Reconstruction of the axis position and primary particle energy of extensive air showers in the SPHERE-2 experiment

Extended abstract

Vaiman Igor, SINP MSU

Introduction

Cosmic rays (CR) are highly energetic charged particles that traverse space. They play a crucial role in understanding the dynamics of space systems. Their energy spectrum spans 12 orders of magnitude (10^9 to 10^{21} eV), providing insights into their origin and propagation within the Solar System, Galaxy, and Metagalaxy. Above 10^{15} eV, indirect methods are employed to study cosmic rays through the detection of extensive air showers (EAS), utilizing Earth's atmosphere as a giant calorimeter.

The SPHERE-2 experiment focuses on detecting EAS using the relatively novel method of reflected Cherenkov light. This work enhances the experiment's data processing by incorporating a deeper understanding of the SPHERE-2 detector gained in recent years. The key improvement lies in modeling the detector as a stochastic system, enabling accurate estimation of instrumental uncertainties. Bayesian statistics serves as the primary mathematical framework for the analysis.

1 SPHERE-2 experiment

1.1 Method of reflected Cherenkov light to detect extensive air showers

Cherenkov light is a valuable tool for studying extensive air showers (EAS) due to its weak dependence on high-energy hadronic interaction models. The traditional direct detection method, similar to charged particle detection, employs ground array of detectors distributed over a large area to point-measure flux density.

However, unlike charged particles, Cherenkov light in the optical and near-optical range can be effectively scattered by the Earth's (for example, snow) surface and observed from above. The method of registering reflected light was first proposed by A. E. Chudakov [Chudakov1972] and further developed in the SPHERE-1 and SPHERE-2 experiments [1, 2].

One of the main advantages of this method is the access to Cherenkov light from the paraxial region of the shower, made possible by the extended fields of view of individual sensitive elements of the detector.

1.2 SPHERE-2 detector

1.2.1 Geometry

The SPHERE-2 detector was lifted by a balloon and suspended at an altitude ranging from 400 to 900 m. It utilized an optical system comprising a diaphragm, a spherical mirror, and a mosaic of photomultiplier tubes (PMTs) to collect Cherenkov light from EAS reflected by the snow surface below. Each PMT collected reflected light from an area with a diameter ranging from 10 to 50 m,

