ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА
2018. Т. 49. ВЫП. 1. С. 124–127

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА В КОЛЛАЙДЕРНЫХ ЭКСПЕРИМЕНТАХ С УЧЕТОМ РАЗЛИЧНЫХ ТИПОВ СИСТЕМАТИЧЕСКИХ И СТАТИСТИЧЕСКИХ ОШИБОК

П. С. Мандрик *

Институт физики высоких энергий Национального исследовательского центра «Курчатовский институт», Протвино, Россия

Приведено описание ряда методов статистического анализа, отвечающих запросам, возникающим в коллайдерных экспериментах в задачах обработки дискретизированных распределений данных с учетом систематических и статистических ошибок измерений.

Descriptions of techniques of statistical analysis are given for common problems of data analysis in collider experiments with incorporating different types of systematic and statistic uncertainties.

PACS: 06.20.Dk; 02.50.-r

ВВЕДЕНИЕ

При анализе данных в коллайдерных экспериментах часто возникают задачи следующего рода: известны полученные, например, из генераторов Монте-Карло формы распределений сигнального и фоновых процессов (далее шаблоны) и требуется произвести с их помощью подгонку распределения экспериментальных данных для установления ограничений на величину сечения сигнального процесса, учтя при этом присутствующие статистические и систематические погрешности в определении данных и шаблонов.

В работе приводится описание одного из методов решения описанной задачи для случая, когда распределение данных и шаблоны дискретны и заданы в виде гистограмм.

*E-mail: Petr.Mandrik@ihep.ru

1. ФУНКЦИЯ ПРАВДОПОДОБИЯ

Отправной точкой в решении задачи подгонки является построение функции правдоподобия $\mathcal{L}(\mathbf{p}) \sim P(\mathbf{x}|\mathbf{t},\mathbf{p})$, отвечающей вероятности наблюдения набора данных $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ при заданных шаблонах $\mathbf{t}$ и значениях параметров $\mathbf{p}$ в некоторой модели. Существуют различные пакеты, такие как HistFactory [1] и Theta [2], предоставляющие инструментарий для построения параметрической функции правдоподобия. Так, в HistFactory функция правдоподобия имеет следующее математическое представление:

$$\mathcal{L}(\mathbf{p}) = \prod_b \text{Poisson}\,(n_b|N_b) \prod_i f(p_i|P_i),$$

где $N_b = N_b(\mathbf{t},\mathbf{p})$ — ожидаемое значение в бине b, в простейшем случае равное сумме значений шаблонов в соответствующем бине; $f(p_i|P_i)$ — функция, накладывающая ограничение на значение параметра p_i.

После построения функция $\mathcal{L}(\mathbf{p})$ может быть использована для нахождения функции плотности вероятности и доверительных интервалов для параметров модели $\mathbf{p}$. При этом конкретный метод зависит от подхода к понятию вероятности и сложности рассматриваемой модели [3]. Так, в байесовском подходе $f(p_i|P_i)$ интерпретируется как априорное распределение плотности вероятности параметра p_i и совершается переход к обратной вероятности $P(\mathbf{p}|\mathbf{t},\mathbf{x}) \sim \mathcal{L}(\mathbf{p})$. Тогда набор параметров $\mathbf{p}$, при котором функция правдоподобия достигает максимума, также является наиболее вероятным при имеющемся измерении $\mathbf{x}$.

2. УЧЕТ ОШИБКИ ИЗМЕРЕНИЯ

Статистическая ошибка данных учитывается при выборе конкретного математического представления модели. Для учета статистической ошибки дискретных шаблонов может быть использован метод Барлоу–Бистон [4], в котором предлагается видоизмененная функция правдоподобия:

$$L(\mathbf{p}) \sim P(\mathbf{x}|\mathbf{t},\mathbf{p}) \to L(\mathbf{p}) \sim P(\mathbf{x}|\mathbf{T},\mathbf{p}) \cdot P(\mathbf{t}|\mathbf{T}),$$

где $P(\mathbf{t}|\mathbf{T})$ — дополнительный член, отвечающий вероятности наблюденного набора событий $\mathbf{t}$ в шаблонах при некотором неизвестном «истинном» числе событий в шаблонах $\mathbf{T}$. Таким образом, в функцию правдоподобия вводится набор параметров, число которых соответствует числу бинов/каналов и которые подобно другим параметрам могут быть найдены на стадии минимизации или же аналитически, при подходящей структуре функции правдоподобия. Как правило, в пакетах для построения функции правдоподобия используются различные вариации данного метода.

