A Bayesian analysis of pentaquark signals from CLAS data

D.G. Ireland,[1] B. McKinnon,[1] D. Protopopescu,[1] P. Ambrozewicz,[13] M. Anghinolfi,[18] G. Asryan,[38] H. Avakian,[33] H. Bagdasaryan,[28] N. Baillie,[37] J.P. Ball,[3] N.A. Baltzell,[32] V. Batourine,[22] M. Battaglieri,[18] I. Bedlinskiy,[20] M. Bellis,[6] N. Benmouna,[15] B.L. Berman,[15] A.S. Biselli,[6,12] L. Blaszczyk,[14] S. Bouchigny,[19] S. Boiarinov,[33] R. Bradford,[6] D. Branford,[11] W.J. Briscoe,[15] W.K. Brooks,[33] V.D. Burkert,[33] C. Butuceanu,[37] J.R. Calarco,[25] S.L. Careccia,[28] D.S. Carman,[33] L. Casey,[7] S. Chen,[14] L. Cheng,[7] P.L. Cole,[16] P. Collins,[3] P. Coltharp,[14] D. Crabb,[36] V. Crede,[14] N. Dashyan,[38] R. De Masi,[8,19] R. De Vita,[18] E. De Sanctis,[17] P.V. Degtyarenko,[33] A. Deur,[33] R. Dickson,[6] C. Djalali,[32] G.E. Dodge,[28] J. Donnelly,[1] D. Doughty,[9,33] M. Dugger,[3] O.P. Dzyubak,[32] H. Egiyan,[33] K.S. Egiyan,[38] L. El Fassi,[2] L. Elouadrhiri,[33] P. Eugenio,[14] G. Fedotov,[24] G. Feldman,[15] A. Fradi,[19] H. Funsten,[37] M. Garçon,[8] G. Gavalian,[28] N. Gevorgyan,[38] G.P. Gilfoyle,[31] K.L. Giovanetti,[21] F.X. Girod,[8,33] J.T. Goetz,[4] W. Gohn,[10] A. Gonenc,[13] R.W. Gothe,[32] K.A. Griffioen,[37] M. Guidal,[19] N. Guler,[28] L. Guo,[33] V. Gyurjyan,[33] K. Hafidi,[2] H. Hakobyan,[38] C. Hanretty,[14] N. Hassall,[1] F.W. Hersman,[25] I. Hleiqawi,[27] M. Holtrop,[25] C.E. Hyde-Wright,[28] Y. Ilieva,[15] B.S. Ishkhanov,[24] E.L. Isupov,[24] D. Jenkins,[35] H.S. Jo,[19] J.R. Johnstone,[1] K. Joo,[10] H.G. Juengst,[28] N. Kalantarians,[28] J.D. Kellie,[1] M. Khandaker,[26] W. Kim,[22] A. Klein,[28] F.J. Klein,[7] M. Kossov,[20] Z. Krahn,[6] L.H. Kramer,[13,33] V. Kubarovsky,[33,29] J. Kuhn,[6] S.V. Kuleshov,[20] V. Kuznetsov,[22] J. Lachniet,[28] J.M. Laget,[33] J. Langheinrich,[32] D. Lawrence,[23] K. Livingston,[1] H.Y. Lu,[32] M. MacCormick,[19] N. Markov,[10] P. Mattione,[30] B.A. Mecking,[33] M.D. Mestayer,[33] C.A. Meyer,[6] T. Mibe,[27] K. Mikhailov,[20] M. Mirazita,[17] R. Miskimen,[23] V. Mokeev,[24,33] B. Moreno,[19] K. Moriya,[6] S.A. Morrow,[8,19] M. Moteabbed,[13] E. Munevar,[15] G.S. Mutchler,[30] P. Nadel-Turonski,[15] R. Nasseripour,[32] S. Niccolai,[19] G. Niculescu,[21] I. Niculescu,[21] B.B. Niczyporuk,[33] M.R. Niroula,[28] R.A. Niyazov,[33] M. Nozar,[33] M. Osipenko,[18,24] A.I. Ostrovidov,[14] K. Park,[22] E. Pasyuk,[3] C. Paterson,[1] S. Anefalos Pereira,[17] J. Pierce,[36] N. Pivnyuk,[20] O. Pogorelko,[20] S. Pozdniakov,[20] J.W. Price,[5] S. Procureur,[8] Y. Prok,[36] B.A. Raue,[13,33] G. Ricco,[18] M. Ripani,[18] B.G. Ritchie,[3] F. Ronchetti,[17] G. Rosner,[1] P. Rossi,[17] F. Sabatié,[8] J. Salamanca,[16] C. Salgado,[26] J.P. Santoro,[7] V. Sapunenko,[33] R.A. Schumacher,[6] V.S. Serov,[20] Y.G. Sharabian,[33] D. Sharov,[24] N.V. Shvedunov,[24] E.S. Smith,[33] L.C. Smith,[36] D.I. Sober,[7] D. Sokhan,[11] A. Stavinsky,[20] S.S. Stepanyan,[22] S. Stepanyan,[33] B.E. Stokes,[14] P. Stoler,[29] S. Strauch,[32] M. Taiuti,[18] D.J. Tedeschi,[32] U. Thoma,[33] A. Tkabladze,[15] S. Tkachenko,[28] C. Tur,[32] M. Ungaro,[10] M.F. Vineyard,[34] A.V. Vlassov,[20] D.P. Watts,[11] L.B. Weinstein,[28] D.P. Weygand,[33] M. Williams,[6] E. Wolin,[33] M.H. Wood,[32] A. Yegneswaran,[33] L. Zana,[25] J. Zhang,[28] B. Zhao,[10] and Z.W. Zhao[32]