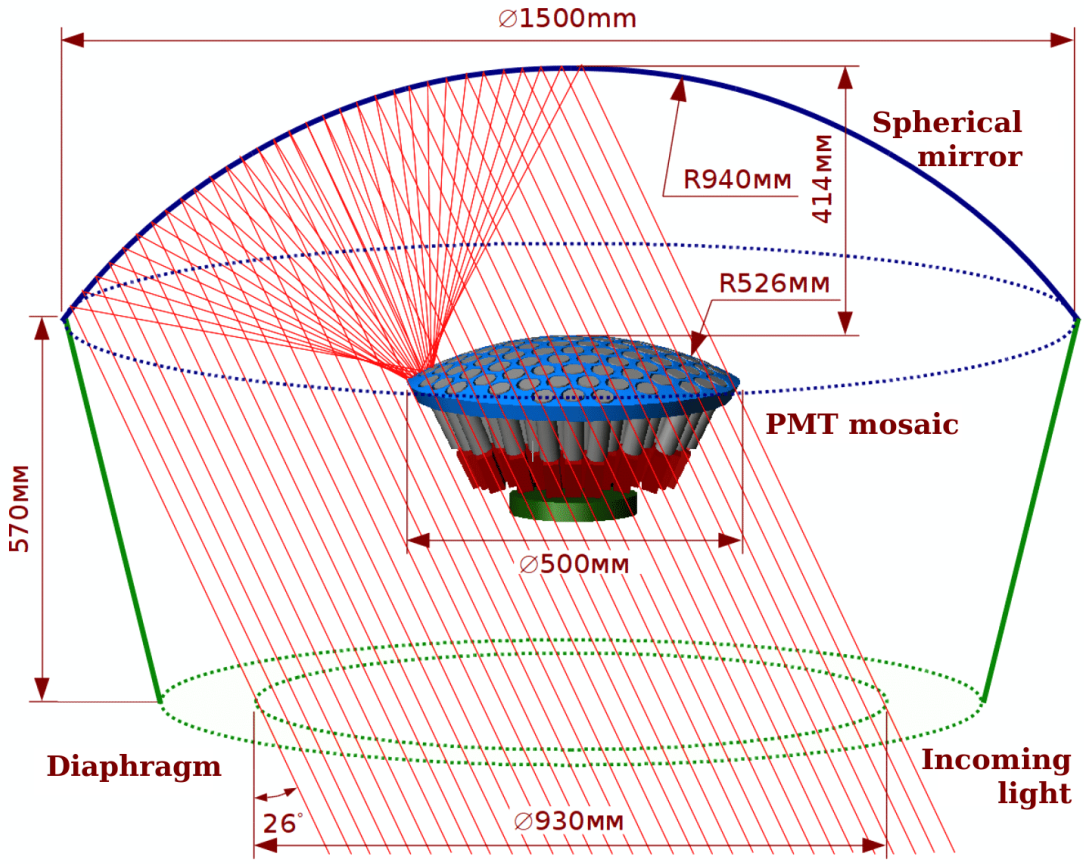

Figure 1.1: SPHERE-2 detector optical scheme

depending on the detector altitude. Figure 1.1 illustrates the optical scheme of the SPHERE-2 detector.

The PMT mosaic, arranged in an approximately hexagonal grid, was located near the focal surface of the spherical mirror. It consisted of 109 PMTs, with the central PMT being the Hamamatsu R3886, serving as a reference for detector calibration [3], and the remaining PMTs being Soviet-made PMT 84-3.

1.2.2 Intermediate PMT mode

When light strikes the PMT photocathode, it emits a photoelectron with a certain probability. This photoelectron then triggers a cascade of secondary photoelectrons through the dynode system, resulting in an observable charge at the anode.

The electron cascade within a PMT is a fundamentally stochastic process, as verified by direct laboratory measurements of anode charge fluctuations [3, Fig. 9]. Each dynode undergoes random multiplication, and due to the cascading nature, small variations in the early stages can lead to significant differences in amplification.

The stochastic nature of PMT amplification is not apparent when measuring high fluxes due to averaging. Likewise, it is not significant at low fluxes when the PMT operates in photon counting mode, allowing for the observation of well-resolved individual pulses. However, in the SPHERE-2 detector, PMTs operate in an intermediate mode: the flux is too high to resolve individual photons, yet insufficient for effective averaging. Consequently, the deconvolution problem needs to be formulated in statistical terms.

2 Bayesian deconvolution

Early stages of the SPHERE-2 analysis included an attempt to construct the process of estimating the parameters of an EAS photons in each PMT (the photon count and the arrival time distribution) directly from the recorded signal, but it was not successful for a number of reasons. Instead, the idea arose to carry out a full deconvolution, that is, to extract information about the photon flux at PMTs input at each moment of time. This deconvolution should take into account the stochastic properties of the PMT. Similar problems are also considered in other areas, for example, in the processing of [4] spectral measurements and [5] images. In such problems, Bayesian statistics is fruitfully used, based on the interpretation of probability as a measure of information about the random variable or confidence in its value [6].

2.1 Problem setup

The input signal is modeled as a set of δ-functions offsetted in time. Time bins (12.5 nsec in the SPHERE-2 detector) provide natural time scale, and we assume that within each time bin, δ-functions are distributed uniformly. We limit ourselves with N time bins. The deconvolution will yield estimation for n_i, photon counts per $[i-1, i]$ time bin, where $i = 1 \dots N$.

We assume that the PMT is a linear system with stochastic impulse response function $\tilde{h}(t)$ in the sense that each of the δ-functions that make up the input signal is convolved with an independent sample $h(t) \sim \tilde{h}(t)$. This corresponds to the idea of fluctuations in the PMT amplification happening independently for each electron cascade. We will assume that any sample from $\tilde{h}(t)$ is causal, i.e. $\forall t < 0 \;\; \forall h \sim \tilde{h} \;\; h(t) = 0$, and finite in time, i.e. $\exists \tilde{L}$ such that $\forall t > \tilde{L} \;\; \tilde{h}(t) = 0$. We will call this random function, which gives a new sample for each input δ-function, the system's randomized impulse response (RIR).

We pose the problem of statistical deconvolution using Bayesian terminology

> Given the randomized impulse response of the system $\tilde{h}(t)$ and the output signal $s_j, \;\; j = 1, \dots, N + L$, find the posterior probability density functions for the values n_i, $i = 1, \dots, N$.

Note that, unlike regular deconvolution, we are not trying to estimate the full original signal (sum of δ-functions), but only its integrated in each time bin.

