On the use of double quantile regression and visual assessment to estimate performance constraints

We recently published evidence that male Adelaide's warblers' vocal performance improves over the course of the dawn chorus (Vazquez-Cardona et al., 2023). In that study, vocal performance was measured as the distance from an estimated performance limit in acoustic space (Logue, Sheppard, Walton, Brinkman, & Medina, 2020). We used mixed quantile regression (QR) to estimate performance limits (Logue & Bonnell, 2023; Wilson, Bitton, Podos, & Mennill, 2014). Visual inspection of scatterplots showed the QR lines approximated the slopes of the constrained edges of the distributions. We took this as evidence that the lines accurately represent how the two variables trade off at the limits of performance. A recent paper by Cardoso (2024) argues that there were two important flaws with our approach.

Cardoso's (2024) first critique is that our performance metrics ("deviation scores") incorrectly emphasize one of the two acoustic variables from which they are derived. He writes that deviation scores derived from QR lines tend to correlate strongly with one of the two acoustic variables that trade-off at the limit of performance. He gives the example of our deviation score *recovery time*, which is based on the trade-off between the duration of a note, and the duration of the subsequent silent gap. *Recovery time* correlated more strongly with gap duration ($r = 0.94$), than it did with note duration ($r = 0.24$), because the slope of the QR line was low ($\beta = 0.17$). According to Cardoso (2019, 2024), the assignment of variables to axes is largely arbitrary, and if the axes were switched, the deviation score would emphasize the other variable (note duration, in this case). Following this logic, *recovery time's* emphasis on gap duration is arbitrary and does not reflect the species' biology. Cardoso's preferred solution is to perform double quantile regression (DQR). To conduct a DQR, one first conducts a QR of *x* on *y*, then reverses the axes and runs a QR of *y* on *x*. If the edge of interest has a negative slope we can use the same tau for both QRs. If, however, it has a positive slope, we should use $_{\text{inverted}} = 1 – _{\text{original}}$ so that QR can estimate the constrained edge in both orientations (Cardoso, Abreu, Archer, Crottini, & Mota, 2020). The deviation scores from these two regression lines can then be averaged to generate deviation scores that are not biased by the assignment of variables to axes.

Cardoso (2024) is also concerned that uneven sampling is responsible, in part or in whole, for the appearance of performance constraints in our data plots. As evidence, he provides several plots of unconstrained, unevenly sampled simulated data (Fig. 2 in Cardoso 2024). These plots, he argues, spuriously indicate the existence of a constraint. The acoustic data from Vazquez-Cardona et al. (2023) are unevenly sampled (i.e., data are denser in some regions of the x-axis than others), so the appearance of constraints on our scatterplots may also be misleading. If that is true, then visual assessment would produce biased estimates of performance constraints and the deviation scores derived from them. Cardoso (2024) argues that visual assessment is therefore not appropriate for estimating the slopes of performance constraints. Instead, he writes, we should use DQR to estimate constraints.

What are we trying to do with quantile regression?

Before we address Cardoso's critique, it is worthwhile to explain what QR does and why we used it. Quantile regression produces a straight line that best estimates some quantile *y* (defined by) over the range of *x*. Students of performance constraints, however, do not want to estimate a particular quantile, *per se*. Rather, we want to estimate the slope of one edge of a two-dimensional distribution of data for which there is *a priori* evidence of constraint (Logue & Bonnell 2023). Cardoso (2024) seems to agree with this goal, as evidenced by his Fig. 1, in which two datapoints (× and □) that are equidistant from the edge of the distribution are said to have "similar performance." The edge of the distribution, however, is not necessarily well-described by QR (Logue et al. 2020, Vazquez-Cardona et al. 2023).

Consider two distributions that have the same, positively sloping, upper boundary. If the lower boundary of Distribution 1 is parallel to the upper boundary, but the lower boundary of Distribution 2 slopes in the opposite direction, QR analysis of the upper boundary will produce different slope estimates, even though the upper boundaries of the two distributions (i.e., the slopes we want to estimate) are identical. Although QR is an imperfect tool to estimate the slope of an edge, it can often do a passable job when applied to a large dataset and combined with visual inspection.

Our methods follow Vazquez-Cardona et al. (2023) except where stated. All analyses were conducted in R 4.3.2 (Team, 2013) running in the R Studio environment (RStudio Team, 2020). We used *lqmm* (Geraci, 2014) for mixed quantile regression, *quantreg* (Koenker et al., 2018) for quantile regression without random terms, *patchwork* (Pedersen, 2019) for composite graphs, and *tidyverse* (Wickham et al., 2019) for data manipulation and visualization.

