Akademie věd České republiky
ÚSTAV TEORIE INFORMACE A AUTOMATIZACE

Academy of Sciences of the Czech Republic
INSTITUTE OF INFORMATION THEORY AND AUTOMATION

RESEARCH REPORT

J.J. Hench

Toward a Convergent Entropic Identification Algorithm

No. 1830 February 1995

ÚTIA AV ČR P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42)(2)6641 4903
e-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Toward a Convergent Entropic Identification Algorithm

J.J. Hench
Institute of Information and Control Theory
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
182 08 Praha 8

March 2, 1995

Abstract

An analogy between an entropic identification scheme proposed by Klán and an identification scheme based on an under–determined least–squares formulation is made. This formulation produces an estimate of the parameter vector in an ARMA setting which is of minimum two–norm while satisfying the constraint that the sum of the elements of the parameter vector is equal to one and that the inner product of the parameter and regression vector at any given time is equal to the measured output one time–step in the future.

Continuing along this thread, another formulation is produced, one which is akin to Kaczmarz's method. This formulation provides an estimate of the parameter vector such that it satisfies the same constraints as the previous formulation, but minimizes the two-norm between each successive parameter estimate, providing an additive update of the parameter vector at each time–step. This formulation provides insight into the entropic identification scheme proposed by Klán; namely, that it satisfies similar constraints, but produces a parameter estimate which minimizes an entropic "norm" instead of the standard two–norm. By replacing the entropic "norm" in their algorithm by the Kullback–Leibler metric measuring the distance between discrete probability distributions, an new entropic identification scheme is produced. This algorithm has a number of interesting properties. First, parameter convergence is similar, under certain conditions, to Kaczmarz's method. Second, the parameter update is multiplicative rather than additive. Finally, numerical experiments suggest that the entropic algorithms are less prone to the bursting phenomenon than the least-squares algorithms.

Keywords 1 Entropy, Identification, Convex Combination, Kullback-Leibler, Kalman Filter

1 Introduction

In this paper, we draw parallels between various standard (least–squares) methods for system identification, and one based on an entropic measure. More specifically, a under-determined variant of a least-squares identification procedure is reformulated to include an algebraic constraint, and an *a priori* estimate. Such a formulation provides a starting point for the consideration of Klán's method [4, 5], as it differs only in the parameter domain and metric used in the minimization procedure. This perspective is used to suggest possible modifications to the procedure.

To begin, the focus in this paper is restricted to the ARMA (or DELTA) [6, 10] model. For a linear system with inputs u_k and outputs y_k, the underlying relation may be written as

$$y_k = \sum_{j=0}^{n} \alpha_j u_{k-j} + \sum_{j=1}^{m} \beta_j y_{k-j}. \tag{1}$$

A parametric model is used to describe this relationship. Let

$$Y(k,n) = \begin{pmatrix} y_k \\ \vdots \\ y_{k-n} \end{pmatrix} \quad \text{and} \quad U(k,m) = \begin{pmatrix} u_k \\ \vdots \\ u_{k-m} \end{pmatrix} \tag{2}$$

Also, let the parameters α_j and β_j be collected in a vector θ, where

$$\theta(m,n) = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix}. \tag{3}$$

Further, let the (Hankel) matrix $\Phi(k,n,m,p)$ be defined as

$$\Phi_k(n,m,p) = \begin{bmatrix} y_{k-1} & \cdots & y_{k-m} & u_k & \cdots & u_{k-n} \\ y_{k-2} & \cdots & y_{k-m-1} & u_{k-1} & \cdots & u_{k-n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{k-p} & \cdots & y_{k-m} & u_k & \cdots & u_{k-n-p} \end{bmatrix}. \tag{4}$$

For the sequence $\{y\}$ and $\{u\}$ fitting the model proposed in (1), we have the relation

$$y(k,p) = \Phi_k(n,m,p)\theta(m,n). \tag{5}$$

The problem of system identification, however, requires that an estimate of $\theta(m,n)$ is produced given possibly noise data. For the purposes of this paper, noisy data may be formulated as follows. Let

$$z_k = y_k + \delta_k \quad \text{and} \quad v_k = u_k + \epsilon_k, \tag{6}$$

where δ_k and ϵ_k are error sequences, possibly unmodelled, with the vectors $z(m,n)$ and $v(m,n)$ defined analogously. Also, the matrix $\Psi_k(m,n,p)$ is defined to be the (Hankel) matrix in (4), where z and v are substituted for y and u, respectively.

Thus, the problem of least–squares linear system identification may be written as

$$\vartheta_k(m,n) = \underset{x(m,n)}{\operatorname{argmin}} \|\Psi_k(m,n,p)x(m,n) - z(k,p)\|_2. \tag{7}$$

The least–squares problem, however, may be expanded, allowing for additional information to be included in the formulation. One variant might be that the estimate $\vartheta(m,n)$ is required to be somehow close to a (possibly arbitrary or time–varying) *a priori* estimate $\rho(m,n)$, where

$$\rho_k(m,n) = \begin{pmatrix} \rho_{k[1]} \\ \vdots \\ \rho_{k[m+n+1]} \end{pmatrix}. \tag{8}$$

In such a case, the identification problem is reduced to finding the deviation vector $\xi_k(m,n)$, where

$$\vartheta_k(m,n) = \rho_k(m,n) + \xi_k(m,n) \tag{9}$$

Another variant might be that the parameter estimate estimate is somehow constrained. While many constraints are feasible [2], we restrict our attention to constraints of the linear type, namely

$$A\vartheta = b. \tag{10}$$

With these considerations, a generalized formulation of the least–squares problem which estimates $\theta(m,n)$, *i.e.,* $\vartheta(k,m,n)$ at time k given sequences $\{z_j\}$ and $\{v_j\}$, and constraint parameters A, b, and $\rho_k(m,n)$ may be written as follows:

$$\vartheta_k(m,n) = \rho_k(m,n) + \xi_k(m,n) \tag{11}$$

where

$$\xi_k(m,n) = \underset{x(m,n)}{\operatorname{argmin}} \|\Psi_k(m,n,p)(\rho_k(m,n) + x(m,n)) - Z_k(p)\|_2, \tag{12}$$

subject to

$$A(\rho(m,n) + x(m,n)) = b. \tag{13}$$

The solution to this problem may be obtained using the singular value decomposition [2]. In the simpler case where $A = 0$ and $b = 0$, the solution may be written as

$$\vartheta_k(m,n) = \rho_k(m,n) + \Psi_k^+(m,n,p)(Z_k(p) - \Psi_k(m,n,p)\rho_k(m,n)), \tag{14}$$

where $\Psi_k^+(m,n,p)$ is the Morse–Penrose pseudo-inverse of $\Psi_k(m,n,p)$. When p is chosen such that the system (14) is overdetermined, *i.e.,* $\operatorname{rank}[\Psi_k(m,n,p)] > m+n+1$, then

$$\Psi_k^+(m,n,p)\Psi_k(m,n,p) = I,$$

which gives the standard least–squares solution [6, 8, 11]

$$\vartheta_k(m,n) = \Psi_k^+(m,n,p)Z_k(p). \tag{15}$$

For the sake of clarity, we provide two tables which contain the nomenclature used in this paper. The parameters which denote matrix and vector size, *i.e.,* those which are within parenthesis, will in general be omitted.

