Tina Memo No. 2012-003
Internal Report (second year PhD transfer report)

Quantitative Prior Estimation and Independent Component Analysis for Linear Poisson Models.

P. Tar, N.A. Thacker.

Last updated
13 / 6 / 2012

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

Quantitative Prior Estimation and Independent Component Analysis for Linear Poisson Models

Paul Tar, Neil Thacker

Abstract

This document presents a solution to the problem of determining the composition of linearly combined statistical samples (histograms) formed by multiple data generating processes. In this problem the probability distributions of individual processes are either known, or need to be estimated, and their relative proportions need to be determined. The method is based upon Bayes Theorem implemented in the form of an Expectation Maximisation (EM). Components of the theory have been developed which deal with uncertainty due to perturbation in incoming (measured) data and errors in the model. This allows the techniques to be used for the quantitative analysis of data (i.e. quantity estimates and associated measurement covariances). Background theory is presented along with Monte-Carlo tests which illustrate quantitative agreement within statistical limits.

1 Introduction

Expectation Maximisation (EM) is often used to estimate parameters for Gaussian mixture models. Principle Component Analysis (PCA) is used if orthogonal parameterisations of distributions are needed. Alternatively, Independent Component Analysis (ICA) is commonly used to separate signals in multivariate data to construct (non-orthogonal) linear models. Objective functions can be based upon non-Gaussianity of components, Mutual Information (MI), likelihood and others. In our opinion, statistical linear models of the sort described above are often used to summarise data, but rarely (even never) as an approach to quantitative measurement. The process of noise is generally treated as a negligible perturbation (broadening) of the underlying distributions. If the methods explicitly model the process of measurement an assumption of homogenous Gaussian noise is made. However, in many real situations data is generated as samples with Poisson, as opposed to Gaussian, statistics. Transform methods such as kernel methods, which might then be used to regain approximate Gaussian behaviour, will also modify the behaviour of the linear model. This would be inadvisable when the physical data generation process is known to be strictly linear.

A method of ICA applicable to Poisson noise on data points is developed here, appropriate for those situations where data is generated in the form of counted frequencies, such as histograms. A key aspect of the theory is the ability to predict

uncertainties in estimated quantities, making the approach suitable for scientific analysis tasks.

If the bin frequencies within histograms are expressed using the function $H(X|k)$, where $X \in \{0, 1, 2, \ldots\}$ represents a specific bin and $k \in \{0, 1, 2, \ldots\}$ is a label indicating a particular process, then an unknown linear combination of histograms

$$H(X) = \sum_k H(X|k)$$

may be approximated in terms of probability mass functions (PMFs) of data generating processes, $P(X|k)$, weighted by the integrals of their respective contributing histograms

$$H(X) \approx \sum_k P(X|k)Q(k)$$

where $Q(k)$ is the unknown quantity (integral) of entries in $H(X)$ associated with generation process k, i.e. $Q(k) \approx \sum_X H(X|k)$. The aim is to find a vector of quantities $\mathbf{Q} = \{Q(k = 0), Q(k = 1), \ldots\}$ which best describes the composition of $H(X)$ and also to estimate the confidence with which these quantities can be determined, i.e. estimate the associated errors on $\mathbf{Q}$. As these quantities describe the relative proportions of each constituent histogram they also provide the prior probabilities, $P(k)$, of each generation process

$$P(k) = \frac{Q(k)}{\sum_l Q(l)}$$

making the problem equivalent to the issue of how priors should be objectively selected to best describe incoming data and knowing quantitatively how accurately such priors can be determined.

A complete solution is presented here which provides a maximum likelihood estimate of $\mathbf{Q}$ and a covariance matrix, $\mathbf{C}$, giving the accuracy to within which quantities can be estimated. Solutions will be presented for increasing levels of generality, starting with an ideal case where PMFs of data generating processes are known. This will be extended to cases where PMFs have to be estimated from exemplar histograms where each data generation process can be isolated and sampled independently. The most general case will show how PMFs can be estimated when a histogram generation process is itself a linear combination of sub-processes which can not be isolated, but multiple example histograms containing different proportions of these sub-processes are available. Results using Monte-Carlo simulated data will also be presented to illustrate the validity of the method.

2 Quantity Estimation

Given a histogram composed from multiple processes the quantity of data from any individual process can be estimated iteratively using

$$\hat{Q}(k) = Q_{\infty}(k)$$

$$Q_t(k) = \sum_X P_t(k|X)H(X) \tag{1}$$

$$P_t(k|X) = \frac{P(X|k)Q_{t-1}(k)}{\sum_l P(X|l)Q_{t-1}(l)} \tag{2}$$

where $P_t(k|X)$, via Bayes Theorem, is the estimated probability at step t that process k is the source of data in bin X; and $P(X|k)$ is the PMF for process k, and quantities from previous iterations are denoted with the subscript $t-1$. Equation [1] is part of the iterative parameter update stage (Maximisation) of Expectation Maximisation (EM). Other estimation processes may be necessary here, depending upon the application, and we provide an algorithm for iterative (non-parametric) estimation of $P(X|k)$ below. Equation [2] is the expectation step.

This algorithm maximises the EM log-likelihood [1] [5]

$$\ln L = \sum_X \ln \left[\sum_k P(X|k)Q(k) \right] H(X) - \sum_k Q(k)$$

which is an extended maximum likelihood problem [2] making assumptions of independent Poisson errors on histogram bins, i.e. $\sigma^2_{H(X)} =< H(X) >$, and Poisson errors on the total normalisation $\sigma^2_{\Sigma} =< \sum_k Q(k) >$.

3 Covariance Estimation: Correct Model with Noisy Data

Likelihood methods assume the correct model is being fitted to incoming data, which in the problem domain means the PMFs, $P(X|k)$, are assumed to be correct making the only uncertainty the Poisson noise on incoming histogram bins. In cases where these PMFs can be known exactly the accuracy to within which quantities can be estimated can be determined using the Cramer Rao Bound (CRB) [3]

[1] The second term here may generally be neglected during operation of EM estimation, as it is fixed during the maximisation step. It is however, important for the valid estimation of error covariances (see below).

$$\frac{\partial^2 \ln L}{\partial Q(i)\partial Q(j)} \geq \mathbf{C}_{ij}^{-1}$$

where $\mathbf{C}_{ij}^{-1}$ is the inverse covariance between quantities $Q(i)$ and $Q(j)$. Computing the first and second order derivatives gives

$$\frac{\partial \ln L}{\partial Q(j)} = \sum_X \frac{P(X|j)H(X)}{\sum_k P(X|k)Q(k)} - 1$$

$$\frac{\partial^2 \ln L}{\partial Q(i)\partial Q(j)} = \sum_X \frac{P(X|i)P(X|j)H(X)}{[\sum_k P(X|l)Q(k)]^2}$$

then the similarity with Bayes Theorem can be used to give

$$\mathbf{C}_{ij}^{-1} \approx \frac{\sum_X P(i|X)P(j|X)H(X)}{Q(i)Q(j)} \tag{3}$$

From this the covariance matrix can be computed using a standard matrix inversion algorithm. In this ideal case the true distributions of the data generating processes must be known. If these distributions are not known then they must be estimated from training examples which introduces noise into the model. The CRB can be used with noisy models only to place a lower bound on quantity estimation accuracy due to errors on incoming data, but additional steps must be taken to account for additional variation introduced by model inaccuracies which will be explained in the next section.

