BAYESIAN SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

John Skilling
Department of Applied Mathematics and Theoretical Physics
Silver Street
Cambridge CB3 7HJ
England

ABSTRACT. In the numerical solution of ordinary differential equations, a function $y(x)$ is to be reconstructed from knowledge of the functional form of its derivative: $dy/dx = f(x,y)$, together with an appropriate boundary condition. The derivative f is evaluated at a sequence of suitably chosen points (x_k, y_k), from which the form of $y(\cdot)$ is estimated. This is an inference problem, which can and perhaps should be treated by Bayesian techniques. As always, the inference appears as a probability distribution $\text{prob}(y(\cdot))$, from which random samples show the probabilistic reliability of the results. Examples are given.

1. Introduction

The classic ordinary differential equation problem is the simplest initial value problem

$$\text{"Given } dy/dx = f(x,y) \text{ and } y(0) = 0, \text{ find } Y = y(X)\text{"}. \tag{1}$$

A variant on this is to require

$$y(x) \;\forall\; 0 \le x \le X \tag{2}$$

We are to solve equation (1) numerically and approximately, using as few evaluations of f as possible. The special case where f is independent of y reduces to simple integration, but when f depends on y as well, the problem is more awkward. The difficulty is that we never know the correct value of y at any selected abscissa x_k. When we estimate it, as y_k say, we will almost certainly be wrong. Hence our evaluations of data,

$$D_k = f(x_k, y_k) \tag{3}$$

will not measure dy/dx accurately. In other words, we remain uncertain of precisely what we are integrating. Clearly this is an inference problem of some subtlety. Probability calculus being the only calculus for consistently reasoning from incomplete information, we should nevertheless recommend it and try to use it in such problems.

This paper suggests ways of setting up a Bayesian treatment. Far from considering these suggestions to be the last word, the author considers them to be merely pointers to what could, and perhaps should, develop into an open field of research, complementing and if appropriate superseding the traditional algorithms.

C. R. Smith et al. (eds.), Maximum Entropy and Bayesian Methods, Seattle, 1991, 23–37.

As always, Bayesian analysis requires

a) a hypothesis space $\{h\}$ which will here involve the set of admissible functions $y(\cdot)$, augmented by such extra parameters as may be required;
b) a prior probability distribution $\mathrm{pr}(h)$ over these hypotheses;
c) the likelihood $\mathrm{pr}(D|h)$, being the probability of acquiring any supplied data constraints D, given one of the hypotheses. Bayes' theorem, that

$$\begin{aligned}\mathrm{pr}(h,D) &= \mathrm{pr}(h)\,\mathrm{pr}(D|h) \equiv \text{prior} \times \text{likelihood} \\ &= \mathrm{pr}(D)\mathrm{pr}(h|D) \equiv \text{evidence} \times \text{inference}\end{aligned} \tag{4}$$

(conditional on whatever background information $\mathcal{I}$ is currently assumed) allows us to draw probabilistic conclusions in the form of the overall evidence

$$\mathrm{pr}(D) = \int dh\, \mathrm{pr}(h,D) \tag{5}$$

and the posterior inference

$$\mathrm{pr}(h|D) = \mathrm{pr}(h,D)/\mathrm{pr}(D) \tag{6}$$

The evidence, more properly written as $\mathrm{pr}(D|\mathcal{I})$, allows us to assess any rival choices for $\mathcal{I}$.

It may at first seem optimistic to attempt this approach, which immediately involves integration over large spaces, and we may surmise this to be why numerical analysis did not historically take this path. However, we shall try to obey the Bayesian rules.

2. Spatial Correlation in the Hypothesis Space

Presumably f is reasonably smooth in x and y, otherwise its sampled values would offer inadequate guidance to construct $y(x)$, and the problem would be hopeless from the start. This is a fundamental property which must be part of our treatment. We assign

$$W = \text{length-scale with respect to } x, \tag{7}$$

representing the variability in f, which will ultimately be reflected in our inference about the solution $y(\cdot)$. Similarly, we assign a stiffness constant

$$c = \|\partial f/\partial y\| \tag{8}$$

to quantify the sensitivity of f to variations in y. Although c has dimensions formally inverse to W, it seems better to keep the two parameters independent.

Experience in other areas of scientific data analysis suggests that the natural way of imposing a preference for smoothness in x is to introduce an intrinsic correlation function (ICF) which explicitly smooths some sharper, hidden function by averaging it over the width W. Here, it is the differential y' which is assumed to be smooth, rather than y itself, so we take y' to be a smoothed version of some hidden function $h(x)$. Specifically, we shall suppose

$$y'(x) = \int_{-\infty}^{\infty} dx\, R(x,x)\, h(x)$$

or, in shorter matrix notation,

$$y' = Rh \tag{10}$$

where

$$R(x,u) = e^{-(x-u)^2/2W}/\sqrt{2\pi W} \tag{11}$$

Purely for convenience, this ICF R is given the form of a Gaussian convolution of width W, which incidentally requires h to be defined for negative as well as positive x. This approach to smoothness differs from the traditional approach of assuming some limit on some higher differential, typically $\|y''\|_\infty$.

There are reasons for preferring the ICF approach.

a) We seek the macroscopic structure of $y(\cdot)$, for which local irregularities in a high differential would scarcely be relevant, so that integration is a more natural representation of the smoothness property we need.
b) Explicit limits on high differentials are often unavailable, especially in non-trivial problems, where the algebraic form of a high differential may be excessively long.
c) Any prior probability distribution for a high differential would have to be supplemented by awkward subsidiary conditions on the corresponding constants of integration in order to reach a normalised prior for y itself.
d) Provided only that h is integrable ($h \in \mathcal{L}_1$), its smoothed form Rh shares all the continuity and differentiability properties of R, so that when R is Gaussian, Rh is automatically differentiable to all orders. On the other hand, the (realistic) error bars on the results from this approach lack the absolute force of the (pessimistic) worst case bounds which the traditional approach can give in favourable examples.

