Tina Memo No. 2014-004
Internal Report

Tutorial: Using Tina Vision's Quantitative Pattern Recognition Tool.

P.D.Tar.

Last updated
07 / 06 / 2014

ISBE, Medical School,
University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT, UK.

Tutorial: Using Tina Vision's Quantitative Pattern Recognition Tool

Contents

1 The QPR Tool

1.1 Getting Started

To follow this tutorial you will first need to download the QPR Tool example files and make sure that a local version of Tina Vision is available. This tutorial makes use of the following files:

- class_a_examples.cls
- class_b_examples.cls
- multi_class_examples.cls
- multi_class_errors.cls
- corr_example.cls
- header

1.2 Linear Poisson Models

The Quantitative Pattern Recognition Tool (QPR Tool) is accessible through the Histogram tool. This tool is intended to work with histograms which can be modelled as linear combinations of Probability Mass Functions (PMFs):

$$\mathbf{H}_X \approx \mathbf{M}_X = \sum_k P(X|k)\mathbf{Q}_k \tag{1}$$

where $\mathbf{H}_X$ is the observed frequency in bin X of histogram $\mathbf{H}$; $\mathbf{M}_X$ is the modelled frequency of bin X; $P(X|k)$ is the probability of an event occurring in bin X, originating from independent component k; and $\mathbf{Q}_k$ is the quantity of component k within the histogram. Solutions to the linear model are based upon maximising the following Extended Maximum Likelihood function:

$$\ln \mathcal{L} = \sum_X \ln \left[\sum_k P(X|k)\mathbf{Q}_k \right] \mathbf{H}_X - \sum_k \mathbf{Q}_k \tag{2}$$

which assumes independent Poisson noise on each histogram bin. The analysis methods are based upon Linear Poisson Models [Tina Memos 2012-003, 2013-006]. The tool provides facilities for:

- converting images found in the Image Calculator into grey level histograms;
- performing Independent Component Analysis (ICA);
- combining related PMFs into different classes of data;
- fitting models to new histograms to estimate the quantities of constituent classes;
- analysing errors on estimated quantities;
- inspecting model-data residuals.

This tutorial will guide users through the process of generating example data, importing data into the QPR Tool, building a multi-class model, then applying the model to new data in order to make quantity measurements.

2 Creating Training Data

2.1 Manual Creation

A source of histogram data is required before the QPR Tool can be demonstrated. The easiest way to generate example histograms is by making test images, via the Image Create Tool, and importing their grey level histograms into the Histogram Tool:

1. From a command line, start Tina using "./tinaTool -f setup". This will open all necessary windows and tools, including the Imcalc Tool and the Create Tool.
2. In the Image Create Tool, click on the **chequer** button. A black and white pattern should appear in the imcalc window.
3. Enter the value 100 in the **noise** text field and the value 1 in both the **ax** and **ay** text fields, then click on the **noise** button. The black and white pattern should move into the imcalc2 window and white noise should appear in the imcalc window.
4. In the Imcalc Tool, click on the + button. The pattern and noise should now be superimposed in the imcalc window.
5. In the Imcalc Parameters window, enter the value 300 in the **const** text field and click on the **+k** Imcalc Tool button. Nothing visual should change, but this will shift the origin of the image grey levels away from zero.
6. In the QPR Tool, make sure the **Results** field contains the value 0, **Min** contains 0, **Max** contains 1200 and **Bins** contains 100. These values tell the tool to place resulting histograms into the Histogram Tool starting at position 0, with an x-axis ranging from 0 to 1200, divided into 100 bins.
7. Click on the QPR Tool **Pop** button, then the Histogram Tool **plot hist** button. A plot should appear in the histogram window showing the distribution of grey levels found within the simulated data.

