Rheinisch-Westfälische Technische Hochschule Aachen

Master Thesis

Denoising Methods for Multi-Dimensional Photoemission Spectroscopy

SUBMITTED BY

Muhammad Zain Sohail

Rheinisch-Westfälische Technische Hochschule Aachen

Christian-Albrechts-Universität zu Kiel

Deutsches Elektronen Synchrotron

SUPERVISORS

Prof. Dr. Benjamin Berkels

Rheinisch-Westfälische Technische Hochschule Aachen

Prof. Dr. Kai Rossnagel

Christian-Albrechts-Universität zu Kiel

Aachen, June 2024

Summary

In the realm of photoemission spectroscopy, the exploration of large multi-dimensional phase spaces necessitates time-intensive data acquisition to ensure statistical robustness. Despite the unparalleled capabilities of free-electron lasers (FELs), in peak brightness and ultra- short pulsed X-rays, the limitations of low repetition rates prolong the data acquisition process. This impedes the agility of decision making that could otherwise enhance experimental results in the limited and valuable beamtime. By employing denoising strategies to mitigate noise while preserving intrinsic information, our proposed approach aims to streamline the data acquisition process, and effectively manage the escalating size and complexity of multi-dimensional photoemission data.

Contents

List of Acronyms

BAM beam arrival monitor. 23

DESY Deutsches Elektronen-Synchrotron. 3
DLD delay line detector. 23

FEL free-electron laser. ix

GMD gas monitor detector. 23

HDF5 hierarchical data format version 5. 23

OpenCOMPES open community of multidimensional photoemission spectroscopy. 23

PES photoemission/photoelectron spectroscopy. 3

SASE self-amplified spontaneous emission. ix
SED Single Event DataFrame. 23

Glossary

Free-Electron Laser is an x-ray radiation source; fundamentally comprising of a linear particle accelerator and an Undulator (or a series of undulators). The accelerator produces a bunched electron beam similar to that of a synchrotron, which can be compressed to reach ultrashort pulse duration (femtosecond) with peak brightness many orders of magnitude above synchrotrons. Since the electrons move in a vacuum, it is termed Free-Electron in comparison to traditional lasers which are bound by the materials energy levels. Whereas, it is called a Laser due to there being light amplification and the shared properties with traditional optical lasers such as high pulse energy and being coherent. Taken from [1]. 1

Microbunch Microbunches are produced by the interaction between the oscillating electrons in the undulator and the radiation that they produce (due to the oscillatory acceleration) leads to periodic longitudinal density modulation known as Microbunching. The in-phase emitted radition adds coherently, increasing intensity and enhancing microbunching. Adapted from [**?**]. ix

Pulse Also known as *Microbunch*. Each Train contains about 500 pulses produced from the self-amplified spontaneous emission (SASE) process (See Self-Amplified Spontaneous Emission). These are which are used as a secondary index for the data reduction process. 1, 23

Self-Amplified Spontaneous Emission is a process where the electron beam in the accelerator, when passing through an undulator, starts emitting radiation due to acceleration. The interaction between the emitted radiation and the charge distribution leads to microbunching. These microbunches emit radiation coherently, leading to the intense, coherent radition, characteristic of an free-electron laser (FEL). For more information see [**?**]. ix, 1

Train Also known as *Macrobunch*. A train represents a group of closely spaced electron bunches produced and accelerated by the FEL (or more generally, any accelerator). Each train is associated with a unique identifier called `trainId`, which is used as the primary index for much of the data reduction process. ix, 1, 23

Undulator Magnets arranged periodically to produce a periodic magnetic field. It is used to produce coherent radiation by accelerating electrons through it. ix, 1

Introduction

Conjecturing hypotheses based on observations has always been an important aspect of the scientific methodology. Especially in natural sciences, where the aim is to describe natural phenomena, the role of observations can not be overstated.

After formulating a hypothesis, an experiment is designed to test it. From the evidence gathered through the experiment and through the aid of statistical principles, these hypotheses can be accepted (or rejected) with a defined level of confidence.

Hence, science is inherently linked to data. Data science helps us to formally handle this data, regardless of the domain. We want testable outcomes

With the dimensionality and size of data increasing, more sophisticated tools are necessary.

Especially in the field of neuroscience where experiements can not be performed easily or directly, a lot of statisitcal tools are employed.

