Aggregated Expectile Regression by Exponential Weighting

YUWEN GU AND HUI ZOU*

University of Connecticut and University of Minnesota

Abstract

Various estimators have been proposed for estimating conditional expectiles, including those from the multiple linear expectile regression, local polynomial expectile regression, boosted expectile regression, and so on. It is a common practice that several plausible candidate estimators are fitted and a final estimator is selected from the candidate list. In this paper we advocate using an exponential weighting scheme to adaptively aggregate the candidate estimators into a final estimator. We show that the aggregated estimator enjoys an oracle inequality. Simulations and a real data example show that the aggregated estimator outperforms the cross-validated estimator when cross-validation exhibits selection uncertainty.

Keywords: Cross-validation, Expectile regression, Oracle inequality, Model aggregation.

1 Introduction

Expectiles (Newey and Powell, 1987) are informative location measures of probability distributions. For each $\tau \in (0,1)$, the τth expectile of a probability distribution F is defined as the quantity e_τ that

*Correspondence to: School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, USA. Tel: (612)625-4005; Fax: (612)624-8868; Email address: zouxx019@umn.edu (Hui Zou).

satisfies

$$\int_{-\infty}^{e_\tau} |x - e_\tau| \, \mathrm{d}F(x) = \tau \int_{-\infty}^{\infty} |x - e_\tau| \, \mathrm{d}F(x). \tag{1}$$

Denote $\mathscr{E}^\tau$ the expectile operator at level τ such that $\mathscr{E}^\tau(F) = e_\tau$. For any random variable $Y \sim F$, we will also write $\mathscr{E}^\tau(Y) = \mathscr{E}^\tau(F)$. It can be shown that the 0.5th expectile coincides with the mean, $\mathscr{E}^{0.5}(F) = \int_{-\infty}^{\infty} x \mathrm{d}F(x)$ and moreover, all the expectiles exist as long as the mean is finite. The financial meaning of the expectiles is transparent: $\mathscr{E}^\tau(F)$ is the amount of money that should be added to a position in order to have a pre-specified gain-loss ratio (Bernardo and Ledoit, 2000). Specifically, suppose $Y \sim F$ and let $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$. From definition (1) of the expectiles one has

$$\frac{\mathbb{E}(Y - e_\tau)_+}{\mathbb{E}(Y - e_\tau)_-} = \frac{1 - \tau}{\tau},$$

where $\mathbb{E}(Y - e_\tau)_+$ and $\mathbb{E}(Y - e_\tau)_-$ can be interpreted as the expected values of the gain and loss respectively and $(1 - \tau)/\tau$ the targeted gain-loss ratio. An equivalent definition views expectiles as solutions to the minimization problem

$$\mathscr{E}^\tau(F) = \arg\min_{a \in \mathbb{R}} \int_{-\infty}^{\infty} \Psi_\tau(x - a) \, \mathrm{d}F(x), \; \tau \in (0, 1),$$

where $\Psi_\tau(u) = |\tau - I(u < 0)|u^2$ is the asymmetric squared error loss and $I(\cdot)$ represents the indicator function. This definition reveals the fact that expectiles are very sensitive to the magnitude of extreme losses, which leads to favorable applications of expectiles in econometrics and finance as a downside risk measure (Kuan et al., 2009). The special role of expectiles in risk management has been further recognized by many researchers recently. In risk management, Value at Risk (VaR) and expected shortfall (ES) are the two most popular risk measures in use. However, it has been shown that VaR lacks the desired property of coherence since the seminal work by Artzner et al. (1999). Specifically, VaR is not sub-additive, which contradicts the diversification principle that merging portfolios together should reduce the risk. ES is coherent (Acerbi and Tasche, 2002), but

nevertheless fails to enjoy elicitability (Gneiting, 2011), another desired property of risk measures for which meaningful point forecasts and forecast performance comparisons are possible. Expectiles are the only risk measure that is both coherent and elicitable (Ziegel, 2014; Bellini and Bignozzi, 2015).

Expectile regression estimates the conditional expectiles of a response variable given a set of covariates and is a useful extension to the mean regression. It has been widely applied to finance, demography, and education (see Taylor, 2008; Schnabel and Eilers, 2009a; Sobotka et al., 2013b). Since its advent (Aigner et al., 1976), a variety of expectile regression methods have been proposed. The multiple linear expectile regression was systematically studied in Newey and Powell (1987), in which theoretical properties were rigorously derived and applications to testing heteroscedasticity and conditional symmetry were described. Nonparametric and semi-parametric expectile estimation methods have also been considered in the literature to allow for more flexibility. Among others, Yao and Tong (1996) provided kernel smoothing estimators of the conditional expectiles based on local polynomial regression and they further presented the asymptotic behavior of these estimators. Their work was followed and extended by Guo and Härdle (2012), in which simultaneous confidence bands were established for the expectile functions. A nonparametric expectile estimation method based on spline smoothing was introduced in Schnabel and Eilers (2009b) and similar to that, Sobotka et al. (2013a) proposed a semi-parametric expectile estimation approach using splines. In practice, both kernel and spline smoothing methods suffer from the curse of dimensionality. Yang and Zou (2015) proposed a nonparametric multiple expectile regression method using gradient boosting with regression tree base learners.

With the availability of various expectile regression methods, a practical problem is to choose the right method for the data at hand. The topic of model selection in the context of mean regression has been heavily studied in the literature. For example, lots of work has been devoted to the so-called model selection information criteria such as AIC (Akaike, 1974) and BIC (Schwarz et al., 1978). To our knowledge, there is no AIC- or BIC-like model selection criterion for expectile

regression that has been justified theoretically. Moreover, these information criteria are often not applicable when comparing a parametric model with a nonparametric alternative. As such, cross-validation has been widely applied in practice and of course can be used in the context of expectile regression. The model selection process by information criteria or cross-validation is always stochastic. Consequently, the uncertainty in model selection is inherently a part of the stochastic error in the final chosen model. Therefore, when the model selection uncertainty is large, the selected model tends to suffer.

When several plausible expectile regression estimators are present, instead of trying to select the best one, another good alternative is aggregation. In the literature, this idea is also known as model averaging or model combining. One can use these three names interchangeably wherever no confusion arises. There are multiple ways to do aggregation. We refer the interested readers to a review article by Hoeting et al. (1999) on Bayesian model averaging. In this article, we take an exponential weighting scheme to combine different expectile regression estimators. Our estimator is a weighted average of these candidate estimators and the weight of each candidate estimator is inversely proportional to the exponential of its cumulative empirical prediction risk. Such an exponential weighting scheme has a solid information-theoretic justification in the context of conditional mean regression (Yang, 2001, 2004; Catoni and Picard, 2004). We prove an oracle inequality for the aggregated expectile regression estimator by exponential weighting in terms of both prediction risk and squared loss. The theory implies that the aggregated expectile regression estimator at least behaves like the best candidate expectile regression estimator. We further compare the aggregated estimator and the cross-validated estimator by extensive simulations. It is shown that the aggregated estimator significantly outperforms the cross-validated estimator when there is selection uncertainty.

The article is organized as follows. In the next section, we present the aggregated expectile regression estimator and study its theoretical properties. The applications of the aggregated expectile regression are introduced in Section 3 through several simulation examples. We apply the aggregated

expectile regression to study a real personal computer data example in Section 4. The technical proofs are relegated to the appendix.