4 Covariance Estimation: Noisy Model with Noisy Data

In many real world situations the PMFs of the data generating processes are not available a-priori and so must be estimated. PMFs for individual processes can be sampled from exemplar histograms assuming the data generators can be isolated

$$P(X|k) = \frac{H(X|k)}{H(X|k) + H(\bar{X}|k)}$$

where $H(X|k)$ is an exemplar histogram frequency for bin X and process k; and $H(\bar{X}|k)$ is the frequency of the sum over all bins other than X, i.e. the frequency of 'not X'. Independent errors within histograms are assumed to be Poisson allowing bin frequencies to be used directly as approximations to their variances. This applies to both exemplar histograms and incoming data

$$\sigma^2_{H(X|k)} =< H(X|k) > \approx H(X|k)$$

$$\sigma^2_{H(X)} =< H(X) > \approx H(X)$$

In cases where data generating processes can not be easily isolated for sampling purposes an Independent Component Analysis (ICA) stage is required which will be described in section 5. It is sufficient for the current section to assume $\sigma^2_{H(X|k)}$ is valid in the ICA case also, as $H(X|k)$ can be an extracted independent component with equivalent Poisson bin errors.

If the EM algorithm is seeded with the ground truth, i.e. $\mathbf{Q}$ is set to match the proportions of the data generating processes which actually generated incoming data, then noise in the model and data will cause convergence to occur some distance away from the ground truth. Beginning at the true values for the prior quantities $Q(k)$ (equation (2)), an iteration of the estimation process (equation (1)) will cause these values to change. This error will bias the next and subsequent estimates amplifying the initial effect, making the final error potentially much larger. In general we do not have the true values with which to seed the EM algorithm, but the convergence point (assuming there are no local minima) will be the same distance away from the true values irrespective of where the algorithm started. This insight supports a possible approach for estimation of the deviation. Below we approximate the deviation using a convergent geometric series, and so generate a linear approximation for the amplification process, suitable for use in error propagation.

With knowledge of the underlying sources of error, error propagation can be applied to the EM update function to estimate uncertainties in the estimated quantity vector $\mathbf{Q}$ using a two step approach.

1. Single EM Step Error: a single EM step estimate of how noise affects one instance of the update function;
2. Error Amplification: an amplification stage accounting for the accumulative effects of the iterative feedback of errors in subsequent EM steps.

4.1 Single EM Step Error

A single step in the EM algorithm can be stated in terms of incoming (Q) and model histograms (H) giving

$$Q'(k) = P(k|X)H(X) + P(k|\bar{X})H(\bar{X})$$

$$P(k|X) = \frac{\left(\frac{H(X|k)Q(k)}{H(X|k)+H(\bar{X}|k)}\right)}{\left(\frac{H(X|k)Q(k)}{H(X|k)+H(X|k)} + \frac{H(X|\bar{k})Q(\bar{k})}{H(X|k)+H(\bar{X}|k)}\right)}$$

$$P(k|\bar{X}) = \frac{\left(\frac{H(\bar{X}|k)Q(k)}{H(X|k)+H(\bar{X}|k)}\right)}{\left(\frac{H(\bar{X}|k)Q(k)}{H(X|k)+H(\bar{X}|k)} + \frac{H(\bar{X}|\bar{k})Q(\bar{k})}{H(X|\bar{k})+H(\bar{X}|\bar{k})}\right)}$$

where $Q'(k)$ is the updated quantity; $Q(k)$ is the previous quantity; and $\bar{k}$ represents all data generating processes other than k. The covariance matrix must now take account of two sources of error: noise in incoming data (previously accounted for using the CRB); and noise in the model. After a single EM step these sources combine to give

$$\mathbf{C}_{EMStep} = \mathbf{C}_{data} + \mathbf{C}_{model} \tag{4}$$

$$\mathbf{C}_{ij(data)} = \sum_X \left[\left(\frac{\partial Q(i)}{\partial H(X)}\right)\left(\frac{\partial Q(j)}{\partial H(X)}\right)\sigma^2_{H(X)}\right]$$

$$\mathbf{C}_{ij(model)} = \sum_X \left[\sum_k \left(\frac{\partial Q(i)}{\partial H(X|k)}\right)\left(\frac{\partial Q(j)}{\partial H(X|k)}\right)\sigma^2_{H(X|k)}\right] \tag{5}$$

where $\mathbf{C}_{data}$ is the equivalent to the CRB (or will be after amplification) for incoming histogram data; and $\mathbf{C}_{model}$ combines errors on each model histogram bin. The contribution from the incoming histogram data is simply

$$\mathbf{C}_{ij(data)} \sum_X P(i|X)P(j|X)H(X) \tag{6}$$

The contribution from exemplar histogram model errors involves relatively complex derivatives. The derivative of a quantity with respect to an exemplar histogram bin can be divided into two terms

$$\frac{\partial Q(j)}{\partial H(X|k)} = \frac{\partial P(j|X)H(X)}{\partial H(X|k)} + \frac{\partial P(j|\bar{X})H(\bar{X})}{\partial H(X|k)}$$

Defining the total quantity of data (integrated over all X) for class k as $Q_T(k)$, in the cases where $j = k$ these two terms are given by

$$\frac{\partial P(k|X)H(X)}{\partial H(X|k)} =$$

$$\frac{\left(\frac{H(X|k)Q(k)}{Q_T(k)} + \frac{H(X|\bar{k})Q(\bar{k})}{Q_T(\bar{k})}\right)\left(\frac{P(\bar{X}|k)Q(k)}{Q_T(k)}\right) - \left(\frac{H(X|k)Q(k)}{Q_T(k)}\right)\left(\frac{P(\bar{X}|k)Q(k)}{Q_T(k)}\right)}{\left(\frac{H(X|k)Q(k)}{Q_T(k)} + \frac{H(X|\bar{k})Q(\bar{k})}{Q_T(\bar{k})}\right)^2} H(X)$$

$$= \frac{P(X|\bar{k})Q(\bar{k})P(\bar{X}|k)Q(k)H(X)}{Q_T(k)[P(X|k)Q(k) + P(X|\bar{k})Q(\bar{k})]^2} \times \frac{P(X|k)}{P(X|k)}$$

$$= \frac{P(k|X)P(\bar{k}|X)P(\bar{X}|k)H(X)}{Q_T(k)P(X|k)}$$

and

$$\frac{\partial P(k|\bar{X})H(\bar{X})}{\partial H(X|k)} =$$

$$\frac{-\left(\frac{H(\bar{X}|k)Q(k)}{Q_T(k)} + \frac{H(\bar{X}|\bar{k})Q(\bar{k})}{Q_T(\bar{k})}\right)\left(\frac{P(\bar{X}|k)Q(k)}{Q_T(k)}\right) + \left(\frac{H(\bar{X}|k)Q(k)}{Q_T(k)}\right)\left(\frac{P(\bar{X}|k)Q(k)}{Q_T(k)}\right)}{\left(\frac{H(\bar{X}|k)Q(k)}{Q_T(k)} + \frac{H(\bar{X}|\bar{k})Q(\bar{k})}{Q_T(\bar{k})}\right)^2} H(\bar{X})$$

$$= -\frac{H(\bar{X}|\bar{k})Q(\bar{k})P(\bar{X}|k)Q(k)H(\bar{X})}{Q_T(\bar{k})Q_T(k)[P(\bar{X}|k)Q(k) + P(\bar{X}|\bar{k})Q(\bar{k})]^2}$$

$$= -\frac{P(\bar{k}|\bar{X})P(k|\bar{X})H(\bar{X})}{Q_T(k)}$$

giving

$$\frac{\partial Q(k)}{\partial H(X|k)} = \frac{P(k|X)P(\bar{k}|X)P(\bar{X}|k)H(X)}{Q_T(k)P(X|k)} - \frac{P(\bar{k}|\bar{X})P(k|\bar{X})H(\bar{X})}{Q_T(k)} \tag{7}$$

