Robust and Non parametric estimation of state price densities with machine learning

Julian Beatty*

First draft: May 4, 2024
This version: Decemter 15, 2024

Abstract

This paper develops a novel two step approach for estimating option implied distributions. Combining kernel ridge regression, with kernel density estimation, we produce realistic option implied distributions. We conduct a Monte-Carlo experiment to show the effectiveness of our machine learning approach across a variety of statistical metrics, compared to approaches using other curve fitting procedures such as local regression and parametric fitting. We emphasize not only the accuracy of our approach in recovering these densities, but also the robustness when faced with exceptionally noisy data that many financial practitioners may face in real life markets.

Keywords: inflation, deflation, cryptocurrency, options.
JEL codes: D12, D84, E31, G13, G51.

1 Introduction

This paper introduces a convenient yet effective methodology for estimating the risk-neutral density (RND) from option prices. Our approach leverages kernel ridge regression (KRR) to smooth the implied volatility (IV) surface, from which a pilot RND is extracted. This pilot RND is then post-processed using a weighted kernel density estimator to obtain the final, well-behaved RND.

From a theoretical perspective, our framework offers several key advantages:

1. The interpolation is model-free, allowing it to capture complex features in the RND such as multiple peaks or broad shoulders.

2. The hyperparameters in KRR enable the model to handle outliers in the IV curve more effectively, preventing unrealistic behavior in the RND such as discontinuities or excessive oscillations.

3. The final kernel smoothing step ensures that the estimated RND is always a valid and interpretable probability density, regardless of the quality or noise level in the underlying IV data.

*University of Houston. Address: 4250 Martin Luther King Blvd, Houston, TX 77204.
E-mail: jtbeatty@bauer.uh.edu.

Empirically, we distinguish our contribution from prior work by benchmarking our method against several established alternatives—including splines, polynomial fitting, and local polynomial regression—within a Monte Carlo framework. By simulating option prices under the SABR model, we are able to generate a known "true" RND, allowing us to quantify the accuracy of each method using a variety of statistical distance measures commonly used in the literature.

In addition to evaluating accuracy, we emphasize the robustness of our method in the presence of challenging data. To this end, we complement our simulation results with empirical applications on real-world options data known for its irregularities—such as bitcoin and gold. These examples highlight the practical advantages of our approach and suggest that robust RND extraction may pave the way for deeper empirical investigations into such assets.

1.1 Options and Risk Neutral Density

The risk neutral density is the central topic of this paper. The derivation by Breeden and litzenberger [2] follows: Let $C(K,T)$ denote the price of a European call option with strike K and maturity T. Under the risk-neutral measure $\mathbb{Q}$, the price is given by:

$$C(K,T) = e^{-rT}\mathbb{E}^{\mathbb{Q}}[(S_T - K)^+] = e^{-rT}\int_K^\infty (S_T - K) f(S_T)\, dS_T, \tag{1}$$

where $f(S_T)$ is the risk-neutral density of the terminal asset price S_T, and r is the constant risk-free rate.

Differentiating under the integral sign using Leibniz's Rule:

$$\frac{\partial C(K)}{\partial K} = \frac{d}{dK}\left[e^{-rT}\int_K^\infty (S_T - K) f(S_T)\, dS_T\right] \tag{2}$$

$$= e^{-rT}\left[-\int_K^\infty f(S_T)\, dS_T\right] = -e^{-rT}\mathbb{Q}(S_T > K). \tag{3}$$

Differentiating again:

$$\frac{\partial^2 C(K)}{\partial K^2} = \frac{d}{dK}\left[-e^{-rT}\mathbb{Q}(S_T > K)\right] \tag{4}$$

$$= e^{-rT} f(K). \tag{5}$$

Multiplying both sides by e^{rT}, we obtain the risk-neutral density:

$$f(K) = e^{rT}\frac{\partial^2 C(K)}{\partial K^2}. \tag{6}$$

The second derivative of the call price with respect to strike extracts the marginal state price density. The first derivative gives the negative of the risk-neutral cumulative distribution function, and the second derivative reveals how probability mass is distributed around strike K.

1.2 SABR pricing model

The SABR model (Hagan et al., 2002) is a stochastic volatility model designed to capture the volatility smile observed in options markets. It describes the dynamics of the forward price F_t and its volatility σ_t under the risk-neutral measure $\mathbb{Q}$:

Model Dynamics

$$dF_t = \sigma_t F_t^{\beta}\, dW_t, \tag{7}$$
$$d\sigma_t = \nu \sigma_t\, dZ_t, \tag{8}$$
$$dW_t\, dZ_t = \rho\, dt, \tag{9}$$

where:

- F_t: forward price of the underlying asset (e.g., forward rate),
- σ_t: stochastic volatility process,
- $\beta \in [0, 1]$: elasticity parameter controlling the dependence of volatility on the level of F_t,
- $\nu > 0$: volatility of volatility,
- $\rho \in [-1, 1]$: instantaneous correlation between the Brownian motions W_t and Z_t,
- W_t, Z_t: standard Brownian motions under the risk-neutral measure.

Hagan et al. provided an approximation for the implied Black-Scholes volatility $\sigma_{\text{BS}}(K)$ for a European option with strike K:

$$\sigma_{\text{BS}}(K) \approx \frac{\sigma_0}{(F_0 K)^{(1-\beta)/2}} \left\{ 1 + \left[\frac{(1-\beta)^2}{24} \left(\ln \frac{F_0}{K} \right)^2 + \frac{(1-\beta)^4}{1920} \left(\ln \frac{F_0}{K} \right)^4 \right] \right\} \cdot \frac{z}{x(z)},$$

where:

$$z = \frac{\nu}{\sigma_0} (F_0 K)^{(1-\beta)/2} \ln \left(\frac{F_0}{K} \right),$$

$$x(z) = \ln \left(\frac{\sqrt{1 - 2\rho z + z^2} + z - \rho}{1 - \rho} \right).$$

This approximation is valid for small ν and small $\ln(F_0/K)$, and is widely used for calibrating SABR parameters to implied volatility surfaces.

