Nonparametric Sharpe Ratio Function Estimation in Heteroscedastic Regression Models via Convex Optimization

Seung-Jean Kim
Cubist Systematic Strategies

Johan Lim
Seoul National University

Joong-Ho Won
Seoul National University

Abstract

We consider maximum likelihood estimation (MLE) of heteroscedastic regression models based on a new "parametrization" of the likelihood in terms of the Sharpe ratio function, or the ratio of the mean and volatility functions. While with a standard parametrization the MLE problem is not convex and hence hard to solve globally, our parametrization leads to a functional that is jointly convex in the Sharpe ratio and inverse volatility functions. The major difficulty with the resulting infinite-dimensional convex program is the shape constraint on the inverse volatility function. We propose to solve the problem by solving a sequence of finite-dimensional convex programs with increasing dimensions, which can be done globally and efficiently. We demonstrate that, when the goal is to estimate the Sharpe ratio function directly, the finite-sample performance of the proposed estimation method is superior to existing methods that estimate the mean and variance functions separately. When applied to a financial dataset, our method captures a well-known covariate-dependent effect on the Shape ratio.

1 Introduction

We consider a regression model of the form

$$y_i = \mu(x_i) + \sigma(x_i) z_i, \quad i = 1, \ldots, n, \tag{1}$$

where $z_1, \ldots, z_n$ are independent and identically distributed (iid) Gaussian random variables with zero mean and unit variance. This model states that the conditional distribution of y given x is Gaussian with mean $\mathbf{E}(y \,|\, x) = \mu(x)$ and variance $\mathbf{V}(y \,|\, x) = \sigma(x)^2$. The conditional variance depends on x, so the model (1) is called a heteroscedastic regression model. This model has been widely used in financial econometrics, since it allows us to take into account nonlinearity and conditional heteroscedasticity in financial time series. In particular, this model arises as the discretized version of the continuous-time stochastic diffusion model which is commonly used in financial derivative pricing.

We are interested in estimating the ratio between the mean and volatility (standard deviation) functions, or the Sharpe ratio function $f(x) = \mu(x)/\sigma(x)$, from the observed data points $(x_1, y_1), \ldots, (x_n, y_n)$. In finance, the Sharpe ratio is the most popular measure of risk-adjusted return, and often used as a gold standard to compare different assets or trading strategies. Covariate-dependent Sharpe ratio (Sharpe ratio function) has been motivated in many different contexts of financial literature. Since static Sharpe ratio does not reflect the asset dynamics, it may oversimplify risk, be distorted by serial correlation, or affected by the phases of business cycle (Lo, 2002). Diffusion process modeling (Leung et al., 2013) thus uses time-varying Sharpe ratio in order to outperform a (time-varying) benchmark. In this regard, Tang and Whitelaw (2011) points out that, "if [market Sharpe ratio] shows substantial predictable variation, then this variation needs to be accounted for when using the market as a performance benchmark." Also, "time-variation in the Sharpe ratio might provide clues to the fundamental economics underlying the economy and asset pricing" (Tang and Whitelaw, 2011), and "proxies for the net change in the investment opportunity set" (Maio and Santa-Clara, 2012). In the latter work, the Sharpe ratio is modeled as a linear function of time-varying market variables such as credit spreads and yields.

The literature on estimating heteroscedastic regression models, and the related diffusion process models, is vast; see Härdle and Tsybakov (1997), Hall and Carroll (1989), and Cai and Wang (2008), to name a few.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

However, the existing literature is limited to the estimation of the mean and variance functions, not the Sharpe ratio function itself. With existing methods, the mean and variance functions are estimated separately using, *e.g.*, the local polynomial estimation procedure (Fan and Gijbels, 1996), and combine them to construct an estimate of the Sharpe ratio function. This approach is less efficient than estimating the target function directly.

In this paper, we propose a direct estimation method for the Sharpe ratio function, assuming that this function is smooth. To do this, we reparametrize the regression model (1) with the Sharpe ratio function $f(x) = \mu(x)/\sigma(x)$ and the inverse volatility function $g(x) = 1/\sigma(x)$. In heteroscedastic settings, separating the scale from the target function is, in our view, more appealing. We then show that the maximum likelihood estimation problem in (f, g) is convex (in an infinite-dimensional space), and that $f(x)$ and $g(x)$ can be expanded as series of basis functions under reasonable assumptions on their smoothness and shape. The major obstacle is the shape constraint that the inverse volatility function g should be nonnegative. We resolve this by solving a sequence of finite-dimensional convex optimization problems, each of which can be solved globally and efficiently.

We organize this paper as follows. Section 2 reviews existing approaches for the estimation problems in the heteroscedastic regression model (1), in the context of Sharpe ratio function estimation. In Section 3, we introduce our proposal for the Sharpe ratio function estimation based on the reparametrization (f, g), the function space we consider, and the associated convex optimization procedure. In Section 4, we carry out extensive simulation studies to investigate the performance of the proposed method in estimating the Sharpe ratio function, and compare this to existing methods. In Section 5, we apply our procedure to a three-month US Treasury Bill interest rate dataset, and make a connection to continuous-time diffusion process models in finance. We conclude the paper in Section 6 with discussion on extensions of the proposed method, *e.g.*, to non-Gaussian data.

2 Related methods

Many methods have been proposed to estimate the mean and/or variance functions in the heterogeneous regression model (1). Estimation of the smooth mean function has been studied for many decades; methods include smoothing splines (Wahba, 1990), kernel regression (Wand and Jones, 1994), and local polynomial regression (Fan and Gijbels, 1996). For the estimation of the variance function, there are two major approaches: residual-based and difference-based.

