46

Deconvolution

▶ 46.1 Traditional image reconstruction methods

Optimal linear filters

In many imaging problems, the data measurements $\{d_n\}$ are linearly related to the underlying image $\mathbf{f}$:

$$d_n = \sum_k R_{nk} f_k + n_n. \tag{46.1}$$

The vector $\mathbf{n}$ denotes the inevitable noise that corrupts real data. In the case of a camera which produces a blurred picture, the vector $\mathbf{f}$ denotes the true image, $\mathbf{d}$ denotes the blurred and noisy picture, and the linear operator $\mathbf{R}$ is a convolution defined by the point spread function of the camera. In this special case, the true image and the data vector reside in the same space; but it is important to maintain a distinction between them. We will use the subscript $n = 1, \ldots, N$ to run over data measurements, and the subscripts $k, k' = 1, \ldots, K$ to run over image pixels.

One might speculate that since the blur was created by a linear operation, then perhaps it might be deblurred by another linear operation. We can derive the *optimal linear filter* in two ways.

Bayesian derivation

We assume that the linear operator $\mathbf{R}$ is known, and that the noise $\mathbf{n}$ is Gaussian and independent, with a known standard deviation σ_ν.

$$P(\mathbf{d} \,|\, \mathbf{f}, \sigma_\nu, \mathcal{H}) = \frac{1}{(2\pi\sigma_\nu^2)^{N/2}} \exp\left(- \sum_n \left(d_n - \textstyle\sum_k R_{nk} f_k\right)^2 \Big/ \left(2\sigma_\nu^2\right) \right). \tag{46.2}$$

We assume that the prior probability of the image is also Gaussian, with a scale parameter σ_f.

$$P(\mathbf{f} \,|\, \sigma_f, \mathcal{H}) = \frac{\det^{-\frac{1}{2}} \mathbf{C}}{(2\pi\sigma_f^2)^{K/2}} \exp\left(- \sum_{k,k'} f_k C_{kk'} f'_k \Big/ \left(2\sigma_f^2\right) \right). \tag{46.3}$$

If we assume no correlations among the pixels then the symmetric, full rank matrix $\mathbf{C}$ is equal to the identity matrix $\mathbf{I}$. The more sophisticated 'intrinsic correlation function' model uses $\mathbf{C} = [\mathbf{G}\mathbf{G}^{\mathsf{T}}]^{-1}$, where $\mathbf{G}$ is a convolution that takes us from an imaginary 'hidden' image, which is uncorrelated, to the real correlated image. The intrinsic correlation function should not be confused with the point spread function $\mathbf{R}$ which defines the image-to-data mapping.

A zero-mean Gaussian prior is clearly a poor assumption if it is known that all elements of the image $\mathbf{f}$ are positive, but let us proceed. We can now write down the posterior probability of an image $\mathbf{f}$ given the data $\mathbf{d}$.

$$P(\mathbf{f} \,|\, \mathbf{d}, \sigma_\nu, \sigma_f, \mathcal{H}) = \frac{P(\mathbf{d} \,|\, \mathbf{f}, \sigma_\nu, \mathcal{H}) P(\mathbf{f} \,|\, \sigma_f, \mathcal{H}))}{P(\mathbf{d} \,|\, \sigma_\nu, \sigma_f, \mathcal{H})}. \tag{46.4}$$

In words,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}. \tag{46.5}$$

The 'evidence' $P(\mathbf{d} \,|\, \sigma_\nu, \sigma_f, \mathcal{H})$ is the normalizing constant for this posterior distribution. Here it is unimportant, but it is used in a more sophisticated analysis to compare, for example, different values of σ_ν and σ_f, or different point spread functions $\mathbf{R}$.

Since the posterior distribution is the product of two Gaussian functions of $\mathbf{f}$, it is also a Gaussian, and can therefore be summarized by its mean, which is also the *most probable image*, $\mathbf{f}_{\rm MP}$, and its covariance matrix:

$$\mathbf{\Sigma}_{\mathbf{f}|\mathbf{d}} \equiv \left[-\nabla\nabla \log P(\mathbf{f} \,|\, \mathbf{d}, \sigma_\nu, \sigma_f, \mathcal{H}) \right]^{-1}, \tag{46.6}$$

which defines the joint error bars on $\mathbf{f}$. In this equation, the symbol ∇ denotes differentiation with respect to the image parameters $\mathbf{f}$. We can find $\mathbf{f}_{\rm MP}$ by differentiating the log of the posterior, and solving for the derivative being zero. We obtain:

$$\mathbf{f}_{\rm MP} = \left[\mathbf{R}^{\mathsf{T}}\mathbf{R} + \frac{\sigma_\nu^2}{\sigma_f^2}\mathbf{C} \right]^{-1} \mathbf{R}^{\mathsf{T}}\mathbf{d}. \tag{46.7}$$