In the case where $j \neq k$ the same terms become

$$\frac{\partial P(j|X)H(X)}{\partial H(X|k)} = \frac{-\left(\frac{H(X|j)Q(j)}{Q_T(j)}\right)\left(\frac{P(\bar{X}|k)Q(k)}{Q_T(k)}\right)}{\left(\frac{H(X|k)Q(k)}{Q_T(k)} + \frac{H(X|\bar{k})Q(\bar{k})}{Q_T(k)}\right)^2} H(X)$$

$$= -\frac{P(X|j)Q(j)P(\bar{X}|k)Q(k)H(X)}{Q_T(k)[P(X|k)Q(k) + P(X|\bar{k})Q(\bar{k})]^2} \times \frac{P(X|k)}{P(X|k)}$$

$$= -\frac{P(j|X)P(k|X)P(\bar{X}|k)H(X)}{Q_T(k)P(X|k)}$$

and

$$\frac{\partial P(j|\bar{X})H(\bar{X})}{\partial H(X|k)} = \frac{P(\bar{X}|j)Q(j)P(\bar{X}|k)Q(k)H(\bar{X})}{Q(k)[P(\bar{X}|k)Q(k) + P(\bar{X}|\bar{k})Q(\bar{k})]^2}$$

$$= \frac{P(j|\bar{X})P(k|\bar{X})H(\bar{X})}{Q(k)}$$

giving

$$\frac{\partial Q(j)}{\partial H(X|k)} = \frac{P(j|\bar{X})P(k|\bar{X})H(\bar{X})}{Q(k)} - \frac{P(j|X)P(k|X)P(\bar{X}|k)H(X)}{Q_T(k)P(X|k)} \tag{8}$$

Substituting these results back into the covariance calculation provides an estimate of the error on quantities after performing a single EM step. The expected amplification of this initial deviation is modelled below.

4.2 Error Amplification

The process of error amplification generated by the iterative operation of the EM algorithm is complicated. By way of explanation, we start here by considering a single estimated quantity and then generalise to a multi dimensional system.

4.2.1 Single Parameter Quantity

If an error term, Δ_t, is introduced into the EM update function as an initial bias on one of the $Q(k)$ (equation (2)) a feedback loop develops. This can be modelled as follows for a single quantity

$$Q_t(k) = \sum_X \frac{P(X|k)[Q_{t-1}(k) + \Delta_{t-1}]}{P(X)} H(X)$$

The total accumulated error will be $\Delta = \sum_{t=0}^{\infty} \Delta_t$, where each EM step is viewed as a time series over t.

Given an initial change Δ_0 the next step's contribution can be approximated linearly using

$$\Delta_1 \approx \left(\frac{\partial Q_1(k)}{\partial \Delta}\right) \Delta_0$$

and similar approximations can be made for each subsequent step

$$\Delta_2 \approx \left(\frac{\partial Q_2(k)}{\partial \Delta}\right) \Delta_1 \approx \left(\frac{\partial Q_2(k)}{\partial \Delta}\right) \left(\frac{\partial Q_1(k)}{\partial \Delta}\right) \Delta_0$$

assuming that the derivative does not change significantly we can drop the t subscript, and this can be generalised to any step

$$\Delta_t \approx \left(\frac{\partial Q(k)}{\partial \Delta}\right)^t \Delta_0$$

The accumulated error becomes a geometric series

$$\Delta \approx \sum_{t=0}^{\infty} \left(\frac{\partial Q(k)}{\partial \Delta}\right)^t \Delta_0 = \left(1 - \frac{\partial Q(k)}{\partial \Delta}\right)^{-1} \Delta_0$$

To amplify the covariance matrix derived for a single EM step this amplification theory needs to be extended to multiple dimensions.

4.2.2 Vector Quantities

If $\mathbf{\Delta}$ is a vector of quantity errors evaluated at a particular time, t, then the accumulation of error from one step to the next is given by

$$\boldsymbol{\Delta}|_t \approx (\boldsymbol{\Delta}|_{t-1})^T \nabla \mathbf{Q}|_{t-1}$$

where $\nabla \mathbf{Q}$ is the Jacobian

$$\nabla \mathbf{Q}_{ij}|_{t-1} = \frac{\partial \mathbf{Q}_i|_{t-1}}{\partial \boldsymbol{\Delta}_j}$$

with quantities $\mathbf{Q}_i = Q(i)$. The diagonal terms of the Jacobian are given by

$$\nabla \mathbf{Q}_{ii} = \frac{\partial \mathbf{Q}_i}{\partial \boldsymbol{\Delta}_i}$$

$$= \sum_X \frac{P(X|i)[P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}}] - P(X|i)^2\mathbf{Q}_i}{[P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}}]^2} H(X)$$

where $P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}} \approx H(X)$ giving

$$= \sum_X \frac{P(X|i)H(X) - P(X|i)^2\mathbf{Q}_i}{H(X)^2} H(X)$$

$$= \sum_X P(X|i) - \frac{P(X|i)^2\mathbf{Q}_i}{H(X)}$$

which via Bayes Theorem becomes

$$\nabla \mathbf{Q}_{ii} = \sum_X P(X|i) - P(X|i)P(i|X) \tag{9}$$

Similar treatment gives the off-diagonal terms

$$\nabla \mathbf{Q}_{ij} = \frac{\partial \mathbf{Q}_i}{\partial \boldsymbol{\Delta}_j}$$

$$= \sum_X \frac{-P(X|i)\mathbf{Q}_i P(X|j)}{[P(X|j)\mathbf{Q}_j + P(X|\bar{j})\mathbf{Q}_{\bar{j}}]^2} H(X)$$

$$= \sum_X \frac{-P(X|i)\mathbf{Q}_i P(X|j)}{H(X)^2} H(X)$$

$$= \sum_X -P(X|j)\frac{P(X|i)\mathbf{Q}_i}{H(X)}$$

i.e.

$$\nabla \mathbf{Q}_{ij} = \sum_{X} -P(X|j)P(i|X) \tag{10}$$

Assuming again that the derivative computed at any time is approximately equal, i.e. $\frac{\partial \mathbf{Q}_i|_{t-1}}{\partial \mathbf{\Delta}_j} \approx \frac{\partial \mathbf{Q}_i|_t}{\partial \mathbf{\Delta}_j}$, such that for all t then $\nabla \mathbf{Q}|_t \approx \nabla \mathbf{Q}$, then the error accumulation from the first step onwards becomes

$$\mathbf{\Delta}|_1 \approx \mathbf{\Delta}|_0 \nabla \mathbf{Q}$$

$$\mathbf{\Delta}|_2 \approx \mathbf{\Delta}|_1 \nabla \mathbf{Q} \approx \mathbf{\Delta}|_0 \nabla \mathbf{Q}^2$$

$$\mathbf{\Delta}|_t \approx \mathbf{\Delta}|_0 \nabla \mathbf{Q}^t$$

The total vector change in the quantity $\mathbf{Q}$ is then given by

$$\mathbf{\Delta} \approx \sum_{t=0}^{\infty} \mathbf{\Delta}|_t = \mathbf{\Delta}|_0 + \sum_{t=1}^{\infty} \mathbf{\Delta}|_0 \nabla \mathbf{Q}^t$$

$$\mathbf{\Delta} \approx \mathbf{\Delta}^T|_0 \left[\mathbf{I} + \sum_{t=1}^{\infty} \nabla \mathbf{Q}^t \right] \;=\; \mathbf{\Delta}^T|_0 \left[\mathbf{I} - \nabla \mathbf{Q} \right]^{-1}$$

so that the total error amplification can be approximated by a single linear process

$$\mathbf{\Delta} \approx \mathbf{\Delta}^T|_0 \mathbf{A}$$

where the amplification matrix is

$$\mathbf{A} = \left[\mathbf{I} - \nabla \mathbf{Q} \right]^{-1} \tag{11}$$

4.3 Error Propagation via Amplification of the Single Step Covariance

Now, using error propagation the covariance matrix for the quantity vector can be given by scaling the one step covariance by the amplification matrix

$$\mathbf{C} \;=\; \mathbf{A}^T \mathbf{C}_{EMStep} \mathbf{A} \tag{12}$$

Whilst this formulation looks considerably different from the original CRB covariance estimate they are consistent. It will be shown numerically in section 8 that when model error terms are all zero the error propagation version described above produces equivalent results to the CRB version. Importantly, this confirms that the approach taken is an appropriate approximation of this amplification effect.