2 Aggregated Expectile Regression by Exponential Weighting

2.1 Setup and notation

Consider the standard regression setting with i.i.d. observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are p-dimensional covariate vectors and y_i are scalar responses. Assume these observations are realizations of the random pair $(\mathbf{X}, Y)$, where $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. Let $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ and $\sigma^2(\mathbf{x}) = \mathrm{var}(Y|\mathbf{X} = \mathbf{x})$ be respectively the conditional mean and variance functions. Assume both $m(\mathbf{X})$ and $\sigma(\mathbf{X})$ exist and $\sigma(\mathbf{X}) > 0$ almost surely. Define $\varepsilon = (Y - m(\mathbf{X}))/\sigma(\mathbf{X})$. It follows immediately that $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$ and $\mathrm{var}(\varepsilon|\mathbf{X}) = 1$ almost surely. For ease of exposition, let us write Y in terms of $\mathbf{X}$ and ε as

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon. \tag{2}$$

It should be noted that (2) by no means implies that we restrict ourselves to additive models only, although additive models are obviously included. As a matter of fact, a model with multiplicative error, for instance, $Y = f(\mathbf{X})\varepsilon$ can be easily cast into (2) as long as the conditional mean and variance functions of Y given $\mathbf{X}$ exist. Denote the τth conditional expectile by $e_\tau(\mathbf{x}) = \mathscr{E}^\tau(Y|\mathbf{X} = \mathbf{x})$, $0 < \tau < 1$. The goal of the expectile regression is to estimate $e_\tau(\mathbf{x})$. For an estimator $\hat{e}_\tau(\mathbf{x})$ of the expectile function $e_\tau(\mathbf{x})$, define the prediction risk and squared loss of $\hat{e}_\tau(\mathbf{x})$ by $\mathbb{E}\Psi_\tau(Y - \hat{e}_\tau(\mathbf{X}))$ and $\mathbb{E}(\hat{e}_\tau(\mathbf{X}) - e_\tau(\mathbf{X}))^2$ respectively. If the τth conditional expectile of ε given $\mathbf{X}$ is $b_\tau(\mathbf{X})$, then $e_\tau(\mathbf{x}) = m(\mathbf{X}) + \sigma(\mathbf{X})b_\tau(\mathbf{X})$. When $e_\tau(\mathbf{x})$ is approximately linear in $\mathbf{x}$, the linear expectile regression is expected to perform quite well. However, when complicated nonlinear pattern exists in $e_\tau(\mathbf{x})$, the linear expectile regression can result in very large bias. As a remedy, nonparametric expectile

regression methods can be used to accommodate the non-linearity. Of course, nonparametric expectile regression often has higher estimation variance than the linear expectile regression.

2.2 The aggregation algorithm

Suppose we have a sequence of estimating procedures $\Delta = \{\delta_j, j \geq 1\}$, all of which can provide estimates of $e_\tau(\mathbf{x})$. Specifically, the estimate of $e_\tau(\mathbf{x})$ from procedure $\delta_j \in \Delta$ fitted on data with sample size n is denoted by $\hat{e}_{\tau,j,n}(\mathbf{x})$, $j \geq 1$. Note that we allow the number of procedures to be either finite or countably infinite. Following Yang (2001), we impose no special assumptions on the procedures and they can be either model-based or non-model based. The goal is to construct an estimating procedure δ_a by adaptively aggregating this sequence of candidate estimating procedures in the hope of achieving a small estimation risk. The algorithm for this aggregation is displayed in Algorithm 1.

Algorithm 1 about here.

In Algorithm 1, n_0 is often chosen such that both n_0 and $n - n_0$ are of the same order as n. See more discussion in the next subsection. The tuning parameter λ is a properly chosen constant which controls the effect of the performance of the candidate estimators on the weights. On one hand, when λ is very small, Algorithm 1 will assign almost equal weights to the candidate estimators. In the extreme case $\lambda = 0$, Algorithm 1 is merely a simple average of the candidate estimators. On the other hand, when λ is large enough, Algorithm 1 will almost put all the weights on the procedure with best performance upon evaluation on $D^{(1)}$. We comment on the choice of λ in Remark 3 in the next subsection.

Note that in Algorithm 1, the weights $W_{j,i}, n_0 \leq i \leq n, j \geq 1$ depend on the order of the observations from the random partition. Multiple splits can be carried out as shown in Algorithm 2 to avoid large variance in the weights. From the computational point of view, these multiple splits can be carried out in parallel to accelerate the computation. We recommend Algorithm 2 for

practical use and according to our empirical studies the algorithm often works quite well when the number of splits B is taken to be several hundred.

Algorithm 2 about here.

2.3 Oracle inequalities for AEREW

We provide the oracle inequalities for AEREW in terms of both prediction risk and squared loss in the following theorem. To facilitate the discussion, let us introduce some notation. Denote $\underline{c} = \min(\tau, 1-\tau)$ and $\bar{c} = \max(\tau, 1-\tau)$. For a random variable Z, define the sub-exponential norm of Z by $\|Z\|_{\text{SEXP}} \equiv \sup_{k\geq 1} k^{-1}(\mathbb{E}|Z|^k)^{1/k}$. If $\|Z\|_{\text{SEXP}}$ is finite, we call Z a sub-exponential random variable (Vershynin, 2010).

Theorem 1. *Under the general model* (2) *and assume that the candidate estimators satisfy the following conditions:*

(C1) With probability one, $\sup_{i,j} |\hat{e}_{\tau,j,i}(\mathbf{X}) - e_\tau(\mathbf{X})| \leq A_\tau$ *and* $|e_\tau(\mathbf{X})| \leq B_\tau$, *where* $A_\tau, B_\tau \in (0,\infty)$ *are positive constants depending on* τ.

(C2) With probability one, $|\sigma(\mathbf{X})| \leq C_0$, *where* $C_0 \in (0,\infty)$ *is also a positive constant.*

(C3) With probability one, the sub-exponential norm of ε *given* $\mathbf{X}$ *is bounded by a positive constant* $K \in (0,\infty)$.

Let $K_\tau = 2\bar{c}(K + B_\tau)$ *and* $D_\tau = 4eK_\tau$, *where* $e = \exp(1)$. *Define the two functions* $\mathscr{M}_0(t) = 2\exp(2e^2K_\tau^2t^2)$ *and* $\mathscr{M}_2(t) = 16\sqrt{2}\exp(4e^2K_\tau^2t^2)$. *When the tuning parameter* λ *is chosen such that*

$$\lambda \leq \min\left\{ \frac{1}{2C_0A_\tau D_\tau}, \frac{\underline{c}\exp\left(-\bar{c}A_\tau(C_0D_\tau)^{-1}\right)}{C_0^2\mathscr{M}_2(D_\tau^{-1}) + 16\bar{c}^2A_\tau\mathscr{M}_0(D_\tau^{-1})} \right\}, \tag{3}$$

the risk of the combined estimator by AEREW (Algorithm 1 and Algorithm 2) has the following upper bound

$$\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))\leq \inf_{j\geq 1}\left\{\frac{\log(1/\pi_j)}{\lambda(n-n_0)}+\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j,n_0}(\mathbf{X}))\right\}. \tag{4}$$

and the squared loss of the combined estimator satisfies

$$\mathbb{E}(\hat{e}_{\tau,\cdot,n}(\mathbf{X})-e_\tau(\mathbf{X}))^2\leq \inf_{j\geq 1}\left\{\frac{\log(1/\pi_j)}{\lambda\underline{c}(n-n_0)}+(\bar{c}/\underline{c})\mathbb{E}(\hat{e}_{\tau,j,n_0}(\mathbf{X})-e_\tau(\mathbf{X}))^2\right\}, \tag{5}$$

where $(\mathbf{X},Y)$ *is taken to be a random observation from* (2) *that is independent of the observations* $(\mathbf{X}_i,Y_i)_{i=1}^n$.

Remark 1. The assumption of conditions (C1-C2) is mild and can be easily satisfied with high probability if the mean function $m(\mathbf{X})$ as well as the variance function $\sigma^2(\mathbf{X})$ are bounded almost surely. This assumption is fairly common in related work for aggregation.

Remark 2. The class of sub-exponential random variables covers all random variables for which the moment generating functions exist in a neighborhood of zero and hence is quite large to encompass commonly used error distributions. As a consequence, condition (C3) does not restrict the response Y to be bounded, which is however a condition often assumed in the machine learning literature for simplicity.

Remark 3. The oracle inequality (4) tells us that the aggregated estimator achieves a prediction risk that is smaller than the smallest prediction risk offered by the candidate estimators plus an additional risk term. Assume there are M candidate estimators all of which are assigned equal prior weights, the extra risk term in the oracle inequality (4) becomes $\log(M)/\{\lambda(n-n_0)\}$. Although λ has an upper bound by (3) in order for us to prove the oracle inequalities, there should also be a lower bound for λ in order to make the extra term $\frac{\log(M)}{\lambda(n-n_0)}$ much smaller than $\mathbb{E}\Psi_\tau(Y-e_\tau(\mathbf{X}))$. Otherwise the oracle inequality offers no meaningful conclusions. Typically, $\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j,n_0}(\mathbf{X}))$

converges to $\mathbb{E}\Psi_\tau(Y - e_\tau(\mathbf{X}))$ at rate n_0^{-1} for parametric estimators and at a slower rate than n_0^{-1} for nonparametric estimators. So if one only cares about the absolute prediction risk, then we only need to require $\frac{\log(M)}{\lambda(n-n_0)} \ll 1$. Of course, we often also care about the rate of convergence. If n_0 is chosen such that both n_0 and $n-n_0$ are of order n, then as long as $\lambda \geq O(\log(M))$ the extra risk term $\frac{\log(M)}{\lambda(n-n_0)}$ does not affect the rate of convergence, i.e., the aggregated estimator by AEREW will achieve the same rate of convergence as the best candidate estimator. We recommend using $\max(1, \lfloor \log(M) \rfloor)$ as the default value for λ. In all of our numerical experiments this default choice works very well and we have also found that the performance of AEREW is insensitive to the choice of λ in a fairly wide range.