Mathematically, we write

$$P(\vec{n}|\vec{s}) = \frac{P(\vec{s}|\vec{n})\, P(\vec{n})}{P(\vec{s})} \tag{2.1}$$

Using uninformative prior $P(\vec{n}) = Const$ (it can't be normalized, but we will use only proportionality, not an absolute value), denoting likelihood function $\mathcal{L}(\vec{s}, \vec{n}) = P(\vec{s}|\vec{n})$, we get

$$P(\vec{n}|\vec{s}) \propto \mathcal{L}(\vec{s}, \vec{n}) \tag{2.2}$$

2.2 Likelihood function

PMT output signal $\vec{s}$ is a sample from some multivariate random variable. Denoting this variable $\vec{S}$, we can write

$$S_j = \sum_{l=0}^{L} C(n_{j-l}, l) \tag{2.3}$$

Here $C(n, l)$ is a random variable describing the contribution of n δ-functions with the delay of l bins. It is easy to see from the chosen bin indexing scheme and the conditions of causality and boundedness in time of the RIH that $l \in [0, L]$, since the contribution from δ-functions from earlier bins is equal to zero.

The distribution of $C(n, l)$ can be studied with Monte-Carlo. Sampling it goes as follows. If we sample RIR $h_k(t) \sim \tilde{h}(t)$ and in-bin time $t_{inbin} \sim U(0, 1)$, and get a sample from $C(1, l) = h_k(l + 1 - t_i nbin)$. Sampling from $C(n, l)$ is trivial, since the system is linear and we can just add n independent samples from $C(1, l)$.

To calculate likelihood function, one needs to find a probability to find a particular $\vec{s}$ sampling from $\vec{S}$. Naive Monte-Carlo likelihood estimation requires sampling a large number of $\vec{n}$ and computing histogram in $\vec{s}$ space. This method is computationally unfeasible, because of the curse of dimensionality: in practice we are interested in signals at least 50 time bins long.

2.2.1 Multivariate normal distribution approximation

We approximate likelihood function with the multivariate normal distribution. There is no formal proof for the applicability for such approximation, but we estimate that it is fairly close for input intensities as low as 4-5 δ-functions per bin, which is the characteristic background intensity in SPHERE-2 detector. For partial univariate s_j distribution, the probability density function is within 1-2% of its normal approximation.

The general form of multivariate normal distribution is

$$p(\vec{s}) = \left((2\pi)^{N+L} \det \Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}(\vec{s}-\vec{\mu})^T \Sigma^{-1} (\vec{s}-\vec{\mu})\right) \tag{2.4}$$

Here, mean vector $\vec{\mu}$ and covariance matrix Σ depend on input $\vec{n}$ and RIR $\tilde{h}(t)$. Specifically, $\vec{\mu} \equiv \mathbb{E}\vec{S}$ can be obtained from 2.3 by computing mean of left and right terms and using the fact that $\mathbb{E}C(n,l) = n\ \mathbb{E}C(1,l)$. The matrix Σ can be expressed in terms of $C(n,l)$ autocovariance, which can be further expressed in terms of $C(1,l)$ autocovariance. The latter depend only on RIR and not on $\vec{n}$, and can be pre-calculated for computational effectiveness.

$$\Sigma_{ij} = \operatorname{cov}(S_i, S_j) = \sum_{l=0}^{L-(i-j)} \operatorname{cov}(C(n_{i-l}, l), C(n_{i-l}, l + (i-j))) \tag{2.5}$$

2.3 Output signal error

Multivariate normal approximation does not account for the error of measuring the output signal. The largest contribution to this error is ADC discretization. We assume that ADC floors its input, i.e. for each recorded s_j, the real input value was somewhere between s_j and $s_j + \delta$ and the distribution is uniform. Such error can be accounted for by integrating the $\mathcal{L}_{\text{exact}}(\vec{s}, \vec{n})$ defined by 2.4 over the N-dimensional cube with side δ and «lower» corner at $\vec{s}$:

$$\mathcal{L}(\vec{s}, \vec{n}) = \int_{s_1}^{s_1+\delta} \int_{s_2}^{s_2+\delta} \dots \int_{s_N}^{s_N+\delta} \mathcal{L}_{\text{exact}}(\vec{s}, \vec{n})\, d\vec{s} \tag{2.6}$$

To perform this integration numerically, we use the ready-made adaptive integration algorithm described in [7], implemented in the `scipy.stats` [8] `Python` package.

2.4 Posterior distribution sampling

Having obtained the likelihood function $\mathcal{L}$, we can estimate $\vec{n}$ given the recorded output signal $\vec{s}$ and RIR $\tilde{h}(t)$. We want not only to find the optimal value of $\vec{n}$, but also characterize our confidence in it.

