Lecture 3

Parameter/reconstruction errors, Non-linear least-squares, correlated noise

Least Squares: $\chi^2=(d-Am)^T N^{-1}(d-Am)$

- To find best-fitting model, minimize χ^2. Calculus on matrices works like regular calculus, as long as no orders get swapped.

- $\partial\chi^2/\partial m = -A^T N^{-1}(d-Am)+\ldots=0$ (at minimum)

- We can solve for m: $A^T N^{-1} A m = A^T N^{-1} d$. Or, $m=(A^T N^{-1} A)^{-1} A^T N^{-1} d$

Error Bars

- What is the (co)variance of my model parameters?
- d_t=Am_t, where d_t=true (noiseless data), m_t=true model.
- $A^TN^{-1}Am_t=A^TN^{-1}d_t$
- $A^TN^{-1}A(m-m_t)=A^TN^{-1}(d-d_t)=A^TN^{-1}n$ where n is the actual noise I got
- $(m-m_t)=(A^TN^{-1}A)^{-1}A^TN^{-1}n$.
- $<(m-mt)(m-mt)T>=<(A^TN^{-1}A)^{-1}A^TN^{-1}n((A^TN^{-1}A)^{-1}A^TN^{-1}n)T>$
  $=<(A^TN^{-1}A)^{-1}A^TN^{-1}nnN^{-1}A^T(A^TN^{-1}A)^{-1}>$
- But, $<nn> = N$, so goes to $(A^TN^{-1}A)^{-1}A^TN^{-1}NN^{-1}A^T(A^TN^{-1}A)^{-1}=(A^TN^{-1}A)^{-1}$
- Parameter covariance is just the (inverse) of the matrix we may have already inverted!

Error Bars ctd.

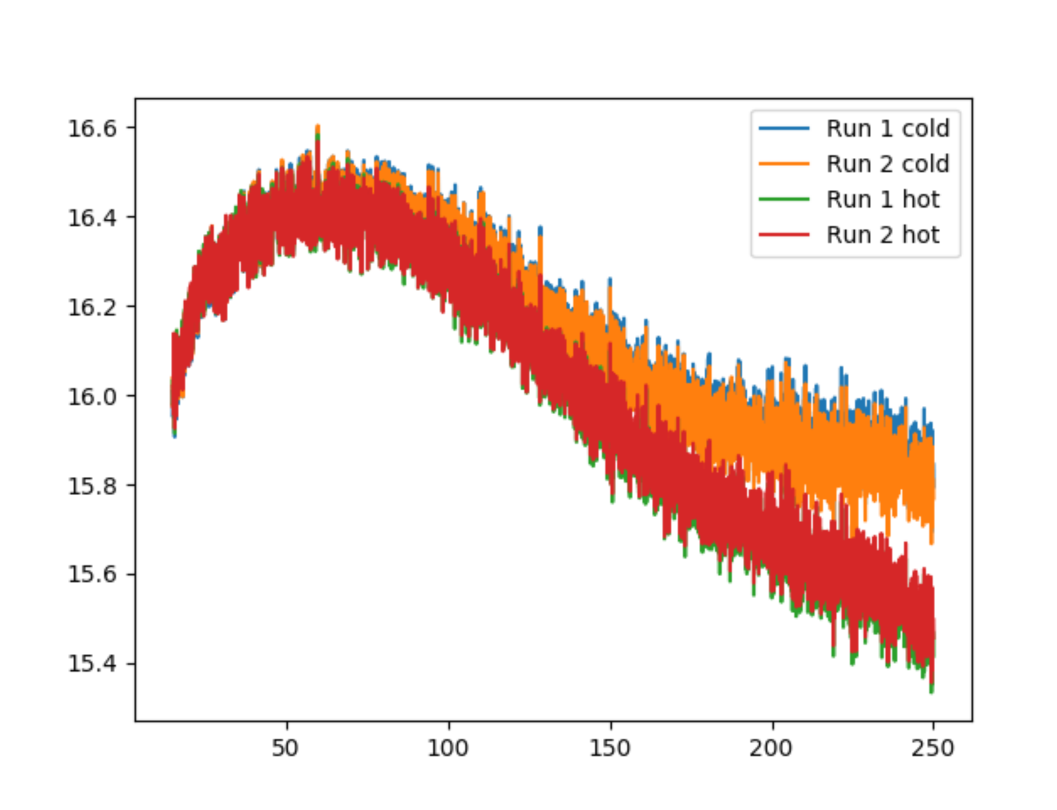
- Sometimes we want the uncertainty in our predicted data rather than model parameters.

- For this we need $<(d_t\text{-}Am)(d_t\text{-}Am)^T> = <(Am_t\text{-}Am)(Am_t\text{-}Am)^T>$ $=<A(m_t\text{-}m)(m_t\text{-}m)^T A^T>$

- We already know <(mt-m)(mt-m)T> - it's $(A^T N^{-1} A)^{-1}$. Our data errors are then just $A(A^T N^{-1} A)^{-1} A^T$.

