Tina Memo No. 2015-002
Internal.

Understanding and Improving Residual Distributions for Linear Poisson Models

P.D. Tar, N.A. Thacker

Last updated
6 / 3 / 2015

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Abstract

The method of Linear Poisson Models (LPMs) is able to construct approximating linear models of histogram data behaviour based upon the assumption of independant Poisson noise. Crucially the method has been developed with associated techniques for the assessment of the effects of noise on both model construction and subsequent use of these models in quantitative data analysis. As a consequence it is possible to apply LPM's in the form of a pattern recognition system (i.e. trained to classify data on the basis of ground truth data), while at the same time extracting estimates of associated uncertainty generally only available to more conventional regression based model fitting. Crucial to correct working of these methods is the apprpriateness of the inital assumptions of inherrent Poisson-like behaviour of the input data. This document provides statistical tests for the assessment of these assumptions as well as a methods which reduce the correlations which can arise due to poor generalisation of the trained linear models. These methods are tested using Monte-Carlo test data, in future work we intend to demonstrate the practical value of these methods in real world analyses.

1 Introduction

Linear Poisson Models [Tina Memos 2012-003, 2013-006] are 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. This Likelihood is maximised using Expectation Maximisation. Given a successful solution, error covariances are estimated for model weighting quantities, $\mathbf{Q}$, via error propagation:

$$\mathbf{C} = \mathbf{C}_{stat} + \mathbf{C}_{sys} \tag{3}$$

$$\mathbf{C}_{ij(stat)} = \sum_X \left[\left(\frac{\partial \mathbf{Q}_i}{\partial \mathbf{H}_X} \right) \left(\frac{\partial \mathbf{Q}_j}{\partial \mathbf{H}_X} \right) \sigma^2_{\mathbf{H}_X} \right] \tag{4}$$

$$\mathbf{C}_{ij(sys)} = \sum_X \left[\sum_k \left(\frac{\partial \mathbf{Q}_i}{\partial P(X|k)} \right) \left(\frac{\partial \mathbf{Q}_j}{\partial P(X|k)} \right) \sigma^2_{P(X|k)} \right] \tag{5}$$

where $\mathbf{C}_{stat}$ is a statistical error originating from incoming histogram data; $\mathbf{C}_{sys}$ is a systematic error originating from training data from which PMFs are sampled; $\sigma^2_{\mathbf{H}_X}$ is the variance of independent histogram bin X; and $\sigma^2_{P(X|k)}$ is the variance of the probability estimate $P(X|k)$. The summation over X assumes independent noise within each bin.

When a model is fitted to a new histogram, a goodness-of-fit, based upon a chi-square per degree of freedom, is used to quantify the level of agreement between model and data:

$$\chi^2_{(m-n)} = \frac{1}{m-n} \sum_X \frac{\left(\sqrt{\mathbf{M}_X} - \sqrt{\mathbf{H}_X} \right)^2}{\frac{1}{4} + \sigma^2_{\sqrt{\mathbf{M}_X}}} \tag{6}$$

$$\sigma^2_{\sqrt{\mathbf{M}_X}} = \sum_k \frac{\mathbf{Q}_k^2}{4\mathbf{M}_X} \sigma^2_{P(X|k)}$$

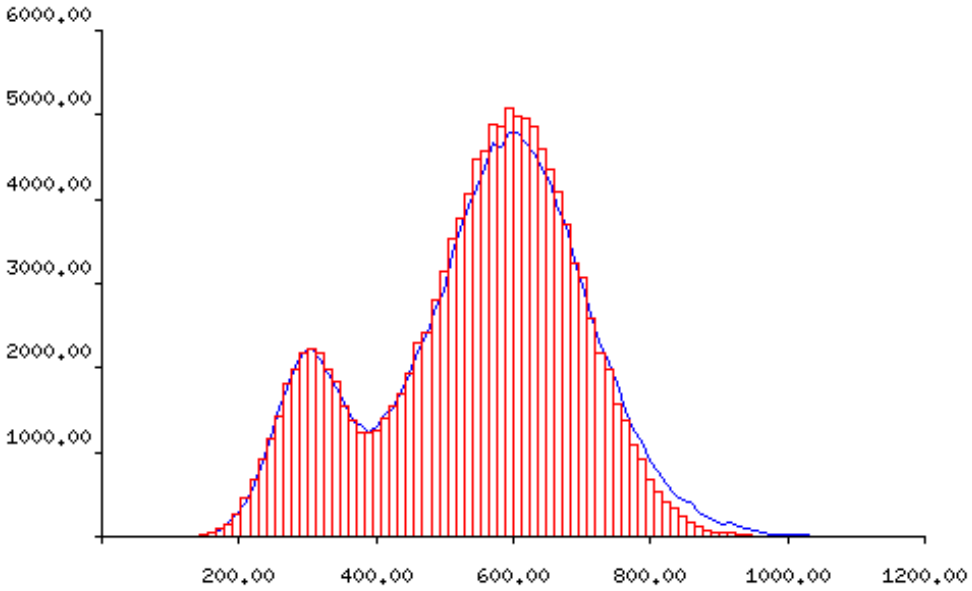

Figure 1: Correlated residuals caused by a poor model fit. The red bars represent the histogram bins, $\mathbf{H}_X$, and the blue curve represents the model, $\mathbf{M}_X$. Note that the model is systematically below the histogram data around the peak on the right, whereas the inverse is true at the tail on the right.

where m is the number of bins in the histogram and n is the number of model components, i.e. number of PMFs. The square-roots stabilise the variance of the Poisson bin frequencies, making them more Gaussian with approximate fixed widths. This goodness-of-fit should be approximately unity when a model has been successfully fitted to data.

Linear Poisson Models can fail to describe data, fail to estimate errors and fail to spot problems via the goodness-of-fit when the properties of data violate model assumptions and when approximations reach their limits. This document explores modes of failure involving poorly behaved residuals, $\mathbf{M}_X - \mathbf{H}_X$. Methods of quantifying residual problems are presented and some improvements suggested. In particular, a correlation test is developed. Monte-Carlo results will show that some problems can be fixed with minor modifications to the LPM method.

2 Measuring residual correlations

If the modelled description $\mathbf{M}$ of histogram $\mathbf{H}$ is fitted correctly and there are no correlations between noise in bins, then the model-data residuals, $\mathbf{R}_X = \mathbf{M}_X - \mathbf{H}_X$, should be independent. However, poorly fitted models or correlations in the original data can produce residuals which vary together. Figure 1 shows how groups of adjacent residuals can jointly become positive or negative due to fit failures, whereas Figure 2 shows how correlations can exist in data naturally.

The method of runs might be used to spot these correlations if they affect adjacent bins. However, if correlations exist between distant bins or the adjacency between bins is unclear, e.g. if the x-axis is categorical, then a more sophisticated check is required. Such a check can make use of a correlation matrix.