Instead of maximum likelihood method, where we would maximize the value of $\mathcal{L}(\vec{n}, \vec{s})$ with respect to $\vec{n}$, we use Marko Chain Monte Carlo sampling to draw a large sample from $P(\vec{n}|\vec{s}) \propto \mathcal{L}(\vec{n}, \vec{s})$. Using this sample of possible $\vec{n}$ values, we will then get the bets-fitting value for system's input, and its error.

The idea of the Markov Chain Monte Carlo (MCMC) family of methods [9] is to launch a Markov process (random walk) in the parameter space $\vec{n}$ with transition probability chosen in such a way that the «trace» of the walk will yield a sample from the target probability density functions. A common characteristic of this family is that they require only knowing only a ratio of probability density functions at different points in the parameter space, which is why in the expression (2.2) and in all subsequent calculations we were able to drop the marginal probability and the normalization of the prior distribution.

The particular method used in this work is affine-invariant MCMC [10], implemented in `Python` package `emcee` [11].

2.5 Deconvolution results

A detailed description of the PMT connection and power supply circuits, the amplification circuit, as well as the characteristics of the used ADC converters can be found in [12]. For purposes of this work, we define detector's randomized impulse response as

$$\tilde{h}(t) = C\,\tilde{C}_{PMT}\,h_I(t) \tag{2.7}$$

Where

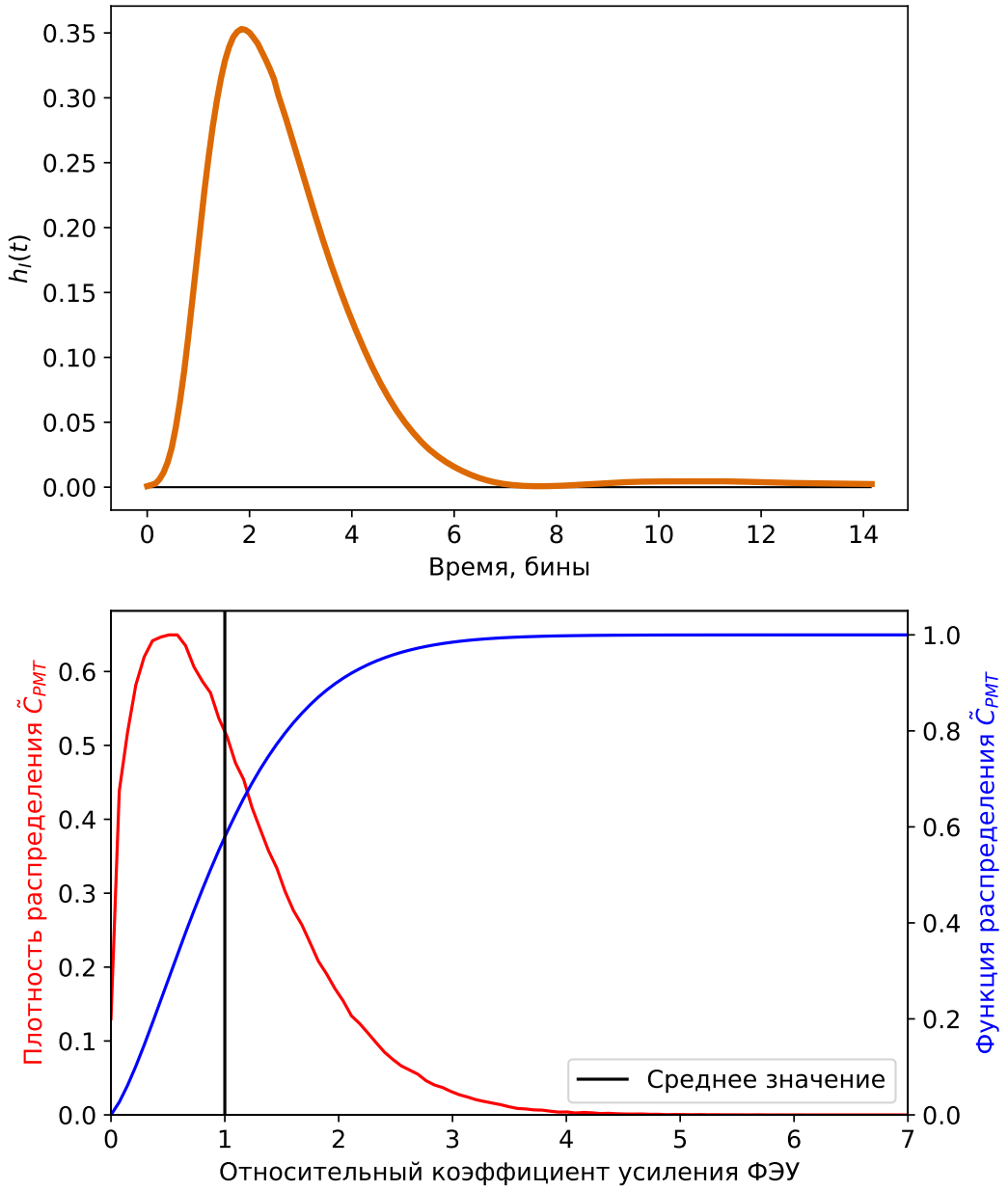

Figure 2.1: SPHERE-2 PMT's randomized impulse response. $h_I(t)$ (top panel) is the signal's shape in time domain, $\tilde{C}_{PMT}$ (bottom panel: PDF and CDF) — dimensionless random PMT amplification coeffient.

1. $h_I(t)$ — shape of the impulse response, defined by time characteristics of the PMT and signal amplification circuit. Measured in the laboratory and plotted on the top panel of Fig. 2.1. Normalized to have integral 1.

2. $\tilde{C}_{PMT}$ — dimensionless random PMT amplification coefficient, normalized to have a mean value of 1. It's PDF and CDF are plotted on the bottom panel of Fig. 2.1.

3. C — scale coefficient. It is measured during detector calibration and applied to signals earlier in data processing pipeline.

We illustrate deconvolution procedure on toy input data: Poisson background with mean $\lambda = 20$ photons per time bin and a «signal» photon packet: 3 bins with additional Poisson signal with $\lambda_{\text{signal}} = 40$. Fig. 2.2 illustrates the system's input and output on the top panel and the deconvolution results on the bottom panel.

3 Extensive air shower parameters estimation

3.1 Shower photons discrimination

During deconvolution, no assumptions are made about the presence or absence of an EAS photons in the signal. From optical detector simulation, we expect EAS photons to reach the mosaic in packets — groups of photons with an approximately normal arrival time distribution. Each packet can be described by three parameters: n_{EAS} — total number of photons in the packet, μ_t — average photon arrival time, σ_t — standard deviation of arrival times. In the following, we define $\Theta \equiv (n_{EAS}, \mu_t, \sigma_t)$. The background (primarily, star and zodiacal light reflected from the snow) is modeled as Poisson-distributed photons with the average number λ_n known from the calibration.

