

Constrained Polynomial Likelihood

Caio Almeida, Ricardo Masini & Paul Schneider

To cite this article: Caio Almeida, Ricardo Masini & Paul Schneider (2025) Constrained Polynomial Likelihood, Journal of Business & Economic Statistics, 43:2, 482-493, DOI: [10.1080/07350015.2024.2394587](https://doi.org/10.1080/07350015.2024.2394587)

To link to this article: <https://doi.org/10.1080/07350015.2024.2394587>



Published online: 25 Oct 2024.



Submit your article to this journal [↗](#)



Article views: 256



View related articles [↗](#)



View Crossmark data [↗](#)



Constrained Polynomial Likelihood

Caio Almeida^a, Ricardo Masini^b, and Paul Schneider^c

^aDepartment of Economics and BCF, Princeton University, Princeton, NJ; ^bDepartment of Statistics, UC Davis, Davis, CA; ^cUSI Lugano and SFI, Lugano, Switzerland

ABSTRACT

We develop a nonnegative polynomial minimum-norm likelihood ratio (PLR) of two distributions of which only moments are known. The sample PLR converges to the unknown population PLR under mild conditions. The methodology allows for additional shape restrictions, as we illustrate with two empirical applications. The first develops a PLR for the unknown transition density of a jump-diffusion process, while the second extracts a positive density directly from option prices. In both cases, we show the importance of implementing the non-negativity restriction.

ARTICLE HISTORY

Received September 2022
Accepted July 2024

KEYWORDS

Likelihood ratio; Positive polynomial

1. Introduction

In Hilbert spaces, orthogonality and minimum-norm problems are tightly related. As a consequence, orthogonal polynomials have a prominent role in minimum-norm approximation of unknown likelihood ratios and numeric integration. However, extant approaches based on orthogonality only do not preserve important properties of the approximated objects. In this article, we develop projections of likelihood ratios onto positive polynomials, thus, *preserving positivity*, and if desired, additional structural constraints.

The literature considers approximations of likelihood ratios with orthogonal polynomials foremost due to the link between polynomials and the possibility of expressing moments of a distribution as expectations of polynomials (Aït-Sahalia 2002; Filipović, Mayerhofer, and Schneider 2013; Kato and Kuriki 2013; Renner and Schmedders 2015). While many other approaches exist in the literature to approximate likelihood ratios,¹ polynomials are a natural choice, as moments are known for many models, such as the large affine class (Duffie, Filipović, and Schachermayer 2003). Moreover, under mild technical conditions, polynomials generate weighted L^2 spaces that arise naturally when working with expectations and sample averages. In our article, we work with positive polynomials with the smallest modification to the extant expansions mentioned above.²

Our framework, rooted in Grenander (1981) sieve estimation, is built upon several steps. First, we consider the conventional projection of a likelihood ratio on polynomials. Through an optimization problem, we then obtain our polynomial minimum-norm likelihood ratio (PLR) as the sum of this element and the minimum-norm polynomial that guarantees pointwise positivity. We show that this optimization problem, if feasible, has the PLR as its unique solution. Subsequently, we prove consistency and rate of convergence of the PLR based on sample moments. Importantly, these coefficients of the PLR can be obtained rapidly as the solution of a conic optimization program, allowing for additional constraints that can modify the shape of the PLR.

The sieves literature nonparametrically approximates a function using a sequence of finite-dimensional spaces and corresponding basis functions, such as polynomials, splines, wavelets, and trigonometric functions, that grow with the sample size to asymptotically eliminate the bias in the estimation process (see Chen 2007). In contrast, we opt to keep the degree of the PLR polynomial fixed, approximating the pseudo-true likelihood with (potential) bias. Our motivation stems from the fact that preserving positivity is more important when a finite sample size caps the maximum degree of the PLR than in a limiting large sample/large polynomial degree environment, where the

CONTACT Caio Almeida  calmeida@princeton.edu  Department of Economics and BCF, Princeton University, Princeton, NJ 08544.

¹The literature on density approximations ranges from saddlepoint approximations (Aït-Sahalia and Yu 2006), small time expansions (Yu 2007), to simulation-based methods (Mijatović and Schneider 2010; Giesecke and Schwenkler 2018) to cite a small subset of the econometrics and statistics literature. In machine learning, starting from Berlinet and Thomas-Agnan (2004), a sequence of papers embeds distributions in Reproducing Kernel Hilbert Spaces (Song et al. 2009; Grünewälder et al. 2012; Park and Muandet 2020; Klebanov, Schuster, and Sullivan 2020), as well as likelihood ratios (Schuster et al. 2020). While some of the aforementioned papers preserve positivity and normalization of distributions, they do not provide the flexibility to include additional shape constraints.

²More generally, we operate within the generic Hilbert space problem

$$\underset{g \in \mathcal{H}}{\text{minimize}} \quad \|g\|, \text{ subject to } (g, f) \in C, f \in \mathcal{H}$$

where C is a convex set, and $\mathcal{H}$ is a Hilbert space along with its inner product $(\cdot, \cdot)$, encompassing function approximation, interpolation, and many other applications. In this article, we specialize in likelihood ratios with probabilistic models in mind.

estimated PLR would converge to its positive population counterpart.

A related literature beginning with Hansen and Scheinkman (2009) emphasizes the importance of positive eigenfunctions in asset pricing problems.³ Preserving positivity is crucial in such problems, as illustrated by Christensen (2017), who uses Hermite polynomial approximations to estimate a nonparametric decomposition of the Stochastic Discount Factor into transitory and permanent components. In this context, our methodology could be adopted to ensure a nonnegative approximation to the positive eigenfunctions that define the SDF decomposition in finite samples, where a projection onto polynomials may not guarantee positivity preservation.

We illustrate the usefulness of PLR with two simple applications. In the first, we expand the likelihood ratio of the transition density of a continuous-time jump-diffusion process with respect to a Gamma density.⁴ Unlike the conventional orthogonal polynomial approach, which yields approximations that are negative near zero, our PLR method ensures non-negativity across its entire range. This feature makes our approach particularly suitable for applications that depend on the structural properties of probability measures such as normalization and positivity, including accept-reject sampling and Bayesian modeling.

In our second application, we investigate how to extract probability densities directly from option prices. This is particularly interesting in the context of pricing exotic and/or over-the-counter derivatives based on the probability density extracted from observed vanilla option prices. We illustrate this procedure in a controlled environment using a bivariate continuous-time jump-diffusion process from Duffie, Pan, and Singleton (2000). Our experiments reveal how the non-negativity of the PLR method crucially translates into meaningful estimates of positive probability densities, rather than projections of densities that do not satisfy the basic defining properties of probability measures.

The article is structured as follows. In Section 2.1, we introduce the necessary notation. The traditional approach to minimum-norm expansions is discussed in Section 2.2 for reference. In Section 2.3, we propose modifications to the standard program to ensure positivity. Section 2.4 presents the uniqueness, consistency, and rate of convergence of the polynomial coefficients defining the PLR. Applications are dis-

cussed in Section 3.1 (density approximations) and Section 3.2 (option pricing). Section 4 provides the concluding remarks. The Appendix reviews sum-of-squares (SOS) polynomials and includes the proofs for the theoretical results.

2. Constrained Polynomial Likelihood

2.1. Set-Up and Notation

Denoting by $\mathcal{P}$ a probability measure on $D \subseteq \mathbb{R}^d$, let $L_{\mathcal{P}}^2$ be the equivalence class of functions $f : D \mapsto \mathbb{R}$ such that $\int_D f^2(t) d\mathcal{P}(t) < \infty$ with inner product

$$(f, g) = \int_D f(t)g(t) d\mathcal{P}(t), f, g \in L_{\mathcal{P}}^2, t := [t_1, \dots, t_d], \quad (1)$$

where we allow in subsequent sections to extend the inner product to act element-wise on matrices.

Denote by $\mathbb{R}[t]$ the ring of square-integrable polynomials on $\mathbb{R}^d$, and by $\mathbb{R}[t]_n$ the subset of polynomials $\xi \in \mathbb{R}[t]$ with $\deg(\xi) \leq n$. Denote further by $P_{\mathcal{P},n}$ the set of polynomials $\mathbb{R}[t]_n$ endowed with inner product (1), a finite-dimensional Hilbert space.

We denote the standard canonical monomial basis of $P_{\mathcal{P},n}$ by

$$\tau_n(t) := [1, t_1, \dots, t_d, t_1^2, t_1 t_2, \dots, t_d^2, \dots, t_1^n, t_1^{n-1} t_2, \dots, t_d^n]^\top, \quad (2)$$

as well as multi-index powers by $t^\beta := t_1^{\beta_1} \dots t_d^{\beta_d}$ for $\beta \in \mathbb{N}_0^d$, where the length of the multi-index is $|\beta| = \beta_1 + \dots + \beta_d$.

There are $\binom{n+d}{d}$ elements in the monomial basis, and we denote by $\alpha_0, \dots, \alpha_N$, with $N = \binom{n+d}{d} - 1$, the multi-indices corresponding to their order of appearance in (2). For example, the second element is $t^{\alpha_1} = t_1^{\alpha_{11}} \dots t_d^{\alpha_{1d}} = t_1^1 t_2^0 \dots t_d^0 = t_1$. Any polynomial $\xi_n \in P_{\mathcal{P},n}$ can be written as $\xi_n = \mathbf{x}^\top \tau_n$, where $\mathbf{x} \in \mathbb{R}^{N+1}$ is a coefficient vector.

With projections on positive polynomials in mind, we parameterize a particularly tractable cone $M_n(D) \subset P_{\mathcal{P},n}$ of positive polynomials as follows.

Definition 2.1 (Positive s.o.s. polynomials). Let $S_+^q, q \in \mathbb{N}$, denote the set of symmetric positive semidefinite (p.s.d.) matrices of dimension q . We denote by $M_n(D)$ the set of positive polynomials $\xi_n(t) = \mathbf{x}^\top \tau_n(t) \geq 0, \forall t \in D$ for matrices $V \in S_+^{q_V}$ and $W \in S_+^{q_W}$ on a case-by-case basis.

1. Case $D = \mathbb{R}^d$: $\xi_n(t) = \tau_{n/2}^\top(t) V \tau_{n/2}(t)$,
2. Case $D = \mathbb{R}_+$: $\xi_n(t) = \tau_{n/2}^\top(t) V \tau_{n/2}(t) + t \left(\tau_{n/2-1}^\top(t) W \tau_{n/2-1}(t) \right)$,
3. Case $D = [a, b]$: $\xi_n(t) = \tau_{n/2}^\top(t) V \tau_{n/2}(t) + (b - t)(t - a) \left(\tau_{n/2-1}^\top(t) W \tau_{n/2-1}(t) \right)$.

By collecting coefficients in the relation above, we can identify a linear relation $\mathbf{x} = T(V, W)$ between the coefficient vector $\mathbf{x}$ of a positive polynomial in $M_n(D)$ to elements of the symmetric positive semidefinite (p.s.d.) matrices V and W , whose dimensions are $q_V = \binom{n/2+d}{d}$ and $q_W = 0$ in the first case, and $q_V = \binom{n/2+1}{1}$ and $q_W = \binom{n/2}{1}$ in the latter two cases. As

³Hansen and Scheinkman (2009) proposes an operator approach to study the long-run risk-return tradeoff properties of Stochastic Discount Factors (SDFs) in Markovian environments by solving a Perron-Frobenius eigenfunction problem. The SDF is decomposed into a permanent (martingale) and a transitory component (see Alvarez and Jermann 2005). Ross (2015) uses the Perron-Frobenius approach to recover investors' beliefs, that is, subjective probabilities under rational expectations. Borovička, Hansen, and Scheinkman (2016) discuss the conditions under which the probability recovered with Perron-Frobenius coincides with the actual probabilities of investors under rational expectations. Christensen (2017) proposes using Sieves to empirically identify this SDF decomposition, establishing the asymptotic theory for Perron-Frobenius eigenfunction and eigenvalue estimators. A non-exhaustive list of additional papers on the SDF decomposition and on recovery probabilities include Chabi-Yo and Bakshi (2012), Chabi-Yo, Bakshi, and Gao (2018), Backus, Chernov, and Zin (2014), Qin and Linetsky (2016), and Qin and Linetsky (2017).

