium that drives this result, but the variance estimates themselves. Finally, the hump also gets larger if $\mathcal{Q}$ skewness becomes more negative. Jointly, we conclude that the hump shape strengthens in economic environments that are characterized by low interest rates, low variance and considerable risk-neutral skewness.

IV.4.3 Expected Pricing Kernel Variance

Having described time variation in pricing kernel shapes in detail in the previous subsection, we now turn to the conditional variance of the log pricing kernel. Variance is defined as the second central moment of a random variable, which can be obtained via integration:

$$\sigma^2_{m,t,\tau} = \int_{-\infty}^{\infty} \left(m_t(r_{t,\tau}) - E(m_t(r_{t,\tau}))\right)^2 p_t(r_{t,\tau})dr_{t,\tau} = \int_{-\infty}^{\infty} \nu(r_{t,\tau})dr_{t,\tau}. \tag{IV.21}$$

Since we have daily empirical estimates of the $\mathcal{P}$ density $p_t(r_{t,\tau})$ and the pricing kernel $m_t(r_{t,\tau})$ as functions of market returns for a dense grid of monthly returns $\{r_1, ..., r_N\}$, we can numerically approximate the integral in IV.21. To that end, we apply the trapezoidal rule and calculate

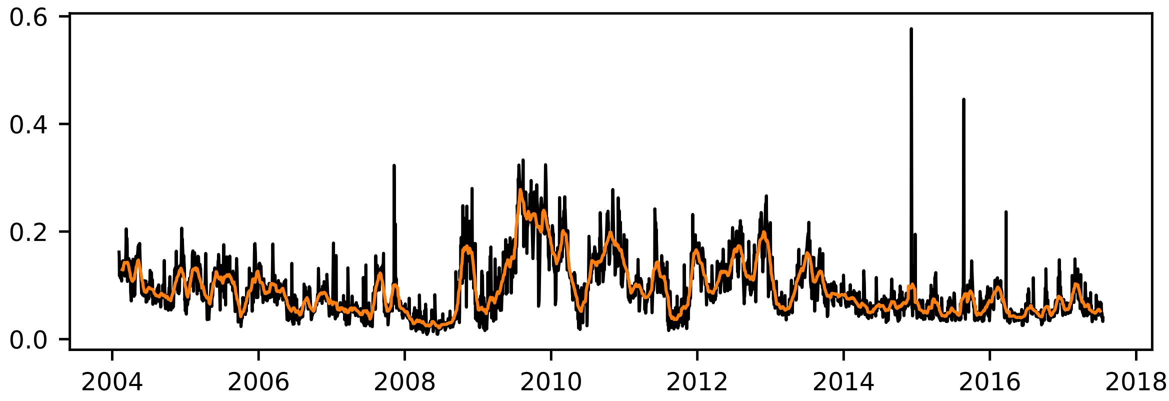
$$\hat{\sigma}^2_{m,t,\tau} = \sum_{i=2}^{N} \frac{\nu(r_i) + \nu(r_{i-1})}{2}(r_i - r_{i-1}), \tag{IV.22}$$

where the return grid reaches from $r_1 = -0.45$ to $r_N = 0.45$ in a step size of $r_i - r_{i-1} = 0.0025$.[6] Our tests showed that both truncation and discretization errors are negligible

---

[6] In parallel to the definition of log pricing kernel variance, we estimate the expected value of the log pricing kernel by numerically evaluating the integral $E(m_{t,\tau}) = \int_{-\infty}^{\infty} m_t(r_{t,\tau})\, p(r_{t,\tau})\, dr_{t,\tau}$ with the same

Figure IV.6: Expected Variance of the Log Pricing Kernel



The figure shows the daily one-month-horizon expected variance of the pricing kernel in black over time. The orange line shows a 22-trading-day moving average.

at these values. In order to mitigate the impact of outliers, we winsorize our log pricing kernel estimates at the 0.5% and 99.5% percentiles.

A time series chart of the evolution of log pricing kernel variance can be found in figure IV.6. Conditional variance was especially low during the first months of the financial crisis. This finding matches our observation from section IV.4.1 that the pricing kernel is much flatter during the crisis than during calm times. Directly following the bankruptcy of Lehman brothers on September 15, 2008, pricing kernel variance increases strongly, before dropping again to relatively low values for the final months of the financial crisis. Only after the S&P 500 had reached its minimum in March 2009, pricing kernel variance persistently increased. In general, we observe similar patterns of low variance during the event and an increase in variance after the event for the two major waves of the Greek debt crisis in April/May 2010 and June to August 2011. It took until 2014 for the variance of the log pricing kernel to stabilize again close to its pre-crisis levels.

For the purpose of further analysing the evolution of pricing kernel variance, we follow Bollerslev and Todorov (2011), who decompose return variance into a continuous and a jump component. The authors first estimate the conditional standard deviation of continuous shocks, label realized returns as jump or continuous shocks based on whether the absolute return is below some multiple of the continuous volatility and then separately calculate variance for jump and continuous returns. Effectively, this procedure splits the return distribution at predetermined percentiles. The central part of the distribution,

---

parametrization of the trapezoidal rule.

between the splitting percentiles, is considered as continuous and the tails, outside of the splitting percentiles, are considered as jumps. We follow their notion for the variance of the log pricing kernel and decompose equation IV.21 as

$$\begin{aligned}\sigma^2_{m,t,\tau} = &\int_{a_t}^{b_t} \left(m_t(r_{t,\tau}) - E(m_t(r_{t,\tau}))\right)^2 p_t(r_{t,\tau}) dr_{t,\tau} + \\ &\int_{-\infty}^{a_t} \left(m_t(r_{t,\tau}) - E(m_t(r_{t,\tau}))\right)^2 p_t(r_{t,\tau}) dr_{t,\tau} + \\ &\int_{b_t}^{\infty} \left(m_t(r_{t,\tau}) - E(m_t(r_{t,\tau}))\right)^2 p_t(r_{t,\tau}) dr_{t,\tau} \\ = &\, \sigma^2_{mc,t,\tau} + \sigma^2_{mlj,t,\tau} + \sigma^2_{mrj,t,\tau},\end{aligned} \tag{IV.23}$$

where $\sigma^2_{mc,t,\tau}$ measures the variance of the central part of the distribution, the expected continuous variation, and $\sigma^2_{mlj,t,\tau}$ ($\sigma^2_{mrj,t,\tau}$) measures the variation due to downward (upward) jumps.

In line with Bollerslev and Todorov (2011), we use the bipower variation of Barndorff-Nielsen and Shephard (2004) to estimate realized continuous return variation based on daily realized S&P 500 returns $\mathbf{r}_{t,1}$ over a monthly time interval:

$$\left(\hat{\sigma}^{CV}_{r,t,\tau}\right)^2 = \sum_{i=1}^{22} |\mathbf{r}_{t+i-1,1}|\,|\mathbf{r}_{t+i,1}|. \tag{IV.24}$$

We estimate expected continuous variation via an AR(1) model. Andersen et al. (2003) strongly recommend to use the log of variance estimates for such prediction models as the log is approximately normally distributed, while actual variance estimates exhibit strong positive skewness, which may bias regression coefficient estimates. Consequently, our prediction model for the log of continuous variance reads

$$ln\ \left(\hat{\sigma}^{CV}_{r,t+\tau,\tau}\right)^2 = \alpha + \beta\ ln\ \left(\hat{\sigma}^{CV}_{r,t,\tau}\right)^2 + \epsilon_{r,t+\tau,\tau}, \tag{IV.25}$$

which we estimate via OLS to obtain parameter estimates $\hat{\alpha}$ and $\hat{\beta}$. Let $ECV_{t,\tau} = exp\left(\hat{\alpha} + \hat{\beta}\ ln\ \left(\hat{\sigma}^{CV}_{r,t,\tau}\right)^2\right)$ stand for the expected continuous variation of returns at time t for the time horizon τ. We follow Bollerslev and Todorov (2011) and set the splitting percentiles to $a_t = -3\sqrt{ECV_{t,\tau}}$ and $b_t = 3\sqrt{ECV_{t,\tau}}$.

