SORT 31 (1) January-June 2007, 55-74

Parameter estimation of S-distributions with alternating regression

I-Chun Chou[1], Harald Martens[2], Eberhard O. Voit[1,*]

[1] *Georgia Institute of Technology and Emory University*
[2] *CIGENE/IKBM, Norwegian U. of Life Sciences*

Abstract

We propose a novel *3-way alternating regression* (3-AR) method as an effective strategy for the estimation of parameter values in S-distributions from frequency data. The 3-AR algorithm is very fast and performs well for error-free distributions and artificial noisy data obtained as random samples generated from S-distributions, as well as for traditional statistical distributions and for actual observation data. In rare cases where the algorithm does not immediately converge, its enormous speed renders it feasible to select several initial guesses and search settings as an effective countermeasure.

MSC: 62G05, 62E17, 62J02, 62J05.

Keywords: Alternating regression, Parameter estimation, S-distribution, S-system.

1 Introduction

Motivated by a distribution family based on S-systems (Savageau, 1982), the S-distribution was introduced in the early 1990s as a convenient univariate, unimodal four-parameter distribution that is capable of modelling a wide range of shapes and skewness (Voit, 1992). Due to its rich shape flexibility and relatively simple mathematical

* *Address for correspondence:* I-Chun Chou, Eberhard O. Voit. The Wallace H. Coulter Department of Biomedical Engineering at Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA, 30332, U.S.A. E-mail: gtg392p@mail.gatech.edu, Eberhard.Voit@bme.gatech.edu. Harald Martens. CIGENE/IKBM, Norwegian U. of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway. Email: harald.martens@matforsk.no.
Received: January 2007
Accepted: April 2007

format, the S-distribution has been shown to constitute a good general-purpose default distribution, especially for data of unknown structure. The S-distribution may also be used in lieu of the traditional distributions, because it always has the same structure and, with an appropriate choice of parameter values, rather accurately approximates many continuous central and non-central distributions, as well as a wide variety of discrete distributions (Voit, 1992; Voit and Yu, 1994; Yu and Voit, 1996). In addition, the S-distribution allows for combinations of parameter values that do not correspond to traditional distributions and permits a spectrum of distributions with long or heavy tails and with skewness to the left or right. Thus, one might in many cases expect a better fit than is possible with traditional distributions. As a specific application of the combination of its flexibility and small number of parameters, the S-distribution is well suited for the non-trivial characterization of trends of distributions that change mean, variance, shape, and even skewness over time (Voit, 1996; Sorribas, March and Voit, 2000; Voit and Sorribas, 2000).

The S-distribution is formulated as a differential equation, which renders the estimation of parameter values from data a challenge. Several methods have been suggested for this task, including nonlinear regression (Voit, 1992; Sorribas, March and Voit, 2000), a graphical method (Voit, 1992), constrained maximum likelihood estimation (Voit, 2000), and techniques based on quantiles (Voit and Schwacke, 2000; Hernández-Bermejo and Sorribas, 2001). Here, we propose an entirely different method called *3-way alternating regression* (3-AR), which was motivated by a 2-way alternating regression method used for the estimation of parameters in multivariate S-systems (Chou, Martens and Voit, 2006). The main appeal of 3-AR is its enormous speed and robustness. In this article, we discuss the method and apply it to several artificial and actual examples.

2 S-distribution

The S-distribution is a four-variable distribution that emphasizes the cumulative density function (*cdf*) F, which is formulated as a differential equation with respect to random variable X and reads

$$f = \frac{dF}{dX} = \alpha\left(F^g - F^h\right), \qquad F_0 = F(X_0) \in (0,1). \tag{1}$$

Because the probability density function (*pdf*) f is the derivative of F, the S-distribution can be seen as an algebraic function $f(F)$. The first parameter of the distribution, X_0, characterizes the location of the distribution. The second parameter, α, is a positive real number, which determines the scale. The remaining two parameters,

g and h, may be any real numbers as long as $g < h$; they determine the shape of the distribution.[1]

3 Alternating Regression

Suppose the S-distribution is characterized through N values of the random variable, X_1, X_2,..., X_k,..., X_N, and that X_k, $F(X_k)$ and $f(X_k)$ are observed or obtainable for each k (see later sections for further discussion on the construction of *pdf*s and *cdf*s). For the purpose of parameter estimation, the original differential equation can then be analyzed in the form of N uncoupled algebraic equations as

$$\begin{aligned} f(X_1) &\approx \alpha\left(F^g(X_1) - F^h(X_1)\right), \\ f(X_2) &\approx \alpha\left(F^g(X_2) - F^h(X_2)\right), \\ &\vdots \\ f(X_k) &\approx \alpha\left(F^g(X_k) - F^h(X_k)\right), \\ &\vdots \\ f(X_N) &\approx \alpha\left(F^g(X_N) - F^h(X_N)\right). \end{aligned} \tag{2}$$

The $\approx$ symbol is used because the data may only be representable in approximation by the S-distribution format. As a consequence of this decoupling step, substitution of the derivative of F with f allows us to estimate the S-distribution parameters α, g, and h in a purely algebraic system (*cf.* Voit and Almeida, 2000). We propose for this estimation purpose a new method called *3-way alternating regression* (3-AR).

In previous work, we have shown that alternating regression (AR), applied to S-system models of the form

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}, \; i = 1, 2, ..., n, \tag{3}$$

and combined with methods for slope estimation and decoupling systems of differential equations, provides a fast tool for identifying parameter values from time series data (Chou, Martens and Voit, 2006). The key feature of AR is the reduction of the nonlinear inverse problem of parameter estimation into iterative steps of two phases of linear regression. In the first phase, the parameters of the β-term, β_i and h_{ij}, are set to some reasonable values. Given measurements of all X_i at N time points and estimates $S_i(t_k)$ of

1. Throughout the paper, random variables and *cdf*s are represented as upper-case italics, while *pdf*s are given by the corresponding lower-case italic symbols (X, F, f). An upper-case boldface variable (**L**) represents a matrix of regressor columns and a lower-case boldface variable (**y**) represents a regressand column in a linear statistical regression model.

the slope of X_i at these points, the β-term becomes a number at each time point, and this number is added to both sides of Equation (3). Taking the logarithm of the equation for each time point, one obtains a linear regression problem with the slope and the β-term as a real number on the left-hand side, and a linear expression on the right hand side:

$$\log\left(S_i(t_k) + \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}(t_k)\right) \approx \log(\hat{\alpha}_i) + \sum_{j=1}^{n} \hat{g}_{ij} \log(X_j(t_k)) + \varepsilon_{i,k} \tag{4}$$

The regression with the N equations of this type at time points t_k now yields estimates for α_i and all g_{ij}. In the second step of AR, these estimates are used in an analogous fashion to compute β_i and h_{ij}. The algorithm switches back and forth and usually converges fast (see Chou, Martens and Voit (2006) for details).

