The Maximum Likelihood Estimation Method

(last modified August 14, 2009)

Xianguo Lu

xianguo.lu@desy.de

Motivation & Outlines

Analytically illustrate the maximum likelihood method, from the basic to the advanced, with MC examples; implement the method to practical data analysis.

- Part 1: all about the standard method
  1. Concepts
  2. Analysis
  3. MC Examples
- Part 2: "weighting" and the extended method
  1. Concepts
  2. Analysis
  3. Application
- Part 3: for the combined analysis

Part 1:

THE STANDARD

Concepts

- Observable

$$x \in \mathbf{X}. \tag{1}$$

- Data set

$$x_1,\ x_2, \ldots,\ x_N. \tag{2}$$

Remark.
"Experiment", consist of N observations, obtaining $x_1,\ x_2, \ldots,\ x_N$.

•

$$\text{Prob}\left(x' < x < x' + \mathrm{d}x\right) = p\left(x'\right)\mathrm{d}x, \tag{3}$$

$p\left(x\right)$, **Probability density function (p.d.f.)**; $p\left(x|\theta_1,\ \theta_2, \ldots,\ \theta_M\right)$;
$\theta_1,\ \theta_2, \ldots,\ \theta_M$, **parameters**, $\underline{\theta} \equiv \text{col}\left(\theta_1,\ \theta_2, \ldots,\ \theta_M\right)$;
$\underline{\theta}^*$, the **true parameters** corresponding to the data;
(Assuming p is twice differentiable with respect to θ_i) $\frac{\partial}{\partial \underline{\theta}} \equiv \text{col}\left(\frac{\partial}{\partial\theta_1},\ \frac{\partial}{\partial\theta_2},\ \ldots,\ \frac{\partial}{\partial\theta_M}\right)$,
gradient operator in the parameter space.

$$\int_{\mathbf{X}} p\left(x|\underline{\theta}\right)\mathrm{d}x = 1, \qquad \langle f(x)\rangle \equiv \int_{\mathbf{X}} f(x) p\left(x|\underline{\theta}^*\right)\mathrm{d}x. \tag{4}$$

Remark.

- Expectation for repeated *observations*.
- $\langle f(x)\rangle_{\underline{\theta}} \equiv \int_{\mathbf{X}} f(x) p\left(x|\underline{\theta}\right)\mathrm{d}x$.

- Joint p.d.f./likelihood function

$$L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \equiv \prod_{i=1}^{N} p\left(x_i|\underline{\theta}\right).$$

$$\int\cdots\int_{\mathbf{X}^N} L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_N = 1. \tag{5}$$

$$\langle g(x_1,\ x_2,\ldots,\ x_N)\rangle \equiv$$
$$\int\cdots\int_{\mathbf{X}^N} g(x_1,\ x_2,\ldots,\ x_N) L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}^*\right) \mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_N. \tag{6}$$

Remark.

- Expectation for repeated $experiments$.
- $\langle g\left(x_1,\ x_2,\ldots,\ x_N\right)\rangle_{\underline{\theta}} \equiv$
  $\int\cdots\int_{\mathbf{X}^N} g\left(x_1,\ x_2,\ldots,\ x_N\right) L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_N.$

- Log-likelihood function

$$l\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \equiv \ln L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \tag{7}$$

$$= \sum_{i=1}^{N} \ln p(x_i|\underline{\theta}). \tag{8}$$

- Estimators

$$\underline{\hat{\theta}}\left(x_1,\ x_2,\ldots,\ x_N\right). \tag{9}$$

Remark.
Also random.

- Bias

$$\underline{b}_{\underline{\hat{\theta}}} \equiv \left\langle \underline{\hat{\theta}} \right\rangle - \underline{\theta}^*. \tag{10}$$

- Covariance matrix

$$\underline{\underline{V}}_{\underline{\hat{\theta}}} \equiv \left\langle \left(\underline{\hat{\theta}} - \left\langle \underline{\hat{\theta}} \right\rangle\right) \left(\underline{\hat{\theta}} - \left\langle \underline{\hat{\theta}} \right\rangle\right)^T \right\rangle. \tag{11}$$

- Minimum variance bound, MVB

$$\underline{\underline{V}}_{\underline{\hat{\theta}}} \geqslant \left(\left.\frac{\partial \left\langle \hat{\theta}_i \right\rangle_{\underline{\theta}}}{\partial \theta_j}\right|_{\underline{\theta}^*} \right) \left(\left\langle \left.\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right|_{\underline{\theta}^*} \right\rangle \right)^{-1} \left(\left.\frac{\partial \left\langle \hat{\theta}_j \right\rangle_{\underline{\theta}}}{\partial \theta_i}\right|_{\underline{\theta}^*} \right). \tag{12}$$

[Hint: Using Schwarz's Inequality (generalized, cf. Appendix),
$\left\langle \underline{a}\ \underline{a}^T \right\rangle \geqslant \langle \underline{a}\ \underline{b}^T \rangle \langle \underline{b}\ \underline{b}^T \rangle^{-1} \langle \underline{b}\ \underline{a}^T \rangle$, let $\underline{a} = \underline{\hat{\theta}} - \langle \underline{\hat{\theta}} \rangle,\ \underline{b} = \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}^*}$. (If matrix $\underline{\underline{A}}$ is positive definite (semidefinite), i.e. $\forall \underline{v} \neq \underline{0},\ \underline{v}^T \underline{\underline{A}}\ \underline{v} > (\geqslant)\, 0$, we write $\underline{\underline{A}} > (\geqslant)\, \underline{0}$).]

Remark.

– For unbiased estimators, $\left\langle \hat{\theta}_i \right\rangle_{\underline{\theta}} = \theta_i$,

$$\therefore \underline{\underline{V}}_{\underline{\hat{\theta}}} \geqslant (\delta_{ij}) \left(\left\langle \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \bigg|_{\underline{\theta}^*} \right\rangle \right)^{-1} (\delta_{ji}) \tag{13}$$

$$= \left(\left\langle \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \bigg|_{\underline{\theta}^*} \right\rangle \right)^{-1} = \frac{1}{N} \left(\left\langle \frac{\partial \ln p}{\partial \theta_i} \frac{\partial \ln p}{\partial \theta_j} \bigg|_{\underline{\theta}^*} \right\rangle \right)^{-1}. \tag{14}$$

Let $\underline{v} = \mathrm{col}(0, \ \ldots, \ \underbrace{1}_{i\text{th component}}, \ \ldots, \ 0)$, the **variance**

$$\sigma^2_{\hat{\theta}_i} \equiv \left[\underline{\underline{V}}_{\underline{\hat{\theta}}} \right]_{ii} = \underline{v}^T \underline{\underline{V}}_{\underline{\hat{\theta}}} \underline{v} \geqslant \underline{v}^T \frac{1}{N} \left(\left\langle \frac{\partial \ln p}{\partial \theta_i} \frac{\partial \ln p}{\partial \theta_j} \bigg|_{\underline{\theta}^*} \right\rangle \right)^{-1} \underline{v} = \frac{1}{N} \left[\left(\left\langle \frac{\partial \ln p}{\partial \theta_i} \frac{\partial \ln p}{\partial \theta_j} \bigg|_{\underline{\theta}^*} \right\rangle \right)^{-1} \right]_{ii}. \tag{15}$$

In the single parameter case,
$$\sigma^2_{\hat{\theta}} \geqslant \frac{1}{N \left\langle \left(\frac{\partial \ln p}{\partial \theta} \right)^2_{\theta^*} \right\rangle}. \tag{16}$$

[Hint: $\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} = \sum_m \frac{\partial \ln p(x_m)}{\partial \theta_i} \frac{\partial \ln p(x_m)}{\partial \theta_j} + \sum_m \sum_{n \neq m} \frac{\partial \ln p(x_m)}{\partial \theta_i} \frac{\partial \ln p(x_n)}{\partial \theta_j}$, the second term having an expectation of 0.]

- Point estimation

Criteria,

1. **Consistency**: $\underline{\hat{\theta}} \xrightarrow[N\to\infty]{P} \underline{\theta}^*$, ($\xrightarrow[N\to\infty]{P}$ means converging in probability as $N \to \infty$).

   Remark.

   If the statistics is large enough, the estimation gives the true values;

2. **Unbiasedness**: $\underline{b_{\hat{\theta}}} = 0$,

   Remark.

   If the experiment is repeated for many times, even with low statistics, the estimators can give the true values;

3. **Efficiency**: $\sigma^2_{\hat{\theta}_i}$ is as small as possible,

   Remark.

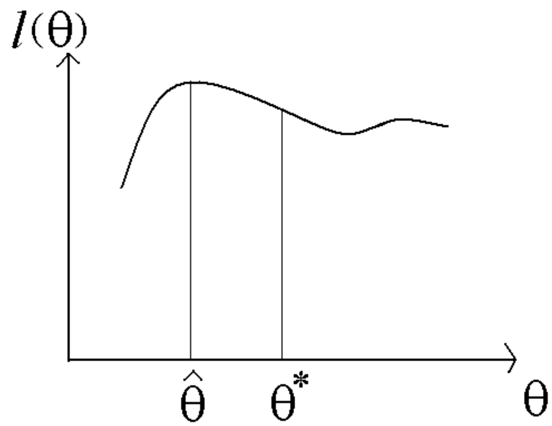
   The larger efficiency, the less repeated experiments are required.

...

Analysis

Maximum Likelihood Estimators:

the ones maximizing the likelihood function L.



Why? The one happened is the most probable?

Need analytical proofs of the criteria; common sense is NOT always correct.

Analysis

- Strong Law of Large Numbers

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \xrightarrow[N\to\infty]{P} \langle f(x)\rangle . \tag{17}$$

Remark.

– Large sample limit.

– $l = \sum_{i=1}^{N} \ln p\left(x_i|\underline{\theta}\right) \xrightarrow[N\to\infty]{P} N\left\langle \ln p\left(x|\underline{\theta}\right)\right\rangle$, independent of x in the large sample limit.

-

$$\left\langle \left.\frac{\partial \ln p}{\partial \underline{\theta}}\right|_{\underline{\theta}^*}\right\rangle = \underline{0}, \qquad \left\langle \left.\frac{\partial^2 \ln p}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right\rangle = -\left\langle \left(\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\theta_j}\right)_{\underline{\theta}^*}\right\rangle . \tag{18}$$

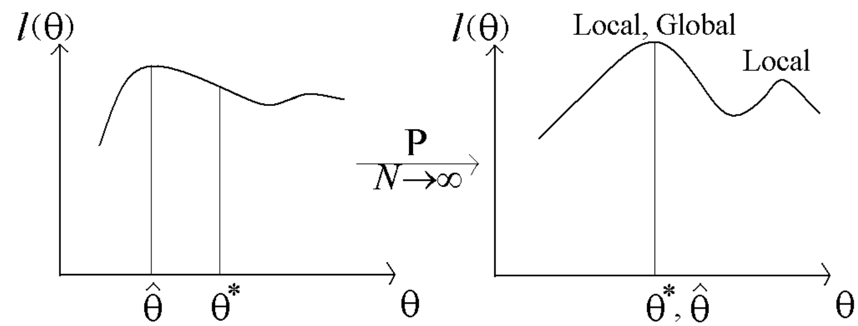
[Hint: $\left\langle \frac{\partial \ln p}{\partial \underline{\theta}}\Big|_{\underline{\theta}^*}\right\rangle = \int \left(\frac{\partial \ln p}{\partial \underline{\theta}} p\right)_{\underline{\theta}^*} \mathrm{d}x = \int \frac{\partial p}{\partial \underline{\theta}}\Big|_{\underline{\theta}^*} \mathrm{d}x = \left(\frac{\partial}{\partial \underline{\theta}} \int p \mathrm{d}x\right)_{\underline{\theta}^*}$ ($\mathbf{X}$ needs to be independent of the parameters), $\frac{\partial^2 \ln p}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\left(\frac{\partial \ln p}{\partial \theta_j}\right) = -\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j} + \frac{1}{p}\frac{\partial^2 p}{\partial \theta_i \partial \theta_j}$.]

Analysis

1. Consistency:

$$\underline{\hat{\theta}} \xrightarrow[N\to\infty]{P} \underline{\theta}^*$$

$$\Leftrightarrow \forall \underline{\delta} \neq \underline{0},\ l\left(\underline{\theta}^* + \underline{\delta}\right) < l\left(\underline{\theta}^*\right).$$



(19)

Locally ($\underline{\delta} \longrightarrow \underline{0}$), approximated to the second order,

$$l\left(\underline{\theta}^* + \underline{\delta}\right) = l\left(\underline{\theta}^*\right) + \left(\frac{\partial l}{\partial \underline{\theta}}\right)^T_{\underline{\theta}^*} \underline{\delta} + \frac{1}{2}\underline{\delta}^T \left(\left.\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right|_{\underline{\theta}^*}\right) \underline{\delta}. \tag{20}$$

$$\Rightarrow \begin{cases} \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}^*} = \underline{0}, \\ \text{Hessian } \underline{\underline{H}}\left(l\right) \equiv \left(\left.\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right|_{\underline{\theta}^*}\right) < 0 \qquad \text{(Second Derivative Test)} \\ (\textit{Reminder.}\ \text{negative definite, i.e. } \forall\ \underline{v} \neq \underline{0},\ \underline{v}^T \underline{\underline{H}}\ \underline{v} < 0). \end{cases} \tag{21}$$

Proof. of local maximum

$$l \xrightarrow[N\to\infty]{P} N\left\langle \ln p \right\rangle \tag{22}$$

$$\therefore \begin{cases} \frac{\partial l}{\partial \underline{\theta}}\Big|_{\theta^*} \xrightarrow[N\to\infty]{P} N\left\langle \frac{\partial \ln p}{\partial \underline{\theta}}\Big|_{\theta^*} \right\rangle = 0 \\ \\ H_{ij} \xrightarrow[N\to\infty]{P} N\left\langle \frac{\partial^2 \ln p}{\partial \theta_i \partial \theta_j}\Big|_{\underline{\theta}^*} \right\rangle = -N\left\langle \left(\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\partial \theta_j} \right)_{\underline{\theta}^*} \right\rangle . \end{cases} \tag{23}$$

Denote $\frac{\partial \ln p}{\partial \underline{\theta}}\big|_{\underline{\theta}^*} = \underline{d}\left(x\right)$, so

$$\underline{\underline{H}} \xrightarrow[N\to\infty]{P} -N\left\langle \underline{d}\ \underline{d}^T \right\rangle , \tag{24}$$

$$\forall \underline{v} \neq 0,\ \underline{v}^T \underline{\underline{H}}\ \underline{v} \xrightarrow[N\to\infty]{P} -N\left\langle \underline{v}^T \underline{d}\ \underline{d}^T \underline{v} \right\rangle \tag{25}$$

$$= -N\left\langle \left(\underline{v}^T \underline{d} \right)^2 \right\rangle < 0 \text{ if } \exists x' \in \mathbf{X},\ \underline{v}^T \underline{d}\left(x'\right) \neq 0. \tag{26}$$

□

Remark.