Residual-based methods estimate the mean function $\mu(x)$ first and then estimate the variance function by estimating the mean of the squared residual $r(x) = (y - \mu(x))^2$ using the fact $\mathbf{E}(r(x)) = \sigma^2(x)$. For the estimation of the mean function, The squared residuals are evaluated at the "data points" $(x_i, \widehat{r}_i)$, where $\widehat{r}_i = (y_i - \widehat{\mu}(x_i))^2$ for $i = 1, \ldots, n$. Local linear regression estimator is one of the most popular methods, which estimate the mean of the squared residuals by solving for each i

$$\big(\widehat{\alpha}_i, \widehat{\beta}_i\big) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \big(\widehat{r}_i - \alpha - \beta\big(x_i - x\big)\big)^2 W\left(\frac{x_i - x}{h}\right),$$

where $W(\cdot)$ is a kernel on $\mathbb{R}$ and $h > 0$ is the bandwidth. Then the residual-based variance function estimator is defined as $\widehat{\sigma^2}(x_i) = \widehat{\alpha}_i$. Fan and Yao (1998) propose a two-stage method based on the optimal bandwidth selection procedure for local polynomial regression (Fan and Gijbels, 1996). In case the distribution of ϵ_i is known, a smoothing spline estimator of the variance function is considered by Liu et al. (2007), using the reparametrization $\sigma(x) = \exp(\tilde{g}(x)/2)$ with $\tilde{g} \in \mathcal{W}^{2,2}$ to remove the nonnegativity constraint. We note that this reparametrization is somewhat restrictive as far as the nonnegativity is concerned. With any method, if the Sharpe ratio function is of interest, it should be estimated indirectly as $\hat{f}(x) = \hat{\mu}(x)/(\widehat{\sigma^2}(x))^{1/2}$.

Difference-based methods possess an advantage that they do not require the estimate of the mean function. We should remark that if the Sharpe ratio function is the object of concern, the mean function has to be estimated separately using the aforementioned methods, e.g., Fan and Gijbels (1996). These methods utilize that fact that in the homoscedastic case, when the data points are sorted so that $x_1 \le x_2 \le \cdots \le x_n$, the pseudo-residual

$$\widehat{\sigma^2}(x_i) = \left(\textstyle\sum_{j=-r}^{r} w_j y_{i+j}\right)^2,$$

where $r > 0$ is a fixed constant and coefficients $\{w_j\}$ satisfies $\sum_{j=-r}^{r} w_j = 0$ and $\sum_{j=-r}^{r} w_j^2 = 1$, forms an unbiased estimator of the variance $\sigma^2(x) \equiv \sigma^2$. For example, if $r = 1$, $w_1 = 1/\sqrt{2}$, $w_0 = -w_1$, and $w_{-1} = 0$, and $w_j = 0$, the estimator becomes $\widehat{\sigma^2}(x_i) = (y_{i+1} - y_i)^2/2$. In the heteroscedastic case, Brown and Levine (2007) consider applying local linear regression to the pseudo-residuals to estimate the variance function $\sigma^2(x)$; see Tong et al. (2013) and Dai et al. (2015) for theoretical analysis of many difference-based estimators and their extension to repeatedly measured data. The effects of the mean function esti-

mator on the estimation of the variance function are reviewed in Wang et al. (2008).

3 Sharpe ratio function estimation

3.1 Main problem

Under model (1), the negative log-likelihood of the data is given by

$$l(\mu,\sigma)=\sum_{i=1}^{n}\left\{\frac{1}{2}\left(\frac{y_i-\mu(x_i)}{\sigma(x_i)}\right)^2+\log\sigma(x_i)\right\}, \quad (2)$$

ignoring additive constants. Of course, it is required that the standard deviation function $\sigma(x)$ (or the variance function $\sigma^2(x)$) to be positive. As our interest is to estimate the Sharpe ratio function $f(x)=\mu(x)/\sigma(x)$, we reparametrize (2) with f and additionally with $g(x)=1/\sigma(x)$ so that

$$l(f,g)=\sum_{i=1}^{n}\left\{\frac{1}{2}\big(g(x_i)y_i-f(x_i)\big)^2-\log g(x_i)\right\}. \quad (3)$$

We want to estimate functions f and g from appropriate vector spaces of functions, under some reasonable assumptions on their shape. As our primary interest is in f, function g can be considered as a nuisance parameter.

In the homoscedastic case, *i.e.*, g is a constant function, estimating f is equivalent to estimating μ, and it is customary to find f over the second-order Sobolev space that consists of twice differentiable functions on $[0,1]$ (without loss of generality) equipped with the norm

$$\|f\|_{2,2}=\left(\sum_{i=0}^{2}\int_0^1 (f^{(i)}(x))^2\,dx\right)^{1/2},$$

where $f^{(i)}$ is the ith derivative of f. Throughout, this space is denoted by $\mathcal{W}^{2,2}([0,1])$ or simply $\mathcal{W}^{2,2}$. It is also customary to penalize the roughness of f, where it is measured by the size of its second derivative, in order to have a smooth estimate. The resulting optimization problem is thus

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\sum_{i=1}^{n}(y_i-f(x_i))^2\\ \text{subject to} & f\in\mathcal{W}^{2,2},\ \int_0^1(f^{(2)}(x))^2\,dx\le B\end{array} \quad (4)$$

for some roughness bound $B>0$. It is well known that the solution of this problem is given by the natural cubic spline with knots at the data points $x_1,x_2,\ldots,x_n$ (Wahba, 1990; Schölkopf et al., 2001). That is, the solution f to (4) has the form

$$f^\star(x)=\sum_{j=1}^{n}\alpha_j^\star N_j(x), \quad (5)$$

where $\{N_1(x),\ldots,N_n(x)\}$ is the set of basis functions for representing the family of natural cubic splines having knots at $x_1,x_2,\ldots,x_n$. The optimal coefficient vector $\alpha^\star=(\alpha_1^\star,\ldots,\alpha_n^\star)$ is obtained by solving the following (quadratically constrained) quadratic program (QP)

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\alpha^T P\alpha-q^T y\\ \text{subject to} & \alpha^T\Omega\alpha\le B,\end{array}$$

where $P=(P_{jk})$ with $P_{jk}=\sum_{i=1}^{n}N_j(x_i)N_k(x_i)$; $q=(q_1,\ldots,q_n)$ with $q_i=\sum_{j=1}^{n}N_j(x_i)$; $\Omega=(\Omega_{jk})$ with $\Omega_{jk}=\int_0^1 N_j^{(2)}(x)N_k^{(2)}(x)\,dx$; and $y=(y_1,\ldots,y_n)$. Equivalently, we can minimize the Lagrangian form of the above QP

$$(1/2)\alpha^T P\alpha-q^T y+(\lambda/2)\alpha^T\Omega\alpha, \quad (6)$$

for the regularization parameter $\lambda>0$ corresponding to B. Problem (6) is the conventional (linearly constrained) QP, which is easier for numerical optimization.