2 Underdetermined Least Squares

In most least squares algorithms p is chosen such that the system (14) is over-determined; however, there is no theoretical reason why this must be so. Since our eventual goal is to relate the mechanism of entropic identification to least squares method, it is important to consider the case where Ψ_k is underdetermined. This goal also motivates the consideration of two *a priori* parameter estimates. One *a priori* parameter estimate of interest is the (non-time-varying) vector

$$\rho_k(m,n) = h(m,n)/(m+n+1) = h/\|h\|_2. \tag{16}$$

The second *a priori* parameter estimate of interest is time-varying. It is none other that that which is used in Kaczmarz's method:

$$\rho_k = \vartheta_{k-1}. \tag{17}$$

An additional constraint to be considered is the requirement that the sum of the elements of the parameter vector be equal to one. The choice of these constraints will have

Perfect Data	Description	Noisy Data
u_k	Input	v_k
y_k	Output	z_k
$\theta(m,n)$	$(n+m+1)$ Parameter Vector	$\vartheta_k(m,n)$
—	Convergent Constr. Underdet. LS (CCULS) estimator	$\tilde{\vartheta}_k(m,n)$
—	Static Constr. Underdet. LS (SCULS) estimator	$\bar{\vartheta}_k(m,n)$
—	Convergent Entropic (CE) Estimator	$\tilde{\theta}_k(m,n)$
—	Static Entropic (SE) [Klán's] Estimator	$\bar{\theta}_k(m,n)$
$\Phi_k(m,n,p)$	$(p \times n+m+1)$ Hankel Matrix	$\Psi_k(m,n,p)$
—	$(1 \times n+m+1)$ Hankel Matrix (regressor vector)	$\psi_k^T(m,n)$
—	$(2 \times n+m+1)$ Constrained Hankel Matrix	$H_k(m,n)$

Table 1: Data Conventions

Variable	Description
A	Constraint Matrix
b	Constraint Vector
$\rho_k(m,n)$	*A Priori* Estimate
$\xi_k(m,n)$	Deviation Vector
$h(m,n)$	Vector of ones
$\Delta(\cdot,\cdot)$	Kullback–Leibler Metric
P_k	Diagonal Matrix from ρ_k
Θ_k	Diagonal Matrix from θ_k
Ψ_k	Diagonal Matrix from ψ_k

Table 2: Other Conventions

obvious meaning in the following section. It suffices to mention now that this constraint fits within our original framework, where $A = h^T(m,n)$ and $b = 1$, where $h(m,n)$ is the vector of ones of length $m+n+1$. Parenthetically, this constraint may seem unduly restrictive, but it is a reasonable constraint for parameter vector for systems whose inputs and outputs are normalized to fall within the range from zero to one [4]. Furthermore, if parameter vector is strictly positive, then this restriction corresponds to a rather desirable class of steady–state unit–gain, minimum–phase filters.

Since the underlying problem is underdetermined, there exists an elegant solution to both of these variants. Since the problem is underdetermined, it is possible to include the constraint directly in the minimization procedure. Let

$$\psi_k^T = \Psi_k(m,n,1), \tag{18}$$

$$H_k = \begin{bmatrix} \psi_k^T \\ h^T \end{bmatrix}, \tag{19}$$

and

$$\zeta_k = \begin{pmatrix} z_k \\ 1 \end{pmatrix}. \tag{20}$$

Both constrained estimation problems may then be written as:

$$\vartheta_k = \rho_k + \xi_k \tag{21}$$

where

$$\xi_k = \underset{x}{\operatorname{argmin}} \|H_k(\rho_k + x) - \zeta_k\|_2. \tag{22}$$

If ρ_k is defined by (16), then the resulting estimation algorithm is called the Static Constrained Underdetermined Least-Squares (SCULS) identification algorithm, and if ρ_k is defined by (17), then the resulting estimation algorithm is called the Convergent Constrained Underdetermined Least-Squares (CCULS) identification algorithm. The solutions to these two problems may be written as follows. The SCULS parameter estimator $\bar{\vartheta}_k$

$$\bar{\vartheta}_k = h/||h||_2 + H_k^+(\zeta_k - H_k h/\|h\|_2), \tag{23}$$

and the CCULS parameter estimator $\tilde{\vartheta}_k$

$$\tilde{\vartheta}_k = \tilde{\vartheta}_{k-1} + H_k^+(\zeta_k - H_k\tilde{\vartheta}_{k-1}). \tag{24}$$

3 Entropic Identification

In the previous section, the both the CCULS and the SCULS parameter estimators were produced by minimizing the two norm distance between an *a priori* parameter and constraint. In this section, essentially the same problem is considered, except that the Kullback–Leibler "norm" [1] replaces the two-norm that is used to measure the distance between the *a priori* estimate and the estimator. In the most usual sense, the Kullback–Leibler metric is used to measure distances between probability distributions. In this section, the parameter vector estimate θ_k is supposed in some sense to look like a distribution, *i.e.,*

$$h^T\theta_k = 1 \quad \text{and} \quad \theta_j \geq 0. \tag{25}$$

The Kullback–Leibler metric $\Delta(\theta, \rho)$, used to measure the distance between a parameter vector θ and an *a priori* estimate ρ, may be written as

$$\Delta(\theta, \rho) = \text{trace}\left[\Theta_k \ln\left(P_k^{-1}\Theta_k\right)\right], \tag{26}$$

where

$$\begin{aligned} P_k &= \text{diag}(\rho_k) \\ \Theta_k &= \text{diag}(\theta_k). \end{aligned} \tag{27}$$

For the purpose of identification, we have the additional constraint [4] that the inner–product of the regression and parameter vector is equal to the output, akin to (21) and (22). This constraint is basically the expected value constraint used in the Maximum Entropy Principle [3]. Thus, the generalized entropic identification problem may be written as:

$$\theta_k = \underset{x \in \Omega}{\text{argmin}} \Delta(x, \rho_k). \tag{28}$$

subject to

$$\psi_k^T\theta_k = z_k \tag{29}$$

where Ω is the space of all vectors satisfying (25).

The solution to this problem may be found by using the method of Lagrange Multipliers [3]. The Lagrangian may be written as

$$L(\theta, \alpha, \beta) = \text{trace}\left[\Theta_k \ln\left(P_k^{-1}\Theta_k\right)\right] + \alpha\left(\text{trace}\left[\Theta_k - I\right]\right) + \beta\left(\text{trace}\left[\Psi_k\Theta_k - z_k I\right]\right). \tag{30}$$

Here, P_k and Θ_k are defined as in (27) with

$$\Psi_k = \text{diag}(\psi_k). \tag{31}$$

The Euler–Lagrange necessary conditions are

$$\begin{aligned} \frac{\partial}{\partial \Theta} L(\theta,\alpha,\beta) &= 0 \\ \frac{\partial}{\partial \alpha} L(\theta,\alpha,\beta) &= 0 \\ \frac{\partial}{\partial \beta} L(\theta,\alpha,\beta) &= 0. \end{aligned} \tag{32}$$

Solving the first equation in (32) produces

$$\begin{aligned} I + \ln(P_k^{-1}\Theta_k) + \alpha I + \beta \Psi_k &= 0 \\ \Rightarrow \qquad \Theta_k &= e^{\alpha-1} P_k e^{\beta \Psi_k}. \end{aligned} \tag{33}$$

Substituting this result back in the second equation of (32) produces

$$\begin{aligned} \text{trace}\left[\Theta_k\right] &= 1 \\ \Rightarrow \quad \text{trace}\left[P_k e^{\beta\Psi_k}\right] &= \tfrac{1}{e^{\alpha-1}} \\ \Rightarrow \qquad \Theta_k &= P_k e^{\beta\Psi_k}/\text{trace}\left[P_k e^{\beta\Psi_k}\right]. \end{aligned} \tag{34}$$

Returning to the third equation of (32) produces

$$\begin{aligned} \text{trace}\left[\Psi_k\Theta_k - z_k I\right] &= 0 \\ \Rightarrow \quad \text{trace}\left[\Psi_k P_k e^{\beta\Psi_k} - z_k I \left(\text{trace}\left[P_k e^{\beta\Psi_k}\right]\right)\right] &= 0 \\ \Rightarrow \qquad \text{trace}\left[\left(\Psi_k - z_k I\right) P_k e^{\beta\Psi_k}\right] &= 0 \\ \Rightarrow \qquad \text{trace}\left[\left(\Psi_k - z_k I\right) P_k e^{\beta(\Psi_k - z_k I)}\right] &= 0. \end{aligned} \tag{35}$$

As in Section 2, we will consider two *a priori* parameter estimates, *i.e.,* those defined by (16) and (17). In keeping with the nomenclature of the previous section, the Static Entropic (SE) parameter estimator $\bar{\theta}_k$ is produced by considering $\rho_k = h/\|h\|$ and the Convergent Entropic (CE) parameter estimator $\tilde{\theta}_k$ is produced by considering $\rho_k = \tilde{\theta}_{k-1}$. At this point, a number of interesting properties of these estimators may be noted. First, the SE parameter estimator produces an expression for θ_k which is identical to that produced by Klán's method [4, 5]. Second, update for θ_k is multiplicative, *i.e.,*

$$\theta_k = \rho_k \odot \theta_{k-1}, \tag{36}$$

where $\odot$ indicates element by element multiplication.

Perhaps one last remark is in order. There exists no closed form solution to (35); however, it is trivial to write down the derivative of this function with respect to β:

$$\frac{d^n}{d\beta^n}\left(\Psi_k - z_k I\right) P_k e^{\beta(\Psi_k - z_k I)} = \left(\Psi_k - z_k I\right)^{n+1} P_k e^{\beta(\Psi_k - z_k I)} \tag{37}$$

This fact allows for straight–forward computation of β via Newton's method [7, 9].

