J Exp Biol Advance Online Articles. First posted online on 3 January 2017 as doi:10.1242/jeb.148775
Access the most recent version at http://jeb.biologists.org/lookup/doi/10.1242/jeb.148775

Estimating monotonic rates from biological data using local linear regression

Colin Olito*, Craig R. White, Dustin J. Marshall, Diego R. Barneche

Centre for Geometric Biology, School of Biological Sciences, Monash University, Victoria 3800, Australia.

*Corresponding author e-mail: colin.olito@gmail.com

Keywords: Biological rates, time series, local linear regression, autocorrelation, linearity, reproducible research

Abstract

Accessing many fundamental questions in biology begins with empirical estimation of simple monotonic rates of underlying biological processes. Across a variety of disciplines, ranging from physiology to biogeochemistry, these rates are routinely estimated from non-linear and noisy time series data using linear regression and *ad hoc* manual truncation of non-linearities. Here, we introduce the R package `LoLinR`, a flexible toolkit to implement local linear regression techniques to objectively and reproducibly estimate monotonic biological rates from non-linear time series data, and demonstrate possible applications using metabolic rate data. `LoLinR` provides methods to easily and reliably estimate monotonic rates from time series data in a way that is statistically robust, facilitates reproducible research, and is applicable to a wide variety of research disciplines in the biological sciences.

Introduction

Living organisms, the communities they form, and the environments they inhabit, are all temporally dynamic systems. Consequently, many fundamental questions in biology begin with estimating the rates of underlying biological processes. When nonlinearity is of biological interest and accurately represents the rate of interest, then nonlinear, function valued approaches may be most appropriate (Marshall et al., 2013; Stinchcombe et al., 2012). Non-linear approaches can be further generalized using function regression if the question also requires accounting for change in parameters as a function of other variables (Yen et al., 2015). However, in many cases, there is a putatively linear rate that we wish to estimate free of artifactual nonlinear regions (e.g., Fig. 1). In these cases, researchers often reduce experimental complexity by holding state variables constant in order to estimate monotonic rates, which are then used in subsequent analyses. For example, the Metabolic Theory of Ecology (MTE) seeks to explain ecological patterns by scaling the size- and temperature-dependence of metabolic rates from individuals to ecosystems (Brown et al., 2004; Gillooly et al., 2001). The study of MTE therefore begins with estimates of metabolic rate at standardized temperatures (e.g. Barneche et al., 2014; Brown et al., 2004; Gillooly et al., 2001; White et al., 2011). Similarly, predicting the impacts of climate change on ecosystem processes, such as community primary productivity, begins with estimates of O_2 production rates (Tanaka et al., 2013; Yvon-Durocher et al., 2015). Other examples include estimates of leaf respiration rate (Shapiro et al., 2004), ecosystem functioning (Ross et al., 2013), and components of biogeochemical cycles including denitrification (Song et al., 2011), CO_2, and CH_4 gas emissions (e.g. Larmola et al., 2013). Thus, accurate estimates of biological rates provide the foundation for many branches of biology but surprisingly, there are few systematic approaches to estimating monotonic rates from biological data.

Biological rates are routinely estimated from non-linear or noisy time series using linear regression. For example, many studies in physiology, ecosystem ecology, and biogeochemistry monitor reactant consumption in closed chambers at standardized temperatures, and use the resulting, often non-linear, time series to estimate the rate of interest (e.g. Larmola et al., 2013; Ross et al., 2013; Song et al., 2011; Tanaka et al., 2013; White et al., 2011; Yvon-Durocher et al., 2015). There are several problems with this approach. First, the data used rarely meet the criteria for linear regression to be an appropriate analytic tool. The first measures in a time series are often noisy as equipment/samples/organisms equilibrate after setup, while at the end of the time series, rates can change because of saturation effects or the exhaustion of a limiting resource (e.g., Fig. 1). Consequently, naive linear regression of a full time series can conflate the biological rate of interest with undesired effects. Second, common *ad hoc* methods to ameliorate this problem, such as manually truncating non-linear portions of the time series, introduce subjectivity into the analysis, and may reduce statistical power by removing useful data. Last, published studies rarely provide both the raw data and the specific methods necessary to reproduce reported results. This makes it difficult or impossible to evaluate the appropriateness of the methods, and is particularly problematic in a new era that demands scientific transparency and reproducibility (Fang et al., 2012; Grieneisen & Zhang, 2012). The need to estimate monotonic rates from time series data will only increase as technological advances continue to make collection of high-resolution data easy and cost effective. This presents biologists with a non-trivial challenge: to reliably estimate biological rates in a way that is statistically robust, and fully reproducible.

Here, we introduce the `LoLinR` package for R (R Development Core Team 2016), which provides a suite of simple functions to implement local linear regressions to estimate monotonic rates from time series data. We describe the general approach to reproducible and statistically robust estimation of monotonic rates, and the specific methods used in the package. We then walk through two example analyses to illustrate the utility of the package, as well as important analytic considerations and pitfalls (additional examples are available through **`https://github.com/colin-olito/LoLinR`**).

Materials and Methods

We first describe the base function around which the `LoLinR` package is built, then detail the component linearity metrics underpinning the function. The methods provided in the `LoLinR` package are a modification of traditional Loess regression built around the wrapper function `rankLocReg`. While Loess techniques are designed primarily for data interpolation and visualization, `rankLocReg` stores relevant data from all possible local regressions of ordered subsets (adjacent points) of a full time series, given a minimum window size. The function returns an object of class `rankLocReg`, which includes a ranked list of the most linear subsets of the data, as well as corresponding data for each local regression. A call to `rankLocReg` implements three steps: 1) define the minimum window size, 2) fit local regressions, and 3) rank local regressions.

The only user-defined constraint imposed on the analysis is `alpha`, which defines the minimum window size used to fit local regressions, expressed as a proportion of the total number of observations in the full data set (analogous to a traditional Loess smoothing parameter). At a minimum, `alpha` must take into account the total number of observations in the full data set, *N*, such that $(\mathtt{alpha} \times N) \geq 15$ (Harrell, 2001). Ideally, $(\mathtt{alpha} \times N)$ should also represent a biologically meaningful interval for the given data set. A call to `rankLocReg` fits all possible local regressions with $n \geq (\mathtt{alpha} * N)$ adjacent observations using Ordinary Least Squares.