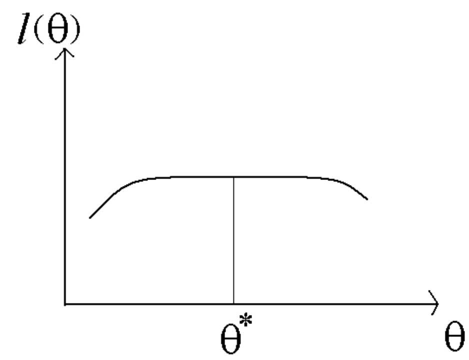
- For a proof of global maximum, please refer to Wald's (1949) proof, e.g. in Section 17.15 of *The advanced theory of statistics*, by M. Kendall et al., 4th ed. of Vol. 2 of the 3-volume ed. (1979) ISBN: 0 85264 255 5.

- This proof is **Conditional**; it fails if

$$\exists \underline{v} \neq 0,\ \forall x \in \mathbf{X},$$

$$\underline{v}^T \underline{d} = v_1 \frac{\partial \ln p}{\partial \theta_1}\Big|_{\underline{\theta}^*} + v_2 \frac{\partial \ln p}{\partial \theta_2}\Big|_{\underline{\theta}^*} + \cdots + v_M \frac{\partial \ln p}{\partial \theta_M}\Big|_{\underline{\theta}^*} = 0,$$

i.e. $\frac{\partial \ln p}{\partial \theta_1}\Big|_{\underline{\theta}^*},\ \frac{\partial \ln p}{\partial \theta_2}\Big|_{\underline{\theta}^*},\ \ldots,\ \frac{\partial \ln p}{\partial \theta_M}\Big|_{\underline{\theta}^*}$ are linearly dependent functions of x; in the single parameter case, $\frac{\partial \ln p}{\partial \theta}\Big|_{\underline{\theta}^*} = 0.$

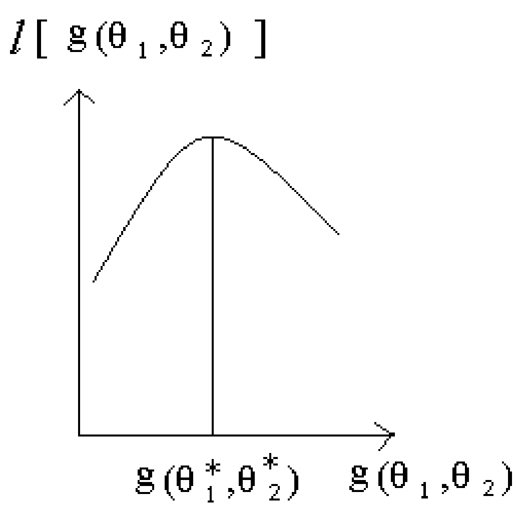


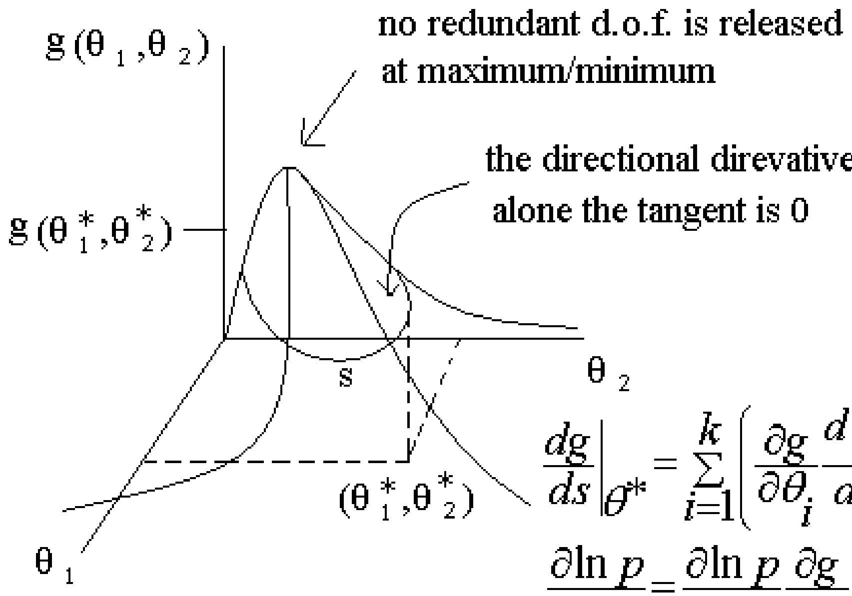
- The condition ensures that the maximum likelihood method is justified and imposes some **constraints on**
  - the **parameterization/construction of the p.d.f.**, e.g., the following are inappropriate:

  (a) $\forall x,\ \exists i,\ \frac{\partial p}{\partial \theta_i}|_{\underline{\theta}^*} = 0$;

  (b) $\forall x,\ \exists i,\ \frac{\partial p}{\partial \theta_i} \equiv 0$, that is, p is independent of θ_i, typically when normalizing the p.d.f.;

(c) $p\left(x|\theta_1,\ \theta_2,\ldots,\ \theta_M\right) = p\left[x|g\left(\theta_1,\ \theta_2,\ldots,\ \theta_k\right),\ \ldots\right],\ 2 \leq k \leq M$, that is, the dependence of $\theta_1,\ \theta_2,\ldots,\ \theta_k$ degenerates into a whole via $g\left(\theta_1,\ \theta_2,\ldots,\ \theta_k\right)$ and thus redundant degree(s) of freedom are(is) released. (Except that $g\left(\theta_1^*,\ \theta_2^*,\ldots,\ \theta_k^*\right)$ is maximum/minimum, which belongs to Situation a.)
Suppose $p\left(x|\theta_1,\ \theta_2\right) = \left[1 - \left(\theta_1^2 + \theta_2^2\right)x\right]/\mathrm{nor}$, and thus $g = \theta_1^2 + \theta_2^2$.
In maximizing $l = l\left[g\left(\theta_1,\ \theta_2\right)\right]$, the program enters the region $g\left\langle\theta_1,\ \theta_2\right) = g\left\langle\theta_1^*,\ \theta_2^*\right)$ where l has its largest value, but still can not decide which $\left(\theta_1,\ \theta_2\right)$ is the true one:



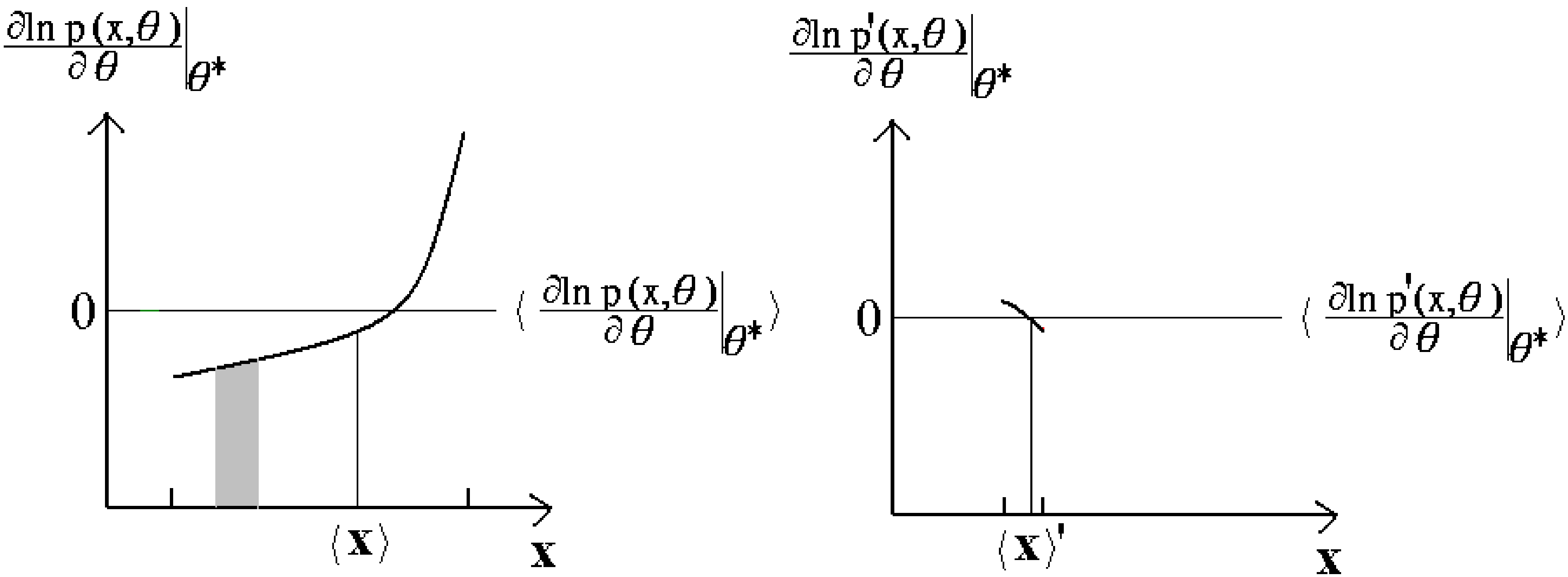


$$\left.\frac{dg}{ds}\right|_{\theta^*} = \sum_{i=1}^{k}\left(\frac{\partial g}{\partial \theta_i}\frac{d\theta_i}{ds}\right)_{\theta^*} = 0$$

$$\frac{\partial \ln p}{\partial \theta_i} = \frac{\partial \ln p}{\partial g}\frac{\partial g}{\partial \theta_i}$$

– the **selection of data samples**, e.g.,

the "spread" of x, measured by its $\mathrm{RMS} \equiv \sqrt{\left\langle (x - \langle x \rangle)^2 \right\rangle}$, should be large, since as $x \longrightarrow$ certain fixed value, $\forall x$, $\frac{\partial \ln p}{\partial \theta_i} \longrightarrow \left\langle \frac{\partial \ln p}{\partial \theta_i} \right\rangle = 0$; and thus $\forall x$, $\frac{\partial \ln p}{\partial \theta_i} \equiv 0$ holds approximately, making the method unjustified. (This constrain holds no matter whether $\langle x \rangle$ is zero.)



– the **grouping of the data sample**:
The data set can be divided into several, say K, sub-sets, each of which has its p.d.f.: $p_1\left(x|\underline{\theta}\right),\ p_2\left(x|\underline{\theta}\right),\ \ldots,\ p_K\left(x|\underline{\theta}\right)$, and correspondingly $L_J = \prod_i p_J\left(x_i|\underline{\theta}\right)$, $J = 1,\ 2,\ \ldots,\ K$. For the whole data set, the likelihood function reads

$$L_s = L_1 L_2 \cdots Ł_K \tag{27}$$

$$\therefore l_s = l_1 + l_2 + \cdots + l_K \tag{28}$$

$$= \sum_i^{N_1} \ln p_1(x_i|\underline{\theta}) + \sum_i^{N_2} \ln p_2(x_i|\underline{\theta}) + \cdots + \sum_i^{N_K} \ln p_K(x_i|\underline{\theta}) \tag{29}$$

$$\neq l = \left(\sum_i^{N_1} + \sum_i^{N_2} + \cdots + \sum_i^{N_K} \right) \ln p(x_i|\underline{\theta}),\ \text{since generally } p_J \neq p. \tag{30}$$

One can verify (by following the same approaches as above) that maximizing l_s also gives consistency results conditionally and since

$$[H_s]_{mn} \xrightarrow[N\to\infty]{P} -\sum_{J=1}^{K} N_J \left\langle \left(\frac{\partial \ln p_J}{\partial \theta_m} \frac{\partial \ln p_J}{\partial \theta_n} \right)_{\underline{\theta}^*} \right\rangle, \tag{31}$$

the grouping may be unjustified if all the sub-sets are close to any of the above inappropriate situations.

2. Bias and Efficiency

Using Taylor's Theorem,

$$\left.\frac{\partial l}{\partial \theta_i}\right|_{\hat{\underline{\theta}}} = \left.\frac{\partial l}{\partial \theta_i}\right|_{\underline{\theta}^*} + \sum_j^M \left.\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^\triangle} \left(\hat{\theta}_j - \theta_j^*\right) \Rightarrow \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\hat{\underline{\theta}}} = \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}^*} + \left(\left.\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^\triangle}\right) \left(\hat{\underline{\theta}} - \underline{\theta}^*\right), \quad (32)$$

where $\underline{\theta}^\triangle$ are some values between $\hat{\underline{\theta}}$ and $\underline{\theta}^*$. Since $\frac{\partial l}{\partial \underline{\theta}}|_{\hat{\underline{\theta}}} = 0$ and $\underline{\theta}^\triangle \xrightarrow[N\to\infty]{P} \underline{\theta}^*$,

$$\because \begin{cases} \left(\hat{\underline{\theta}} - \underline{\theta}^*\right) = \left(-\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}|_{\underline{\theta}^\triangle}\right)^{-1} \frac{\partial l}{\partial \underline{\theta}}|_{\underline{\theta}^*}, \text{ if } \left(-\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}|_{\underline{\theta}^\triangle}\right) \text{ is nonsingular} \\ \left(-\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}|_{\underline{\theta}^\triangle}\right) \xrightarrow[N\to\infty]{P} \left(-\left.\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right), \text{ independent of } x, \text{ nonsingular} \\ \qquad\qquad\text{(assuming the condition for consistency holds)} \end{cases} \quad (33)$$

$$(34)$$

$$\underline{b}_{\hat{\underline{\theta}}} = \left\langle \left(-\left.\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^\triangle}\right)^{-1} \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}^*} \right\rangle \xrightarrow[N\to\infty]{P} \left(-\left.\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1} \left\langle \left.\frac{\partial l}{\partial \underline{\theta}}\right|_{\underline{\theta}^*} \right\rangle = \underline{0}. \quad (35)$$

Similarly, expanding at $\left\langle \underline{\hat{\theta}} \right\rangle$, with $\underline{\theta}_2^{\triangle}$ between $\underline{\hat{\theta}}$ and $\left\langle \underline{\hat{\theta}} \right\rangle$,

$$\underline{\underline{V}}_{\underline{\hat{\theta}}} = \left\langle \left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \Big|_{\underline{\theta}_2^{\triangle}} \right)^{-1} \left(\frac{\partial l}{\partial\theta_i} \frac{\partial l}{\partial\theta_j} \Big|_{\left\langle \underline{\hat{\theta}} \right\rangle} \right) \left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \Big|_{\underline{\theta}_2^{\triangle}} \right)^{-1} \right\rangle \tag{36}$$

$$\xrightarrow[N\to\infty]{P} \left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \Big|_{\underline{\theta}^*} \right)^{-1} \left(\left\langle \frac{\partial l}{\partial\theta_i} \frac{\partial l}{\partial\theta_j} \Big|_{\underline{\theta}^*} \right\rangle \right) \left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \Big|_{\underline{\theta}^*} \right)^{-1}. \tag{37}$$

$$\underline{\underline{V}}_{\underline{\hat{\theta}}} \xrightarrow[N\to\infty]{P} \frac{1}{N} \left(\left\langle \frac{\partial \ln p}{\partial\theta_i} \frac{\partial \ln p}{\partial\theta_j} \Big|_{\underline{\theta}^*} \right\rangle \right)^{-1} = \left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \Big|_{\underline{\theta}^*} \right)^{-1}. \tag{38}$$

Remark.