4 Numerical Examples

In this section, we provide a number of numerical examples to illustrate the discussions of the previous sections.

Example 1: In this example, two sets of input and output sequences, u_k and y_k, of length $l = 500$ are produced by the linear system characterized by (5) with $\theta = [1/10, 2/5, 1/10, 2/5]^T$. These sequences were then corrupted by Gaussian white noise sequences δ_k and ϵ_k with zero–mean and covariance $\sigma = .01$, producing the sequences v_k and z_k as per (6), which are then used by algorithm. In the first example the input u_k was a uniformly distributed random sequence on $[0, 1]$, while in the second example the input u_k was a triangular wave. In all tests the size of the estimated parameter vector was equal to the actual parameter vector. All four algorithms discussed in this paper are plotted in Appendix A, in Figures 1-8.

Some interesting properties of the algorithm are apparent. First, note the convergence properties (in parameters) of the CCULS and CE algorithms, and the lack of convergence in their static counterparts, the SCULS and SE algorithms. Second, note the overall similarities in both of the static and convergent algorithms.

Example 2: In this example, a set of input and output sequences, y_k and u_k, of length $l = 500$ are produced by the time–varying system,

$$y_{k+1} = c_k y_k + (1 - c_k) u_k$$

where

$$c_k = \frac{1}{2} + \frac{1}{10} \tan^{-1}(\frac{t - 25}{5}).$$

These sequences were then corrupted by Gaussian white noise sequences δ_k and ϵ_k with zero–mean and covariance $\sigma = .02$, producing the sequences v_k and z_k as per (6), which are then used by algorithm. The input u_k was a uniformly distributed random sequence on $[0, 1]$. In all tests the size of the estimated parameter vector was of length 3, while the actual parameter vector was of length 2. All four algorithms discussed in this paper are again plotted in Appendix A, in figures 9-12.

The motivation for this numerical was to demonstrate the CULS type algorithms were more susceptible to bursting than the entropic ones. Note that the maximum error in the parameter estimate norm for the CULS type algorithms was close to 4 (Figures 9 and 10), while the maximum error in the parameter estimate norm for the entropic type algorithms was about 0.8 (Figures 11 and 12).

5 Conclusion

In this paper, an effective updating entropic system identification scheme is presented. Although the primary intent is focused on identifying linear systems whose parametric form is restricted to the space where the elements are positive and sum to one, it has many other potential usages. One possible usage is providing a method for assigning reasonable weights to identification "channels" which employ different methods of system identification. Numerical experiments also suggest that similar results may be achieved by applying a corresponding restricted least-squares approach, which could be in addition to the "sum" constraint to limit the absolute values of the parameter elements. Such a restriction can limit the magnitude of the "bursting" phenomenon from which the least–squares approach may suffer under the condition of insufficient excitation.

References

[1] Cover, T.M and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[2] Golub, G.H. and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1983, (second edition: 1989).

[3] Guiasu, S. and A. Shenitzer, "The Maximum Entropy Principle," *Mathematical Intellegence*, 7 (1985), pp. 42–48.

[4] Klán, P., *On Adaptive Forecasting of Control System Outputs*, PhD thesis, Czech Technical University, Prague, CSFR, 1993, (in Czech).

[5] Klán, P. and J. Maršík, "Short–Horizon Forecasts by a New Method," Internal departmental report of the Adaptive Control Department of the Institute of Infromation Theory and Automation, Prague, Czech Republic.

[6] Kumar, P.R., "A Survey of some Results in Stochastic Adaptive Control," *SIAM J. Control Opt.*, 25 (1985), pp. 329–380.

[7] Laub, A.J., "Invariant Subspace Methods for the Numerical Solution of Riccati Equations," in *The Riccati Equation*, S. Bittanti, A.J. Laub, and J.C. Willems (eds.), chapter 7, Springer-Verlag, Berlin, 1991.

[8] Ljung, L., *System Identification: Theory for the User*, Prentice-Hall International, London, 1987.

[9] Luenberger, D.G., *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts, 1973, (second edition: 1984).

[10] Peterka, V., "Control of Uncertain Processes: Applied Theory and Algorithms," *Kybernetica (Supplement)*, 22 (1986), pp. 1–57.

[11] Söderström, T. and T. Stoica, *System Identification*, Prentice-Hall International, London, 1989.

Appendix A

This appendix contains the plots that are referred to in section 5. The first eight figures pertain to example one in the numerical examples section. The last four figures pertain to example two in the numerical examples section.

