Evaluation of the Black–Scholes equation via Crank–Nicolson and SINDy

Stephen Cornelius & Andy Mapes & Bill Xu

December 15, 2025

1 Introduction

The objective of this project was to empirically validate the Black–Scholes partial differential equation (PDE) using observed option market data. In particular, we sought to recover the governing PDE directly from data via Sparse Identification of Nonlinear Dynamics (SINDy), a framework designed to infer parsimonious differential equations from time-dependent observations. Such an approach offers a data-driven alternative to traditional model verification and provides a natural test of whether classical pricing equations are supported by empirical evidence.

A central challenge of this task is that the Black–Scholes PDE depends on local derivatives of the option price with respect to both the underlying asset price and time to maturity. Accurate estimation of these derivatives requires smooth, high-frequency observations of the option price surface. However, the empirical data available for this study consisted of option prices sampled at a monthly frequency, with additional noise arising from market microstructure effects, discrete strike grids, and volatility smiles. As a result, direct numerical differentiation produced unstable and unreliable derivative estimates.

To address these limitations, we explored a wide range of interpolation schemes and feature-based filtering strategies aimed at constructing a smoother approximation of the option price surface. These methods sought to reduce extraneous variability while preserving the essential dynamics implied by the Black–Scholes framework. Despite these efforts, the resulting surfaces remained too irregular for SINDy to consistently recover meaningful governing equations, highlighting fundamental identifiability constraints imposed by the data resolution.

To separate methodological limitations from data limitations, we supplemented the empirical analysis with synthetic option price data generated using a Crank–Nicolson finite difference scheme applied to the Black–Scholes PDE. In this controlled setting, SINDy successfully recovered the expected PDE structure, thereby validating the system identification pipeline itself. This contrast demonstrates that while data-driven PDE discovery is feasible in principle, its application to empirical option data is strongly constrained by sampling frequency and noise.

Overall, this study illustrates both the promise and the limitations of data-driven model discovery in quantitative finance, emphasizing the critical role of data quality and resolution in validating continuous-time pricing models.

2 Black–Scholes Derivation

The Black–Scholes equation links the stochastic dynamics of a traded asset to the deterministic pricing of derivatives written on that asset. Its derivation rests on a single central idea: by dynamically hedging a derivative with the underlying asset, all market risk can be eliminated over an infinitesimal time interval. The absence of arbitrage then forces the hedged portfolio to grow at the risk-free rate, yielding a partial differential equation for the option price.

2.1 Asset Dynamics and Option Dependence

We assume the underlying asset price S_t follows a geometric Brownian motion,

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{1}$$

where μ is the expected return, σ is the volatility, and W_t is a standard Wiener process. Let $V(S,t)$ denote the value of a derivative written on the asset. Since V depends on the random variable S_t, its evolution is itself stochastic.

2.2 Propagation of Randomness

Applying Itô's Lemma to $V(S,t)$ gives

$$dV = \left(V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + \mu S V_S \right) dt + \sigma S V_S \, dW_t. \tag{2}$$

The key observation is that both the asset and the derivative are driven by the same Brownian motion. This shared source of randomness makes it possible to construct a hedging strategy that removes risk.

2.3 Delta Hedging and Risk Elimination

Consider a portfolio that is short one derivative and long Δ shares of the underlying asset,

$$\Pi = -V + \Delta S. \tag{3}$$

Choosing

$$\Delta = V_S$$

eliminates the stochastic term proportional to dW_t, rendering the portfolio locally risk-free. This step is the conceptual core of the Black–Scholes argument: risk is neutralized through trading rather than priced directly.

2.4 No-Arbitrage and the Black–Scholes PDE

A risk-free portfolio must earn the risk-free interest rate r. Enforcing this condition yields the Black–Scholes partial differential equation,

$$\boxed{V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0.} \tag{4}$$

Notably, the expected return μ does not appear in the equation, reflecting the risk-neutral nature of derivative pricing.

2.5 Interpretation and Boundary Conditions

Introducing the option Greeks

$$\theta = V_t, \qquad \Delta = V_S, \qquad \Gamma = V_{SS},$$

the equation may be written as

$$\theta + \frac{1}{2}\sigma^2 S^2 \Gamma = r\big(V - S\Delta\big), \tag{5}$$

which equates the intrinsic evolution of the option to the required return on the delta-hedged position.