In the heteroscedastic case, it is natural to also seek g in $\mathcal{W}^{2,2}$, yielding the following formulation

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^{n}\left\{\frac{1}{2}(g(x_i)y_i-f(x_i))^2-\log g(x_i)\right\}\\ \text{subject to} & f\in\mathcal{W}^{2,2},\ \int_0^1(f^{(2)}(x))^2\,dx\le B_f,\\ & g\in\mathcal{W}^{2,2},\ g\in\mathcal{P},\ \int_0^1(g^{(2)}(x))^2\,dx\le B_g,\end{array} \quad (7)$$

where $\mathcal{P}=\{g: g(x)\ge 0,\forall x\in[0,1]\}$ is the collection of all positive functions. Problem (7) is jointly convex in $(f,g)\in\mathcal{W}^{2,2}\times\mathcal{W}^{2,2}$, because $\mathcal{P}$ is a convex cone and the L^2 norm constraints are convex. Unlike (4), however, the infinite number of constraints in $\mathcal{P}$ prevent us from obtaining an easy-to-solve, finite-dimensional convex program as (6). Nevertheless, (7) can be efficiently solved by solving a sequence of finite-dimensional convex programs. This is the main subject of the next subsection.

Note that, if we are to find μ and σ (or σ^2) $\in\mathcal{P}$ jointly over $\mathcal{W}^{2,2}\times\mathcal{W}^{2,2}$, then the objective of (7) shall be replaced by (2), and the constraints be retained with f and g substituted respectively by μ and σ. While the constraint set is convex, the objective in this situation is not convex in general.

3.2 Solution procedure

We first show that the optimal f for problem (7) attains the form (5). That is, the optimal f is a natural cubic spline having knots on the data points.

Proposition 1. *Let $(f^\star,g^\star)$ be a solution of (7). Then, $f^\star(x)$ has the form (5).*

Proof. Fix $g\in\mathcal{W}^{2,2}\cap\mathcal{P}$ such that $\int_0^1(g^{(2)}(x))^2\,dx\le B_g$. Then (7) is equivalent to finding f_g minimizing $(1/2)\sum_{i=1}^{n}(y_i^g-f(x_i))^2$ subject to $f\in\mathcal{W}^{2,2}$

and $\int_0^1 (f^{(2)}(x))^2\,dx \le B_f$, where $y_i^g = g(x_i)y_i$, $i = 1,\ldots,n$. This subproblem is precisely the smoothing spline problem (4) with "data" $\{(x_i, y_i^g)\}_{i=1}^n$ and thus the minimizing f_g must have the form (5). The optimal $f^\star$ is among those f_g for all g satisfying the constraints. □

If there were no nonnegativity constraint $g \in \mathcal{P}$, then the optimal g could have also been characterized in the same fashion as Proposition 1, and (7) should have been reduced to a finite-dimensional convex program on the spline coefficients. With $g \in \mathcal{P}$, however, no such simple characterization of the optimal g is available. While Utreras (1985) provides an optimality condition for the solution to smoothing splines with nonnegativity, monotonicity, or convexity (*i.e.*, (4) with an additional constraint such as $f \in \mathcal{P}$), no constructive algorithm, especially for the nonnegativity constraint, has yet to be devised. To circumvent this difficulty, one could think of restricting the class of functions, e.g., to a nonnegative combination of nonnegative basis functions or a log-linear parametrization (Ramsay, 1998), or starting from the unconstrained solution to (7) and adaptively adding local constraints until the desired shape is obtained (Turlach, 2005). In any case, the search space is much smaller than the allowed $\mathcal{W}^{2,2} \cap \mathcal{P}$.

Here we adopt the method proposed by Papp and Alizadeh (2014), which explores the cone $\mathcal{K} = \mathcal{W}^{2,2} \cap \mathcal{P}$ sequentially by sieves. In other words, we form a sequence of subsets $\{\mathcal{K}_m\}$ of $\mathcal{K}$ such that $\cup_{m=1}^\infty \mathcal{K}_m$ is dense in $\mathcal{K}$. Then for each m, we solve an approximate version of (7) in which $\mathcal{K}$ is replaced by $\mathcal{K}_m$. We design the sieve $\{\mathcal{K}_m\}$ so that the tolerated roughness of the functions increases with m so that an appropriate value of m can be determined by cross-validation. Specifically, we let $\mathcal{K}_m = \mathcal{S}_m \cap \mathcal{P}$ where $\mathcal{S}_m$ is the space of cubic splines with k_m knots at $t_{m1} < \cdots < t_{mk_m}$. (Hence $\mathcal{S}_m$ is finite-dimensional with dimension $k_m + 4$.) If the length $\max_{i=0,\ldots,k_m} |t_{m,i+1} - t_{mi}|$ of the longest interval between adjacent knots ($t_{m0} = 0$ and $t_{m,k_m+1} = 1$) approaches zero as $m \to \infty$, then $\cup_{m=1}^\infty \mathcal{S}_m$ is dense in $\mathcal{W}^{2,2}$ (Schumaker, 1981). Consequently, $\cup_{m=1}^\infty \mathcal{K}_m$ is dense in $\mathcal{W}^{2,2} \cap \mathcal{P}$. With this choice of a sieve (together with Proposition 1) and for each $\mathcal{K}_m$, g is restricted to have the form $g(x) = \beta^T\psi(x)$, where $\psi(x) = (\psi^{(0)}(x), \ldots, \psi^{(k_m)}(x))^T$, $\beta = (\beta^{(0)}, \ldots, \beta^{(k_m)})^T \in \mathbb{R}^{4k_m}$ with

$$\psi^{(j)}(x) = \mathbb{I}_{\{t_j \le x \le t_{j+1}\}}\left(1, \tfrac{x-t_j}{t_{j+1}-t_j}, \tfrac{(x-t_j)^2}{(t_{j+1}-t_j)^2}, \tfrac{(x-t_j)^3}{(t_{j+1}-t_j)^3}\right),$$
$$\beta^{(j)} = (\beta_0^{(j)}, \beta_1^{(j)}, \beta_2^{(j)}, \beta_3^{(j)}),$$