To quantify linearity for each of the local regressions, we define the combined linearity metric *L*, which represents a weighted sum of three component metrics. The first metric is the skewness of the standardized residuals for the local regression, estimated as the Fisher-Pearson Standardized Third Moment Coefficient

$$S = \frac{n}{(n-1)*(n-2)} \times \sum \left[\left(\frac{x - \bar{x}}{\sigma(x)} \right)^3 \right], \quad (1)$$

where $\sigma(x)$ is the sample standard deviation of *x*. The second metric is the range of the 95% confidence interval for the slope of the local regression β_1,

$$C.I.range = \left(\beta_1 + t^*_{0.975} \times \frac{\sigma}{\sqrt{n}} \right) - \left(\beta_1 - t^*_{0.025} \times \frac{\sigma}{\sqrt{n}} \right), \quad (2)$$

where σ is the sample standard deviation, and *n* is the number of observations used in the local regression. The third and final metric is a modified Breusch-Godfrey statistic

$$R^2_{BG} = \frac{nR^2}{n}, \quad (3)$$

for serial correlation of the standardized residuals of the local regression up to order $(n-k-1)$ (where *k* is the number of parameters in the fitted model -- usually 2). We divide the traditional nR^2 Breusch-Godfrey statistic by *n* to remove the multiplicative effect of sample size. We do this because we wish to compare the variance explained by local regressions with different sample sizes, rather than perform a significance test for an asymptotically $\chi^2_{d.f.=n}$ distributed variable with fixed sample size *n*. It is also possible to account for autocorrelation using generalized linear models with a specified correlation stucture. However, the R^2_{BG} metric accounts for serial correlation up to the maximum lag of $(n-k-1)$ inclusive, and does not require additional assumptions made by alternative correlation structures. Each of the three component metrics, *x*, are **Z** standardized against the minimum value (or minimum absolute value for *S*) obtained from all *i* fitted local regressions as

$$\mathbf{Z}_{min}[x_i] = \frac{x_i - min(x)}{\sigma(x)}, \quad \mathbf{Z}_{min(abs)}[x_i] = \frac{x_i - min(abs(x))}{\sigma(x)}, \quad (4)$$

where σ is the sample standard deviation, ensuring that all component metrics have a common scale, and smaller values of each correspond to greater linearity of the associated local regression. Thus, the combined linearity metric without any further weighting of the component metrics is defined as

$$L_Z = \mathbf{Z}_{min(abs)}[S] + \mathbf{Z}_{min}[C.I.range] + \mathbf{Z}_{min}[R^2_{BG}]. \tag{5}$$

L_Z implicitly weights the contributions of each component metric by the relative magnitudes of their empirical variances for a given data set. For the many cases where the empirical distributions of the component metrics differ, we define and strongly recommend two alternative weighting methods: L_{eq} and $L_{\%}$. L_{eq} enforces equal weights by dividing the $\mathbf{z}_{min}$ scores for each metric by their maximum value before summing. $L_{\%}$ sums the percentile-ranks of the $\mathbf{z}_{min}$ scores for each component metric. The choice of weighting method will ultimately depend on the specific characteristics of each data set, the `alpha` value used for the `rankLocReg` analysis, and the biology of the system being studied.

When used with `rankLocReg` objects, the `plot` function generates several diagnostic plots to help determine the most appropriate method for a given analysis. Users can examine results from alternative *L* metrics by using the `reRank` function. Fig. 1 provides a schematic overview of a typical workflow using `LoLinR` to estimate biological rates. Crucially, analyses using `LoLinR` can be fully reproduced from (1) the time series data and (2) any one of the following: summary `plots`, `summary` tables, or the R code used to perform the analysis. All are easily included as appendices or supplementary material to published articles, making `LoLinR` analyses extremely easy to reproduce.

Results and Discussion

Larval Metabolic Rate

The first example is from a study allometric scaling of metabolic rate during larval development in two bryozoan species (*Bugula neritina* and *Watersipora subquortata*; Pettersen et al., 2015). Metabolic rate was estimated for individual larva from O_2 saturation time series collected using closed chambers. Fig. 2A provides an example of the analytical challenge presented by these data. The full time series is clearly non-linear: the rate of O_2 consumption initially decelerates as the chamber and larva equilibrate after handling. There is also a subtle acceleration towards the end of the time series, probably resulting from a physiological response by the larva to declining O_2 availability (Lagos et al., 2015), or accumulation of bacterial biofilm that began to consume oxygen. Any estimate of O_2 consumption rate including these non-linearities would be conflated with these other processes. However, truncating the data to exclude these non-linearities is subjective and difficult, especially for the subtle curvature towards the end of the data set. Ultimately, we wish to identify the region where the relationship between O_2 concentration and time is most linear, and estimate its slope.

Using the `LoLinR` package, we can analyze this data set with the call

```
library(LoLinR)
data(BugulaData)
BugulaRegs <- rankLocReg(xall=BugulaData$Time.s,
yall=BugulaData$D1, alpha=0.2, method="eq").
```

which implements the `rankLocReg` function with a minimum window size of `alpha` = 0.2, and uses the linearity metric L_{eq} to rank local regressions. This `alpha` value results in a minimum window size of $(\texttt{alpha} \times N) = 22$ for this data set. This call returns an object of class `rankLocReg` which includes a ranked list of all possible local regressions, the number of local regressions fitted, and several summary statistics. Examination of the `summary` output and the distribution of local regression slopes (Fig. 2D, density plot) indicate that both the L_Z and L_{eq} weighting methods return the same rank 1 local regression, while the $L_{\%}$ method returns a slightly different result. However, all three methods identify local regressions in the later half of the time series, where the rate of O_2 consumption has stabilized. The $L_{\%}$ rank 1 local regression includes a larger subset of the data (*n*=44 observations) than the other two methods (*n*=26), and all three rank 1 local regressions have nearly identical slopes (L_Z, L_{eq}: $\beta_1 = -0.00133$; $L_{\%}$: $\beta_1 = -0.00132$). Given the similarity of the results, we would recommend using the $L_{\%}$ method in this case, because it provides greater statistical power for the estimation of β_1, the parameter of interest. Inspection of the chosen local regression and accompanying residual plots also suggests that other than some autocorrelation, which is expected in time series data, there are no other major concerns (Fig. 2C-F).

A comparison of this result with common alternative approaches highlights the usefulness of the methods. Naive linear regression of the full time series yields an estimate of $\beta_1 = -0.00119$, indicating that non-linearities, particularly early in the time series, result in under-estimation of the metabolic rate. Estimates obtained by linear regression of manually truncated subsections of this data (with the same window size of *n*=22 observations) can range from $\beta_1 = -0.00212$ to $\beta_1 = -0.00067$, more than a threefold difference in the estimate of metabolic rate. In addition to being methodologically opaque and conflating the desired rate with other experimental and biological processes, these common *ad hoc* methods can give highly inaccurate estimates.

Flow-through respirometry

The second example is a study of the metabolic costs of living in the Arctic for Great Cormorants (*Phalacrocorax carbo*) (White et al., 2011). In this study, metabolic rate was estimated for individual birds using flow-through respirometry protocols (White et al., 2011; *Supplementary Material*). These techniques generate time series of the rate of O_2 consumption ($\dot{V}O_2$, mL O_2kg^{-1} min^{-1}) rather than O_2 saturation or concentration. For this data, the analytic goal was to estimate resting metabolic rate, which should correspond to the subset of the time series where $\dot{V}O_2$ is lowest and most linear. Conventional methods for estimating $\dot{V}O_2$ from flow-through respirometry data are based on analysis of the distribution of sequences of adjacent data points, and the minimum running average of subsets of adjacent points with varying numbers of included observations (e.g. Withers, 2001). Here, we illustrate how `rankLocReg` can be used to estimate resting metabolic rate from these data by leveraging the statistical framework of local linear regression and examining the distribution of standardized residuals.