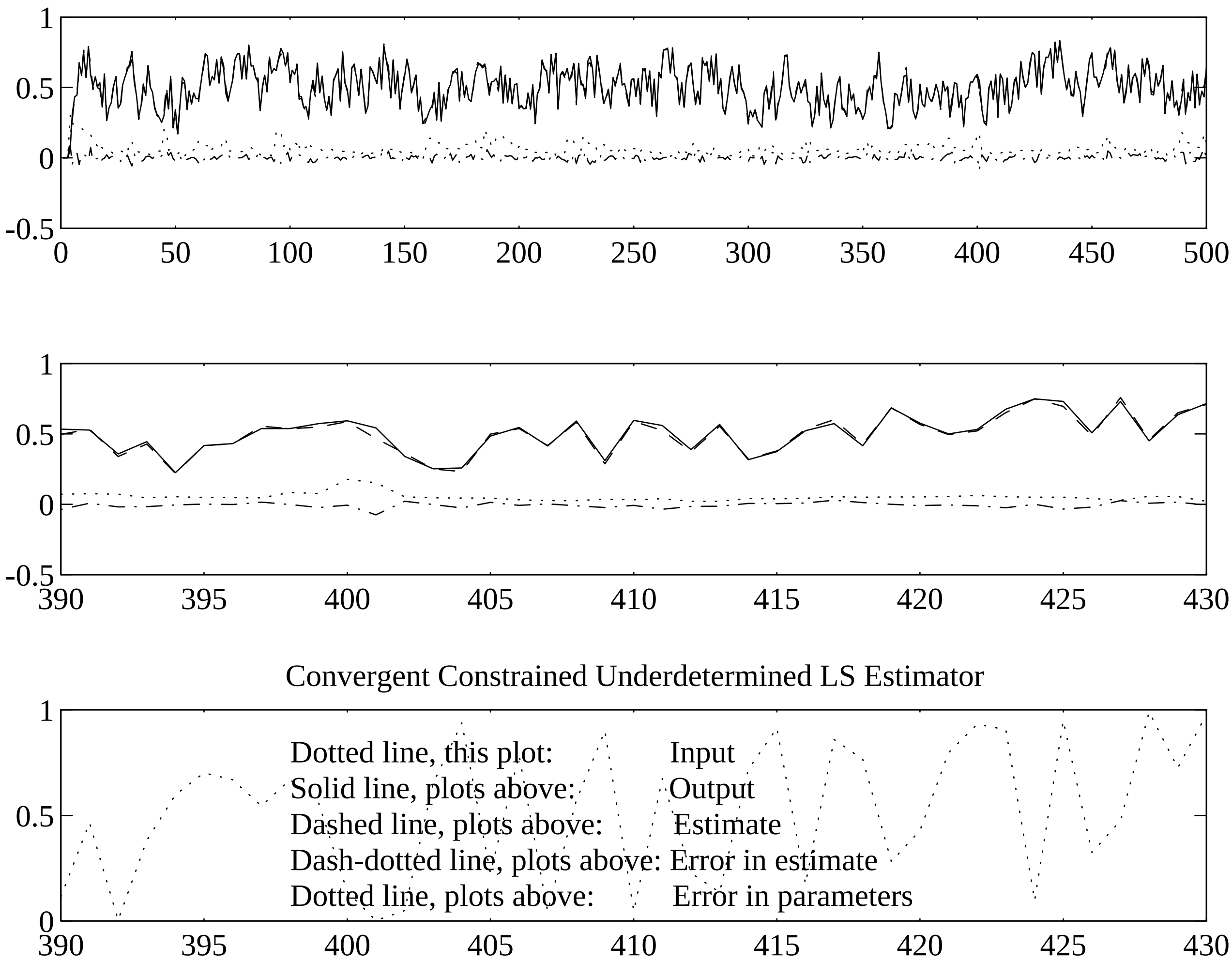

Figure 1: Random Input to CCULS Algorithm

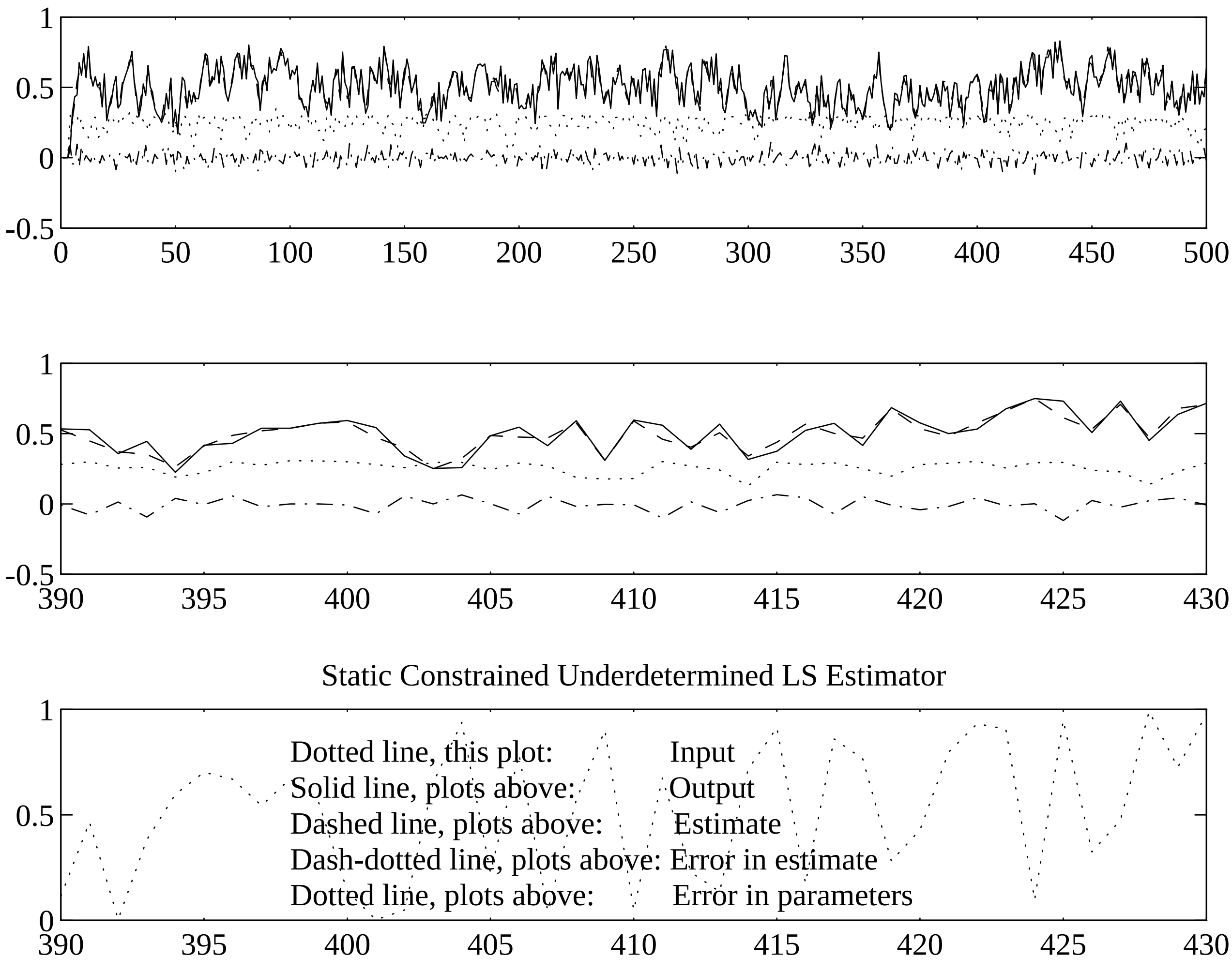

Figure 2: Random Input to SCULS Algorithm

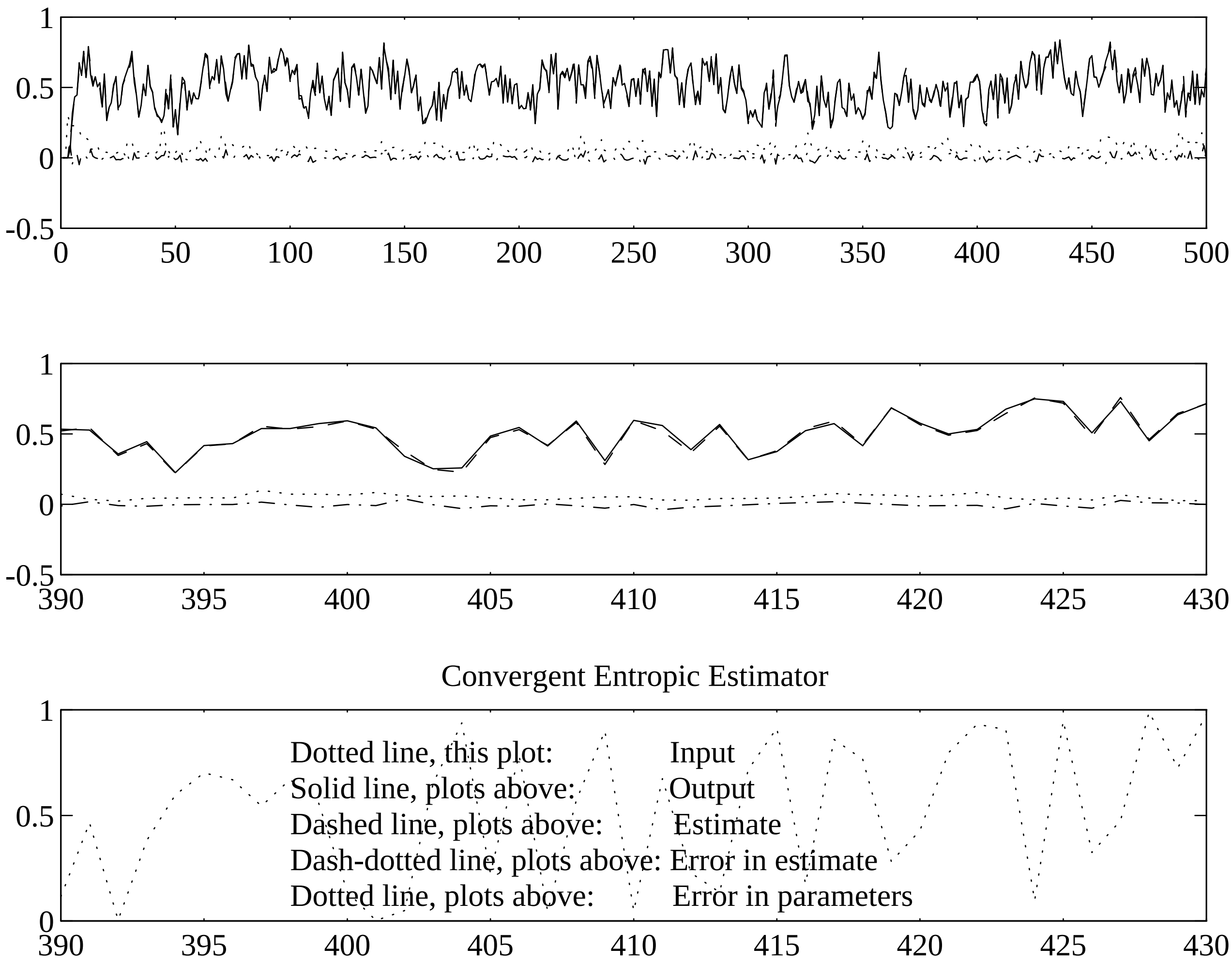