The idea behind the definition of a_t and b_t is that it is highly unlikely to observe continuous shocks, which follow a normal distribution, beyond the 3 standard deviation bound. Virtually all variation due to shocks outside of these bounds must therefore be attributable to jumps. We progress on that thought by assuming that a jump in the market return is accompanied by a jump in the pricing kernel. The same assumption is made by factor models that use the market return as a risk factor. While we closely follow the procedure of Bollerslev and Todorov (2011), our approach differs from theirs in one aspect. With their daily return observations, Bollerslev and Todorov (2011) only observe a single draw from each day's conditional density, thus making it necessary to assume a parametric structure for the evolution of variance and the relationship between continuous and jump variation to obtain time-varying estimates of continuous and jump variation. As we estimate the full density of returns for each day, we do not need to make this assumption, but obtain a daily measure of expected continuous and jump variation model-free, by separately evaluating the integrals in equation IV.23. As before, we use the trapezoidal rule to numerically approximate these integrals.

Unconditional and conditional sample estimates for the different components of pricing kernel variance and their relative share are given in panel (a) of table IV.8. Overall, jump variation represents about one fifth to one fourth of log pricing kernel variance. This relative share is approximately the same in the crisis and calm sample. However, differences arise for the components of jump variation. In absolute terms, negative jump variation during the financial crisis was less than half its value during non-crisis times. Also, its contribution is lower in relative terms. The reverse is true for positive jump variation, which is higher during the crisis, both in absolute and relative terms.

Figure IV.7 shows the time series decomposition of pricing kernel variance into its continuous, its negative jump and its positive jump component. The bulk of time variation in pricing kernel variance can be attributed to changes in continuous risk. In contrast, negative jump variation appears to stick relatively closely to its unconditional mean throughout our sample, with the exception of the financial crisis, where we measure lower estimates. In relative terms, the contribution of negative jump variation varies strongly though, taking values between 2.3% and 54.8% of total pricing kernel variance. However, changes in the relative contribution of negative jump variation are largely governed by the moves of continuous variation: If negative jump variation remains constant and continuous variation drops, the relative contribution of negative jump variation to the overall variance of the pricing kernel must rise. Consequently, as panel (b) of table IV.8 shows, the correlation between absolute continuous variation and relative left jump variation is at -0.7, while the correlation between the two absolute variation components

Table IV.8: Decomposition of Expected Log Pricing Kernel Variance

(a) Sample Averages

Sample		$\hat{\sigma}^2_{mlj,t,\tau}$	$\hat{\sigma}^2_{mc,t,\tau}$	$\hat{\sigma}^2_{mrj,t,\tau}$
Full	Absolute	0.0156 (0.0001)	0.0718 (0.0008)	0.0051 (0.0001)
	Relative	16.8 (0.2)	77.7 (0.9)	5.5 (0.2)
Crisis	Absolute	0.0078 (0.0006)	0.0571 (0.0026)	0.006 (0.0004)
	Relative	11.1 (0.8)	80.4 (3.6)	8.5 (0.5)
Non-Crisis	Absolute	0.0166 (0.0001)	0.0738 (0.0009)	0.0049 (0.0002)
	Relative	17.4 (0.1)	77.4 (1.0)	5.2 (0.2)

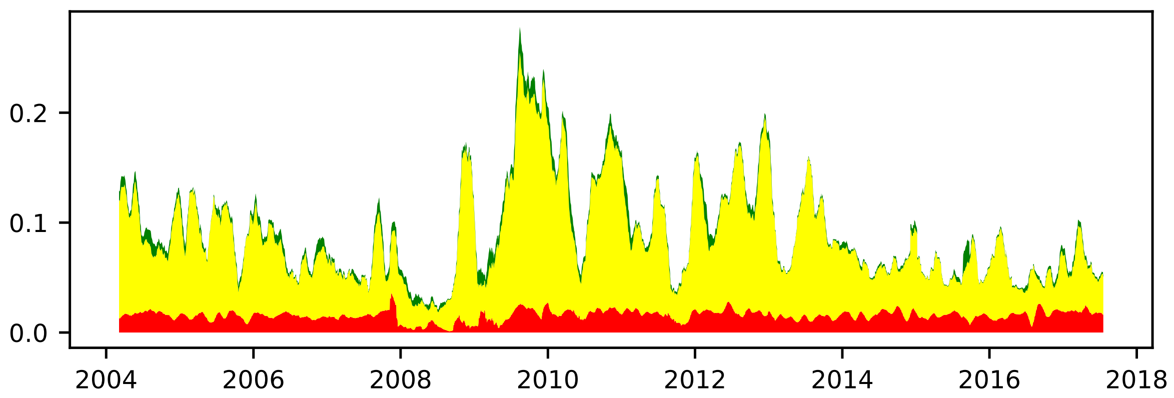
(b) Correlations

		Absolute			Relative		
		$\hat{\sigma}^2_{mc,t,\tau}$	$\hat{\sigma}^2_{mlj,t,\tau}$	$\hat{\sigma}^2_{mrj,t,\tau}$	$\hat{\sigma}^2_{mc,t,\tau}$	$\hat{\sigma}^2_{mlj,t,\tau}$	$\hat{\sigma}^2_{mrj,t,\tau}$
Absolute	$\hat{\sigma}^2_{mc,t,\tau}$	1.0	0.26	0.4	0.71	-0.7	-0.22
	$\hat{\sigma}^2_{mlj,t,\tau}$		1.0	0.31	-0.3	0.37	-0.06
	$\hat{\sigma}^2_{mrj,t,\tau}$			1.0	-0.06	-0.27	0.7
Relative	$\hat{\sigma}^2_{mc,t,\tau}$				1.0	-0.9	-0.48
	$\hat{\sigma}^2_{mlj,t,\tau}$					1.0	0.05
	$\hat{\sigma}^2_{mrj,t,\tau}$						1.0

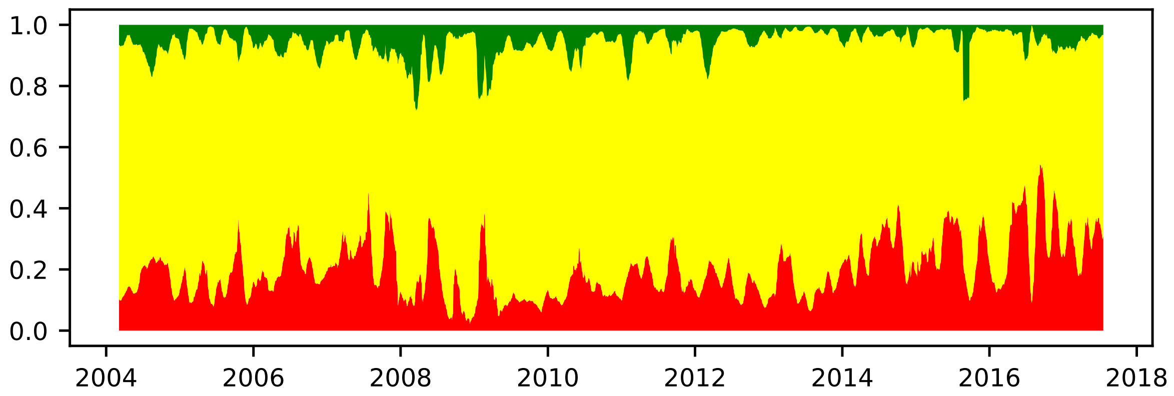
The table shows sample statistics of our daily estimates for the components of log pricing kernel variance. $\hat{\sigma}^2_{mc,t,\tau}$, $\hat{\sigma}^2_{mlj,t,\tau}$ and $\hat{\sigma}^2_{mrj,t,\tau}$ are the expected variation due to continuous shocks, negative jumps and positive jumps. Our crisis sample covers the time period between 12/2007 and 06/2009, the non-crisis sample consists of the remaining data. *Absolute* values show the estimates themselves, *relative* figures show the estimates divided by the conditional total log pricing kernel variance. Relative figures are given in percent. Panel (a) shows the sample averages of our estimates along with standard errors in parenthesis, panel (b) their time series correlations.

