A pedagogical derivation of the matrix element method in particle physics data analysis

To cite this article: Suharyo Sumowidagdo 2018 J. Phys.: Conf. Ser. 988 012003

View the article online for updates and enhancements.
A pedagogical derivation of the matrix element method in particle physics data analysis

Suharyo Sumowidagdo
Research Center for Physics, Indonesian Institute of Sciences Serpong, Banten, INDONESIA
E-mail: Suharyo.Sumowidagdo@lipi.go.id

Abstract. The matrix element method provides a direct connection between the underlying theory of particle physics processes and detector-level physical observables. I am presenting a pedagogically-oriented derivation of the matrix element method, drawing from elementary concepts in probability theory, statistics, and the process of experimental measurements. The level of treatment should be suitable for beginning research student in phenomenology and experimental high energy physics.

The matrix element method as known in high energy physics, was first proposed by Kondo [1, 2] and later by Dalitz and Goldstein [3, 4]. The method provides direct connection between the underlying theory of particle physics processes and physical observables measured by the particle detector. An early successful application of the method is the re-measurement of the top quark mass using D0 Run I data [5] which yields the most precise single-measurement from Tevatron Run I data.

Despite the method has been proposed almost 30 years ago, an elementary and pedagogical derivation of the method is not easy to find. Discussions and treatments of the method are scattered in original research papers, preprints, and dissertation. Furthermore, most of the existing treatments are written assuming the reader has a good knowledge of maximum likelihood estimation method. The treatment in this article is intended to be an entry-level treatment of the subject, after reading which beginning research students are ready to read more specialized treatments on the subject.

1. The concept of likelihood for single-variable probability distributions
To understand the matrix element method, one need to first understand the concept of likelihood as it is defined in statistics and probability theory. The treatment in this article is not meant to be comprehensive and rigorous, but rather a quick introduction into the key necessary concepts. A more elaborate discussions on probability theory, statistics, and likelihood can be found elsewhere, e.g. in the text by James [6].

Consider a set of measurements of a continuous variable $x$. Each measurement is performed under controlled, identical experimental situation. The outcome of each measurement is independent of the outcomes of other measurements. Then $x$ is a random variable. The full set of measurements is defined as the sample. A single measurement in the sample is named an event.

If the continuous variable $x$ is hypothesized to follow a probability distribution function $P(x)$,
then the likelihood of an individual measurement which has value \( x_1 \) is defined as

\[
\mathcal{L}(x_1) = P(x_1). \tag{1}
\]

The likelihood that a sample which has \( n \) events to have measured a set of values \( \{x_1, x_2, \ldots, x_n\} \) is the product of all likelihoods of the individual events

\[
\mathcal{L}(\{x_1, x_2, \ldots, x_n\}) = P(x_1) \times P(x_2) \times \cdots \times P(x_n) = \prod_{i=1}^{n} P(x_i). \tag{2}
\]

For computational purpose, the negative log-likelihood of a sample is preferred, which is defined as the negative of the sum of logarithms of all likelihoods

\[
- \ln \mathcal{L}(\{x_1, x_2, \ldots, x_n\}) = - (\ln P(x_1) + \ln P(x_2) + \cdots + \ln P(x_n)) = - \sum_{i=1}^{n} \ln P(x_i). \tag{3}
\]

Since the logarithm of a function is monotonically increasing function, the extremum points (maxima and minima) of the log-likelihood function will be located at the same extremum points of the likelihood function.

It is important to note that the probability distribution function \( P(x) \) is a priori. In the context of experimental physics, \( P(x) \) represent the theory to be tested; or used as theoretical input, in the experiment.

2. Multivariate probability distributions

The discussion in the previous section can be generalized for multivariate probability distributions. Consider a simultaneous measurement of \( k \) independent variable in each event. Thus \( x \) in the previous section becomes a vector \( \mathbf{x} = (x_1, x_2, x_3, \ldots, x_k) \). If all the measured variables are independent of one another, the generalization of Equations (1), (2), and (3) are

\[
\mathcal{L}(\mathbf{x}_1) = P(\mathbf{x}_1), \tag{4}
\]

\[
\mathcal{L}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = P(\mathbf{x}_1) \times P(\mathbf{x}_2) \times \cdots \times P(\mathbf{x}_n) = \prod_{i=1}^{n} P(\mathbf{x}_i), \tag{5}
\]

\[
- \ln \mathcal{L}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = - (\ln P(\mathbf{x}_1) + \ln P(\mathbf{x}_2) + \cdots + \ln P(\mathbf{x}_n)) = - \sum_{i=1}^{n} \ln P(\mathbf{x}_i). \tag{6}
\]

In physics process, \( \mathbf{x} \) is an independent set of variables which uniquely characterizes each event. In the context of scattering process, \( \mathbf{x} \) is the set of final-state particles’ momenta, subject to the constraint of energy-momentum conservation.