The S-distribution is obviously a special case of an S-system, with the notable feature that by definition $\alpha = \beta$. This feature is important for AR methods, because α and β are no longer independent of each other, and it turns out to be inconvenient to constrain α to be the same in both phases of the regression. Therefore, we modify the 2-way AR approach here into a three-cycle 3-AR method specifically for S-distribution estimation.

Similar to the original AR, 3-AR works by iteratively cycling between phases of linear regression. The first phase begins with guesses of the values of g and h and uses these to solve for the value of parameter α. Experience has shown that it is more expedient to start the algorithm with g and h, rather than g and α or h and α, presumably due to the fact that the typical ranges of g and h are much smaller than that of α and because h is per definition constrained by g. The second phase takes estimates of α and h to solve for g, while the third phase takes estimates of α and g to solve for h and thus improve the parameter guesses or estimates from the previous phases. The phases are iterated until a solution is found or AR terminates for other reasons. The overall flow of the method is shown in Figure 1, and specific steps of the 3-AR algorithm are detailed below.

Steps of the 3-AR Algorithm

{1} Define $\mathbf{L_f}$ and $\mathbf{L_F}$ as $2 \times N$ matrices of logarithms of regressors f and F, respectively:

$$\mathbf{L_f} = \begin{bmatrix} 1 & \log(f(X_1)) \\ 1 & \log(f(X_2)) \\ \vdots & \vdots \\ 1 & \log(f(X_k)) \\ \vdots & \vdots \\ 1 & \log(f(X_N)) \end{bmatrix} \tag{5}$$

$$\mathbf{L_F} = \begin{bmatrix} 1 & \log(F(X_1)) \\ 1 & \log(F(X_2)) \\ \vdots & \vdots \\ 1 & \log(F(X_k)) \\ \vdots & \vdots \\ 1 & \log(F(X_N)) \end{bmatrix} \tag{6}$$

$\mathbf{L_f}$ is used in the first phase of AR to determine α, and $\mathbf{L_F}$ is used in the second and third phases of AR to determine g and h.

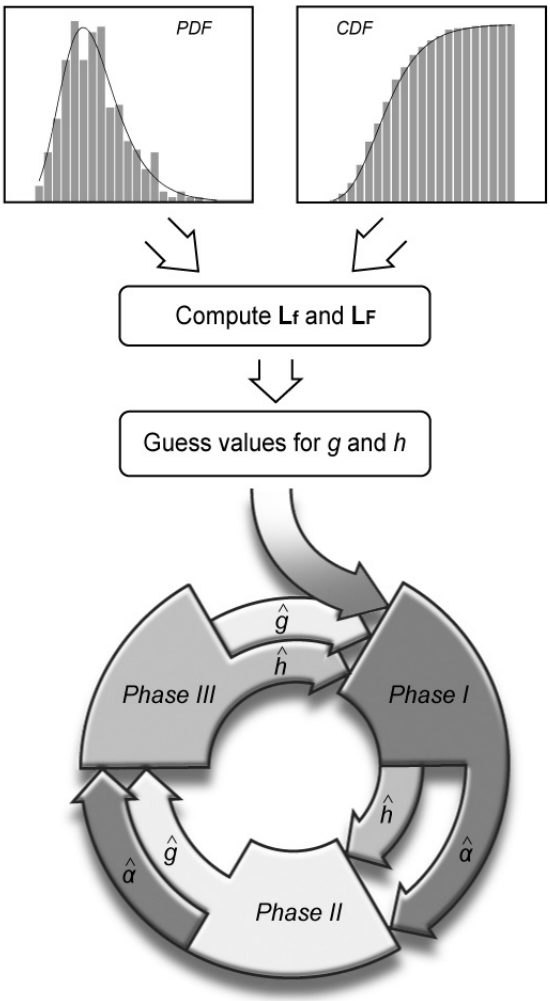

Figure 1: *Flow of parameter estimation by 3-way alternating regression.*

{2} Select values for g and h in accordance with experience about S-distribution parameters (see Voit (1992) for relationships between parameter values and distributional shape).

{3} For all X_k, $k = 1, 2, \ldots, N$, compute $F^{\hat{g}}(X_k) - F^{\hat{h}}(X_k)$, using values $F(X_k)$ from the data distribution. Here $\hat{g}$ and $\hat{h}$ denote the estimators of g and h after the 2^{nd} iteration, while during the 1^{st} iteration, $\hat{g}$ and $\hat{h}$ are the initial guesses for g and h, respectively. Determine the index I_α of all positive quantities $F^{\hat{g}}(X_k) - F^{\hat{h}}(X_k)$. The number of *qualified points* then becomes N_α, where N_α is the length of I_α. Quantities restricted to N_α instead of all N points are identified in the following with an additional subscript α. Note: Theoretically $F^g(X_k)$ should always be greater than $F^h(X_k)$, because $g < h$, or at most equal, for $F = 0$ and $F = 1$. However, because of noise, this may not always be true, suggesting temporary exclusion of some data points.

{4} After logarithmic transformation and rearrangement, Equation (1) can be written as $\log\left(\frac{f}{\alpha}\right) = \log\left(F^g - F^h\right)$. Therefore, compute the N_α-dimensional vector $\mathbf{y_\alpha} = \log\left(F_\alpha^{\hat{g}} - F_\alpha^{\hat{h}}\right)$ for N_α points, as well as $\mathbf{L_{f_\alpha}}$, where the subscript α limits the computation to qualified points.

{5} Based on the linear regression model

$$\mathbf{y_\alpha} = \mathbf{L_{f_\alpha}}\hat{\mathbf{b}}_\alpha + \boldsymbol{\varepsilon}_\alpha, \tag{7}$$

estimate the regression coefficient vector $\hat{\mathbf{b}}_\alpha = [\hat{b}_{\alpha_1}, \hat{b}_{\alpha_2}]^{\mathrm{T}}$ over the N_α qualified points, to obtain an estimate of α. In other words, this equation may be written as $\mathbf{y_\alpha} \approx \log\left(\frac{1}{\hat{\alpha}}\right) + \log(f_\alpha) + \varepsilon_\alpha$ so that $\hat{b}_{\alpha_1}$ is equivalent to $\log\left(\frac{1}{\hat{\alpha}}\right)$ and $\hat{b}_{\alpha_2}$ is the coefficient of $\log(f_\alpha)$, which is expected to converge to 1. Thus, $\hat{\mathbf{b}}_\alpha$ is estimated with any of the methods of linear regression, *e.g.*, by ordinary least squares regression (OLSR) as

$$\hat{\mathbf{b}}_\alpha = \left(\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha}\right)^{-1}\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{y_\alpha}. \tag{8}$$