One way to test Cardoso's (2019) assertion that DQR produces a less-biased estimate than QR would be to compare the edge of the distribution to the slope estimates produced by QR and DQR. The version of double quantile regression described in Cardoso (2019), however, does not generate a single consensus slope estimate. Rather, it generates two slope estimates based on QR, calculates deviation scores from each one, and averages them. We hypothesized that a line that bisects the angles created by the two QR lines would describe a consensus DQR line whose deviation scores would be equivalent to the average scores from the constituent QRs. To test this hypothesis, we compared the average deviation scores of the two DQR lines (the output of a DQR), to the deviation scores from one of the two lines that bisects the DQR lines. We found that the deviation scores from the bisecting line scaled linearly with the average of the deviation scores from the two QR lines (Fig. 1). Thus, the bisecting line characterizes how DQR sees the edge of a data distribution and we can use its slope to help determine whether DQR is appropriate for a given dataset. Typically, researchers will not know the true slope of the edge, so they will not have a "gold standard" against which to compare the estimated slopes (but see analysis of simulated data below). Visual inspection of a scatterplot, however, can tell us whether the estimate approximates the edge.

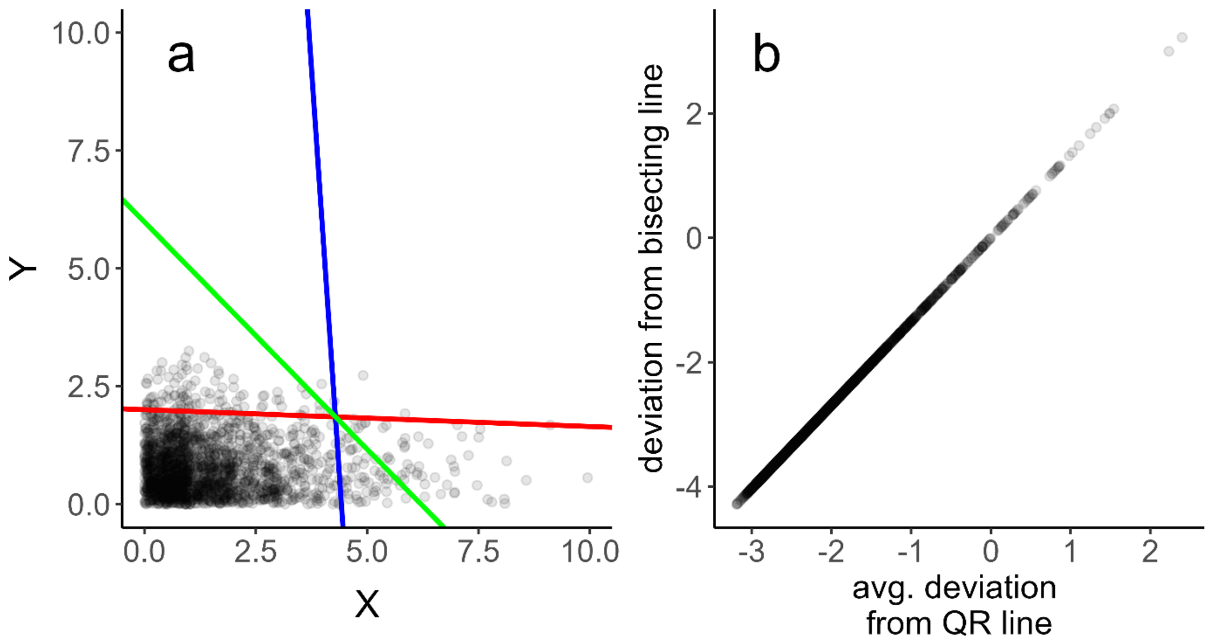

Fig. 1. (a) A scatterplot of simulated, unevenly sampled, unconstrained data from Cardoso (2024) showing the original (red) and inverted (blue) quantile regression lines, and their bisector (green). (b) Plot comparing the average of deviations from the two QR lines to the deviations from the bisecting line.

We conducted DQR on the acoustic data from Vazquez-Cardona et al. (2023) to see whether the bisecting lines estimated the slope of the distributions' constrained edges more accurately than the lines from our original QR analysis. Based on our visual comparison of the slope of the bisecting line and the slope of the constrained edge of the distribution, DQR estimates were as good as or better than estimates based on QR (Fig. 2). For two of the comparisons, however, the QR lines derived from the inverted data (blue lines in Figs. 2b and 2c) appeared to have radically different slopes than the edges of their respective distributions. We will return to this finding later.