Additional constraints on β to ensure continuity of derivatives

$$\beta_0^{(j)} + \beta_1^{(j)} + \beta_2^{(j)} + \beta_3^{(j)} = \beta_0^{(j+1)}$$
$$\beta_1^{(j)} + 2\beta_2^{(j)} + 3\beta_3^{(j)} = \beta_1^{(j+1)}$$
$$2\beta_2^{(j)} + 6\beta_3^{(j)} = 2\beta_2^{(j+1)} \qquad (8)$$

are necessary to be imposed for $j = 0,\ldots,k_m - 1$. Then, the approximate version of (7) corresponding to each $\mathcal{K}_m$ is the following convex program:

$$\begin{aligned}
\text{minimize}\quad & \sum_{i=1}^n \left(\frac{1}{2}(\alpha^T a_i - \beta^T b_i)^2 - \log(\beta^T c_i)\right) \\
\text{subject to}\quad & \alpha^T \Omega \alpha \le B_f, \qquad (9)\\
& \beta^T\psi(x) \ge 0,\ \forall x \in [0,1], \\
& \sum_{l=0}^3 \beta_l^{(j)} = \beta_0^{(j+1)},\ j = 0,\ldots,k_m-1, \\
& \sum_{l=1}^3 l\beta_l^{(j)} = \beta_1^{(j+1)},\ j = 0,\ldots,k_m-1, \\
& \beta_2^{(j)} + 3\beta_3^{(j)} = \beta_2^{(j+1)},\ j = 0,\ldots,k_m-1, \\
& \|\beta\|_2 \le B_m,
\end{aligned}$$

where $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^{4k_m}$ are the optimization variables; $a_i = (N_1(x_i), \ldots, N_n(x_i))^T$, $b_i = y_i\psi(x_i)$, and $c_i = \psi(x_i)$. The norm constraint on the last line is due to the equivalence of all norms in a finite-dimensional space. Increasing the bound B_m slowly with m regularizes the roughness of $g(x) = \beta^T\psi(x)$. The third line of (9) still involves infinite number of constraints. In the sequel, we describe how these constraints can be represented by a small number of second-order cone constraints on β and auxiliary variables. Thus (9) reduces to a finite-dimensional convex program.

3.3 Nonnegative splines

Papp and Alizadeh (2014) utilize the fact that a necessary and sufficient condition for a univariate polynomial to be nonnegative over an interval can be expressed as a set of semidefinite constraints. Focusing on cubic splines, this fact can be written:

Proposition 2 (Papp and Alizadeh (2014))**.** *Let* $p^{(j)}(x) = \sum_{l=0}^3 \beta_l^{(j)} \left(\frac{x-t_j}{t_{j+1}-t_j}\right)^l$. *Then* $p^{(j)}(x) \ge 0$ *for all* $x \in [t_j, t_{j+1}]$ *if and only if*

$$U^{(j)} = \begin{bmatrix} u_0^{(j)} & u_1^{(j)} \\ u_1^{(j)} & u_2^{(j)} \end{bmatrix},\ V^{(j)} = \begin{bmatrix} v_0^{(j)} & v_1^{(j)} \\ v_1^{(j)} & v_2^{(j)} \end{bmatrix} \in \mathbb{S}_+^2, \qquad (10)$$

where $\mathbb{S}_+^d$ *is the set of symmetric, positive semidefinite matrices in* $\mathbb{R}^{d\times d}$, *and*

$$\begin{aligned}
\beta_0^{(j)} &= v_0^{(j)} \\
\beta_1^{(j)} &= 2v_1^{(j)} + u_0^{(j)} - v_0^{(j)} \\
\beta_2^{(j)} &= v_2^{(j)} + 2u_1^{(j)} - 2v_1^{(j)} \\
\beta_3^{(j)} &= u_2^{(j)} - v_2^{(j)}. \qquad (11)
\end{aligned}$$

It is obvious that $\beta^T\psi(x) \geq 0$ for all $x \in [0,1]$ if and only if $p^{(j)}(x) \geq 0$ for $x \in [t_j, t_{j+1}]$, $j = 0, \ldots, k_m$. Because the 2×2 semidefinite constraints (10) are equivalent to the second-order cone constraints

$$\begin{aligned} \|(u_0^{(j)} - u_2^{(j)}, 2u_1^{(j)})^T\|_2 &\leq u_0^{(j)} + u_2^{(j)}, \\ \|(v_0^{(j)} - v_2^{(j)}, 2v_1^{(j)})^T\|_2 &\leq v_0^{(j)} + v_2^{(j)} \end{aligned} \tag{12}$$

(see, *e.g.*, Lobo et al. (1998)), the third line in (9) can be replaced by (11) and (12), with additional variables $u^{(j)} = (u_0^{(j)}, u_1^{(j)}, u_2^{(j)})$ and $v^{(j)} = (v_0^{(j)}, v_1^{(j)}, v_2^{(j)})$, $j = 0, \ldots, k_m$. Then problem (9) can be efficiently solved using existing interior-point solvers. Modern optimization software such as `SDPT3` (Toh et al., 1999) and `Knitro` (Byrd et al., 2006) can handle second-order cone constraints (which subsumes linear constraints) with a convex objective.

