SLAC-PUB-1358
(E)
December 1973

MEASUREMENT OF MULTIVARIATE SCALING AND FACTORIZATION IN EXCLUSIVE MULTIPARTICLE PRODUCTION*

Jerome H. Friedman

Stanford Linear Accelerator Center

Stanford, California

November 1973

ABSTRACT

Several new model independent techniques for the analysis of multidimensional data are presented and applied to exclusive multiparticle production. For the reaction $pp \to pp\pi^+\pi^+\pi^-\pi^-$ from 12 to 28 GeV/c, an algorithm that directly compares two multidimensional point distributions is used to show that the shape of $d^{12}\sigma/d(\vec{p}_\perp)^{12}$ is energy independent while $d^6\sigma/dx^6$ ($x=p_{||}/p_{max}$) varies dramatically with beam energy. A multidimensional test for independence is used to show that the multivariate differential cross-section approximately factors into its cylindrical momentum components,

$$d^{18}\sigma/d(\vec{p})^{18} \simeq d^6\sigma/dp_{||}^6 \cdot d^6\sigma/d(p_\perp^2)^6 \cdot d^6\sigma/d\varphi^6.$$

The shape of $d^6\sigma/d\varphi^6$ is also shown to be compatible with that predicted solely by kinematics.

(Submitted to The Physical Review)

*Work supported by the U.S. Atomic Energy Commission.

Multiparticle production presents a difficult problem in data analysis due to the large number of independent observables necessary to completely describe the data. Excluding spin information, an n-particle final state requires $3n-4$ independent measurables for a complete description. The normalized multivariate differential cross section

$$\rho(x_1, x_2, \cdots, x_{3n-4}) = \frac{1}{\sigma} \frac{d^{3n-4}\sigma}{dx_1 dx_2 \cdots dx_{3n-4}}$$

can be thought of as a probability density function in the chosen measurables, $x_1 x_2 \cdots x_{3n-4}$, defined over their physically allowed values.

Each event can be represented as a point in the multidimensional space whose coordinates are the measurables of the event. This point contains all the information in the event and, thus, the collection or swarm of experimental points in this space contains all the information from the experiment. The purpose of data analysis is to infer the properties of the unknown probability density function (p.d.f.) from the experimental point sample in the multidimensional space. For exclusive experiments where all $3n-4$ independent observables are measured, this p.d.f. is directly related to the transition matrix element squared for the reaction.

This report describes the results of applying several newly developed nonparametric (model independent) techniques for investigating the properties of multidimensional point swarms[1,2] to the reaction $pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$ from from 12 to 28 GeV/c.[3] An algorithm that directly compares the shapes of two multidimensional point distributions is used to study the energy dependence of the multivariate differential cross-section, while a multivariate independence test is used to study its factorization properties.

Comparing Two Multidimensional Point Distributions

This algorithm tests the hypothesis that two multidimensional point samples were drawn from the same unknown p.d.f. It is completely described in Reference 1 and only its essentials are discussed here. The two samples (classes) with N_1 and N_2 events, respectively, are combined into a single sample of N_1+N_2 events with each event tagged as to its origin. Associated with each event in the combined sample is a measure of the composition of the events closest to it. Specifically, the closest k points to each point are examined and the number that come from class one, k_1, is observed. In Reference 1, it is shown that for the null hypothesis (both samples from the same p.d.f.), this quantity k_1, as measured for all of the events, is distributed according to a binomial distribution $B_p^k(k_1)$ with probability $p = N_1/(N_1 + N_2)$. For alternate hypotheses, this quantity will not be so distributed. A Pearson χ^2 statistic[4] is formed between a histogram of the frequency of the various possible values of k_1 ($0 \leq k_1 \leq k$) and the corresponding binomial distribution. Since this χ^2 is a measure of the deviation of the experimentally measured frequency of k_1 from its expected distribution when the two samples have the same multidimensional shape, it provides a measure of the disagreement between the multidimensional shapes of the two samples.