In modern days, a paradigm shift has occurred where instead of trying to explicitly model the system we are trying to learn about, has also brought revolutions such as machine learning where we don't need to know the model itself and basically make the machine learn the non-linear model. This doesn't explain the process but allows us to perform for example inference.

Expensive aspect! People don't use math methods first. Then we go into denoising

With the necessity to acquire data in so many dimensions, because we have low electron counts.

Free-Electron Laser Undulator Train Pulse Self-Amplified Spontaneous Emission

Photoemission Spectroscopy

In the seminal paper by Einstein [2], that laid foundations to Quantum Mechanics, Einstein postulated that light is made of discrete quanta of energy $E = h\nu$ to explain the observations by Hertz and J.J. Thompson; explaining the photoelectric effect. The effect can be described by the Equation 2.1 where E_e the emitted kinetic energy, h is the plank's constant, ν the frequency of the incoming photon and ϕ the material-specific work function (also known as binding energy).

$$E_e = h\nu - \phi \tag{2.1}$$

The equation describes how incident photons on a surface eject photoelectrons, provided the photon energy $h\nu$ exceeds ϕ. This also highlights that the emitted kinetic energy E_e does not depend on the photon flux (photon counts per second). However, the flux increases the total amount of electrons released from the material.

It is then apparent that the binding energy of electrons can be found by irradiating light onto the material and measuring the E_e of photoelectrons. photoemission/photoelectron spectroscopy (PES), is exactly such a technique that leverages this principle to probe the electronic structure of materials.

2.1 Where PES is performed

Table top, Synchrotrons, FELs. Time resolved PES and how it relates to acquisition times.

2.2 HEXTOF Setup at FLASH

Deutsches Elektronen-Synchrotron (DESY) is a national

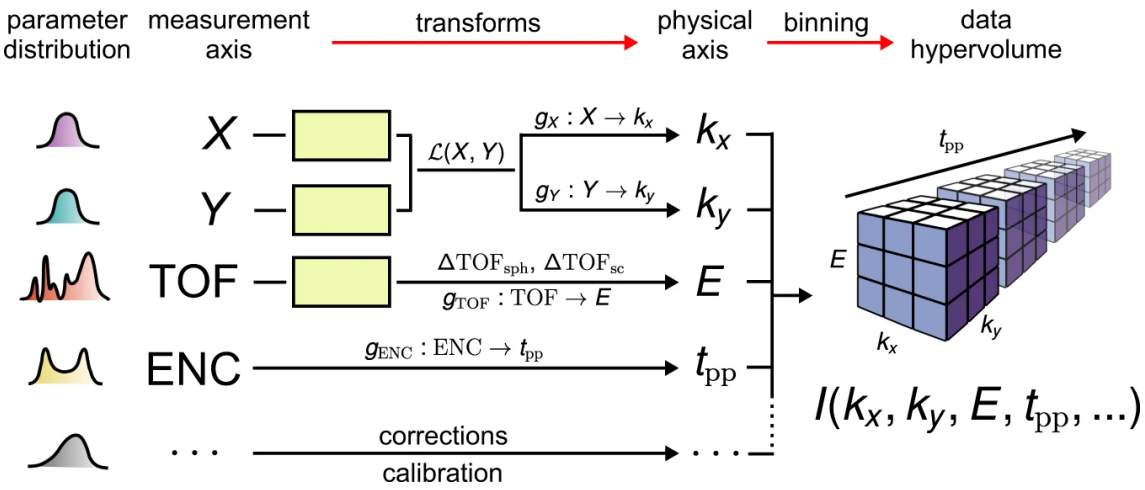

Figure 2.1: MPES taken from [3]