Calibration Method

The standard approach is to minimize the squared error between model and market implied volatilities:

$$\min_{\alpha,\rho,\nu} \sum_{i=1}^{n} \left[\sigma_{\text{SABR}}(K_i; \alpha, \beta, \rho, \nu) - \sigma_{\text{market}}(K_i) \right]^2.$$

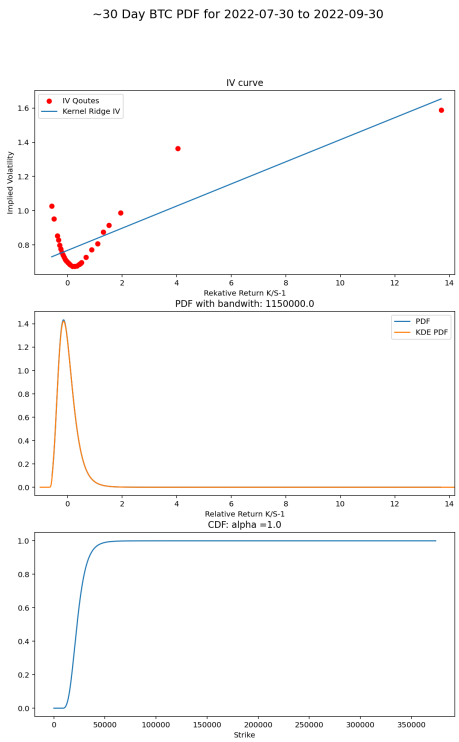
To ensure stability, realistic constraints are enforced:

$$-1 < \rho < 1, \quad \nu > 0, \quad \alpha > 0.$$

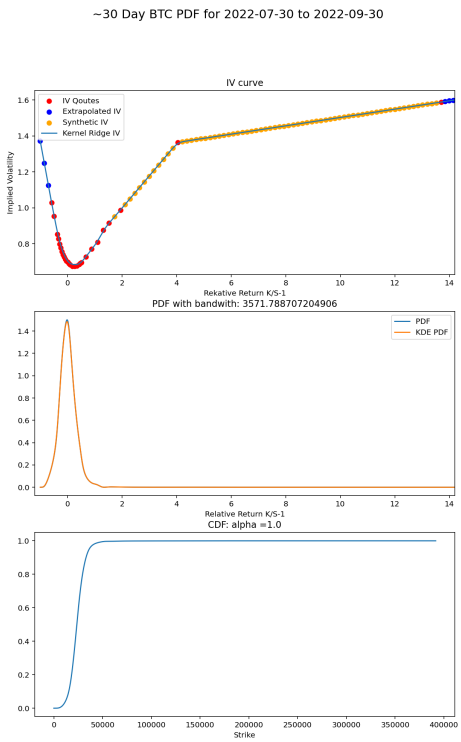
2 Practical considerations

Before discussing the implementation of the methodologies above, we first outline some practical considerations. As noted in papers such as [3], the similarity between implied volatility (IV) and risk-neutral density (RND) estimates across different methodologies tends to diminish when fewer option prices are available. Likewise, the quality of the RND estimates deteriorates when the number of available option quotes is limited, especially for non parametric methods. To improve the number and quality of our option surface, we replace ITM call options with OTM put options using the put-call parity relationship.

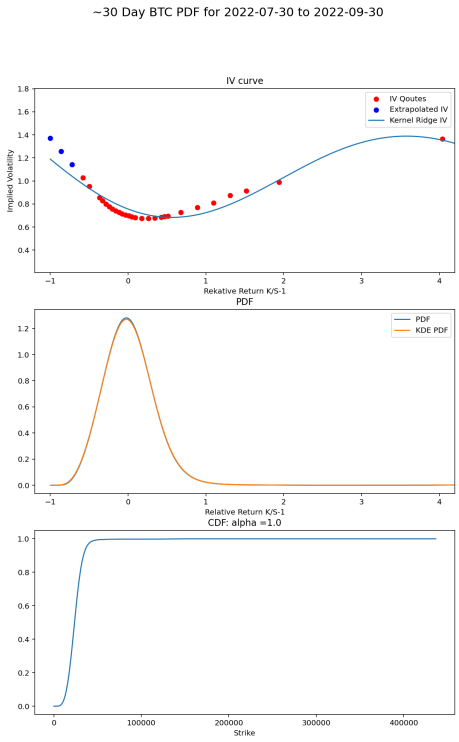
Another important consideration, also noted by [5, 6], is the dispersion of the option data, which significantly influences the quality of the implied volatility (IV) surface and the resulting risk-neutral density (RND) estimates. In most datasets, option strikes are unevenly distributed, with wider spacing particularly evident among deep out-of-the-money (OTM) or in-the-money (ITM) options. This irregular dispersion can lead to anomalous behavior in the IV surface and RND, as illustrated in Figure 1. We alleviate this issue by linearly interpolating the IV curve whenever the spacing between two options is more than five times the median spacing. The linear interpolation is uniformly at the average strike spacing until the the spacing between the options is less than 3 times the Median distance.