We analyze a representative $\dot{V}O_2$ time series for an individual Cormorant. The time series is non-linear with large spikes occurring when the animal is physically active inside the chamber, but there appears to be a region of relative stability between 2.5–5.25 hours (Fig 3). We analyze the thinned data with a call to `rankLocReg` using `alpha=0.1`. This ensures that the minimum window size corresponds to approximately 30 min, and a minimum of 15 observations for the local regressions. Although we are not necessarily interested in the slopes of the local regressions (β_1), an examination of the distribution of β_1 highlights the fact that the $L_{\%}$ metric returns a different rank 1 local regression than the other two L metrics (Fig 3A). For these data, the L_Z and L_{eq} metrics misidentify the most stable subset of these data (a consequence of strongly skewed empirical distribution of R^2_{BG}) (Fig 3B, Fig A1). However,

$L_{\%}$ identifies a period of approximately 2 hours where $\dot{V}O_2$ is most stable (Fig 3C). The average (or median) $\dot{V}O_2$ during this period is easily recovered using the `summary` for this analysis, and returns an estimate of $\dot{V}O_2 = 36.45\,ml\,m^{-1}$. This estimate is similar to those obtained using conventional methods ($\dot{V}O_2 = 33.90$ and 35.57; Fig A2; see Withers (2001) for detailed methods), as well as the median of the full time series ($36.71\,ml\,m^{-1}$), but has two distinct advantages. First, the *L* metrics provide an objective measure of linearity to identify periods of stability in the $\dot{V}O_2$ time series. Second, the *L* metrics do not preferentially select the smallest possible subset of the time series to estimate resting metabolic rate, resulting in an estimate that is based on more observations, and therefore has more statistical power (*n*=47 using `LoLinR`; *n*=11 or 18 using conventional methods).

Data accessibility

All data used in this study are included in the `LoLinR` package and available at **`https://github.com/colin-olito/LoLinR`**.

Acknowledgments

The authors thank A. Pettersen and G. Yvon-Durocher for graciously providing data, and two anonymous reviewers.

Funding

This work was funded by a Monash University Dean's International Postgraduate Student Scholarship to CO, ARC grants to DJM and CRW, and a Monash University Centre for Geometric Biology Post-Doctoral Fellowship to DRB.

Competing Interests

The authors declare no competing or financial interests.

Author contributions

CO and DRB wrote the package. CO wrote the manuscript with assistance from CRW, DJM, and DRB.

References

Barneche, D. R., M. Kulbicki, S. R. Floeter, A. M. Friedlander, J. Maina, and A. P. Allen. 2014. Scaling metabolism from individuals to reef-fish communities at broad spatial scales. Ecology Letters 17:1067–1076.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. Toward a metabolic theory of ecology. Ecology 85:1771–1789.

Fang, F. C., R. G. Steen, and A. Casadevall. 2012. Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences of the United States of America 109:17028–17033.

Grieneisen, M. L., and M. Zhang. 2012. A Comprehensive Survey of Retracted Articles from the Scholarly Literature. PLoS One 7:e44118.

Gillooly, J. F., J. H. Brown, G. B. West, V. M. Savage, and E. L. Charnov. 2001. Effects of size and temperature on metabolic rate. Science 293:2248–2251.

Halsey, L. G., and C. R. White. 2010. Measuring energetics and behaviour using accelerometry in cane toads *Bufo marinus*. PLoS One 5(4):e10170.

Harrell F. E. 2001. Regression modelling strategies with applications to linear models, logistic regression and survival analysis. Springer, New York, USA.

Lagos, M. E., C. R. White, and D. J. Marshall. 2015. Avoiding log-oxygen environments as a mechanism of habitat selection in a marine environment. Marine Ecology Progress Series 540:99–107.

Larmola, T., J. L. Bubier, C. Kobyljanec, N. Basiliko, S. Juutinen, E. Humphreys, M. Preston, and T. R. Moore. 2013. Vegetation feedbacks of nutrient addition lead to a weaker carbon sink in an ombrotrophic bog. Global Change Biology 19:3729–3739.

Marshall, D. J., M. Bode, and C. R. White. 2013. Estimating physiological tolerances - a comparison of traditional approaches to nonlinear regression techniques. Journal of Experimental Biology 216:2178–2182.

Pettersen, A. K., C. R. White, and D. J. Marshall. 2015. Why does offspring size affect performance? Integrating metabolic scaling with life-history theory. Proceedings of the Royal Society B 282:20151946. (doi:10.1098/rspb.2015.1946)

Ross, D. J., A. R. Longmore, and M. J. Keough. 2013. Spatially variable effects of a marine pest on ecosystem function. Oecologia 172:525–538.

Shapiro, J. B., K. L. Griffin, J. D. Lewis, and D. T. Tissue. 2004. Response of *Xanthium strumarium* leaf respiration in the light to elevated CO_2 concntration, nitrogen availability, and temperature. New Phytologist 162:377–386.

Song, K., S.-H. Lee, and H. Kang. 2011. Denitrification rates and community structure of denitrifiying bacteria in newly constructed wetland. European Journal of Soil Biology 47:24–29.

Stinchcombe, J. R., Function-valued Traits Working Group, and M. Kirkpatrick. 2012. Genetics and evolution of funciton-valued traits: understanding environmentally responsive phenotypes. Trends in Ecology and Evolution 27:637–647.

Tanaka, T., S. Alliouane, R. G. B. Bellerby, J. Czerny, A. de Kluijver, U. Riebesell, K. G. Schulz, A. Silyakova, and J. P. Gattuso. 2013. Effect of increased $p\mathrm{CO}_2$ on the phytoplanktonic metabolic balance during a mesocosm experiment in an Arctic fjord. Biogeosciences 10:315–325.

White, C. R., D. Grémillet, J. A. Green, G. R. Martin, and P. J. Butler. 2011. Metabolic rate throughout the annual cycle reveals the demands of an Arctic existence in Great Cormorants. Ecology 92:475–486.

White, C. R., M. R. Kearney, P. G. D. Matthews, S. A. L. M. Kooijman, and D. J. Marshall. 2011. A manipulative test of competing theories for metabolic scaling. American Naturalist 178:746–754.

Withers, P. C. 2001. Design, calibration, and calculation for flow-through respirometry systems. Australian Journal of Zoology 49:445–461.

Yen, D. L. J., J. R. Thomson, D. M. Paganin, J. M. Keith, and R. MacNally. 2015. Function regression in ecology and evolution: FREE. Methods in Ecology and Evolution: 6:17–26.

Yvon-Durocher, G., A. P. Allen, M. Cellamare, M. Dossena, K. J. Gaston, M. Leitao, J. M. Montoya, D. C. Reuman, G. Woodward, and M. Trimmer. 2015. Five years of experimental warming increases the biodiversity and productivity of phytoplankton. PLoS Biology 13:e1002324.