As an alternative to OLSR, weighted or robust estimators could be used. If $\mathbf{L_{f_\alpha}}$ does not have full column rank, *i.e.*, if $\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha}$ has a small eigenvalue, one could also use a small ridge regression constant κ for stabilization and compute $\hat{\mathbf{b}}_\alpha$ as

$$\hat{\mathbf{b}}_\alpha = \left(\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{L}_{\mathbf{f}_\alpha} + \kappa\mathbf{I}\right)^{-1}\mathbf{L}_{\mathbf{f}_\alpha}^{\mathrm{T}}\mathbf{y_\alpha}. \tag{9}$$

{6} For the estimation of g, reformulate Equation (1) as $\frac{f}{\alpha} + F^h = F^g$. Thus, using values of $f(X_k)$ and $F(X_k)$ that are directly obtained from the data (see later sections), compute $\frac{f(X_k)}{\hat{\alpha}} + F^{\hat{h}}(X_k)$ for all X_k, $k = 1, 2, \ldots, N$. Here $\hat{h}$ denotes the estimator of h after the 2^{nd} iteration, while during the 1^{st} iteration, $\hat{h}$ is the initial guess for h. Find the index I_g of

positive quantities $\frac{f(X_k)}{\hat{\alpha}} + F^{\hat{h}}(X_k)$. The number of qualified points for this step becomes N_g, where N_g is the length of I_g.

{7} Compute the N_g-dimensional vector $\mathbf{y_g} = \log\left(\frac{f_g}{\hat{\alpha}} + F_g^{\hat{h}}\right)$ for N_g points, as well as $\mathbf{L_{F_g}}$.

{8} Based on the linear regression model

$$\mathbf{y_g} = \mathbf{L_{F_g}}\hat{\mathbf{b}}_\mathbf{g} + \boldsymbol{\varepsilon}_\mathbf{g}, \tag{10}$$

and in analogy to step {5}, estimate the regression coefficient vector $\hat{\mathbf{b}}_\mathbf{g} = [\hat{b}_{g_1}, \hat{b}_{g_2}]^{\mathrm{T}}$ by regression over the N_g time points as

$$\hat{\mathbf{b}}_\mathbf{g} = \left(\mathbf{L}_{\mathbf{F_g}}^{\mathrm{T}}\mathbf{L}_{\mathbf{F_g}}\right)^{-1}\mathbf{L}_{\mathbf{F_g}}^{\mathrm{T}}\mathbf{y_g}, \tag{11}$$

or with an alternative regression method. The estimator $\hat{b}_{g_2}$ is the parameter of interest, $\hat{g}$; estimator $\hat{b}_{g_1}$ is expected to be zero in the model.

{9} For the estimation of h, reformulate Equation (1) as $F^g - \frac{f}{\alpha} = F^h$ and compute $F^{\hat{g}}(X_k) - \frac{f(X_k)}{\hat{\alpha}}$ for all X_k, $k = 1, 2, \ldots, N$, again using the values of $f(X_k)$ and $F(X_k)$. Determine the index I_h of positive quantities $F^{\hat{g}}(X_k) - \frac{f(X_k)}{\hat{\alpha}}$. The number of qualified points for this step becomes N_h, where N_h is the length of I_h.

{10} Compute the N-dimensional vector $\mathbf{y_h} = \log\left(F_h^{\hat{g}} - \frac{f_h}{\hat{\alpha}}\right)$ for N_h points, as well as $\mathbf{L_{F_h}}$.

{11} Based on the linear regression model

$$\mathbf{y_h} = \mathbf{L_{F_h}}\hat{\mathbf{b}}_\mathbf{h} + \boldsymbol{\varepsilon}_\mathbf{h}, \tag{12}$$

and in analogy to steps {5} and {8}, estimate the regression coefficient vector $\hat{\mathbf{b}}_\mathbf{h} = [\hat{b}_{h_1}, \hat{b}_{h_2}]^{\mathrm{T}}$ by regression over the N_h time points as

$$\hat{\mathbf{b}}_\mathbf{h} = \left(\mathbf{L}_{\mathbf{F_h}}^{\mathrm{T}}\mathbf{L}_{\mathbf{F_h}}\right)^{-1}\mathbf{L}_{\mathbf{F_h}}^{\mathrm{T}}\mathbf{y_h}, \tag{13}$$

or with an alternative regression method. The estimator $\hat{b}_{h_2}$ is the parameter of interest, $\hat{h}$; estimator $\hat{b}_{h_1}$ is expected to be zero in the model.

{12} Iterate steps {3} – {11} until a solution is found or some termination criterion is satisfied.

At each phase of 3-AR, lack-of-fit criteria are estimated and used for monitoring the iterative process and to define termination conditions. We use here specifically the logarithm of the sums of squared y-errors (SSE_α, SSE_g, and SSE_h) as optimization criteria

for the three regression phases. Upon convergence, we also compute the residual error *SSE* of the fit and the standard deviation $S.D. = \sqrt{SSE/(N-p)}$ of the *pdf*, as well as the *cdf* and f-F plots, where p is the number of estimated parameters, which in all cases here is 3.

The location parameter X_0 is not explicit in the method, because it does not appear in the algebraic formulation of the *pdf* as a function of the *cdf*. However, it is easily estimated directly as the observed or estimated median or by optimizing the horizontal position of the distribution with parameters $\hat{\alpha}$, $\hat{g}$, and $\hat{h}$ (Voit, 2000).

4 Results

We tested the 3-AR method with a large number of representative cases, including estimations based on "data" from error-free distributions, artificial noisy data obtained as random samples generated from S-distributions with known parameters, traditional statistical distributions (using Matlab®), and from actual observation data. Representative details of each case are discussed in this section.

4.1 Fitting distributions without noise

In order not to confuse the features of 3-AR with possible effects of noise in the data, we begin the exploration of convergence properties by using true S-distribution *cdf*s and *pdf*s, which are evaluated directly from Equation (1) at a number of values for the random variable. Specifically, we choose 50 equally spaced instances of the random variable and compute the corresponding f and F values from Equation (1) to obtain the "true" *pdf* and *cdf*. Figure 2 shows an example of a typical convergence pattern. Starting from the (essentially arbitrary) initial guesses $g = 3$ and $h = 6$, it takes the 3-AR algorithm just 51 iterations to converge to the true solution, requiring 0.0742 seconds on a Pentium® D (~3.4GHz) machine. Since we use noise-free data, the residual error should approach 0, which corresponds to $-\infty$ in logarithmic coordinates. We use -9 instead as one of the termination criteria, which corresponds to a result very close to the true value, but allows for issues of machine precision and numerical inaccuracies. The low error tolerance causes the algorithm to need 51 iterations. However, as Figure 2 indicates, the estimates are already very close to the true optimum after just a few initial iterations. Big jumps in the beginning do not negatively affect convergence time. For instance, using the same error tolerance and initial guesses $g = 10$, $h = 10.5$ or $g = 100$ and $h = 120$, respectively, the algorithm needs 57 iterations (0.0535 second) or 63 iterations (0.0567 second) to converge to the true parameter values. Thus, somewhat different from results for general S-systems (Chou, Martens and Voit, 2006),

the speed of convergence here does not depend much on initial guesses. Also in contrast to observations with S-systems, the convergence patterns for α, g, and h are often not monotonic, and each parameter may temporarily increase or decrease during the initial iterations.

