Published in final edited form as:
Phys Rev E Stat Nonlin Soft Matter Phys. 2011 September ; 84(3-1): 031914.

Objective method for estimating asymptotic parameters, with an application to sequence alignment

Sergey Sheetlin, **Yonil Park**, and **John L. Spouge***
National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA

Abstract

Sequence alignment is an indispensable computational tool in modern molecular biology. The model underlying biological sequence alignment is of interest to physicists because it approximates the statistical mechanics of DNA and protein annealing, while bearing an intimate relationship to models of directed polymers in random media. Recent methods for determining the statistics of random sequence alignments have reduced the computation time to less than 1 s, opening up some interesting possibilities for online computation with biological search engines. Before implementation, however, the methods required an objective technique for computing regression coefficients pertinent to an asymptotic regime. Typically, physicists estimate parameters pertinent to an asymptotic regime subjectively: They eyeball their data; estimate the asymptotic regime where the regression model holds with reasonable accuracy; and then regress data only within the estimated asymptotic regime. Our publicly available computer program ARRP replaces the subjective assessment of the asymptotic regime with an objective change-point detection method, increasing confidence in the scientific objectivity of the parameter estimates. Asymptotic regression has potential applications across most of physics.

I. INTRODUCTION

Sequence alignment is an indispensable computational tool in modern molecular biology, and some physicists have even regarded it as a problem worthy of study in its own right [1,2]. Sequence alignment has intimate connections to physical models of nucleic acid and protein annealing because the models of sequence alignment described below approximate the equilibrium statistical mechanics of annealing, essentially by replacing the ensemble average of a Boltzmann distribution with its mode [3–5]. The models also bear interesting mathematical analogies to familiar models of condensed matter in physics, such as first-passage percolation or interacting particle systems [6,7]. Moreover, with numerically unimportant differences [8,9], they are intimately related to models for a directed polymer in random media (DPRM) [10–13]. The DPRM problem itself also maps onto other important models in physics [14], such as the Kardar-Parisi-Zhang equation [15], which describes random growth at a surface, an aggregation process of long-standing interest [16], and the noisy Burgers equation [17,18].

Sequence alignment permits biologists to infer the functional, structural, and evolutionary relationships of a novel protein or nucleic acid sequence by searching a database for similar sequences of known functions. Local alignment, which compares subsequences in sequences [19], usually provides the method of choice to determine the relationships because the specific functionalities inherent in a biological sequence can usually be attributed to specific

*spouge@ncbi.nlm.nih.gov.

subsequences within it [20–22]. In contrast to local alignment, global alignment compares whole sequences, so it is probably more effective than local alignment when classifying sequences.

The theory of sequence alignment relies heavily on path optimization [23]. The following abbreviated presentation of sequence alignment as a path optimization is available in greater detail elsewhere [24]. (Note: Later sections on asymptotic regression are mostly independent of the theory of sequence alignment, should the reader wish to skip ahead.)

Let $\mathbf{A} = A_1A_2 \ldots$ and $\mathbf{B} = B_1B_2 \ldots$ be two infinite sequences drawn from a finite alphabet L, e.g., from the amino acid alphabet {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}. Let $S: L \times L \mapsto \mathbb{R}$ denote a scoring matrix. In a physical application, $S(a,b)$ represents free energy dissipated when, e.g., amino acids a and b from two different protein molecules form hydrogen bonds. In database applications, $S(a,b)$ is a similarity matrix, quantifying the similarity between a and b.

The alignment graph Γ of the sequence pair (**A,B**) is a directed weighted lattice graph in two dimensions. (See Fig. 1.) The vertices v of Γ are non-negative integer points (i,j). (Below, := denotes a definition; and i, j, k, m, n, and g are integers throughout the paper.) Three sets of directed edges e come out of each vertex $v = (i,j)$: northward, northeastward, and eastward. One northeastward edge goes into $(i + 1, j + 1)$ with weight $S(A_{i+1}, B_{j+1})$. For each $g > 0$, one eastward edge goes into $(i + g, j)$ and one northward edge goes into $(i, j + g)$; both are assigned the same weight $-\Delta(g) < 0$, where $\Delta(g)$ is the so-called gap penalty.

A directed path $\pi = (v_0, e_1, v_1, e_2, \ldots, e_k, v_k)$ in Γ is a finite alternating sequence of vertices and edges that starts and ends with a vertex. For each $i = 1,2, \ldots,k$, the directed edge e_i comes out of vertex v_{i-1} and goes into vertex v_i. Thus, path π starts at v_0 and ends at v_k.

Denote subsequences of a sequence **A** by $\mathbf{A}(i,m] = A_{i+1}A_{i+2} \ldots A_m$. Every gapped alignment of the subsequences $\mathbf{A}(i,m]$ and $\mathbf{B}(j,n]$ corresponds to exactly one directed path that starts at $v_0 = (i,j)$ and ends at $v_k = (m,n)$. The alignment's score is the path weight $W_\pi := \sum_{i=1}^{k} W(e_i)$. The definitions extend naturally to trivial cases: for an empty path π consisting of a point $W_\pi := 0$, and $\mathbf{A}(i,i] = \varnothing$ (the empty string).

Define the global score $S_{i,j} := \max_\pi W_\pi$, where the maximum is taken over all paths π starting at $v_0 = (0,0)$ and ending at $v_k = (i,j)$. The paths π starting at v_0 and ending at v_k with weight $W_\pi = S_{i,j}$ are optimal global paths and correspond to optimal global alignments between $\mathbf{A}(0,i]$ and $\mathbf{B}(0,j]$. Therefore, the global alignment score is closely related to a first-passage percolation time. Let $\mathbb{Z}_+ := \{0,1,2, \ldots\}$ denote the non-negative integers. The edge maximum is $E_n := \max(\{S_{n,j}: 0 \le j \le n\} \cup \{S_{i,n}: 0 \le i \le n\})$; the global maximum is $M := \max\{E_n: n \in \mathbb{Z}_+\}$; and the set $\{(i, j) \in \mathbb{Z}_+^2 : S_{i,j} = s\}$ of vertices with global score s contains $N(s)$ vertices. Note: Although $N(s)$, in principle, can be infinite, it is finite with probability 1, under the conditions of interest (given below).

