Pseudo-measurement simulations and bootstrap for the experimental cross-section covariances estimation with quality quantification

S. Varet^{1,a}, P. Dossantos-Uzarralde¹, N. Vayatis², and E. Bauge¹

¹ CEA-DAM-DIF, 91 Arpajon Cedex

² ENS Cachan

Abstract. The classical use of a generalized χ^2 -distance to determine the evaluated cross section uncertainty requires the values of the experimental cross sections covariance matrix. The usual propagation error method to estimate the covariances is hardly usable and the lack of data prevents from using the direct empirical estimator. Thus we present an alternative which exploits a regression model of the experimental cross section to generate pseudo-measurements and thereby allows an estimation of experimental covariances. The problem of assessing the quality of the estimate still remains. In our approach, we propose to determine the estimation quality through the means of the bootstrap method. We show on numerical examples that the bootstrap allows to have an order of magnitude of the estimation quality through a matrix norm. All the results are illustrated with a toy model (where all quantities are known) and also with real cross-section data measurements.

1 Introduction

Generalized χ^2 -distance for the evaluated cross section uncertainty determination needs to take into account the correlations and the uncertainties of the experimental measurements. As the covariances are not provided, an estimation of the experimental cross section covariance matrix Σ_F and its inverse is necessary. Since in many cases only one measure per energy is available, the standard empirical estimator is not available and the classical method of error propagation [1] is rarely usable due to the lack of information on the experimental parameters and due to the linearity assumption that is not fulfilled. Thus we propose in this article a statistical approach for the covariances evaluation based on the generation of gaussian pseudo-measurements around a regression model which is assimilated to the mean. The main advantage of this approach is that we do not impose any structure to the matrix. It remains to quantify the quality of the obtained matrix. We then propose to quantify the estimator quality via a bootstrap approach. In order to test our alternative and the bootstrap approach we use a toy model for which we know all the quantities we try to estimate. We present our alternative for the covariance evaluation in the next section and we give a way to evaluate the estimator quality in the third section. The entire approach is tested with the toy model and also illustrated in the case of the ^{25}Mn nucleus in the corresponding sections.

Notations Let $F = (F_1, \dots, F_N)^t$ the random vector of the measured cross section¹ at N distinct energies E_j , $j = 1, \dots, N$. We assume that F is gaussian. The covariance matrix of

^a e-mail: suzanne.varet@cea.fr

¹ $()^t$ denotes the transpose of a vector.

F is denoted by Σ_F . We assume that we can observe n realisations of the F vector written $F^{(1)}, \dots, F^{(n)}$ (we recall that, most of the time, for the cross-sections experiments $n = 1$).

Toy model description For the toy model construction we simulate n realisations of a gaussian vector, that can be assimilated to the experimental cross sections, with a fixed mean m_M and a fixed covariance matrix Σ_M . In the numerical examples of this article we have taken $N = 5$ and $n = 1$.

2 Experimental covariance matrix evaluation

Our alternative for the estimation of Σ_F first consists in summarizing the experimental measurements with a regression model. We have chosen a non-linear SVM (Support Vector Machine) model [2]. We assume that in the energy range of interest ($> 1\text{MeV}$) the response is a smooth and flat function of the energy. We denote by $h(E)$ the SVM regression model evaluated at the energy E . Assuming that the measurements are gaussian (section 1), we generate r vectors $S^{(1)}, \dots, S^{(r)}$ such as $S^{(i)}$ is a gaussian random vector of N independant random variables with expectation corresponding to the value of the regression model. For the noise variances, we refer to the square of the experimental uncertainties σ_i , $i = 1, \dots, N$. The final sample $(F^{(1)}, \dots, F^{(n)}, S^{(1)}, \dots, S^{(r)})$ is denoted by $X = (X^{(1)}, \dots, X^{(n+r)})$. With the pseudo-measurements, we compute the empirical estimator of Σ_F where the empirical mean is replaced

