Tina Memo No. 2018-006
Internal.

The Stability of Probability Mass Function Estimation for Linear Poisson Modelling.

Paul D. Tar, Neil A. Thacker.

Last updated
25 / 6 / 2018

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

The Stability of Probability Mass Function Estimation for Linear Poisson Modelling

P.D. Tar and N.A.Thacker

1 Introduction

Linear Poisson Modelling (LPM) has successfully been applied for the estimation of total quantities associated with classes of linearly composed histograms. These estimated quantities, (i.e. $Q_c = \sum_{k \in c} Q(k)$), are accompanied by error covariances that account for statistical and systematic sources of uncertainty, ultimately stemming from the Poisson sampling processes from which the histograms were populated.

The modelling process also produces a set of Probability Mass Functions (PMFs), $P(X|k)$, that thus far have only been empirically assessed regarding their stability. This is for good reason as, unlike the total Qs, the potential for linear degeneracies makes the unique estimation of $P(X|k)$ and $Q(k)$ impossible using the Likelihood function alone. However, our estimation process involves several additional sources of information, such as positive only $P(X|k)$, and the MAX SEP method, which together can help to regain a unique linear model. In the case of mass spectra the use of MAX SEP is expected to generate ICA components which are close to the true underlying data generators.

This document applies the minimum variance bound to the LPM extended maximum likelihood function to estimate PMF bin errors, assuming that the linear degeneracies have been avoided. These errors are important if the bins are to be interpreted as physical measurements, e.g. if they represent mass spectra peaks. The MVB error prediction are validated using Monte Carlo data.

2 MVB estimation of bin uncertainties

Histograms are typically accumulated by counting discrete Poisson events which occur with some expected mean frequency per bin. The statistical modelling of this scenario is usually attributed to Fermi in the form of Extended Maximum Likelihood, which in terms of PMFs and quantities becomes:

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

where $P(X|k)$ (i.e. $\mathbf{P}$) and $\mathbf{Q}$ are parameters to be estimated. These can co-vary producing potential linear degeneracies in the model. In particular linear combinations of ICA components can often be used to descibe the same data subject to selection of an appropriate set of Q_ks. Under these circumstances the expected errors on $P(X|k)$s could be considered infinite.

However, it can be observed that there are some circumstances where this can be avoided. Data which can only be explained by extremes of the possible linear combination force a unique interpretation of (positive only) ICA components. This will occur whenever a data histogram has zero entries for one value of X which is non-zero in all but one of the extracted ICA components. This observation led previosuly to the MAX SEP algorithm. As the name suggests, the algorthm seeks to maximise the linear separation of ICA vectors by subtracting as much of each ICA component from all the rest whilst keeping all $P(X)$ positive. As we remain on the original ICA linear manifold, any Likelihood function is completely unaffected. However, the resulting linear model has greater volume of applicability, so it is more likely to be able to describe new data. Also this regains a unique interpretation for $P(X)$, so we can assume that this additional information generates a linear model in which the resulting corrrelations with Q_h play only a minor role. In this case the contribution to uncertainties in $P(X|k)$s from variations in $Q(k)$ can be neglected. This is mathematically the same as the $Q(k)$s being eliminated as free parameters from the model covariance.

In order to separate useful terms, the likelihood function can be stated in terms of two components, a and b, and all other components, k, that are not a or b, i.e. $k \neq a, b$. Also, by approximating with a less efficient function L', for histogram bin X and all other bins that are not X, i.e. $\bar{X}$: we can simplify the Likelihood expression. The function can be separated into,

$$\ln L' = \sum_h \ln \left[P_{Xa} Q_{ah} + P_{Xb} Q_{bh} + \sum_{k \neq a,b} P_{Xk} Q_{kh} \right] H_{Xh}$$

$$+\sum_h \ln\left[(1-P_{Xa})Q_{ah}+(1-P_{Xb})Q_{bh}+\sum_{k\neq a,b}(1-P_{Xk})Q_{kh}\right]H_{\bar{X}h}$$

$$-\sum_h\left[Q_{ah}+Q_{bh}+\sum_{k\neq a,b}Q_{kh}\right]$$

where the sum over h is a sum over histograms within the training cohort; P_{Xa} is the probability of an event in bin X given the source was component a; Q_{ah} is the quantity of data associated with component a in histogram h; and H_{Xh} is the frequency recorded in bin X in histogram h.