Also, define the local score $\hat{S}_{i,j} := \max_\pi W_\pi$, where the maximum is taken over all paths π ending at $v_k = (i,j)$, regardless of their starting point. Define the local maximum $\hat{M}_{m,n} := \max\{\hat{S}_{i,j}: 0 \le i \le m, 0 \le j \le n\}$. The paths π ending at $v_k = (i,j)$ with local score $W_\pi = \hat{S}_{i,j} = \hat{M}_{m,n}$ are optimal local paths, corresponding to the optimal local alignments between subsequences of $\boldsymbol{A}(0,m]$ and $\boldsymbol{B}(0,n]$.

Introduce randomness with the so-called independent letters model: Choose each letter in sequences $\boldsymbol{A}$ and $\boldsymbol{B}$ independently from a fixed distribution on the alphabet L [25]. Under appropriate conditions, in the so-called logarithmic regime [26], as $m,n \to \infty$, the

distribution of the random local maximum $\hat{M}_{m,n}$ approximates a Gumbel extreme value distribution [27–29]. The Gumbel distribution corresponds to the tail probability,

$$\mathbb{P}(\widehat{M}_{mn} > s) \approx 1 - \exp[-Kmn\exp(-\lambda s)], \tag{1}$$

where the random letters model determines the probability $\mathbb{P}$[and, hence, the corresponding expectation $\mathbb{E}$, in Eq. (2) below]. The Gumbel distribution in Eq. (1) has prefactor K and scale parameter λ.

Over the Web, biologists search the sequence databases at the National Center for Biotechnology Information about three times a second. BLASTP is the standard computer program for searching protein databases. Its defaults are the so-called BLOSUM62 scoring matrix [30], the affine gap penalty $\Delta(g) = 11 + g$, and the so-called Robinson and Robinson random amino acid letter frequencies [31]. All simulations in the results used the BLASTP defaults because the corresponding Gumbel parameters $\lambda \approx 0.2666 \pm 0.0003$ and $K \approx 0.0409 \pm 0.0003$ were known for exquisite accuracy [32].

The Gumbel distribution in Eq. (1) provides a limit as the sequence length $m,n \to \infty$. As an approximation, its accuracy generally degrades as m and n decrease or as the score s increases. A finite-size correction (FSC) to the Gumbel distribution enhances the approximation's accuracy, however [33]. The FSC currently implemented in BLASTP estimates $\mathbb{P}(\hat{M}_{mn} > s)$ within a factor of 2 for $m = n = 40$ and for p values as small as 10^{-7}, accuracy good enough for many practical purposes. The theory of the FSC is well beyond the scope of this paper [34,35], but the accuracy the FSC provides makes the Gumbel parameters λ and K fundamental to protein database searches with the BLASTP program.

Presently, BLASTP needs to compute the Gumbel parameters λ and K offline, so it gives users a narrow choice of about six standard sets of alignment parameters. One research goal in bioinformatics over the last decade has been to compute λ and K online in less than 1 s, so a user could choose alignment parameters arbitrarily. The computer program ALP (ascending ladder point) fulfilled the aim (along with the computation of the finite-size correction); the ALP program is publicly available [36].

ALP relied on several novelties, including different mathematical expressions for λ and K, given in Eqs. (2) and (3) below. The equations permit simulations of global alignment to estimate of the statistical parameters λ and K for local alignment. The theory behind the equations is beyond the scope of this paper but is explained in detail elsewhere [24,37]. The theory, in its entirety, likely extends to directed polymers in random media.

All statements in the remainder of this section are well-grounded speculations, but they have no mathematical proof. The logarithmic regime corresponds to the condition $\limsup_{n\to\infty} \mathbb{E}[E_n - E_{n-1}] < 0$, where E_n is the edge maximum defined above. In the following, let k be an arbitrary fixed positive integer, and let $\lambda_{n-k,n} > 0$ satisfy:

$$\mathbb{E}[\exp(\lambda_{n-k,n} E_n)] = \mathbb{E}[\exp(\lambda_{n-k,n} E_{n-k})]. \tag{2}$$

In the logarithmic regime, the following speculations pertain. For every n, the root $\lambda_{n-k,n} > 0$ exists, and for n greater than some n_0, the root $_{n-k,n} > 0$ is unique. Moreover, the Gumbel scale parameter λ satisfies $\lambda = \lim_{n\to\infty} \lambda_{n-k,n}$ for every arbitrary fixed k. To derive the Gumbel prefactor K, if M is the global maximum defined above, and the set

$\{(i, j) \in \mathbb{Z}_+^2 : S_{i,j} = s\}$ of vertices with global score s contains $N(s)$ vertices (also defined above), and if

$$K_s = \frac{e^{\lambda s}}{1 - e^{-\lambda}} \frac{\mathbb{P}(M = s)^2}{\mathbb{E}N(s)}, \tag{3}$$

then $K = \lim_{s\to\infty} K_s$.

Simulations can estimate both sides of Eq. (2), yielding estimates of $\lambda_{n-k,n}$ for substitution into the limit $\lambda = \lim_{n\to\infty} \lambda_{n-k,n}$. Similarly, simulations can estimate the right side of Eq. (3), yielding estimates of K_s for substitution into the limit $K = \lim_{s\to\infty} K_s$. The ALP program needed to extract the constant limits λ and K from Eqs. (2) and (3) objectively, motivating the asymptotic regression methods described below.