It follows from the assumed form of y' that y is yet smoother, effectively being h convolved with a smoothed step function instead of the smoothed delta function represented by the Gaussian function R. More precisely, we integrate (using $y(0) = 0$) to obtain

$$y(x) = \int_0^x dz\, y'(z) = \int_{-\infty}^{\infty} du\, Q(x,u)\, h(u) \tag{12}$$

or, in shorter matrix notation,

$$y = Qh \tag{13}$$

where

$$Q(x,u) = \int_0^x dz R(z,u) = \tfrac{1}{2}\Big(\mathrm{erf}\big((x-u)/\sqrt{2}W\big) + \mathrm{erf}\big(u/\sqrt{2}W\big)\Big) \tag{14}$$

Q may be viewed as the ICF for y itself, as opposed to y'.

As an alternative, we can assume that y is usefully twice differentiable, meaning that it is y'' which is the convolution with the original ICF R. In order to fix the second constant of integration, we then use the initial datum $D_0 = f(0,0)$. This gives us a different hypothesis about the form of solution y to be expected. Indeed, we could go further, and let yet higher derivatives be the convolution with R (though the treatment of the extra constants of integration would become less clear). The point at issue is not whether or not the derivatives *exist*, because that will be determined by the differentiability of R, but rather whether or not the derivatives are *useful*. The "order" of these Bayesian algorithms can

Thus the *first order* Bayesian algorithm has $y' = Rh$ as in (9), the *second order* Bayesian algorithm has $y'' = Rh$ instead, and so on.

3. The Prior

Assuming, as above, that the only useful differential of y is the first, we need a quantified prior on the hidden function $h(\cdot)$. For tractability, we adopt a Gaussian with parameter α on a flat (unit) measure. In matrix notation, using a large number M of closely spaced values of x,

$$\mathrm{pr}(h)\, d^M h = \sqrt{\det(\alpha I/2\pi)}\; e^{-\alpha h^T h/2}\; d^M h \tag{15}$$

where

$$h^T h = \int du\; h(u)^2, \tag{16}$$

$$\det(\alpha I/2\pi) = (\alpha/2\pi)^M \tag{17}$$

I being the unit matrix. By placing h almost certainly in $\mathcal{L}_2$, it must almost certainly be in $\mathcal{L}_1$ as well, as required. As usual, the value of α needs to be discussed. Its dimensions are

$$[\alpha] = 1/[h^T h] = 1/[x h^2] = [x/y^2]. \tag{18}$$

Properly, α should be given a normalised prior biassed towards values of the order of

$$\alpha \sim \frac{(\text{expected scale in } x)}{(\text{expected scale in } y)^2} \tag{19}$$

In practice, once several data have been obtained the posterior distribution of α tends to be fairly sharply peaked around its maximum, so that it suffices to fix α at this maximising value. This completes the preliminaries. The hypothesis space has been set up, involving a hidden function h related to the required function y by an ICF of assumed Gaussian form and assumed width W. On this space, the prior on h has been defined, conditional upon the parameter α which can effectively be fixed *a posteriori*. The stiffness parameter c is also available to enter the analysis at a later stage.

4. The Likelihood

At any stage in the estimation of y, we have a list of coordinates (x_k, y_k) at which f has been evaluated as $D_k = f(x_k, y_k)$. Even at the very beginning, we may allow ourselves $D_0 = f(0,0)$. These data will (somehow) modify our estimates of the hidden function h, and thence of y and the ultimate required value $Y = y(X)$. However, we need to know how accurately the data represent the slope y' of the true curve $y(x)$.

Suppose that we can (somehow) use the data to estimate a mean $\hat{y}(x)$ and standard deviation $\delta\hat{y}(x)$, allowing us to infer

$$y(x_k) = \hat{y}_k \pm \delta\hat{y}_k \tag{20}$$

at each measurement abscissa x_k. Awkwardly, the measurement ordinate y_k will usually be different again. Although one hopes that it was chosen intelligently, it will

have been selected on the basis of less data than are currently available. Hence we may write the positional error of y_k, relative to the true curve, as

$$\Delta y_k \equiv y_k - y(x_k) = \big(y_k - \hat{y}_k\big) + \big(\hat{y}_k - y(x_k)\big) \tag{21}$$

On the right, $y(x_k)$ is the only unknown. Its estimated mean and variance lead us to the variance

$$\langle(\Delta y_k)^2\rangle = (y_k - \hat{y}_k)^2 + (\delta\hat{y}_k)^2 \tag{22}$$

describing our uncertainty about the true curve, relative to the measurement ordinate y_k. This uncertainty in position y can be rescaled through the stiffness constant to give a corresponding uncertainty in f:

$$\Delta f = c\Delta y \tag{23}$$

This represents the difference between the true slope $f\big(x_k, y(x_k)\big)$ at abscissa x_k and the measured slope $D_k = f(x_k, y_k)$. Accordingly, D_k measures the true slope at abscissa x_k, with an error σ_k given by

$$\sigma_k^2 = c^2\big((y_k - \hat{y}_k)^2 + (\delta\hat{y}_k)^2\big) \tag{24}$$

Datum D_k becomes interpreted as a noisy measurement of the true slope, just as in ordinary data analysis. The main difference from ordinary data analysis is that the data uncertainties are not fixed in advance. The errors depend on the current inference through $\hat{y}$ and $\delta\hat{y}$, and these are likely to evolve as more data are acquired. However, we shall shortly see how to construct $\hat{y}$ and $\delta\hat{y}$ from a given dataset, from any given errors σ. A few iterations of this allows $\hat{y}$ and $\delta\hat{y}$ to be determined self-consistently along with σ: there seems no reason to expect any instability in this scheme.

Given a set of errors σ, albeit provisional, as well as the data D, we can finally construct the likelihood. In its usual Gaussian form,

$$\mathrm{pr}(D|h) = \sqrt{\det(\sigma^{-2}/2\pi)}\; e^{\frac{1}{2}(Rh-D)^T\sigma^{-2}(Rh-D)} \tag{25}$$

where σ^{-2} is the $n\times n$ inverse covariance matrix of the errors. In this preliminary study, we take σ^{-2} to be diagonal, ignoring any possible correlation between the errors of different data: it is likely that this simplification detracts from the quality of the results.

