

Using training samples for eliminating nuisance parameters

L'uso dei campioni di prova per l'eliminazione dei parametri di disturbo

Sara Antignani and Fulvio Spezzaferri¹

Dip. di Statistica, Probabilità e Statistiche Applicate

Università di Roma "La Sapienza"

sara.antignani@uniroma1.it, fulvio.spezzaferri@uniroma1.it

Riassunto: L'eliminazione dei parametri di disturbo dalla funzione di verosimiglianza è un problema ben noto nella statistica inferenziale. In questo lavoro intendiamo presentare un nuovo approccio basato sull'utilizzo dei campioni di prova. Esamineremo, illustrandoli con esempi, differenti modi di impiego dei campioni di prova suggeriti dalla vasta letteratura relativa ai fattori di Bayes ed effettueremo qualche confronto tra la nostra proposta e i metodi standard.

Keywords: Likelihood function, nuisance parameters, training sample, fractional likelihood, geometric training likelihood, imaginary training sample.

1. Introduction

Let $f(\mathbf{x}; \theta)$ denote the joint probability density function of the sample $\mathbf{X} = (X_1, \dots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the likelihood function for θ is defined by $l(\theta; \mathbf{x}) \propto f(\mathbf{x}; \theta)$, where the proportional factor does not depend on θ . Often θ may be written as $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is the nuisance parameter. Of course they can be vectors again. Making inference on ψ using the likelihood is not possible, because in general conclusions will depend on the value of λ , which is unknown. For this reason the relevant question is how to eliminate the nuisance parameters from the argument of the likelihood function. Many methods have been proposed for eliminating λ from the likelihood in order to obtain certain functions of the data and ψ that may not be genuine likelihoods (since they are not derived from density functions), but that sometimes have properties similar to those of a likelihood. For this reason they are called *pseudo-likelihood* functions. Some commonly used elimination approaches are, for instance, maximization, conditioning or marginalization (see Severini (2000)). The Bayesian approach to the problem of eliminating nuisance parameters consists in integrating out λ from the likelihood using an appropriate weighting function, that is

$$\ell_I(\psi) = \int_{\Lambda} \ell(\psi, \lambda; \mathbf{x}) \pi(\lambda|\psi) d\lambda. \quad (1)$$

The introduction of a proper conditional prior density for λ allows to take into account the uncertainty due to the nuisance parameter and implies that the ℓ_I is a genuine likelihood. However this is a weak point of the method because the necessary elicitation of $\pi(\lambda|\psi)$ is the result of a subjective choice (Berger *et al.* (1999)). On the other hand, the use of an improper conditional prior does not lead to an "exact" likelihood.

¹Indirizzo per corrispondenza: Dipartimento di Statistica, Università "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy

2. A new approach to the elimination problem: using training samples

The aim of this work is to study how to use training samples for eliminating nuisance parameters from the likelihood function. Considerations that have led to our proposal are the following. Suppose that we want to eliminate the parameter λ from the model $f(\mathbf{x}; \psi, \lambda)$. Denote with (ψ_0, λ_0) the true parameter values, those that have generated the data. If we knew these values, a likelihood for the parameter of interest could be simply obtained substituting λ with λ_0 , that is $\ell(\psi) = f(\mathbf{x}; \psi, \lambda = \lambda_0)$. Since we do not know the true value of the nuisance parameter, we have to obtain information on λ in some way. For this purpose we split the sample into two parts $\mathbf{x} = (\mathbf{x}(l), \mathbf{x}(n-l))$ and, assuming to know the value of ψ_0 , we use the training sample of dimension l , $\mathbf{x}(l)$, in order to compute a posterior distribution for λ from which we get the necessary information. Let $\pi(\lambda|\psi_0)$ be a non informative, typically improper, conditional prior distribution of λ . We assume that we can choose $\mathbf{x}(l)$ to assure the posterior distribution of λ , $\pi(\lambda|\mathbf{x}(l), \psi_0) \propto \ell(\lambda, \psi_0; \mathbf{x}(l))\pi(\lambda|\psi_0)$, to be proper. The last step is to construct a likelihood based on the remaining data $\mathbf{x}(n-l)$, and to integrate out the nuisance parameter from this likelihood weighted with the posterior distribution of λ , that is

$$\ell_{\mathbf{x}(l)}^*(\psi, \psi_0) = \int_{\Lambda} \ell(\psi, \lambda; \mathbf{x}(n-l))\pi(\lambda|\mathbf{x}(l), \psi_0)d\lambda. \quad (2)$$