II. ASYMPTOTIC REGRESSION

To start simply, consider the hypothetical relationship $y = \beta + f(t)$ for times t, where β is a constant and $\lim_{t\to\infty} f(t) = 0$ [e.g., $f(t) = ae^{-bt}$ with $a,b > 0$]. The times t might be either discrete (e.g., $t = 0,1,2, \ldots$) or continuous. In many cases, the asymptotic regression coefficient $\beta = \lim_{t\to\infty} y$, characterizing the permanent long-time behavior of y is of primary interest, while the transient function $f(t)$ is relatively uninteresting. In many applications, each time measurement t is essentially free from error, although each measurement of y has a random error e with known (or approximately known) distribution. (The distribution of e might depend on t and/or y, however.) Thus, examination of the constant asymptotic regression model $y = \beta + \varepsilon + f(t)$, where the asymptotic error ε has standard deviation $\sigma(\varepsilon)$, is interesting. Sometimes, data collection is expensive, or the need for accuracy is paramount, so our aim is to regress the data to estimate the asymptotic coefficient β as efficiently as possible.

In applications, the asymptotic control variable t controlling the asymptotic regime $t \to \infty$ is often time, but other asymptotic control variables appear in the examples below. Note also that reparametrization makes no essential change in the asymptotic regime, e.g., $\tau = 1/t \to 0$, so unless stated otherwise, throughout this paper, the asymptotic regime is $t \to \infty$. The complement of the asymptotic regime is called the transient regime (e.g., with the asymptotic regime $t \to \infty$, the transient regime consists of the time interval before y manifests its asymptotic behavior).

In many scientific fields other than physics, the estimation of an asymptotic coefficient is important, e.g., in agricultural statistics, the effect of fertilizer on mature growth and final crop yields [38]; in animal ecology, the upper asymptote of the number of prey eaten per predator as prey density increases [39]; and in pharmacology, the estimation of the maximum drug efficacy in the dose-response curve [40].

In many areas in physics, however, the estimation of an asymptotic coefficient (e.g., a critical exponent) is the actual focus of research. Examples follow: in chemical physics, the estimation of the energy value for Hartree-Fock equations [41]; in polymer science, the estimation of the asymptotic exponents characterizing the radius of gyration in polymers [42]; in condensed matter physics, the spectral properties of liquid crystals driven by an external electric field [43] or theoretical investigations of the validity of Fourier's law of heat conduction in one dimension [44]; and in relativistic physics, the exponents characterizing diffusion laws in relativistic turbulent media [45]. A formal method of

asymptotic regression would benefit these and many other studies by strengthening confidence in the objectivity of their scientific conclusions. Section I indicates our specific motivation for developing such a method, namely, a desire to estimate the statistical parameters of local sequence alignment in less than 1 s, without human supervision [24,46–48].

Section III examines simple linear regression, while Sec. V considers general linear regression. For the time being, however, let us focus on the constant asymptotic regression model $y = \beta + \varepsilon + f(t)$, with the aim of estimating the asymptotic regression coefficient β. If detailed knowledge of the transient function $f(t)$ is available, e.g., if $f(t)$ is a function of known form with unknown parameters, standard regression techniques applied to all the data can estimate the asymptotic regression coefficient β. In many applications, however, the form of the transient function $f(t)$ is unknown. In the absence of an objective methodology, many scientists presently estimate β by eyeballing their data; estimating the asymptotic regime where $f(t) \approx 0$, so the ordinary linear regression model $y = \beta + \varepsilon$ holds with reasonable accuracy; censoring the transient regime; and then regressing only the asymptotic data. A typical summary phrase is, e.g., “The asymptote was fitted by eye…” [49,50]. In our context, the need for an independent computer program rendered subjective methods useless.

Section III details our method, which superficially resembles change-point regression, also known as segmented regression or two-phase regression [51]. In present terminology, change-point regression requires two regression models, one for the transient regime and another for the asymptotic regime. A cumulative sum (CUSUM) procedure, similar to the procedure described later in Sec. III, estimates the change point separating the regimes. It then regresses each of the two separate data sets with the appropriate model. Thus, a change-point regression fits all data points. In its initial step, our asymptotic regression method uses a similar CUSUM procedure to estimate the change point between the transient and the asymptotic regimes, effectively discovering when the transient bias from $f(t)$ starts to have a magnitude comparable to the asymptotic standard deviation $\sigma(\varepsilon)$. In contrast to change-point regression, however, our asymptotic regression only regresses the data belonging to the asymptotic regime.

A review [52] of methods for estimating the parameters governing an asymptotic power law presents a competing maximum likelihood method. The competing method concerns itself primarily with model rejection, however. Moreover, it cannot readily use information about errors. Therefore, it is not adapted to physicists’ needs because, in physics, usually the relevant asymptotic model is known in advance from mathematical considerations anyway, while the experiments or simulations producing the data points provide accompanying errors. The error structure of the data has value in determining model parameters, so if available, clearly it should be exploited.

Because data points in the transient regime usually look like outliers in the asymptotic linear regression model, asymptotic regression superficially resembles robust regression [53]. Robust regression (described briefly in the next section) considers outliers one by one, without regard to the contiguity of any regime and reduces their influence in determining the regression line appropriately, whereas, asymptotic linear regression determines the transient regime as a whole and then censors it completely. Because robust regression is an accepted method and a possible alternative for determining asymptotic parameters, we compare asymptotic regression to it.

The organization of this paper follows. Section III gives our asymptotic regression method, briefly summarizes robust regression, and then describes our computer programs. Section IV

then illustrates asymptotic regression with two examples that use simulation data to calculate the statistical parameters pertinent to local sequence alignment in bioinformatics. Finally, Sec. V gives our concluding remarks.