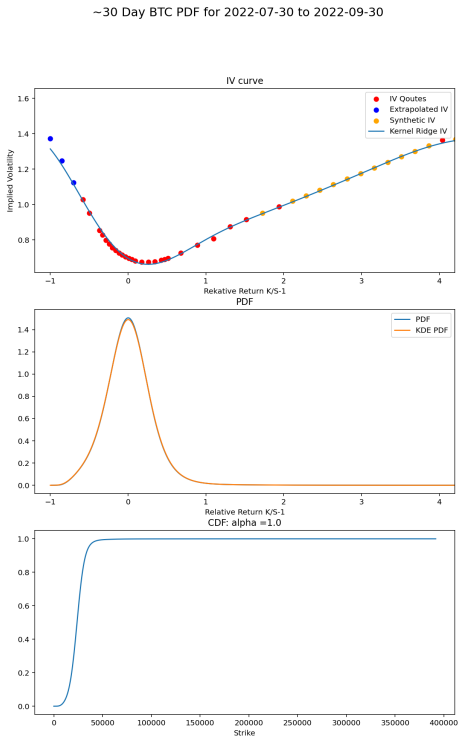
(a) Local Polynomial



(b) Local Polynomial



(c) Kernel Ridge



(d) Kernel Ridge

Figure 1: Exhibit of anomalous behavior when fitting options with extreme moneyness. Linearly interpolating between options whenever the strike gap is more than 5 times the average. Subfigure 1 demonstrates that the Cross validation algorithm will pick an excessively large bandwidth in order to accommodate the deep OTM options.

For this experiment, the risk-free rate was obtained by interpolating the U.S. Treasury spot yield curve using the Svensson five-factor model, as provided on the Federal Reserve's website. In cases where the interpolation resulted in negative yields at shorter maturities, we imposed a lower bound and capped the rate at 0.001.

2.1 Locally Linear Kernel Regression

Local polynomial regression is a commonly employed non-parametric smoothing approach that has notably been implemented by [1]. Consider a sample $\{(X_i, Y_i)\}_{i=1}^n$ drawn from the non-parametric regression model:

$$Y_i = m(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid X_i] = 0.$$

The objective is to estimate the conditional expectation function $m(x) = \mathbb{E}[Y \mid X = x]$.

Local Linear Approximation

Unlike the Nadaraya–Watson estimator, which assumes a locally constant approximation, we adopt a locally linear approach. Specifically, in a neighborhood of a fixed evaluation point x, the function $m(z)$ is approximated by a first-order Taylor expansion:

$$m(z) \approx a + b(z - x),$$

where a approximates $m(x)$, and b represents the local slope.

The coefficients (a, b) are obtained by solving a weighted least squares problem:

$$\min_{a,b} \sum_{i=1}^{n} K_h(X_i - x) \left[Y_i - a - b(X_i - x)\right]^2,$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ is a kernel function with bandwidth h. This weighting ensures that observations closer to the target point x exert greater influence on the local fit.

Bandwidth Selection

We employ the Gaussian kernel in all estimations; however, the choice of bandwidth h plays a more critical role than the kernel shape itself. The bandwidth controls the effective size of the local neighborhood and thus governs the bias–variance trade-off.

To select an appropriate value of h, we conduct 5-fold cross-validation over a grid of candidate bandwidths. The search grid spans values ranging from half the minimum strike spacing to five times the average spacing between strikes. For each candidate bandwidth, we interpolate the entire implied volatility (IV) curve. If any region of the curve fails to include data points within the kernel window—resulting in undefined or zero interpolated values—the corresponding candidate is assigned an infinite cross-validation error. This constraint ensures the selected bandwidth provides global coverage across the strike domain.

2.2 Kernel Ridge Regression

Let $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, denote the dataset. Kernel ridge regression seeks to estimate a function $f : \mathbb{R}^d \to \mathbb{R}$ that balances empirical fit with smoothness by incorporating both kernel methods and ℓ_2-regularization.

Ridge Regression in Feature Space

Suppose the input space is mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi : \mathbb{R}^d \to \mathcal{H}$. Ridge regression in this space solves:

$$\min_{w \in \mathcal{H}} \left\{ \sum_{i=1}^{n} \left(y_i - \langle w, \phi(x_i) \rangle \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \right\},$$

where $\lambda > 0$ is the regularization parameter.

Representer Theorem

By the Representer Theorem, the solution $w \in \mathcal{H}$ admits the finite-dimensional representation:

$$w = \sum_{i=1}^{n} \alpha_i \phi(x_i),$$

which implies the fitted function is:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x),$$

where $K(x, x') = \langle \phi(x), \phi(x') \rangle$ denotes the kernel function.

Optimization in Terms of Coefficients

Substituting the representer form into the objective yields an optimization problem in $\boldsymbol{\alpha} \in \mathbb{R}^n$:

$$\min_{\boldsymbol{\alpha}} \left\{ \|y - K\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} \right\},$$

where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix with entries $K_{ij} = K(x_i, x_j)$, and $y \in \mathbb{R}^n$ is the response vector.

Taking the derivative and setting it to zero yields:

$$(K + \lambda I)\boldsymbol{\alpha} = y.$$

Thus, the solution is given in closed form by:

$$\boldsymbol{\alpha} = (K + \lambda I)^{-1} y.$$

Prediction

For a new input x, the predicted value is:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) = \mathbf{k}(x)^\top \boldsymbol{\alpha},$$

where $\mathbf{k}(x) = [K(x_1, x), \ldots, K(x_n, x)]^\top$ is the vector of kernel evaluations.

Radial Basis Function Kernel

Throughout, we use the Radial Basis Function (RBF) kernel, also known as the Gaussian kernel:

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2 \right),$$

where $\gamma > 0$ determines the kernel width.

Role of γ The parameter γ controls the locality of the kernel:

- **Large γ**: Produces a narrow kernel with fast decay, leading to a highly flexible model that may overfit.
- **Small γ**: Yields a wide kernel, producing a smoother fit with greater bias and risk of underfitting.