(The CLAS Collaboration)

[1] *University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[2] *Argonne National Laboratory, Argonne, IL 60439*
[3] *Arizona State University, Tempe, Arizona 85287-1504*
[4] *University of California at Los Angeles, Los Angeles, California 90095-1547*
[5] *California State University, Dominguez Hills, Carson, CA 90747*
[6] *Carnegie Mellon University, Pittsburgh, Pennsylvania 15213*
[7] *Catholic University of America, Washington, D.C. 20064*
[8] *CEA-Saclay, Service de Physique Nucléaire, 91191 Gif-sur-Yvette, France*
[9] *Christopher Newport University, Newport News, Virginia 23606*
[10] *University of Connecticut, Storrs, Connecticut 06269*
[11] *Edinburgh University, Edinburgh EH9 3JZ, United Kingdom*
[12] *Fairfield University, Fairfield CT 06824*
[13] *Florida International University, Miami, Florida 33199*
[14] *Florida State University, Tallahassee, Florida 32306*
[15] *The George Washington University, Washington, DC 20052*
[16] *Idaho State University, Pocatello, Idaho 83209*
[17] *INFN, Laboratori Nazionali di Frascati, 00044 Frascati, Italy*
[18] *INFN, Sezione di Genova, 16146 Genova, Italy*
[19] *Institut de Physique Nucleaire ORSAY, Orsay, France*
[20] *Institute of Theoretical and Experimental Physics, Moscow, 117259, Russia*
[21] *James Madison University, Harrisonburg, Virginia 22807*
[22] *Kyungpook National University, Daegu 702-701, South Korea*
[23] *University of Massachusetts, Amherst, Massachusetts 01003*
[24] *Moscow State University, General Nuclear Physics Institute, 119899 Moscow, Russia*
[25] *University of New Hampshire, Durham, New Hampshire 03824-3568*
[26] *Norfolk State University, Norfolk, Virginia 23504*