5 Independent Component Analysis: Extending the Method to Unknown Sub-Processes

The previous section assumed that data generating processes could be isolated in order to estimate their PMFs. Many real world situations involve unknown mixtures of data generating sub-processes which can not easily be isolated requiring an independent component analysis to separate signals from multiple examples. Given a source of histogram data, K, which is a set of multiple unknown generation processes, $k \in K$, that combine linearly, and given a set of independent histograms sampled from K

$$H_r(X) = \sum_{k \in K} H_r(X|k)$$

where $H_r(X)$ is the rth independent histogram, then the EM algorithm can be extended to estimate the constituent PMFs of the sub-processes. This will first be done for the case where the number of subcomponents is known, then extended later for cases where the number of subcomponents must be determined through model selection. The covariance matrix will also be extended to account for this nesting of processes.

5.1 Poisson ICA When Number of Components is Known

The EM algorithm extension begins with initial arbitrarily randomised PMF estimates for each sub-process, $P_0(X|k)$, at iteration 0. Using these random PMF starting points the most likely quantities of each sub-process is determined for each of the examples and used to give a-posteriori probabilities, $P_{r,0}(k|X)$, for each histogram. In line with the EM algorithm these probabilities are used to provide a Likelihood estimate of the total contributions from each sub-process across each example histogram r

$$R_r(X|k) \;=\; P_r(k|X)H_r(X) \tag{13}$$

In order to combine these estimates we need to know their expected distribution around the true value. This can be determined via a two step argument.

Assuming that the true value of $P(k|X)$ is known (which is true at the solution), if the quantity of data $H_r(X)$ were fixed, then the variance on $R_r(X|k)$ would follow a Binomial distribution, with each bin entry having a probability of $P(k|X)$ of being included in the sample and a probability of $1-P(k|X)$ of being excluded. This variance is given by

$$H_r(X)P(k|X)[1 - P(k|X)] = H_r(X)P(k|X) - H_r(X)P(k|X)^2$$

However, as $H_r(X)$ is not fixed, there is also an independent Poisson contribution from the originating histogram sample

$$\left(\frac{\partial R_r(X|k)}{\partial H_r(X)}\right)^2 H_r(X) = P(k|X)^2 H_r(X)$$

giving a total variation of

$$\sigma^2_{R_r(X|k)} = P(k|X)H_r(X) = R_r(X|k)$$

i.e. $R_r(X|k)$ behaves as an independent Poisson variable, allowing the individual estimates to be appropriately combined via addition into an overall weighted estimate of the underlying distribution[2]

$$H(X|k) = \sum_r R_r(X|k) \tag{14}$$

These combined histograms are normalised to provide new Likelihood estimates for each PMF at each maximisation step of the EM algorithm, until the algorithm converges

$$\begin{aligned} P_1(X|k) &= \frac{H_0(X|k)}{\sum_r \hat{Q}_{r,0}(k)} \\ P_t(X|k) &= \frac{H_{t-1}(X|k)}{\sum_r \hat{Q}_{r,t-1}(k)} \end{aligned} \tag{15}$$

The final converged results are the independent components sought

$$P(X|k) = P_\infty(X|k)$$

The accuracy to within which these PMFs can be estimated is driven predominantly by the Poisson noise within the example histograms. The variances of PMFs and their originating histograms are therefore given by

$$\sigma^2_{H(X|k)} = H(X|k) \tag{16}$$

and

$$\begin{aligned} \sigma^2_{P(X|k)} &= \left(\frac{\partial P(X|k)}{\partial H(X|k)}\right)^2 \sigma^2_{H(X|k)} \\ &= \frac{H(X|k)}{(\sum_r \hat{Q}_r(k))^2} \end{aligned} \tag{17}$$

These can now be substituted into the original covariance calculation (equation (5)) in place of their directly sampled counterparts.

[2]The conventional methods for estimating densities using EM are also consistent with this result.

5.2 ICA Model Selection

If the number of sub-processes within a set is unknown the above algorithm can still be applied. The algorithm can be executed multiple times to extract different numbers of components until a sufficient number of components has been reached satisfying the criteria that the resulting linear model provides high enough fidelity to fully describe the data whilst not over-fitting. To avoid the risk of converging into a local minimum the algorithm can be restarted multiple times from different random PMF initialisations. However, to achieve this a goodness-of-fit function is required.

Assuming normally distributed residuals, the fit between a model and data believed to be generated by that model will have an ideal Chi-squared per degree of freedom of 1. As the linear histogram models sought are Poisson distributed the residuals between model and data will not be uniform nor will they be normally distributed. A square-root transform [4] can be performed to both model and data to approximately transform the residuals into something better approximating a Gaussian with uniform width of $\sigma^2 = \frac{1}{4}$. This property of the square-root transform can be seen via error propagation

$$\begin{aligned}\sigma^2_{\sqrt{H(X)}} &= \left(\frac{\partial\sqrt{H(X)}}{\partial H(X)}\right)^2 \sigma^2_{H(X)} \\ &= \left(\frac{1}{2}H(X)^{-\frac{1}{2}}\right)^2 H(X) \\ &= \frac{1}{4}H(X)^{-1}H(X) = \frac{1}{4}\end{aligned}$$

A Chi-squared per D degrees of freedom function χ^2_D can then be defined as

$$\chi^2_{(R(N-c))} = \frac{1}{R}\sum_r \frac{4}{N-c}\sum_X \left(\sqrt{M_r(X)} - \sqrt{H_r(X)}\right)^2 \tag{18}$$

$$M_r(X) = \sum_{k\in K} P(X|k)\hat{Q}_r(k) \tag{19}$$

where N is the number of bins in the histograms; c is the number of components in the model, i.e. $c = |K|$; and $M_r(X)$ is the modelled frequency in the rth example histogram's X bin accounted for by sub-processes $k \in K$. The EM ICA algorithm can be repeated with different numbers of components until this fit function approaches as close to unity as possible, at which point the ideal number of components will have been extracted. This does however only work correctly for perfect Poisson data. In cases where the data is only approximately Poisson, or scaled Poisson then AICc [1] may be required or an equivalent technique.

6 Hierarchical Covariance Estimation

The labelling of data generating processes can be extended hierarchically to account for processes which are themselves linear combinations of other processes, such as those extracted using ICA. At the bottom level are the individual processes, k, which belong to superordinate categories of data generation processes, K. In terms of estimated quantities of data

$$Q(K) = \sum_{k \in K} Q(k)$$

The covariance terms of subclass quantities can be combined using a binary mapping matrix to give the covariance terms for the superordinate class quantities

$$\mathbf{C_{super}} = \mathbf{DCD}^T \tag{20}$$

where the mapping matrix, $\mathbf{D}$, contains a row for each superordinate class and a column for each subclass. A 1 is set in elements where the subclass belongs to the corresponding class, i.e. $\mathbf{D}_{Kk} = \delta(k \in K)$.

7 Handling Additional Errors and Filtering Inappropriate Datasets

The goodness-of-fit function can be utilised in two additional ways: if the Poisson errors on histogram bins have been underestimated, due to double counting for instance, the resulting fit can be used to scale the covariance matrix giving a more accurate estimate of the real errors in incoming data; if the fit is deemed significantly poor then it can be used to reject the incoming data entirely as being unrepresentative of the data upon which the algorithm was trained. The rescaling to additional errors can be divided into two additional parts: during ICA; or during analysis of new data.

