risks

MDPI

Article

A Note on Combining Machine Learning with Statistical Modeling for Financial Data Analysis

José María Sarabia [1], Faustino Prieto [1], Vanesa Jordá [1] and Stefan Sperlich [2,*]

[1] Department of Economics, University of Cantabria, 39005 Santander, Spain; jose.sarabia@unican.es (J.M.S.); faustino.prieto@unican.es (F.P.); vanesa.jorda@unican.es (V.J.)

[2] Geneva School of Economics and Management, University of Geneva, 1211 Geneva, Switzerland

* Correspondence: stefan.sperlich@unige.ch; Tel.: +41-22-379-8223

Received: 19 December 2019; Accepted: 1 April 2020; Published: 3 April 2020

Abstract: This note revisits the ideas of the so-called semiparametric methods that we consider to be very useful when applying machine learning in insurance. To this aim, we first recall the main essence of semiparametrics like the mixing of global and local estimation and the combining of explicit modeling with purely data adaptive inference. Then, we discuss stepwise approaches with different ways of integrating machine learning. Furthermore, for the modeling of prior knowledge, we introduce classes of distribution families for financial data. The proposed procedures are illustrated with data on stock returns for five companies of the Spanish value-weighted index IBEX35.

Keywords: semiparametric modeling; machine learning; VaR estimation; analyzing financial data

JEL Classification: C14; C53; C58; G17; G22; C45

1. Introduction

The Editors of this Special Issue pointed out that machine learning (ML) has no unanimous definition. In fact, the term "ML", coined by Samuel (1959), is quite differently understood in the different communities. The general definition is that ML is concerned with the development of algorithms and techniques that allow computers to learn. The latter means a process of recognizing patterns in the data that are used to construct models; cf. "data mining" (Friedman 1998). These models are typically used for prediction. In this note, we speak about ML for data based prediction and/or estimation. In such a context, one may say that ML refers to algorithms that computer codes apply to perform estimation, prediction, or classification. As said, they rely on pattern recognition (Bishop 2006) for constructing purely data-driven models. The meaning of "model" is quite different here from what "model" or "modeling" means in the classic statistics literature; see Breiman (2001). He speaks of the data modeling culture (classic statistics) versus the algorithmic modeling coming from engineering and computer science. Many statisticians have been trying to reconcile these modeling paradigms; see Hastie et al. (2009). Even though the terminology comes from a different history and background, the outcome of this falls into the class of so-called semi-parametric methods; see Ruppert et al. (2003) or Härdle et al. (2004) for general reviews. According to that logic, ML would be a non-parametric estimation, whereas the explicit parametrization forms the modeling part from classic statistics.

Why should this be of interest? The Editors of this Special Issue also urge practitioners not to ignore what has already been learned about financial data when using presumably fully automatic ML methods. Regarding financial data for example, Buch-Larsen et al. (2005), Bolancé et al. (2012), Scholz et al. (2015, 2016), and Kyriakou et al. (2019) (among others) have shown the significant gains in estimation and prediction when including prior knowledge in nonparametric prediction. The first two showed how knowledge-driven data transformation improves nonparametric estimation of distribution and operational risk; the third paper used parametrically-guided ML for stock return prediction; the fourth imputed bond returns to improve stock return predictions; and the last proposed comparing different theory-driven benchmark models regarding their predictability. Grammig et al. (2020) combined the opposing modeling philosophies to predict stock risk premia. In this spirit, we will discuss the combination of purely data-driven methods with smart modeling, i.e., using prior knowledge. This will be exemplified along the analysis of the distributions (conditional and unconditional) of daily stock returns and calculations of their value-at-risk (VaR).

Notice finally that in the case of estimation, methods are desirable that permit practitioners to understand, maybe not perfectly, but quite well, what the method is doing to the data. This will facilitate the interpretation of results and any further inference. Admittedly, this is not always necessary and certainly also depends on the knowledge or imagination of the user. Yet, we believe it is often preferable to analyze data in a glass box than in a black box. This aspect is respected in our considerations.

Section 2 revisits the ideas of semiparametric statistics. Section 3 provides an intensive treatment of the distribution modeling followed by its combination with local smoothing. In Section 4, we give empirical illustrations. Section 5 concludes. In the Appendix are given additional details.

2. Preliminary Considerations and General Ideas

It is helpful to first distinguish between global and local estimation. Global means that the parameter or function applies at any point and to the whole sample. Local estimation applies only to a given neighborhood, like for kernel regression. It is clear that localizing renders a method much more flexible; however, the global part allows for an easy modeling, and its estimation can draw on the entire sample. Non-parametric estimators are not local by nature; for example, power series based estimators are not. Unless you want to estimate a function with discontinuities, local estimators are usually smoothing methods; see Härdle et al. (2004). This distinction holds also for complex methods (cf. the discussion about extensions of neural networks to those that recognize local features), which often turn out to be related to weighted nearest neighbor methods; see Lin and Jeon (2006) for random forests or Silverman (1984) for splines. The latter is already a situation where we face a mixture of global and local smoothing; another one is orthogonal wavelet series (Härdle et al. 1998).

Those mixtures are interesting because the global parts can borrow the strength from a larger sample and have a smoothing effect, while the local parts allow for the desired flexibility to detect local features. Power series can offer this only by including a (in practice unacceptable) huge number of parameters. This is actually a major problem of many complex methods, but mixtures allow substantially reducing this number. At the same time, they allow us to include prior knowledge about general features. For example, imposing shape restrictions is much simpler for mixtures (like splines) than it is for purely local smoothers; see Meyer (2008) and the references therein. Unless the number of parameters is pre-fixed, their selection happens via reduction through regularization, which can be implemented in many ways. Penalization methods like P-splines (Eilers et al. 2015) or LASSO (Tibshirani 1996) are popular. The corresponding problem for kernel, nearest neighbors, and related methods is the choice of the neighborhood size. In any case, one has to decide about the penalization criterion and a tuning parameter. The latter is until today an open question; presently, cross-validation-type methods are the most popular ones. For kernel based methods, see Heidenreich et al. (2013) and Köhler et al. (2014) for a review or Nielsen and Sperlich (2003) in the context of forecasting in finance.

The first question concerns the kind of prior information available, e.g., whether it is about the set of covariates, how they enter (linearly, additively, with interactions), the shape (skewness, monotonicity, number of modes, fat tails), or more generally, about smoothness. This is immediately followed by the question of how this can be included; in some cases, this is obvious (like if knowing the set of variables to be included); in some others, it is more involved (like including parameter information via Bayesian modeling). Knowledge about smoothness is typically supposed in order to justify a particular estimator and/or the selection method for the smoothing parameter. Information about the shape or how covariates enter the model comprises the typical ingredients of semiparametric modeling (Horowitz 1998) to improve nonparametric estimation (Glad 1998).