III. METHODS

A. The constant asymptotic regression model $y = \beta + \varepsilon + f(t)$ as $t \to \infty$

Consider the data $(t_i, y_i \pm s_i)(i = 1, \ldots, n)$, where $t = t_i$ is the asymptotic control variable. Reorder the data if necessary, so $t_1 \le t_2 \le \ldots \le t_n$. Define the true normalized residuals $e_i^* := (y_i - \beta)/s_i$ (where:= denotes a definition). Assume that each true normalized residual e_i^* is drawn independently from a common distribution (possibly approximate and usually standard normal in applications). As is typical in applications, the change-point time when the asymptotic regime starts, i.e., when the limiting model $y = \beta + \varepsilon$ starts to pertain, is unknown. (See Fig. 2.)

To estimate the change point from the transient to the asymptotic regime, we require a criterion function $\rho(e^*)$, which could be $\rho(e^*) = \frac{1}{2}(e^*)^2$ (least-squares normalized errors) or a robust alternative, such as $\rho(e^*) = |e^*|$ (L^1-normalized errors), etc. [53]. A common alternative in robust regression is Andrews' wave function $\rho(e^*) = a[1 - \cos(e^*/a)]$ for $|e^*| \le \pi a$ and $\rho(e^*) = 2a$ otherwise, where $a > 0$ is an arbitrary parameter. Among popular robust criterion functions, the Andrews wave function behaves most like asymptotic regression because it assigns weight 0 to every point with $|e^*| > a\pi$, effectively censoring the point completely.

Assume that, under the common approximate distribution of the true normalized residuals, the criterion function has a finite mean $\bar{\rho} := \mathbb{E}\rho(e^*) < \infty$. For example, the mean of Andrews' (bounded) wave function is finite under any distribution.

Given a criterion function for the regression, we then estimate the change point from the transient regime to the asymptotic regime as follows. Given some arbitrary $c > \bar{\rho}$, define the criterion function difference $X_i = \rho(e_i^*) - c$ so that, in the asymptotic regime, $\mathbb{E}X = \bar{\rho} - c < 0$. For any given estimate $\beta = \hat{\beta}$, define the estimated normalized residuals $\widehat{e_i^*} = (y_i - \widehat{\beta})/s_i$,

$$k_* := \underset{k=0,\ldots,n}{\arg\min} \left\{ \min_{\widehat{\beta}} \sum_{i=k+1}^{n} [\rho(\widehat{e_i^*}) - c] \right\}, \tag{4}$$

and

$$\widehat{\beta}_* = \underset{\widehat{\beta}}{\arg\min} \sum_{i=k^*+1}^{n} [\rho(\widehat{e_i^*}) - c], \tag{5}$$

where empty sums are 0 as usual. [Note: The computation of $\hat{\beta}_*$ in Eq. (5) is already implicit in the computation of k_* in Eq. (4).]

Intuitively, Eq. (4) extends the asymptotic regime as far as is sensible, and then Eq. (5) minimizes the sum of the criterion function differences within the extended region. A good experimental design should yield many points in the asymptotic regime, yielding a value of k_* much less than n. On the other hand, untoward results, such as $k_* = n - 1$ or $k_* = n$

indicate a poor experimental design (by leaving k_* and $\hat{\beta}_*$ undefined). We refer to Eqs. (4) and (5) as single-sided asymptotic regression.

In some data sets (e.g., some of the simulation data in Sec. IV), sampling for extremely large values of the asymptotic control variable t can be sparse, so the true normalized residuals e_i^* no longer comply with the standard normal approximation in a noncompliant regime. (See Fig. 3.) Asymptotic regression can then augment the single change point described above with a second change point, one marking the time t at which noticeable deviations from the normal approximation start to occur. Define

$$(k_*, k^*) := \underset{0 \le k_* \le k^* \le n}{\arg\min} \left\{ \min_{\widehat{\beta}} \sum_{i=k_*+1}^{k^*} [\rho(\widehat{e_i^*}) - c] \right\}, \tag{6}$$

and

$$\widehat{\beta} = \underset{\widehat{\beta}}{\arg\min} \sum_{i=k_*+1}^{k^*} [\rho(\widehat{e_i^*}) - c], \tag{7}$$

where $\hat{\beta}$ again is implicit in the computation of k_*, k^* in Eq. (6). We refer to Eqs. (6) and (7) as double-sided asymptotic regression.

The selection of parameter c is a compromise between contamination with too many points in the transient regime (c too large) and exclusion of too many points in the asymptotic regime (c too small). For a bounded criterion function ρ with finite supremum $\hat{\rho} := \sup \rho < \infty$, the selection $c = \frac{1}{2}(\widehat{\rho} + \overline{\rho})$ is reasonable. In the transient regime, $X_i = \rho(e_i^*) - c$ typically approximates $\widehat{X} := \widehat{\rho} - c = \widehat{\rho} - \frac{1}{2}(\widehat{\rho} + \overline{\rho}) = \frac{1}{2}(\widehat{\rho} - \overline{\rho}) > 0$, whereas, in the asymptotic regime, X_i has the expectation,

$$\mathbb{E}X_i = \mathbb{E}\rho(e_i^*) - \tfrac{1}{2}(\widehat{\rho} - \overline{\rho}) = \overline{\rho} - \tfrac{1}{2}(\widehat{\rho} + \overline{\rho}) = -\widehat{X}, \tag{8}$$

so typical values of X_i in the transient and asymptotic regimes have similar magnitudes but opposite signs.

In the case of the (unbounded) least-squares criterion function $\rho(e^*) = \frac{1}{2}(e^*)^2$, data are often considered typical if they lie within 2 standard deviations of the corresponding regression estimate. Thus, for $\rho(e^*) = \frac{1}{2}(e^*)^2$, the cutoff $c = 2$ in Eqs. (4) –(7) is a reasonable default value.