[27]Ohio University, Athens, Ohio 45701
[28]Old Dominion University, Norfolk, Virginia 23529
[29]Rensselaer Polytechnic Institute, Troy, New York 12180-3590
[30]Rice University, Houston, Texas 77005-1892
[31]University of Richmond, Richmond, Virginia 23173
[32]University of South Carolina, Columbia, South Carolina 29208
[33]Thomas Jefferson National Accelerator Facility, Newport News, Virginia 23606
[34]Union College, Schenectady, NY 12308
[35]Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0435
[36]University of Virginia, Charlottesville, Virginia 22901
[37]College of William and Mary, Williamsburg, Virginia 23187-8795
[38]Yerevan Physics Institute, 375036 Yerevan, Armenia
(Dated: February 23, 2008)

We examine the results of two measurements by the CLAS collaboration, one of which claimed evidence for a Θ^+ pentaquark, whilst the other found no such evidence. The unique feature of these two experiments was that they were performed with the same experimental setup. Using a Bayesian analysis we find that the results of the two experiments are in fact compatible with each other, but that the first measurement did not contain sufficient information to determine unambiguously the existence of a Θ^+. Further, we suggest a means by which the existence of a new candidate particle can be tested in a rigorous manner.

PACS numbers: 13.60.Rj; 12.39.Mk; 14.20.Jn; 14.80.-j; 02.50.-r; 02.70.Uu
Keywords: pentaquark; CLAS; Bayesian

The debate about the existence of the $S = +1$ $\Theta^+(1540)$ baryon state is still raging at this point in time in spite of results from dedicated, high-statistics measurements. One of these, [1], from the CLAS collaboration at the Thomas Jefferson National Accelerator Facility used the reaction $\gamma d \to pK^+K^-n$. It showed convincing evidence that production cross sections for such a state are nowhere near the levels implied by an earlier CLAS measurement [2] of the same channel, which had seen a peak in the pK^- missing mass spectrum at 1.542 GeV/c^2 with a 5.2σ statistical significance. The salient point is that the work of Ref. [1] was a dedicated, high statistics repeat of Ref. [2], where the experimental running conditions were as similar as practically possible.

In the whole history of Θ^+ pentaquark searches, there were several independent experiments that claimed to have found evidence, whilst a similar number claimed to have found nothing. It is impractical to examine the results of all such experiments in a consistent fashion, but the similarity of the two CLAS experiments provides us with an ideal opportunity to investigate apparently contradictory results.

One can examine in detail whether any discrepancy arose from the data quality of the two experiments by making systematic tests on, for example, the effects of different cuts. In the original work for both measurements, however, parallel analyses were carried out to confirm the final spectra, and different internal reviews verified the correctness of the analysis procedures. We therefore assume that the quality of the data in both the experiments was consistent, and that the analyses of both experiments were carried out correctly. We concentrate solely on the end-points of the analyses: namely, the events passing all cuts, which contribute to missing mass spectra.

To get a feel for the problem, we took the data set from Ref. [1] (hereafter referred to as "g10" after the CLAS running period in which the data was obtained) which had been analyzed in exactly the same way as the data from Ref. [2] (hereafter referred to as "g2a"). The g10 data contained a factor of just under six more events, which could be directly compared. The g10 data were then split into five independent subsamples, each containing the same number of counts as the g2a data set, and pK^- missing mass spectra were produced. These missing mass spectra would be where a Θ^+ might be expected to appear. The g10 subsample spectra are depicted in figure 1a-e, and the g2a spectrum is depicted in figure 1f.

Peak-like features appear in several of the g10 subsamples, but the shapes are by no means consistent. As mentioned previously and in keeping with current convention, the g2a result quoted a "significance" of about 5σ, which was similar to other experiments claiming evidence of discovery. However, 5σ means that the probability that a feature is a fluctuation is of the order of 10^{-6}. This is a very small number; it does not appear to match the relative ease of generating peak-like features in the subsample spectra. How do we quantify the intuitive feeling that the odds of obtaining the observed g2a peak from fluctuations are not as small as 1 in 10^6 ?