⁴When expanding a transition density, it is necessary to choose an appropriate auxiliary density function on the same support. For positive support, the Gamma distribution is suited (see Filipović, Mayerhofer, and Schneider (2013)).

outlined in [Appendix A, Definition 2.1](#) completely characterizes the set of positive polynomials in the univariate case. However, this characterization is only sufficient for the multivariate case, where positive polynomials exist that cannot be represented as sums of squares. The benefit of the above sum-of-squares (s.o.s.) parameterization is that positive polynomials can be effectively modeled using semidefinite programming. Nonetheless, it is important to note that without further modifications, the linear transformation T is not generally a bijection, as illustrated by the following example.

Example 1 (Selection matrices). Suppose that $n = 4$, and $D = \mathbb{R}$. Then positivity requires

$$\begin{aligned}\tau_4^\top(t)\mathbf{x} &= \tau_2^\top(t) \mathbf{V} \tau_2(t) = \text{tr}(\mathbf{V} \tau_2(t) \tau_2^\top(t)) \\ &= \text{tr} \left(\begin{bmatrix} V_{11} & V_{21} & V_{31} \\ V_{21} & V_{22} & V_{32} \\ V_{31} & V_{32} & V_{33} \end{bmatrix} \begin{bmatrix} 1 & t & t^2 \\ t & t^2 & t^3 \\ t^2 & t^3 & t^4 \end{bmatrix} \right).\end{aligned}$$

Collecting coefficients, we can thus express the selection matrices as

$$\begin{aligned}V_0 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, V_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\ V_2 &= \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, V_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, V_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},\end{aligned}$$

such that $[x_0, x_1, x_2, x_3, x_4] = T^\top(\mathbf{V}) = [\text{tr}(\mathbf{V}V_i)]_{i=0}^4 = [V_{11}, 2V_{21}, V_{22} + 2V_{31}, 2V_{32}, V_{33}]$.

Before presenting the full type of problems we intend to solve on $M_n(D)$, we briefly review projections of Radon-Nikodym derivatives on polynomials through equality-constrained minimum-norm problems. While this problem is standard, and its solution is well-known, it helps inform the optimization program for the shape-constrained version in the subsequent section.

2.2. Minimum-Norm Likelihood Ratio Projection

For projections in $L_{\mathbb{P}}^2$, with or without shape constraints, the Gram matrix is defined by

$$\begin{aligned}\mathbf{H}_n &:= (\tau_n, \tau_n^\top) = \begin{bmatrix} \mu_{0,0}^{\mathcal{P}} & \mu_{0,1}^{\mathcal{P}} & \cdots & \mu_{0,N}^{\mathcal{P}} \\ \mu_{1,0}^{\mathcal{P}} & \mu_{1,1}^{\mathcal{P}} & \cdots & \mu_{1,N}^{\mathcal{P}} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{N,0}^{\mathcal{P}} & \mu_{N,1}^{\mathcal{P}} & \cdots & \mu_{N,N}^{\mathcal{P}} \end{bmatrix}, \text{ with} \\ \mu_{ij}^{\mathcal{P}} &:= \int_D t^{\alpha_i + \alpha_j} d\mathcal{P}(t)\end{aligned}\quad (3)$$

Our first assumption regarding this symmetric and positive semidefinite matrix greatly simplifies further analysis. It is mild, easily checked, and ensures in particular that the inverse of the Gram matrix $\mathbf{H}_n^{-1}$ is readily available, greatly facilitating the analysis below.

Assumption 2.2 (Rank of $\mathbf{H}_n$). Gram matrix $\mathbf{H}_n$'s smallest eigenvalue $\lambda_{\min}(\mathbf{H}_n) > 0$.

We are particularly interested in projections on polynomials of the Radon-Nikodym derivative of a probability measure $\mathcal{Q}$ on D that is absolutely continuous with respect to $\mathcal{P}$. Finding the minimum-norm of such projection on polynomials, where $\mathcal{Q}$ is only known from its moment vector $\boldsymbol{\mu}_{\mathcal{Q}} := [\mu_0^{\mathcal{Q}}, \dots, \mu_N^{\mathcal{Q}}]^\top$ with $\mu_i^{\mathcal{Q}} := \int_D t^{\alpha_i} d\mathcal{Q}(t)$, solves the equivalent problems (Luenberger 1997)

$$\begin{aligned}\underset{\xi_n \in P_{\mathcal{P},n}}{\text{minimize}} \quad & \left\| \frac{d\mathcal{Q}}{d\mathcal{P}} - \xi_n \right\|^2 \quad \Leftrightarrow \quad \underset{\mathbf{x} \in \mathbb{R}^N}{\text{minimize}} \quad \left\| \mathbf{H}_n^{1/2} \mathbf{x} \right\|_2^2, \text{ subject to} \\ & \mathbf{H}_n \mathbf{x} = \boldsymbol{\mu}_{\mathcal{Q}},\end{aligned}\quad (4)$$

where $\mathbf{H}_n^{1/2}$ denotes the Cholesky factor of $\mathbf{H}_n$. The well-known solution to problem (4) is

$$\xi_n^*(t) := \boldsymbol{\mu}_{\mathcal{Q}}^\top \mathbf{H}_n^{-1} \tau_n(t). \quad (5)$$

On $P_{\mathcal{P},n}$, the projection of a Radon-Nikodym derivative on polynomials evidently corresponds to moment-matching. However, there is no mechanism that ensures the structural properties of Radon-Nikodym derivatives—normalization and positivity—in the projection.

To project on *positive* polynomials, with $M_n(D)$ not being a linear space, using it as a hypothesis space requires additional considerations. In the next section, we, therefore, introduce an optimization program akin to (4), where membership to $M_n(D)$ is implemented as a conic constraint. The formulation as an optimization program additionally allows to generalize the equality constraints in [Section 2.2](#) to inequalities and to add further restrictions, as demanded by each application.

2.3. Constrained Minimum-Norm Likelihood Ratio Projection

In this section, we introduce a minimization program that accommodates the constraints in (4) as a special case and includes additionally the conic constraint $\xi_n \in M_n(D)$ from [Definition 2.1](#).

$$\begin{aligned}\underset{\xi_n \in M_n(D)}{\text{minimize}} \quad & \|\xi_n\|^2, \text{ subject to} \\ & (\xi_n, f_i) = c_i, \quad i = 1, \dots, m, \text{ and } (\xi_n, g_j) \leq d_j, \quad j = 1, \dots, l,\end{aligned}\quad (6)$$

where $f_1, \dots, f_m, g_1, \dots, g_l \in L_{\mathcal{P}}^2$ are linearly independent functions generating the subspace $K := [f_1, \dots, f_m, g_1, \dots, g_l]$. While in general $K \subsetneq P_{\mathcal{P},n}$, we nevertheless routinely set $f_1 = 1$ so that with $(\xi_n, f_1) = 1$, ξ_n represents a normalized and positive likelihood ratio with respect to a probability distribution $\mathcal{P}$. Together with the additional constraints, it is a constrained polynomial likelihood ratio (PLR). Note that the constraint $\xi_n \in M_n(D)$ ensures pointwise positivity. This is important for evaluating the PLR with any argument while maintaining positivity, for instance, in out-of-sample exercises or simulations.

It is easy to construct an example for which program (6) is infeasible. The constraint $(\xi_n, 1) = -1$ can not be satisfied by any nonnegative polynomial ξ_n , with $\mathcal{P}$ being a probability measure. Aside from these cases, the feasibility of equation (6) becomes more likely as the order n of ξ_n increases, due to the higher number of coefficients involved. In light of this

observation, denoting by $\tilde{K} = \{\mathbf{x} \in \mathbb{R}^{N+1} : \{(\xi_n, f_i) = c_i\}_{i=1,\dots,m}, \text{ and } \{(\xi_n, g_j) \leq d_j\}_{j=1,\dots,l}\}$ the feasible set of program (6), we state our second assertion.

Assumption 2.3 (Feasibility). There exists $n \in \mathbb{N}$ such that $\tilde{K}$ has non-empty interior.

In the next section, we discuss a generic solution algorithm to the optimization problem (6) and show the consistency of the solution in the case when the inner product (1) is estimated from sample averages.

2.4. Properties of Solutions on $P_{\mathcal{P},n}$ and $M_n(D)$

The present section discusses the properties of the solution to program (6).

2.4.1. Uniqueness of the Primal Optimization Problem

As a first step, we argue the uniqueness of the solution to program (6), if the constraints are feasible, as a standard result in finite-dimensional convex optimization. To develop a solution, we first exploit that we work in a finite-dimensional Hilbert space of polynomials, which allows us to express the functional inequalities as matrix equations. The (in)equalities in (6) are linear in ξ_n , and defining $\mathbf{c} := [c_1, \dots, c_m]^\top$, $\mathbf{d} := [d_1, \dots, d_l]^\top$, $\mathbf{f} := [f_1, \dots, f_m]^\top$, and $\mathbf{g} := [g_1, \dots, g_l]^\top$, we can express them as

$$(\xi_n, \mathbf{f}) = \mathbf{F}_n \mathbf{x} = \mathbf{c}, \text{ and } (\xi_n, \mathbf{g}) = \mathbf{G}_n \mathbf{x} \leq \mathbf{d}, \quad (7)$$

where $\leq$ represents generalized, conic, inequality, and the inner product is naturally applied element-wise, such that $\mathbf{F}_n \in \mathbb{R}^{m \times N+1}$ and $\mathbf{G}_n \in \mathbb{R}^{l \times N+1}$.