B. The linear asymptotic regression model $y = \beta_0 + \beta_1 x + \varepsilon + f(x)$ as $t = x \to \infty$

Our methods extend to a linear asymptotic regression model in a straightforward manner, even though the asymptotic control variable t here is the explanatory variable x. Equations (4)–(7) remain the same, except that the normalized residuals are $\widehat{e_i^*} := (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)/s_i$. We mention the linear asymptotic regression model here, only to make the reader aware that we implemented it in a computer program.

C. Robust regression

The robust regression method is an alternative to the least-squares regression when outliers contaminate the data or the random errors ε are not Gaussian [53]. In the notation above, for any given initial estimate $(\hat{\beta}_0,\hat{\beta}_1)$ and criterion function $\rho(e^*)$, we get

$$(\widehat{\beta}_0,\widehat{\beta}_1)=\arg\min_{\widehat{\beta}_0,\widehat{\beta}_1}\sum_{i=1}^{n}\rho(\widehat{e}_i^*). \tag{9}$$

Least-squares regression corresponds to the criterion function $\rho(e^*)=\frac{1}{2}(e^*)^2$. All robust regressions in this paper were performed with the Andrews function $\rho(e^*)$, defined above.

D. Implementation

We implemented the C++ code for the constant and linear asymptotic regression models above, using the least-squares criterion function $\rho(e^*)=\frac{1}{2}(e^*)^2$. The software can be found at the reference URL [54].

IV. RESULTS

The asymptotic response variable $y = \beta$ in the constant asymptotic regression model pertains to λ in Eq. (2) and K in Eq. (3) . For the data pertinent to $\beta = \lambda$ in Fig. 2, the control variable t on the x axis is actually the subsequence length n; for the data pertinent to $\beta = K$ in Fig. 3, it is actually the integer score s. The simulations also yielded error bars for Figs. 2 and 3 as described elsewhere (Refs. [24,37], respectively). To compare asymptotic regression with another regression method, we also used robust regression with the Andrews wave criterion function to estimate the parameters. If robust regression assigned a non-0 weight to a point, we considered the point to be part of the asymptotic regime for robust regression. To compress the presentation in Tables I and II, parameter a in the Andrews wave function was set equal to cutoff c in asymptotic regression. (Note: Robust regression cannot use cutoff a = 0.)

Figure 2 indicates that, as t increases, the estimate of y approaches the asymptotic value with exponential rapidity, but unfortunately, also with a progressive increase in the sampling error. For different values of cutoff c and parameter a, Table I compares estimates of y from robust regression with estimates from single-sided asymptotic regression.

Figure 2 of the present paper corresponds to Fig. 2 of Ref. [24], where we (also the authors of Ref. [24]) applied robust regression to a constant model. Table I shows that estimates of $\beta = \lambda$ from robust regression of the constant model, which were sensitive to the choice of parameter a, tended to overestimate $\beta = \lambda$, and were usually of inferior accuracy when compared to single-sided asymptotic regression.

Figure 3 indicates that, as t increases, the accuracy of the estimate of $\beta = K$ at first improves [as the bias from the transient function $f(t)$ decreases] but then degrades (as sparse sampling causes both sampling error and bias to increase). For different values of cutoff c and parameter a, Table II compares estimates of $\beta = K$ from robust regression with estimates from double-sided asymptotic regression. Because of large random fluctuations in the estimates y_i of $\beta = K$ at large values of t, robust regression did not yield a connected asymptotic regime. Therefore, its asymptotic regime is given as a complicated union of disjoint intervals, highlighting the intrinsic unsuitability of applying robust regression to determine an asymptotic regression model.

Estimates of $\beta = K$ from robust regression of the constant model, which were sensitive to the choice of parameter a, tended to underestimate $\beta = K$ and usually were of inferior accuracy when compared to double-sided asymptotic regression.

V. DISCUSSION

This paper applies change-point analysis to asymptotic regression. Its methods bear some resemblance to change-point regression and robust regression, but it tailors its regression methods specifically to the problem of estimating asymptotic parameters and defining an asymptotic regime. Unlike change-point regression, it requires no specific model for the transient function $f(t)$, and unlike robust regression, it exploits the connectedness of the asymptotic regime. In fact (as demonstrated in Table II), because robust regression considers outliers one by one, it sometimes gives no precise indication of where the asymptotic regime begins. On the other hand, asymptotic regression delimits the boundary of the asymptotic regime by locating the point where the bias from the transient function begins to lose itself in the statistical noise.

Our methods generalize readily to any asymptotic linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \mathbf{f}(t)$, where $\mathbf{f}(t)$ is a column vector whose elements $f_i(t)$ satisfy $\lim_{t\to\infty} f_i(t) = 0$ for $i = 1, \ldots, n$. The variables $\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ in the asymptotic linear regression model retain their conventional roles, as in the ordinary linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$: $\mathbf{y}$ is an n-dimensional column vector representing n independent samples of a response variable y; $\mathbf{X}$ is an $n \times k$ matrix representing k explanatory variables in each of the n independent samples; and $\boldsymbol{\beta}$ is a k-dimensional column vector representing the regression coefficients. In the asymptotic linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \mathbf{f}(t)$, however, the ordinary linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is pertinent only in the asymptotic regime, i.e., in the limit $t \to \infty$. The transient function $\boldsymbol{f}(t)$ is relatively unrestricted: It might be random or deterministic; a function of $\mathbf{y}$, $\mathbf{X}$, or $\boldsymbol{\beta}$ or not, etc. Moreover, the asymptotic control variable t might easily be one of the explanatory variables in model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (a case occurring in many applications).

Our simple objective change-point method can replace the subjective censoring of data that occurs when estimating an asymptotic regime by eye [49,50]. The method has particularly broad applications in physics, where it will increase confidence in the scientific objectivity of estimating asymptotic parameters.