5. Evidence and Inference

Both the prior $\mathrm{pr}(h)$ and the likelihood $\mathrm{pr}(D|h)$ are now available as Gaussians in h, albeit dependent on the iteratively-derived errors σ. Accordingly, the joint distribution is also Gaussian. Its integral gives the evidence

$$\mathrm{pr}(D) = e^{-\frac{1}{2}D^T B^{-1} D}/\sqrt{\det(2\pi B)} \tag{26}$$

where

$$B = \sigma^2 + RR^T/\alpha \tag{27}$$

The parameter α is chosen to maximise the value of the evidence. Also, the mean and covariance give

$$\langle \delta\hat{h}\delta\hat{h}^T \rangle = \left(I - R^T B^{-1} R/\alpha\right)/\alpha \tag{29}$$

¿From h, we can estimate the result $y = Qh$ and its derivative $y' = Rh$ as

$$\hat{y} = QR^T B^{-1} D/\alpha \tag{30}$$

$$\langle \delta\hat{y}\delta\hat{y}^T \rangle = \left(QQ^T - QR^T B^{-1} RQ^T/\alpha\right)/\alpha \tag{31}$$

$$\hat{y}' = RR^T B^{-1} D/\alpha \tag{32}$$

$$\langle \delta\hat{y}'\delta\hat{y}'^T \rangle = \left(RR^T - RR^T B^{-1} RR^T/\alpha\right)/\alpha \tag{33}$$

These results evaluated at the measurement abscissae define the errors σ self-consistently.

As a consequence of using Gaussian distributions, we have been able to write the "visible" results y and y' without using M-dimensional "hidden" space at all: the matrices Q and R which contain this dimension only appear in the combinations

$$\begin{aligned} RR^T(x,z) &= \int_{-\infty}^{\infty} du\, R(x,u)\, R(z,u) \\ &= e^{-(x-z)^2/4W^2}/2\sqrt{\pi}W \end{aligned} \tag{34}$$

$$\begin{aligned} QR^T(x,z) &= \int_{-\infty}^{\infty} du\, Q(x,u)\, R(z,u) \\ &= \tfrac{1}{2}\left(\mathrm{erf}((z-x)/2W) + \mathrm{erf}(x/2W)\right) \end{aligned} \tag{35}$$

$$\begin{aligned} QQ^T(x,z) &= \int_{-\infty}^{\infty} du\, Q(x,u)\, Q(z,u) \\ &= \tfrac{1}{2}\left(x\,\mathrm{erf}(x/2W) + z\,\mathrm{erf}(z/2W) - (x-z)\,\mathrm{erf}((x-z)/2W)\right) \\ &\quad + \pi^{-\frac{1}{2}} W\left(e^{-x^2/4W^2} + e^{-z^2/4W^2} - e^{-(x-z)^2/4W^2} - 1\right) \end{aligned} \tag{36}$$

These combinations operate on the n-dimensional vector D and $n\times n$ matrix B. Thus the dimensionality of the matrix calculations is only n (the number of data), augmented perhaps by the number of abscissae at which results are required.

We can now estimate the desired result $Y = y(X)$, with its probabilistic uncertainty, as

$$\hat{Y} = \hat{y}(X) \pm \delta\hat{y}(X) \tag{37}$$

If the uncertainty $\delta\hat{y}(X)$ is acceptably small, all well and good. But if the uncertainty is unacceptably large, we must acquire more data, by evaluating more values of $f(x,y)$. This raises the question of where such samples should be taken.

6. Sampling Strategy

In numerical analysis, as opposed to more examples of scientific data analysis, we often have freedom to place our data samples where we wish. In our ordinary differential equation problem, we can select our next sample, or samples, at any arbitrary coordinates x and y. Presumably these samples should be chosen sensibly, rather than just being purely random. Indeed, having selected a value ξ for x, then the selection $\hat{y}(\xi)$ seems appropriate for y, in order to minimise the error of σ of this new measurement, as expected on the basis of the currently available data. However, the choice of ξ is less ob-

One could use some pre-assigned strategy. The simplest would be to set some fairly small step-length Δx, and steadily increase ξ by this amount. For extra safety, one could re-sample y a few times at each chosen x. One could also emulate Runge-Kutta by returning to the interior of the current step and re-sampling within. In fact any traditional strategy could be incorporated into the Bayesian formulation, because any strategy whatever will produce a dataset which the Bayesian can interpret.

However it is more interesting, and may ultimately prove more productive, to seek that sample abscissa ξ for which the new datum is expected to be *most informative* about our result Y. In this way, we may hope to gain information as rapidly as possible, and thus converge to an accurate result with fewest samples.

If we were to measure at a particular place ξ, we would refine our result from (say)

$$Y = \mu_0 \pm \sigma_0 \tag{38}$$

to (say)

$$Y = \mu_1 \pm \sigma_1 \tag{39}$$

both uncertanties being Gaussian. The natural measure of information gleaned in this refinement is the (positive) cross-entropy

$$\begin{aligned} S &= \int dY \ \mathrm{pr}_1(Y) \log\left(\mathrm{pr}_1(Y)\mathrm{pr}_0(Y)\right) \\ &= \left((\mu_1-\mu_0)^2+\sigma_1^2-\sigma_0^2\right)/2\sigma_0^2 - \log(\sigma_1/\sigma_0) \end{aligned} \tag{40}$$

As one would hope, this formula demonstrates a gain both for the reduction in variance σ^2 occasioned by acquiring more data, and for any change in the expectation value μ.

Suppose that we were to obtain a new datum $D^* = f(\xi, \hat{y}(\xi))$ at our selected abscissa ξ. Ignoring any iterative refinement of errors σ in the light of the new value D^*, we could already assign its uncertainty σ^* on the basis of the expected misfit $\delta\hat{y}(\xi)$ between our current estimate $\hat{y}(\xi)$ and the true curve $y(\xi)$. Accordingly, we could update the inference for Y, and obtain the corresponding information gain S^*. Until we perform the evaluation, we do not know the actual value of D^*, but we *can* predict what D^* measures, namely the slope $y'(\xi)$ which we expect to observe at ξ. Its mean and variance are the results (31) and (32) for y', evaluated at ξ. Averaging over this expected range of values for D^* gives us the expectation information $\langle S^*(\xi)\rangle$ which we *expect* to gain *if* we were to sample at ξ. The averaging can be carried out analytically, so that $\langle S^*(\xi)\rangle$ can be computed reasonably quickly for any ξ. It is then straightforward to evaluate it at sufficient trial values of ξ to locate its maximum.

