Linex and double-linex regression for parameter estimation and forecasting

Mike G. Tsionas*

Abstract

The choice of an estimation method has received considerable attention in the Operations Research literature. In this paper we depart from the standard use of linex and double-linex loss functions which are widely used in parameter estimation and forecasting problems and we propose a non-standard use for them. Specifically, we propose to use the corresponding linex and double-linex error densities as models for the errors of a regression problem when more emphasis should be placed on over-estimation or under-estimation of errors. The new techniques are applied to synthetic as well real data concerning the role of management in production as well as to an application of forecasting volatility in intradaily data.

Key Words: Decision Analysis; Linex loss functions; Regression Problems; Estimation; Forecasting.

Acknowledgments: The author is grateful to two anonymous reviewers for many useful comments on an earlier version.

*Montpellier Business School, France & Lancaster University Management School, LA1 4YX, U.K., m.tsionas@montpellier-bs.com

1 Introduction

The choice of estimation method in regressions is a time-honored problem and has been considered extensively in Operations Research and related fields (Bottmer et al., 2021; Liang et al., 2021; Ulrich et al., 2019; Bajestani et al., 2017, inter alia). The problem we address is how estimation should be addressed when (in terms of forecasting for example) we do not have a quadratic loss function which would yield the least squares solution. For example, overestimating and underestimating a forecast relative to the true value may have different costs for the decision maker, a fact that should be reflected in his / her objective function for estimation.

Linex loss functions proposed by Varian (1975) are asymmetric loss or cost functions which have been studied extensively in the area of forecasting as well as optimal parameter estimation, see Zellner (1986), Granger, 1999), Christoffersen and Diebold (1997) and Hwang et al. (2001). "The linex loss function introduces preference asymmetries by being approximately exponential for under- (over-) forecasting and approximately linear for over- (under-) forecasting, depending on the sign of its parameter" (Christodoulakis, 2005). The double-linex loss function allows for more flexibility as it contains two parameters and it can, therefore, represent preferences over forecast errors more accurately. To summarize the problem let $\hat{\theta}$ denote the estimate of a parameter θ and $\Delta = \hat{\theta} - \theta$ denote the estimation error. Varian (1975) introduced the linex loss function $\Lambda(\Delta) = \tau \left(e^{a\Delta} - a\Delta - 1\right)$ where $a \neq 0$, $\tau > 0$, and τ is a scale parameter. For $a > 0$ the loss function can be quite asymmetric reflecting the fact that overestimation is more costly than underestimation. For $|a|$ near zero, since $e^{a\Delta} \simeq 1 + a\Delta + \frac{1}{2}a^2\Delta^2$, we have $\Lambda(\Delta) \simeq \frac{1}{2}a^2\Delta^2$ corresponding to a quadratic loss function. However, when $|a|$ assumes appreciable values, optimal point estimates and predictions will be quite different from those obtained with a symmetric squared error loss function" (Zellner, 1986, p. 446). For different values of a, the linex loss functions are presented in Figure 1 (for $\tau = 1$).

With data Y, if the posterior of the parameter is $p(\theta|Y)$,[1] a point estimate relative to the linex cost function is based on minimizing expected posterior loss $\mathbb{E}_{\theta|Y}\Lambda(\hat{\theta} - \theta)$ with respect to $\hat{\theta}$.

The use of linex loss functions as models for the data themselves has not been considered before although there is clearly scope for such considerations. For example, in forecasting problems the objective is to determine optimal forecasts $\hat{y}_{T+h}$ of a quantity y_{T+h} where T is the number of observations and h is the forecast horizon. In forecasting problems, the objective is, therefore, to minimize expected loss $\Lambda(\varepsilon_{T+h})$ where

$$\varepsilon_{T+h} = \hat{y}_{T+h} - y_{T+h}, \tag{1}$$

is the forecast error. Corresponding to quadratic loss, $\Lambda(\varepsilon) \propto \varepsilon^2$ and in relation to a regression problem, say $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ $(i = 1, \ldots, n)$, where $\mathbf{x}_i \in \mathbb{R}^k$ is a vector of explanatory variables whose coefficients are $\beta \in \mathbb{R}^k$, this corresponds to a normal distribution for the ε_is yielding the least squares (LS) estimator as the maximum likelihood (ML) estimator in this instance.

An interesting issue is a previously unconsidered problem in the literature, viz. considering linex distributions resulting from linex cost functions. In this study we consider this problem for the linex as well as the more general double-linex loss

[1] If the likelihood function is $L(\theta; Y)$ and the prior is $p(\theta)$ then by Bayes' theorem the posterior distribution has density $p(\theta|Y) \propto L(\theta; Y)p(\theta)$. Including the normalizing constant, we have $p(\theta|Y) \propto \frac{L(\theta;Y)p(\theta)}{\int_\Theta L(\vartheta;Y)p(\vartheta)\,\mathrm{d}\vartheta}$, where Θ is the parameter space.

functions and we show that parameters like a can be estimated from the data. The reason that the data is informative about such parameters is that the ε_is themselves may be skewed and the direction of skewness may be unknown. Of course, parameter a can be fixed in advance when the researcher wants to place more emphasis on under-estimation or over-estimation of errors. The problem is important and arises, inter alia, in engineering control problem where the trade-off between quality and quantity is important (Chang and Hung, 2007 building on earlier work by Chen and Chou (2004), Huang (2001), and Taguchi (1986). Another area where skewness aspects are important include stochastic frontier analysis for efficiency and productivity analysis (for a detailed presentation, see Kumbhakar and Lovell, 2000).

We show that the extension of linex or double-linex cost functions to error densities allows interesting likelihood functions (and posterior distributions in a Bayesian context) and that ML estimation reduces to orthogonality between the regressors $\mathbf{x}_i$ and "generalized residuals" which are functions of the ε_is. This opens up the way for wider scope analyses when the regressors fail to satisfy such orthogonality conditions, in which case these orthogonality conditions can be replaced by instruments resulting in instrumental variables (IV) or Generalized Method of Moments (GMM) estimation of the parameters of the model.

2 Linex loss functions and densities

The linex loss function is defined as

$$\Lambda(\varepsilon) = \tau\left(e^{a\varepsilon} - a\varepsilon - 1\right),\ a \neq 0,\ \tau > 0,\ \varepsilon \in \mathbb{R}. \tag{2}$$

We can set $\tau = 1$ although, in some cases, we retain τ for generality. Clearly, $\Lambda(0) = 0$ but other than that it is asymmetric in errors ε depending on the value of the parameter a. Although we have to exclude the case $a = 0$, when $|a| \to 0$ then we obtain a quadratic loss function as we showed in the previous section.