Figures

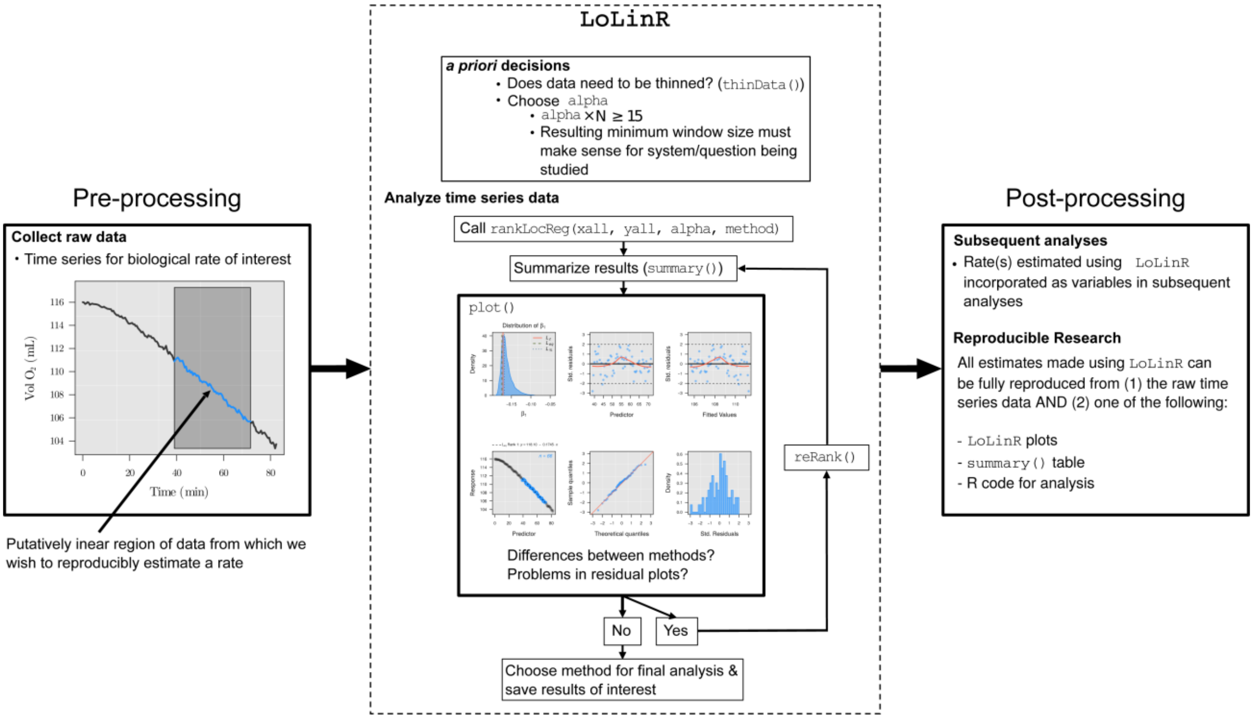
LoLinR
a priori decisions
• Does data need to be thinned? (thinData())
• Choose alpha
• alpha × N ≥ 15
• Resulting minimum window size must make sense for system/question being studied
Pre-processing
Collect raw data
• Time series for biological rate of interest
Vol O2 (mL)
Time (min)
Putatively inear region of data from which we wish to reproducibly estimate a rate
Analyze time series data
Call rankLocReg(xall, yall, alpha, method)
Summarize results (summary())
plot()
Differences between methods?
Problems in residual plots?
reRank()
No
Yes
Choose method for final analysis & save results of interest
Post-processing
Subsequent analyses
• Rate(s) estimated using LoLinR incorporated as variables in subsequent analyses
Reproducible Research
All estimates made using LoLinR can be fully reproduced from (1) the raw time series data AND (2) one of the following:
- LoLinR plots
- summary() table
- R code for analysis

Figure 1: Schematic showing both a typical LoLinR workflow, as well as how using LoLinR to estimate biological rates would fit into a experimental project workflow. In this example, O_2 saturation data is being used to estimate metabolic rate for an individual sea urchin (*Heliocidaris erythrogramma*; Olito and Marshall *unpublished data*). The putatively linear region of interest occurs after equipment equilibration, but before the urchin begins to show physiological responses to declining oxygen. Crucially, any estimate made using LoLinR can be easily reproduced from two pieces of information: (1) the raw time series data, and (2) LoLinR plots *OR* summary tables *OR* the R code used to implement the analysis. Each of these can be very easily compiled into appendices for published studies, making LoLinR analyses very easy to reproduce.