Given N independent histograms, $\mathbf{H}_{(\mathbf{1})}, \mathbf{H}_{(\mathbf{2})}, \ldots, \mathbf{H}_{(\mathbf{N})}$ and fitted models, $\mathbf{M}_{(\mathbf{1})}, \mathbf{M}_{(\mathbf{2})}, \ldots, \mathbf{M}_{(\mathbf{N})}$, a correlation coefficient can be computed allowing the relationships between each residual and each other to be inspected:

$$\rho_{XY} = \frac{1}{N} \sum_{r=1}^{N} \frac{R_{Xr} R_{Yr} + cov_{dof}(R_{Xr}, R_{Yr})}{\sigma_{Xr} \sigma_{Yr}} \tag{7}$$

where $R_{Xr} = \sqrt{\mathbf{H}_{X(r)}} - \sqrt{\mathbf{M}_{X(r)}}$ is the residual between data and model at bin X for sample r, transformed to stabilise the Poisson errors of the histogram to constant Gaussian errors; σ_{Xr} and σ_{Yr} are the standard deviations of the residual; and $cov_{dof}(R_{Xr}, R_{Yr})$ is a degree of freedom correction. The correction term is required as the estimation of quantities, $\mathbf{Q}$, during model fitting minimises residuals, removing additional variation which would otherwise be observed if the true values of $\mathbf{Q}$ were used. The degree of freedom correction reintroduces this missing variance and can be computed using error propagation:

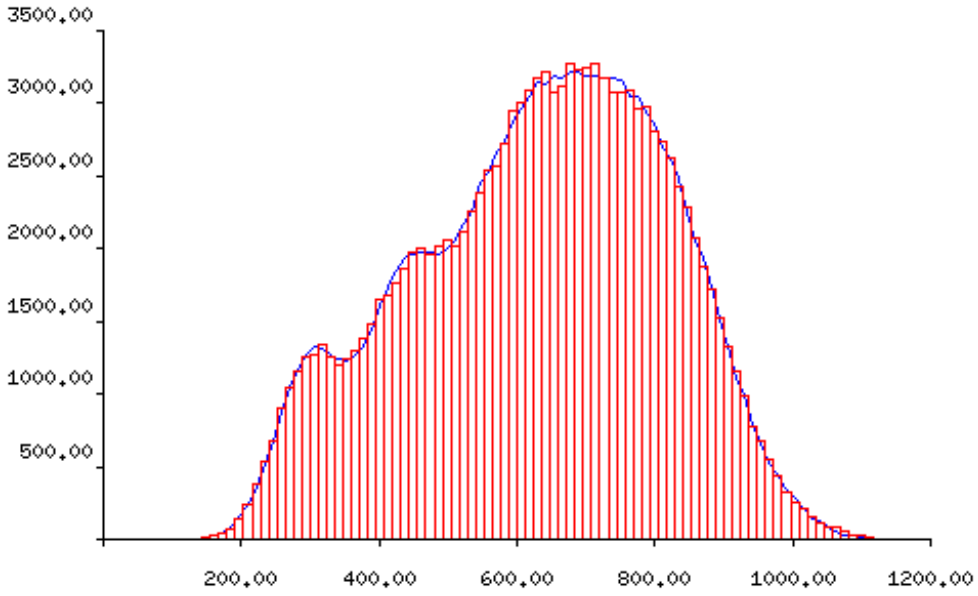

Figure 2: Correlated residuals caused by underlying histogram data. The red bars represent the histogram bins, $\mathbf{H}_X$, and the blue curve represents the model, $\mathbf{M}_X$. Note that immediately adjacent residuals are often the same sign due to localised correlations in the original histogram.

$$cov_{dof}(R_X, R_Y) = \sum_i \sum_j \left(\frac{\partial R_X}{\partial \mathbf{Q}_i}\right)\left(\frac{\partial R_Y}{\partial \mathbf{Q}_j}\right) cov(\mathbf{Q}_i, \mathbf{Q}_j) \tag{8}$$

where

$$\frac{\partial R_X}{\partial \mathbf{Q}_i} = \frac{P(X|i)}{2\sqrt{\mathbf{M}_X}} \tag{9}$$

On average these correlations should be zero if there are no offending residuals. However, due to noise, these terms will rarely be exactly zero. A hypothesis test is then required in order to use the correlation matrix to determine the statistical significance of any correlations.

The stability of the elements of ρ vary as a function of the number of samples used to estimate them and the size of any correlations. However, the distribution of ρ can be transformed into something approximately Gaussian using:

$$z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{10}$$

which is also known as Fisher's z-transform. The width of this distribution is given by:

$$\sigma_z = \frac{1}{\sqrt{N-3}} \tag{11}$$

where N is the sample size. This transformation and the inverse

$$\rho = \frac{e^{2z}-1}{e^{2z}+1} \tag{12}$$

allows confidence intervals to be computed for z and ρ.

As an alternative to computing confidence intervals for individual z or ρ values, full distributions of z scores can be inspected. If there are no correlations expected in a dataset and the correlation matrix contains many off-diagonal elements, all of which should be approximately zero, then a distribution of z scores, normalised to their error (equ 11) should produce a Gaussian of unit width. The distribution of ρ for independent residuals can be seen in figure

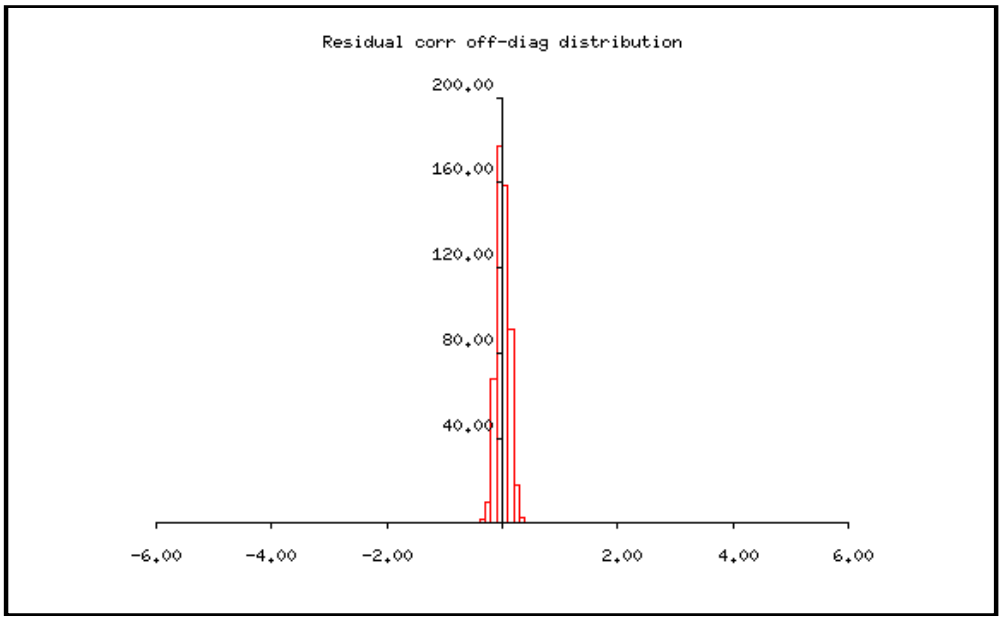

Figure 3: Distribution of independent residual ρ values. The standard deviation of this distribution is 0.11