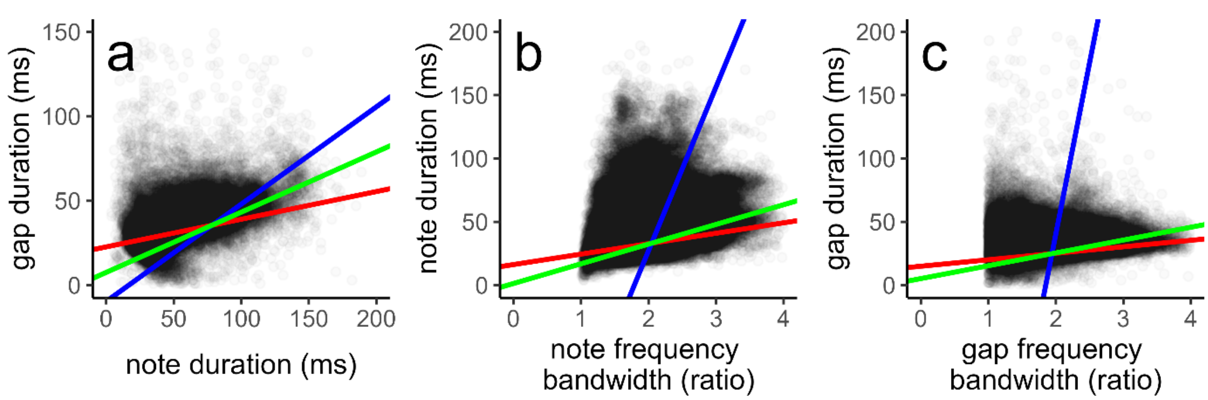

Figure 2. Scatterplots comparing pairs of acoustic variables from the notes of male Adelaide's warbler songs. In each case, the lower boundary of the distribution is hypothesized to be constrained. Lines show the results of quantile regression (= 0.05) in the original orientation

(red) and the inverted orientation (blue), as well as their bisecting line (green). Axes have been cropped to aid visualization, so some extreme outliers are missing. The bisecting lines do not appear to bisect the angles on these graphs because the axes have different scales.

Is it problematic when QR lines emphasize one variable?

We now address Cardoso's (2024) first argument, which is that single QR arbitrarily emphasizes one of the two variables that trade off at the limit of performance. Visual assessment indicates that the constrained edges of our data distributions form a shallow angle with the x-axis. We therefore expect the deviation scores to be influenced by (and thus correlate with) the *y*-variable more than the *x*-variable. This makes sense biologically. Consider the performance metric *unvoiced FM* which is a deviation score derived from the trade-off between the frequency bandwidth and duration of the silent gaps between notes. The data indicate that, at the limit of performance, an increase of 1 SD in frequency bandwidth requires an increase in gap duration of < 1 SD. We see no biological reason to expect that constraints should impose one-to-one trade-offs in general.

Cardoso (2024, 2019) writes that the assignment of variables to the *x*- and *y*-axes is largely arbitrary. For some distributions, however, only one orientation is amenable to QR. For QR to accurately estimate the slope of the constrained edge of a distribution, the constrained edge must comprise the entire top or bottom of the distribution. For some distributions, this only occurs in one of the two possible orientations. Our *recovery time* data is amenable to lower boundary (low) QR in the original orientation and upper boundary (high) QR in the inverted orientation (Fig. 3a, 3a'). Both orientations permit QR to approximate the slope of the constrained edge, and the bisecting line fits the edge better than either of the single QR lines (Fig. 2a). However, the same cannot be said of our other two comparisons. Our *unvoiced FM* data permits lower boundary QR when gap frequency BW is assigned to the x-axis and gap duration is assigned to the y-axis (Fig. 3c). When we inverted the axes, however, the constrained edge comprised neither the top nor the bottom of the distribution (Fig. 3c'). QR on the inverted data produces a line that does not approximate the slope of the constrained edge. The same is true for the *voiced FM* data. Cardoso's (2024) suggestion to transform the data (e.g., by multiplying by -1) does not eliminate this problem.

We can generalize from these examples to predict whether inversion will expose the constrained edge to QR. If the facet of the distribution opposite the constrained edge has the opposite slope as the constrained edge (e.g., the constrained edge is positively sloped and the opposite edge is negatively sloped), a QR on the inverted distribution will "see" both the constrained edge and the opposite edge, and so return a line that does not accurately characterize the slope of the constrained edge (Figs. 2,3).