The operator $\left[\mathbf{R}^{\mathsf{T}}\mathbf{R} + \frac{\sigma_\nu^2}{\sigma_f^2}\mathbf{C} \right]^{-1} \mathbf{R}^{\mathsf{T}}$ is called the optimal linear filter. When the term $\frac{\sigma_\nu^2}{\sigma_f^2}\mathbf{C}$ can be neglected, the optimal linear filter is the pseudoinverse $\left[\mathbf{R}^{\mathsf{T}}\mathbf{R}\right]^{-1}\mathbf{R}^{\mathsf{T}}$. The term $\frac{\sigma_\nu^2}{\sigma_f^2}\mathbf{C}$ regularizes this ill-conditioned inverse.

The optimal linear filter can also be manipulated into the form:

$$\text{Optimal linear filter} = \mathbf{C}^{-1}\mathbf{R}^{\mathsf{T}} \left[\mathbf{R}\mathbf{C}^{-1}\mathbf{R}^{\mathsf{T}} + \frac{\sigma_\nu^2}{\sigma_f^2}\mathbf{I} \right]^{-1}. \tag{46.8}$$

Minimum square error derivation

The non-Bayesian derivation of the optimal linear filter starts by assuming that we will 'estimate' the true image $\mathbf{f}$ by a linear function of the data:

$$\hat{\mathbf{f}} = \mathbf{W}\mathbf{d}. \tag{46.9}$$

The linear operator $\mathbf{W}$ is then 'optimized' by minimizing the expected sum-squared error between $\hat{\mathbf{f}}$ and the unknown true image . In the following equations, summations over repeated indices k, k', n are implicit. The expectation $\langle \cdot \rangle$ is over both the statistics of the random variables $\{n_n\}$, and the ensemble of images $\mathbf{f}$ which we expect to bump into. We assume that the noise is zero mean and uncorrelated to second order with itself and everything else, with $\langle n_n n_{n'} \rangle = \sigma_\nu^2 \delta_{nn'}$.

$$\langle E \rangle = \frac{1}{2} \left\langle (W_{kn} d_n - f_k)^2 \right\rangle \tag{46.10}$$

$$= \frac{1}{2} \left\langle (W_{kn} R_{nj} f_j - f_k)^2 \right\rangle + \frac{1}{2} W_{kn} W_{kn} \sigma_\nu^2. \tag{46.11}$$

Differentiating with respect to $\mathbf{W}$, and introducing $\mathbf{F} \equiv \left\langle f_{j'} f_j \right\rangle$ (cf. $\sigma_f^2 \mathbf{C}^{-1}$ in the Bayesian derivation above), we find that the optimal linear filter is:

$$\mathbf{W}_{\rm opt} = \mathbf{F}\mathbf{R}^{\mathsf{T}} \left[\mathbf{R}\mathbf{F}\mathbf{R}^{\mathsf{T}} + \sigma_\nu^2 \mathbf{I} \right]^{-1} . \tag{46.12}$$

If we identify $\mathbf{F} = \sigma_f^2 \mathbf{C}^{-1}$, we obtain the optimal linear filter (46.8) of the Bayesian derivation. The ad hoc assumptions made in this derivation were the choice of a quadratic error measure, and the decision to use a linear estimator. It is interesting that without explicit assumptions of Gaussian distributions, this derivation has reproduced the same estimator as the Bayesian posterior mode, $\mathbf{f}_{\rm MP}$.

The advantage of a Bayesian approach is that we can criticize these assumptions and modify them in order to make better reconstructions.

Other image models

The better matched our model of images $P(\mathbf{f} \,|\, \mathcal{H})$ is to the real world, the better our image reconstructions will be, and the less data we will need to answer any given question. The Gaussian models which lead to the optimal linear filter are spectacularly poorly matched to the real world. For example, the Gaussian prior (46.3) fails to specify that all pixel intensities in an image are positive. This omission leads to the most pronounced artefacts where the image under observation has high contrast or large black patches. Optimal linear filters applied to astronomical data give reconstructions with negative areas in them, corresponding to patches of sky that suck energy out of telescopes! The *maximum entropy* model for image deconvolution (Gull and Daniell, 1978) was a great success principally because this model forced the reconstructed image to be positive. The spurious negative areas and complementary spurious positive areas are eliminated, and the quality of the reconstruction is greatly enhanced.