We use this maximum as the new sample abscissa ξ, and proceed to the actual new measurement, augmenting the current dataset by

$$D_{n+1} = f\left(\xi, \hat{y}(\xi)\right) \tag{41}$$

(An alternative procedure would be to treat $S^*(\xi)$ as an additive distribution, and sample ξ randomly from it.) We now have a full, usable algorithm for solving the original problem (1). It involves evaluating f at successive points $\left(\xi, \hat{y}(\xi)\right)$ which can be chosen to be optimal for the specific f in hand, and then repeating the process until the estimate of the final

7. Deletion of data

As the number n of data increases, the cost of their acquisition in terms of evaluations of f clearly increases in direct proportion. However, the overheads increase much faster, with the inversion of B being $\mathcal{O}(N^3)$ and the evaluations of $\langle S^* \rangle$ being $\mathcal{O}(N^2 n)$. If the cost of the overheads is not to overwhelm the cost of evaluating f, there must be a mechanism for deleting points from the dataset, as well as including new ones. Presumably, the point to be deleted should be that one which causes least damage to the result Y. Again, such degradation can be quantified in terms of the cross-entropy between the initial (with all data) and the degraded (with one datum deleted) probability distributions of Y. Given a choice, we shall choose to delete that datum for which the cross-entropy damage is *least.*

According to this scheme, the algorithm will proceed to loop indefinitely if the most recently added point turns out to be the least informative. When that happens, it seems that the algorithm has proceeded as far as it can with its restricted size of dataset. This gives a natural termination criterion.

8. Probabilistic displays

Finally, it is always interesting, and can be useful, to display samples from the current posterior probability distribution for y. Indeed, this seems to be the only way of readily visualising the current state of the algorithm.

The first step is to compute a sample of the hidden function h on a discrete grid of M abscissae x. Introduce vectors a and b whose M and n components (respectively) are each independently drawn from the unit normal distribution $\mathcal{N}(0,1)$, so that their statistics are

$$\langle a \rangle = 0, \quad \langle aa^T \rangle = I \tag{42}$$

$$\langle b \rangle = 0, \quad \langle bb^T \rangle = I \tag{43}$$

Then it can be straightforwardly verified that the M-dimensional vector

$$\tilde{h} = R^T B^{-1} D/\alpha + \alpha^{-\frac{1}{2}} \left(a - R^T B^{-1} (Ra + \alpha^{\frac{1}{2}} \sigma b)/\alpha \right) \tag{44}$$

has the correct statistics (27) and (28) to be a random sample of y. After this has been calculated, application of $M \times M$ versions of the matrices Q and R yield the required sample

$$\tilde{y} = Q\tilde{h} \tag{45}$$

and, should it be needed, the corresponding derivative

$$\tilde{y}' = R\tilde{h} \tag{46}$$

The accompanying figures show overlaid plots of a few such random samples $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3, \ldots$ from the full result $\mathrm{pr}(y(\cdot))$ in particular examples.

However, a more dynamic and visually effective presentation can be achieved on a computer screen. Instead of superposing the original samples, generate a related sequence

$$\tilde{y}^*_1 = \tilde{y}_1, \quad \tilde{y}^*_{i+1} = \tilde{y}_i \cos\theta + \tilde{y}_{i+1} \sin\theta$$

where $0 < \theta < \pi/2$. Each $\tilde{y}^*(\cdot)$ is itself a random sample from $\mathrm{pr}(y(\cdot))$, but successive samples are correlated according to $\cos\theta$. The movie, in which successive frames show successive $\tilde{y}^*(\cdot)$, shows a single sample from $\mathrm{pr}(y(\cdot))$ undergoing Brownian diffusion through the probability cloud, with mean free path governed by $\sin\theta$. Both the structure and the variability of the results are clearly seen by watching $\tilde{y}^*$ in such movies.

9. Results

Table 1 shows three test problems, all with $X = 1$, and with $n = 20$ evaluations of f without deletion.

Problem	Differential $f(x,y)$	Solution $y(x)$	Result $Y = y(1)$
A	$(1+y)/(1+x^2)$	$(1+y_0)\exp(\tan^{-1}x) - 1$	1.1933
B	$2\pi\cos(2\pi x)$	$y_0 + \sin(2\pi x)$	0
C	$5(y + .08 - x^2)$	$x^2 + .4 * x + y_0 e^{5x}$	1.4000

Table 1: Test Problems

Problem A is a preliminary easy test, with a stable, nearly straight solution. Problem B is a little harder, in that the solution is oscillatory. Problem C is stiff: although the desired solution is a simple quadratic, any error becomes amplified by up to $e^5 \simeq 150$ by the time x reaches 1.

Firstly, each problem was solved by the first order Bayesian algorithm, in which y' is obtained from the hidden function h by smoothing it as in equation (8). Problems A and B were both assigned width $W = 0.2$ and stiffness $c = 1$. The rationale for the choice of width was that the functional form of f seemed fairly smooth, so that W might be fairly large. On the other hand, the solution y should not be artificially forced to be too smooth, so that W should not be too large. The rationale for the choice of stiffness was simply that a number of order unity seemed appropriate for a problem in which all variables were of that order. Unusually in a Bayesian calculation, it is not possible to tune the parameter values W and c directly from the evidence $\mathrm{pr}(D|W,c)$. The reason is that different choices of parameters lead the algorithm to select different sample locations (x_k, y_k). With the datasets being different, the relative evidence values are meaningless. It might be possible to tune the parameters as the algorithm proceeds, using the existing data to refine their values, rather as α and σ are already refined: that is a matter for future investigation. In the present study, parameter values were accepted if they gave consistent and reasonably accurate results, for which the result $\hat{Y} \pm \delta\hat{Y}$ converged sensibly as data were acquired. The original values $W = 0.2$ and $c = 1$ were acceptable for Problems A and B.