Consider the problem of estimating a distribution, starting with the unconditional case. In many situations, you are more interested in those regions for which data are hardly available. If you used then a standard local density estimator, you would try to estimate interesting parameters like VaR from only very few observations, maybe one to five, which is obviously not a good idea. Buch-Larsen et al. (2005) proposed to apply a parametric transformation using prior knowledge. Combining this way a local (kernel density) estimator with such a global one, however, allowed them to borrow strength from the model and from data that were further away. Similarly, consider conditional distributions. Locally, around a given value of the conditioning variable, you may have too few observations to estimate a distribution nonparametrically. Then, you may impose on this neighborhood the same probability law up to some moments, as we will do in our example below.[1]

Certainly, a good mixture of global and local fitting is problem-adapted. Then, the question falls into two parts: which is the appropriate parametric modeling, and how to integrate it with the flexible local estimator. For the former, you have to resort to expertise in the particular field. For the latter, we will discuss some popular approaches. All this will be exemplified with the challenges of modeling stock returns for five big Spanish companies and predicting their VaR.

In our example, the first step is to construct a parametric guide for the distribution of stock returns Y. To this aim, we introduce the class of generalized beta-generated (BG) distributions (going back, among others, to Eugene et al. (2002) and Jones (2004)), as this distribution class allows modeling skewed distributions with potentially long or fat tails. While this is not a completely new approach, we present it with an explicit focus on the above outlined objectives including the calculation of VaRs and combining it with nonparametric estimation and/or validation. Our validation is more related to model selection and testing, today well understood and established, and therefore kept short. The former, i.e., the combination with nonparametric estimation, is discussed for the problem of analyzing conditional distributions and can be extended to the combination with methods for estimating in high dimensions. For this example, in which the prior knowledge enters via a distribution class, we discuss two approaches: one is based on the method of moments, the other one on maximum likelihood. The latter is popular due to Rigby and Stasinopoulos (2005) and Severini and Staniswalis (1994). Rigby and Stasinopoulos (2005) considered a fully parametrized model in which each distribution parameter (potentially transformed with a known link) is written as an additive function of covariates, typically including a series of random effects. They proposed a backfitting algorithm (implemented in the R library GAMLSS) to maximize a penalized likelihood corresponding to a posterior mode estimation using empirical Bayesian arguments. Severini and Staniswalis (1994) started out with the parametric likelihood, but localized it by kernels. This is maximized then for some given values of the covariates.

3. A Practical Example

We now discuss the technical steps for the announced practical example. While this section focuses on the technical part, the empirical exercise is done in the next section.

3.1. Distribution Modeling

Often, the Student t distribution was used in financial econometrics and risk management to model the conditional asset returns, going back to Bollerslev (1987). However, it is well known that it does not describe very well the empirical features of most financial data. Therefore, several proposals have been

[1] Further advantages are that semiparametric modeling can help to overcome the curse of dimensionality and that semiparametric models are more robust to the choice of smoothing parameters.

made of skewed Student t distributions; see Theodossiou (1998) and Zhu and Galbraith (2010) for the context of finance or Jones and Faddy (2003) and Azzalini and Capitanio (2003) in (applied) statistics. For a more general discussion and compendium, see Rigby et al. (2019).

Let us consider two classes of skewed t distributions. Both are derived from the (generalized) BG classes; see Appendix A. These allow for the generation of many flexible distribution classes. The fist version depends on two parameters, whereas the second one depends on three. One may directly go for the second one. Remember, however, later on, that this is just a parametric guide for the semiparametric estimator. There is certainly a trade-off between the gain of flexibility, on the one hand, and the loss of its regularization, on the other hand. Moreover, each additional parameter in the global part may raise the costs of implementation and computation to an unacceptable degree. Both skewed t distributions are generated by taking a standard Student t as the baseline distribution in (A1) and (A2), respectively. Specifically, plugging in $F_Y(y;a,b) = \frac{1}{2}\left(1 + y/\sqrt{a+b+y^2}\right)$, a so-called scaled Student t_2, into (A1), gives our skewed t of Type 1 with the probability density function (pdf):

$$g_{T_1}(y;a,b) = k_1\left(1 + \frac{y}{\sqrt{a+b+y^2}}\right)^{a+1/2}\left(1 - \frac{y}{\sqrt{a+b+y^2}}\right)^{b+1/2}, \tag{1}$$

where $k_1^{-1} = B(a,b)\sqrt{a+b}2^{a+b-1}$ and $y \in \mathbb{R}$. We write $Y \sim T_1(a,b)$; see Jones and Faddy (2003). Plugging the baseline cumulative density (cdf) into (A2) gives our skewed t of Type 2 with the pdf:

$$g_{T_2}(y;a,b,c) = k_2\frac{1}{(a+b+y^2)^{3/2}}\left(1 + \frac{y}{\sqrt{a+b+y^2}}\right)^{ac-1}\left(1 - \frac{1}{2^c}\left(1 + \frac{y}{\sqrt{a+b+y^2}}\right)^c\right)^{b-1}, \tag{2}$$

where $k_2 = \frac{c(a+b)}{B(a,b)2^{ac}}$ and $y \in \mathbb{R}$. We write $Y \sim T_2(a,b,c)$; see Alexander and Sarabia (2010) and Alexander et al. (2012). The densities g_{T_1}, g_{T_2} are illustrated in Figure 1.

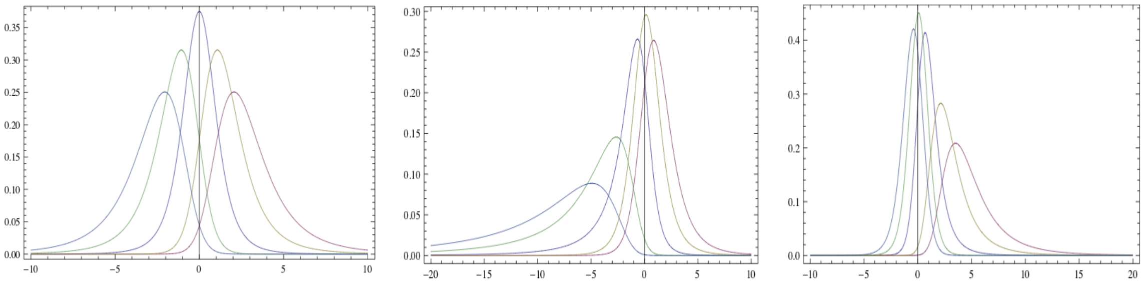

Figure 1. Graphics of the probability density function: left for the skewed $t1$ (1) when $(a,b) =$ (2,2), (5,2), (8,2), (2,5), and (2,8); center for the skewed $t2$ (2) with $(a,b,c) =$ (2,2,0.5), (8,2,0.5), (5,2,0.5), (2,5,0.5), and (2,8,0.5); and on the right, (2,2,2), (8,2,2), (5,2,2), (2,5,2), and (2,8,2).