Proposition 2.4. If Assumptions 2.2 and 2.3 hold, problem (6) has a unique solution from the primal problem

$$\begin{aligned} & \underset{\mathbf{x} \in \mathbb{R}^{N+1}}{\text{minimize}} && \frac{1}{2} \|\mathbf{H}_n^{1/2} \mathbf{x}\|_2^2, \\ & \text{subject to} && \mathbf{F}_n \mathbf{x} = \mathbf{c}, \mathbf{G}_n \mathbf{x} \leq \mathbf{d}, \\ & && \mathbf{x} = T(\mathbf{V}, \mathbf{W}); \mathbf{V} \in S_+^{mv}, \mathbf{W} \in S_+^{mw}, \end{aligned} \quad (8)$$

where the linear map $T : S_+^{mv} \times S_+^{mw} \mapsto \mathbb{R}^{N+1}$,

$$T(\mathbf{V}, \mathbf{W}) := [\text{tr}(\mathbf{V}_0 \mathbf{V}) + \text{tr}(\mathbf{W}_0 \mathbf{W}), \dots, \text{tr}(\mathbf{V}_N \mathbf{V}) + \text{tr}(\mathbf{W}_N \mathbf{W})]^\top, \quad (9)$$

for fixed symmetric selection matrices $\mathbf{V}_0, \dots, \mathbf{V}_N$ and $\mathbf{W}_0, \dots, \mathbf{W}_N$ of the same dimension as $\mathbf{V}$ and $\mathbf{W}$, respectively. The corresponding dual reads

$$\begin{aligned} & \underset{\eta \in \mathbb{R}^m, \boldsymbol{\varepsilon} \in \mathbb{R}_+^l, \mathbf{v} \in \mathbb{R}^{N+1}}{\text{maximize}} && -\frac{1}{2} \|\mathbf{H}_n^{-1/2} (\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon})\|_2^2 - \boldsymbol{\varepsilon}^\top \mathbf{d} + \boldsymbol{\eta}^\top \mathbf{c} \\ & \text{subject to} && \sum_{i=0}^N v_i \mathbf{V}_i \geq 0, \sum_{i=0}^N v_i \mathbf{W}_i \geq 0, \end{aligned} \quad (10)$$

where the primal and dual solutions are linked as

$$\mathbf{x}_0 = \mathbf{H}_n^{-1} (\mathbf{F}_n^\top \boldsymbol{\eta}_0 + \mathbf{v}_0 - \mathbf{G}_n^\top \boldsymbol{\varepsilon}_0). \quad (11)$$

It is noteworthy that the dual formulation is strictly convex in the sum $\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon}$, rather than in the Lagrange multipliers $\boldsymbol{\eta}, \mathbf{v}, \boldsymbol{\varepsilon}$ corresponding to the equality, positivity, and inequality constraints, separately. This observation becomes important for deriving results about the behavior of optimization problems (8) and (10) when the population matrices $\mathbf{F}_n, \mathbf{G}_n, \mathbf{H}_n$, and possibly also $\mathbf{c}$ and $\mathbf{d}$ are estimated from data. Conveniently, the ingredients of the program (6) depend entirely and exclusively on expectations induced by $\mathcal{P}$. This feature is useful in applications where the weight function $\mathcal{P}$ in (1) is unknown, but, for instance, its moments can be estimated.

Furthermore, note that if K is generated by polynomials, we have a unique orthogonal decomposition of the solution $\xi_n(t) = \mathbf{x}_0^\top \boldsymbol{\tau}_n(t)$ of the program (8) into

$$\xi_n = \xi_n^* \oplus \xi_n^\circ, \quad (12)$$

where ξ_n^* is the minimum-norm polynomial on K , and ξ_n° is the minimum-norm polynomial lifting the solution into $M_n(D)$.

2.4.2. Consistency

In this section, we are concerned with the behavior of the optimal coefficients when the inner product can only be evaluated from the sample distribution induced by identically distributed draws $x_1, \dots, x_k$ from $\mathcal{P}$. We derive two types of asymptotic results, the first being a consistency result on the coefficients $\mathbf{x}$, and the second being a nonstandard result accounting for the non-uniqueness of the arguments that minimize the dual (10).

We first define

$$\langle f, g \rangle_k := \frac{1}{k} \sum_{i=1}^k f(x_i) g(x_i) \text{ for } f, g \in L_{\mathcal{P}}^2. \quad (13)$$

With the empirical inner product, we can define $\hat{\mathbf{H}}_{n,k} := \langle \boldsymbol{\tau}_n, \boldsymbol{\tau}_n^\top \rangle_k$, and analogously $\hat{\mathbf{F}}_{n,k} := \langle \xi_n, \mathbf{f} \rangle_k$ and $\hat{\mathbf{G}}_{n,k} := \langle \xi_n, \mathbf{g} \rangle_k$. Denoting by $\tilde{K}_k$ the feasible set induced by $\hat{\mathbf{F}}_{n,k}, \hat{\mathbf{G}}_{n,k}$, and $\hat{\mathbf{H}}_{n,k}$, we are in the position to introduce the sample versions of problem (6), of its corresponding mixed-conic primal and dual matrix versions (8) and (10), and to establish consistency of the estimators in these problems. Let $\mathbf{x}_0$ represent the coefficients of the polynomial ξ_n , that are the solution of problem (6) with restrictions based on the population moments $\{(\xi_n, f_i), (\xi_n, g_j)\}$, and $\hat{\mathbf{x}}_k$ represent the corresponding coefficients from the sample moments $\{(\xi_n, f_i)_k, (\xi_n, g_j)_k\}$, $i = 1, \dots, m; j = 1, \dots, l$, with analogous notation for the dual problem.

To provide the two types of consistency results, we introduce the population and sample objective functions

$$\begin{aligned} \mathcal{R}_k(\mathbf{x}) &:= -\mathbf{x}^\top \hat{\mathbf{H}}_{n,k} \mathbf{x} \quad \text{and} \\ \mathcal{R}_0(\mathbf{x}) &:= -\mathbf{x}^\top \mathbf{H}_n \mathbf{x} \end{aligned} \text{ of the primal problem.}$$

To establish consistency of the estimators, as well as statements about their behavior with increasing k , we make a (mild) assumption concerning the matrices depending on estimated moments.

Assumption 2.5 (Rate of Convergence). Let $\hat{\mathbf{H}}_{n,k}, \hat{\mathbf{F}}_{n,k}, \hat{\mathbf{G}}_{n,k}, \hat{\mathbf{c}}_k$, and $\hat{\mathbf{d}}_k$ be estimators of $\mathbf{H}_n, \mathbf{F}_n, \mathbf{G}_n, \mathbf{c}$, and $\mathbf{d}$, respectively. Set $\mathbf{s}^\top := [(\text{vec } \mathbf{H}_n)^\top, (\text{vec } \mathbf{F}_n)^\top, (\text{vec } \mathbf{G}_n)^\top, \mathbf{c}^\top, \mathbf{d}^\top]$ and define $\hat{\mathbf{s}}$ in the same manner. Suppose that $\|\hat{\mathbf{s}} - \mathbf{s}\|_2 = O_{\mathcal{P}}(r_k)$ for some nonnegative sequence $r_k \rightarrow 0$ as $k \rightarrow \infty$.

In most cases, the estimators $\hat{H}_{n,k}$, $\hat{F}_{n,k}$, $\hat{G}_{n,k}$, $\hat{c}_k$, and $\hat{d}_k$ will be sample averages. Since we already assume square integrability, the Weak Law of Large numbers will ensure [Assumption 2.5](#) is valid with $r_k = 1/\sqrt{k}$. Uncorrelatedness of the realizations suffices, but several other forms of weak dependencies could be accommodated, such as mixingales and mixing sequences, provided that the serial correlation can be controlled. Note that [Assumptions 2.2](#) and [2.5](#) together also ensure that the symmetric p.s.d. matrix $\hat{H}_{n,k}$ will be positive definite with high probability for k large enough.

Proposition 2.6 (Consistency). Under [Assumptions 2.2](#), [2.3](#), and [2.5](#), $\hat{x}_k \xrightarrow{\mathcal{P}} x_0$ as $k \rightarrow \infty$ for fixed n .

[Proposition 2.6](#) is based on mild assumptions and describes the asymptotic behavior of the estimator of the optimal coefficients x_0 . However, it does not account for the p.s.d. matrices Λ_V and Λ_W , which appear as Lagrange multipliers of the s.o.s. cones (see the Lagrangian in (B1)).

To describe the joint behavior of the estimated primal and dual parameters, let $\theta^\top := [x^\top, (\text{vec } V)^\top, (\text{vec } W)^\top, \eta^\top, \varepsilon^\top, v^\top, (\text{vec } \Lambda_V)^\top, (\text{vec } \Lambda_W)^\top] \in \Theta$ where Θ is the Cartesian product of the domain of each component of θ defined above, and $\Psi : \Theta \rightarrow \mathbb{R}^e$ is given by

$$\Psi(\theta) := \begin{bmatrix} H_n x - F_n^\top \eta + G_n^\top \varepsilon - v \\ \text{vec}(\Lambda_V - \sum_{i=0}^n v_i V_i) \\ \text{vec}(\Lambda_W - \sum_{i=0}^n v_i W_i) \\ F_n x - c \\ \varepsilon \odot (G_n x - d) \\ x - T(V, W) \\ \text{tr}(\Lambda_V V) \\ \text{tr}(\Lambda_W W) \end{bmatrix},$$

where $\odot$ denotes the Hadamard product. The solutions to problems (8)–(10) can be characterized by $\Theta_0 := \{\theta \in \Theta : \Psi(\theta) = 0\}$ due to strong duality. Note that Ψ is closely related to the KKT conditions, in the sense that the first three equations of $\Psi(\theta) = 0$ are stationary conditions, and the remaining are non-slackness-type, and primal and dual feasibility conditions.

Since Θ_0 is a level set of a continuous function, it is closed. It is also convex since it is the solution set of a convex optimization problem, but it is not necessarily bounded. However, for optimal $x_0 \in \mathbb{R}^{N+1}$, we can regularize such that all other entries are bounded by a (large) constant $C > 0$. Take, for instance, [Example 1](#), where $x_2 = V_{22} + 2V_{31}$. We can impose a large upper bound C for the magnitudes of V_{22} and V_{31} . Following this principle, we consider only solutions in the compact $\Theta_* := \Theta_0 \cap \bar{B}_C(0)$, where $\bar{B}_C(0)$ denotes the closed ball in Θ with radius C .

Let $\hat{\Psi}$ be the sample analogue of the function Ψ , that is, $\hat{\Psi}$ is the function Ψ with H_n, F_n, G_n, c , and d replaced by estimators $\hat{H}_{n,k}, \hat{F}_{n,k}, \hat{G}_{n,k}, \hat{c}_k$, and $\hat{d}_k$, respectively, and define $\hat{\Theta}_k := \{\theta \in \bar{B}_C(0) : \|\hat{\Psi}(\theta)\| \leq \delta_k\}$ for some nonnegative sequence such that $\delta_k \rightarrow 0$ as $k \rightarrow \infty$, keeping in mind [Assumption 2.5](#) for the estimators.

Due to the partial identification of all coefficients (except for x that is unique), both the estimator $\hat{\Theta}_k$ and the parameter of interest Θ_* are sets (subsets of Θ). Therefore, we need to

measure the distance between sets and, to that end, we use the Hausdorff distance defined by a pair of sets of a metric space $A, B \subseteq (M, d)$ as

$$d_H(A, B) := \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\}.$$

Proposition 2.7 (Set rate of convergence). Under [Assumptions 2.3](#) and [2.5](#), for every $\epsilon > 0$, there are constants $C_\epsilon, M_\epsilon > 0$ such that for every $k \in \mathbb{N}$,

$$\mathcal{P}(d_H(\hat{\Theta}_k, \Theta_*) \leq M_\epsilon r_k) \geq 1 - \epsilon,$$

provided that $\delta_k = Cr_k$ for $C \geq C_\epsilon$ where d_H is the Hausdorff distance.

[Proposition 2.7](#) above complements [Proposition 2.6](#) by establishing the rate of convergence of $\hat{\Theta}_k$ to Θ_* as $k \rightarrow \infty$ in the Hausdorff distance, which coincides with the usual Euclidean distance when both $\hat{\Theta}_k$ and Θ_* are singletons. In general, we expect $r_k = 1/\sqrt{k}$, and in that case, $\hat{\Theta}_k$ is $\sqrt{k}$ -consistent, which implies that $\hat{x}_k - x_0 = O_{\mathcal{P}}(1/\sqrt{k})$. From the proof of [Proposition 2.7](#), the following related result follows immediately.