Interaction with λ The regularization parameter λ interacts with γ to determine model complexity:

- **Small λ, large γ**: High variance, potential overfitting.
- **Large λ, small γ**: High bias, potential underfitting.

Optimal values of λ and γ are selected using cross-validation.

2.3 Smoothing Splines

Smoothing splines provide an alternative nonparametric regression method that balances data fidelity and smoothness. Spline based approaches have been implemented by [4]. Given observations $(x_1, y_1), \ldots, (x_n, y_n)$, the estimator $\hat{f}(x)$ minimizes:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b \left(f''(x)\right)^2 \, dx.$$

This objective consists of:

1. **Residual Sum of Squares (RSS)**: Measures goodness-of-fit to the data.
2. **Roughness Penalty**: Controls the curvature of the function by penalizing the integrated squared second derivative.

The smoothing parameter $\lambda \geq 0$ governs the trade-off:

- $\lambda = 0$: Exact interpolation (natural cubic spline).
- $\lambda \to \infty$: Linear regression (maximal smoothness).

The solution is a *natural cubic spline* with knots at $x_1, \ldots, x_n$, characterized by piecewise cubic segments joined smoothly with continuous second derivatives and natural boundary conditions.

Selection of the smoothing parameter λ is typically performed using cross-validation or generalized cross-validation (GCV), ensuring an optimal balance between overfitting and underfitting.

Distance and Divergence Measures

A random sample from the target distribution is needed in order to calculate several of our mentioned statistical metrics, such as the energy distance. We use the method of inverse-cdf-transforms to generate samples as described below. Once we have the implied probability density function (PDF), we compute the corresponding cumulative distribution function (CDF).

To generate random samples from this implied distribution, we apply the Inverse Transform Sampling method:

1. Generate a uniform random variable $U \sim \text{Uniform}(0, 1)$.
2. Apply the inverse CDF $X = F^{-1}(U)$ to obtain samples from the implied distribution.

These generated samples can then be used for further analysis, such as statistical testing.

Hellinger Distance

The **Hellinger distance** between two probability distributions P and Q with probability density functions $p(x)$ and $q(x)$ is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \left(\int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \right)^{1/2}$$

Alternatively, the square of the Hellinger distance is expressed as:

$$H^2(P,Q) = \frac{1}{2} \int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx$$

Key Features:

- Ranges from 0 to 1.
- Symmetric, i.e., $H(P,Q) = H(Q,P)$.
- Measures the difference between the square root densities of two distributions.
- Bounded between 0 and 1.

Interpretation: The Hellinger distance quantifies the shape difference between two distributions by comparing their square root densities. A value of 0 indicates that the distributions are identical, and a value of 1 indicates that the distributions are maximally different (i.e., have no overlap in their support).

Wasserstein Distance

The **Wasserstein distance** of order p between two probability distributions P and Q is defined as:

$$W_p(P,Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} d(x,y)^p d\gamma(x,y) \right)^{1/p}$$

Where $\Gamma(P,Q)$ represents the set of all joint distributions γ whose marginal distributions are P and Q, and $d(x,y)$ is the distance between points x and y. For discrete distributions, this becomes a minimization problem over all possible couplings of P and Q.

Key Features:

- Sensitive to both the shape and spread (variance) of distributions.
- Can be interpreted as the minimum "work" required to transform one distribution into another.
- Ranges from 0 to infinity.
- Sensitive to the distribution support and how mass must be transported between distributions.

Interpretation: The Wasserstein distance measures the optimal transport needed to transform one distribution into another. It quantifies the effort required to move mass from one distribution to match another, and it captures the difference in their shapes and spreads (such as variance).

Total Variation Distance

The **Total Variation distance** between two probability distributions P and Q is defined as:

$$d_{TV}(P, Q) = \frac{1}{2} \int_{-\infty}^{\infty} |p(x) - q(x)| dx$$

Key Features:

- Ranges from 0 to 1.
- Measures the maximum difference between the probabilities assigned by two distributions.
- Sensitive to the **maximum discrepancy** between the distributions.
- Symmetric, i.e., $d_{TV}(P, Q) = d_{TV}(Q, P)$.

Interpretation: The Total Variation distance measures the maximum possible difference between two distributions. It quantifies how much one distribution needs to be altered to match the other, reflecting the greatest discrepancy in probability assignments for any event. A distance of 0 means the distributions are identical, while a distance of 1 means they are completely disjoint.

Energy Distance

The **Energy distance** between two distributions P and Q is defined as:

$$D_E(P, Q) = 2\mathbb{E}_{X \sim P, Y \sim Q} \|X - Y\| - \mathbb{E}_{X, X' \sim P} \|X - X'\| - \mathbb{E}_{Y, Y' \sim Q} \|Y - Y'\|$$

Where $\| \cdot \|$ represents the Euclidean distance between points, and $\mathbb{E}$ represents the expectation over the distributions P and Q.

Key Features:

- Sensitive to the second-order moments (variances) of the distributions.
- Measures the average pairwise distances between samples from each distribution.
- Ranges from 0 to infinity.
- Symmetric, i.e., $D_E(P, Q) = D_E(Q, P)$.

Interpretation: The Energy distance quantifies the average pairwise distance between samples drawn from each distribution. It is especially sensitive to the spread (variance) of the distributions, capturing how far apart the points are in each distribution. It is useful for comparing the overall shape and spread of the distributions.