3. Parameter estimation using likelihood

In using the likelihood for parameter estimation, the likelihood for an event is expressed as a function of a set of free parameters \( \mu \), \( \mathcal{L} = \mathcal{L}(\mathbf{x}, \mu) \). The negative log-likelihood to observe a sample of \( n \) events is expressed as

\[
- \ln \mathcal{L}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \mu) = - \sum_{i=1}^{n} \ln \mathcal{L}(\mathbf{x}_i, \mu). \tag{7}
\]

Minimizing the negative log-likelihood over the parameter space of \( \mu \) will yield an estimator for \( \mu, \tilde{\mu} \). Assuming the central limit theorem applies, the likelihood for a sample will have a Gaussian shape,

\[
\mathcal{L} \propto \exp \left( -\frac{1}{2}(\mu - \tilde{\mu})^T \Sigma^{-1}(\mu - \tilde{\mu}) \right) \tag{8}
\]

2
in the neighbourhood of $\tilde{\mu}$. The region bounded by the surface where the value of $-\ln L$ is higher by 0.5 relative to value of $-\ln L$ evaluated at $\tilde{\mu}$ is considered to be the one-sigma confidence region of the parameters.

If there is only one parameter to be estimated, $\mu$, the likelihood will have an univariate Gaussian shape,

$$L \propto \exp \left( -\frac{(\mu - \tilde{\mu})^2}{2\sigma} \right). \tag{9}$$

The negative log-likelihood $-\ln L$ curve has the shape of a concave upward parabola. The estimator of $\mu$, $\hat{\mu}$, is the value of $\mu$ corresponding to the minimum of this curve. The one-sigma confidence interval of $\mu$ is taken between the points where the value of the $-\ln L$ is higher by 0.5 relative to the value of $-\ln L$ evaluated at $\tilde{\mu}$.

4. Hypothesis testing using likelihood

In using likelihood for hypothesis testing, two likelihoods are calculated each using two different hypotheses. The first likelihood is calculated using the *null hypothesis* $H_0$ as prior. The null hypothesis is a proposition which is expected to be true before the hypothesis testing is performed.

$$L_0 = L_0(x|H_0) \tag{10}$$

The second likelihood is calculated using the *alternative hypothesis* $H_c$ as prior. The alternative hypothesis is a proposition which is not expected to be true before the hypothesis testing. Depends on the results of hypothesis testing, the alternative hypothesis may or may not be accepted as alternative to the null hypothesis.

$$L_c = L_c(x|H_c) \tag{11}$$

The Neyman-Pearson lemma [7] states that the ratio between the two likelihoods, $\Lambda$, is the optimal discriminating variable between the two hypothesis,

$$\Lambda = \frac{L_0}{L_c} \tag{12}$$

The hypothesis testing is done by comparing the likelihood ratio with a pre-determined value $c$, which corresponds to a certain significance level. If it is found that the likelihood ratio is less than $c$,

$$\Lambda < c \tag{13}$$

then the null hypothesis is rejected and the alternative hypothesis is accepted. If the likelihood ratio is equal or larger than $c$, the null hypothesis is not rejected.

5. Derivation of the matrix element method

In experimental particle physics, one observed the outcomes from collisions of pre-determined initial condition. Each individual collision is the equivalent of event in the previous discussions. For each event, there are multiple quantities measured by the particle detector, in the forms of final particles’ momenta. Thus the likelihood of an event is a function of more than one variable.

Consider a physics process $v_1 + v_2 \rightarrow p_1 + p_2 + \ldots + p_k$, where $v_1, v_2, p_1, p_2, \ldots, p_k$ are all four-momenta. We shall use the notation $v_i$ to indicate the initial set of four-momenta $\{v_1, v_2\}$, and the notation $p_f$ to indicate the final set of four-momenta $\{p_1, p_2, \ldots, p_k\}$. The *parton-level* differential cross-section for this process can be expressed as

$$d\sigma(p_f|v_i) = \frac{1}{F(v_i)}(2\pi)^4|M(v_i \rightarrow p_f)|^2 \delta^4(v_1 + v_2 - \sum_{f=1}^{k} p_f) \prod_{f=1}^{k} \frac{d^3p_f}{(2\pi)^3(2E_f)} \tag{14}$$
as stated in standard field theory textbooks (e.g. [8]). Here $F$ is the flux factor which may includes convolution over the parton distribution functions, and $\mathcal{M}$ is the matrix element or Lorentz-invariant scattering amplitude of the process. In practice, the detector will smear a parton-level four-momentum $p_f$ into detector-level four-momentum $q_f$. To express the differential cross-section as a functions of $q_f$, instead of $p_f$, we must unsmear each individual detector-level four-momenta $q_f$, i.e., express $q_f$ as a convolution of $p_f$.