Any loss function can be converted to a likelihood function or density (Chernozhukov and Hong, 2003):

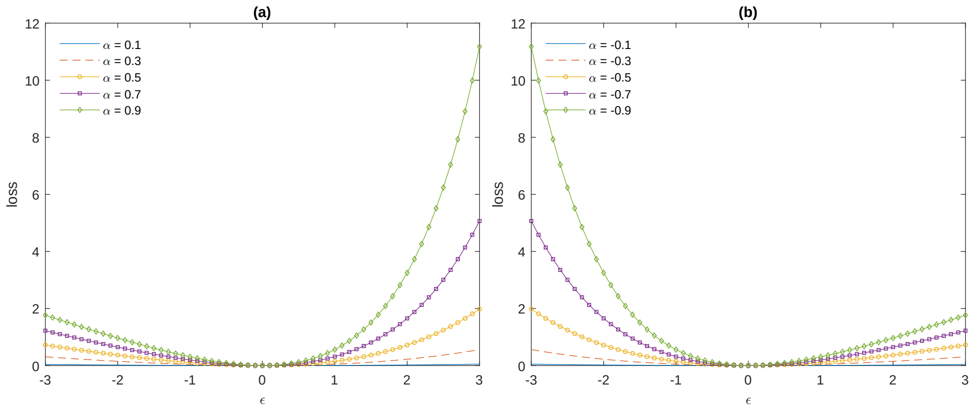
$$f(\varepsilon) = \frac{e^{-\Lambda(\varepsilon)}}{\int_{-\infty}^{\infty} e^{-\Lambda(\epsilon)}\,\mathrm{d}\epsilon}, \tag{3}$$

provided the integral in the denominator converges.

The constant of integration can be computed in closed form and we have to evaluate

$$\int_{-\infty}^{\infty} e^{-\tau\left(e^{a\varepsilon} - a\varepsilon - 1\right)}\,\mathrm{d}\varepsilon. \tag{4}$$

Figure 1: Linex loss functions



Using the change of variables $e^{a\varepsilon} = u$, the integral becomes

$$\tfrac{1}{a}\int_0^\infty e^{-\tau(u-\ln u-1)}\,\tfrac{1}{u}\,\mathrm{d}u = \tfrac{e^\tau}{a}\int_0^\infty u^{\tau-1}e^{-\tau u}\,\mathrm{d}u = \tfrac{e^\tau\Gamma(\tau)}{a\,\tau^\tau}, \tag{5}$$

using properties of the gamma distribution, where $\Gamma(\cdot)$ denotes the gamma function. Therefore, the full form of the density is

$$f(\varepsilon) = \tfrac{a\,\tau^\tau}{\Gamma(\tau)}e^{-\tau(e^{a\varepsilon}-a\varepsilon)}. \tag{6}$$

Suppose $\mathbf{x}_i \in \mathbb{R}^k$ is a vector of explanatory variables whose coefficients are $\beta \in \mathbb{R}^k$ and

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i,\; i = 1,\ldots,n. \tag{7}$$

We denote the data $Y = [y_i, x_i, i = 1,\ldots,n]$. In the following analysis we depart from the usual definition of error in linex loss function which would have been $\varepsilon_i = \mathbf{x}_i'\beta - y_i$.[2]

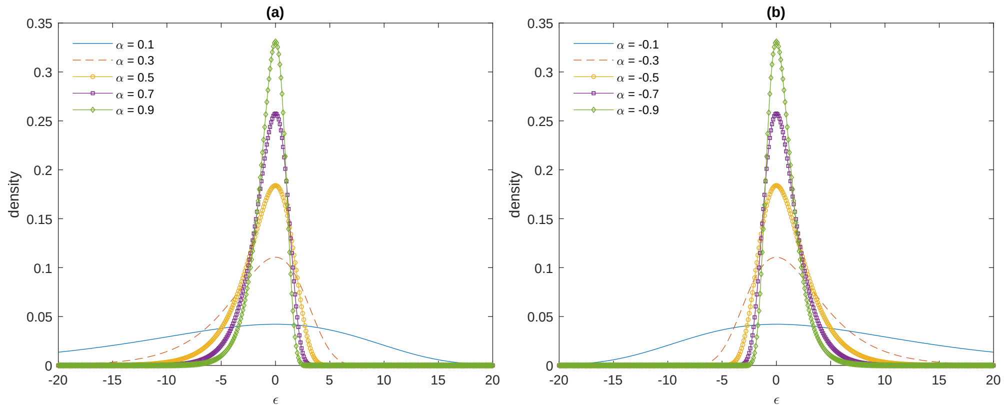
2.1 Estimation by maximum likelihood

Although it is likely that applied researchers will, most likely, want to set a value for a in advance and examine how the estimates of β change with different values of a, in this paper we examine estimation of a as well. The likelihood function is given by:

$$L(\beta, a; \tau, Y) \propto \tau^{n\tau}a^n\Gamma(\tau)^{-n}e^{-\tau\left(\sum_{i=1}^n e^{a\varepsilon_i} - a\sum_{i=1}^n \varepsilon_i\right)}. \tag{8}$$

[2]The expected value of (6) is non-zero. The fact that the expectation of error is non-zero affects only the estimate of the intercept in the regression which is, usually, of lesser importance. Another problem is that the introduction of a mean zero assumption for the error would complicate significantly ML and Bayesian analysis for parameter τ.

Figure 2: Different linex densities ($\tau = 1$)



The log-likelihood function:

$$l(\beta, a, \tau; Y) = \ln l(\beta, a; Y) = n\tau \ln \tau + n \ln a - n \ln \Gamma(\tau) - \tau \left(\sum_{i=1}^{n} e^{a\varepsilon_i} - a \sum_{i=1}^{n} \varepsilon_i \right). \tag{9}$$

We have

$$\frac{\partial l(\beta, a; \tau, Y)}{\partial \beta} = \tau a \sum_{i=1}^{n} \left(e^{a\varepsilon_i} - 1\right) \mathbf{x}_i, \tag{10}$$

$$\frac{\partial l(\beta, a; \tau, Y)}{\partial a} = na^{-1} - \tau \left(\sum_{i=1}^{n} \varepsilon_i e^{a\varepsilon_i} - \sum_{i=1}^{n} \varepsilon_i \right), \tag{11}$$

$$\frac{\partial^2 l(\beta, a; \tau, Y)}{\partial \beta \partial \beta'} = -\tau a^2 \sum_{i=1}^{n} \mathbf{x}_i e^{a\varepsilon_i} \mathbf{x}_i', \tag{12}$$

$$\frac{\partial^2 l(\beta, a; \tau, Y)}{\partial a^2} = -\left(\frac{n}{a^2} + \tau \sum_{i=1}^{n} \varepsilon_i^2 e^{a\varepsilon_i} \right), \tag{13}$$

$$\frac{\partial^2 l(\beta, a; \tau, Y)}{\partial \beta \partial a} = \tau \sum_{i=1}^{n} (a\varepsilon_i + 1) e^{a\varepsilon_i} \mathbf{x}_i - \tau \sum_{i=1}^{n} \mathbf{x}_i. \tag{14}$$

From the first-order conditions for maximum likelihood estimation (MLE) we have

$$\sum_{i=1}^{n} \left\{ e^{a\hat{\varepsilon}_i} - 1 \right\} \mathbf{x}_i = \mathbf{0}_{(k \times 1)}, \tag{15}$$

where $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\beta}$. These equations provide the so-called orthogonality conditions. For $a \to 0$, we have $e^{a\hat{\varepsilon}_i} \simeq 1 + a\hat{\varepsilon}_i$ and we recover the orthogonality condition corresponding to the least squares (LS) estimator. From (15) it is clear that the regressors $\mathbf{x}_i$ should be all orthogonal to the (scalar) generalized residual $\hat{u}_i$ where

$$\hat{u}_i = e^{a\hat{\varepsilon}_i} - 1,\ i = 1, \ldots, n, \tag{16}$$

rather than the LS residuals (to which they are orthogonal, however, when $|a| \to 0$, to first-order of approximation. Alternatively, the columns of the $n \times k$ matrix $\mathbf{X} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$ should be orthogonal to the $n \times 1$ generalized residual vector. To second-order, however, the regressors should be orthogonal to $\tilde{u}_i = \hat{\varepsilon}_i \left(1 + \frac{a}{2}\hat{\varepsilon}_i^2\right)$.