7.1 Scaling Errors

During ICA the goodness-of-fit function assumes histograms contain simple Poisson errors, however if there are additional errors the EM ICA algorithm may converge with a fit value greater than 1. If α_{ICA} is the fit value upon convergence, the variances on PMFs can be adjusted to

$$\sigma^2_{P(X|k)} \approx \alpha_{ICA} \frac{H(X|k)}{(\sum_r \hat{Q}_r(k))^2} \tag{21}$$

which can be used within the covariance calculation to provide a more honest assessment of the expected uncertainties in $\mathbf{Q}$.

The goodness-of-fit function, as used in ICA, does not take into account errors on the model, which is appropriate during training given the model is being defined. However, if the goodness-of-fit function is to be used to assess the quality of a fit to new incoming data the squared difference between each model and data point must be normalised to the model error

$$\chi^2_{(R(n-c))} = \frac{1}{R}\sum_r \frac{1}{N-c}\sum_X \frac{\left(\sqrt{M_r(X)} - \sqrt{H_r(X)}\right)^2}{1/4 \; + \; \sigma^2_{\sqrt{M_r(X)}}} \tag{22}$$

where

$$\sigma^2_{\sqrt{M_r(X)}} = \sum_{k\in K}\left(\frac{\partial\sqrt{M_r(X)}}{\partial P(X|k)}\right)^2 \sigma^2_{P(X|k)}$$

$$= \sum_{k\in K}\frac{\hat{Q}_r(k)^2}{4M_r(X)}\sigma^2_{P(X|k)}$$

If the goodness-of-fit is still larger than 1 for incoming data, but deemed to be within acceptable levels, then the covariance matrix can be scaled by the new fit, α_{data}

$$\mathbf{C}' = \alpha_{data}\mathbf{C} \tag{23}$$

which by this point takes into consideration both additional noise during training and additional noise in new data.

7.2 Dataset rejection

For theoretically computed covariances to be of value they must ideally account for the dominant sources of error within a data analysing scenario. If the rescaling required on incoming data is larger than a factor of 2 on the variance then the theory can no longer be trusted as being a true account of the uncertainties within the data. By definition if the rescaling required is more than a factor of 2 then the main source of variation must be something other than that which has been modelled by the error theory. In this case the goodness-of-fit function can be used to reject a dataset as being unanalysable.

8 Monte-Carlo Simulation Results

Three differently shaped distributions were used to generate multiple shifted and overlapping histogram components which were mixed in different proportions to

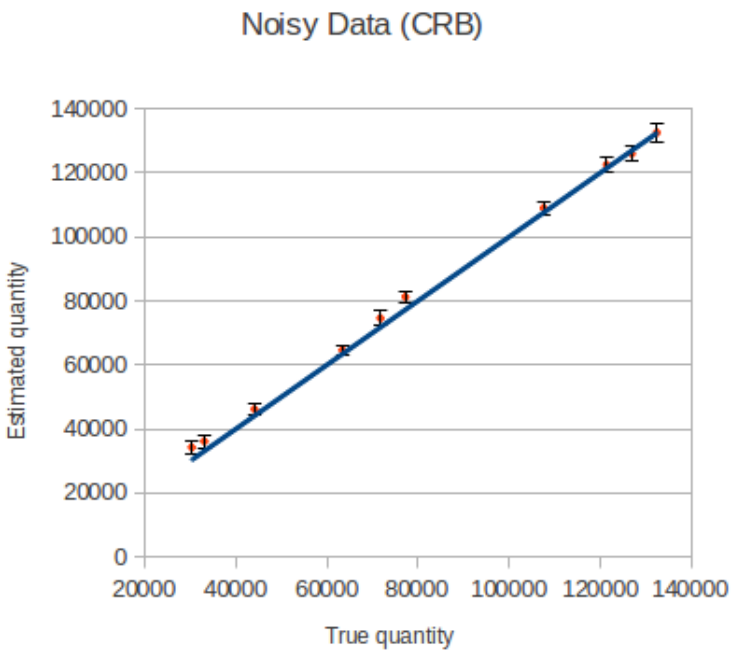

Figure 1: Example Monte-Carlo results when using true model (i.e. noise free) with noisy data. The Cramer Rao Bound was used to determine the error bars (equation(3)). Plot shows estimated quantities with 1 sigma error bars compared to true quantities.

test the EM quantity estimation, error estimation and ICA techniques. The simulated PMF distributions can be seen in figure 10.

The first test generated 1000 incoming datasets using mixtures of 9 subclasses grouped into 3 sets combined using a wide range of random quantities. Estimates of these quantities were made using:

- known PMFs, with errors computed via the CRB;
- sampled PMFs, with errors computed using error propagation and amplification;
- and PMFs extracted using the EM ICA algorithm, with error propagation and amplification

Figures 1, 2 and 3 plot the respective estimated quantities against ground truth with +/- 1 standard deviation error bars for a small number of selected trials. In this selection the error bars (computed using equations (2),(8),(9) and (10)) encompass the true values within expected statistical limits.

Figure 4 shows the average ratio of observed to predicted errors over the 1000 trials in the case where the CRB was used with known model components. Each point in this plot corresponds to a class of data generation process which are shifted versions of the 3 simulated distributions. As can be seen the ratio is close to unity, showing good agreement between theory (equation(3)) and observed errors for each process.

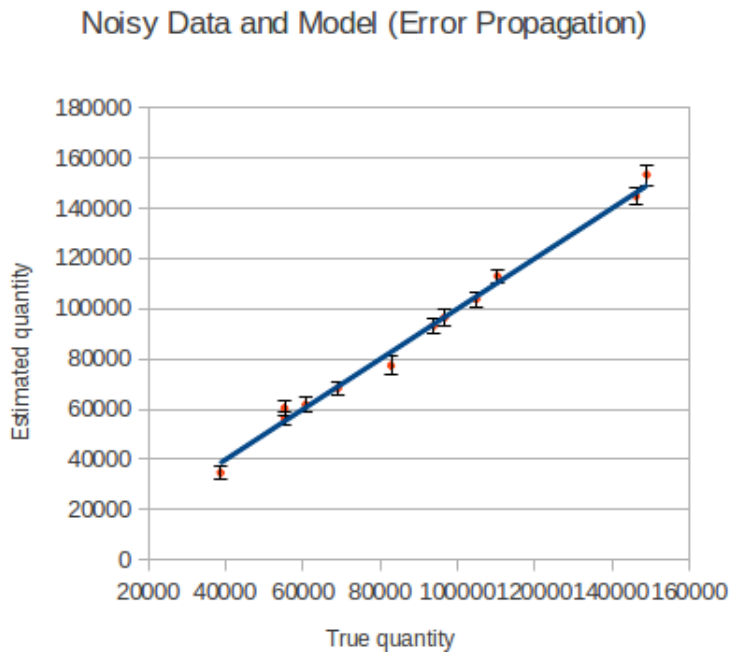

Figure 2: Example Monte-Carlo results when using model based upon sampled exemplar histograms. Error propagation on incoming data and noisy model was used to determine the error bars (equation(4)). Plot shows estimated quantities with 1 sigma error bars compared to true quantities.