Problem C, though, was different, because of its stiffness. Using the original values for W and c, the results as more data were acquired became seriously inconsistent. Thus Y was estimated as -1.0735 ± 0.0664 after 5 evaluations of f, but $+1.8701 \pm 0.0120$ after 20 evaluations. Meanwhile, the value of α bounced by a factor of 100. These are symptoms of assumptions which are badly matched to the actual problem in hand. Stability returned when the stiffness constant c was increased to 25, after which an increase in W to 0.6 gave lower estimated errors. These latter parameter values were used for Problem C. The "y' Bayes" column of Table 2 gives the results, which can be compared with the correct answers, and with the results from 5 applications of standard 4th-order Runge-Kutta integration,

Secondly, each problem was solved using the alternative second order model in which it is the second derivative y'' which is obtained from the hidden function h by smoothing with R. One might expect this method to prove better for problems in which the solution is reasonably straight, and this expectation is borne out. The results are shown in the "y'' Bayes" column of Table 2. The results for A and C have their quoted errors reduced by factors of about 3. On the other hand, the result for the oscillatory problem B has its quoted error increased instead. One can also note that all the Bayesian quoted errors are reasonably in accord with the actual deviations from the true answers, 4 out of 6 being within one standard deviation and all being within two.

Problem	W	c	y' Bayes	y'' Bayes	Runge-Kutta	True Y
A	0.2	1.	1.1924±.0028	1.1923±.0008	1.1933	1.1933
B	0.2	1.	0.0054±.0195	-0.0147±.0550	0 †	0
C	0.6	25.	1.1680±.1437	1.3554±.0533	1.3294	1.4000

Table 2: Test Results. († Errors at intermediate steps $\simeq 0.0009$)

Some insight into the operation of the Bayesian algorithms can be gained from the Figures, which plot ten random samples from the posterior inference $\mathrm{pr}(y(\cdot))$ as data are acquired. Figure 1 shows the evolution of the inference for Problem A, according to the original first order "y'" algorithm. With a Gaussian prior underlying y', the solution y is expected *a priori* to be on average flat, but to random-walk away from the initial value as x increases. The initial datum D_0 modifies this within a width W or so of the origin, as seen in the first frame of Figure 1. Given this distribution of probable solutions, the most informative place to select the next sample is at (0.54,0.33). The next frame shows the effect on the posterior inference of this new datum. It turned out to have a slope rather more positive than would have been expected on average, so that the distribution of solutions was sloped (and consequently shifted) upwards. The subsequent sample, at (0.89, 0.81), behaves similarly, and completes a first, rough, sampling of the interval [0,1]. The most informative next sample interleaves the earlier ones, and appears at (0.30, 0.30). Again, the posterior distribution narrows, until after 20 samples it forms a narrow band barely distinguishable visually from the true analytic solution.

Figure 2 shows similar behaviour. With an oscillatory equation, though, more samples are needed to localise the posterior adequately. After two additional samples, the overall "up-down-up" behaviour of the final solution has already begun to appear, but both these points are a long way from the analytic solution. Ultimately being the least informative about the final solution, they would be the first samples to be deleted from the dataset if deletion were necessary.

Figure 3 shows the evolution of the posterior inference for Problem C. Here, the behaviour is rather different. Because the stiffness constant is relatively high, any sample which lies substantially away from the true solution will be almost useless as a measurement of the slope. Accordingly, the only safe strategy is to proceed in small, careful steps. Indeed, the strategy outlined above does this, incrementing x fairly steadily by about 0.03 each step, with only 2 out of the first 20 samples interleaving previous ones. After 20 samples, x has only been able to reach 0.46, so that the extrapolation to $x = 1$ relies purely on the relatively large ICF width W. Whether or not the extrapolation is accurate depends on whether or not the very precise linear combination Qh of error functions defined by the 20 closely spaced samples is sufficient to describe the solu-

Figures 4, 5, 6 show the posterior inferences for Problems A, B, C respectively, using the alternative second order "y''" algorithm. Whereas the first order algorithm favours solutions $y(\cdot)$ which random-walk away from some arbitrary constant, the second order algorithm favours solutions which random-walk away from an arbitrary straight line. This behaviour can clearly be seen in the Figures. The second order algorithm is better for Problems A and C, when the solution incorporates a linear trend, but worse for Problem B which does not have a net trend.

10. Conclusions

Solving differential equations numerically should be seen as an inference problem, in which results are to be estimated on the basis of incomplete data. Ideally, inference problems should be solved by Bayesian probabilistic methods. The advantage of this is that one obtains the probability distribution of solutions, including the uncertainties, whereas traditional algorithms compute a single solution only, with any error analysis grafted on. This paper shows that it is *possible* to compute with the full probability distribution, and it is possible to set up properly Bayesian algorithms, although the extra computational overheads from the matrix calculations are relatively severe in test problems.

There is a wide choice of Bayesian algorithms, and this paper is merely a preliminary foray. Indeed, there is an obvious inefficiency in the preliminary algorithms presented here, in that the sample points visibly cluster too closely in x. Presumably this is due to the practically convenient (but wrong) assumption that the errors σ_k on the data D_k are independent. If the correlations were allowed for, it could no longer be thought advantageous to place a new sample arbitrarily close to an old one, because duplicating an evaluation of f does not actually make it any more accurate. Thus the sampling strategy clearly needs improvement. There may be another inefficiency in the choice of a Gaussian for the form of the intrinsic correlation function. A Cauchy distribution, having larger wings, might well give better extrapolation, whereas some form of multiquadric might be more appropriate if y were to have isolated irregularities. The questions of the best Bayesian order, and of methods of choosing the width and stiffness constants, also remain open. Much work remains to be done, so that it is hardly surprising that the algorithms presented here are not yet competitive with traditional methods: Table 2 shows simple Runge-Kutta obtaining more accurate results in 5 out of 6 cases.