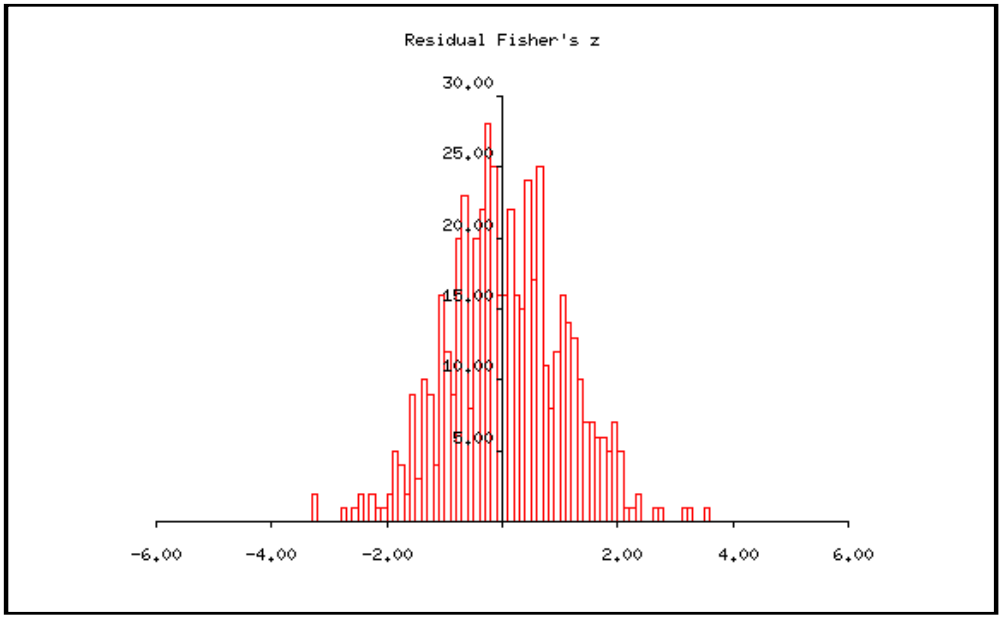

Figure 4: Distribution of independent residual z values. The standard deviation of this distribution is 1.03, showing that residuals are independent.

3 and the associated distribution of z can be seen in figure 4. These have standard deviations of 0.11 (width of ρ) and 1.03 (width of z), respectively, as expected. To illustrate the effects of correlations, figure 5 shows the z distribution for data which has been intentionally correlated by smoothing. The width of this distribution is 1.15, with a clear asymmetry showing an abundance of positively correlated bins, reaching close to 5 standard deviations along the positive x-axis.

3 Reducing residual correlations

There are several strategies for reducing correlations:

- Reduce model fit failures (e.g. fixing Figure 1) by increasing the volume of the LPM subspace which weighted PMFs can span;

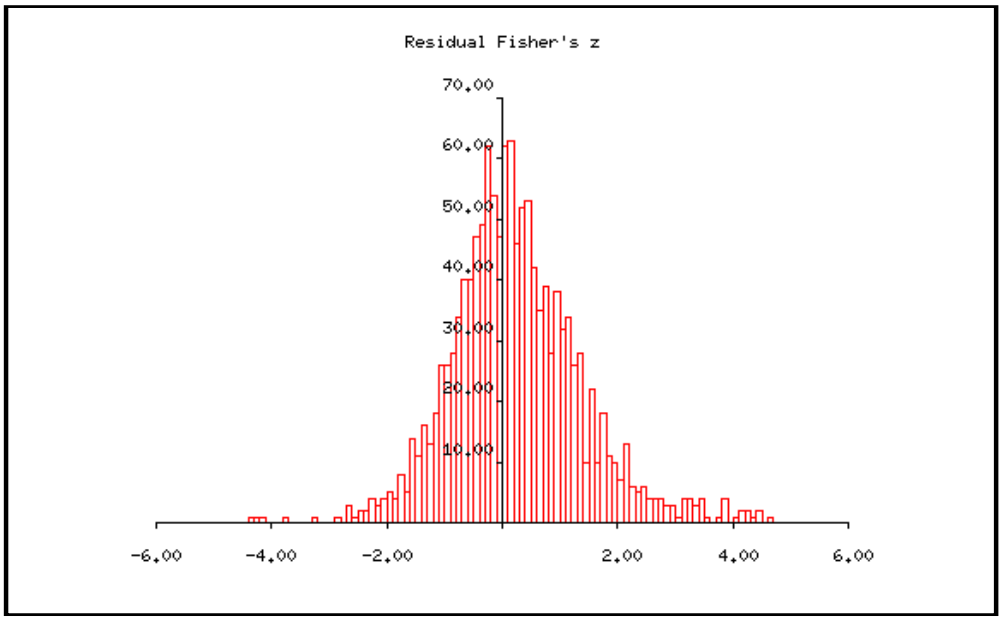

Figure 5: Distribution of correlated residual z values, where correlations are caused by smoothed data. The standard deviation of this distribution is 1.15 with skew towards positive values, indicating the presence of correlations.

- Group highly correlated residuals into common histogram bins (e.g. fixing Figure 2);
- Whiten the residual distribution

The first option of increasing the span of linear components is related to issues discussed in section 3.8 and Figure 9 of memo 2013-08 (Quantitative Planetary Image Analysis via Machine Learning). There is was noted that the training of LPMs should include examples of histograms with the most extreme compositions of components covering the full range of variability likely to exist in future data. This was to ensure situations like Figure 1 were rare. However, if such example histograms are not available inspiration can be found in the Factor Analysis literature, relating to 'rotations', for selecting alternative components. The second option, spotting then grouping correlated bins, can be applied in conjunction with the first option to remove correlations in the underlying data. Both of these options are investigated below. The final option of whitening the residuals is more difficult to achieve, but is a possible route for future investigation.

3.1 Increasing the span of PMFs

Factor Analysis aims to describe data as a linear combination of base vectors, analogously to LPMs which describe positive only histogram data as linear combinations of probabilities. Like LPMs, FA models can have many solutions. Principles such as VARIMAX rotate base vectors so that their elements are all large or near zero, with few intermediate values, such that each base vector is as different as possible from one another. These exact methods are not applicable to histogram models, due to positive only coefficients, but the ICA training algorithm and model selection criteria for LPMs can be extended to incorporate similar ideas.

To make PMFs as independent as possible, if any portion of a PMF can be used to describe any corresponding portions of others, then that corresponding portion is subtracted from the others. In this way, the overlapping regions then contain high values in one PMF and near zero value in the others, in line with the goals of VARIMAX. More precisely, during each iteration of the EM ICA algorithm, the following adjustment is applied to each component, l, with respect to possible overlaps with component k:

$$H'(X|l) = H(X|l) - \gamma H(X|k) \tag{13}$$

where γ is the largest amount of $H(X|k)$ which can be removed, for all X, such that all values in $H'(X|l)$ remain non-negative. This is performed during each EM iteration. Multiple models can be constructed in this way from different random seeds. The model which achieve a satisfactory goodness-of-fit, i.e. close to unity, and also maximise separation of components can then be selected for use. Various measures of separation can be considered, including:

- Minimising the total number of non-zero bins in each component
- Minimising the trace of the quantity error covariance: $Trace(\mathbf{C})$
- Minimising the determinant of the quantity error covariance: $Det(\mathbf{C})$

Minimising the total non-zero bins will reduce overlaps between components but will not be sensitive to the differences between near zero bins and large bins, as any finite bin values will be kept. Minimising the trace of the quantity error covariance will select components which individually can be estimated with higher accuracies, but will not account for correlated quantity errors. Minimising the determinant will minimise the total 'volume' of error in the covariance. It was explained in section 4.1 of memo 2013-08 that components which overlap the least will have the smallest quantity errors, therefore minimising errors will indirectly minimise overlap. All three alternatives should have the effect of increasing independence of components and increasing their span.