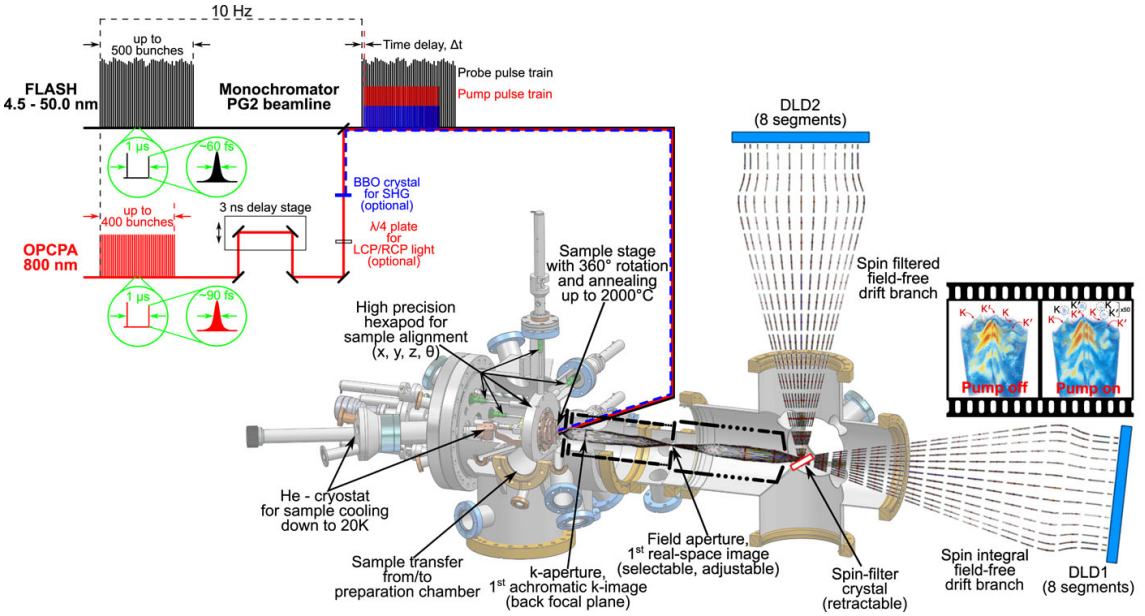

Figure 2.2: HEXTOF taken from [4]

2.3 Describing the data GdW, WSe2, GrIr, and new one

2.4 Creating the dataset Corrections Calibrations etc

2.5 Using CLT

The **Central Limit Theorem** states that for a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with mean μ and variance σ^2, the normalized sum of these variables approaches a standard normal distribution as n tends to infinity:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty.$$

This convergence is in distribution.

Imaging

3.1 Binning as means of Imaging

digital imaging from lecture sampling etc Binning is a way to find underlying distribution.

3.2 SASE fluctuations

Check the dld detector electron copies causing problem

3.3 Poisson process checking

There's a lot of parameters that need to be tested to determine what sort of counting statistics the dataset has.

Controls for the test: lets see

- Total time being looked at (like 1000 s or 20 hours)
  - distribution might change due to overtime FEL intensity changes
- Time bins being used (like 2 s vs 20 s and so on).
  - Seems like distribution changes based on that too
- Looking at individual pixels on X and Y
- Looking at Energy axis as it behaves weirder
- Looking at a larger region in X and Y
  - should follow same statisitics as single pixels
- Check after removing correlated electrons within each pulse
  - It is possible that electrons are correlated between different pulses because the time delay is long enough. But seems highly unlikely!
  -