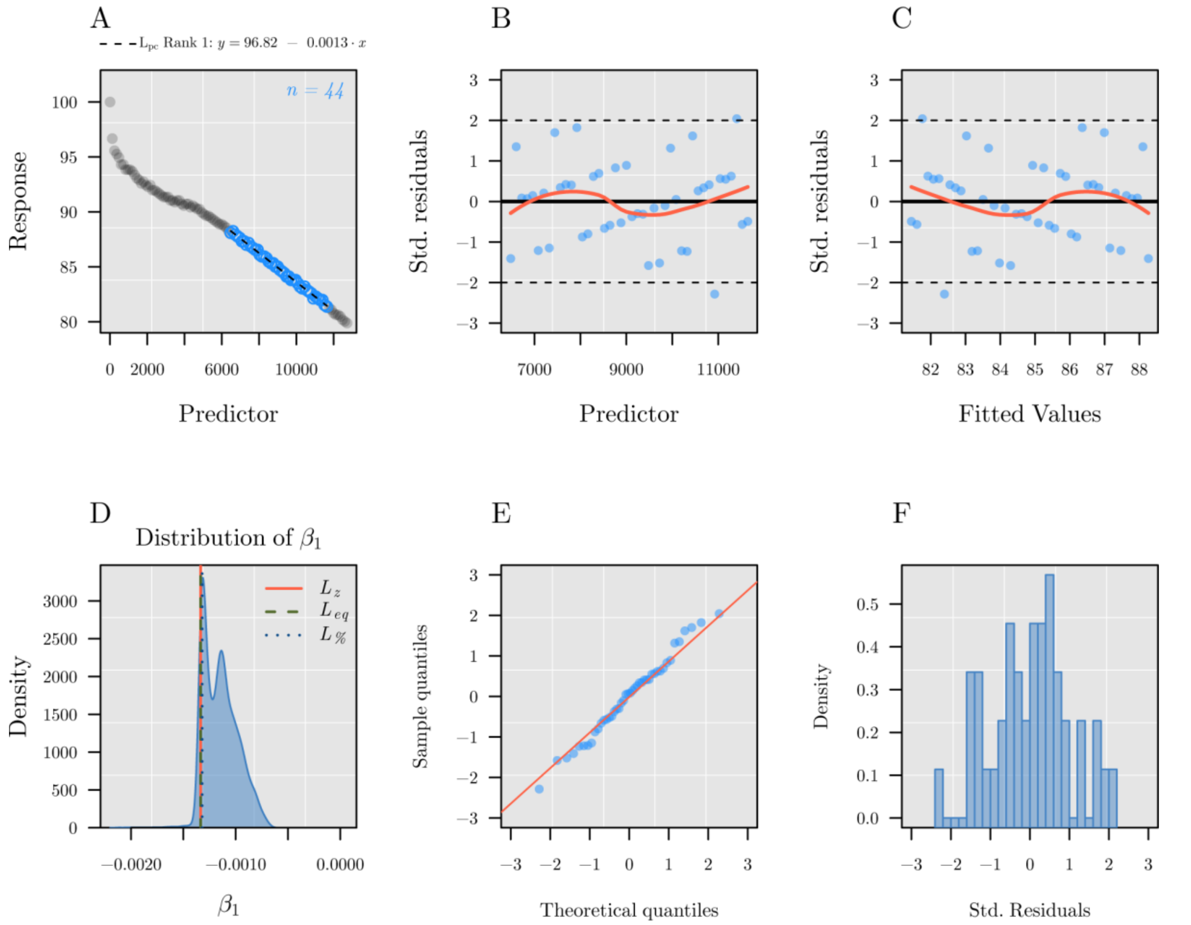

Figure 2: Diagnostic plots generated by the plot command for the $L_{\%}$ rank 1 local regression for the *Bugula* larva respiration time series data (Pettersen et al., 2015). The output plots show A) The full time series, with the rank 1 local regression highlighted in blue, along with the associated regression equation and number of observations; B) Standardized residuals for the chosen local regression, regressed against the predictor variable (time in seconds for the *Bugula* data set); C) Standardized residuals regressed against the fitted values; D) Density plot showing the empirical distribution of local regression slopes (β_1), benchmarked against the slopes of the rank 1 local regressions for each L metric; E) Normal-Quantile-Quantile plot for the rank 1 local regression; F) A histogram showing the distribution of the standardized residuals for the rank 1 local regression.

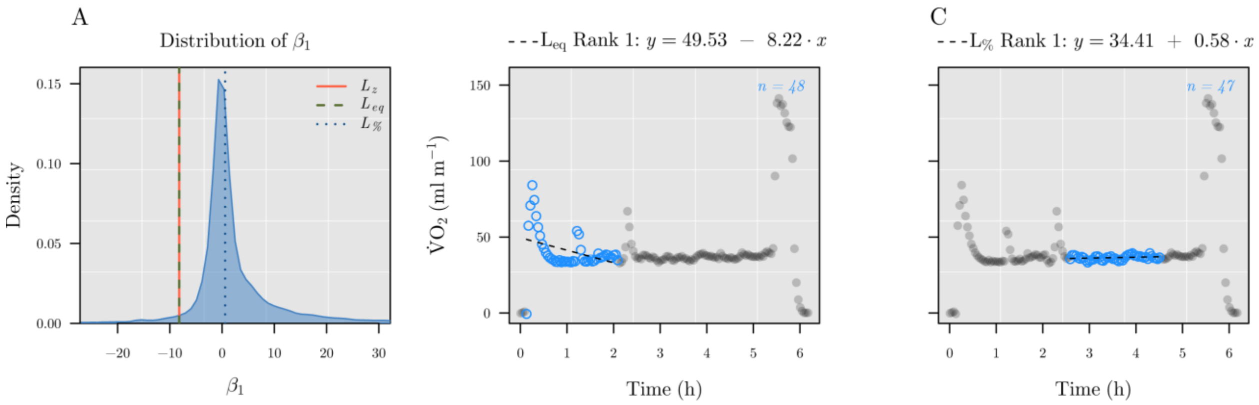

Figure 3: Diagnostic plots of the flow-through respirometry data (White et al., 2011) showing A) The distribution of local regression slopes (β_1) (Note that long tails on the distribution have been truncated to emphasize differences in the L metric benchmarks); B) the L_{eq} rank 1 local regression, which badly misidentifies the subset of the data where O_2 consumption was most stable; and C) The $L_\%$ rank 1 local regression, which identifies a reasonable subset of the full time series where the rate of O_2 consumption is most stable.

Supplementary figures

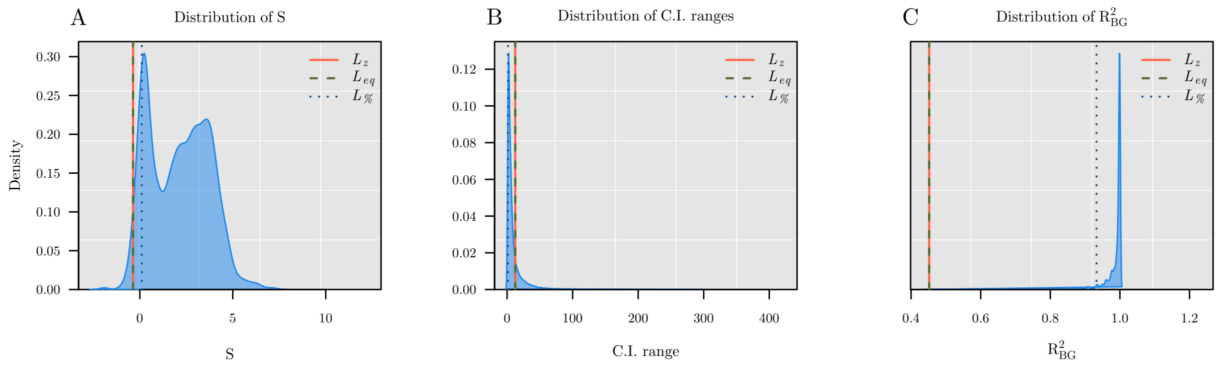

Figure A1: Empirical distributions of the component metrics A) *S*; B) *C.I. range*; and C) R^2_{BG} for the analysis of resting metabolic rate in Great Cormorants (raw data thinned using `thinData(CormorantData, by=3)`, and a call to `rankLocReg` using `alpha=0.1`). Note the extremely long tail in the distribution of R^2_{BG} (panel C), and the large discrepancy between the $L_{\%}$ benchmark and the other metrics.