- $\underline{\hat{\theta}}$ is not unbiased. As N tends to infinity, the biases tend to zero and $\underline{\underline{V}}_{\underline{\hat{\theta}}}$ tends to the MVB.
- $\underline{\underline{V}}_{\underline{\hat{\theta}}}$ can be estimated by $\left(- \frac{\partial^2 l}{\partial\theta_i \partial\theta_j} \big|_{\underline{\hat{\theta}}} \right)^{-1}$ in the large sample limit.

MC Examples

1. MINUIT
   - Searching *local* minima of FCN as a function of par: $\texttt{FCN} = -l$, $\texttt{par}_i = \theta_i$;
   - **Error matrix**: $2\left(\frac{\partial^2\texttt{FCN}}{\partial\texttt{par}_i\partial\texttt{par}_j}\right)^{-1} \times \texttt{UP} : \texttt{UP} = 0.5$;
   - The MINUIT command

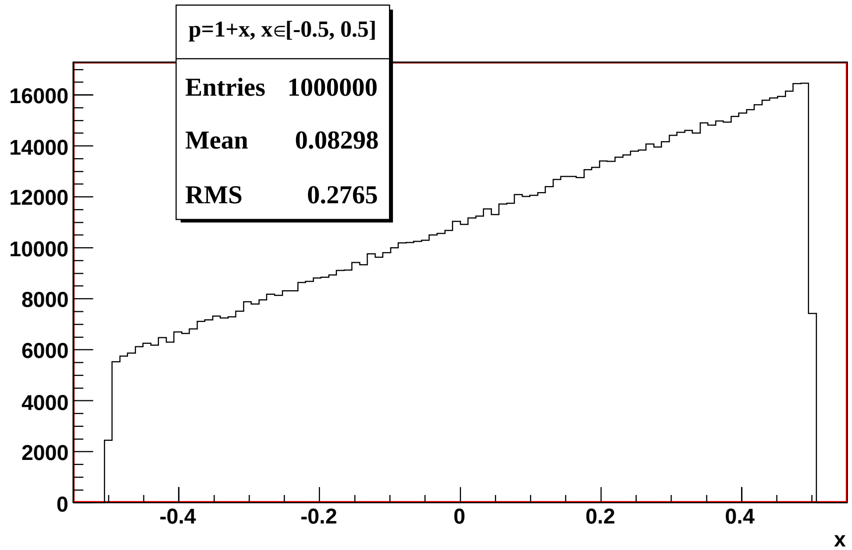
     gMinuit->mnexcm("MIGRAD", arglist ,2, ierflg)

     searches a *local* minimum of FCN and produces the error matrix.
2. $p = (1 + \theta x)\,/\mathrm{nor}$

   The prototype-p.d.f. in asymmetry analysis. E.g., in beam-spin asymmetry, $x = \mathrm{beam\ pol.}$, $\theta = A_{LU}^{c0}$.

MC Examples

- Inappropriate parameterization $\frac{\partial p}{\partial \theta_i}|_{\underline{\theta}^*} = 0$:



1. $p = 1 + (1+\theta)\,x,\ \ \theta^* = 0$: $\frac{\partial p}{\partial \theta}\Big|_{\theta^*} = x$, $\hat{\theta} = (-4.5 \pm 3.2) \times 10^{-3}$;
2. $p = 1 + \left(1+\theta^2\right)x,\ \ \theta^* = 0$: $\frac{\partial p}{\partial \theta}\Big|_{\theta^*} = 0$, $\hat{\theta} = (0.2 \pm 33.4) \times 10^{-3}$.

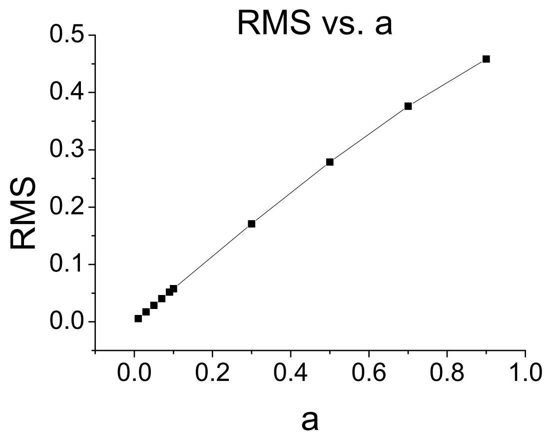
- The smaller RMS, the larger error:
  - MC samples: $$p = \frac{1+x}{\mathrm{nor}},\ x \in [x_0 - a, x_0 + a],\ x_0 = 0.1,\ N = 10^6;$$ varying a (N fixed) to get samples with different RMS.

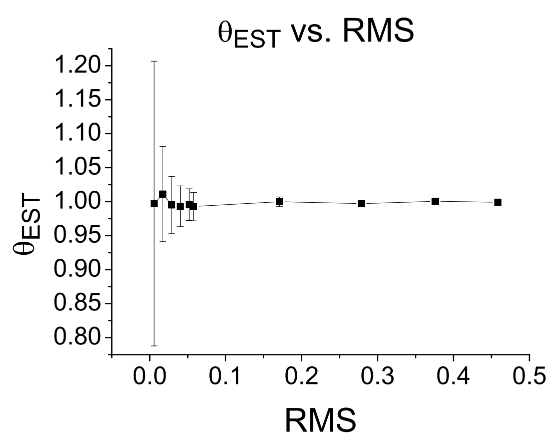
    (Cf. *RMS vs. a*).
  - Parameterization: $$p = \frac{1+\theta x}{\mathrm{nor}(\theta)},\ \theta^* = 1.$$

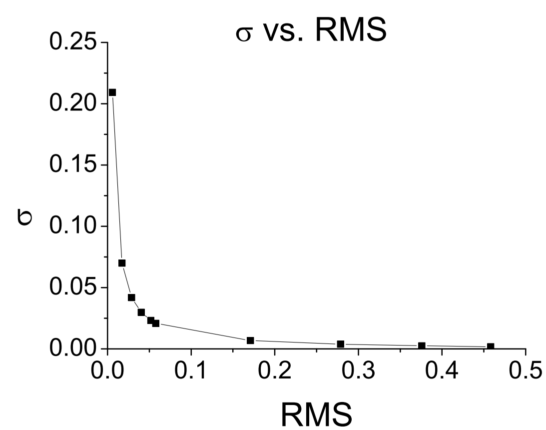
    (Cf. θ_{EST} *vs. RMS* for fitting results.)
  - Conclusion:

    The error gets larger with smaller RMS.

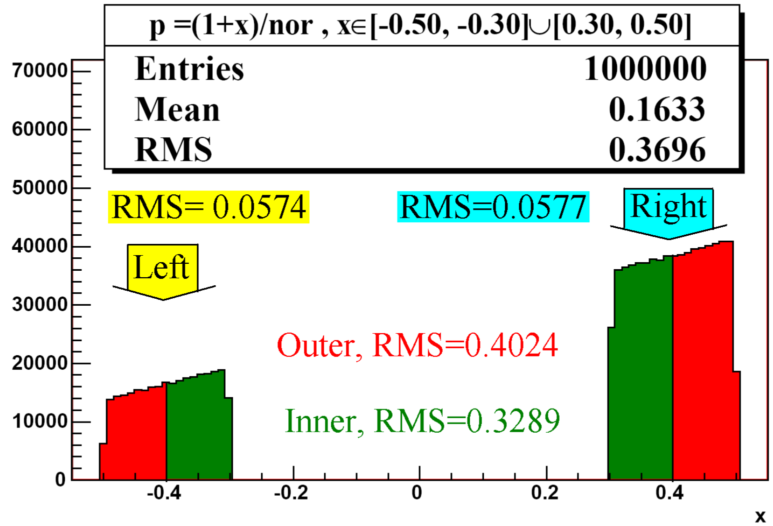
    (Cf. σ *vs. RMS*).







- Inappropriate data grouping:



  - Parameterization: $p = \frac{1+\theta x}{\mathrm{nor}(\theta)},\ \ \theta^* = 1;$
  - Results:

	no grouping	Left-Right	Inner-Outer
$\hat{\theta}$	1.0003±0.0022	1.0146±0.0107	1.0003±0.0022

  - Conclusion: inappropriate data grouping can lead to larger fitting errors.

(This explains why in BSA analysis, the constant term of the asymmetry by $l_s = l_{\text{beam pol.}>0} + l_{\text{beam pol.}<0}$ has much larger error.)

Summary for Part 1

1. Introduce *useful formulas* to deal with theoretical problems of maximum likelihood;
2. Conclude the *condition* for the justified maximum likelihood method, illustrate with MC examples and explain certain issue on practical data analysis;
3. Show the *standard procedure to obtain the covariance matrix* of the estimators, which will be used directly in the following parts.

Part 2:

WEIGHTS
&
THE EXTENDED

Weights – Concepts

- Weight Function $w(x)$:
  the multiplicity of x;
  - *Naturally* $w(x) = p(x)$ ($\mathrm{Prob}(x)$ if x is discrete), e.g. $N = 3$,
    $$x_1 = A,\ x_2 = A,\ x_3 = B,$$
    $w_0(A) = 2/3$. (may not be precise since N is very small)
  - *Artificially* weights – any non-negative number – can be assigned to x: "**weighting**", e.g. assigning weights $w(A) = 3,\ w(B) = 5$ to *each data point*, we have a "virtual" data sample:

$$x_1^1 = A,\ x_1^2 = A,\ x_1^3 = A, \tag{39}$$

$$x_2^1 = A,\ x_2^2 = A,\ x_2^3 = A, \tag{40}$$

$$x_3^1 = B,\ x_3^2 = B,\ x_3^3 = B,\ x_3^4 = B,\ x_3^5 = B. \tag{41}$$

$\mathrm{Prob}_w(A) = \frac{\mathrm{Prob}(A)w(A)}{\mathrm{Prob}(A)w(A)+\mathrm{Prob}(B)w(B)} = \frac{6}{11}$, $N_w = w(x_1) + w(x_2) + w(x_3) = 11.$

– Weighted p.d.f.

$$\mathrm{Prob}_w\left(A\right) = \frac{\mathrm{Prob}\left(A\right) w\left(A\right)}{\mathrm{Prob}\left(A\right) w\left(A\right) + \mathrm{Prob}\left(B\right) w\left(B\right)} \tag{42}$$

$$\Rightarrow p_w\left(x|\underline{\theta}\right) \equiv \frac{p\left(x|\underline{\theta}\right) w\left(x\right)}{\int_{\mathbf{X}} p\left(x|\underline{\theta}\right) w\left(x\right) \mathrm{d}x}. \tag{43}$$

– Weighted likelihood function

$$N_w = w\left(x_1\right) + w\left(x_2\right) + w\left(x_3\right) \tag{44}$$

$$\Rightarrow L_w\left(x_1,\ x_2, \ldots,\ x_N|\underline{\theta}\right) \equiv \prod_{i=1}^{N} \left[p_w\left(x_i|\underline{\theta}\right)\right]^{w\left(x_i\right)} \tag{45}$$

– Weighted log-likelihood function

$$l_w\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \equiv \ln L_w\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}\right) \tag{46}$$

$$= \sum_{i=1}^{N} w\left(x_i\right) \ln p_w\left(x_i|\underline{\theta}\right) \tag{47}$$

$$\xrightarrow[N\to\infty]{P} N\left\langle w\left(x\right) \ln p_w\left(x|\underline{\theta}\right)\right\rangle \tag{48}$$

Remark.
Since by $\sum_i^N$ we can only deal with real data, by the Strong Law of Large Numbers, the definitions of expectations do NOT change:

$$\langle f(x)\rangle \equiv \int_{\mathbf{X}} f(x) p\left(x|\underline{\theta}^*\right) \mathrm{d}x, \tag{49}$$

$$\langle g(x_1,\ x_2,\ldots,\ x_N)\rangle \equiv$$

$$\int\cdots\int_{\mathbf{X}^N} g(x_1,\ x_2,\ldots,\ x_N) L\left(x_1,\ x_2,\ldots,\ x_N|\underline{\theta}^*\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_N. \tag{50}$$

Weights – Analysis

1.

$$\left(\textit{Reminder.} \quad \left\langle \frac{\partial \ln p}{\partial \underline{\theta}}\bigg|_{\underline{\theta}^*} \right\rangle = \underline{0}, \qquad \left\langle \frac{\partial^2 \ln p}{\partial \theta_i \partial \theta_j}\bigg|_{\underline{\theta}^*} \right\rangle = -\left\langle \left(\frac{\partial \ln p}{\partial \theta_i}\frac{\partial \ln p}{\theta_j} \right)_{\underline{\theta}^*} \right\rangle . \right) \tag{51}$$

$$\left\langle w \frac{\partial \ln p_w}{\partial \underline{\theta}}\bigg|_{\underline{\theta}^*} \right\rangle = 0, \qquad \left\langle w \frac{\partial^2 \ln p_w}{\partial \theta_i \partial \theta_j}\bigg|_{\underline{\theta}^*} \right\rangle = -\left\langle w \left(\frac{\partial \ln p_w}{\partial \theta_i}\frac{\partial \ln p_w}{\partial \theta_j} \right)_{\underline{\theta}^*} \right\rangle . \tag{52}$$