While convergence is almost always extremely fast, as in the example described above, some initial values cause 3-AR not to converge at all. In such rare cases, the value of α typically increases without bound, while g and h converge toward each other and ultimately become the same. This case corresponds to the trivial solution $\frac{f}{\alpha} \to 0 \leftarrow F^g - F^h$ in Equation (1) and is easy to detect and discard.

Figure 3 combines results for several noise-free S-distributions and essentially exhaustive sets of initial guesses for g and h satisfying $g < h$, as required. The selected distributions are representative for different shapes and skewness, which are reflected in different categories of parameter values (*cf.* Voit, 2000):

1. $g > 0$ and $h > 0$: as exemplified in Figure 3A and 3B;
2. $g < 0$ and $h > 0$: as exemplified in Figure 3C;
3. $g < 0$ and $h < 0$.

In addition, samples from all categories must by definition satisfy the condition $g < h$.

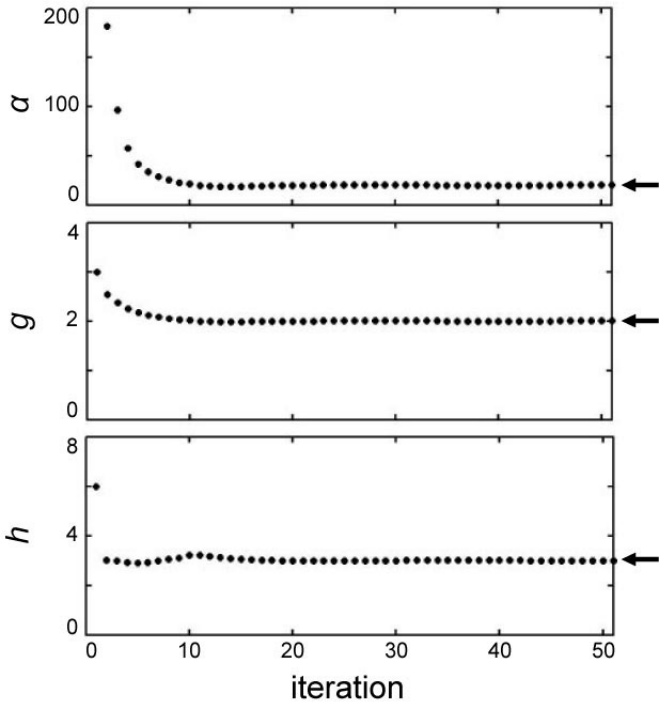

Figure 2: *Convergence pattern of 3-AR. For this example, 50 instances of the random variable were chosen from a parent distribution with parameters* $\alpha = 20$, $g = 2$, $h = 3$, *and* $F_0 = 0.01$. *Initial guesses were chosen as* $g = 3$ *and* $h = 6$, *but do not affect convergence much. No initial guess for* α *is needed in 3-AR.*

The left panels in Figure 3 exhibit the *cdf* and *pdf* of each distribution. Inserts show the so-called f-F plots, where the *pdf* is plotted against the corresponding *cdf*. These plots are important because they are the basis for 3-AR and many other estimation methods for S-distributions. The right-hand panels present "heat maps" of convergence: the x- and y-axes represent the initial guesses of h and g, respectively, and the gray bar represents the logarithm (base 10) of the number of iterations needed for convergence. Once the predetermined error level is reached, 3-AR stops and the number of iterations is recorded as a measure for the speed of convergence. In each case shown here, 25 instances of the random variable were chosen and the corresponding noise-free f and F values were obtained according to the selected random variables. Black areas represent divergence to the trivial solution $\alpha \approx \infty$, $g \approx h$.

As discussed above, the convergence time for a given distribution does not vary much with different initial guesses, and the basin of convergence within each heat map is therefore almost monochrome. However, the heat maps of different distributions are quiet different. For instance, the times needed to generate the heat maps in Figures 3A, 3B, and 3C for a total of 57,600 initial values shown are 14,957, 1,197, and 1,094 seconds on a single PC, respectively, thus yielding average convergence times of 0.26, 0.021, and 0.019 seconds per case. While reasons for the wide variations in convergence times among distributions are unclear, the convergence *patterns* are similar in all cases: 3-AR takes big steps during the first few iterations, already coming very close to the true solution, and then spends many iterations on fine-tuning. The convergence area in each case is relatively large, and it seems to be a good general strategy to choose rather large, similar initial values for g and h, such as 10 and 10.5, to avoid divergence. Of importance is that each iteration consists essentially of three linear regressions, which are very fast. Thus, even if one encounters a rare case of divergence, the choice of alternative initial settings is computationally cheap and provides for effective estimation results.

Examples with $g < 0$ and $h < 0$ or with different α values are not shown in Figure 3, but 3-AR performed in a similar fashion for all cases tested. Most of the estimation tasks were solved very effectively, except for cases where the difference between g and h is large, for instance, $g = 0.1$ and $h = 6$. In such cases, the algorithm sometimes converges to sets of values between the true g and h and oscillates between them. A possible reason for this behaviour may be that in the 3^{rd} phase of regression (estimation of h), the slope of the regression line in the $\mathbf{y_h}$-$\mathbf{L_{F_h}}$ plot (which is reflected in the high value of h) is large and greatly affected by small errors, especially when f and F values are small so that their logarithms dominate the regression. In this case, the algorithm may not converge to exactly the right solution, but the oscillation happens within a reasonable range of parameter values. If it is desirable to obtain only one g and h, instead of ranges of oscillation that bound these values, a possible solution is to exclude some of the small F values. In the cases we tested, this omission heuristically resulted in the algorithm converging to the true optimum.