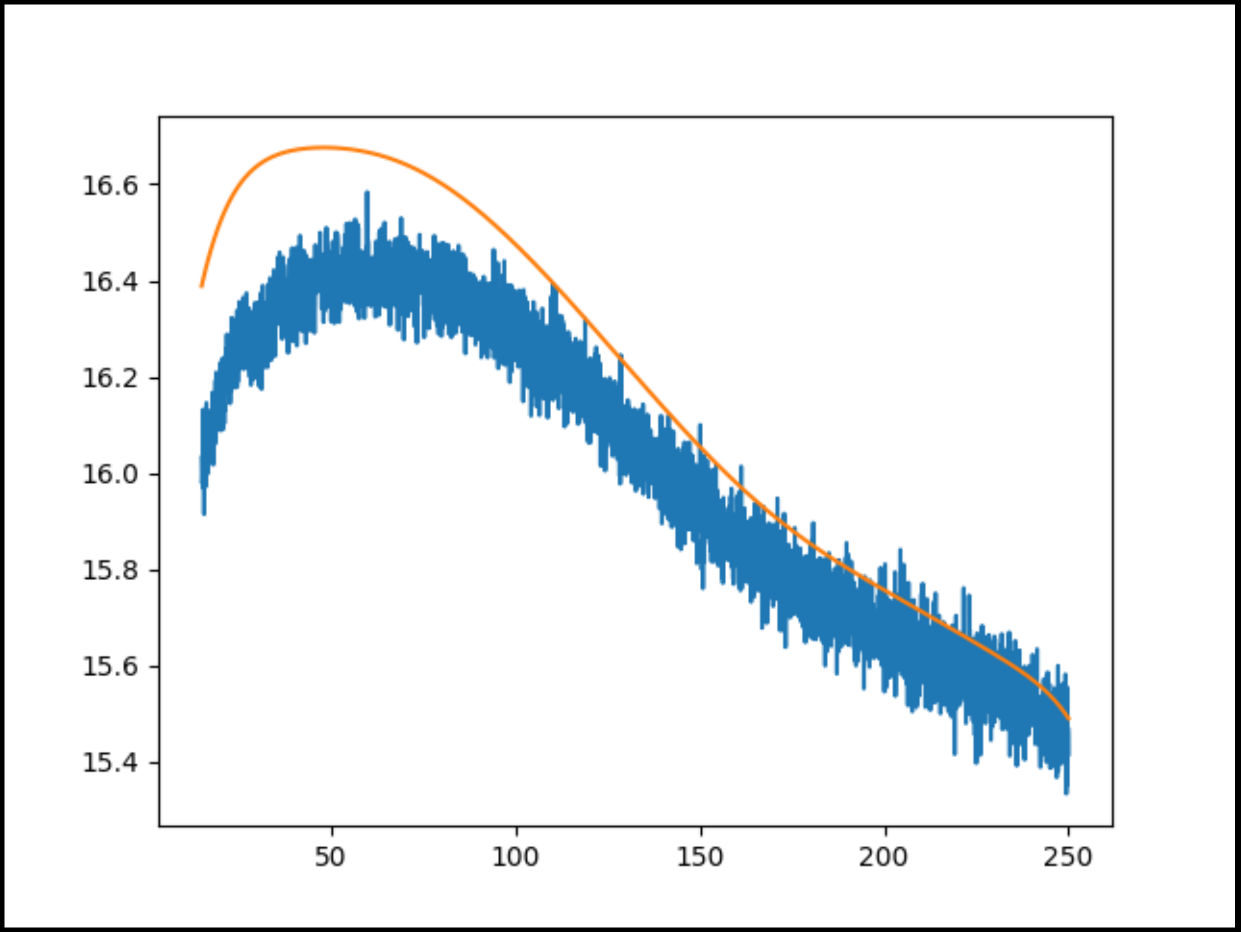
Example: Amplifier Gain

- We have amplifiers for radio telescopes which amplify incoming signals.
- The gain varies as a function of frequency - if I want to know absolute signal level, need to correct for that.
- Plot shows two different runs for an amplifier at each of two different temperatures.
- How would I model the gain vs. frequency? vs. temperature?
- How would I decide on error bars?



Higher Order

- More polynomials should be more accurate, right?
- Let's fit same problem but now to 10th order.
- Other than changing order, code is identical.
- What happens now?



Condition # and Roundoff

- Recall that the eigenvalues of a symmetric matrix are real, and the eigenvectors are orthogonal. So, $(A^TN^{-1}A)$ can be re-written $V^T\Lambda V$, where Λ is diagonal and V is orthognal (so $V^{-1}=V^T$).

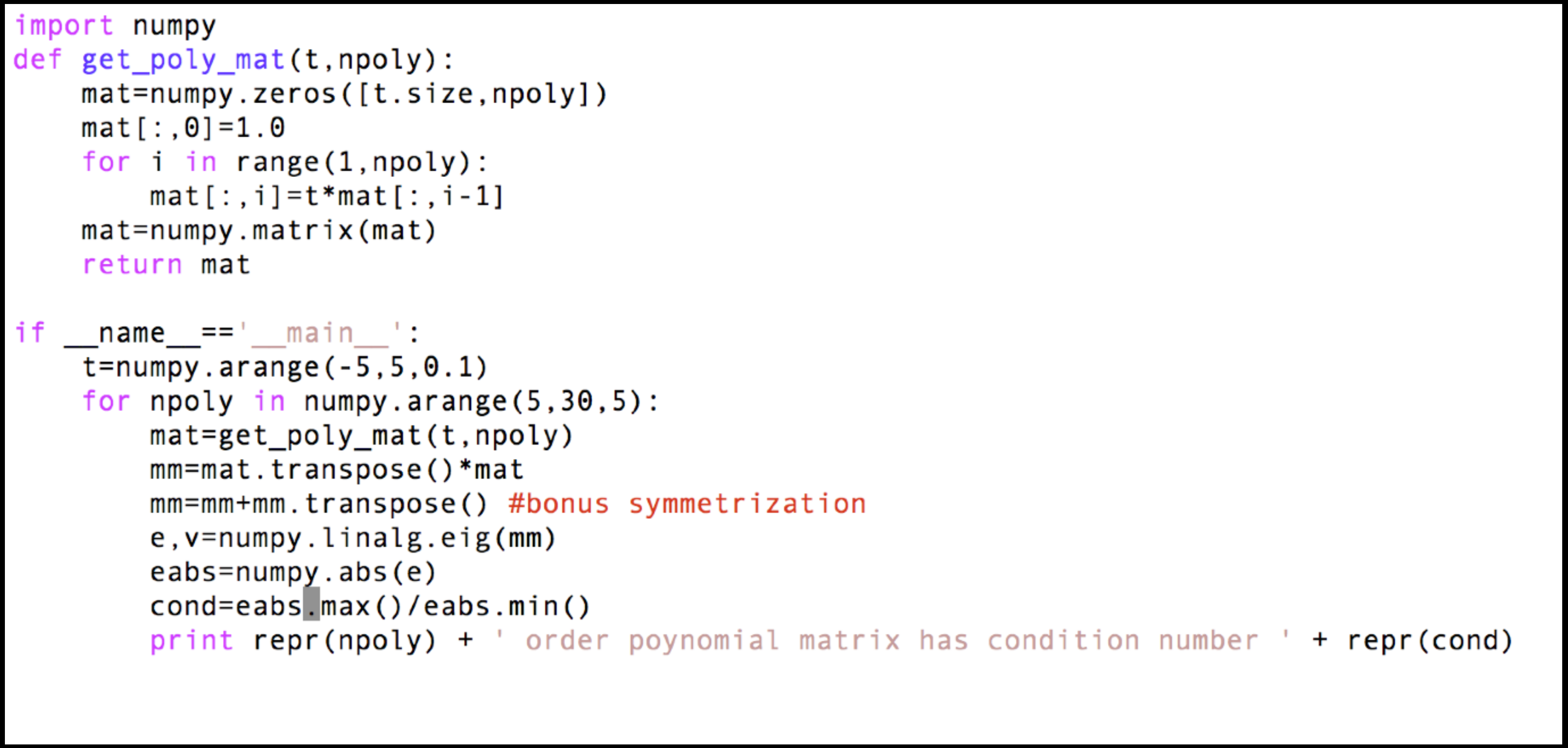
- $(ABC)^{-1}=C^{-1}B^{-1}A^{-1}$, so inverse$=V^{-1}\Lambda^{-1}(V^T)^{-1}=V^T\Lambda^{-1}V$.