Figure IV.7: Decomposition of Expected Pricing Kernel Variance

(a) Absolute Contribution to Overall Variance Expectation



(b) Relative Contribution to Overall Variance Expectation



The plots show how log pricing kernel variance expectations decompose into their continuous and jump components over time. The continuous component is depicted in yellow, the negative jump component in red and the positive jump component in green. All time series use a 22-trading-day moving average to smooth out noise.

is at 0.25. Right jump variation plays a less important role, with a contribution to overall variance between 0.2% and 28.3%. The correlation between the two jump components is 0.31, indicating that they tend to rise and fall simultaneously. Together, the two jump components contribute between 4.9% and 60% to the expected variance of the log pricing kernel. In summary, our results suggest that jumps add a non-negligible portion to the variance of the pricing kernel overall and at times even contribute more than the continuous part.

Table IV.9: Sample Statistics of Economic Shocks

			Percentiles				
Sample	Skewness	Kurtosis	5%	25%	50%	75%	95%
Full	-1.48	7.05	-1.47	-0.47	0.23	0.79	1.15
Crisis	-0.04	5.73	-1.57	-0.33	0.31	0.78	1.42
Non-Crisis	-1.82	7.11	-1.45	-0.5	0.21	0.79	1.12

Sample statistics for the time series of economic shocks, $\hat{\varepsilon}_{t,\tau}$. The shocks refer to the one-month horizon and are standardized to zero mean and unit variance in expectation. The crisis sample spans the NBER recession period between 12/2007 and 06/2009, the non-crisis sample contains the remaining estimates between 01/2004 and 07/2017.

IV.5 Shocks to the Economy

The pricing kernel is a random variable that is determined by shocks to fundamental economic state variables (Cochrane, 2001, ch. 1). We can therefore express the log pricing kernel as

$$m_{t,\tau} = \delta_{t,\tau} - \gamma\, \varepsilon_{t,\tau}, \tag{IV.26}$$

where $\delta_{t,\tau}$ captures the predictable component and $\varepsilon_{t,\tau}$ is a random variable with $E_t(\varepsilon_{t,\tau}) = 0$, which aggregates the shocks to all economic state variables. We call ε_t the *economic shock*. As described in section IV.3.1, given our measurement of the pricing kernel as a function of S&P 500 returns, $m_t(r_{t,\tau})$, we can obtain a time series of realized log pricing kernel values by evaluating $m_t(r_{t,\tau})$ at the observed log return $\mathbf{r}_{t,\tau}$: $\mathbf{m}_{t,\tau} = m_t(\mathbf{r}_{t,\tau})$. Combining this realization with our measurement for the expected value and variance of the log pricing kernel from section IV.4.3, we obtain standardized estimates for the realized economic shock as

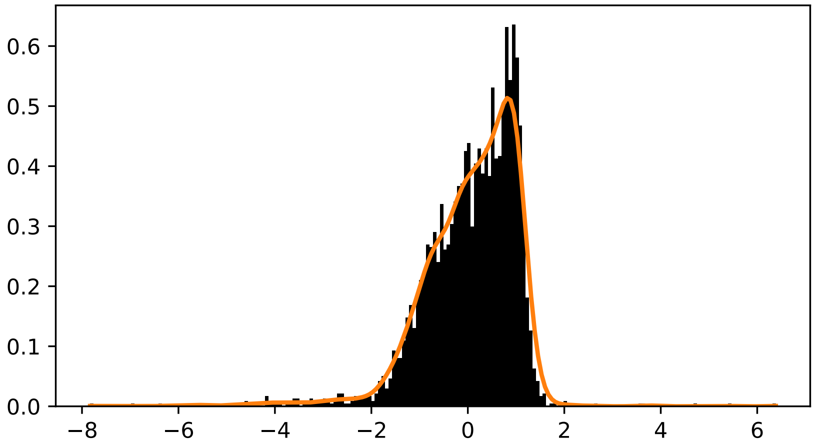
$$\hat{\varepsilon}_{t,\tau} = -\frac{\mathbf{m}_{t,\tau} - E_t(m_{t,\tau})}{\hat{\sigma}_{m,t,\tau}}. \tag{IV.27}$$

The remaining subsections of this chapter are devoted to the analysis of the time series of economic shocks.

IV.5.1 Distribution and Time Series of Economic Shocks

The empirical distribution of economic shocks is depicted in figure IV.8. Table IV.9 provides summarizing sample statistics. By design, the time series of economic shocks

Figure IV.8: Distribution of Economic Shocks



This chart depicts a histogram of realized economic shocks. The orange line shows a kernel density estimate for the distribution of economic shocks. The shocks are normalized to have $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = 1$.

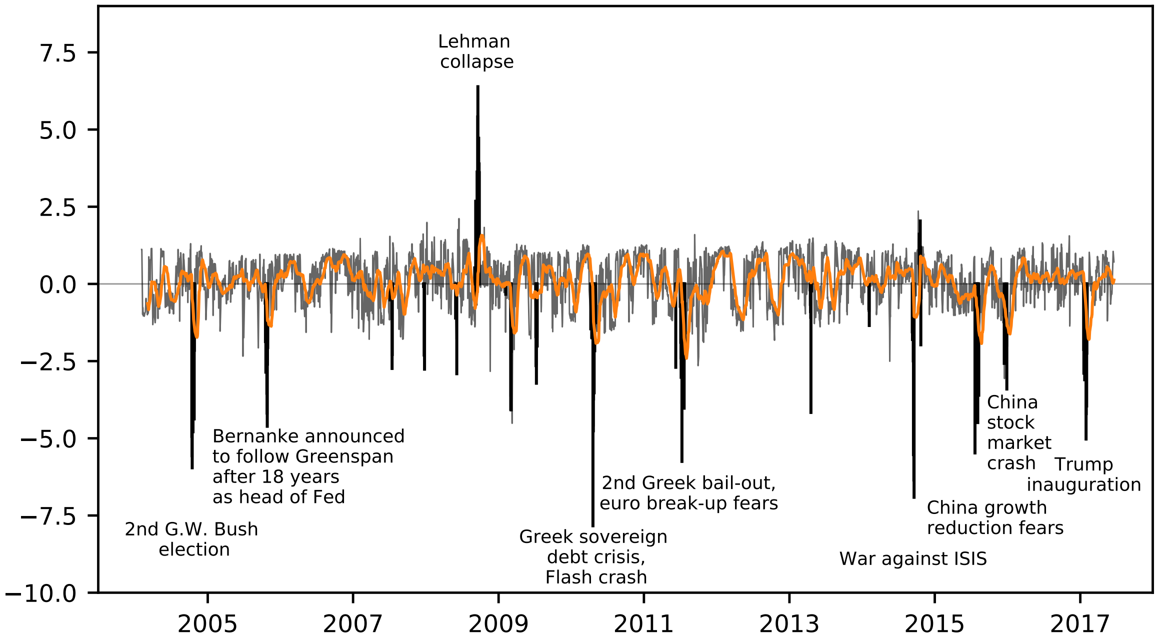
has in expectation zero mean and unit volatility.[7] Clearly, economic shocks are left skewed and heavy tailed. In our sample, 59% of the economic shocks realized higher than expected ($\hat{\varepsilon}_{t,\tau} > 0$), however, a positive economic shock has an average size of 0.69 and is thus smaller in absolute terms than the average negative economic shock of -0.82. The distribution of economic shocks appears less skewed during the crisis sample than during the calm sample, which is puzzling as we would expect more large negative events to have realized during the financial crisis than large positive events. We will return to this issue in detail shortly.

We first take a look at the time series chart of economic shocks in figure IV.9. In the chart, daily economic shock estimates are plotted in grey and the filled black spikes mark tail events, which we identify based on the realized S&P 500 return. More precisely, if the ex ante probability of a realized return was below 2.5% or above 97.5% according to the MDN conditional $\mathcal{P}$ density estimator, we classify the return and the corresponding pricing kernel shock as a tail event. Generally, we observe that a tail event in the S&P 500 also comes along with a spike in the economic shock. The spikes can be traced very well to historic events, figure IV.9 contains labels for some of them. Interestingly, while about 43% of the tail events are generated by large positive S&P 500 returns, the pricing kernel nearly always spikes downwards. This is a direct result of the U-shape in the pricing

---

[7] To be precise, since we use the conditionally expected log pricing kernel mean and variance in the estimation, instead of their realized counterparts, mean and variance of our economic shock time series differ slightly from these expected values. Our empirical mean of economic shocks is 0.064 and volatility is 0.983.