The examples shown here have been one-dimensional, with only one component of y. However, the extension to multi-dimensional y, with several components, is as straightforward here as it is with traditional methods. Moreover, the extension to second or higher order differential equations is entirely natural in these Bayesian schemes. An interesting test example for this would be the two-point boundary problem with the second-order Airy equation, $y'' = xy$ with $y(a)$ and $y(b)$ given.

Beyond ordinary differential equations lie partial differential equations. These too could be investigated from a Bayesian viewpoint. The future is wide open.

ACKNOWLEDGMENTS. This paper arose from long exposure to Laplace/Cox/Jaynes probabilistic reasoning, combined with the University of Cambridge's desire that the author teach some (traditional) numerical analysis. The rest is common sense. In this circumstance, the author hopes he will be forgiven for not making the usual explicit selection of formal references from the burgeoning literature. Simply, Bayesian ideas are "in the air", to the extent that the author judged

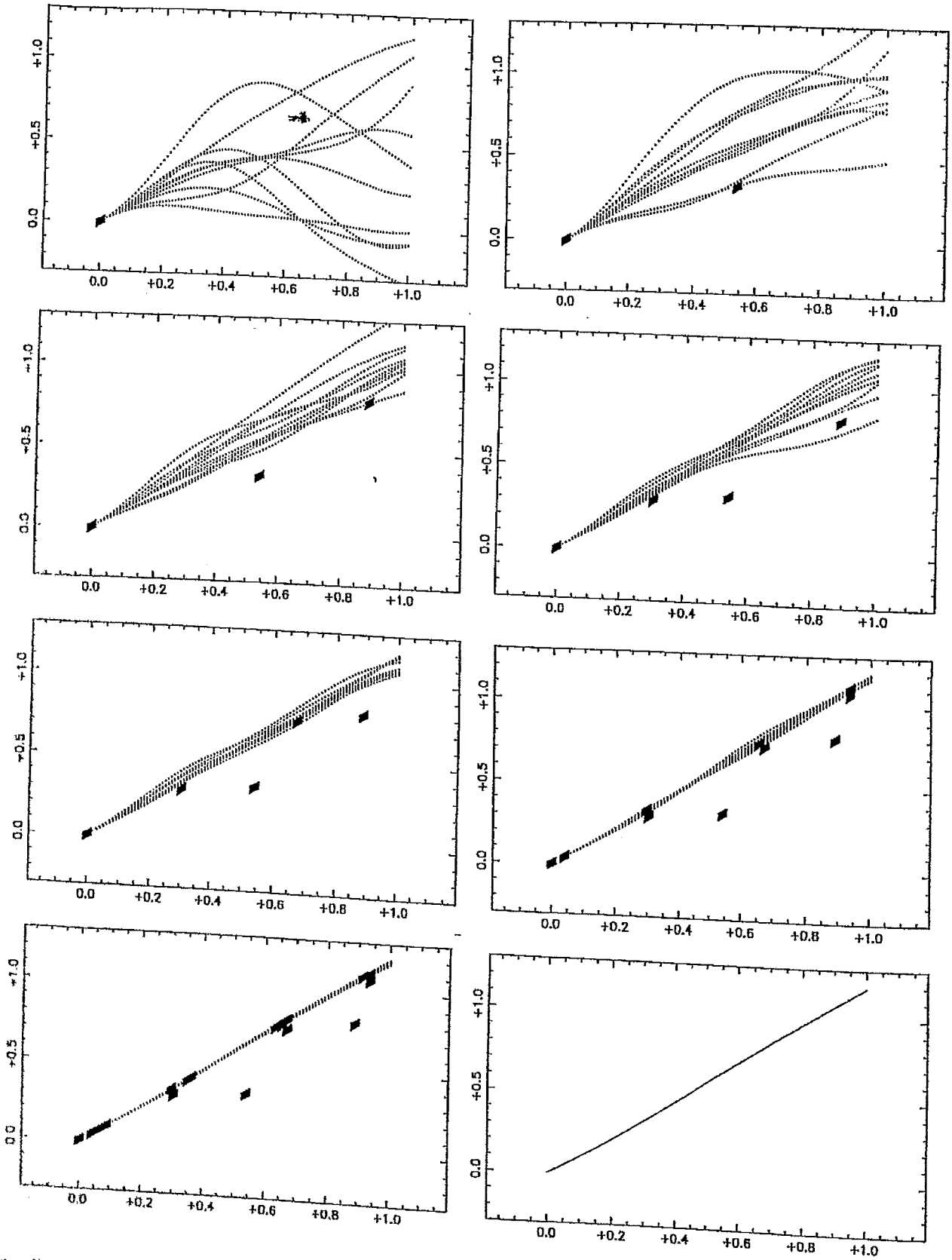

Fig. 1. Problem A, solved with the first order "y'" algorithm. The posterior distribution of $y(\cdot)$ after 1, 2, 3, 4, 5, 10, 20 samples, with the analytic solution.