To study the energy dependence of the shape of the multivariate differential cross-section for the reaction $pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$, 203 events at 12 GeV/c were compared to 196 events at 28 GeV/c.[5] Figure 1 shows frequency histograms of the number of 12 GeV/c events in a 20 event neighborhood about every event in the combined sample for various coordinate subspaces. Figure 1a compares the two samples in the six-dimensional subspace[6] of scaled center of mass momentum components[7] parallel to the incident beam. The frequency histogram, for this case, deviates considerably from the binomial distribution (open circles) expected for the null hypothesis, indicating

that these two samples differ considerably in their multidimensional shapes. Figure 1b shows the results of comparing these two samples in the 12 dimensional momentum subspace[8] transverse to the incident beam direction. Figures 1c and 1d show the comparison in the two six-dimensional cylindrical coordinate subspaces of transverse momentum. In contrast to the scaled longitudinal momenta, the frequency histogram for these cases does not deviate significantly from the expected binomial distribution. The comparison is slightly better in the azimuthal angle subspace than the subspace of the transverse momenta squared, however, neither deviates strongly from the expected binomial distribution.

These results show that the 12 dimensional shape of the differential cross-section transverse to the beam direction, $d^{12}\sigma/d(\vec{p}_\perp)^{12}$, is either independent of or, at most, varies slowly with energy for this reaction in this energy range. By contrast, the shape of the six-dimensional differential cross-section parallel to the incident beam, $d^6\sigma/dx^6$, is changing considerably; thus, the energy dependence of the dynamics manifests itself mostly, if not completely, in the longitudinal variables.

Determining the Factorization Properties of a Multidimensional Point Distribution

This algorithm seeks to test the hypothesis that the unknown multidimensional p.d.f., from which the data points were sampled, factors into the product of two density functions each defined over exclusive orthogonal subspaces of the full dimensional space,

$$\rho\,(x_1,x_2,\ldots,x_n) = \rho\,(x_1,x_2,\ldots,x_m)\cdot\rho\,(x_{m+1},x_{m+2},\ldots,x_n) \qquad (1)$$

For this test, two procedures were employed.

The first procedure is based on the mutual information measure[9] of the relationship between coordinate pairs. The mutual information (MI) between

a pair of coordinates (x_i, x_j) is defined as

$$M_{ij} = [H(x_i) + H(x_j) - H(x_i, x_j)] \;/\; \text{Minimum}\,[H(x_i), H(x_j)], \qquad (2)$$

where $H(x_i)$ is an estimate of the information (negentropy) of the data as projected onto the ith coordinate axis. Similarly, $H(x_i, x_j)$ is the estimated information of the data as projected onto the $x_i x_j$ plane. If there is no relationship between x_i and x_j in the data (null hypothesis), then M_{ij} will have a small value[10] while if there is a complete dependence (x_i determines x_j), then M_{ij} will take on its maximum value, unity. For partial relationships, M_{ij} will have values between these extremes. The MI measure is a generalization of the correlation coefficient

$$C_{ij} = \langle (x_i - \langle x_i \rangle)\,(x_j - \langle x_j \rangle) \rangle \Big/ \sqrt{\langle (x_i - \langle x_i \rangle)^2 \rangle \langle (x_j - \langle x_j \rangle)^2 \rangle}$$

in that the MI measures any relationship between the coordinates, whereas the correlation coefficient only measures linear relationships.

The MI measures, M_{ij} $(1 < i \leq 18,\ 1 \leq j < i)$ were evaluated in the following manner. Each axis of the coordinate pair was divided into ten channels with equal number of counts in each channel. These channels define a 100 cell grid on the $x_i x_j$ plane and the number of counts in each of these cells is determined. The information is defined as

$$H = -\sum_{\ell=1}^{M} \frac{n_\ell}{N} \log \frac{n_\ell}{N}$$

where M is the number of cells or channels, n_ℓ is the number of counts in each cell and N is the total number of events. For the one-dimensional axis projections, $n_\ell = N/M$ by construction so that $H = \log M$ and the MI (Eqn 2) for this case becomes

$$M_{ij} = 2 - (1/\log 10) \sum_{\ell=1}^{100} \frac{n_\ell}{N} \log \frac{n_\ell}{N} \quad . \qquad (3)$$

For the factorization hypothesis (Eqn 1) to be true, it is necessary for the MI measures for all pairs of coordinates between the two subspaces (M_{ij}, $1 < i \leq m$, $m + 1 \leq j \leq n$) to be consistent with the minimum value expected for the null hypothesis.