[Hint: $\left\langle w \frac{\partial \ln p_w}{\partial \underline{\theta}}\big|_{\underline{\theta}^*} \right\rangle = \int w \left(\frac{\partial \ln p_w}{\partial \underline{\theta}} p \right)_{\underline{\theta}^*} \mathrm{d}x = \left(\int p w \mathrm{d}x \cdot \frac{\partial}{\partial \underline{\theta}} \int p_w \mathrm{d}x \right)_{\underline{\theta}^*} .$]

2. Consistency *fails* if

$$\left(\textit{Reminder.} \quad \frac{\partial \ln p}{\partial \theta_1}\bigg|_{\underline{\theta}^*}, \ \frac{\partial \ln p}{\partial \theta_2}\bigg|_{\underline{\theta}^*}, \ \ldots, \ \frac{\partial \ln p}{\partial \theta_M}\bigg|_{\underline{\theta}^*}, \ \text{are linearly dependent functions of } x. \right) \tag{53}$$

$$\frac{\partial \ln p_w}{\partial \theta_1}\bigg|_{\underline{\theta}^*}, \ \frac{\partial \ln p_w}{\partial \theta_2}\bigg|_{\underline{\theta}^*}, \ \ldots, \ \frac{\partial \ln p_w}{\partial \theta_M}\bigg|_{\underline{\theta}^*} \ \text{are linearly dependent functions of } x. \tag{54}$$

3. (*Reminder.* $\underline{\underline{V}}_{\underline{\hat{\theta}}} \xrightarrow[N\to\infty]{P} \left(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} \left(\left\langle \frac{\partial l}{\partial\theta_i}\frac{\partial l}{\partial\theta_j}\Big|_{\underline{\theta}^*}\right\rangle\right) \left(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} = \left(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1}$.)

$$\underline{\underline{V}}_{\underline{\hat{\theta}}_w} \xrightarrow[N\to\infty]{P} \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} \left(\left\langle \frac{\partial l_w}{\partial\theta_i}\frac{\partial l_w}{\partial\theta_j}\Big|_{\underline{\theta}^*}\right\rangle\right) \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} \tag{55}$$

$$= \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} \left(\sum_{k=1}^{N} w^2(x_k) \left[\frac{\partial \ln p_w(x_k|\underline{\theta})}{\partial\theta_i}\frac{\partial \ln p_w(x_k|\underline{\theta})}{\partial\theta_j}\right]_{\underline{\theta}^*}\right) \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1}. \tag{56}$$

Remark.

(a) Suppose $w(x) \equiv w_0 > 0$ constant,

$$p_w = p,\; l_w = w_0 l,\; N_w = \sum_i^N w_0 = w_0 N, \tag{57}$$

$$\Rightarrow \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1} = \frac{1}{w_0 N}\left(\left\langle \frac{\partial \ln p}{\partial\theta_i}\frac{\partial \ln p}{\partial\theta_j}\Big|_{\underline{\theta}^*}\right\rangle\right)^{-1} = \frac{1}{N_w}\left(\left\langle \frac{\partial \ln p}{\partial\theta_i}\frac{\partial \ln p}{\partial\theta_j}\Big|_{\underline{\theta}^*}\right\rangle\right)^{-1}. \tag{58}$$

$\left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*}\right)^{-1}$, defined as the covariance matrix by MINUIT, which takes into account the artificially arbitrary statistical inflation, is *not* the corresponding matrix in Eq. 56.

(b) If, e.g., $\frac{\partial^2 p_w}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}^*} = 0$,

$$\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_k^N w^2\left(x_k\right)\ln p_w\left(x_k|\underline{\theta}\right)\right]_{\underline{\hat{\theta}}} \xrightarrow[N\to\infty]{P} \sum_{k=1}^N w^2\left(x_k\right)\left[\frac{\partial \ln p_w\left(x_k|\underline{\theta}\right)}{\partial\theta_i}\frac{\partial \ln p_w\left(x_k|\underline{\theta}\right)}{\partial\theta_j}\right]_{\underline{\theta}^*}. \tag{59}$$

This provides a method to estimate the covariance matrix in Eq. 56.

[Hint: $\langle w^2\frac{\partial^2 \ln p_w}{\partial\theta_i\partial\theta_j}|_{\underline{\theta}^*}\rangle = -\langle w^2(\frac{\partial \ln p_w}{\partial\theta_i}\frac{\partial \ln p_w}{\partial\theta_j})_{\underline{\theta}^*}\rangle + \langle w^2(\frac{1}{p_w}\frac{\partial^2 p_w}{\partial\theta_i\partial\theta_j})_{\underline{\theta}^*}\rangle$,
$\langle w^2(\frac{1}{p_w}\frac{\partial^2 p_w}{\partial\theta_i\partial\theta_j})_{\underline{\theta}^*}\rangle = \int w^2(p\frac{1}{p_w}\frac{\partial^2 p_w}{\partial\theta_i\partial\theta_j})_{\underline{\theta}^*}\mathrm{d}x = (\int pw\mathrm{d}x)\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}(\int wp_w\mathrm{d}x) \neq 0$ generally.]

(c) Generally the weighted covariance matrix in Eq. 56 does *not* equal to the matrix by the standard method, $\underline{\underline{V}}_{\hat{\theta}_w} \neq \underline{\underline{V}}_{\hat{\theta}}$. The difference arises from the inhomogeneity of the weights.

$$\underline{\underline{V}}_{\hat{\theta}_{w\equiv w_0}} = \underline{\underline{V}}_{\hat{\theta}}, \qquad \underline{\underline{V}}_{\hat{\theta}_{k\cdot w}} = \underline{\underline{V}}_{\hat{\theta}_w}, \quad k > 0 \text{ constant.}$$

Example:
MC sample: $p\left(x\right) = \frac{1+x}{\mathrm{nor}}$, $x \in \mathbf{X}$; $p\left(x|\theta\right) = \frac{1+\theta x}{\mathrm{nor}(\theta)}$, $\theta^* = 1$, $w\left(x\right) = x$. For different $\mathbf{X}$ the homogeneity of $w\left(x\right)$ changes and thus $\sigma_{\hat{\theta}_w} \simeq \sigma_{\hat{\theta}}$ does not hold in the same approximation level:

$\mathbf{X}$	$\hat{\theta}_{w(x)=x}$	$\hat{\theta}$
$[0.00,\ 0.20]$	0.9785 ± 0.0245	0.9928 ± 0.0209
$[0.07,\ 0.13]$	1.0000 ± 0.0698	1.0113 ± 0.0700

Since $\underline{\underline{V}}_{\hat{\theta}}$ is the minimum variance bound, the larger the difference between $\underline{\underline{V}}_{\hat{\theta}_w}$ and $\underline{\underline{V}}_{\hat{\theta}}$ is, the worse is the estimation of this weighted method. Only when

$$\frac{w_i}{w_j} \simeq 1, \ \forall \ i \neq j, \tag{60}$$

the estimation is acceptable.

Weights – Application

1. Eliminating the parameter-dependence of the normalization factor:
   - If $p\left(x|\underline{\theta}\right) = f\left(\underline{\theta}\right) P\left(x|\underline{\theta}\right)$,
   $$p\left(x|\underline{\theta}\right) = \frac{P\left(x|\underline{\theta}\right)}{\int_{\mathbf{X}} P\left(x|\underline{\theta}\right) \mathrm{d}x}. \quad (61)$$
     $P\left(x|\underline{\theta}\right)$, the **Extended p.d.f.**;
     $\mathrm{nor}\left(\underline{\theta}\right) \equiv \int_X P\left(x|\underline{\theta}\right) \mathrm{d}x$, the **normalization factor**.
     Usually $P\left(x|\underline{\theta}\right)$ can be easily obtained while $\mathrm{nor}\left(\underline{\theta}\right)$ needs more complex calculation. A parameter-$independent$ nor can be neglected in the minimization procedure
     ($\frac{\partial \ln p(x|\underline{\theta})}{\partial \underline{\theta}} = \frac{\partial \ln P(x|\underline{\theta})}{\partial \underline{\theta}} - \frac{\partial \ln \mathrm{nor}}{\partial \underline{\theta}}$, $\frac{\partial \ln \mathrm{nor}}{\partial \underline{\theta}} = 0$).
   - $$p_w\left(x|\underline{\theta}\right) = \frac{w\left(x\right) P\left(x|\underline{\theta}\right)}{\int_{\mathbf{X}} w\left(x\right) P\left(x|\underline{\theta}\right) \mathrm{d}x}. \quad (62)$$
     $w\left(x\right)$ can be suitably chosen so that $\mathrm{nor}_w\left(\underline{\theta}\right) \equiv \int_{\mathbf{X}} w\left(x\right) P\left(x|\underline{\theta}\right) \mathrm{d}x$ is independent of $\underline{\theta}$ and thus equivalently $l_w = \sum_i w\left(x_i\right) \ln P\left(x_i|\underline{\theta}\right)$.
     Remark.
     Since we would not trade "efficiency" (the criterion!) for convenience, we require that the weights should be as homogenous as possible.

- Example

  MC sample: $\mathrm{Prob}\left(x\right)=\frac{1+x}{\mathrm{nor}},\ x\in\left\{-0.3,\ 0.2\right\},\ N=10^{6}$, which indicates

$$N_{-0.3}=\frac{1-0.3}{1-0.3+1+0.2}N=368,421,\ N_{0.2}=\frac{1+0.2}{1-0.3+1+0.2}N=631,579.$$

  Method 1: Simple SML fitting, i.e. without weight:

$$\mathrm{Prob}\left(x|\theta\right)=\frac{1+\theta x}{\mathrm{nor}\left(\theta\right)},\ \theta^{*}=1;\ \hat{\theta}=1.0000\pm0.0035. \tag{63}$$

  Method 2: Weighting:

  (a) $\mathrm{nor}_{w}\left(\theta\right)=\sum_{x}\left(1+\theta x\right)w\left(x\right)=$
  $w\left(-0.3\right)+w\left(0.2\right)+\theta\left[-0.3w\left(-0.3\right)+0.2w\left(0.2\right)\right]$,

  we can choose $w\left(-0.3\right)=2,\ w\left(0.2\right)=3$ so that $\mathrm{nor}_{w}=5$, independent of θ.

$$l_{w}\left(\theta\right)=\sum_{i=1}^{N}w\left(x_{i}\right)\ln\left(1+\theta x_{i}\right),\ \theta^{*}=1;\ \hat{\theta}_{w}=1.0000\pm0.0023. \tag{64}$$

  This error is given by $\left(-\frac{\partial^{2}l_{w}}{\partial\theta^{2}}\right)_{\hat{\theta}_{w}}^{-1}$, which is *not* the corresponding weighted error (cf. Weights – Analysis Remark. 3a).

(b) Since

$$\text{Prob}_w\left(\underline{\theta}\right) \propto (1+\theta x)\,,\ \frac{\partial^2 \text{Prob}_w}{\partial \theta^2} = 0, \tag{65}$$

we can use

$$l_{w^2}(\theta) = \sum_{i=1}^{N} w^2(x_i) \ln(1+\theta x_i) \tag{66}$$

to evaluate the weighted error (cf. Weights – Analysis Remark 3b):

$$\sigma^2_{\hat{\theta}_w} = \frac{-\frac{\partial^2 l_{w^2}}{\partial \theta^2}\Big|_{\hat{\theta}_w}}{\left(-\frac{\partial^2 l_w}{\partial \theta^2}\Big|_{\hat{\theta}_w}\right)^2}, \qquad \sigma_{\hat{\theta}_w} = 0.0035. \tag{67}$$

(c) Because the weights $w(-0.3) = 2,\ w(0.2) = 3$ are approximately homogenous, this estimation is acceptable (cf. Weights – Analysis Remark 3c).

(d) To get the correct $\sigma^2_{\hat{\theta}_w}$ (Eq. 67), one needs

i. "MIGRAD" $-l_w$ to estimate $\hat{\theta}_w$ and record the output error σ_M;

ii. Squared the weights;

iii. "HESSE" $-l_{w^2}$ and record the output error σ_H;
The MINUIT command

gMinuit->mnexcm("HESSE", arglist, 0, ierflg)

calculates the error matrix with the *current* parameter values *without minimization*.

iv. Finally, $\sigma_{\hat{\theta}_w} = \sigma^2_M / \sigma_H$.

v. For multi-parameter cases, one just needs to use the covariance matrix instead and follow $\underline{\underline{V}}_{\hat{\underline{\theta}}} \xrightarrow[N\to\infty]{P} \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\underline{\theta}}}\right)^{-1} \left(-\frac{\partial^2 l_{w^2}}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\underline{\theta}}}\right) \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\underline{\theta}}}\right)^{-1}$.

(e) The condition Eq. 65 is *important*. One can see that by the same procedure as above using another parameterization $\mathrm{Prob}_w(\underline{\theta}) \propto \left(1+\theta^2 x\right) \quad \left(\frac{\partial^2 \mathrm{Prob}_w}{\partial\theta^2} \propto x \neq 0\right)$, which makes the estimation in Weights – Analysis Remark 3b unjustified. Since the weights are approximately homogeneous, we expect $\sigma_{\hat{\theta}_w} = \sigma_{\hat{\underline{\theta}}}$ if the weighted covariance matrix is correctly estimated, which however contradicts with the following results:

i. No weighting: $\mathrm{Prob}(x|\theta) = \frac{1+\theta^2 x}{\mathrm{nor}(\theta)}$, $\theta^* = 1$; $\hat{\theta} = 1.0000 \pm 0.0017$.

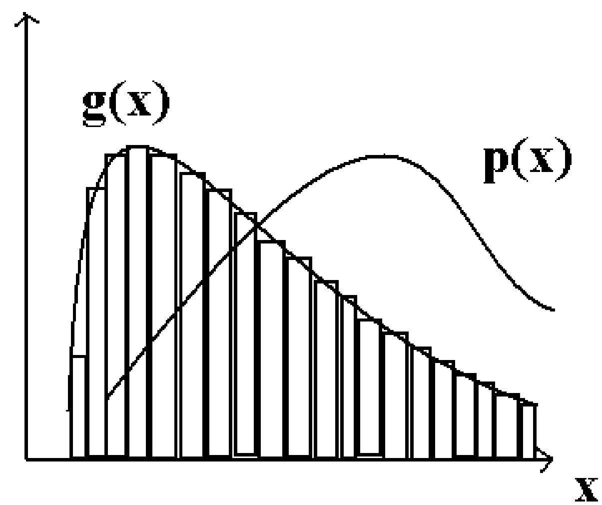
ii. Weighting: $l_w(\theta) = \sum_{i=1}^N w(x_i) \ln\left(1+\theta^2 x_i\right)$, $\theta^* = 1$; $\hat{\theta}_w = 1.0000 \pm 0.0014$ (final).