3 Applications and Simulation Examples

In this section, we demonstrate several useful applications of aggregation in expectile regression. These applications are illustrated through two simulation examples.

3.1 Local expectile regression: bandwidth selection or aggregation?

When there is a single covariate, nonparametric expectile regression can be done via the local fitting scheme as shown in Yao and Tong (1996). It was argued by Yao and Tong (1996) that the local linear fit automatically corrects the boundary effects inherited from the local constant fit (see also Fan, 1992) and the estimator of the derivative plays an important role in monitoring the reliability of non-linear prediction and in detecting chaos. To be specific, given a random sample $(X_i, Y_i)_{i=1}^n$, the local linear estimators of $e_\tau(x) = \mathscr{E}^\tau(Y|X=x)$ and $e'_\tau(x) = \mathrm{d}e_\tau(x)/\mathrm{d}x$ are defined as

$$\big(\hat{e}_\tau(x;h), \hat{e}'_\tau(x;h)\big) = \underset{a,b\in\mathbb{R}}{\arg\min} \sum_{i=1}^{n} \Psi_\tau(Y_i - a - b(X_i - x)) h^{-1} K\left(\frac{X_i - x}{h}\right), \tag{6}$$

where $K(\cdot)$ is the kernel density and $h > 0$ is the bandwidth. Though many kernel densities are available for the local linear regressions, we chose the Gaussian kernel $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ for illustration purpose.

The choice of h reflects the tradeoff between the bias and variance of the estimators and has a high impact on the performance of the prediction. The theoretically optimal bandwidth is $h = C_\tau n^{-1/5}$ where C_τ depends on unknown quantities (Yao and Tong, 1996). In practice, we can select the bandwidth through cross-validation. See Heidenreich et al. (2013) for a recent review of bandwidth selection methods.

Alternatively, we can combine the kernel estimators at different bandwidths. Specifically, for a sequence of candidate bandwidths $h_1, \ldots, h_M$, we obtain the local linear fit $\hat{e}_\tau(x; h_j)$ for each bandwidth h_j, $1 \le j \le M$ and combine these estimators using AEREW.

We illustrate this application through a simulation study. Consider the following heteroscedastic model

$$Y = 0.5\{X + 2\exp(-16X^2)\} + \{0.4\exp(-2X^2) + 0.2\}\varepsilon, \tag{7}$$

where the scalar covariate X is independent of the random error ε. Moreover, suppose $X \sim$ Uniform$(-2,2)$ and $\varepsilon \sim$ Laplace$(0, 1/\sqrt{2})$. The density of ε is $f_\varepsilon(u) = \exp(-\sqrt{2}|u|)/\sqrt{2}$. Note that ε is a sub-exponential random variable satisfying $\mathbb{E}(\varepsilon) = 0$ and $\text{var}(\varepsilon) = 1$. A similar model to (7) was considered in Fan and Yao (1998) under a different error distribution. For the simulation study, a training set of $N = 200$ observations were randomly generated from model (7) and local linear regressions (6) were fitted to the training data with five candidate bandwidths $\mathbf{h} = (0.1, 0.3, 0.5, 0.7, 0.9) \times (200)^{-1/5}$.

To demonstrate the benefit of aggregation and compare it with cross-validation in bandwidth selection, we applied a five-fold cross-validation to select the best bandwidth and also combined the five local linear expectile regression estimators using AEREW (Algorithm 2), for which $B = 200$ splits were conducted and the splitting size $N_0 = 160$ and prior weights $\pi_j = 1/5$, $1 \le j \le 5$ were

chosen. We also set $\lambda = 1$. To compare the estimation performance of different procedures, we independently simulated a test set of $N_1 = 10000$ observations from model (7) and calculated the following two performance measures based on the test data. Assume the true expectile function is $e_\tau(\cdot)$ and its estimate from a specific procedure is $\hat{e}_\tau(\cdot)$. The two measures: (i) the estimated prediction risk, and (ii) estimated squared deviation (MSD) for $\hat{e}_\tau(\cdot)$ are respectively defined as

$$\text{risk}(\tau) = \frac{1}{N_1}\sum_{i=1}^{N_1}\Psi_\tau(Y_i - \hat{e}_\tau(X_i)) \quad \text{and} \quad \text{MSD}(\tau) = \sqrt{\frac{1}{N_1}\sum_{i=1}^{N_1}(\hat{e}_\tau(X_i) - e_\tau(X_i))^2}. \tag{8}$$

For model (7), the true expectile function is $e_\tau(x) = 0.5\{x + 2\exp(-16x^2)\} + \{0.4\exp(-2x^2) + 0.2\}b_\tau$, where $b_\tau = \mathscr{E}^\tau(\varepsilon)$ is the τth expectile of the Laplace random error. The simulations were repeated $M = 100$ times under the above setting. For illustration purpose, we also presented the proportion p_{CV} of each candidate estimator being selected by the five-fold cross-validation among these 100 runs. The results are summarized in Table 1.

Table 1 about here.

In Table 1, the performance measures were calculated by averaging over the 100 replicates and their respective standard errors were reported in the parentheses. It is clear from Table 1 that the optimal bandwidths are different for different expectile levels. Smaller bandwidths are preferred for expectile levels around 0.5 while at extreme expectile levels (τ close to 0 and 1), slightly larger bandwidths are favored. Also it is evident from Table 1 that AEREW compares quite favorably with the five-fold cross-validation. Indeed, AEREW outperforms the cross-validation for all expectile levels other than 0.5 and its performance there is very close to or even better than that of the best candidate estimator. On the other hand, the cross-validation gives slightly better estimation for the mean function ($\tau = 0.5$) than AEREW, but still AEREW performs quite well in this case. From the simulation result, it can be seen that when cross-validation is uncertain about the best estimator (several estimators have reasonably large p_{CV} values), AEREW can outperform the cross-validated

estimator.

3.2 Multiple expectile regression: parametric or nonparametric?

We now demonstrate the application of AEREW in multiple expectile regression where more than one covariate is available. For such models, the local linear etimator, such as (6), is not very useful in practice when there are more than five covariates. In the current toolbox for multidimensional expectile regression, we have the multiple linear expectile regression (Newey and Powell, 1987; Efron, 1991) as well as the regression tree based nonparametric gradient boosting (Yang and Zou, 2015). The question one often encounters in practice is which one to use. In fact, it is well known that the nonparametric regression is quite flexible to accommodate non-linearity but loses efficiency when the linear parametric model is correctly specified. In the case of expectile regression, when linear and nonlinear effects coexist in an underlying model, it is also possible that the expectile function is nearly linear at certain expectile levels and becomes highly nonlinear at other levels. Therefore, when multiple expectile levels need to be inspected together, it is beneficial to consider adaptively aggregating both parametric and nonparametric methods for better estimation.

As an illustration, let us consider the following heteroscedastic model with multiple covariates

$$Y = \mathbf{X}^{\mathrm{T}}\beta + 2\varepsilon\exp(-0.35X_2 - 1.1X_4), \tag{9}$$

where $\beta = (1.5, 2.5, 1.0, 0.5, 2.0, 1.5)^{\mathrm{T}}$, $\mathbf{X} = (X_1, \ldots, X_6)^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{I}_6)$, $\varepsilon \sim N(0,1)$ and $\mathbf{X}$ is independent of ε. The true expectile function in this model is $\mathbf{X}^{\mathrm{T}}\beta + 2b_\tau\exp(-0.35X_2 - 1.1X_4)$, where b_τ is the τth expectile of $N(0,1)$. Intuitively, for some expectile levels (such as τ near 0.5), the performance of parametric estimators may dominate that of nonparametric estimators, while the opposite is true at the other expectile levels (such as larger or small τ values).

For the simulation study, we considered three candidate estimators. The multiple linear expectile regression and nonparametric multiple expectile regression via gradient boosting arise as the two

natural candidates. For illustration purpose, we also considered a slightly more complicated version of the linear expectile regression by including an additional interaction term between X_2 and X_4 besides all the main effects $X_1, \ldots, X_6$. We included such a linear interaction model because it is also a practically popular model in applications. The simulations were repeated $M = 100$ times. For each simulation, we first generated a training set of $N = 500$ observations and applied the three aforementioned candidate estimators, plus a five-fold cross-validation to select the best procedure, as well as an aggregation of the candidate estimators using AEREW (Algorithm 2) to obtain the estimates. In AEREW, the weights were averaged over $B = 200$ splits with split size $N_0 = 400$. Equal prior weights were assigned for the three candidate estimators for each split. We set $\lambda = 1$. The estimated prediction risks and MSDs, defined in (8), were reported in Table 2 for independent test sets of $N_1 = 10000$ observations from model (9). In addition, we also included the proportion p_{CV} of each candidate estimator being selected by the cross-validation among the 100 independent simulations.

