Model selection for $K^+\Sigma^-$ photoproduction within an isobar model

P. Bydžovský,[1] A. Cieplý,[1] D. Petrellis,[1] D. Skoupil,[1] and N. Zachariou[2]
[1]*Nuclear Physics Institute, CAS, Řež/Prague, Czech Republic*
[2]*University of York, York YO10 5DD, United Kingdom*

(Received 30 June 2021; accepted 16 November 2021; published 13 December 2021)

We utilize an isobar model to investigate the $K^+\Sigma^-$ photoproduction off a neutron in the resonance region. Except for the Born terms, we include high-spin (spin-3/2 and spin-5/2) nucleon resonances in the consistent formalism together with a few Δ and kaon resonances to achieve an acceptable agreement with data. Interestingly, we reveal that no hyperon resonances are needed to achieve a reasonable description of data. On the other hand the $N(1720)3/2^+$ resonance was found to be very important for correct description of data. The free parameters of the model were fitted to experimental data from the LEPS and CLAS Collaborations on either differential cross sections or photon beam asymmetry. The novel feature of the fitting procedure is the use of a regularization method, the least absolute shrinkage selection operator, and information criteria in order to choose the best fit.

DOI: 10.1103/PhysRevC.104.065202

I. INTRODUCTION

The study of the kaon-hyperon photo- and electroproduction from nucleons in the third nucleon resonance region provides important information about the spectrum of baryon resonances and interactions in the systems of hyperons and nucleons, which arise from quantum chromodynamics. Not only do we aim at studying the reaction mechanism, we also focus on obtaining more information about the existence and properties of the so-called missing resonances, which have been predicted by quark models but have not been seen in the pion production of πN scattering processes [1,2]. These states may have escaped experimental confirmation due to their stronger decay coupling to $K\Lambda$ and $K\Sigma$ rather than to the more well-known pion final states; see results of the coupled-channel analysis [3] and outcomes of the partial-wave analysis [4].

A plethora of theoretical studies on hyperon production have been performed over the past decades with focus primarily given to the $K^+\Lambda$ production channel off the proton due to the large amount of available experimental data; see, e.g., Ref. [5] and references therein. The analyses before 2004 unfortunately suffered from a lack of high-quality experimental data [6] but the situation changed dramatically after new high-duty-factor accelerators, providing good quality, high-current, and polarized continuous beams, were constructed in Jefferson Lab (CEBAF) and Bonn University (ELSA).

The $K\Sigma$ photoproduction channels were also studied, but, similarly to the case of the $K^+\Lambda$ channel, only after 2004 could these studies be based on more good-quality data, mainly due to the data in the resonance region of the $K^+\Sigma^0$ channel from the CEBAF, ELSA, MAMI (University in Mainz), and SPring-8 (Japan) facilities; see, e.g., Ref. [7] and references therein. However, data in the other channels with the Σ hyperon are still quite scarce in comparison with the number of data in the $K^+\Lambda$ and $K^+\Sigma^0$ channels; see Tables 11–15 in a recent overview by Ireland, Pasyuk, and Strakovsky [8]. A combined analysis of all four channels with Σ was performed by Mart and Kholili in Ref. [7]. The background part of the amplitude was constructed using an isobar model and the resonant part using a multipole formulation. The problem with unbalanced data sets for the $K^+\Sigma^0$ and the other channels (about a factor of 10) was solved by introducing a weighting factor and its optimum value was used in the analysis.

For the time being, the database of the channels using neutron targets is very limited, with available measurements of the differential cross section for $K^+\Sigma^-$ [9,10] and $K^0\Lambda$ reactions [10]. Inclusive momentum spectra in K^0 photoproduction off deuteron were measured in the threshold region at LNS of Tohoku University [11]. There are only two measurements of the beam asymmetry Σ: from the LEPS Collaboration [9] associated with very limited kinematical coverage, and just recently a precise measurement from the CLAS Collaboration [12], which covers a wide range of kinematics. There are also recent results on beam-target helicity asymmetry E, also from the CLAS Collaboration [13].

In the present paper we reanalyze the new CLAS data on the beam asymmetry [12] and the other older data in the $K^+\Sigma^-$ channel using an isobar model and the LASSO (least absolute shrinkage and selection operator) method of adjusting free parameters of the model (fit L). Our previous analysis of these data, denoted here as fit M, was done using an ordinary χ^2 method, similarly to the work done in Ref. [5] for the $K^+\Lambda$ channel, and this fit was already presented in Ref. [12] in comparison with the new data. Here we will give more details on the model fit M and compare it with the new model fit L mainly in view of their resonance content

TABLE I. Characteristics of included resonances with their masses and widths taken as the Particle Data Group (PDG) Breit-Wigner averages. The available branching ratios to the $K\Lambda$ and $K\Sigma$ channels are also taken from the PDG [17]. For the nucleon and Δ resonances, the values of coupling constants, g_1 and g_2, show the baryon-$K\Sigma$ scalar and tensor couplings obtained in our fit, while for the K^* and K_1 states they represent the vector and tensor couplings, respectively. We show values of g_1 and g_2 achieved with MINUIT only (denoted as fit M) and with the LASSO method (fit L).

		Mass	Width	Branching ratio		Fit M		Fit L	
Tag	Resonance	(MeV)	(MeV)	$K\Lambda$	$K\Sigma$	g_1	g_2	g_1	g_2
K*	$K^*(892)$	891.7	50.8			0.366 ± 0.024	1.103 ± 0.198	0.310 ± 0.019	
K1	$K_1(1270)$	1270	90			-1.448 ± 0.189	0.473 ± 0.156		
N3	$N(1535)\ 1/2^-$	1530	150			-0.709 ± 0.071			
N4	$N(1650)\ 1/2^-$	1650	125	0.07	0.00	0.314 ± 0.034		-0.085 ± 0.006	
N8	$N(1675)\ 5/2^-$	1675	145			-0.013 ± 0.001	0.022 ± 0.003	-0.010 ± 0.001	0.003 ± 0.002
N6	$N(1710)\ 1/2^+$	1710	140	0.15	0.01	-0.940 ± 0.093			
N7	$N(1720)\ 3/2^+$	1720	250	0.05	0.00	-0.098 ± 0.017	-0.082 ± 0.002	-0.187 ± 0.004	-0.126 ± 0.002
P4	$N(1875)\ 3/2^-$	1875	200	0.01	0.01	-0.220 ± 0.023	-0.223 ± 0.023	-0.042 ± 0.015	0.025 ± 0.013
P1	$N(1880)\ 1/2^+$	1880	300	0.16	0.14	-0.050 ± 0.064			
Mx	$N(1895)\ 1/2^-$	1895	120	0.18	0.13	-0.063 ± 0.005		0.019 ± 0.002	
P2	$N(1900)\ 3/2^+$	1920	200	0.11	0.05	-0.051 ± 0.005	-0.004 ± 0.001	0.027 ± 0.003	0.010 ± 0.001
M4	$N(2060)\ 5/2^-$	2100	400	0.01	0.03	-0.00001 ± 0.0001	0.003 ± 0.0003	-0.003 ± 0.0001	0.004 ± 0.0002
M1	$N(2120)\ 3/2^-$	2120	300			-0.034 ± 0.014	-0.010 ± 0.013	0.0003 ± 0.001	0.0 ± 0.0001
D1	$\Delta(1900)\ 1/2^-$	1860	250		0.01	0.298 ± 0.028			
D2	$\Delta(1930)\ 5/2^-$	1880	300						
D3	$\Delta(1920)\ 3/2^+$	1900	300						
D4	$\Delta(1940)\ 5/2^-$	1950	400						
S1	$\Sigma(1660)\ 1/2^+$	1660	100						
S2	$\Sigma(1750)\ 1/2^-$	1750	90						
S3	$\Sigma(1670)\ 3/2^-$	1670	60						
S4	$\Sigma(2010)\ 3/2^-$	1940	220						

and a quality of data description. We deem that the more elaborate statistical method is more sensitive to a selected resonant content of the model. This advanced method will be also used in our further more robust analysis of data in the photoproduction channels with the Σ hyperon.

The paper is organized as follows: In Sec. II, we discuss the isobar model which we use for describing the Σ^- photoproduction reaction off the neutron. Section III deals with the free parameters in the model and with the new method of their fitting to the data. In Sec. IV we discuss the obtained results and Sec. V provides a brief summary and conclusions.

II. MODEL DESCRIPTION

The current model based on an effective Lagrangian in the tree-level approximation is constructed to describe data only in the $K^+\Sigma^-$ channel assuming no final-state interaction.

The nonresonant part of the amplitude consists of the Born terms and exchanges of resonances in the t channel (K^* and K_1 kaon resonances) and u channel (Σ^* hyperon resonances). The main coupling constant $g_{K^+\Sigma^-n} = \sqrt{2}\, g_{K^+\Sigma^0 p}$, that determines the strength of the Born terms, was taken from the model constructed for the $K^+\Lambda$ channel [5] and kept unchanged in the present fit. The resonant part is modeled by s-channel exchanges of nucleon and Δ resonances with masses from around the process threshold up to about 2 GeV. Hadronic form factors included in the strong vertexes account for a hadron structure and regularize the amplitude at large energies. The form factors are introduced in a way that keeps gauge invariance in analogy with the method used in Refs. [5] and [14]. The relevant formulas to show gauge invariance of the amplitude in the case of $K^+\Sigma^-$ photoproduction are given in Appendix A.