The inverse error covariance between a PMF bin one component and its counterpart in another component is then approximately given by considering only the probabilities in bin X, i.e. all possible a and b combinations across the model vector

$$C_{ab}^{-1}\approx\frac{\partial^2-\ln L'}{\partial P_{Xa}\partial P_{Xb}}$$

The first derivative is

$$\frac{\partial-\ln L'}{\partial P_{Xa}}=\sum_h\frac{-Q_{ah}H_{Xh}}{P_{Xa}Q_{ah}+P_{Xb}Q_{bh}+\sum_{k\neq a,b}P_{Xk}Q_{kh}}$$

$$+\frac{Q_{ah}H_{\bar{X}h}}{(1-P_{Xa})Q_{ah}+(1-P_{Xb})Q_{bh}+\sum_{k\neq a,b}(1-P_{Xk})Q_{kh}}$$

The second derivative taken w.r.t. another component, b, for the same bin, X is then

$$\frac{\partial^2-\ln L'}{\partial P_{Xa}\partial P_{Xb}}=\sum_h\frac{Q_{ah}Q_{bh}H_{Xh}}{[P_{Xa}Q_{ah}+P_{Xb}Q_{bh}+\sum_{k\neq a,b}P_{Xk}Q_{kh}]^2}$$

$$-\frac{Q_{ah}Q_{bh}H_{\bar{X}h}}{[(1-P_{Xa})Q_{ah}+(1-P_{Xb})Q_{bh}+\sum_{k\neq a,b}(1-P_{Xk})Q_{kh}]^2}$$

As the denominators sums to the squared LPM model, M_X, and the modelled value is approximately equal to the histogram value, H_X, this can be stated as

$$=\sum_h\frac{Q_{ah}Q_{bh}}{M_{Xh}^2}H_{Xh}-\frac{Q_{ah}Q_{bh}}{M_{\bar{X}h}^2}H_{\bar{X}h}$$

$$\approx\sum_h\frac{Q_{ah}Q_{bh}}{H_{Xh}}-\frac{Q_{ah}Q_{bh}}{H_{\bar{X}h}}$$

This can also be rewriten in terms of Bayes Theorem, which can then be interpreted in terms of the probability of component membership

$$=\sum_h\frac{P_h(a|X)P_h(b|X)H_{Xh}}{P_{Xa}P_{Xb}}-\frac{P_h(a|\bar{X})P_h(b|\bar{X})H_{\bar{X}h}}{P_{\bar{X}a}P_{\bar{X}b}}$$

3 Monte Carlo validation

To test the error predictions, a simple 3 component model was repeatedly estimated using Monte Carlo datasets with known generators. The true PMFs of the model can be seen in Figures 1 to 3, consisting of an upward ramp, a downward ramp and a top hat. The opposing ramps allow histogram bins to each be constructed from a wide range of relative component contributions. The top hat, padded with zeros at the start and end of the PMF, allowed the central bins to be more ambiguous that those at the edges where only data from the ramps would otherwise be observed.

10 datasets were created with independent Poisson noise. To minimize the chances of variation due to linear degeneracy, the LPM was seeded with the 'true' answers then allowed to drift away during convergence under the effect of the sampling noise. The difference between each PMF bin and its estimate was normalised to its MVB error to form a pull distribution:

$$\delta = \frac{< P(X|k) > - P(X|k)}{\sqrt{C_{kk}}}$$

Scatter plots of 'true' to estimated PMFs can be seen in Figures 4 to 5. The mean and standard deivations of the pull distributions can be seen in Figure 7.