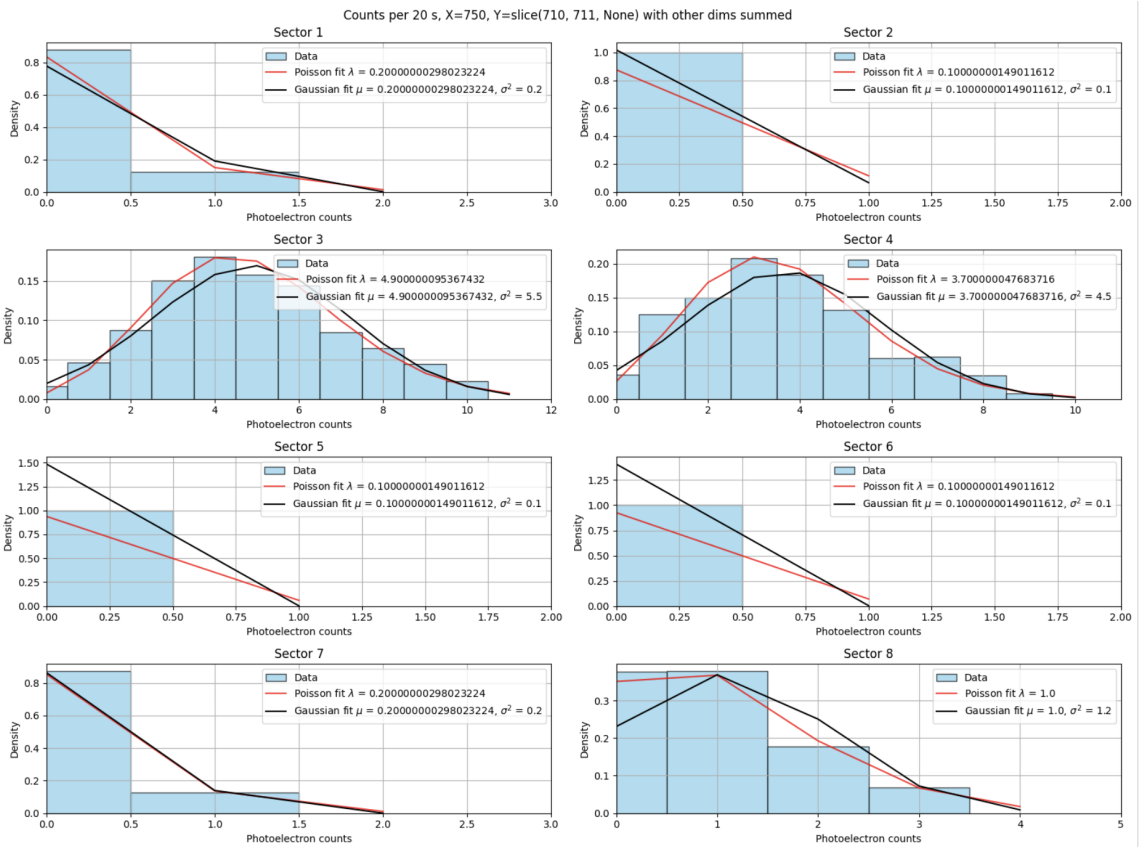

Figure 3.1: Enter Caption

3.3.1 Before filtering

3.3.2 After filtering

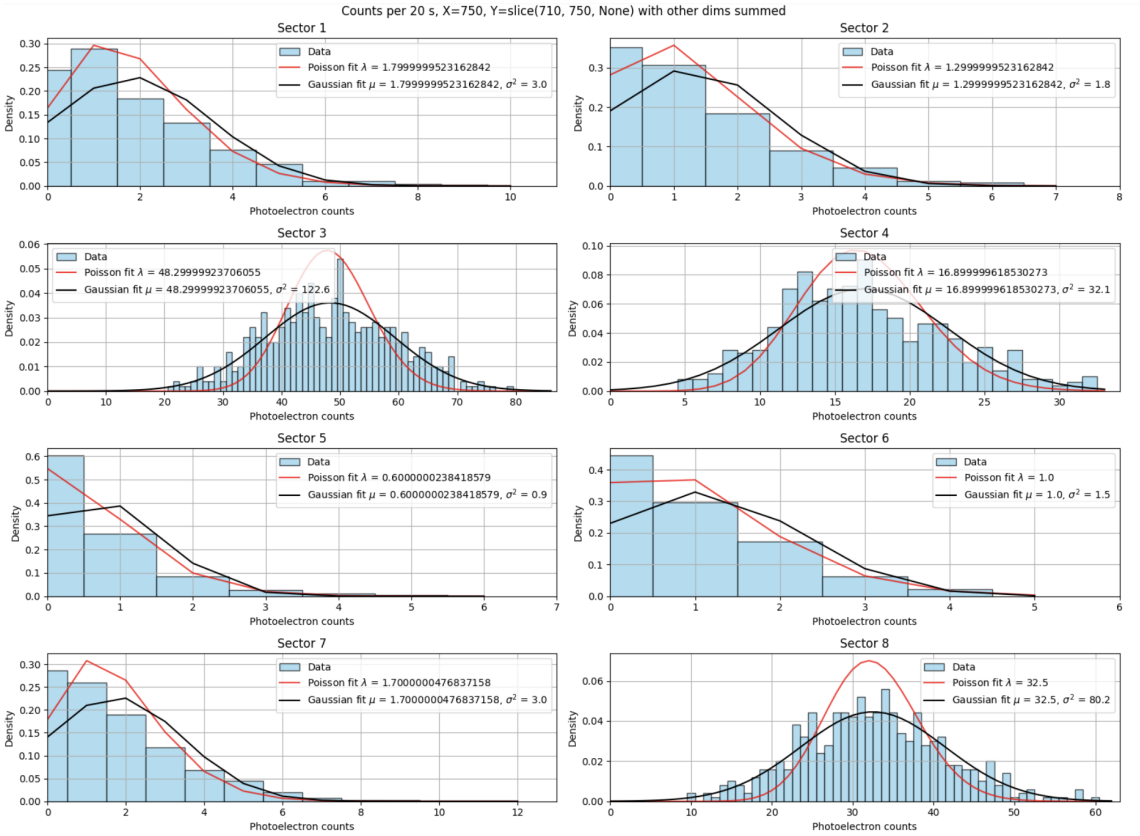

Figure 3.2: Image over 40 pixels and summed over energy

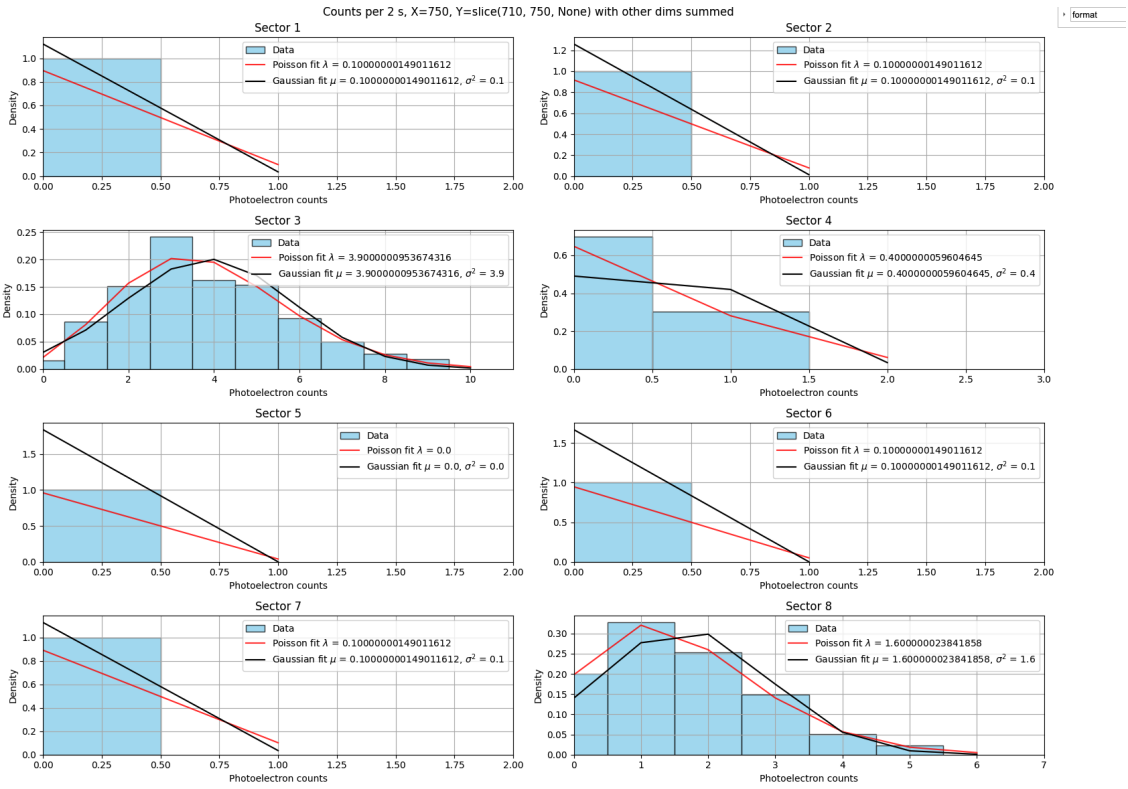

Figure 3.3: Data is filtered here with the KNN and looking at small region

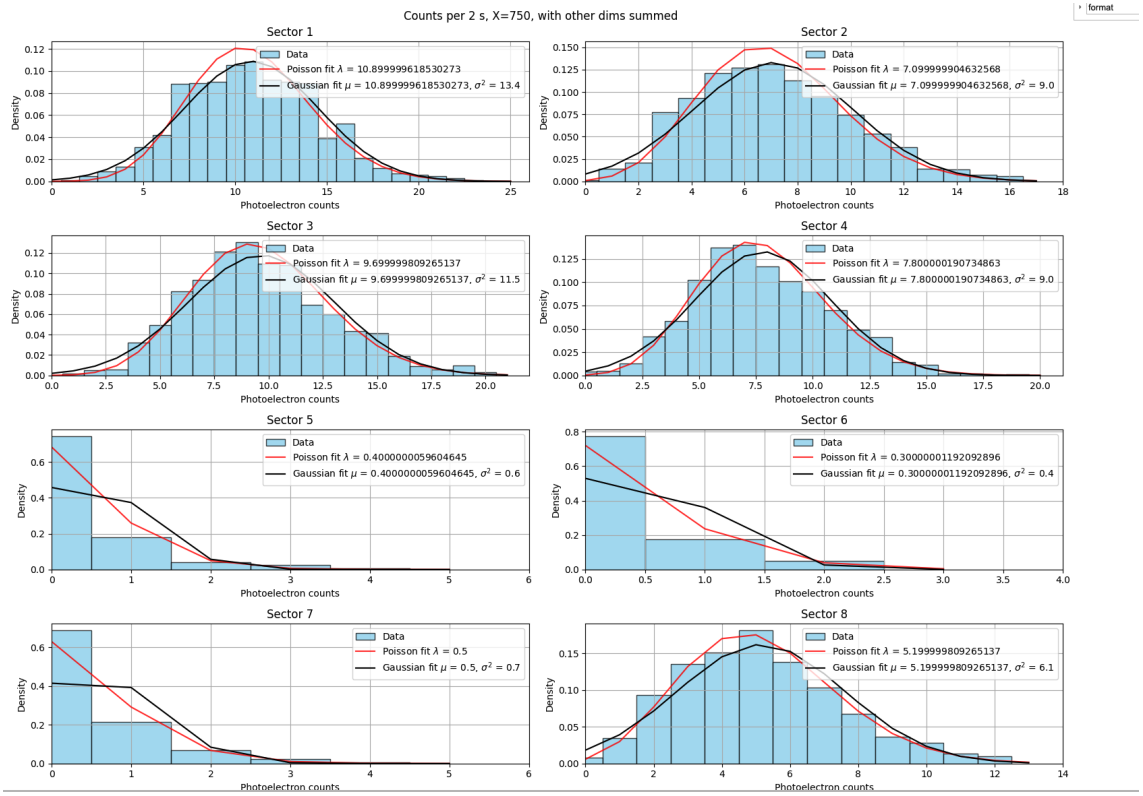

Figure 3.4: Enter Caption

Image Reconstruction

Rather than calling it denoising, better word in image reconstruction because Image Reconstruction:

Purpose: To reconstruct an image from incomplete, noisy, or indirect measurements. This is often used in medical imaging (e.g., MRI, CT scans), computational photography, and computer vision applications.

Reconstruction involves generating a complete image from partial or indirect data, which can include denoising and deblurring as sub-tasks.