Acknowledgments

The authors would like to acknowledge helpful discussions with Yi-Kuo Yu. This research was supported by the Intramural Research Program of the NIH, NLM.

References

1. Chia N, Bundschuh R. Phys Rev E. 2004; 70:021906.
2. Hartmann AK. Phys Rev E. 2002; 65:056102.
3. Yeramian E, Debonneuil E. Phys Rev Lett. 2007; 98:078101. [PubMed: 17359063]
4. Wolfsheimer S, Melchert O, Hartmann AK. Phys Rev E. 2009; 80:061913.
5. Sardiu ME, Alves G, Yu YK. Phys Rev E. 2005; 72:061917.
6. Bundschuh R. Phys Rev E. 2002; 65:031911.
7. Uchiyama M, Sasamoto T, Wadati M. J Phys A. 2004; 37:4985.
8. De Los Rios P, Zhang YC. Phys Rev Lett. 1998; 81:1023.
9. Hwa T, Lässig M. Phys Rev Lett. 1996; 76:2591. [PubMed: 10060738]
10. Huse DA, Henley CL. Phys Rev Lett. 1985; 54:2708. [PubMed: 10031417]
11. Kardar M. Nucl Phys B. 1987; 290:582.

12. Fisher DS, Huse DA. Phys Rev B. 1991; 43:10728.
13. Wolfsheimer S, Melchert O, Hartmann AK. Phys Rev E. 2009; 80:061913.
14. Yu YK. Phys Rev E. 2004; 69:061904.
15. Kardar M, Parisi G, Zhang YC. Phys Rev Lett. 1986; 56:1810. [PubMed: 10033551]
16. Halpin-Healy T, Zhang YC. Phys Rep. 1995; 254:215.
17. Huse DA, Henley CL, Fisher DS. Phys Rev Lett. 1985; 55:2924. [PubMed: 10032275]
18. Burgers, JM. The Nonlinear Diffusion Equation: Asymptotic Solutions and Statistical Problems. Reidel; Boston: 1974.
19. Smith TF, Waterman MS. J Mol Biol. 1981; 147:195. [PubMed: 7265238]
20. Altschul SF, et al. J Mol Biol. 1990; 215:403. [PubMed: 2231712]
21. Altschul SF, et al. Nucleic Acids Res. 1997; 25:3389. [PubMed: 9254694]
22. Schäffer AA, et al. Nucleic Acids Res. 2001; 29:2994. [PubMed: 11452024]
23. Needleman SB, Wunsch CD. J Mol Biol. 1970; 48:443. [PubMed: 5420325]
24. Park Y, Sheetlin S, Spouge JL. J Phys A. 2005; 38:97.
25. Park Y, Spouge JL. Bioinformatics. 2002; 18:1236. [PubMed: 12217915]
26. Arratia R, Waterman MS. Ann Probab. 1985; 13:1236.
27. Aldous, D. Probability Approximations Via the Poisson Clumping Heuristic. 1. Springer-Verlag; New York: 1989.
28. Galombos, J. The Asymptotic Theory of Extreme Order Statistics. 1. Wiley; New York: 1978.
29. Olsen, R.; Bundschuh, R.; Hwa, T. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology; Menlo Park, CA: AIAA; 1999. p. 211-222.
30. Henikoff S, Henikoff JG. Proc Natl Acad Sci USA. 1992; 89:10915. [PubMed: 1438297]
31. Altschul SF, et al. Nucleic Acids Res. 2001; 29:4793. [PubMed: 11726688]
32. Spouge, JL. [http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/htmlncbi/blast/blastdisplay.html?5]
33. Altschul SF, Gish W. Methods Enzymol. 1996; 266:460. [PubMed: 8743700]
34. Spouge JL. J Math Anal Appl. 2005; 301:401.
35. Spouge JL. J Appl Probab. 2001; 38:1.
36. [http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/index/software.html#6].
37. Sheetlin S, Park Y, Spouge JL. Nucleic Acids Res. 2005; 33:4987. [PubMed: 16147981]
38. Stevens WL. Biometrics. 1951; 7:247.
39. Hassell MP, Lawton JH, Beddington JR. J Anim Ecol. 1977; 46:249.
40. Jumbe N, et al. J Clin Invest. 2003; 112:275. [PubMed: 12865415]
41. Seeger R, Pople JA. J Chem Phys. 1976; 66:3045.
42. Pandey RB, Anderson KL, Farmer BL. J Polym Sci, Part B: Polym Phys. 2005; 43:1041.
43. Silvestri L, Fronzoni L, Grigolini P, Allegrini P. Phys Rev Lett. 2009; 102:014502. [PubMed: 19257199]
44. Garrido PL, Hurtado PI, Nadrowski B. Phys Rev Lett. 2001; 86:5486. [PubMed: 11415282]
45. Dettmann CP, Frankel NE. Phys Rev E. 1996; 53:5502.
46. Altschul SF, et al. Nucleic Acids Res. 2001; 29:4793. [PubMed: 11726688]
47. Park Y, Sheetlin S, Spouge JL. Ann Stat. 2009; 37:3697. [PubMed: 20148197]
48. Bundschuh R. J Comput Biol. 2002; 9:243. [PubMed: 12015880]
49. Moses RA. Ophthalmol Vis Sci. 1983; 24:1079.
50. Benetatos C, et al. Ann Geophys. 2002; 45:513.
51. Julious SA. J R Stat Soc Ser, D, Stat. 2001; 50:51.
52. Clauset A, Shalizi CR, Newman MEJ. SIAM Rev. 2009; 51:661.
53. Montgomery, DC.; Peck, EA.; Vining, GG. Introduction to Linear Regression Analysis. Wiley; New York: 2001.
54. [http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/index/software.html#3].

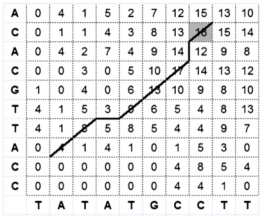

FIG. 1.