We tested the performance of QR and DQR on distributions in which the edge of interest and the opposite edge have opposite slopes. We simulated two-dimensional data with opposite-sloping constraints on the lower (intercept = 0, slope = 0.25) and upper (intercept = 8, slope = -0.5) facets of the distribution. We then conducted QR and DQR on the lower boundary (= 0.05). The QR analysis returned an estimated slope of 0.15, which was 0.1 less than the true slope. The DQR analysis, however, estimated the slope as -0.2, which was 0.45 less than the true slope. When tasked with estimating the slope of the lower boundary in a simulated dataset with opposite sloped boundaries on the upper and lower facets neither estimate was very accurate, but QR was much more accurate than DQR.

Uneven sampling and the illusion of constraint

Up to this point, we have assumed that visual inspection of the scatterplot is an accurate way to estimate the slope of the constrained edge. Cardoso (2024), however, argues that visual inspection is a poor way to estimate the slope of a distribution's edge because uneven sampling alone can produce the illusion of constraint. In support of this argument, Cardoso presents unevenly sampled distributions of simulated data that are unconstrained but nevertheless appear to have sloped edges (e.g., Fig. 2a in Cardoso 2024). He points out that the data in Vazquez-Cardona (2023) are also unevenly sampled, so the appearance of constraint may be an artifact of uneven sampling. We tested this hypothesis.

If the sloping lower boundary of a distribution is an artifact of uneven sampling, it should disappear when sampling is even. Rarefication ensures even sampling by drawing a fixed number of samples from each of several equal-sized bins. We rarefied our data, leaving 1000 samples in each bin (*recovery time*: bin size = 5 ms; *voiced FM* and *unvoiced FM*: bin size = 0.2). Extreme values of *x* did not include enough samples in each bin, so we truncated the data. We then compared plots of the rarefied data to plots of the whole dataset (Fig. 4). Rarefication had little effect on the visibility or slope of the lower boundaries of the scatterplots.