by the regression model value. Let, for each $i = 1, \dots, N$, $\widehat{V}_i = \frac{1}{n+r-1} \sum_{k=1}^{n+r} (X_i^{(k)} - h(E_i))^2$ be the

estimation of the diagonal terms. Finally with the diagonal of the empirical estimator taken as $(\sigma_1^2, \dots, \sigma_N^2)$, we obtain the estimator $\widehat{\Sigma}_F$. The $\widehat{\Sigma}_F$ term at the i^{th} line and j^{th} column C_{ij} , defined by the estimate of $\text{cov}(F_i, F_j)$, is given by $C_{ii} = \sigma_i^2$ and for $i \neq j$ by the following equation:

$$C_{ij} = \frac{\sigma_i \sigma_j}{n+r} \sum_{k=1}^{n+r} \frac{(X_i^{(k)} - h(E_i))(X_j^{(k)} - h(E_j))}{\sqrt{\widehat{V}_i} \sqrt{\widehat{V}_j}} \quad (1)$$

As for each $k = 1, \dots, r$ and $i, j = 1, \dots, N$, $i \neq j$, the two realisations $S_i^{(k)}$ and $S_j^{(k)}$ are assumed to be independent, if r is too large (in eq. (1)) compared with n , then $\widehat{\Sigma}_F$ is close to a diagonal matrix. Moreover, in many applications, the inverse covariance matrix is required. Thus, r must be large enough to ensure the invertibility of $\widehat{\Sigma}_F$. As a consequence, to determine the number r of generated vectors, we distinguish two cases: with $r = 0$, $\widehat{\Sigma}_F$ is invertible (case 1) or not (case 2). In the first case, the direct empirical estimator of the covariance matrix of $F = (F_1, \dots, F_N)$ is invertible. In the second case, we generate as many vectors as needed to ensure the invertibility of the covariance matrix.

Fig. 1. Σ_F (left) and its estimation with $r=1$ (right).

Application to the toy model Using one pseudo-measurement, we can see on the figure 1 that the covariance matrix estimation (eq. (1)) is close to the real one.

Application to the ^{55}Mn nucleus The ^{55}Mn (n,2n) cross section is measured at $N = 11$ energies in the range 12 MeV to 20 MeV with $n = 1$. The experimental data, extracted from [3], are plotted with their uncertainty on the figure 1(a) (green). We build a SVM regression model on the experimental data (red curve figure 1(a)). We generate $r = 1$ pseudo-measurements (blue circles figure 1(a)) per energie with a gaussian noise $\mathcal{N}(h(E_i), \sigma_i^2)$ for $i = 1, \dots, N$.

Fig. 1. ^{55}Mn (n,2n) measurements (right) and color map of experimental covariances estimation (left).

The $N = 400$ experimental data (with $n = 1$) (green circles) of the total ^{55}Mn cross section extracted from the EXFOR database³, the SVM regression model (red curve) and the $r = 1$ pseudo-measurement (blue circles) are plotted on the figure 2(a).

Fig. 2. Total ^{55}Mn cross section measurements and color map of the Σ_F estimation.

3 Quality assessment

Providing an estimation of the experimental covariance matrix, how far is the estimation from the real matrix? To answer this question we have chosen the following scalar criterion:

$$\|\widehat{\Sigma_F} - \Sigma_F\| = \|\widehat{\Sigma_F} - \Sigma_F\|_F \quad (2)$$