For implementation, estimation, model selection, and in particular, for the possible integration of ML, we take a closer look at the basic properties of these two distributions. Note that for $a = b$ in (1), one obtains the classical Student t distribution with $2a$ degrees of freedom. The same is true for the three-parameter case (2) when setting $a = b$ with $c = 1$. Furthermore, the cdfs are given by:

$$G_{T_1}(y;a,b) = I\left(\frac{1}{2}\left\{1 + \frac{y}{\sqrt{a+b+y^2}}\right\};a,b\right), \tag{3}$$

and:

$$G_{T_2}(y;a,b,c) = I(F_Y^c(y;a,b);a,b), \tag{4}$$

respectively. Having explicit expressions for the pdf, one may think that likelihood based methods are straight-forward, including those for tests and model selection. Note, however, that for a sample of size m, you need to know the joint distribution of $(y_1, \ldots, y_m)$, which are not independent. In practice, many work with $\log \prod_{i=1}^{m} g_{T_k}(y_i; \theta)$, ($k = 1, 2$ with $\theta = (a, b)$ for $k = 1$, $\theta = (a, b, c)$ else) as a likelihood approximate, ignoring potential dependencies. This weakens both the efficiency of estimation and the validity of likelihood based inference. Therefore, it is interesting to look at alternatives. For estimation, recall the moment based approach; then, we need to to express parameters a, b, and if applicable c, in terms of estimable moments. For the case of the skewed t distribution of the first kind, we have:

$$E(Y^r) = \frac{(a+b)^{r/2}}{B(a,b)} \sum_{j=0}^{r} \binom{r}{j} 2^{-j} (-1)^j B\left(a - \frac{r}{2} - j, b - \frac{r}{2}\right) \tag{5}$$

if $a, b > r/2$. An advantage of this expression is that we do not need to estimate the centered, standardized moments. This is even more an advantage when we try to estimate these moments nonparametrically. Similarly, for the skewed t distribution of the second kind:

$$E(Y^r) = \frac{(a+b)^{r/2}}{B(a,b)} \sum_{j=0}^{r} (-1)^j \binom{r}{j} 2^{-j} \sum_{k=0}^{\infty} \binom{-r/2}{k} (-1)^k B\left(a - \frac{r/2 + j - k}{c}, b\right) . \tag{6}$$

Clearly, this reveals a problem because we cannot solve this (easily) for a, b, and c.

3.2. Financial Risk Measures

The parametric part provides us with explicit formulas that we can use to derive closed expressions for the risk measures, namely the VaR and tail moments. In fact, it can be shown that:

$$\mathrm{VaR}_{T_1}[p; a, b] = \frac{\sqrt{a+b}(2\mathrm{VaR}_B[p; a, b] - 1)}{2\sqrt{\mathrm{VaR}_B[p; a, b](1 - \mathrm{VaR}_B[p; a, b])}}, \text{ and} \tag{7}$$

$$\mathrm{VaR}_{T_2}[p; a, b, c] = \frac{\sqrt{a+b}(2\mathrm{VaR}_B^{1/c}[p; a, b] - 1)}{2\sqrt{\mathrm{VaR}_B^{1/c}[p; a, b](1 - \mathrm{VaR}_B^{1/c}[p; a, b])}}, \tag{8}$$

with $0 \le p \le 1$, where $\mathrm{VaR}_B[p; a, b]$ denotes the VaR of a classical $\mathcal{B}e(a, b)$ distribution. Furthermore, if $Y \sim T_1(a, b)$, then the for a given y_p, its corresponding tail moments can we written as:

$$\begin{aligned} E\{Y^k | Y \le y_p\} &= E\left\{ \frac{(a+b)^{k/2}(2B-1)^k}{2^k B^{k/2}(1-B)^{k/2}} \mid \frac{\sqrt{(a+b)}(2B-1)}{2\sqrt{B(1-B)}} \le y_p \right\} \\ &= \frac{(a+b)^{k/2}}{2^k} \sum_{j=0}^{k} a_j E\{B^{k/2-j}(1-B)^{k/2} \mid B \le h(x_p)\}, \end{aligned}$$

where $a_j = (-1)^j \binom{k}{j} 2^{k-j}$ and $h(z) = \{a + b + z^2 + \sqrt{(a+b)z^2 + z^4}\}/\{2(a + b + z^2)\}$, which is increasing in z, with $a, b \in \mathbb{R}^+$. For $k = 1$, we get the T(ail)VaR. If $B \sim \mathcal{B}e(a, b)$, then one has:

$$E\{B^{k/2-j}(1-B)^{k/2} \mid B \le h(x_p)\} = \frac{1 - F_{\tilde{B}}(h(y_p))}{F_B(h(y_p))} \frac{B(k/2 + a - j, b - k/2)}{B(a, b)}, \tag{9}$$

where $\tilde{B} \sim \mathcal{B}e(k/2 + a - j, b - k/2)$, $b > k/2$, and $h(z)$ as above. In sum, for the Type 1 class, we obtain explicit expressions for the tail moments that do not change when (a, b) are functions of covariates.

3.3. Combining The Prior with Nonparametric Estimation

First, let us briefly consider the case of being only interested in the estimation of the unconditional distribution $g(\cdot)$, say, in that of stock returns Y. Once a proper parametric choice (say G_θ) is found and the parameters θ are estimated, we know that $\tilde{Y} := G_{\hat{\theta}}(Y)$ has a pretty smooth density $\tilde{g}$, which is not too far from the uniform $[0,1]$ distribution. Then, as $g(y) = \tilde{g}\{G_{\hat{\theta}}(y)\}$, you obtain the final estimate. A next step to proceed could be to apply a nonparametric estimator like, for example, a kernel density for $\tilde{g}$. The prior estimation served here to stretch data where we had many observations, and contract them where we had only a few. You may say that this could also be done either by taking the empirical cdf for transformation or by taking the local bandwidth that always includes the same number of neighbors. This would be purely nonparametric and therefore renounce the use of prior information. In practice, these alternatives suffer from various problems like giving wiggly results, a much harder bandwidth choice, etc. For details and applications in actuarial sciences, consult Bolancé et al. (2012) and Martínez Miranda et al. (2009).

