Data Driven Sensing and Computation

KIMAS 2003

Cambridge MA

October 2, 2003

Roger Brockett

Engineering and Applied Sciences

Harvard University

Data Driven Sensing, Data Driven Computation

1. In many situations ranging from medical diagnosis to analytical chemistry,the sequence of sensing procedures (tests) are determined sequentially, guided by the outcome of earlier tests. Acknowledging that each sensing procedure has an associated cost, finding the optimal choice of a sensing sequence usually involves solving a dynamic programming problem. Sensing may be deterministic or probabilistic.

2. Although algorithms are designed to compute specific things such as eigenvalues, when the algorithm is computing something like a Fourier transform or Radon transform, it is simply as transforming data. When the outcome of such transformation is to be used to help classify the original data, we may think of the application of the transform as being a type of sensing procedure, used for the purpose of detecting the presence or absence of a type of structure. The decision procedure may be deterministic or probabilistic.

Optimal Sensing Requires Optimal of Excitation and Tuning

1. When a sensing procedure involves simple observation there is no need to choose an excitation but it may be necessary to choose certain parameters. More frequently, sensing involves the introduction of signals or altered states intended to increase the usefulness of the subsequent measurements. Optimization when active sensing is involved involves additional complexity.

2. Is there a computational analog of active sensing? Many algorithms depend on parameters. These must be set correctly if the algorithm is to be maximally useful. When applying a discrete Fourier transform it is necessary to sample the continuous signal at some density. This parameter must be selected before the algorithm is applied even though this usually involves (human) guess work. Secondly, the “experimenter” must select which subset of the data to process. This is closely analogous to the the choice of an optimal excitation.

Two Examples

We now more concrete examples of problems that can serve to motivate the point of view we have adopted. Suitably generalized, these two examples are, in themselves, quite important and have been the subject of whole conferences.

The first comes from the field of image understanding and the second from nuclear resonance spectroscopy, a key tool in the determination of the structure of proteins. The first example serves to help frame the discussion of automatic algorithm selection whereas the second puts in concrete form questions of optimal sensing in a situation in which we must choose the right excitation if we are to get the informative answers.

The Problem

Find a line drawing that represents the framing present in the image and identifies the parts of the image that are not well represented by a line drawing.

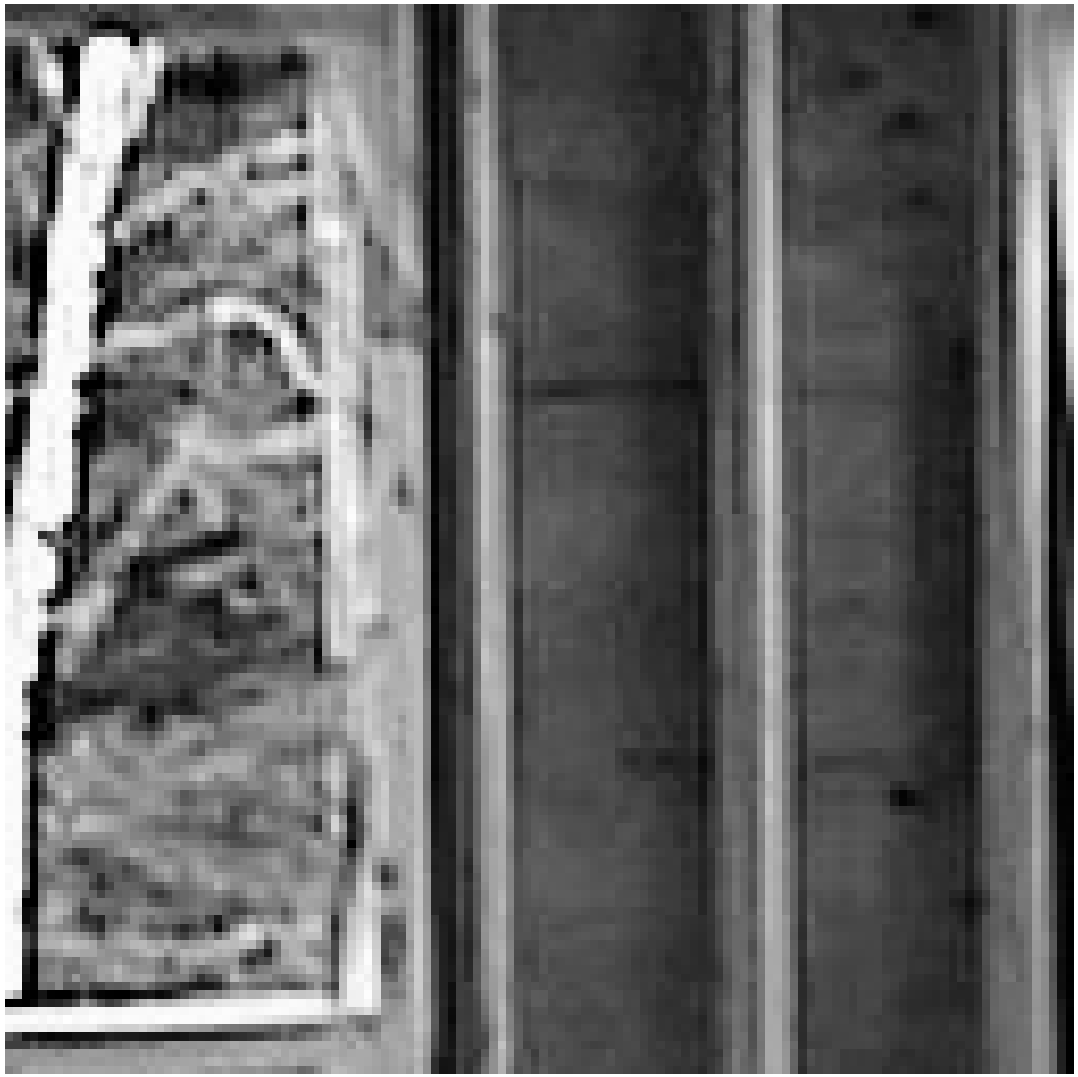
Issues to be clarified include the identification of “correct” scale for analysis of the image and the role to be played by line detectors such as the Radon transform.