2. Deal with observation efficiency:

- **Efficiency**, the probability that a process with x is observed, $e(x)$.

$$p(x|\underline{\theta}), \text{ p.d.f. for the } \textit{happened} \text{ processes,}$$
$$\int_{\mathbf{X}} p(x|\underline{\theta})\,\mathrm{d}x = 1;$$

$$g(x|\underline{\theta}), \text{ p.d.f. for the } \textit{observed} \text{ processes,}$$
$$g(x|\underline{\theta}) \equiv \frac{e(x)p(x|\underline{\theta})}{\int_{\mathbf{X}} e(x)p(x|\underline{\theta})\mathrm{d}x}. \tag{68}$$



- If we need to investigate the observed data sample but g is not known analytically, we can weight the observed data to generate a virtual one – the one from the happened processes – and then try to obtain information corresponding to the observed.

Let $w(x) = \frac{1}{e(x)}$,

$$g_w = \frac{wep}{\int wep \mathrm{d}x} = p, \tag{69}$$

$$l_w = \sum_{i=1}^{N} w(x_i) \ln g_w(x_i|\underline{\theta}) = \sum_{i=1}^{N} w(x_i) \ln p(x_i|\underline{\theta}), \tag{70}$$

$$\underline{\underline{V}}_{\underline{\hat{\theta}}_w} \xrightarrow[N\to\infty]{P} \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\bigg|_{\underline{\hat{\theta}}}\right)^{-1} \left(\sum_{k=1}^{N} w^2(x_k) \left[\frac{\partial \ln p(x_k|\underline{\theta})}{\partial\theta_i}\frac{\partial \ln p(x_k|\underline{\theta})}{\partial\theta_j}\right]_{\underline{\hat{\theta}}}\right) \left(-\frac{\partial^2 l_w}{\partial\theta_i\partial\theta_j}\bigg|_{\underline{\hat{\theta}}}\right)^{-1}. \tag{71}$$

Again, this method is inappropriate if $e(x)$ is severely inhomogeneous.

The Extended

- If $P\left(x|\underline{\theta}\right)$ is so chosen that $\mathbb{N}\left(\underline{\theta}\right)\left[\equiv \mathrm{nor}\left(\underline{\theta}\right)=\int P\mathrm{d}x\right]$ is the **theoretical number of the observed events**, by assuming that N follows a Poisson distribution,

$$\frac{e^{-\mathbb{N}(\underline{\theta})}\left[\mathbb{N}\left(\underline{\theta}\right)\right]^N}{N!}, \tag{72}$$

we have the **extended likelihood function**,

$$L_{\mathrm{ext}}=\frac{e^{-\mathbb{N}(\underline{\theta})}\left[\mathbb{N}\left(\underline{\theta}\right)\right]^N}{N!}\prod_{i=1}^{N}p\left(x_i|\underline{\theta}\right)=\frac{e^{-\mathbb{N}(\underline{\theta})}\left[\mathbb{N}\left(\underline{\theta}\right)\right]^N}{N!}\prod_{i=1}^{N}\frac{P\left(x_i|\underline{\theta}\right)}{\mathbb{N}\left(\underline{\theta}\right)}, \tag{73}$$

and thus the **extended log-likelihood function** (equivalently),

$$l_{\mathrm{ext}}=\sum_{i=1}^{N}\ln P\left(x_i|\underline{\theta}\right)-\mathbb{N}\left(\underline{\theta}\right). \tag{74}$$

- The weighted extended likelihood function:

$$L_{\mathrm{ext},w} = \frac{e^{-\mathbb{N}_w(\underline{\theta})}\left[\mathbb{N}_w\left(\underline{\theta}\right)\right]^{N_w}}{N_w!}\prod_{i=1}^{N}\left[p_w\left(x_i|\underline{\theta}\right)\right]^{w\left(x_i\right)} \tag{75}$$

$$= \frac{e^{-\mathbb{N}_w(\underline{\theta})}\left[\mathbb{N}_w\left(\underline{\theta}\right)\right]^{N_w}}{N_w!}\prod_{i=1}^{N}\left[\frac{w\left(x_i\right)P\left(x_i|\underline{\theta}\right)}{\mathbb{N}_w\left(\underline{\theta}\right)}\right]^{w\left(x_i\right)}, \tag{76}$$

where $N_w = \sum_{i=1}^{N} w\left(x_i\right)$.

-

$$N\left\langle w\frac{\partial \ln P}{\partial \underline{\theta}}\bigg|_{\underline{\theta}^*}\right\rangle - \frac{\partial \mathbb{N}_w}{\partial \underline{\theta}}\bigg|_{\underline{\theta}^*} \xrightarrow[\mathbb{N}\to\infty]{P} 0, \tag{77}$$

$$N\left\langle w\frac{\partial^2 \ln P}{\partial \theta_i \partial \theta_j}\bigg|_{\underline{\theta}^*}\right\rangle - \frac{\partial^2 \mathbb{N}_w}{\partial \theta_i \partial \theta_j}\bigg|_{\underline{\theta}^*} \xrightarrow[\mathbb{N}\to\infty]{P} -N\left\langle w\left(\frac{\partial \ln P}{\partial \theta_i}\frac{\partial \ln P}{\partial \theta_j}\right)_{\underline{\theta}^*}\right\rangle. \tag{78}$$

[Hint: $\frac{N}{\mathbb{N}} - 1 \xrightarrow[\mathbb{N}\to\infty]{P} 0$.]

Remark.

For unweighted case, let $w\left(x\right) = 1$.

- Consistency *fails* if

$$\left.\frac{\partial \ln P}{\partial \theta_1}\right|_{\underline{\theta}^*}, \left.\frac{\partial \ln P}{\partial \theta_2}\right|_{\underline{\theta}^*}, \ldots, \left.\frac{\partial \ln P}{\partial \theta_M}\right|_{\underline{\theta}^*} \text{ are linearly dependent.} \tag{79}$$

Remark.
Different from the one for the standard method and thus has certain unique applications (cf. Example).

-

$$\underline{\underline{V}}_{\underline{\hat{\theta}}_{\mathrm{ext}}} \xrightarrow[\mathbb{N}\to\infty]{P} \frac{1}{N}\left(\left\langle \left.\frac{\partial \ln P}{\partial \theta_i}\frac{\partial \ln P}{\partial \theta_j}\right|_{\underline{\theta}^*}\right\rangle\right)^{-1} = \left(-\left.\frac{\partial^2 l_{\mathrm{ext}}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1}, \tag{80}$$

$$\underline{\underline{V}}_{\underline{\hat{\theta}}_{\mathrm{ext},w}} \xrightarrow[\mathbb{N}\to\infty]{P} \left(-\left.\frac{\partial^2 l_{\mathrm{ext},w}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1} \left(\sum_{k=1}^{N} w^2(x_k)\left[\frac{\partial \ln P(x_k)}{\partial \theta_i}\frac{\partial \ln P(x_k)}{\partial \theta_j}\right]_{\underline{\theta}^*}\right) \left(-\left.\frac{\partial^2 l_{\mathrm{ext},w}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1} \tag{81}$$

$$= \left(-\left.\frac{\partial^2 l_{\mathrm{ext},w}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1} \left(-\left.\frac{\partial^2 l_{\mathrm{ext},w^2}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right) \left(-\left.\frac{\partial^2 l_{\mathrm{ext},w}}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}^*}\right)^{-1}. \tag{82}$$

Remark.
$\underline{\underline{V}}_{\underline{\hat{\theta}}_{\mathrm{ext},w}}$ is more accessible than $\underline{\underline{V}}_{\underline{\hat{\theta}}_w}$ in general cases.

- Example
  1. MC sample:
$$p\left(x\right)=\frac{1+x}{\mathrm{nor}},\ x\in\left[-0.6,\ 0.8\right],\ N=10^{6}. \tag{83}$$
  2. Fitting with the standard method, SML,
$$p\left(x|\theta\right)=\frac{1+\theta x}{\mathrm{nor}\left(\theta\right)},\ \theta^{*}=1;\ \hat{\theta}=1.0004\pm 0.0026. \tag{84}$$
  3. Fitting with the extended method, EML,
$$P\left(x|\theta_0,\ \theta_1\right)=\theta_0+\theta_1 x,\ \theta_0^{*}=\theta_1^{*}=\frac{N}{\int 1+x\mathrm{d}x}=6.4935\times 10^{5}; \tag{85}$$
$$\begin{array}{l}\hat{\theta}_0=\left(6.4933\pm 0.0067\right)\times 10^{5}\\ \hat{\theta}_1=\left(6.4959\pm 0.0165\right)\times 10^{5}\end{array},\ \underline{\underline{V}}_{\underline{\hat{\theta}}}=\begin{pmatrix}4.443\times 10^{5} & 1.911\times 10^{5}\\ 1.911\times 10^{5} & 2.738\times 10^{6}\end{pmatrix}; \tag{86}$$
$$\therefore\frac{\hat{\theta}_1}{\hat{\theta}_0}=1.0004\pm 0.0026. \tag{87}$$

Remark.

$$p=\frac{P}{\int P\mathrm{d}x},\ \text{with } P\left(x|\theta_0,\ \theta_1\right)=\theta_0+\theta_1 x, \tag{88}$$

is an inappropriate parameterization (Situation c) for the standard method, but not for the extended.

- Comparison between SML and EML:
  - Procedure:

    a. Apply SML and EML fits to a MC sample of a fixed / Poisson-fluctuated size.

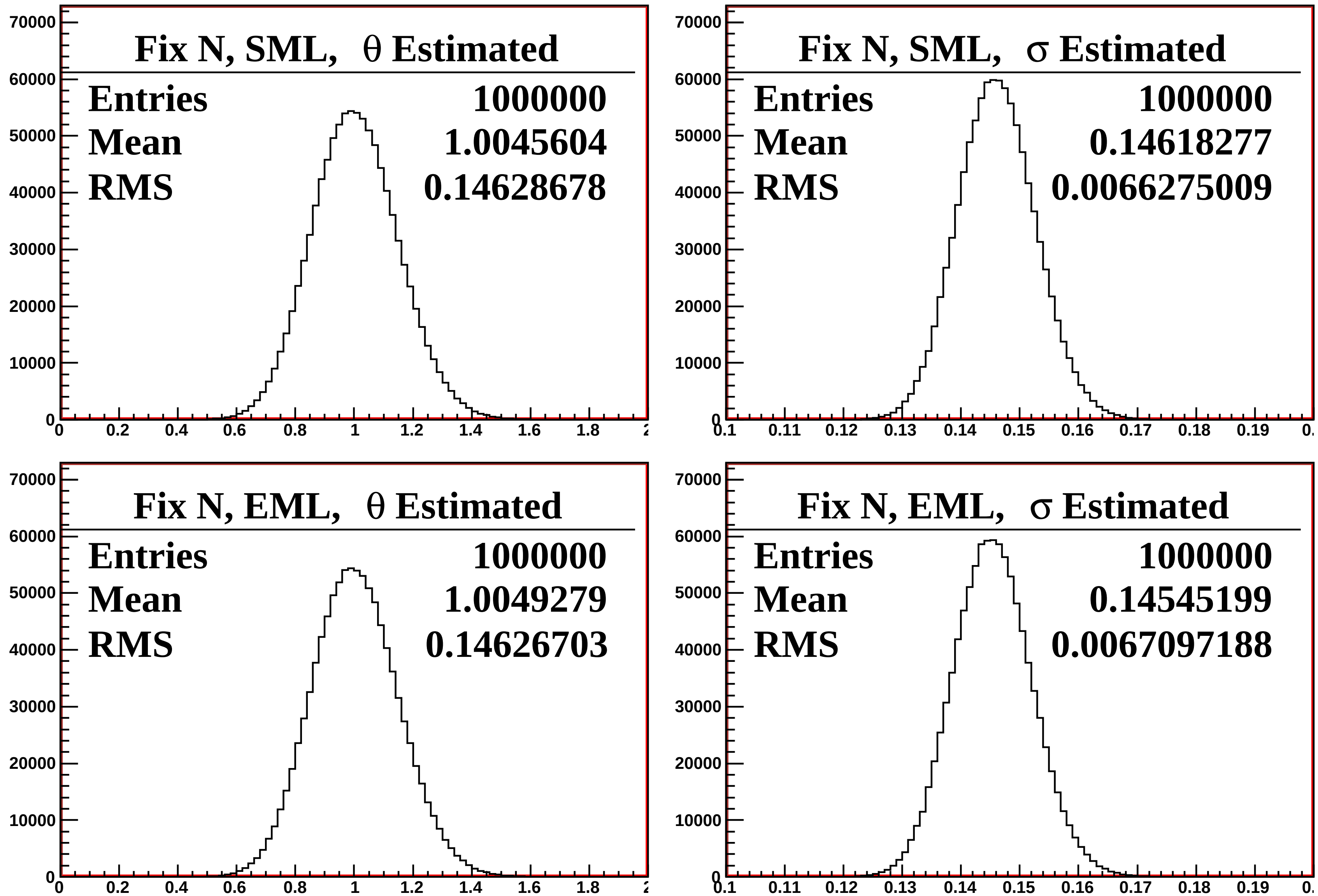
$$\text{MC: } p\left(x\right) = \frac{1+x}{\text{nor}},\ x \in \left[-0.3,\ 0.8\right],\ N = 10^3\ /\ \mathbb{N} = 10^3; \tag{89}$$