We turn now to the slightly more challenging problem of considering conditional distributions. For example, what is the stock return distribution of Telefónica and its VaR given a certain value of the Spanish IBEX35? Having huge datasets, you may try to estimate this with a fully nonparametric estimator. However, if you doubt that stationarity holds over a long period, you may want to restrict your dataset to not include more than twelve months (as an example). In such a case, you better resort to a parametric guide like above and turn a, b (and c if applicable) into flexible functions of the conditioning covariate X. Again, a likelihood based approach may now look at $\log prod_{i=1}^{m} g_{T_k}(y_i; \theta(x_i))$ ($k = 1, 2$ with $\theta(x_i) = (a(x_i), b(x_i))$ for $k = 1$, etc.) ignoring potential dependencies. Rigby and Stasinopoulos (2005) specified the elements of θ as additive, parametrized functions of x_i with some random coefficients, and maximized a penalized version of this. In contrast, Severini and Staniswalis (1994) did not parametrize the elements of θ, but maximized (along θ) the smoothed version, i.e., $sum_{i=1}^{m} \log g_{T_k}(y_i; \theta(x)) K_h(x - x_i)$ for any x. Here, $K_h(v) = h^{-1}K(v/h)$ for a kernel function $K(\cdot)$ with bandwidth h. That is, they obtained for given x an estimate of the value that θ takes at x. Notice that the latter method is usually implemented for θ containing only one element (typically the mean).

As said before, an interesting alternative is to estimate nonparametrically the moments of Y and then derive the elements of θ. Thanks to Formula (5), we can do this for the Type 1 skewed t distributions. More specifically, the algorithm would look as follows: For sample (x_i, y_i), $i = 1, 2, \ldots, m$ with y_i being the return of the stock and x_i the conditioning variable (in our application, the market return) conduct:

Step 1: Estimate the conditional first two moments μ_1 and μ_2 by:

$$\mu_1(x_i) = E(y|x_i), \quad \text{and} \quad \log \mu_2(x_i) = E(y^2|x_i)\ ,$$

where the nonparametric functions can be estimated, e.g., by kernel regression, splines, etc.

Step 2: You now may either take a grid over the range of X, with M grid points x_j, or you may calculate the estimates for each observation x_i. Let us call them $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$, $j = 1, 2, \ldots, M$ for either case.

Step 3: Calculate estimates $(\hat{a}_j, \hat{b}_j)$ by solving in (a_j, b_j), $j = 1, 2, \ldots, M$, the non-linear system:

$$\hat{\mu}_{1j} = \frac{(a_j - b_j)\sqrt{a_j + b_j}}{2} \cdot \frac{\Gamma(a_j - \frac{1}{2})\Gamma(b_j - \frac{1}{2})}{\Gamma(a_j)\Gamma(b_j)}, \tag{10}$$

$$\hat{\mu}_{2j} = \frac{a_j + b_j}{4} \cdot \frac{(a_j - b_j)^2 + a_j - 1 + b_j - 1}{(a_j - 1)(b_j - 1)}. \tag{11}$$

Step 4: Then, the estimate of the conditional distribution is of the form:

$$\hat{g}(y|x_j) = g_F(y|\hat{a}_j, \hat{b}_j)\ .$$

Following Nielsen and Sperlich (2003), Kyriakou et al. (2019), and Mammen et al. (2019), you could use local linear regression in Step 1 for both functions, combined with the validated R^2 for the bandwidth choice. Obviously, any method known as ML, including LASSO variable selection, can be applied in this step. However, it is less evident how these methods could be combined with the aforementioned likelihood based approaches. Note further that the conditional distribution obtained by this strategy can also be used for semiparametric prediction of the unconditional distribution:

$$\hat{g}(y) = \frac{1}{m} \sum_{i=1}^{m} g_F(y|\hat{a}_i, \hat{b}_i) \, . \tag{12}$$

This shows that with $\hat{\mu}_1$, $\hat{\mu}_2$, you can predict the marginal distribution of Y for scenarios in which the distribution of X is changing (Dai et al. 2016). For example, we could predict the unconditional distribution of stocks for different distributions of the IBEX35. Note finally that Step 3 cannot easily be applied to the skewed t distribution of Type 2; recall (6). In such a case, you could only try to apply the idea of Rigby and Stasinopoulos (2005), but with procedures and algorithms that are still to be developed.

4. Empirical Illustration

Consider daily stock returns from 1 January 2015 to 31.12.2015 for five companies of the Spanish value-weighted index IBEX35; namely Amadeus (IT solutions for tourist industry), BBVA (global financial services), Mapfre (insurance market), Repsol (energy sector), and Telefónica (information and communications technology services); see Table 1. The returns are negatively skewed in four of the five companies considered, but positively skewed in the last one.

Table 1. Summary statistics. The sample size is $m = 261$ for all sets.

Stocks	Amadeus	BBVA	Mapfre	Repsol	Telefónica
Maximum daily return	0.046286	0.040975	0.050847	0.073466	0.062264
Minimum daily return	−0.097367	−0.060703	−0.067901	−0.0877323	−0.051563
Mean	0.000900	−0.000452	−0.000623	−0.001416	−0.000408
Standard deviation	0.014601	0.016249	0.015942	0.021349	0.016301
Skewness	−1.163797	−0.465779	−0.723655	−0.166165	0.130372
Kurtosis	10.292160	3.824688	4.873980	5.435928	4.422885

We first fit the data by the maximum likelihood (ignoring dependence) to both distribution classes, working with standardized data. Tables 2 and 3 show the parameter estimates with standard errors.

Table 2. Maximum likelihood estimates for the skewed t model of Type 1, standardized data. Standard errors are in parenthesis.

Stocks	Amadeus	BBVA	Mapfre	Repsol	Telefónica
$\hat{a}$	6.194309	10.773980	7.271484	5.009976	7.083988
	(2.378890)	(8.474473)	(3.684818)	(1.980271)	(3.810294)
$\hat{b}$	6.171897	10.76088	7.250156	5.005015	7.086958
	(2.378415)	(8.477441)	(3.686769)	(1.980433)	(3.810390)

Table 3. Maximum likelihood estimates for the skewed t model of Type 2, standardized data. Standard errors are in parenthesis.

Stocks	Amadeus	BBVA	Mapfre	Repsol	Telefónica
$\hat{a}$	1.050617	0.935678	0.8685684	0.808572	1.120804
	(0.443072)	(0.407211)	(0.329048)	(0.363157)	(0.650834)
$\hat{b}$	5.126098	7.144007	6.217545	2.998354	3.497796
	(2.091549)	(5.104088)	(3.184219)	(0.917090)	(1.110353)
$\hat{c}$	2.973896	3.653026	3.617017	2.721519	2.331579
	(0.761879)	(0.733523)	(0.668503)	(0.698227)	(0.774010)

To choose between these two models, one may use the Bayesian information criterion (BIC). However, as our (working) likelihood neglects the dependence structure, it might not be reliable. While the three-parameter model presents the largest values for BIC (not shown), the gain, however, is always close to, or smaller than, 1%.