3.2 Grouping correlated bins

Even with improved models with wider linear spans it is possible that underlying data correlations can continue to cause problems. If there are still significant correlations between bins those bins can be merged and scaled. The merging has the effect of converting correlated histogram entries spread over multiple bins into double (or higher) counted entries within individual bins. The scaling then corrects for the double counting effect introduced by the multiple correlated entries into the merged bins.

4 Experiments / Results

The technique of increasing the LPM span given in section 3.1 was first tested by generating simple histograms containing two overlapping Gaussian components spread over 100 bins. 90 different mixtures of the Gaussians were produced, with fixed means and widths, so the only changes in distributions were the relative quantities of each component in each example. LPMs were created, with and without using equ 13, with models selected which minimised non-zero bins, the trace and the determinant of the quantity error covariance. Figure 6 shows the correlation between minimising non-zero bins within components and the trace of $\mathbf{C}$, and Figure 7 shows the correlation with minimising the determinant of $\mathbf{C}$.

Attempts to extract the Gaussians using the original LPM ICA algorithm can be seen in Figures 8 and 9. Here, both extracted components have non-Gaussian shapes, as the overlapping regions of the true Gaussian generators have interfered with one another. Attempts to extract the Gaussians using equation 13 can be seen in Figures 10 to 15. These show the improved determination of the true Gaussians which minimise the number of non-zero bins, trace of $\mathbf{C}$ and determinant of $\mathbf{C}$. Upon visual inspection the extracted components show less overlap using the improved method.

Additional testing checked improvements in model fits and residual correlations. LPMs were trained using mixtures of Gaussian distributions with only 20 examples and a limited range of mixing quantities. LPMs were trained using both the old method and the new method with minimised determinant. Resulting models were fitted to previously unseen mixing quantities to assess if the improved models had a greater span and generally improved behaviour. Figures 8 and 9 show poorly fitted examples before improvements to the method. Figures 19 and 20 show the same data fitted using the improved components. Figure 18 shows the distribution of z scores for the original method, showing positive skew and widths of 1.59, whereas figure 21 shows the improvement in this distribution achieved using the components with a greater span. The improved z distribution is not skewed and has a standard deviation of 0.94.

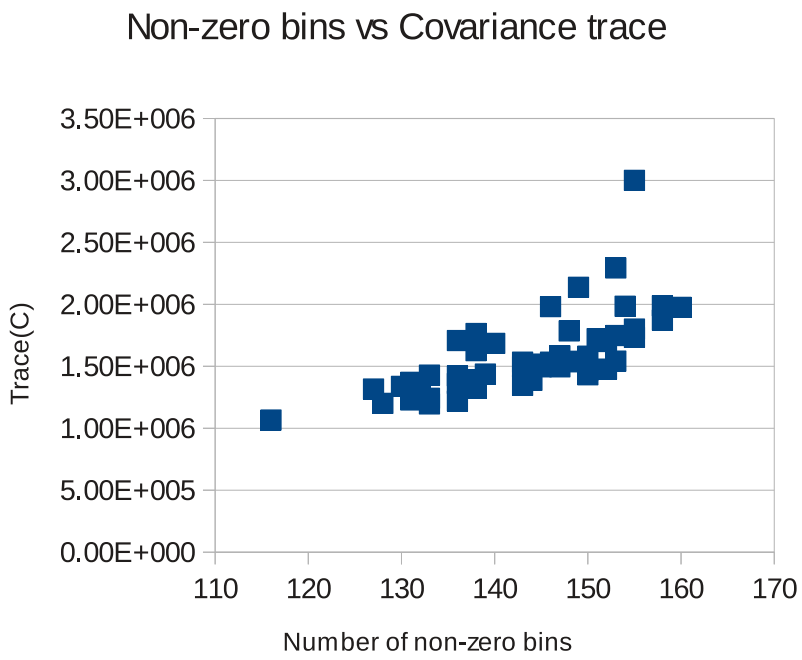

Figure 6: Relationship between number of non-zero histogram bins and the trace of $\mathbf{C}$.

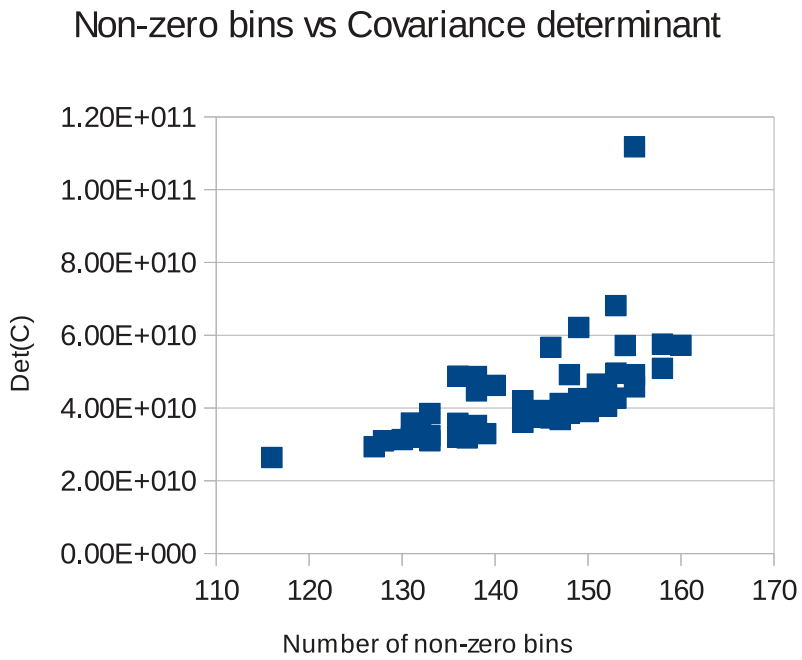

Figure 7: Relationship between number of non-zero histogram bins and the determinant of $\mathbf{C}$.

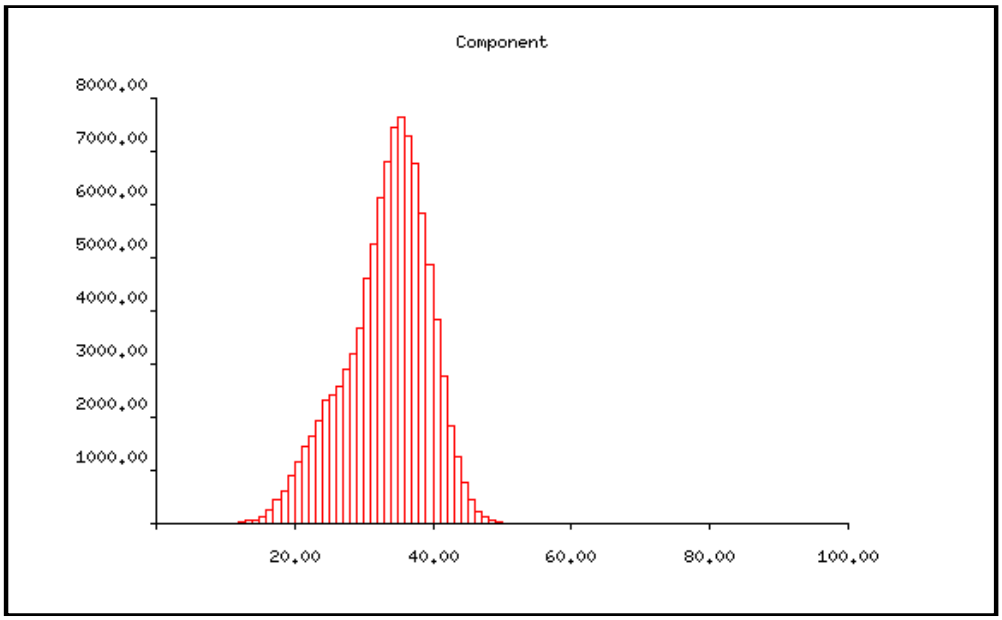

Figure 8: First extracted component using LPM ICA without equation 13, showing significant deviation away from the Gaussian generator of the data.

