Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by adisn123 on Tue, 08 Aug 2006 19:30:23 GMT
View Forum Message <> Reply to Message

I used to have a similar problem. One of the simpliest thing that I did
was using a simple
linear equation such as y =ax + b.

Overplot the linear eqaution in your original plot in such a way that
the linear line is placed
just above the red poligon (the data points that you want to throw out)
then

simply you can throw out whatever the y values are below the linear
line.

rdellsy@gmail.com wrote:
> http://photos1.blogger.com/blogger/4016/2263/320/graphroi.pn g
>
> The above is a plot of my data (minus the red polygon). I need to
> seperate the data inside the red polygon (real data) from the data
> outside the red polygon (noise, for lack of a better term) All of these
> points are already containted in an array. I'm just trying to figure
> out a way for the computer to automatically figure out what is noise
> and what isn't based on that plot distribution. Each data set is
> slightly different, but has the same overall distribution, and, for
> properly dialed in data, there is always that characteristic seperation
> between the good stuff and the bad stuff. Currently, we are manually
> setting x-boundaries and y-boundaries on our data.
> Thanks inadvance,
> Rob

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by rdellsy on Tue, 08 Aug 2006 19:57:02 GMT
View Forum Message <> Reply to Message

I considered that. Unfortunately, ambient conditions can vary the x and
y positions of the data by as much as a factor of ten. That is why I am
trying to figure out a method to compute it on the fly, since going
through the process for just five movies can take up to half an hour,
and dealing with fifty movies can be a full day's work.
Thanks,
Rob

adisn123@yahoo.com wrote:
> I used to have a similar problem. One of the simpliest thing that I did
> was using a simple
> linear equation such as y =ax + b.
>
> Overplot the linear eqaution in your original plot in such a way that
> the linear line is placed
> just above the red poligon (the data points that you want to throw out)
> then
>
> simply you can throw out whatever the y values are below the linear
> line.
>
>
>
>
>
>
> rdellsy@gmail.com wrote:
>> http://photos1.blogger.com/blogger/4016/2263/320/graphroi.pn g
>>
>> The above is a plot of my data (minus the red polygon). I need to
>> seperate the data inside the red polygon (real data) from the data
>> outside the red polygon (noise, for lack of a better term) All of these
>> points are already containted in an array. I'm just trying to figure
>> out a way for the computer to automatically figure out what is noise
>> and what isn't based on that plot distribution. Each data set is
>> slightly different, but has the same overall distribution, and, for
>> properly dialed in data, there is always that characteristic seperation
>> between the good stuff and the bad stuff. Currently, we are manually
>> setting x-boundaries and y-boundaries on our data.
>> Thanks inadvance,
>> Rob

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by btt on Tue, 08 Aug 2006 20:57:28 GMT

View Forum Message <> Reply to Message

rdellsy@gmail.com wrote:
> I considered that. Unfortunately, ambient conditions can vary the x and
> y positions of the data by as much as a factor of ten. That is why I am
> trying to figure out a method to compute it on the fly, since going
> through the process for just five movies can take up to half an hour,
> and dealing with fifty movies can be a full day's work.
> Thanks,
> Rob
>

> adisn123@yahoo.com wrote:
>> I used to have a similar problem. One of the simpliest thing that I did
>> was using a simple
>> linear equation such as y =ax + b.
>>
>> Overplot the linear eqaution in your original plot in such a way that
>> the linear line is placed
>> just above the red poligon (the data points that you want to throw out)
>> then
>>
>> simply you can throw out whatever the y values are below the linear
>> line.
>>
>>
>>
>>
>>
>> rdellsy@gmail.com wrote:
>>> http://photos1.blogger.com/blogger/4016/2263/320/graphroi.pn g
>>>
>>> The above is a plot of my data (minus the red polygon). I need to
>>> seperate the data inside the red polygon (real data) from the data
>>> outside the red polygon (noise, for lack of a better term) All of these
>>> points are already containted in an array. I'm just trying to figure
>>> out a way for the computer to automatically figure out what is noise
>>> and what isn't based on that plot distribution. Each data set is
>>> slightly different, but has the same overall distribution, and, for
>>> properly dialed in data, there is always that characteristic seperation
>>> between the good stuff and the bad stuff. Currently, we are manually
>>> setting x-boundaries and y-boundaries on our data.

