Discussion of "Beyond mean regression" (T. Kneib)

Jochen Einbeck, Durham University, `jochen.einbeck@durham.ac.uk`

This very enjoyable article by Thomas Kneib looks at (semiparametric) regression from a very wide perspective, allowing us to model a wide range of features in the response distribution. Specifically, the features under study are quantiles and expectiles, which give access to almost any aspect of the response distribution that one may possibly be interested in, and which include mean and median regression as special cases. In the conclusion of his manuscript, the author mentions that modal regression has purposely been left out of this presentation. The main objective of this discussion will be to provide some hopefully useful contribution in this respect. I will begin with discussing (weighted) modal regression in general (and investigate whether it makes sense to define "modiles"), and proceed with some thoughts on the feasibility of *semiparametric* modal regression. I will finish this commentary with some general remarks on other aspects of the paper.

Modes and "Modiles". For a pair of real random variables (X, Y), conditional quantiles and expectiles of Y at $X = x$ are theoretically obtained as solutions of the minimization problem

$$\arg\min_q E\left(w_\tau(Y, q) l(Y - q) | X = x\right) \tag{1}$$

where $w_\tau(Y, q)$ is defined as in Section 5 of the main paper, and loss functions $\ell(\cdot) = |\cdot|$ and $\ell(\cdot) = (\cdot)^2$ for quantiles and expectiles, respectively. In order to extract the conditional mode, Matzner-Løber et al. (1998) proposed using a "non-convex loss function with a unique minimizer $l(u) = 0$ when $u = 0$ and $l(u) = 1$ otherwise". Equivalently, and slightly more elegantly, this can be formulated as

$$l(\cdot) = -\delta(\cdot)$$

where $\delta(\cdot)$ is the delta function, i.e. it takes the value 0 for each input except 0, and it integrates to 1. Plugging this loss function tentatively into (1), one obtains

$$\begin{aligned} & \arg\min_q E\left(w_\tau(Y,q)\{-\delta(Y-q)\}|X=x\right) = \\ = & \arg\max_q E\left(w_\tau(Y,q)\delta(Y-q)|X=x\right) \\ = & \arg\max_q \int w_\tau(y,q)\delta(y-q)f(y|x)\,dy \\ = & \arg\max_q w_\tau(q,q)f(q|x) \end{aligned} \tag{2}$$

Interestingly, this expression features the weight $w_\tau(q,q)$, the definition of which is "basically arbitrary" (Section 5) for expectile and quantile regression, and which is set equal to 0 in the paper under discussion. In the context of modal regression, we see that this would be an unfortunate choice, as in this case the entire argument of the minimization problem vanishes! Hence, for the sake of a flexible use of the weights w_τ over a wider range of loss functions, I would recommend to settle on a different convention, say $w_\tau(q,q) = 1/2$, or $w_\tau(q,q) = \tau$ as in Schulze Waltrup et al. (2013).

We see that all choices of $w_\tau(q,q) \equiv c$, for some constant $c > 0$, will lead to the same result (the overall mode, or the overall mode*s*, if there are several values of $y|x$ which simultaneously achieve the density maximum). So, in this sense the addition of asymmetric weights for modal regression is rather pointless, as they do not lead to new information; or, to put it in other words: all modiles are equal to the overall mode(s).

Beside the overall mode $m = \arg\max_q f(q|x)$, a conditional density $f(y|x)$ may have further local modes, associated with smaller densities than $f(m|x)$. An example is given, for the Munich rental data, in Figure 1a. Perhaps slightly disappointingly, we have seen that modiles are not a suitable tool to identify these directly. In the absence of strong overall trends, one may still be able to extract these local modes through a variant of the above approach, by defining weights

$$W_\tau(q) \equiv w_\tau(q,q)$$

which not only depend on τ but also indeed on q. For instance, the choice $W_\tau(q) = 1_{\{q \geq 0\}}$ would provide the conditional mode of all $Y \geq 0|x$. This way, the local modes could be found by examining the response distribution separately and successively in several vertical layers. This idea is in spirit somewhat similar to having several (vertically spread) starting points for the conditional mean shift procedure for modal regression, as explained in Einbeck & Tutz (2006).