- If a bunch of eigenvalues are really small, they will be huge in the inverse. Double precision numbers are good to ~16 digits, so if spread gets bigger than 10^{16}, we'll lose information in the inverse.

- Ratio of largest to smallest eigenvalue is called the condition number. If it is large, matrices are ill-conditioned, and will present problems.

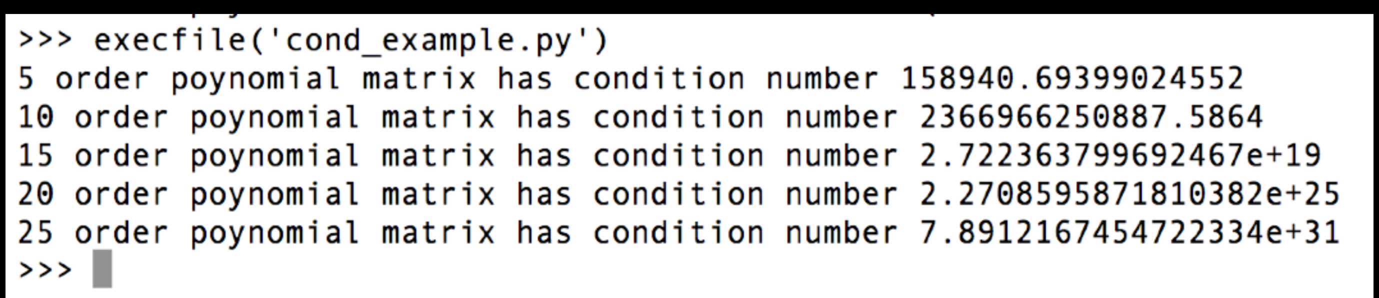
Condition # of Polynomial Matrices

- Condition # quickly blows up. So, we should have expected problems.



```python
import numpy
def get_poly_mat(t,npoly):
    mat=numpy.zeros([t.size,npoly])
    mat[:,0]=1.0
    for i in range(1,npoly):
        mat[:,i]=t*mat[:,i-1]
    mat=numpy.matrix(mat)
    return mat

if __name__=='__main__':
    t=numpy.arange(-5,5,0.1)
    for npoly in numpy.arange(5,30,5):
        mat=get_poly_mat(t,npoly)
        mm=mat.transpose()*mat
        mm=mm+mm.transpose() #bonus symmetrization
        e,v=numpy.linalg.eig(mm)
        eabs=numpy.abs(e)
        cond=eabs.max()/eabs.min()
        print repr(npoly) + ' order poynomial matrix has condition number ' + repr(cond)
```



```
>>> execfile('cond_example.py')
5 order poynomial matrix has condition number 158940.69399024552
10 order poynomial matrix has condition number 2366966250887.5864
15 order poynomial matrix has condition number 2.722363799692467e+19
20 order poynomial matrix has condition number 2.2708595871810382e+25
25 order poynomial matrix has condition number 7.8912167454722334e+31
>>> 
```

One Possibility: SVD

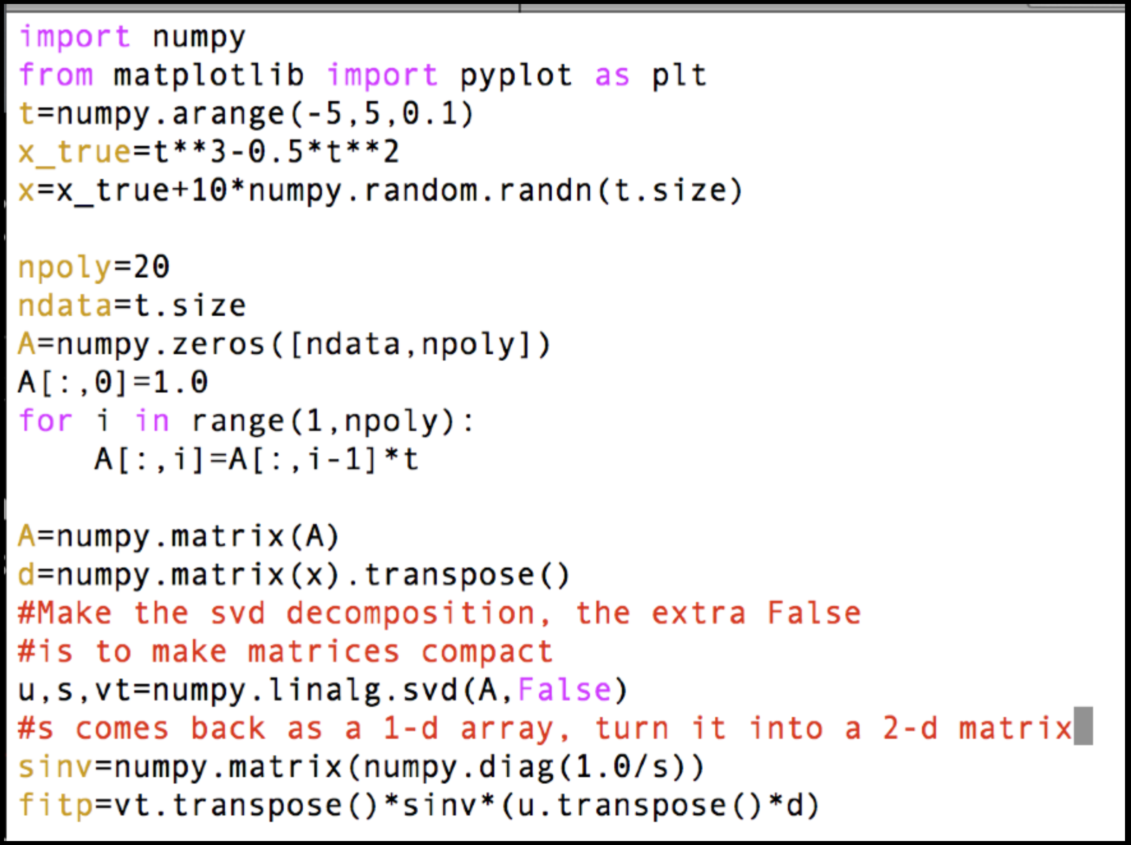
- Take noiseless case. Then solving $A^TAm=A^Tx$.
- Singular value decomposition (SVD) factors matrix $A=USV^T$, where S is diagonal, and U and V are orthogonal, and V is square. For symmetric, U=V, S=eigenvalues, but SVD works for any matrix.
- Solutions: $(USV^T)^TUSV^Tm=(USV^T)^Tx$. $VSU^TUSV^Tm=VSU^Tx$
- U^TU=identity, so cancels. $VS^2V^Tm=VSU^Tx$. S^2 squares the condition number, so that was bad. We can analytically cancel left-hand V and one copy of S: $SV^Tm=U^Tx$. Then $m=VS^{-1}U^Tx$
- NB - this can be done even faster with QR