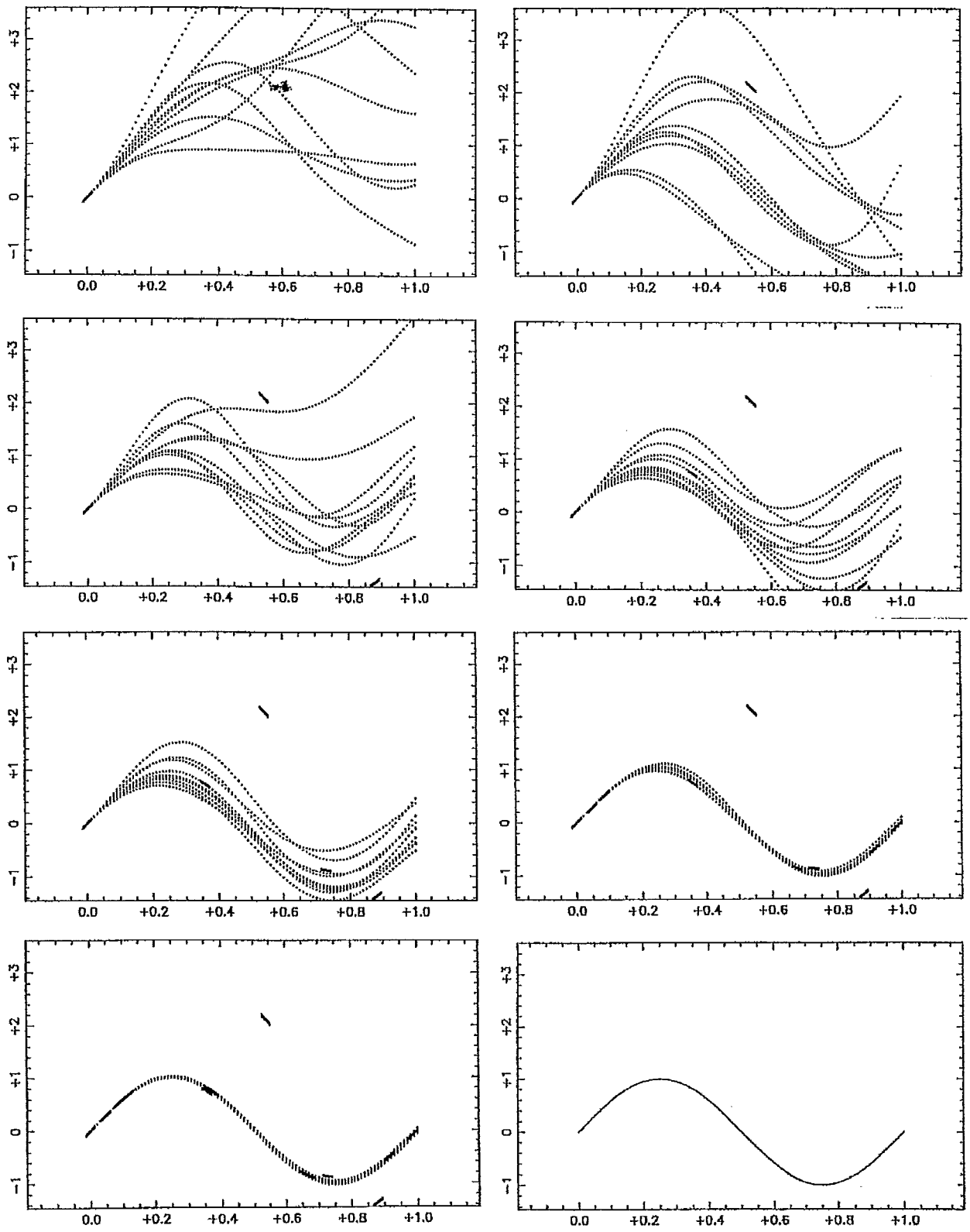

Fig. 2. Problem B, solved with the first order "y'" algorithm. The posterior distribution of $y(\cdot)$ after 1, 2, 3, 4, 5, 10, 20 samples, with the analytic solution.

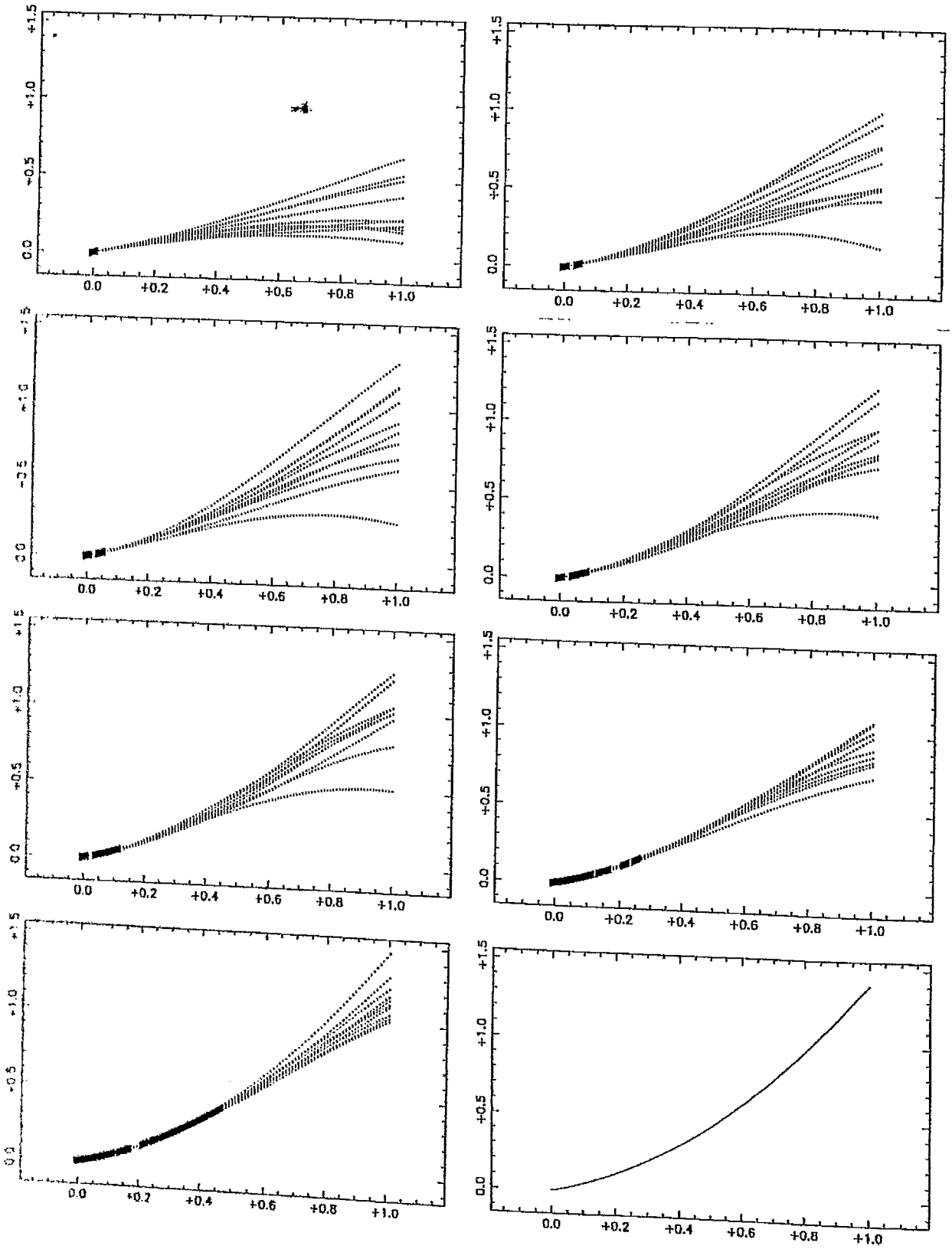

Fig. 3. Problem C_1 solved with the first order "y'" algorithm. The posterior distribution of $y(\cdot)$ after 1, 2, 3, 4, 5, 10, 20 samples, with the analytic solution.