Figure 3: Random Input to Convergent Entropic Algorithm

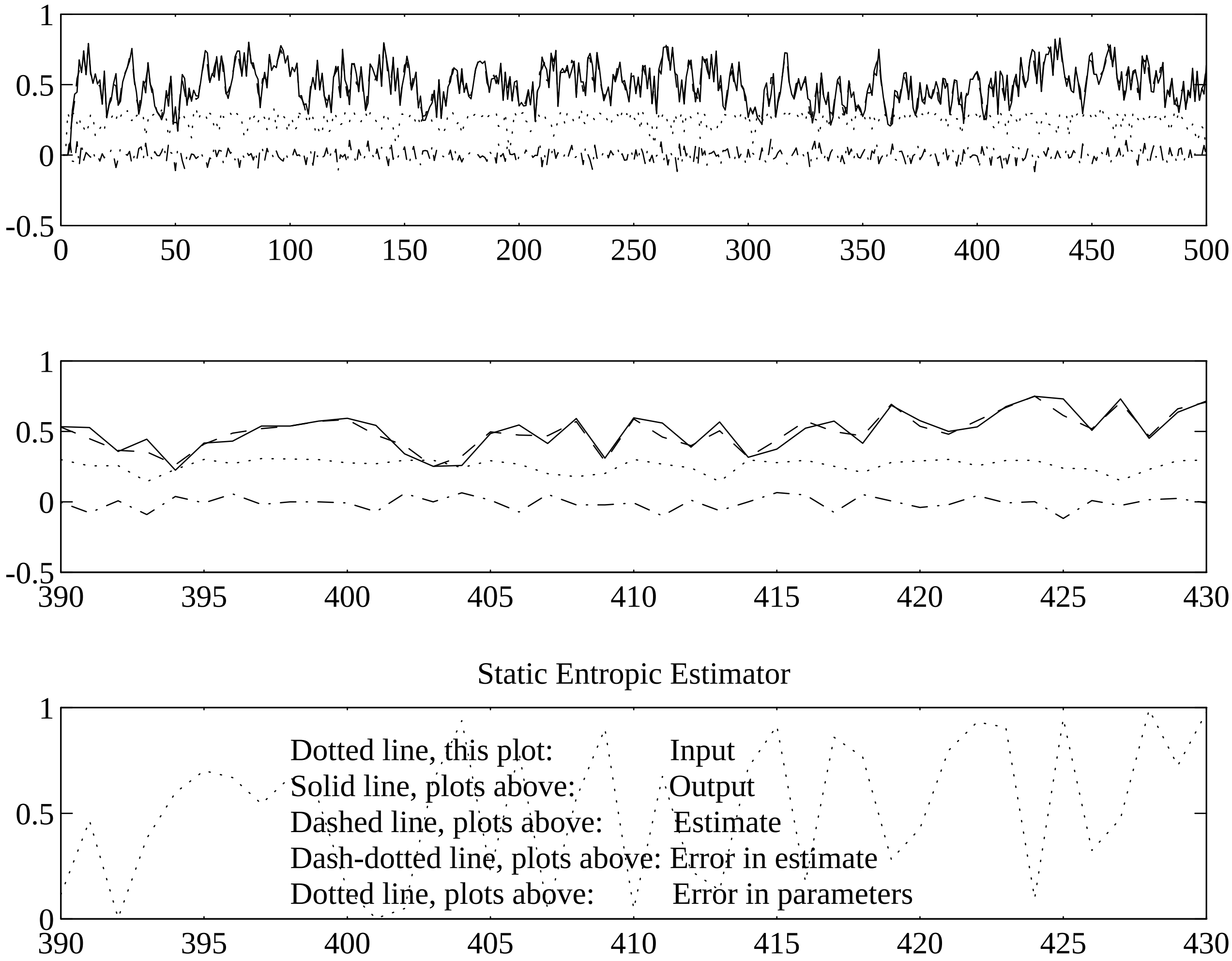

Figure 4: Random Input to Static Entropic Algorithm

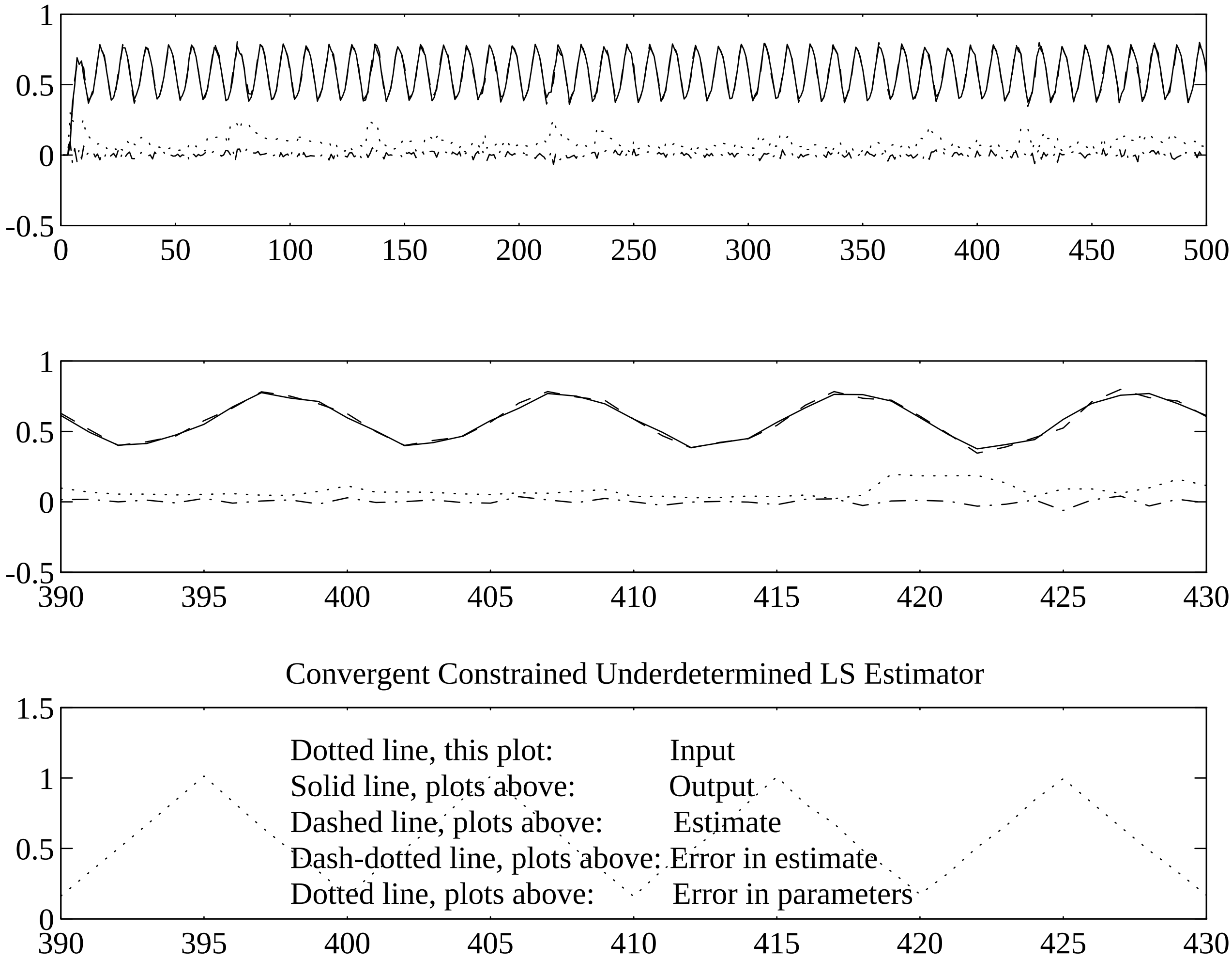

Figure 5: Triangular Input to CCULS Algorithm