Alternatively, we may use nonnegative piecewise cubic polynomials that span $\mathcal{S}_m$ and combine them with nonnegative coefficients (*i.e.*, conic combination) to represent $g(x)$. In this case, the constraint $\beta^T\psi(x) \geq 0$ in (7) is replaced simply by $\beta \geq 0$, with the dimensions of β and $\psi(x)$, and the components of $\psi(x)$ appropriately modified. The continuity constraints (8) should also be modified accordingly, while preserving linearity. Possible choices of $\psi(x) = (\psi^{(0)}(x), \ldots, \psi^{(r_m-1)}(x))$ include the cubic B-spline basis ($r_m = k_m + 4$), and the piecewise cubic Bernstein polynomials $\psi^{(j)}(x) = \mathbb{I}_{\{t_j \leq x \leq t_{j+1}\}}(\psi_0^{(j)}(x), \psi_1^{(j)}(x), \psi_2^{(j)}(x), \psi_3^{(j)}(x))$, where

$$\psi_l^{(j)}(x) = \left(\tfrac{x-t_j}{t_{j+1}-t_j}\right)^l \left(1 - \tfrac{x-t_j}{t_{j+1}-t_j}\right)^{3-l}, \quad l = 0, \ldots, 3.$$

($r_m = k_m$). While the set of functions, say, $\mathcal{P}_m$ represented by conic combination of these functions is a subset of $\mathcal{K}_m$, it can be shown that $\{\mathcal{P}_m\}$ forms a sieve under an additional assumption on the sequence of knots (Papp and Alizadeh, 2014, Thorem 1). The key advantage of this approach is that it makes the constraint set in (9) polyhedral. If the objective in (7) is quadratic, then we need to solve a sequence of QPs, which are in general solved more efficiently than the problems involving the second-order cone constraints (12). In our problem, however, the objective also contains non-quadratic convex functions, and we do not see a merit over the formulation in the previous paragraph, which can handle the full $\mathcal{K}_m$.

4 Simulation results

In this section we carry out extensive simulations to compare our Sharpe ratio estimation method with existing methods of separately estimating mean and variance functions. The focus of this study is estimation accuracy of the Sharpe ratio function, which our method estimates directly and globally, while the other methods do indirectly via the mean and variance functions. We use two simulated models previously studied in the literature.

Example 1 We simulated 100 random samples of size $n = 200$ from the model considered by Fan and Yao (1998)

$$Y_i = a(X_i + 2\exp(-16X_i^2)) + (0.4\exp(-2X_i^2) + 0.2)\varepsilon_i,$$

where $X_i \overset{\text{iid}}{\sim} \text{Unif}[-2, 2]$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Thus the Sharpe ratio function is given by

$$f(x) = a(x + 2\exp(-16x^2))/(0.4\exp(-2x^2) + 0.2).$$

Four different values of $a = 0.5, 1, 2$, and 4 were used in the simulation. For each simulated sample, the estimation accuracy of an estimator was evaluated by the mean absolute deviation

$$\mathcal{E}_{\text{MAD}} = (1/n_{\text{grid}}) \sum_{j=1}^{n_{\text{grid}}} |\hat{f}(x_j) - f(x_j)|,$$

where x_j, $j = 1, \ldots, n_{\text{grid}}$ are uniform grid points on $(-2, 2)$ with $n_{\text{grid}} = 99$. We compared our estimator with a residual-based one (Fan and Yao, 1998) and a difference-based one (Brown and Levine, 2007, $r = 1$). For the former, $\hat{f}$ was estimated by solving (7) with the regularization parameters chosen by a leave-one-out cross-validation over a grid of parameters. For the latter two, $\hat{f}$ was determined by $\hat{f}(x) = \widehat{\mu}(x)/(\widehat{\sigma^2}(x))^{1/2}$, where $\widehat{\mu}$ and $\widehat{\sigma^2}$ were estimated separately using local polynomial fitting with the Epanechnikov kernel, where the bandwidth was selected by the method of Fan and Gijbels (1996). We used `CVX` (Grant and Boyd, 2014) with `SDPT3` as the backend to solve (9) and R package `locpol` for the other two methods. The results are summarized in Fig. 1, in which the boxplots of $\mathcal{E}_{\text{MAD}}$ are presented for each value of the four as. (The corresponding numbers are in the supplement.) The proposed method tends to have smallest median absolute deviation for all the scenarios considered. The residual-based method comes the next, with a larger average and dispersion but some overlap; the difference-based method follows with some margin. We remark that there were some discarded values (`NaN`s) in computing $\mathcal{E}_{\text{MAD}}$ for the residual- and difference-based methods; see the next paragraph for the details.

An investigation into the case $a = 0.5$ gives a further insight into our estimation method. (The situation is similar in other cases as well.) In Fig. 2, we plot the estimated Sharpe ratio function $\hat{f}(x)$ and variance function $\widehat{\sigma^2}(x)$ together with the true functions $f(x)$ and $\sigma^2(x)$. Note the lesser variability of the

estimated Sharpe ratio functions $\hat{f}$ (dotted gray) of the proposed method compared to the others. Highlighted for the residual- and difference-based methods are an estimated variance function $\widehat{\sigma^2}$ that has negative values and the corresponding $\hat{f}$. Both the residual- and difference-based methods occasionally had negative variance estimates, particularly at xs close to the boundary: the difference-based method had nine out of 100 samples with negative variance estimates, while the residual-based method had one. In these cases, `NaN`s were generated in computing $\hat{f}(x)$ for xs with negative variances. The unstable estimates of the variance function are one reason that these two methods have a large average and dispersion of $\mathcal{E}_{\mathrm{MAD}}$ in estimating $f(x) = \mu(x)/\sigma(x)$. This phenomenon is due to the fact that these methods estimate $\sigma^2(x)$ by local polynomial regression, where the nonnegativity constraint is difficult to be imposed. As the estimation of $f(x)$ is indirectly done using the estimated mean and variance functions with these methods, a negative value in the variance function may cause an undefined behavior in the estimation of the Sharpe ratio function. (Local *constant* regression does not suffer this drawback, with less satisfactory theoretical properties; see Xu and Phillips (2011).) On the contrary, in our joint estimation approach $\hat{f}$ is estimated directly, and the nonnegativity constraint in (7) ensures that $\widehat{\sigma^2}(x) = 1/(\hat{g}(x))^2$ is positive (there was no exact zeros). While the variance function is estimated indirectly in this case, Fig. 2 demonstrates a quite acceptable estimation result.