The 'classic maximum entropy' model assigns an entropic prior

$$P(\mathbf{f} \,|\, \alpha, \mathbf{m}, \mathcal{H}_{\rm Classic}) = \exp(\alpha S(\mathbf{f}, \mathbf{m}))/Z, \tag{46.13}$$

where

$$S(\mathbf{f}, \mathbf{m}) = \sum_i (f_i \ln(m_i/f_i) + f_i - m_i) \tag{46.14}$$

(Skilling, 1989). This model enforces positivity; the parameter α defines a characteristic dynamic range by which the pixel values are expected to differ from the default image $\mathbf{m}$.

The 'intrinsic-correlation-function maximum-entropy' model (Gull, 1989) introduces an expectation of spatial correlations into the prior on $\mathbf{f}$ by writing $\mathbf{f} = \mathbf{G}\mathbf{h}$, where $\mathbf{G}$ is a convolution with an intrinsic correlation function, and putting a classic maxent prior on the underlying hidden image $\mathbf{h}$.

Probabilistic movies

Having found not only the most probable image $\mathbf{f}_{\rm MP}$ but also error bars on it, $\mathbf{\Sigma}_{\mathbf{f}|\mathbf{d}}$, one task is to visualize those error bars. Whether or not we use Monte Carlo methods to infer $\mathbf{f}$, a correlated random walk around the posterior distribution can be used to visualize the uncertainties and correlations. For a Gaussian posterior distribution, we can create a correlated sequence of unit normal random vectors $\mathbf{n}$ using

$$\mathbf{n}^{(t+1)} = c\mathbf{n}^{(t)} + s\mathbf{z}, \tag{46.15}$$

where $\mathbf{z}$ is a unit normal random vector and $c^2 + s^2 = 1$ (c controls how persistent the memory of the sequence is). We then render the image sequence defined by

$$\mathbf{f}^{(t)} = \mathbf{f}_{\mathrm{MP}} + \mathbf{\Sigma}_{\mathbf{f}|\mathbf{d}}^{1/2}\mathbf{n}^{(t)} \tag{46.16}$$

where $\mathbf{\Sigma}_{\mathbf{f}|\mathbf{d}}^{1/2}$ is the Cholesky decomposition of $\mathbf{\Sigma}_{\mathbf{f}|\mathbf{d}}$.

▶ 46.2 Supervised neural networks for image deconvolution

Neural network researchers often exploit the following strategy. Given a problem currently solved with a standard algorithm: interpret the computations performed by the algorithm as a parameterized mapping from an input to an output, and call this mapping a neural network; then adapt the parameters to data so as to produce another mapping that solves the task better. By construction, the neural network can reproduce the standard algorithm, so this data-driven adaptation can only make the performance better.

There are several reasons why standard algorithms can be bettered in this way.

1. Algorithms are often not designed to optimize the real objective function. For example, in speech recognition, a hidden Markov model is designed to model the speech signal, and is fitted so as to to maximize the generative probability given the known string of words in the training data; but the real objective is to *discriminate* between different words. If an inadequate model is being used, the neural-net-style training of the model will focus the limited resources of the model on the aspects relevant to the discrimination task. Discriminative training of hidden Markov models for speech recognition does improve their performance.

2. The neural network can be more flexible than the standard model; some of the adaptive parameters might have been viewed as fixed features by the original designers. A flexible network can find properties in the data that were not included in the original model.

▶ 46.3 Deconvolution in humans

A huge fraction of our brain is devoted to vision. One of the neglected features of our visual system is that the raw image falling on the retina is severely blurred: while most people can see with a resolution of about *1 arcminute* (one sixtieth of a degree) under any daylight conditions, bright or dim, *the image on our retina is blurred through a point spread function of width as large as 5 arcminutes* (Wald and Griffin, 1947; Howarth and Bradley, 1986). It is amazing that we are able to resolve pixels that are twenty-five times smaller in area than the blob produced on our retina by any point source.