Hi,

Just an end-of-the-day wildcard, but I would bin the data into a 2d histogram (ala JD's HIST_ND or the built-in HIST_2D). Then I would try to find the "saddle" between the data and noise. You'll have to fiddle with the binsize a bit to balance "lumping" and "splitting" - maybe that can be done dynamically. I dunno. But it should be quick.

It is an interesting problem that we have face here with flow cytometry - but we work the region manually as you do. I'll be interested to see what your final solution is.

Cheers,
Ben

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by JD Smith on Tue, 08 Aug 2006 21:45:42 GMT
View Forum Message <> Reply to Message

On Tue, 08 Aug 2006 16:57:28 -0400, Ben Tupper wrote:
> Hi,
>
> Just an end-of-the-day wildcard, but I would bin the data into a 2d
> histogram (ala JD's HIST_ND or the built-in HIST_2D). Then I would try to
> find the "saddle" between the data and noise. You'll have to fiddle with
> the binsize a bit to balance "lumping" and "splitting" - maybe that can be
> done dynamically. I dunno. But it should be quick.
>
> It is an interesting problem that we have face here with flow cytometry -
> but we work the region manually as you do. I'll be interested to see what
> your final solution is.

A related concept would be to:

1. Bin the original data into a 2D image, with HIST_ND, with using
   REVERSE_INDICES (call this RI#1).
2. Threshold this binned image so that it's zero below, and 1 above
   some threshold value representing the "no data" saddle. This
   threshold could be zero, but doesn't have to be (e.g. to take care
   of random noisy points in the distribution). As Ben mentions,
   you'll have to experiment to pick a good bin size.
3. Use LABEL_REGION to find all contiguous blobs of data in the
   bi-valued, thresholded, binned image.
4. Use HISTOGRAM with REVERSE_INDICES (RI#2) on the resulting "label
   image" to find the extents/centroid/etc. of the data in each "blob"
   (either roughly via the bin positions present in the blob, or more
   precisely using RI#2 and RI#1 to locate the original un-binned data
   which fall in the blob, performing an average over the data).
5. Pick the blob which is at the lower-right, and is large enough,
   etc. The criteria you use here can be quite flexible, assuming the
   "blobs" always arrive in the same pattern. You might even choose
   just to exclude certain blobs that have a given shape and relative
   position, and then take everything else.
6. Find the bins which belong to the chosen blob(s), using RI#2, and
   then locate the data points within these original bins, with RI#1.
7. Give yourself a raise.

This is actually a very good exercise to try if you want to know
everything about HISTOGRAM and REVERSE_INDICES.

JD

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by rdellsy on Tue, 08 Aug 2006 22:20:13 GMT
View Forum Message <> Reply to Message

I'm a tad confused about what you're suggesting. I'll try and work it out, but I'm still fairly new to IDL, so if you could give an IDL or pseudo-code example of what you're trying to explain, I would appreciate. If that's too much work, I understand, and I'll just try to puzzle it out on my own.
Thanks,
Rob

JD Smith wrote:
> On Tue, 08 Aug 2006 16:57:28 -0400, Ben Tupper wrote:
>> Hi,
>>
>> Just an end-of-the-day wildcard, but I would bin the data into a 2d
>> histogram (ala JD's HIST_ND or the built-in HIST_2D). Then I would try to
>> find the "saddle" between the data and noise. You'll have to fiddle with
>> the binsize a bit to balance "lumping" and "splitting" - maybe that can be
>> done dynamically. I dunno. But it should be quick.
>>
>> It is an interesting problem that we have face here with flow cytometry -
>> but we work the region manually as you do. I'll be interested to see what
>> your final solution is.
>
> A related concept would be to:
>
> 1. Bin the original data into a 2D image, with HIST_ND, with using
> REVERSE_INDICES (call this RI#1).
> 2. Threshold this binned image so that it's zero below, and 1 above
> some threshold value representing the "no data" saddle. This
> threshold could be zero, but doesn't have to be (e.g. to take care
> of random noisy points in the distribution). As Ben mentions,
> you'll have to experiment to pick a good bin size.
> 3. Use LABEL_REGION to find all contiguous blobs of data in the
> bi-valued, thresholded, binned image.
> 4. Use HISTOGRAM with REVERSE_INDICES (RI#2) on the resulting "label
> image" to find the extents/centroid/etc. of the data in each "blob"
> (either roughly via the bin positions present in the blob, or more
> precisely using RI#2 and RI#1 to locate the original un-binned data
> which fall in the blob, performing an average over the data).
> 5. Pick the blob which is at the lower-right, and is large enough,
> etc. The criteria you use here can be quite flexible, assuming the
> "blobs" always arrive in the same pattern. You might even choose
> just to exclude certain blobs that have a given shape and relative
> position, and then take everything else.
> 6. Find the bins which belong to the chosen blob(s), using RI#2, and

