Bootstrap-based fit of proton dipole scalar polarizabilities from real Compton scattering data

S Sconfietti$^{1,2}$, and P Pedroni$^2$
$^1$ Dipartimento di Fisica, Università degli studi di Pavia - Pavia, Italy
$^2$ INFN, Sezione di Pavia - Pavia, Italy
E-mail: stefano.sconfietti@pv.infn.it

Abstract. We describe the fitting technique that we recently applied to the analysis of the proton real Compton scattering data to extract the dipole scalar polarizabilities $\alpha_{E1}$ and $\beta_{M1}$. This fitting technique is based on the parametric bootstrap, which allows one to propagate the probability distribution of the experimental data to the fit results, with no a-priori assumption. It has never been applied before to the analysis of Compton data, and it offers several advantages: for instance, the automatic inclusion of the systematic errors and the propagation of those parameters that enter the model as inputs but that are not fitted. The most important result of our work is the definition of a realistic goodness-of-fit distribution even for those cases where the experimental points cannot be considered as independent Gaussian variables. We define a method that allows one, in different Physics domains, to define the correct $p$-values associated to a fit result.

1. Proton real Compton scattering
Elastic scattering of a real photon off a proton target, i.e., real Compton scattering (RCS), provides information on the electromagnetic structure of the proton. At very low energy, the incoming photon carries a quasi-static electromagnetic field and the response of the nucleon can be parametrized in terms of static polarizabilities. This is a reflection of the composite structure of the nucleon: the nucleon system cannot be described as a point-like particle characterized by only its static properties (mass, electric charge and anomalous magnetic moment), but its internal constituents cause the appearance of a collective response to the external field.

In this work, we show the new fitting technique that was used in Ref. [1] for the extraction of the static dipole scalar polarizabilities, $\alpha_{E1}$ and $\beta_{M1}$, from the proton RCS data below the pion production threshold. We use the theoretical framework of subtracted fixed-$t$ dispersion relations (DRs) [2–6]. Our fitting technique is based on the parametric bootstrap (see, for instance, [7] and references therein) and provides some advantages in the data analysis, as it will be shown in the following sections.

2. A new fitting strategy
The idea underlying our fit technique [8] is to assume that each point $E_i$, measured at a given set of parameters $x$, is the maximum likelihood estimate of its true (and unknown) value $E_i$. The true probability distribution $p(x,E_i)$ is approximated with the (known) probability distribution of the measured value $E_i$, i.e.

$$p(x,E_i) \simeq p(x,E_i).$$

(1)
Then, in case of a common scaling factor uncertainty, we define the bootstrap sampling as

$$B_{ij} = (1 + \delta_{ij})(E_i + \gamma_{ij}\sigma_i), \quad (2)$$

where $B_{ij}$ is a generic bootstrapped point with the index $i$ running over the number of data points ($n$) and the index $j$ indicating the $j^{th}$ bootstrap cycle. The $\gamma_{ij}$ parameters are sampled from the standard Gaussian distribution $N[0, 1]$, while the $\delta_{ij}$ are random numbers uniformly distributed, in each sub-set, as $U[-\Delta_k, +\Delta_k]$, being $\pm\Delta_k$ the percentage systematic uncertainty of each sub-set $k$ ($k$ runs from 1 to the number of the different data sub-sets $n_s$). Note that the probability distribution assigned to the systematic errors can be arbitrarily chosen on the basis of the experimental setup. If only statistical uncertainties have been taken into account, the systematic sources can be easily excluded from this procedure by just imposing $\delta_{ij} \equiv 0$.