A cylindrical coordinate representation was chosen to describe the 18 dimensional momentum space for the reaction $pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$ at 23 GeV/c.[11] The first six coordinates are the transverse momentum azimuthal angles, φ, for each of the particles, followed by the squares of the transverse momentum, $p_\perp^2$, for each of them and finally the six center of mass longitudinal momenta, $p_{||}$, for each of the particles. The 153 MI measures, M_{ij}, between each pair of these coordinates were calculated using Eqn 3. These MI measures were then standardized[12]

$$\hat{M}_{ij} = (M_{ij} - 0.0186) \; / \; 0.0030 \quad . \qquad (4)$$

Here 0.0186 is the expected value of M_{ij}, and 0.0030 is the expected standard deviation about that value, for the case where x_i and x_j are independent (null hypothesis). These $\hat{M}_{ij}$ values were arranged in a lower triangular matrix ($2 \leq i \leq 18$, $1 \leq j \leq i - 1$). Table 1a summarizes this matrix by averaging its elements over the regions that correspond to the three six-dimensional subspaces of the φ, $p_\perp^2$ and $p_{||}$ coordinates. A second $\hat{M}_{ij}$ matrix with the rapidity, η,[13] replacing the $p_{||}$ is also shown.

Inspecting Table 1a shows that the data is consistent with no pairwise MI between the six dimensional subspaces of the φ and both those of the $p_\perp^2$ and $p_{||}$. In addition, there is no pairwise MI within the subspace of the $p_\perp^2$. There is measurable MI within the subspaces of both the φ and $p_{||}$ as well as between the subspaces of the $p_\perp^2$ and $p_{||}$.

Momentum and energy conservation require non-zero MI among some of the coordinates. Specifically, momentum conservation requires non-zero MI within the φ and $p_{\parallel}$ subspaces, while energy conservation requires non-zero MI within both the $p_{\perp}^2$ and $p_{\parallel}$ subspaces as well as between the $p_{\parallel}$ and $p_{\perp}^2$ subspaces. Owing to relative average smallness of the $|\vec{p}_{\perp}|$ as compared to the $|p_{\parallel}|$, this energy conservation effect should appear most strongly in the $p_{\parallel}$ subspace, with a small contribution to the MI between the $p_{\parallel}$ and $p_{\perp}^2$, and an even smaller contribution within the $p_{\perp}^2$ subspace. Energy conservation should also produce more MI between a particle's $p_{\perp}^2$ and its own $p_{\parallel}$, than with the other particles' $p_{\parallel}$. A Table 1a insert subdivides the $p_{\parallel}$ - $p_{\perp}^2$ region of the MI matrix into its diagonal elements, which represent the MI between the $p_{\perp}^2$ and $p_{\parallel}$ of the same particle and the off diagonal region which represents the MI between the $p_{\perp}^2$ and $p_{\parallel}$ of different particles. As can be seen, the average MI between the $p_{\perp}^2$ and $p_{\parallel}$ of the same particle is much larger than that with different particles.

Another quasi-kinematic mechanism for the non-zero MI between the $p_{\perp}^2$ and $p_{\parallel}$ could be the inverse energy of each particle that appears in the Lorentz invariant phase space volume element

$$dv = \prod_{i=1}^{n} \left(\frac{dp_{\parallel}\, dp_{\perp}^2 d\varphi}{E} \right)_i \quad .$$

This would produce MI only between $p_{\perp}^2$ and $p_{\parallel}$ of the same particle. To investigate this possibility, Table 1a also shows the corresponding MI matrix for the measurables $p_{\perp}^2$, φ, and $\eta = \sinh^{-1} \left(p_{\parallel} / \sqrt{m^2 + p_{\perp}^2} \right)$ where m is the particle's rest mass. In terms of these variables, the Lorentz invariant volume element becomes

$$dv = \prod_{i=1}^{n} \left(d\eta dp_{\perp}^2 d\varphi \right)_i \quad .$$

The results in Table 1a show, however, that the phase space density, if anything, is less factorable when the rapidity, η, is used instead of $p_{||}$. In particular, the MI between the same particle's $p_{\perp}^2$ and η is considerably larger than between it's $p_{\perp}^2$ and $p_{||}$.