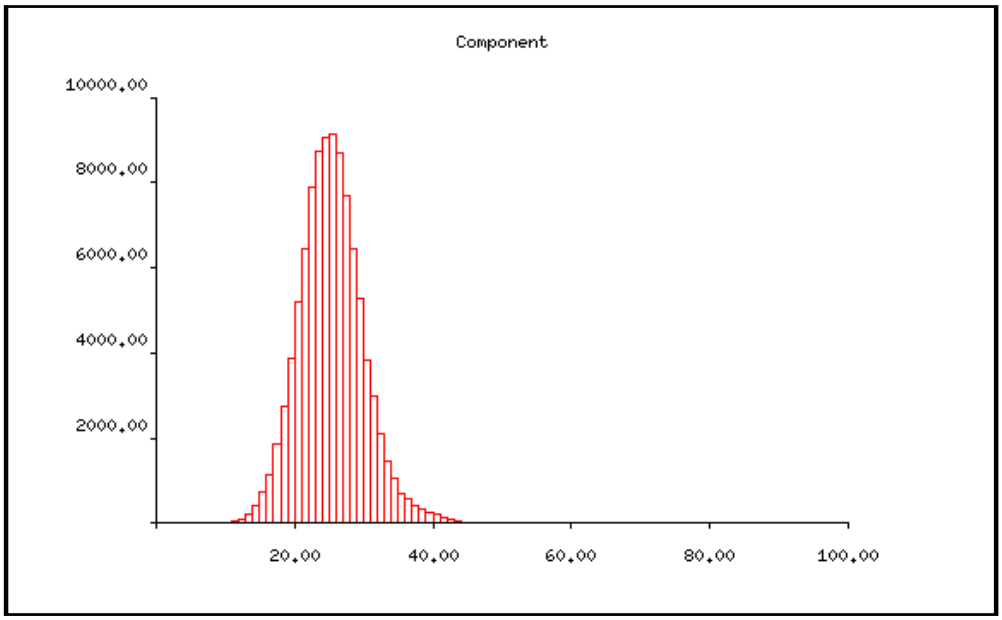

Figure 9: Second extracted component using LPM ICA without equation 13, showing deviation away from the Gaussian generator of the data with a slight positive skew.

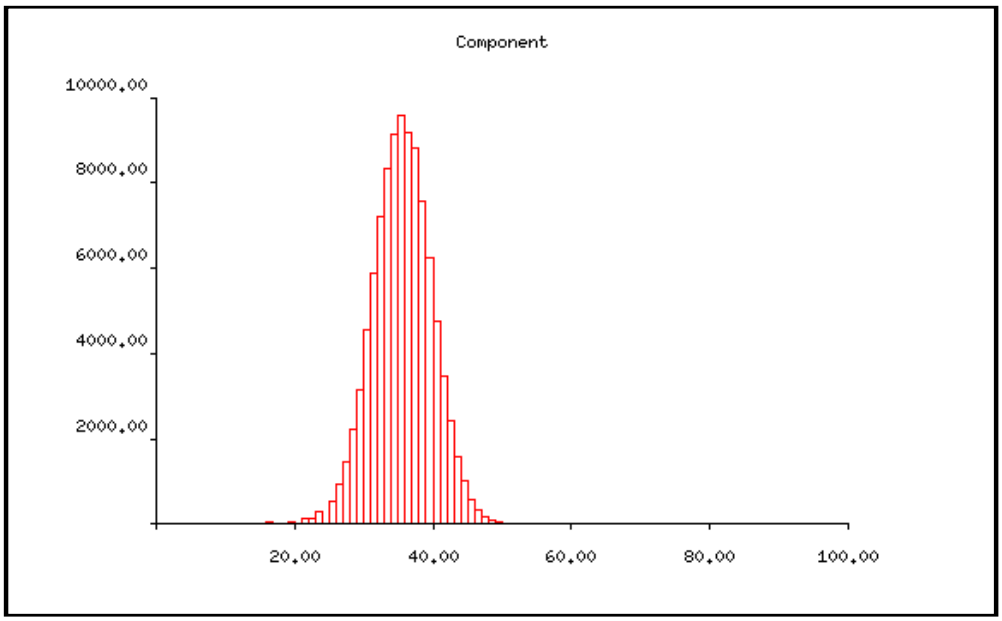

Figure 10: First extracted component using LPM ICA with equation 13, selected to minimise number of non-zero bins.

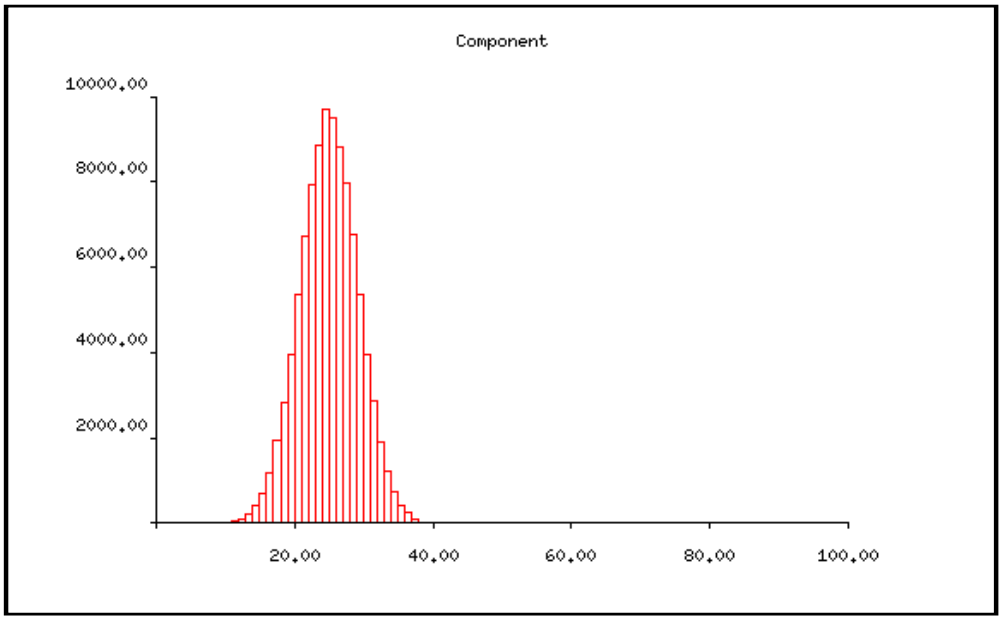

Figure 11: Second extracted component using LPM ICA with equation 13, selected to minimise number of non-zero bins.