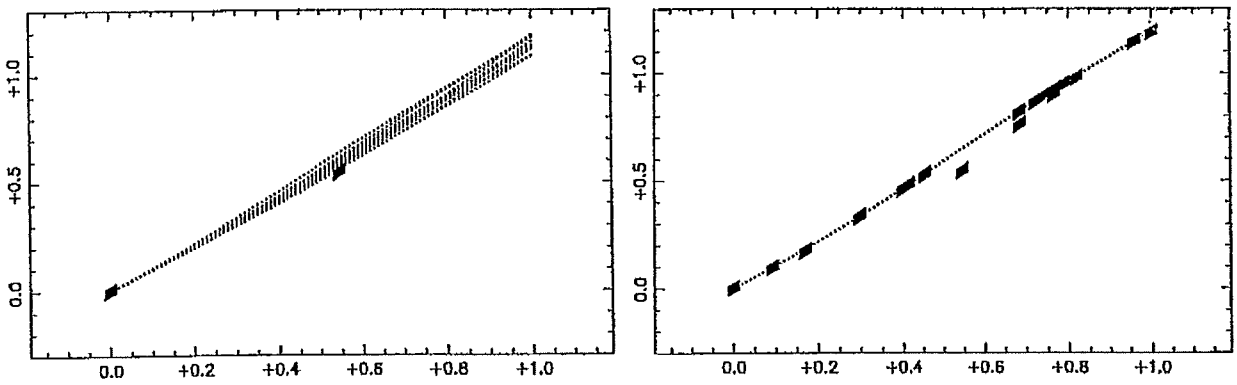

Fig. 4. Problem A, with the "y''" algorithm. The distribution of $y(\cdot)$ after 2 and 20 samples.

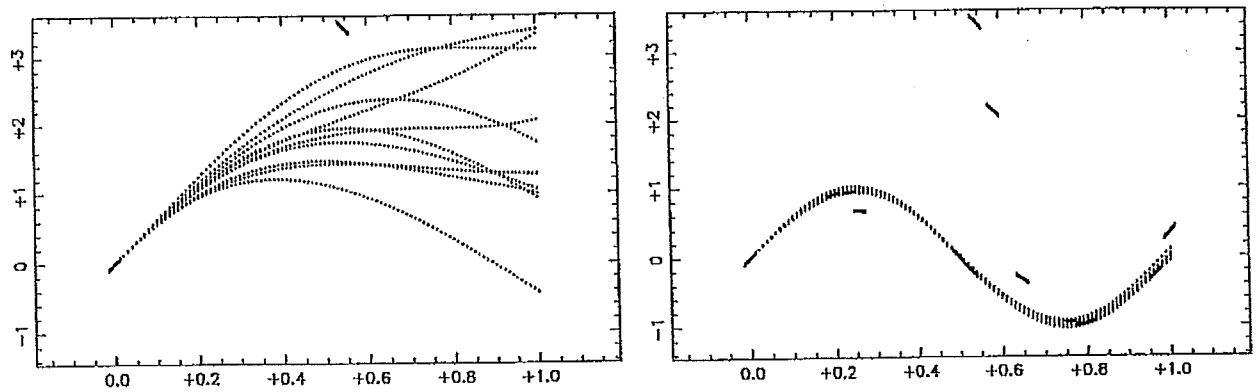

Fig. 5. Problem B, with the "y''" algorithm. The distribution of $y(\cdot)$ after 2 and 20 samples.

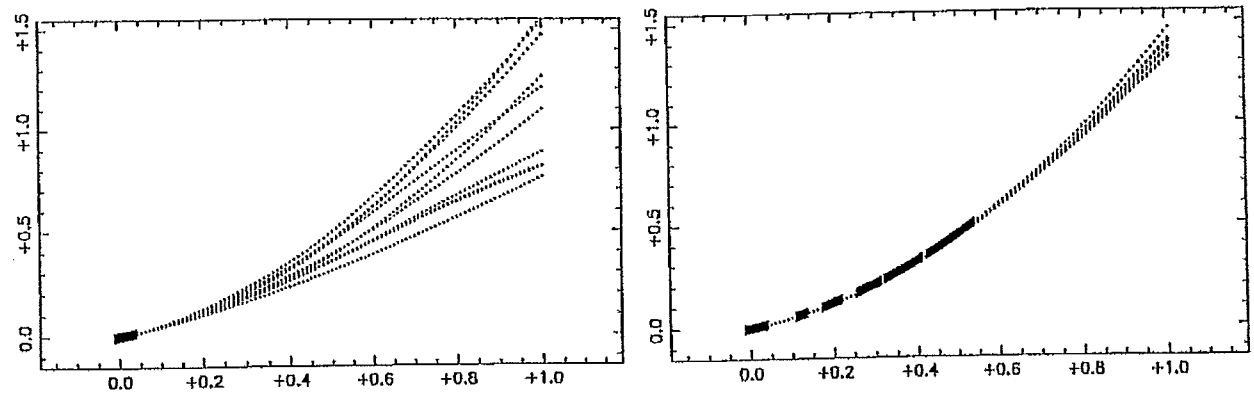

Fig. 6. Problem C, with the "y''" algorithm. The distribution of $y(\cdot)$ after 2 and 20 samples.