The MLE $\hat{a}$ of a solves the equation:

$$\tfrac{n}{\hat{a}} - \tau \sum_{i=1}^{n} \hat{\varepsilon}_i \left(e^{\hat{a}\varepsilon_i} - 1\right) = 0. \tag{17}$$

For $\hat{a} \to 0$ the solution is, approximately, $\tau \simeq \frac{n}{\hat{a}^2 \sum_{i=1}^{n} \hat{\varepsilon}_i^2}$ from which we see that τ (or, more precisely, τa) can be interpreted as a precision (inverse variance) parameter. Solving the nonlinear equation (17) provides the solution for the ML estimator $\hat{a}$.

It is worth noting that if we define $e^{(a/2)\varepsilon_i}\mathbf{x}_i = \tilde{\mathbf{x}}_i$, then from (12) the covariance matrix of $\hat{\beta}$ will be $\text{cov}(\hat{\beta}) = \frac{1}{\tau\hat{a}^2}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ where $\tilde{\mathbf{X}}$ is the n×k matrix which contains the $\tilde{\mathbf{x}}_i$s.

2.2 Bayesian analysis

Using the flat prior $p(\beta, a) \propto a^{-1}$ (in our case, $\tau = 1$ as we mentioned before) the posterior is given by Bayes' theorem as the product of the likelihood in (8) and the prior:

$$p(\beta, a|\tau, Y) \propto a^{n-1}\tau^{n\tau-1}\Gamma(\tau)^{-n} e^{-\tau\left(\sum_{i=1}^{n} e^{a\varepsilon_i} - a\sum_{i=1}^{n}\varepsilon_i\right)},\ \varepsilon_i = y_i - \mathbf{x}_i'\beta\ \forall i = 1, \ldots, n. \tag{18}$$

Posterior analysis by Markov Chain Monte Carlo (MCMC) is performed using two facts. First, the log-posterior is concave with respect to β because (12) is negative definite, so fact rejection algorithms can be used to draw from the posterior conditional distribution of $\beta|a, \tau, Y$ (Gilks and Wilde, 1992). As the mode is required, this can be obtained by solving the orthogonality conditions. Second, although the conditional $a|\beta, \tau, Y$ is not in any known family, from (18) it is a product of a gamma density, $\mathscr{G}(n, \sum_{i=1}^{n}\varepsilon_i)$ with the term $e^{-\sum_{i=1}^{n} e^{a\varepsilon_i}}$. To obtain a draw from $p(a|\beta, \tau, Y)$ we use as a candidate generating density an exponential distribution[3], viz. $q(a) = \lambda e^{-\lambda a}$, where λ is an unknown parameter. The ratio of the target to the candidate generating density is

$$R(a, \lambda) = \tfrac{p(a|\beta, Y)}{q(a)} = \lambda^{-1} e^{\lambda a} a^{n-1} e^{-\tau\left(\sum_{i=1}^{n} e^{a\varepsilon_i} - a\sum_{i=1}^{n}\varepsilon_i\right)} \tag{19}$$

[3]A gamma distribution of the form $\mathscr{G}\left(n, \sum_{i=1}^{n}\varepsilon_i\right)$ would seem to be more reasonable but the second-order conditions to solve the saddle point problem in (21) are not always met. In this case, one can use the Metropolis-Hastings algorithm.

and its logarithm is (keeping in mind that $\tau = 1$):

$$r(a, \lambda) = \ln R(a, \lambda) = -\ln \lambda + \lambda a + (n-1)\ln a - \tau \left(\sum_{i=1}^{n} e^{a\varepsilon_i} - a \sum_{i=1}^{n} \varepsilon_i \right). \tag{20}$$

We select the optimal value of λ by solving the minimax problem:

$$\min_{\lambda} \max_{a} r(a, \lambda). \tag{21}$$

We have

$$\tfrac{\partial r(a,\lambda)}{\partial a} = \lambda + \tfrac{n-1}{a} - \tau \left(\sum_{i=1}^{n} \varepsilon_i e^{a\varepsilon_i} - \sum_{i=1}^{n} \varepsilon_i \right), \tag{22}$$

Moreover

$$\tfrac{\partial r(a,\lambda)}{\partial \lambda} = a - \lambda^{-1}. \tag{23}$$

Setting $\frac{\partial r(a,\lambda)}{\partial \lambda} = 0$ we obtain the optimal value $\lambda = a^{-1}$. In turn from (22) the optimal value $\tilde{a}$ of a solves the nonlinear equation:

$$n = a\tau \left(\sum_{i=1}^{n} \varepsilon_i e^{a\varepsilon_i} - \sum_{i=1}^{n} \varepsilon_i \right). \tag{24}$$

As $\frac{\partial^2 r(a,\lambda)}{\partial \lambda^2} = \frac{1}{\lambda^2} > 0$, and $\frac{\partial^2 r(\tilde{a},1)}{\partial a^2} = -\frac{n-1}{\tilde{a}^2} - \tau \sum_{i=1}^{n} \varepsilon_i^2 e^{\tilde{a}\varepsilon_i} < 0$ then $\tilde{a}$ and $\tilde{\lambda} = \frac{1}{\tilde{a}}$ corresponds to a solution of the saddle point problem in (21). In turn, we generate a candidate draw $a^{(c)}$ from an exponential distribution with parameter $\lambda = \frac{1}{\tilde{a}}$ (where $\tilde{a}$ solves (24)) and we accept the draw with probability

$$\tfrac{r(a^{(c)},\tilde{\lambda})}{r(\tilde{a},\tilde{\lambda})}, \tag{25}$$

where the denominator is the maximal value of the acceptance rate in (21).[4] An alternative is to notice that the conditional posterior of a is log-concave. Indeed, we have:

$$\tfrac{\partial \ln p(a|\beta,\tau,Y)}{\partial a} = \tfrac{n-1}{a} - \sum_{i=1}^{n} \varepsilon_i - \sum_{i=1}^{n} \varepsilon_i e^{a\varepsilon_i}, \tag{26}$$

$$\tfrac{\partial^2 \ln p(a|\beta,\tau,Y)}{\partial a^2} = -\tfrac{n-1}{a^2} - \tau \sum_{i=1}^{n} \varepsilon_i^2 e^{a\varepsilon_i} < 0. \tag{27}$$

To provide a drawing from $p(a|\beta, \tau, Y)$ we use, again, rejection designed for log-concave densities (Gilks and Wilde,

[4]This procedure is quite fast as it, usually, requires no more than two or three rejections per acceptance.