The considered set of nucleon resonances was motivated by previous analyses of $K^+\Lambda$ and $K\Sigma$ photoproduction [5,14] and [15], respectively. Some additional nucleon resonances decaying strongly into the $K\Sigma$ channel were also considered in our analysis, together with Δ and Σ resonances used to complement the model in the s-channel and u-channel sectors. The free parameters of a particular model, the coupling constants, and ranges of hadronic form factors for a given set of resonances, were fitted to the data, and the quality of the model was checked by comparing its prediction with the data. In the end, a variant with the smallest χ^2/n.d.f. and reasonable values of the parameters was selected. In our current best fit with the CERN MINUIT library [16], which is also presented in Ref. [12] and which was aimed at description of the new CLAS data on asymmetry, we have used 14 resonances which are shown in Table I. It includes the *** and **** baryon states with most of them decaying into the $K\Lambda$ and $K\Sigma$ channels [17] and also some other Δ and Σ resonances which we have taken into account in our analysis. We note that we used only the statistical errors when computing the χ^2, which results in a relatively large value $\chi^2/\text{n.d.f.} = 2.39$ obtained for the selected solution. The reason for using only statistical errors was missing systematic errors in some data sets. When

systematics is taken into account, the χ^2 value usually drops, but the inclusion of systematics does not change the quality of results.

In the following section we will introduce the fitting procedure and its extension—the least absolute shrinkage selection operator (LASSO)—which is a more sophisticated method for adjusting the free model parameters. The LASSO method was used in a recent analysis of pion photoproduction and study of baryon resonances [18]. The advantage of this method lies in its capability of removing redundant parameters and thus limiting the number of contributing resonances.

III. ADJUSTING MODEL PARAMETERS

The model used in our study of strangeness production is an effective model, with coupling constants and cutoff values of hadron form factors not determined. Because of this, experimental data play a crucial role in fixing these parameters that enhance the predictive power of our model.

The free parameters to be adjusted are the main coupling constant, $g_{K^+\Sigma^-n}$, cutoff parameters for the hadron form factors of background and resonant terms, and the couplings of resonances introduced. Please note that the $g_{K^+\Sigma^-n}$ coupling was kept unchanged during the fitting procedure with MINUIT but it was allowed to vary within the boundaries shown in Eq. (2) during the fitting process with the LASSO method. There is one free parameter for spin-1/2 resonance and two free parameters for spin-3/2 and -5/2 (nucleon and hyperon) resonances, while each kaon resonance introduces two additional free parameters (vector and tensor couplings).

We calculate the χ^2, in order to check whether a given hypothesis describes the given data well. The optimum set of free parameters $(c_1, \ldots, c_n)$ for a given set of data points $(d_1, \ldots, d_N)$ is obtained by minimizing the χ^2, calculated as follows:

$$\chi^2 = \sum_{i=1}^{N} \frac{[d_i - p_i(c_1, \ldots, c_n)]^2}{\left(\sigma_{d_i}^{\text{stat}}\right)^2}, \tag{1}$$

where N is the number of data points, n is the number of free parameters, and p_i represents the theoretical prediction of observables (differential cross sections and photon beam asymmetry in this case) for the measured data point d_i.

The minimization was done with the help of least-squares fitting method making use of the MINUIT library [16]. The $g_{K^+\Sigma^-n}$ coupling constant was kept inside the limit of the 20% broken SU(3) symmetry [19],

$$\sqrt{2} \times 0.8 \leqslant \frac{g_{K^+\Sigma^-n}}{\sqrt{4\pi}} \leqslant \sqrt{2} \times 1.3. \tag{2}$$

Moreover, the cutoff parameters of the hadron form factors for both background and resonant terms were kept inside the limits from 0.5 to 3.0 GeV, in order to avoid too soft or too hard form factors.

One of the problems that arise when fitting a theoretical model to experimental data is that of overfitting the data. This means that, although a more complex model (one with more parameters) may improve the fitting to the existing data, that model may fail to generalize to new data, resulting thus in a poor description of reality. The set of techniques that have been developed to combat this problem is known in the machine learning literature with the name "regularization" [20,21]. Typically, regularization involves the addition of a penalty term in the error function that prevents the parameters of the model from taking large values, when the total error function is minimized. The penalty term may take various forms and the amount by which it contributes to the total error function is determined by the coefficient multiplying it, called the regularization parameter and commonly denoted by λ. Higher values of λ tend to push more parameters close to zero, or even exactly to zero, thus favoring simpler models (with fewer parameters), which may underfit the data. With different values of λ leading to different sets of parameters, and hence different models, the choice of the optimal λ becomes a problem of model selection. For this choice we intend to use criteria based on information theory, like the Akaike and the Bayesian information criteria [22], that have been recently used in similar problems [18].

Generally speaking, the χ^2 is a good measure to determine underfitting but it says nothing about overfitting [18]. For this reason we turn to the LASSO method in order to select the simplest model that can describe the data with the minimal amount of resonances. In order to do so, we introduce a penalty term

$$P(\lambda) = \lambda^4 \sum_{i=1}^{N_{\text{res}}} |g_i|, \tag{3}$$

where λ is the regularization parameter, which we determine using the information criteria, g_i represents couplings of resonances, and N_{res} is the number of assumed resonances. We opt for the fourth power of λ, as a higher power enables us to move quickly through the region of large values of λ and give more weight to the region of small λ. The power affects also the step in λ, resulting in a finer sampling of the region of small λ. The reason why we want to stress the region with small λ is that in this region more and more resonances are allowed to contribute and the results can change abruptly with only slight changes in λ. Moreover, in Eq. (3) each resonance is penalized through its coupling g_i, on top of the standard definition of the χ^2 in Eq. (1). In order to incorporate the penalty term, we define the penalized χ_T^2 as

$$\chi_T^2 = \chi^2 + P(\lambda). \tag{4}$$

In practical calculations, we scan a range of λ values and in each step minimize the χ_T^2. With help of the χ_T^2 values we then turn to several information criteria, which serve as a tool to determine the optimal λ value and select the most suitable model for the description of the given data. The three information criteria used in this work are the Akaike information criterion (AIC) [23], a finite sample size corrected version of the AIC (further on referred to as AICc) [24], and the Bayesian information criterion (BIC) [25], which are respectively defined as

$$\text{AIC} = 2n + \chi_T^2, \tag{5a}$$

$$\text{AICc} = \text{AIC} + \frac{2n(n+1)}{N-n-1}, \tag{5b}$$

$$\text{BIC} = n\ln(N) + \chi_T^2. \tag{5c}$$

where n is the number of parameters which changes as a function of λ and N is the number of data points.

Even though it may be self-evident, let us stress that the information criteria are useful in selecting the best model in the particular set, and that models which are not included in the set remain out of consideration. If all the models of the set are poor, the information criteria will still guide us towards the best model, but even that relatively best model might be poor in the absolute sense. Therefore, while using the information criteria, every effort must be made to ensure that the set of models is well founded. For detailed information on the regularization method and a brief derivation of the information criteria, see Appendix B.

A. Experimental data

In the fitting procedure, we used altogether 674 data points to fit around 20 free parameters of our model. The currently available experimental data in the $K^+\Sigma^-$ channel are data on the differential cross section, photon beam asymmetry, and beam-target asymmetry only. In the last two decades, data on the differential cross section and the beam-spin asymmetry were determined by the CLAS [10,12] and LEPS [9] Collaborations. The CLAS Collaboration provided measurements for a wide range of kinematics. The differential cross sections were measured for photon laboratory energies from the near-threshold value $E_\gamma^{\rm lab} = 1.15$ to 3.55 GeV and for $\cos\theta_K^{\rm c.m.}$ in the range from -0.85 to 0.85, whereas the beam-asymmetry data span the region of photon laboratory energies $E_\gamma^{\rm lab}$ from 1.1345 to 2.276 GeV and $\cos\theta_K^{\rm c.m.}$ from -0.7687 to 0.7484. The LEPS Collaboration provided complementary measurements at forward angles as they focused only on the region of $\cos\theta_K^{\rm c.m.}$ from 0.65 to 0.95. The most recent results on the beam-spin asymmetry from the CLAS Collaboration provided tight constraints due to their precision and kinematical coverage [12]. We have exploited these data, except for the beam-target asymmetry data, in a similar fashion to what we have done in Refs. [5,14] for the $K^+\Lambda$ channel.

In the fit M we considered the experimental data for energies smaller than 2.6 GeV whereas the fit L was fitted up to energies 2.95 GeV.

B. The course of the fitting procedure

The goal of the fitting process is to find the global minimum, i.e. the set of parameters which describe the data in the best way and produce the smallest χ^2. Unfortunately, this is not an easy task as we work in a very large parameter space with numerous local minima scattered around. Thus, the results of the fitting process depend also on the starting values of the parameters that are being adjusted.

What makes the situation even worse is the fact that the χ^2 is only a mathematical tool that illustrates the goodness of fit. Hence, the results with similar (or even identical) χ^2 values can give rather different predictions of the observables as we may end up in different local minima. In order to distinguish satisfactory results from the unreliable ones, we pay attention not only to the final χ^2 value but also to the values of fitted parameters. Moreover, we briefly check the agreement of the fit with data.

Extremely helpful in recognizing valuable outcomes of the fitting procedure is the LASSO method as described above. The first step in using this technique is initializing the resonance parameters with random values in the range from -1 to $+1$. The main coupling constant, $g_{K^+\Sigma^-n}$, was initialized with a random value within the range shown in Eq. (2) and the initial values for the cutoff parameters were chosen inside the range of (0.8,3.0) GeV. We use the forward LASSO technique, which means that we decrease the value of λ in the penalty term [see Eq. (3)]. The starting value of λ is chosen to be 3 and we decrease it by subsequent steps of either 0.2 or 0.1 until we reach zero. The role of this parameter, and the penalty function as a whole, rests in turning off model parameters. The larger the λ the more model parameters are turned off. This in turn means that with large λ we can produce very economical models with a very few parameters but their agreement with data (as illustrated by the χ^2 value) is rather clumsy. This, in our opinion, is a clear sign of underfitting the data as the model includes too few parameters which do not allow it to capture the data. Usually, we arrive at reasonable fits when λ decreases to around 1. For smaller values of λ, the values of the information criteria [Eqs. (5)] tend to increase which is an indication of overfitting, i.e., introducing more parameters than is needed. In some sense, the value of λ for the minimal value of the information criterion shows how far we are from the ideal number of parameters; i.e., $\lambda \approx 0$ would mean that we are close to the ideal number of parameters, but when λ is significantly larger than zero we have considered too many parameters in the beginning of the process.