To price a specific derivative, the PDE is solved backward in time subject to a terminal condition at expiration $t = T$,

$$V(S, T) = \Phi(S),$$

where $\Phi(S)$ is the payoff function. This yields the unique arbitrage-free option price.

3 Crank-Nicolson Simulation

3.1 Theoretical Framework

The Crank–Nicolson method is a widely-used finite difference scheme for solving parabolic partial differential equations. It combines implicit and explicit updates to achieve second-order accuracy in both space and time while maintaining unconditional stability. This section describes the application of Crank–Nicolson to the Black–Scholes equation.

3.2 Discretization of the Black–Scholes PDE

We discretize the spatial domain $[0, S_{\max}]$ into $M + 1$ points with spacing ΔS, and the temporal domain $[0, T]$ into $N + 1$ points with spacing Δt. Let $V_{n,j} \approx V(t_n, S_j)$ denote the numerical approximation at time $t_n = n\Delta t$ and asset price $S_j = j\Delta S$.

The Black–Scholes PDE is rewritten in the form

$$V_t = \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV.$$

3.3 Crank–Nicolson Scheme

The Crank–Nicolson method evaluates the right-hand side at both time levels n and $n + 1$, averaging to achieve balanced accuracy:

$$\frac{V_{n+1,j} - V_{n,j}}{\Delta t} = \frac{1}{2}\left[\mathcal{L}(V_{n+1,j}) + \mathcal{L}(V_{n,j})\right],$$

where $\mathcal{L}$ denotes the spatial differential operator. Rearranging yields a system of linear equations at each time step:

$$(I - \tfrac{\Delta t}{2}\mathcal{L})\, V^{n+1} = (I + \tfrac{\Delta t}{2}\mathcal{L})\, V^{n}.$$

For the Black–Scholes equation, this becomes a tridiagonal system that can be solved efficiently using the Thomas algorithm.

3.4 Boundary Conditions and Initial Setup

At each time level, boundary conditions are enforced at $S = 0$ and $S = S_{\max}$. For a European call option, these are typically

$$V(t, 0) = 0, \qquad V(t, S_{\max}) \approx S_{\max} - Ke^{-r(T-t)},$$

where K is the strike price and T is the expiration time.

The scheme is initialized with the option payoff at maturity:

$$V(T, S) = \max(S - K, 0) \quad \text{for a call option.}$$

Time integration then proceeds backward from $t = T$ to $t = 0$, generating a complete spatio-temporal solution matrix.

3.5 Advantages for SINDy

The Crank–Nicolson scheme produces smooth, accurate numerical solutions with minimal oscillation. This smoothness is essential for SINDy, since derivative estimation is highly sensitive to noise and irregularity. By solving the known Black–Scholes PDE numerically, we generate synthetic data with known ground truth, enabling validation of the SINDy recovery process before application to empirical market data.

4 Construction of a Smooth Option Surface from Sparse Market Data

The empirical option data used in this study consists of prices sampled at a monthly frequency across a discrete grid of underlying prices and times to maturity. While sufficient for descriptive analysis, this data presents substantial challenges for derivative-based system identification methods such as SINDy. The Black–Scholes PDE depends on local derivatives of the option price surface, which are highly sensitive to noise, irregular sampling, and market microstructure effects.

To make the data more amenable to analysis, we explored interpolation and filtering strategies aimed at constructing a smoother approximation to the option price surface

$$V = V(S, t),$$

where S denotes the underlying price and t denotes time to maturity.

4.1 Interpolation in the Underlying Price

At each fixed maturity t_j, option prices are observed only at a discrete set of underlying prices $\{S_i\}$. To enable stable estimation of spatial derivatives V_S and V_{SS}, we interpolated across S using smooth basis representations. The approaches considered were:

- **Spline interpolation:** Cubic splines were fit to $V(S, t_j)$ for each maturity slice, ensuring continuity of the function and its first two derivatives with respect to S.
- **Local polynomial regression:** Low-order local polynomial fits were applied in S to reduce sensitivity to outliers and irregular spacing.
- **Implied-volatility–space interpolation:** In some cases, interpolation was performed on the implied volatility surface $\sigma_{\text{imp}}(S, t)$ rather than directly on prices, leveraging its empirically smoother structure.