1992). The mode of the density is required and this can be obtained by solving the nonlinear equation:

$$\tfrac{n-1}{a} - \tau\left(\sum_{i=1}^{n} \varepsilon_i e^{a\varepsilon_i} - \sum_{i=1}^{n} \varepsilon_i\right) = 0,$$

for a.

3 Double-linex loss functions and densities

The double-linex loss function involves a second parameter, b, introduced in the interest of flexibility:

$$\Lambda(\varepsilon) = \tau\left[e^{a\varepsilon} + e^{-b\varepsilon} - (a-b)\varepsilon - 2\right],\ a, b, \tau > 0. \tag{28}$$

If we turn the loss function into a density we have

$$f(\varepsilon) = \tfrac{1}{J(a,d,\tau)} \cdot \exp\left\{-\tau\left[e^{a\varepsilon} + e^{-b\varepsilon} - (a-b)\varepsilon - 2\right]\right\}, \tag{29}$$

where $J(a,d,\tau)$=$\int_{-\infty}^{\infty} \exp\left\{-\tau\left[e^{a\varepsilon} + e^{-b\varepsilon} - (a-b)\varepsilon - 2\right]\right\} \mathrm{d}\varepsilon$, and $d = \frac{b}{a}$. Finally, the form of the density becomes

$$f(\varepsilon) = \tfrac{1}{J(a,d,\tau)} \exp\left\{-\tau\left[e^{a\varepsilon} + e^{-b\varepsilon} - (a-b)\varepsilon - 2\right]\right\}. \tag{30}$$

It is not possible to evaluate $J(a, d, \tau)$ in closed form but we can compute it numerically. When $\tau = 1$, for example, a third-degree polynomial can be used to approximate the values shown in panel (a) of Figure 3. Different double-linex densities (with $\tau = 1$) are shown in panel (b) of Figure 3.

For known values of a and d, the log-likelihood function is

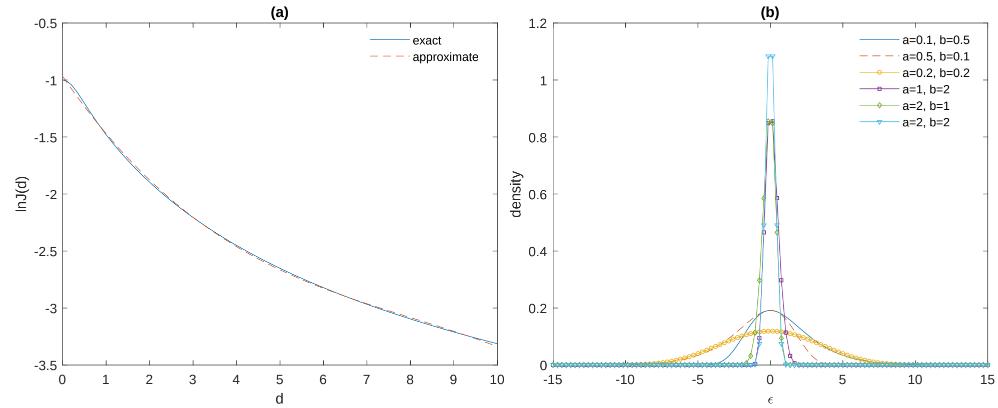
$$l(\beta; Y) = -n \ln J(d,\tau) - \tau \sum_{i=1}^{n} \left\{e^{a\varepsilon_i} + e^{-ad\varepsilon_i} + a(1-d)\varepsilon_i\right\},\ \varepsilon_i = y_i - \mathbf{x}_i'\beta. \tag{31}$$

Therefore,

$$\tfrac{\partial l(\beta;\tau,Y)}{\partial \beta} = \tau a \sum_{i=1}^{n} \left\{e^{a\varepsilon_i} - de^{-ad\varepsilon_i} + (1-d)\right\} \mathbf{x}_i = \mathbf{0}_{(k\times 1)}, \tag{32}$$

$$\tfrac{\partial^2 l(\beta;\tau,Y)}{\partial \beta \partial \beta'} = -\tau a^2 \sum_{i=1}^{n} \left\{e^{a\varepsilon_i} + d^2 e^{-b\varepsilon_i}\right\} \mathbf{x}_i \mathbf{x}_i'. \tag{33}$$

Figure 3: Exact and approximate values of $J(a, d, 1)$ and different double-linex densities



From (32) it is clear that the regressors $\mathbf{x}_i$ should be orthogonal to the generalized errors

$$u_i = e^{a\varepsilon_i} - de^{-ad\varepsilon_i} + (1-d),\ i = 1, \ldots, n. \tag{34}$$

Finally, from (33) it is clear that the Hessian of the log likelihood is negative definite and it can be used to obtain the asymptotic covariance matrix of the ML estimator.

4 Illustration

We use a sample size of $n = 10$ and we generate random observations from $\varepsilon_i \sim \mathcal{N}(0, 1)$, a standard normal distribution. All regression parameters are equal to 1.[5] Our prior is $p(a, d) \propto (ad)^{-1}$. Our prior for the regression parameters is flat. However, for the regression parameters one can also use a multivariate normal prior. We fix $d = 2$ and we present the marginal posterior densities of β_2 and β_3 in panel (a) of Figure 4. Then we fix $a = 0.3$ and we present the marginal posterior of d in panel (c) of Figure 4.

Although the marginal posterior of a given d and d given a are well defined, in this instance, the bivariate posterior in panel (c) of Figure 4, indicates that as a gets larger and $d \to 1$ the marginal posteriors converge to the marginal posteriors corresponding to those of the normal distribution for the error terms, as the case should be.

Next, we construct an example with $n = 100$, $k = 3$, $y_i = \mathbf{x}_i'\beta + u_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i3}$, $u_i \sim \mathcal{N}(0, 1)$. The regressors are generated from standard normal distributions, with $k = 3$ (the first regressor is a vector of ones corresponding to having an intercept in the model and true coefficients β equal to one. We present the marginal posterior densities of β_2 and β_3 in panels (a) and (b) of Figure 5 for various values of a. The marginal posterior density of d is reported in panel (c) of the same

[5] We use the numerically computed value of the normalizing constant. Results using the polynomial approximation were the same.