³ <http://www.oecd-neia.org/dbdata/x4/>

where $\|\cdot\|_F$ is the Frobenius norm (for any $q \times m$ matrix A with elements a_{ij} , $\|A\|_F = \left(\sum_{i=1}^q \sum_{j=1}^m |a_{ij}|^2 \right)^{\frac{1}{2}}$) which is easily computable and interpretable. Even if (2) can be estimated by $\|\widehat{\Sigma}_F - \Sigma_F\|$, we propose an overestimation $\mathbb{E}(\|\widehat{\Sigma}_F - \Sigma_F\|)$ knowing that $\|\widehat{\Sigma}_F - \Sigma_F\| \leq \mathbb{E}(\|\widehat{\Sigma}_F - \Sigma_F\|)$ due to the norm convexity. However, given that our interest is the estimation of Σ_F^{-1} , we introduce the Kullback-Leibler measure of $\widehat{\Sigma}_F$ which is defined by

$$\text{KL}(\widehat{\Sigma}_F, \Sigma_F) := \text{tr}(\widehat{\Sigma}_F^{-1} \cdot \Sigma_F) - \log(\det(\widehat{\Sigma}_F^{-1} \cdot \Sigma_F)) - N. \quad (3)$$

To estimate the norm of the difference between the two matrix and the Kullback-Leibler measure we propose to use a parametric bootstrap methodology. Thus, we distinguish again the two cases: with $r = 0$, whether $\widehat{\Sigma}_F$ is invertible or not. In the first case, we create B bootstrap samples $(F^{(1)}, \dots, F^{(n)})_1, \dots, (F^{(1)}, \dots, F^{(n)})_B$ following a gaussian law $\mathcal{N}(H, \widehat{\Sigma}_F)$ where $H = (h(E_1), \dots, h(E_N))^t$. In the second case, we create, in addition to $(F^{(1)}, \dots, F^{(n)})_1, \dots, (F^{(1)}, \dots, F^{(n)})_B$, B bootstrap samples $(S^{(1)}, \dots, S^{(r)})_1, \dots, (S^{(1)}, \dots, S^{(r)})_B$ following a gaussian law $\mathcal{N}(h(E_i), \sigma_i^2)$. The sampling $(S^{(1)}, \dots, S^{(r)})_1, \dots, (S^{(1)}, \dots, S^{(r)})_B$ with a diagonal matrix allows to reproduce the initial generation of $(S^{(1)}, \dots, S^{(r)})$. Then, in both cases we compute the B covariance empirical estimators $\widehat{\Sigma}_F^b$, $b = 1, \dots, B$ where the $\widehat{\Sigma}_F^b$ term at the i^{th} line and j^{th} column, C_{ij}^b is defined by $C_{ii}^b = \sigma_i^2$ and for $i \neq j$ by a similar expression as eq.(1). We estimate $\|\widehat{\Sigma}_F - \Sigma_F\|$ by the bootstrap estimation of $\mathbb{E}(\|\widehat{\Sigma}_F - \Sigma_F\|)$:

$$\frac{1}{B} \sum_{b=1}^B \|\widehat{\Sigma}_F^b - \widehat{\Sigma}_F\| \quad (4)$$

and the Kullback-Leibler measure by

$$\frac{1}{B} \sum_{b=1}^B \text{KL}(\widehat{\Sigma}_F^b, \widehat{\Sigma}_F) = \frac{1}{B} \sum_{b=1}^B \left(\text{tr}(\widehat{\Sigma}_F^b)^{-1} \cdot \widehat{\Sigma}_F) - \log(\det(\widehat{\Sigma}_F^b)^{-1} \cdot \widehat{\Sigma}_F) - N \right). \quad (5)$$

Application to the toy model The quality estimators (eq. (4) and (5)) can be tested now with the toy model (section 1). We take the case 2 with $N = 5$ and $n = 1$. We compute the svm regression model of the vector $F^{(1)}$ and we generate $r = 1$ pseudo-vector $S^{(1)}$ of $N = 5$ independant random variables. In order to average the results, we repeat this operation fifty times. That is we generate $50 * r$ vectors noted $S^{(1)1}, \dots, S^{(1)50}$. For each sample $(F^{(1)}, S^{(1)k})$, $k = 1, \dots, 50$, we compute the norm of the difference between the empirical estimator and the real matrix $\|\widehat{\Sigma}_F - \Sigma_F\|$ and the Kullback-Leibler measure of $\widehat{\Sigma}_F$ (eq. (3)). We compare the mean and the variance of these two values with the mean and the variance of their bootstrap estimations (with $B=30$) (eq. (4) and (5) respectively). The mean and the variance of the bootstrap estimations are obtained in repeating the bootstrap estimation about fifty times. We realise this computations for ten 1-sample of measurements.