4.1 Address reconstruction/denoising schemes

VST with BM3D: BM3D uses collaborative filtering, which is also used in recommender systems [citation need] PnP iterative stuff maybe non local sometime UNET noise2noise

4.2 What is poisson

Taken from [5] which cites [6]

4.3 Comparsion of bm3d

with and without anscombe on different datasets from flash, lab and fhi Testing s

4.4 Algorithms

ERM Deep learning is just linear sepeartor problem with a non linear function applied to it like RELU

Poisson data occur in all imaging processes where images are obtained by means of the count of particles, in general photons, arriving in the image domain. In the case of radioactive decay, fluorescence emission or similar phenomena, the arrival of particles is described by a *Poisson process*, i.e. a continuous stochastic process that is a collection of independent random variables $\{N(t); t \geqslant 0\}$, where $N(t)$ is the number of particles arrived up to time t. The number of particles arriving within a given time interval T is a random variable (r.v.) with a *Poisson distribution* [52, 118], i.e. the probability of receiving n particles is given by

$$p(n) = \frac{e^{-\lambda}\lambda^n}{n!}, \qquad n = 0, 1, 2, \ldots, \tag{1}$$

where λ, proportional to T, is the expected value of the counts. This statistical model is appropriate to describe data acquired in fluorescence microscopy, emission tomography, optical/infrared astronomy, etc. Even if the energy of the photons (hence their wavelength) is different in the different applications, the statistics of the data is the same. The imaging system, however, significantly varies from an application to another. It is described by a sparse matrix in tomography, and by a convolution matrix in microscopy and astronomy.