An alternative is to apply ML methods comparing the parametric estimates with purely nonparametric ones. This is not recommendable any longer when switching to conditional distributions due to the moderate sample sizes. In Figure 2, you see how our models (in red) adapted to the empirical cdf (blue) for the stocks of Amadeus and BBVA. As expected, the three-parameter model gave slightly better fits. In practice, the interesting thing to see is where improvements occurred, if any. The practitioner has to judge then what is of interest for his/her problem; ML cannot do this for him/her. However, ML can offer specification tests; see Gonzales-Manteiga and Crujeiras (2013) for a review. For example, a test that formalizes our graphical analysis is the Kolmogorov–Smirnov (KS) test:

$$KS = \sup |F_m(y_i) - F(y_i;\hat{\theta})|, \; i = 1, 2, \ldots, m,$$

where $F_m(y_i)$ is the empirical cdf and $F(y;\hat{\theta})$ is the cdf of the particular model class with $\hat{\theta}$ from Tables 2 and 3. To calculate the p-value, we can use the parametric bootstrap:

Step 1: For the observed sample, find the maximum likelihood estimator, $F(y;\widehat{\boldsymbol{\theta}})$, $\hat{F}_m(y)$, and KS.

Step 2: Generate J bootstrap samples $y_1^{(j)}, \ldots, y_m^{(j)} \sim F(y;\widehat{\boldsymbol{\theta}})$ under H_0 (the data follow model F); fit them; and compute $F(y;\widehat{\boldsymbol{\theta}}^{(j)})$, $\hat{F}_m^{(j)}(y)$, and $KS^{(i)}$ for each bootstrap sample $j = 1, 2, \ldots, J$.

Step 3: Calculate the p-value as the fraction of synthetic bootstrap samples with a KS statistic greater than the empirical KS statistic obtained from the original data.

To obtain an approximate accuracy of the p-value for ϵ =0.01, we generated $J = \frac{1}{4}\epsilon^{-2} = 2500$ bootstrap samples. Table 4 shows the results for both distribution classes and all datasets. It can be seen that with all p-values larger than 0.499, both models could not be rejected at any reasonable significance level in any of the considered datasets.

Finally, let us see how different the VaR are, when calculated on the base of one model compared to the other; recall Equations (7) and (8). They were calculated at the 95% confidence level for all datasets; see Table 5. The T_1 model with two parameters provided slightly higher VaR values than the T_2 model. The difference again seemed to be somewhat marginal, except maybe for Amadeus.

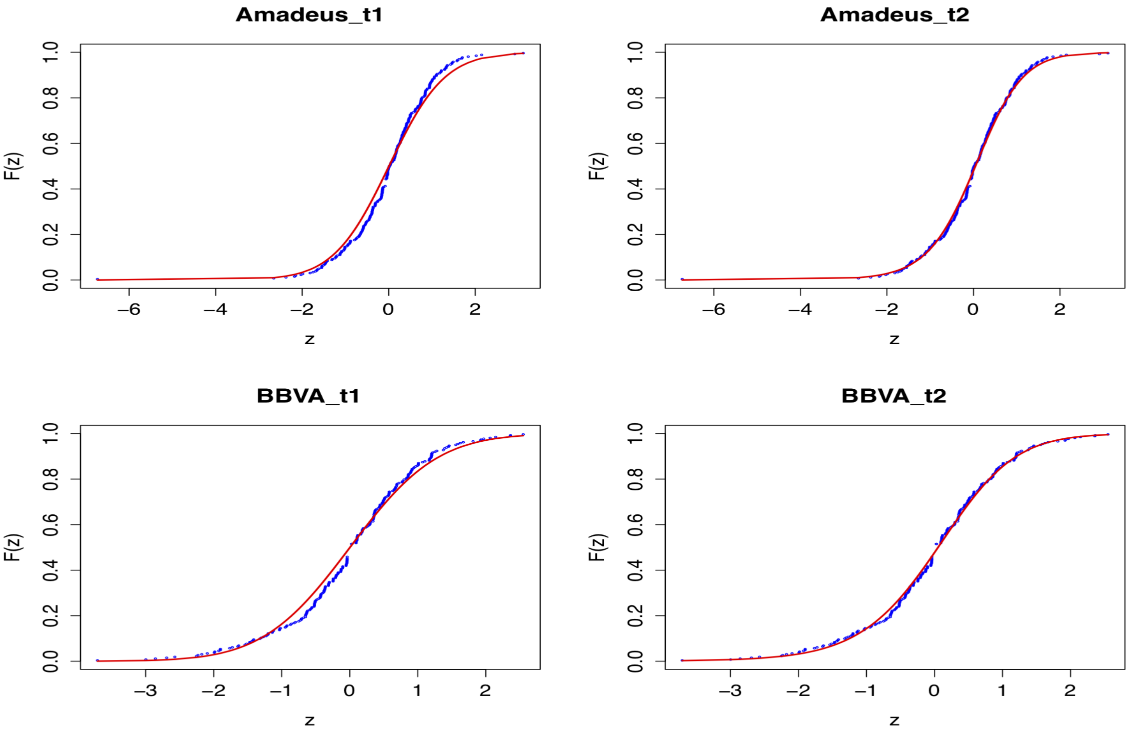

Figure 2. Plots of the theoretical cdfs of the skewed t models (LEFT: T_1 model; RIGHT: T_2 model) and the empirical cdf. Stocks: Amadeus; BBVA.

Table 4. Bootstrap p-values for both models and all five datasets.

Stocks	Amadeus	BBVA	Mapfre	Repsol	Telefónica
Skewed t T_1	0.593	0.676	0.499	0.761	0.829
Skewed t T_2	0.732	0.908	0.733	0.732	0.915

Table 5. Values at risk $\text{VaR}_{T1}[0.05; a, b)]$ and $\text{VaR}_{T2}[0.05; a, b, c]$ for the five stocks considered.