What is more, there seems to be a strong influence of the λ value on the fit results and the convergence of the fitting process. It seems that large values of λ prevent the minimizer from converging as they turn off too many parameters. In the MINUIT library, there are several different minimizers available. We decided to use the Minimize minimizer as it combines the merits of Migrad and Simplex minimizers. When we include also the hyperon resonances, we observe that for $\lambda = 3$ there is no convergence by either Migrad or Simplex minimizers; for λ approximately between 0.6 and 2.5 only the Simplex method converges; below 0.5 Simplex and Migrad alternate; and only for $\lambda < 0.05$ the Migrad converges without Simplex being called.

In the LASSO method, we have to consider also the number of decimal digits for each parameter, i.e., from which value the parameter is considered numerically zero and therefore does not appear in the total number of fitted parameters. For example, when we do calculations with individual resonances whose coupling parameters are at the order of 10^{-5} or 10^{-6}, their contributions are almost zero. Thus we reckon that we should not take so many decimal places into account. From where we stand, it seems that taking into account four decimal places for each resonance is enough. When we use more, we artificially turn on parameters that are not needed and include resonances which do not contribute. In this way we would artificially increase the number of parameters n and decrease the precision of the information criteria.

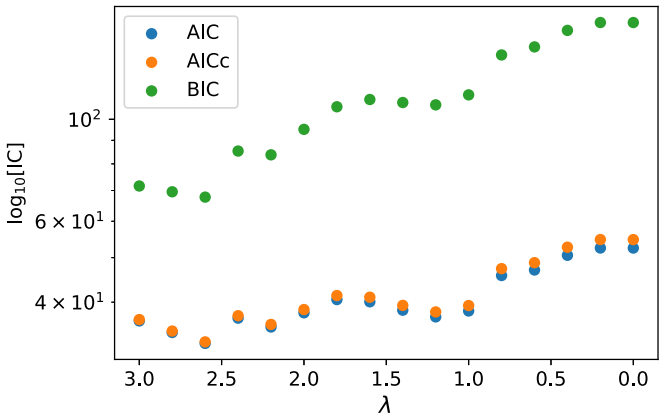

FIG. 1. Information criteria (IC) values in dependence on the parameter λ. We show the Akaike (blue), corrected Akaike (orange), and Bayesian (green) information criteria. Please note that the scale of the vertical axis is logarithmic. Also note that it is not the absolute size of the individual information criteria which is important, it is the differences between IC values for different λ.

The number of parameters n encompasses all of the parameters we introduce, including the main coupling constant $g_{K\Sigma N}$ and the cutoffs for the background, $\Lambda_{\rm bgr}$, and resonant terms, $\Lambda_{\rm res}$. The values of these three parameters are not included in the penalty term. The reason for doing so is that we do not want these parameters to vanish. On the contrary, the rest of the fitted parameters do appear in the penalty term which then pushes their values to zero and tends to turn off their contributions.

Similarly to what we have done in our analysis of the $K^+\Lambda$ channel, we tried to modify the masses and widths of the nucleon resonances within the ranges provided by the PDG tables. Unfortunately, this does not lead to any better agreement with data (even though in some cases it can produce smaller values of the χ^2).

As mentioned above in the case of using pure MINUIT, even when we exploit the LASSO method we may end up in one of many local minima. From what we observe, we can conclude that even with the LASSO method the minimum which we find strongly depends on the initial values of the fitted parameters. Therefore, we cannot avoid the pitfall of local minima and we cannot guarantee that the fit we end up with is the best fit which exists and can be found (which would correspond to a truly global minimum). We rather say that the models we show are among the best fits we could find.

The initial set of resonances for the LASSO method was the one that was first found with help of MINUIT. We took those resonances and ran the forward LASSO, i.e., we introduced the penalty term to the χ^2 definition, as can be seen in Eq. (4), set the λ at 3.0 and reduced it in subsequent steps of 0.2 until we reached $\lambda = 0$. When we plot values of the information criteria in dependence on the λ, Fig. 1, we observe two distinct minima. One is around $\lambda = 2.5$ and the other somewhere around $\lambda = 1$. We were discouraged from taking the first minimum, the deeper one, as the best fit since its χ^2 is larger than the χ^2 of the other minimum and the correspondence with data is also much worse. Clearly, the first minimum is a result of underfitting. The other minimum was, on the other hand, acceptable as its χ^2 value appeared to be reasonable, $\chi^2/\text{n.d.f} = 3.2$, and its agreement with data is good.

In subsequent fits, we added hyperon and Δ resonances on top of this core set. It seems though that no hyperon resonances are needed for data description in the $K^+\Sigma^-$ channel. None of them was found to be substantial in either the pure χ^2 fitting or using the LASSO method since the couplings of hyperon resonances were zero for every value of λ. What is more, with hyperon resonances included and values of $\lambda > 0.05$ MINUIT was unable to reach convergence.

We did not rely only on the LASSO outcomes but we also compared several results with experimental data. All of the fits with the hyperon resonances tend to diverge at very forward angles. We deem this is so basically because there are not enough data in this kinematical region, especially for $\cos\theta_K^{\rm c.m.} = 0.95$, and thus there is nothing that can control the behavior of the models in this region. At central angles, the majority of fits with the hyperon resonances do not capture the data at the very threshold and some of the fits overestimate the first peak in the backward angles. Besides, in the fits where the λ allows some of the hyperon resonance couplings to acquire nonzero values, the information criteria (both the Akaike and Bayesian ones) tend to have larger values. This is a clear indication that these parameters (resonances) are not substantial for data description. Moreover, in the plot of the information criterion values against the λ parameter, we see a significant drop once the hyperon resonances do not contribute. This is a very interesting observation and it strongly corroborates the claim of hyperon resonances not being important for this channel.

The results with additional Δ resonances are slightly better than the results with hyperon resonances. However, once we add the Δ resonances, i.e., we add more resonances to the core set of resonances, the χ^2 gets worse. In other words, we add more free parameters and thus the model has more freedom to adapt to data but contrary to common sense the χ^2 becomes larger not smaller. We again deem this to be a clear indication of unimportance of the Δ resonances for reliable data description in this channel.

IV. DISCUSSION

We concluded the fitting process with two distinct models. One of them, which we will refer to in the subsequent text by fit M, was achieved using solely the MINUIT procedure for minimizing the χ^2 (see Sec. II and Ref. [12]). The other one, which we will call fit L, is a result of using MINUIT together with the LASSO technique as presented in Sec. III.

The fit M incorporates altogether 14 resonances: two kaon resonances, multiple nucleon resonances, one Δ resonance and no hyperon resonances. The latter feature is rather surprising given our experience with describing the $K^+\Lambda$ production channel where a plethora of hyperon resonances contribute in a significant way (see Ref. [5]). The obtained couplings g_1 and g_2 listed in Table I are all reasonable and the same can be said about the hadronic form factor ranges $\Lambda_{\rm bgr} = 0.87$ GeV and $\Lambda_N = 1.45$ GeV, see [14] for a description of these

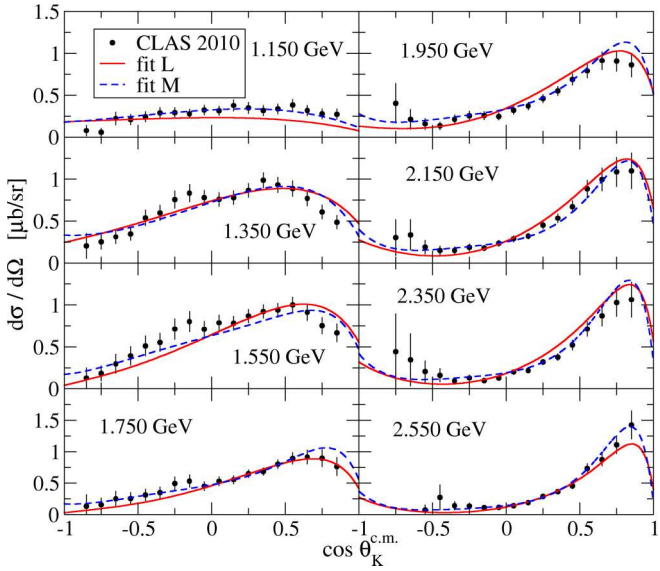

FIG. 2. The differential cross section as a function of the kaon center-of-mass angle $\theta_K^{\text{c.m.}}$. The points represent data from CLAS [10]. The solid red line and dashed blue lines show the models with parameters determined applying the LASSO method and MINUIT alone, respectively.

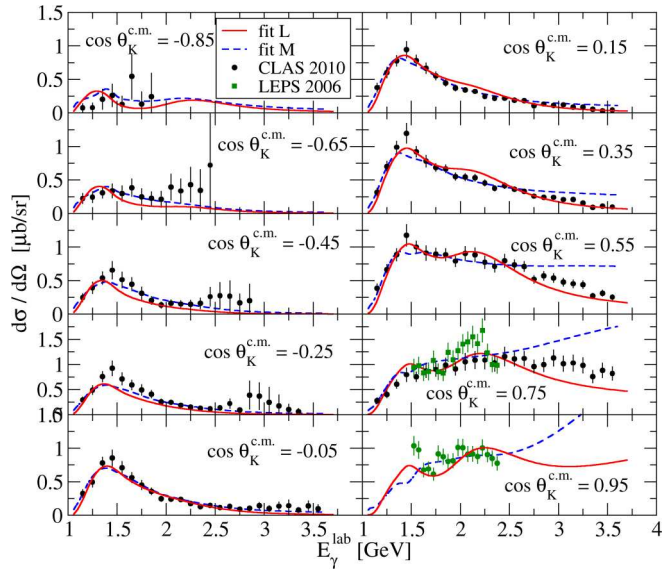

FIG. 3. The differential cross section as a function of the incident photon energy E_γ^{lab}. CLAS and LEPS data are shown with the black and green points, respectively. The curves indicate the two methods used for obtained the best description of the data as described in Fig. 2.

parameters. The fit produces results which are in a very good agreement with the cross section and beam asymmetry data. This analysis also showed that hyperon resonances do not play a key role in the description of the available data, and their inclusion results in negligible effects.