Each method imposes different smoothness assumptions on the option surface, directly affecting the accuracy of higher-order derivative estimates.

4.2 Interpolation in Time to Maturity

The monthly sampling frequency is insufficient to directly approximate the time derivative V_t. To mitigate this limitation, intermediate surfaces were introduced between observed maturities using temporal interpolation. The strategies examined were:

- **Linear interpolation:** Option prices were interpolated linearly in time for fixed S, imposing minimal temporal structure.
- **Higher-order interpolation:** Polynomial and spline-based interpolants were also considered to produce smoother estimates of V_t, at the cost of introducing additional modeling assumptions.

These interpolations create a pseudo-continuous time dimension, enabling finite-difference approximations of temporal derivatives.

4.3 Feature-Based Filtering

Rather than smoothing the option surface directly, we applied feature-based filtering to reduce variance in variables not explicitly represented in the Black–Scholes PDE. Specifically, filtering was performed over combinations of

$$S - K, \quad K, \quad \sigma_{\text{imp}}, \quad t,$$

where K denotes the strike price.

By restricting attention to subsets of the data with limited variation in these features—such as narrow moneyness bands, strike ranges, implied volatility windows, or maturity intervals—we aimed to suppress cross-sectional heterogeneity arising from volatility smiles, term structure effects, and discrete quoting behavior. This conditioning reduces extraneous variability while preserving the raw structure of the data.

4.4 Modeling Implications

Interpolation and filtering introduce implicit assumptions about smoothness and local behavior of the option surface. While necessary for derivative-based analysis, these regularization steps constrain the extent to which governing equations can be recovered from empirical data. In practice, aggressive regularization suppresses genuine market structure, while insufficient regularization leads to unstable derivative estimates.

Accordingly, the constructed surfaces should be viewed as regularized approximations suitable for exploratory PDE identification rather than exact representations of the underlying market dynamics.

5 SINDy Method

5.1 Introduction and Motivation

In this section we introduce the *Sparse Identification of Nonlinear Dynamics* (SINDy) framework and describe how to validate our numerical solutions of the Black–Scholes equation. SINDy is a data-driven method for discovering governing differential equations directly from time-resolved measurements, under the assumption that the true dynamics can be represented by a small number of active terms in a larger candidate library of functions (for example,).

In our setting, a Crank–Nicolson scheme is used to compute approximate option values $V(t, S)$ on a discrete grid of asset prices and times to maturity. The resulting spatio-temporal data set serves as input to SINDy, which attempts to recover the underlying partial differential equation from the data alone. By comparing the identified model to the classical Black–Scholes equation, we are able to obtain a data-driven consistency check on both the numerical scheme and the model assumptions.

5.2 Overview of the SINDy Framework

SINDy provides a systematic approach for uncovering the governing equations of a dynamical system directly from data. The central idea is to represent the unknown dynamics in terms of a library of candidate functions and to determine a parsimonious model by selecting only the few terms that actively contribute to the evolution of the system This sparsity assumption reflects the structure of many physical and financial models, where the underlying equations typically involve only a small subset of all possible functional combinations.

In the context of partial differential equations (PDEs), SINDy is extended to the PDE-FIND methodology introduced in where spatio-temporal data are used to estimate both temporal and spatial derivatives. A function library consisting of candidate terms such as u, u_x, u_{xx}, and nonlinear products is constructed, and the algorithm identifies the subset of terms that best explains the observed dynamics.