As mentioned above, for the multivariate factorization hypothesis (Eqn 1) to be true, it is necessary that all bivariate MI measures, $\hat{M}_{ij}$ (Eqn 4) between the two subspaces be compatible with zero. This necessity is, however, not sufficient to insure the factorization hypothesis. It is possible for there to exist relationships between the aggregates of coordinates in the two subspaces that average to zero when projected onto all two-dimensional subspaces spanning them. An algorithm that provides a necessary and sufficient test of the factorization hypothesis is described in Reference 2 and is recounted briefly here.

The procedure for comparing two multidimensional point distributions is used to compare the data to another point set of the same sample size constructed as follows. The first m coordinates (Eqn 1) of each constructed point are identical to those of each data point, while the remaining n - m coordinates of each constructed point are obtained by randomly selecting a set of corresponding coordinates from a different data point. Thus, this constructed point set is identical to the data in the subspace of the first m coordinates and a random permutation of the data in the subspace of the last n - m coordinates. The data points are then compared to the constructed points in the full dimensionality using the algorithm discussed in the previous section. For the factorization hypothesis to be true, it is necessary and sufficient that the multidimensional shapes of these two point sets agree. The statistic for testing this hypothesis is less straightforward, however, owing to the non-independence of the two point sets. The quantity k_1 is not generally distributed as a binomial distribution $B_{\frac{1}{2}}^{k}(k_1)$, as for the independent case. The k_1 distribution must, however, be symmetric about a mean of k/2 for the null hypothesis.

Table 1b shows the results of applying this multivariate independence test to the data in terms of the same cylindrical momentum coordinates by showing the mean and third central moment of the frequency histogram of the number of true data points in a 20 point neighborhood of the composite sample of data and constructed points. This mean must be compatible with 1/2 and third central moment must be compatible with zero for the factorization hypothesis to be true. For this data, it is seen from Table 1b that the results of the multivariate tests merely confirm the results of the bivariate tests and no additional relationships seem to be present among these coordinates. The hypothesis $d^{18}\sigma/d(\vec{p})^{18} = d^{12}\sigma/d(\vec{p}_\perp)^{12} \cdot d^6\sigma/dp_\parallel^6$ was also tested with the multivariate MI algorithm, resulting in a mean of .512 ± .004 and third central moment -.00003 ± .0001. Combining this with the results in Table 1, the following factorization properties of the multivariate differential cross-section can be inferred from the data:

$$d^{12}\sigma/d(\vec{p}_\perp)^{12} = d^6\sigma/d(p_\perp^2)^6 \cdot d^6\sigma/d\varphi^6$$

$$d^{12}\sigma/dp_\parallel^6\, d\varphi^6 = d^6\sigma/dp_\parallel^6 \cdot d^6\sigma/d\varphi^6$$

$$d^{12}\sigma/d(p_\perp^2)^6 dp_\parallel^6 \simeq d^6\sigma/d(p_\perp^2)^6 \cdot d^6\sigma/dp_\parallel^6$$

$$d^{18}\sigma/d(\vec{p})^{18} \simeq d^{12}\sigma/d(\vec{p}_\perp)^{12} \cdot d^6\sigma/dp_\parallel^6$$

And thus, $d^{18}\sigma/d(\vec{p})^{18} \simeq d^6\sigma/d(p_\perp^2)^6 \cdot d^6\sigma/d\varphi^6 \cdot d^6\sigma/dp_\parallel^6$.