In this letter we attempt to address this problem within a Bayesian analysis framework, and to suggest an alternative means of quantifying the evidence for discovery. What is specifically required is a quantitative comparison between two hypotheses: "the spectrum contains a peak", and "the spectrum does not contain a peak". One can model the shape of a spectrum as the addition of sim-

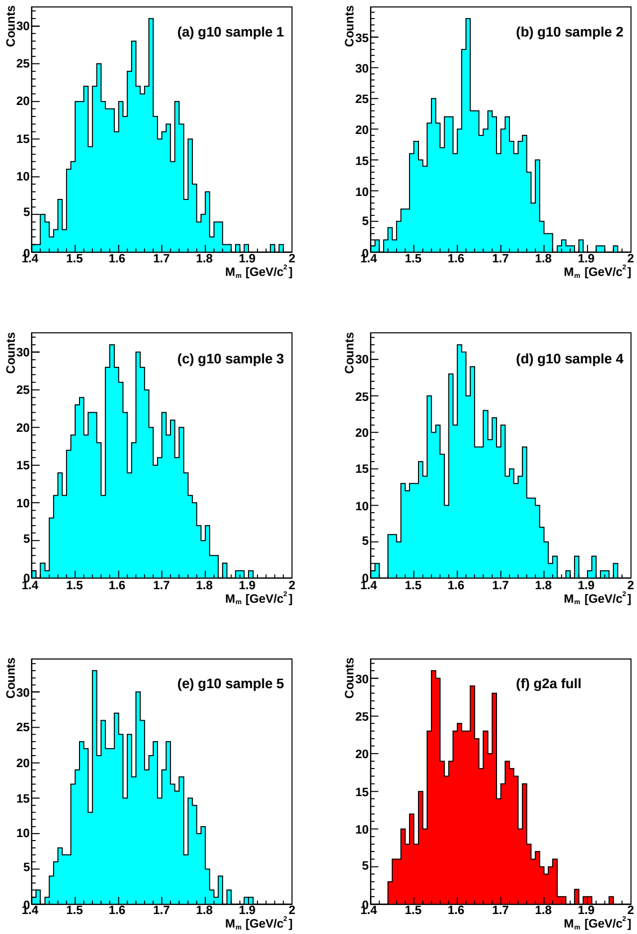

FIG. 1: (Color online) pK^- missing mass spectra from the five g10 subsamples and the original g2a data. The data are sorted into bins of width 10 MeV/c^2.

ple functions, provided that they appear to describe the shape of the spectrum reasonably well, and have plausible physical origins (e.g. Gaussians for resolution effects, etc.). We refer to these as "data models", to distinguish them from theoretical models. The posterior probability that a data model (M) is true given some observed data (D) is given by Bayes' theorem,

$$P(M \mid D) = \frac{P(D \mid M)\, P(M)}{P(D)}, \qquad (1)$$

where $P(D \mid M)$ is the probability of the data being observed given the model, and $P(M)$ represents the prior probability of the model being correct. $P(D)$ is a normalizing constant, which will cancel out in the ratio, that compares the posterior probabilities of two models.

Now the data model will depend on some parameters ξ, and the posterior probability of these taking on specific values is

$$P(\xi \mid D, M) = \frac{P(D \mid \xi, M)\, P(\xi \mid M)}{P(D \mid M)}, \qquad (2)$$

where $P(D \mid \xi, M)$ is the probability of the data being observed given the model and its parameters, and $P(\xi \mid M)$ is the prior probability of the parameters. Fitting parameters to data is a matter of maximizing this posterior. The quantity in the denominator of Eq. (2) is known as the *evidence* for a model and is obtained by marginalizing (integrating) over the parameters:

$$P(D \mid M) = \int d\xi P(D \mid \xi, M)\, P(\xi \mid M). \qquad (3)$$

Since the evidence is an integral over the model parameters, it implicitly implements Occam's razor. Evidence ratios provide a balance between favouring on the one hand the simpler model, and on the other hand the model that better fits the data.