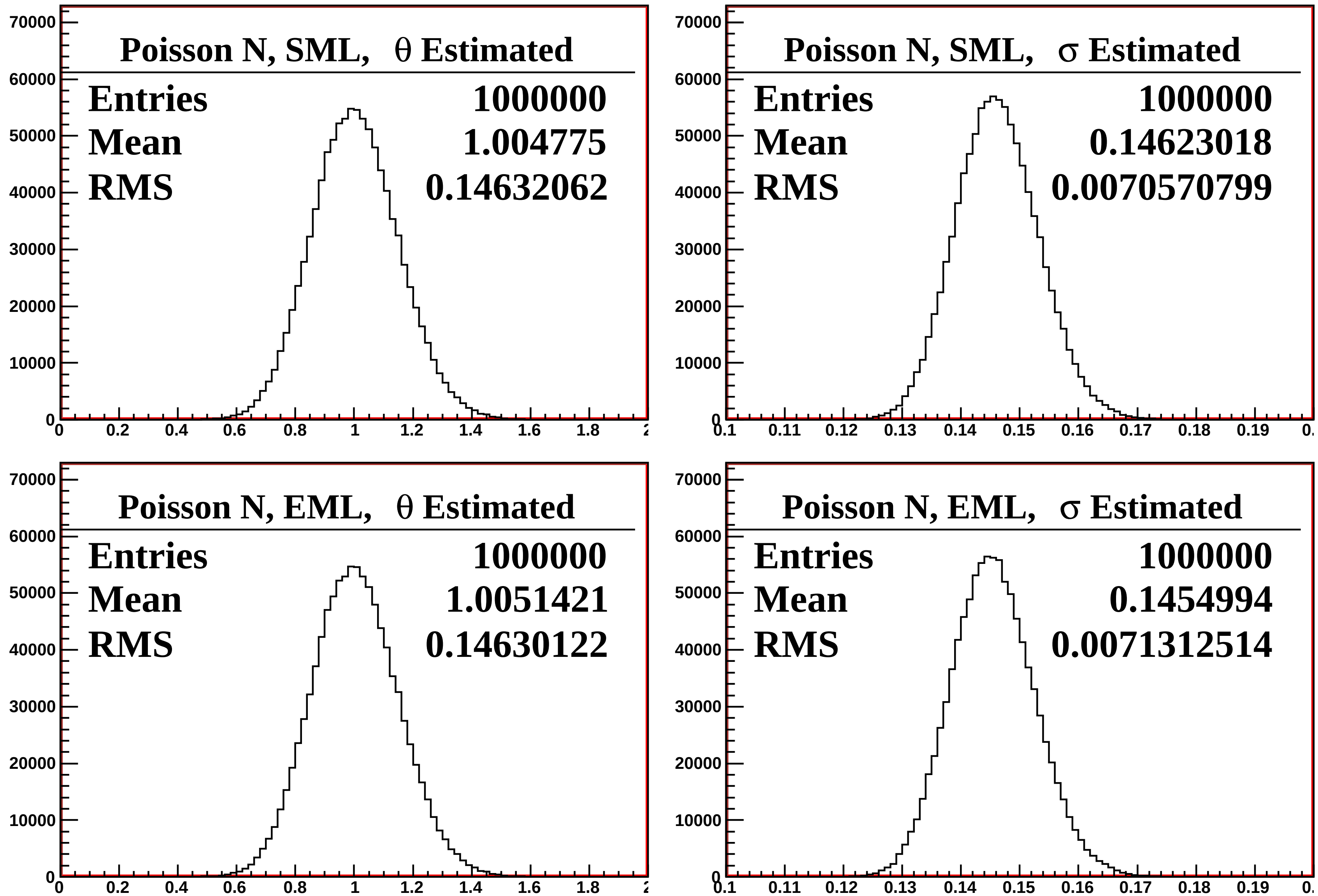
$$\text{SML :} p\left(x|\theta\right) = \frac{1+\theta x}{\text{nor}\left(\theta\right)},\ \theta_{\text{SML}} \equiv \theta,\ \theta^*_{\text{SML}} = 1; \tag{90}$$

$$\text{EML :} P\left(x|\theta_0, \theta_1\right) = \theta_0 + \theta_1 x,\ \theta_{\text{EML}} \equiv \frac{\theta_1}{\theta_0},\ \theta^*_{\text{EML}} = 1. \tag{91}$$

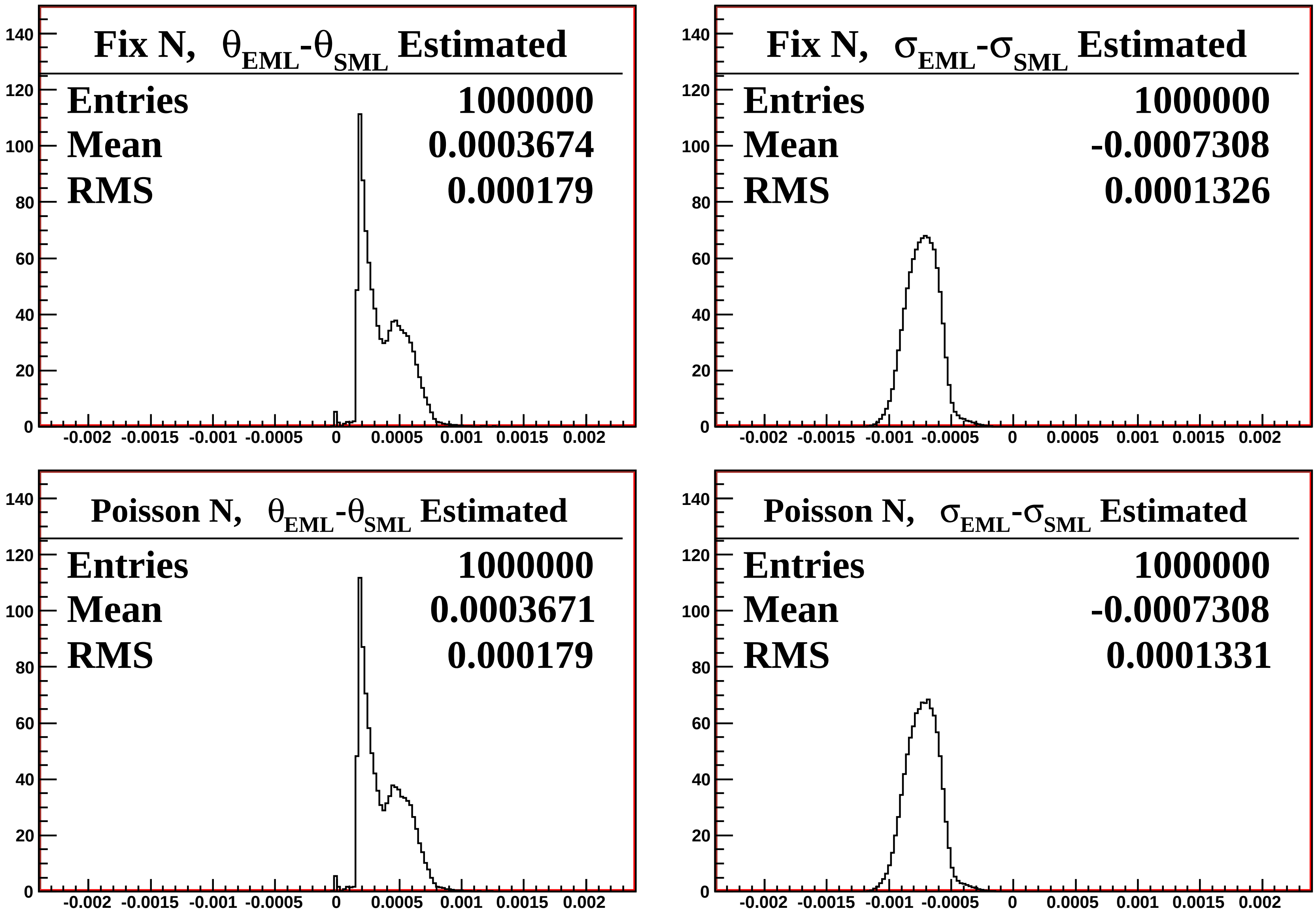
    b. Repeat a. with different random samples (10^6 times for each type of sample).

    c. Compare the distributions of the estimated parameters θ_{SML}, θ_{EML} and the estimated errors.


Fix N, SML, θ Estimated
Entries 1000000
Mean 1.0045604
RMS 0.14628678
Fix N, SML, σ Estimated
Entries 1000000
Mean 0.14618277
RMS 0.0066275009
Fix N, EML, θ Estimated
Entries 1000000
Mean 1.0049279
RMS 0.14626703
Fix N, EML, σ Estimated
Entries 1000000
Mean 0.14545199
RMS 0.0067097188



Poisson N, SML, θ Estimated
Entries 1000000
Mean 1.004775
RMS 0.14632062
Poisson N, SML, σ Estimated
Entries 1000000
Mean 0.14623018
RMS 0.0070570799
Poisson N, EML, θ Estimated
Entries 1000000
Mean 1.0051421
RMS 0.14630122
Poisson N, EML, σ Estimated
Entries 1000000
Mean 0.1454994
RMS 0.0071312514


Fix N, θEML-θSML Estimated
Entries 1000000
Mean 0.0003674
RMS 0.000179
Fix N, σEML-σSML Estimated
Entries 1000000
Mean -0.0007308
RMS 0.0001326
Poisson N, θEML-θSML Estimated
Entries 1000000
Mean 0.0003671
RMS 0.000179
Poisson N, σEML-σSML Estimated
Entries 1000000
Mean -0.0007308
RMS 0.0001331

– Conclusion: SML and EML are consistent to a large extend.

Part 3:

COMBINED ANALYSIS

Combined Analysis – p.d.f.

- **Cross section** for exclusive process with longitudinally polarized beam and unpolarized target:

$$\sigma_{\mathrm{LU}}(\phi, \mathbf{x}) = \sigma_{\mathrm{UU}}^{0}(\phi, \mathbf{x}) \left[1 + \eta A_{\mathrm{C}}(\phi, \mathbf{x}) + \lambda A_{\mathrm{LU}}^{\mathrm{DVCS}}(\phi, \mathbf{x}) + \eta\lambda A_{\mathrm{LU}}^{\mathcal{I}}(\phi, \mathbf{x})\right], \tag{92}$$

  η, **beam charge;** λ, **beam pol.**;
  ϕ, the angle between the production and scattering planes;
  $\mathbf{x}$, the **set of kinematic variables** $\{-t_c,\ x_B,\ Q^2,\ \ldots\}$ ***excluding*** ϕ.

- The **number density** of exclusive events in the τ-ϕ-$\mathbf{x}$ **space** (not considering the acceptance effect and detection efficiency):

$$n(\tau, \phi, \mathbf{x}) = \mathcal{L}(\tau)\sigma_{\mathrm{LU}}(\phi, \mathbf{x}) = \mathcal{L}(\tau)\sigma_{\mathrm{UU}}^{0}(\phi, \mathbf{x})\underline{b}^{T}(\tau)\underline{a}(\phi, \mathbf{x}), \tag{93}$$

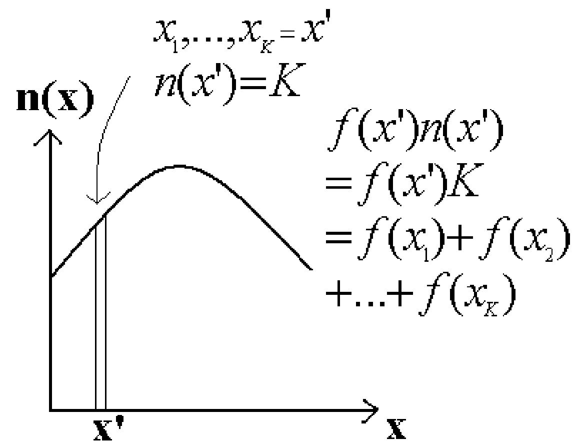
  τ, **time**; $\mathcal{L}(\tau)$, **luminosity**;

$$\underline{b}(\tau) \equiv \mathrm{col}\left[1,\ \eta(\tau),\ \lambda(\tau),\ \eta(\tau)\lambda(\tau)\right], \tag{94}$$

$$\underline{a}(\phi, \mathbf{x}) \equiv \mathrm{col}\left[1,\ A_{\mathrm{C}}(\phi, \mathbf{x}),\ A_{\mathrm{LU}}^{\mathrm{DVCS}}(\phi, \mathbf{x}),\ A_{\mathrm{LU}}^{\mathcal{I}}(\phi, \mathbf{x})\right]. \tag{95}$$

- Sum rules for discrete variables.
  x, y are independent variables, coupled as data point (x, y).

$$\sum_x f(x)\, n(x) = \sum_{i=1}^{\mathcal{N}} f(x_i);$$



$$\sum_y \sum_x f(x, y)\, n(x, y) = \sum_{i=1}^{N} f\left[(x, y)_i\right];$$

$$\overset{\text{OR}}{=} \sum_y \left[\sum_{i=1}^{\mathcal{N}(y)} f(x_i, y)\right] = \sum_y \mathcal{N}(y) \left[\frac{1}{\mathcal{N}(y)} \sum_{i=1}^{\mathcal{N}(y)} f(x_i, y)\right]$$

$$= \sum_y \mathcal{N}(y)\, \mathbb{f}(y) = \sum_{j=1}^{N} \mathbb{f}(y_j),$$

$$\mathbb{f}(y) \equiv \frac{1}{\mathcal{N}(y)} \sum_{i=1}^{\mathcal{N}(y)} f(x_i, y), \text{(x, y are independent)}$$

$$= \frac{1}{\mathfrak{N}} \sum_{i=1}^{\mathfrak{N}} f(x_i, y), \text{ the "}x\text{-averaged } f\text{".}$$

- **P.d.f.** for the exclusive process

$$p(\tau, \phi, \mathbf{x}) = \frac{n(\tau, \phi, \mathbf{x})}{\sum_\tau \sum_\phi \sum_{\mathbf{x}} n(\tau, \phi, \mathbf{x})}, \tag{96}$$

$$\mathbb{p}(\tau, \phi) = \frac{\sum_{\mathbf{x}} n(\tau, \phi, \mathbf{x})}{\sum_\tau \sum_\phi \sum_{\mathbf{x}} n(\tau, \phi, \mathbf{x})}. \tag{97}$$

For the application of the likelihood method, the parameterization of the p.d.f., which originates from that of $\underline{a}(\phi, \mathbf{x})$, requires that the number density and the total number $\boldsymbol{N} \equiv \sum_\tau \sum_\phi \sum_\mathbf{x} n(\tau, \phi, \mathbf{x})$ preserve their dependence on $\underline{a}(\phi, \mathbf{x})$. The following is to calculate the analytical form of $\boldsymbol{N}(\underline{a})$.