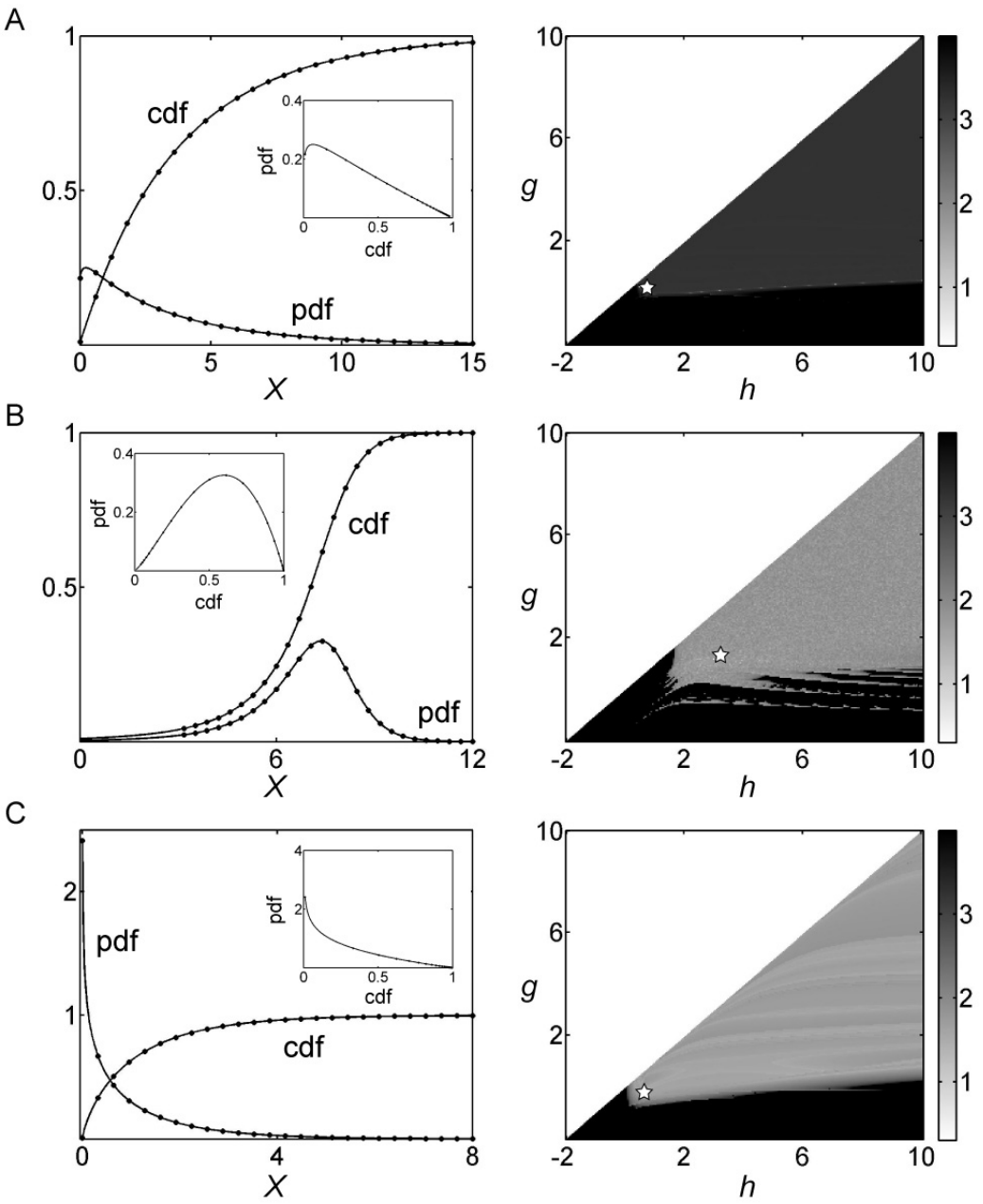

Figure 3: *Summary of convergence patterns of 3-AR. Panels on the left show the pdf, cdf, and f-F plot (insert) of each distribution. Panels on the right present heat maps of convergence as functions of starting values of g and h, with gray bar indicating the logarithm (base 10) of the number of iterations needed for convergence. Each asterisk represents the true value of g or h. Case A:* $\alpha = 1$, $g = 0.25$, $h = 0.5$, $F_0 = 0.01$. *Case B:* $\alpha = 1$, $g = 1.2$, $h = 3$, $F_0 = 0.01$. *Case C:* $\alpha = 1$, $g = -0.2$, $h = 0.5$, $F_0 = 0.01$. *Twenty-five instances of the random variable were chosen in each case.*

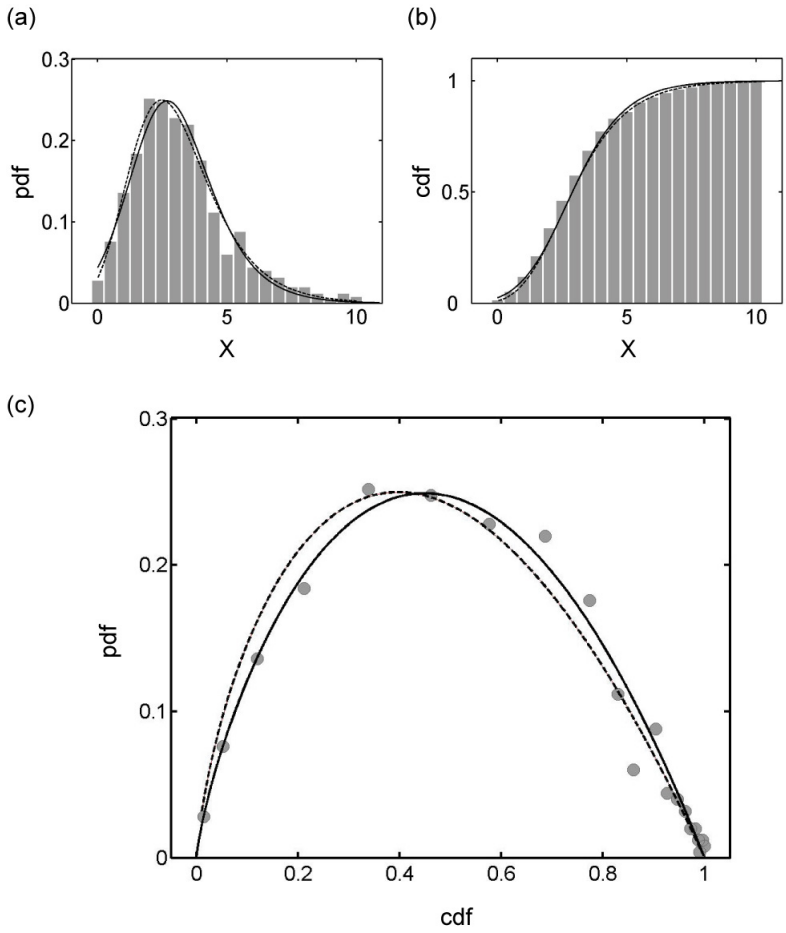

Figure 4: *Data sampled from an S-distribution with parameter values* $\alpha = 1$, $g = 0.75$, $h = 1.5$ *and fits with the parent S-distribution (dashed lines) and with an S-distribution obtained with 3-AR and initial guesses* $g = 10$ *and* $h = 10.5$ *(solid lines). Optimal parameter estimates are obtained as* $\alpha = 0.80$, $g = 0.78$, $h = 1.87$. *(a) pdfs; (b) cdfs; (c)* f*-*F *plot showing the pdf as algebraic function of the cdf. SSE of the 3-AR optimized distribution is 0.0041 (*$S.D. = 0.0151$*), while SSE for the parent S-distribution is 0.0064 (*$S.D. = 0.0189$*).*