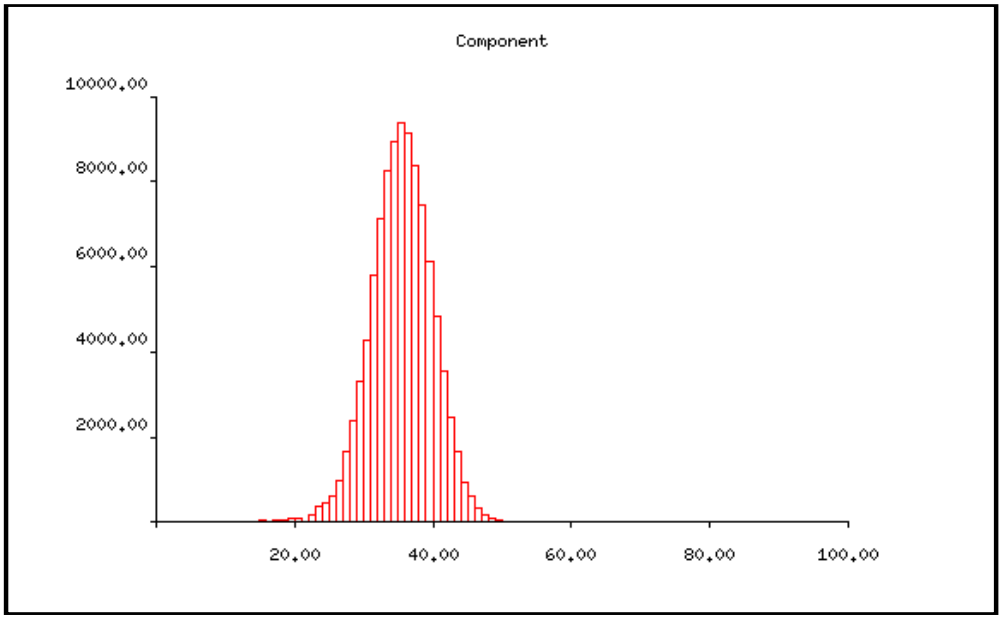

Figure 12: First extracted component using LPM ICA with equation 13, selected to minimise the trace of **C**.

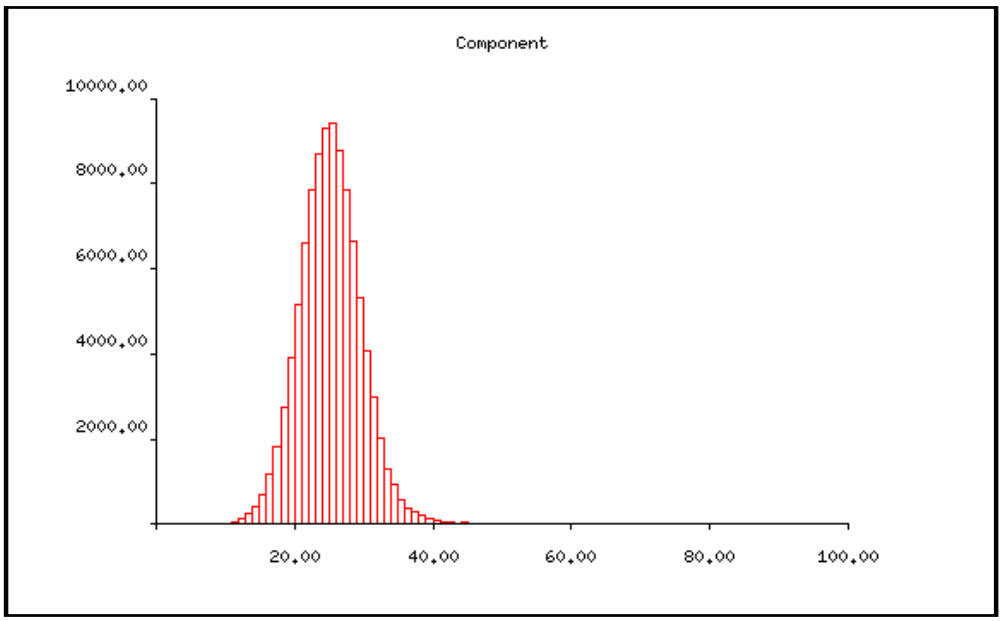

Figure 13: Second extracted component using LPM ICA with equation 13, selected to minimise the trace of **C**.

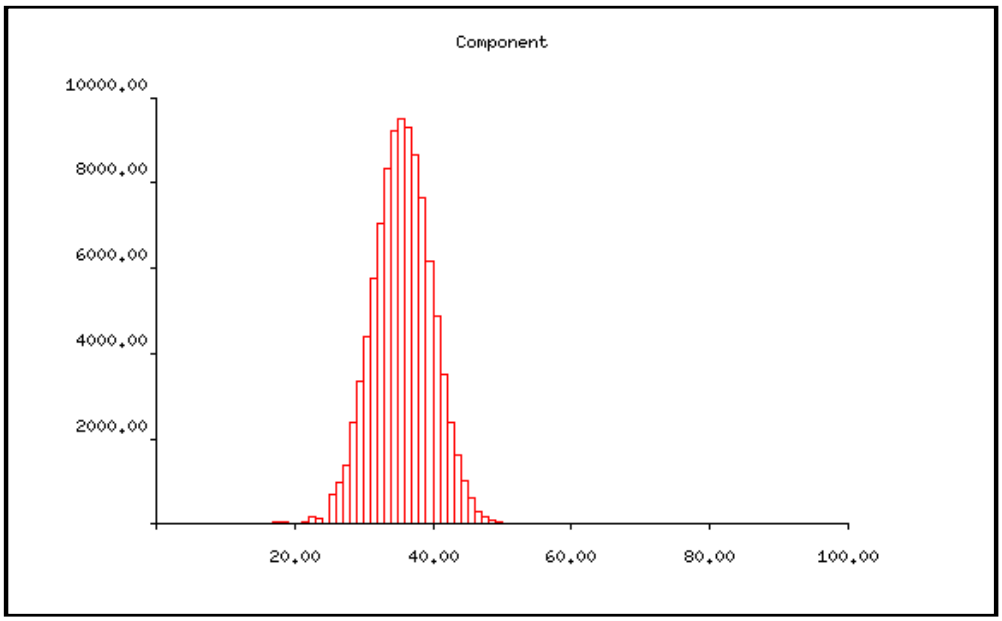

Figure 14: First extracted component using LPM ICA with equation 13, selected to minimise the determinant of **C**.