We construct two very simple data models of the missing mass spectra obtained from experiment:

- Model M_0: The spectrum can be described by a 3^{rd} order polynomial in the region of interest. This represents the assumption that there is no new particle. A 3^{rd} order polynomial was employed in the original analysis to model the background shape. This model depends on four parameters.

- Model M_P: The spectrum can be described by a "narrow" Gaussian peak sitting atop a 3^{rd} order polynomial background in the region of interest. "Narrow" in this case meaning that the width is significantly less than the region of interest in the mass spectrum. This model depends on seven parameters.

To compare the different models, a ratio of their probabilities in the light of data can be formed:

$$R_E = \frac{P(M_P \mid D)}{P(M_0 \mid D)} = \frac{P(D \mid M_P)}{P(D \mid M_0)} \times \frac{P(M_P)}{P(M_0)}, \qquad (4)$$

where Bayes' theorem has been used to obtain the final expression. This is the ratio of evidences for the models multiplied by the ratio of prior probabilities of the models. If there is no prior preference for either model, the final factor is unity, so the ratio of model probabilities becomes a ratio of evidences. R_E is known as the "Bayes' Factor" or "evidence ratio".

It is computationally convenient and equivalent to examine the logarithms of the evidence ratios:

$$\ln(R_E) = \ln P(D \mid M_P) - \ln P(D \mid M_0)\,. \qquad (5)$$

Determining what value of $\ln(R_E)$ to use in deciding between data models is somewhat arbitrary, but Jeffreys established [3] a rough evidence scale versus written descriptors: $|\ln(R_e)| < 1$ is *weak*, $1 < |\ln(R_e)| < 2.5$ is *substantial*, $2.5 < |\ln(R_e)| < 5$ is *strong* and $|\ln(R_e)| > 5$ is *decisive*. So model comparison is quantified by R_E, and as constructed means that data favouring a data model with a peak have positive $\ln(R_e)$.

To evaluate evidences, we see from Eq. (3) that an integral over a likelihood $P(D \mid \xi, M)$ and a prior $P(\xi \mid M)$ is required. We calculate the likelihood by evaluating for each bin in a spectrum an "ideal" number of counts, $S_i(\xi)$, for a given set of parameters. The probability of this being correct given the measured counts n_i is calculated using a Poisson distribution. The total likelihood is then a product of these probabilities for each bin:

$$P(D \mid \xi, M) = \prod_i \frac{S_i^{n_i} \exp(-S_i)}{n_i!}. \quad (6)$$

Here, the prior probability is constructed by assuming no initial correlations between parameters, so it is simply a product of priors for each separate parameter. We assume that each prior is a uniform distribution between a lower and upper limit since this represents the least initial bias. The prior parameter ranges were established by performing an initial fit and setting the limits to be $\pm 50\%$ of the values found. This resulted in a large flexibility in the shapes of both background and peak.

To perform the integrations over the many parameters in the models, we utilized the technique of "nested sampling" developed by Skilling [4, 5]. This is a Monte Carlo integration method for which we outline a brief sketch, but refer to the original reference for details and to Ref. [6] for an example application. An initial N sample points are drawn with uniform probability from the full prior parameter space. The likelihood function is evaluated at each of the points, and the "worst" point with the lowest likelihood is saved. An iteration consists of replacing the previous worst point by another sample point drawn with uniform probability from the parameter space, *provided that its evaluated likelihood is greater than the previous worst.* The algorithm thus iterates its way towards the region of highest likelihood. The successive points converge, and a stopping criterion based on the desired accuracy can be applied. The likelihoods of the stored points and the remaining points L_j are then used in a weighted sum to give a Monte Carlo estimate of the integral. The weight of the point L_j represents half the volume of the parameter space whose likelihood value lies between L_{j-1} and L_{j+1}, and is estimated probabilistically.