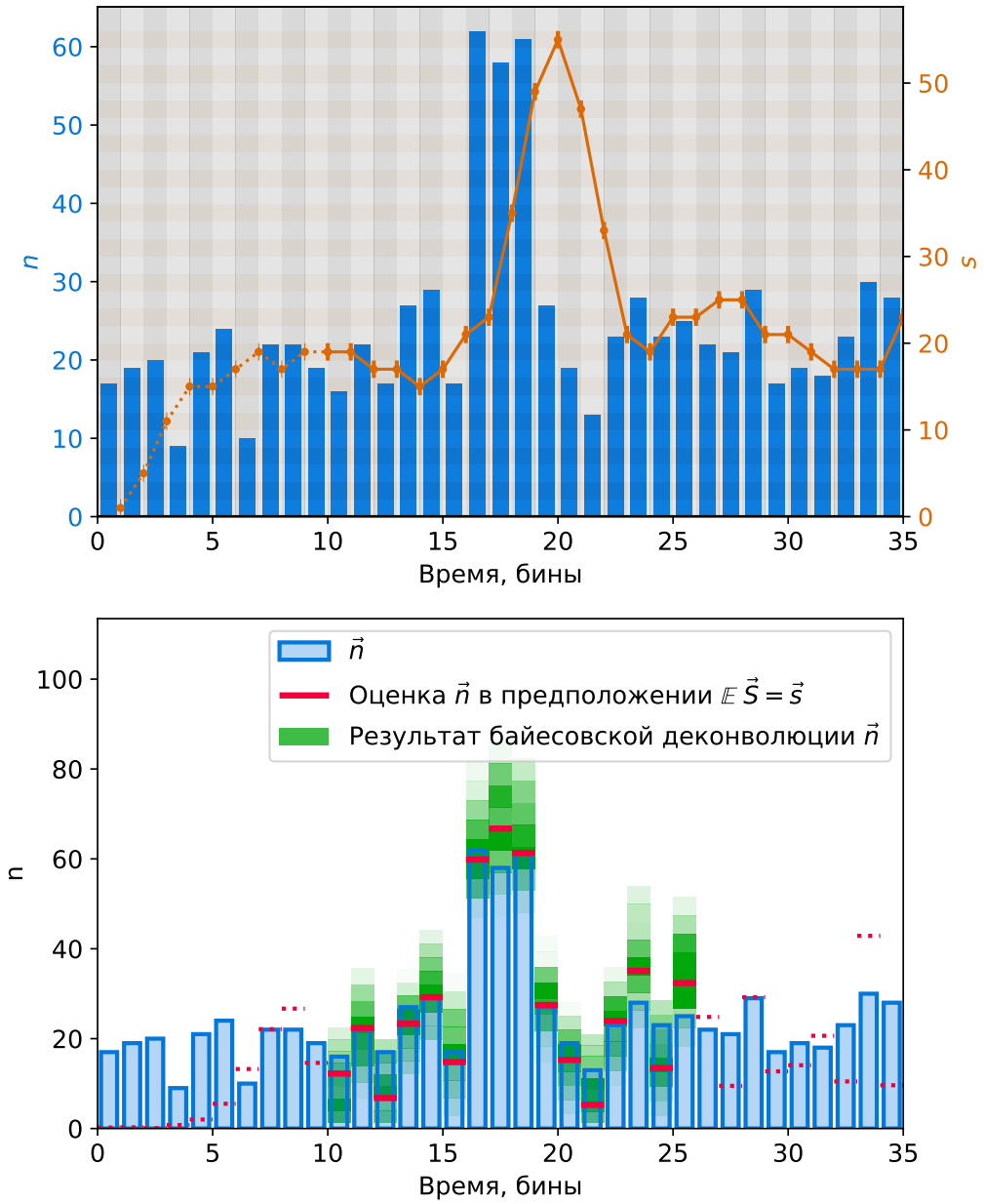

Figure 2.2: Bayesian deconvolution performed on synthetic toy input data. Top panel: toy input data $\vec{n}$ (blue) and corresponding system's output (convolution with randomized impulse response from fig. 2.1) $\vec{s}$, orange (X axis: «Time, bins»). Bottom pannel: same input data $\vec{n}$ (blue); rough $\vec{n}$ estimation from mean values, MCMC sampling starting point (red); deconvolution result, marginal posterior distributions in each time bin (green).

This problem can also be formulated as a Bayesian problem of finding a posterior distribution or EAS photons packet given the distribution of $\vec{n}$ obtained from the previous step (deconvolution).

First, let's consider a constant value of $\vec{n}$ and define the likelihood function for Θ in this case, i.e. the probability that the particular signal $\vec{n}$ was generated from the known background parameter λ_n and an EAS packet with parameters Θ. This can be easily done with the following Monte-Carlo: we sample a large number of photon arrival times following Θ (n_{EAS} times $\sim \mathcal{N}(\mu_t, \sigma_t)$), convert each of them to histogram $\vec{n}_{\text{signal}}$, calculate $\vec{n}_{\text{background}} \equiv \vec{n} - \vec{n}_{\text{signal}}$ and, finally, compute the probability that this set of background photon counts is a sample from Poisson distribution with mean λ_n using the standard Poisson likelihood function.

$$\mathcal{L}(\Theta) \equiv P(\vec{n}|n_{EAS}, \mu_t, \sigma_t) = \prod_i \frac{e^{-\lambda_n} \lambda_n^{n^{(i)}_{\text{background}}}}{(n^{(i)}_{\text{background}})!} \tag{3.1}$$

Finally, to get the full likelihood function, given that $\vec{n}$ is defined by its posterior distribution, we have to average this exact likelihood over this distribution. Luckily, since we have not the posterior distribution itself, but a sample from it, we can just average the exact likelihood value over this sample.

Unlike the uninformative prior used for deconvolution, we can select meaningful priors Θ from Monte-Carlo simulation data: exponential distribution with $\lambda = 40$ for n_{EAS}, and $\mathcal{N}(2.4, 1)$ truncated at zero for σ_t. For μ_t, a uniform prior over the signal frame was chosen.

After that, the same MCMC technique was used to get the final signal reconstruction result: posterior distribution of photon packet parameters Θ. Performing both steps of the analysis within the same Bayesian framework lets us effectively chain these two analysis together, while also keeping them separate.