After a complete cycle (defined as when the number of iterations is equal to the total number of experimental points) and once defined the rescaled statistical uncertainty $\sigma_{ij} \equiv (1 + \delta_{ij})\sigma_i$, the minimization procedure is performed on the function:

$$\chi^2_j = \sum_{i=1}^{n} \left( \frac{B_{ij} - T_i(\theta)}{\sigma_{ij}} \right)^2, \quad (3)$$

where $T_i$ is the prediction of a theoretical model depending on a set of unknown parameters $\theta$ to be adjusted to the data. The fit results from each cycle $j$ are stored and are considered as entries for the (unknown) probability distribution of the fit parameters. In this way, the probability distribution of the fit parameters is not assumed to be a-priori Gaussian, but is reconstructed directly from the probability distribution assigned to the experimental data. Furthermore, if we consider the more general situation where the model $T$ depends on some additional parameters $\theta_f$ not fitted to the data, but taken as input from other experiments, we can conveniently modify the minimization function to include the propagation of the statistical error of $\theta_f$ on the fit parameters $\theta$ as:

$$\chi^2_{b,j} = \sum_{i=1}^{n} \left( \frac{B_{ij} - T_i(\theta_f, \hat{\theta}_j)}{\sigma_{ij}} \right)^2, \quad (4)$$

whose minimum value can be written as

$$\hat{\chi}^2_{b,j} = \sum_{i=1}^{n} \left( \frac{M_{ij} - T_i(\theta_f, \hat{\theta}_j)}{\sigma_{ij}} \right)^2. \quad (5)$$

The $\hat{\chi}^2$ value, corresponding to the minimum of a standard $\chi^2$ variable, can be related to the $\hat{\chi}^2_{b,j}$ variable, as shown in Ref. [8] and a $p$-value has to be assigned to it. Usually, such a value is assumed to be distributed according to a standard $\chi^2$ distribution, which does not correctly describe a situation where there are correlated and/or non Gaussian variables. On the other hand, the goodness-of-fit probability distribution can be reconstructed also in this case using the parametric bootstrap technique. This distribution can be evaluated considering an ideal situation in which the experimental points are exactly the values predicted by our model. The sampling shown in (2) can be performed replacing each experimental data with $T_i(\theta_f, \hat{\theta})$, thus obtaining

$$M_{ij} = (1 + \delta_{ij})(T_i(\theta_f, \hat{\theta}) + \gamma_{ij}\sigma_i). \quad (6)$$

Based on this new sampling, the minimization function and its minimum (at every $j^{th}$ bootstrap cycle) can be respectively defined as

$$\hat{\chi}^2_{ib,j} = \sum_{i=1}^{n} \left( \frac{M_{ij} - T_i(\theta_f, \hat{\theta})}{\sigma_{ij}} \right)^2, \quad (7)$$
\[ \chi^2_{th,j} = \sum_{i=1}^{n} \left( \frac{M_{ij} - T_i(\theta_{f,j}, \hat{\theta}_j)}{\sigma_{ij}} \right)^2. \] (8)

The sampled parameters \( \theta_{f,j} \) are exactly the same as in (4), while the fit values of the parameters at every bootstrap cycle are, in general, different from the ones obtained from the fit of the bootstrapped data: for this reason we use \( \hat{\theta}_j \) instead of \( \hat{\theta}_j \).

Defining the variables
\[
\begin{align*}
\epsilon'_{ij} &= \frac{1}{\sigma_i} \left[ T_i(\theta_f, \hat{\theta}) - T_i(\theta_{f,j}, \hat{\theta}_j) \right], \\
\eta'_{ij} &= \frac{1}{(1 + \delta_{ij})\sigma_i} \left[ T_i(\theta_f, \hat{\theta}_j) - T_i(\theta_{f,j}, \hat{\theta}_j) \right], \\
D'_{ij} &= 2\epsilon'_{ij}\gamma_{ij}, \\
\Phi'_{ij} &= \eta'_{ij}^2 + 2\eta'_{ij} (\epsilon'_{ij} + \gamma_{ij}),
\end{align*}
\]
the \( \chi^2_{th,j} \) value can be decomposed as:
\[ \hat{\chi}^2_{th,j} = \chi^2_{th} + \sum_i \gamma_{ij}^2 + \sum_i \epsilon'_{ij}^2 + \sum_i D'_{ij} + \sum_i \Phi'_{ij}. \] (10)

Here, \( \hat{\chi}^2_{th} \) is defined as \( \hat{\chi}^2 \), but replacing \( E_i \) with \( T_i(\theta_{f,j}, \hat{\theta}_j) \). This parameter is identically zero by construction. Therefore, as a cross-check of the minimization procedure, we can verify that
\[ 0 = \hat{\chi}^2_{th,j} - \left[ \sum_i \gamma_{ij}^2 + \sum_i \epsilon'_{ij}^2 + \sum_i D'_{ij} + \sum_i \Phi'_{ij} \right]. \] (11)