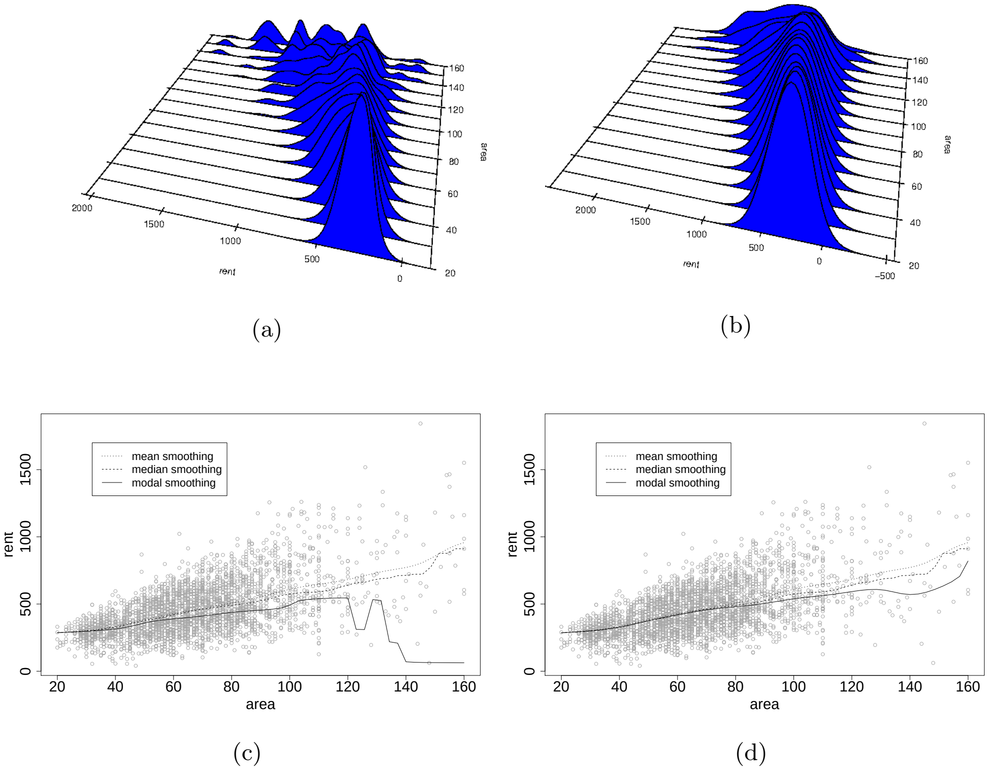
Figure 1 provides some examples for modal regression in the context of the Munich rental data (which were used in the discussed paper), where, as in Figure 1 (left) of the main paper, $x =$'living area' is the only predictor, and $y =$'rent' is the response variable. The left top panel shows estimated conditional densities $\hat{f}(y|x) = \hat{f}_{h,b}(x,y)/\hat{f}_h(x)$ for the rent data, where $\hat{f}_h(x)$ is a univariate kernel density estimator with bandwidth h, and $\hat{f}_{h,b}$ is a bivariate (product) kernel density estimator with bandwidths h and b in horizontal and vertical direction, respectively (using Gaussian kernels in each case). Using firstly the bandwidths $h = 8$ and $b = 50$ (these are automatic bandwidths as suggested by Bashtannyk and Hyndman's (2001) hybrid rule), the conditional densities, computed for $x = 20, 40, \ldots, 160$, show clear multimodality for $x \geq 120$. Starting the mean shift procedure from the bottom of the data range, the smallest of the conditional modes is identified, which corresponds to the overall mode up to $x \approx 120$ (Fig. 1c). The other conditional modes could be detected by using different starting points for the mean shift procedure, but the resulting modal curves are not provided for the sake of clarity of the graphical representation. The modal smoother is compared with a local median smoother (see e.g. Fried et al, 2007) and a local constant mean (Nadaraya–Watson) smoother, using in either case the same horizontal bandwidth $h = 8$.