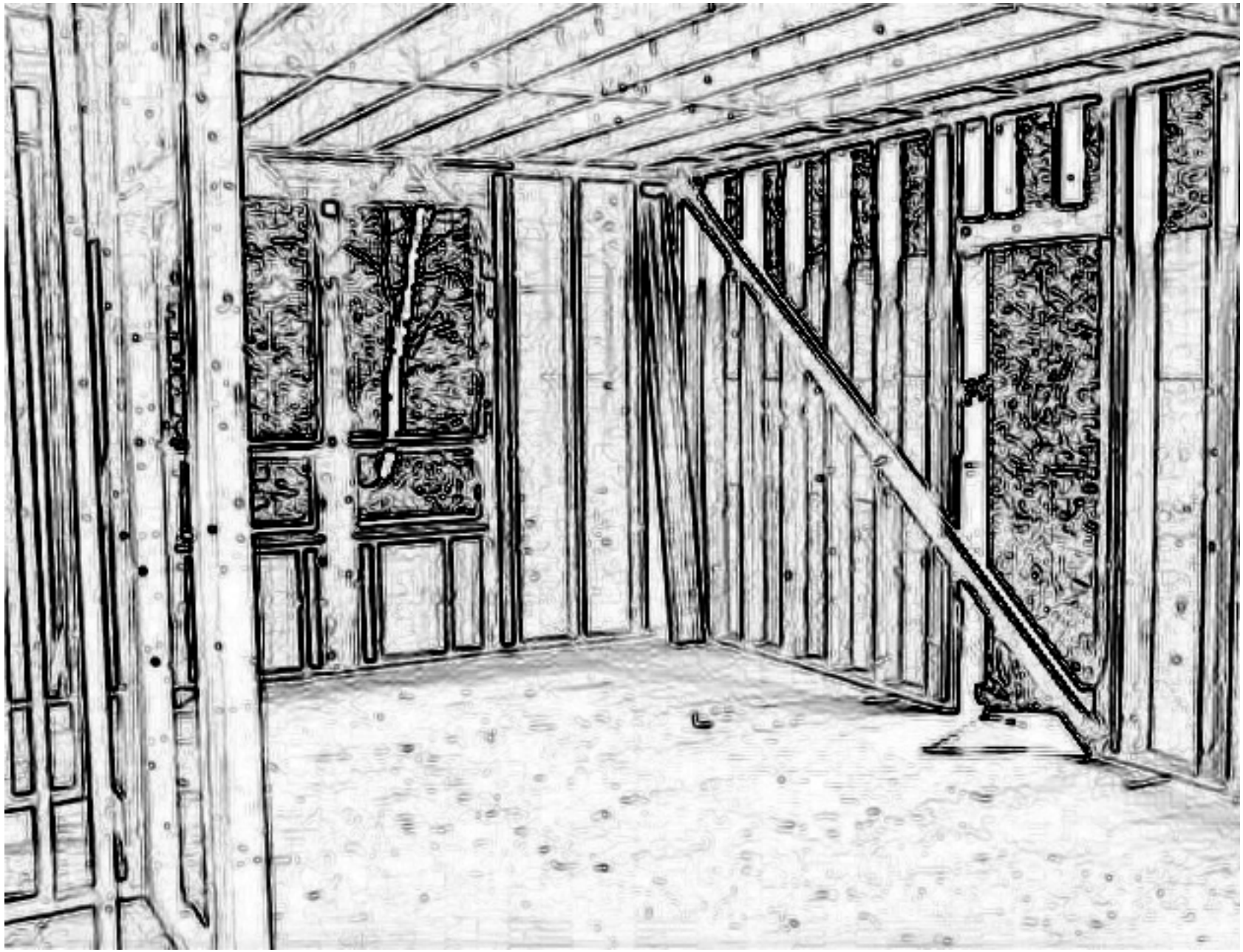
Applying the Radon transform to the entire image is clearly not a good thing to do because there are no lines that extend over the whole image. Applying the Radon transform to blocks should be better but we need to determine a suitable block size. How should this be done?

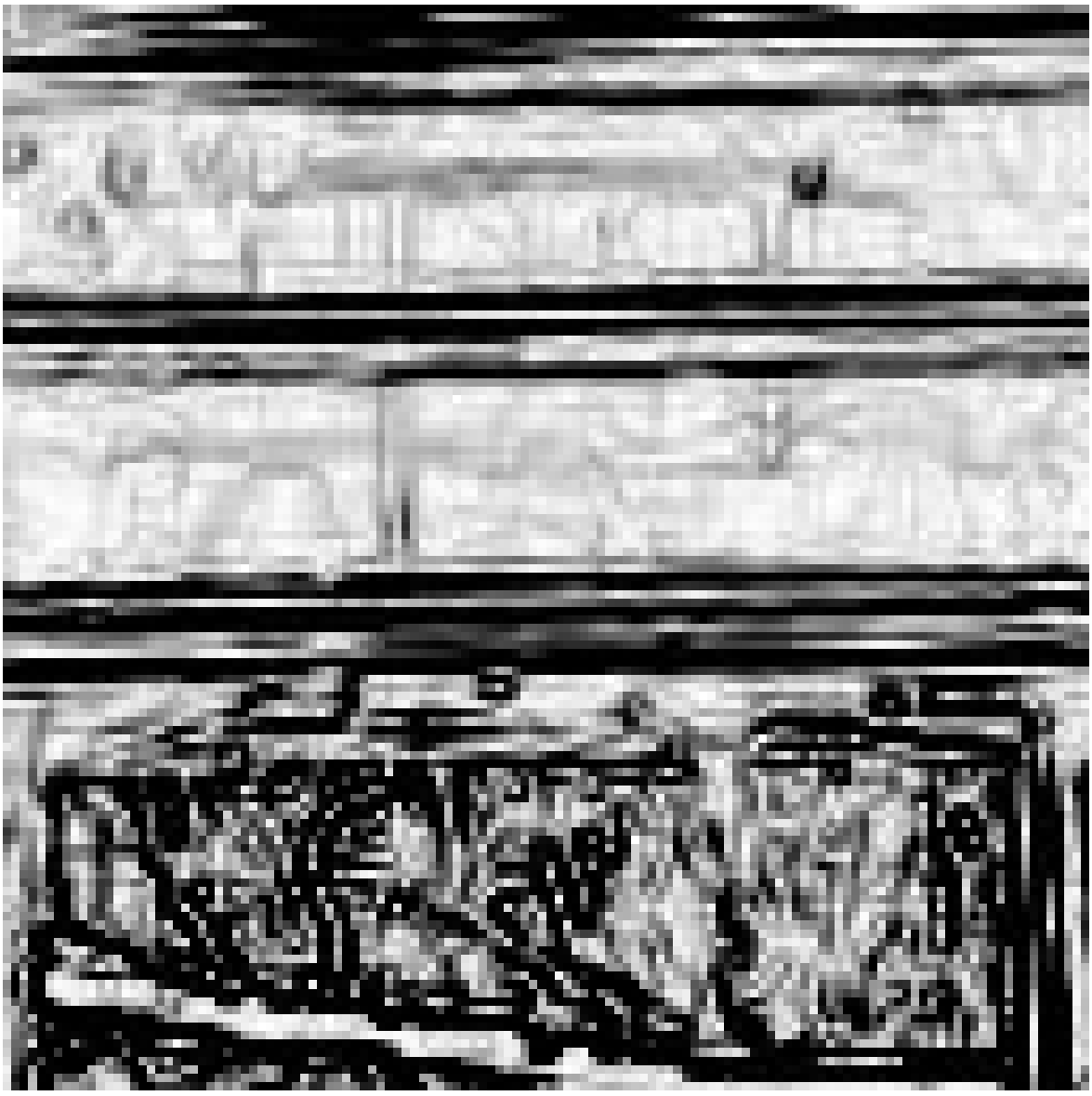
Transforms that Minimize the Description Length

The goal of sensing is to reduce uncertainty. When we have correctly identified an object or scene we can describe it more briefly than we can when it is not yet identified. The goal of algorithmic processing is to find a description of the scene that is accurate and brief in terms of a pre determined vocabulary.

Of course the length of the description of a situation depends on the vocabulary that is available. In vision processing by animals, the vocabulary can be thought of as the set of mutually understood signals used by different parts of the visual cortex. On the time scale of day-to-day events, it can be thought of as being fixed. It might include a signal for “horizontal line”, “to bright”, etc. However, the vocabulary might also include highly composite objects such as “tree” and “mother”.