We applied the model comparison framework to all the spectra shown in figure (1). In addition we analyzed the spectra shown in figure (2), which consisted of: (a) the full g10 spectrum; (b) a "fake" spectrum, constructed by sampling from a combination of signal and background functions in the data model with the peak (M_P), which had the same signal-to-background ratio as the g2a spectrum. This was done to show what the results of this analysis would have been, had a resonance been there; (c) and (d) pK^- missing mass spectra from the g2a and g10 data sets, but showing the $\Lambda(1520)$ signal, in order to test how the technique fared for the case of a well-established particle.

The results are quoted in table (I), and displayed graphically in figure 3. We omit the results for the $\Lambda(1520)$ from the figure, as they would render the scale unusable. To estimate the uncertainty in the Monte Carlo integrals, we ran at least 20 independent calculations for each spectrum analysed. The errors listed in the table represent the standard error of the samples.

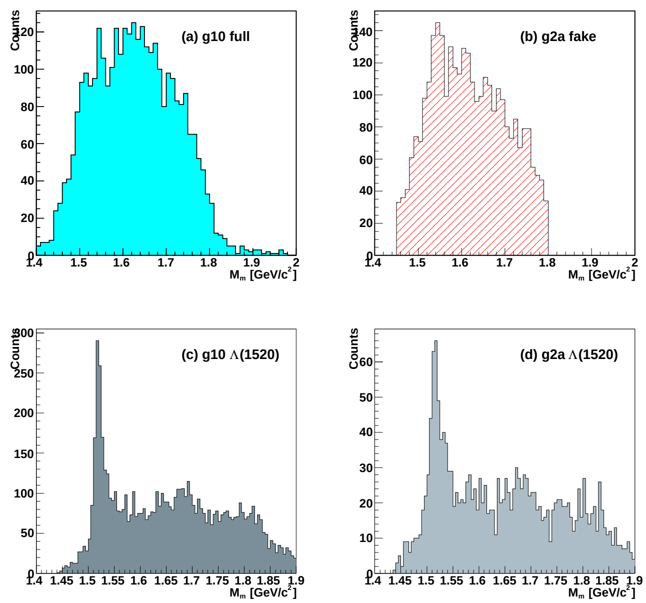

FIG. 2: (Color online) Missing mass histograms for Θ^+ from a) g10, b) fake, and $\Lambda(1520)$ from c) g10, and d) g2a data. The data in a) and b) are sorted into bins of width 10 MeV/c^2, and the bins in c) and d) have width 5 MeV/c^2.

With the splitting of the g10 data set, we have shown

Data sample	$\ln(R_E)$		
g10 sample 1	-1.56	±	0.07
g10 sample 2	-1.09	±	0.13
g10 sample 3	-1.64	±	0.09
g10 sample 4	-1.11	±	0.11
g10 sample 5	-1.82	±	0.07
g10 full	-2.87	±	0.11
g2a	-0.41	±	0.10
fake	5.78	±	0.27
g2a $\Lambda(1520)$	96.70	±	0.70
g10 $\Lambda(1520)$	549.12	±	2.17

TABLE I: Evidence ratios. Calculations are done by nested sampling, hence the need to include standard errors.

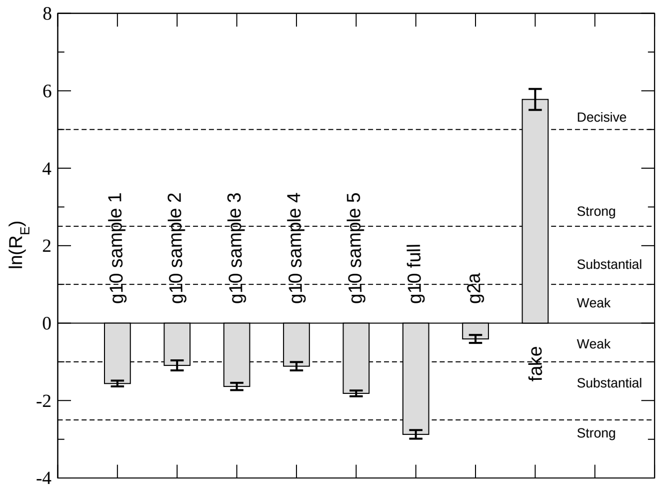

FIG. 3: Graphical representation of the values of evidence ratios from table I, on a logarithmic scale. The horizontal lines correspond to the limits of the regions associated with the different descriptors of the Jeffreys scale.