3.1.1 Shower photon packet significance

One of the advantages of a completely statistical approach is the ability to use the concept of significance to disciriminate the EAS and the background photons. We use Bayesian information criterion (BIC), introduced by Schwartz [13] (in a sense, this is an Akaike's information criterion [14] adapted for Bayesian analysis). It states that several models describing data can be compared by the amount of information that is lost when replacing data with a model. The smaller the loss of information, the smaller the value of the criterion, and the better the model performs. The calculation is carried out according to the formula

$$\mathrm{BIC} = k \ln n - 2 \ln \mathcal{L}_{max} \tag{3.2}$$

Here k is the number of model parameters, n is the number of data sample elements, $\mathcal{L}_{max}$ is the maximum value of the likelihood function for the given model. As can be seen from the formula, BIC penalizes large number of parameters k.

To determine the significance of the found EAS signal, we compare two models: the «background only» model ($n_{EAS} = 0$) without parameters ($k = 0$), and the «background + signal» model, described in the previous section, with $k = 3$. The difference $\Delta\mathrm{BIC} = \mathrm{BIC}_{noise} - \mathrm{BIC}_{background+EAS}$ shows the significance of the reconstructed signal. $\Delta\mathrm{BIC} \leq 0$ means the noise-only model is better, and such channels are considered empty. $0 < \Delta\mathrm{BIC} < 4$ means weak signal significance [15]; in practice, we find that most deconvolution artifacts or strong background fluctuations fall into this region. Channels with $\Delta\mathrm{BIC} > 4$ are accepted as reliable (remaining false-positives are later eliminated during the shower geometry reconstruction). Figure 3.1 shows an example of full EAS photon packet parameters reconstruction for an experimental event.