Example 2 We simulated 100 random samples of size $n = 200$ from the model considered by Wang et al. (2008)

$$Y_i = (3/4)\sin(b\pi X_i) + ((X_i - 1/2)^2 + 1/2)^{1/2}\varepsilon_i,$$

where $X_i \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Thus the Sharpe ratio function is given by

$$f(x) = (3/4)\sin(b\pi x)/\sqrt{(x-1/2)^2 + 1/2}.$$

Four different values of $b = 0, 10, 20$, and 40 were used in the simulation. For each simulated sample, the estimation accuracy of estimator was evaluated by the mean absolute deviation as above. We compared our estimator with a residual-based (Fan and Yao, 1998) and a difference-based ones (Brown and Levine, 2007) in the same manner as Example 1. The results are summarized in Fig. 3, in which the boxplots of $\mathcal{E}_{\mathrm{MAD}}$ are presented for each value of the four bs. (The corresponding numbers are in the supplement.) Our joint optimization method performs significantly better than the others for $b = 10$, 20, and 40. For $b = 0$, even though the accuracy of our method is markedly worse, the range between the first and third quartiles of the $\mathcal{E}_{MAD}$ overlaps with those of the other two methods. Also the median MAD values are quite small (between 0.05 and 0.1) for all the three methods. Thus the performance does not deteriorate as it might appear at first glance. The reason for this relatively worse performance is likely that the cubic B-spline basis does not contain a constant function, thus requires a precise linear combination to represent a perfectly flat function. Unlike Example 1, the difference-based method exhibits a better or comparable performance than the residual-based method. This is due to that the former is known to perform well when the true curve is close to non-smooth (Wang et al., 2008).

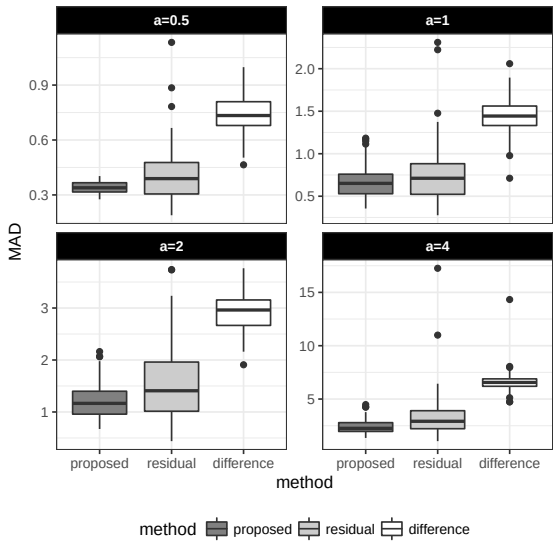

Figure 1: Boxplots of the mean absolute deviation $\mathcal{E}_{\mathrm{MAD}}$ from the estimation of the Sharpe ratio function $f(x) = a(X_i + 2\exp(-16X_i^2))/(0.4\exp(-2X_i^2) + 0.2)$ in Example 1.

5 Term structure modeling

In this section, we apply the proposed estimation method to interest rate modeling. The analyzed dataset consists of 1735 weekly observations of the yields of the three-month US Treasury Bill from the secondary market rates, taken from January 5, 1962 to March 31, 1995. These time series data are presented in the left panel of Fig. 4. The data have been analyzed by various authors, *e.g.*, Andersen and Lund (1997), Gallant and Tauchen (1997), and Fan and Yao (1998). Following Andersen and Lund (1997) and Fan and Yao (1998), we first fitted a fifth-order autoregressive model to the time series, denoted by z_t, and regressed the residuals, denoted by y_t, against $x_t \equiv z_{t-1}$.

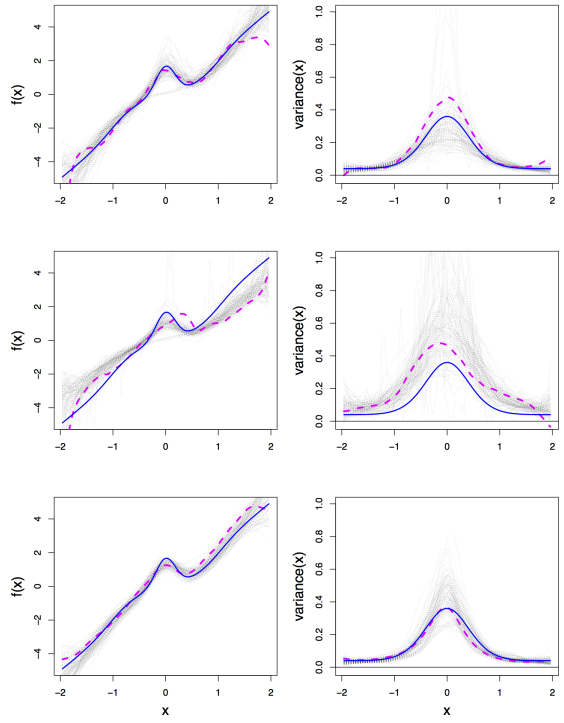

Figure 2: Simulation results for the estimation methods in Example 1 ($a = 0.5$). Top: residual-based method. Middle: difference-based method. Bottom: proposed method. Left: estimated regression curves (dotted gray) of the Sharp ratio function $f(x)$ (solid thick, blue curve). Right: estimated regression curves (dotted gray) of the variance function $\sigma^2(x)$ (solid thick, blue curve). An estimated variance function with a negative value is highlighted in the right, and so is the corresponding Sharpe ratio function (dashed thick, magenta curves) in the left.