2.2 Macros

Dozens of histograms are required to demonstrate most features of the QPR Tool. It would be very tedious to repeat the steps above for each histogram. To speed things up, macros have been provided to create test images in batches. To use the macros follow these steps:

1. From a command line, start Tina using "./tinaTool -f setup". This will open all necessary windows and tools, including the Imcalc Tool and the Create Tool.
2. In the Macro Tool, enter "class_a_examples" in the **Macro File** field, then click **run**. This will create 9 sample images which will be stored on the stack.
3. In the QPR Tool, make sure the **Results** field contains the value 0, **Min** contains 0, **Max** contains 1200 and **Bins** contains 100.
4. Click on the QPR Tool **Pop** button, then the Histogram Tool **plot hist** button. Use the < and > buttons to scan through the histograms. 9 histograms (from index 0 to 8) should be populated.
5. Repeat the above steps, this time using the macro "class_b_examples".

The QPR Tool can work with many sources of histograms, not just grey level distributions. An alternative method for quickly populating histograms is using the Histogram Tool **Load CSV** button. This form of input will be demonstrated later.

2.3 Questions

The QPR Tool can only successfully analyse histograms if they can be approximated using equation 1 and solved using equation 2. It is therefore important to understand the links between data and the terms of these equations:

1. Equation 1 and 2 both contain a sum over k, where each k denotes a particular model component. How many components would you expect to find in the above training data and what might they relate to?

2. The $P(X|k)$ terms approximate the probability distributions of components, without making any assumptions regarding their shapes. What shape would you expect the distributions in the training data to have and how are these shapes generated by the Image Create Tool?

3. Which physical quantities might be related to the $\mathbf{Q}_k$ terms?

4. Given how the training data is generated, why should the resulting histogram bins be (approximately) independent Poisson variables?

5. What aspect of equation 2 assumes independence between histogram bins?

3 Independent Component Analysis

3.1 ICA Models

The QPR Tool can inspect multiple independent examples of histograms, which were generated by the same process, in order to extract a set of common probability distributions, i.e. $P(X|k)$. The aim is to extract a sufficient number of common components to allow a satisfactory reconstruction of the data via the linear model, whilst not over-fitting to patterns in the noise. This process is known as Independent Component Analysis and can be achieved by following these steps:

1. From a command line, start Tina using "./tinaTool -f setup".
2. Run the macro "class_a_examples" and follow similar steps as section 2.2 to load the 9 histograms into the Histogram Tool.
3. The QPR Tool needs to know which histograms are to be analysed. Enter the value 0 in the QPR Tool **First** text field and 8 in the **Last** text field.
4. The QPR Tool will plot extracted components as histograms. Enter the value 9 in the **Results** text field to indicate that they should be plotted immediately after the training data.
5. Enter the value 1 into the **Comp count** field then click on the **ICA** button. This will attempt to describe all 9 training histograms using only 1 component.
6. Click on the Histogram Tool **plot hist** button and use the $<$ and $>$ buttons to scan through the histograms. There should be curves superimposed over histograms 0 to 8, which are the bin frequencies predicted by equation 1, i.e. $\mathbf{M}_X$. Histogram 9 should show the extracted component.

   The agreement between histogram bins, $\mathbf{H}_X$, and modelled frequencies, $\mathbf{M}_X$, will be rather poor, which can be seen visually.
7. Enter the value 2 into the **Comp count** field (not to be confused with the similarly named **Comp** field) then click on the **ICA** button. This will attempt to describe all 9 training histograms using 2 components. Inspecting the results again in the Histogram Tool should reveal much better agreement between data and model. Histograms 9 and 10 should show the extracted components.
8. Repeat ICA again, this time extracting 10 components (this may take several minutes of run-time). The agreement between data and model should now be *too* good. The plot of $\mathbf{M}_X$ will be less smooth than earlier, as the system has been over-trained. Histograms 9 to 18 should show very noisy extracted components.

Knowing how many components to extract either requires prior knowledge, or some goodness-of-fit score. The QPR Tool provides a goodness-of-fit score for each modelled histogram which should be approximately 1 for the "correct" number of components. This score will be investigated later.

