1 Notes about point-source fitting and noise

1.1 Definitions and units

A point source of flux density f (SI units $Wm^{-2}Hz^{-1}$) gives an intensity I (SI units $Wm^{-2}Hz^{-1}sr^{-1}$) at a position $\mathbf{r}$ relative to the centre of the source in the focal plane of the telescope I, where

$$I(\mathbf{r}) = P(\mathbf{r})f \tag{1}$$

and the Point Spread Function P is an inverse solid angle (SI units of sr^{-1} and is normalised so

$$\int_{\infty} P\, d\Omega = 1. \tag{2}$$

When recorded by a detector we receive a signal $s = \int_{\rm pix} I\, d\Omega$ (units of flux density). Usually the detectors have a small solid angle $\Delta\Omega$ so $s \approx I\Delta\Omega$ and in any case the function P and calibration can usually defined in such a way that that is exact. Data is often represented either as an intensity or as a flux per pixel

$$I(\mathbf{r}) = \frac{s}{\Delta\Omega} = P(\mathbf{r})f. \tag{3}$$

A common approach for estimating fluxes is aperture photometry in which one sums up the flux per pixel and we can see

$$\sum s = f \sum P(\mathbf{r})\Delta\Omega. \tag{4}$$

So if the sum is taken over all pixels this tends to the true flux. It is often convenient to use an alternative normalisation of PSF: $P' = P(\mathbf{r})\Delta\Omega$, where P' (with units of per pixel) has

$$\sum P' = 1 \tag{5}$$

in which case

$$s(\mathbf{r}) = P'(\mathbf{r})f \tag{6}$$

which has the same form as Equation 1

1.2 Optimal Weighting

We will consider our data $\mathbf{d}$ is a measurement of the signal (which might be the intensity I or the flux per pixel s) subject to an error σ. Any element of d_i can be used to estimate the flux $f' = d_i/P_i$ (where P_i should normalised as appropriate for flux-per-pixel, Eqn. 5, or intensity normalisation, Eqn. 2). The error in that estimate $\sigma_f = \sigma_i/P_i$

A simple method for combining data to estimate the flux of a single source would be to construct a weighted sum of these estimates

$$\hat{f} = \frac{\sum_i w_i\, d_i/P_i}{\sum_i w_i}. \tag{7}$$

For un-correlated errors the error the maximum likelihood estimator for f is when the weights are set to the inverse of the variance i.e. $w_i = P_i^2/\sigma_i^2$. Then this estimator is

$$\hat{f} = \frac{\sum_i P_i\, d_i/\sigma_i^2}{\sum_i P_i^2/\sigma_i^2} \tag{8}$$

with variance

$$\sigma_{\hat{f}}^2 = \frac{1}{\sum_i w_i^2} = \left(\sum_i \frac{P_i^2}{\sigma_i^2}\right)^{-1} \tag{9}$$

If the variances per detector element are constant

$$\hat{f} = \frac{\sum_i P_i\, d_i}{\sum_i P_i^2} \tag{10}$$

and

$$\sigma_{\hat{f}}^2 = \frac{\sigma_i^2}{\sum_i P_i^2}. \tag{11}$$

This explains the basis of many source extraction methods where and image is filtered by the point spread function and shows that the required scaling between such a convolved map and point source flux and the noise per detector and the resulting variance in the flux is $1/\sum_i P_i^2$, if the point-spread function has been properly normalised.

When I have some spare time I might play around and see what this normalisation factor is for certain PSF geometries and pixel scales.