4.2 Fitting distributions with noise

The preceding section discussed 3-AR for error-free samples from S-distributions. In this section we analyze finite random samples from S-distributions, which result in artificial datasets that appear noisy. To create these data, we use the quantile method, as discussed in Voit (2000). Specifically, we consider the inverted *cdf* equation

$$\frac{dX}{dF} = \frac{1}{\alpha\,(F^g - F^h)}, \qquad F(0.5) = median \tag{14}$$

and draw random numbers R_i from the uniform distribution over (0,1), which are used as quantiles. Solving Equation (14) numerically upwards or downwards from the median to $F = R_i$ yields in X_i the desired S-distributed random number. The S-distributed random numbers are collected and form the equivalent of an observed data sample, whose "noise" depends on the sample size.

The performance of 3-AR in fitting these artificial data is shown in Figure 4 with an example, where five hundred random numbers were generated from an S-distribution and categorized into 21 bins of a relative frequency histogram (Figure 4a). The *pdf* was constructed from the resulting histogram without smoothing and easily yielded the *cdf* (Figure 4b). The 3-AR algorithm converged within 47 iterations from the initial guesses $g = 10$ and $h = 10.5$ to the estimated solution. Interestingly, the fit with this solution is associated with a lower *SSE* than a fit with the parent S-distribution, from which the "data" were sampled, which confirms similar earlier observations (*e.g.*, Sorribas, March and Voit, 2000). To assess dependence on sample size, we also tested the algorithm with smaller sample sizes, *e.g.*, $n = 100$, and 3-AR performed similarly well.

To explore the flexibility of the S-distribution, we repeated the example shown in Figure 4 several times with 500 points each. The results (Figure 5) show slightly different fits with *SSE*s around 0.0045-0.0047 (Figure 5A), 0.0054-0.0057 (Figure 5B), and 0.0096 (Figure 5C), which are driven by the degree with which each random sample truly represents the underlying distribution. Within each class, the relationships between the estimates α, g, and h are similar, again confirming earlier results (Sorribas, March and Voit, 2000), where classes of quasi-equivalent S-distributions with quite similar *SSE*s were produced by fixing the value of α and fitting g and h. In each class, g and h exhibit an almost linear relationship between each other and with $\log(\alpha)$ and converge to each other when α becomes larger. Even though the parameter sets within each class are clearly different, the resulting distributions are essentially indistinguishable.

In some cases, the 3-AR algorithm does not converge to a single value. Instead, it oscillates between reasonable candidate solutions. This is probably due to noise in the data, causing 3-AR to find the best "local" fit for each phase, which however is not the best fit for other phases. This behaviour is commonly seen in nonlinear algorithms. It is easy to find a suitable solution by choosing from among the candidate solutions, based on their *SSE*s.

4.3 Fitting traditional statistical distributions

The selection of a traditional distribution for fitting data is often difficult because the "true" parent distribution is typically not known. Testing candidate distributions one by one is cumbersome, and all-encompassing distribution families (*e.g.*, Savageau, 1982) often contain so many parameters that over-fitting and redundancy become complicating issues. Instead, the S-distribution may be used as an inclusive model that is capable of

A

α	g	h	SSE
1.134	0.884	1.748	0.0045
2.219	1.028	1.375	0.0046
1.005	0.848	1.805	0.0046
0.743	0.767	1.982	0.0047

B

α	g	h	SSE
0.611	0.704	2.098	0.0054
1.318	0.898	1.555	0.0055
1.830	0.959	1.429	0.0056
4.151	1.055	1.254	0.0057

C

α	g	h	SSE
1.221	0.899	1.785	0.0096
1.537	0.946	1.472	0.0096
2.403	1.004	1.360	0.0096

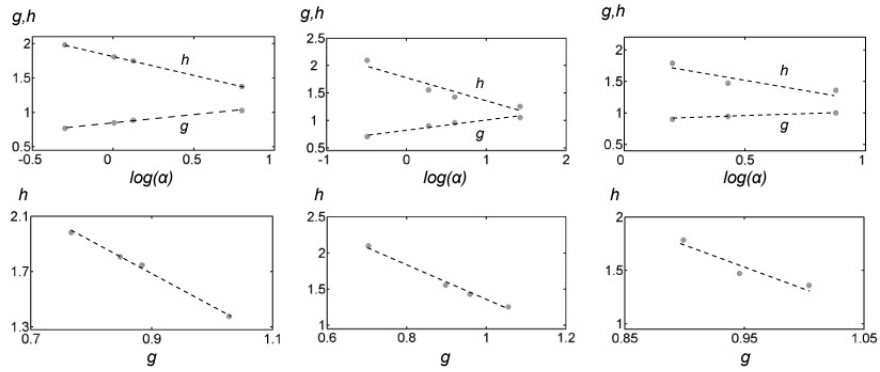

Figure 5: *Quasi-equivalent S-distributions. Parameters are estimated for different samples randomly generated from a given distribution ($\alpha = 1$, $g = 0.75$, $h = 1.5$). The residual errors SSEs are recorded and classified into three classes based on the value of SSE. The plots of g or h versus* $\log(\alpha)$ *and of g versus h are generated in each class. A: SSE between 0.0045 and 0.0047; B: SSE between 0.0054 and 0.0057; C: SSE equal to 0.0096.*