Here the equal sign (=) represents no measurable relationship between the subspaces, whereas the approximately equal (≃) indicates small but measurable relationship between them. Furthermore, the replacement of the longitudinal momenta, $p_\parallel$, by rapidity, η, does not cause the multivariate differen cross-section to be more factorable and, if anything, renders it less factorable.

As mentioned above, momentum conservation introduces interrelationships within the six-dimensional subspace of the transverse angles. To test whether the relationships measured within this subspace (Table 1a) is due mainly to this mechanism or to dynamical effects, the shape of $d^6\sigma/d\varphi^6$ was compared to that predicted solely by kinematics. For this purpose, 970 Monte Carlo events of the reaction $pp \rightarrow pp(4\pi)$ at 23 GeV/c were generated according to peripheral phase space.[14] These Monte Carlo events were compared to the data in the six-dimensional azimuthal angle subspace. Figure 2 shows the results of the comparison. The frequency of data events within a 20 event neighborhood of each point in the combined sample is clearly compatible with the corresponding binomial distribution. Thus, to the statistical accuracy of this test, the shape of $d^6\sigma/d\varphi^6$ is compatible with that predicted solely by kinematics. In particular, jets in the transverse plane would require the data to approximately lie on a lower dimensional manifold in this six-dimensional space and would be easily detectable.

Helpful discussions with S. Steppel and J.W. Tukey are gratefully acknowledge, as is the excellent work of D.B. Smith in the data reduction phases of the experiment.

FOOTNOTES AND REFERENCES

1. Friedman, J.H., Steppel, S. and Tukey, J.W., "A Nonparametric Procedure for Comparing two Multivariate Point Sets," Stanford Linear Accelerator Center, Computation Group Technical Memo No. 153, November 1973.

2. Friedman, J.H., "A Multivariate Test for Independence," Stanford Linear Accelerator Center, Computation Group Technical Memo No. 154, November 1973.

3. The six-pronged events resulting from a proton exposure in the Brookhaven National Laboratory 80-in. H_2 bubble chamber were fit to the reaction $pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$ in the beam momentum range 12-28 GeV/c. For a detailed description of the experiment see D.B. Smith, Ph.D. thesis, Lawrence Berkeley Laboratory, Report No. UCRL-20632, 1971 (unpublished).

4. Eadie, W.T., Drijard, D., James, F.E., Roos, M., Sadoulet, B., "Statistical Methods in Experimental Physics," North-Holland (1971), pp 257.

5. For the analyses on this data, identical particles were distinguished by the value of their longitudinal momenta.

6. The data is really five-dimensional in this subspace because longitudinal momentum conservation requires all of these data points to lie on a five dimension linear manifold within the six-dimensional space.

7. Feynman, R.P., Phys. Rev. Letters 23, 1415 (1969).

8. The data is really ten-dimensional in this subspace because transverse momentum conservation requires all of these data points to lie on a ten-dimensional linear manifold within the twelve-dimensional space.

9. Lewis, P.M., IEEE Trans. Info. Theory, IT-8, 171-178, February 1962.

10. This minimum value for the MI estimate expected for the null hypothesis depends upon the sample size and approaches zero asymptotically.

11. For these tests, the data from the higher momentum samples (18 to 28 GeV/c) were combined into a single sample of 970 events with mean beam momentum 23 GeV/c.

12. A standardized variable is one whose expected mean is zero and whose expected variance is unity.

13. Wilson, K.B., Acta, Phys. Austr. 17, 37 (1963).

14. Friedman, J.H., Risk, C., Smith, D.B., Phys. Rev. Letters 28, 191 (1971).

TABLE AND FIGURE CAPTIONS

Table 1: Measurements of the factorability of the multivariate differential cross-section in terms of a cylindrical coordinate decomposition.

Figure 1: Tests of the energy dependence of the multivariate shape of the differential cross-section for various coordinate combinations.

Figure 2: Comparison of the multivariate shape of the six-dimensional transverse momentum azimuthal angular distribution to that predicted by kinematics.