Jensen-Shannon Divergence

The **Jensen-Shannon (JS) divergence** is a symmetrized version of the Kullback-Leibler divergence. It is defined as:

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

Where $M = \frac{1}{2}(P + Q)$ is the average distribution, and $D_{KL}(P||Q)$ is the Kullback-Leibler divergence between P and Q, defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Key Features:

- Symmetric, i.e., $D_{JS}(P, Q) = D_{JS}(Q, P)$.
- Ranges from 0 (when $P = Q$) to $\log 2$ (when the distributions are maximally different).
- Measures the average divergence from each distribution to the average of the two.
- Useful for measuring the similarity between two distributions in a way that is less sensitive to outliers.

Interpretation: The Jensen-Shannon divergence provides a symmetrized measure of the difference between two distributions, making it more stable and less sensitive to the tails of the distributions compared to the Kullback-Leibler divergence. It is often used to assess how similar two distributions are, with a value of 0 indicating identical distributions.

Measure	Definition	Interpretation	Best for
Wasserstein Distance (Earth Mover's Distance)	Measures the "cost" of transporting probability mass to transform one distribution into another.	Captures both shape and location differences.	Comparing empirical distributions and capturing shape differences.
Energy Distance	A distance metric based on expected differences in pairwise distances between samples from two distributions.	Related to Wasserstein but considers **distribution spread** more effectively.	Multivariate distributions, clustering, shape differences.
Jensen-Shannon Divergence (JSD)	Measures the difference between two distributions using the average of their KL divergences with respect to their mean distribution.	A smoothed, symmetric version of KL divergence.	Comparing probability distributions where interpretability and symmetry are needed.
Total Variation Distance	Measures the largest possible difference in probabilities assigned by two distributions.	Maximum difference in probabilities across all events.	Probability mass comparisons and discrete distributions.
Hellinger Distance	Measures the similarity between probability distributions based on their square root transformations.	Similar to Bhattacharyya distance, emphasizing probability mass similarity.	Comparing PDFs and PMFs, especially when probability mass concentration matters.
Kolmogorov-Smirnov (KS) Test	Measures the maximum absolute difference between two cumulative distribution functions (CDFs).	Used for hypothesis testing (whether two samples come from the same distribution).	Hypothesis testing, comparing empirical CDFs.

Table 1: Comparison of Distribution Similarity Measures

3 Empirical Strategy

The primary objective of this study is to accurately and reliably recover the *state price density* (SPD) from a discrete set of option prices. To evaluate the performance of different estimation methods, we conduct a Monte Carlo experiment.

Our experimental design begins by calibrating a stochastic (SABR) model to an observed implied volatility curve. Using the estimated SABR parameters, we generate simulated implied volatility at the observed strike prices. To introduce realistic uncertainty, we perturb these simulated volatilities by adding a noise term drawn from a uniform distribution of $\pm 5\%$ of the implied volatility value.

Each estimation method is then applied to the perturbed implied volatility curve to recover the state price density (SPD). Since the true SPD is known from the calibrated SABR model—obtained by simulating a continuum of implied volatilities—we evaluate the accuracy of each method by comparing the estimated SPDs to the true SPD using a range of statistical distance measures, summarized in Table 1.

To ensure robustness and realism, we conduct 200 trials of this experiment for every Bitcoin option with a maturity of 14, 30, and 60 days that has ever been traded as of August 2024. The total number of simulations per maturity is 30,080 for 14-day options and 7,200 for 30-day and 60-day options. Results are aggregated by maturity to provide a comprehensive assessment of estimation accuracy across different time horizons. The hyperparameter range used for kernel ridge regression was: $\alpha \in (-1.5, 3, 10)$ and $\gamma \in (-0.1, 1.5, 20)$.

4 Results

The monte-carlo experiment was performed on bitcoin options, and GLD options, and the mean statistics were reported in tables. The experiment was also performed at 2.5 and 7.5 percent noise thresholds but very similar results were obtained. The main difference between bitcoin options and traditional equity options were the number of options, and the spacing between. Whereas bitcoin had relatively few options with large spacings between them, equity options often have hundreds, with very narrow spacing that become larger as moneyness decreases. In our experiment, Kernel Ridge Regression (KRR) consistently outperformed quartic polynomials, locally linear regression and splines across all metrics and datasets as can be seen in tables 2, 3, 6, 5.

Table 2: Statistical Distance Summary for BTC 30

Metric	Kernel Ridge	Quartic	Spline	Local Linear
Jensen-Shannon Div.	0.0010848	0.00160489	0.00163265	0.00443763
Hellinger Dist.	0.0297074	0.0364546	0.0366047	0.0583029
Total Variation Dist.	0.0285947	0.0323478	0.0407278	0.0503389
Energy Dist.	0.00892426	0.0100368	0.0133425	0.0124074
Wasserstein Dist.	0.00522356	0.008508594	0.00730496	0.00701566
Kolm. P. Value	0.0105105	0.00807813	0.0034668	0.00058176

Table 3: Statistical Distance summary for GLD 30 day

Metric	Kernel Ridge	Quartic	Spline	Local Linear
Jensen-Shannon Div.	0.0226439	0.0410863	0.049293	0.0390436
Hellinger Dist.	0.128847	0.177646	0.197497	0.18774
Total Variation Dist.	0.130602	0.206533	0.23709	0.1484
Energy Dist.	0.0177713	0.0360062	0.0469422	0.0430471
Wasserstein Dist.	0.00525485	0.0109267	0.0157975	0.0178283
Kolm. P.Value	0.00260681	6.6718e-11	8.74068e-56	6.14061e-05

The performance of KRR was largely driven by its lack of outliers which were particularly harmful for local regression, and to some extent smoothing splines for the more densely packed equity options which can be seen in summary tables 4. Looking at total variation distance for instance, we can observe that the 75 percentile for Kernel Ridge was only 0.039 whereas for splines and locally linear it was 0.057 and 0.064. Local regression had particular difficulty with outliers that were induced by the noise. As these points entered and exit the the weighted average calculation as the window moves, the IV curve experiences tiny abrupt changes which leads to multi modal or jagged anomalies in the density curve. The smoothing splines encountered a similar issue; when cross validating the regularization, the algorithm tended to select the parameter at the lower boundary of the grid search which occasionally results in an IV curve with tiny oscillations that were transformed into large spikes or anomalous in the state price density. These anomalous behaviors result in poor scores in the monte carlo situation; which can be visualized in figure 2.