Corollary 1 (Asymptotic coverage). Under [Assumptions 2.3](#) and [2.5](#) we have $\mathcal{P}(\Theta_* \subseteq \hat{\Theta}_k) \rightarrow 1$ provided that $\delta_k = O(r_k \log k)$ as $k \rightarrow \infty$.

[Corollary 1](#) provides a useful guarantee for the estimated coefficients of the model. It ensures that the $(r_k \log k)$ -level set of $\hat{\Psi}$ contains the true set Θ_* with high probability as k increases. Once again, we expect $r_k = 1/\sqrt{k}$, consequently $\hat{\Theta}_k$ is a (nearly parametric rate) confidence set for Θ_* with a confidence level approaching one as the sample size increases. Note that the log term is arbitrary in the sense that any strictly increasing function could replace it without changing the result.

3. Applications

In this section, we present two applications of the PLR developed in this article. The first, in [Section 3.1](#), approximates the unknown transition density of a jump-diffusion process whose moments are known. The second, in [Section 3.2](#), extracts densities from option prices. For both applications, both $\mathcal{P}$, as well as $\mathcal{Q}$ possess Lebesgue densities p and q , respectively.

3.1. Density Expansions

As a natural application and as a continuation of [Section 2.2](#), we investigate here approximations of an unknown distribution $\mathcal{Q}$ given moments $\mu_{\mathcal{Q}} := \int_D \tau_m d\mathcal{Q}$, where we allow for $m \leq n$ (rather than $m = n$ as in [Section 2.2](#)). To indicate the dependence of ξ_n on the number of monomials spanning K , we denote the optimal n -order polynomial likelihood expansion with moment constraints up to order m by $\xi_{n(m)}^{(m)}$ or more concisely $\xi_n^{(m)}$, where n is chosen sufficiently high according to [Assumption 2.3](#). We begin by showing that $\xi_n^{(m)}$ converges in $L^2_{\mathcal{P}}$ to its unknown counterpart $\frac{d\mathcal{Q}}{d\mathcal{P}}$. For this asymptotic anal-

ysis formulated with population moments, we need a technical assumption.⁵

Assumption 3.1 (Polynomial basis). The ring of polynomials $\mathbb{R}[t]$ is a basis of $L^2_{\mathcal{P}}$.

Assumption 3.1 is justified in particular for compact state spaces, as well as unbounded state spaces with the tails of p decaying sufficiently quickly (Filipović, Mayerhofer, and Schneider 2013).

Theorem 3.2. If $\frac{dQ}{dP} \in L^2_{\mathcal{P}}$, and **Assumptions 2.2, 2.3**, and **3.1** hold, the nonnegative expansion $\xi_n^{(m)}$ converges in $L^2_{\mathcal{P}}$,

$$\lim_{m \rightarrow \infty} \left\| \frac{dQ}{dP} - \xi_n^{(m)} \right\| = 0. \quad (14)$$

To put our PLR to work, we now confront our density approximation approach with the one proposed in Filipović, Mayerhofer, and Schneider (2013, FMS). For this purpose, we consider the basic affine jump diffusion (BAJD) solving the stochastic differential equation

$$dY_t = (\kappa\theta - \kappa Y_t) dt + \sigma \sqrt{Y_t} dW_t + dL_t, \quad (15)$$

where W_t is a Brownian motion, the intensity of the compound Poisson process L is $\lambda \geq 0$, and the expected jump size of the exponentially distributed jumps is $\nu \geq 0$. The transition density of $Y_{\Delta} | Y_0$, $\Delta > 0$ is not known in closed form, but its existence is assured if $2\kappa\theta > \sigma^2$ (Filipović, Mayerhofer, and Schneider 2013, Theorem 2) on its domain $D = \mathbb{R}_+$. Note that since the BAJD is a polynomial process, its conditional moments $\mu_i := \mathbb{E}^Q[Y_{\Delta}^i | Y_0]$ are known in closed form for $\Delta > 0$, however, even though the transition density is not. This process, as well as some of its variations, have been adopted to model the dynamics of stock index prices (Bates 2000) and to represent the intensity of the first jump to default when pricing CDS options and other related credit derivatives (Brigo and Mercurio 2006).

In the following, we develop a likelihood ratio to approximate the unknown distribution Q of the BAJD by tilting a Gamma distribution $\Gamma(1 + \tilde{p}, 1)$ with density $p(x; \tilde{p}) = \frac{e^{-x} x^{\tilde{p}}}{\Gamma(1 + \tilde{p})}$, where $\tilde{p} = \mu_1^2 / (\mu_2 - \mu_1^2) - 1$ and Γ denotes the Gamma function. This auxiliary density, here representing $\mathcal{P}$, pertains to the scaled random variable $\tilde{Y}_{\Delta} := Y_{\Delta} \cdot c$, with $c := \frac{\mu_1}{\mu_2 - \mu_1^2}$ such that the object of interest $q^{(m)}(y) = \xi_n^{(m)}(\tilde{y}) p(\tilde{y}) \cdot c$ accounts for this change of variables, where the notation $\xi_n^{(m)}(\tilde{y})$ emphasizes the dependence of the PLR on the scaled random variable.

We match moments $\mu_0, \dots, \mu_5$, where the choice of $m = 5$ is arbitrary, leaving sufficiently many free coefficients ($n = 8$) to obtain a PLR. The corresponding program reads

$$\text{minimize } \|\xi_8\|, \text{ subject to } (\xi_8, t^i) = \mu_i, i = 0, \dots, 5,$$

and we denote by $\xi_8^{(5)}$ the solution to the program above. We confront our PLR approximation with the one proposed in FMS using the same weight function $\tilde{p}$, and matching the same moments $\mu_0, \dots, \mu_5$. The corresponding program reads

$$\text{minimize } \|\eta_5\|, \text{ subject to } (\eta_5, t^i) = \mu_i, i = 0, \dots, 5,$$

and we denote by $\eta^{(5)}$ its solution. FMS solves the program using the projection theorem via orthogonal polynomials. Ours and their solutions are easily related: As $\xi_n^{(m)} = \eta^{(m)}$ for every non-negative integer m by construction, we have that $\xi_n^{(m)} = \eta^{(m)} + \xi_n^{o(m)}$ (see decomposition (12)). From elementary arguments, the PLR is farther away in the $L^2_{\mathcal{P}}$ norm from the true likelihood ratio than the FMS one for every m . This is the price of imposing non-negativity, which we will see in our next illustration might be worth paying, depending on the application.

We perform the comparison with the parameters $\kappa\theta = 0.05, \kappa = 1, \sigma = 0.2, \lambda = 1, \nu = 0.05, \gamma_0 = 0.05$, roughly describing the dynamics of a stochastic equity volatility process, with the true transition density obtained numerically from Fourier inversion using the exponentially-affine characteristic function of the BAJD.

Figure 1 shows that the FMS density becomes negative close to zero; the more negative, the smaller Δ , as Panels 1a and 1b indicate. Thus, in any application demanding positivity, such as derivatives pricing (Aït-Sahalia 1999), likelihood ratio tests, or MCMC sampling, it is imperative to use the PLR proposed in this article rather than the FMS projections without modifications.

3.2. Distributions from Options

Deviations from the positivity of probability distributions may have sizable economic ramifications. In this section, we explore PLR in the context of option pricing and the concept of no-arbitrage. The absence of arbitrage, or free lunch, implies that any payoff $X_{t+\Delta}$ has forward price $F_t(X_{t+\Delta})$, where F_t is a linear operator that has a representation in terms of an expectation⁶

$$F_t(X_{t+\Delta}) = \mathbb{E}^Q[X_{t+\Delta} | \mathcal{F}_t], \quad (16)$$

where Q is a so-called forward measure. As before, we assume in this section that Q is absolutely continuous with respect to the Lebesgue measure with density q (and likewise $\mathcal{P}$ with density p). Note that here $\mathcal{P}$ represents an auxiliary possibly misspecified parametric probability measure approximating Q .

A call option written on the payoff $X_{t+\Delta}$ with strike price K pays $(X_{t+\Delta} - K)^+$, while for puts the payoff is $(K - X_{t+\Delta})^+$. Given a set $\mathcal{K} := \{K_1, \dots, K_{\Omega}\}$ of strike prices along with bid-ask spreads, or for simplicity, mid quotes of option forward prices $P_{t,\Delta}(K_1), \dots, P_{t,\Delta}(K_{\omega}), C_{t,\Delta}(K_{\omega+1}), \dots, C_{t,\Delta}(K_{\Omega})$ of out-of-the-money options (meaning that $K_1 < \dots < K_{\omega} \leq F_t(X_{t+\Delta}) \leq K_{\omega+1} < \dots < K_{\Omega}$), our goal is to find a density q that is compatible with these prices. If strike prices were continuously quoted, q could be computed from second derivatives of option prices (Breedon and Litzenberger 1978),

⁵It is important to keep in mind that for fixed n the framework from **Sections 2.2** and **2.3** is applicable also to distributions with finite moments, but no moment-generating function, like the log-normal distribution. However, for the asymptotic analysis on the polynomial dimension n , as shown in this section, only $\mathcal{P}$ random variables with a moment-generating function are permissible.

⁶We conduct our study in the forward market. Alternatively, in the spot market, we would assume zero interest rates for simplicity.

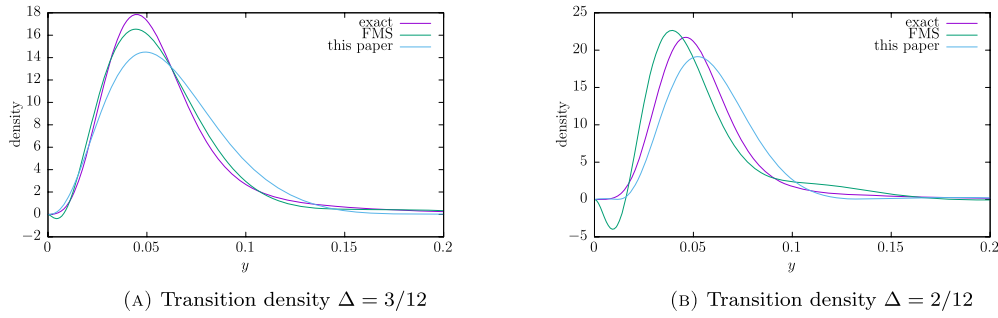


Figure 1. Comparison of density approximations. This figure shows transition density approximations of the Basic Affine Jump-Diffusion (BAJD) solving the stochastic differential equation $dY_t = (\kappa\theta - \kappa Y_t)dt + \sigma\sqrt{Y_t}dW_t + dL_t$. The parameters used are $\kappa\theta = 0.05, \kappa = 1, \sigma = 0.2, \lambda = 1, \nu = 0.05, y_0 = 0.05$. Panels a and b show the approximation for a time span of $\Delta = 3/12$, and $\Delta = 2/12$, respectively. For both pictures, the exact density is obtained from Fourier inversion, while the density approximation FMS facilitates the approach from Filipović, Mayerhofer, and Schneider (2013).

but the presence of only a discrete number of observations in $\mathcal{K}$ requires additional considerations to obtain q .