Figure IV.9: Time Series of Economic Shocks



The figure shows our estimates of daily realized one-month-horizon shocks to the economy in gray. The black areas are realized tail events, with a 2-sided ex ante probability of less than 5%, i.e., less than 2.5% for either side of the return distribution. The orange line shows a 22-trading-day moving average.

kernel: Large positive returns, just like large negative returns, lead to high pricing kernel realizations, which translate into negative economic shocks.

The upward spike following the bankruptcy of Lehman Brothers on September 15, 2008, clearly stands out. While our estimators anticipate a large economic shock following that event, its positive sign is highly unrealistic. At the same time, this high positive shock is the reason for our close-to-zero skewness measurement of economic shocks during the crisis sample. If we exclude September and October 2008, the two months around the Lehman collapse, from our crisis sample, the skewness of economic shocks in the remaining crisis sample drops to -1.47, suggesting that negative economic shocks dominated at that time. Due to the unrealistic measurement of the Lehman shock, we are cautious about our estimates of economic shocks around that event.

Especially considering these concerns, it remains to show whether our recovered economic shocks are actually connected to shocks of known risk factors from the literature. We continue to do so in the next subsection.

IV.5.2 Drivers of Economic Shocks

In order to identify links between our economic shock estimates and established sources of risk, we run linear regressions of the type

$$\hat{\varepsilon}_{t,\tau} = \alpha + \beta\, X_{t,\tau} + \eta_{t,\tau}. \tag{IV.28}$$

Here, $X_{t,\tau}$ is a vector of realized risk variables, that are measured at the same time and over the same horizon as $\hat{\varepsilon}_{t,\tau}$. $\eta_{t,\tau}$ is an error term with $E(\eta_{t,\tau}) = 0$. Prior to estimation, we normalize all variables to zero mean and unit variance and estimate the regressions via OLS. Throughout our analysis, we will present all results twice: The first estimation excludes the two months of September and October 2008 from our data sample due to our unrealistic measurement of the economic shock around the Lehman collapse. A similar approach is followed by Joslin et al. (2014), who even exclude the entire financial crisis sample in their model estimation due to concerns about structural breaks. The two months stand for 1.3% of all observations in our full data set. For completeness, our second set of estimation results uses our full data sample. If not specified differently, our comments below relate to the first estimation set-up.

We start our analysis with the market return and variance. Cuesdeanu and Jackwerth (2018b) emphasize that a U-shape in the pricing kernel can be rationalized by investors that are averse against both market and variance shocks. In line with this notion, some recent studies model the pricing kernel as a function of the market return and its variance (Christoffersen et al., 2013; Schneider et al., 2019). Such a set-up implies that both variables should be closely connected to economic shocks. For robustness, we do not use the S&P 500 returns, that we used as target variable in the estimation of our $\mathcal{P}$ density estimator, to proxy for the shocks to the overall market, but use the market risk factor from Kenneth French's website instead. We compute daily realizations of one-month log returns of the market factor and square them to obtain measures for realized variance.

Table IV.10 shows the β coefficients for our first set of regressions. Market returns are clearly connected to economic shocks. A one standard deviation decrease in market returns leads to a 0.35 standard deviation decrease in the economic shock. For variance, the relationship is even stronger. Here, a one standard deviation positive variance shock relates to a 0.62 standard deviation downward economic shock. Both effects appear to be independent of each other as the coefficients are estimated at nearly the same values

Table IV.10: Regression Results of Economic Shocks on Market Returns and Variance

(a) Excluding Lehman Shock

	(1)	(2)	(3)
$r_{t,\tau}^{market}$	0.35***		0.33***
	(3.27)		(3.84)
$(r_{t,\tau}^{market})^2$		-0.62***	-0.59***
		(-4.6)	(-3.24)
Adj. R^2	0.099	0.112	0.2

(b) Full Sample

	(1)	(2)	(3)
$r_{t,\tau}^{market}$	0.2		0.22
	(1.65)		(1.82)
$(r_{t,\tau}^{market})^2$		-0.05	0.05
		(-0.35)	(0.26)
Adj. R^2	0.04	0.002	0.042

Each column shows the beta estimates of an explaining regression for economic shocks. $r_{t,\tau}^{market}$ is the log Fama and French (1993) market return and $(r_{t,\tau}^{market})^2$ is the squared return as measurement for realized variance. The returns are measured over the same period that $\varepsilon_{t,\tau}$ covers and normalized to zero mean and unit variance. Newey-West adjusted t-statistics with 22 lags are given in parenthesis. '*' marks parameter estimates at the 5% significance level, '**' marks the 1% significance level and '***' marks the 0.1% significance level. Panel (a) excludes September and October 2008 from the estimation, panel (b) uses our full data sample.

in a joint regression. In all of that, all β coefficient are highly significant. Including the Lehman shock into the estimation changes the picture dramatically, as panel (b) of table IV.10 reveals. Neither market returns nor variance are estimated to have a significant connection with economic shocks, the β coefficient for variance is even close to 0. The effect is especially remarkable given the small amount of data points that constitute the Lehman shock.

We proceed with the regression of economic shocks on the Fama and French (1993) market, size and value and the Carhart (1997) momentum risk factor log returns. β estimates of these regressions are given in table IV.11. While economic shocks only seem to be significantly connected to the size factor in the individual regressions, this relationship breaks down after controlling for the market return. The size and value factor are cross-sectional risk factors. Our measurement of economic shocks is based on the projection of the pricing kernel on market returns, though. The projection marginalizes over cross-sectional elements in the pricing kernel, thus yielding a measure for economic shocks with respect to the overall market return dimension. These shocks should in fact be unrelated to cross-sectional risk factors, which are orthogonal to the market factor. Our result confirm this notion with low and insignificant β estimates when we control for market returns. However, our results do not imply that shocks to the pricing kernel are unrelated to shocks to the size or value factor, as the pricing kernel may well have cross-sectional components, which we just cannot capture in our set-up.

Table IV.11: Regression Results of Economic Shocks on Risk Factors

(a) Excluding Lehman Shock

	(1)	(2)	(3)	(4)	(5)
$r_{t,\tau}^{market}$				0.36***	0.41***
				(3.57)	(4.31)
$r_{t,\tau}^{SMB}$	0.17***			0.05	0.04
	(3.54)			(1.17)	(0.98)
$r_{t,\tau}^{HML}$		0.03		-0.08	-0.02
		(0.52)		(-1.41)	(-0.35)
$r_{t,\tau}^{MOM}$			0.05		0.17**
			(0.72)		(2.46)
Adj. R^2	0.028	0.001	0.002	0.107	0.128

(b) Full Sample

	(1)	(2)	(3)	(4)	(5)
$r_{t,\tau}^{market}$				0.2	0.25*
				(1.6)	(2.01)
$r_{t,\tau}^{SMB}$	0.13*			0.05	0.04
	(2.3)			(1.17)	(1.01)
$r_{t,\tau}^{HML}$		0.01		-0.06	0.01
		(0.19)		(-1.13)	(0.18)
$r_{t,\tau}^{MOM}$			0.07		0.18**
			(1.1)		(2.57)
Adj. R^2	0.016	0.000	0.004	0.045	0.067

Each column shows the beta estimates of an explaining regression for economic shocks. $r_{t,\tau}^{market}$, $r_{t,\tau}^{SMB}$ and $r_{t,\tau}^{HML}$ are the log Fama and French (1993) market, size and value return. $r_{t,\tau}^{MOM}$ is the Carhart (1997) log momentum return. The returns are measured over the same period that $\varepsilon_{t,\tau}$ covers and normalized to zero mean and unit variance. Newey-West adjusted t-statistics with 22 lags are given in parenthesis. '*' marks parameter estimates at the 5% significance level, '**' marks the 1% significance level and '***' marks the 0.1% significance level. Panel (a) excludes September and October 2008 from the estimation, panel (b) uses our full data sample.