Stocks	Amadeus	BBVA	Mapfre	Repsol	Telefónica
VaR_{T1}	−0.024941	−0.028328	−0.028521	−0.040059	−0.029110
VaR_{T2}	−0.023089	−.02817	−0.027794	−0.038330	−0.028029

Let us turn to the estimation of the conditional distribution. Here, the integration of the ML happens by incorporating the covariates nonparametrically. For the sake of presentation and brevity, we restricted ourselves here to the moment based approach; for more details on the likelihood based one, we refer (besides the above cited literature) to the recent compendium of Rigby et al. (2019) and the references therein. Limiting ourselves to the moment based method automatically limited us to the skewed $t1$ class; recall Equation (6).[2] Furthermore, for the sake of illustration, we limited the exercise to the estimation of all distribution parameters as nonparametric functions of one given covariate X, namely the IBEX35. It was obvious that estimates for μ_1, μ_2 could equally well be the result of a complex multivariate regression or a variable selection procedure like LASSO. We estimated μ_1(IBEX35) and μ_2(IBEX35) using different methods provided by standard software; the presented results were obtained from penalized (cubic) spline regression with data-driven penalization. For details,

[2] You may develop numerical approximations working with (6), but this is clearly beyond the scope of this note. However, the above studies insinuate that the gain by using the more complex Type 2 class is rather marginal. Those advantages get easily compensated by the local estimator.

consult Ruppert et al. (2003) and the SemiPar project. Figure 3 gives an example of how this performed for the BBVA stocks.

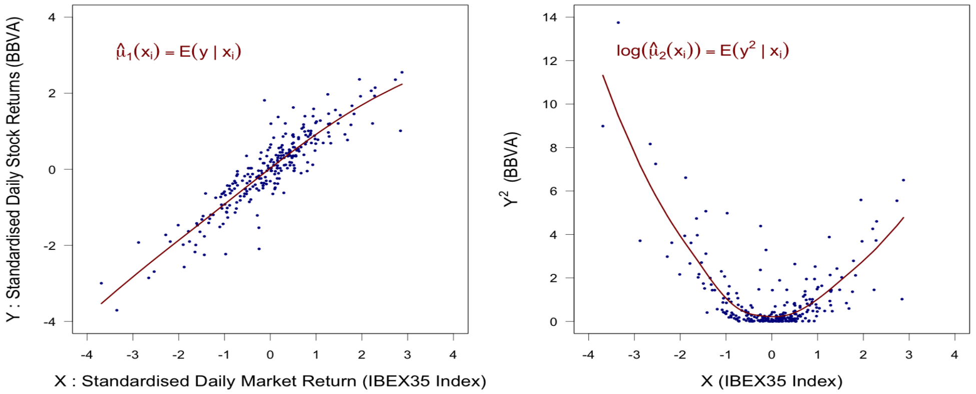

Figure 3. Estimates $\hat{\mu_1}$, $\widehat{\log \mu_2}$ for BBVA stock returns as functions of IBEX35.

Table 6 gives for the IBEX35 quantiles $Q_1 = -0.00768$, $Median = Q_2 = 0.00079$, $Q_3 = 0.00687$ the corresponding moment estimates of the different stock returns; Table 7 the corresponding (a_j, b_j) for Formulas (10) and (11). Table 8 gives the corresponding conditional VaRs obtained from Formula (7).

Table 6. First two moments of stock returns for given IBEX35 values (looking at its quantiles).

IBEX35	Amadeus		BBVA		Mapfre		Repsol		Telefónica	
Quartile	$\hat{\mu}_{1j}$	$\hat{\mu}_{2j}$	$\hat{\mu}_{1j}$	$\hat{\mu}_{2j}$	$\hat{\mu}_{1j}$	$\hat{\mu}_{2j}$	$\hat{\mu}_{2j}$	$\hat{\mu}_{1j}$	$\hat{\mu}_{2j}$	$\hat{\mu}_{1j}$
Q_1	−0.323500	2.336970	−0.493668	1.489340	−0.481228	2.215213	−0.430198	1.654110	−0.509551	1.941169
Q_2	0.025464	2.385443	0.094427	1.240531	0.060678	1.487417	0.056462	1.474321	0.034497	1.253373
Q_3	0.280358	2.492818	0.503731	1.523203	0.439138	1.623399	0.405799	1.698132	0.433284	1.421627

Table 7. Parameter (a_j, b_j) of the conditional stock return distributions for given IBEX35 values.

IBEX35	Amadeus		BBVA		Mapfre		Repsol		Telefónica	
Quartile	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$
Q_1	1.742906	2.136783	5.397247	6.904428	1.990996	2.690554	3.143186	4.048275	2.490487	3.398563
Q_2	1.736629	1.709150	5.492494	5.226327	3.133551	3.019128	3.184576	3.076590	5.016808	4.924298
Q_3	1.953575	1.639172	6.478997	5.008818	4.327494	3.356619	3.640170	2.851049	6.809153	5.483973

Table 8. The conditional value at risk $VaR_{T1}(IBEX35)$ for given IBEX35 values.

IBEX35	Amadeus	BBVA	Mapfre	Repsol	Telefónica
Q_1	−0.037728	−0.038910	−0.044802	−0.053801	−0.043939
Q_2	−0.031199	−0.027981	−0.030278	−0.041117	−0.029327
Q_3	−0.026029	−0.020886	−0.022744	−0.032601	−0.021913

In Figure 4, you can see the entire conditional distributions for the three IBEX35 quantiles. Finally, in Figure 5, we plotted the resulting unconditional distributions when you integrate over the observed IBEX35 values; recall Equation (12). They reflect quite nicely the asymmetries and some fat tails.

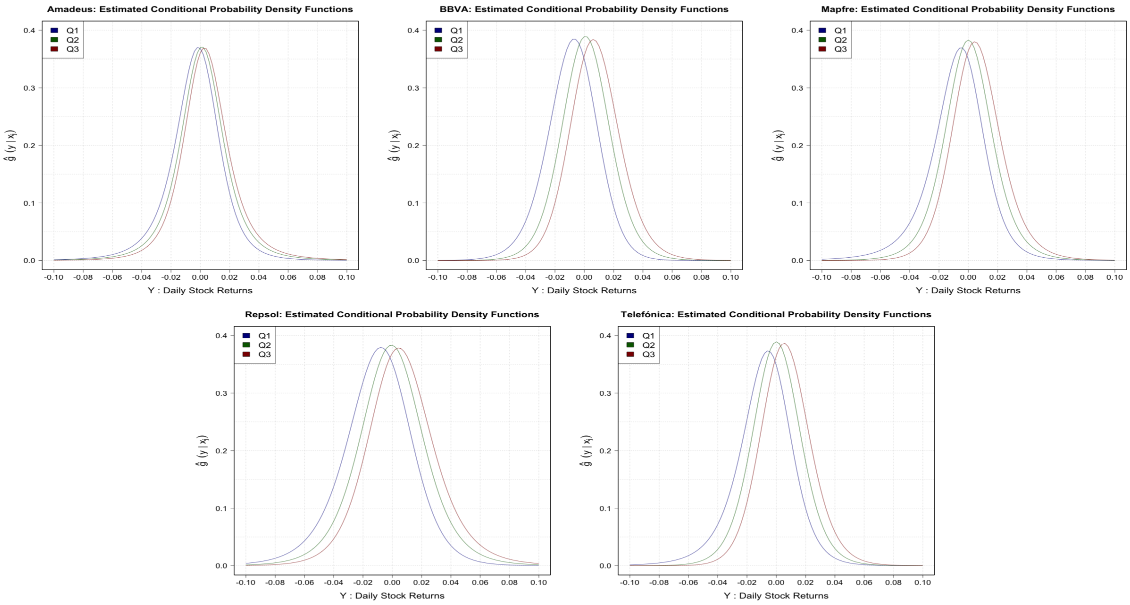

Figure 4. Conditional densities of stock returns at the quantiles of IBEX35 for Amadeus (**upper left**), BBVA (**upper center**), Mapfre (**upper right**), Repsol (**lower left**), and Telefónica (**lower right**).