Table 4: Total Variation distance across different methods for BTC 60.

Statistic	Quartic	S-spline	KRR-RBF	LP
Mean	0.042865	0.0526707	0.0402903	0.0612312
Std	0.0673586	0.0776597	0.0701154	0.054817
Min	0.00044793	0.00237292	0.00326762	0.0096048
25%	0.0171311	0.0255207	0.0172151	0.0345938
50%	0.026078	0.0374655	0.0258027	0.0461731
75%	0.0329333	0.057085	0.0392315	0.0641463
Max	0.491894	0.506839	0.459028	0.375166

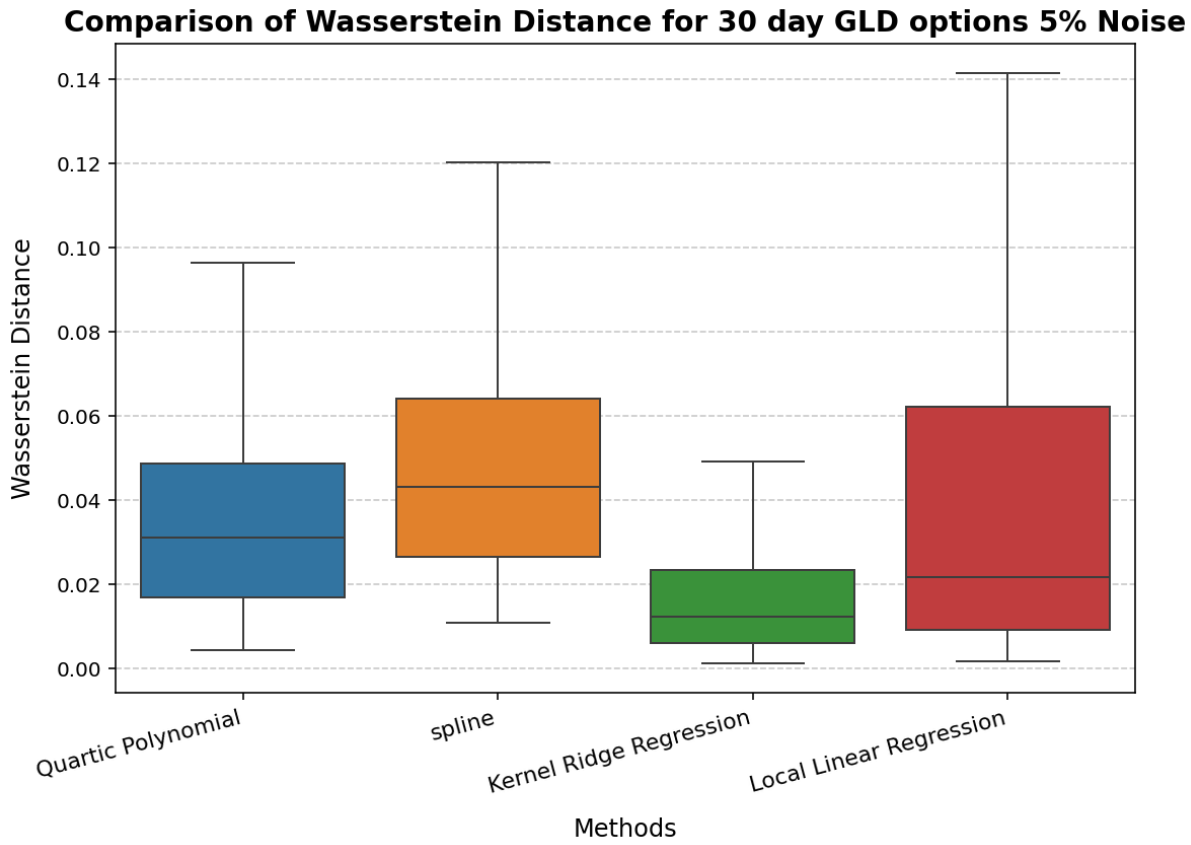


Figure 2: Box plot of Wasserstein Distance for GLD options with 30 days of expiration.

The only way to avoid this issue in our simulations was to raise the lower bound of the grid search. However, doing so was functionally equivalent to arbitrarily selecting a level of regularization, as the algorithm consistently chose the smallest allowable regularization parameter. This presents an undesirable dilemma: cross-validation cannot be relied upon to select a suitable parameter because it tends to overfit the implied volatility (IV) curve, with no consideration for the resulting state price density.

By contrast, kernel ridge regression (KRR) circumvents this overfitting problem. One key reason is that KRR is built on radial basis function (RBF) kernels, which remain smooth across a wide range of gamma and alpha values. As a result, the interpolated IV curves generated by KRR consistently yield well-behaved densities without spurious discontinuities. Moreover, the gamma parameter in KRR modulates the influence of each training point, effectively allowing the model to ignore outliers rather than forcing a tight fit. This is fundamentally different from regularization alone: while parameters like alpha (in KRR) or lambda (in splines) help smooth the curve, they do not prevent the curve from contorting around outliers—behavior that often leads to unnecessarily multimodal densities.