Besides the best MINUIT fit, we revealed another noteworthy fit using only the MINUIT procedure. We do not show its comparison with data as it is hardly distinguishable from the fit M. Its χ^2 is 2.33, there are two Δ resonances, and one of these, the D3 = $\Delta(1920)\,3/2^+$ one, has a significant effect on data description. When we omit this Δ state, the beam asymmetry falls and is in accordance with data only at very forward kaon angles. Moreover, as there is a hyperon resonance S1 = $\Sigma(1660)\,1/2^+$ included, we could observe its effect on the beam asymmetry data description, which is negligible. This corroborates our observation with the fit M. The N7 = $N(1720)\,3/2^+$ state in this fit behaves in a similar way to the fit M, i.e., when we leave it out the beam asymmetry drops substantially, in some angular regions even to negative values of the beam asymmetry.

In comparison to the fit M, the χ^2 of the fit L, $\chi^2 = 3.42$ is significantly larger. The large χ^2 value is the price we had to pay for a smaller number of parameters. In the fit L, there are mere 17 parameters and 9 resonances while in the fit M there are 25 parameters and 14 resonances; see Table I.

In Fig. 2, we compare the two best fits with one another and with the experimental data on differential cross sections from the CLAS Collaboration. The overall trend of the data is captured by both models. The fit M produces a slightly sharper peak at forward angles and for photon laboratory energies around 2 GeV overshoots the data, while the fit L is more moderate and tends to underestimate the data at highest photon laboratory energies shown. Unfortunately, neither of the models is able to capture the two-peak structure of the differential cross section above the threshold at energies 1.35 and 1.55 GeV as both models produce a smooth differential cross section at central kaon angles.

The differential cross sections in dependence on the photon laboratory energy E_γ^{lab} are shown in Fig. 3. We again compare our best fits with the experimental data from the CLAS and LEPS Collaborations. In all angular bins, the models are in a satisfactory agreement with the data. The fit M tends to diverge quickly at very forward angles, while the fit L, on the other hand, underestimates the data at energies above 2.5 GeV in the $\cos\theta_K^{\text{c.m.}} = 0.75$ bin. Also noteworthy are the structures that the fit L produces at forward angles. Whereas the fit M produces more or less flat cross sections, the fit L shows two broad peaks which are also supported by the data.

Among the set of included nucleon resonances in the fit M, the most noteworthy is the contribution of the $N(1720)\,3/2^+$ nucleon resonance whose omission leads to substantially decreased cross sections. An important effect of this resonance was also observed in the $K^0\Sigma^+$ channel [26] and in the combined analysis of all Σ channels performed in Ref. [7]. We also note a significant contribution of the $N(1895)\,1/2^-$ state with a relatively large $K\Sigma$ branching ratio. This state was found in Ref. [7] to be among the most significant states for dynamics in the $K\Sigma$ channels. The role of the $\Delta(1900)\,1/2^-$ resonance is in modeling the peak in the cross-section data, but it slightly modifies the beam asymmetry description as well. This can be seen once a contribution of this resonance is switched off (see Figs. 4 and 7). This Δ state was found to be the most significant one in Ref. [7]. We see, therefore, that the results of our single-channel analysis are in a very good agreement with the results of a multichannel analysis.

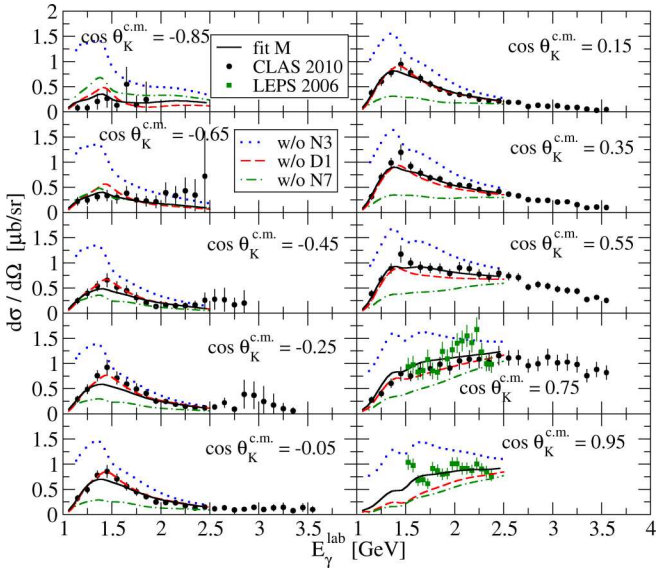

FIG. 4. The differential cross section as described by the full fit M (solid line) and by the same fit with the N3 (dotted line), D1 (dashed line), and N7 (dash-dotted line) resonances omitted. The data are from CLAS [10] and LEPS [9] experiments.

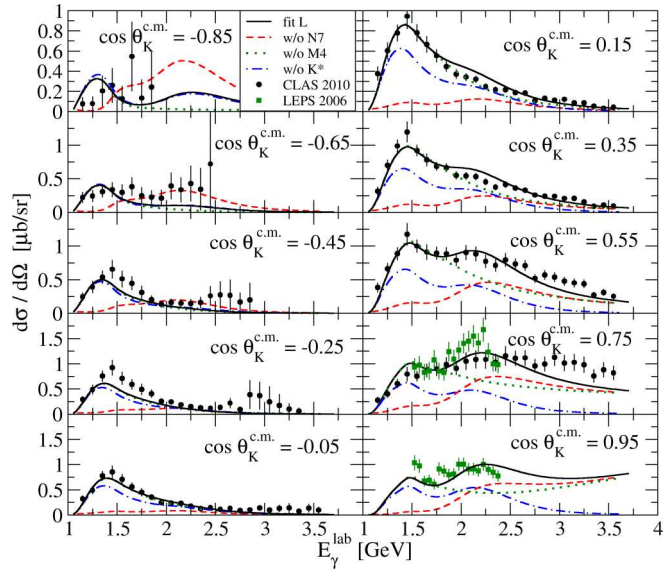

FIG. 5. The differential cross section as described by the full fit L (solid line) and by the same fit with the N7 (dashed line), M4 (dotted line), and K* (dash-dotted line) resonances omitted. The data are from CLAS [10] and LEPS [9] experiments.

In the fit L, the most important contributions stem from the K^*, N7, and M4 resonances. The kaon resonance K^* helps us to capture the experimental data predominantly at forward angles. When we omit this resonance, the cross section falls substantially (the smaller the kaon angle, the more notable the cross-section drop becomes). The N7 nucleon resonance on its own creates the first peak and near this peak clearly dominates the model. Once we omit this resonance, the fit produces a plateau instead of a peak. In contrast, the M4 nucleon resonance contributes in a substantial way to the description of the second peak, which is more tangible at small kaon angles. Besides these three resonances, no other resonance can produce such a substantial effect in the differential cross sections (see Fig. 5).

As the real reason why we began investigating the $K^+\Sigma^-$ channel was the new data on the photon beam asymmetries, this paper would not be complete without showing the results for this observable, which we could achieve with merely the MINUIT fitting procedure and also with the LASSO method. Figure 6 shows the photon beam asymmetry in dependence on the cosine of the kaon center-of-mass angle $\theta_K^{\text{c.m.}}$. The recent CLAS data [12] have a distinctive shape; they are large, positive, and almost uniform for central kaon angles, and gradually decrease at backward kaon angles. At forward kaon angles, the data are typically slightly larger than the data at backward angles (this is more notable for higher energies). The figure is complemented with the LEPS data [9] which are limited to forward going kaons and have slightly larger uncertainties than the CLAS data. Above the threshold, the two fits are hardly distinguishable as both of them produce beam asymmetry which is large and positive around central kaon angles and fall off gradually at backward angles and rather abruptly at forward angles, i.e., both models can capture the experimental data well. The sole difference can be seen at energies above 2 GeV around $\cos\theta_K^{\text{c.m.}} = -0.5$ where the fit M creates a peak while the fit L shows a dip. The latter behavior is more in concert with experimental data which indicate a plateau or a dip rather than a peak in this kinematical region.

In Figs. 7 and 8, we show description of the photon beam asymmetry data by both of our fits with some of the most contributing resonances omitted. The strongest contributions to the fit M, Fig. 7, stem from N7 and D1 resonances. The omission of the N7 nucleon resonance is tangible in all energy

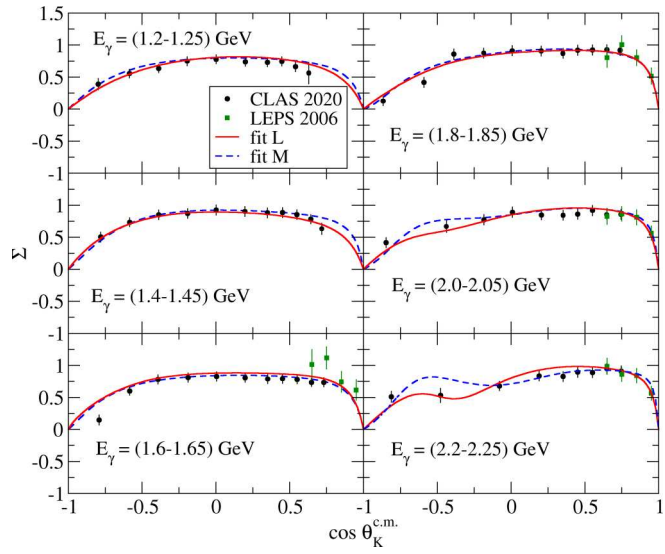

FIG. 6. The photon beam asymmetry for several energetic bins in dependence on the kaon center-of-mass angle $\theta_K^{\text{c.m.}}$ as calculated by the best fit achieved by the MINUIT alone and by the LASSO method and compared with the CLAS [12] data. Notation of the curves is the same as in Fig. 2.