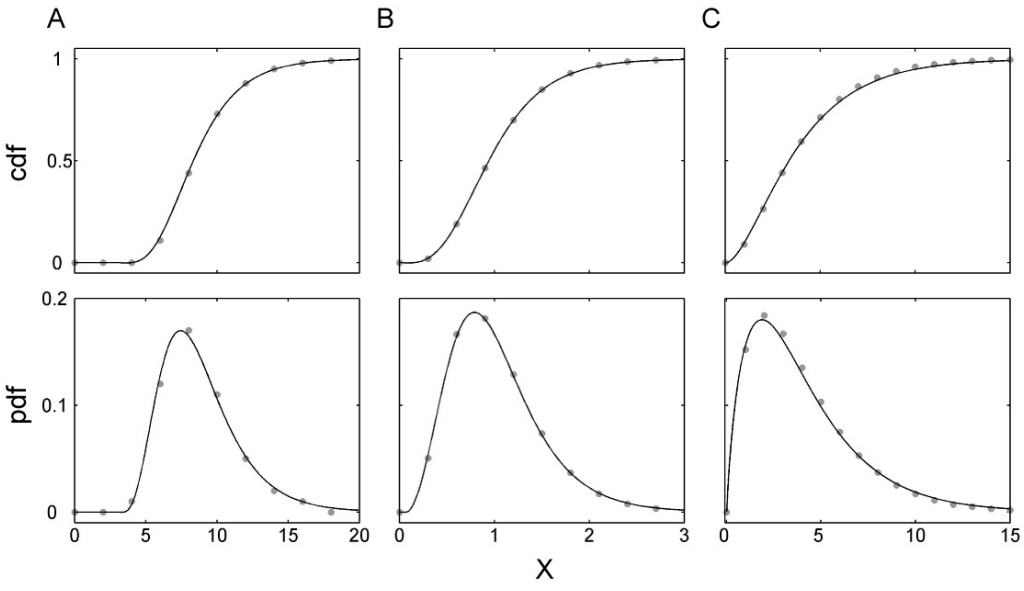

Figure 6: *Fitting traditional distributions. The gray dots represent data used in the regressions, while the solid curves represent the estimated S-distributions. The SSEs are calculated for the f-F plot.A: noncentral $t_{8,8}$-distribution, SSE = 0.00007, S.D. =* 0.0032*; B: $F_{10,100}$-distribution, SSE = 0.00066, S.D. =* 0.0097*; C: χ^2_4-distribution, SSE = 0.00026, S.D. =* 0.0045.

representing many traditional statistical distributions in sufficiently close approximation. The strategy thus becomes to fit data of unknown structure with an S-distribution and to identify which traditional distributions have similar shapes (Voit, 1992; Voit and Yu, 1994; Yu and Voit, 1996). This section explores how well 3-AR identifies S-distributions for random samples from traditional distributions.

The S-distribution contains only two classical distributions as special cases: the exponential distribution for $g = 0$ and $h = 1$ and the logistic distribution for $g = 1$ and $h = 2$. Fitting these two distributions yield *SSE*s equal to 0 (results not shown). All other classical distributions incur some unavoidable approximation error when modelled as S-distributions. Figure 6 shows the results of 3-AR fitting of three examples that are not special cases, namely a noncentral t-distribution, an F-distribution, and a χ^2-distribution; the initial guesses were again chosen as $g = 10$ and $h = 10.5$. As before, 3-AR converges to a solution within a few iterations for these and many other examples. The only convergence problems occurred when fitting traditional distributions requiring $g \approx h$ (see Voit (1992) for these uncommon cases). A possible reason is presumably that the S-distribution is not a very good model for such distributions.

4.4 Fitting observed data

The ultimate measure of success of any fitting algorithm is the modelling of actual data. Figure 7 shows the performance of 3-AR in fitting an S-distribution to weight data of males ages 20 to 29 (data from *NHANES III* (National Center for Health Statistics, 1996)). The observed distribution contains 574 males, classified into bins of 3 kg. The *pdf* and *cdf* histograms were constructed in the same fashion as in Section 4.2. The *SSE* of the fit is similar to the result of using a constrained maximum likelihood estimator (Voit, 2000), although the parameter values are somewhat different, exhibiting again the flexibility and quasi-redundancy inherent in S-distributions. Visually, and judged by the *SSE*, the fit obtained here is satisfactory and obtained in less than a second.

5 Discussion

The S-distribution is a four-variable distribution that combines mathematical simplicity with superior flexibility in modelling data. A crucial prerequisite for using the distribution in practical applications is the availability of effective methods for estimating optimal parameter values from observed frequency data. Addressing this issue, we introduced here a method called *3-way alternating regression* (3-AR) that is extremely fast and robust. The 3-AR method constitutes a modification of a 2-way alternating regression method that was recently proposed for parameter estimation in S-systems (Chou, Martens and Voit, 2006), of which S-distributions are special cases.

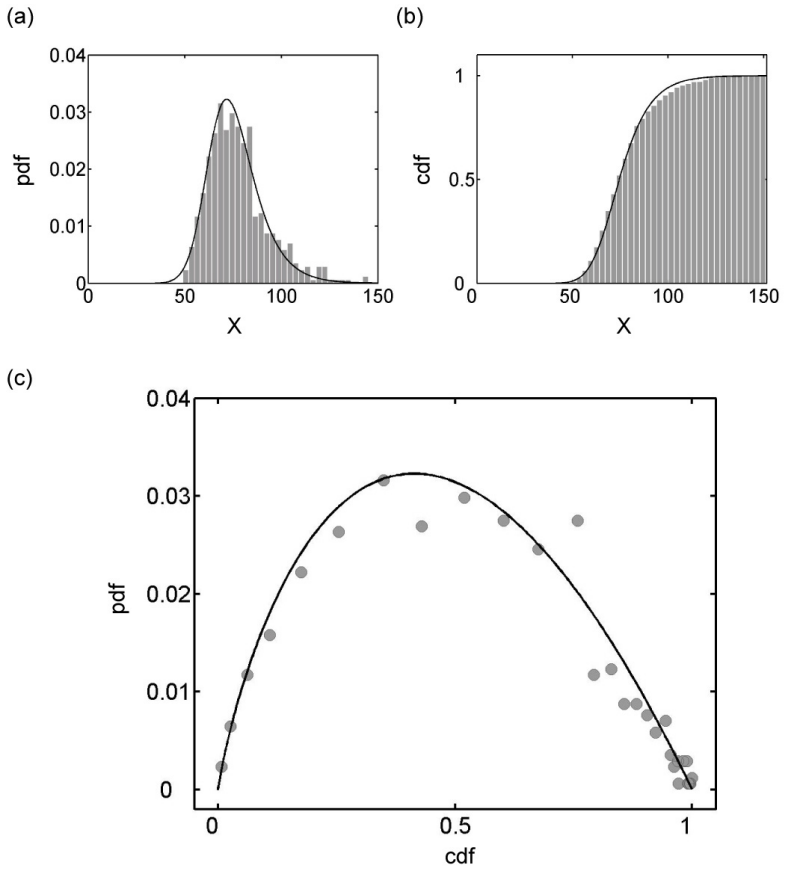

Figure 7: *Fitting observed data. Observed distribution (bars and dots) of weights of 574 males, ages 20-29 (National Center for Health Statistics, 1996) and S-distribution fit (lines) obtained with 3-AR and initial guesses* $g = 10$, $h = 10.5$. *Estimated parameter values:* $\alpha = 0.270$, $g = 0.958$, $h = 1.328$, $X_{0.5} = 74.37$. *(a) pdf (SSE = 0.000143,* $S.D. = 0.0023$*); (b) cdf (SSE =0.009629,* $S.D. = 0.0189$*); (c)* f*-*F *plot (SSE = 0.000187,* $S.D. = 0.0026$*).*

The 3-AR method performs well in all typical scenarios, namely for estimating parameters from error-free distributions, from random samples generated from S-distributions, from traditional statistical distributions, and from actual data. The basin of convergence is rather large, and convergence speed is essentially independent of initial guesses that are selected to start the 3-AR algorithm. Therefore, even if one selects initial guesses quite far away from the true optimum, the algorithm only takes a few iterations to converge to points very close to the true solution and refines this solution with a relatively small number of further cycles. An exception is the situation where 3-AR converges to the trivial solution where α increases without bound and g approaches h. This scenario is easy to spot and the choice of another initial guess typically remedies the situation. A second exception to rapid convergence may occur if the true g and h are very different. In this rather unusual case, the algorithm sometimes converges to values

between the true g and h and oscillates between them. In this case, one may select values from within the oscillation range or redo the estimation by omitting some of the very small values of the *pdf* and *cdf*.