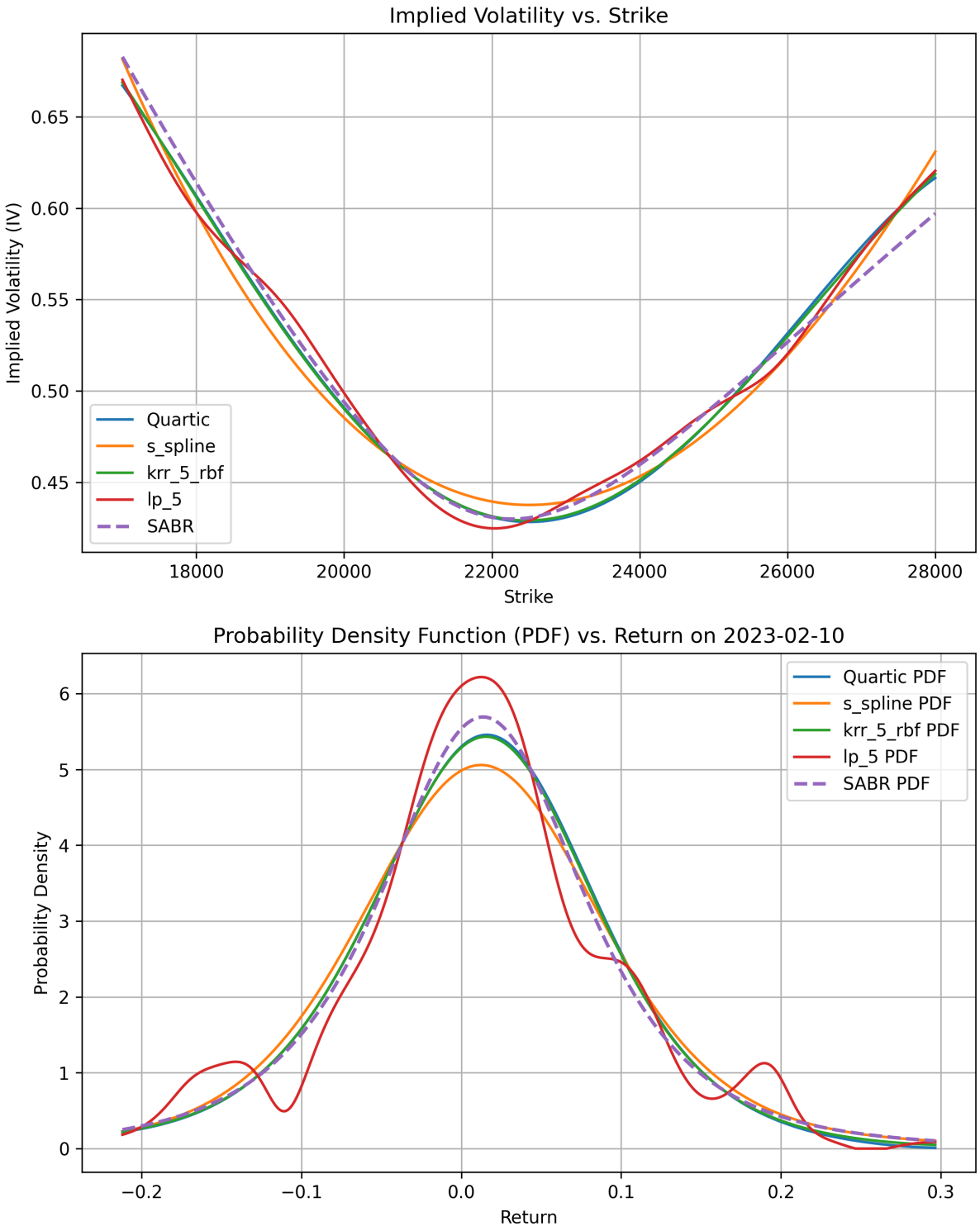


Figure 3: Monte Carlo simulation on bitcoin options.

The quartic polynomial, while simple, also performed reasonably well. By design, polynomials are smooth, which helps avoid the most severe fitting failures. However, in contrast to KRR, the quartic polynomial tended to underfit the IV curve and lacked the flexibility needed to accommodate complex structures in the data.

To further substantiate our simulation results, we reproduce some of the key challenges using real-world option data. These examples highlight the practical difficulties of extracting meaningful state price densities, especially when hyperparameters must be tuned manually to obtain feasible results. In contrast, our kernel ridge regression approach consistently demonstrates robustness, even when applied to some of the most poorly behaved option surfaces.