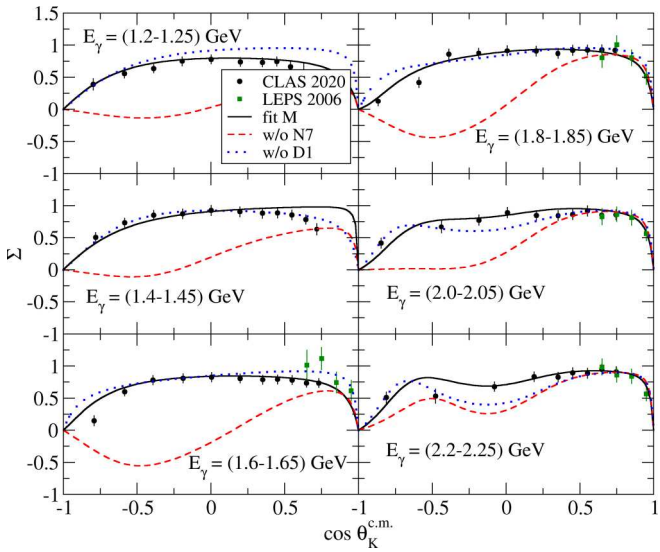

FIG. 7. The photon beam asymmetry described by the full fit M (solid line) and by the same fit with the N7 (dashed line) an D1 (dotted line) resonances omitted. The data are from CLAS [12] and LEPS [9] experiments.

bins as it leads to a significant drop of the photon beam asymmetry to zero and in some cases even below zero. When we leave out the D1 Δ resonance, on the other hand, we observe only minor modifications and in most energy bins we can still find agreement with data.

In the case of the fit L, we identified the K^*, N7, and M4 resonances to be the ones that contribute most to the photon beam asymmetry; see Fig. 8. When we omit the K^* kaon resonance, we cannot capture the magnitude of the data as the model outcomes lie below the data in almost all energy

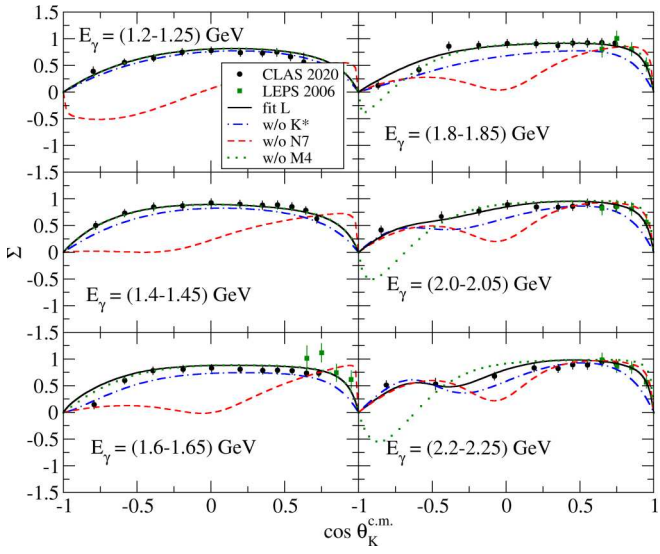

FIG. 8. The photon beam asymmetry described by the full fit L (solid line) and by the same fit with the K* (dash-dotted line), N7 (dashed line), and M4 (dotted line) resonances omitted. The data are from CLAS [12] and LEPS [9] experiments.

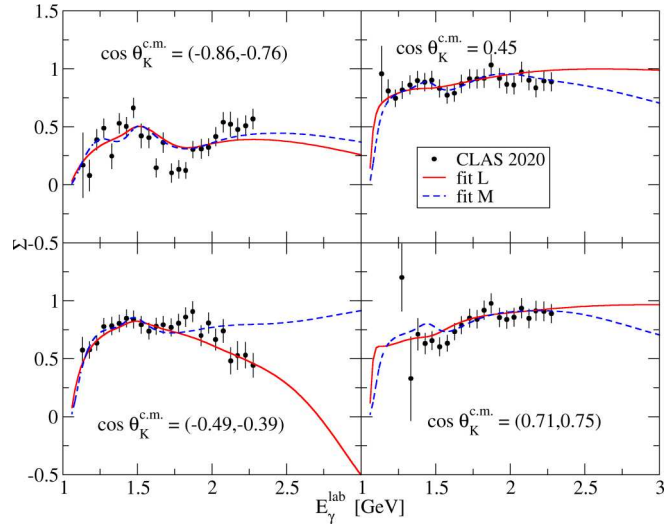

FIG. 9. The photon beam asymmetry in dependence on the photon laboratory energy $E_\gamma^{\rm lab}$ as calculated by the best fit achieved by MINUIT alone and by the LASSO method and compared with the CLAS [12] data. Notation of the curves is the same as in Fig. 2.

bins. Similar but more pronounced outcomes are produced when we turn off the N7 nucleon resonance: Near the energy threshold, the photon beam asymmetry even changes its sign. The omission of the M4 nucleon resonance leads to a minor corrections up to $E_\gamma^{\rm lab} = 1.8$ GeV, while beyond this energy we observe a dip at backward angles.

We show also the energy dependence of the photon beam asymmetry for four angular bins in Fig. 9. The two fits produce again very similar results, the only difference being the behavior at energies above 2 GeV, which is most notable in the $\cos\theta_K^{\rm c.m.} = (-0.49, -0.39)$ angular bin. Please note that the model results are calculated for the middle value in each angular bin whereas the data are scattered within the whole bin.

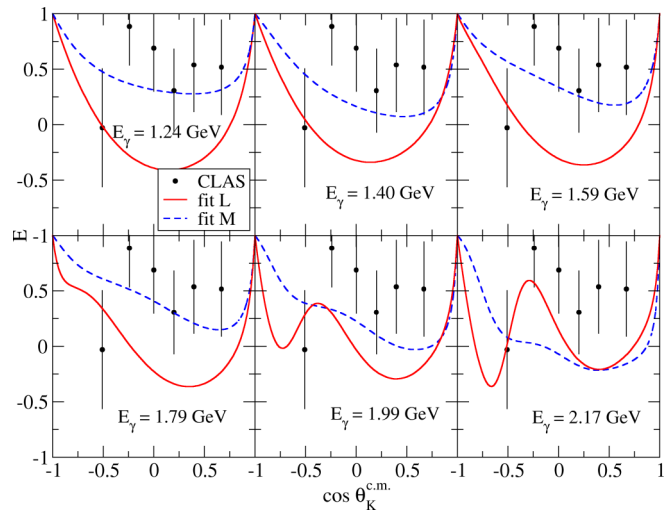

FIG. 10. The CLAS beam-target asymmetry data [13] in dependence on $\cos\theta_K^{\rm c.m.}$ compared with predictions of fit L (solid line) and fit M (dashed line).

In Fig. 10, we show the angular dependence of the beam-target asymmetry E. This observable was not included among the data and thus Fig. 10 illustrates merely the predictive power of our new fits. The data are positive in all energy bins shown and for energies below 1.8 GeV the fit M can capture their shape as it produces decreasing E. The fit L, on the other hand, can produce acceptable beam-target asymmetry only in the $E_\gamma^{\rm lab} = 2.17$ GeV bin; in other energy bins it underestimates all the data except for the data point at $\cos\theta_K^{\rm c.m.} = -0.5$.

V. CONCLUSION

Exploiting the isobar model, we performed an investigation of the $K^+\Sigma^-$ photoproduction off a neutron target in the resonance region. In order to reach an acceptable correspondence with experimental data, we used the tree-level Feynman diagrams with exchanges of particles in their ground as well as excited states. For high-spin nucleon states, we have used the consistent formalism where spurious lower-spin modes vanish in the amplitude.

The cornerstone of this analysis was, however, the upgrade to the fitting method. In our previous studies, we used only the plain χ^2 minimization. Such a technique, unfortunately, cannot prevent us from overfitting the data, i.e., introducing more parameters (and thus resonances) than are needed for the data description. In the third nucleon-resonance region, where the process under study in this paper occurs, there are a plenty of resonant states which overlap each other. It is therefore of crucial importance to limit the number of states which we consider in our analysis. A solution to this issue seems to be a method called regularization, where we introduce a penalty term which in effect restricts the number of nonzero parameters and in this way hinders us from overfitting the data.

With the help of both plain χ^2 minimization and the more advanced LASSO method, we could arrive at two models which both give us satisfactory agreement with data. We discussed the course of the fitting process, commented on the outcomes, and identified the resonant states which contribute most to the differential cross sections and photon beam asymmetries in various kinematic regions. We observe only slight differences in the data description by our models, the most notable ones being the description of differential cross sections at very forward angles where the fit L produces two broad peaks while the fit M is flat, and the photon beam asymmetries beyond 2 GeV at backward angles where the fit M produces a bump which is not supported by the data. In both fits, the $N(1720)3/2^+$ resonance was found to be important for correct description of data.

A natural step forward for us will be using the LASSO method to perform the multichannel analysis of the photoproduction as we deem this method to be of immense importance for selecting the optimal resonances contributing to the process.

ACKNOWLEDGMENTS

The authors would like to thank H. Haberzettl for drawing their attention to the LASSO method. This work was supported by the Czech Science Foundation GACR Grants No. 19-19640S (P.B., A.C., and D.S.) and No. 19-14048S (D.P.).

APPENDIX A: CONTRIBUTIONS TO THE INVARIANT AMPLITUDE

We consider the process

$$\gamma_V(k) + n(p) \to K^+(p_K) + \Sigma^-(p_\Sigma) \tag{A1}$$

with the corresponding four-momenta given in the parentheses; the four-momentum of the intermediate particle is denoted by q. In the following sections, we summarize the invariant amplitudes and include also the hadron form factors, $f_x(x)$, $x = s, t, u$, which turn out to be essential for the gauge invariance of the full amplitude.

