Software Energy Compensation for the AHCAL.

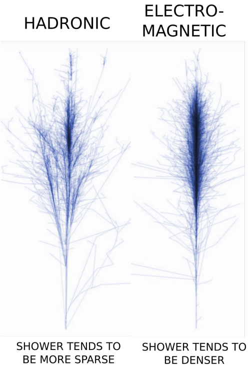
Jack Rolph
University of Hamburg
April 11, 2019

SPONSORED BY THE
Federal Ministry of Education and Research

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Software Energy Compensation Procedure

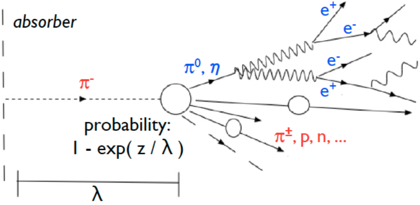
- AHCAL is a **sampling calorimeter**;
- AHCAL composed of **alternating** layers of **active and passive material**;
- **Calorimeter response different** for **electromagnetic** and **hadronic** energy deposition;
- **Pure hadronic showers** contain **more 'invisible energy'** ($\nu_{e,\mu,\tau}$, $\mu^{\pm}$, binding energy), **not seen by active layers.**



A comparison of an electromagnetic and hadronic shower
(Credit to Christian Winter's presentation!)

Software Energy Compensation Procedure

- **The $\frac{e}{h}$ fraction** of each shower **fluctuates significantly** from event to event;
- **At best**, **energy** is distributed in calorimeter as a **Gaussian** about mean of shower under various assumptions;
  - **full shower containment** w/i 5λ calorimeter;
  - **Only one type of particle observed;**
  - **Invisible energy fluctuations are distributed as Gaussian.**
- **Software compensation** possible by **weighting energy distribution** according to **Monte Carlo Simulation (MC) fit**;



A hadron may interact via EM or Strong Force.

- EM interactions initiate EM subshowers (i.e. $\frac{1}{3}$ interactions $\pi_0 \rightarrow \gamma\gamma$)
- Secondary/tertiary nuclear reactions lead to hadronic cascade
- Some observable energy lost to 'invisible energy'

Software Energy Compensation Procedure

Local Compensation

Energy is weighted on a 'hit-by-hit' basis:

$$E_{\text{Shower, Reco}} = \sum_{i=0}^{N_{\text{Hits}}} \omega(E, \vec{O}) E_{\text{Hit},i} \tag{1}$$

- $E_{\text{Shower, Reco}}$ = **Reconstructed energy** of a **shower**;
- $\omega(E, \vec{O})$ = **Weight** as a function of:
  - **particle energy** (E) **(GeV)**;
  - observables $\vec{O}$ to **describe** $\frac{e}{h}$ **proportion in shower**;
- $E_{\text{Shower Hit},i}$ = **Energy** of a **shower hit**;
- N_{Hits} = Number of hits.

Software Energy Compensation Procedure

Initial Questions

- **Which shower variables** are **most optimal** to choose **weights** in order to perform **energy compensation**?
- How does **choosing weights** based on **these variables** affect **the result of the energy compensation**?
- Is this **question** best suited for **machine learning**?

On Choosing Weights for Local Compensation: Preamble

- **Distributions of recorded observables** contain **variances between hits**;
- Technique called **Principal Component Analysis** used to **analyze and project to dimensions** of **significant variance** between hits in data;
- Allows bins to be chosen in most **statistically significant dimensions possible;**
- In this projection, hits are **maximally separated** from one another other in **variance space**.

On Choosing Weights for Local Compensation: Preamble

Definitions of Observables:

- I - **Index** of hit scintillator tile in **X direction** (1 - 24);
- J - **Index** of hit scintillator tile in **Y direction** (1 - 24);
- K - **Index** of hit scintillator tile in **Z direction** (1 - 40);
- E_{Sum} - **Energy of each process hit (MIPs) contributing to calorimeter cell**, Pedestal-subtracted and MIP-calibrated;
- $< t >$ - Reference-subtracted and NS-calibrated **time of the average hit (nanoseconds)** in calorimeter cell;.

On Choosing Weights for Local Compensation: Preamble

Standard Score (Z-Score)

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{2}$$

- z_i = **Standardized Z-score** of variable x at index i (# Standard Deviations from Mean);
- x_i = **Variable x in hit**;
- $\bar{x}$ = **Mean of variable** x over all hits of all events;
- σ_x = **Standard Deviation of variable** x over all hits over all events;

On Choosing Weights for Local Compensation: Preamble

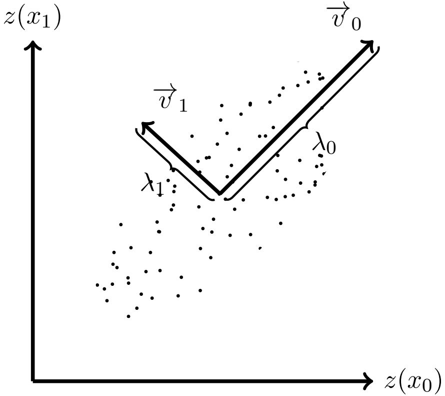
Principal Component Analysis (PCA)