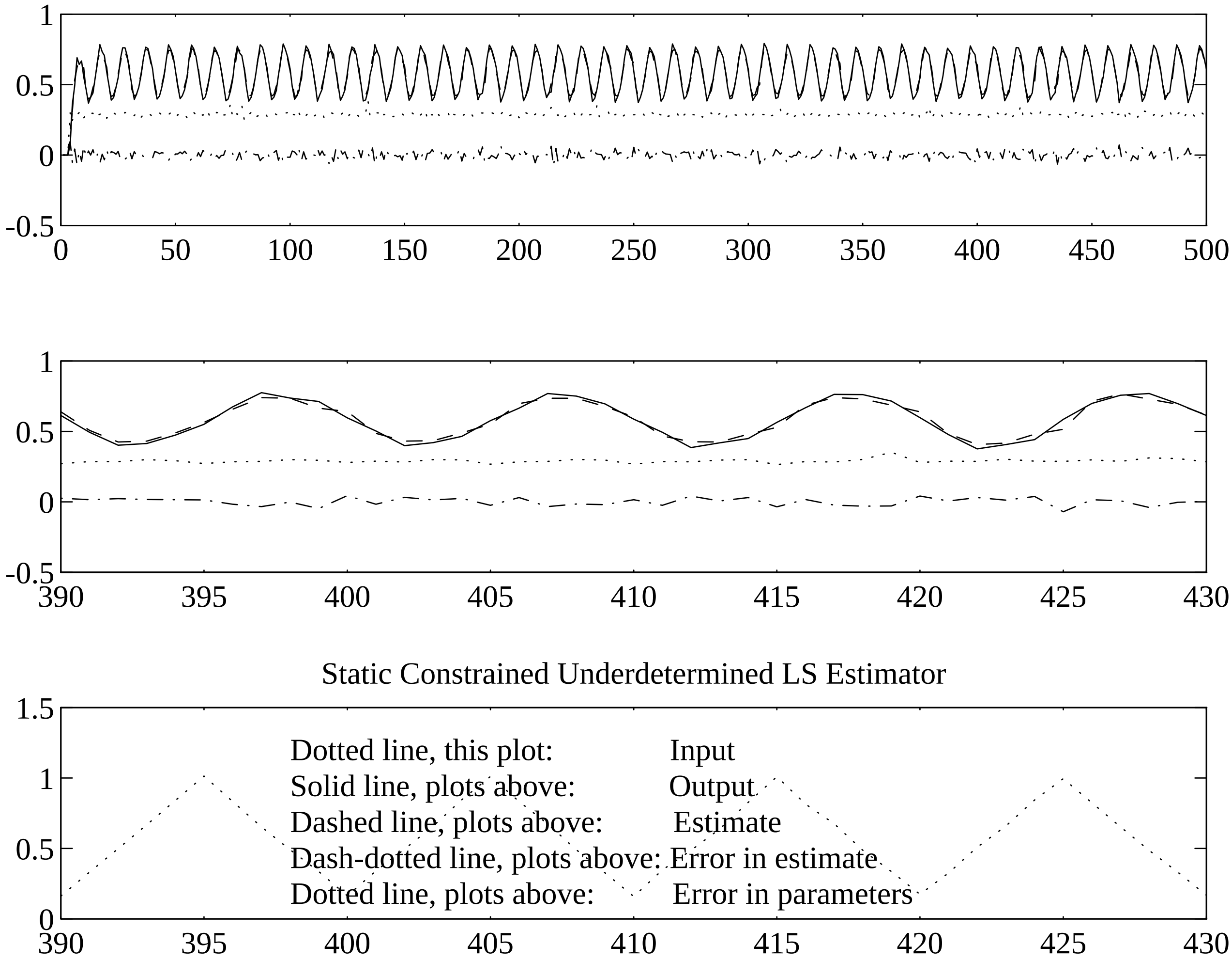

Figure 6: Triangular Input to SCULS Algorithm

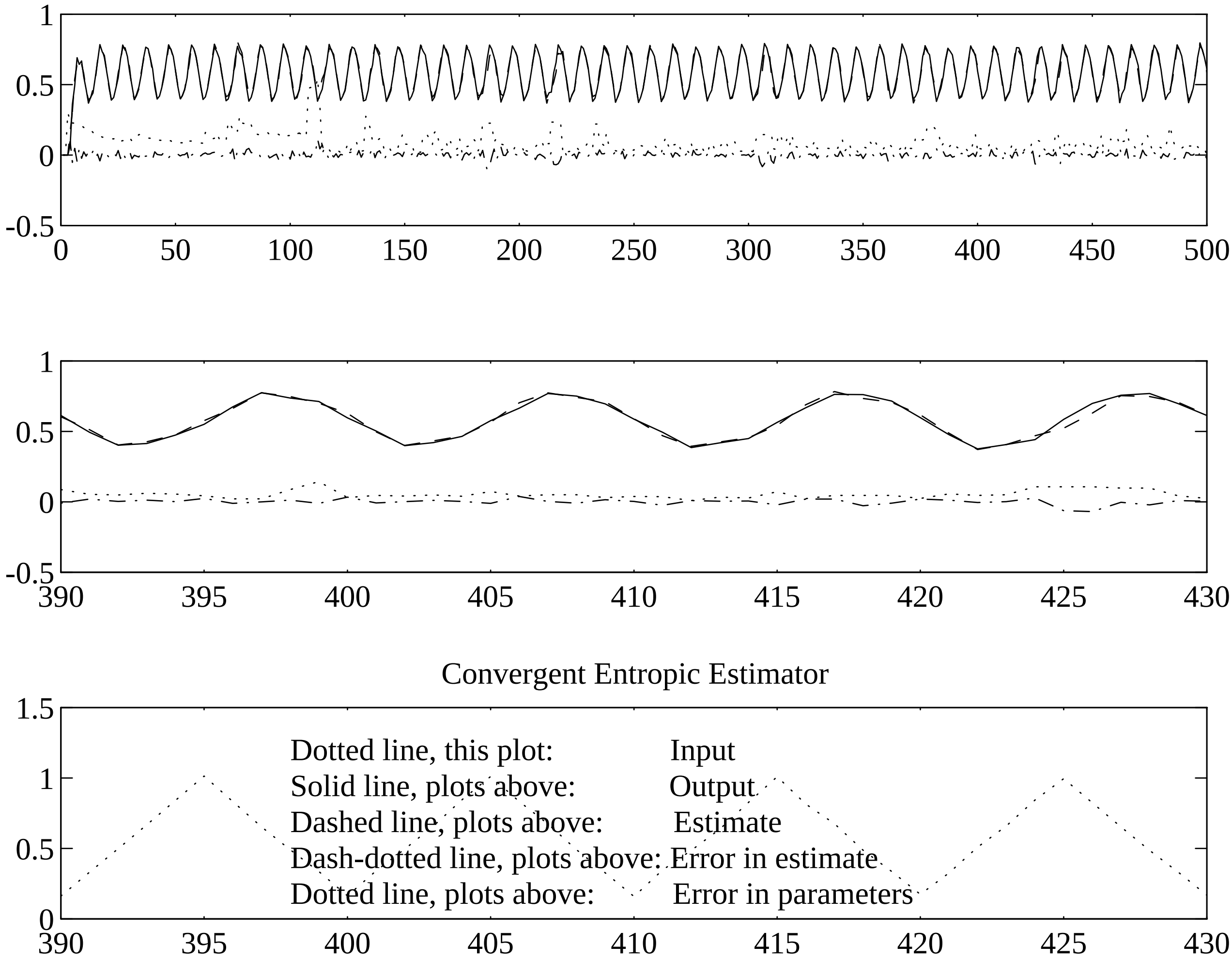

Figure 7: Triangular Input to Convergent Entropic Algorithm

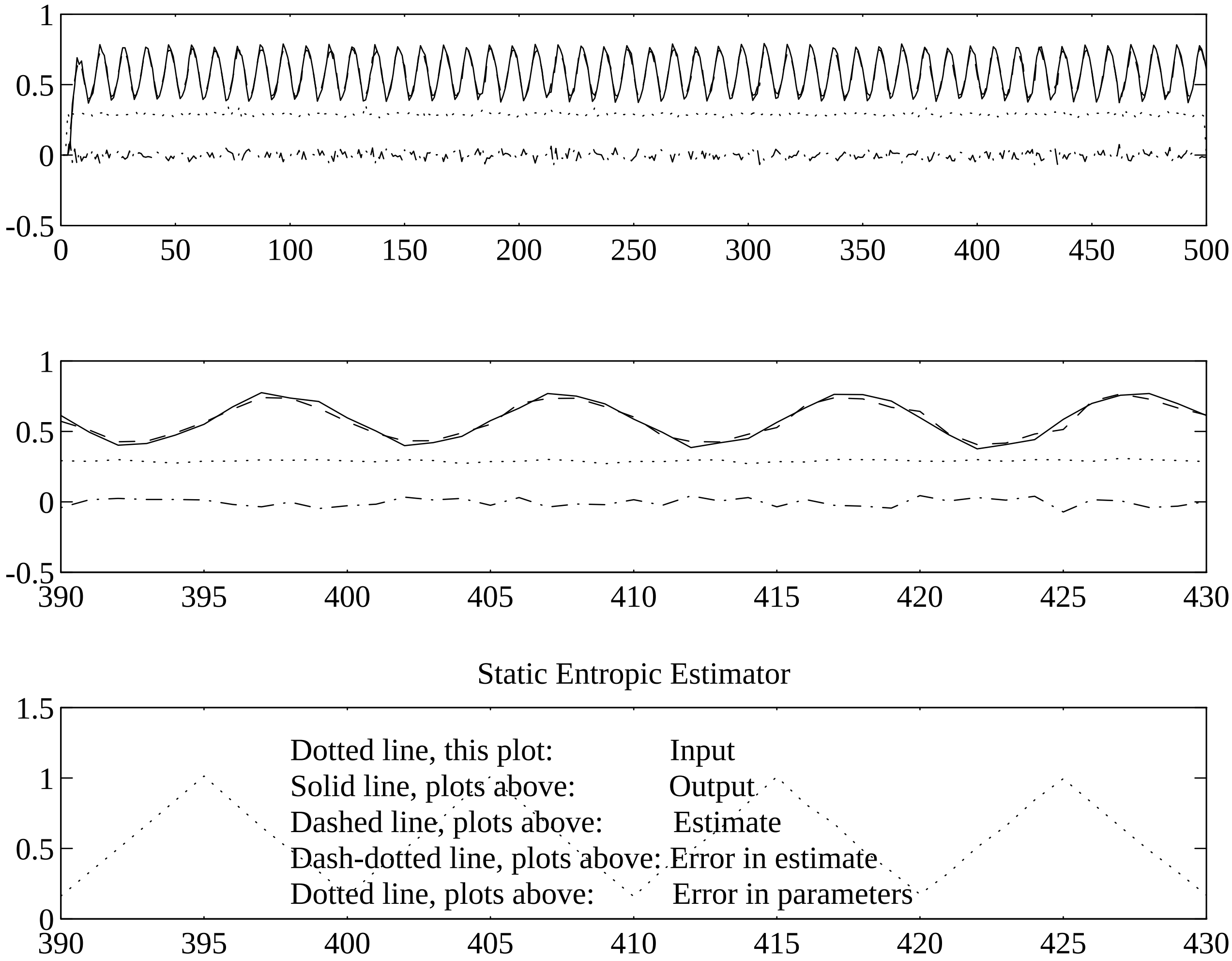