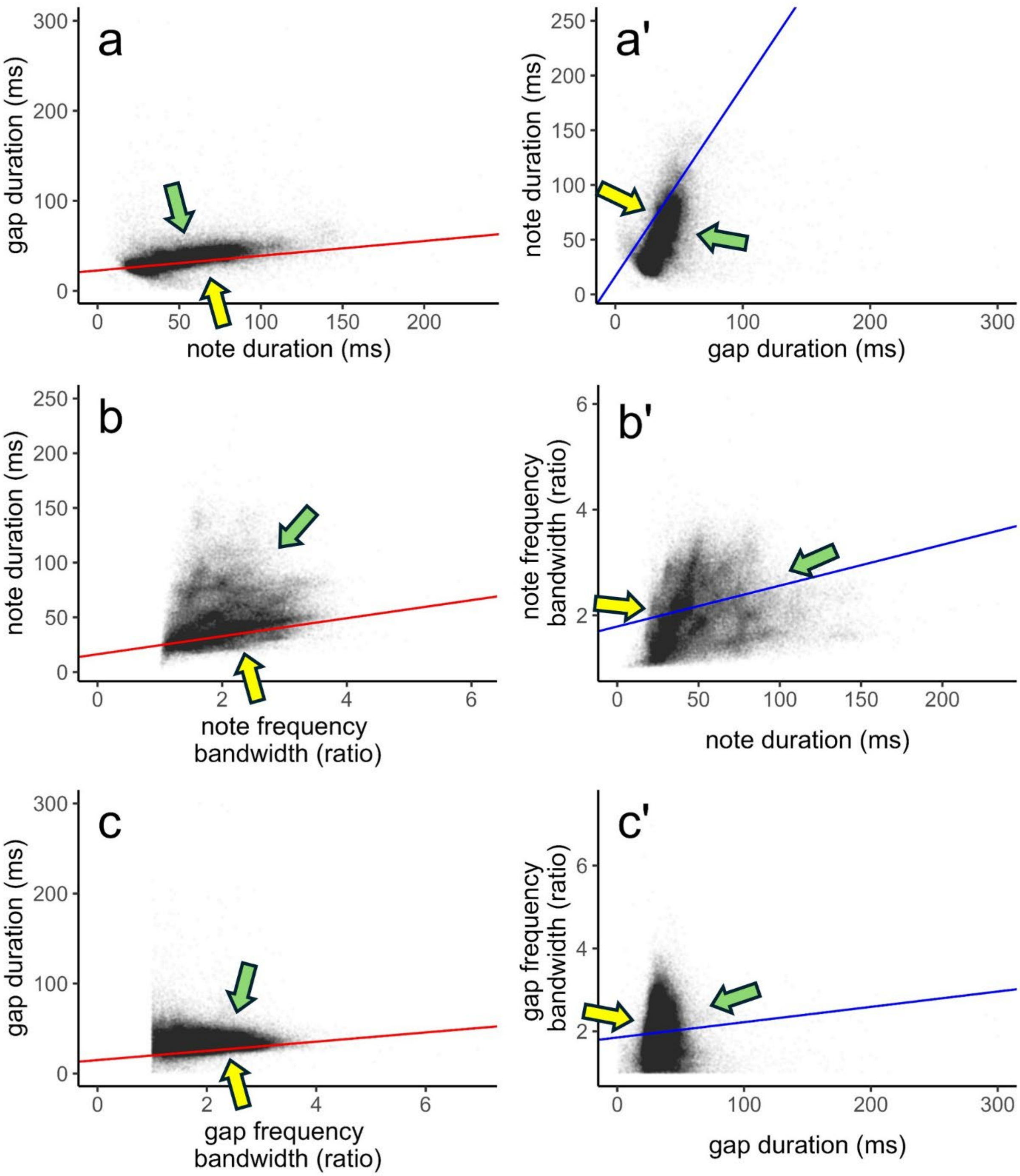

Figure 3. Scatterplots of note-level acoustic data from male Adelaide's warblers. (a-c) Scatterplots in the orientation they were analyzed for previous studies (Logue et al. 2020, Vazquez-Cardona et al. 2023). (a'-c') Scatterplots with the axes reversed relative to previous studies. Lines are from quantile regression with = 0.05 (red lines) or = 0.95 (blue lines). Arrows indicate the constrained edge of the distribution (yellow) and the opposite edge (green). Points are smaller and lighter than in Figure 2.