The residuals are plotted against the yields x_t in the right panel of Fig. 4. The fitted AR(5) model coefficients are $(1.3252, -0.2800, -0.0263, 0.0276, -0.0472)$, thus the model that we estimate is

$$x_{t+1} - 1.3252x_t + 0.2800x_{t-1} + 0.0263x_{t-2} - 0.0276x_{t-3} + 0.0472x_{t-4} = y_t = \tilde{\mu}(x_t) + \tilde{\sigma}(x_t)\varepsilon_t,$$

in which we assume $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

The above equation can be considered as a discrete-time approximation to the continuous-time diffusion process model

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \tag{13}$$

where μ is the drift function, σ is the diffusion function, and W_t is the standard Brownian motion, which has

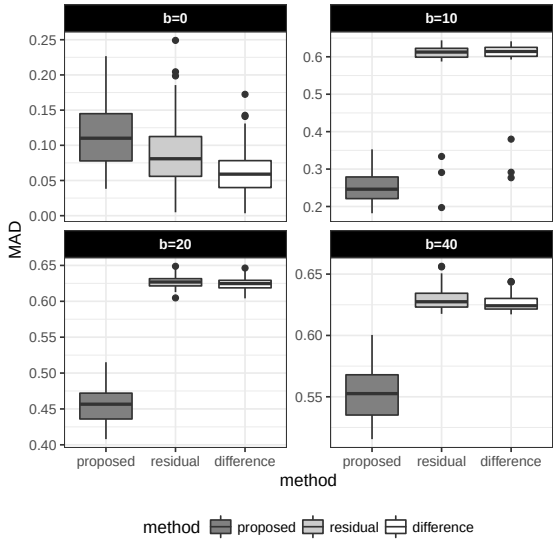

Figure 3: Boxplots of the mean absolute deviation $\mathcal{E}_{\text{MAD}}$ from the estimation of the Sharpe ratio function $f(x) = \frac{3}{4}\sin(b\pi x)/\sqrt{(x-1/2)^2 + 1/2}$ in Example 2.

been widely used to model the stochastic behavior of economic variables, such as interest rates, exchange rates, and stock prices; see, *e.g.*, Andersen and Lund (1997). In this case, μ represents the instantaneous expected rate of return, and σ the volatility.

We are particularly interested in nonparametric estimation of the Sharpe ratio function $f(x) = \mu(x)/\sigma(x)$ directly by solving (7). As Examples 1 and 2, we compare the results with those using a residual-based method (Fan and Yao, 1998) and a difference-based one (Brown and Levine, 2007, $r = 1$). Shown in Fig. 5 are the estimated Sharpe ratio functions (left column, thick blue curves) and the squared volatility, or conditional variance functions (right column, thick blue curves), together with the 95% two-sided pointwise confidence band obtained by using the regression bootstrap (dotted curves). These functions are scaled versions of $\mu(x)/\sigma(x)$ and $\sigma^2(x)$ in (13). It appears that the proposed method captures the well-known low price effect of the Sharpe ratio (Gilbertson et al., 1982; Branch and Chang, 1990), that low-priced assets outperform high-priced ones, more accurately than the others. Specifically, only the estimate by the proposed method shows that the Sharpe ratio function $f(x)$ increases as the price x gets close to 0, whereas the those by the other two methods are flat around $x = 0$. The estimated $\hat{f}(x)$ also suggests some resemblance to the famous Cox-Ingersoll-Ross model (Cox et al., 1985) for interest rate term structure, whose Sharpe ratio function has the form $f(x) = ax^{-1/2} + bx^{1/2}$.

Recall that for the proposed method $f(x)$ is estimated

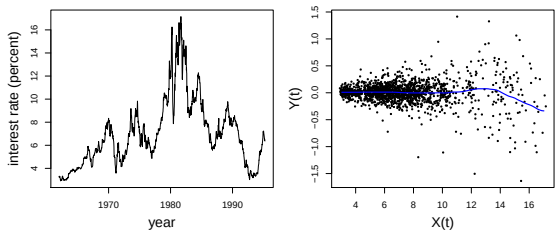

Figure 4: 3-month US Treasury Bill data. Left: time series plot of the yields (z_t). Right: scatter plot of residuals after an AR(5) model fit (y_t) vs $x_t = z_{t-1}$.

directly, while for the other methods, $\sigma^2(x)$ is estimated directly. As observed in Example 1, the two existing methods compared may result in negative variance, as indicated by the bootstrap confidence bands in the right panels of Fig. 5, yielding `NaN`s in some values of estimated $f(x)$ (discarded in the computation of the confidence interval).

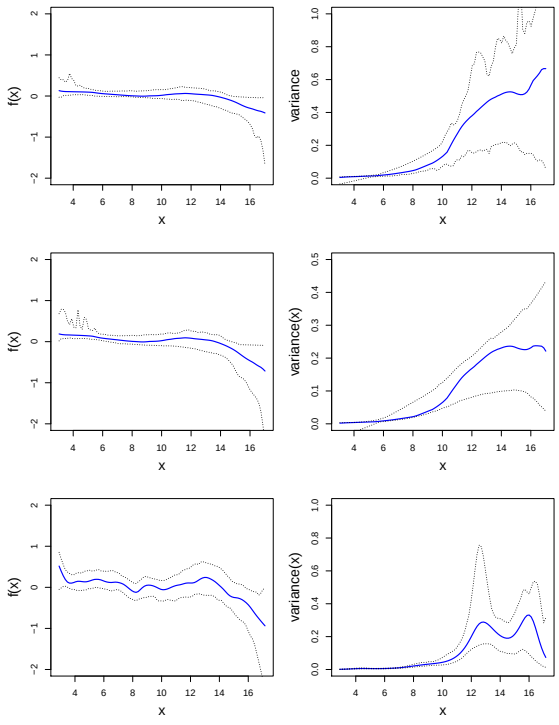

Figure 5: Estimation of the Sharpe ratio function from the 3-month US Treasury Bill data. Top: residual-based method. Middle: difference-based method. Bottom: proposed method. Left: estimated regression curve of the Sharp ratio function $f(x)$ (solid blue curve). Right: estimated regression curve of the variance function $\sigma^2(x)$ (solid blue curve).