1. Born s channel

The invariant amplitude of the proton exchange reads

$$\mathbb{M}_{Bs} = \bar{u}(p_\Sigma) V_S f_s(s) \frac{\not{p} + \not{k} + m_p}{s - m_p^2} V_\mu^{EM} \varepsilon^\mu(k) u(p), \tag{A2}$$

where

$$V_S = i g_{K\Sigma n} \gamma_5 \tag{A3}$$

and

$$V_\mu^{EM} = i \frac{\kappa_n}{2m_n} \sigma_{\mu\nu} k^\nu \tag{A4}$$

are strong and electromagnetic vertex functions, respectively, and $\sigma_{\mu\nu} = \frac{i}{2}[\gamma_\mu, \gamma_\nu]$.

When we recast the amplitude into the compact form

$$\mathbb{M} = \bar{u}(p_\Lambda) \sum_{j=1}^{6} \mathcal{A}_j(k^2, s, t, u) \mathcal{M}_j u(p), \tag{A5}$$

we can extract the scalar amplitudes

$$\mathcal{A}_1 = \frac{g_{K\Sigma n}}{s - m_n^2} \kappa_n f_s(s), \tag{A6a}$$

$$\mathcal{A}_4 = \frac{g_{K\Sigma n}}{s - m_n^2} \frac{\kappa_n}{m_n} f_s(s) = -2\mathcal{A}_6. \tag{A6b}$$

The operators $\mathcal{M}_j$ appearing in Eq. (A5) are defined in Eqs. (17) in Ref. [5].

2. Born t channel

The invariant amplitude of the kaon exchange in the t channel reads

$$\mathbb{M}_{Bt} = \bar{u}(p_\Sigma) V_S f_t(t) \frac{1}{t - m_K^2} V_\mu^{EM} \varepsilon^\mu(k) u(p) \tag{A7}$$

which can be cast into the compact form

$$\mathbb{M}_{Bt} = \bar{u}(p_\Sigma)\gamma_5 \left(\mathcal{A}_2 \mathcal{M}_2 + \mathcal{A}_3 \mathcal{M}_3 - g_{K\Sigma n} f_t(t) \frac{k \cdot \varepsilon}{k^2} \right) u(p). \tag{A8}$$

The last term in the brackets is a gauge-invariance breaking term. There are only two nonzero scalar amplitudes,

$$\mathcal{A}_2 = -\mathcal{A}_3 = 2\frac{g_{K\Sigma n}}{t - m_K^2} f_t(t). \tag{A9}$$

3. Born *u* channel

The electromagnetic $\gamma\Sigma\Sigma$ vertex factor has the form

$$V_\mu^{EM} = \gamma_\mu + i\frac{\kappa_\Sigma}{2m_\Sigma}\sigma_{\mu\nu}k^\nu \tag{A10}$$

and the strong vertex function is the same as in Eq. (A3). The hyperon exchange in the u channel reads

$$\mathbb{M}_{Bu} = \bar{u}(p_\Sigma)V_\mu^{EM}\frac{\not{p}_\Sigma - \not{k} + m_\Sigma}{u - m_\Sigma^2}V_S f_u(u)\varepsilon^\mu(k)u(p) \tag{A11}$$

and can be again recast into the compact form

$$\mathbb{M}_{Bu} = \bar{u}(p_\Sigma)\Big(\mathcal{A}_1\mathcal{M}_1 + \mathcal{A}_5\mathcal{M}_5 + \mathcal{A}_6\mathcal{M}_6 + g_{K\Sigma n}f_u(u)\frac{k\cdot\varepsilon}{k^2}\Big)u(p). \tag{A12}$$

Similarly to the t-channel exchange, the last term in the brackets is the gauge-invariance breaking term. In case we do not assume hadron form factors, i.e., $f_t = f_u = 1$, the gauge-invariance terms in t and u channels cancel each other and the resulting amplitude is gauge invariant.

The scalar amplitudes are

$$\mathcal{A}_1 = \frac{g_{K\Sigma n}}{u - m_\Sigma^2}f_u(u), \tag{A13a}$$

$$\mathcal{A}_5 = \frac{g_{K\Sigma n}}{u - m_\Sigma^2}\frac{\kappa_\Sigma}{m_\Lambda}f_u(u) = 2\mathcal{A}_6. \tag{A13b}$$

4. Contact current

In case we introduce the hadron form factors, the full amplitude contains gauge noninvariant terms. In order to get rid of them, we consider a contact term which acquires the form

$$\mathbb{M}_{cc} = \bar{u}(p_\Sigma)V_S e\Big[-(2p_K - k)^\mu\frac{f_t - 1}{t - m_K^2}f_u + (2p_\Sigma - k)^\mu\frac{f_u - 1}{u - m_\Sigma^2}f_t\Big]\varepsilon_\mu(k)u(p) \tag{A14}$$

and can be recast into the compact form

$$\mathbb{M}_{cc} = ieg_{K\Sigma n}u(p)\gamma_5\Big\{\Big[-2\mathcal{M}_2 + 2\mathcal{M}_3 + \big(t - m_K^2\big)\frac{k\cdot\varepsilon}{k^2}\Big]\frac{f_t - 1}{t - m_K^2}f_u + \Big[2\mathcal{M}_3 - \big(u - m_\Sigma^2\big)\frac{k\cdot\varepsilon}{k^2}\Big]\frac{f_u - 1}{u - m_\Sigma^2}f_t\Big\}u(p). \tag{A15}$$

Its contributions to the scalar amplitudes read

$$\mathcal{A}_2 = -2ieg_{K\Sigma n}\frac{f_t - 1}{t - m_K^2}f_u, \tag{A16a}$$

$$\mathcal{A}_3 = 2ieg_{K\Sigma n}\Big[\frac{f_t - 1}{t - m_K^2}f_u + \frac{f_u - 1}{u - m_\Sigma^2}f_t\Big]. \tag{A16b}$$

and the gauge-invariance breaking terms abolish those terms in the t (A7) and u (A11) channel exchanges.

APPENDIX B: ASPECTS OF THE FITTING PROCEDURE

1. Ridge and LASSO regularization

Typically, the regularization term is a norm of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ multiplied by a regularization parameter λ that determines the amount of the penalty on the parameter values,

$$P(\lambda) = \lambda\sum_{j=1}^{n}|\theta_j|^q. \tag{B1}$$

The two most common types of regularization correspond to $q = 1$ and $q = 2$. The $q = 1$ case, or L_1-norm regularization, is also known as LASSO [27], while the $q = 2$ case, or L_2-norm regularization, is also known as Ridge [28]. Minimizing the new function $\chi_T^2 = \chi^2 + P(\lambda)$ is equivalent to minimizing χ^2, subject to the constraint $\sum_{j=1}^{n}|\theta_j|^q \leqslant c$ (where $c > 0$). Due to the geometry of the constraint in parameter space (hyperoctahedron for LASSO, as opposed to hypersphere for Ridge) LASSO forces some of the parameters to zero, thus favoring a sparser model. This characteristic of LASSO makes it more suitable for feature selection and this is why we use it in our approach.

2. Parameter selection procedure

In machine learning, a model's accuracy is not assessed by the error calculated on the *training set*, which is used to fit its parameters, but on an independent *test set* that is intentionally held out of the original dataset. In the process called k-fold cross-validation, the original dataset is partitioned in k samples of equal size, each of which serves as the test set against which the model trained on the rest of the data (the remaining $k - 1$ samples) will be evaluated and the results of these runs are then averaged [20].

When more parameters are added to a model, the training set error will invariably decrease, but if the test set error increases it is a sign that the model is overfitted to the training set and fails to generalize to new data. So, using only a subset of a model's parameters may improve its prediction accuracy, as well as its interpretability.

Choosing the best subset can be done through either forward or backward stepwise selection [21]. Forward selection consists in sequentially adding the parameter that most improves the fit, while backward selection consists in sequentially removing the parameter that, when removed, least worsens the fit. In both cases, it is the test set (prediction) error that decides which is the best parameter subset. Forward selection has a wider applicability than backward selection, as it can be used also in cases where the number of parameters exceeds the number of data points, while backward selection cannot be used in such cases. Naturally, these selection procedures become highly impractical when a large number of

parameters are involved, resulting in extremely large number of subsets to be evaluated.

In the present work the parameter subset selection is done automatically by changing the regularization parameter λ accordingly; e.g., to implement forward selection we start from a large value of λ that leads to a very sparse model and gradually decrease it to zero, leading to the full unregularized model. The optimal value of λ and, hence the best parameter subset, is chosen based on information criteria which are commonly used in model selection. Furthermore, it has been shown that model selection by cross-validation is asymptotically equivalent to Akaike's information criterion [29].

3. Maximum likelihood estimation

In a measurement process, an observation d_i can be represented by a random variable D_i characterized by a probability density function $f(D_i|\boldsymbol{\theta})$, which depends on a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. The joint probability of N independent and identically distributed measurements is given by the product of the individual densities,

$$P(\boldsymbol{D}|\boldsymbol{\theta}) = P(D_1, \ldots, D_N|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(D_i|\boldsymbol{\theta}). \tag{B2}$$

For a sample of data $\boldsymbol{d} = (d_1, \ldots, d_N)$, this joint probability distribution (now only a function of $\boldsymbol{\theta}$) defines the likelihood function

$$L_{\boldsymbol{d}}(\boldsymbol{\theta}) = P(D = \boldsymbol{d}|\boldsymbol{\theta}) \tag{B3}$$

for that particular sample and expresses how probable the observed data $\boldsymbol{d}$ is, for given values of the parameters $\boldsymbol{\theta}$. In maximum likelihood estimation (MLE), we seek the parameter values $\hat{\boldsymbol{\theta}}$ that maximize the likelihood, $(L_{\boldsymbol{d}})_{\max} = L_{\boldsymbol{d}}(\hat{\boldsymbol{\theta}})$ and hence the probability of observing the specific sample of data. For computational reasons, it is the natural logarithm of the likelihood, $\ln L_{\boldsymbol{d}}(\hat{\boldsymbol{\theta}})$, that is commonly maximized. Furthermore, the fundamental assumption is made that the outcome d_i of each measurement follows a Gaussian distribution characterized by a given variance σ_i^2 (given by experiment) and a mean value μ_i. In fitting a theoretical model that depends on a set of parameters $\boldsymbol{c} = (c_1, \ldots, c_n)$ to the data, we adjust the parameters of the model so that its predictions $p_i(\boldsymbol{c})$ provide the mean values that maximize the likelihood of the particular sample. The likelihood function to be maximized thus becomes

$$L_{\boldsymbol{d}}(\boldsymbol{c}) = \prod_{i=1}^{N} \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left(-\frac{[d_i - p_i(c)]^2}{2\sigma_i^2}\right) \tag{B4}$$

and the log-likelihood

$$\ln L_{\boldsymbol{d}}(\boldsymbol{c}) = \text{const} - \sum_{i=1}^{N} \frac{(d_i - p_i(c))^2}{2\sigma_i^2} = \text{const} - \chi^2. \tag{B5}$$

Thus, under the above assumptions, maximizing the log-likelihood is equivalent to minimizing χ^2.