Систематические ошибки измерения в рамках данной задачи можно разделить на три категории:

1) влияющие на нормализацию шаблонов;

2) изменяющие значения в бинах шаблонов скореллированно;

3) изменяющие значения в бинах шаблонов нескоррелированным образом.

Учет систематических ошибок первых двух типов схож с методом Барлоу–Бистон для статистической ошибки и заключается в введении в функцию правдоподобия дополнительных «незначительных» параметров.

К примеру, пусть имеется модель сигнал–фон с параметрами S и B нормализации сигнального и фонового шаблона. Наличие ошибки σ_L в нормализации шаблонов изменит имеющуюся функцию правдоподобия следующим образом:

$$P(\mathbf{x}|S \cdot t_S, B \cdot t_B) \to P(\mathbf{x}|L \cdot t_S, L \cdot t_B) \cdot P(L|L_0, \sigma_L),$$

где $P(L|L_0, \sigma_L)$ — дополнительный множитель, ограничивающий изменение «незначительного» параметра L исходя из величины его номинального значения L_0 и неопределенности σ_L.

Для учета ошибки второго типа чувствительные к ней шаблоны создаются путем интерполяции между своими номинальными шаблонами и шаблонами, полученными при вариации измеренной с ошибкой величины. При этом в функции правдоподобия появляется новый «незначительный» параметр интерполяции. Например, если имеется некоторая величина C, измеренная с ошибкой σ_C и влияющая на значения бинов в фоновом шаблоне, то

$$P(\mathbf{x}|S \cdot t_S, B \cdot t_B) \to P(\mathbf{x}|S \cdot t_s, B \cdot I(\delta, t_B, t^C_{B_-}, t^C_{B_+})),$$

где $I(\delta, t_{B_0}, t^C_{B_-}, t^C_{B_+})$ — интерполирующая функция. Шаблоны $t^C_{B_+}$ и $t^C_{B_-}$ получены при значении величины C, равной $C_0 + \sigma_C$ и $C_0 - \sigma_C$ соответственно, аналогично тому, как ранее в анализе был получен номинальный шаблон t_B, отвечающий измеренному значению C_0.

К третьей категории относятся ошибки, учет которых путем скоррелированного изменения значений в бинах шаблонов приведет к недооценке величины погрешности. Их вклад в итоговые значения измерений может быть оценен посредством псевдоэкспериментов. Для этого строится дополнительная модель, шаблоны в которой создаются в измененных условиях, отвечающих вкладу рассматриваемой ошибки. Затем производится подгонка номинальной модели дополнительной моделью

$$L(\mathbf{p}) \sim P(\mathbf{t}_0, \mathbf{p}_0|\mathbf{t}, \mathbf{p})$$

и по величине смещения полученных параметров $\mathbf{p}$ от $\mathbf{p}_0$ оценивается неопределенность в их определении на реальных данных.

ЗАКЛЮЧЕНИЕ

В работе приведено краткое описание ряда методов, находящих применение в коллайдерных экспериментах при статистическом анализе данных. Так, в частности, в работе [5] рассматривались распределения дискриминанта выхода нейронной сети для данных и сигнального и фоновых процессов, полученных из моделирования Монте-Карло. Вклад неопределенности в измерении светимости оценивается путем введения влияющего на нормализацию шаблонов «незначительного» параметра, неопределенность в измерении энергии струй — как ошибка второго типа из представленной выше классификации, неопределенность в партонных функциях распределения — путем псевдоэксперимента. Подобным образом учитываются и другие ошибки.

СПИСОК ЛИТЕРАТУРЫ

1. *Cranmer K. et al.* CERN-OPEN-2012-016. 2012.
2. *Müller Th., Ott J., Wagner-Kuhr J.* http://www-ekp.physik.uni-karlsruhe.de/∼ott/theta/theta.pdf.
3. *Sinervo P.* eConf C. 030908. 2003. TUAT004.
4. *Barlow R., Beeston C.* // Comp. Phys. Commun. 1993. V. 77. P. 219.
5. *CMS Collab.* CMS-PAS-TOP-14-007.