3.2 Saving Models

The example macros "class_a_examples" and "class_b_examples" contain different classes of data. Both classes contain two independent components. Before the QPR Tool can be used to make quantity measurements it needs to be trained to know which components belong to which classes. This is achieved by building ICA models for each class of data, saving each class to a different file, then constructing a new model containing all classes of interest. Saving ICA models is straightforward:

1. From a command line, start Tina using "./tinaTool -f setup".
2. Run the "class_a_examples" macro, load the histograms into the Histogram Tool, then run ICA to extract 2 components. Refer back to section 3.1 if necessary.
3. In the QPR Tool **File** text field, enter the file name "class_a" then click on the **Store** button. Make sure this is done in the QPR Tool, **not** the Histogram Tool, which has some very similar fields and buttons.
4. Repeat the process again for the "class_b_examples" macro, saving the resulting ICA model as "class_b".

3.3 Questions

Having the ability to create ICA models of data is important, as data distributions are often not fixed, i.e. their shape can change depending upon the context. The resulting linear models might then be used to describe previously unseen combinations of components, dynamically adjusting model weights ($\mathbf{Q}$) to best fit new data. This is known as transduction. Using these dynamic models in practice requires an understanding of the relationships between training data, modelled components (PMFs) and new data:

1. The original generator of the training data, i.e. the Image Create Tool, produces Gaussian distributions. However, when extracted components are viewed in the Histogram Tool they appear to be skewed. Why is this?
2. If it was important for ICA to extract the *true* generating distributions, how could the training data be changed to improve the resulting PMFs?
3. Beyond issues of over-fitting, why would it be pointless to extract more than 9 components from the training data created by the class a and b macros?
4. In general, is it safe to attribute physical meaning to individual components?
5. Is it safe to attribute physical meaning to individual classes?
6. What kind of problems might arise if an ICA model is applied to new data?

4 Multi-Class Quantity Measurements

4.1 Building A Multi-Class Model

The main purpose of the QPR Tool is to make quantity measurements. To demonstrate this, the previously saved ICA models ("class_a" and "class_b" files) can be loaded into the QPR Tool as part of a combined linear model with 4 components and 2 classes in total. This model can then be fitted to new data, which contains a combination of both classes, to estimate how much of each class is present. To create the new multi-class model the following steps can be taken:

1. From a command line, start Tina using "./tinaTool -f setup".
2. In the QPR Tool, make sure **Min** contains 0, **Max** contains 1200 and **Bins** contains 100. This is to ensure that the parameters of the new model are the same as those used during training.
3. Enter the value 4 in **Comp Count**, and 2 in **Class count**, then click on the **New** button. This will create a new empty model. The tinatool text window should display a brief summary, with a 2 by 4 "Class map" matrix at the bottom containing all zeros. This matrix has a row per class and a column per component.
4. Enter the file name "class_a" into the QPR Tool **File** field, then make sure that the value 0 is in both the **Comp** and **Class** fields. This indicates that the components which are about to be loaded should fill the model from index 0 and belong to class 0.
5. Click on **Fetch** button. The tinatool text window should display an updated summary, with the "Class map" now showing the value 1 in the first two positions. This confirms that the first 2 components have been associated with class 0.
6. Enter the file name "class_b" into the **File** field, then make sure that the value 2 is in the **Comp** field and 1 is in the **Class** field. This indicates that the components which are about to be loaded should fill the model from index 2 and belong to class 1.
7. Click on **Fetch** button. The tinatool text window should display an updated summary, with the "Class map" now showing the value 1 in the last two positions of the bottom row. This confirms that the next 2 components have been associated with class 1.
8. The components which have been loaded can be viewed in the Histogram Tool, starting from whichever index was chosen in the QPR Tool **Results** field.

4.2 Making Measurements

Once a multi-class model has been loaded, via the steps of section 4.1, new histogram data can be entered containing combinations of classes for analysis. Again, the Image Create Tool can be used to generate these multi-class testing histograms. A macro has been provided to speed this process up:

1. Enter "multi_class_examples" into the Macro Tool and run it. This should produce 9 new test images containing combinations of class a and b data.
2. Pop the images' histograms into the QPR Tool, making sure they are loaded starting from position 0.
3. In the QPR Tool, set the **First** and **Last** fields to select all of the popped histograms, then click on the **fit** button.
4. Scan through the histograms in the Histogram Tool to confirm that the model has been fitted to each example.
5. Scroll through the tinatool text window to view the results of the fits. In particular, find the "Block 0", "Block 1" and "Block 2" results.
6. Interpret the descriptions as follows:

   Within each block, the first line of output gives the block number and the goodness-of-fit. A fit score close to unity indicates a successful analysis. Blocks 0 and 2 should indicate success, whereas block 1 should indicate a fit failure.