4. Derivations of information criteria

The following two sections present the basic steps in deriving the Akaike (AIC) and the Bayesian (BIC) information criteria and are based on [30] and [20,22] respectively.

a. Akaike information criterion

The Akaike information criterion [31] is an extension of the maximum likelihood principle based on the notion of relative entropy, or Kullback-Leibler (K-L) divergence, from information theory [32].

The K-L divergence is defined as

$$I(f, g) = \int f(x) \ln\left(\frac{f(x)}{g(x|\boldsymbol{\theta})}\right) dx \equiv E_x\left[\ln\left(\frac{f(x)}{g(x|\boldsymbol{\theta})}\right)\right] \tag{B6}$$

and expresses the degree of dissimilarity between the true (but unknown) probability distribution $f(x)$ that generates the data and an approximating distribution $g(x|\boldsymbol{\theta})$ that is specified by a set of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. It can also be regarded as the expected value of $\ln[f(x)/g(x|\boldsymbol{\theta})]$ with respect to the true distribution $f(x)$, for the whole population x. Using the likelihood function in place of $g(x|\boldsymbol{\theta})$, one observes that the value $\boldsymbol{\theta_0}$ that maximizes likelihood also minimizes the K-L divergence, meaning that $g(x|\boldsymbol{\theta_0})$ is as close as possible to the true $f(x)$.

However, since our data $\boldsymbol{d}$ is a sample taken from the statistical population x, we can only make an estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{d})$ of the true $\boldsymbol{\theta_0}$, based on this sample. Hence, the aim of seeking the minimum $I(f, g(\cdot|\boldsymbol{\theta_0}))$ is replaced by finding the (larger) minimum of the average $E_{\boldsymbol{d}}[I(f, g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{d})))]$ over repeated samples $\boldsymbol{d}$. It should be noted that with increasing sample size the likelihood maximizing $\boldsymbol{\theta}_{\text{MLE}} = \hat{\boldsymbol{\theta}}$ approaches the true $\boldsymbol{\theta_0}$.

For $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\boldsymbol{d})$, Eq. (B6) can be written as

$$\begin{aligned} I(f, g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{d}))) &= \int f(x) \ln f(x) dx - \int f(x) \ln g[x|\hat{\boldsymbol{\theta}}(\boldsymbol{d})] dx \\ &= \text{constant} - E_x[\ln g(x|\hat{\boldsymbol{\theta}}(\boldsymbol{d}))] \end{aligned} \tag{B7}$$

where the first term does not depend on $\boldsymbol{\theta}$, so the focus will be on $E_x[\ln g(x|\hat{\boldsymbol{\theta}}(\boldsymbol{d}))]$.

If we Taylor expand $\ln g(x|\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta_0}$ up to second order we get

$$\begin{aligned} \ln g(x|\hat{\boldsymbol{\theta}}) \approx \ln g(x|\boldsymbol{\theta_0}) &+ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^{\intercal} \nabla \ln g(x|\boldsymbol{\theta_0}) \\ &+ \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^{\intercal} \nabla^2 \ln g(x|\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \end{aligned} \tag{B8}$$

with the second term containing the gradient vector and the third term the Hessian matrix of second derivatives, both evaluated at $\boldsymbol{\theta_0}$. Taking the expected value $E_x[\cdots]$ of both sides of Eq. (B8), as defined in Eq. (B6), it is easily shown that the gradient term vanishes since $\boldsymbol{\theta_0}$ is the true minimum of $I(f, g(\cdot|\boldsymbol{\theta}))$. So, from Eq. (B8) we have

$$E_x[\ln g(x|\hat{\boldsymbol{\theta}})] \approx E_x[\ln g(x|\boldsymbol{\theta_0})] - \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^{\intercal} \mathrm{I}(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}), \tag{B9}$$

where $\mathrm{I}(\boldsymbol{\theta_0})$ is the $n \times n$ matrix with matrix elements

$$\mathrm{I}(\boldsymbol{\theta_0})_{ij} = E_x\left[-\frac{\partial^2 \ln g(x|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta_0}}\right] \tag{B10}$$

known as the Fisher information matrix. As already mentioned, the quantity we seek to minimize is the average $E_{\boldsymbol{d}}[I(f, g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{d})))]$ over different samples $\boldsymbol{d}$, which can be written as $E_{\hat{\boldsymbol{\theta}}}[I(f, g(\cdot|\hat{\boldsymbol{\theta}}))]$, i.e., as the average over the different $\hat{\boldsymbol{\theta}}$'s of each sample.

Therefore, the relevant quantity to maximize [because of the minus sign in Eq. (B7)] is

$$T \equiv E_{\hat{\theta}} E_x[\ln g(x|\hat{\boldsymbol{\theta}})]. \tag{B11}$$

Using Eq. (B9), T can be written

$$T \approx E_x[\ln g(x|\boldsymbol{\theta_0})] - \tfrac{1}{2} E_{\hat{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^\intercal \mathrm{I}(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})] \tag{B12}$$

since the first term does not depend on $\hat{\boldsymbol{\theta}}$. Taking into account the matrix identity involving the trace $z^\intercal A z = \mathrm{tr}(A z z^\intercal)$ and the fact that $I(\boldsymbol{\theta_0})$ is independent of $\hat{\boldsymbol{\theta}}$, Eq. (B12) becomes

$$T \approx E_x[\ln g(x|\boldsymbol{\theta_0})] - \tfrac{1}{2}\mathrm{tr}(\mathrm{I}(\boldsymbol{\theta_0}) E_{\hat{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^\intercal]). \tag{B13}$$

In the large sample limit, $E_{\hat{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^\intercal]$ is the covariance matrix of the maximum likelihood estimate, denoted by Σ, so T is written

$$T \approx E_x[\ln g(x|\boldsymbol{\theta_0})] - \tfrac{1}{2}\mathrm{tr}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma]. \tag{B14}$$

In the next steps, the first term, $E_x[\ln g(x|\boldsymbol{\theta_0})]$, is approximated in a similar way as $E_x[\ln g(x|\hat{\boldsymbol{\theta}})]$ in Eqs. (B8) and (B9), but with $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta_0}$ switched. So, $\ln g(x|\boldsymbol{\theta_0})$ is expanded around $\hat{\boldsymbol{\theta}}$

$$\begin{aligned}\ln g(x|\boldsymbol{\theta_0}) \approx{}& \ln g(x|\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal \nabla \ln g(x|\hat{\boldsymbol{\theta}}) \\ &+ \tfrac{1}{2}(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal \nabla^2 \ln g(x|\hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})\end{aligned} \tag{B15}$$

only, this time, the gradient vanishes before taking the expected value, as it is evaluated at $\hat{\boldsymbol{\theta}}$, which is the value that maximizes the log-likelihood $\ln g(x|\boldsymbol{\theta})$, leading to

$$\ln g(x|\boldsymbol{\theta_0}) \approx \ln g(x|\hat{\boldsymbol{\theta}}) + \tfrac{1}{2}(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal \nabla^2 \ln g(x|\hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}}). \tag{B16}$$

Taking the expected value with respect to x, Eq. (B16) becomes

$$E_x[\ln g(x|\boldsymbol{\theta_0})] \approx E_x[\ln g(x|\hat{\boldsymbol{\theta}})] - \tfrac{1}{2} E_x[(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal \hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})] \tag{B17}$$

with $\hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})$ the $n \times n$ matrix with matrix elements

$$\hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})_{kl} = -\left.\frac{\partial^2 \ln g(x|\boldsymbol{\theta})}{\partial\theta_k \partial\theta_l}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{B18}$$

and employing again the matrix property $z^\intercal A z = \mathrm{tr}(A z z^\intercal)$

$$\begin{aligned}&E_x[\ln g(x|\boldsymbol{\theta_0})] \\ &\quad\approx E_x[\ln g(x|\hat{\boldsymbol{\theta}})] - \tfrac{1}{2}\mathrm{tr}(E_x[\hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal]).\end{aligned} \tag{B19}$$

For large samples, the approximation $\mathrm{I}(\boldsymbol{\theta_0}) \approx \hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})$ is valid and the expected value inside the trace becomes

$$\begin{aligned}&E_x[\hat{\mathrm{I}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal] \\ &\quad\approx \mathrm{I}(\boldsymbol{\theta_0}) E_x[(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta_0} - \hat{\boldsymbol{\theta}})^\intercal] \\ &\quad= \mathrm{I}(\boldsymbol{\theta_0}) E_x[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^\intercal] = \mathrm{I}(\boldsymbol{\theta_0})\Sigma\end{aligned} \tag{B20}$$

leading to

$$E_x[\ln g(x|\boldsymbol{\theta_0})] \approx E_x[\ln g(x|\hat{\boldsymbol{\theta}}(x))] - \tfrac{1}{2}\mathrm{tr}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma]. \tag{B21}$$

Using this result in Eq. (B14) we get

$$T \approx E_x[\ln g(x|\hat{\boldsymbol{\theta}}(x))] - \mathrm{tr}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma]. \tag{B22}$$

Maximizing this expression can be used as a criterion for model selection when we have many and large samples to average over.