The \( \Phi'_{ij} \) term includes all the dependence on the additional parameter set \( \theta_f \) and plays an important role in the determination of the goodness-of-fit distribution. Since the \( \chi^2 \) variable does not depend on the \( \theta_f \) parameters, the correct probability distribution computed in this framework is
\[ \chi^2_{u,j} = \sum_{i=1}^{n} \left( \frac{M_{ij} - T_i(\theta_{f,j}, \theta)}{\sigma_{ij}} \right)^2, \] (12)
with a minimum at
\[ \hat{\chi}^2_{u,j} = \sum_{i=1}^{n} \left( \frac{M_{ij} - T_i(\theta_{f,j}, \hat{\theta}_j)}{\sigma_{ij}} \right)^2 = \hat{\chi}^2_{th,j} - \sum_i \Phi'_{ij}. \] (13)

In this way, the parameters \( \theta_f \) affect the determination of the \( \hat{\theta} \) vector, but they do not introduce any bias in the probability distribution which the variable \( \hat{\chi}^2 \) belongs to. The \( \chi^2_{2,j} \) values thus give the goodness-of-fit distribution without the need of any a-priori assumption. When the systematic errors are not included in the fit procedure, the expression in (13) basically coincides with the standard \( \chi^2 \) distribution, i.e. with the sum \( \sum_i \gamma_{ij}^2 \). Viceversa, when systematic errors are included, the bootstrapped point are generated from the convolution \( \mathcal{U}[{-\Delta_i, \Delta_i}] \ast \mathcal{G}[ T_i(\theta_f, \hat{\theta}), \sigma_i^2] \), thus introducing a distortion with respect to the previous case. These considerations can be summarized as:

\[ \begin{align*}
\text{without systematics: } \hat{\chi}^2_{u,j} &= \left. \sum_i \gamma_{ij}^2 + \sum_i \epsilon'_{ij}^2 \right|_{\delta_{ij}=0} + \sum_i D'_{ij} \left|_{\delta_{ij}=0} \sim \sum_i \gamma_{ij}^2, \\
\text{with systematics: } \hat{\chi}^2_{u,j} &= \left. \sum_i \gamma_{ij}^2 + \sum_i \epsilon'_{ij}^2 \right|_{\delta_{ij} \neq 0} + \sum_i D'_{ij} \left|_{\delta_{ij} \neq 0} . \end{align*} \] (14)
Provided that the number of bootstrap replicas is sufficiently high, the $\hat{\chi}_E$ we label the expected value of such quantities as $\alpha$. In our recent work [1], we extracted the electromagnetic polarizabilities $\alpha_{E1}$ and $\beta_{M1}$ from the available RCS data below pion production threshold. The other four leading-order polarizabilities that enter the RCS differential cross section (namely, the spin polarizabilities) were taken from experimental evaluations, as described in detail in Ref. [1]. In this way, the minimization function that we used for the fit is given by the expression (4). In this section, we will focus on the goodness-of-fit probability distribution related to the fit results.

3. Results

In our recent work [1], we extracted the electromagnetic polarizabilities $\alpha_{E1}$ and $\beta_{M1}$ from the available RCS data below pion production threshold. The other four leading-order polarizabilities that enter the RCS differential cross section (namely, the spin polarizabilities) were taken from experimental evaluations, as described in detail in Ref. [1]. In this way, the minimization function that we used for the fit is given by the expression (4). In this section, we will focus on the goodness-of-fit probability distribution related to the fit results.
In Fig. 1 we show the probability distributions of the quantities in (18), which determine the goodness-of-fit distribution. We observe that: (I) the errors on the fitted parameters cannot be small, due to the relatively big values of the $E[\epsilon'^2]$ and $E[D']$ terms; (II) the systematic errors give a relevant contribution to the data analysis, being the $E[\epsilon'^2]$ term increased by a factor 5 as soon as the systematic errors are included in the data analysis and (III) the sampling of the non-fitted parameters is under control, being the $E[\Phi']$ small. Furthermore, if we do not propagate the uncertainties of the non-fitted polarizabilities, the goodness-of-fit distributions get closer to a standard reduced $\chi^2$ distribution, thus enlightening once more the role of the uncertainties of the non-fitted polarizabilities. The goodness-of-fit CDFs are shown in Fig. 2,