SVD Code

- Here's how to take singular value decompositions with numpy.

- This will work better than before, but still won't get us to e.g. 100th order polynomials.

- Main issue is that simple polynomials are ill-conditioned: x^{20} looks a lot like x^{22}.



```
import numpy
from matplotlib import pyplot as plt
t=numpy.arange(-5,5,0.1)
x_true=t**3-0.5*t**2
x=x_true+10*numpy.random.randn(t.size)

npoly=20
ndata=t.size
A=numpy.zeros([ndata,npoly])
A[:,0]=1.0
for i in range(1,npoly):
    A[:,i]=A[:,i-1]*t

A=numpy.matrix(A)
d=numpy.matrix(x).transpose()
#Make the svd decomposition, the extra False
#is to make matrices compact
u,s,vt=numpy.linalg.svd(A,False)
#s comes back as a 1-d array, turn it into a 2-d matrix
sinv=numpy.matrix(numpy.diag(1.0/s))
fitp=vt.transpose()*sinv*(u.transpose()*d)
```

(Bad) Solution: pinv

- Since small eigenvalues in $A^TN^{-1}A$ are the problem, we could zero them out in the inverse.
- np.linalg.pinv will do this, with adjustable threshold for small eigenvalues.
- This will give a sensible answer. Will it give a good answer?
- Probably not! We've thrown away information by zeroing out eigenvalues. Our fit does not actually have as much freedom as it should.
- Is this a problem? That is for you to decide…

Solution: Different Poly Basis

- There are several families of polynomials that have better properties (Legendre, Chebyshev…). Usually defined on (-1,1) through recursion relations.

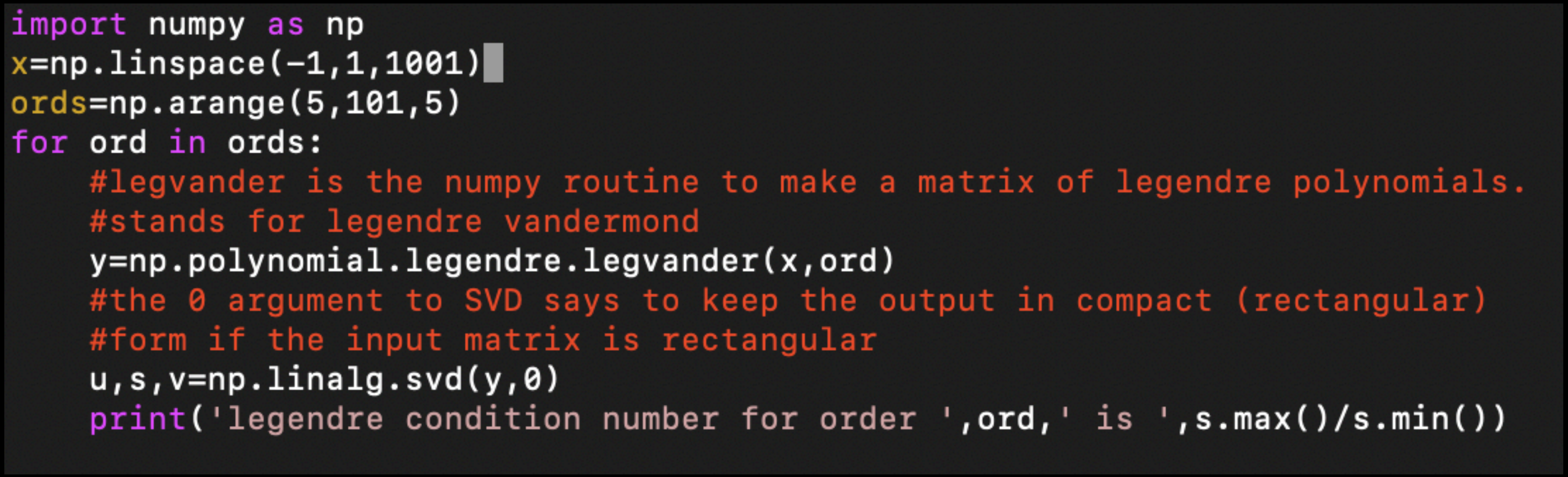
- Legendre polynomials are constructed to be orthogonal on (-1,1), so condition number should be good. If our t range is different from (-1,1), rescale so that it is.

- Key relation: $(n+1)P_{n+1}(t)=(2n+1)tP_n(t)-nP_{n-1}(t)$ with $P_0=1$ and $P_1=t$.