TABLE 1

$pp \rightarrow pp\pi^+\pi^+\pi^-\pi^-$

18-28 GeV/c 970 EVENTS

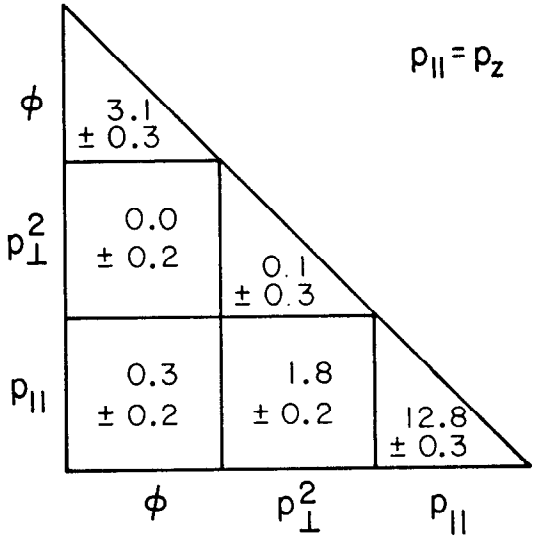
MUTUAL INFORMATION COMPARISONS

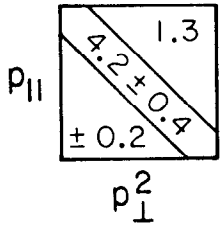
(a) COMPOSITE BIVARIATE

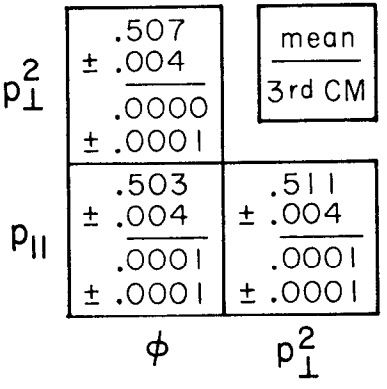
(b) MULTIVARIATE

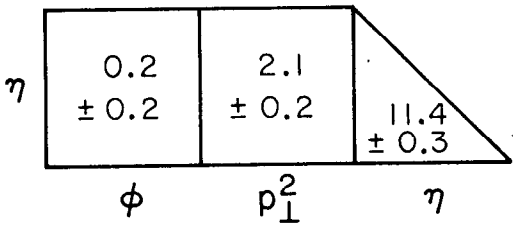
$\phi = \tan^{-1}(p_y/p_x)$ $\quad p_\perp^2 = p_x^2 + p_y^2$

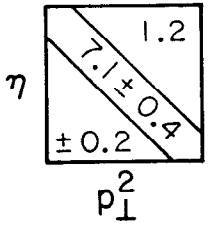
$p_{||} = p_z$ $\quad \eta = \sinh^{-1}\left(p_{||}/\sqrt{m^2 + p_\perp^2}\right)$