We see that, due to the skewness of the conditional response distribution, we have the ordering mean $\geq$ median $\geq$ mode also over the range where the conditional density is in fact unimodal. If we increase the vertical bandwidths to $b = 150$, we see that the multimodalities disappear and the modal estimator becomes generally closer to the mean and the median smoother (Fig. 1b, 1d). In fact, if $b \longrightarrow \infty$, then the solid line will fall onto the dotted line; or, in other words, for infinite vertical bandwidth, the mean–shift–based modal regression estimator becomes equivalent to the Nadaraya–Watson–estimator (Taylor, 2012).

Modal smoothing beyond the scatterplot? As the author states fully correctly, modal smoothing is yet restricted to simple scenarios such as bivariate scatterplot smoothing. While simple extensions have been attempted, such as modal smoothing with bivariate or multivariate predictors (Taylor and Einbeck, 2011), there is still a long way to go to fit semiparametric models even of simple type as, say,

$$\text{mode}(Y|\boldsymbol{x}, z) = \eta = \boldsymbol{x}^T\beta + f(z), \tag{3}$$

with a univariate covariate z. Before attempting this, it should be clarified whether there is actually a necessity to go this way. There are two possible reasons why one may want to do this: Firstly, as the author suggests, this may be useful in situations where the response distribution is multimodal (and, hence, a multi–valued "function" f needs to be estimated). From a statistical modelling point of view, it should be noted here that a multimodal response distribution will often be a sign that "something is wrong", for instance that an important predictor has been omitted from the model, or that the experiment during which the data was collected was not sufficiently controlled and, hence, led to the creation of multiple latent regimes in the response distribution. While I do not want to exclude the case that in some situations a semiparametric model with multimodal response distribution could justifiably be fitted, it seems to me that it is another property of the mode which could make modal semiparametric regression indeed attractive: its robustness and edge–preserving properties. So, for instance, in cases where one explicitly wants to identify sudden changes or 'breakpoints' in the behavior of the system, (uni–!) modal regression may be an attractive choice as it will not smooth over the edges. How could such a model then be fitted? In the context of penalized regression, one approach would be to formulate the analogue of Kneib's expression (6) but with $l(\cdot)$ replaced by an



(a)

(b)

(c)

(d)

Figure 1: Munich rent data: Conditional densities (top), and localized mean–, median–, and modal smoothers (bottom), using horizontal bandwidths $h = 8$, and vertical bandwidths $b = 50$ (left) and $b = 150$ (right).

approximated version of the delta-function, say

$$l(\cdot) = -\frac{1}{a}K(\cdot/a),$$

where K is a kernel function and a a small "bandwidth". The penalized minimization problem corresponding to model (3) then takes the shape

$$-\sum_{i=1}^{n} K\left(\frac{y_i - \eta_i}{a}\right) + \lambda \text{pen}(f), \tag{4}$$

(where weights have been omitted given the considerations in the first part of this discussion, and the irrelevant constant $1/a$ has been omitted too). If K is chosen to be a Gaussian kernel function, then this fulfils the "twice continuously differentiable" condition mentioned in Section 3. Using an appropriate quadratic penalty pen(f), optimization problem (4) could then in principle be solved using penalized Fisher scoring, as well as functional gradient descent boosting, as discussed later in the same section of Kneib's paper.

Further comments. One issue that I would like to raise is the behavior of the discussed techniques for extreme choices of τ, such as $\tau \longrightarrow 1$ or $\tau \longrightarrow 0$. To motivate this point, observe that for the "100% expectile", with $\tau = 1$, one clearly has

$$\min_e \sum_{i=1}^{n} w_1(y_i, e)(y_i - e)^2 = \min_e \sum_{i=1}^{n} 1_{\{y_i > e\}}(y_i - e)^2 = 0, \tag{5}$$

which is solved for all $e \geq \max\{y_i, i = 1, \ldots, n\}$. Obviously, this is a conceptual characteristic of expectiles (as well as quantiles) and not an issue of *expectile regression* as such. While Schulze Waltrup et al. (2013) look into the question of extreme quantiles and expectiles from a theoretical point of view, I am interested in the more practical question of whether the techniques provided in this paper – and particularly the boosting approach – still work at or close to this limit, and whether they provide stable and meaningful results? More specifically, I wonder whether some additional structure or model assumptions are necessary to perform expectile regression at or close to this limit? In this context, I would like to point to a publication by Hall & van

Keilegom (2009), who consider nonparametric regression for data $Y = a(X) + \epsilon$ where errors are anchored at the "endpoints" of the error distribution, i.e. $P(\epsilon > 0) = 1$ (this corresponds to our $\tau = 0$) or $P(\epsilon > 0) = 0$ ($\tau = 1$). Estimation (in the first case) is carried out locally as the maximum intercept over all lines which lie underneath all data points in a window centered at the target point (with several additional technicalities needed to deal with existence and identifiability issues).