Figure 8: Triangular Input to Static Entropic Algorithm

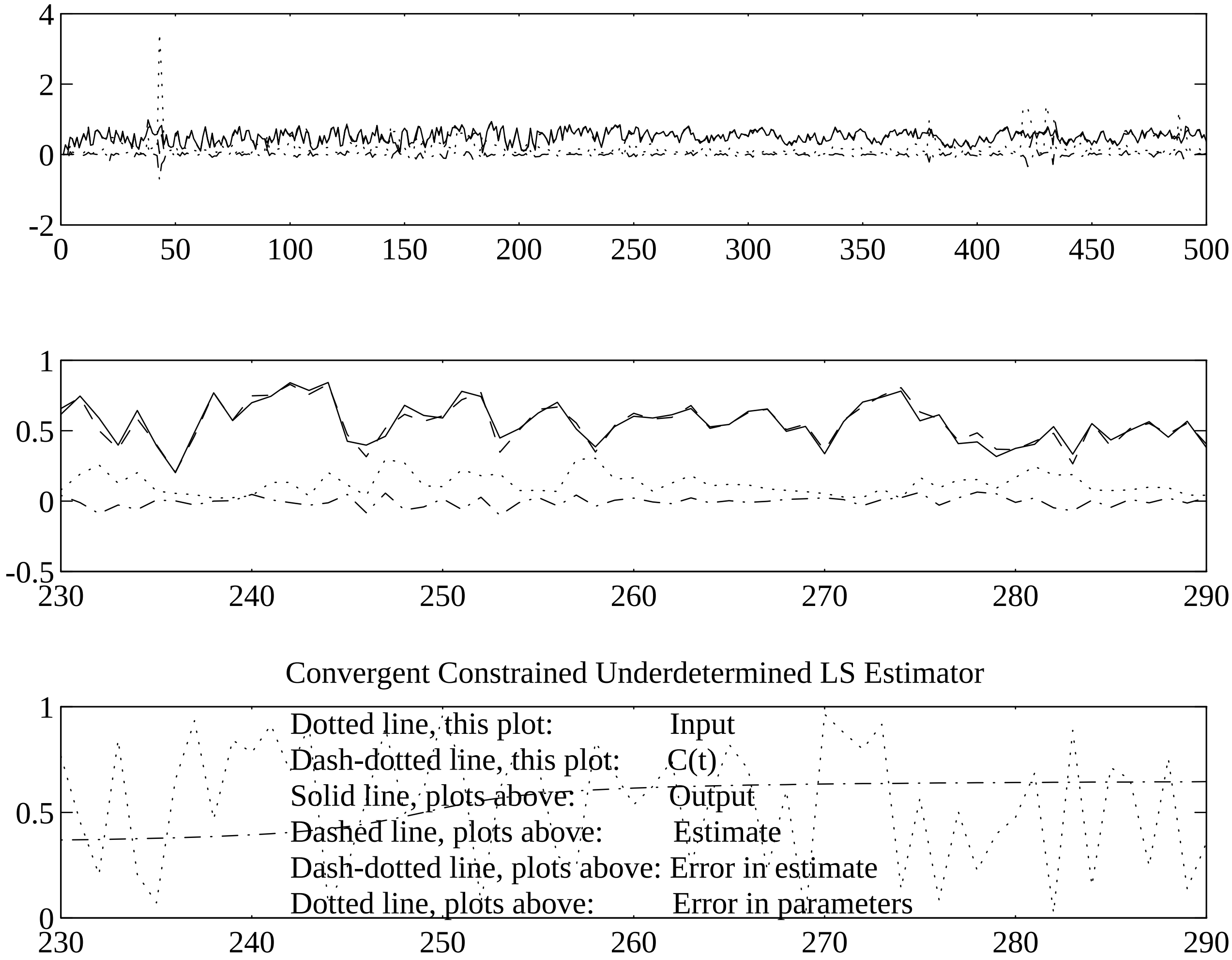

Figure 9: Bursting Phenomenon in CCULS Algorithm

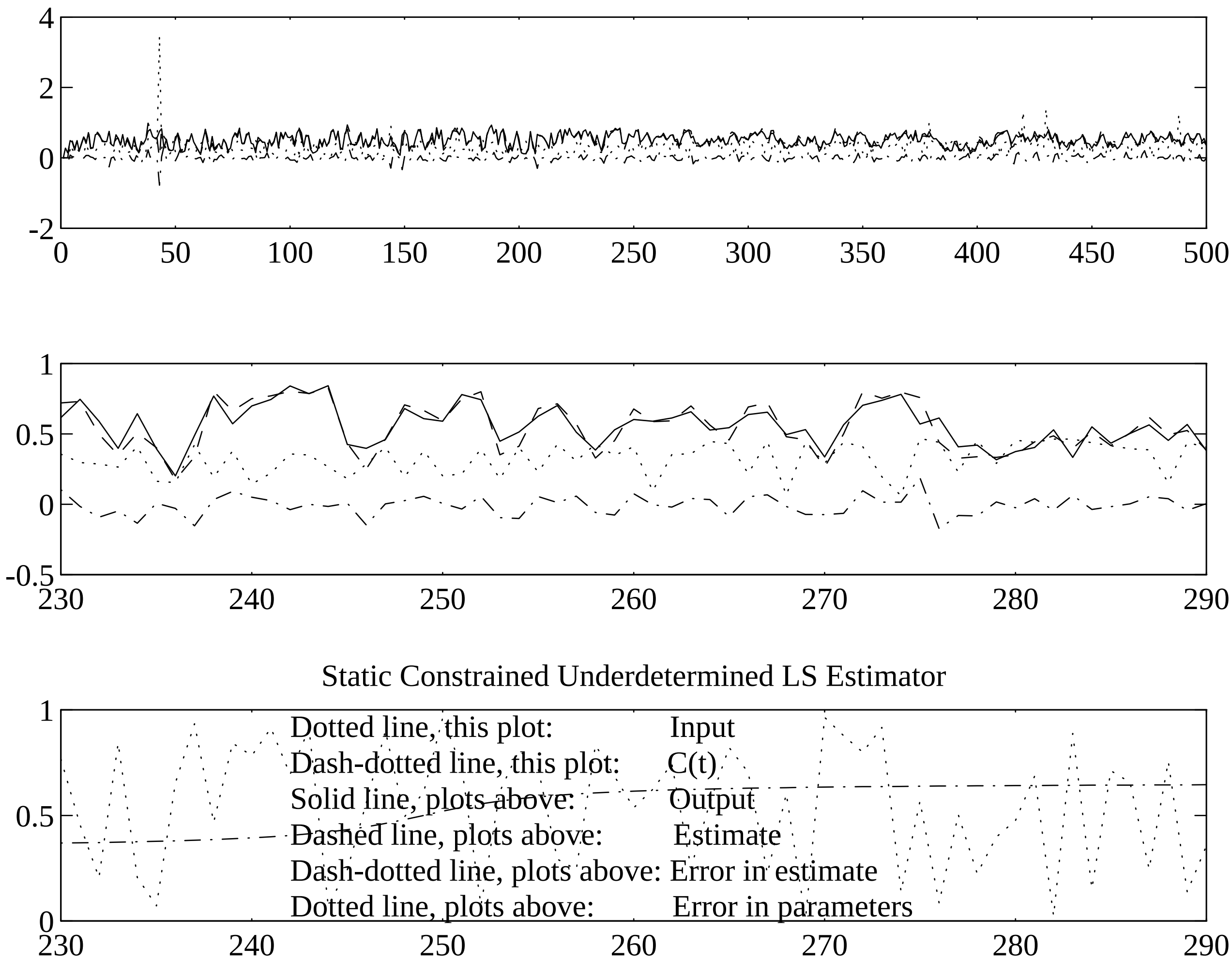

Figure 10: Bursting Phenomenon in SCULS Algorithm