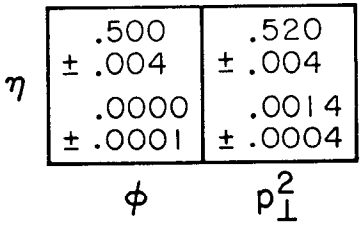












2444A1

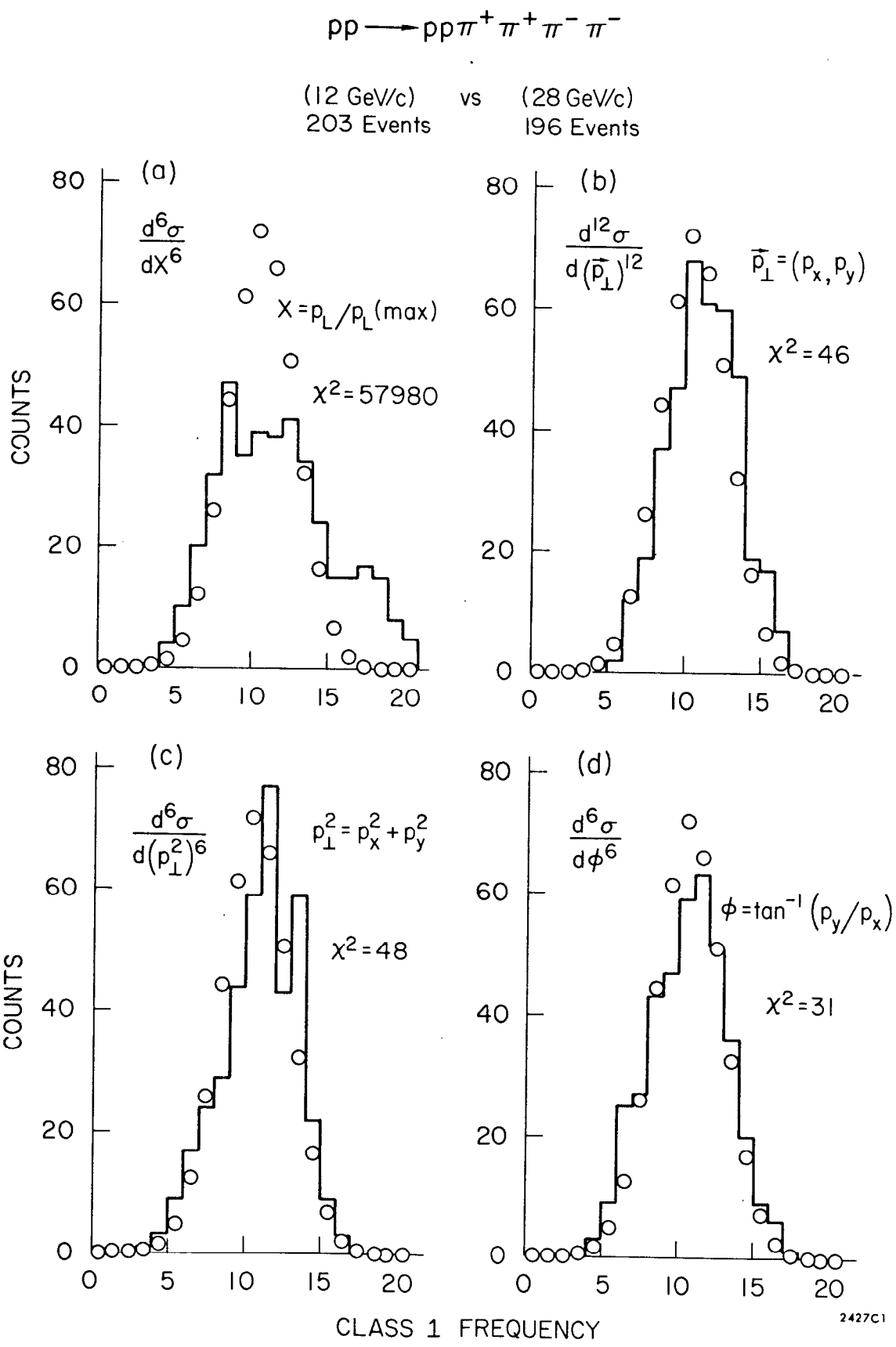

$pp \longrightarrow pp\pi^+\pi^+\pi^-\pi^-$
(12 GeV/c) 203 Events
vs
(28 GeV/c) 196 Events
(a)
$\frac{d^6\sigma}{dX^6}$
$X = p_L/p_L(max)$
$\chi^2 = 57980$
(b)
$\frac{d^{12}\sigma}{d(\vec{p}_\perp)^{12}}$
$\vec{p}_\perp = (p_x, p_y)$
$\chi^2 = 46$
(c)
$\frac{d^6\sigma}{d(p_\perp^2)^6}$
$p_\perp^2 = p_x^2 + p_y^2$
$\chi^2 = 48$
(d)
$\frac{d^6\sigma}{d\phi^6}$
$\phi = \tan^{-1}(p_y/p_x)$
$\chi^2 = 31$
COUNTS
CLASS 1 FREQUENCY
2427C1