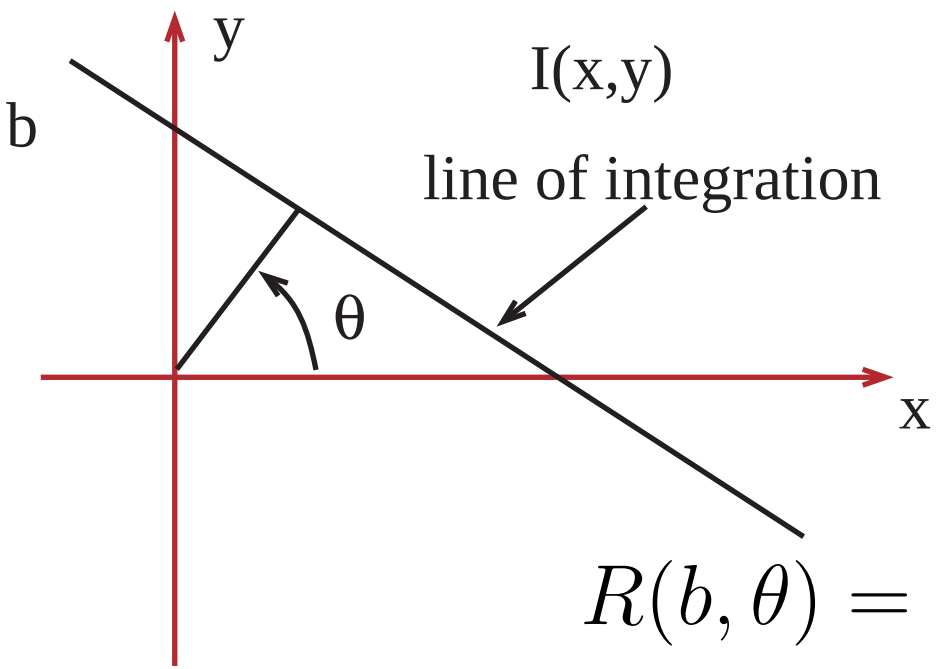




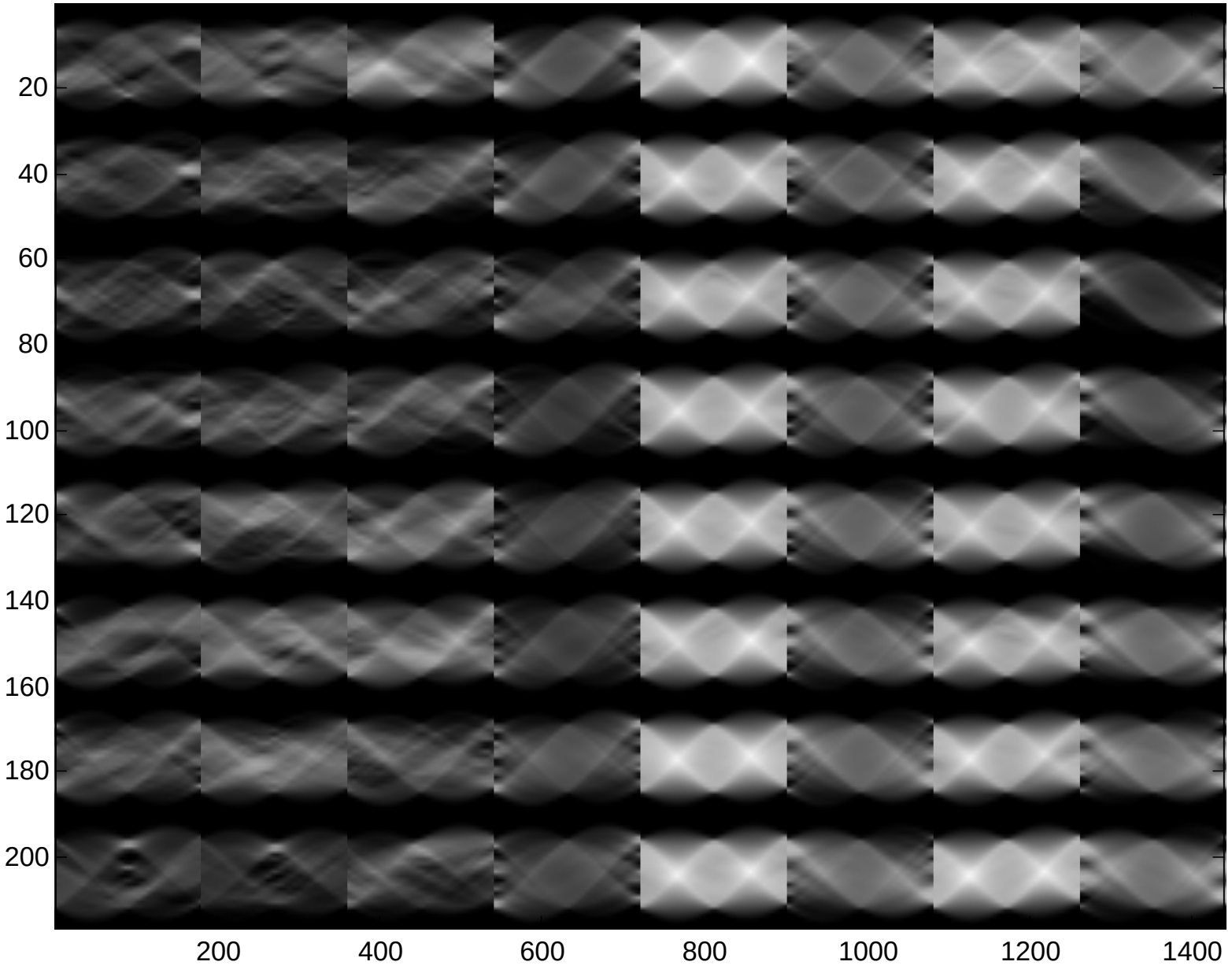


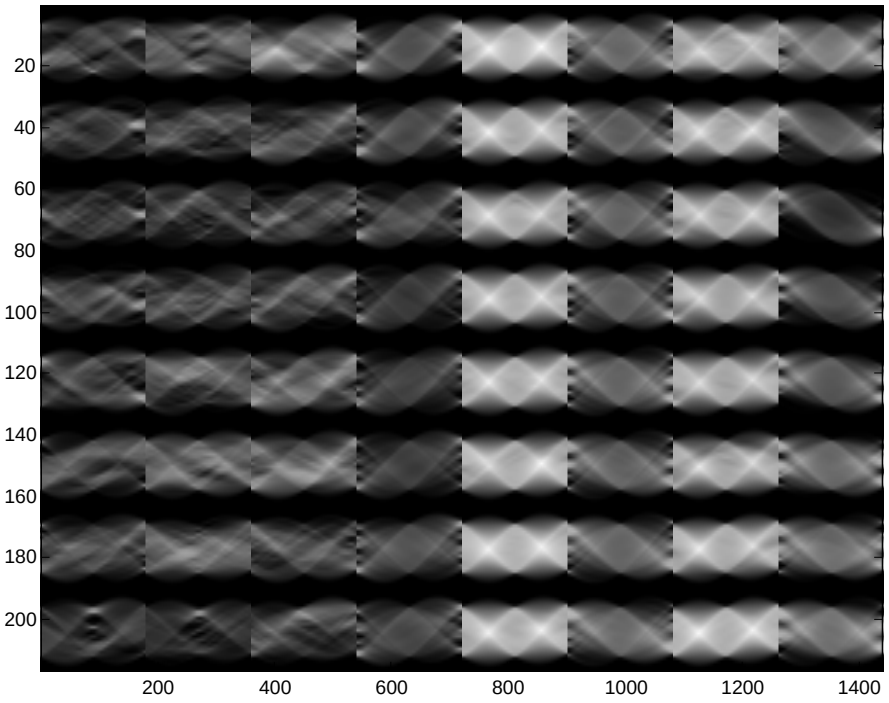
The Block Radon Transform Requires a Choice of Scale

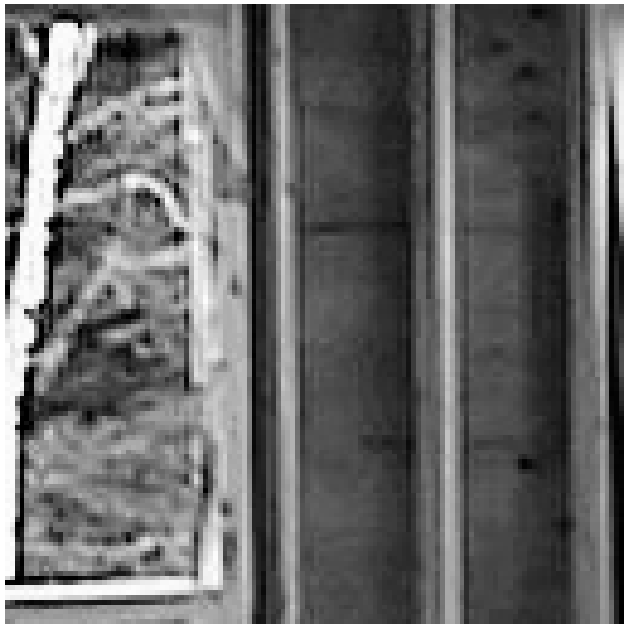
The sub image is 128 by 128 pixels. Because it is not homogeneous we divide it into blocks before processing it. The size of the blocks is a parameter that must be selected before the algorithm can be run but only after the algorithm is run can we judge the success of a certain parameter choice. The next two images show the results of the choices 16 by 16 and a 32 by 32.