4 Discussion

Despite the approximation of the Likelihood function L with L', the functional form of the error predictions have theoretical properties consistent with the estimation uncertainties of PMF bins. Firstly, there are ambiguity terms $P(a|X)P(b|X)$, that are largest when components are equally probable (0.5 vs 0.5), and zero when a bin only exists in one component but not the other (1.0 vs 0.0). Secondly, the errors scale with the total sample size available per bin, i.e. the H_{Xh} terms.

The Monte Carlo simulated data shows that error predictions are consistent with the observed repeatability of PMF estimates. Figure 7 shows that the mean and width of the pull distribution of δ are consistent with being zero and one, respectively, as expected. However, Figure 4 and Figure 5 show evidence of some local solutions being found that are significantly away from the 'true' answers. This is despite the models being seeded with the true PMF shapes. This lack of stability is almost certainly due to the potential linear degeneracies in the model. The problem of such local solutions is already known. Here, as the Likelihood function contains no information with which to drive a specific degenerate solution, the parameters generally stayed close to their original values.

The degenerate nature of LPM solutions can be mitigated using MAX SEP, which was not applied in these cases. Despite the local solutions, there is reason to believe that the error predictions (based upon the assumption that the Qs are no longer free parameters), are still representative of the shape of the cost function at these local solutions. The global solutions follow predictions well, including the wider spread of estimates around the top hat region, which is the most ambiguous.

In the general case, the errors on ICA components $P(X|k)$, defined as the differences between estimated values and true underlying generators of data, will be difficult to predict. However, when either re-estimating the ICA components from a true starting point, or constraining them via use of MAX SEP, we would expect the variations in $P(X|k)$ to contain a contribution from the uncertainty implied by the Likelihood function. This suggests that the associated MVB estimates are at least a lower bound on ICA parameter stability, if not a direct estimate of the error from ground truth.

5 Summary

Knowledge of how precisely PMFs can be estimated in theory provides an additional tool towards a solution to the local minima problem. This analytical approach can be compared to randomly seeded models to separate uncertainty contributions from sampling noise and those from poorly selected models. Current work on mass spectra, where MAX SEP appears to perform well, can immediately benefit from these new error predictions, as they allow more relieable peak ratios to be identified.

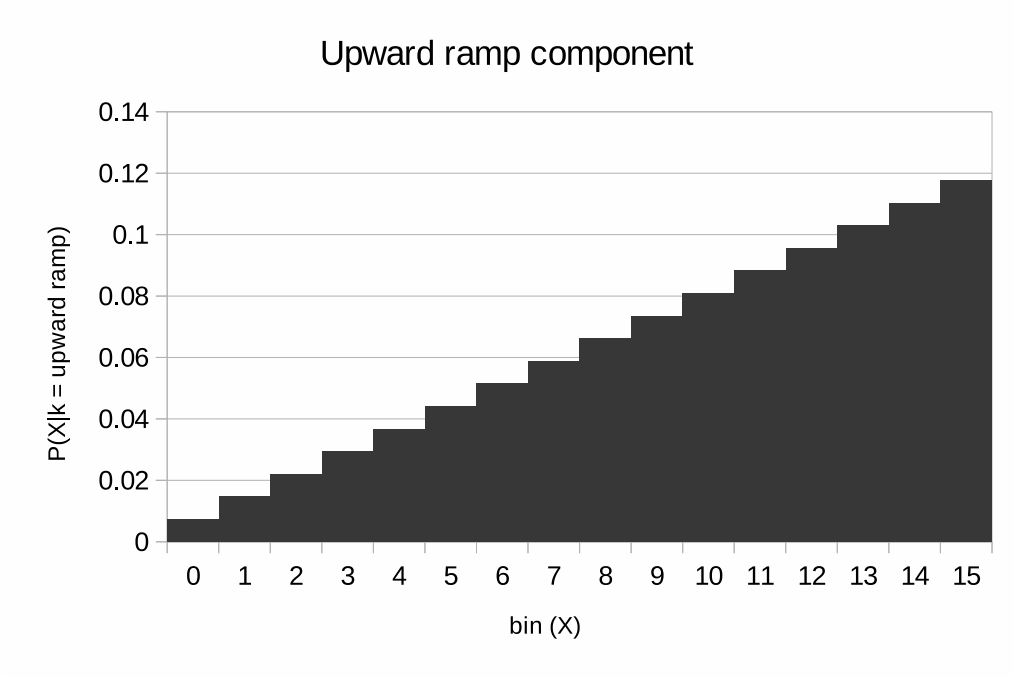

Figure 1: The PMF of the upward ramp shaped component.