- Procedure of finding the eigenvalues and eigenvectors of the covariance matrix;
- Useful for:
  - Dimensionality reduction;
  - Co-variance analysis;
  - Dimensional analysis.

$$\mathrm{Cov}(\vec{O}) = \sum_{i=0}^{dim(\vec{O})} \lambda_i \vec{V}_i \tag{3}$$

- $\mathrm{Cov}(\vec{O})$ = **Covariance Matrix** of Observables ($\vec{O}$)
- λ_i = **Eigenvalue/Length of Principal Component** in σ^2 Space (Significance)
- $\vec{V}_i$ = **Eigenvector/Direction of Principal Component in σ^2 Space** (Axis of Signficance)

On Choosing Weights for Local Compensation: Preamble



- $z(x_0)$ - Z-Score of Variable x_0
- $z(x_1)$ - Z-Score of Variable x_1
- $\vec{v}_0$ - Principal Component 0
  (Eigenvector of Greatest Variance)
- $\vec{v}_1$ - Principal Component 1
  (Eigenvector of Next Greatest Variance)
- λ_0 - Significance of Principal Component 0 (Unnormalized)
  (Eigenvalue of Greatest Variance)
- λ_1 - Significance of Principal Component 1 (Unnormalized)
  (Eigenvalue of Next Greatest Variance)

Convention:

- PC0 refers to normalized $\vec{v}_0$, the most significant principal component;
- PC1 refers to normalized $\vec{v}_1$, the most significant principal component etc.;

On Choosing Weights for Local Compensation: Method

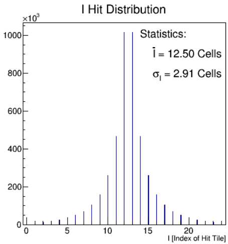
- 20,000 π^- events were produced using Geant4 Simulation of AHCAL, with each contributing process and energy deposition recorded;
  - Energy:
    - Poisson Smearing of 14 pixels applied;
    - 0.5 MIP cut on Hit Energy;
  - Time:
    - Gaussian Smearing of 5 ns applied.
  - Hits:
    - Reject # Hits < 50 ($\mu^\pm$/Punch-through π^- Cut)

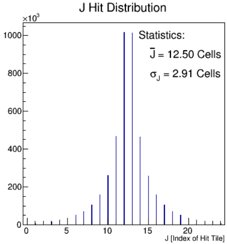
(Credit to Eldwan Brianne for the simulation!)

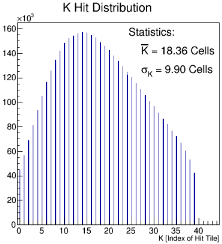
- Standard scores (z) were calculated in I, J, K, E_{Sum} and $< t >$;

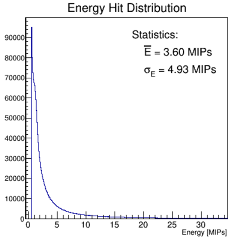
- PC dimensions/significances calculated by eigendecomposing the covariance matrix of the calculated z-scores.

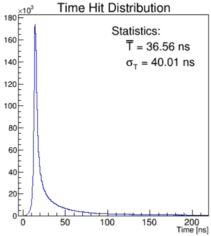
On Choosing Weights: Cell Energy Deposition Results





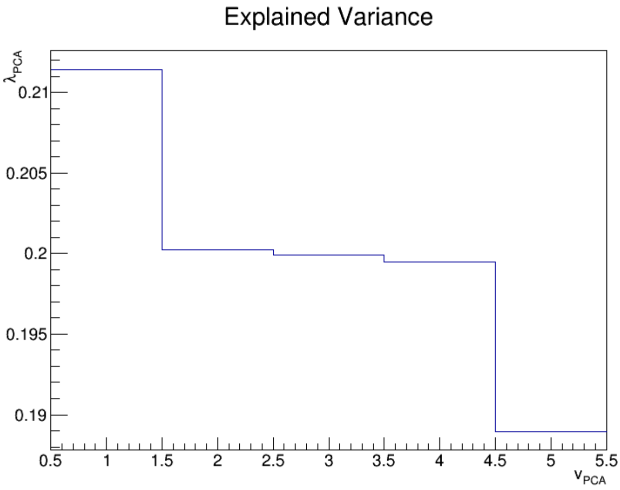






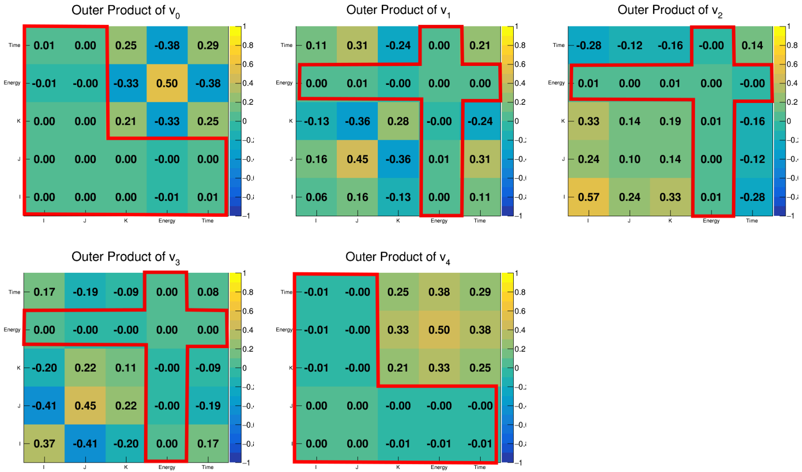
Distributions of different observables from MC with means and variances.