2 Basic Linear Method

Our data $\mathbf{d}$ is a vector of M elements from M (good) detector readouts. For example $\mathbf{d}$ might come from an $n_1 \times n_2$ image i.e. with $n_1 \times n_2 = M$ pixels. The detectors are located at positions $(\mathbf{x}, \mathbf{y})$. Our model assumes this data to be formed by a number N of point sources with known positions, $(\mathbf{u}, \mathbf{v})$, and with unknown flux, f_i, all the fluxes are thus a N element vector $\mathbf{f}(\mathbf{u}, \mathbf{v})$. Each source i makes a contribution to the data given by the response function $P(\mathbf{x} - u_i, \mathbf{y} - v_i)$ and the data is given by

$$d_j = \sum_i P(x_j - u_i, y_j - v_i) f_i + n_j \tag{12}$$

where δ_j is an additional noise contribution i.e.

$$\mathbf{d} = P(\mathbf{\Delta X}, \mathbf{\Delta Y})\mathbf{f} + \mathbf{n} \tag{13}$$

where $\mathbf{\Delta X}$ and $\mathbf{\Delta Y}$ are $(M \times N)$ matrices giving the offset between detectors and sources. This is a linear equation of the form

$$\mathbf{d} = \mathbf{A}\mathbf{f} + \mathbf{n} \tag{14}$$

and thus the maximum likelihood solution is

$$\hat{\mathbf{f}} = (\mathbf{A}^T\mathbf{N}^{-1}\mathbf{A})^{-1}\,\mathbf{A}^T\mathbf{N}^{-1}\mathbf{d} \tag{15}$$

Where $\mathbf{N}$ is the covariance matrix. This matrix equation can be solved efficiently by a Conjugate Gradient method.

Let us write the above more explicitely, by defining the theoretical value $\hat{\mathbf{d}}$ for the data given the flux $\mathbf{f}$ and the model as $\hat{\mathbf{d}} = A\mathbf{f}$. The covariance matrix is then

$$\langle(\mathbf{d} - \hat{\mathbf{d}})(\mathbf{d} - \hat{\mathbf{d}}^T)\rangle = \langle\mathbf{n}\mathbf{n}^T\rangle = \mathbf{N}$$

as expected. To derive the maximum likelihood solution, we write down the likelihood as the Gaussian probability function for the data given the fluxes,

$$L(f) = p(d|f) \propto |\mathbf{N}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{d} - \hat{\mathbf{d}})^T\mathbf{N}^{-1}(\mathbf{d} - \hat{\mathbf{d}})\right\}$$

Defining $\chi^2 = (\mathbf{d} - \hat{\mathbf{d}})^T\mathbf{N}^{-1}(\mathbf{d} - \hat{\mathbf{d}})$ we see at the maximum of the likelihood we require

$$\frac{\partial\chi^2}{\partial f} = 0.$$

Written explicitely in terms of components (with summation over repeated indices implied), this becomes

$$\frac{\partial}{\partial f_m}\left[(d_i - A_{ij}f_j)(N^{-1})_{ik}(d_k - A_{kl}f_l)\right] = -2A_{im}(N^{-1})_{ik}(d_k - A_{kl}f_l) = 0$$

The solution of this equation is the one given above.

3 Advances

3.1 Background or other source terms

We can add in other additive model contributions to the signal in an obvious way, by including extra terms in equation 13 and calculating the matrix $\mathbf{A}$ accordingly with the vector $\mathbf{f}$ then representing all the model parameters, not just the point source fluxes. E.g. a constant background would be a single extra element in the sources 'fluxes' vector $\mathbf{f}$ with a corresponding row of $(1, 1, 1, ..., 1)$ in the transfer matrix $\mathbf{A}$. More complicated model backgrounds with more parameters can be included by adding extra terms to $\mathbf{f}$ and $\mathbf{A}$.

3.2 Solving for unknown PRF

If the PRF, P function is unknown.

I can see how to do this iteratively with the same linear methods. i.e. a) assume PRF, b) fit fluxes c) fix flux d) fit PRF e) fix PRF f) fit fluxes etc. Ideally they could be done simultaneously but I think that is non-linear.