Fig. 3 $\|\widehat{\Sigma}_F - \Sigma_F\|$ (blue) and of its bootstrap estimation

$$\frac{1}{B} \sum_{b=1}^B \|\widehat{\Sigma}_F^b - \widehat{\Sigma}_F\| \quad (\text{red}).$$

We can see (figure 3) that the bootstrap allows to have a fine estimation both in terms of mean and variance of the deviation from the real matrix.

Fig. 4 $KL(\widehat{\Sigma}_F, \Sigma_F)$ (blue) and of its bootstrap estimation

$$\frac{1}{B} \sum_{b=1}^B KL(\widehat{\Sigma}_F^b, \widehat{\Sigma}_F) \text{ (red)}$$

Application to the ^{55}Mn nucleus We have estimated the deviation of the $(n, 2n)$ cross-section covariance matrix (figure 1(b)) from the real matrix. The bootstrap gives $\|\widehat{\Sigma}_F - \Sigma_F\| \approx 0.019$ with a variance equals to $1.2 \cdot 10^{-5}$ and $KL(\widehat{\Sigma}_F, \Sigma_F) \approx 6.834$ with a variance equals to 2.351. As this case is, in dimension, close to the toy model, we can assume, with the estimation of $\|\widehat{\Sigma}_F - \Sigma_F\|$ that our estimation of the covariance matrix is good enough and, with the estimation of $KL(\widehat{\Sigma}_F, \Sigma_F)$, the inverse seems not too far from the real inverse.

For the total cross-section (figure 2(b)) the bootstrap gives $\|\widehat{\Sigma}_F - \Sigma_F\| \approx 2.385$ with a variance equals to $4.3 \cdot 10^{-5}$ and $KL(\widehat{\Sigma}_F, \Sigma_F) \approx 386.36$ with a variance equals to 6.19. We can conclude for this case that even if $\widehat{\Sigma}_F$ seems to be close to true matrix (due to a low norm value) the estimation $\widehat{\Sigma}_F^{-1}$ is probably far from the true inverse (a high K-L measure).

4 Conclusion

In this article a new protocol for experimental covariance matrix estimation has been presented. We focus on the situation where the lack of measurements prevents from using the direct empirical estimator and the informations on the experimental protocol are not available. This approach is based on the construction of a regression model to replace the empirical mean and the generation of pseudo-measurements as a noise around this regression model. The main advantage of this approach is that we can quantify the estimator quality thanks to a bootstrap sampling. Moreover, as it is a statistical approach, only measurements are needed and the only assumption is the gaussianity of the measurements which seems to be reasonable for experimental data. The numerical validation of our approach has been done by the construction of a gaussian toy model. Indeed the numerical tests have pointed out that we can easily supply an estimation of the covariance matrix. Moreover the toy model has illustrated that the bootstrap can be a good way to quantify the matrix estimation quality.

References

1. M. Ionescu-Bujor and D. G. Cacuci, *A comparative review of sensitivity and uncertainty analysis of large-scale systems – I: Deterministic methods*, Nuclear Science and Engineering **147** (2004), 189–203
2. A.J. Smola and B. Schölkopf, *A tutorial on support vector regression*, Statistics and Computing **14** (2004), 199–222.
3. A. Milocco, A. Trkov, and R. Capote Noy *Nuclear data evaluation of ^{55}Mn by the EMPIRE code with emphasis on the capture cross-section*, Nuclear Engineering and Design **241** (2011), 1071–1077.