On Choosing Weights: Cell Energy Deposition Results



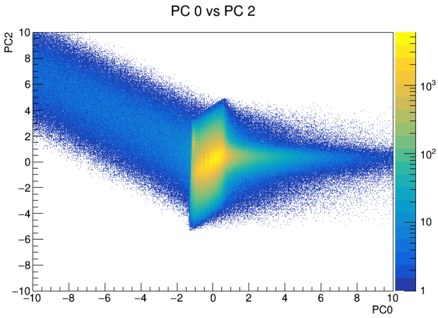
Explained variance as a function of PC in 20,000 Event 20 GeV π^- MC.

On Choosing Weights: Cell Energy Deposition Results

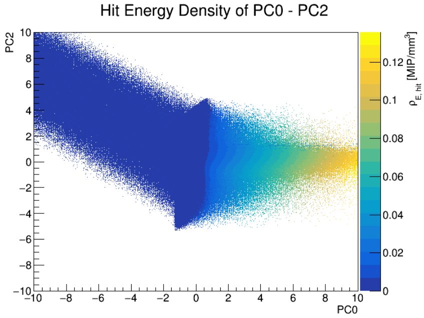


Outer products ($|v_i\rangle\langle v_i|$) as a function of input parameters in 20,000 Event 20 GeV π^- MC. Red lines show negligible components.

On Choosing Weights: Cell Energy Deposition Results

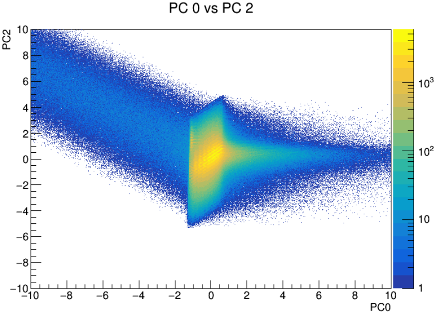


Distribution of hits projected on PC0 and PC2 for 20 GeV π^- simulation

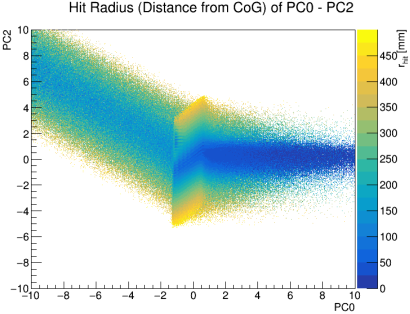


Hit energy density projected on PC0 and PC2 for 20 GeV π^- simulation

On Choosing Weights: Cell Energy Deposition Results

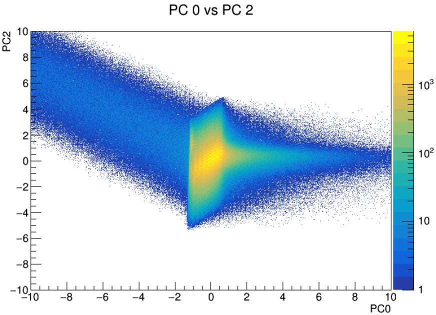


Distribution of hits projected on PC0 and PC2 for 20 GeV π^- simulation

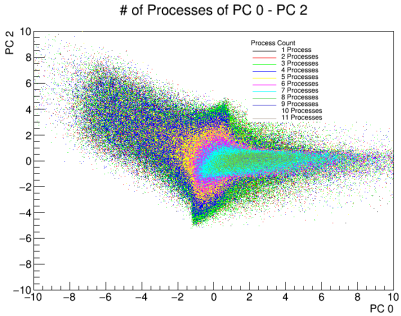


Hit radius projected on PC0 and PC2 for 20 GeV π^- simulation

On Choosing Weights: Cell Energy Deposition Results



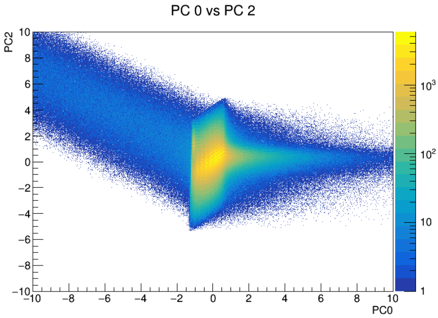
Distribution of hits projected on PC0 and PC2 for 20 GeV π^- simulation



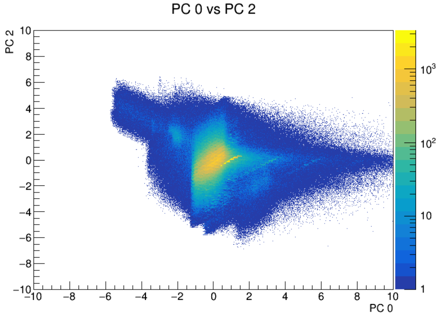
Distribution of # processes responsible for cell energy deposition projected on PC0 and PC2 for 20 GeV π^- simulation

Deposition Processes in Simulation include: Ionization, Multiple Scattering, Compton Scattering, Coloumb Scattering, Photoelectric Effect, Bremsstrahlung, Inelastic/Elastic Scattering, Neutron Capture, Positron Annihilation etc.