Table 2 about here.

It can be seen from Table 2 that none of the three candidate estimator is universally better than the others for all expectile levels. Indeed, when τ is in the middle of its range (around 0.5), the linear expectile regression with only main effects compares favorably with the other two procedures. This is expected since the true expectile function is mainly linear. For extreme expectile levels (τ away from 0.5), the linear expectile regression with interaction and the gradient boosting outperform the linear expectile regression with main effects only. From the patterns of p_{CV}, it can be seen that for expectile levels that are close to 0, 1, or 0.5, there is a clear dominating candidate estimator which is selected by cross-validation with high probability. While for moderate expectile levels, there are usually two competing candidate estimators (see e.g. $\tau = 0.25$ and 0.75). AEREW outperforms the three candidate estimators as well as the cross-validated estimator at all expectile levels. Moreover, we observe a greater gain by using AEREW for moderate expectile levels when cross-validation

experiences difficulty in selecting a clear winner.

4 Personal Computer Data

We apply AEREW to a data set described in Stengos and Zacharias (2006). The data set contains monthly price information of personal computers from January 1993 to November 1995 and was analyzed using a hedonic analysis. There are $n = 6259$ observations with 10 variables. The response variable is *Price*, and the hedonic variables *Speed*, *HD*, *RAM*, *Screen*, *CD*, *Multi*, and *Premium* directly describe the major hedonic characteristics that make up a computer. The other two explanatory variables *ADs* and *Trend* are not directly related to the personal computer characteristics, but are believed to be associated with the price. For example, it might be interesting to see if aggressive advertising is associated with lower price, or if an intertemporal effect exists among the hedonic components of personal computers. After inspecting the data, we decided to take the logarithmic transformation on all continuous variables except *Trend*. We considered a hedonic analysis at different price levels using expectile regression. Three candidate models were considered: the multiple linear expectile regression with main effects only, the linear expectile regression with main effects and two-way interactions, and the nonparametric approach via boosting. We applied a five-fold cross-validation to select the best procedure among the three candidates. Finally, we used AEREW to aggregate the three candidate models.

For the analysis, we randomly sampled $N = 3129$ observations from the data to form a training set, on which the three aforementioned candidate estimators were fitted and a five-fold cross-validation was applied to select the best procedure. To aggregate the candidate estimators through AEREW (Algorithm 2), we averaged the weights over $B = 200$ splits with split size $N_0 = 2503$. Equal prior weights were assigned and we set $\lambda = 1$. The prediction risks of the procedures, as defined in (8), were evaluated on the remaining $N_1 = 3130$ observations. This procedure was repeated $M = 100$ times and the results are summarized in Table 3 and Figure 1. The proportions of

the candidate estimators being selected by the five-fold cross-validation are reported in p_{CV}.

Table 3 about here.

Figure 1 about here.

It can be seen from both Table 3 and Figure 1 that at all expectile levels, the linear expectile regression with interactions outperforms the linear expectile regression with main effects only. The boosting estimator performs better for middle range of the expectile levels, while the linear expectile regression with interactions is better for extreme expectile levels. At 0.10 level, boosting and the linear model with interactions are very comparable to each other. This can be seen from the values of p_{CV}. However, it is also very clear that if we consider multiple expectile levels, there is no clear winner among the candidate estimators. It is interesting to see that AEREW gives smaller prediction errors than all candidates and the cross-validated estimator at all expectile levels in Table 3.

Appendix: Proofs

We present here the proofs of all theoretical results in previous sections along with a few technical lemmas. The first lemma concerns the smoothness of the asymmetric squared error loss. For ease of notation, let $w_\tau(u) = |\tau - I(u < 0)|$. Observe that $\underline{c} \le w_\tau(u) \le \bar{c}$ for all $u \in \mathbb{R}$.

Lemma 1. *The asymmetric squared error loss $\Psi_\tau(u)$ has Lipschitz continuous derivative,*

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| \le 2\bar{c}|u-u_0|,\ \forall u, u_0 \in \mathbb{R}. \tag{10}$$

Moreover, $\Psi_\tau(u)$ satisfies

$$\underline{c}(u-u_0)^2 \le \Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \le \bar{c}(u-u_0)^2,\ \forall u, u_0 \in \mathbb{R}. \tag{11}$$

Proof. We first prove the inequalities in (10). Note that $\Psi'_\tau(u) = 2w_\tau(u)u$. If $u = 0$ or $u_0 = 0$, then the inequalities in (10) hold trivially. If $uu_0 > 0$, we must have $w_\tau(u) = w_\tau(u_0)$. It follows that

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| = 2w_\tau(u)|u-u_0| \le 2\bar{c}|u-u_0|.$$

If instead $uu_0 < 0$, there are two cases: $u > 0, u_0 < 0$ or $u < 0, u_0 > 0$. For the first case, we have

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| = 2\tau u - 2(1-\tau)u_0 \le 2\bar{c}|u-u_0|.$$

For the second case, we have

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| = -2(1-\tau)u + 2\tau u_0 \le 2\bar{c}|u-u_0|.$$

This establishes the inequalities in (10).

Next we prove the inequalities in (11). Note that the second inequality in (11) follows from the

second inequality in (10) by Theorem 2.1.5 of Nesterov (2004). To prove the first inequality in (11), note that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \\
=& w_\tau(u)u^2 - w_\tau(u_0)u_0^2 - 2w_\tau(u_0)u_0(u-u_0) \\
=& w_\tau(u_0)(u-u_0)^2 + \{w_\tau(u) - w_\tau(u_0)\}u^2.
\end{aligned}$$

If $w_\tau(u) \geq w_\tau(u_0)$, then obviously we get

$$\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \geq w_\tau(u_0)(u-u_0)^2 \geq \underline{c}(u-u_0)^2.$$

If $w_\tau(u) < w_\tau(u_0)$, then we have $\underline{c} = w_\tau(u)$, $\bar{c} = w_\tau(u_0)$ and $u_0 u \leq 0$. It follows that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \\
=& \underline{c}u^2 - 2\bar{c}u_0(u-u_0) - \bar{c}u_0^2 \\
\geq& \underline{c}u^2 - 2\underline{c}u_0 a + \underline{c}u_0^2 = \underline{c}(u-u_0)^2.
\end{aligned}$$

Therefore, we have established the first inequality in (11). This completes the proof of Lemma 1. □

The second lemma deals with sub-exponential random variables. See Vershynin (2010) for a thorough treatment of sub-exponential random variables.

Lemma 2. *Let ζ be a centered sub-exponential random variable, whose sub-exponential norm satisfies $K = \|\zeta\|_{SEXP} = \sup_{k\geq 1} k^{-1}(\mathbb{E}|\zeta|^k)^{1/k} \in (0,\infty)$. Then,*

(a). $\mathbb{E}\exp(t|\zeta|) \leq 2\exp(CK^2t^2), \forall |t| \leq c/K$, *where* $C = 2e^2, c = 1/(2e)$ *and* $e = \exp(1)$.

(b). Let $\eta_\tau = \Psi'_\tau(\zeta - \mathscr{E}^\tau(\zeta)) = 2(\zeta - \mathscr{E}^\tau(\zeta))|\tau - I(\zeta < \mathscr{E}^\tau(\zeta))|$ *for* $\tau \in (0,1)$. *Then* η_τ *is also*

centered and satisfies

$$\mathbb{E}\exp(t|\eta_\tau|) \leq 2\exp(CK_\tau^2 t^2), \forall |t| \leq c/K_\tau,$$

and

$$\mathbb{E}\{|\eta_\tau|^2 \exp(t|\eta_\tau|)\} \leq 16\sqrt{2}K_\tau^2 \exp(2C^2K_\tau^2 t^2), \forall |t| \leq c/(2K_\tau),$$

where $K_\tau = \|\eta_\tau\|_{SEXP} = \sup_{k\geq 1} k^{-1}(\mathbb{E}|\eta_\tau|^k)^{1/k}$ *is the sub-exponential norm of* η_τ *satisfying that* $K_\tau \leq 2\bar{c}\{K + |\mathscr{E}^\tau(\zeta)|\}$.