5.3 Mathematical Foundations of SINDy

SINDy is built on the assumption that dynamical systems—including those governed by partial differential equations—has simple underlying structure. Despite the large number of admissible functional forms, the true governing equation typically contains only a small subset of all possible terms. The goal of SINDy is to exploit this sparsity to recover the governing equation directly from data

5.3.1 Representation of Dynamics Using a Function Library

Consider a dynamical system with state variable $x(t) \in \mathbb{R}^d$ governed by an unknown differential equation

$$\dot{x}(t) = f(x(t)).$$

SINDy approximates $f(x)$ by expressing it as a linear combination of its candidate nonlinear functions. To do so, one constructs a library $\Theta(x)$ consisting of functions such as constants, polynomial terms, trig functions, or some other nonlinearities:

$$\Theta(x) = \left[1,\ x,\ x^2,\ x^3,\ \sin(x),\ \ldots\right].$$

The dynamics are then assumed to satisfy

$$\dot{x}(t) \approx \Theta(x(t))\,\Xi,$$

where $\Xi \in \mathbb{R}^{K \times d}$ is a matrix of coefficients. Here, SINDy seeks a *sparse* coefficient matrix: most rows of Ξ are zero.

5.3.2 Formulation as a Sparse Regression Problem

Given measurement data $\{x(t_j)\}_{j=1}^{N}$, one computes numerical estimates of the time derivative $\dot{x}(t_j)$. Evaluating the library at all data samples yields the matrix equation

$$\dot{X} = \Theta(X)\,\Xi,$$

where $\dot{X} \in \mathbb{R}^{N\times d}$ and $\Theta(X) \in \mathbb{R}^{N\times K}$. Recovering the model becomes a sparse regression problem: identify the few active columns of $\Theta(X)$ that best explain the observed dynamics.

A common approach is *sequential thresholded least squares* (STLSQ), introduced in method alternates between:

1. solving a standard least-squares regression, and
2. thresholding (setting to zero) coefficients whose magnitudes fall below a prescribed sparsity level.

This procedure is analogous to performing an ℓ_0-type model selection step but avoids the combinatorial complexity of exhaustive search. Connections to LASSO and other sparse regression techniques further justify the stability and parsimony properties of the method.

5.3.3 Extension to Partial Differential Equations (PDE-FIND)

To identify partial differential equations, SINDy is extended to incorporate spatial derivatives. Let $u(t,x)$ denote the solution of an unknown PDE. The PDE-FIND approach introduced in estimates temporal and spatial derivatives from a spatio-temporal data set and constructs a function library of candidate terms such as

$$u, \quad u_x, \quad u_{xx}, \quad u^2, \quad u\,u_x, \quad x\,u_x, \quad \text{etc.}$$

The governing equation is assumed to have the form

$$u_t = \Theta(u, u_x, u_{xx}, \ldots)\,\Xi,$$

and sparse regression is again used to determine the active terms.

5.3.4 Conditions for Successful Model Recovery

The accuracy of SINDy depends on several identifiability conditions:

- **Rich data:** the solution must sufficiently explore the dynamics so that active terms in Θ can be distinguished.
- **Derivative accuracy:** numerical differentiation errors should be small relative to the magnitudes of the terms in the PDE.
- **Low correlation in the function library:** highly collinear candidate terms make sparse regression unstable.
- **Correct library specification:** the true terms must be included in the candidate dictionary.
- **Appropriate sparsity thresholding:** prevents overfitting and guards against noise amplification.

When these conditions are met, SINDy provides a mathematically justified and computationally efficient framework for discovering interpretable governing equations from data.

5.4 Why SINDy is Well Suited for the Black–Scholes Equation

SINDy is well suited for our case because the Black–Scholes equation is sparse in its natural variables and can be expressed using a small number of linear terms involving V, V_S, and V_{SS}. By SINDy, our goal is not to derive the equation from first principles, but to demonstrate that the numerical solution generated by the Crank–Nicolson method is consistent with the equation identified by SINDy.

The Black–Scholes equation possesses a structure that makes it especially compatible with the SINDy framework. The PDE is linear in the option value V and involves only a small number of active terms:

$$V_t = -\frac{1}{2}\sigma^2 S^2 V_{SS} - rSV_S + rV.$$

Although financial models can be highly nonlinear, the Black–Scholes equation depends only on three fundamental components: V, V_S, and V_{SS}, along with multiplicative factors involving the asset price S. This sparsity matches one of the core assumptions of SINDy: that the governing equation can be represented using only a few elements from a larger library of candidate terms.