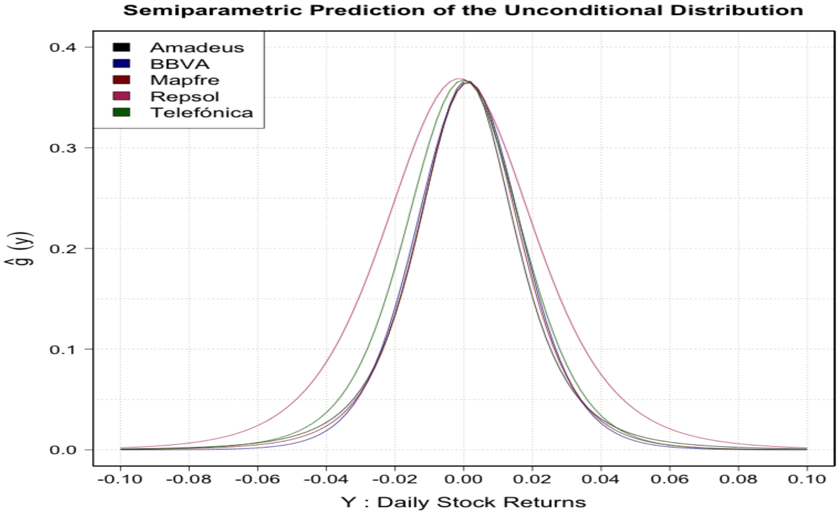

Figure 5. Unconditional densities of stock returns obtained from integrating the conditional ones over all observed IBEX35 values.

5. Discussion and Conclusions

In this note, we revisited the ideas of semiparametric modeling to propose for ML what one could call glass box modeling. It integrates ML in a mixture of a global and a local part in which the global one is as a parametric guide for the nonparametric estimate. We discussed different advantages of such a kind of smart modeling and the steps to be performed. In our illustration (analyzing financial data), we proposed as a parametric guide some (generalized) beta-generated distributions. In particular, we considered two classes of skewed t distributions. This allowed us to work with analytical expressions for the pdf, cdf, moments, and quantile functions, including the VaR, even on a local level. An empirical application with five datasets of stock returns was performed for illustration.

Author Contributions: All authors (J.M.S., F.P., V.J., S.P.) contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the Institute and Faculty of Actuaries in the U.K. for funding their research through the grant "Minimizing Longevity and Investment Risk while Optimizing Future Pension Plans" and the Spanish Ministerio de Economía y Competitividad, Project ECO2016-76203-C2-1-P, for partial support of this work.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Some Classes of Beta-Generated Distributions

This is to recall the class of BG distributions and to present some basic properties of this class. Let us call pdf $f(y)$ the baseline probability function with $F(y)$ being the corresponding cdf. The class of BG distributions is defined in terms of the pdf by:

$$g_F(y;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} f(y)F(y)^{a-1}[1-F(y)]^{b-1}, \tag{A1}$$

where $\Gamma(y)$ denotes the gamma function and $a, b > 0$ some real numbers. If for m being the sample size, we set $a = i$ and $b = m - i + 1$ in (A1), one obtains the pdf of the ith order statistic from F (Jones 2004). For $a \neq b$, we obtain a family of skewed distributions, for $a = m$ and $b = 1$ the distribution of the maximum, for $a = 1$ and $b = m$ the one of the minimum, and for $a = b = 1$ obviously $g_F = f$. In our context, the first property is the most interesting one. The parameters a and b control the tailweight of the distribution, where a controls the left-hand and b the right-hand tailweight. Consequently, for $a = b$, one obtains a symmetric sub-family, but still with $a = b$ controlling the tailweight. The BG distribution accommodates several kinds of tails like potential and exponential ones (Jones 2004).

For the moment method and in order to express the VaR in terms of moments or (a, b), we need to relate a, b to (directly) estimable moments. Let us now denote the cdf associated with (A1) by $G_F(y; a, b) = I(F(y); a, b)$ with $I(F(y); \cdot, \cdot)$ denoting the incomplete beta-function ratio:

$$B_x(a,b)/B_1(a,b)\ , \text{ where } B_y(a,b) = \int_0^y t^{a-1}(1-t)^{b-1}dt$$

where $0 \leq y \leq 1$, such that $B_1(a,b) = (\Gamma(a)\Gamma(b)/\Gamma(a+b)$. For a random variable $B \sim \mathcal{B}e(a,b)$, i.e., following the classical beta distribution, a simple stochastic representation of (A1) is $Y = F^{-1}(B)$. This allows for a direct simulation of the values of a random variable with pdf (A1). The raw (i.e., not centered, not normalized) moments of a BG distribution can be obtained by:

$$E[Y^r] = E[\{F^{-1}(B)\}^r] \qquad \text{for integers } r > 0\ .$$

Recently, some extensions have been proposed, e.g., by Alexander and Sarabia (2010), Alexander et al. (2012) and Cordeiro and de Castro (2011), of which we consider the one towards three parameters: The generalized BG (GBG) distribution is defined for $a, b, c > 0$ by the pdf:

$$g_F(y;a,b,c) = \frac{c\Gamma(a+b)}{\Gamma(a)\Gamma(b)} f(y)F(y)^{ac-1}[1-F(y)^c]^{b-1}. \tag{A2}$$

For $c = 1$, we get the BG distribution; for $a = c = 1$, one obtains the so-called proportional hazard model; and setting $a = 1$ yields the so-called Kumaraswamy generated distribution.

References

Alexander, Carol, and José-María Sarabia. 2010. Generalized Beta-Generated Distributions. In *ICMA Centre Discussion Papers in Finance DP2010-09*. Reading: ICMA Centre.

