PHY 657 — Module 3 Homework: Regression & Matched Filtering

Caleb Fink Spring 2026

Introduction

In this assignment you will analyze simulated data from a simplified particle detector. The detector records short time-domain traces containing noise and occasional pulses produced by energy depositions in the sensor. Your goal is to reconstruct the deposited energy as accurately as possible and evaluate how well your reconstruction method performs.

This assignment mirrors the workflow of a real experimental analysis:

1. Understand likelihoods and parameter estimation
2. Understand why noise weighting matters
3. Construct an estimator for a physical quantity
4. Calibrate the detector response
5. Measure resolution and detection efficiency

The early problems introduce statistical tools that will later be used in the full pulse reconstruction analysis. You are encouraged to reuse your earlier code.

Unless otherwise stated, all uncertainties are Gaussian and independent. *Derivations should be clear but do not need to be excessively formal — emphasize reasoning.*

Reading Assignment: Bishop Sections 3.1- 3.3

Due: *Monday March 2nd.*

Notation / reminders

- Vectors are bold lower case, $\mathbf{x}$. Matrices are upper case, A.
- For multivariate Gaussian: $p(\mathbf{x} \mid \mu, \Sigma) = \mathcal{N}(\mu, \Sigma)$.
- For an $N \times d$ data matrix X: N samples, d dimensions.

Problem 1: Likelihood and Parameter Uncertainty

You observe a detector that triggers with probability θ on each trial. You perform N independent trials and observe k triggers.

(a) Maximum likelihood estimate

Write the likelihood function

$$p(k \mid \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Compute the maximum likelihood estimator $\hat{\theta}$.

(b) Numerical likelihood scan

For $N = 40$ and $k = 26$:

1. Plot the log-likelihood $\log p(k|\theta)$ for $0 < \theta < 1$
2. Mark the maximum

(c) Uncertainty from likelihood curvature

Define the 1σ confidence interval using

$$\log L(\theta) = \log L_{\max} - \frac{1}{2}.$$

Numerically determine the upper and lower bounds.

(d) Interpretation

Explain in words why the width of the likelihood peak represents uncertainty in the parameter.

Problem 2: Why Noise Weighting Matters

You measure a signal that should follow a linear relation

$$y = ax + b$$

but each data point has a different uncertainty σ_i.

You are given arrays (x_i, y_i, σ_i).

(a) Unweighted fit

Fit the model using ordinary least squares (ignore σ_i). Plot the data and best-fit line.

(b) Weighted fit

Now minimize

$$\chi^2 = \sum_i \frac{(y_i - ax_i - b)^2}{\sigma_i^2}.$$

Plot the new best-fit line.

(c) Comparison

1. Which points influenced the weighted fit more strongly?
2. Why does inverse-variance weighting improve parameter estimation?

(d) Conceptual connection

Suppose instead of fitting a line, you wanted to estimate the amplitude of a known waveform buried in noise with covariance matrix Σ.

Explain why you would expect the optimal estimator to weight the data by Σ^{-1}. (No derivation required.)

Problem 3: Matched Filtering, Calibration, and Detection Efficiency

In this assignment you will analyze simulated detector data containing non–white noise, three monoenergetic nuclear decay lines, and a flat background. Your goal is to extract optimal energy estimates, calibrate the detector response, and measure detection efficiency.

You will be provided with two waveform datasets which will be measurements of voltage vs time. The datasets given consist of 'noise' data, and 'events' consisting of triggered pulses. The pulses are embedded in colored (non–white) noise and include:

- Three spectral lines from nuclear decays
- A broad "noise blob" population
- A flat background of random triggers

Your task is to determine the detector performance using three different estimators:

1. Matched/Optimal Filter amplitude estimator (MF)
2. Naive peak estimator (maximum sample or fixed index estimator)
3. Integral estimator (sum/integral over pulse window)

You must compare their energy resolution and detection efficiency.

Part I: Matched Filter Amplitudes

1. Construct a noise power spectral density (PSD), $J(f)$, estimate using the noise-only dataset. Plot the (one-sided) PSD on a log-log plot.