Figure 4.1: Enter Caption

Conclusion and Outlook

We use SASE currently but with seeded FEL, might be more robust poisson statistics. Maybe even sub-poissonian.

Bibliography

[1] M. Z. Sohail, "Ultrafast dynamic studies in spin-filter material Europium Oxide with Resonant Inelastic X-ray Scattering," Bachelor's Thesis, Carl von Ossietzky Universität Oldenburg, 2021.

[2] A. Einstein, "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," *Annalen der Physik*, vol. 322, no. 6, pp. 132–148, 1905, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19053220607. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19053220607

[3] R. P. Xian, Y. Acremann, S. Y. Agustsson, M. Dendzik, K. Bühlmann, D. Curcio, D. Kutnyakhov, F. Pressacco, M. Heber, S. Dong, T. Pincelli, J. Demsar, W. Wurth, P. Hofmann, M. Wolf, M. Scheidgen, L. Rettig, and R. Ernstorfer, "An open-source, end-to-end workflow for multidimensional photoemission spectroscopy," *Scientific Data*, vol. 7, no. 1, p. 442, Dec. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-020-00769-8

[4] D. Kutnyakhov, R. P. Xian, M. Dendzik, M. Heber, F. Pressacco, S. Y. Agustsson, L. Wenthaus, H. Meyer, S. Gieschen, G. Mercurio, A. Benz, K. Bühlman, S. Däster, R. Gort, D. Curcio, K. Volckaert, M. Bianchi, C. Sanders, J. A. Miwa, S. Ulstrup, A. Oelsner, C. Tusche, Y.-J. Chen, D. Vasilyev, K. Medjanik, G. Brenner, S. Dziarzhytski, H. Redlin, B. Manschwetus, S. Dong, J. Hauer, L. Rettig, F. Diekmann, K. Rossnagel, J. Demsar, H.-J. Elmers, P. Hofmann, R. Ernstorfer, G. Schönhense, Y. Acremann, and W. Wurth, "Time- and momentum-resolved photoemission studies using time-of-flight momentum microscopy at a free-electron laser," *Review of Scientific Instruments*, vol. 91, no. 1, p. 013109, Jan. 2020. [Online]. Available: https://pubs.aip.org/rsi/article/91/1/013109/1018655/Time-and-momentum-resolved-photoemission-studies

[5] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, "Image deblurring with Poisson data: from cells to galaxies," *Inverse Problems*, vol. 25, no. 12, p. 123006, Dec. 2009. [Online]. Available: https://iopscience.iop.org/article/10.1088/0266-5611/25/12/123006

[6] W. Feller, *An introduction to probability theory and its applications*, third ed.

rev ed., ser. Wiley series in probability and mathematical statistics. New York Chichester Brisbane [etc.]: J. Wiley, 1968.

Acknowledgements

Thank Dima especially Dataset providers Thank Lorenz Kruger, Martin and others in group (Lukas Weigand). Maxwell of course

Contributions

A preliminary analysis during the early works of this thesis titled *Efficient Data Acquisition in Multi-Dimensional Photoemission Spectroscopy using Denoising*, was presented, in the form a of poster, at the DPG Spring Meetings 2024, as well as NanoMat Science Day 2024 at DESY.

Declaration

Insert declaration here Date & Signature: _____

Appendices

Transforming Raw Data to structured format

Extract, Transform, Load

Raw data from the experiment is stored in hierarchical data format version 5 (HDF5) files. This includes many Beamline diagnostic information such as beam arrival monitor (BAM), gas monitor detector (GMD), the delay stage readings, the monochromator energy, sample specific information such as extractor voltage, and the three-dimensional electron counting by the delay line detector (DLD). This information is resolved at each bunch of electrons coming from the accelerator called a Train, which are further microbunched into pulses.

The open community of multidimensional photoemission spectroscopy (OpenCOMPES) was established to develop tools and infrastructure to make analysis easier. To this end, a modular Python library called `Single Event DataFrame (SED)` was created that provides the entire pipeline from easy data ingestion to common calibration and corrections, Multidimensional binning to create images, and saving the images standard formats, with data provenance.

The data pipeline is designed to extract raw data from HDF5 files, transform the data into a structured format for analysis, and load the transformed data into buffer files. These buffer files are subsequently used for downstream processing, analysis, and visualization. The following sections provide a detailed description of each stage in the ETL process.