My next remark concerns the existence of features in expectile curves. Comparing the 90% quantile and expectile curves for the year of construction, which are provided in Kneib's paper in figures 3 (bottom right) and 5 (bottom right), respectively, we observe that the quantile curve gives a strong indication for a dip at about 1950 (which appears to be backed up by the pointwise credible intervals), while the expectile curve does not identify this dip clearly. This, of course, raises the question of "whether this dip is really there"? In the context of mean regression, this question has been discussed by Chauduri & Marron (1999), who argue that, for a dip being "really there", the derivative of the curve must make a "significant" crossing of 0 (from negative to positive, in this case). A certain dip is declared as significant if the simultaneous confidence band crosses, in its full width, the zero line at the location of the dip. This procedure is then repeated over all bandwidths, and the results are visualized in 2D maps, indicating, at a glance, whether there does exist any degree of resolution for which the dip is significant. Of course, the same can be done analogously to test for the existence of peaks.

The interesting question is then whether an "expectile SiZer" could be developed, where one tests for significant zero crossings of derivatives of expectile curves? Clearly, this would require the ability to estimate derivatives of expectile curves, as well as (simultaneous) confidence bands for these estimates. The former should be relatively straightforward under the propagated setup, since simply the derivative of the basis function expansion could be calculated. Confidence intervals for expectile regression are now well developed (Sobotka et al, 2013), which I would trust to be straightforwardly extendible to derivatives of expectiles. However, the extension from *pointwise* to the required *simultaneous* confidence bands (i.e., a band containing the entire expectile curve with some level of confidence) may still pose some challenges. A simulation–based approach to this problem, in the context of penalized spline (mean) regression, is provided in Ruppert, Wand, and Carroll (2003), pp 142ff.

Finally, I wish to thank the author for the healthy attitude of putting the estimation problem and the modelling aspects in the foreground, and considering Bayesian and frequentist estimation strategies with equal weight, and without philosophical burden, next to each other. This allows to focus attention on the actual inferential problem, and to get a comprehensive and unbiased view on the strengths and weaknesses of the individual methods. It would be desirable if this broader view on statistical inference were adapted by more authors in the future.

References

Bashtannyk, D.M. and Hyndman, R.J. (2001) Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis* **36**, 279–298.

Chauduri, P., and Marron, J. S. (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807 – 823.

Einbeck, J. and Tutz, G. (2006) Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society C* **55**, 461–475.

Fried, R., Einbeck, J., and Gather, U. (2007) Weighted Repeated Median Smoothing and Filtering. *Journal of the American Statistical Association* **102**, 1300–1308.

Hall, P., and van Keilegom, I. (2009) Nonparameetric "regression" when the errors are positioned at end–points. *Bernoulli* **15**, 614–633.

Matzner-Løber, E., Gannoun, A., and Gooijer, J.D.G. (1998) Nonparametric forecasting: a comparison of three kernel–based methods. *Communications in Statistics — Theory and Methods* **27**, 1593–1617.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003) *Semiparametric regression.* Cambridge University Press:New York.

Sobotka, F., Kauermann, G., Schulze Waltrup. L., and Kneib, T. (2013) On confidence intervals for semiparametric expectile regression. *Statistics and Computing* **23**, 135–148.

Taylor, J. (2012) Strategies for mean and modal multivariate local regression. PhD thesis, Durham University, 2012.

Taylor, J., and Einbeck, J. (2011): Multivariate regression smoothing through the 'fallling net' . In: Conesa et al. (Eds.): 26th International Workshop on Statistical Modelling, Valencia, 11-15 July 2011, pages 597-602.

Schulze Waltrup, L., Sobotka, F., Kneib, T., Kauermann, G. (2013) Quantile or expectile regression – is there a favorite? Technical Report.