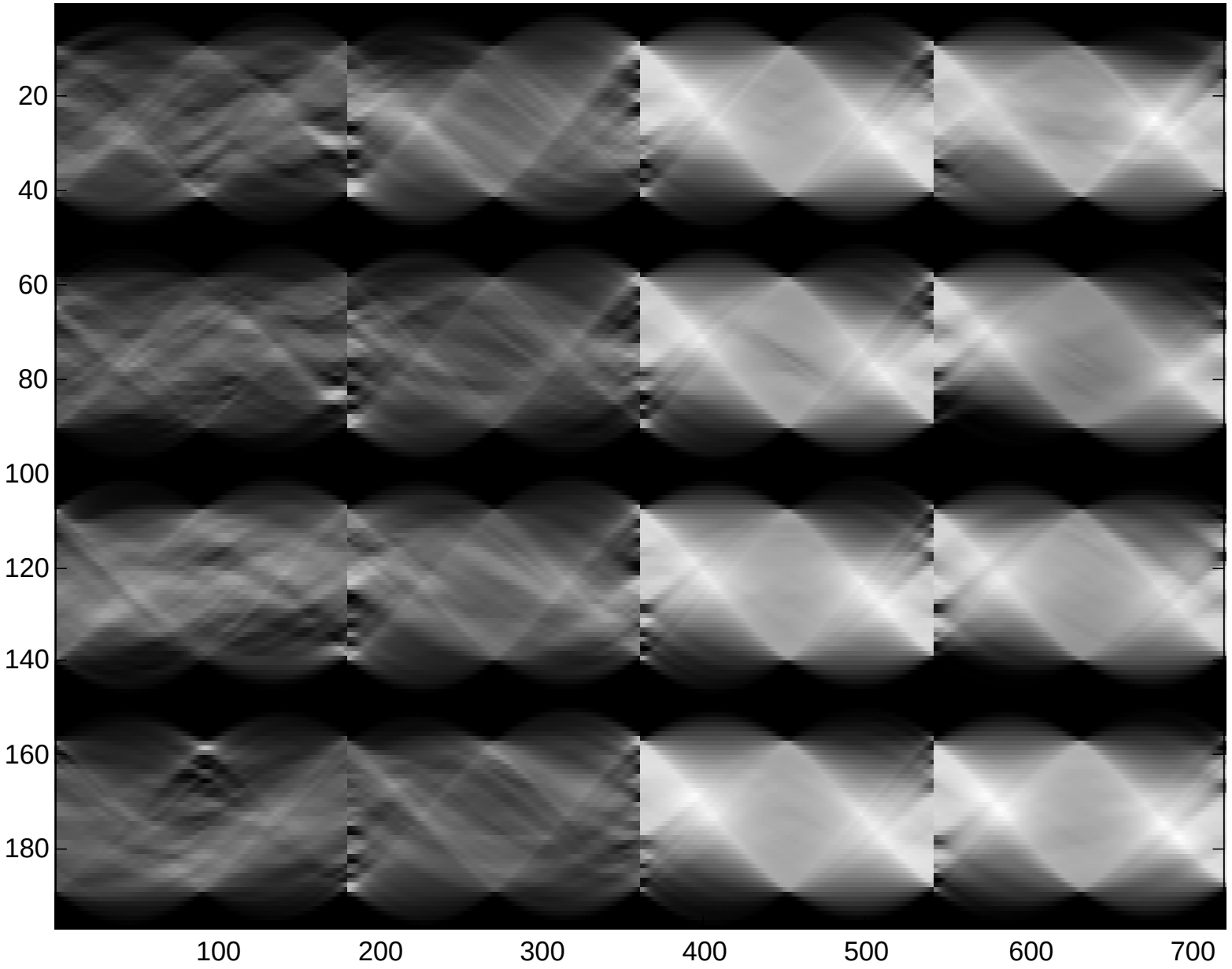


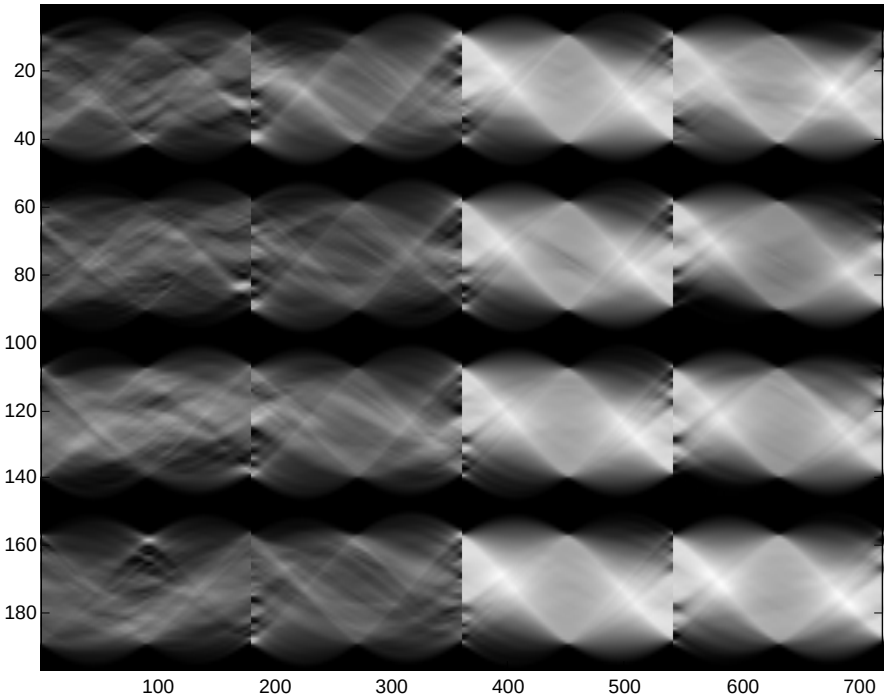
$$R(b,\theta) = \int_{y=mx+b} I(x,y)ds$$