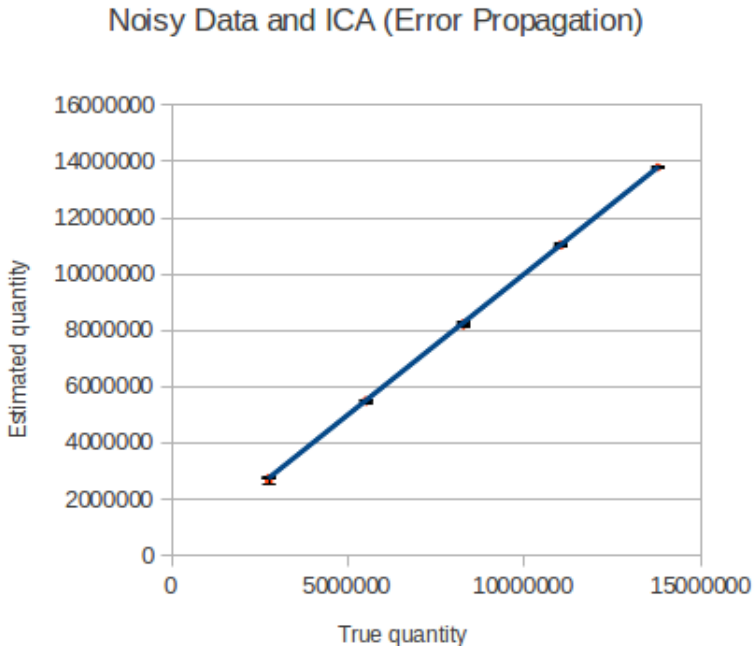

Figure 3: Example Monte-Carlo results when using model based upon Independent Component Analysis (equation(15)). Error propagation on incoming data and noisy model was used to determine the error bars (equations(16 and 17)). Plot shows estimated quantities with 1 sigma error bars compared to true quantities. Note that the large quantities tested were required to make the model noise contributions the dominant source or error.

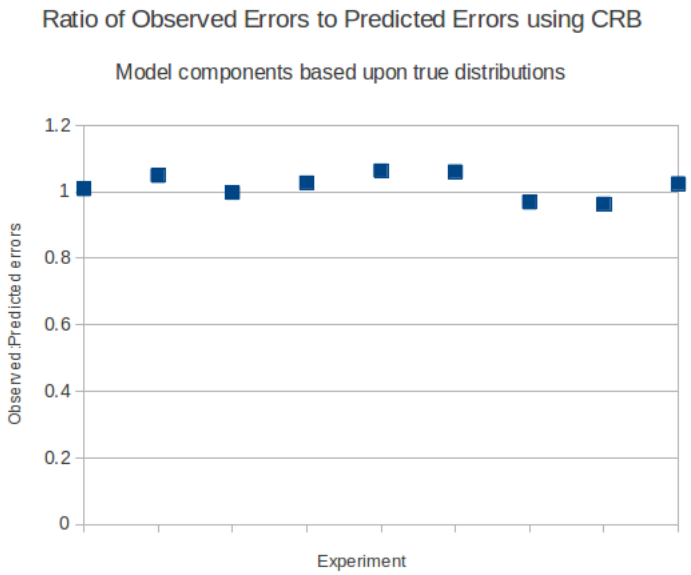

Figure 4: Ratio of observed to predicted errors when the true distributions were used as model components and the Cramer Rao Bound (equation (3)) was used to estimate the errors. Each point represents an average over 1000 trials using 3 classes generated for each of the 3 simulated distributions.

Figure 5 shows the ratio of observed to predicted errors when model components are sampled as the relative quantity of incoming data in comparison to training data increases. As the quantity of incoming data increases the relative error contribution from the data, $\mathbf{C}_{data}$, decreases as the contribution from the noisy model, $\mathbf{C}_{model}$, increases. This change in contributions between the two error components is shown in figure 6. Agreement across the entire dynamic range of samples therefore corroborates the approach taken for estimation of the two components of predicted variation (equation(4)).

The second test demonstrated the equivalence between using the CRB and the Error Propagation methods when using known PMFs, i.e. when there is only noise on the incoming histograms. Figure 7 shows the agreement between these two methods over a range of quantities.

The final test illustrates the extraction of sub-processes using ICA. Figure 11 shows the independent components found within 100 simulated histogram mixtures of shifted simulated distributions. For this test an additional set of simulated components (S0 to S2) are derived from distribution A by applying a scaling factor of 1, 2 and 3 respectively. The shape of the distributions extracted approximate those of figure 10 which were the original processes used to generate the simulated data. Differences are possible here due to the degenerate nature of the linear model, as any components which allow the prediction of the underlying linear behaviour over the full range of the data are statistically equivalent (as shown by the quantitative agreement of the theory in preceeding plots). Unique generating components are only expected if the data sample contains examples close to the

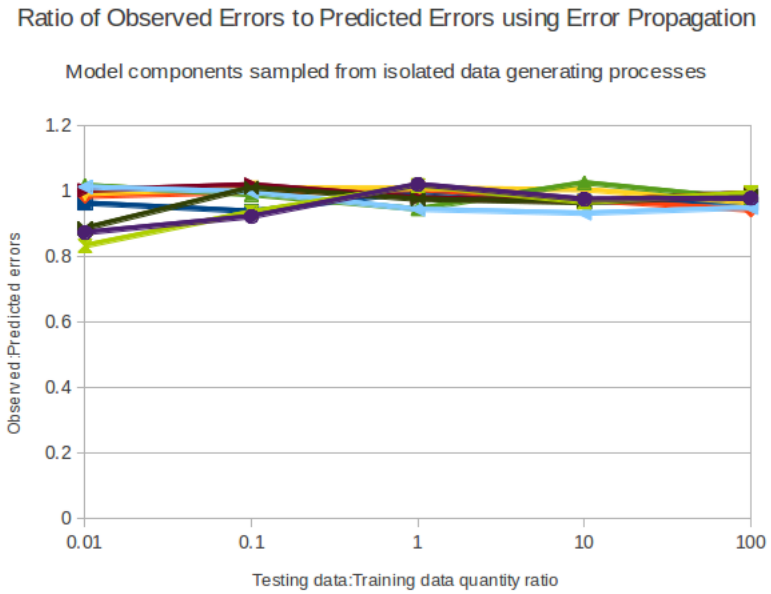

Figure 5: Ratio of observed to predicted errors when model components were sampled randomly introducing noise into the model. In this case error propagation was used to predict the errors. Each line represents a different class generated from the 3 simulated distributions. The quantity of incoming testing data relative to the quantity of training data increases along the X axis.

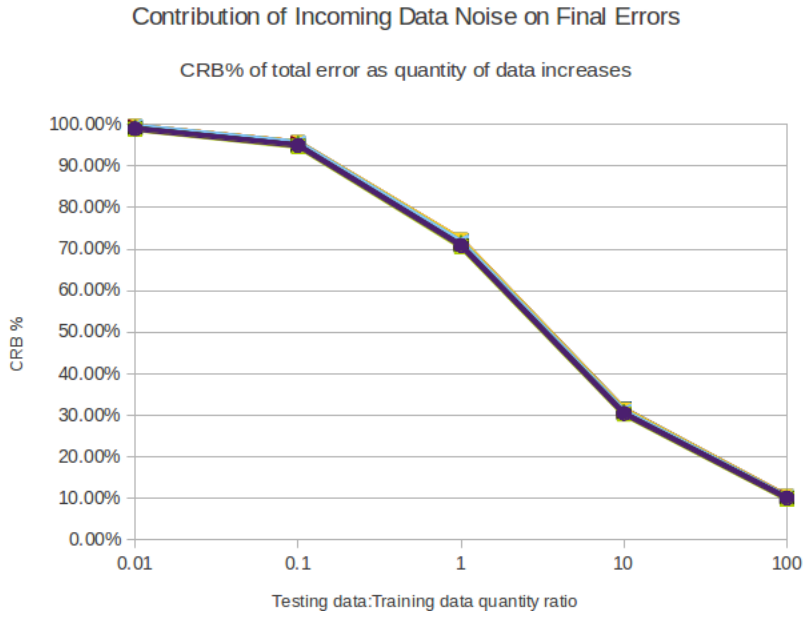

Figure 6: The CRB contribution (equation(6)) to the total error (equation(4)) when using error propagation as quantity of testing data increases.