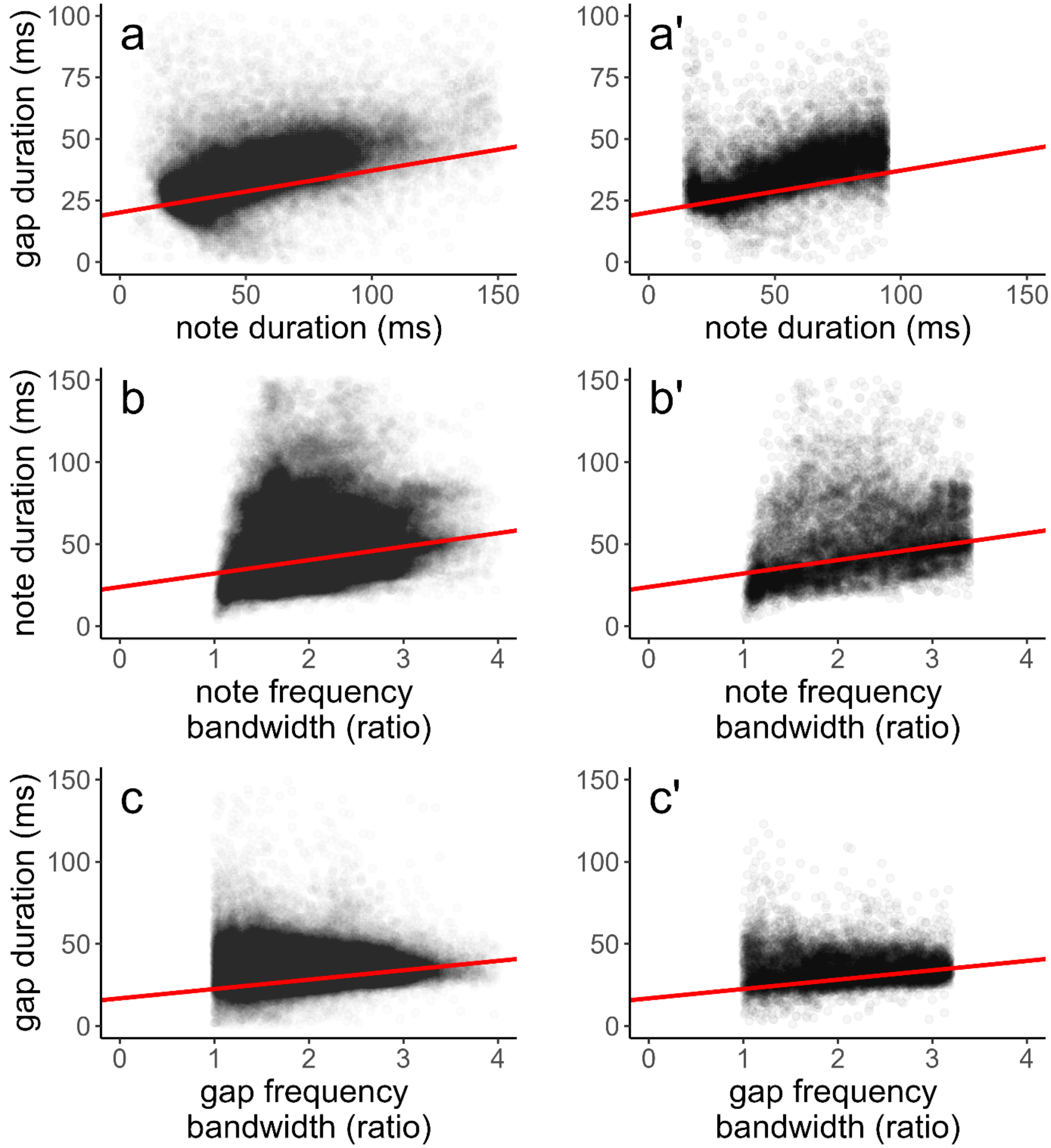

Figure 4. Scatterplots of note-level acoustic variables from male Adelaide's warblers. Plots a-c are based on the complete dataset, whereas plots a' – c' are based on rarefied data. Points are semi-transparent to show density (a-c: alpha = 0.01; a' – c': alpha = 0.03). Red lines are the quantile regression lines used to estimate the edge of the distribution in Vazquez-Cardona et al. (2023).

We noticed an important difference between the scatterplots of our bird song data and the plots of simulated data in Cardoso (2024). The bird song plots show distributions with sharply defined lower edges, whereas the simulated distributions do not show edges at all, but rather smooth density gradients (compare our Fig. 4a-4c with Fig. 2 from Cardoso 2024). Our group has previously shown that constrained edges are characterized by steep clines, not diffuse gradients (Logue et al. 2020, Logue and Bonnell 2023). We conclude that motor constraints generate sloped edges on trade-off plots, but uneven sampling generates sloping gradients. This difference can be detected by visual inspection of scatterplots (Logue & Bonnell 2023).

We evaluated QR and DQR's performance on unconstrained data (Fig. 1a). When we applied single QR to the unconstrained, unevenly sampled, simulated data that is the basis of Cardoso's (2024) Fig. 2a, we found little evidence of a sloped edge, regardless of the orientation of the data ($slope_{original}$ = - 0.04, $slope_{inverted}$= - 0.07; red and blue lines in Fig. 1a). In contrast, the bisecting line that shows how DQR characterizes the edge of the distribution is strongly sloped ($slope_{bisecting}$ = - 0.96; green line in Fig. 1a). In this case, single QR correctly estimates that there is not a sloped edge, but DQR treats an unbounded distribution as if it were bounded. When using DQR, one should be careful not to identify boundaries where none exist. Another potential shortcoming of DQR is that, unlike single QR, it does not provide any measure of confidence (e.g., confidence intervals or a p-value) that the edge of the slope is not flat.

Finally, we tested Cardoso's (2024) hypotheses that the effect of warming up on *recovery time* reported in Vazquez-Cardona et al. (2023) was exaggerated due to uneven sampling, and that substituting DQR for single QR would weaken or eliminate the pattern. We also tested the uneven sampling hypothesis by running the same Bayesian mixed model from Vazquez-Cardona et al. (2023) using *recovery time* deviation scores based on the QR line from the rarefied data. If uneven sampling exaggerated the warm-up effect by biasing the estimate of the trade-off between gap duration and note duration, we should see a reduced effect when deviation scores are measured from a QR line derived from rarefied data. This was not the case. The slope of the QR line based on the rarefied dataset was identical to the slope based on the full dataset (β = 0.17). When we ran the deviation scores through the Bayesian mixed model from Vazquez-Cardona et al. (2023), we found a nearly identical pattern to the original: Performance was lowest first thing in the morning, improved rapidly during the dawn chorus, and decreased gradually thereafter (Fig. 5).

We tested whether DQR would eliminate or weaken the observed pattern by using DQR to calculate *recovery time* and running the results through our model. We still saw the same basic pattern: performance was low when birds began singing, improved rapidly during the dawn chorus, and gradually diminished thereafter. We quantified the "warm-up effect" as the estimated improvement in performance from 2460 seconds before sunrise to the time of sunrise. The estimated warm-up effect when we used QR to measure recovery time (1.2 SD) was slightly greater than the one based on DQR (0.9 SD). The decrease in performance between sunrise and 12,000 seconds after sunrise was greater for the estimates based on DQR (1.4 SD) than for those based on QR (0.9 SD). Thus, the performance estimates based on DQR indicated a slightly weaker warm-up effect, and a slightly stronger cool-down effect. Notably, however, both effects are well within the credible intervals estimated by the QR-based model. We ran four iterations of the DQR model with shuffled data. None of these indicated a warm-up effect (average warmup = -0.1 SD; Fig. 5), indicating that the estimated warm-up effect in the DQR-based model exceeds null expectations.