Moreover, the solution $V(t, S)$ of the Black–Scholes equation is smooth in both variables under standard boundary and payoff conditions, ensuring that numerical derivatives such as V_S, V_{SS}, and V_t can be estimated with sufficient accuracy. This smoothness is an important assumption for the PDE-FIND extension of SINDy since derivative estimation is typically the most sensitive part of the identification process.

Crucially, the terms in the Black–Scholes PDE can be expressed using a simple candidate library of the form

$$\{ V,\ V_S,\ V_{SS},\ SV_S,\ S^2V_{SS} \},$$

all of which fall within the class of polynomial and product terms that SINDy handles efficiently. Therefore, SINDy can be expected to identify the correct combination of active terms and recover coefficients that correspond to the parameters r and σ.

From a practical perspective, SINDy provides a data-driven mechanism for verifying that the numerical solution generated by the Crank–Nicolson scheme is consistent with the theoretical Black–Scholes model. Instead of assuming the PDE, we attempt to rediscover it directly from the computed solution surface. Agreement between the identified PDE and the classical equation serves as a validation of both the numerical method and the suitability of the Black–Scholes assumptions for the given input data.

5.5 Application of SINDy to Crank–Nicolson Data

To apply the SINDy framework to the Black–Scholes equation, we will convert the numerical solution produced by the Crank–Nicolson method into a form suitable for sparse regression. This section summarizes the steps required to construct the data matrices, compute numerical derivatives, assemble the function library, and perform the regression.

5.5.1 Discrete Data Generated by the Crank–Nicolson Scheme

Let the asset price domain $[0, S_{\max}]$ be discretized into $M + 1$ grid points

$$S_0,\ S_1,\ \ldots,\ S_M,$$

and let the time interval $[0, T]$ be discretized into $N + 1$ time levels

$$t_0,\ t_1,\ \ldots,\ t_N.$$

The Crank–Nicolson scheme produces numerical approximations of the option price at each grid point, yielding a matrix

$$V \in \mathbb{R}^{(N+1)\times(M+1)},$$

where

$$V_{n,j} \approx V(t_n, S_j).$$

Each row corresponds to a fixed time level and each column corresponds to a fixed asset price.

5.5.2 Numerical Derivatives

To construct the regression matrix for SINDy, we must estimate the temporal and spatial derivatives of V.

Temporal derivative. For each spatial index j, we approximate V_t using a backward difference:

$$(V_t)_{n,j} \approx \frac{V_{n,j} - V_{n-1,j}}{\Delta t}, \qquad n = 1, \ldots, N.$$

This results in a derivative matrix

$$V_t \in \mathbb{R}^{N\times(M+1)}.$$

First spatial derivative. Using centered differences for interior points,

$$(V_S)_{n,j} \approx \frac{V_{n,j+1} - V_{n,j-1}}{2\,\Delta S}, \qquad j = 1, \ldots, M - 1.$$

At boundaries, one-sided differences are used. The resulting matrix lies in

$$V_S \in \mathbb{R}^{(N+1)\times(M+1)}.$$

Second spatial derivative. For interior points,

$$(V_{SS})_{n,j} \approx \frac{V_{n,j+1} - 2V_{n,j} + V_{n,j-1}}{(\Delta S)^2}.$$

Again, boundary handling is applied where necessary. The matrix satisfies

$$V_{SS} \in \mathbb{R}^{(N+1)\times(M+1)}.$$

5.5.3 Vectorization and Data Alignment

SINDy expects the data in vectorized form. For each quantity $Q \in \{V, V_S, V_{SS}, V_t\}$, we reshape the matrix into a column vector by stacking rows or columns:

$$q = \mathrm{vec}(Q) \in \mathbb{R}^L,$$

where

$$L = N \times (M+1)$$

for quantities defined only on time levels $1, \dots, N$, such as V_t, and

$$L = (N+1) \times (M+1)$$

for quantities defined at all time levels. To ensure consistency, we truncate all variables to use the same $L = N(M+1)$ data points corresponding to time levels where V_t is defined.

5.5.4 Construction of the Library Matrix

For each data point (t_n, S_j), we evaluate the candidate functions in the SINDy library. A typical library for the Black–Scholes equation includes

$$\Theta = \left[V,\ V_S,\ V_{SS},\ SV_S,\ S^2V_{SS},\ V_t \text{ (target)}\right].$$

To build the regression matrix, we vectorize each candidate term and form the column-stacked library matrix

$$\Theta(X) \in \mathbb{R}^{L\times K},$$

where each column corresponds to a candidate function and K is the number of functions in the library.