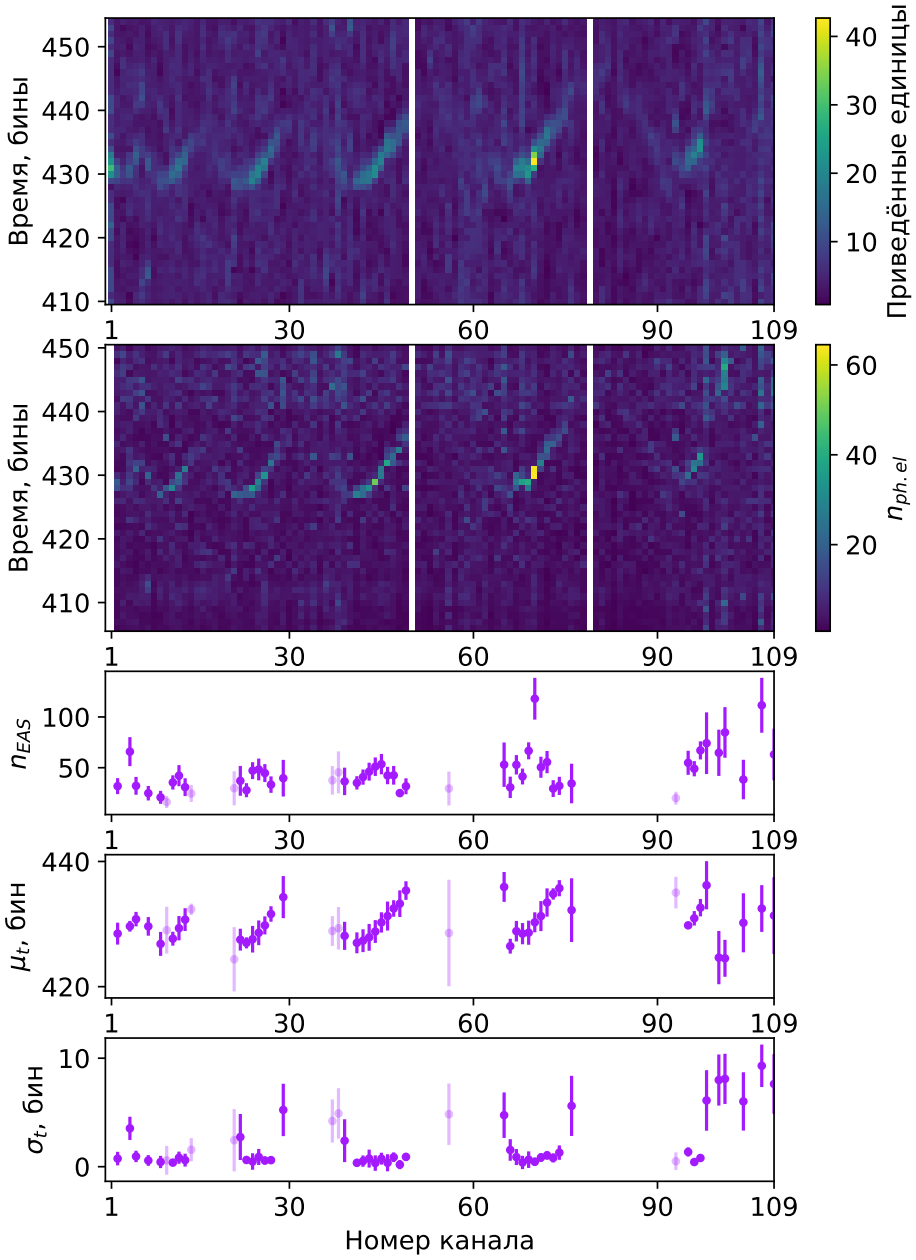

Figure 3.1: Full signal reconstruction for experimental event #10675. Common X axis: channel number, see [3] for numbering scheme. Top panel: experimental data with calibration coefficiens applied. Second-to-top panel: deconvolution output, photon count in each time bin. Three bottom panels: reconstructed photon packet parameters n_{EAS}, μ_t, σ_t — dots and error bars represent marginal means and standard deviations; dimmed values are signals with $\Delta\text{BIC} \in (0, 4]$, negative ΔBIC (background-only channels) are omitted.

3.2 Shower parameter reconstruction

The rest of the reconstruction performed here in this work standard methods, with the exception of being applied to the output of Bayesian reconstruction.

- Shower arrival direction is determined by a plain fit of arrival times. Ground level arrival times for each channel are determined from μ_t by subtracting ground-to-detector light travel time, accounting for the detector altitude and inclination.
- Shower axis position is determined by finding a point on the surface that maximizes the probability (with respect to each channel's posterior distribution) of monotonously decreasing lateral distribution function.
- Lateral distribution function for EAS Cherenkov light is determined by fitting a parametrized function to the data. In this work, a simplified LDF from [16] is used. The fitting is done by convolving the LDF with each channel's light collection function (which is calculate from detector's optic system model). Shower energy is then derived from LDF parameters.

Bibliography

[1] R.A. Antonov et al. "Primary cosmic ray spectrum measured using Cherenkov light reflected from the snow surface". In: Nuclear Physics B - Proceedings Supplements 52.3 (Feb. 1997), pp. 182–184. DOI: 10.1016/s0920-5632(96)00876-6.