The 3-AR fitting of data from traditional distributions works well in most cases, except for distributions that are not well approximated by S-distributions and where the relatively best fit requires $g \approx h$, as described in Section 4.3.

For finite random samples, the estimated solution is also obtained very quickly, but its parameters depend on the particular sample. As a consequence, the computed estimates may be rather different, even though the *SSE*s are very similar and the shapes of the resulting distributions are essentially indistinguishable. This finding is a manifestation of the shape flexibility and quasi-redundancy of S-distributions and confirms similar observations in the literature (*e.g.*, Sorribas, March and Voit, 2000).

The 3-AR algorithm provides a strategy for parameter estimation with S-distributions that is genuinely different from all other published methods. While some issues associated with the basin of convergence should be investigated further, our results shown here provide strong indication that this algorithm is much faster than the currently available alternatives.

An issue that seems generic to S-distributions and has been observed in other contexts is the covariance among the parameters α, g, and h (*e.g.*, Sorribas, March and Voit, 2000). While each set of these parameters determines a unique distribution, the covariance permits distinct sets leading to solutions that are so similar that their differences are often smaller than the noise in the data. This quasi-equivalence will require future work. For instance, it might be possible to specify the theoretical uncertainty variances of the estimated parameters or analytically study the uncertainty variance by principal component analysis or linear series expansion of the model around the convergence point (α, g and h).

Quasi-equivalence also poses problems when it is necessary to determine the uncertainty in the estimated parameters, for instance in the context of significance testing. The quasi-equivalent different parameter sets, which yield essentially indistinguishable distributions, are not arbitrary, but form slightly curved, essentially one-dimensional manifolds in the parameter space, as we and others have discussed in the literature several times. These manifolds may be similar to quasi-solution sets recently derived from Newton flow methods (see Dedieu and Shub, 2005). Whatever the structure of the quasi-solution sets may be, it is quite evident that equivalence tests focusing on one parameter at a time will not be useful. Instead, one will have to compare solutions globally, for instance based on Hellinger or Kullbach-Leibler distances (see Balthis, 1998) or on some measure of maximal distance, such as $Q_2 = sup_X|F_1(X) - F_2(X)|$. To calculate a confidence interval for these distances, one would probably use the bootstrap. One could similarly use bootstrap methods to calculate p-values for the null hypothesis that two S-distributions are the same, although the bootstrap sampling for hypothesis testing would be slightly different than that used

for confidence intervals. Furthermore, one could use Monte Carlo simulation methods to construct power curves for the alternative significance tests, under different true scenarios.

A related issue needing future attention will be the characterization of the intrinsic features of the 3-AR estimator, including its biasedness, consistency, and efficiency. These characterizations appear to be complex and may have to be postponed until the convergence behaviour of 3-AR is more fully understood.

Finally, a future extension of 3-AR might be its generalization to the more comprehensive GS-distribution (Muiño, Voit and Sorribas, 2006), which is characterized by increased flexibility in shape, in particular, for symmetric distributions, at the cost of one additional parameter. The inclusion of this additional parameter will require modifications to the 3-AR algorithm that need to be investigated in detail.

Acknowledgments

This work was supported in part by a National Heart, Lung and Blood Institute Proteomics Initiative (Contract N01-HV-28181; D. Knapp, PI), a grant from the National Science Foundation (MCB 0517135; E. O. Voit, PI), and an endowment from the Georgia Research Alliance. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring institutions.

References

Balthis, W. L. (1998). *Application of Hierarchical Monte Carlo Simulation to the Estimation of Human Exposure to Mercury via Consumption of King Mackerel (Scomberomorus cavalla)*, Ph.D. Dissertation, Medical University of South Carolina, Charleston, SC.

Chou, I. C., Martens, H. and Voit, E. O. (2006). Parameter estimation in biochemical systems models with alternating regression. *Theoretical Biology and Medical Modelling*, 25.

Dedieu, J.-B. and Shub, M. (2005). Newton flow and interior point methods in linear programming. *International Journal of Bifurcation and Chaos*, 15, 827-840.

National Center for Health Statistics. (1996). Analytic and Reporting Guidelines: The Third National Health and Nutrition Examination Survey, NHANES III (1988-1994). U.S. Department of Health and Human Services, Public Health Service, Center for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

Hernández-Bermejo, B. and Sorribas, A. (2001). Analytical quantile solution for the S-distribution, random number generation and statistical data modelling. *Biometrical Journal*, 43, 1007-1025.

Muiño, J. M., Voit, E. O. and Sorribas, A. (2006). GS-distributions: a new family of distributions for continuous unimodal variables. *Computational Statistics and Data Analysis*, 50, 2769-2798.

Savageau, M. A. (1982). A suprasystem of probability distributions. *Biometrical Journal*, 24, 323-330.

Sorribas, A., March, J. and Voit, E. O. (2000). Estimating age-related trends in cross-sectional studies using S-distributions. *Statistics in Medecine*, 19, 697-713.

Voit, E. O. (1992). The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions. *Biometrical Journal*, 34, 855-878.

Voit, E. O. (1996). Dynamic trends in distributions. *Biometrical Journal*, 38, 587-603.

Voit, E. O. (2000). A maximum likelihood estimator for the shape parameters of S-distributions. *Biometrical Journal*, 42, 471-479.

Voit, E. O. and Almeida, J. (2000). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 20, 1670-81.

Voit, E. O. and Schwacke, L. H. (2000). Random number generation from right-skewed, symmetric, and left-skewed distribution. *Risk Analysis*, 20, 59-71.

Voit, E. O. and Sorribas, A. (2000). Computer modelling of dynamically changing distributions of random variables. *Mathematical and Computer Modelling*, 31, 217-225.

Voit, E. O. and Yu, S. (1994). The S-distribution. Approximation of discrete distributions. *Biometrical Journal*, 36, 205-219.

Yu, S. S. and Voit, E. O. (1996). A graphical classification of survival distributions. In: *Lifetime Data: Models in Reliability and Survival Analysis*, Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A., Kluwer Academic Publishers, Dordrecht, 385-392.