Option cross sections are informative financial instruments, as from Carr and Madan (2001), for any twice-differentiable almost-everywhere function f , its forward-neutral expectation is a particular option portfolio,

$$\begin{aligned} \mathbb{E}^Q[f(X_{t+\Delta})|\mathcal{F}_t] &= \int_0^{F_t(X_{t+\Delta})} f''(K)P_{t,\Delta}(K)dK \\ &\quad + \int_{F_t(X_{t+\Delta})}^{\infty} f''(K)C_{t,\Delta}(K)dK, \end{aligned} \quad (17)$$

so that, for instance, forward-neutral moments can be approximated from observed option cross-sections. We will use formula (17) to compute $\mu_i^Q := \mathbb{E}^Q[(\log X_{t+\Delta})^i|\mathcal{F}_t]$ and $\sigma^Q := \sqrt{\mu_2^Q - \mu_1^Q}$ to define

$$\begin{aligned} Y &:= \frac{\log \frac{X_{t+\Delta}}{F_t(X_{t+\Delta})} - \mu_1^Q}{\sigma^Q}, \text{ and similarly} \\ w &:= \frac{\log \frac{K}{F_t(X_{t+\Delta})} - \mu_1^Q}{\sigma^Q}. \end{aligned} \quad (18)$$

From this change of variables, forward option prices can be written as

$$\begin{aligned} P_{t,\Delta}(K) &= \mathbb{E}^Q[(K - X_{t+\Delta})^+] \\ &= e^{\mu_1^Q} F_t(X_{t+\Delta}) \int_{-\infty}^{\log w} (e^{\sigma^Q w} - e^{\sigma^Q y}) \tilde{q}(y) dy, \text{ and} \\ C_{t,\Delta}(K) &= \mathbb{E}^Q[(X_{t+\Delta} - K)^+] \\ &= e^{\mu_1^Q} F_t(X_{t+\Delta}) \int_{\log w}^{\infty} (e^{\sigma^Q y} - e^{\sigma^Q w}) \tilde{q}(y) dy, \end{aligned} \quad (19)$$

where the density $\tilde{q}$ pertains to Y . From $\tilde{q}$, we can recover the original density q , the object of interest, from

$$q(x) = \frac{1}{x\sigma^Q} \tilde{q}\left(\frac{\log \frac{x}{F_t(X_{t+\Delta})} - \mu_1^Q}{\sigma^Q}\right).$$

The above centering and scaling facilitate the object of interest

$$\xi_n = \frac{\tilde{q}}{p} \in L^2_{\mathcal{P}},$$

such that the options are priced and martingale and normalization restrictions are satisfied. In terms of the auxiliary density p , the martingale and density normalization restrictions that should be satisfied under the forward measure read

$$\begin{aligned} 1 &= e^{\mu_1^Q} \int_{-\infty}^{\infty} e^{\sigma^Q y} \xi_n(y) p(y) dy, \text{ and} \\ 1 &= \int_{-\infty}^{\infty} \xi_n(y) p(y) dy, \text{ respectively.} \end{aligned}$$

In the context of fat-tailed forward-neutral $\tilde{q}$ densities, the choice of the auxiliary density p becomes particularly important. Lee (2004) investigates the tail of the distribution of forward-neutral log returns, showing that it is exponential and, therefore, ticker than the Gaussian tail. Thus, in the context of option pricing, the absolute continuity of p with respect to $\tilde{q}$ puts strong restrictions on the choice of p , excluding, in particular, the normal distribution, as its tails are too thin.

From these considerations, we solve the problem using the generalized hyperbolic distribution (Barndorff-Nielsen and Halgreen 1977) as auxiliary density p . The generalized hyperbolic distribution has four parameters. We choose the parameters to fit the first four moments of Y , such that $\mu_1^Q = 0, \mu_2^Q = 1, \mu_3^Q, \mu_4^Q$. These moments can be computed either from a model under consideration or using the formula (17) with observed option prices. Importantly, the generalized hyperbolic distribution features exponential tails, such that the approximation in $L^2_{\mathcal{P}}$ is meaningful. In particular, for the parameterizations used below, the tails of the generalized hyperbolic distribution satisfy the $L^2_{\mathcal{P}}$ integrability condition.

Specializing program (6) for the task of extracting a density from option prices, we

$$\begin{aligned} &\text{minimize } \|\xi_n\|^2, \\ &\quad \xi_n \in M_n(D) \\ &\text{subject to } (\xi_n, 1) = 1, (\xi_n, e^{\mu_1^Q} e^{\sigma^Q \cdot Y}) = 1, \\ &1 - \varepsilon \leq \left(\xi_n, \frac{e^{\mu_1^Q} F_t(X_{t+\Delta}) (e^{\sigma^Q \cdot w_j} - e^{\sigma^Q \cdot Y})^+}{P_{\Delta}(K_j)} \right) \\ &\leq 1 + \varepsilon, j = 1, \dots, \omega \\ &1 - \varepsilon \leq \left(\xi_n, \frac{e^{\mu_1^Q} F_t(X_{t+\Delta}) (e^{\sigma^Q \cdot Y} - e^{\sigma^Q \cdot w_j})^+}{C_{\Delta}(K_j)} \right) \\ &\leq 1 + \varepsilon, j = \omega + 1, \dots, \Omega \end{aligned} \quad (20)$$

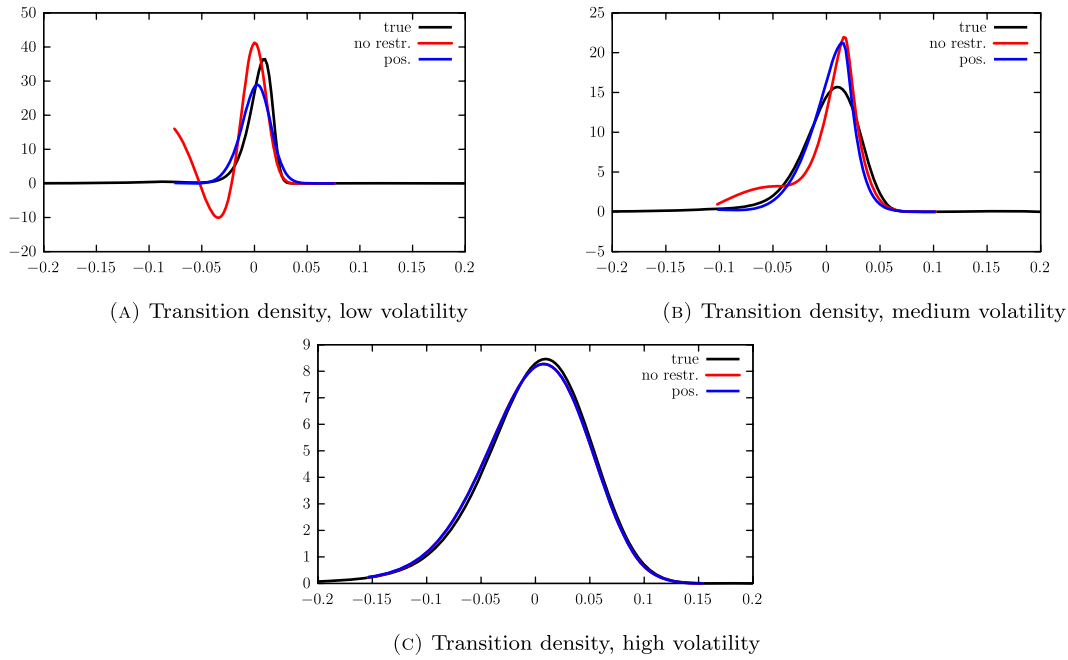


Figure 2. Densities implied from options. The figure shows the true density, the density solving program (20) without (no restr.), and with (pos.) positivity restriction for the Duffie, Pan, and Singleton (2000) double-jump model estimated from options. Option data is generated for low ($\sqrt{V_t} = 0.02$), medium ($\sqrt{V_t} = 0.087$), and high ($\sqrt{V_t} = 0.17$) volatility at valuation time. Both types of PLR are computed with polynomial degree $n = 8$. In the horizontal axis, we observe log returns $\log \frac{X_{t+\Delta}}{X_t}$ of the underlying asset.

Tolerance $\varepsilon > 0$ bounds the maximal relative option pricing error from above. Lu and Qu (2021) perform a similar task with polynomials by minimizing the squared pricing errors. Differently from their approach, we obtain a *positive* density q , that furthermore satisfies the martingale restriction $(\xi_n, e^{\mu_1^Q} e^{\sigma^Q \cdot Y}) = 1$. Additionally, the generalized hyperbolic distribution features exponential tails, while Lu and Qu (2021) employs the normal distribution, for which the approximating expansion diverges asymptotically.

We illustrate our approach using the Duffie, Pan, and Singleton (2000) double-jump model for $\log X_t$, and its stochastic variance process V_t with joint dynamics

$$d \begin{pmatrix} \log X_t \\ V_t \end{pmatrix} = \begin{pmatrix} r - \bar{\lambda} \bar{\mu} - \frac{1}{2} V_t \\ \kappa_v (\bar{v} - V_t) \end{pmatrix} dt + \sqrt{V_t} \begin{pmatrix} \rho & \sqrt{1 - \rho^2} \\ \sigma_v & 0 \end{pmatrix} d \begin{pmatrix} W_{1,t}^Q \\ W_{2,t}^Q \end{pmatrix} + dJ_t^Q, \quad (21)$$

where $W_{1,t}^Q, W_{2,t}^Q$ are Q Brownian motions, and J^Q is a pure jump process in $\mathbb{R}^2$ with constant mean arrival rate $\bar{\lambda}$, and $\bar{\mu}$ is the mean jump size in the $\log X_t$ direction under Q .

For fixed Δ , the characteristic function of $\log X_{t+\Delta}, V_{t+\Delta} | \log X_t, V_t$ is readily available in closed form, such that option prices can be obtained via the dampened transform as proposed by Lee (2004). Using the estimated parameters from Duffie, Pan, and Singleton (2000, Table I), we consider three scenarios for the value of V_t : low ($\sqrt{V_t} = 0.02$), medium ($\sqrt{V_t} = 0.087$, the value quoted in the original paper), and high ($\sqrt{V_t} = 0.17$), and compute the resulting densities with positivity restriction, and without. For both the positive and the unrestricted PLR, we find the smallest possible tolerance ε that makes program (20) feasible through trial and error. This process takes a fraction of a second.

Figure 2 shows transition densities computed from Fourier inversion of model-implied option prices (true), without (no restr.), and with (pos.) positivity restriction imposed when solving program (20). Panels 2a and 2c show reasonably accurate q densities for the low and high volatility scenarios, while the medium volatility in Panel 2b scenario is estimated too peaked. The reason for this is the auxiliary generalized hyperbolic distribution p , which, despite matching the first four moments of the original distribution $\tilde{q}$, features a peak that an order eight polynomial can not tilt sufficiently. Note also on Panel 2a that the unrestricted transition density achieves large negative values for negative returns close to zero in the low volatility scenario.

Figure 3 illustrates the stark differences between the unrestricted PLR and the positive one induced by the positivity restriction. It also shows that the tilting is moderate in magnitude, thanks to the minimum-norm objective function. The unrestricted ξ_n , that is, the one with no positivity restriction, takes negative values for all scenarios but the high-volatility one and presents an excessive oscillatory behavior when volatility is low or medium.