Figure 4: Marginal posteriors in an example with $n = 10$

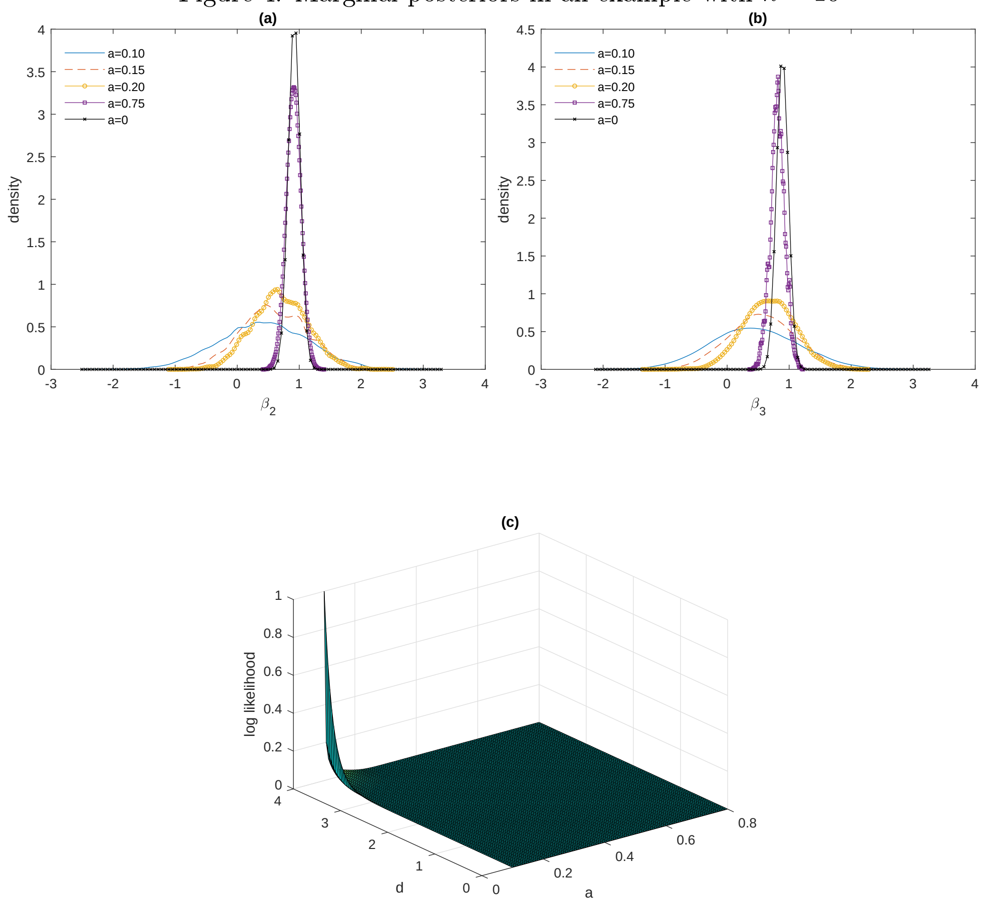


Figure 5: Marginal posteriors, synthetic data

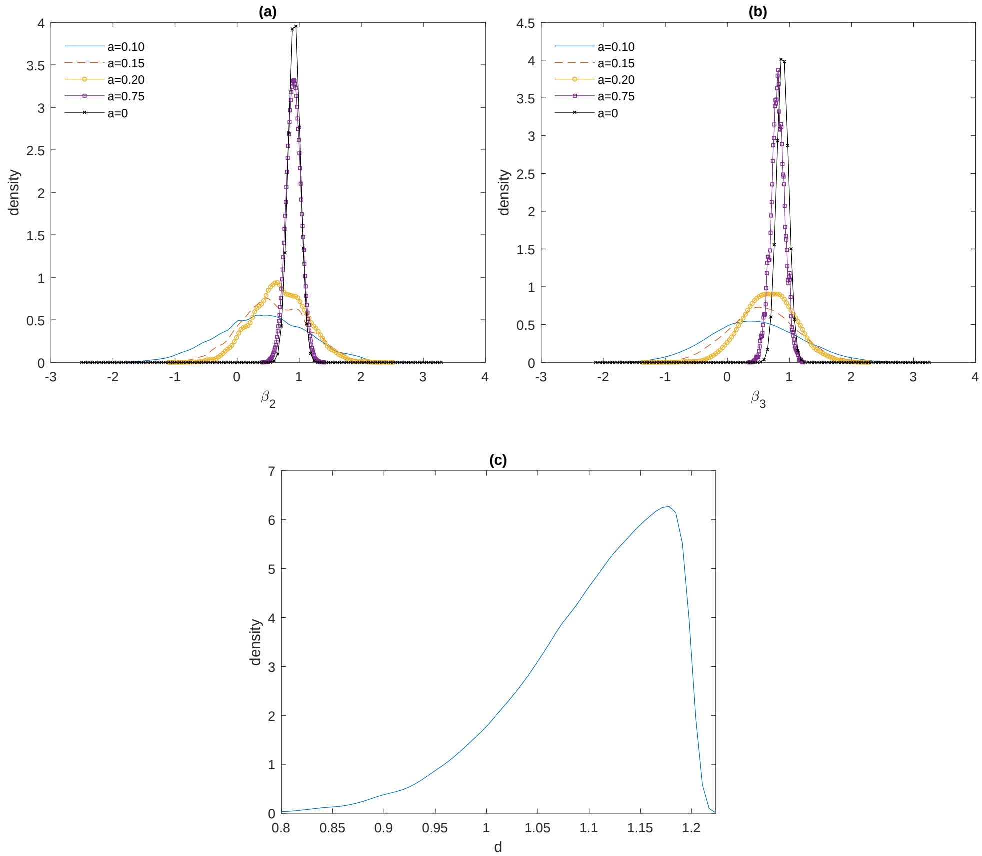


Figure. The posterior density of d is highly skewed to the left, making maximum likelihood inferences somewhat unreliable as they depend on asymptotic normality of parameter estimates.

5 Endogeneity

Often, some of the regressors $\mathbf{x}_i$ are endogenous in the sense that they are correlated with the error term in (7). Previously we did not make assumptions about the regressors. In fact, the moment conditions regrading orthogonality to so-called "generalized residuals" was prominent leading directly to the discussion on endogeneity. We think this is important precisely as it deviates from the practice of deterministic regressors or regressors that are orthogonal to "generalized residuals". If this not so, then the orthogonality conditions in (15) or (33) do not provide consistent estimators. Instead, suppose $\mathbf{z}_i \in \mathbb{R}^p$ is a vector of instrumental variables which are orthogonal to the errors but, nevertheless, correlated with the regressors (and, in fact, contain any regressors that are thought to be uncorrelated with the errors). In this case, (15) or (33) can be replaced by

$$\sum_{i=1}^{n} \left(1 - e^{a\hat{\varepsilon}_i}\right) \mathbf{z}_i = \mathbf{0}_{(p\times 1)}, \tag{35}$$

Table 1: Empirical results for managerial practices

	LSDV	$a = 5$	$a = 4$	$a = 3$	$a = 0.4$	$a = 0.25$	$a = 0.1$
labor	0.287	0.287	0.287	0.288	0.289	0.307	0.324
	(0.007)	(0.052)	(0.039)	(0.029)	(0.019)	(0.010)	(0.009)
capital	0.689	0.694	0.695	0.696	0.698	0.699	0.689
	(0.007)	(0.066)	(0.050)	(0.036)	(0.024)	(0.012)	(0.011)
management	0.076	**0.072**	**0.072**	0.071	0.069	0.064	0.059
	(0.009)	**(0.055)**	**(0.042)**	(0.033)	(0.020)	(0.010)	(0.008)

Notes: LSDV is the least squares estimator with dummy variables. Standard errors are reported in parentheses. Numbers in bold mean that the respective estimates are not statistically significant at the 5% significance level.

for the linex case or

$$\sum_{i=1}^{n} \left\{ e^{a\varepsilon_i} - de^{-ad\varepsilon_i} + (1-d) \right\} \mathbf{z}_i = \mathbf{0}_{(p\times 1)}, \tag{36}$$

for the double-linex case. Generally, $p \geq k$. Given these orthogonality conditions, the Generalized Method of Moments (GMM) estimator can be used (Hansen, 1982). Bayesian versions can be constructed using the Bayesian empirical likelihood / posterior (Lazar, 2003; Chaudhuri and Mondai, 2017).