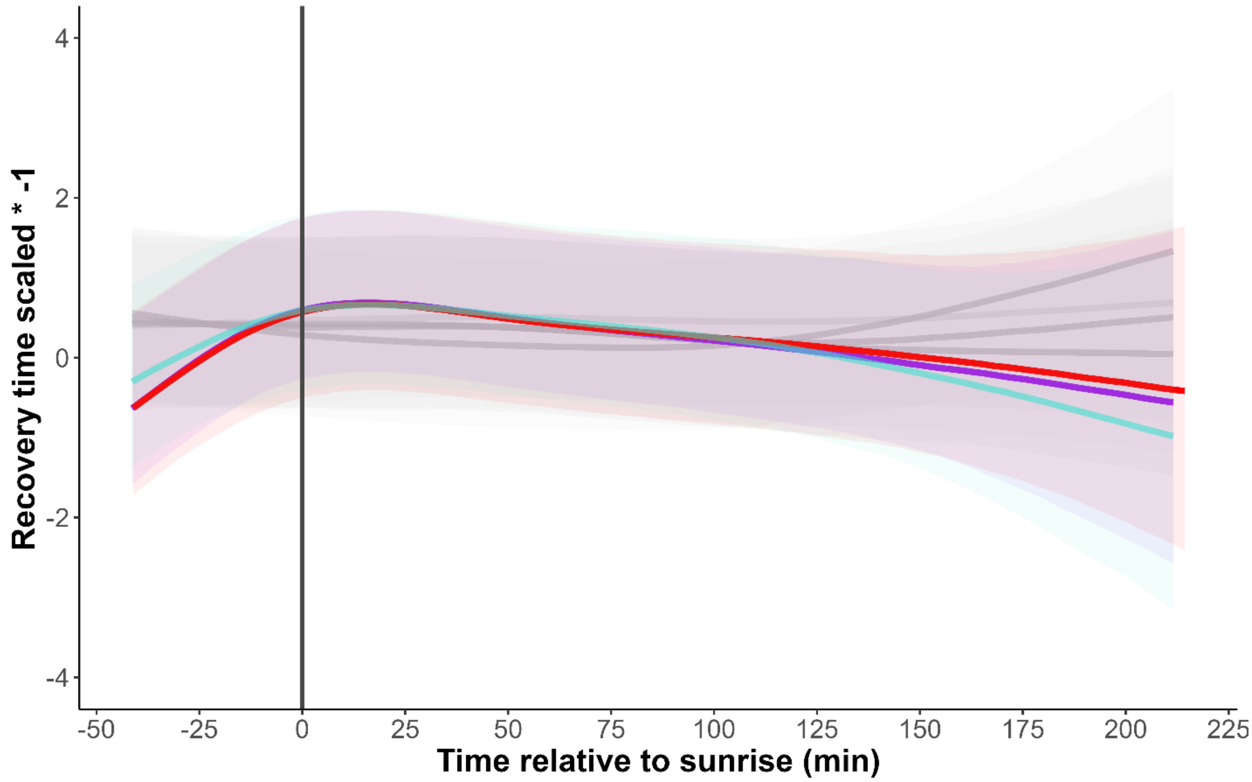

Figure 5. A plot showing the estimated conditional effect of time on the performance metric mean *recovery time. Recovery time* was measured with quantile regression applied to the full dataset (red line, as in Vazquez-Cardona et al. 2023), quantile regression applied to rarefied data (purple line), and double quantile regression applied to the full dataset (turquoise line). Gray lines use the same data as the turquoise line, but data were shuffled prior to analysis. Pale bands represent 95% credible intervals.

Summary and Discussion

We conducted a series of analyses to address Cardoso's (2024) critique of Vazquez-Cardona et al. (2023). We found mixed evidence for Cardoso's (2024) assertions that the deviation scores in Vazquez-Cardona et al. (2023) incorrectly emphasize one of the two acoustic variables from which they are derived, and that DQR should be used to reduce this bias. While we reject his argument that asymmetric correlations between deviation scores and the variables that comprise them are evidence of bias, our analyses indicate that DQR can improve the fit of the regression line to the edge for some distributions (Fig. 2).