If, however, we have only one sample at our disposal, we can assume that an estimator of T, $\hat{T}$, will be of the same form as T, without the expected value and with an estimator for the trace,

$$\hat{T} \approx \ln g(x|\hat{\boldsymbol{\theta}}) - \hat{\mathrm{tr}}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma]. \tag{B23}$$

Instead of maximizing T, it is a convention to minimize the quantity

$$-2\hat{T} \approx -2\ln g(x|\hat{\boldsymbol{\theta}}) + 2\,\hat{\mathrm{tr}}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma]. \tag{B24}$$

It can be shown [30] that if $g(x|\boldsymbol{\theta})$ is a good approximation for the true distribution $f(x)$, then $\mathrm{I}(\boldsymbol{\theta_0}) = \Sigma^{-1}$ and $\hat{\mathrm{tr}}[\mathrm{I}(\boldsymbol{\theta_0})\Sigma] = \mathrm{tr}(\mathbb{I}_{n\times n}) = n$, where n is the number of parameters. Thus, the quantity to be minimized is

$$\mathrm{AIC} = -2\ln g(x|\hat{\boldsymbol{\theta}}) + 2n, \tag{B25}$$

where $\hat{\theta}$ is the maximum likelihood estimate and n the number of parameters of the model.

b. Bayesian information criterion

From Bayes's theorem, the posterior probability of model M_i from a set of r candidate models $\{M_1, \ldots, M_r\}$, given a set of observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is

$$P(M_i|\boldsymbol{x}) = \frac{g(\boldsymbol{x}|M_i)P(M_i)}{\sum_{j=1}^{r} g(\boldsymbol{x}|M_j)P(M_j)}, \tag{B26}$$

where $P(M_i)$ is the prior probability of model M_i. The likelihood of M_i, $g(\boldsymbol{x}|M_i)$ is given by

$$g(\boldsymbol{x}|M_i) = \int g(\boldsymbol{x}|\boldsymbol{\theta}, M_i)\pi(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta}, \tag{B27}$$

where $g(\boldsymbol{x}|\boldsymbol{\theta}, M_i)$ is the likelihood of the model parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ and $\pi(\boldsymbol{\theta}|M_i)$ the corresponding prior probabilities, and is referred to as *marginal likelihood* since the parameters are marginalized (integrated) out. It is also known as *model evidence* as it expresses the probability of the data $\boldsymbol{x}$ being generated from model M whose parameters are sampled from the specified prior distribution.

It is obvious from Eq. (B26) that the model with the largest posterior probability is the one with the largest $g(\boldsymbol{x}|M)P(M)$ and, assuming that all models have the same prior probabilities $P(M)$, it is the model with the largest $g(\boldsymbol{x}|M)$. The goal is, therefore, to maximize the quantity $g(\boldsymbol{x}|M)$ as it hints at the most probable model from the set of r candidate models $\{M_1, \ldots, M_r\}$.

The integral in Eq. (B27) can be evaluated using the Laplace approximation [20,33], whereby the logarithm of $g(\boldsymbol{x}|\boldsymbol{\theta}, M_i)$ in the integrand is Taylor expanded to second order around its maximum $\hat{\boldsymbol{\theta}}$, leading to a Gaussian integral. The expansion of the log-likelihood (M_i is omitted for simplicity) yields

$$\ln g(\boldsymbol{x}|\boldsymbol{\theta}) \approx \ln g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) - \tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\intercal \mathrm{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{B28}$$

since the gradient vanishes at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, with $\mathrm{H} = -\nabla\nabla \ln g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ the $n \times n$ Hessian matrix of the log-likelihood. After exponentiation, the likelihood function becomes

$$g(\boldsymbol{x}|\boldsymbol{\theta}) \approx g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}\mathrm{H}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\}. \tag{B29}$$

Assuming a uniform prior, $\pi(\boldsymbol{\theta})$ can be considered constant and $\pi(\boldsymbol{\theta}) \approx \pi(\hat{\boldsymbol{\theta}})$, so that the right-hand side of Eq. (B27) can be calculated as an n-dimensional Gaussian integral resulting in

$$g(\boldsymbol{x}|M) \approx g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})\pi(\hat{\boldsymbol{\theta}})\frac{(2\pi)^{n/2}}{|\mathrm{H}|^{1/2}}, \tag{B30}$$

where $|\mathrm{H}|$ is the determinant of H. Since the likelihood of a sample is the product of the likelihoods of each observation, the log-likelihood that appears in the elements of the matrix H is the sum of log-likelihoods of the observations and the elements of H become

$$\begin{aligned}\mathrm{H}_{kl} &= -\left.\frac{\partial^2 \ln g(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\left.\frac{\partial^2 \ln \prod_{j=1}^{N} g(x_j|\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right|_{\theta=\hat{\theta}} \\ &= -\left.\frac{\partial^2 \sum_{j=1}^{N} \ln g(x_j|\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right|_{\theta=\hat{\boldsymbol{\theta}}} \\ &= -N\left.\frac{\partial^2 E[\ln g(x_j|\boldsymbol{\theta})]}{\partial\theta_k\partial\theta_l}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = N\mathrm{I}(\hat{\boldsymbol{\theta}})_{kl}\end{aligned} \tag{B31}$$

where N is the sample size and $\mathrm{I}(\hat{\boldsymbol{\theta}})$ the Fisher information matrix. Thus, the determinant of H equals $|\mathrm{H}| = N^n|\mathrm{I}|$, where n is the dimensionality of the parameter space. Finally, if we take again the logarithm of Eq. (B30) we get

$$\begin{aligned}\ln g(\boldsymbol{x}|M) \approx \ln g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + \ln \pi(\hat{\boldsymbol{\theta}}) + \frac{n}{2}\ln 2\pi \\ - \frac{1}{2}n\ln N - \frac{1}{2}\ln|\mathrm{I}|\end{aligned} \tag{B32}$$

and, keeping only terms that vary at least linearly with sample size N, we end up with

$$\ln g(\boldsymbol{x}|M) \approx \ln g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) - \frac{n}{2}\ln N. \tag{B33}$$

Let us recall that $g(\boldsymbol{x}|M)$ is the quantity we sought to maximize and that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. As a matter of convention we define

$$\mathrm{BIC} = -2\ln g(\boldsymbol{x}|M) \approx \ln g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + n\ln N \tag{B34}$$

as the quantity that, when minimized, will yield the most probable model given the data.

[1] S. Capstick and W. Roberts, Prog. Part. Nucl. Phys. **45**, S241 (2000).
[2] U. Loring, B. C. Metsch, and H. R. Petry, Eur. Phys. J. A **10**, 395 (2001); **10**, 447 (2001).
[3] D. Rönchen, M. Döring, and U.-G. Meißner, Eur. Phys. J. A **54**, 110 (2018).
[4] A.V. Anisovich, V. Burkert, M. Hadzimehmedovic, D.G. Ireland, E. Klempt, V.A. Nikonov, R. Omerovic, H. Osmanovic, A.V. Sarantsev, J. Stahov, A. Svarc, and U. Thoma, Phys. Rev. Lett. **119**, 062004 (2017).
[5] D. Skoupil and P. Bydžovský, Phys. Rev. C **93**, 025204 (2016).
[6] R. A. Adelseck and B. Saghai, Phys. Rev. C **42**, 108 (1990).
[7] T. Mart and M. J. Kholili, J. Phys. G: Nucl. Part. Phys. **46**, 105112 (2019).
[8] D. G. Ireland, E. Pasyuk, and I. Strakovsky, Prog. Part. Nucl. Phys. **111**, 103752 (2020).
[9] H. Kohri *et al.* (LEPS Collaboration), Phys. Rev. Lett. **97**, 082003 (2006).
[10] S. A. Pereira *et al.* (CLAS Collaboration), Phys. Lett. B **688**, 289 (2010).
[11] K. Tsukada *et al.*, Phys. Rev. C **78**, 014001 (2008); **83**, 039904(E) (2011).
[12] N. Zachariou *et al.*, arXiv:2106.13957.
[13] N. Zachariou *et al.*, Phys. Lett. B **808**, 135662 (2020).
[14] D. Skoupil and P. Bydžovský, Phys. Rev. C **97**, 025202 (2018).
[15] J. C. David, C. Fayard, G. H. Lamot, and B. Saghai, Phys. Rev. C **53**, 2613 (1996).
[16] F. James and M. Roos, MINUIT, CERN Report No. D506, 1981 (unpublished).
[17] P. Zyla *et al.*, Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).
[18] J. Landay, M. Döring, C. Fernández-Ramírez, B. Hu, and R. Molina, Phys. Rev. C **95**, 015203 (2017); J. Landay, M. Mai, M. Döring, H. Haberzettl, and K. Nakayama, Phys. Rev. D **99**, 016001 (2019).
[19] J. de Swart, Rev. Mod. Phys. **35**, 916 (1963).
[20] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2006).
[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, Berlin, 2009).
[22] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling* (Springer, Berlin, 2008).
[23] H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).
[24] J. E. Cavanaugh, Stat. Probab. Lett. **33**, 201 (1997).
[25] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[26] T. Mart, Phys. Rev. C **62**, 038201 (2000).
[27] R. Tibshirani, J. R. Stat. Soc., Ser. B **58**, 267 (1996).
[28] A. E. Hoerl and R. Kennard, Technometrics **12**, 55 (1970).
[29] M. Stone, J. R. Stat. Soc., Ser. B **39**, 44 (1977).
[30] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer, Berlin, 2002).
[31] *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa, Springer Series in Statistics (Springer, Berlin, 1998).
[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, New York, 2006).
[33] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).