Proof. Let us first show result (a). It follows directly from Lemma 5.15 of Vershynin (2010) that $\mathbb{E}\exp(t\zeta) \leq \exp(CK^2t^2), \forall |t| \leq c/K$. Let F be the CDF of ζ. For $|t_0| \leq c/K$ and $t_0 \geq 0$, we have $\mathbb{E}\exp(t_0\zeta) \leq \exp(CK^2t_0^2)$ and $\mathbb{E}\exp(-t_0\zeta) \leq \exp(CK^2t_0^2)$. Then it follows that

$$\int_0^\infty \exp(t_0 z)\,\mathrm{d}F(z) \leq \exp(CK^2t_0^2) \quad \text{and} \quad \int_{-\infty}^0 \exp(-t_0 z)\,\mathrm{d}F(z) \leq \exp(CK^2t_0^2).$$

Thus, we have

$$\mathbb{E}\exp(t_0|\zeta|) = \int_0^\infty \exp(t_0 z)\,\mathrm{d}F(z) + \int_{-\infty}^0 \exp(-t_0 z)\,\mathrm{d}F(z) \leq 2\exp(CK^2t_0^2).$$

Now for any $t \in [-c/K, c/K]$, we have $\mathbb{E}\exp(t|\zeta|) \leq \mathbb{E}\exp(|t|\cdot|\zeta|) \leq 2\exp(CK^2t^2)$. This completes the proof of result (a).

For result (b), first note that by definition of $\mathscr{E}^\tau(\zeta)$, we conclude that $\mathbb{E}(\eta_\tau) = 0$. By Minkowski inequality, we have $K_\tau \leq 2\bar{c}\{K + |\mathscr{E}^\tau(\zeta)|\} < \infty$. Thus, η_τ is also a sub-exponential random variable. The upper bound on the moment generating function of $|\eta_\tau|$ follows naturally from result (a). For $\mathbb{E}\{|\eta_\tau|^2\exp(t|\eta_\tau|)\}$, note that by Cauchy-Schwarz inequality we have

$$\mathbb{E}\{|\eta_\tau|^2 \exp(t|\eta_\tau|)\} \leq (\mathbb{E}|\eta_\tau|^4)^{1/2}\{\mathbb{E}\exp(2t|\eta_\tau|)\}^{1/2},$$

for which $(\mathbb{E}|\eta_\tau|^4)^{1/2} = \{(\mathbb{E}|\eta_\tau|^4)^{1/4}\}^2 \le (4K_\tau)^2$ and $\{\mathbb{E}\exp(2t|\eta_\tau|)\}^{1/2} \le \sqrt{2}\exp(2CK_\tau^2 t^2)$ for any $|t| \le c/(2K_\tau)$. Result (b) then follows. □

Proof of Theorem 1. We first prive the oracle inequality for AEREW by Algorithm 1. The same proof works for AEREW by Algorithm 2 with a small modification which will be explained later.

Let $q_{n_0}^n = \sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^n \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\right\}$. Observe that

$$\begin{aligned} q_{n_0}^n &= \sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \Psi_\tau(y_{n_0+1} - \hat{e}_{\tau,j,n_0}(\mathbf{x}_{n_0+1}))\right\} \\ &\quad \times \frac{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n_0+2} \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\right\}}{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \Psi_\tau(y_{n_0+1} - \hat{e}_{\tau,j,n_0}(\mathbf{x}_{n_0+1}))\right\}} \\ &\quad \times \cdots \times \frac{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n} \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\right\}}{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n-1} \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\right\}} \\ &= \prod_{i=n_0+1}^n \left(\sum_{j=1}^\infty W_{j,i} \exp\left\{-\lambda \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\right\} \right). \end{aligned}$$

Fix $i \in \{n_0+1, \ldots, n\}$. Let J be the discrete random variable such that $\mathbb{P}(J=j) = W_{j,i}$, $j \ge 1$. Let ν be the discrete measure induced by J on $\mathbb{Z}^+$ such that $\nu(j) = \mathbb{P}(J=j) = W_{j,i}$, $j \ge 1$. For ease of notation, denote $h(J) = -\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))$. It follows that

$$\sum_{j=1}^\infty W_{j,i} \exp\{-\lambda \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\} = \mathbb{E}_\nu \exp\{\lambda h(J)\}.$$

By Lemma 3.6.1 of Catoni and Picard (2004, p. 85), we have

$$\log \mathbb{E}_\nu \exp\{\lambda h(J)\} \le \lambda \mathbb{E}_\nu(h(J)) + \frac{\lambda^2}{2} \mathrm{var}_\nu(h(J)) \exp\left[\lambda \max\left\{0, \sup_{\gamma \in [0,\lambda]} \frac{M^3_{\nu_\gamma}(h(J))}{\mathrm{var}_{\nu_\gamma}(h(J))}\right\}\right], \quad (12)$$

where the induced measure $\nu_\gamma, \gamma \in [0,\lambda]$ is given by

$$\nu_\gamma(j) = \frac{W_{j,i}\exp(\gamma h(j))}{\sum_{j'=1}^{\infty} W_{j',i}\exp(\gamma h(j'))},\ j \geq 1,$$

and $M^3_{\nu_\gamma}(h(J)) = \mathbb{E}_{\nu_\gamma}\{h(J) - \mathbb{E}_{\nu_\gamma}h(J)\}^3$ is the third central moment.

To facilitate the presentation, let $b_\tau(\mathbf{x}) = \mathscr{E}^\tau(\varepsilon|\mathbf{x})$ be the conditional τth expectile of the random error ε given $\mathbf{X} = \mathbf{x}$. It can be seen that the expectile function $e_\tau(\mathbf{x}) = m(\mathbf{x}) + \sigma(\mathbf{x})b_\tau(\mathbf{x})$. By Lemma 1, it can be shown that

$$\begin{aligned}
&\sup_{\gamma\in[0,\lambda]} \frac{M^3_{\nu_\gamma}(h(J))}{\mathrm{var}_{\nu_\gamma}(h(J))} \leq \sup_{\gamma\in[0,\lambda]}\sup_{j\geq 1}|h(j) - \mathbb{E}_{\nu_\gamma}(h(J))| \leq \sup_{j_1,j_2\geq 1}|h(j_1) - h(j_2)| \\
&\leq 2\sup_{j\geq 1}|\Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - e_\tau(\mathbf{x}_i))| \\
&\leq 2\sigma(\mathbf{x}_i)|\Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}))|\sup_{j\geq 1}|\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i)| + 2\bar{c}\sup_{j\geq 1}(\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i))^2
\end{aligned}$$

and that

$$\begin{aligned}
&\mathrm{var}_\nu(h(J)) \leq \mathbb{E}_\nu\left\{\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\}^2 \\
&\leq \sup_{j\geq 1}\left(|\Psi'_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))| + \bar{c}|\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)|\right)^2\mathbb{E}_\nu\left(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)\right)^2 \\
&\leq \left\{\sigma(\mathbf{x}_i)|\Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}))| + 4\bar{c}\sup_{j\geq 1}|\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i)|\right\}^2\mathbb{E}_\nu\left(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)\right)^2.
\end{aligned}$$

Also from Lemma 1, we get that

$$\begin{aligned}
&\Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) \\
&\geq \Psi'_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))(\mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i)) + \underline{c}(\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))^2.
\end{aligned}$$

Taking expectation with respect to J of both sides of the above inequality, we have

$$\mathbb{E}_v(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_v\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))^2 \le \underline{c}^{-1}\big\{\mathbb{E}_v\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_v\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\}.$$

Let $\xi_i = \Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}))$. It follows from inequality (12) and assumptions (C1-C3) that with probability one

$$\begin{aligned}
&\log \mathbb{E}_v \exp\{\lambda h(J)\} \\
&\le \lambda \mathbb{E}_v(h(J)) + 2^{-1}\lambda^2(C_0|\xi_i| + 4\bar{c}A_\tau)^2 \exp\big\{2\lambda C_0 A_\tau|\xi_i| + 2\lambda\bar{c}A_\tau^2\big\} \\
&\times \underline{c}^{-1}\big\{\mathbb{E}_v\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_v\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\} \qquad (13)\\
&\le -\lambda\mathbb{E}_v\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) + \lambda^2\underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left(C_0^2|\xi_i|^2 + 16\bar{c}^2A_\tau^2\right)\exp(2\lambda C_0A_\tau|\xi_i|) \\
&\times \big\{\mathbb{E}_v\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_v\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\}.
\end{aligned}$$