On Choosing Weights: Cell Energy Deposition Results



Distribution of hits projected on PC0 and PC2 for 20 GeV π^- simulation



PC0 - PC2 distribution in 20 GeV π^- Runs from June SPS Testbeam using MC Eigenvectors, Averages and Variances (Runs 061269 - 73)

On Choosing Weights for Local Compensation: Choosing Bins

Current Binning Argument:

interaction processes (and other variables) seem to vary in pattern space in a way that describes the shower core

$\downarrow$

Ansatz: Shower core likely to contain highest # processes, also a discrete, projection-independent variable ($\therefore$ useful to seperate/classify co-ordinates);

$\downarrow$

EM-dominated showers $\rightarrow$ tighter, denser shower core
Hadronic-dominated showers $\rightarrow$ looser, more sparse shower core;

$\downarrow$

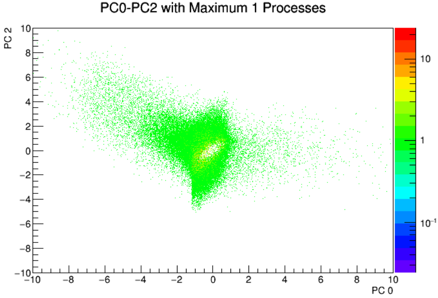
Weights should reflect 'core' and 'periphery' hits differently

On Choosing Weights for Local Compensation: Choosing Bins

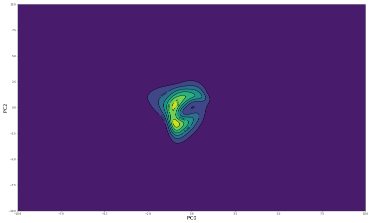
Cut Extraction Procedure:

- **Check max. # of processes taking place in pattern space**, and the # hits of this bin.
- **Fit PDF to each distribution of max. # of hits** using **2D Kernel Density Estimation**;
- **Combine and weight PDFs** in a manner that provides appropriate trade-off between **classification error, # bins and # hits found bins**;
- Extract cut from Hessian matrix of PDF error of final PDF combination.

On Choosing Weights for Local Compensation: Choosing Bins

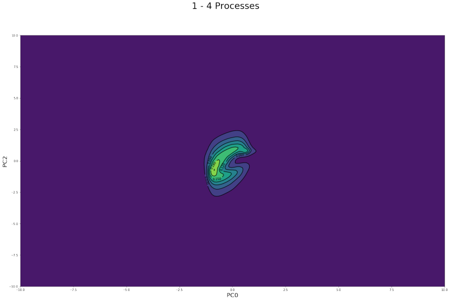


Hit distribution for co-ordinates with a possible maximum of one process taking place there,

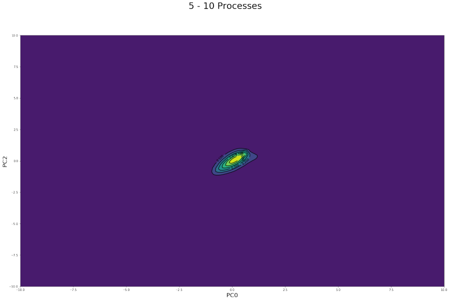


Fitted PDF distribution for hit cells with a possible maximum of one process taking place there.

On Choosing Weights for Local Compensation: Choosing Bins



Weighted PDF for 1-4 processes bin.



Weighted PDF for 5-10 processes bin.

On Choosing Weights for Local Compensation: Choosing Bins

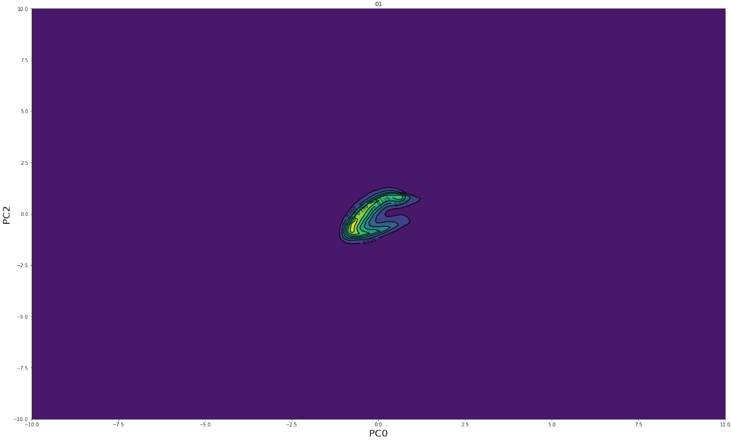
Probability Overlap

$$P_{\text{Error, 12}} = \sum_{i=0}^{N=\text{Bins, x}} \sum_{j=0}^{N=\text{Bins, y}} \min\left(\text{PDF}_1(i,j), \text{PDF}_2(i,j)\right) \tag{4}$$