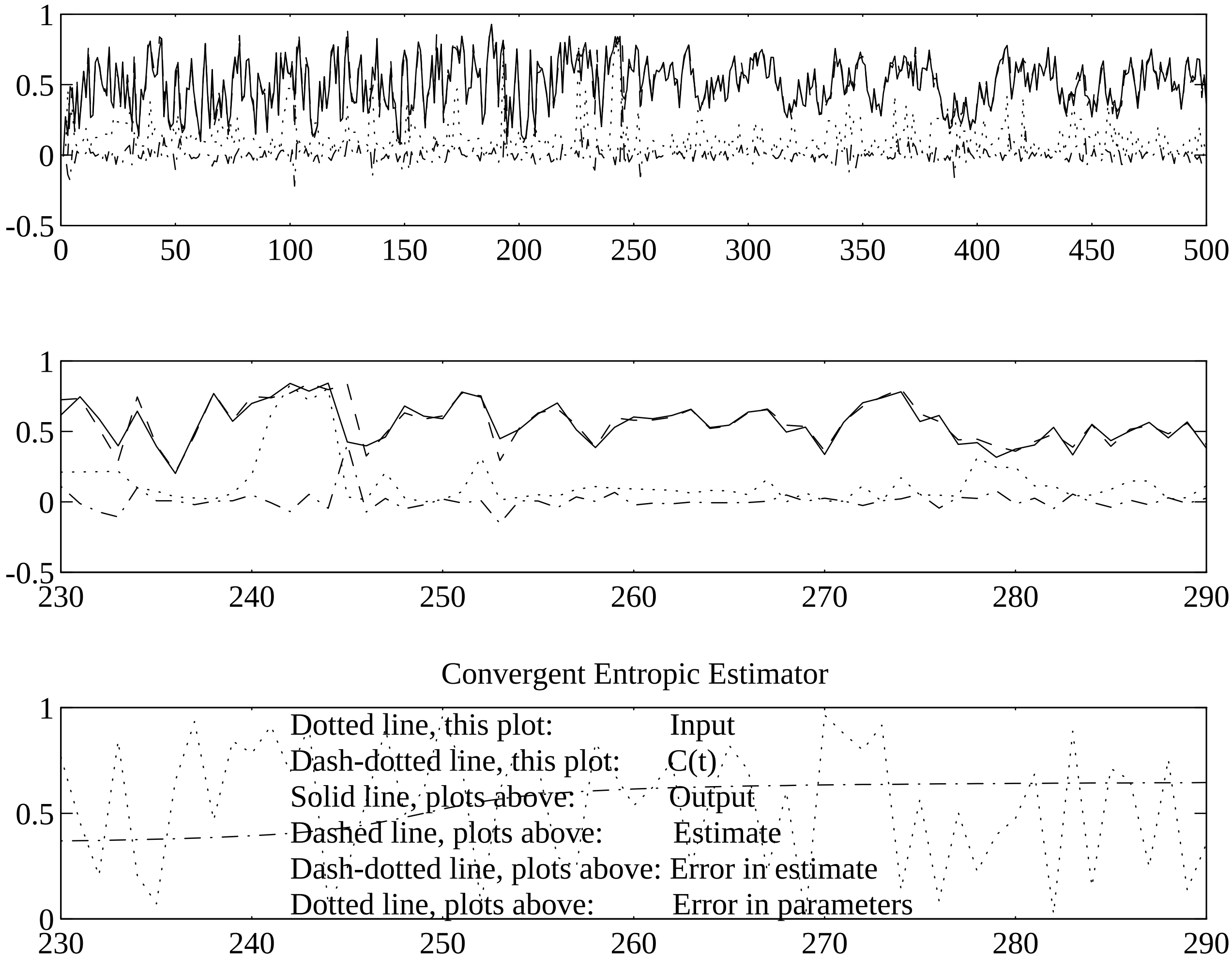

Figure 11: Bursting Phenomenon in Convergent Entropic Algorithm

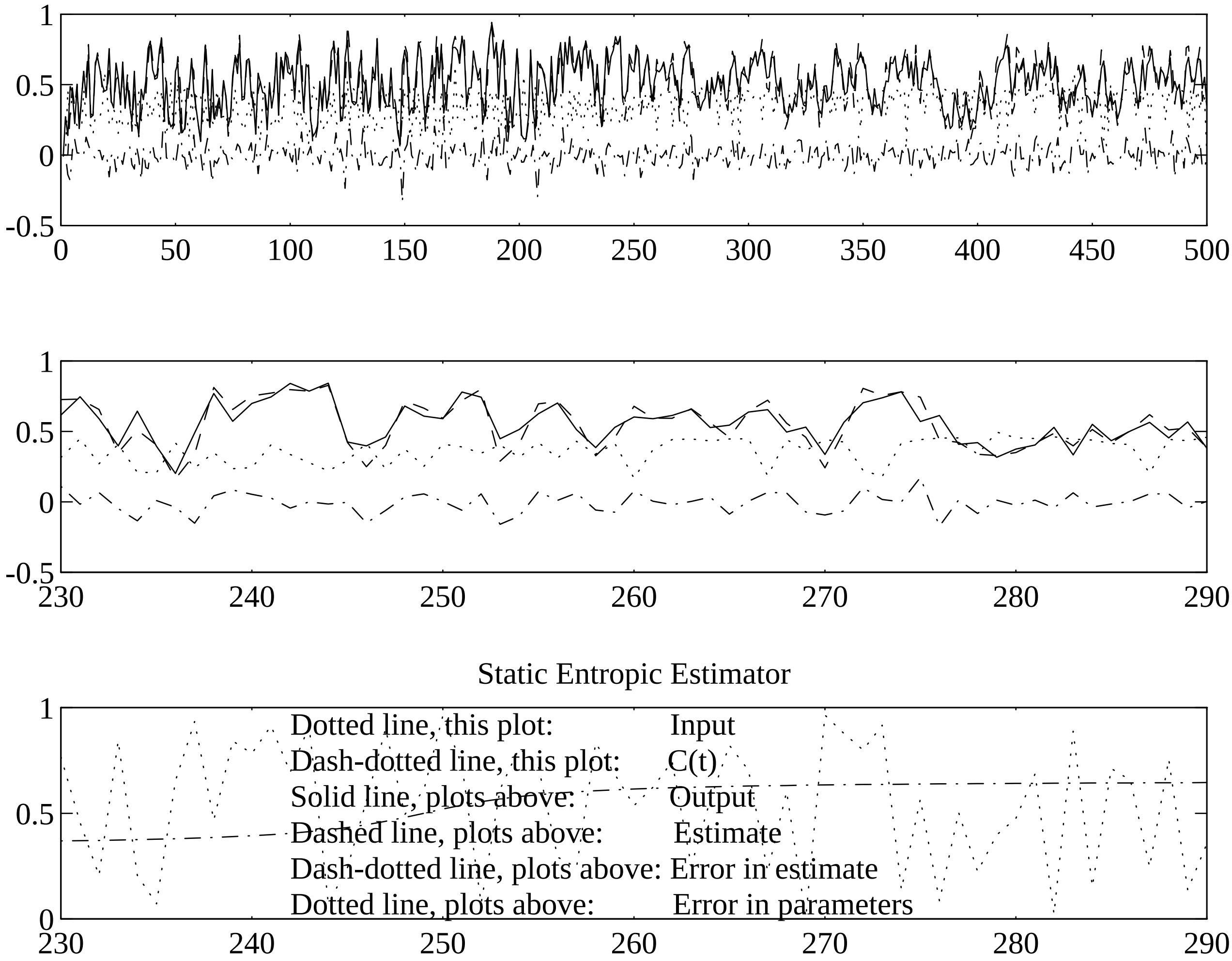

Figure 12: Bursting Phenomenon in Static Entropic Algorithm