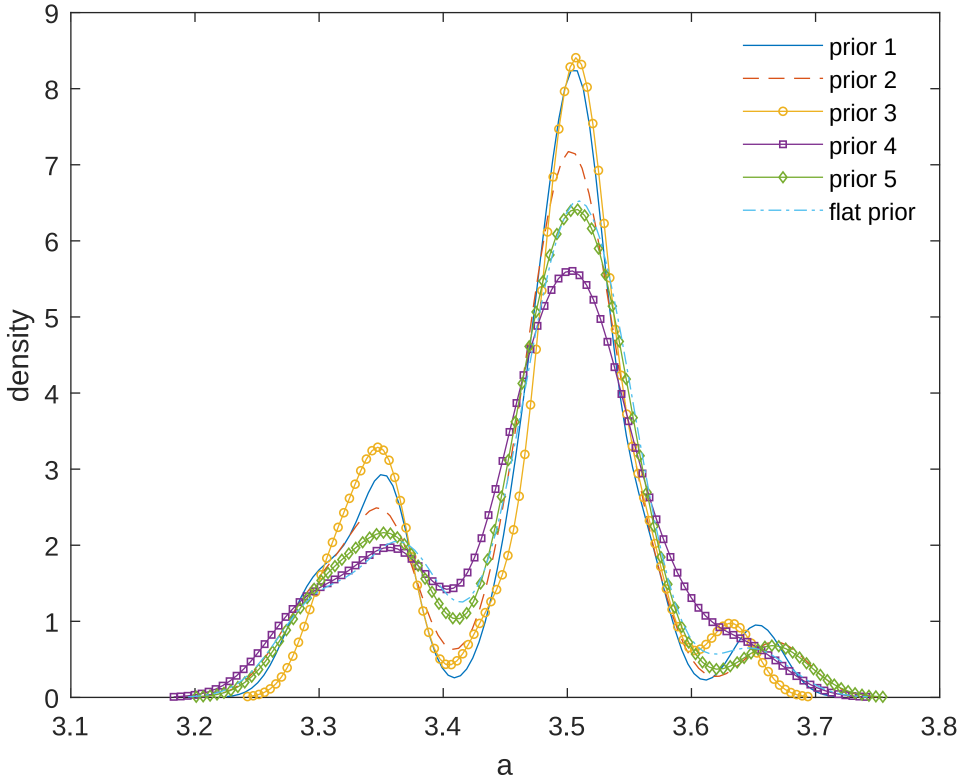
6 Empirical applications

6.1 Management in production

We apply the new techniques to a data set of Bloom and van Reenen (2007) which examines the role of managerial practices in production. The dependent variable is log deflated sales, and the regressors include log labor, log capital, the index of managerial practices as well as fixed effects. Our empirical results are reported in Table 1. For this application, we keep the original interpretation of the error $\varepsilon_i = \mathbf{x}_i'\beta - y_i$ rather than the opposite. So, as $\alpha \to 0$ we would get estimates close to the LS results.

Numbers in bold mean that the respective estimates are not statistically significant at the 5% significance level. This happens for estimates of management contribution for $a = 3, 4$ and 5. The question is whether one should "believe" more these estimates that try to avoid over-estimating the regression errors. As we are estimating a production function it is quite likely that technical inefficiency is present, a fact that would give negative skewness to the errors. Therefore, over-estimating the regression errors should be avoided and estimates corresponding to the higher values of a, would be more closely to reality casting some doubt on the role of the particular managerial index in production. Indeed, for this data set, the ML estimate of a turns out to be $\hat{a} = 3.472$ (with standard error 0.105). The marginal posterior density of a from Bayesian analysis is reported in Figure 6. Moreover, in Table 2 we report differences in estimates from LSDV, linex and double-linex posteriors. Apparently, the results are quite different as posterior means and posterior standard deviations differ between linex and LS, as well as linex versus double-linex. Therefore, in practice, we should not expect that LSDV, linex and double-linex will deliver the same results.

Figure 6: Marginal posterior density of a



For negative values of a we have not managed to obtain convergence in solving the system of orthogonality conditions (15). Moreover, posterior results are very similar to what we report here so, we avoid stating them in the interest of space although they are available on request. The marginal posterior densities of a from Bayesian analysis are reported in Figure 6. We use five different normal priors. All of them are normal $a \sim \mathcal{N}(0, \underline{\sigma}_a^2)$ for $\underline{\sigma}_a = 1$ (prior 1), $\underline{\sigma}_a = 5$ (prior 2), $\underline{\sigma}_a = 10$ (prior 3), $\underline{\sigma}_a = 100$ (prior 4), and $\underline{\sigma}_a = 10^3$ (prior 5). The sixth prior is the flat prior for β. All priors seem to yield the same posteriors as in Figure 6. Notice that as $\underline{\sigma}_a \to \infty$, the normal prior reduces to a flat prior. The densities appear to be bimodal indicating that asymptotic inferences may be somewhat misleading in this instance, although the different modes are not too far apart. So, marginal posterior densities of a appear to be robust. The posterior mean (for prior 5) is 3.47 and the posterior standard deviation is 0.092 (somewhat lower compared to its ML counterpart).

6.2 Stock returns

We use 43,151 intraday (tick-by-tick) observations of the stock of Johnson & Johnson (JNJ), traded on October 5, 2010 between 9:30 AM and 4:00 PM Eastern time. Aspects of the data are reported in Figure 7.

Table 2: Empirical results for managerial practices using different estimation techniques

	LSDV	linex	double- linex
labor	0.287	0.126	0.177
	(0.007)	(0.022)	(0.018)
capital	0.689	0.494	0.415
	(0.007)	(0.036)	(0.010)
management	0.076	0.055	0.035
	(0.009)	(0.017)	(0.005)

Notes: LSDV is the least squares estimator with dummy variables. Standard errors are reported in parentheses. For linex and double-linex we report posterior means with posterior standard deviations in parentheses.