- $P_{\text{Error, 12}}$ = Probability Overlap between PDFs 1 and 2
- i, j = x, y index
- $\text{PDF}_1(i,j)$ = PDF_1 at index (i,j)
- $\text{PDF}_2(i,j)$ = PDF_2 at index (i,j)

On Choosing Weights for Local Compensation: Choosing Bins

P_{error}



Overlap PDF between bins for 1-4 processes and 5-10 processes.
This constitutes a **38.34 % classification error**.

On Choosing Weights for Local Compensation: Choosing Bins

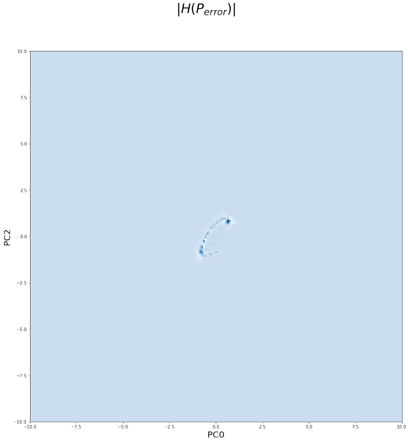
Hessian Matrix

$$\mathrm{Hess}(P_{Error}(i,j)_{k,l}) \equiv \frac{\partial^2 P_{Error}(i,j)}{\partial x_k \partial x_l} \tag{5a}$$

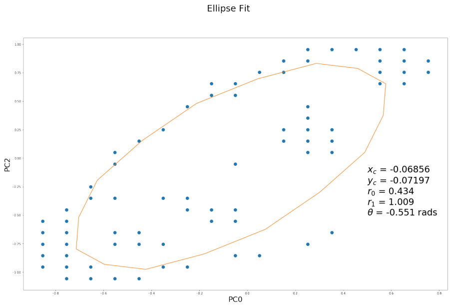
$$P_{Error}(i,j) \begin{cases} \text{Local Maximum} & \text{if } \det \mathrm{Hess}(P_{Error}(i,j)_{k,l}) > 0 \\ \text{Turning Point} & \text{if } \det \mathrm{Hess}(P_{Error}(i,j)_{k,l}) = 0 \\ \text{Local Minimum} & \text{if } \det \mathrm{Hess}(P_{Error}(i,j)_{k,l}) < 0 \end{cases} \tag{5b}$$

- i, j, k, l = x, y, coordinate k, coordinate l index;
- $P_{\text{Error}(i,j)}$ = Probability Overlap between PDFs;

On Choosing Weights for Local Compensation: Choosing Bins

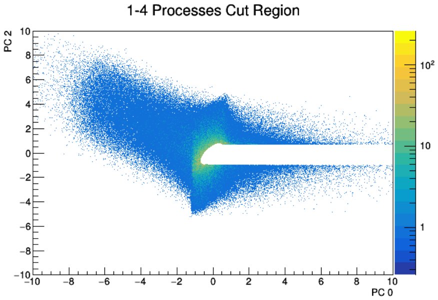


Determinant of Hessian of P_{Error}

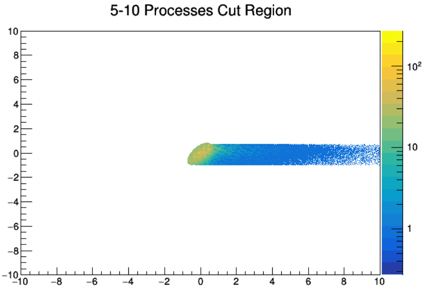


Ellipse fit of $|H(P_{error})|$

On Choosing Weights for Local Compensation: Choosing Bins

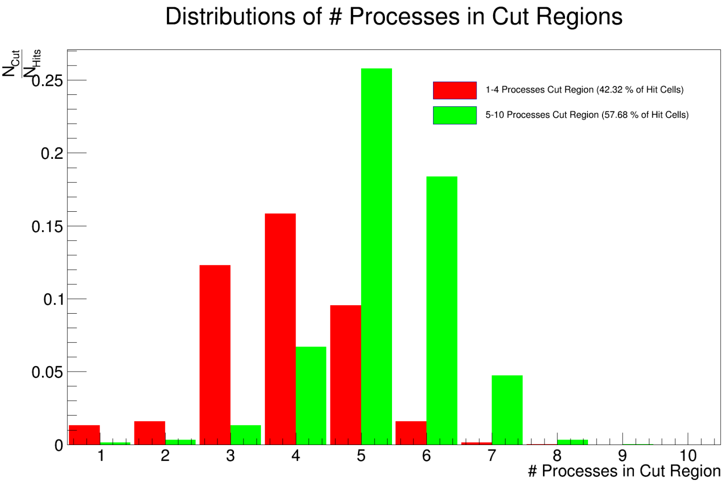


Cut distribution for the 1-4 processes bin.