Local alignment. Gapped local alignment scores and the corresponding directed paths for two subsequences **A**(0,10] = TATATGCCTT and **B**(0,10] = CCATTGCACA of sequences drawn from the nucleotide alphabet {A,C,G,T}. The nucleotide scoring matrix in the figure is commonly used in applications: $S(a,b) = 4$ if $a = b$ and -5 otherwise and the affine gap penalty $\Delta(g) = 2 + g$. The vertex $(i,j) = (10,10)$ is in the northeast corner of the figure with the origin (0,0) at the southwest corner. The cell (i,j) corresponding to the letter pair (A_i,B_j) displays the corresponding local score $\hat{S}_{i,j}$. A local score of 0 indicates that no path of positive weight ends at the corresponding point. The optimal local path, which ends at point (8,9), e.g., consists of eight edges, in order: 2 northeast, 1 east, 3 northeast, 1 north, and 1 northeast. It corresponds to the optimal local sequence alignment of **A**(0,10] and **B**(0,10],

$$\begin{array}{cccccccc} A & T & A & T & G & C & - & C \\ A & T & - & T & G & C & A & C \end{array}.$$

The local score $\hat{S}_{8,9} = 4 + 4 - 3 + 4 + 4 + 4 - 3 + 4 = 18$ is the sum of the corresponding edges and represents the path of greatest weight ending at point (8,9). The score is also the greatest weight of any path within the rectangle displayed, so it also corresponds to the local maximum score $\hat{M}_{10,10} = 18$.

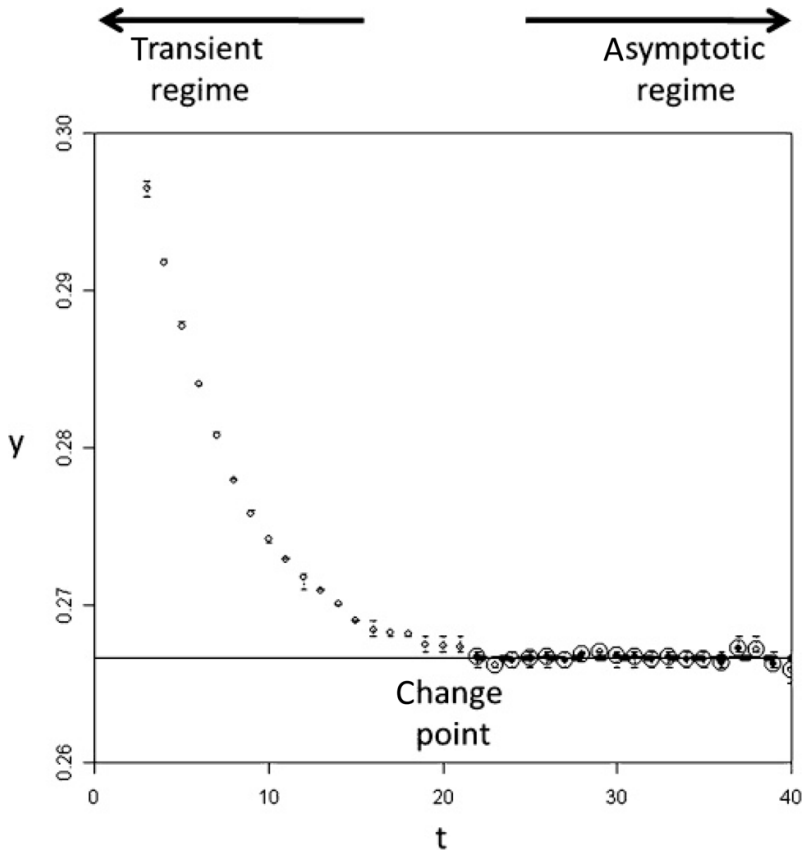

FIG. 2.

Single-sided asymptotic regression. The figure indicates a typical plot relevant to single-sided asymptotic regression, with the transient regime on the left, the asymptotic regime on the right, and an ill-defined change point separating the two regimes. Once in the asymptotic regime, the data point approximates a horizontal line $y = \beta$. The rest of this caption is pertinent only to the Results section, as follows. This figure displays estimates y_i for $\beta = \lambda_{n-5,n}$ from Eq. (2) against sequence length $t = n$ as small open circles (○) with error bars. (Some error bars are small and are not readily visible). The error bars indicate standard errors from 1 000 000 random sequence pairs. The solid horizontal line shows the present best estimate $\beta = \lambda \approx 0.2666 \pm 0.0003$ [32]. For single-sided asymptotic regression, data in the asymptotic regime are indicated by large open circles (○); the estimated $\hat{\beta}$ is indicated by a thick dotted horizontal line (indistinguishable from the solid horizontal line). For robust regression with Andrews' wave function, data in the asymptotic regime are indicated by closed circles (●); the estimated $\hat{\beta}$ is indicated by a dashed horizontal line (barely distinguishable from the solid horizontal line). The values $c = 2$ and $a = 0.5$ were chosen for single-sided asymptotic regression and robust regression with Andrews' wave function, respectively, because they minimized the absolute difference between the estimated β and the present best estimate $\beta = \lambda \approx 0.2666 \pm 0.0003$. (Table I summarizes the different estimates of β from different cutoff values c and parameters a.)