1. $\mathcal{L}$ can be evaluated by the number density of the DIS events,

$$n_{\mathrm{DIS}}(\tau) = \mathcal{L}(\tau)\sigma_{\mathrm{DIS}}, \tag{98}$$

σ_{DIS}, the total cross section of the DIS process. Thus

$$n(\tau, \phi, \mathbf{x}) = n_{\mathrm{DIS}}(\tau) r(\phi, \mathbf{x}) \underline{b}^T(\tau) \underline{a}(\phi, \mathbf{x}), \; r(\phi, \mathbf{x}) \equiv \frac{\sigma^0_{\mathrm{UU}}(\phi, \mathbf{x})}{\sigma_{\mathrm{DIS}}}. \tag{99}$$

2. Define **summation operators** with respect to different ranges of time τ,

$$\underline{\sum}_\tau \equiv \begin{pmatrix} \sum_{\forall\tau, \eta(\tau)=+1, \lambda(\tau)>0} \\ \sum_{\forall\tau, \eta(\tau)=+1, \lambda(\tau)<0} \\ \sum_{\forall\tau, \eta(\tau)=-1, \lambda(\tau)>0} \\ \sum_{\forall\tau, \eta(\tau)=-1, \lambda(\tau)<0} \end{pmatrix}. \tag{100}$$

The number densities in ϕ and $\mathbf{x}$ over the corresponding τ ranges,

$$\underline{\mathcal{N}}(\phi, \mathbf{x}) \equiv \underline{\sum}_\tau n(\tau, \phi, \mathbf{x}) = \begin{pmatrix} \overrightarrow{\mathcal{N}}^+(\phi, \mathbf{x}) \\ \overleftarrow{\mathcal{N}}^+(\phi, \mathbf{x}) \\ \overrightarrow{\mathcal{N}}^-(\phi, \mathbf{x}) \\ \overleftarrow{\mathcal{N}}^-(\phi, \mathbf{x}) \end{pmatrix}, \tag{101}$$

$\longrightarrow$ ($\longleftarrow$), $^+$ ($^-$): positive (negative) beam polarization and charge.

The **beam-state matrix** – a constant matrix encoding the beam-charge and -polarization state, ($\underline{b}(\tau) \equiv \mathrm{col}\,[1,\ \eta(\tau),\ \lambda(\tau),\ \eta(\tau)\lambda(\tau)]$)

$$\underline{\underline{\mathbb{B}}} \equiv \underline{\sum}_{\tau} n_{\mathrm{DIS}}(\tau)\underline{b}^T(\tau) = \begin{pmatrix} \overrightarrow{N}^{+}_{\mathrm{DIS}} & \overrightarrow{N}^{+}_{\mathrm{DIS}} & \overrightarrow{N}^{+}_{\mathrm{DIS}}\left\langle \overrightarrow{\lambda}^{+} \right\rangle & \overrightarrow{N}^{+}_{\mathrm{DIS}}\left\langle \overrightarrow{\lambda}^{+} \right\rangle \\ \overleftarrow{N}^{+}_{\mathrm{DIS}} & \overleftarrow{N}^{+}_{\mathrm{DIS}} & \overleftarrow{N}^{+}_{\mathrm{DIS}}\left\langle \overleftarrow{\lambda}^{+} \right\rangle & \overleftarrow{N}^{+}_{\mathrm{DIS}}\left\langle \overleftarrow{\lambda}^{+} \right\rangle \\ \overrightarrow{N}^{-}_{\mathrm{DIS}} & -\overrightarrow{N}^{-}_{\mathrm{DIS}} & \overrightarrow{N}^{-}_{\mathrm{DIS}}\left\langle \overrightarrow{\lambda}^{-} \right\rangle & -\overrightarrow{N}^{-}_{\mathrm{DIS}}\left\langle \overrightarrow{\lambda}^{-} \right\rangle \\ \overleftarrow{N}^{-}_{\mathrm{DIS}} & -\overleftarrow{N}^{-}_{\mathrm{DIS}} & \overleftarrow{N}^{-}_{\mathrm{DIS}}\left\langle \overleftarrow{\lambda}^{-} \right\rangle & -\overleftarrow{N}^{-}_{\mathrm{DIS}}\left\langle \overleftarrow{\lambda}^{-} \right\rangle \end{pmatrix}. \tag{102}$$

[Hint: $\sum_x n(x)f(x) = \sum_{i=1}^{N} f(x_i) = N\langle f\rangle$.]
N_{DIS} the number of the corresponding DIS events,
$\langle\lambda\rangle$ the average beam polarization over the DIS events.
Generally $\underline{\underline{\mathbb{B}}}$ is non-singular.

$$\underline{\sum}_{\tau} n(\tau,\phi,\mathbf{x}) = \underline{\sum}_{\tau} n_{\mathrm{DIS}}(\tau) r(\phi,\mathbf{x})\underline{b}^T(\tau)\underline{a}(\phi,\mathbf{x}), \tag{103}$$

$$\underline{\mathcal{N}}(\phi,\mathbf{x}) = \underline{\underline{\mathbb{B}}}\, r(\phi,\mathbf{x})\underline{a}(\phi,\mathbf{x}). \tag{104}$$

So the number density summed over the whole range of τ:

$$\mathcal{N}(\phi,\mathbf{x}) \equiv \underline{s}^T\, \underline{\mathcal{N}}(\phi,\mathbf{x}) \tag{105}$$

$$= \underline{s}^T\, \underline{\underline{\mathbb{B}}}\, r(\phi,\mathbf{x})\underline{a}(\phi,\mathbf{x}),\quad \underline{s} \equiv \mathrm{col}\,(1,\ 1,\ 1,\ 1)\,. \tag{106}$$

3. $r(\phi, \mathbf{x})$ can be solved:

$$\because \underline{a}(\phi,\mathbf{x}) \equiv \mathrm{col}\left[1,\ A_{\mathrm{C}}(\phi,\mathbf{x}),\ A_{\mathrm{LU}}^{\mathrm{DVCS}}(\phi,\mathbf{x}),\ A_{\mathrm{LU}}^{\mathcal{I}}(\phi,\mathbf{x})\right] \tag{107}$$

$$\therefore r(\phi,\mathbf{x}) = \left[r(\phi,\mathbf{x})\underline{a}(\phi,\mathbf{x})\right]_1 \tag{108}$$

$$= \left[\underline{\underline{\mathbb{B}}}^{-1}\underline{\mathcal{N}}(\phi,\mathbf{x})\right]_1 \tag{109}$$

$$= \sum_{k=1}^{4}\left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{1k}\left[\underline{\mathcal{N}}(\phi,\mathbf{x})\right]_k . \tag{110}$$

$$\therefore \mathcal{N}(\phi,\mathbf{x}) = \sum_{k=1}^{4}\left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{1k}\underline{s}^{T}\,\underline{\underline{\mathbb{B}}}\,\underline{a}(\phi,\mathbf{x})\left[\underline{\mathcal{N}}(\phi,\mathbf{x})\right]_k . \tag{111}$$

4. The total number (with dependence on $\underline{a}$) over the whole range of τ, ϕ and $\mathbf{x}$:

$$\boldsymbol{N} = \sum_{\phi}\sum_{\mathbf{x}} \mathcal{N}(\phi, \mathbf{x}) = \sum_{\phi}\sum_{\mathbf{x}}\sum_{k=1}^{4} \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{1k} \underline{s}^{T}\, \underline{\underline{\mathbb{B}}}\, \underline{a}(\phi, \mathbf{x})\, [\underline{\mathcal{N}}(\phi, \mathbf{x})]_{k} \tag{112}$$

$$= \underline{s}^{T}\, \underline{\underline{\mathbb{B}}} \sum_{i=1}^{\boldsymbol{N}} W_i\, \underline{a}(\phi_i, \mathbf{x}_i) \overset{\text{OR}}{=} \underline{s}^{T}\, \underline{\underline{\mathbb{B}}} \sum_{i=1}^{\boldsymbol{N}} W_i\, \underline{\mathbb{a}}(\phi_i). \tag{113}$$

$$\underline{\mathbb{a}}(\phi) \equiv \frac{1}{\mathfrak{N}} \sum_{j=1}^{\mathfrak{N}} \underline{a}(\phi, \mathbf{x}_j) \tag{114}$$

$$= \mathrm{col}\left[1,\ \mathbb{A}_{\mathrm{C}}(\phi),\ \mathbb{A}_{\mathrm{LU}}^{\mathrm{DVCS}}(\phi),\ \mathbb{A}_{\mathrm{LU}}^{\mathcal{I}}(\phi)\right],\ \text{the "}\mathbf{x}\text{-averaged } \underline{a}\text{".} \tag{115}$$

The upper bound $\boldsymbol{N}$ is the total number (simply the sample size, no dependence on $\underline{a}$);

$$W_i = \begin{cases} \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{11} & (\eta_i = +1,\ \lambda_i > 0) \\ \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{12} & (\eta_i = +1,\ \lambda_i < 0) \\ \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{13} & (\eta_i = -1,\ \lambda_i > 0) \\ \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{14} & (\eta_i = -1,\ \lambda_i < 0) \end{cases}. \tag{116}$$

□

- Parameterized p.d.f.,

$$p\left(\tau,\phi,\mathbf{x};\underline{\theta}\right)=\frac{n(\tau,\phi,\mathbf{x};\underline{\theta})}{\boldsymbol{N}\left(\underline{\theta}\right)} \tag{117}$$

$$=\frac{\mathcal{L}(\tau)\sigma^{0}_{\mathrm{UU}}(\phi,\mathbf{x})\underline{b}^{T}(\tau)\underline{a}(\phi,\mathbf{x};\underline{\theta})}{\underline{s}^{T}\ \underline{\underline{\mathbb{B}}}\sum_{i=1}^{\boldsymbol{N}}W_i\ \underline{a}(\phi_i,\mathbf{x}_i;\underline{\theta})} \tag{118}$$

$$\mathbb{p}\left(\tau,\phi;\underline{\theta}\right)=\sum_{\mathbf{x}}p\left(\tau,\phi,\mathbf{x};\underline{\theta}\right) \tag{119}$$

$$=\frac{\mathcal{L}(\tau)\bar{\sigma}^{0}_{\mathrm{UU}}(\phi)\underline{b}^{T}(\tau)\underline{\mathbb{a}}'(\phi;\underline{\theta})}{\underline{s}^{T}\ \underline{\underline{\mathbb{B}}}\sum_{i=1}^{\boldsymbol{N}}W_i\ \underline{\mathbb{a}}(\phi_i;\underline{\theta})} \tag{120}$$

$$\simeq\frac{\mathcal{L}(\tau)\bar{\sigma}^{0}_{\mathrm{UU}}(\phi)\underline{b}^{T}(\tau)\underline{\mathbb{a}}(\phi;\underline{\theta})}{\underline{s}^{T}\ \underline{\underline{\mathbb{B}}}\sum_{i=1}^{\boldsymbol{N}}W_i\ \underline{\mathbb{a}}(\phi_i;\underline{\theta})}. \tag{121}$$

$$\bar{\sigma}^{0}_{\mathrm{UU}}(\phi)\equiv\sum\nolimits_{\mathbf{x}}\sigma^{0}_{\mathrm{UU}}(\phi,\mathbf{x}),\ \ \underline{\mathbb{a}}'(\phi;\underline{\theta})\equiv\sum\nolimits_{x}\frac{\sigma^{0}_{\mathrm{UU}}(\phi,\mathbf{x})}{\bar{\sigma}^{0}_{\mathrm{UU}}(\phi)}\underline{a}(\phi,\mathbf{x};\underline{\theta})\simeq\underline{\mathbb{a}}(\phi;\underline{\theta}).$$

- For a sample of a certain beam charge,

$$\sigma_{\mathrm{LU}}(\tau, \phi, \mathbf{x}) = \sigma_{\mathrm{UU}}(\phi, \mathbf{x}) \left[1 + \lambda A_{\mathrm{LU}}(\phi, \mathbf{x})\right], \tag{122}$$

one just needs to modify the following definitions:

$$\underline{a}(\phi, \mathbf{x}) \equiv \begin{pmatrix} 1 \\ A_{\mathrm{LU}}(\phi, \mathbf{x}) \end{pmatrix}, \qquad \underline{\underline{\mathbb{B}}} \equiv \begin{pmatrix} \overrightarrow{N}_{\mathrm{DIS}} & \overrightarrow{N}_{\mathrm{DIS}} \left\langle \overrightarrow{\lambda} \right\rangle \\ \overleftarrow{N}_{\mathrm{DIS}} & \overleftarrow{N}_{\mathrm{DIS}} \left\langle \overleftarrow{\lambda} \right\rangle \end{pmatrix}, \tag{123}$$

$$\underline{s} \equiv \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad W_i \equiv \begin{cases} \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{11} & (\lambda_i > 0) \\ \left[\underline{\underline{\mathbb{B}}}^{-1}\right]_{12} & (\lambda_i < 0) \end{cases}, \tag{124}$$

$$\underline{\mathbb{a}}(\phi) \equiv \begin{pmatrix} 1 \\ \mathbb{A}_{\mathrm{LU}}(\phi) \end{pmatrix}. \tag{125}$$

- Corrections for Detection Efficiencies

  Consideration above does not take the detection efficiencies for the exclusive events and the DIS events ($e\left(\tau, \phi, \mathbf{x}\right)$ and $e_{\mathrm{DIS}}\left(\tau\right)$ respectively) into account. The following shows that under certain general conditions, no corrections are needed.

  Considering the efficiencies, similar to Eq. (99), we have

$$n(\tau, \phi, \mathbf{x}) = \frac{e\left(\tau, \phi, \mathbf{x}\right)}{e_{\mathrm{DIS}}\left(\tau\right)} n_{\mathrm{DIS}}(\tau) r(\phi, \mathbf{x}) \underline{b}^T(\tau) \underline{a}(\phi, \mathbf{x}), \tag{126}$$

  where n and n_{DIS} are the *detected* number densities. If $\frac{e(\tau,\phi,\mathbf{x})}{e_{\mathrm{DIS}}(\tau)}$ depends on τ, ϕ, $\mathbf{x}$ weekly, it drops out when Eq. (126) is normalized.

Combined Analysis – Estimation Methods

- The **likelihood function** and **the extended** one for $\mathbb{p}\,(\tau, \phi; \underline{\theta})$

$$l = \sum_{i=1}^{\boldsymbol{N}} \ln \left[\underline{b}_i^T \underline{\mathbb{a}}(\phi_i; \underline{\theta}) \right] - \boldsymbol{N} \ln \left[\underline{s}^T \, \underline{\underline{\mathbb{B}}} \sum_{i=1}^{\boldsymbol{N}} W_i \, \underline{\mathbb{a}}(\phi_i; \underline{\theta}) \right], \text{ SML,} \tag{127}$$

$$l_{\text{ext}} = \sum_{i=1}^{\boldsymbol{N}} \ln \left[\underline{b}_i^T \underline{\mathbb{a}}(\phi_i; \underline{\theta}) \right] - \underline{s}^T \, \underline{\underline{\mathbb{B}}} \sum_{i=1}^{\boldsymbol{N}} W_i \, \underline{\mathbb{a}}(\phi_i; \underline{\theta}), \text{ EML.} \tag{128}$$

where $\underline{b}_i \equiv (1,\ \eta_i,\ \lambda_i,\ \eta_i\lambda_i)^T$ and $(1,\ \lambda_i)^T$ respectively.