Fig. 1

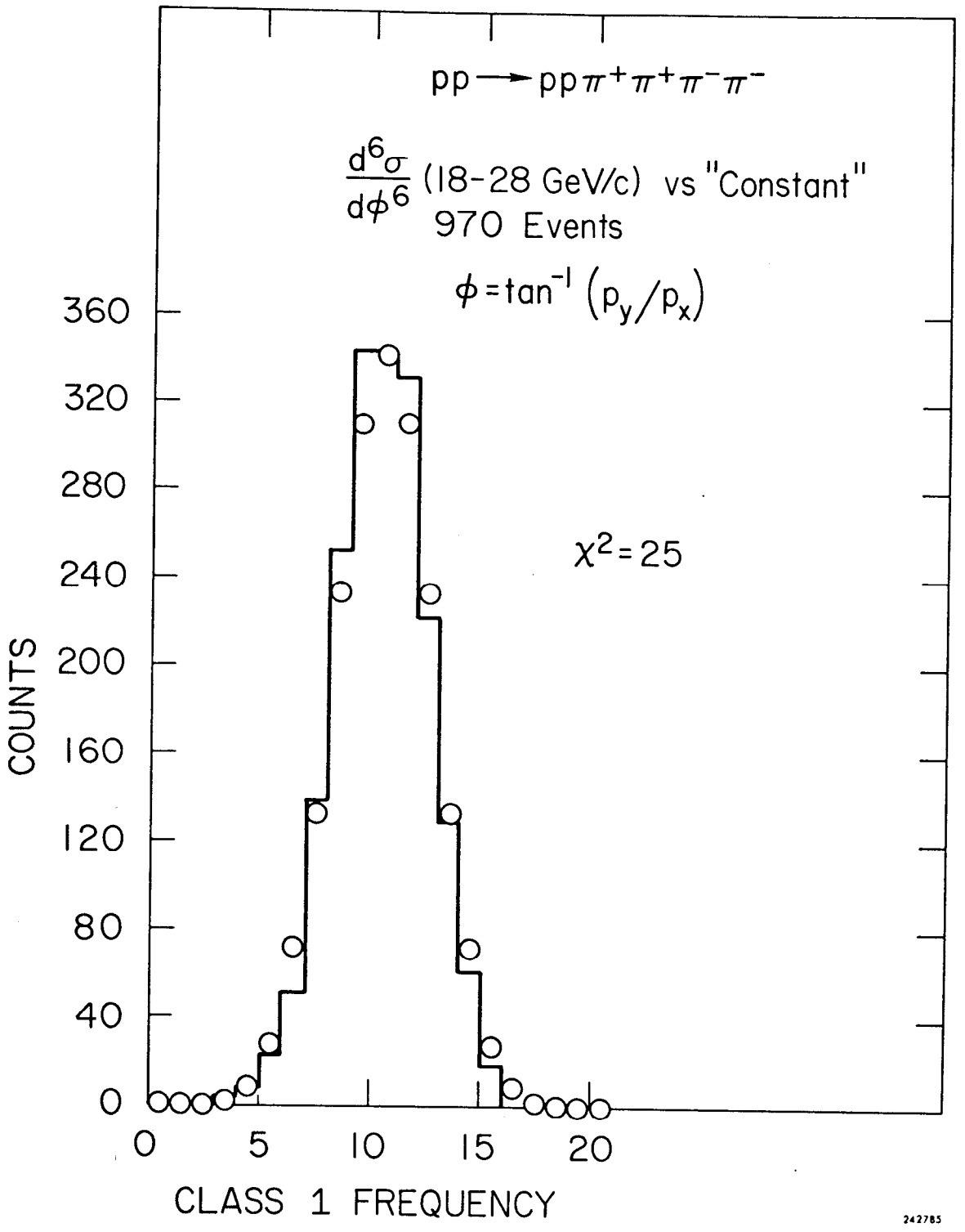

pp → pp π+ π+ π− π−
d6σ/dφ6 (18-28 GeV/c) vs "Constant"
970 Events
φ = tan−1 (py/px)
χ2 = 25
COUNTS
360
320
280
240
200
160
120
80
40
0
0
5
10
15
20
CLASS 1 FREQUENCY
242785

Fig. 2