6 Conclusions

We have described a method for estimating the Sharpe ratio function in heteroscedastic regression model (1) using a scale-separating "reparametrization." The resulting maximum likelihood estimation problem is convex in an appropriate function space; we have devised a method that solves this problem efficiently by using splines. The simulation results show that this method is suitable in estimating the Sharpe ratio function directly, rather than indirectly by combining separate estimates of the mean and variance functions. The results from the interest rate data analysis suggest that it is possible to apply our nonparametric technique to identify the underlying stochastic process in terms of the Sharpe ratio, capturing well-known empirical evidences.

A variety of extensions of the estimation method described in this paper can be considered. First, our method can be immediately extended to non-Gaussian cases (*i.e.*, z_i in (1) are iid but not Gaussian) as long as the employed loss function is jointly convex in f and g; squared loss of the standardized error (*i.e.*, (3) without the logarithm terms) and the Huber loss are two such examples. For the squared standardized error loss, in particular, the alternative method using nonnegative combinations of nonnegative polynomials of Section 3 has a computational advantage, as the problem reduces to a sequence of QPs. Second, we can easily add shape constraints such as monotonicity and convexity. For example, if we believe that the diffusion function is monotonically increasing, a constraint $g^{(1)}(x) \leq 0$ can be included; this constraint is linear for each $\mathcal{S}_m$. Indeed, shape-constrained function estimation is gaining interest in the statistics and machine learning communities as a way to bridge statistical estimation and deterministic optimization (Hannah and Dunson, 2013; Balázs, 2016; Mazumder et al., 2017; Guntuboyina and Sen, 2017). Third, in the above model, x_i could be multivariate. In this case, we can change the search space in (7) to a product space of reproducing kernel Hilbert spaces and replace the L_2 norms by appropriate ones. Proposition 1 should still hold by the representer theorem, while the nonnegativity constraint remains a subject of future research.

We hope that this paper spurs interest in shape-constrained estimation and related theoretical issues.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2014R1A4A1007895 and NRF-2017R1A2B2012264).

References

Andersen, T. and J. Lund (1997). Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics 77*, 343–377.

Balázs, G. (2016). *Convex Regression: Theory, Practice, and Applications*. Ph. D. thesis, University of Alberta.

Branch, B. and K. Chang (1990). Low price stocks and the January effect. *Quarterly Journal of Business and Economics*, 90–118.

Brown, L. D. and M. Levine (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics 35*(5), 2219–2232.

Byrd, R. H., J. Nocedal, and R. A. Waltz (2006). Knitro: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*, pp. 35–59. Springer.

Cai, T. and L. Wang (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics 36*(5), 2025–2054.

Cox, J. C., J. E. Ingersoll Jr, and S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica*, 385–407.

Dai, W., Y. Ma, T. Tong, and L. Zhu (2015). Difference-based variance estimation in nonparametric regression with repeated measurement data. *Journal of Statistical Planning and Inference 163*, 1–20.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. London: Chapman and Hall.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*, 645–660.

Gallant, A. R. and G. Tauchen (1997). Estimation of continuous-time models for stock returns and interest rates. *Macroeconomic Dynamics 1*(01), 135–168.

Gilbertson, R., J. Affleck-Graves, and A. Money (1982). Trading in low priced shares: An empirical investigation 1968–1979. *Investment Analysts Journal 11*(19), 21–29.

Grant, M. and S. Boyd (2014, March). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Guntuboyina, A. and B. Sen (2017). Nonparametric shape-restricted regression. *arXiv preprint arXiv:1709.05707*.

Hall, P. and R. Carroll (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society: Series B (Methodological) 51*(1), 3–14.

Hannah, L. A. and D. B. Dunson (2013). Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research 14*(1), 3261–3294.

Härdle, W. and A. Tsybakov (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics 81*, 223–242.

Leung, T., Q. Song, and J. Yang (2013). Outperformance portfolio optimization via the equivalence of pure and randomized hypothesis testing. *Finance and Stochastics 17*(4), 839–870.

Liu, A., T. Tong, and Y. Wang (2007). Smoothing spline estimation of variance functions. *Journal of Computational and Graphical Statistics 16*(2), 312–329.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal 58*(4), 36–52.

Lobo, M. S., L. Vandenberghe, S. Boyd, and H. Lebret (1998). Applications of second-order cone programming. *Linear algebra and its applications 284*(1), 193–228.

Maio, P. and P. Santa-Clara (2012). Multifactor models and their consistency with the ICAPM. *Journal of Financial Economics 106*(3), 586–613.

Mazumder, R., A. Choudhury, G. Iyengar, and B. Sen (2017+). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association* (to appear).

Papp, D. and F. Alizadeh (2014). Shape-constrained estimation using nonnegative splines. *Journal of Computational and Graphical Statistics 23*(1), 211–231.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(2), 365–375.

Schölkopf, B., R. Herbrich, and A. J. Smola (2001). A generalized representer theorem. In D. Helmbold and B. Williamson (Eds.), *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings*, pp. 416–426. Springer.

Schumaker, L. (1981). *Spline functions: basic theory*. Cambridge University Press.

Tang, Y. and R. F. Whitelaw (2011). Time-varying Sharpe ratios and market timing. *Quarterly Journal of Finance 1*(03), 465–493.

Toh, K.-C., M. J. Todd, and R. H. Tütüncü (1999). SDPT3–a MATLAB software package for semidefinite programming. *Optimization methods and software 11*(1-4), 545–581.

Tong, T., Y. Ma, Y. Wang, et al. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli 19*(5A), 1839–1854.

Turlach, B. A. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistics 20*(1), 81–104.

Utreras, F. I. (1985). Smoothing noisy data under monotonicity constraints: existence, characterization and convergence rates. *Numerische Mathematik 47*(4), 611–625.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wand, M. P. and M. C. Jones (1994). *Kernel smoothing*. CRC Press.

Wang, L., L. D. Brown, T. T. Cai, and M. Levine (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics 36*(2), 646–664.

Xu, K.-L. and P. C. Phillips (2011). Tilted nonparametric estimation of volatility functions with empirical applications. *Journal of Business & Economic Statistics 29*(4), 518–528.