Pipeline Overview

The ETL pipeline is divided into several key stages, each critical to the preparation and validation of the data. These stages include:

- **Data Extraction:** Retrieving raw H5 files from experimental runs.
- **Buffer File Creation:** Generating interim buffer files that facilitate further processing.

- **Data Transformation and Validation:** Applying domain-specific transformations, such as forward-filling non-electron channels and splitting sector IDs, while validating data integrity against predefined schemas.
- **Data Structuring:** Organizing the processed data into structured Parquet files suitable for downstream analysis.

Detailed Pipeline Stages

Data Extraction

The initial stage of the pipeline involves the extraction of raw data from H5 files. These files contain experimental data with various channels, such as electron and timed information, stored in a hierarchical structure. The paths to these H5 files are provided as input to the pipeline, along with configuration settings that dictate the subsequent processing steps.

Buffer File Creation

To streamline the data processing, the pipeline first creates buffer files for each type of data (electron and timed dataframes). This is achieved through the `BufferFilePaths` class, which initializes paths for the raw H5 files and corresponding buffer files. The class checks for existing buffer files and determines whether new files need to be generated or existing ones should be reused, based on the `force_recreate` flag.

For each H5 file, the pipeline generates two buffer files:

- **Electron Buffer File:** Contains data relevant to individual electron events.
- **Timed Buffer File:** Aggregates data at pulse and train levels, resolving timing information without individual electron data.

Data Transformation and Validation

Once the buffer files are established, the pipeline proceeds to transform the data. The key transformation steps include:

- **Forward-Filling Non-Electron Channels:** Missing values in non-electron channels are filled using a forward-fill strategy to ensure data continuity.
- **Schema Validation:** The schema of the generated Parquet files is validated against the expected schema derived from configuration files. This ensures that all required channels are present and correctly formatted. The schema check involves:
  - Reading the schema from existing Parquet files.
  - Comparing the actual schema with the expected schema.

  - Raising errors if discrepancies are found, prompting a review of the configuration or a forced recreation of buffer files.

- **Splitting Sector ID from DLD Time:** A custom transformation is applied to separate the sector ID from the DLD (Delay Line Detector) time within the electron dataframe. This operation is essential for accurately resolving electron events.

Data Structuring and Finalization

After transforming and validating the data, the pipeline structures it into Parquet files, which are columnar storage formats optimized for analytical queries. The steps involved include:

- **Electron Dataframe Structuring:** The electron-resolved dataframe is processed to drop non-electron data and reset the index. The processed data is then saved as a Parquet file.

- **Timed Dataframe Structuring:** The timed dataframe is derived by aggregating data at pulse and train levels, excluding electron-specific data. This structured data is also saved as a Parquet file.

- **Metadata Generation:** Metadata related to file statistics, filling operations, and schema checks is compiled and saved, providing crucial information for downstream analysis and data auditing.

Mathematical Background

B.1 Probablity Measure

B.1.1 Measurable Space

Let $\Omega = \emptyset$, $P(\Omega)$ the power set of Ω and $\mathcal{A} \subset P(\Omega)$. If the following conditions are met, $\mathcal{A}$ is a called σ-algebra, and the $(\Omega, \mathcal{A})$ pair make up a measurable space.

1. $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$.
3. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of sets where $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$, then

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}.$$

B.1.2 Positive Measure

Let $(\Omega, \mathcal{A})$ be a measurable space as defined in B.1.1. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a *positive measure* if it satisfies the following conditions:

1. $\mu(\emptyset) = 0$, (the measure of the empty set is zero)
2. μ is *countably additive*: For any countable collection of disjoint sets $(A_i)_{i\in\mathbb{N}} \subset \mathcal{A}$, we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If these conditions are met, then μ is called a *measure* on the measurable space $(\Omega, \mathcal{A})$.

B.1.3 Probability Measure

A *probability measure* $p : \mathcal{A} \to [0, 1]$ satisfies all the properties of a positive measure, with the additional property known as the Normalization condition:

$$p(\Omega) = 1$$

With the measure space being $(\Omega, \mathcal{A}, p)$

B.2 Landau Notation

Landau notation, commonly referred to as Big O notation and its relatives (Big Omega, Big Theta, etc.), is used to describe the asymptotic behavior of functions. This is particularly important in the analysis of algorithms and in expressing the growth rates of functions.

Big O Notation ($\mathcal{O}$)