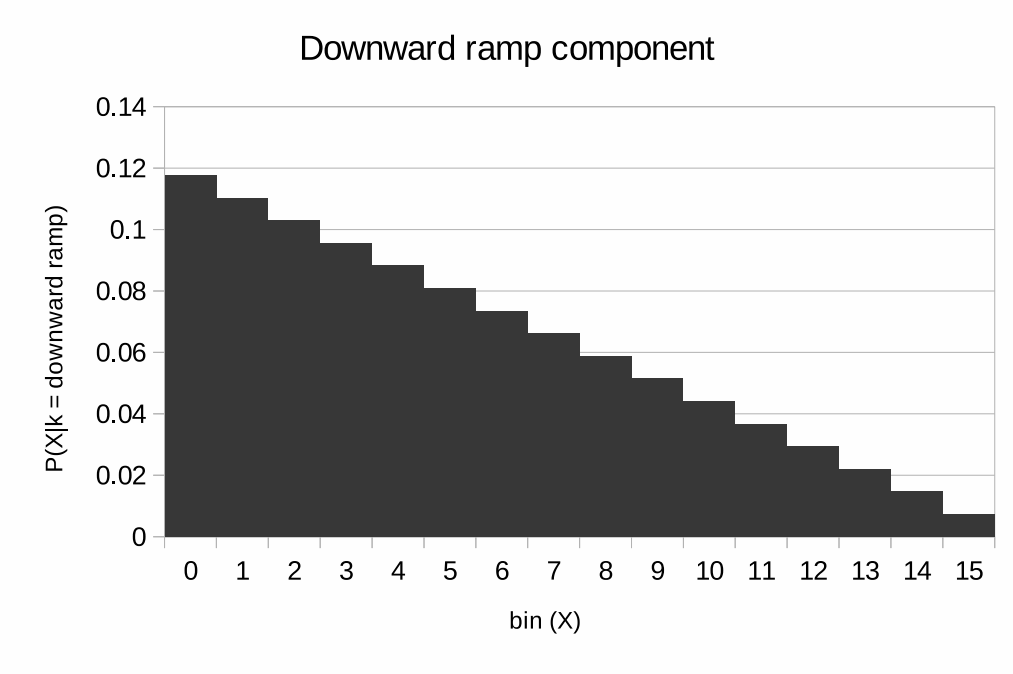

Figure 2: The PMF of the downward ramp shaped component.