Figure 4 shows Black-Scholes implied volatilities from the true model, as well as the polynomial approximations. From Panel 4a, it becomes apparent that the low volatility scenario is the most challenging. In this scenario, OTM call options have low prices dominating the relative pricing error on both polynomial likelihood approximations, which forces them to sacrifice part of their ability to price OTM puts. The consequence for the implied volatility curve is that both models underprice puts (negative log-moneyness in the picture), with the unconstrained model having a smaller bias and an intermediary region of log-moneyness where it overprices puts. The positive model trades precision to avoid the excessive oscillatory behavior produced by the unconstrained PLR on the implied

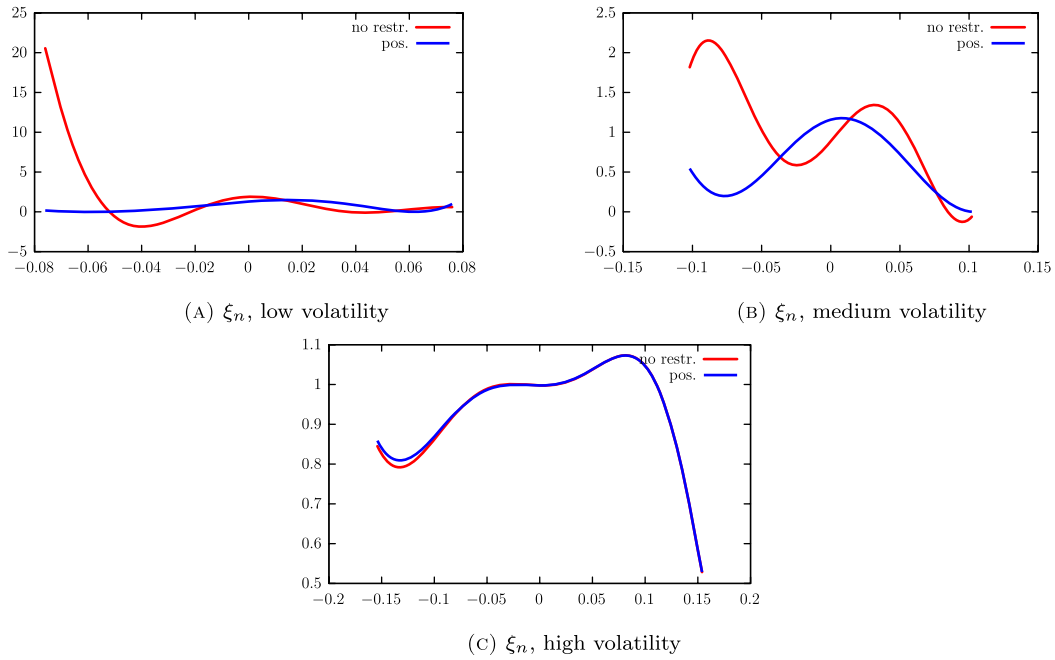


Figure 3. Polynomial tilts. The figure shows the resulting polynomial ξ_n from the solution of the program (20) without (no restr.), and with (pos.) positivity restriction for the Duffie, Pan, and Singleton (2000) double-jump model estimated from options. Option data is generated for low ($\sqrt{V_t} = 0.02$), medium ($\sqrt{V_t} = 0.087$), and high ($\sqrt{V_t} = 0.17$) volatility at valuation time. Both types of PLR are computed with polynomial degree $n = 8$. In the horizontal axis, we observe log returns $\log \frac{X_{t+\Delta}}{X_t}$ of the underlying asset.

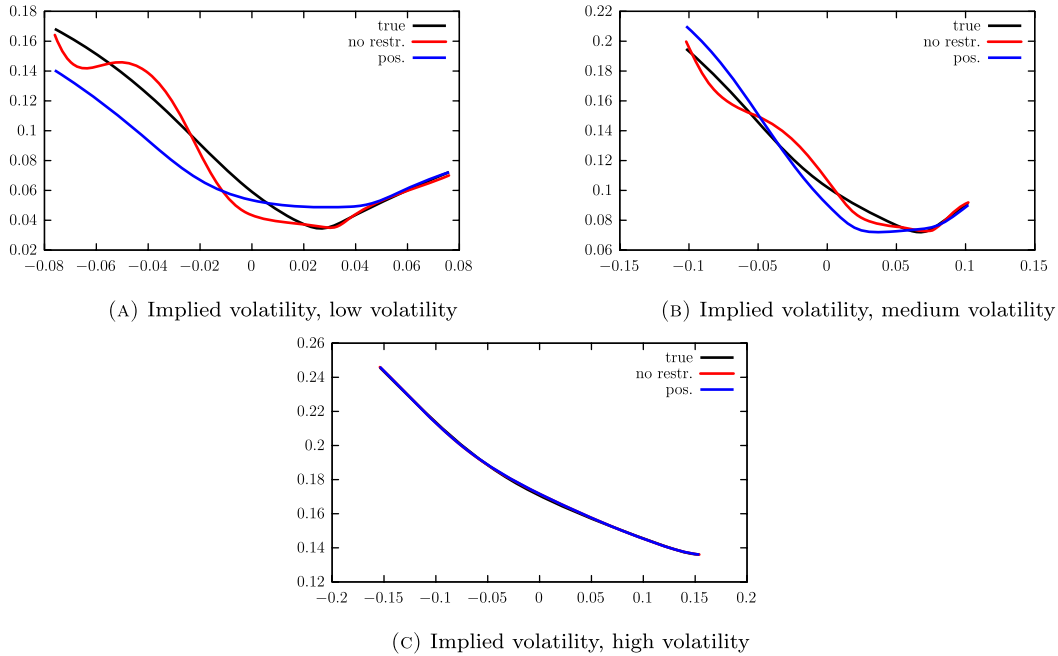


Figure 4. Comparison of implied volatilities. The figure shows Black-Scholes implied volatilities of the option prices generated by the true distribution (true), the density solving program (20) without (no restr.), and with (pos.) positivity restriction for the Duffie, Pan, and Singleton (2000) double-jump model estimated from options as a function of log moneyness. Option data is generated for low ($\sqrt{V_t} = 0.02$), medium ($\sqrt{V_t} = 0.087$), and high ($\sqrt{V_t} = 0.17$) volatility at valuation time. Both types of PLR are computed with polynomial degree $n = 8$. In the horizontal axis, we observe log moneyness $\log \frac{K}{X_t}$.

volatility curve, which comes from the oscillatory behavior of its estimated transition density noted in Figure 2. This excessive oscillatory behavior of implied volatilities also appears in the medium volatility scenario for the unconstrained model, contrasting with well-behaved implied volatilities under positivity constraint. Finally, note that both approximations work quite well in the high volatility scenario.

4. Conclusion

We develop projections of likelihood ratios onto polynomials that preserve positivity. We term them positive polynomial likelihood ratio (PLR). PLR can accommodate shape restrictions, are fast and robust to compute as solutions to conic programs and come with asymptotic theory for their use with sample moments.

We illustrate PLR with two applications. The first is an approximation of the unknown transition density of a jump-diffusion process. The second is a martingale density estimated from the observation of a finite number of option prices only. In both cases, the virtue of positivity of the PLR becomes evident through the positivity of the resulting distributions, which, without restriction, becomes negative and with excessive oscillatory behavior.

PLR lend themselves to nonparametric structurally constrained settings, such as, for instance, Christensen (2017) discount factor decomposition. We leave this and related applications to future research.

Appendix A: Positive Polynomials

In this section, we review results in the literature about positive polynomials. In the univariate case, we have

Proposition A1. (Schmüdgen 2017) For any positive integer n ,

1. $D = \mathbb{R} : \Omega_n := M_{2n} = \{f^2(t) + g^2(t) : f, g \in \mathbb{R}[t]_n\}$,
2. $D = \mathbb{R}_+ : M_{2n} = \{f(t) + tg(t) : f \in \Omega_n, g \in \Omega_{n-1}\}$,
3. $D = \mathbb{R}_+ : M_{2n+1} = \{f(t) + tg(t) : f, g \in \Omega_n\}$,
4. $D = [a, b] : M_{2n} = \{f(t) + (b-t)(t-a)g(t) : f \in \Omega_n, g \in \Omega_{n-1}\}$,
5. $D = [a, b] : M_{2n+1} = \{(b-t)f(t) + (t-a)g(t) : f, g \in \Omega_n\}$.

The set of positive polynomials on any other (continuous) state space can be extracted from Proposition A1 from a change of variables. For instance, $D = [a, \infty)$ can be obtained from parameterization (2) through the change of variables $t \mapsto t - a$ for $p \in M_{2n}$ on $D = \mathbb{R}_+$.

In the multivariate case, nonnegative polynomials exist that are not s.o.s. Since we merely want to assure non-negativity and a s.o.s. polynomial is certainly nonnegative, it is sufficient for our purpose to work with s.o.s. polynomials. Any such polynomial has a representation as a quadratic form (the proof is in Schmüdgen (2017) for Proposition 13.2).

Proposition A2. A polynomial $\xi_{2n} \in \mathbb{R}[t]_{2n}$ is s.o.s. if and only if $\xi_{2n} = \tau_n^\top V \tau_n$, with $V \succeq 0$.

Note that we do not make a distinction between different supports in the multivariate case. Statements A1 and A2 lead to our parameterization on Definition 2.1 of the cone of positive polynomials.

Appendix B: Proofs

Proposition 2.4

Proof. For the minimization, it is convenient to use one-half of the squared norm as an objective function rather than the norm itself. This does not change the result since the norm is nonnegative, and we can write

$$\|\xi_n\|^2/2 = \frac{1}{2}(\xi_n, \xi_n) = \frac{1}{2} \int_D \mathbf{x}^\top \tau_n(t) \tau_n^\top(t) \mathbf{x} d\mathcal{P}(t) = \frac{1}{2} \mathbf{x}^\top \mathbf{H}_n \mathbf{x}.$$

Together with the constraints in coordinate form and Proposition A2, this yields primal (8) that can be solved as a mixed conic semidefinite program.⁷ Since with Assumption 2.2, the objective function is strictly convex, and the constraint set is an intersection of closed convex sets, the solution with Assumption 2.3 is unique in $\mathbf{x}$, and strong duality obtains. With strong duality at hand, we next consider the dual form

to (8). Furthermore, the cone of the symmetric positive semidefinite matrix is self-dual. From these observations, we can write the Lagrangian of system (8) as

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\eta}, \mathbf{v}, \boldsymbol{\varepsilon}, \boldsymbol{\Lambda}_V, \boldsymbol{\Lambda}_W) := & \frac{1}{2} \mathbf{x}^\top \mathbf{H}_n \mathbf{x} - \boldsymbol{\eta}^\top (\mathbf{F}_n \mathbf{x} - \mathbf{c}) \\ & - \mathbf{v}^\top (\mathbf{x} - T(\mathbf{V}, \mathbf{W})) \\ & + \boldsymbol{\varepsilon}^\top (\mathbf{G}_n \mathbf{x} - \mathbf{d}) - \text{tr}(\boldsymbol{\Lambda}_V \mathbf{V}) - \text{tr}(\boldsymbol{\Lambda}_W \mathbf{W}) \end{aligned} \quad (\text{B1})$$

for $\boldsymbol{\eta} \in \mathbb{R}^m, \boldsymbol{\varepsilon} \in \mathbb{R}_+^l, \mathbf{v} \in \mathbb{R}^{N+1}$, as well as p.s.d. matrices $\boldsymbol{\Lambda}_V, \boldsymbol{\Lambda}_W$ of the same dimension as $\mathbf{V}$ and $\mathbf{W}$, respectively. From the first-order condition (on $\mathbf{x}$) we can then deduce $\mathbf{x}_0^\top \mathbf{H}_n - \boldsymbol{\eta}_0^\top \mathbf{F}_n - \mathbf{v}_0^\top + \boldsymbol{\varepsilon}_0^\top \mathbf{G}_n = 0$, and with $\mathbf{H}_n$ invertible from Assumption 2.2, any optimal solution must satisfy (11). The Lagrangian is linear-quadratic in $\mathbf{V}$ and $\mathbf{W}$, and taking matrix derivatives, we therefore have the conditions $\mathbf{v}_0 \mathbf{V}_0 + \dots + \mathbf{v}_N \mathbf{V}_N = \boldsymbol{\Lambda}_V \succeq 0$, and analogously $\mathbf{v}_0 \mathbf{W}_0 + \dots + \mathbf{v}_N \mathbf{W}_N = \boldsymbol{\Lambda}_W \succeq 0$. Plugging these relations into the Lagrangian yields

$$\begin{aligned} g(\boldsymbol{\eta}, \mathbf{v}, \boldsymbol{\varepsilon}) = & \frac{1}{2} (\boldsymbol{\eta}^\top \mathbf{F}_n + \mathbf{v}^\top - \boldsymbol{\varepsilon}^\top \mathbf{G}_n) \mathbf{H}_n^{-1} (\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon}) \\ & - \boldsymbol{\eta}^\top (\mathbf{F}_n \mathbf{H}_n^{-1} (\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon}) - \mathbf{c}) \\ & - \mathbf{v}^\top \mathbf{H}_n^{-1} (\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon}) \\ & + \boldsymbol{\varepsilon}^\top (\mathbf{G}_n \mathbf{H}_n^{-1} (\mathbf{F}_n^\top \boldsymbol{\eta} + \mathbf{v} - \mathbf{G}_n^\top \boldsymbol{\varepsilon}) - \mathbf{d}). \end{aligned}$$