(figure 1) the relative ease with which one can obtain a peak-like feature, given a small number of events. The evidence ratios calculated for the individual subsamples in g10 generally suggest a bias against a peak, which perhaps mirrors an intuitive feeling about how significant such features really are. However, two of the five subsamples (2 and 4) are compatible with the "weak" category, meaning that the results are essentially inconclusive. Whilst the g2a result is more of an outlier, it also falls in the weak category and is inconclusive; the results of the two measurements are therefore compatible with each other.

The $\ln(R_E)$ value for g2a (-0.408) indicates weak evidence in favour of the data model without a peak in the spectrum. What this means is that whilst a data model including a peak gives a better fit by eye to the spectrum, it does not compensate for having had to introduce additional parameters for the peak. This is Occam's razor in action; simpler models are preferable unless more complex models do much better. One must be careful what to conclude from the g2a spectrum, however, since the evidence ratio does *not* conclusively rule out a peak; it is simply inconclusive.

We now turn to the question of whether the g10 experiment could conclusively discriminate between the two possibilities. The log of the evidence ratio for the full g10 spectrum is -2.9. This makes it *strong* evidence against a peak in the spectrum. Another way of looking at this is that with this evidence ratio, the odds against a peak in this spectrum are about 17 to 1. Whilst this cannot completely rule out a discovery, another measurement of this channel is probably not necessary. By comparison, the odds in favour of a peak in the fake spectrum are about 320 to 1, meaning that had a signal really been there in g10, the experimental result would have been *decisive*.

The study of the $\Lambda(1520)$ shows that when a resonance is there, this method picks it out rather readily, with both g2a and g10 data sets yielding a *decisive* result. We take this as a positive test that our method works.

In summary, we have applied a Bayesian model comparison method to analyzing the missing mass spectra produced in pentaquark searches. This has been used to study the relationship between the results of two CLAS measurements, which were taken under almost identical conditions. We have shown that there is no conflict between the results of the two experiments, and that the low number of counts in the first experiment resulted in an ambiguous signal. Furthermore we have shown that the g10 result shows strong evidence against the discovery of a pentaquark in this channel. More generally, this method could be applied to any data set where a search for a new state has been carried out, and can provide a quantitative measure with which to judge whether or not a result represents a discovery.

Acknowledgments

We would like to thank G. Woan (Glasgow) for useful discussions. We would also like to thank the staff of the Accelerator and Physics Divisions at Jefferson Lab who made the experiments possible. Acknowledgments for the support of these experiments go also to the Italian Istituto Nazionale de Fisica Nucleare, the French Centre National de la Recherche Scientifique and Commissariat à l'Energie Atomique, the Korea Research Foundation, the U.S. Department of Energy and the National Science Foundation, and the U.K. Engineering and Physical Science Research Council. The Jefferson Science Associates (JSA) operates the Thomas Jefferson National Accelerator Facility for the United States Department of Energy under contract DE-AC05-06OR23176.

[1] B. McKinnon et al. (CLAS), Phys. Rev. Lett. **96**, 212001 (2006).

[2] S. Stepanyan et al. (CLAS), Phys. Rev. Lett. **91**, 252001 (2003).

[3] H. Jeffreys, *Theory of Probability* (Oxford University Press, 1961).

[4] J. Skilling, in *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics* (2006).

[5] D. Sivia and J. Skilling, *Data Analysis - A Bayesian Tutorial* (Oxford University Press, Oxford, UK, 2006), 2nd ed.

[6] P. Mukherjee, D. Parkinson, and A. R. Liddle, Astrophys. J. **638**, L51 (2006).