Figure 7: Aspects of Johnson & Johnson stock

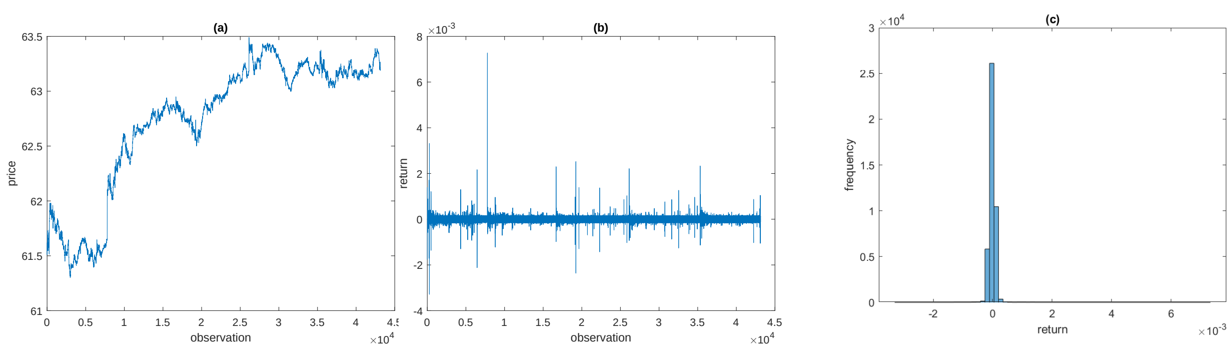
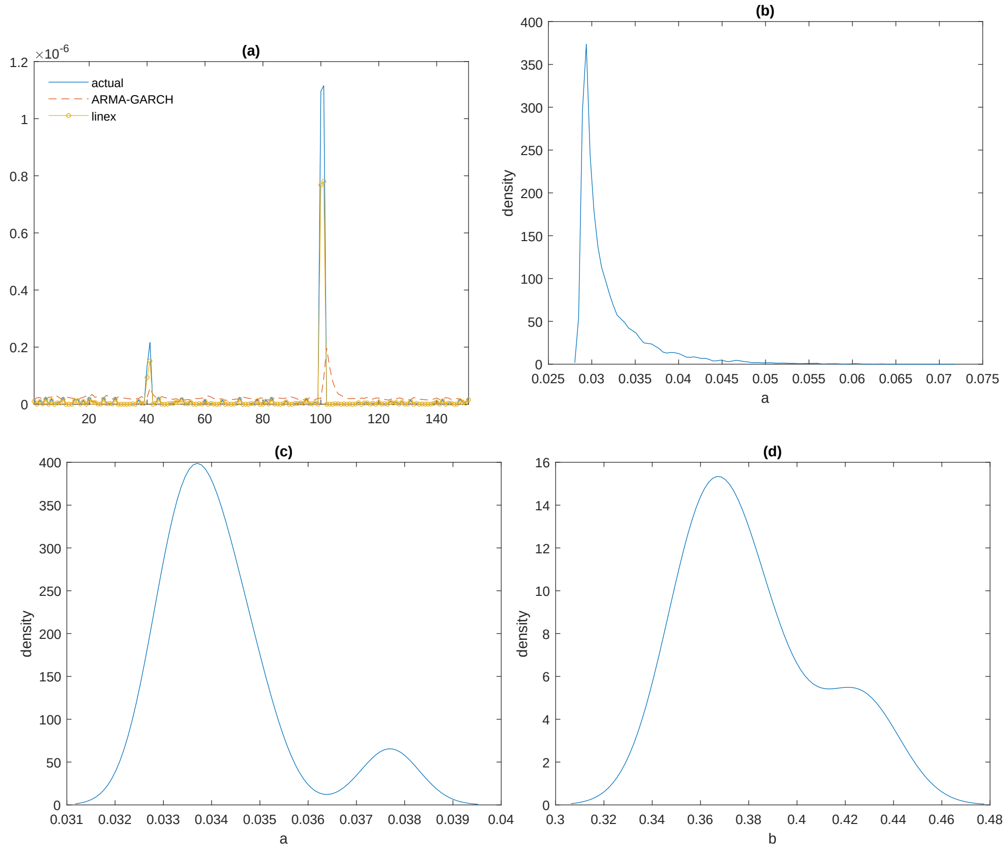


Figure 8: Volatility forecasts



We use an ARMA(1,1)-GARCH-M(1,1) model of the following form:

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 \log \sigma_{t-1}^2 + u_t, \\
u_t &= \beta_2 u_{t-1} + \beta_3 e_{t-1} + e_t, \\
e_t | \mathcal{F}_{t-1} &\sim \mathcal{T}(0, \sigma_t^2, \nu), \\
\sigma_t^2 &= \beta_4 + \beta_5 \sigma_{t-1}^2 + \beta_6 (y_{t-1} - \beta_0 - \beta_1 \log \sigma_{t-1}^2)^2,
\end{aligned}
\tag{37}
$$

where $y_t = \log \frac{p_t}{p_{t-1}}$ denotes returns (the stock price being p_t), $\mathcal{F}_{t-1}$ denotes information up to date $t-1$, $\mathcal{T}(0, \sigma_t^2, \nu)$ is the Student-t distribution with location zero, scale σ_t^2 and unknown degrees of freedom ν. We use the ARMA specification to capture the significance of long lags in returns and the GARCH-M (GARCH-in-mean) specification to allow for the effects of volatility on returns. We estimate the model using observations 1 through 43,000 (using maximum likelihood) and we perform out-of-sample forecasting for the remaining 151 observations. From Figure 8, volatility forecasts (e.g. Chen et al., 2022) slightly underestimate large shocks when they occur. We can correct for this by using linex posterior analysis following the definition of forecast errors in (1). To do this, we assume a linex distribution for e_t instead of the Student-t specification in (37). The forecasts are one-step ahead and the scale parameter σ_t is modeled using a GARCH(1,1) specification.

The results are presented in panel (a) of Figure 8. The posterior distribution of the linex parameter (a) is presented in panel (b). In panels (c) and (d) we report the marginal posterior densities of a and b from double-linex posterior analysis. The posterior mean of a (panel (b), linex case) is 0.032 (posterior s.d. 0.004). From panels (c) and (d), in the double linex case, the marginal posterior moments of a and b are 0.034 (posterior s.d. 0.0014) and 0.381 (posterior s.d. 0.027). The positive sign indicates that, based on the definition of forecast error in (1), underestimation of volatility is the more important problem in this data set (assuming the order of difference is reversed in (1)). Under-estimating volatility would be compatible with most behavioral rules or portfolio managers. Regarding, forecasting and comparison with LS, we present some relevant statistics in Table 3. RMSE stands for "root mean squared error", and MAE for "mean absolute error". DM is the Diebold and Mariano (1995) test for comparing two forecasts (here we use it three times to compare linex and double-linex, linex and LS, as well as double-linex versus LS). The null hypothesis of the DM test is that the two forecasts are the same so, when the p-value is less than, say, 0.01, we can reject the null. The DM test is asymptotically normally distributed. Clearly, RMSE and MAE favor the double-linex model, and the DM test whose p-values are very close to zero, indicate that the the differences in forecasts are statistical significant at the 1% significance level. Maximum likelihood estimates and posterior moments for the JNJ data are reported in Table 4. For linex and double-linex we also report the posterior moments of a and b.

Table 3: Forecast comparisons, Johnson and Johnson data

	linex	double-linex	LS
RMSE	0.041	0.032	0.055
MAE	0.030	0.027	0.067
	DM test		
linex vs. double-linex p-value	0.003		
linex vs. LS	0.000		
double-linex vs. LS, p-value	0.000		

Notes: RMSE stands for "root mean squared error", and MAE for "mean absolute error". DM is the Diebold and Mariano (1995) test for comparing two forecasts (here we use it twice to compare linex and double-linex, as well as double-linex versus LS).