Our last regression exercise relates our estimates of economic shocks to several more fundamental risk measures. Four of these measures relate to bond markets, namely the 3-month US treasury yield, the term spread, the credit spread and the TED spread. Further, we use the sentiment index of Baker and Wurgler (2006), the text-based policy uncertainty index of Baker et al. (2016) and the illiquidity measure of Amihud (2002). For all of these measures, we compute the change over the time period that $\varepsilon_{t,\tau}$ covers, i.e., $\Delta x_t = x_{t+\tau} - x_t$, and regress $\varepsilon_{t,\tau}$ on these changes. The β coefficients of the regressions are given in table IV.12.

At the 5% significant level, a downward economic shock is linked to a decreasing term spread, an increasing credit spread, an increase in the Baker et al. (2016) policy uncertainty index and higher Amihud (2002) illiquidity. These relationships generally match intuition. However, the β estimates of the four measures differ in their size and significance. Illiquidity appears to have the strongest impact on the economic shock in our regressions, with a one standard deviation increase coming along with a 0.27 standard deviation lower economic shock. The connection between the credit spread and our economic shocks appears to be similarly strong, however diminishes and turns insignificant after controlling for illiquidity. The term spread has a somewhat smaller impact on the economic shock, but remains significant with a nearly equal coefficient in a joint regression with illiquidity. Together, the two variables can explain 8% of the variation in our eco-

Table IV.12: Regression Results of Economic Shocks on Economic State Variable Shocks

(a) Excluding Lehman Shock

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\Delta r_{f,t}^{3m}$	-0.03									-0.12**
	(-0.84)									(-2.78)
$\Delta Term_t$		0.17**						0.13**	0.15**	
		(2.92)						(2.76)	(2.94)	
$\Delta Credit_t$			-0.22***					-0.08		-0.27***
			(-3.54)					(-1.59)		(-3.88)
ΔTED_t				-0.05						
				(-0.95)						
$\Delta Sent_{BW,t}$					0.03					
					(0.81)					
ΔEPU_t						-0.06*		-0.04		
						(-2.22)		(-1.87)		
$\Delta ILLIQ_t$							-0.27***	-0.22***	-0.25***	
							(-4.45)	(-4.34)	(-4.77)	
Adj. R^2	0.001	0.029	0.036	0.001	0.001	0.003	0.058	0.085	0.08	0.046

(b) Full Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\Delta r_{f,t}^{3m}$	-0.03									-0.07
	(-0.9)									(-1.39)
$\Delta Term_t$		0.16**						0.16***	0.16***	
		(2.95)						(3.38)	(3.24)	
$\Delta Credit_t$			-0.08					0.03		-0.11
			(-0.94)					(0.37)		(-1.06)
ΔTED_t				-0.02						
				(-0.52)						
$\Delta Sent_{BW,t}$					0.01					
					(0.2)					
ΔEPU_t						-0.06*		-0.06*		
						(-2.32)		(-2.25)		
$\Delta ILLIQ_t$							-0.15	-0.16***	-0.15	
							(-1.76)	(-3.13)	(-1.95)	
Adj. R^2	0.001	0.026	0.007	0.000	0.000	0.003	0.022	0.052	0.048	0.011

Each column shows the beta estimates of an explaining regression for economic shocks. We use the changes in the 3-month treasury yield ($\Delta r_{f,t}^{3m}$), the term spread (10-year US bond yield minus 3-month treasury yield, $\Delta Term_t$), the credit spread (Moody's Baa yield minus 10-year US bond yield, $\Delta Credit_t$), the TED spread (3-month Libor minus 3-month treasury yield, ΔTED_t), the Sentiment index of Baker and Wurgler (2006) ($\Delta Sent_{BW,t}$), the economic policy uncertainty index of Baker et al. (2016) (ΔEPU_t) and the illiquidity measure of Amihud (2002) ($\Delta ILLIQ_t$). All changes are measured over the same period the ϵ_t covers and normalized to zero mean and unit variance. Newey-West adjusted t-statistics with 22 lags are given in parenthesis. '*' marks parameter estimates at the 5% significance level, '**' marks the 1% significance level and '***' marks the 0.1% significance level. Panel (a) excludes September and October 2008 from the estimation, panel (b) uses our full data sample.

nomic shock measures. The economic policy uncertainty index is weakly significant, with low β estimates and low explanation power for economic shocks. It turns insignificant when controlling for illiquidity, too. Finally, a joint regression of the 3-month treasury yield and the credit spread can also explain 4.6% of the variation in economic shocks, indicating that economic shocks are linked to corporate refinancing rates.

Overall, our regressions reveal that the realized economic shock is connected to a range of market-related and economic state variables during nearly our entire data sample. The β estimates in our regressions appear intuitive and economically relevant. We find clear evidence that shocks to the pricing kernel are linked to both market returns and variance, speaking in favor of at least a two-dimensional structure of the pricing kernel.

IV.6 Summary

In this chapter, we constructed model-free daily estimates of the pricing kernel. Our estimates are forward-looking and non-parametric, thus requiring only minimal exogenous assumptions about the shape of the pricing kernel, investor preferences or the dynamics of these characteristics.

Despite their high flexibility, our estimates for the realized pricing kernel appear to robustly discount equity market returns and are closely connected to economic state variables. Jointly, market returns and variance can explain 20% of the variation in shocks to the pricing kernel, with market variance playing an even more important role than returns. These results align with our finding of a clear U-shape in the pricing kernel, which persists throughout our entire data sample. We also identify a hump in the pricing kernel around the zero return region, which strongly reduced or even vanished during the financial crisis. Further, we find clear signs for time variation in the expected variance of the pricing kernel. Jumps appear to account for 22% of expected pricing kernel variance on average, but their relative contribution sometimes reaches up to 60% of total expected variance, indicating that they represent an important source of risk in the pricing kernel.

Our findings are in line with the notion of a multi-dimensional pricing kernel. Such a model could explain the relationship of pricing kernel shocks to returns, variance and further economic state variables and at the same time produce non-monotonicities in the projection of the pricing kernel on returns. However, reasons for the changes in the conditional variance of the pricing kernel and the mechanics behind the time variation in the shape of the pricing kernel remain to be explored. Here, our method for recovering the forward-looking pricing kernel in a model-free fashion represents a tool for future research for the precise measurement of the actual object of interest.

V Conclusion and Outlook

This thesis develops and examines tools for the model-free, precise and robust estimation of the risk-neutral and physical distribution of equity returns and the pricing kernel.

In a first step, chapter II compares a range of popular state-of-the-art estimators for the implied volatility surface, and thus the risk-neutral return distribution, with two newly proposed techniques. In a thorough statistical assessment, we inspect how accurately the volatility surfaces of Figlewski (2008), OptionMetrics (2016) and Beber and Brandt (2006) represent option market information and find that they perform considerably worse than a simple one-dimensional kernel regression for the end-of-day set-up. For high-frequency applications, our proposed Bayesian volatility surface is more than 6 times more accurate than the second-best alternative, thanks to its ability to infer information about changes in the entire volatility smile from single trade observations. Our finding of severe differences in volatility surfaces is especially troublesome as we document considerable sensitivity in option-implied measures like Bakshi et al. (2003) implied volatility or skewness with respect to the employed volatility surface. These measures are popular inputs to financial analyses in the literature and biases in them may directly affect the implications of these analyses, such that the choice of the construction method for the implied volatility surface is of high practical importance. Our proposed methods render themselves as natural candidates for this choice, as they reflect option market information most accurately.

We turn to the estimation of the physical density of equity returns in chapter III. Here, we propose a novel forward-looking predictor of the entire return distribution, that is based on a neural network architecture. Our predictor bases its forecasts on risk-neutral moment estimates and is model-free in the sense that it can approximate any continuous conditional return density solely by learning from presented samples. We benchmark our estimator against a battery of forward-looking and backward-looking alternatives and find a higher conditional log likelihood and lower prediction errors for realized moments of equity returns. In contrast to the alternatives in our study and forward-looking estimators in the literature (Jensen et al., 2019; Jackwerth and Menner, 2018), our physical return density predictions cannot be rejected in statistical tests. In a sensitivity analysis, we identify that our estimator builds on significant nonlinear relationships between risk-neutral and physical moments. Exploiting these nonlinearities appears to be an important ingredient for the robust translation of risk-neutral to physical expectations.