The objective function in (8) can be obtained by completing the square. $\square$

Propositions 2.6 and 2.7

In what follows, for a lighter notation, we denote by $\|\cdot\|$ the Euclidean norm rather than the $L^2_{\mathcal{P}}$ norm as in the main text. In addition, for a set $A \subseteq \mathbb{R}^{N+1}$ and $\mathbf{x} \in \mathbb{R}^{N+1}$, define the distance $d(\mathbf{x}, A) := \inf_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$ and the ϵ -enlargement of A as $A^\epsilon := \{\mathbf{x} \in \mathbb{R}^{N+1} : d(\mathbf{x}, A) \leq \epsilon\}$.

Proposition 2.6

Proof. For convenience, define $\mathcal{R}_k(\mathbf{x}) := \mathbf{x}^\top \hat{\mathbf{H}}_{n,k} \mathbf{x}$, $\mathcal{R}_0(\mathbf{x}) := \mathbf{x}^\top \mathbf{H}_n \mathbf{x}$ and denote by $\mathbf{x}_0$ the solution of (8). Let $\mathbf{v}_0 := (\text{vec } \mathbf{H}_n, \text{vec } \mathbf{F}_n, \dots)$ denote a vector collecting all the vectorized population quantities to be estimated from the data and $\hat{\mathbf{v}}_k$ its respective estimator. Finally For $\epsilon, \rho > 0$, define the event $\mathcal{A}_k := \{\|\hat{\mathbf{v}}_k - \mathbf{v}_0\| \leq \epsilon, \max(\|\mathbf{x}_0\|, \|\hat{\mathbf{x}}_k\|) \leq \rho, \sup_{\|\mathbf{x}\| \leq \rho} |\mathcal{R}_k(\mathbf{x}) - \mathcal{R}_0(\mathbf{x})| \leq \epsilon\}$.

We claim that the event $\mathcal{A}_k$ implies

- (i) $\tilde{K} \subseteq \tilde{K}_k^{\rho\epsilon}$;
- (ii) $\Delta_\epsilon := \sup_{\mathbf{x} \in \tilde{K}_k^{\rho\epsilon}, \mathbf{y} \in \tilde{K}_k} |\mathcal{R}_k(\mathbf{x}) - \mathcal{R}_k(\mathbf{y})| \leq 2\rho^2(\|\mathbf{H}_n\| + \epsilon)\epsilon$, for every $\epsilon > 0$;
- (iii) $\mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) \leq 2[1 + \rho^2(\|\mathbf{H}_n\| + \epsilon)]\epsilon$.

Indeed, for $\mathbf{x} \in \tilde{K}$ we have $d(\mathbf{x}, \tilde{K}_k) \leq \|\hat{\mathbf{v}}_k - \mathbf{v}_0\| \|\mathbf{x}\| \leq \rho\epsilon$ which shows (i). For (ii), we have for $\mathbf{x} \in \tilde{K}_k^{\rho\epsilon}$ and $\mathbf{y} \in \tilde{K}_k$

$$\begin{aligned} |\mathcal{R}_k(\mathbf{x}) - \mathcal{R}_k(\mathbf{y})| &= |(\mathbf{x} + \mathbf{y})^\top \hat{\mathbf{H}}_{n,k} (\mathbf{x} - \mathbf{y})| \\ &\leq \|\mathbf{x} + \mathbf{y}\| \|\hat{\mathbf{H}}_{n,k}\| \|\mathbf{x} - \mathbf{y}\| \leq 2\rho^2 \|\hat{\mathbf{H}}_{n,k}\| \epsilon, \end{aligned}$$

and $\|\hat{\mathbf{H}}_{n,k}\| \leq \|\mathbf{H}_n\| + \|\hat{\mathbf{H}}_{n,k} - \mathbf{H}_n\| \leq \|\mathbf{H}_n\| + \|\hat{\mathbf{v}}_k - \mathbf{v}_0\| \leq \|\mathbf{H}_n\| + \epsilon$. Finally, for (iii) we have

$$\begin{aligned} \mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) &= \mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_k(\mathbf{x}_0) + \mathcal{R}_k(\mathbf{x}_0) - \mathcal{R}_0(\mathbf{x}_0) \\ &\leq \mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_k(\hat{\mathbf{x}}_k) + \mathcal{R}_k(\mathbf{x}_0) - \mathcal{R}_0(\mathbf{x}_0) + \Delta_\epsilon \\ &\leq 2 \left[\sup_{\mathbf{x}} |\mathcal{R}_k(\mathbf{x}) - \mathcal{R}_0(\mathbf{x})| + \rho^2(\|\mathbf{H}_n\| + \epsilon)\epsilon \right] \\ &\leq 2[1 + \rho^2(\|\mathbf{H}_n\| + \epsilon)]\epsilon, \end{aligned}$$

⁷We use the Mosek optimizer to solve the program in practice.

where the first equality follows because $\mathcal{R}_k(\mathbf{x}_0) \geq \mathcal{R}_k(\mathbf{x}^*)$ for some $\mathbf{x}^* \in \tilde{K}_k^{\rho\epsilon} \supseteq \tilde{K}$ by (i) and $\mathcal{R}_k(\mathbf{x}^*) \geq \mathcal{R}_k(\tilde{\mathbf{x}}) - \Delta_\epsilon$ for some $\tilde{\mathbf{x}} \in \tilde{K}_k$ by (ii). Finally $\mathcal{R}_k(\tilde{\mathbf{x}}) \geq \mathcal{R}_k(\hat{\mathbf{x}}_k)$ by the optimality of $\hat{\mathbf{x}}_k$ for $\mathcal{R}_k$ restricted to $\tilde{K}_k$.

Fix an arbitrary $\eta > 0$, due to the uniqueness of the solution (Proposition 2.4) and continuity of the objective function, the event $\|\hat{\mathbf{x}}_k - \mathbf{x}_0\| > \eta$ implies that $\mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) > \delta$ for some $\delta > 0$. For any given $\rho > 0$, we always take ϵ small enough in (iii) to have $\mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) \leq \delta$ on $\mathcal{A}_k$. Therefore,

$$\begin{aligned} \mathcal{P}(\|\hat{\mathbf{x}}_k - \mathbf{x}_0\| > \eta) &\leq \mathcal{P}(\mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) > \delta) \\ &\leq \mathcal{P}(\{\mathcal{R}_0(\hat{\mathbf{x}}_k) - \mathcal{R}_0(\mathbf{x}_0) > \delta\} \cap \mathcal{A}_k) + \mathcal{P}(\mathcal{A}_k^c) \\ &\leq \mathcal{P}(\|\hat{\mathbf{v}}_k - \mathbf{v}_0\| > \epsilon) + \mathcal{P}(\max(\|\mathbf{x}_0\|, \|\hat{\mathbf{x}}_k\|) > \rho) \\ &\quad + \mathcal{P}(\sup_{\|x\| \leq \rho} |\mathcal{R}_k(x) - \mathcal{R}_0(x)| > \epsilon). \end{aligned}$$

The first term in the last expression can be made arbitrarily small by taking k large enough by Assumption 2.5. Also, since $\mathcal{R}_0$ is strictly convex by Assumption 2.2 and $\hat{\mathbf{H}}_{n,k} \rightarrow \mathbf{H}_n$ in probability as $k \rightarrow \infty$ for fixed n by Assumption 2.5, we have that $\mathcal{R}_k$ is strongly convex with high probability (for k large enough). Therefore, there exist a closed all $B \subseteq \mathbb{R}^{N+1}$ of radius $\rho > 0$ and $N_0 \in \mathbb{N}$ such that $\mathbf{x}_0, \hat{\mathbf{x}}_k \in B$ for all $k \geq N_0$ with high probability. Thus, the second term can be made arbitrarily small by taking the radius ρ large enough. Finally, for k large enough, $\{\mathcal{R}_k\}$ is a sequence of strictly convex functions (with high probability), such that, under Assumption 2.5, it converges to $\mathcal{R}_0$ in probability uniformly on a compact set. That concludes the proof of consistency. $\square$

Proposition 2.7

Proof. Fix $\epsilon > 0$ and pick a compact set $\mathcal{E} \subseteq \Theta_* \cup \hat{\Theta}_k$ such that $\|\hat{\mathbf{s}} - \mathbf{s}\| \leq C_{\mathcal{E}} r_k$ with probability at least $1 - \epsilon$, with $\mathbf{s}, \hat{\mathbf{s}}, r_k$ being variables defined in Assumption 2.5. First we note that Ψ uniformly converges in probability over $\mathcal{E} \subseteq \Theta$, because

$$\sup_{\theta \in \mathcal{E}} \|\hat{\Psi}(\theta) - \Psi(\theta)\| \leq C_1 \|\hat{\mathbf{s}} - \mathbf{s}\| \sup_{\theta \in K} \|\theta\| \leq C_2 \|\hat{\mathbf{s}} - \mathbf{s}\| \leq C_{\mathcal{E}} r_k.$$

where C_1 is a constant depending on the dimensions N, ℓ, m ; $C_2 := C_1 \text{diam}(\mathcal{E})$ and $C_{\mathcal{E}} := C_2 C_{\mathcal{E}}$. From here on and below, we are conditioning on the event that $\{\|\hat{\mathbf{s}} - \mathbf{s}\| \leq C_{\mathcal{E}} r_k\}$. For $\theta \in \Theta_*$ we have $\|\hat{\Psi}(\theta)\| = \|\hat{\Psi}(\theta) - \Psi(\theta)\| \leq C r_k$ thus, $\Theta_* \subseteq \hat{\Theta}$ provided that $\delta \geq C_{\mathcal{E}} r_k$.