> then locate the data points within these original bins, with RI#1.
> 7. Give yourself a raise.
>
> This is actually a very good exercise to try if you want to know
> everything about HISTOGRAM and REVERSE_INDICES.
>
> JD

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by btt on Wed, 09 Aug 2006 12:37:37 GMT
View Forum Message <> Reply to Message

rdellsy@gmail.com wrote:
> I'm a tad confused about what you're suggesting. I'll try and work it
> out, but I'm still fairly new to IDL, so if you could give an IDL or
> pseudo-code example of what you're trying to explain, I would
> appreciate. If that's too much work, I understand, and I'll just try to
> puzzle it out on my own.

Perhaps you could posted some data as a text file (not much) that
contains the "bunching" your graphic shows.

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by edward.s.meinel@aero. on Wed, 09 Aug 2006 14:38:26 GMT
View Forum Message <> Reply to Message

That looks like a typical clustering problem. ENVI has some built-in
clustering algorithms.

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by James Kuyper on Wed, 09 Aug 2006 15:38:18 GMT
View Forum Message <> Reply to Message

edward.s.meinel@aero.org wrote:
> That looks like a typical clustering problem. ENVI has some built-in
> clustering algorithms.

Given the nature of the data, I'd recommend CLUSTER_TREE, with
LINKAGE=0. When you prune the tree down to about three clusters, I'd
expect one of them to be a good match to your polygon.

Subject: Re: Need Some Advice on Seperating Out Some Data

Posted by rdellsy on Wed, 09 Aug 2006 17:11:43 GMT
View Forum Message <> Reply to Message

http://s8.quicksharing.com/v/2874818/bm.gdf.html

There's an example data file. Thanks for your ongoing help so far with
this project of mine.
Rob

kuyper@wizard.net wrote:
> edward.s.meinel@aero.org wrote:
>> That looks like a typical clustering problem. ENVI has some built-in
>> clustering algorithms.
>
> Given the nature of the data, I'd recommend CLUSTER_TREE, with
> LINKAGE=0. When you prune the tree down to about three clusters, I'd
> expect one of them to be a good match to your polygon.

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by JD Smith on Wed, 09 Aug 2006 17:47:43 GMT
View Forum Message <> Reply to Message

On Tue, 08 Aug 2006 15:20:13 -0700, rdellsy wrote:

> I'm a tad confused about what you're suggesting. I'll try and work it
> out, but I'm still fairly new to IDL, so if you could give an IDL or
> pseudo-code example of what you're trying to explain, I would
> appreciate. If that's too much work, I understand, and I'll just try to
> puzzle it out on my own.

You might find much of what you need in the HISTOGRAM tutorial:

http://www.dfanning.com/tips/histogram_tutorial.html

But before you go that route, you might first try the CLUSTER function
in IDL (which I just read up on). Here's an example using a fake
clustered data set with 5 clusters. You'll probably have to experiment
with the number of clusters.

JD

```
 tvlct,[255,0,0,0,255,255],[0,255,0,255,255,0],[0,0,255,255,0 ,255],1
n_clust=5

;; Make some flake clustered data
if n_elements(x) ne 0 then begin
   n=1000
```

```
   clust_fwhm=.2
   cposx=randomu(sd,n_clust) & cposy=randomu(sd,n_clust)
   cind=fix(randomu(sd,n)*n_clust)

   x=clust_fwhm
   fac=2*sqrt(2*alog(2))
   x=randomn(sd,n)*clust_fwhm/fac+cposx[cind]
   y=randomn(sd,n)*clust_fwhm/fac+cposy[cind]
endif

array=transpose([[x],[y]])
w=clust_wts(array,N_CLUSTERS=n_clust)
c=cluster(array,w)
h=histogram(c,REVERSE_INDICES=ri)
nh=n_elements(h)

plot,x,y,PSYM=4,/ISOTROPIC

cen=make_array(2,nh,VALUE=!VALUES.F_NAN)
for i=0,nh-1 do begin
   if ri[i+1] eq ri[i] then continue
   take=ri[ri[i]:ri[i+1]-1]
   oplot,x[take],y[take],PSYM=4,COLOR=i+1
   cen[0,i]=[mean(x[take]),mean(y[take])]
endfor

;; Find the lower right cluster
void=max(cen[0,*]-cen[1,*],lrc,/NAN)

;; Highlight it
keep=ri[ri[lrc]:ri[lrc+1]-1]
oplot,x[keep],y[keep],PSYM=6,SYMSIZE=2

END
```