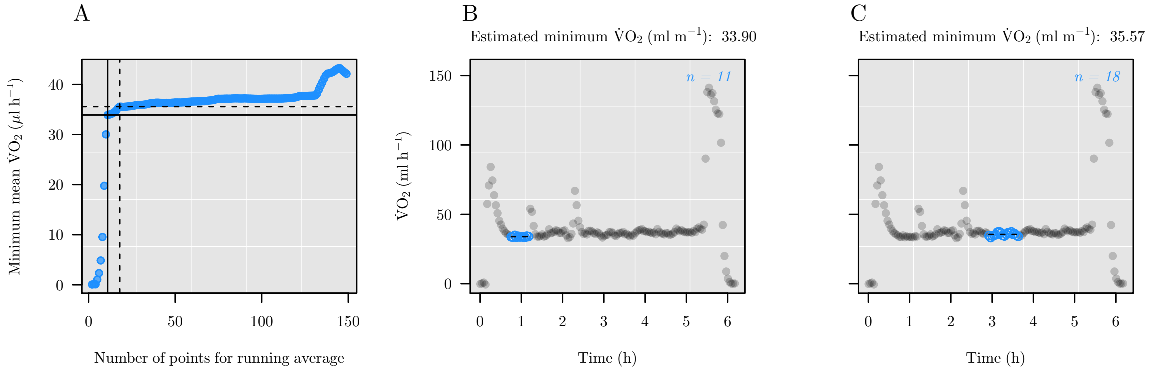

Figure A2: A graphical summary of conventional methods of estimating resting metabolic rate from $\dot{V}O_2$ time series data (after Withers, 2001). Panel A shows the minimum estimated $\dot{V}O_2$ as a function of the number of adjacent observations used to calculate the running average. After a rapid increase in the minimum $\dot{V}O_2$ with increasing observations up to $n = 11$ (solid black cross), there is a brief region with an intermediate slope, before the plot plateaus at $n = 18$ (dotted cross). Using these methods, a researcher could potentially justify using the running average associated with either of these two points as an estimate of resting metabolic rate. Panels A and B show the full time series, with the subsets associated with $n = 11$ and $n = 18$ respectively highlighted in blue. Note that both subsets are much smaller than the number of observations included in the $L_{\%}$ rank 1 local regression identified by `rankLocReg` (Fig. 3).

Quantifying LoLinR performance using simulated data

An informative comparison between the methods provided in LoLinR and common alternatives is deceptively difficult for at least two reasons. First, the most common alternative methods (e.g., eyeballing the data to select a linear region and then running a linear regression, or the methods described in Withers (2001)) simply cannot be reliably compared to LoLinR because they are not objectively reproducible. Second, generating appropriate simulated data for the types of analyses LoLinR is designed to assist with (i.e., where some subset of a time series is indeed linear, or expected to be linear, but the rest is arbitrarily non-linear) is a non-trivial problem. This is particularly problematic when the specific nature of the non-linearity has a strong influence on the behaviour of the methods that LoLinR might be compared against (e.g., naive linear regression of full data sets). As a first step towards providing objective validation and comparison of our methods, we provide functions to generate simulated data and analyze the performance of rankLocReg (see the supplementary files 'functions-supplement.R'). However, we emphasize that this is not a comprehensive sensitivity analysis, but rather a starting point for validating our methods and making future comparisons with other reproducible methods. Here, we briefly describe how simulated data was generated, and how the performance of rankLocReg was assessed.

We generated simulated data that roughly resemble O_2 consumption data (similar to the 'Larval Metabolic Rate' example in the main article), with an initial phase of acceleration/deceleration which then stabilizes as a straight line. Simulated data sets are composed of 100 observations. The first 50 observations are non-linear, following a sine wave from the apex (or trough) at $\pm\pi/2$ to the inflection point at $\pm\pi$. The second 50 observations are linear, following $y = \beta_0 \pm x\beta_1$, where β_0 is equal to the 50^{th} observation (the last of the non-linear subset), and the regression slope is uniformly distributed on the intervals $\beta_1 \in [-0.028, -0.004] \cup [0.004, 0.028]$. We add a small amount of normally distributed noise to the entire data set ($\epsilon \sim N(0, 0.05)$) to simulate random variation in present in real time series data. We analyze each randomly generated data set using `rankLocReg`, with `alpha=0.2`.

To quantify the performance of `rankLocReg`, we use three simple metrics: (1) the difference between the actual slope of the linear subset of the simulated data, and the slope of the local linear regressions identified by `rankLocReg` ($\Delta_i = \beta_i - \beta_{real}$), where i indicates each of the L methods used by `rankLocReg` ($i \in [Z, Eq, \%]$). Since each of the three L metrics perform differently for different data sets, we also compare the difference between the real regression slope and the best of the three local regressions identified by `rankLocReg` (i.e., the one with the smallest Δ_i), which we designate Δ_{best}; (2) the proportion of the linear subset of the data that is correctly included in the local linear regressions identified by `rankLocReg`; and (3) the proportion of the observations included in local linear regressions that correctly include the linear subset of the data.

Results summary

For this paticular type of simulated test data, `rankLocReg` performs remarkably well, particularly in comparison with a naive linear regression of the full data sets. The result of this specific comparison is not surprising, however, as the curvature in the first half of the simulated data results in systematic bias of β_{naive} towards 0. This is reflected in Fig. A3A, where the distribution of Δ_{naive} is right skewed with a thicker tail than the distribution of Δ_{best}. In contrast, Δ_{best} is tightly distributed about 0, with a few outliers in the right tail. Each of the three L metrics performs similarly, although the local regressions identified using the $L_{\%}$ methods are generally better at recovering regression slopes that are more similar to β_{real} (Fig. A3B). As expected, the absolute values of β_{real} also influence the performance of `rankLocReg`. Overall, `rankLocReg` performs better when β_{real} is further from 0 (Fig. A3C). Specifically, when β_{real} is negative but close to 0, `rankLocReg` tends to choose regressions that include the non-linear portion of the data, with slopes that are more steeply negative than β_{real}. When β_{real} is positive but close to 0, `rankLocReg` tends to choose local regressions with more steeply positive slopes than β_{real} for the same reasons. This behaviour makes sense because the 95% C.I. for β_1 is used to calcuate the L metrics used by `rankLocReg`, which becomes increasingly inflated as β_{real} approaches 0.

`rankLocReg` also does a reasonably good job of correctly identifying the truly linear subset of these simulated data. This is encouraging, particularly because the curvature of the simulated data in this example should make this rather difficult. This is because the second half of non-linear portion of the simulated data are increasingly linear (with a slope of $\beta_1 \approx = \pm 0.016$ on the x-scale used for this analysis) as they approach the inflection point of the sine wave at $\pm\pi$. Thus, it should be difficult for `rankLocReg` to distinguish between the end of the non-linear subset of the data, and the truly linear subset. However, at least one of the L methods implemented by `rankLocReg` (i.e., the 'best' local regression with the smallest Δ value) generally included a large fraction of the truly linear subset of the data (Fig. A3D). However, there was quite a bit of variability in the performance of each of the L methods (Fig. A3E). The L_Z and L_{eq} methods in particular either performed very well, or very poorly, at identifying the truly linear subset of the data. In contrast, the $L_{\%}$ method generally identified at least half of the truly linear subset (Fig. A3E). The ability of `rankLocReg` to correctly identify the truly linear subset of the data was again sensitive to the aboslute value of β_{real}. Specifically, even the 'best' local regression mis-identified the truly linear subset more frequently as β_{real} approached 0 (Fig. A3F).