Similarly, for $\theta \in \hat{\Theta}$ we have $\|\Psi(\theta)\| \leq \|\hat{\Psi}(\theta) - \Psi(\theta)\| + \|\hat{\Psi}(\theta)\| \leq C_{\mathcal{E}} r_k + \delta$. Now, fix $\eta > 0$ and let $\theta \in \hat{\Theta} \setminus \Theta_*^\eta$ and θ_* be the (unique) point closest to θ that belongs to Θ_* . Finally, let $\tilde{\theta}$ be the point in the middle of the line segment between θ and θ_* . Since $\tilde{\theta} \notin \Theta_*$, we have that $\|\Psi(\tilde{\theta})\| > 0$ and $\nabla \|\Psi(\tilde{\theta})\|^2 \geq C_0$ for some constant $C_0 > 0$. The last condition holds because $\|\Psi(\theta)\|^2$ is a differentiable convex nonnegative function that attains its minimum when $\Psi(\theta) = 0$ (equivalently when $\theta \in \Theta_*$). Moreover $\Psi(\theta) = 0$ if and only if $\nabla \|\Psi(\theta)\|^2 = 0$. Therefore, $\theta \in \hat{\Theta} \setminus \Theta_*$ implies both $\|\Psi(\theta)\| > 0$ and $\nabla \|\Psi(\theta)\| \neq 0$ as stated. Also, due to the convexity of $\|\Psi(\theta)\|^2$, we have $\|\Psi(\theta)\|^2 \geq \|\Psi(\tilde{\theta})\|^2 + \nabla \|\Psi(\tilde{\theta})\|^2 (\theta - \tilde{\theta}) > 0 + C_0 \eta/2$. Take $\eta \rightarrow 0$ and $k \rightarrow \infty$ and eventually (for large k) $C_0 \eta/2 \leq 1$ and we have $\|\Psi(\theta)\| > \sqrt{C_0 \delta/2} \geq C_0 \eta/2$.

Set $\delta = C r_k$ for $C \geq C_{\mathcal{E}}$ and $\eta = (C/C_0) r_k$ to conclude that both inequalities imply that if $\theta \in \hat{\Theta}$ then $\theta \in \Theta_*^{M r_k}$ for large k and some positive constant $M := M_{\mathcal{E}}$ as $k \rightarrow \infty$. We can choose $M_{\mathcal{E}}$ such that $\hat{\Theta} \subseteq \Theta_*^{M_{\mathcal{E}} r_k}$ for every $k \in \mathbb{N}$. Therefore, $d(\hat{\Theta}, \Theta_*) = \inf\{\omega \geq 0 : \hat{\Theta} \subseteq \Theta_*^\omega, \Theta_* \subseteq \hat{\Theta}^\omega\} \leq M_{\mathcal{E}} r_k$ with probability at least $1 - \epsilon$.

A careful review of the proof of Proposition 2.7 reveals Corollary 1, which can be useful for inferring the coverage of Θ_* by the set estimator $\hat{\Theta}$. $\square$

Theorem 3.2

Proof. The fact that the subspace K is generated by monomials allows for the following direct sum decomposition

$$\xi_n^{(m)} = \xi_n^{\star(m)} \oplus \xi_n^{\circ(m)}, \quad (\text{B2})$$

where $\xi_n^{\star(m)}$ is the minimum-norm polynomial projection from (5), and $\xi_n^{\circ(m)}$ is the minimum-norm polynomial that lifts $\xi_n^{(m)}$ into the cone of pointwise positive polynomials. From Assumption 2.3, the solution is feasible. From Assumption 3.1, $P_{\mathcal{P},n} \oplus P_{\mathcal{P},n}^\perp = L_{\mathcal{P}}^2$. We can, therefore, write $\frac{dQ}{dP} = \xi_n^{\star(m)} + \xi_n^{\circ(m)} + \epsilon$ with $\xi_n^{\star(m)} \in K, \xi_n^{\circ(m)} \in K^\perp, \xi_n^{\star(m)} + \xi_n^{\circ(m)} \in M_n(D) \cap K \subseteq P_{\mathcal{P},n}$, and $\epsilon \in P_{\mathcal{P},n}^\perp$. For each m , $\xi_n^{\star(m)}$ solves the standard Hilbert minimum-norm problem (see Section 2.2), and from Assumption 3.1 it converges in $L_{\mathcal{P}}^2$ to dQ/dP (Filipović, Mayerhofer, and Schneider 2013). From this, we can write

$$\begin{aligned} 0 &= \lim_{m \rightarrow \infty} \left\| \frac{dQ}{dP} - \xi_n^{\star(m)} \right\|^2 = \lim_{m \rightarrow \infty} \left\| \xi_n^{\circ(m)} + \epsilon \right\|^2 \\ &= \lim_{m \rightarrow \infty} \left\| \xi_n^{\circ(m)} \right\|^2 + \lim_{m \rightarrow \infty} \|\epsilon\|^2. \end{aligned}$$

From the non-negativity of the norms $\lim_{m \rightarrow \infty} \left\| \xi_n^{\circ(m)} \right\| = 0$ and $\lim_{m \rightarrow \infty} \|\epsilon\| = 0$. Therefore, $\lim_{m \rightarrow \infty} \left\| \frac{dQ}{dP} - \xi_n^{(m)} \right\| = \lim_{m \rightarrow \infty} \|\epsilon\| = 0$. $\square$

Acknowledgments

We are grateful to Don Andrews, Oleg Bondarenko, Albina Danilova, Marc Van Uffelen, and participants at the 2021 SoFiE Conference (San Diego), 2021 Brazilian Meeting of Finance, Lunchtime Workshop at Boston University, and the Remote Seminar Series on Computational Economics and Finance for helpful comments.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Paul Schneider gratefully acknowledges the SNF grant 100018_189086 “Scenarios”.

References

- Ait-Sahalia, Y. (1999), “Transition Densities for Interest Rate and Other Nonlinear Diffusions,” *Journal of Finance*, 54, 1361–1395. [487]
- (2002), “Maximum Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach,” *Econometrica*, 70, 223–262. [482]
- Ait-Sahalia, Y., Yu, J. (2006), “Saddlepoint Approximations for Continuous-Time Markov Processes,” *Journal of Econometrics*, 134, 507–551. [482]
- Alvarez, F., and Jermann, U. (2005), “Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth,” *Econometrica*, 73, 1977–2016. [483]
- Backus, D. K., Chernov, M., and Zin, S. (2014), “Affine Term Structure Models and the Forward Premium Anomaly,” *Journal of Finance*, 69, 51–99. [483]
- Barndorff-Nielsen, O., and Halgreen, C. (1977), “Infinite Divisibility of the Hyperbolic and Generalized Inverse Gaussian Distributions,” *Probability Theory and Related Fields*, 38, 309–311. [488]

- Bates, D. S. (2000), "Post-'87 Crash Fears in the s&p 500 Futures Option Market," *Journal of Econometrics*, 94, 181–238. [487]
- Berlinet, A., and Thomas-Agnan, C. (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, pp. 1–54, Boston, MA: Springer. [482]
- Borovička, J., Hansen, L. P., and Scheinkman, J. A. (2016), "Misspecified Recovery," *Journal of Finance*, 71, 2493–2544. [483]
- Breeden, D. T., and Litzenberger, R. H. (1978), "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51, 621–651. [487]
- Brigo, D., and Mercurio, F. (2006), *Interest Rate Models: Theory and Practice. With Smile, Inflation and Credit*, Berlin: Springer-Verlag. [487]
- Carr, P., and Madan, D. (2001), *Towards a Theory of Volatility Trading. Option Pricing, Interest Rates and Risk Management*, pp. 458–476, New York: Cambridge University Press. [488]
- Chabi-Yo, F., and Bakshi, G. (2012), "Variance Bounds on the Permanent and Transitory Components of the Stochastic Discount Factor," *Journal of Financial Economics*, 105, 191–208. [483]
- Chabi-Yo, F., Bakshi, G., and Gao, X. (2018), "A Recovery that We Can Trust? Deducing and Testing the Restrictions of the Recovery Theorem," *Review of Financial Studies*, 31, 532–555. [483]
- Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics 6B*, eds. J. Heckman and E. E. Leamer, pp. 5549–5632, Amsterdam: North-Holland. [482]
- Christensen, T. (2017), "Nonparametric Stochastic Discount Factor Decomposition," *Econometrica*, 85, 1501–1536. [483,491]
- Duffie, D., Pan, J., and Singleton, K. (2000), "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," *Econometrica*, 68, 1343–1376. [483,489,490]
- Duffie, D., Filipović, D., and Schachermayer, W. (2003), "Affine Processes and Applications in Finance," *Annals of Applied Probability*, 13, 984–1053. [482]
- Filipović, D., Mayerhofer, E., and Schneider, P. (2013), "Density Approximations for Multivariate Affine Jump-Diffusion Processes," *Journal of Econometrics*, 176, 93–111. [482,483,487,488,492]
- Giesecke, K., and Schwenkler, G. (2018), "Filtered Likelihood for Point Processes," *Journal of Econometrics*, 204, 33–53. [482]
- Grenander, U. (1981). *Abstract Inference / Ulf Grenander*. New York: Wiley. [482]
- Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012), "Conditional Mean Embeddings as Regressors," in *ICML'12*, pp. 1803–1810, Omnipress. [482]
- Hansen, L. P., and Scheinkman, J. (2009), "Long-Term Risk: An Operator Approach," *Econometrica*, 77, 177–234. [483]
- Kato, N., and Kuriki, S. (2013), "Likelihood Ratio Tests for Positivity in Polynomial Regressions," *Journal of Multivariate Analysis*, 115, 334–346. [482]
- Klebanov, I., Schuster, I., and Sullivan, T. J. (2020), "A Rigorous Theory of Conditional Mean Embeddings," *SIAM Journal on Mathematics of Data Science*, 2, 583–606. [482]
- Lee, R. (2004), "Option Pricing by Transform Methods: Extensions, Unification and Error Control," *Journal of Computational Finance*, 7, 51–86. [488,489]
- Lu, J., and Qu, Z. (2021), "Sieve Estimation of Option-Implied State Price Density," *Journal of Econometrics*, 224, 88–112. [489]
- Luenberger, D. G. (1997), *Optimization by Vector Space Methods* (1st ed.), New York: Wiley. [484]
- Mijatović, A., and Schneider, P. (2010), "Globally Optimal Parameters for Non-Linear Diffusions," *Annals of Statistics*, 38, 215–245. [482]
- Park, J., and Muandet, K. (2020), "A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings," working paper, Max Planck Institute for Intelligent Systems. [482]
- Qin, L., and Linetsky, V. (2016), "Positive Eigenfunctions of Markovian Pricing Operators: Hansen-Scheinkman Factorization, Ross Recovery, and Long-Term Pricing," *Operations Research*, 64, 99–117. [483]
- (2017), "Long-Term Risk: A Martingale Approach," *Econometrica*, 85, 299–312. [483]
- Renner, P., and Schmedders, K. (2015), "A Polynomial Optimization Approach to Principal-Agent Problems," *Econometrica*, 83, 729–769. [482]
- Ross, S. (2015), "The Recovery Theorem," *Journal of Finance*, 70, 615–648. [483]
- Schmüdgen, K. (2017), *The Moment Problem*. Graduate Texts in Mathematics, Cham: Springer. [491]
- Schuster, I., Mollenhauer, M., Klus, S., and Muandet, K. (2020), "Kernel Conditional Density Operators," in *Proceedings of Machine Learning Research* (Vol. 108), pp. 993–1004, Online, 26–28 Aug 2020. PMLR. [482]
- Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009), "Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML 2009, pp. 961–968, New York, NY, USA. Association for Computing Machinery. ISBN 9781605585161. [482]
- Yu, J. (2007), "Closed-Form Likelihood Estimation of Jump-Diffusions with an Application to the Realignment Risk of the Chinese Yuan," *Journal of Econometrics*, 141, 1245–1280. [482]