In chapter IV, we analyze the empirical properties of the pricing kernel, which we obtain model-free as the ratio of our risk-neutral and physical density estimates. Unconditionally, we identify a strong overall U-shape in the log pricing kernel and a hump around the zero return region. The magnitude of these shapes varies and while we can link variation in the U-shape to changes in the variance risk premium, the markedness of the hump shape appears to co-move with risk-neutral moments. These hints for time variation in the pricing kernel are supported by fluctuations in conditional expected pricing kernel variance. We further decompose this variance and find that a significant and varying portion of the overall risk in the pricing kernel is due to jumps. Obtaining a time series of realized pricing kernels, we show that they are connected to economic state variables like the term spread, the credit spread or illiquidity. A joint model of market returns and variance explains 20% of the variation in shocks to the pricing kernel, thus indicating that the true pricing kernel is multi-dimensional. Besides their economic implications, our results indicate on a methodological level that our approach is capable of measuring the conditional pricing kernel in a model-free fashion.

While the methods that we describe in this thesis perform well in our statistical assessments, they can still be improved and extended. Describing multiple volatility surfaces with joint latent factors and modeling the evolution of these factors in a filtering framework could be a route to a significant extension of the Bayesian intraday volatility surface towards a stable representation of the volatility surfaces of single stocks. At the same time, such an approach could uncover hidden drivers of time variation in cross-sectional volatility surfaces. Open research questions also remain for our forward-looking physical density estimator. Currently, it is only possible to create a robust estimator for relatively short time horizons up to one month. For higher time horizons, the lack of a sufficient amount of independent training samples increases the probability of over-fitting. It might therefore proof beneficial to explore additional ways of regularization for the neural network estimator, such that it becomes possible to obtain a robust estimator based on only few observations.[1] Approaches for combining the advantages of neural networks and Bayesian modeling (Mullachery et al., 2018) or shrinkage techniques towards an unconditional estimate of the return distribution could lead a path to such advancements.

Our tools and insights can help to answer interesting research questions, that go beyond the scope of this thesis. While we sticked to the equity market as our object of study in this thesis, it is straight-forward to apply our risk-neutral and physical density estimators to other asset classes with only minimal adjustments. For example, it would be interesting to see whether the relationships between expected risk-neutral and physical moments are different in the bond market from what we observed for the equity market. Exploring the dependencies between forward-looking physical densities at different maturities provides a path to the model-free estimation of term structures for the variance and higher moments of interest rates. Further, estimating the pricing kernel with respect to bond returns in addition to our estimates for equity returns might be a promising approach. The two pricing kernel estimates would be projections of the same entity on the returns of different asset classes, thus giving the researcher the opportunity to use different lenses to analyze the same object of interest.

---

[1] Using only a small amount of training data likely leads to a less powerful estimator. Still, such an estimator would be desirable as long as it does not over-fit and improves upon the unconditional physical return density.

Bibliography

Ambrogioni, L., U. Güçlü, M. A. J. van Gerven, and E. Maris. 2017. The Kernel Mixture Network: A Nonparametric Method for Conditional Density Estimation of Continuous Random Variables. *Working Paper*. URL `http://arxiv.org/abs/1705.07111`.

Amihud, Y. 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5:31–56.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys. 2003. Modeling and Forecasting Realized Volatility. *Econometrica*, 71:579–625.

Ang, A. 2014. *Asset Management: A systematic approach to factor investing.* Oxford University Press.

Atkinson, K. E. 1989. *An Introduction to Numerical Analysis.* 2nd ed. John Wiley & Sons.

Audrino, F., R. Huitema, and M. Ludwig. 2015. An Empirical Analysis of the Ross Recovery Theorem. *Working Paper*. URL `https://ssrn.com/abstract=2433170`.

Aït-Sahalia, Y., and A. Lo. 1998. Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices. *The 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.

Backus, D., S. Foresi, and L. Wu. 2004. Accounting for Biases in Black-Scholes. *Working Paper*. Available at SSRN: https://ssrn.com/abstract=585623.

Bahaludin, H., and M. H. Abdullah. 2017. Empirical performance of interpolation techniques in risk-neutral density (RND) estimation. *Journal of Physics: Conference Series*, 819:012026.

Baker, M., and J. Wurgler. 2006. Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, 61:1645–1680.

Baker, S. R., N. Bloom, and S. J. Davis. 2016. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131:1593–1636.

Bakshi, G., F. Chabi-Yo, and X. Gao. 2018. A Recovery that We Can Trust? Deducing and Testing the Restrictions of the Recovery Theorem. *The Review of Financial Studies*, 31:532–555.

Bakshi, G., and N. Kapadia. 2003. Delta-Hedged Gains and the Negative Market Volatility Risk Premium. *The Review of Financial Studies*, 16:527–566.

Bakshi, G., N. Kapadia, and D. Madan. 2003. Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. *The Review of Financial Studies*, 16:101–143.

Bansal, R., and A. Yaron. 2004. Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *The Journal of Finance*, 59:1481–1509.

Barndorff-Nielsen, O. E., and N. Shephard. 2004. Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*, 2:1–37.

Barone-Adesi, G., and H. Dall'O. 2010. Is the Price Kernel Monotone? *Working Paper*. URL `https://ssrn.com/abstract=1539363`.

Barone-Adesi, G., H. Dall'o, and V. Vovchak. 2012. Is the price kernel monotone? *ACRN Journal of Finance and Risk Perspectives*, 1:43–68.

Barone-Adesi, G., R. F. Engle, and L. Mancini. 2008. A GARCH Option Pricing Model with Filtered Historical Simulation. *The Review of Financial Studies*, 21:1223–1258.

Barone-Adesi, G., and R. E. Whaley. 1987. Efficient analytic approximation of American option values. *The Journal of Finance*, 42:301–320.

Beare, B. K., and L. D. W. Schmidt. 2016. An Empirical Test of Pricing Kernel Monotonicity. *Journal of Applied Econometrics*, 31:338–356.

Beber, A., and M. W. Brandt. 2006. The effect of macroeconomic news on beliefs and preferences: Evidence from the options market. *Journal of Monetary Economics*, 53:1997–2039.

Bekaert, G., and M. Hoerova. 2014. The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183:181–192.

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

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.

Bishop, C. M. 1994. Mixture Density Networks. Technical Report. Aston University, Birmingham.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning.* New York, NY: Springer.

Black, F., and M. Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81:637–654.

Bliss, R. R., and N. Panigirtzoglou. 2002. Testing the stability of implied probability density functions. *Journal of Banking & Finance*, 26:381–422.

Bliss, R. R., and N. Panigirtzoglou. 2004. Option-Implied Risk Aversion Estimates. *The Journal of Finance*, 59:407–446.

Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Bollerslev, T., and V. Todorov. 2011. Tails, Fears, and Risk Premia. *The Journal of Finance*, 66:2165–2211.

Borovička, J., L. P. Hansen, and J. A. Scheinkman. 2016. Misspecified Recovery. *The Journal of Finance*, 71:2493–2544.

Breeden, D. T., and R. H. Litzenberger. 1978. Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51:621–651.

Busch, T., B. J. Christensen, and M. Ø. Nielsen. 2011. The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, 160:48–57.

Buss, A., and G. Vilkov. 2012. Measuring Equity Risk with Option-implied Correlations. *The Review of Financial Studies*, 25:3113–3140.

Carhart, M. M. 1997. On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52:57–82.

Carr, P., and J. Yu. 2012. Risk, Return, and Ross Recovery. *The Journal of Derivatives*, 20:38–59.

CBOE. 2019. CBOE Volatility Index. *White Paper.* URL: https://www.cboe.com/micro/vix/vixwhite.pdf.

Chabi-Yo, F., R. Garcia, and E. Renault. 2008. State Dependence Can Explain the Risk Aversion Puzzle. *The Review of Financial Studies*, 21:973–1011.

Chakravarty, S., H. Gulen, and S. Mayhew. 2004. Informed Trading in Stock and Option Markets. *The Journal of Finance*, 59:1235–1257.

Chang, B.-Y., P. Christoffersen, K. Jacobs, and G. Vainberg. 2012. Option-Implied Measures of Equity Risk. *Review of Finance*, 16:385–428.