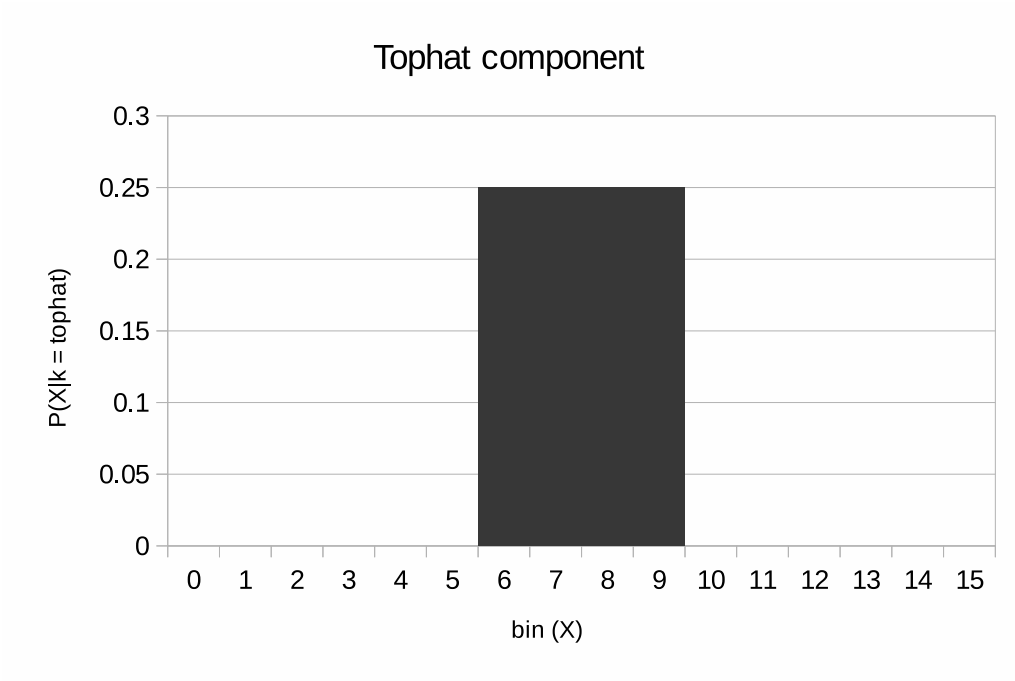

Figure 3: The PMF of the top hap shaped component.

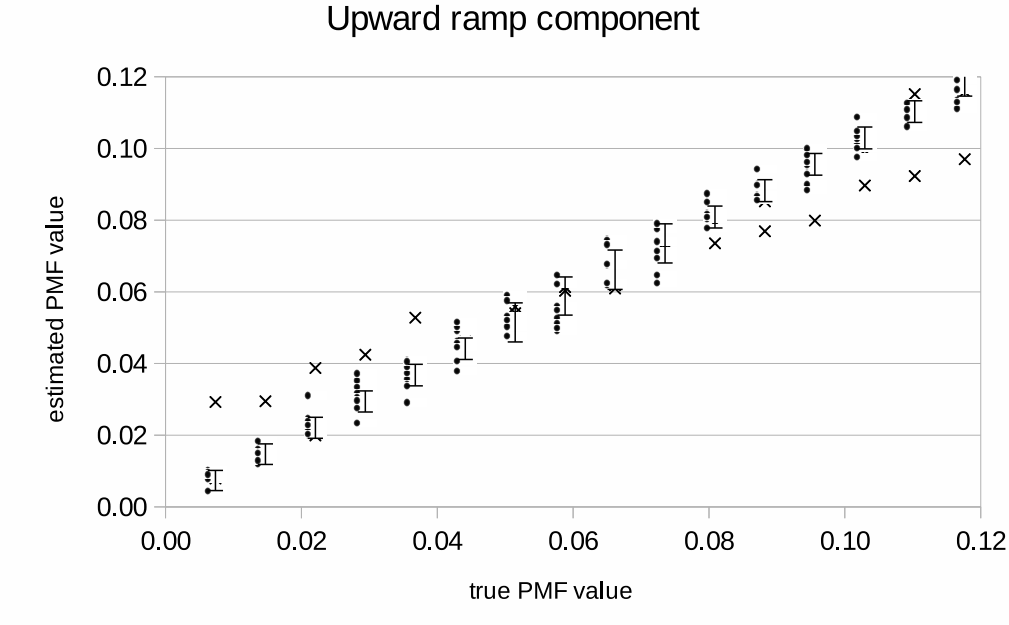

Figure 4: Scatter plot of estimated PMFs against true values. The diagonal set of error bars indicate 1 standard deviation predicted error distributions. The small dots represent the repeated PMF estimates. The crosses indicate attempts to construct models that are believed to have fallen into minima consistent with linear degeneracy.