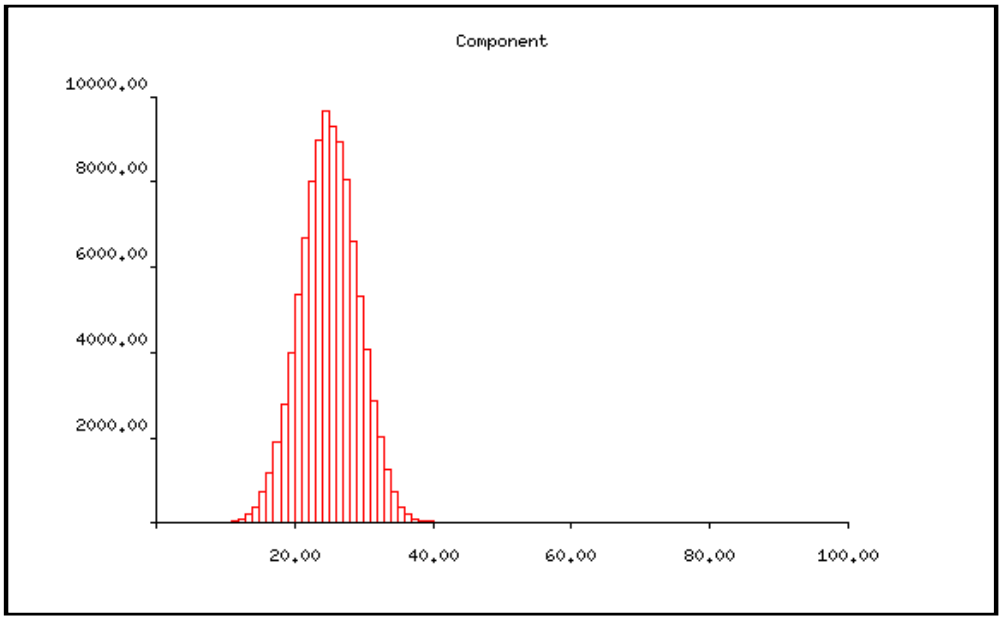

Figure 15: Second extracted component using LPM ICA with equation 13, selected to minimise the determinant of **C**.

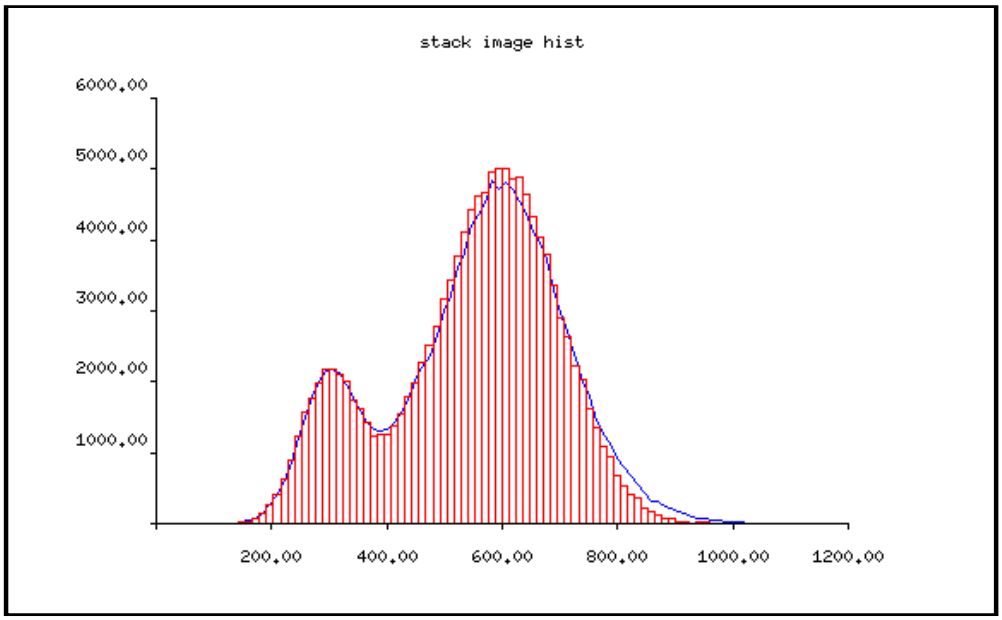

Figure 16: Extracted components fitted to new data, where LPM was trained using old method, i.e. without equation 13

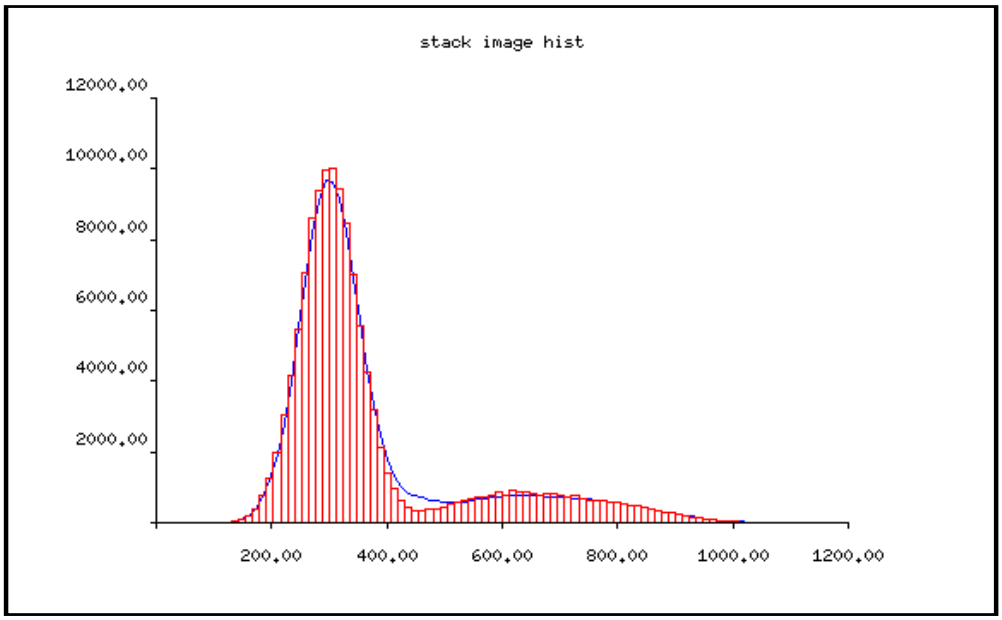

Figure 17: Extracted components fitted to new data, where LPM was trained using old method, i.e. without equation 13

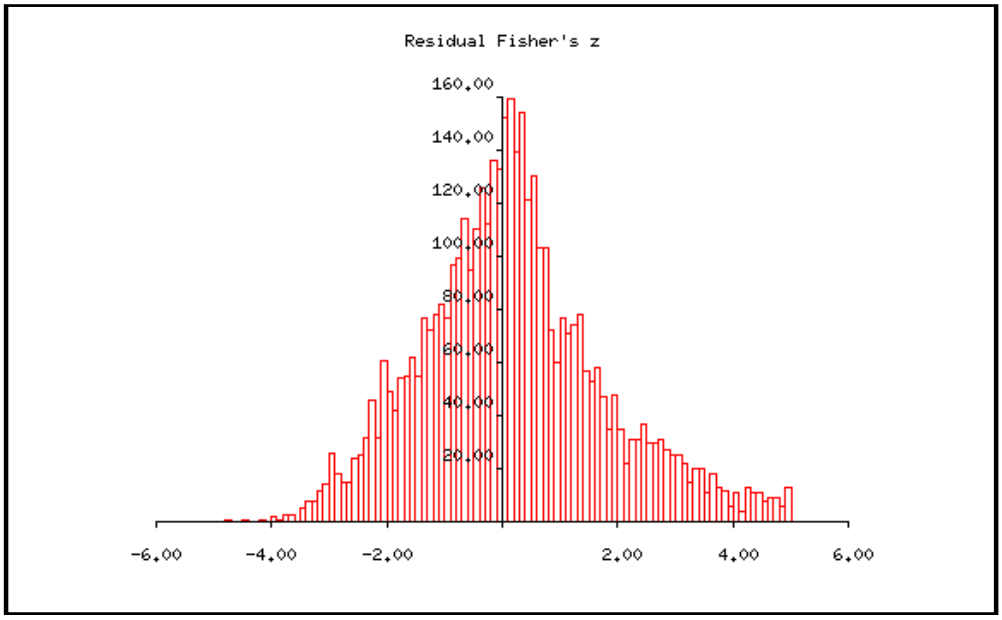

Figure 18: Distribution of z for poorly fitted models. The standard deviation of this z distribution is 1.59