Next, we compare the performance of the linex and double-linex loss functions results with state-of-the-art prediction methods. We use the concept of optimal prediction pools (Geweke and Amisano, 2011). These are weighted linear combinations of prediction models, or linear pools, evaluated using the log predictive scoring rule. An optimal linear combination typically includes several models with positive weights. For a given sample $Y_n = Y_n^o$ (where "o" stands for observed) and model $m \in \{1, ..., M\}$, the log predictive score is

$$LS_m(Y_n^o) = \sum_{t=1}^{n} \log p_m(y_t^o; Y_t^o). \tag{38}$$

Table 4: Empirical results for Johnson & Johnson data

	maximum likelihood	Bayes linex posterior	Bayes double-linex posterior
β_0	0.00072	0.00070	0.00071
	(16.12)	(23.45)	(19.42)
β_1	$4.04\ 10^{-5}$	$5.12\ 10^{-3}$	$6.05\ 10^{-3}$
	(14.03)	(22.35)	(43.14)
β_2	–0.045	–0.057	–0.071
	(–2.14)	(–6.47)	(–13.65)
β_3	–0.155	–0.120	–0.251
	(–6.58)	(–7.18)	(–6.55)
β_4	$4.12\ 10^{-4}$	$3.20\ 10^{-3}$	$4.78\ 10^{-3}$
	(24.88)	(15.44)	(21.56)
β_5	0.522	0.713	0.434
	(30.78)	(28.44)	(19.81)
β_6	0.120	0.192	0.361
	(14.48)	(48.71)	(22.55)
ν	5.032	4.210	2.717
	(44.32)	(87.55)	(55.40)
a	—	0.034	0.045
		(0.001)	(0.007)
a	—	0.032	—
		(0.004)	
b	—	—	0.381
			(0.027)

Notes: z-statistics are reported in parentheses. For Bayesian results the z-statistics are the ratios of posterior means to posterior standard deviations.

Table 5: Log predictive scores

	LPS	Optimal weights
GARCH	–47,881.25	0.000
GARCH-t	–46,335.12	0.000
EGARCH	–45,812.55	0.235
linex	-41,832,44	0.248
double-linex	-40,745.34	0.517

With $p_m(y_t; Y_{t-1}) = p_m(y_t|Y_{t-1})$ and we have

$$LS_m(Y_t^o) = \sum_{t=1}^{n} \log p_m(y_t^o|Y_{t-1}^o) = \log p_m(Y_t^0) = \log \int p_m(Y_t^o, \theta_m)\, \mathrm{d}\theta_m, \tag{39}$$

where θ_m is the parameter vector corresponding to model m. In linear pooling we consider predictive densities of the form $\sum_{m=1}^{M} w_m p_m(y_t; Y_{t-1}^o)$, $\sum_{m=1}^{M} w_m = 1$, $w_m \geq 0$ $(m = 1, ..., M)$. Computing the optimal weights is a relatively straightforward mathematical programming problem. We evaluate these densities using the log predictive score (LPS) function

$$LPS = \sum_{t=1}^{n} \log \sum_{m=1}^{M} w_m p_m(y_t^o; Y_{t-1}^o). \tag{40}$$

In our empirical application we consider several models and we report the LPS in Table 5. Evidently, the double-linex performs best in comparison to GARCH (with normal disturbances), GARCH-t, EGARCH as well as the linex model. From the optimal linear pool weights in the last column of Table 5, we see that the EGARCH is contained in the pool with a weight 0.235, the linex receives 0.248 and the double-linex receives the lion's share (0.517) testifying to the good predictive abilities of the model.

Concluding remarks

In this paper we propose a novel extension of linex and double-linex loss functions to the context of densities as models of the error terms in regression models. We derive the equations for maximum likelihood estimation and we propose efficient MCMC schemes for Bayesian inference. In the linex case, the parameter of the loss function (a) can be estimated using ML or Bayesian methods. The double-linex case presents more difficulties which are, however, immaterial when the researcher wants to compare parameter estimates of the regression model for different known values of the underlying loss parameters (a and b). The new methods are illustrated using simulated as well as real data concerning the role of management in production and volatility forecasts in tick-by-tick data. In terms of future research, it will be worthwhile to use our instrumental variables (IV) estimator in Section 5 to formulate the problem in terms of a Bayesian empirical posterior which is known to perform better than IV or GMM. Additionally, it would be quite interesting to see more applications in the area of quality control where the usual practice is to assume a linear model with normal errors (usually, such models have the form

$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$) and then use the linex loss function for decision making. The techniques developed here can be used to integrate estimation and decision making into a single step through a linex- or double-linex-based error process.

References

Bajestani, N. S., A. V. Kamyad, E. N. Esfahani, A. Zare (2018). Prediction of retinopathy in diabetic patients using type-2 fuzzy regression model. European Journal of Operational Research 264 (3), 859–869. Bloom, N., & John Van Reenen (2007). Measuring and explaining management practices across firms and countries. Quarterly Journal of Economics CXXII (4), 1351–1408. Bottmer, L., C. Croux, I, Wilms (2021). Sparse regression for large data sets with outliers. European Journal of Operational Research, in print. Chang, Y-C, & W-L Hung (2007). LINEX Loss Functions with Applications to Determining the Optimum Process Parameters. Quality & Quantity 41, 291–301. Chaudhuri, S., D. Mondai, T.Yin (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79 (1), 293–320. Chen, C.-H. and Chou, C.-Y. (2004). Set the optimum process parameters based on asymmetric quality loss function. Quality & Quantity 38:75–79. Chen, W-J, J-J Yao, & Y-H Shao (2022). Volatility forecasting using deep neural network with time-series feature embedding. Economic Research-Ekonomska Istraživanja DOI: 10.1080/1331677X.2022.2089192 Chernozhukov, V., & H. Hong (2003). An MCMC approach to classical estimation. Journal of Econometrics 115 (2), 293–346. Christodoulakis, G. (2005). Financial forecasts in the presence of asymmetric loss aversion, skewness and excess kurtosis. Finance Research Letters 2, 227–233. Christoffersen, P. F., and F. X. Diebold (1997). Optimal prediction under asymmetric loss. Econometric Theory 13 (6), 808–817. Diebold, F.X. and R.S. Mariano (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 134–145. Geweke, J., and G. Amisano (2011). Optimal prediction pools. Journal of Econometrics 164 (1), 130-141. Gilks, W. R., & P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337–348. Granger, C.W.J. (1969). Prediction with a generalised cost error function. Operations Research Quarterly 20, 199–207. Granger, C.W.J. (1999). Outline of forecast theory with generalised cost functions. Spanish Economic Review 1, 161–173. Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054. Huang, Y.-F. (2001). Trade-off between quality and cost. Quality & Quantity 35:265–276. Hwang, S., Knight, J., Satchell, S. (2001). Forecasting non-linear functions of returns using Linex loss functions. Annals of Economics and Finance 2, 187–213. Kumbhakar, S.C., & C.A.K. Lovell (2000). Stochastic frontier analysis. Cambridge University Press. Lazar, N. A. (2003). Bayesian empirical likelihood. Biometrika 90(2), 319–326. Liang, C., Z. Zhang, Y. Song, L. Jiang (2021). Kernel-based online regression with canal loss. European Journal of Operational Research, in press. Taguchi, G. (1986). Introduction to Quality Engineering. Tokyo: Asian Productivity Organization. Varian, H. (1975). A Bayesian approach to real estate assessment. In: Fienberg, S., Zelner, A. (Eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. North-Holland, Amsterdam, pp. 195–208. Ulrich, M., H. Jahnke, R. Langrock, R. Pesch, R. Seng (2021). Distributional regression for demand forecasting in e-grocery. European Journal of Operational Research 294 (3), 831–842. Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. Journal of American Statistical Association 81, 446–451.