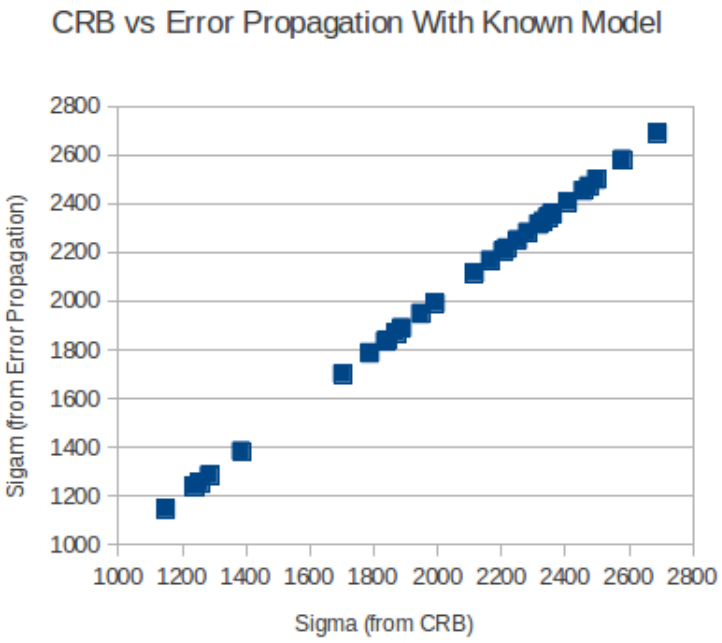

Figure 7: Numerical comparison between data on noise measured using the Cramer Rao Bound and the equivalent Error Propagation and amplification technique (equation(20)).

limits of composition (i.e. 0 and 100 %). From a quantitative perspective what is important is that the noise on model component bins is Poisson (or scaled Poisson) as is assumed in the theory (equation (14)). A chi-squared per degree of freedom can be used to confirm this which compares component residuals to known regional samples

$$\chi^2 = \sum_X^N \sum_k^K \sum_r^R \frac{(R_r(X|k) - M_r(X|k))^2}{R_r(X|k)}$$

$$\chi^2_{NK+(K-1)R+N(K-1)R} = \frac{\chi^2}{NKR - (NK + (K-1)R + N(K-1)R)}$$

with degrees of freedom from NK estimated $P(X|k)$ modelled bins; $N(K-1)R$ regional estimates of $\hat{Q}_r(k)$; $N(K-1)R$ regional estimates of $P_r(k|X)$; and where the observed Poisson count for a given bin and component is given by

$$R_r(X|k) = P_r(k|X)H_r(X)$$

and the expected (modelled) Poisson count is given by

$$M_r(X|k) = P(X|k)\hat{Q}_r(k)$$

Figure 8 shows for all non-scaled (A0 to C2) extracted components the Chi-squared is close to unity and the scaled components (S0 to S2) match their respective scaling factors of 1, 2 and 3. Confirmation that the validity of the overall

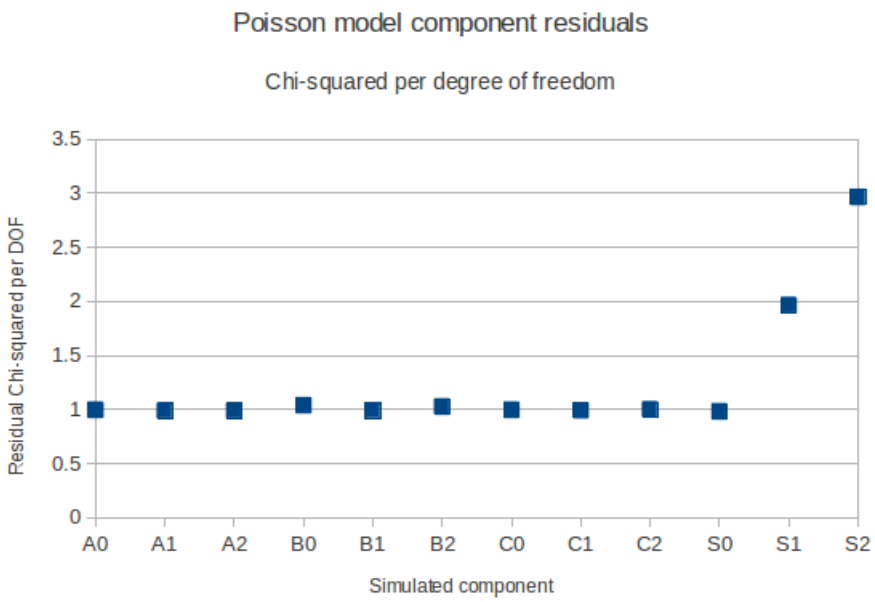

Figure 8: Confirmation of the Poisson behaviour of model component histogram bins when extracted using the EM ICA algorithm. Components A0 to C2 have non-scaled Poisson errors. The components S0 to S2 are consistent with their scaling factors of 1, 2 and 3.

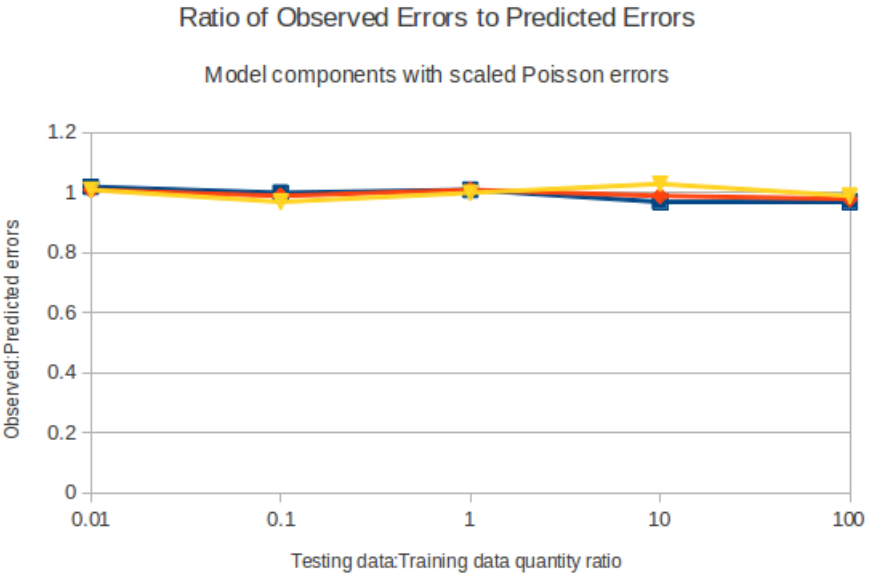

Figure 9: Ratio of observed to predicted errors using scaled components. The quantity of incoming testing data relative to the quantity of training data increases along the X axis.

error model is maintained in the presence of scaling is shown in figure 9 where the scaled components (S0 to S2) where used to estimate quantities using different ratios of training to testing data.

Figure 12 illustrates the convergence process for one of these processes, where the relative probabilities of each histogram bin can be seen to evolved over time from a random starting point. Convergence is consistent with the theory of EM estimation.