If we consider the PRF to be a tabulated, pixelated, function $\mathbf{P}(\mathbf{\Delta x}, \mathbf{\Delta y})$ (e.g. a PRF map) with components $P_k(\Delta x, \Delta y_k)$. We rewrite Equation 12 as

$$d_j = \sum_k f'(x_j - \Delta x_k, y_j - \Delta y_k) P_k + n_j \tag{16}$$

Where $f'(u, v)$ is the sum of the fluxes of sources at the "pixel" with coordinates (u, v). So we get

$$\mathbf{d} = \mathbf{f'(u, v)P} + \mathbf{n} \tag{17}$$

this is the same form as equation 14 but with the PRF function $\mathbf{P}$ replacing the source fluxes f and $\mathbf{A_{jk}}$ now representing the fluxes of all sources at the position that contributes to the detector j according to the PRF defined at P_k. This can be solved using the same linear methods.

3.3 Using prior information on source flux

This could be very powerful (may even remove some of these biases). It would really seriously advance the field. One way of doing this might be to multiply the posterior flux distribution functions by the prior. However, this is unsatisfactory in a number of ways, e.g. it doesn't take into account the simultaneous fitting of all the point sources and it doesn't provide any consistency with the observed data.

A better method is to treat the prior knowledge of the flux as additional (but uncorrelated) data. So we add additional terms to the data vector $\mathbf{d}$, the transfer matrix $\mathbf{A}$ and (diagonal) terms to the covariance matrix $\mathbf{N}$. E.g. if we had prior flux information on the third source $i = 3$ estimating its flux to be s with σ_s^2 we would add s to the end of data vector $\mathbf{d}$ and an extra N element row $(0, 0, 1, 0, 0, ..., 0)$ to the matrix $\mathbf{A}$, so it is now $(M + 1 \times N)$ and we would add an extra row and column to the covariance matrix $\mathbf{M}$ containing just σ^2 on the diagonal point.

3.4 Fitting the noise

Martin mentioned this before Xmas. May be hard. Probably the impact and wow factor of this would be less than 1) so probably lower priority

3.5 Bayesian formalism

In the Baysian formalism, we are interested in the posterior for the fluxes, which is the probability of the fluxes given the data, $p(f|d)$. Using Bayes theorem, we can rewrite this as (I am sorry, I am too lazy to put all the boldface commands in – everything in this paragraph is boldface)

$$p(f|d) = p(d|f)\frac{p(f)}{p(d)} \propto L(f)\pi(f)$$

where we renamed the likelihood $p(d|f) \equiv L(f)$ and the prior $p(f) \equiv \pi(f)$ for easier identification. Additionally we dropped the constant factor $p(d)$ which is relevant only if things depend on further quantities (like e.g. the noise model) but not when estimating f for everything else fixed.

The prior is now an integral part of the final equation, and we can try to find the maximum of the posterior above by deriving it with respect to f. If the prior is independent of f we recover the maximum likelihood solution as above. However, in general the maximum of the posterior will be elsewhere, and also in general the solution will be harder to find (depending on the prior). Also, for a Baysian the *maximum* of the posterior is not really relevant, instead (s)he prefers to derive confidence limits.

In principle we can always solve for f with a Monte Carlo method which basically just tries all possible f (preferably in a somewhat clever way) and so maps the posterior. Confidence limits can then be inferred by integrating out the variables that we are not interested in. However, in my (MK) limited experience, MC methods have be sophisticated *and* fast to get anywhere in such high-dimensional spaces as those here (in other words, I haven't yet found a good one). Remember, the dimensionality of the space here is the number of dimensions of f, i.e. the number of fluxes, and that's for the simplest case. But it's an interesting and generally important problem.

In general we are being simplistic if we write $p(f|d)$, since we also assumed known matrices A and N, and should really have written $p(f|d, A, N)$, and the above changes to

$$p(f|d, A, N) = p(d|f, A, N)\frac{p(f|A, N)}{p(d|A, N)}.$$

Thus, if the noise is unknown, we have to move N to the left hand side, and if the PRF is unknown, we have to move A to the lhs. Then we can try to estimate everything on the lhs simultaneously, by applying Bayes theorem again, and expressing things in terms of the likelihood $p(d|f, A, N)$ times diverse priors.

The last paragraph is very abstract, but I did not want to go into more details without knowing which cases are interesting, and how to model them.