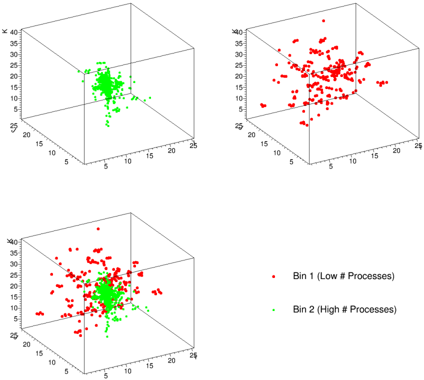
Cut distribution for the 5-10 processes bin.

On Choosing Weights for Local Compensation: Choosing Bins

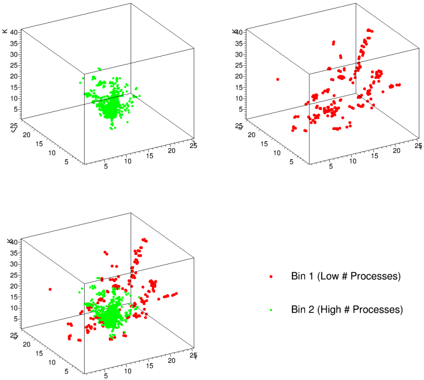


Final cut distribution.

On Choosing Weights for Local Compensation: Choosing Bins



Event display for a random simulated event in I-J-K space.

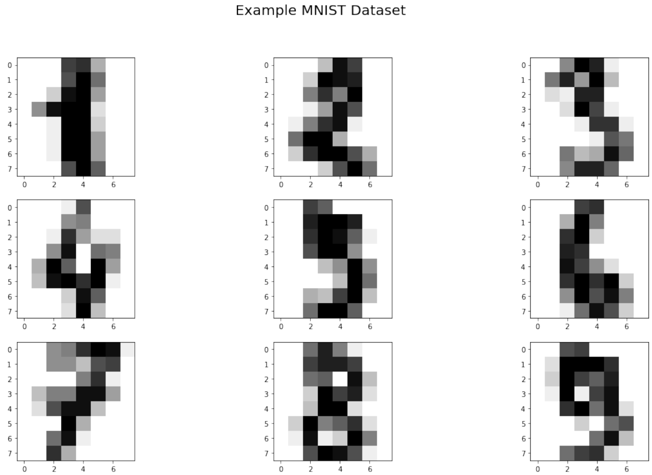


Event display for another random simulated event in I-J-K space.

Summary

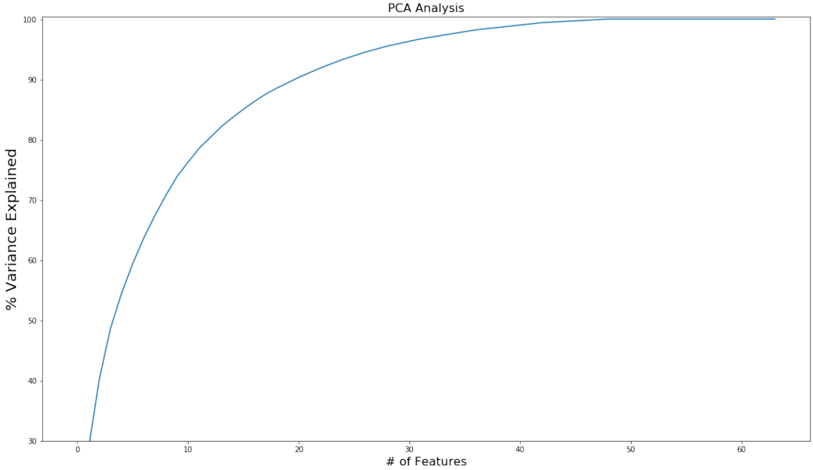
- **Dimensions of most significance to dataset were found;**
- **Hits for 20 GeV π^- vary most** according to two dimensions:
  - energy deposited/depth evolution
  - volume evolution;
- **Cut** has strong relationship with **separating 'core hits' from 'periphery' hits: a low hit energy event in the core can be weighted differently to on the periphery.**
- **Potential link to $\frac{e}{h}$ in shower; denser/more sparse showers will likely have different distributions of hits in each bin.**
- Can we extend this method using E_{Reco} as an input?
- Next $\rightarrow$ perform energy weighting procedure

Backup



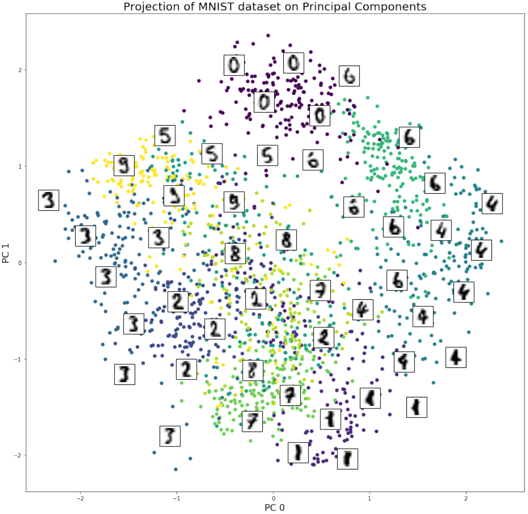
Handwritten digits 1-9 of MNIST dataset (8×8 pixels = 64 dimensions)

Backup



Cumulative variance of MNIST dataset as a function of number of dimensions.