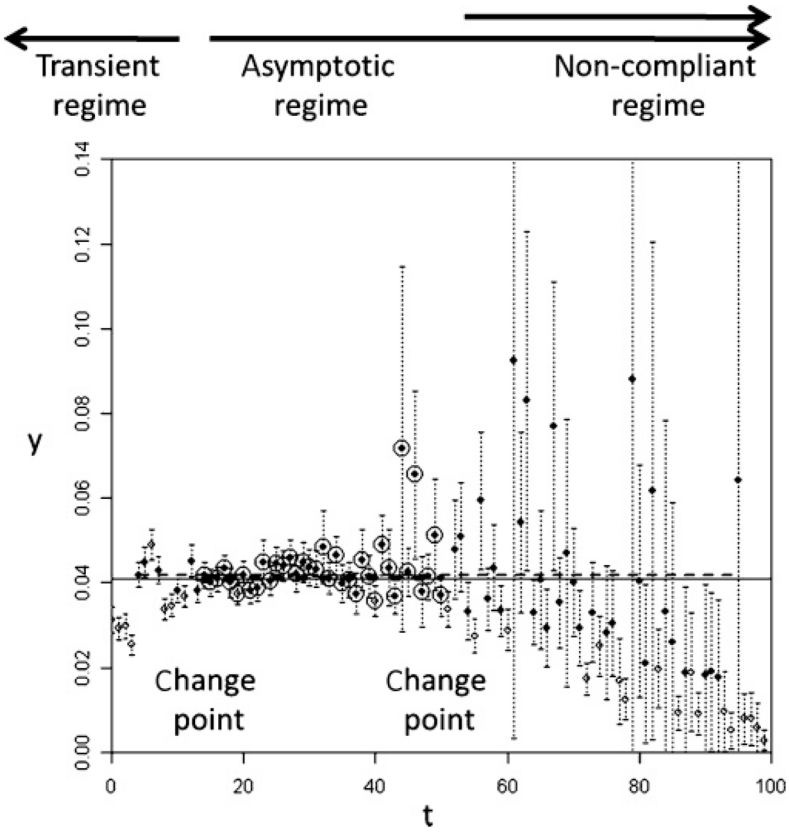

FIG. 3.

Double-sided asymptotic regression. The figure indicates a typical plot relevant to double-sided asymptotic regression, with the transient regime on the left, the asymptotic regime on the right (which is eventually overtaken by a noncompliant regime where the statistical model fails), and two ill-defined change points separating the three regimes. The rest of this caption is pertinent only to the Results section, as follows. This figure displays estimates y_i for $\beta = K$ from Eq. (3) against global score $t = s$ as small open circles (○) with error bars. The error bars indicate standard errors from 1 000 000 random sequence pairs. The solid horizontal line shows the present best estimate $\beta = K \approx 0.0409 \pm 0.0003$ [32]. For double-sided asymptotic regression, data in the asymptotic regime are indicated by large open circles (○); the estimated $\hat{\beta}$ is indicated by a thick dotted horizontal line (barely distinguishable from the solid horizontal line). For robust regression with Andrews' wave function, data are indicated by closed circles (●); the estimated $\hat{\beta}$ is represented by a dashed horizontal line. (Note that, in fact, robust regression outliers break up the single asymptotic interval indicated by the dashed line.) For each method, $c = 0.5$ and $a = 0.5$ minimized the absolute difference between the estimated K and the present best estimate $\beta = K \approx 0.0409 \pm 0.0003$. (Table II summarizes the different estimates of K from different cutoff values c and parameters a.)

TABLE I

Estimates of $\beta = \lambda$ from robust regression and from single-sided asymptotic regression. The first column is either the cutoff value c for double-sided asymptotic regression or a for the Andrews wave function in robust regression. The second and fourth columns give the data range selected as the asymptotic regime; the third and fifth columns give the estimates of $\beta = \lambda$.

	Asymptotic regression $y = \beta + \varepsilon$		**Robust regression with Andrews' function $y = \beta + \varepsilon$**	
c* or *a	**Asymptotic regime**	**$\lambda_{estimate}$**	**Asymptotic regime**	**$\lambda_{estimate}$**
0	[39,40]	0.2661 ± 0.0003	N/A	N/A
0.5	[24,40]	0.2667 ± 0.0001	{22}U[24,28]U[30,37]U{39}	0.2666 ± 0.0001
1	[22,40]	0.2666 ± 0.0001	{19}U[21,40]	0.2667 ± 0.0001
2	[22,40]	0.2666 ± 0.0001	[16,40]	0.2668 ± 0.0001
4	[21,40]	0.2667 ± 0.0001	[14,40]	0.2671 ± 0.0001
10	[18,40]	0.2668 ± 0.0001	[11,40]	0.2676 ± 0.0001

TABLE II

Estimate of $\beta = K$ from robust regression and from double-sided asymptotic regression. The first column is either the cutoff value c for double-sided asymptotic regression or a for the Andrews wave function in robust regression. The second and fourth columns give the data range for each method selected as the asymptotic regime; the third and fifth columns give the estimates of $\beta = K$. Robust regression requires that the asymptotic regime be expressed as a complicated union of sets and intervals, indicating its unsuitability for defining the regime.

	Asymptotic regression $y = \beta + \varepsilon$		Robust regression with Andrews' function $y = \beta + \varepsilon$	
c or a	Asymptotic regime	$K_{estimate}$	Asymptotic regime	$K_{estimate}$
0	[87,88]	0.0188 ± 0.0116	N/A	N/A
0.5	[14,50]	0.0411 ± 0.0007	[4,5]U{7}U{10}U[12,18] U[20,39]U[41,50]U[52,54]U[56, 59]U[61,71]U{73}U[75,76]U[79,82] U[84,85]U{87}U[90,92]U{95}	0.0416 ± 0.0007
1	[10,71]	0.0398 ± 0.0006	{0}U[4,54]U[56,71]U[73,77] U[79,87]U[87,88]U[90,92]U{95}	0.0400±0.0005
2	[4,71]	0.0398 ± 0.0005	[0,85]U[87,93]U[95,98]	0.0384 ± 0.0005
4	[4,85]	0.0388 ± 0.0005	[0,98]	0.0364 ± 0.0005
10	[0,93]	0.0366 ± 0.0005	[0,99]	0.0349 ± 0.0005