Chaudhuri, R., and M. Schroder. 2015. Monotonicity of the Stochastic Discount Factor and Expected Option Returns. *The Review of Financial Studies*, 28:1462–1505.

Chernov, M. 2007. On the Role of Risk Premia in Volatility Forecasting. *Journal of Business & Economic Statistics*, 25:411–426.

Christoffersen, P., B. Feunou, Y. Jeon, and C. Ornthanalai. 2017. Time-Varying Crash Risk: The Role of Stock Market Liquidity. *Rotman School of Management Working Paper.* No. 2797308. Available at SSRN: https://ssrn.com/abstract=2797308.

Christoffersen, P., S. Heston, and K. Jacobs. 2013. Capturing Option Anomalies With a Variance-Dependent Pricing Kernel. *The Review of Financial Studies*, 26:1963–2006.

Cochrane, J. H. 2001. *Asset Pricing.* Princeton University Press.

Colander, D., M. Goldberg, A. Haas, K. Juselius, A. Kirman, T. Lux, and B. Sloth. 2009. The financial crisis and the systemic failure of the economics profession. *Critical Review*, 21:249–267.

Corsi, F. 2009. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, 7:174–196.

Cox, J. C., S. A. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics*, 7:229–263.

Cremers, M., A. Fodor, D. Muravyev, and D. Weinbaum. 2019. How Do Informed Option Traders Trade? Option Trading Activity, News Releases, and Stock Return Predictability. In *11th Annual Mid-Atlantic Research Conference in Finance (MARC).*

Cuesdeanu, H. 2016. Empirical Pricing Kernels: A Tale of Two Tails and Volatility? *Working Paper.* URL `https://ssrn.com/abstract=2870372`.

Cuesdeanu, H., and J. Jackwerth. 2018a. The Pricing Kernel Puzzle in Forward Looking Data. *Review of Derivatives Research*, 21:253—-276.

Cuesdeanu, H., and J. C. Jackwerth. 2018b. The pricing kernel puzzle: survey and outlook. *Annals of Finance*, 14:289–329.

Damghani, B. M. 2015. Introducing the Implied Volatility Surface Parametrization (IVP): Application to the FX Market. *Wilmott*, 2015:68–81.

Damghani, B. M., and A. Kos. 2013. De-arbitraging With a Weak Smile: Application to Skew Risk. *Wilmott*, 2013:40–49.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and their Application.* Cambridge Series on Statistical and Probabilistic Mathematics ; [1]. Cambridge [u.a.]: Cambridge University Press.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, vol. 7(1), pp. 1–68.

Dillschneider, Y., and R. Maurer. 2018. Functional Ross Recovery: Theoretical Results and Empirical Tests. *Working Paper.* URL `https://ssrn.com/abstract=2991984`.

Drechsler, I., and A. Yaron. 2011. What's Vol Got to Do with It. *The Review of Financial Studies*, 24:1–45.

Driessen, J., P. J. Maenhout, and G. Vilkov. 2009. The Price of Correlation Risk: Evidence from Equity Options. *The Journal of Finance*, 64:1377–1406.

Du, J., and N. Kapadia. 2012. Tail and Volatility Indices from Option Prices. *Working Paper.* Available here: https://people.umass.edu/nkapadia/docs/Du_Kapadia_August2012.pdf.

Duffie, D. 2001. *Dynamic Asset Pricing Theory. Third Edition.* Princeton University Press, Princeton and Oxford.

Dunis, C. L., J. Laws, and G. Sermpinis. 2011. Higher order and recurrent neural architectures for trading the EUR/USD exchange rate. *Quantitative Finance*, 11:615–629.

Engle, R., S. Giglio, H. Lee, B. Kelly, and J. Stroebel. 2019. Hedging Climate Change News. *Working Paper*. URL `http://pages.stern.nyu.edu/~jstroebe/PDF/EGKLS_ClimateRisk.pdf`.

Engle, R. F., and A. J. Patton. 2007. What Good is a Volatility Model? In *Forecasting Volatility in the Financial Markets*, pp. 47–63. Elsevier.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Fengler, M. R. 2009. Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, 9:417–428.

Figlewski, S. 2008. Estimating the Implied Risk Neutral Density. In M. W. Tim Bollerslev, Jeffrey R. Russell (ed.), *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*. Oxford, UK: Oxford University Press.

Gatheral, J. 2004. A parsimonious arbitrage-free implied volatility parameterization with application to the valuation of volatility derivatives.

Gatheral, J. 2006. *The Volatility Surface: A Practitioner's Guide*. Hoboken, New Jersey: John Wiley & Sons.

Geisser, S. 1993. *Predictive Inference: An Introduction*. Monographs on Statistics and Applied Probability ; 55, 1st ed. New York [u.a.]: Chapman & Hall.

Gilchrist, S., and E. Zakrajšek. 2012. Credit Spreads and Business Cycle Fluctuations. *American Economic Review*, 102:1692–1720.

Glosten, L. R., R. Jagannathan, and D. E. Runkle. 1993. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 48:1779–1801.

Golubev, Y., W. K. Härdle, and R. Timofeev. 2014. Testing monotonicity of pricing kernels. *AStA Advances in Statistical Analysis*, 98:305–326.

Gordon, M. J. 1962. *The Investment, Financing, and Valuation of the Corporation*. RD Irwin Homewood, IL.

Grith, M., W. Härdle, and J. Park. 2013. Shape Invariant Modeling of Pricing Kernels and Risk Aversion. *Journal of Financial Econometrics*, 11:370–399.

Gu, S., B. Kelly, and D. Xui. 2019. Empirical Asset Pricing via Machine Learning. *Working Paper.* URL `https://ssrn.com/abstract=3159577`.

Géron, A. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems.* Sebastopol, CA: O'Reilly Media.

Hagan, P. S., D. Kumar, A. S. Lesniewski, and D. E. Woodward. 2002. Managing Smile Risk. *The Best of Wilmott*, 1:249–296.

Hansen, L. P., and R. Jagannathan. 1991. Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy*, 99:225–262.

Hansen, L. P., and J. A. Scheinkman. 2009. Long-Term Risk: An Operator Approach. *Econometrica*, 77:177–234.

Hansen, L. P., and K. J. Singleton. 1982. Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica*, 50:1269–1286.

Hansen, L. P., and K. J. Singleton. 1983. Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns. *Journal of Political Economy*, 91:249–265.

Heston, S. L. 1993. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6:327–343.

Hofmann, M., and M. Uhrig-Homburg. 2018. Volatility Noise. *Working Paper.* Available at SSRN: https://ssrn.com/abstract=3130045.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257.

Hull, J. 2018. *Options, Futures, and Other Derivatives.* 10th ed. New York, NY: Pearson.

Hutchinson, J. M., A. Lo, and T. Poggio. 1994. A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. *The Journal of Finance*, 49:851–889.

Härdle, W. 1991. *Applied Nonparametric Regression*. Econometric Society monographs ; 19, 1st ed. Cambridge [u.a.]: Cambridge Univ. Pr.

Iott, J., R. T. Haftka, and H. M. Adelman. 1985. Selecting Step Sizes in Sensitivity Analysis by Finite Differences. *NASA Technical Memorandum 86382*. URL: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19850025225.pdf.

Jackwerth, J. C. 1999. Option-Implied Risk-Neutral Distributions and Implied Binomial Trees: A Literature Review. *The Journal of Derivatives*, 7:66–82.

Jackwerth, J. C. 2000. Recovering Risk Aversion from Option Prices and Realized Returns. *The Review of Financial Studies*, 13:433—-451.

Jackwerth, J. C. 2004. Option-Implied Risk-Neutral Distributions and Risk Aversion. *Working Paper*.

Jackwerth, J. C., and M. Menner. 2018. Does the Ross Recovery Theorem work Empirically? *Working Paper*. URL `https://ssrn.com/abstract=2960733`.

Jackwerth, J. C., and M. Rubinstein. 1996. Recovering Probability Distributions from Contemporaneous Security Prices. *The Journal of Finance*, 51:1611–1631.

Jensen, C. S., D. Lando, and L. H. Pedersen. 2019. Generalized Recovery. *Journal of Financial Economics*, 133:154–174.