$$q_f = \int W(q_f | p_f) \, d^3 p_f$$

(15)

An important assumption about the transfer function warrants a discussion. The transfer function for a parton-level four-momenta $p_i$ into detector-level four-momenta $q_i$ is assumed to be independent of the other parton-level and detector-level four-momenta. In other words, the complete transfer function $W$ can be factorized into individual transfer function $W_i$

$$W(q_f | p_f) = W_1(q_1 | p_1) \times W_2(q_2 | p_2) \times W_3(q_3 | p_3) \times \ldots \times W_k(q_k | p_k)$$

(16)

Thus the detector-level differential cross-section, $d\sigma(q_f | v_i)$ is an integral of the all possible parton-level differential cross-section $d\sigma(p_f | v_i)$ weighed by the transfer function $W(q_f | p_f)$.

$$d\sigma(q_f | v_i) = \int W(q_f | p_f) \, d\sigma(p_f | v_i)$$

$$= \int W(q_f | p_f) \, \frac{1}{F(v_1, v_2)} (2\pi)^4 |\mathcal{M}(v_i \rightarrow p_f)|^2 \delta^4 \left(v_1 + v_2 - \sum_{f=1}^{f=k} p_f\right) \prod_{f=1}^{f=k} \frac{d^3 p_f}{(2\pi)^3 (2E_f)}$$

(17)

We can normalized the expression for detector-level differential cross-section with the detector-level total cross-section, $\sigma_i$ to obtain a normalized probability density (thus likelihood).

$$\mathcal{L}(q_f | v_i) = \frac{1}{\sigma_i} \int W(q_f | p_f) \, \frac{1}{F(v_i)} (2\pi)^4 |\mathcal{M}(v_i \rightarrow p_f)|^2 \delta^4 \left(v_1 + v_2 - \sum_{f=1}^{f=k} p_f\right) \prod_{f=1}^{f=k} \frac{d^3 p_f}{(2\pi)^3 (2E_f)}$$

(18)

Equation 18 is the matrix element method. It express the likelihood to observe an event with detector-level momenta $q_f$, given a prior of the underlying physics process in the matrix element $\mathcal{M}$ and detector effect $W$. The matrix element $\mathcal{M}$ contains the input theoretical model (a priori) and encapsulates the physics of the process under study. In practice sometimes only the dominant processes are calculated, at the lowest-order Feynman diagram. The four-dimensional $\delta$-function represents momentum conservation and will reduce the number of independent variables to $3k - 4$, where $k$ is the number of parton-level particles in the final state. The product term in the Equation 18 represents the final state phase space of parton-level four-momenta $p_f$ after unsmearing of detector effects. To compute the likelihood, the equation must be integrated (numerically) over this phase space. Often, the integration of the phase is the most difficult task in using the matrix element method.

In many applications, there are more than one physics process that can happen and compatible with a given sets of initial and final state particles. One of the process usually is identified as signal process, while the others are identified as background processes. For example, in the study of top quark-antitop quark pair production in the $e^+e^-\nu_\mu\bar{\nu}_\mu b\bar{b}q\bar{q}$ channel at the LHC, the signal process is $pp \rightarrow t\bar{t} \rightarrow W^+bW^-\bar{b} \rightarrow e^+e^-\nu_\mu\bar{\nu}_\mu b\bar{b}q\bar{q}$ while the dominant background process are $pp \rightarrow Zb\bar{b} \rightarrow e^+e^-b\bar{b}$ and $pp \rightarrow Zjj \rightarrow e^+e^-jj$, where $jj$ are either gluon or light-quark jets. If this case, then the likelihood for each event will be a sum of the likelihood for
Figure 1. The likelihood evaluated as a function of five top quark mass values, and the fit of parabolic curve to the negative log-likelihood curve.

each processes, weighted by the fraction of each process in the data sample.

$$\mathcal{L} = \sum_c f_c \mathcal{L}_c$$

(19)

Here $f_c$ is the fraction of events coming from process $c$ in the data sample. The determination of $f_c$ can be done either independently from, or simultaneously with, the application of matrix element method. In measurements, the fractions should be determined independently. In searches, the fraction can be determined simultaneously using the likelihood output from the matrix element method.