   Within each block, the first set of numbers relates to the 4 components of the model. The second set of numbers relates to the 2 classes of the model. Each row of numbers contains the following information:

   (component or class ID), (estimated quantity), (estimated variance), (estimated standard deviation i.e. error).

4.3 Questions

Measurements made using the QPR Tool must be interpreted carefully. Estimated quantities and their predicted errors cannot be trusted if the goodness-of-fit is too far above unity. Furthermore, there are situations when a goodness-of-fit close to unity conceals hidden problems with the data. Understanding the output, and how it was generated, is essential for making trustworthy measurements.

1. What might have caused the fit failure seen in block 1?
2. What is the relationship between component and class errors?
3. The goodness-of-fit score is a chi-square per degree of freedom. What exactly does a goodness-of-fit of unity mean?
4. What type of problems might occur which might not be spotted using the goodness-of-fit?

5 Checking Error Predictions

5.1 Repeating a Measurement

Error bars are essential when interpreting measured quantities, but error bars are only useful if they can be trusted. The validity of estimated errors can be checked by repeating a measurement, then comparing the predicted spread of answers to the one observed. When the QPR Tool is used to make measurements a log file is produced called "qpr_log.csv". This log records the block number, class number, estimated quantity, estimated error and goodness-of-fit of each measurement. If ground-truth is available for the data then the information in the log file can be used to corroborate error predictions. To do this, a measurement first needs to be repeated:

1. Following the process of 4.1, create a 4 component, 2 class model, and load into it the stored models for class a and b data.
2. Enter "multi_class_errors" into the Macro Tool **Macro File** field. **Do not run the macro yet**.
3. Enter the value 1 in the Macro Tool **Start** field, 100 in the **End** field and 1 in the **current** field, then click on the **loop** button. This will create 100 example images containing an identical signal, but perturbed with independent noise.
4. Use the QPR Tool to pop the histograms, placing results starting at position 0, as usual.
5. Select the first 100 histograms to analyse, i.e. set **First** to 0 and **Last** to 99.
6. Delete the existing "qpr_log.csv" file from the local area and then click on the **Fit** button.
7. Inspect the newly created "qpr_log.csv" file to make sure it contains 200 rows (2 class measurements per histogram). The rows are arranged as follows:

   (block ID), (class ID), (estimated quantity), (error), (fit)

5.2 Plotting Measurement Distributions

The Histogram Tool provides an input feature for loading CSV data. The CSV files must be formatted correctly for this option to work. Follow the steps below to convert the log file into a suitable CSV file so the empirical distribution of measurements of class b can be inspected:

1. Using appropriate software, e.g. Libre Office, open the "qpr_log.csv" file for editing.
2. Delete all of the rows which have a class ID of 1.
3. The errors column should contain roughly the save value in each row. Make a note of what this value is.
4. Add a column which computes the estimated quantities, minus the value 105,350, which is the ground-truth measurement associated with the darker area of the example images, outside the lighter ellipse, i.e. class b pixels[1].
5. Delete every column *except* the new column (this may require the use of paste special options to avoid breaking any formulea added). By this point there should be 100 rows and only 1 column.
6. Add a column *after* the existing column and fill it with the value 1 for every row. This is required for the CSV input facility to work correctly.
7. Save the file and call it "class_b_errors.csv"
8. Using a text editor, cut and paste the content of the "header" file and paste it into the top of the "class_b_errors.csv" file. Do not leave any empty rows between the header and the data.
9. In the Histogram Tool **File** field, enter "class_b_errors.csv" and click on **Load CSV**.
10. The data should now appear in position 0. Plot the histogram to visually check the distribution and also click the **hist stats** button to view the mean and standard deviation in the tinatool window.