Isaac Newton was aware of this conundrum. It's hard to make a lens that does not have chromatic aberration, and our cornea and lens, like a lens made of ordinary glass, refract blue light more strongly than red. Typically our eyes focus correctly for the middle of the visible spectrum (green), so if we look at a single white dot made of red, green, and blue light, the image on our retina consists of a sharply focussed green dot surrounded by a broader red blob superposed on an even broader blue blob. The width of the red and blue blobs is proportional to the diameter of the pupil, which is largest under dim lighting conditions. [The blobs are roughly concentric, though most people have a slight bias, such that in one eye the red blob is centred a tiny distance

to the left and the blue is centred a tiny distance to the right, and in the other eye it's the other way round. This slight bias explains why when we look at blue and red writing on a dark background most people perceive the blue writing to be at a slightly greater depth than the red. In a minority of people, this small bias is the other way round and the red/blue depth perception is reversed. But this effect (which many people are aware of, having noticed it in cinemas, for example) is *tiny* compared with the chromatic aberration we are discussing.]

You can vividly demonstrate to yourself how enormous the chromatic aberration in your eye is with the help of a sheet of card and a colour computer screen.

For the most impressive results – I guarantee you will be amazed – use a dim room with no light apart from the computer screen; a pretty strong effect will still be seen even if the room has daylight coming into it, as long as it is not bright sunshine. Cut a slit about 1.5 mm wide in the card. On the screen, display a few small coloured objects on a black background. I especially recommend thin vertical objects coloured pure red, pure blue, magenta (i.e., red plus blue), and white (red plus blue plus green).[1] Include a little black-and-white text on the screen too. Stand or sit sufficiently far away that you can only just read the text – perhaps a distance of four metres or so, if you have normal vision. Now, hold the slit vertically in front of one of your eyes, and close the other eye. Hold the slit near to your eye – brushing your eyelashes – and look through it. Waggle the slit slowly to the left and to the right, so that the slit is alternately in front of the left and right sides of your pupil. What do you see? I see the red objects waggling to and fro, and the blue objects waggling to and fro, through *huge* distances and in opposite directions, while white objects appear to stay still and are negligibly distorted. Thin magenta objects can be seen splitting into their constituent red and blue parts. Measure how large the motion of the red and blue objects is – it's more than 5 minutes of arc for me, in a dim room. Then check how sharply you can see under these conditions – look at the text on the screen, for example: is it not the case that you can see (through your whole pupil) features far smaller than the distance through which the red and blue components were waggling? Yet when you are using the whole pupil, what is falling on your retina must be an image blurred with a blurring diameter equal to the waggling amplitude.

One of the main functions of early visual processing must be to deconvolve this chromatic aberration. Neuroscientists sometimes conjecture that the reason why retinal ganglion cells and cells in the lateral geniculate nucleus (the main brain area to which retinal ganglion cells project) have centre-surround receptive fields with colour opponency (long wavelength in the centre and medium wavelength in the surround, for example) is in order to perform 'feature extraction' or 'edge detection', but I think this view is mistaken. The reason we have centre-surround filters at the first stage of visual processing (in the fovea at least) is for the huge task of deconvolution of chromatic aberration.

I speculate that the *McCollough effect*, an extremely long-lasting association of colours with orientation (McCollough, 1965; MacKay and MacKay, 1974), is produced by the adaptation mechanism that tunes our chromatic-aberration-deconvolution circuits. Our deconvolution circuits need to be rapidly tuneable, because the point spread function of our eye changes with our pupil diameter, which can change within seconds; and indeed the McCollough effect can be induced within 30 seconds. At the same time, the effect is long-lasting

[1] `http://www.inference.phy.cam.ac.uk/mackay/itila/Files.html`

when an eye is covered, because it's in our interests that our deconvolution circuits should stay well-tuned while we sleep, so that we can see sharply the instant we wake up.

I also wonder whether the main reason that we evolved colour vision was not 'in order to see fruit better' but *'so as to be able to see black and white sharper'* – deconvolving chromatic aberration is easier, even in an entirely black and white world, if one has access to chromatic information in the image.

And a final speculation: why do our eyes make micro-saccades when we look at things? These miniature eye-movements are of an angular size bigger than the spacing between the cones in the fovea (which are spaced at roughly 1 minute of arc, the perceived resolution of the eye). The typical size of a microsaccade is 5–10 minutes of arc (Ratliff and Riggs, 1950). Is it a coincidence that this is the same as the size of chromatic aberration? Surely micro-saccades must play an essential role in the deconvolution mechanism that delivers our high-resolution vision.

▶ 46.4 Exercises

Exercise 46.1.[3C] Blur an image with a circular (top hat) point spread function and add noise. Then deconvolve the blurry noisy image using the optimal linear filter. Find error bars and visualize them by making a probabilistic movie.