2. Build a matched filter using a template pulse $s(t)$ and the measured PSD:

$$a_{\mathrm{MF}} = \frac{\sum_k \tilde{s}^*(f_k)\tilde{d}(f_k)/J(f_k)}{\sum_k |\tilde{s}(f_k)|^2/J(f_k)}.$$

   where $\tilde{d}(f_k)$ and $\tilde{s}(f_k)$ are the Fourier transforms of the data and template respectively.

   Our pulse template is normalized to have a maximum amplitude of one, and scales as

$$s(t) \propto \exp(-t/\tau_{\mathrm{fall}}) - \exp(-t/\tau_{\mathrm{rise}})$$

   where $\tau_{\mathrm{rise}} = 100\,\mu\mathrm{s}$ and $\tau_{\mathrm{fall}} = 5\,\mathrm{ms}$.

3. Apply the matched filter to all events and extract an amplitude for each pulse.

4. Plot a histogram of the measured amplitudes.

Part II: Spectral Model and Maximum Likelihood Fit

The amplitude spectrum consists of three spectral lines, a Gaussian noise population, and a flat background.

Model the probability density as

$$p(A) = \sum_{i=1}^{3} w_i \mathcal{N}(A|\mu_i, \sigma_i^2) + w_{\mathrm{noise}} \mathcal{N}(A|\mu_n, \sigma_n^2) + w_{\mathrm{flat}} \frac{1}{A_{\max} - A_{\min}},$$

with

$$\sum_j w_j = 1.$$

1. Perform an unbinned maximum likelihood fit to extract all model parameters:

$$\{w_i, \mu_i, \sigma_i\},\ w_{\mathrm{noise}}, \mu_n, \sigma_n,\ w_{\mathrm{flat}}.$$

2. Identify the three signal peaks and determine their measured amplitudes A_1, A_2, A_3 and uncertainties.

Part III: Energy Calibration

You will be provided with the true energies of the three lines:

$$E_1^{\mathrm{true}} = 25\,\mathrm{keV},\ E_2^{\mathrm{true}} = 65\,\mathrm{keV},\ E_3^{\mathrm{true}} = 100\,\mathrm{keV}.$$

The detector response follows nonlinear saturation model

$$E_{\mathrm{recon}} = a\left(1 - \exp\left(-\frac{E_{\mathrm{true}}}{b}\right)\right).$$

1. Fit for calibration parameters a and b.
2. Construct the inverse calibration function to convert measured amplitudes to reconstructed energy E_{rec}.
3. Convert the full amplitude spectrum into an energy spectrum.
4. comment on the benefits of the above saturation model vs a polynomial/spline fit.

Part IV: Detection Threshold

Define a detection threshold using the noise population:

$$E_{\text{thresh}} = \mu_n + 5\sigma_n.$$

1. Determine the corresponding energy threshold after calibration.
2. Plot the calibrated spectrum and mark the threshold.

Part V: Detection Efficiency

You will now simulate pulses of known amplitude.

1. Inject fake pulses of known true energy into real noise.
2. Process them through the full analysis chain.
3. A pulse is "detected" if $E_{\text{rec}} > E_{\text{thresh}}$.

Measure the efficiency

$$\varepsilon(E) = \frac{N_{\text{detected}}(E)}{N_{\text{injected}}(E)}.$$

Plot efficiency vs energy.

Part VI: Alternative Estimators

Repeat the entire analysis using:

1. Peak estimator: maximum sample value
2. Integral estimator: summed pulse area

For each estimator:

- Recalibrate the energy scale
- Compute resolution $\sigma_E(E)$ at each spectral line
- Compute detection efficiency curve

Final Comparison

Provide a comparison table including:

Estimator	Energy Resolution	Threshold Energy
Matched Filter		
Peak		
Integral		

and overlay the efficiency curves on a single plot.

Discussion: Explain physically why the matched filter performs differently from the other estimators. Discuss how colored noise impacts optimal filtering and detection efficiency.

End of assignment. Good luck — email me with questions about implementation choices!