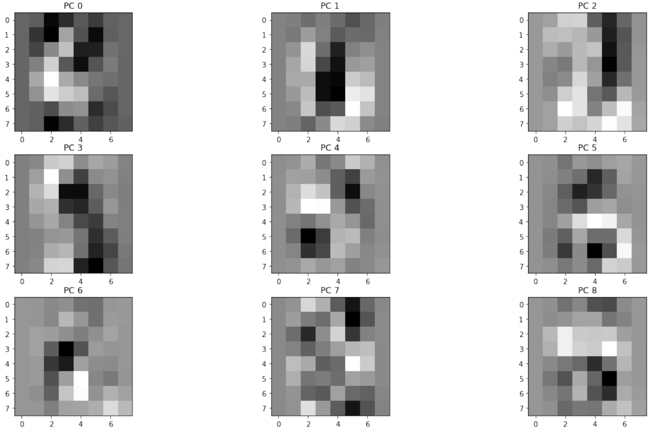
Backup



Distribution of MNIST dataset containing 28.5% of explained variance.

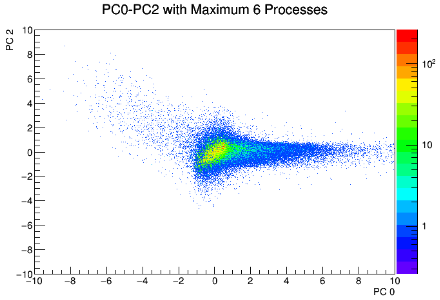
Backup

Principal Components 0-9 (70.8% explained variance)

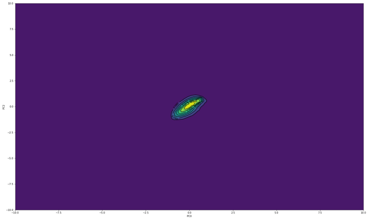


Eigenvectors of MNIST dataset containing 70.8% of variance data.

Backup

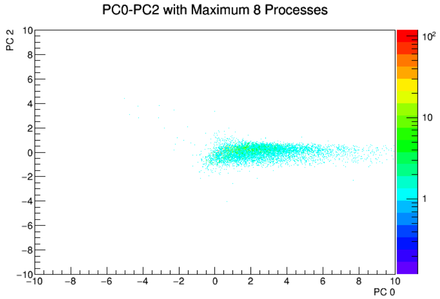


Hit distribution for co-ordinates with a possible maximum of four processes taking place there,

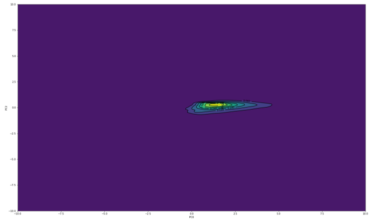


Fitted PDF distribution for hit cells with a possible maximum of four processes taking place there.

Backup



Hit distribution for co-ordinates with a possible maximum of seven processes taking place there,



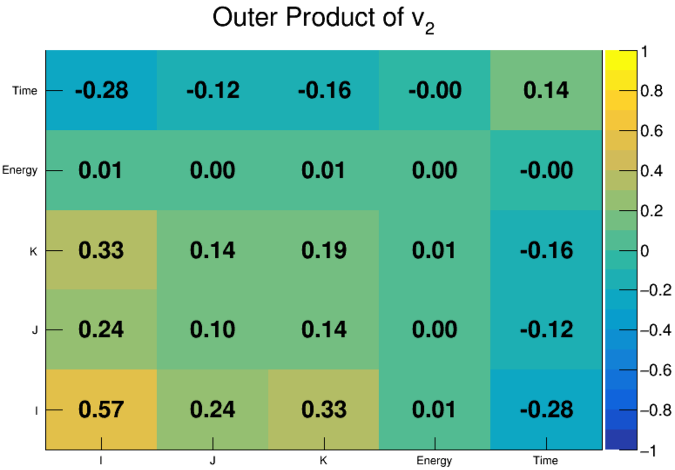
Fitted PDF distribution for hit cells with a possible maximum of seven processes taking place there.