Take the expectation (denoted by $\mathbb{E}_i$) of both sides of (13) with respect to Y_i conditional on $\mathbf{x}_i \cup (\mathbf{x}_k, y_k)_{k=1}^{i-1}$. By Lemma 2, when λ is chosen small enough such that $2\lambda C_0A_\tau \le (4eK_\tau)^{-1}$, with probability one we have

$$\begin{aligned}
&\mathbb{E}_i \log\left(\mathbb{E}_v \exp\{-\lambda\Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\}\right) \\
&\le -\lambda\mathbb{E}_i\big\{\mathbb{E}_v\Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\} \\
&+ \lambda^2\underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2\mathscr{M}_2(2\lambda C_0A_\tau) + 16\bar{c}^2A_\tau\mathscr{M}_0(2\lambda C_0A_\tau)\right\} \\
&\times \mathbb{E}_i\big[\mathbb{E}_v\Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(Y_i - \mathbb{E}_v\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big].
\end{aligned}$$

Moreover, if λ also satisfies

$$\lambda^2\underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2\mathscr{M}_2(2\lambda C_0A_\tau) + 16\bar{c}^2A_\tau\mathscr{M}_0(2\lambda C_0A_\tau)\right\} \le \lambda,$$

with probability one we will have

$$\mathbb{E}_i \log\left(\mathbb{E}_v \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\}\right) \leq -\lambda \mathbb{E}_i \Psi_\tau(Y_i - \mathbb{E}_v \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)),$$

since by convexity of $\Psi_\tau(\cdot)$ and Jensen's inequality we have

$$\Psi_\tau(Y_i - \mathbb{E}_v \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) \leq \mathbb{E}_v \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)).$$

It follows that when λ is small enough such that condition (3) holds, we have

$$\begin{aligned}
\mathbb{E}\log(1/q_{n_0}^n) &= -\sum_{i=n_0+1}^{n} \mathbb{E}\log\left(\sum_{j=1}^{\infty} W_{j,i} \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,j,n_0}(\mathbf{X}_i))\right\}\right) \\
&= -\sum_{i=n_0+1}^{n} \mathbb{E}\left[\mathbb{E}_i \log\left(\mathbb{E}_v \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{X}_i))\right\}\right)\right] \\
&\geq \lambda \mathbb{E}\left[\sum_{i=n_0+1}^{n} \mathbb{E}_i \Psi_\tau\left(Y_i - \sum_{j=1}^{\infty} W_{j,i} \hat{e}_{\tau,j,n_0}(\mathbf{X}_i)\right)\right] \\
&= \lambda \sum_{i=n_0+1}^{n} \mathbb{E}\Psi_\tau\left(Y - \sum_{j=1}^{\infty} W_{j,i} \hat{e}_{\tau,j,n_0}(\mathbf{X})\right).
\end{aligned}$$

The last equality is due to the independence of the observations, i.e., $(\mathbf{X}, Y)$ is independent of $(\mathbf{X}_i, Y_i)_{i=1}^n$. On the other hand, we have, for each $j^* \geq 1$,

$$\begin{aligned}
\mathbb{E}\log(1/q_{n_0}^n) &\leq \log(1/\pi_{j^*}) + \lambda \sum_{i=n_0+1}^{n} \mathbb{E}\Psi_\tau(Y_i - \hat{e}_{\tau,j^*,n_0}(\mathbf{X}_i)) \\
&= \log(1/\pi_{j^*}) + \lambda (n - n_0) \mathbb{E}\Psi_\tau(Y - \hat{e}_{\tau,j^*,n_0}(\mathbf{X})).
\end{aligned}$$

Therefore, for any $j^* \geq 1$, we have

$$\frac{1}{n-n_0} \sum_{i=n_0+1}^{n} \mathbb{E}\Psi_\tau\left(Y - \sum_{j=1}^{\infty} W_{j,i} \hat{e}_{\tau,j,n_0}(\mathbf{X})\right) \leq \frac{\log(1/\pi_{j^*})}{\lambda(n-n_0)} + \mathbb{E}\Psi_\tau(Y - \hat{e}_{\tau,j^*,n_0}(\mathbf{X})). \tag{14}$$

Note that by definition of $\hat{e}_{\tau,\cdot,n}(\mathbf{x})$, we have

$$y-\hat{e}_{\tau,\cdot,n}(\mathbf{x})=\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left(y-\sum_{j=1}^{\infty}W_{j,i}\hat{e}_{\tau,j,n_0}(\mathbf{x})\right).$$

It follows from (14) and convexity of $\Psi_\tau(\cdot)$ that for each $j^*\geq 1$,

$$\begin{aligned}\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))&\leq\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau\left(Y-\sum_{j=1}^{\infty}W_{j,i}\hat{e}_{\tau,j,n_0}(\mathbf{X})\right)\\&\leq\frac{\log(1/\pi_{j^*})}{\lambda(n-n_0)}+\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j^*,n_0}(\mathbf{X})).\end{aligned}$$

This completes the proof of inequality (4). To show (5), note that by Lemma 1

$$\begin{aligned}\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j^*,n_0}(\mathbf{X}))&\leq\mathbb{E}\Psi_\tau(Y-e_\tau(\mathbf{X}))+\bar{c}\mathbb{E}(e_\tau(\mathbf{X})-\hat{e}_{\tau,j^*,n_0}(\mathbf{X}))^2\\\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))&\geq\mathbb{E}\Psi_\tau(Y-e_\tau(\mathbf{X}))+\underline{c}\mathbb{E}(e_\tau(\mathbf{X})-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))^2\end{aligned}$$

due to the fact that $\mathbb{E}\{\Psi_\tau(Y-e_\tau(\mathbf{X}))|\mathbf{X}\}=0$. Inequality (5) then follows from (4).

To prove the same result for AEREW by Algorithm 2, we use the convexity of $\Psi_\tau(\cdot)$ and have

$$\Psi_\tau\big(y-\hat{e}^B_{\tau,\cdot,n}(\mathbf{x})\big)\leq\frac{1}{B}\sum_{k=1}^{B}\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\Psi_\tau\left(y-\sum_{j=1}^{\infty}W^{(k)}_{j,i}\hat{e}^{(k)}_{\tau,j,n_0}(\mathbf{x})\right).$$

The result then follows from the previous proof for AEREW by Algorithm 1. □

References

ACERBI, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance* **26**, 1487–1503.

AIGNER, D. J., AMEMIYA, T. and POIRIER, D. J. (1976). On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review* **17**, 377–396.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Mathematical Finance* **9**, 203–228.

BELLINI, F. and BIGNOZZI, V. (2015). On elicitable risk measures. *Quantitative Finance* **15**, 725–733.

BERNARDO, A. E. and LEDOIT, O. (2000). Gain, loss, and asset pricing. *Journal of political economy* **108**, 144–172.

CATONI, O. and PICARD, J. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, vol. 1851. Springer.

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**, 93–125.

FAN, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association* **87**, 998–1004.

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

GNEITING, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762.

GUO, M. and HÄRDLE, W. K. (2012). Simultaneous confidence bands for expectile functions. *AStA Advances in Statistical Analysis* **96**, 517–541.

HEIDENREICH, N.-B., SCHINDLER, A. and SPERLICH, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis* **97**, 403–433.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science* **14**, 382–401.

KUAN, C.-M., YEH, J.-H. and HSU, Y.-C. (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *Journal of Econometrics* **150**, 261–270.

NESTEROV, Y. (2004). *Introductory lectures on convex optimization*, vol. 87. Springer Science & Business Media.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society* **55**, 819–847.

SCHNABEL, S. K. and EILERS, P. H. (2009a). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research* **21**, 109–134.

SCHNABEL, S. K. and EILERS, P. H. (2009b). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**, 4168–4177.

SCHWARZ, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* **6**, 461–464.

SOBOTKA, F., KAUERMANN, G., WALTRUP, L. S. and KNEIB, T. (2013a). On confidence intervals for semiparametric expectile regression. *Statistics and Computing* **23**, 135–148.

SOBOTKA, F., RADICE, R., MARRA, G. and KNEIB, T. (2013b). Estimating the relationship between women's education and fertility in botswana by using an instrumental variable approach to semiparametric expectile regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**, 25–45.

STENGOS, T. and ZACHARIAS, E. (2006). Intertemporal pricing and price discrimination: a semi-parametric hedonic analysis of the personal computer market. *Journal of Applied Econometrics* **21**, 371–386.

TAYLOR, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* **6**, 231–252.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

YANG, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–588.

YANG, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory* **20**, 176–222.

YANG, Y. and ZOU, H. (2015). Nonparametric multiple expectile regression via er-boost. *Journal of Statistical Computation and Simulation* **85**, 1442–1458.

YAO, Q. and TONG, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics* **6**, 273–292.

ZIEGEL, J. F. (2014). Coherence and elicitability. *Mathematical Finance* .

Tables and Figures

Algorithm 1: The aggregated expectile regression by exponential weighting (AEREW) – Single split.