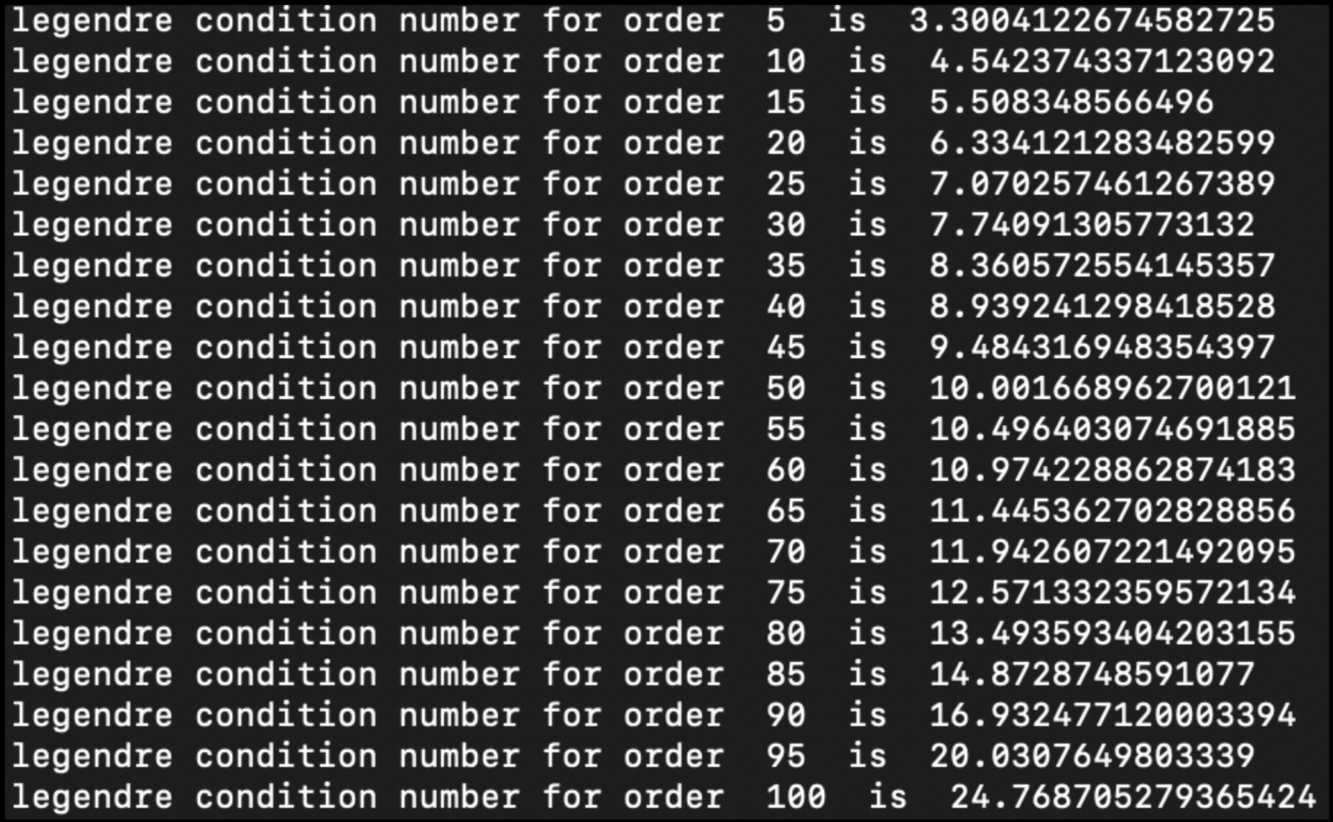
- I pick up a power of *t* each time, so these are also polynomials, just written in linear combinations that have better condition number.

- Strongly encourage you to *never* fit regular polynomials. Always use Legendre, Chebyshev…

Legendre Code



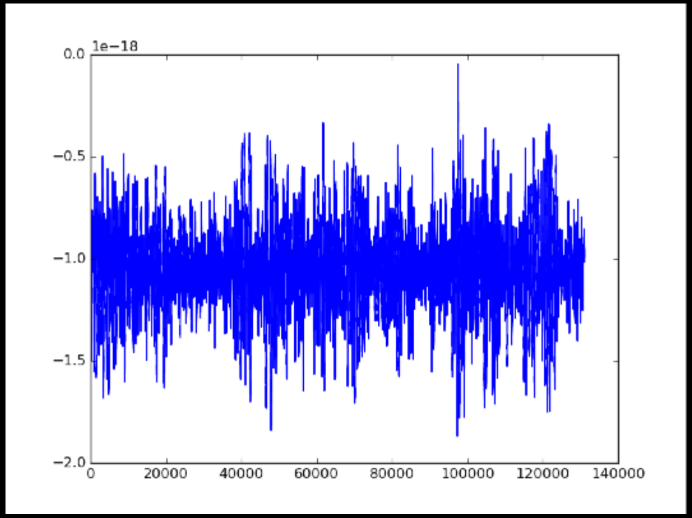
```
import numpy as np
x=np.linspace(-1,1,1001)
ords=np.arange(5,101,5)
for ord in ords:
    #legvander is the numpy routine to make a matrix of legendre polynomials.
    #stands for legendre vandermond
    y=np.polynomial.legendre.legvander(x,ord)
    #the 0 argument to SVD says to keep the output in compact (rectangular)
    #form if the input matrix is rectangular
    u,s,v=np.linalg.svd(y,0)
    print('legendre condition number for order ',ord,' is ',s.max()/s.min())
```

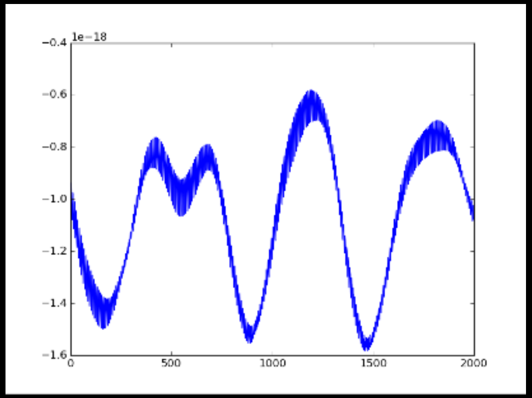


```
legendre condition number for order  5  is  3.3004122674582725
legendre condition number for order  10  is  4.542374337123092
legendre condition number for order  15  is  5.508348566496
legendre condition number for order  20  is  6.334121283482599
legendre condition number for order  25  is  7.070257461267389
legendre condition number for order  30  is  7.74091305773132
legendre condition number for order  35  is  8.360572554145357
legendre condition number for order  40  is  8.939241298418528
legendre condition number for order  45  is  9.484316948354397
legendre condition number for order  50  is  10.001668962700121
legendre condition number for order  55  is  10.496403074691885
legendre condition number for order  60  is  10.974228862874183
legendre condition number for order  65  is  11.445362702828856
legendre condition number for order  70  is  11.942607221492095
legendre condition number for order  75  is  12.571332359572134
legendre condition number for order  80  is  13.493593404203155
legendre condition number for order  85  is  14.8728748591077
legendre condition number for order  90  is  16.932477120003394
legendre condition number for order  95  is  20.0307649803339
legendre condition number for order  100  is  24.768705279365424
```