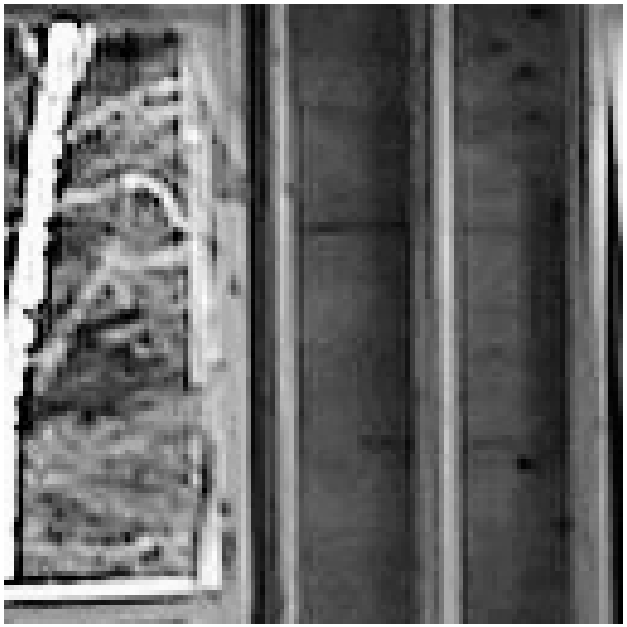

20
40
60
80
100
120
140
160
180
200
200
400
600
800
1000
1200
1400


20
40
60
80
100
120
140
160
180
200
200
400
600
800
1000
1200
1400



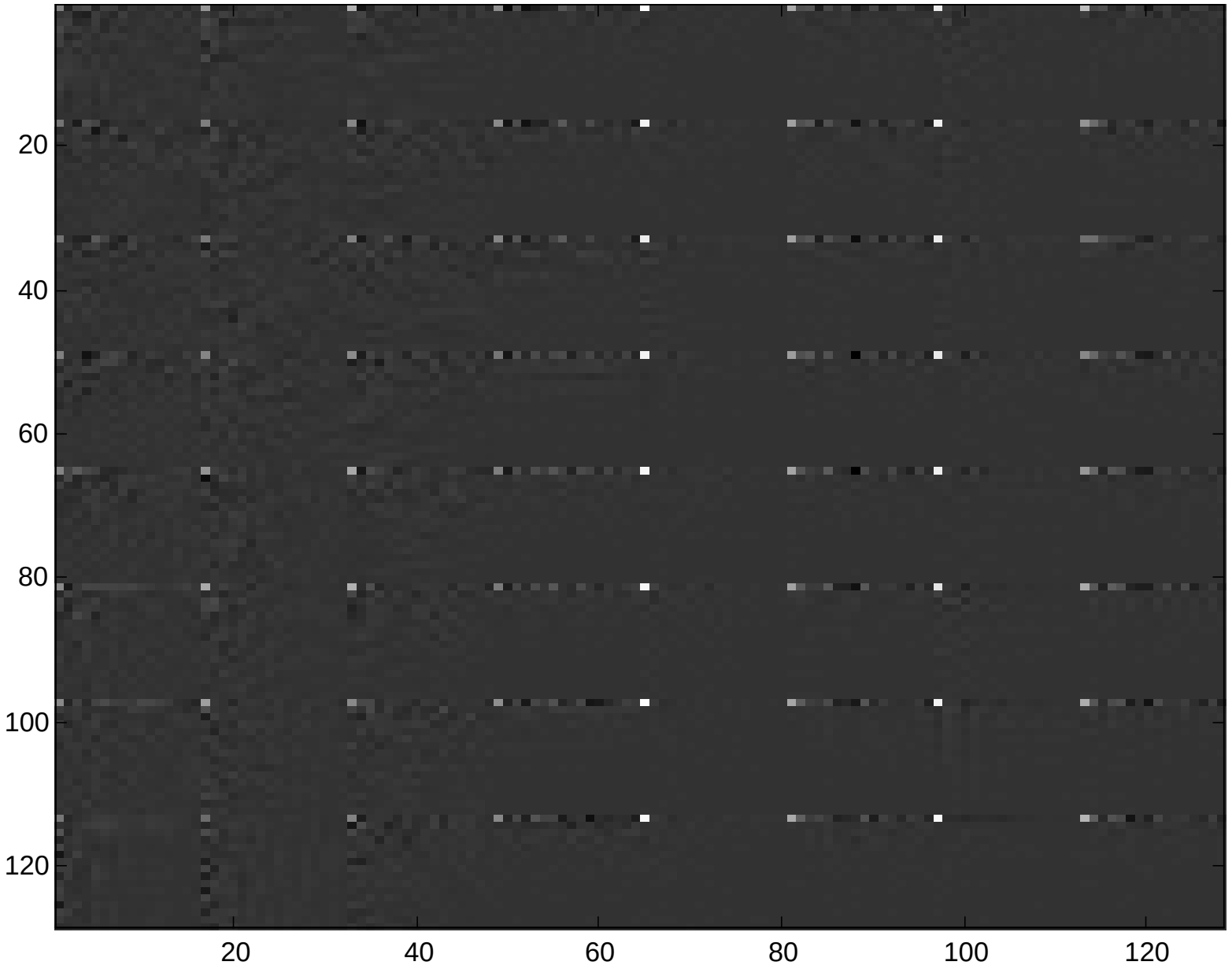

20
40
60
80
100
120
140
160
180
100
200
300
400
500
600
700


20
40
60
80
100
120
140
160
180
100
200
300
400
500
600
700



What about other Transforms?

There is nothing special about the Radon transform and/or the edge detection algorithm we used to prepare for it. What about other transformations of the data? The following slide shows the 16 by 16 block Fourier transform of the 128 by 128 sub image. Notice the lack of structure! Thus if we were running a bank of speculative Computations and applying a minimum description criterion to select The one with the most explanatory power, we would prefer the Radon transform for this particular image.


20
40
60
80
100
120
20
40
60
80
100
120

Signal Processing and Nuclear Magnetic Resonance

1. NMR is the main tool for determining the structure of proteins, key to the utilization of gene sequencing results, and it is now known that the existing methods are far from optimal.

2. NMR is a widely used tool for noninvasive measurement of brain structure and function but higher resolution is needed.

3. There are beautiful things to be learned by studying the methods developed by physicists and chemists working in these fields, especially in the area of nonlinear signal processing.

Abstract Version of the NMR Problem

Consider a stochastic (via W and n) bilinear system of the form

$$dx/dt = (A + W + u(t)B(t))x + b + n(t) \qquad y = cx$$

A given waveform u gives rise to an observation process y. Given a prior probability distribution on the matrices A and B there exists a conditional density for them. Find the input waveform u(t) which makes the entropy of this conditional density as small as possible.

In NMR the matrix A will have complex and lightly damped eigenvalues often in the range 10^7 /sec. Some structural properties of the system will be known and y may have more than one component. A popular idea is to pick u to generate some kind of resonance and get information on the system from the resonant frequency. Compare with optical spectroscopy in which identification is done by frequency.

An Example to Fix Ideas

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & u & 0 \\ -u & -1 & f \\ 0 & -f & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

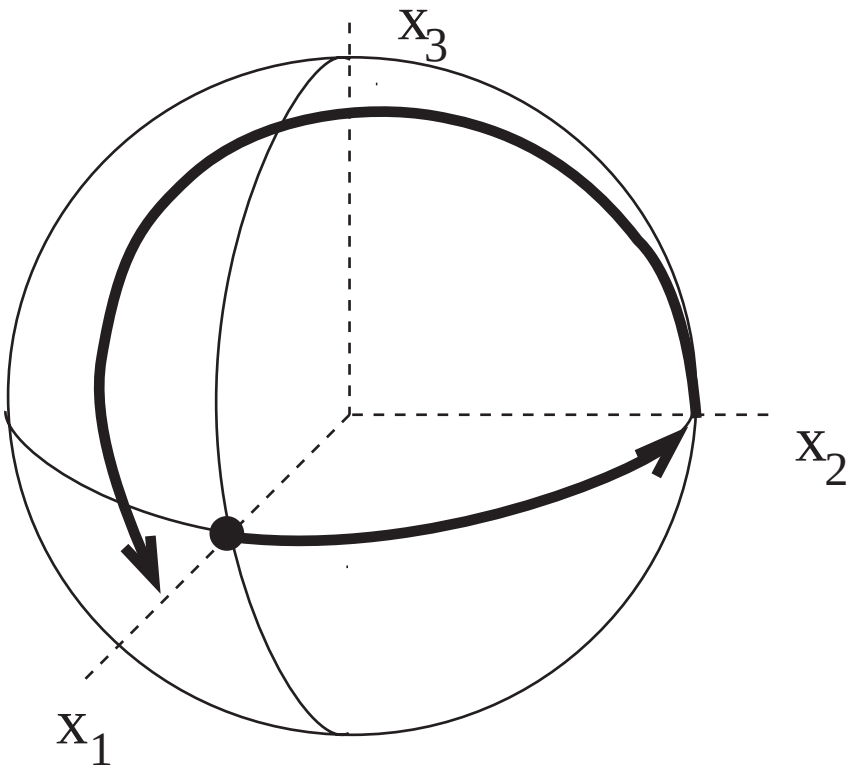
$$y = x_2 + n$$

Let w and n be white noise. The problem is to choose u to reduce the uncertainty in f, given the observation y.