`rankLocReg` also performed well at identifying local regressions that correctly include the truly linear subset of the data. For a large majority of simulated data sets, more than half of the observations included in the 'best' local regression identified by `rankLocReg` were part of the truly linear subset of the data (Fig. A3G). However, there was again significant variability in the performance of each of the three L methods (Fig. A3H). The $L_{\%}$ method clearly performed the best in this respect, generally choosing local regressions with the majority of observations falling within the truly linear subset of the data (Fig. A3H). On the other hand, the L_Z method generally performed very poorly, choosing local regressions that badly mis-identified the linear subset of the data, or choosing local regressions of which only half of the included observations were actually part of the truly linear subset (Fig. A3H). The L_{eq} method also performed poorly, but was better at identifying local regressions with a majority of observations falling within the truly linear subset of the data than the L_Z method (Fig. A3H). Once again, the performance of even the 'best' local regression became worse as β_{real} approached 0. The variability in the performance of the three L methods, is almost certainly a direct consequence of the difficulty in distinguishing between the end of the non-linear and the beginning of the linear subsets of these simulated data. This is particularly clear for $L_{\%}$, which often identified local regressions that spanned both the non-linear and linear portions of the data.

Four main conclusions emerge from this limited test of the performance of `rankLocReg` against simulated data. First, `rankLocReg` performs better than naive linear regression of full time series at estimating the slope of a linear subset of the time series. Second, this analysis strongly supports our recommendation in the main article that $L_{\%}$ be used as the preferred weighting method. $L_{\%}$ is generally more robust, and does a better job than the other methods of choosing local regressions that both accurately estimate β_{real} and correctly include the truly linear subset of these simulated data. Third, `rankLocReg` becomes progressively better at accurately estimating β_{real} that are further from 0 (and presumably further from 1 as well). Last, this analysis highlights that while there can be significant variation in the performance of each of the three L metric weighting methods, it is rare that all three mis-identify the truly linear subset of the time series. Taken together, these results indicate that the methods provided in `LoLinR` perform well for their intended purpose for the type of data simulate here.

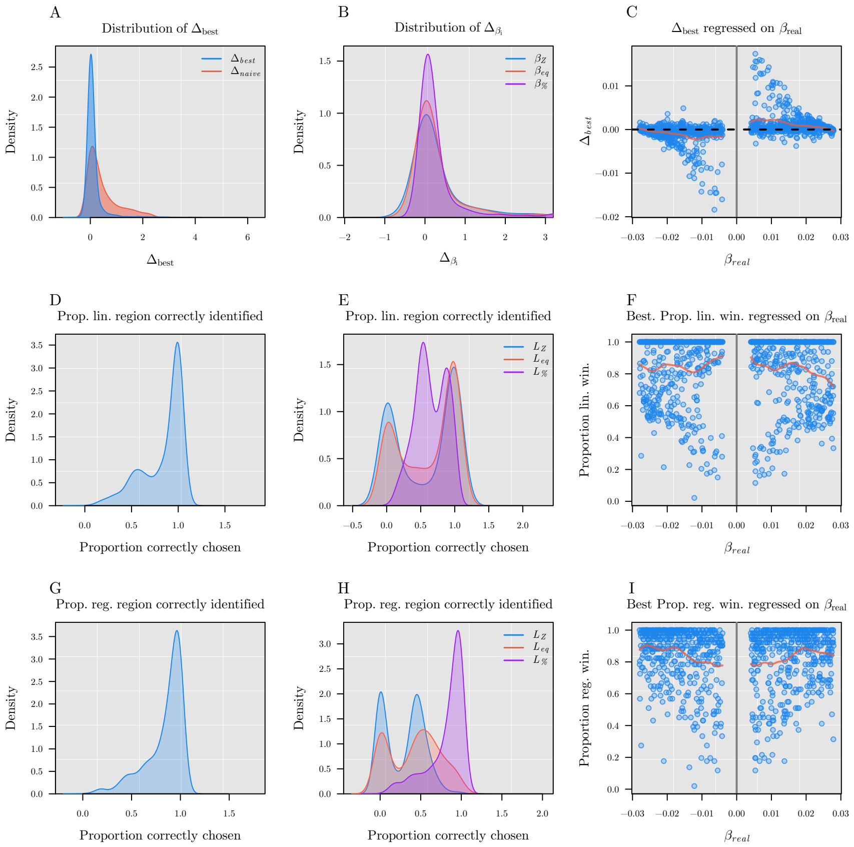

Figure A3: Graphical summary of the performance of `LoLinR` methods using certain types of simulated data. Panels A-C show results for the performance metric Δ; panels D-F show results for the proportion of the linear subset of the data that are correctly included in the local regressions identified by `rankLocReg`; and panels G-I show results for the proportion of the local regression identified by `rankLocReg` that correctly includes the linear subset of the simulated data. The first column of panels (A,D,G) show the distributions for each performance metric of the 'best' local regression identified by `rankLocReg` (the local regression with the smallest Δ). Note that the x-axes are scaled in units of β_{real}; thus, in panel A, a Δ value of 2 indicates that Δ_{best} was twice as large as the real slope of the simulated data (β_{real}. Panel A shows a comparison between the performance of `rankLocReg` and a naive regression by comparing the distributions of Δ_{best} and Δ_{naive}. As expected, the slopes identified by naive regression are systematically biased by the non-linearity present in the simulated data, while `rankLocReg` is better able to identify slopes that are closer to the β_{real}. The second column of panels (B,E,H) show the distributions of each performance metric for each of the three L metric methods provided in `rankLocReg`. The third column of panels (C,F,I) show scatter plots of each performance metrics regressed on the actual slope of the simulated data (β_{real}), with smoothing splines overlaid to help visualize any trends. This is to visualize that, as expected, `rankLocReg` performs better as the slope of the linear region is further from 0. Note that in panel B long right tails have been truncated to better visualize the peaks of the distributions

Additional files

See additional supplementary files 'ReproducePaperAnalyses.R' and 'functions-supplement.R', for R code necessary to reproduce all analyses and figures presented in the paper and supplement.

Click here to Download Reproduced Analyses

Click here to Download Supplementary R functions

References

Withers, P. C. 2001. Design, calibration, and calculation for flow-through respirometry systems. Australian Journal of Zoology 49:445–461.

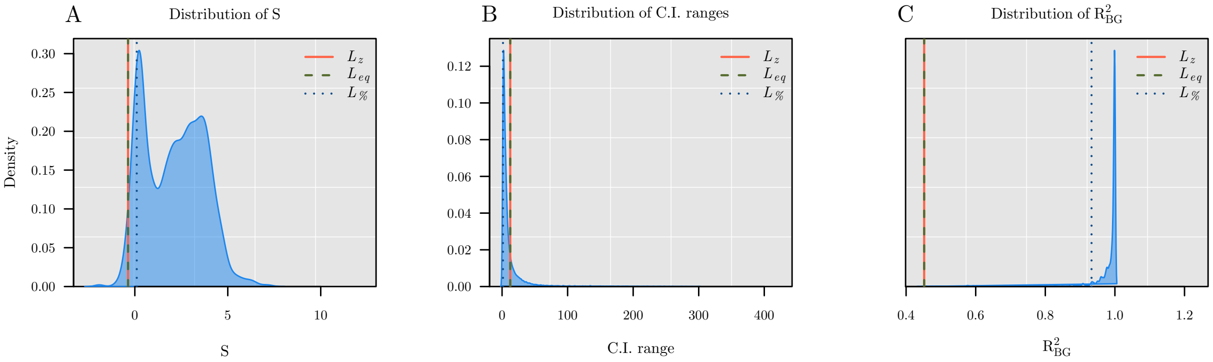

Fig. S1. Empirical distributions of the component metrics A) *S*; B) *C.I. range*; and C) $R^2{}_{BG}$ for the analysis of resting metabolic rate in Great Cormorants (raw data thinned using `thinData(CormorantData, by=3)`, and a call to `rankLocReg` using `alpha=0.1`). Note the extremely long tail in the distribution of $R^2{}_{BG}$ (panel C), and the large discrepancy between the $L_{\%}$ benchmark and the other metrics.