1. Randomly split the data into two parts. Without loss of generality, denote the two parts by $D^{(0)} = (\mathbf{x}_i, y_i)_{i=1}^{n_0}$ and $D^{(1)} = (\mathbf{x}_i, y_i)_{i=n_0+1}^{n}$ respectively, where $D^{(0)}$ is used for training and $D^{(1)}$ is used for evaluation.

2. For each procedure δ_j, obtain the estimate $\hat{e}_{\tau,j,n_0}(\mathbf{x}_i)$ of $e_\tau(\mathbf{x}_i)$ for every $\mathbf{x}_i \in D^{(1)}$ based on the training data $D^{(0)}$, $n_0+1 \le i \le n$, $j \ge 1$.

3. Set $W_{j,n_0+1} = \pi_j$ such that $\pi_j \ge 0$, $j \ge 1$, and $\sum_{j=1}^{\infty} \pi_j = 1$, and calculate the weights

$$W_{j,i} = \frac{\pi_j \exp\left\{-\lambda \sum_{k=n_0+1}^{i-1} \Psi_\tau(y_k - \hat{e}_{\tau,j,n_0}(\mathbf{x}_k))\right\}}{\sum_{j'=1}^{\infty} \pi_{j'} \exp\left\{-\lambda \sum_{k=n_0+1}^{i-1} \Psi_\tau(y_k - \hat{e}_{\tau,j',n_0}(\mathbf{x}_k))\right\}}, j \ge 1, n_0+2 \le i \le n.$$

   Obtain the aggregating procedure δ_a which estimates $e_\tau(\mathbf{x})$ by

$$\hat{e}_{\tau,\cdot,n}(\mathbf{x}) = \sum_{j=1}^{\infty} \left(\sum_{i=n_0+1}^{n} \frac{W_{j,i}}{n-n_0} \right) \hat{e}_{\tau,j,n_0}(\mathbf{x}).$$

Algorithm 2: The aggregated expectile regression by exponential weighting (AEREW) – Multiple splits.

1. Randomly split the data into two parts. Without loss of generality, denote the two parts by $D^{(0)} = (\mathbf{x}_i, y_i)_{i=1}^{n_0}$ and $D^{(1)} = (\mathbf{x}_i, y_i)_{i=n_0+1}^{n}$ respectively, where $D^{(0)}$ is used for training and $D^{(1)}$ is used for evaluation.

2. For each procedure δ_j, obtain the estimate $\hat{e}_{\tau,j,n_0}(\mathbf{x}_i)$ of $e_\tau(\mathbf{x}_i)$ for every $\mathbf{x}_i \in D^{(1)}$ using the training data $D^{(0)}$ for fitting, $n_0+1 \le i \le n$, $j \ge 1$.

3. Set $W_{j,n_0+1} = \pi_j$ such that $\pi_j \ge 0$, $j \ge 1$, and $\sum_{j=1}^{\infty} \pi_j = 1$, and calculate the weights

$$W_{j,i} = \frac{\pi_j \exp\left\{-\lambda \sum_{k=n_0+1}^{i-1} \Psi_\tau(y_k - \hat{e}_{\tau,j,n_0}(\mathbf{x}_k))\right\}}{\sum_{j'=1}^{\infty} \pi_{j'} \exp\left\{-\lambda \sum_{k=n_0+1}^{i-1} \Psi_\tau(y_k - \hat{e}_{\tau,j',n_0}(\mathbf{x}_k))\right\}}, \; j \ge 1, \, n_0+2 \le i \le n.$$

4. Repeat the above three steps $(B-1)$ times. Denote the estimates and weights from the kth random split by $\hat{e}^{(k)}_{\tau,j,n_0}(\mathbf{x})$ and $W^{(k)}_{j,i}$, $n_0 \le i \le n$, $j \ge 1$, $1 \le k \le B$, respectively. Obtain the aggregating procedure δ_a^B which estimates $e_\tau(\mathbf{x})$ by

$$\hat{e}^B_{\tau,\cdot,n}(\mathbf{x}) = \sum_{j=1}^{\infty} \sum_{i=n_0+1}^{n} \sum_{k=1}^{B} \frac{W^{(k)}_{j,i}}{B(n-n_0)} \hat{e}^{(k)}_{\tau,j,n_0}(\mathbf{x}).$$

Table 1: Estimated prediction risks and MADs of local linear regressions with five candidate bandwidths, the five-fold cross-validated kernel estimator, and AEREW ($\lambda = 1$) for the heteroscedastic model (7). The numbers listed are averages over 100 independent runs with their respective standard errors reported in the parentheses. The proportion of each candidate estimator being selected by the five-fold cross-validation among these 100 runs is reported by p_{CV}. All numbers are of order 10^{-2} except those corresponding to p_{CV}

		Bandwidth (h)					Cross-validation	AEREW
τ	Measure	0.0347	0.104	0.173	0.243	0.312		
0.05	risk	3.71 (0.03)	2.98 (0.02)	2.90 (0.02)	2.90 (0.02)	2.92 (0.02)	2.93 (0.02)	2.89 (0.02)
	MSD	25.16 (0.31)	16.56 (0.39)	15.11 (0.36)	15.72 (0.30)	16.48 (0.25)	16.03 (0.33)	14.67 (0.32)
	p_{CV}	0.00	0.23	0.23	0.12	0.42	–	–
0.10	risk	4.98 (0.06)	4.13 (0.03)	4.06 (0.02)	4.12 (0.02)	4.20 (0.02)	4.17 (0.03)	4.07 (0.02)
	MSD	22.30 (0.39)	13.93 (0.37)	13.05 (0.31)	14.42 (0.24)	15.83 (0.20)	14.49 (0.40)	13.04 (0.29)
	p_{CV}	0.03	0.34	0.28	0.14	0.21	–	–
0.25	risk	6.69 (0.06)	5.82 (0.03)	5.85 (0.03)	6.05 (0.03)	6.26 (0.03)	5.88 (0.03)	5.86 (0.03)
	MSD	17.60 (0.39)	10.22 (0.22)	10.71 (0.20)	13.10 (0.18)	15.25 (0.16)	10.99 (0.29)	10.84 (0.21)
	p_{CV}	0.00	0.41	0.46	0.03	0.10	–	–
0.50	risk	7.63 (0.06)	6.72 (0.03)	6.78 (0.03)	7.04 (0.03)	7.37 (0.03)	6.78 (0.04)	6.81 (0.03)
	MSD	16.05 (0.32)	8.96 (0.23)	9.58 (0.21)	12.10 (0.17)	14.55 (0.15)	9.54 (0.25)	9.92 (0.22)
	p_{CV}	0.00	0.54	0.42	0.03	0.01	–	–
0.75	risk	6.76 (0.10)	5.88 (0.03)	5.92 (0.03)	6.13 (0.03)	6.41 (0.04)	5.96 (0.04)	5.94 (0.03)
	MSD	17.62 (0.51)	10.23 (0.25)	10.61 (0.23)	12.77 (0.19)	15.04 (0.16)	10.90 (0.29)	10.81 (0.23)
	p_{CV}	0.01	0.43	0.42	0.12	0.02	–	–
0.90	risk	4.98 (0.06)	4.17 (0.03)	4.09 (0.03)	4.16 (0.03)	4.31 (0.03)	4.15 (0.03)	4.10 (0.03)
	MSD	22.51 (0.35)	14.47 (0.36)	13.08 (0.32)	14.26 (0.26)	16.18 (0.20)	14.02 (0.32)	13.26 (0.31)
	p_{CV}	0.02	0.26	0.42	0.20	0.10	–	–
0.95	risk	3.87 (0.04)	3.08 (0.03)	2.97 (0.03)	3.00 (0.03)	3.08 (0.03)	3.04 (0.03)	2.99 (0.03)
	MSD	26.78 (0.34)	18.41 (0.45)	16.18 (0.46)	16.62 (0.43)	18.20 (0.38)	17.37 (0.45)	16.16 (0.42)
	p_{CV}	0.00	0.29	0.30	0.25	0.16	–	–

Table 2: Estimated prediction risks, MSDs and their respective standard errors (in parentheses) of the linear expectile regression with only main effects, linear expectile regression with interaction, nonparametric expectile regression via boosting, the five-fold cross-validation and AEREW ($\lambda = 1$) for the heteroscedastic model (9) over 100 independent runs. The proportion of each candidate estimator being selected by the cross-validation is summarized by p_{CV}