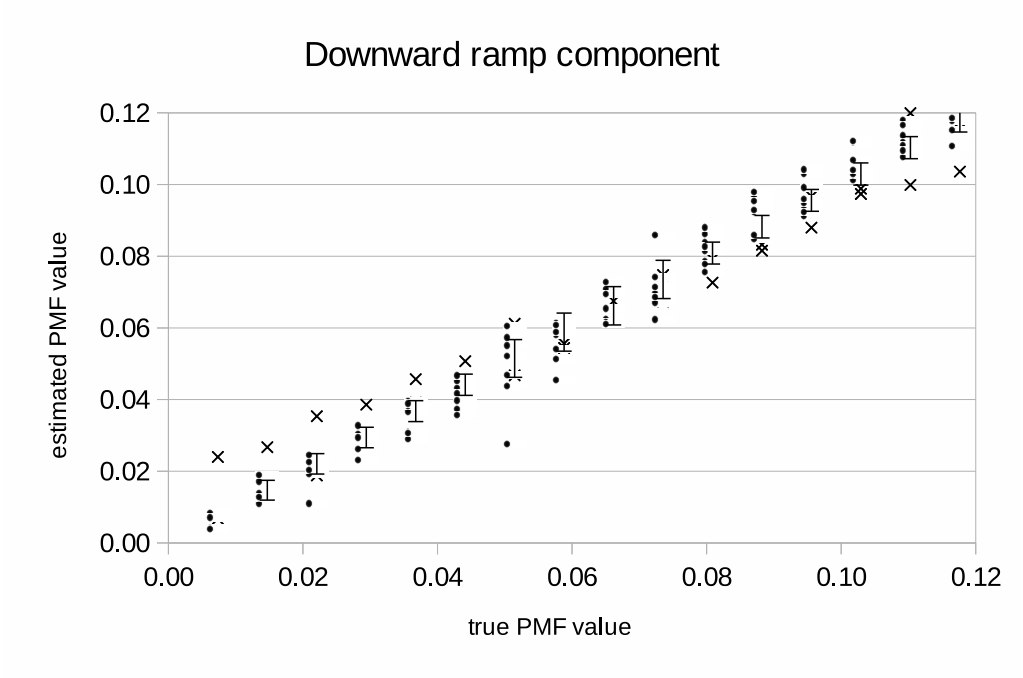

Figure 5: Scatter plot of estimated PMFs against true values. The diagonal set of error bars indicate 1 standard deviation predicted error distributions. The small dots represent the repeated PMF estimates. The crosses indicate attempts to construct models that are believed to have fallen into minima consistent with linear degeneracy.