Outer Product of PC 0 (Projection Operator)

Backup

Outer Product of PC 1 (Projection Operator)

Backup



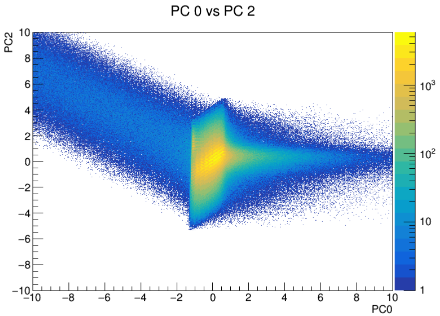
Outer Product of PC 2 (Projection Operator)

Outer Product of PC 3 (Projection Operator)

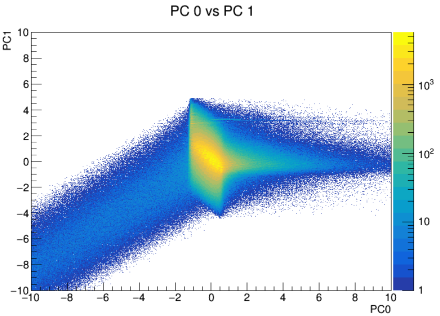
Backup

Outer Product of PC 4 (Projection Operator)

Backup

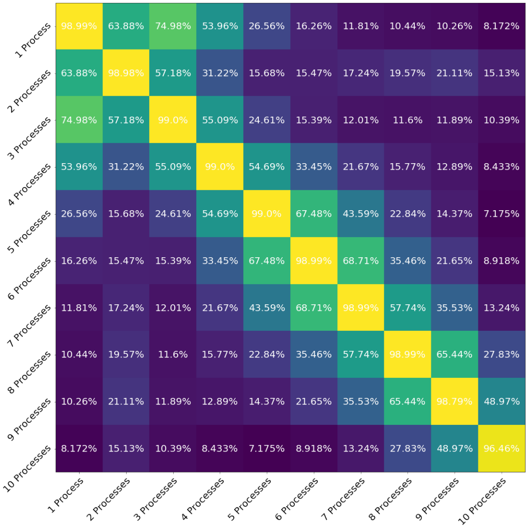


Distribution of hits projected on PC0 and PC2 for 20 GeV π^- simulation



PC0 - PC1 distribution in 20 GeV π^- simulation

Backup



Error matrix of overlaps between different PDFs.

Backup

Definition of Histograms of Maximum # Processes

Summary: make a set of histograms containing only the maximum number of possible processes causing a hit at a given co-ordinate;

$$f(i,j,p)\begin{cases} H_{p+1}(i,j) = H_p(i,j) + H_{p+1}(i,j) & \\ H_p(i,j) = 0, & \text{if } H_{p+1}(i,j) > 0 \\ H_{p+1}(i,j) = 0 & \text{otherwise} \end{cases} \tag{6}$$

- i, j, p = x, y, process index
- $H_{p+1}(i,j)$ = Histogram of next process distribution at bin (i,j).
- $H_p(i,j)$ = Histogram of current process distribution at bin (i,j).

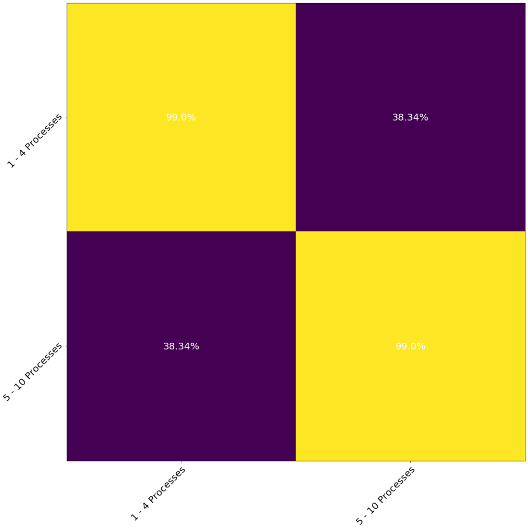
Weighted Sum

$$P_{\mathrm{W}} = \sum_{p=0}^{N=\text{PDFS}} \sum_{i=0}^{N=\text{Bins, x}} \sum_{j=0}^{N=\text{Bins, y}} \frac{N_p}{\sum_{p=0}^{N=\text{PDFS}} N_p} \mathrm{PDF}_p(i,j) \tag{7a}$$

$$N_p = \sum_{i=0}^{N=\text{Bins, x}} \sum_{j=0}^{N=\text{Bins, y}} H_p(i,j) \tag{7b}$$

- P_{W} = Weighted PDF;
- i, j, p = x, y, PDFs in Chosen Binning index;
- N_p = Total number of hit cells in histogram H_p;
- $\sum_{p=0}^{N=\text{PDFS}} N_p$ = Sum over the total number of hit cells in bin.
- $\mathrm{PDF}_p(i,j)$ = PDF value of current process distribution at bin (i,j).

Backup



Error matrix of weighted PDFs summed in bins for 1-4 processes and 5-10 processes.

Backup

Main Observations of PCs:

- **Only two PC dimensions have anything to do with energy.**
- **PC 0 and PC 4** seem to **describe complex energy-depth evolution co-related space** between K, E and t (PC 4 similar to PC 0).
- **PCs 1, 2 and 3** seem to **describe a complex type of volume evolution co-related space** for I, J, K and t.
- **PC 1, PC 2, PC 3 very similar; PC0 and PC2 chosen to analyze, as PC2 has a continuous distribution (no 'discrete peaks' from artifacts of I,J,K) .**