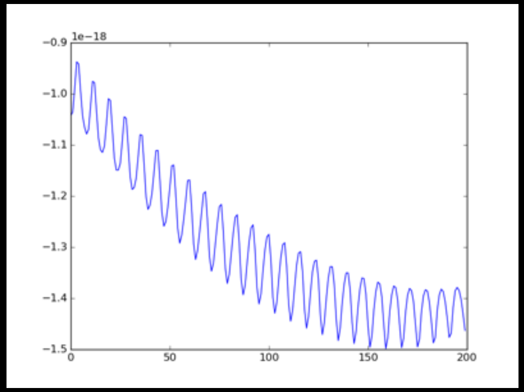
Correlated Noise

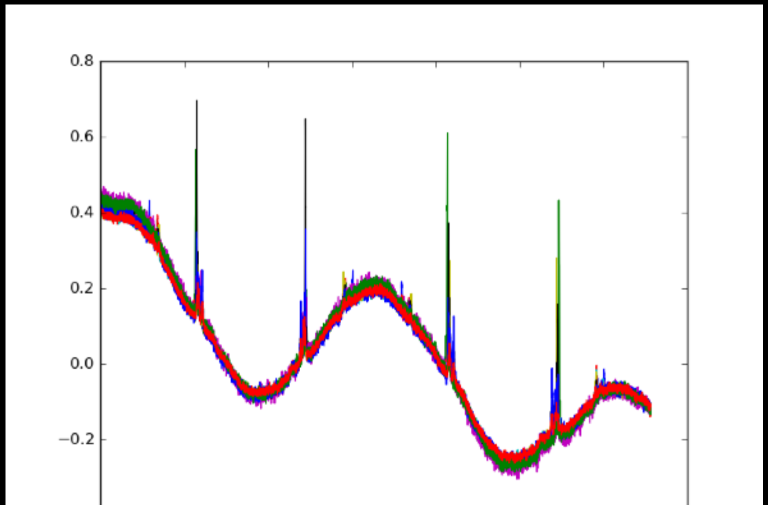
- So far, we have assumed that the noise is independent between data sets.
- Life is sometimes that kind, but very often not. We need tools to deal with this.





Right:LIGO data, with varying levels of zoom. Left: detector timestreams from Mustang 2 camera @GBT





Fortunately…

- Linear algebra expressions for χ^2 already can handle this.

- Let V be an orthogonal matrix, so $VV^T=V^TV=I$, and $d-Am=r$ (for residual)

- $\chi 2=r^TN^{-1}r = r^TV^TVN^{-1}V^TVr$. Let r->Vr, N->$VNV^T$, and $\chi 2$ expression is unchanged in new, rotated space.

- Furthermore, (fairly) easy to show that $<N_{ij}>=<r_ir_j>$.

- So, we can work in this new, rotated space without ever referring to original coordinates. Just need to calculate noise covariances N_{ij}.

Generating Correlated Noise

- Say we have a noise matrix N and want to create realizations from it. How do we do this?

- Same trick in reverse. If $d_{new}=Vd_{old}$, then I can generate d_{old} and rotate to get d_{new}.

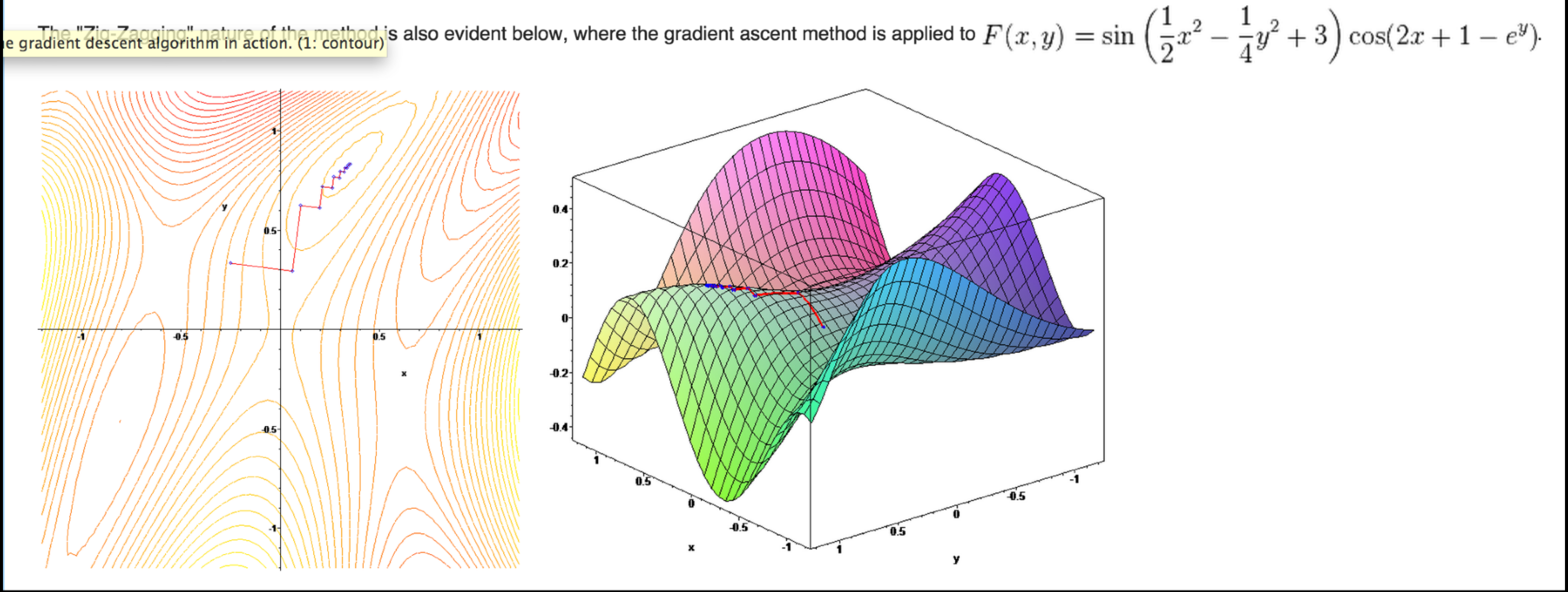
- We can pick any matrix that diagonalizes N, since we know how to generate uncorrelated data.