Observe that there is a constant bias term. Intuitively speaking, one wants to transfer the bias present in x_1 to generate a bias for the signal x_2 which then shows up in y.

Qualitative Analysis Based on the Mean

If we keep u at zero there is no signal. If we apply a pulse, rotating the equilibrium state from $x_1 = 1$, $x_2=0$,$x_3=0$ to $x_1 = 0$, $x_2=1$,$x_3=0$, Then we get a signal that reveals the size of f. The actual signal with noise present can be expected to have similar behavior.



The Continuous Wave Approach

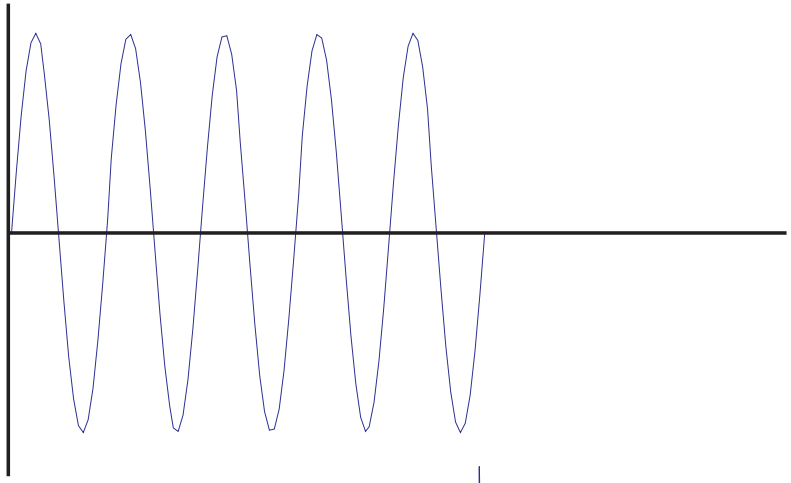
$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & u & 0 \\ -u & -1 & f \\ 0 & -f & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

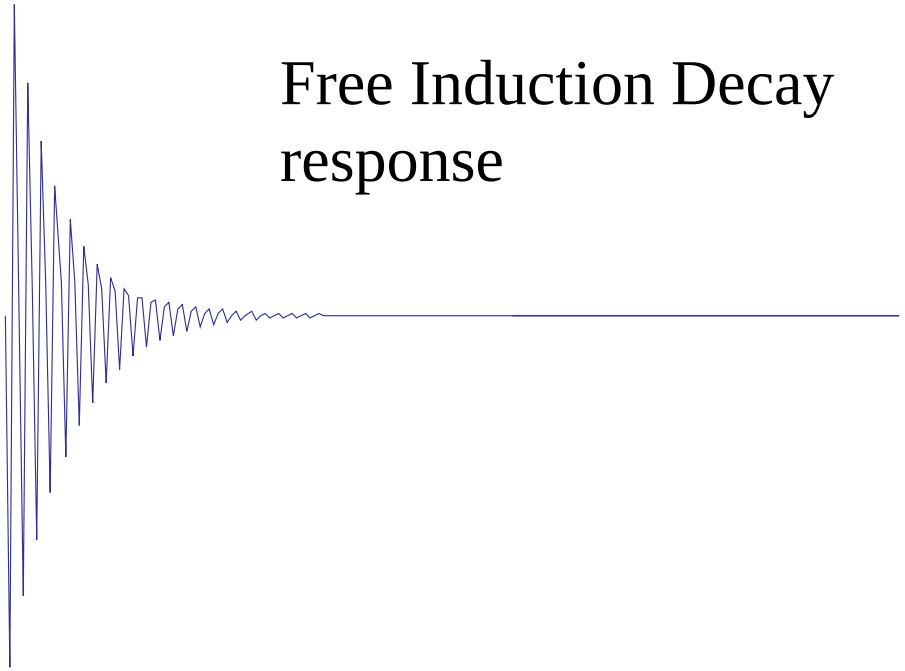
$$y = x_2 + n$$

Let u be “slowly varying sine wave” u=a sin(b(t) t) with b(t) = rt. The benefit of the pulse goes away after the decay--the sine wave provides continuous excitation.

Possible Input-Output Response

Radio Frequency Pulse input

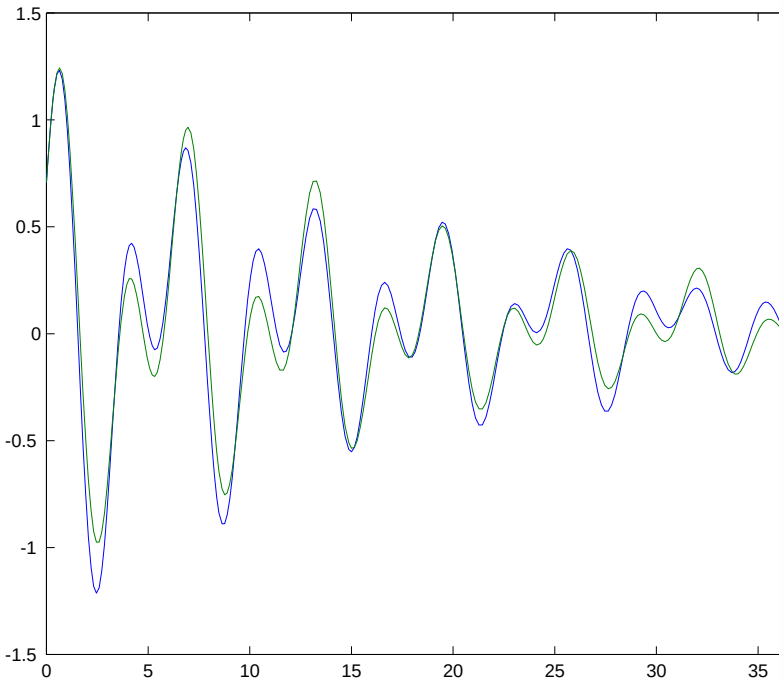


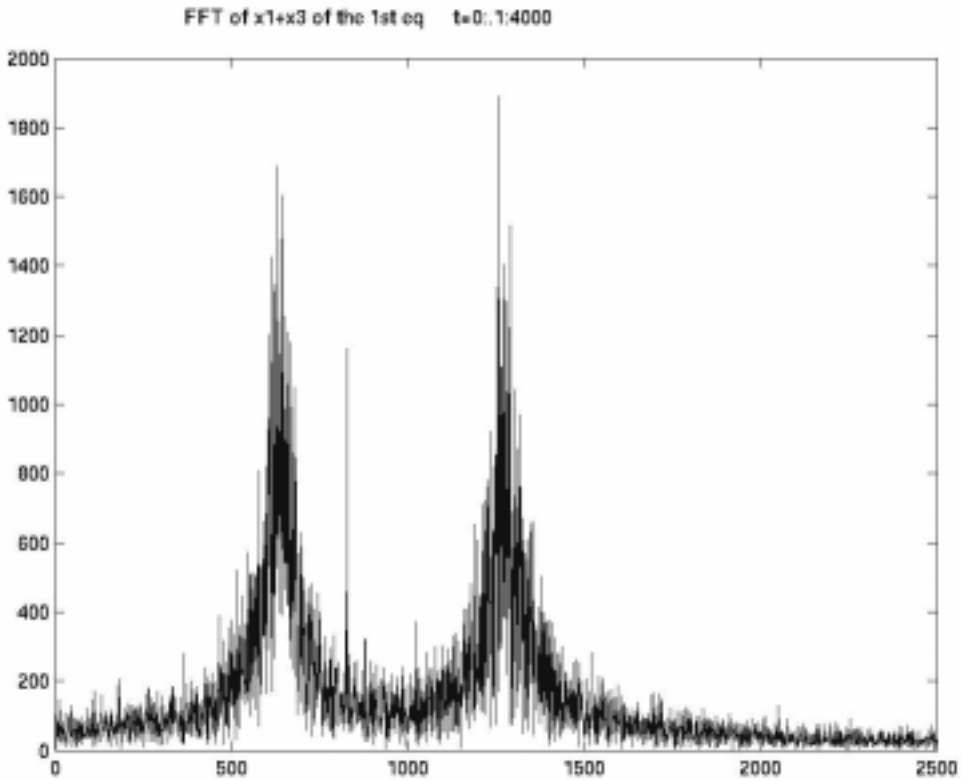


Free Induction Decay response

The Linearization Dilemma

Small input makes linearization valid but gives small signal-to-noise ratio. Large input give higher signal-to-noise ratio but makes nonlinear signal processing necessary.