[1] $wh - \pi r_1 r_2 = 512 * 256 - \pi * 128 * 64 = 105350$

11. Compare the predicted error noted earlier with the actual spread of real measured quantities.

The predicted errors from the QPR Tool include a statistical and systematic component. The statistical component should manifest as a roughly Gaussian spread of values. The systematic effect should manifest as a shift of the mean away from zero. The total deviation from both effects should be no larger than the predicted error.

5.3 Questions

The error analysis performed by the QPR Tool involves several steps which will not be discussed in the tutorial. However, a good understanding of where these errors come from is important for understanding results:

1. Where does the statistical component of the error come from?
2. Where does the systematic component of the error come from?
3. How might the error checking procedure be changed so that the final error distribution has zero mean and a width much closer to the predicted width?

6 Checking Residuals

6.1 Checking Correlations

The QPR Tool will only work correctly on histograms that have independent noise between bins, i.e. the residuals between modelled frequencies and observed histogram frequencies should have no effect on each other. If correlations do exist then the error predictions will be wrong. Unfortunately, the goodness-of-fit score is not sensitive to such correlations, so a fit of unity is not an absolute guarantee of success. However, given enough data, residual correlations can be measured using Pearson's product-moment coefficient (also known as an r-score). The QPR Tool provides a means by which to compute a full correlation matrix, revealing the relationships between bins:

1. Repeat the steps of section 5.1 to fit the previously saved model to 100 example histograms, but this time the deletion and subsequent checking of the "qpr_log.csv" file can be ignored.
2. Enter the value 100 into the QPR Tool **Results** field and then click on **Info**. This should display various pieces of information in the tinatool window and display the 4 model components in histograms 100 to 103. It should also plot a 2D histogram showing residual correlations in histogram 198 and a histogram of goodness-of-fit scores in histogram 199. Histogram 199 can be viewed as normal. Histogram 198 will appear in the imcalc window when it is refreshed.
3. View the correlation matrix (histogram 198) in the imcalc window. Notice the diagonal line, with pixel values of approximately 1, and the off-diagonal terms, with approximate value 0. The off-diagonal terms around 0 indicate no correlations between residuals.
4. View the distribution of goodness-of-fit values (histogram 199). Click on the **hist stats** button to see the spread and mean. The mean value should be very close to unity.

6.2 Correlated Residuals

It is important to be able to spot when residuals are correlated. The following steps provide an example of correlated bins in histogram data:

1. Following a similar process as above, create the multi-class model then use the "corr_example" macro to create 100 example histograms. These examples should have correlated noise.
2. Fit the model to the 100 histograms, then click on the **Info** button to view the residual correlation matrix and goodness-of-fit distribution.
3. The goodness-of-fit distribution should reveal a range of fits from around 2 to 6, and the correlation matrix should show large off-diagonal terms.

6.3 Questions

The measurements produced by the QPR Tool can only be trusted if the data to which it is applied passes certain key tests. The goodness-of-fit and correlation matrix presented above are two such tests, and the corroboration of errors via repeated measurements is another. When there are poor fits, above unity, and/or there are significant off-diagonal covariance terms then quantity errors will typically be underestimated:

1. How poor is the agreement between predicted quantity errors and those observed when the "corr_example" data is used?
2. Why does smoothing the noise in the example images cause correlations within the sampled histograms?
3. The QPR Tool **Samp** field can be used to sub-sample pixel data when the **Pop** button is clicked. For example, to sample only 1 in 10 pixels, without replacement, enter the value 10 before clicking **Pop**. Can sub-sampling be used to reduce the effects of correlated residuals?

7 Summary

The QPR Tool provides a range of facilities for analysing histogram data. Whilst the sources of histograms used within this tutorial were based upon grey level distributions found within example images, many potential sources of histogram data can be used. The Linear Poisson Models, which drive the QPR Tool, can only be applied successfully if data is (or can be approximated) by a linear combination of fixed PMFs with independent Poisson bins. This tutorial has demonstrated how to build models, apply them to new data to make measurements, and to check for problems with distributions, including poor fits and correlated residuals.