- A particularly useful one is Cholesky (LU equivalent for positive-definite): $N=LL^T$. We can generate simulated data just by taking L*g*, where *g* is a vector of zero-mean, unit-variance Gaussian random deviates.

Nonlinear Fitting

- Sometimes data depend non-linearly on model parameters
- Examples are Gaussian and Lorentzian $(a/(b+(x-c)^2)$
- Often significantly more complicated - cannot reason about global behaviour from local properties. May be multiple local minima
- Many methods reduce to how to efficiently find the "nearest" minimum.
- One possibility - find steepest downhill direction, move to the bottom, repeat until we're happy. Called "steepest descent."
- How might this end badly?

Steepest Descent



e gradient descent algorithm in action. (1: contour)

The "Zig-Zagging" nature of the method is also evident below, where the gradient ascent method is applied to $F(x,y)=\sin\left(\frac{1}{2}x^2-\frac{1}{4}y^2+3\right)\cos(2x+1-e^y)$.

From wikipedia.   Zigagging is inefficient.

Worked Example

- What is the best-fit mean and error for a set of uncorrelated gaussian variables with same mean but individual errors?

- A=? Show that $A^T N^{-1} A = \sum(\sigma_i^{-2})$, $A^T N^{-1} d = \sum d_i/\sigma_i^2$.

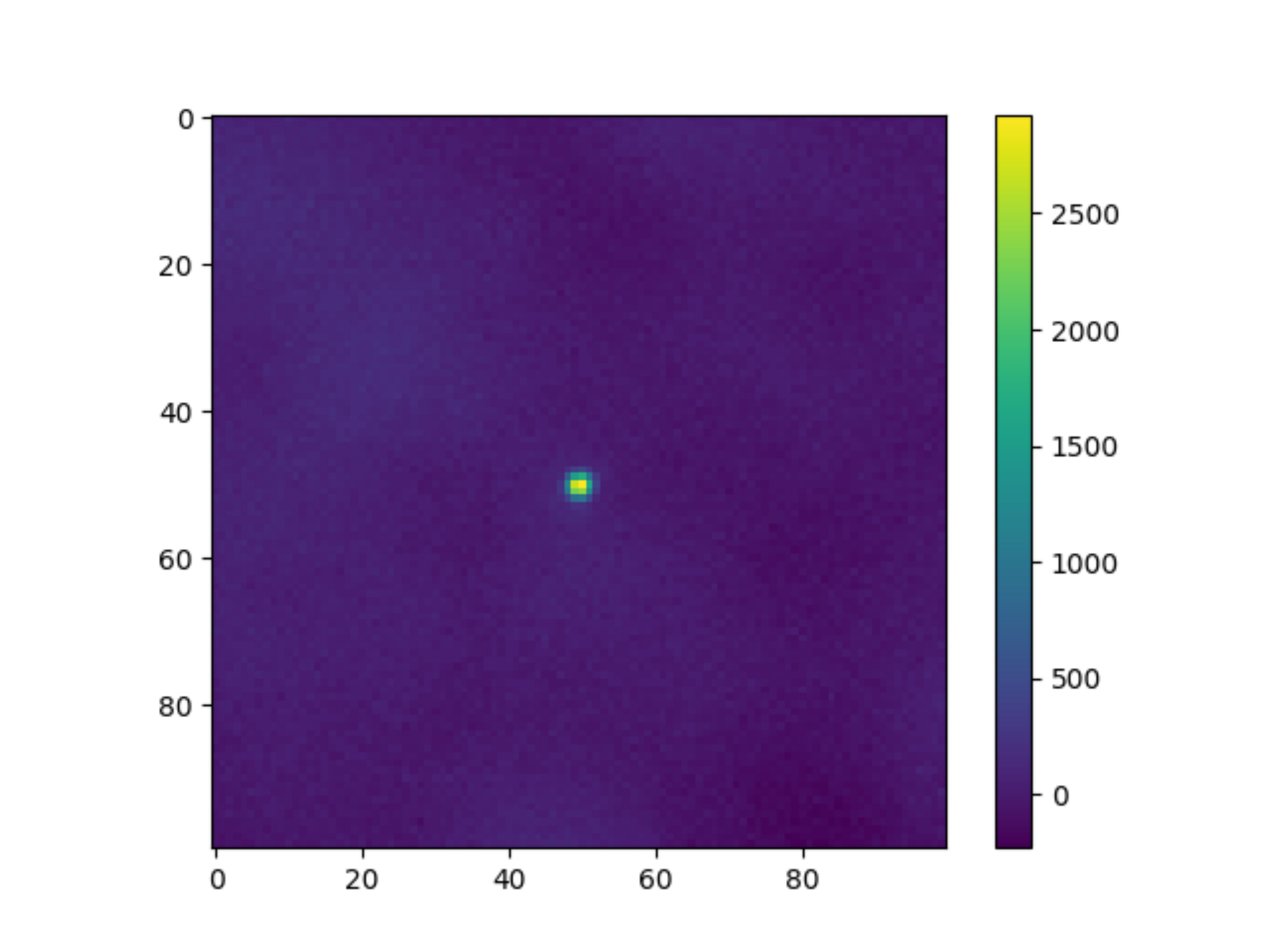
- Define weights $w_i = \sigma_i^{-2}$. Then $m = \sum w_i d_i / \sum w_i$. Variance of our estimator is $1/\sum w_i$.

Worked Example 2

- Let's assume that N is constant and diagonal, and we have a single paramers.
- Show LHS = $\sum(m_i^2/\sigma^2)$
- RHS=$\sum(d_i m_i/\sigma^2)$
- Best-fit amplitude is RHS/LHS = $\sum(d_i m_i)/\sum m_i^2$
- Error=1/sqrt(RHS) = σ/sqrt($\sum m_i^2$). If there are ~n model points with value ~1, this turns into σ/sqrt(n), as roughly expected.