Alexander, Carol, Gauss M. Cordeiro, Edwin M. M. Ortega, and José-María Sarabia. 2012. Generalized beta-generated distributions. *Computational Statistics & Data Analysis* 56: 1880–97.

Azzalini, Adelchi, and Antonella Capitanio. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society. Series B* 65: 367–89. [CrossRef]

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Heidelberg: Springer.

Bolancé, Catalina, Montserrat Guillén, Jim Gustafsson, and Jens Perch Nielsen. 2012. *Quantitative operational Risk Models*. New York: Chapman & Hall/CRC Finance.

Bollerslev, Tim. 1987. A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics* 69: 542–47. [CrossRef]

Breiman, Leo. 2001. Statistical Modeling: The Two Cultures. *Statistical Science* 16: 199–231. [CrossRef]

Buch-Larsen, Tine, Jens Perch Nielsen, Montserrat Guillén, and Catalina Bolancé. 2005. Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics* 39: 503–18. [CrossRef]

Cordeiro, Gauss M., and Mário de Castro. 2011. A new family of generalized distributions. *Journal of Statistical Computation and Simulation* 81: 883–93. [CrossRef]

Dai, Jing, Sefan Sperlich, and Walter Zucchini. 2016. A simple method for predicting distributions by means of covariates with examples from welfare and health economics. *Swiss Journal of Economics and Statistics* 152: 49–80. [CrossRef]

Eilers, Paul H. C., Brian D. Marx, and Maria Durbán. 2015. Twenty years of P-splines. *Statistics and Operation Research Transactions* 39: 149–86.

Eugene, Nicholas, Carl Lee, and Felix Famoye. 2002. The beta-normal distribution and its applications. *Communications in Statistics: Theory Methods* 31: 497–512. [CrossRef]

Friedman, Jerome H. 1998. Data Mining and Statistics: What's the connection? *Computing Science and Statistics* 29: 3–9.

Glad, Ingrid K. 1998. Parametrically guided non-parametric regression. *Scandinavian Journal of Statistics* 25: 649–68. [CrossRef]

Gonzales-Manteiga, Wenceslao, and Rosa M. Crujeiras. 2013. An updated review of goodness-of-fit tests for regression models. *Test* 22: 361–411. [CrossRef] [PubMed]

Grammig, Joachim, Constantin Hanenberg, Christian Schalg, and Jantje Sönksen. 2020. *Diverging Roads: Theory-Based vs. Machine Learning-Implied Stockrisk Premia*. Tübingen, Germany: University of Tübingen Working Papers in Business and Economics, No 130, University of Tübingen.

Härdle, Wolfgang, Gérard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. 1998. *Wavelets, Approximation, and Statistical Applications*. Heidelberg: Springer.

Härdle, Wolfgang, Marlene Müller, Stefan Sperlich, and Alexander Werwatz. 2004. *Nonparametric and Semiparametric Models*. Heidelberg: Springer. [CrossRef]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Heidelberg: Springer.

Heidenreich, Niels-Bastian, Anja Schindler, and Stefan Sperlich. 2013. Bandwidth Selection Methods for Kernel Density Estimation: A review of fully automatic selectors. *AStA Advances in Statistical Analysis* 97: 403–33.

Horowitz, Joel L. 1998. *Semiparametric Methods in Econometrics.* Heidelberg: Springer.

Jones, M. Chris, and M.J. Faddy. 2003. A Skew Extension of the t-Distribution, with Applications. *Journal of the Royal Statistical Society. Series B* 65: 159–74. [CrossRef]

Jones, M. Chris. 2004. Families of distributions arising from distributions of order statistics. *Test* 13: 1–43.

Köhler, Max, Anja Schindler, and Stefan Sperlich. 2014. A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. *International Statistical Review* 82: 243–74. [CrossRef]

Kyriakou, Ioannis, Parastoo Mousavi, Jens Perch Nielsen, and Michael Scholz. 2019. Forecasting benchmarks of long-term stock returns via machine learning. *Annals of Operations Research* doi:10.1007/s10479-019-03338-4. [CrossRef]

Lin, Yi, and Yongho Jeon. 2006. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association* 101: 578–90. [CrossRef]

Mammen, Enno, Jens Perch Nielsen, Michael Scholz, and Stefan Sperlich. 2019. Conditional Variance Forecasts for Long-Term Stock Returns. *Risks* 7: 113. [CrossRef]

Martínez Miranda, M. Dolores, Jens Perch Nielsen, and Stefan Sperlich. 2009. One Sided Cross Validation for Density Estimation. In *Operational Risk Towards Basel III: Best Practices and Issues in Modeling, Management and Regulation*. Edited by Greg N. Gregoriou. Hoboken: John Wiley and Sons, pp. 177–96.

Meyer, Mary C. 2008. Inference using shape-restricted Regression Splines. *Annals of Applied Statistics* 2: 1013–33. [CrossRef]

Nielsen, Jens Perch, and Stefan Sperlich. 2003. Prediction of stock returns: A new way to look at it. *ASTIN Bulletin* 33: 399–417. [CrossRef]

Rigby, Robert A., and D. Mikis Stasinopoulos. 2006. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C* 54: 507–54. [CrossRef]

Rigby, Robert A., Mikis D. Stasinopoulos, Gillian Z. Heller, and Fernanda De Bastiani. 2019. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. New York: Chapman & Hall/CRC Finance.

Ruppert, David, Matt P. Wand, and Raymond J. Carroll. 2003. *Semiparametric Regression*. Cambridge: Cambridge University Press.

Samuel, Arthur. 2006. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3: 210–29. [CrossRef]

Scholz, Michael, Jens Perch Nielsen, and Stefan Sperlich. 2015. Nonparametric prediction of stock returns based on yearly data: The long-term view. *Insurance: Mathematics and Economics* 65: 143–55. [CrossRef]

Scholz, Michael, Stefan Sperlich, and Jens Perch Nielsen. 2016. Nonparametric long term prediction of stock returns with generated bond yields. *Insurance: Mathematics and Economics* 69: 82–96. [CrossRef]

Severini, Thomas A., and Joan G. Staniswalis. 1994. Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* 89: 501–11. [CrossRef]

Silverman, Bernard W. 1984. Spline Smoothing: The Equivalent Variable Kernel Method. *Annals of Statistics* 12: 898–916. [CrossRef]

Theodossiou, Panayiotis. 1998. Financial Data and the Skewed Generalized T Distribution. *Management Science* 44: 1650–61. [CrossRef]

Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B* 58: 267–88. [CrossRef]

Zhu, Dongming, and John Galbraith. 2010. A generalized asymmetric Student-t distribution with application to financial econometrics. *Journal of Econometrics* 157: 297–305.