The Linear System Identification Problem

Given a fixed but unknown linear system

$$dx/dt = Ax+Bw \quad ; y=cx + n$$

Suppose the A belongs to a finite set, compute the conditional probability of the pair (x,A) given the observations y. The solution is well known, in principle. Run a bank of Kalman-Bucy filters, one for each of the models. Each then has its own "mean" and "error variance". There is a key weighting equation associated with each model

$$d(\ln \alpha)/dt = x^T C^T(y-Cx)-(1/2)tr(C^TC- \Sigma^{-1}BB^T\Sigma^{-1)}(xx^T-\Sigma))$$

(weighting equation)

$$dx/dt = Ax-\Sigma C^T(Cx- y)$$ (conditional mean equation)

$$d\Sigma/dt = A\Sigma + \Sigma A^T + B^TB - \Sigma C^T C \Sigma$$ (conditional error variance)

The Mult-Model Identification Problem

Consider the conditional density equation for the joint state-parameter problem

$$\rho_t(t,x,A) = L^* \rho(t,x,A) - (Cx)^2/2\rho\ t,x,A) + yCx\ \rho\ (t,x,A)$$

This equation is unnormalized and can be considered to be vector equation with the vector having a as many components as there are possible models. Assume a solution for a typical component of the form

$$\rho_i(t,x) = \alpha_i(t)(2\pi^n \det\Sigma)^{-1/2} \exp\ (x-x_m)^T \Sigma^{-1}(x-x_m)/2$$

$$d\alpha_i(t)/dt = \ldots$$
$$dx_i(t)/dt = \ldots$$
$$d\Sigma_i(t)/dt = \ldots$$

The Linear System Identification Problem Again

When the parameters depend on a control it may be possible to influence the evolution of the weights in such a way as to reduce the entropy of the conditional distribution for the system identification.

Notice that for the example we could apply a $\pi/2$ pulse to move the the bias to the lower block or we could let u be a sine wave with a slowly varying frequency and look for a resonance. It can be cast as the optimal control (say with a minimum entropy criterion) of

$$d(\ln \alpha)/dt = x^T C^T (y - Cx) - (1/2)\mathrm{tr}(C^T C - \Sigma^{-1} B B^T \Sigma^{-1})(xx^T - \Sigma))$$

$$dx/dt = A(u)x - \Sigma C^T (Cx - y))$$

$$d\Sigma/dt = A(u)\Sigma + \Sigma A(u)^T + B^T B - -\Sigma C^T C \Sigma$$

$$p_i = \alpha_i / (\Sigma \alpha_i)$$

Interpreting the Probability Weighting Equation

The first term rewards α according to the degree of alignment between the "conditional innovations" y-Cx, and the conditional mean of x. It increases α if $x^T C^T(y-Cx)$ is positive. What about

$$(1/2)\mathrm{tr}(C^TC- \Sigma^{-1}BB^T\Sigma^{-1})(xx^T-\Sigma)$$

It compares the sample mean with the error covariance. Notice that

$$C^TC- \Sigma^{-1}BB^T\Sigma^{-1} = -d\Sigma^{-1}/dt - \Sigma^{-1}A-A^T\Sigma^{-1}$$

Thus it measures a difference between the evolution of the inverse error variance with and without driving noise and observation.

Controlling an Ensemble with a Single Control

The actual problem involves many copies with the same dynamics

$$dx_1/dt = A(u)x_1+Bw_1$$

$$dx_2/dt = A(u)x_2+Bw_2$$

……..

$$dx_n/dt = A(u)x_n+Bw_n$$

$$y=(cx_1 + cx_{2+} \ldots+x_n) + n$$

The system is not controllable or observable. There are 10^{23} copies of the same, or nearly the same, system. We can write an equation for the sample mean of the x's, for the sample covariance, etc. Multiplicative control is qualitative different from additive.

A Unified Setting for Problems of this Type

These two problems may appear to be very different. One involves algorithm selection and tuning, the other involves input selection to improve observability. In the second case the optimal data processing involves running a bank of filters generated from a single filter by a parameter choice. Can something of this type be optimal in the first cases well?

Fact 1, We can think of the discrete Fourier transform as being computed by a filter bank and the same is true for the discrete Radon transform. We can think of these as being different parametrized families of computations.

Fact 2. The optimal choice of family, on the other hand, depends on the particular class of system to be identified. Determining this requires a more speculative (read higher level) approach.

Prospects for a Useful Theory

Given the broad scope of the problem area, at what level of generality can one hope for a theory that is grounded in practical algorithms and free of ad hoc assumptions?

Let us at least see what form these questions take if we focus on a class of linear systems and limit ourselves to the state estimation part.

Unification of Sensing and Control

The control can affect both the observation and the dynamics. Consider special cases of

$$\dot{x} = A(u)x + b(u) \quad y = C(u)x$$

e.g., for the case

$$\dot{x} = Ax + bw \;\; y = C(u)x + \dot{\nu}$$

we are led to the problem of controlling the Riccati equation for the error variance

$$\dot{\Sigma} = A\Sigma + \Sigma A^T + bb^T - \Sigma C^T(u)C(u)\Sigma$$

with the goal of minimizing $\Sigma(T)$.

More Generality

The Riccati equation for the error variance takes the form

$$\dot{\Sigma} = A(u)\Sigma + \Sigma A^T(u) + B(u)B^T(u) - \Sigma C^T(u)C(u)\Sigma$$

The special features of individual problems comes in when we express the constrains on $A(u), b(u)$ and $C(u)$.

Example 1: Suppose that the admissible values for $C(u(t))$ are the matrices with one 1. How should we "look" and the state vector so as to minimize $\mathrm{tr}(\Sigma)$ subject to this constraint?

Example 2: Consider the simplified NMR model introduced above

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & u & 0 \\ -u & -1 & f \\ 0 & -f & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$y = x_2 + n$$

In this problem C and B are both constant. The leverage that u has to reduce Σ comes from the fact that it can rotate components of Σ into a subspace where C has an effect.

Does this have any Points of Contact with Biology?

In order to make contact with as wide a circle of ideas as possible, we end with two isolated remarks about some analogous ideas in biology. The connections are lose and will only be useful if they prove to be suggestive.