[2] R. A. Antonov et al. "Balloon-Borne measurements of the CR energy spectrum in the energy range 10^{16}-10^{17} EV". In: International Cosmic Ray Conference. Vol. 1. International Cosmic Ray Conference. Jan. 2001, p. 59.

[3] R.A. Antonov et al. “The LED calibration system of the SPHERE-2 detector”. In: Astroparticle Physics 77 (Apr. 2016), pp. 55–65. DOI: 10.1016/j.astropartphys.2016.01.004.

[4] Cynthia Rhode and Scott Whittenburg. “Bayesian Deconvolution”. In: Spectroscopy Letters 26.6 (July 1993), pp. 1085–1102. DOI: 10.1080/00387019308011596.

[5] David Wipf and Haichao Zhang. “Revisiting Bayesian Blind Deconvolution”. In: arXiv e-prints, arXiv:1305.2362 (May 2013), arXiv:1305.2362. arXiv: 1305.2362 [cs.CV].

[6] Andrew Gelman et al. Bayesian Data Analysis. Chapman and Hall/CRC, Nov. 2013. DOI: 10.1201/b16018.

[7] Alan Genz. “Numerical Computation of Multivariate Normal Probabilities”. In: Journal of Computational and Graphical Statistics 1.2 (June 1992), pp. 141–149. DOI: 10.2307/1390838.

[8] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: Nature Methods 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[9] Sanjib Sharma. “Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy”. In: Annual Review of Astronomy and Astrophysics 55.1 (Aug. 2017), pp. 213–259. DOI: 10.1146/annurev-astro-082214-122339.

[10] Jonathan Goodman and Jonathan Weare. “Ensemble samplers with affine invariance”. In: Communications in Applied Mathematics and Computational Science 5.1 (Jan. 2010), pp. 65–80. DOI: 10.2140/camcos.2010.5.65.

[11] Daniel Foreman-Mackey. “corner.py: Scatterplot matrices in Python”. In: The Journal of Open Source Software 1.2 (June 2016), p. 24. DOI: 10.21105/joss.00024.