		Individual			Cross-validation	Aggregation
τ	Measure	Linear	Interaction	Boosting		
0.05	Risk	53.58 (4.20)	49.07 (4.03)	53.39 (4.03)	49.56 (4.06)	47.71 (3.97)
	MSD	12.08 (0.39)	10.88 (0.39)	11.74 (0.30)	10.95 (0.39)	10.39 (0.30)
	p_{CV}	0.06	0.87	0.07	–	–
0.10	Risk	57.71 (2.86)	55.33 (2.84)	60.13 (2.82)	54.91 (2.76)	53.92 (2.74)
	MSD	8.41 (0.21)	7.74 (0.25)	9.13 (0.20)	7.78 (0.21)	7.59 (0.18)
	p_{CV}	0.17	0.71	0.12	–	–
0.25	Risk	64.87 (2.40)	66.14 (3.21)	69.06 (2.18)	65.86 (3.20)	62.99 (2.14)
	MSD	4.49 (0.16)	4.44 (0.24)	5.77 (0.15)	4.45 (0.24)	4.14 (0.13)
	p_{CV}	0.48	0.47	0.05	–	–
0.50	Risk	71.20 (2.41)	72.37 (2.45)	76.07 (2.45)	71.84 (2.44)	71.52 (2.41)
	MSD	1.51 (0.07)	2.01 (0.11)	3.39 (0.10)	1.75 (0.10)	1.67 (0.08)
	p_{CV}	0.71	0.21	0.08	–	–
0.75	Risk	67.78 (3.23)	67.34 (3.18)	70.07 (3.23)	67.62 (3.19)	66.62 (3.20)
	MSD	4.34 (0.11)	4.18 (0.13)	5.06 (0.11)	4.29 (0.12)	4.07 (0.11)
	p_{CV}	0.47	0.43	0.10	–	–
0.90	Risk	54.98 (3.26)	53.48 (3.50)	53.86 (3.09)	51.44 (3.03)	50.56 (3.05)
	MSD	8.40 (0.26)	7.78 (0.31)	8.41 (0.17)	7.64 (0.17)	7.46 (0.17)
	p_{CV}	0.20	0.66	0.14	–	–
0.95	Risk	46.76 (2.93)	43.32 (2.84)	45.91 (2.97)	43.83 (2.95)	42.22 (2.92)
	MSD	11.50 (0.25)	10.40 (0.24)	10.88 (0.22)	10.46 (0.23)	10.00 (0.20)
	p_{CV}	0.12	0.70	0.18	–	–

Table 3: Estimated prediction risks of the linear expectile regression with main effects only, the augmented linear expectile regression with interactions, the nonparametric expectile regression via boosting, the five-fold cross-validation, and AEREW ($\lambda = 1$) for the personal computer data. The measures are averaged over 100 random splits of the data and their corresponding standard errors are included in the parentheses. Proportions the candidate estimators being selected by the cross-validation are given by p_{CV}. All numbers are of order 10^{-3} except those corresponding to p_{CV}

		Individual			Cross-validation	Aggregation
τ	Measure	Linear	Interaction	Boosting		
0.05	risk	2.074 (0.005)	1.905 (0.005)	2.066 (0.009)	1.905 (0.005)	1.787 (0.005)
	p_{CV}	0.00	1.00	0.00	–	–
0.10	risk	3.327 (0.007)	3.069 (0.006)	3.046 (0.012)	3.065 (0.009)	2.778 (0.007)
	p_{CV}	0.00	0.53	0.47	–	–
0.25	risk	5.634 (0.010)	5.166 (0.009)	4.733 (0.015)	4.733 (0.015)	4.519 (0.011)
	p_{CV}	0.00	0.00	1.00	–	–
0.50	risk	6.973 (0.011)	6.294 (0.010)	5.558 (0.017)	5.558 (0.017)	5.427 (0.011)
	p_{CV}	0.00	0.00	1.00	–	–
0.75	risk	5.946 (0.012)	5.236 (0.010)	4.721 (0.018)	4.721 (0.018)	4.581 (0.012)
	p_{CV}	0.00	0.00	1.00	–	–
0.90	risk	3.728 (0.009)	3.203 (0.007)	3.066 (0.012)	3.086 (0.013)	2.896 (0.009)
	p_{CV}	0.00	0.13	0.87	–	–
0.95	risk	2.436 (0.005)	2.070 (0.005)	2.126 (0.010)	2.071 (0.006)	1.916 (0.006)
	p_{CV}	0.00	0.91	0.09	–	–

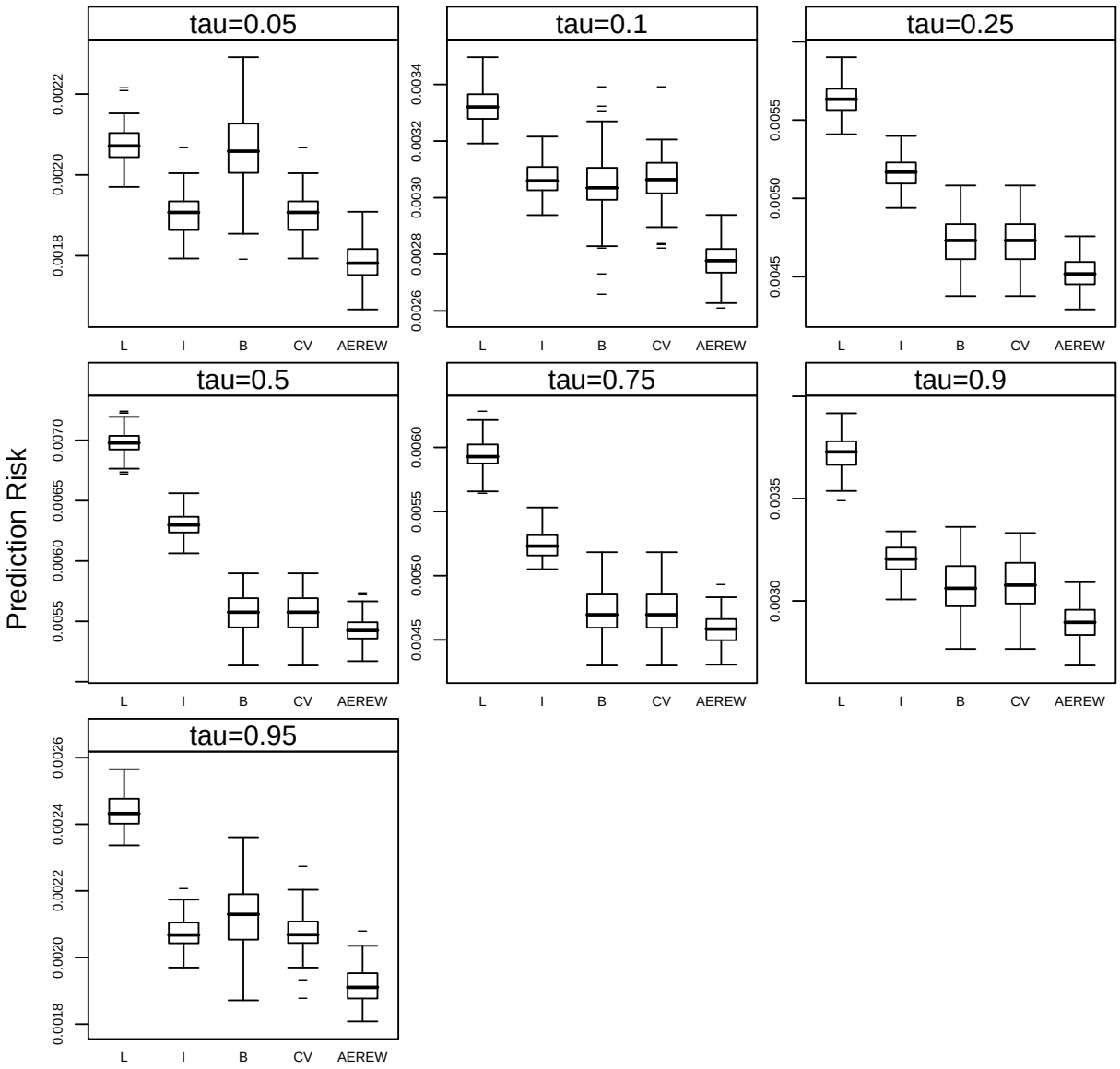

Figure 1: Estimated prediction risks of the linear expectile regression with only main effects, the augmented linear expectile regression with interactions, the nonparametric expectile regression via gradient boosting, the five-fold cross-validation, and AEREW based on 100 independent runs for the personal computer data. On the x-axis of each boxplot, "L" represents the linear expectile regression with only main effects, "I" denotes the augmented linear expectile regression with interactions, and "B" stands for the nonparametric expectile regression via gradient boosting. Each boxplot summarizes the results for one expectile level $\tau \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$.