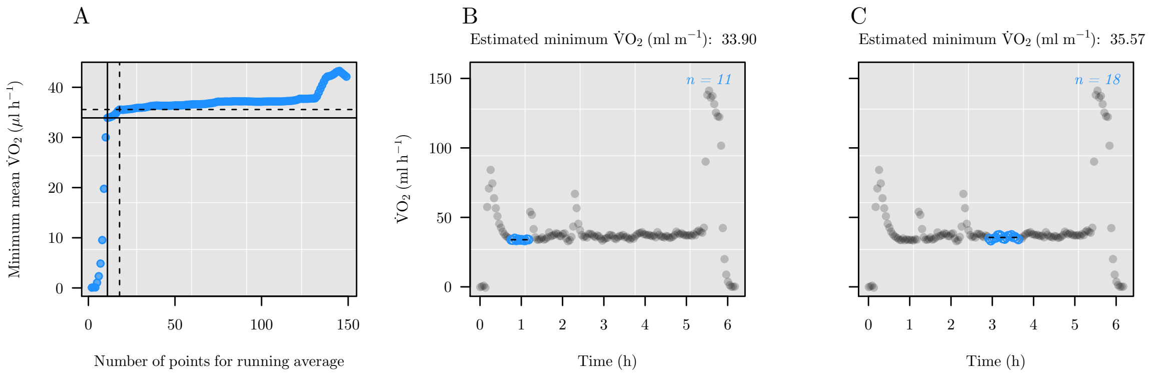

Fig. S2. A graphical summary of conventional methods of estimating resting metabolic rate from $\dot{V}O_2$ time series data (after Withers, 2001). Panel A shows the minimum estimated $\dot{V}O_2$ as a function of the number of adjacent observations used to calculate the running average. After a rapid increase in the minimum $\dot{V}O_2$ with increasing observations up to $n = 11$ (solid black cross), there is a brief region with an intermediate slope, before the plot plateaus at $n = 18$ (dotted cross). Using these methods, a researcher could potentially justify using the running average associated with either of these two points as an estimate of resting metabolic rate. Panels A and B show the full time series, with the subsets associated with $n = 11$ and $n = 18$ respectively highlighted in blue. Note that both subsets are much smaller than the number of observations included in the $L_{\%}$ rank 1 local regression identified by `rankLocReg` (Fig. 3).

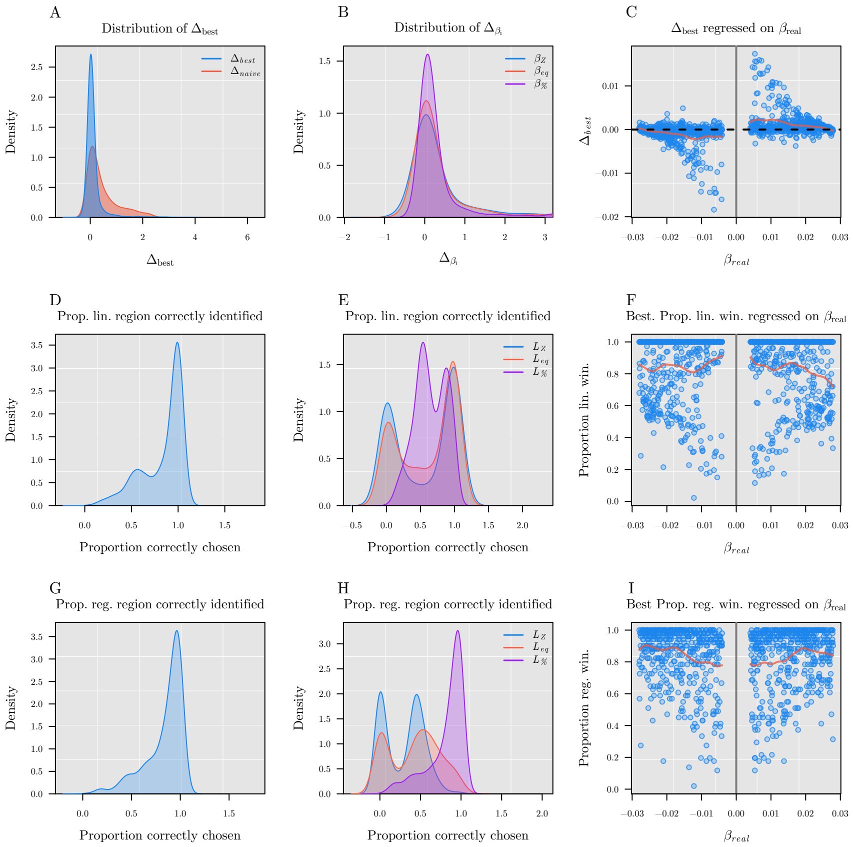

Fig. S3. Graphical summary of the performance of LoLinR methods using certain types of simulated data. Panels A-C show results for the performance metric Δ; panels D-F show results for the proportion of the linear subset of the data that are correctly included in the local regressions identified by rankLocReg; and panels G-I show results for the proportion of the local regression identified by rankLocReg that correctly includes the linear subset of the simulated data. The first column of panels (A,D,G) show the distributions for each performance metric of the 'best' local regression identified by rankLocReg (the local regression with the smallest Δ). Note that the x-axes are scaled in units of β_{real}; thus, in panel A, a Δ value of 2 indicates that Δ_{best} was twice as large as the real slope of the simulated data (β_{real}. Panel A shows a comparison between the performance of rankLocReg and a naive regression by comparing the distributions of Δ_{best} and Δ_{naive}. As expected, the slopes identified by naive regression are systematically biased by the non-linearity present in the simulated data, while rankLocReg is better able to identify slopes that are closer to the β_{real}. The second column of panels (B,E,H) show the distributions of each performance metric for each of the three *L* metric methods provided in rankLocReg. The third column of panels (C,F,I) show scatter plots of each performance metrics regressed on the actual slope of the simulated data (β_{real}), with smoothing splines overlaid to help visualize any trends. This is to visualize that, as expected, rankLocReg performs better as the slope of the linear region is further from 0. Note that in panel B long right tails have been truncated to better visualize the peaks of the distributions

Click here to Download Script S1

Click here to Download Script S2