![Figure 1](image-url)

Figure 1. Decomposition of the $\hat{\chi}^2_{th,j}$ variable when systematic errors are excluded (blue curves) or included (green curves). Upper panels (from left to right): $\hat{\chi}^2$, $\chi^2_b$ and $\gamma^2$. Lower panels (from left to right): $\epsilon^2$, the $D$ term and $\Phi$. See Eq. (18) for the notation.

and the distortion effect caused by the systematic errors is clearly visible. However, even when the systematic errors are excluded, the limit probability distribution for the $\hat{\chi}^2$ variable is not so close to a proper $\chi^2$. This feature is related to the analytical structure of the fit model and to the presence of some offsets in the experimental data; it can be also quantified with the $E[\epsilon'^2]$ and $E[D']$ terms, which are not so small to be discarded even when only the statistical errors are included in the analysis. This difference with the expected $\chi^2$ distribution is not a surprise indeed, but it is well explained by the decomposition given in (14). Numerically, our results [1] read

$$\alpha_{E1} = 12.03^{+0.48}_{-0.53}, \quad \beta_{M1} = 1.77^{+0.52}_{-0.54},$$

$$\rho_{\alpha_{E1}-\beta_{M1}} = -0.72, \quad \hat{\chi}^2 = 1.25, \quad p\text{-value} = 12\%,$$

and are in very good agreement with the ones quoted by the PDG [9].

4. Conclusions

We implemented a new fitting technique for the extraction of the scalar dipole static polarizabilities, $\alpha_{E1}$ and $\beta_{M1}$, from the proton RCS data. Apart from the numerical results, which are new estimates of the electric and magnetic polarizabilities obtained within the framework of subtracted fixed-$t$ DRs, we pointed out the advantages of our fitting technique. We
showed how to consistently include the systematic errors and how do they affect the goodness-of-fit distribution. Furthermore, we discussed how to propagate the uncertainties of those parameters that are not fitted from the data, but enter the model.

We remark that our fitting technique, even if it was applied here to the analysis of the RCS data [1, 10], can be conveniently applied in every different Physics framework.

5. Acknowledgments

We are grateful to the INPC2019 organizers for giving us the opportunity of presenting our work during the conference. We want to thank Barbara Pasquini, that provided all the theoretical framework for the analysis of the RCS data.

References

[1] Pasquini B, Pedroni P and Sconfietti S 2019 J. Phys. G46 104001 (Preprint 1903.07952)
[2] Drechsel D, Gorchtein M, Pasquini B and Vanderhaeghen M 1999 Phys. Rev. C 61 015204 (Preprint hep-ph/9904290)
[3] Holstein B R, Drechsel D, Pasquini B and Vanderhaeghen M 2000 Phys. Rev. C 61 034316 (Preprint hep-ph/9910427)
[4] Pasquini B, Drechsel D and Vanderhaeghen M 2007 Phys. Rev. C 76 015203 (Preprint 0705.0282)
[5] Drechsel D, Pasquini B and Vanderhaeghen M 2003 Phys. Rept. 378 99–205 (Preprint hep-ph/0212124)
[6] Pasquini B and Vanderhaeghen M 2018 Ann. Rev. Nucl. Part. Sci. 68 75–103 (Preprint 1805.10482)
[7] Davidson A C and Hinkley D V 1997 Bootstrap Methods and their Application (Cambridge University Press)
[8] Pedroni P and Sconfietti S 2019 (Preprint 1909.03885)
[9] Patrignani C et al. (Particle Data Group) 2016 Chin. Phys. C 40 100001
[10] Pasquini B, Pedroni P and Sconfietti S 2018 Phys. Rev. C98 015204 (Preprint 1711.07401)