Example - Source in ACT Data

- Let’s fit the amplitude of a source in ACT data.
- Look at find_act_flux.py.
- Let’s try to guess a Gaussian, fit amplitude to it.
- Map is saved as FITS. Ability to read/write/manipulate FITS images extremely useful in astronomy!



Better: Newton's Method

- linear: $<d>=Am$. Nonlinear: $<d>=A(m)$ $\chi^2=(d-A(m))^T N^{-1}(d-A(m))$
- If we're "close" to minimum, can linearize. $A(m)=A(m_0)+\partial A/\partial m*\delta m$
- Now have $\chi^2=(d-A(m_0)-\partial A/\partial m\ \delta m)^T N^{-1}(d-A(m_0)-\partial A/\partial m\ \delta m)$
- What is the gradient?

Newton's Method ctd

- Gradient trickier - $\partial A/\partial m$ depends in general on m, so there's a second derivative

- Two terms: $\nabla\chi^2=(-\partial A/\partial m)^T N^{-1}(d-A(m_0)-\partial A/\partial m\ \delta m)-(\partial^2 A/\partial m_i \partial m_j\ \delta m)^T N^{-1}(d-A(m_0)-\partial A/\partial m\ \delta m)$

- If we are near solution $d\approx A(m_0)$ and δm is small, so first term has one small quantity, second has two. Second term in general will be smaller, so usual thing is to drop it.

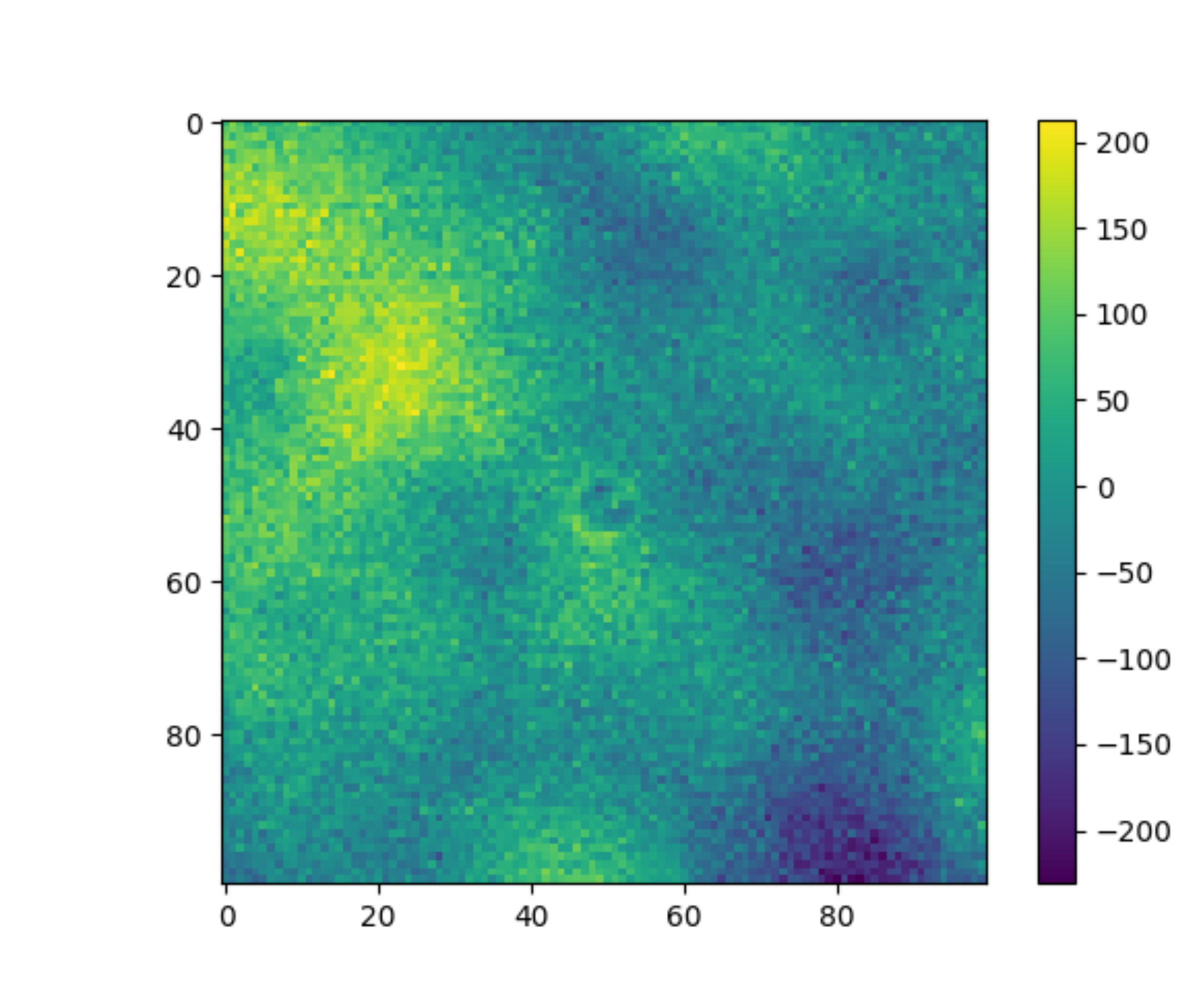
- Call $\partial A/\partial m$ A_m. Call $d-A(m_0)$ r. Then $\nabla\chi^2\approx -A_m^T N^{-1}(r-A_m\delta m)$

- We know how to solve this! $A_m^T N^{-1} A_m\ \delta m = A_m^T N^{-1} r$

How to Implement

- Start with a guess for the parameters: m_0.
- Calculate model $A(m_0)$ and local gradient A_m. Gradient can be done analytically, but also often numerically.
- Solve linear system $A_m^T N^{-1} A_m \, \delta m = A_m^T N^{-1} r$
- Set $m_0 \rightarrow m_0 + \delta m$.
- Repeat until δm is "small". For χ^2, change should be << 1 (why?).

ACT Map Example



- Look at fit_act_flux_newton.py

- This implements numerical derivatives w/ Newton's method to fit a Gaussian (including sigma, dx, dy) to the ACT data.

- How should we estimate the noise, and hence the parameter uncertainties?