For the Black–Scholes example above, $K = 5$, and thus

$$\Theta(X) = \begin{bmatrix} | & | & | & | & | \\ \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 \\ | & | & | & | & | \end{bmatrix},$$

where $\theta_k = \mathrm{vec}(\text{candidate term}_k)$.

The target vector is

$$u_t = \mathrm{vec}(V_t) \in \mathbb{R}^L.$$

5.5.5 Sparse Regression

SINDy seeks a sparse coefficient vector $\Xi \in \mathbb{R}^K$ such that

$$u_t \approx \Theta(X)\,\Xi.$$

Sequential thresholded least squares (STLSQ) is used to determine which columns of $\Theta(X)$ contribute significantly to the dynamics:

$$\Xi^{(k+1)} = \mathcal{T}_\lambda\Big((\Theta^T\Theta)^{-1}\Theta^T u_t\Big),$$

where $\mathcal{T}_\lambda$ is a hard-thresholding operator that zeros out coefficients smaller than a prescribed level λ.

5.5.6 Interpretation of the Identified PDE

Once the sparse vector Ξ is obtained, the corresponding active terms reconstruct the PDE:

$$V_t = \xi_1 V + \xi_2 V_S + \xi_3 V_{SS} + \xi_4 (S V_S) + \xi_5 (S^2 V_{SS}).$$

If the Crank–Nicolson solution is consistent with the theoretical Black–Scholes model, SINDy should identify

$$\xi_4 \approx r, \qquad \xi_5 \approx \tfrac{1}{2}\sigma^2, \qquad \xi_1 \approx -r,$$

with all other coefficients close to zero.

This provides a fully data-driven verification of the numerical solution and the governing dynamics.

6 Conclusion

This study investigated the application of data-driven methods for discovering and validating the Black–Scholes partial differential equation from both synthetic and empirical option price data. We employed two complementary approaches: (i) numerical solution via the Crank–Nicolson finite difference scheme, and (ii) sparse system identification via the SINDy framework.

Our key findings are as follows:

1. **Synthetic Data Validation:** When applied to option price surfaces generated by solving the Black–Scholes PDE numerically with Crank–Nicolson, the SINDy algorithm successfully recovered the governing equation with high fidelity. The identified coefficients matched the theoretical parameters r and σ, demonstrating that the system identification pipeline is sound and that SINDy can reliably discover PDE structure from smooth, high-resolution spatio-temporal data.

2. **Empirical Data Limitations:** Application to real market option data proved substantially more challenging. Despite employing multiple interpolation schemes, smoothing filters, and feature-based conditioning strategies, the empirical option price surface remained too noisy and irregular to support consistent PDE recovery. The monthly sampling frequency and inherent market microstructure effects (volatility smiles, discrete strike grids, bid-ask spreads) limited the accuracy of derivative estimation, which is the most sensitive step in the SINDy procedure.

3. **Methodological Insights:** The contrast between successful recovery from synthetic data and poor recovery from market data highlights a fundamental distinction: data-driven PDE discovery is sensitive not only to algorithmic choices but critically to data quality, resolution, and smoothness. High-frequency, low-noise observations are essential for identifying continuous-time dynamics from discrete samples.

4. **Practical Implications:** For practitioners seeking to validate or discover pricing models in

quantitative finance, these results underscore the importance of data preprocessing and the limitations of inference from sparse, noisy observations. Model validation in finance may require either higher-frequency data acquisition or the integration of domain knowledge to regularize the inverse problem.

Future work could explore advanced smoothing techniques such as Gaussian process regression or kernel methods to construct smoother synthetic surfaces from market data, examine alternative sparse regression formulations robust to noise amplification, or investigate whether SINDy can identify modified or alternative pricing models when the classical Black–Scholes assumptions are relaxed. Additionally, extending this framework to higher-dimensional problems (e.g., multi-asset options) or time-varying parameters may reveal new challenges and opportunities in financial model discovery.