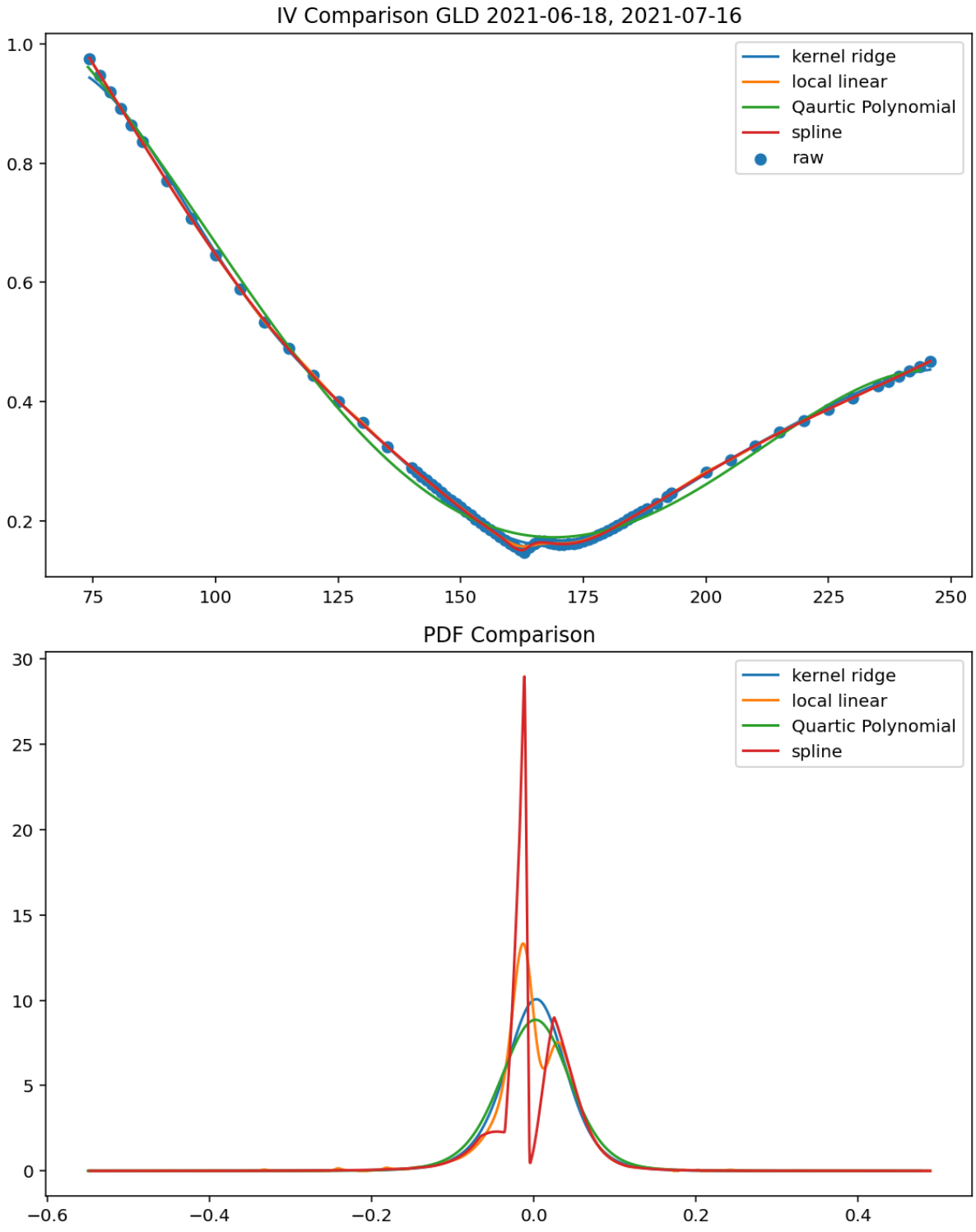


Figure 4: Comparison between methods for an exemplary GLD option. The phenomon seen here is the well document IV jump when transitioning to ITM and ATM options. Kernel Ridge is able to smooth the transition out, whereas other methods either under fit (polynomial), or result in distorted densities.

Conclusion

In conclusion, this paper introduces and explores the use of kernel ridge regression (KRR) as a tool for interpolating volatility surfaces and extracting well-behaved risk-neutral densities. Unlike prior studies, we provide a rigorous benchmark of KRR against several alternative methods—including smoothing splines, local polynomial regression, and polynomial fitting—and evaluate performance using a variety of statistical distance metrics. Our results show that KRR is more effective at handling noisy or irregular option data, a strength we attribute to its use of radial basis function kernels and multiple hyperparameters that mitigate overfitting.

5 Additional Tables

Metric	Kernel Ridge	Quartic	Spline	Local Linear
Jensen-Shannon Div.	0.00095548	0.00132908	0.00116138	0.00338178
Hellinger Dist.	0.0268557	0.0324947	0.0300496	0.0476528
Total Variation Dist.	0.0278001	0.0296267	0.0368198	0.0462245
Energy Dist.	0.00761375	0.00828974	0.0106972	0.00969395
Wasserstein Dist.	0.00331099	0.00367686	0.00430612	0.00421496
Kolm. P. Value	0.00869345	0.00514105	0.00385793	0.00152733

Table 5: This table summarizes the aggregate statistical distance measures for bitcoin options with maturities of 14 days. The total number of simulations was 30,800.

Metric	Kernel Ridge	Quartic	Spline	Local Linear
Jensen-Shannon Div.	0.00476477	0.0054694	0.00651678	0.00774498
Hellinger Dist.	0.0414071	0.0428196	0.0540959	0.0755742
Total Variation Dist.	0.0402903	0.042865	0.0526707	0.0612312
Energy Dist.	0.0148135	0.0157774	0.0201408	0.018211
Wasserstein Dist.	0.0103395	0.010849	0.0142087	0.0125806
Kolm. P. Value	0.0118995	0.0112203	0.00111042	0.00055656

Table 6: This table summarizes the aggregate statistical distance measures for bitcoin options with maturities of 60 days. The total number of simulations was 7200.

References

[1] Y. Aıt-Sahalia and J. Duarte. Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1-2):9–47, 2003.

[2] D. T. Breeden and R. H. Litzenberger. Prices of state-contingent claims implicit in option prices. *Journal of business*, pages 621–651, 1978.

[3] J. M. Campa, P. K. Chang, and R. L. Reider. Implied exchange rate distributions: evidence from otc option markets. *Journal of international Money and Finance*, 17(1):117–160, 1998.

[4] M. R. Fengler. Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, 9(4):417–428, 2009.

[5] J. Jackwerth. Option implied risk-neutral distributions and implied binomial trees: A literature review. 1999.

[6] W.-N. Lai. Comparison of methods to estimate option implied risk-neutral densities. *Quantitative Finance*, 14(10):1839–1855, 2014.