- **WUML** (W: weighting, U: unnormalized):

1. Note that $\underline{s} \equiv \mathrm{col}\,(1,\ 1,\ 1,\ 1)$ is in fact the weighting vector,

$$\underline{s}^T\ \underline{\underline{\mathbb{B}}} = \underline{s}^T \underline{\sum}_\tau n_{\mathrm{DIS}}(\tau)\underline{b}^T(\tau) = \left(\sum_{\eta(\tau)=+1,\lambda(\tau)>0} + \cdots + \sum_{\eta(\tau)=-1,\lambda(\tau)<0} \right) n_{\mathrm{DIS}}(\tau)\underline{b}^T(\tau). \tag{129}$$

   Assigning different $\underline{s}$ components is to weight the data according the beam-state. One can chose $\underline{s}$ so that

$$\underline{s}^T\ \underline{\underline{\mathbb{B}}} = (k,\ 0,\ 0,\ 0)\ ,\ k \text{ is any positive number,} \tag{130}$$

$$\text{i.e. } \underline{s}_w = \left[\underline{\underline{\mathbb{B}}}^T\right]^{-1} \mathrm{col}\,(k,\ 0,\ 0,\ 0), \tag{131}$$

   making $\boldsymbol{N}\,(\underline{\theta})$ independent of $\underline{\theta}$. $\quad (\boldsymbol{N}(\underline{\theta}) = \underline{s}^T\ \underline{\underline{\mathbb{B}}} \sum_{i=1}^{\boldsymbol{N}} W_i\ \underline{\mathbb{a}}(\phi_i; \underline{\theta}))$

2. Warning!!!: When the weights are inhomogeneous because of, e.g., *extremely* unbalanced data sample of different states, this method is unjustified.

3.

$$l_w = \sum_{i=1}^{N} w_i \ln \left[\underline{b}_i^T \underline{\mathbb{a}}(\phi_i; \underline{\theta}) \right], \text{ WUML,} \tag{132}$$

$$w_i = \begin{cases} s_{w_1} & (\eta_i = +1, \ \lambda_i > 0) \\ s_{w2} & (\eta_i = +1, \ \lambda_i < 0) \\ s_{w3} & (\eta_i = -1, \ \lambda_i > 0) \\ s_{w4} & (\eta_i = -1, \ \lambda_i < 0) \end{cases} . \tag{133}$$

Since

$$\mathbb{p}_w (\tau, \ \phi | \underline{\theta}) \propto \underline{b}^T (\tau) \underline{\mathbb{a}}(\phi; \underline{\theta}), \tag{134}$$

$$\tag{135}$$

with parameterizations like $A = \theta_1 + \theta_2 \sin \phi + \ldots,$

$$\frac{\partial^2 \mathbb{p}_w}{\partial \theta_i \partial \theta_j} = 0, \tag{136}$$

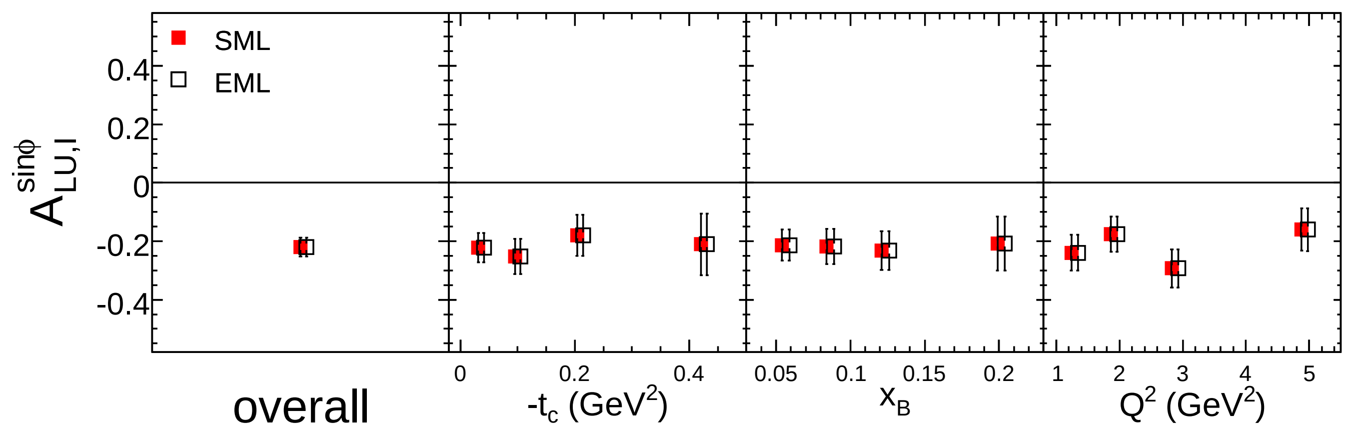
the weighted covariance matrix is accessible.

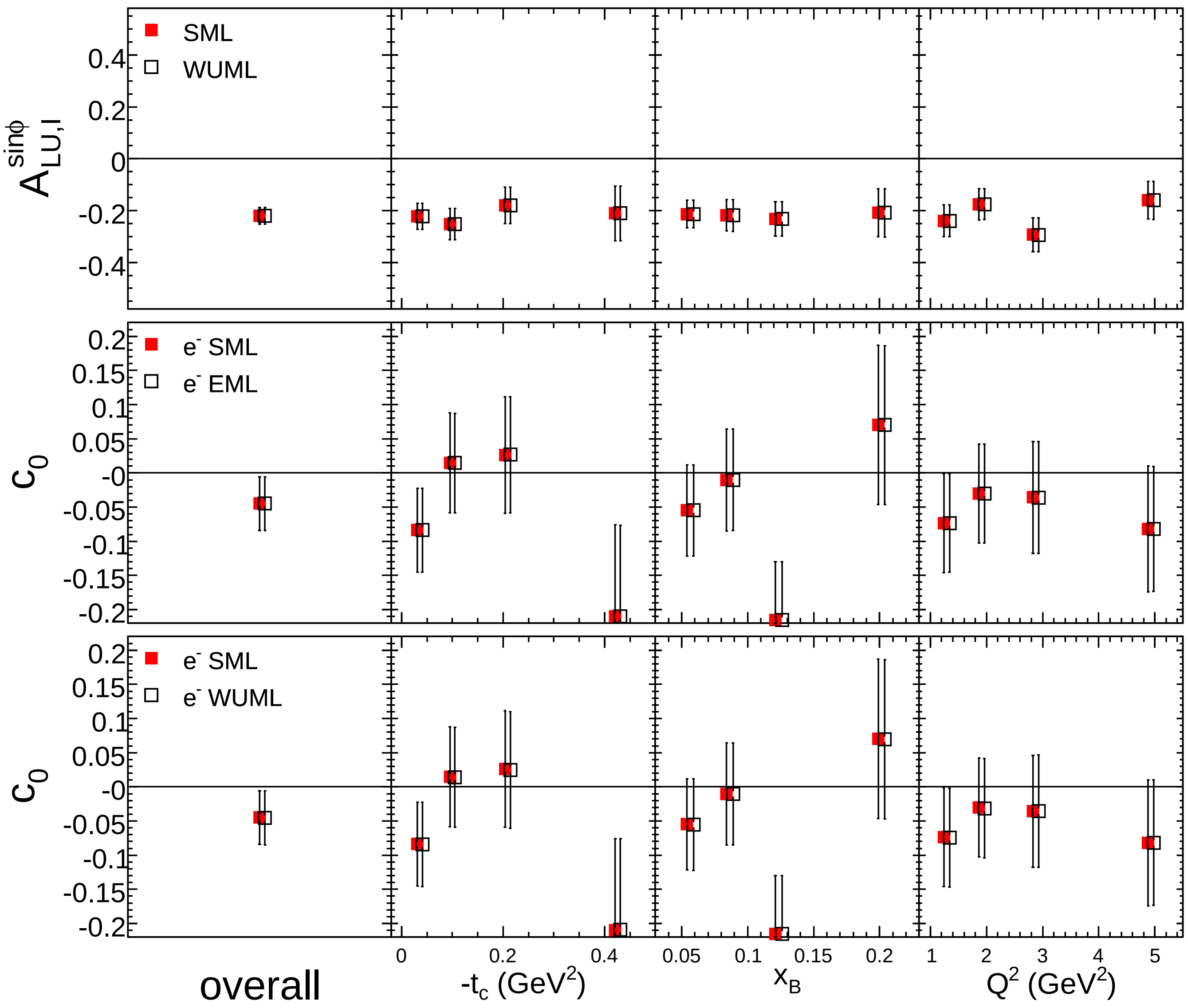
- With parameterizations

$$\mathbb{A}_{\mathrm{LU}}(\phi; c_0, s_1, c_1) = c_0 + s_1 \sin\phi + c_1 \cos\phi \text{ for BSA analysis;} \tag{137}$$

$$\left.\begin{aligned} \mathbb{A}_{\mathrm{C}}(\phi; c_0, s_1, c_1) &= c_0 + s_1 \sin\phi + c_1 \cos\phi; \\ \mathbb{A}_{\mathrm{LU}}^{\mathrm{DVCS}}(\phi; c_0, s_1, c_1) &= c_0 + s_1 \sin\phi + c_1 \cos\phi; \\ \mathbb{A}_{\mathrm{LU}}^{\mathcal{I}}(\phi; c_0, s_1, c_1) &= c_0 + s_1 \sin\phi + c_1 \cos\phi; \end{aligned}\right\} \text{ for combined analysis,} \tag{138}$$

the 3 methods are applied to the 00d0+05b1 Hydrogen DVCS data, results are *very close*:





SML
WUML
$A_{LU,I}^{\sin\phi}$
0.4
0.2
0
-0.2
-0.4
e⁻ SML
e⁻ EML
c_0
0.2
0.15
0.1
0.05
-0
-0.05
-0.1
-0.15
-0.2
e⁻ SML
e⁻ WUML
c_0
0.2
0.15
0.1
0.05
-0
-0.05
-0.1
-0.15
-0.2
0
0.2
0.4
0.05
0.1
0.15
0.2
1
2
3
4
5
overall
$-t_c$ (GeV^2)
x_B
Q^2 (GeV^2)

Summary for Part 2 & 3

1. Systematically illustrate the analytically properties of the weighting and the extended methods.
2. Provide the p.d.f. for combined analysis, application for SML, EML and WUML are described; results are shown.
3. *WUML may be a good choice for combined analysis since its much easier manipulation, while SML is more accurate in general cases.*

References

[1] *The advanced theory of statistics*, by M. Kendall et al., 4th ed. of Vol. 2 of the 3-volume ed. (1979) ISBN: 0 85264 255 5.

[2] A Survey of Maximum Likelihood Estimation, Part 1 & 2, R. H. Norden.

[3] "Probability", "Statistics" and "Monte Carlo techinquess" sections of PDG.

[4] Statistics, Roger Barlow.

[5] ANALYSIS OF EXPERIMENTS IN PARTICLE PHYSICS, FRANKT T. SOLMITZ.

[6] Notes on Statistics for Physicists, Revised, Jay Orear.

[7] MINUIT Reference Manual, Version 94.1, F. James.

[8] The Interpretation of Errors, F. James.

[9] MINUIT User's Guide, F. James.

[10] MINUIT Tutorial, F. James.

Appendix: Schwarz' Inequality (Generalized)

Let $\underline{a}\,(x_1,\ x_2, \dots,\ x_N)\,,\ \underline{b}\,(x_1,\ x_2, \dots,\ x_N)$ be real vector functions with M components, if $\left\langle \underline{b}\ \underline{b}^T \right\rangle$ is nonsingular,

$$\left\langle \underline{a}\ \underline{a}^T \right\rangle \geqslant \left\langle \underline{a}\ \underline{b}^T \right\rangle \left\langle \underline{b}\ \underline{b}^T \right\rangle^{-1} \left\langle \underline{b}\ \underline{a}^T \right\rangle . \tag{139}$$

$$(\underline{\underline{B}} \geqslant \underline{\underline{C}} :\ \underline{\underline{B}} - \underline{\underline{C}} \text{ positive semidefinite.})$$

Proof. Let $\underline{\underline{A}}$ be a constant real $M \times M$ matrix,

$$\begin{aligned}
\forall \underline{v} \neq \underline{0},\ \underline{v}^T & \left\langle \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right) \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right)^T \right\rangle \underline{v} \\
&= \left\langle \underline{v}^T \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right) \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right)^T \underline{v} \right\rangle \\
&= \left\langle \left[\underline{v}^T \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right)\right]^2 \right\rangle \geqslant 0,
\end{aligned}$$

$"="$ holds Iff $\underline{v}^T \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right) \equiv 0$,
i.e. $a_i + (\underline{\underline{A}}\ \underline{b})_i$ are linearly dependent functions of $(x_1,\ x_2, \dots,\ x_N)$.

$$\begin{aligned}
\therefore \quad & \underline{\underline{0}} \leqslant \left\langle \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right) \left(\underline{a} + \underline{\underline{A}}\ \underline{b}\right)^T \right\rangle \\
&= \left\langle \underline{a}\ \underline{a}^T \right\rangle + \left\langle \underline{a}\ \underline{b}^T \right\rangle \underline{\underline{A}}^T + \underline{\underline{A}} \left\langle \underline{b}\ \underline{a}^T \right\rangle + \underline{\underline{A}} \left\langle \underline{b}\ \underline{b}^T \right\rangle \underline{\underline{A}}^T,
\end{aligned}$$

let $\underline{\underline{A}} = - \left\langle \underline{a}\ \underline{b}^T \right\rangle \left\langle \underline{b}\ \underline{b}^T \right\rangle^{-1}$,
we have

$$\left\langle \underline{a}\ \underline{a}^T \right\rangle - \left\langle \underline{a}\ \underline{b}^T \right\rangle \left\langle \underline{b}\ \underline{b}^T \right\rangle^{-1} \left\langle \underline{b}\ \underline{a}^T \right\rangle \geqslant \underline{\underline{0}}.$$

$\square$