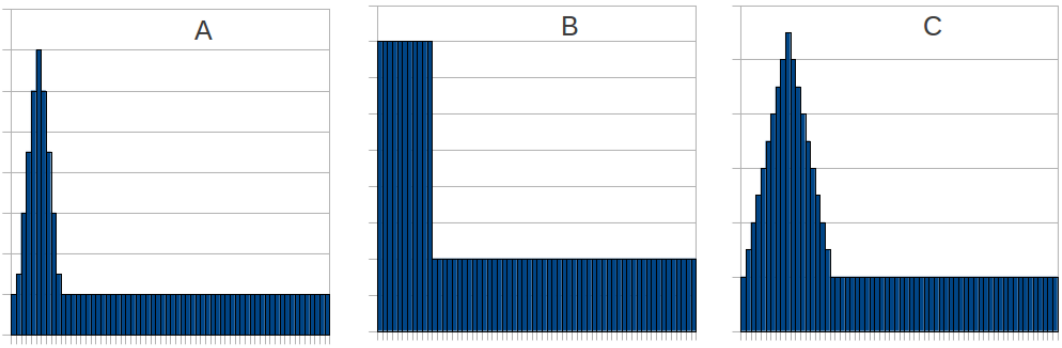

Figure 10: Simulated distributions from which random histograms were drawn within the Monte-Carlo simulation. 9 shifted versions of each distribution were used to generate 3 superordinate classes of data generating processes, each consisting of 3 sub-processes.

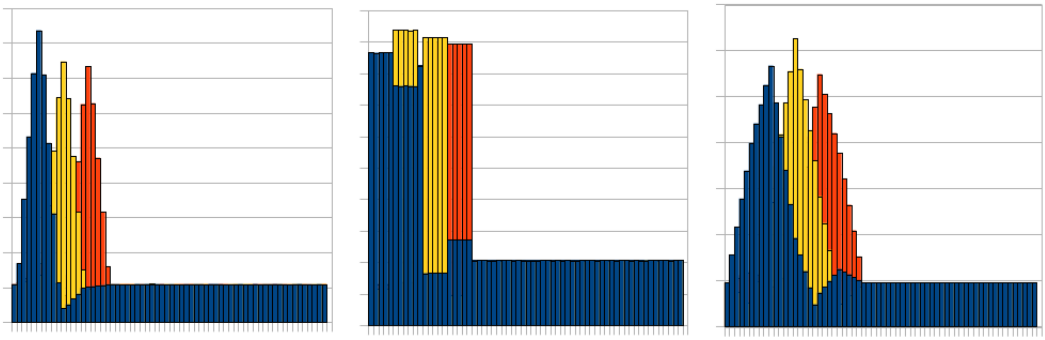

Figure 11: Components extracted using the EM ICA algorithm (equation(15)) from 100 histograms generated using combinations of shifted simulated distributions.

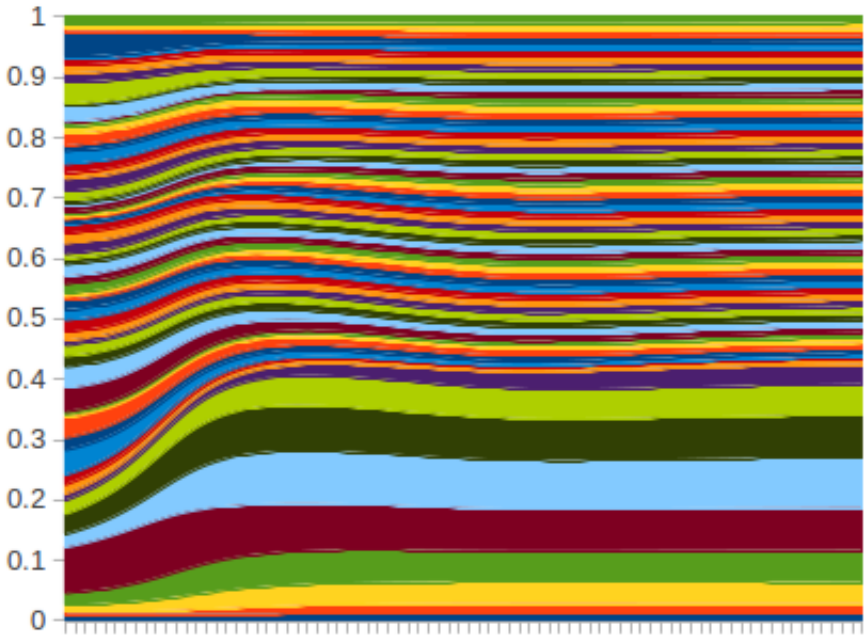

Figure 12: Illustration of the convergence of the Expectation Maximisation Independent Component Analysis algorithm. The EM steps progress along the X axis whilst the stacked bars, summing to 1 on each time step, shows the probabilities for each X bin, $P(X|k)$, for a selected sub-process. An initial random seed is seen to adopt the shape of the data generating sub-process.

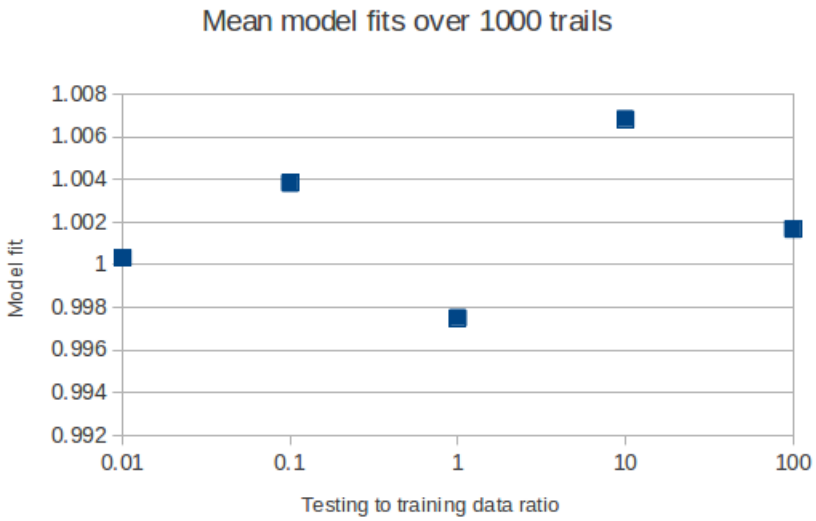

Figure 13: Mean model fits after 1000 trials of each model construction, includes known models, sampled models and ICA models.

9 Conclusions

This document has described the development and testing of a new approach for the modelling of Poisson sampled data, when in the form of linear combinations of fixed underlying distributions. Under these circumstances equations (1) and (2) can be iterated in order to estimate the underlying proportions of generator classes. A theoretical analysis of the EM algorithm has provided estimates of covariances for these estimates (equations (4-12)). We have also presented methods for the extraction of a non-parametric independent component model (equation(15)) and also for the estimation of errors when using this model for subsequent quantitative analysis of data (equations (15), (16), (17)). All key theoretical results have been confirmed via experimental testing. Our Monte-Carlo studies have shown there to be good agreement between the theoretical predictions of distributions and practical estimates. In particular we are able to estimate the accuracy of the extracted components and then to use them in conjunction with the known sample statistics of the data, in order to predict the statistical (data driven) and systematic (model driven) errors associated with overall estimation of component quantities (equation (20)).

Suggestions have been made for scaling this theory in accordance with the observed approximation error (equations(21) and (23)) for those situations where Poisson processed occur at a scale different to the intrinsic unit sample size. We believe that this analysis process is now suitable for use in data analysis applications where the underlying assumptions of a linear combination of Poisson generation of data are expected to hold.

References

[1] H.Akaike, 'A new Look at Statistical Model Identification', IEEE Trans. on Automatic Control, **19**, 716, (1974).

[2] R.J. Barlow. Statistics: A Guide to the use of Statistical Methods in the Physical Sciences. John Wiley and Sons, U.K., 1989.

[3] A. Stuart, K.Ord and S.Arnold,. Kendall's Advanced Theories of Statistics. Volume, 2A, Classical Inference and the Linear Model, Oxford University Press, 1999.

[4] N.A. Thacker and P.A. Bromiley, The Effects of a Square Root Transform on a Poisson Distributed Quantity. Tina memo, 2001-010, 2001.

[5] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction. Tina memo, 2004-005, 2004.