We did not find support for Cardoso's (2023) hypothesis that uneven sampling is responsible for the appearance of performance constraints in our data. Rarefication had little effect on the constrained edge of the data distribution, demonstrating that uneven sampling is not responsible for the appearance of constraint (Fig. 4). Further, basing the QR line on rarefied data had almost no effect on the performance metric *recovery time's* change over time (Fig. 5). Cardoso (2024, p. 3) hypothesized that due to uneven sampling, "the real effect [of warm-up on recovery time] is likely weaker or maybe non-existent," compared to what was indicated in Vazquez-Cardona et al. (2023). We found that DQR does provide a better estimate of the distribution's edge in the *recovery time* comparison, but the improvement does not appear to be attributable to uneven sampling. While the warm-up effect was slightly weaker when we used DQR, the overall pattern of change over time remained the same, confirming our original conclusion that birds warm up their voices during the dawn chorus.

While uneven sampling can bias the slope of a QR line, it does not mimic the effects of constraint. Uneven sampling in the absence of constraint produces a smooth density gradient, whereas constraint produces a sharper cline in density (Logue & Bonnell 2023). Given a sufficiently large sample size, visual assessment can reveal this difference.

We recommend using the bisecting line to characterize the fit of a DQR analysis. Previously, we attempted to characterize how DQR sees the constrained edge by averaging the slopes and intercepts of the QR lines from the original orientation and the inverted orientation (electronic supplemental material for Vazquez-Cardona et al. 2023). Cardoso (2024) was skeptical that the average line characterized DQR, and indeed it does not. One of the two bisecting lines, however, does. Bisecting lines that parallel the constrained edge of the distribution indicate good fit.

When and how, then, should students of performance constraints use DQR? We found that applying QR to inverted distributions can produce slope estimates that do not approximate the constrained edge when the constrained edge slopes in the opposite direction of the opposite edge. Even in these cases, however, DQR can sometimes improve slope estimates (Fig. 2). Our analysis of our own synthesized data, and that of Cardoso, showed that DQR can also produce spurious slope estimates that are inferior to those from QR. We therefore advise caution with DQR when the constrained edge of a distribution slopes away from the opposite edge (Fig. 3).

DQR appears to be a useful approach for estimating the slope of a constrained edge when the opposite facet of the distribution does not slope away from the constrained edge (Cardoso 2019, 2024). However, DQR can return strong slopes when there are no constraints and it does not give an estimate of uncertainty, so it should be used in combination with other methods, like visual inspection of scatterplots, rarefaction, single QR, and analysis of skew (Logue & Bonnell 2023). All these methods are more effective when samples are large. Looking to the future, the field would benefit from the development of tools that characterize the edge of a distribution directly, rather than methods like QR and DQR that estimate quantiles.

Works Cited

Cardoso, G. C. (2019). Double quantile regression accurately assesses distance to boundary trade-offs. *Methods in Ecology and Evolution, 10*(8), 1322-1331.

Cardoso, G. C. (2024). Warm-up and metrics of song performance: a commentary on Vazquez-Cardona et al. *Behavioral Ecology, 35*(3), arae015.

Cardoso, G. C., Abreu, J. M., Archer, J., Crottini, A., & Mota, P. G. (2020). Independent evolution of song diversity and song motor performance in canaries, goldfinches and allies indicates clade-specific trade-offs in birdsong. *Evolution, 74*(6), 1170-1185.

Geraci, M. (2014). Linear quantile mixed models: the lqmm package for Laplace quantile regression. *Journal of Statistical Software, 57*(13), 1-29.

Logue, D. M., & Bonnell, T. R. (2023). Skewed performance distributions as evidence of motor constraint in sports and animal displays. *Royal Society Open Science, 10*(11), 230692.

Logue, D. M., Sheppard, J. A., Walton, B., Brinkman, B. E., & Medina, O. J. (2020). An analysis of avian vocal performance at the note and song levels. *Bioacoustics, 29*(6), 709-730.

Pedersen, T. L. (2019). Package 'patchwork'. *R package http://CRAN. R-project. org/package= patchwork. Cran*.

RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA Retrieved from URL http://www.rstudio.com/

Team, R. C. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. *(No Title)*.

Vazquez-Cardona, J., Bonnell, T. R., Mower, P. C., Medina, O. J., Jiskoot, H., & Logue, D. M. (2023). Vocal performance increases rapidly during the dawn chorus in Adelaide's warbler (Setophaga adelaidae). *Behavioral Ecology*, arad030.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . Hester, J. (2019). Welcome to the Tidyverse. *Journal of open source software, 4*(43), 1686.

Wilson, D. R., Bitton, P. P., Podos, J., & Mennill, D. J. (2014). Uneven sampling and the analysis of vocal performance constraints. *Am Nat, 183*(2), 214-228. doi:10.1086/674379