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by rdellsy on Wed, 09 Aug 2006 20:13:12 GMT

View Forum Message <> Reply to Message

Thanks for that. I took it, and played around with it a bit to get it to work. [Errors I found were: x and y don't concatinate in the line 'array=transpose([[x],[y]])' and I found I had to comment away the /ISOTROPIC in the plotting.) Unfortunately, it seems that cluster seperates on a purely 1 dimensional basis. I tried discarding the histogram related code in favor of a much simpler system in case that was the problem, and it still didn't work. If you look at the data set I provided, the problem should be self evident.

Incidentally, I replaced everything from
h=histogram(c,reverse_indices=ri) down to the second to last line with:
--
plot,x,y,psym=2
bmax=max(array[0,*],maxsubsc)
goodc=c[maxsubsc]
keep=where(c[*] eq goodc)
--
I feel that my code may be a tad more efficient, though I don't know
how efficient the WHERE command is.

Anywho, I'm looking CLUSTER_TREE right now, which shows some more
promise. If I understand it correctly, it works using distance appart,
not coordinates which is a bit more useful, I think, for my problem.
I'm just not sure how I can take the output and turn it into a set of
clusters.

Thanks for all the help!
- Rob

JD Smith wrote:
> On Tue, 08 Aug 2006 15:20:13 -0700, rdellsy wrote:
>
>> I'm a tad confused about what you're suggesting. I'll try and work it
>> out, but I'm still fairly new to IDL, so if you could give an IDL or
>> pseudo-code example of what you're trying to explain, I would
>> appreciate. If that's too much work, I understand, and I'll just try to
>> puzzle it out on my own.
>
> You might find much of what you need in the HISTOGRAM tutorial:
>
> http://www.dfanning.com/tips/histogram_tutorial.html
>
> But before you go that route, you might first try the CLUSTER function
> in IDL (which I just read up on). Here's an example using a fake
> clustered data set with 5 clusters. You'll probably have to experiment
> with the number of clusters.
>
> JD
>
> tvlct,[255,0,0,0,255,255],[0,255,0,255,255,0],[0,0,255,255,0 ,255],1
> n_clust=5
>
> ;; Make some flake clustered data
> if n_elements(x) ne 0 then begin
> n=1000
> clust_fwhm=.2

```
>    cposx=randomu(sd,n_clust) & cposy=randomu(sd,n_clust)
>    cind=fix(randomu(sd,n)*n_clust)
>
>    x=clust_fwhm
>    fac=2*sqrt(2*alog(2))
>    x=randomn(sd,n)*clust_fwhm/fac+cposx[cind]
>    y=randomn(sd,n)*clust_fwhm/fac+cposy[cind]
> endif
>
> array=transpose([[x],[y]])
> w=clust_wts(array,N_CLUSTERS=n_clust)
> c=cluster(array,w)
> h=histogram(c,REVERSE_INDICES=ri)
> nh=n_elements(h)
>
> plot,x,y,PSYM=4,/ISOTROPIC
>
> cen=make_array(2,nh,VALUE=!VALUES.F_NAN)
> for i=0,nh-1 do begin
>    if ri[i+1] eq ri[i] then continue
>    take=ri[ri[i]:ri[i+1]-1]
>    oplot,x[take],y[take],PSYM=4,COLOR=i+1
>    cen[0,i]=[mean(x[take]),mean(y[take])]
> endfor
>
> ;; Find the lower right cluster
> void=max(cen[0,*]-cen[1,*],lrc,/NAN)
>
> ;; Highlight it
> keep=ri[ri[lrc]:ri[lrc+1]-1]
> oplot,x[keep],y[keep],PSYM=6,SYMSIZE=2
>
> END
```

Subject: Re: Need Some Advice on Seperating Out Some Data
Posted by James Kuyper on Thu, 10 Aug 2006 19:30:24 GMT
View Forum Message <> Reply to Message

rdellsy@gmail.com wrote:
> http://s8.quicksharing.com/v/2874818/bm.gdf.html

I don't recognise that file type. How do I read it? I'd prefer an
answer in terms of IDL commands.