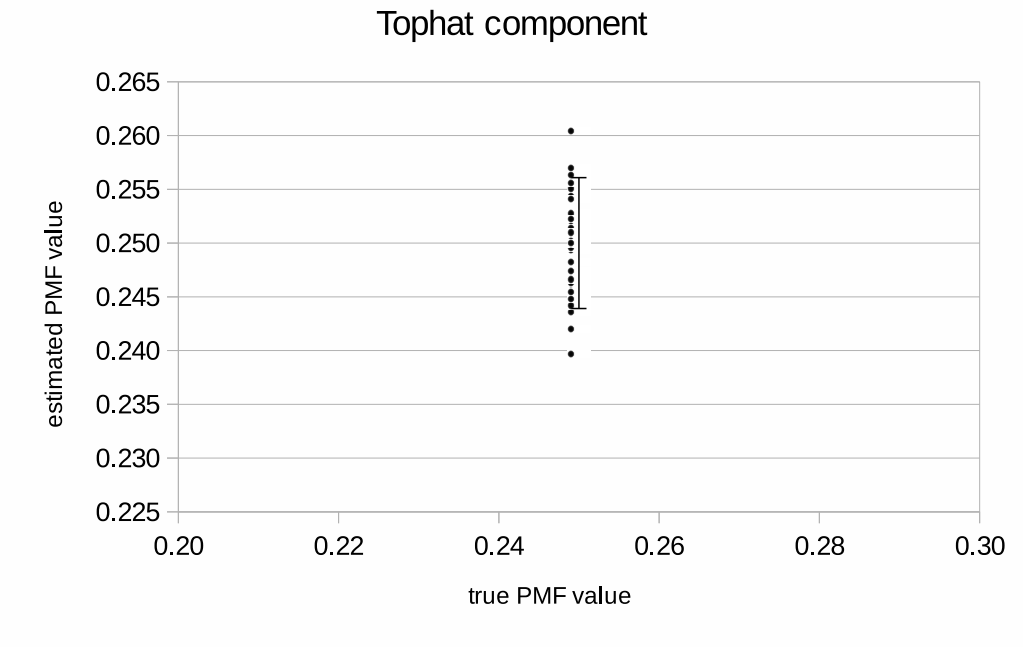

Figure 6: Scatter plot of estimated PMFs against true values. The diagonal set of error bars indicate 1 standard deviation predicted error distributions. The small dots represent the repeated PMF estimates. The crosses indicate attempts to construct models that are believed to have fallen into minima consistent with linear degeneracy.

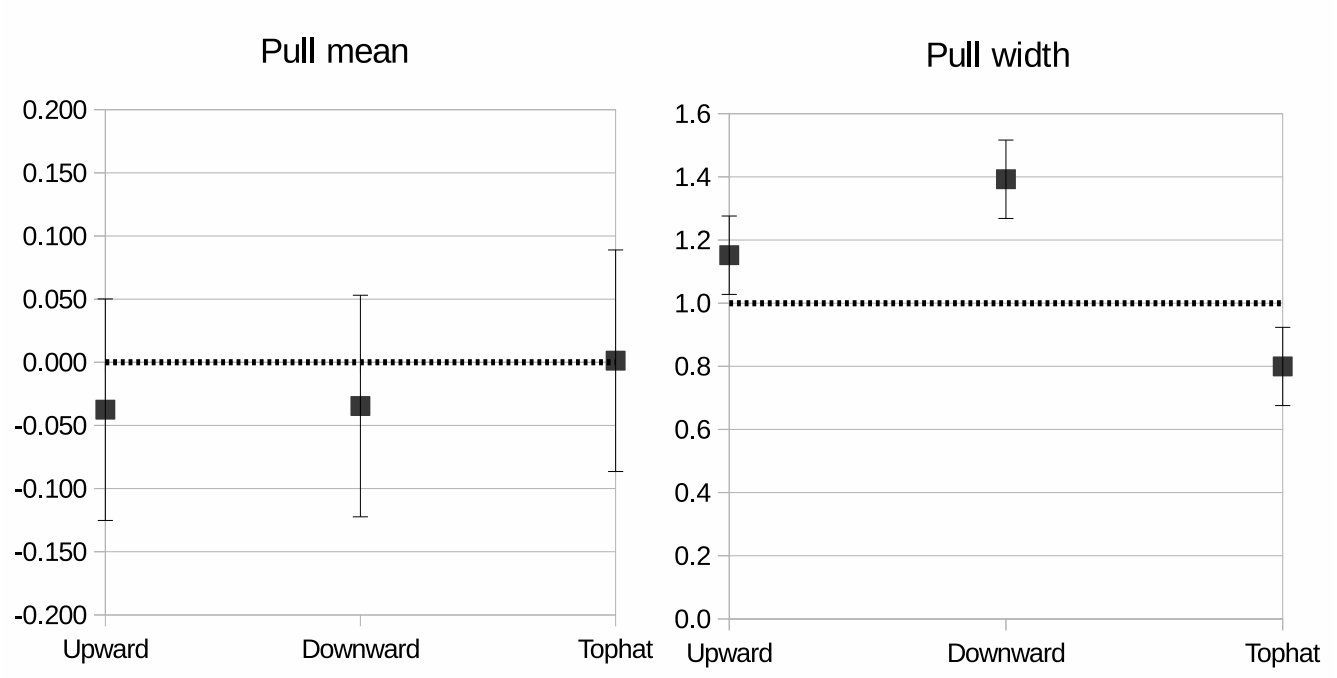

Figure 7: Mean and standard deviation of pull distributions. Dotted lines indicate expected values.