Jiang, G. J., and Y. S. Tian. 2005. The Model-Free Implied Volatility and Its Information Content. *The Review of Financial Studies*, 18:1305–1342.

Jondeau, E., and M. Rockinger. 2000. Reading the smile: the message conveyed by methods which infer risk neutral densities. *Journal of International Money and Finance*, 19:885–915.

Joslin, S., M. Priebsch, and K. J. Singleton. 2014. Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks. *The Journal of Finance*, 69:1197–1233.

Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82:35–45.

Kingma, D. P., and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. URL `http://arxiv.org/abs/1412.6980`.

Knüppel, M. 2015. Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments. *Journal of Business & Economic Statistics*, 33:270–281.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 14:1137–1145.

Lewis, M. 2011. *The Big Short: Inside the Doomsday Machine.* Penguin UK.

Li, Q., and J. S. Racine. 2007. *Nonparametric econometrics: theory and practice.* Princeton University Press.

Linn, M., S. Shive, and T. Shumway. 2018. Pricing Kernel Monotonicity and Conditional Information. *The Review of Financial Studies*, 31:493—-531.

Lucas Jr, R. E. 1978. Asset prices in an exchange economy. *Econometrica*, 46:1429–1445.

Ludwig, M. 2015. Robust Estimation of Shape-Constrained State Price Density Surfaces. *The Journal of Derivatives*, 22:56–72.

Martin, I. 2017. What is the Expected Return on the Market? *The Quarterly Journal of Economics*, 132:367–433.

Martin, I. W. R., and C. Wagner. 2019. What is the Expected Return on a Stock? *The Journal of Finance*, 74:1887–1929.

Matisko, P., and V. Havlena. 2012. Noise covariance estimation for Kalman filter tuning using Bayesian approach and Monte Carlo. *International Journal of Adaptive Control and Signal Processing*, 27:957–973.

McNeil, A. J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools.* revised edition ed. Princeton university press.

Mehra, R. 1970. On the Identification of Variances and Adaptive Kalman Filtering. *IEEE Transactions on Automatic Control*, 15:175–184.

Mehra, R., and E. C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics*, 15:145–161.

Merton, R. C. 1969. Lifetime Portfolio Selection Under Uncertainty: The Continuous-time Case. *The Review of Economics and Statistics*, 51:247–257.

Merton, R. C. 1975. Optimum Consumption and Portfolio Rules in a Continuous-Time Model. In *Stochastic Optimization Models in Finance*, pp. 621–661. Elsevier.

Mullachery, V., A. Khera, and A. Husain. 2018. Bayesian Neural Networks. *arXiv preprint arXiv:1801.07710.*

Næs, R., J. A. Skjeltorp, and B. A. Ødegaard. 2011. Stock Market Liquidity and the Business Cycle. *The Journal of Finance*, 66:139–176.

OptionMetrics. 2016. *IvyDB US File and Data Reference Manual, Version 3.1, Rev. 8/19/2016.* OptionMetrics LLC, 1776 Broadway, Suite 1800, New York, NY 10019.

Parzen, E. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33:1065–1076. URL `http://projecteuclid.org/euclid.aoms/1177704472`.

Patton, A. J., and A. Timmermann. 2010. Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98:605–625.

Politis, D. N., and J. P. Romano. 1994. The Stationary Bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

Ramsey, F. P. 1931. *The Foundations of Mathematics and Other Logical Essays. With a Pref. By G.E. Moore.* K. Paul, Trench, Truber and Co., London.

Rohde, U. L., G. C. Jain, A. K. Poddar, and A. K. Ghosh. 2012. *Introduction to Integral Calculus: Systematic Studies with Engineering Applications for Beginners.* John Wiley & Sons.

Rosenberg, J. V., and R. F. Engle. 2002. Empirical Pricing Kernels. *Journal of Financial Economics*, 64:341–372.

Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27:832–837. URL `http://projecteuclid.org/euclid.aoms/1177728190`.

Ross, S. 2015. The Recovery Theorem. *The Journal of Finance*, 70:615–648.

Rothfuss, J., F. Ferreira, S. Boehm, S. Walther, M. Ulrich, T. Asfour, and A. Krause. 2019a. Noise Regularization for Conditional Density Estimation. *Working Paper.* URL `https://arxiv.org/abs/1907.08982`.

Rothfuss, J., F. Ferreira, S. Walther, and M. Ulrich. 2019b. Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks. *Working Paper.* URL `https://arxiv.org/abs/1903.00954`.

Savage, L. J. 1954. *The Foundations of Statistics.* John Wiley, New York.

Schneider, P., and F. Trojani. 2015. Fear Trading. *Swiss Finance Institute Research Paper.* No. 15-03. Available at SSRN: https://ssrn.com/abstract=1994454.

Schneider, P., and F. Trojani. 2019. (Almost) Model-Free Recovery. *The Journal of Finance*, 74:323–370.

Schneider, P., C. Wagner, and J. Zechner. 2019. Low Risk Anomalies? *Journal of Finance, Forthcoming*, pp. 19–50.

Schönbucher, P. J. 1999. A market model for stochastic implied volatility. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357:2071–2092.

Sharma, J. L., M. Mougoue, and R. Kamath. 1996. Heteroscedasticity in stock market indicator return data: volume versus GARCH effects. *Applied Financial Economics*, 6:337–342.

Shumway, R. H., and D. S. Stoffer. 2017. *Time Series Analysis and Its Applications: With R Examples.* Springer Texts in Statistics, fourth edition ed. Cham, Switzerland: Springer.

Sichert, T. 2020. The Pricing Kernel Is U-shaped. *Working paper.* URL `https://ssrn.com/abstract=3095551`.

Sugiyama, M., and I. Takeuchi. 2010. Conditional density estimation via Least-Squares Density Ratio Estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, pp. 781–788. URL `http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_SugiyamaTSKHO10.pdf`.

Swanson, E. T. 2016. Discussion of Options-Implied Probability Density Functions for Real Interest Rates. Available here: http://www.ijcb.org/journal/ijcb16q3a3_disc.pdf.

Ulrich, M., S. Florig, and C. Wuchte. 2019. A Model-Free Term Structure of U.S. Dividend Premiums. *Working Paper.* URL `https://ssrn.com/abstract=3217096`.

van der Vaart, A. W. 2000. *Asymptotic Statistics.* 1st ed. Cambridge, UK: Cambridge University Press.

Vergote, O., and J. M. P. Gutiérrez. 2012. Interest rate expectations and uncertainty during ECB governing council days: evidence from intraday implied densities of 3-month Euribor. *Journal of Banking & Finance*, 36:2804–2823.

von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior, 2nd rev.* Princeton University Press.

Walden, J. 2017. Recovery with Unbounded Diffusion Processes. *Review of Finance*, 21:1403–1444.

Walther, S., and M. Ulrich. 2019. Option-Implied Information: What's the Vol Surface Got to Do With It?,. *Working Paper.* URL `https://ssrn.com/abstract=3184767`.

Walther, S., M. Ulrich, J. Rothfuss, and F. Ferreira. 2019. Forward-looking P. *Working Paper.* URL `https://ssrn.com/abstract=3437281`.

Welch, G., and G. Bishop. 2001. An Introduction to the Kalman Filter. *Proc of SIGGRAPH, Course*, 8:41.

Welch, I., and A. Goyal. 2008. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21:1455–1508.

World Bank. 2020. Market capitalization of listed domestic companies (current US$). URL `https://data.worldbank.org/indicator/CM.MKT.LCAP.CD`.

Wright, J. H. 2016. Options-Implied Probability Density Functions for Real Interest Rates. *International Journal of Central Banking*, 12:129–149.

Yao, J., Y. Li, and C. L. Tan. 2000. Option price forecasting using neural networks. *Omega*, 28:455–466.

Zhao, B., and S. D. Hodges. 2013. Parametric modeling of implied smile functions: a generalized SVI model. *Review of Derivatives Research*, 16:53–77.

Zhao, Y., C. Stasinakis, G. Sermpinis, and Y. Shi. 2018. Neural network copula portfolio optimization for exchange traded funds. *Quantitative Finance*, 18:761–775.