6. Example: Top quark mass measurement

A recent results is the measurement of top quark mass in the dilepton channel by D0 Collaboration [9]. The likelihood used in the measurement is

$$P(q_f|f_{tt}, m_t) = f_{tt}P_{tt}(q_f|m_t) + (1 - f_{tt})P_{bkg}(q_f)$$

(20)

A comprehensive treatment of the application of the matrix element method in measurements of top quark mass can be found in [10].

The fraction of $tt$ events in the data sample is $f_{tt}$, the fraction of background events is $f_{bkg} = 1 - f_{tt}$. The probability for signal events, $P_{tt}$, is a function of the top quark mass, while the probability for background events, $P_{bkg}$, is independent on the value of top quark mass. The signal probability is taken to be the quark fusion process $q\bar{q} \rightarrow t\bar{t} \rightarrow W^+W^-b\bar{b} \rightarrow \ell^+\ell^-\nu\bar{\nu}\ell^+\ell^-\nu\bar{\nu}$, while the background matrix element is taken to be the $Z\gamma^* \rightarrow \ell^+\ell^- + 2$ jets process. The gluon fusion $t\bar{t}$ production process and the $Z\gamma^* \rightarrow \ell^+\ell^-b\bar{b}$ are both neglected due to their cross sections are relative smaller to the included processes. Both the signal and background matrix elements are calculated at the leading order.

The sample likelihood is calculated as a function of several top quark mass values. To correct for bias, the method is calibrated using $t\bar{t}$ Monte Carlo samples generated at the same set of top quark mass values used in the calculation of the sample likelihood. A parabolic curve (corresponding to Gaussian shape of the likelihood) is then fitted to negative log-likelihood curve, as shown in Figure 1.

7. Example: Spin correlation in top quark-antiquark production

The top quark decays through weak interactions before it hadronizes, thus its spin state information would be transferred cleanly to its decay products. Furthermore, due to its very
short-lifetime, top quark doesn’t experience spin flips after it is produced. In top quark-antiquark production thus there is correlation information between the spin of the top quark and the spin of the top antiquark [11].

The D0 Collaboration has measured the fraction of $t\bar{t}$ with spin correlation using the matrix element method [12]. The details of the method was elaborated in this reference [13]. In the analysis, a variable $\mathcal{R}$, which is related to the likelihood ratio, is introduced with the definition

$$\mathcal{R} = \frac{P_{\text{signal (correlated)}}}{P_{\text{signal (correlated)}} + P_{\text{signal (uncorrelated)}}}$$

A binned fit of the $\mathcal{R}$ variable in data is performed, using templates from Monte Carlo (simulated) sample of $t\bar{t}$ events with and without spin correlation, to determine the fraction of $t\bar{t}$ events in data with spin correlation. Figure 2 shows the templates used, and the results of fitting the templates to data.

8. Acknowledgments
The author would like to thank D0 Collaboration for granting permission to use figures from their publications [9, 12] in this article.

References
[1] Kondo K 1988 J. Phys. Soc. Jap. 57 4126–4140
[2] Kondo K 1991 J. Phys. Soc. Jap. 60 836–844
[3] Dalitz R H and Goldstein G R 1992 Phys. Rev. D45 1531–1543
[4] Goldstein G R, Sliwa K and Dalitz R H 1993 Phys. Rev. D47 967–972 (Preprint hep-ph/9205246)
[5] Abazov V M et al. (D0 Collaboration) 2004 Nature 429 638–642 (Preprint hep-ex/0406031)
[6] James F 2006 Statistical methods in experimental physics (World Scientific) ISBN 9789812567956
[7] Neyman, J and Pearson, Egon S 1933 Phil. Trans. Roy. Soc. Lond. A231 289–337
[8] Peskin M E and Schroeder D V 1995 An Introduction to Quantum Field Theory (Perseus Books Publishing)
[9] Abazov V M et al. (D0) 2016 Phys. Rev. D94 032004 (Preprint 1606.02814)
[10] Fiedler F, Grohsjean A, Haefner P and Schieferdecker P 2010 Nucl. Instrum. Meth. A624 203–218 (Preprint 1003.1316)
[11] Barger V D, Ohnemus J and Phillips R J N 1989 Int. J. Mod. Phys. A4 617
[12] Abazov V M et al. (D0) 2011 Phys. Rev. Lett. 107 032001 (Preprint 1104.5194)
[13] Melnikov K and Schulze M 2011 Phys. Lett. B700 17–20 (Preprint 1103.2122)