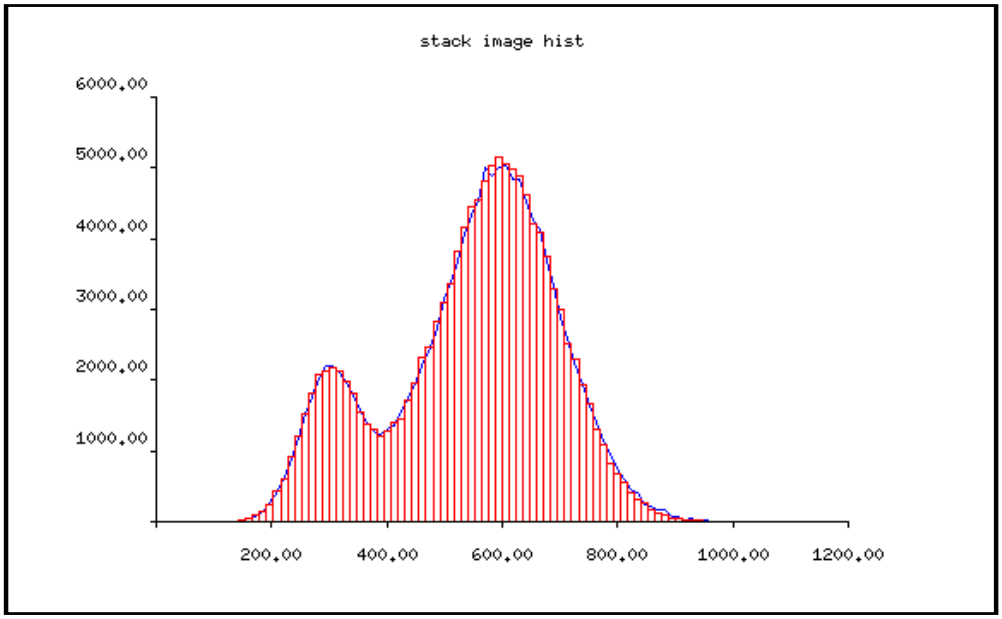

Figure 19: Extracted components fitted to new data, where LPM was trained using new method, i.e. with equation 13 and minimising determinant of $\mathbf{C}$

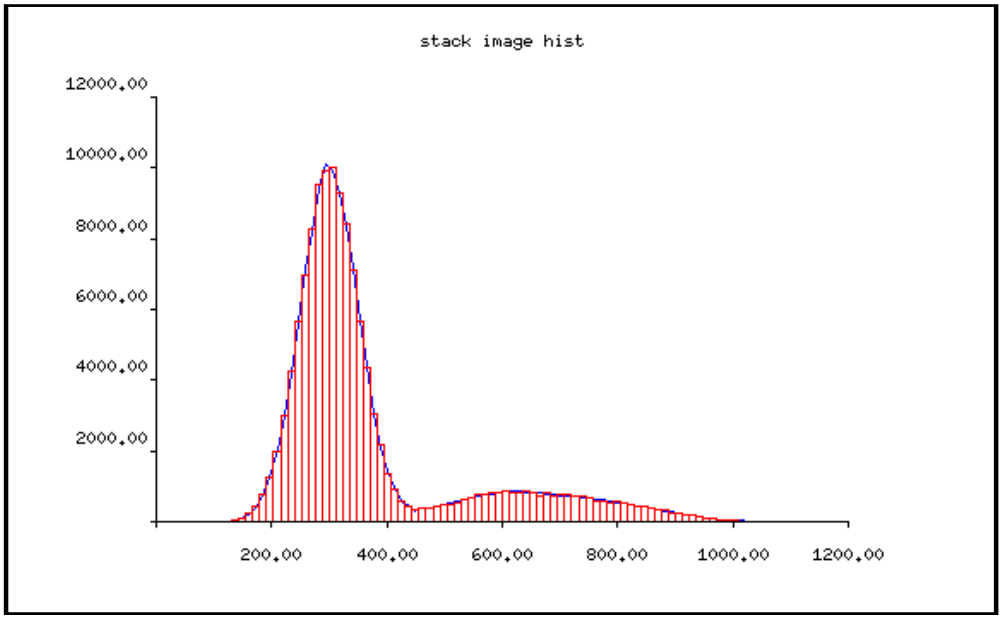

Figure 20: Extracted components fitted to new data, where LPM was trained using new method, i.e. with equation 13 and minimising determinant of $\mathbf{C}$

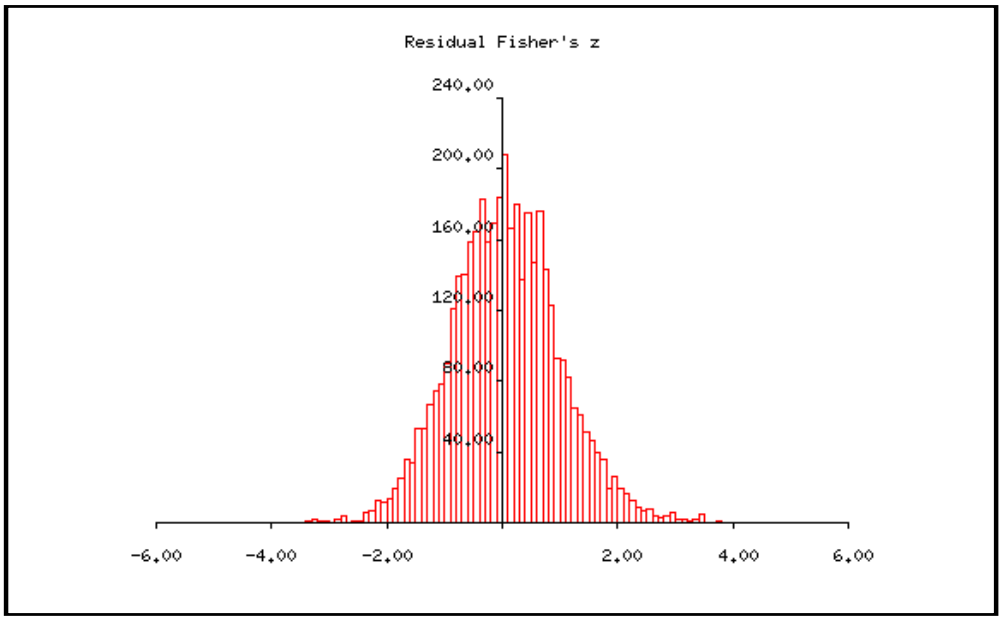

Figure 21: Distribution of z for improved fitting models. The standard deviation of this z distribution is 0.94