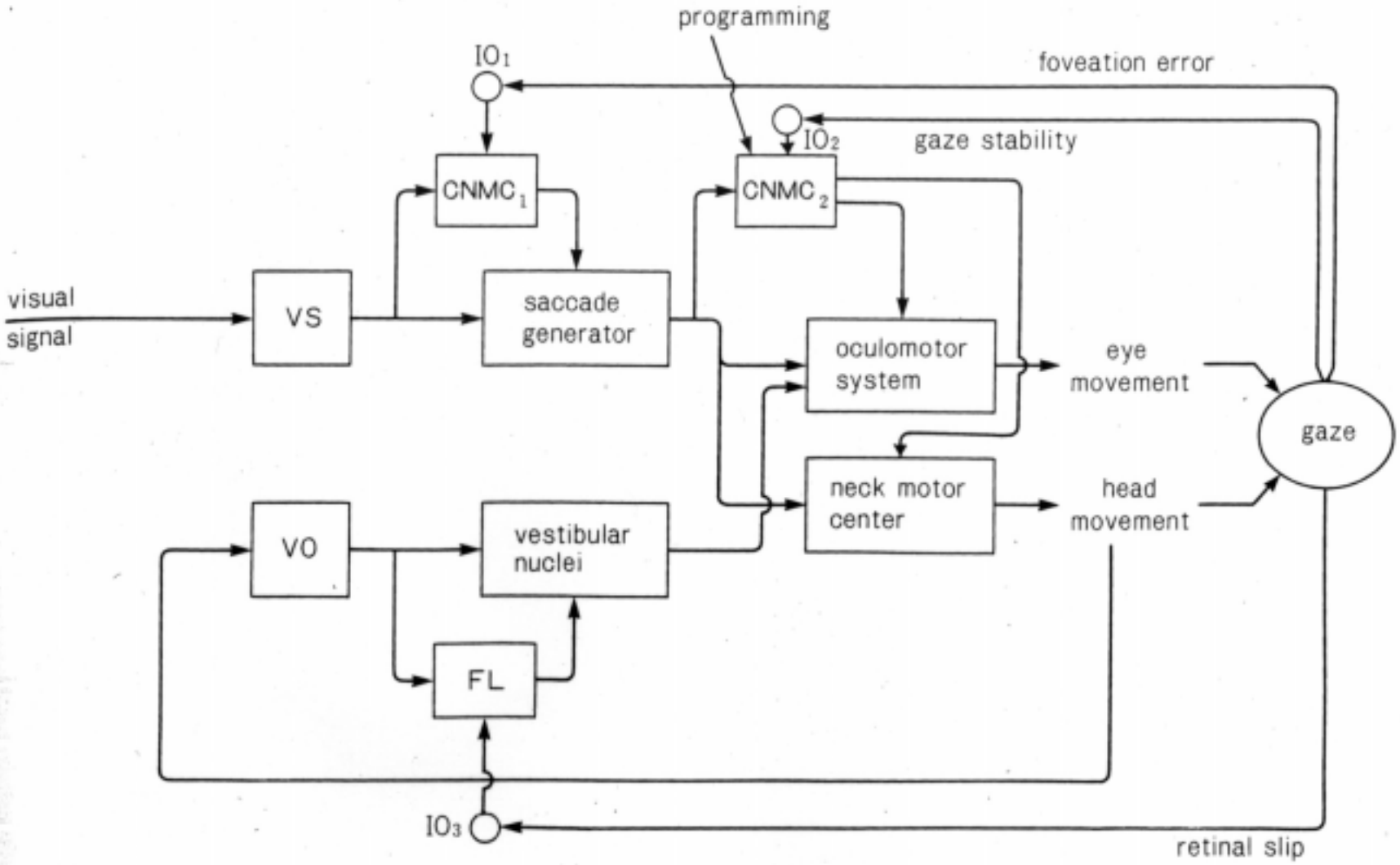
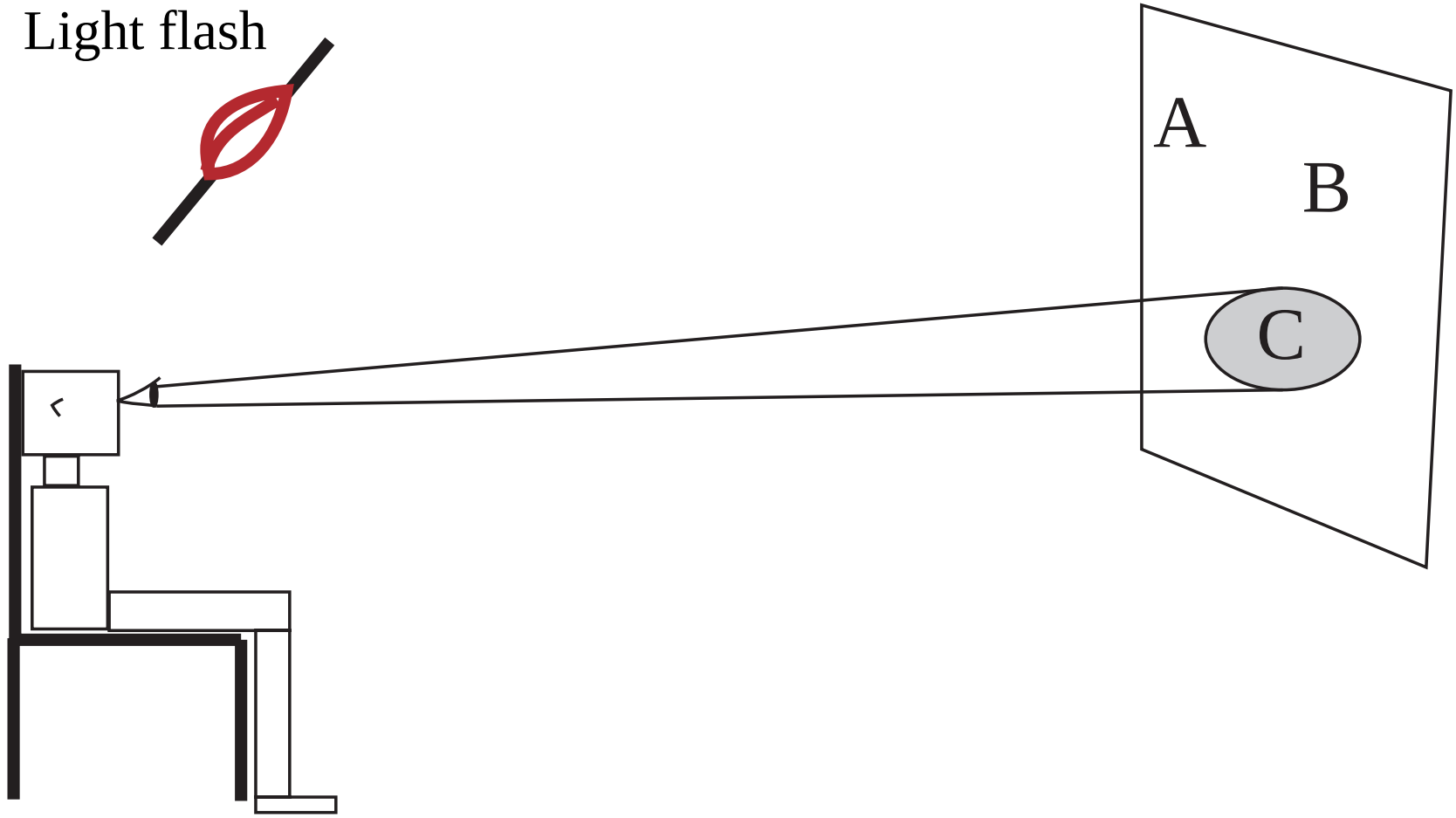


FIG. 195. Hypothetical block diagram for cerebellar contribution to eye–head coordination. VS, visual system; VO, vestibular organ; FL, flocculus; IO_1, IO_2, IO_3, areas of the inferior olive.

Hermann von Helmholtz, 1821-1894

"Theoretical natural science must, therefore, if it is not to rest content with a partial view of the nature of things, take a position in harmony with the present conception of simple forces and the consequences of this conception. Its task will be completed when the reduction of phenomena to simple forces is completed…"

Helmholtz, 1894



The famous Holmholtz experiment showing that humans can direct visual attention without physical motion.

Conclusions

1. We developed the point of view that running an algorithm on data is just a type of sensing and in this way reduced the main steps in system identification to a common framework.

2. Each identification step has a cost associated with it. In some cases this is the cost of making a measurement (cost of a test) and in some cases this is the cost of running an algorithm.

3. In some settings we see that the conditional density is generated by running a bank of filters. In special case this bank of filters can be thought of as computing a discrete Fourier transform but more often, simply something analogous to it.

4. A universal performance measure associated with such systems involves the length of the minimum description.