The $\ell_{\mathbf{x}(l)}^*$, that depends on the specific training sample $\mathbf{x}(l)$, will be called *trained likelihood*. It is worthy pointing out that $\ell_{\mathbf{x}(l)}^*$ is a genuine likelihood (since it derives from the density of $\mathbf{x}(n-l)$ given $\mathbf{x}(l)$) even if we have used an improper conditional prior distribution for λ avoiding the introduction of any type of prior information on the nuisance parameter.

Evidently this procedure is affected by two different problems: firstly the arbitrariness of the choice of a specific training sample; secondly, the use of the unknown value ψ_0 . To solve the former, we can refer to different techniques suggested by the wide literature concerning Bayes factor. We could, for instance, average the $\ell_{\mathbf{x}(l)}^*$ in (2) over the set of all possible training samples, following the idea on which is based the intrinsic Bayes factor (Berger and Pericchi (1996)). We instead focus our attention on the approach founded on the consideration that, for i.i.d. samples, when l and n are large enough, each $\mathbf{x}(l)$ gives essentially the same information and $f(\mathbf{x}; \psi, \lambda)^{1/n} \approx f(\mathbf{x}(l); \psi, \lambda)^{1/l}$ (O'Hagan (1995)). With this approximation, the likelihood (2) assumes the following form:

$$\ell_F^*(\psi, \psi_0) = \int_{\Lambda} \ell^{1-\frac{1}{n}}(\psi, \lambda; \mathbf{x})\ell^{\frac{1}{n}}(\psi_0, \lambda; \mathbf{x})\pi(\lambda|\psi_0)d\lambda, \quad (3)$$

where $\ell^{\frac{1}{n}}(\psi_0, \lambda; \mathbf{x})\pi(\lambda|\psi_0)$ is proportional to the posterior distribution of λ and l is large enough to make proper this posterior distribution. The index F indicates that we have used the so called *fractional likelihood*. Note that ℓ_F^* depends also on l , but for the sake of simplicity we have dropped it from the notation. We will consider this aspect later. Once we have avoided this drawback, we can solve the second problem estimating ψ_0 with $\hat{\psi}_0$, the MLE obtained using ℓ_F^* , that is $\hat{\psi}_0 \equiv \arg \sup_{\psi} \ell_F^*(\psi, \hat{\psi}_0)$. Then $\ell_F^*(\psi, \hat{\psi}_0)$, that is now a function of the only parameter of interest, can be used as likelihood for ψ . It can be shown that, when the conditional prior distribution $\pi(\lambda|\psi)$ in (1) does not depend on ψ , as it is typical with noninformative priors, then $\hat{\psi}_0$ coincides with the MLE of the integrated likelihood ℓ_I .

After these arrangements the resulting ℓ_F^* is not a genuine likelihood any more, but it is

an approximation of it.

Example 1. *Variance of a normal distribution.*(Severini (2000)) Suppose that $X_1, \dots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2 and suppose that σ^2 is the parameter of interest. Imposing a uniform prior conditional density for μ , easy calculations yield to

$$\ell_F^*(\sigma^2, \sigma_0^2) = \sigma^{-(n-l)} \left(\frac{n-l}{\sigma^2} + \frac{l}{\sigma_0^2} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{(n-l)}{2\sigma^2} s^2 \right\}$$

where $\hat{\sigma}_0^2$, the estimate of σ_0 , is obtained as the solution of $\frac{\partial}{\partial \sigma^2} \ell^*(\sigma^2, \sigma_0^2)|_{\sigma^2=\sigma_0^2} = 0$. It results that $\hat{\sigma}_0^2 = \frac{n}{n-1} s^2$, where s^2 is the sample variance. We only observe here that the fractional trained likelihood has a maximum at the unbiased estimator of σ^2 . $\square$

When the data are independent but not identically distributed, De Santis and Spezzaferri (2001) have shown that the fractional likelihood can fail to convert improper priors into proper posteriors leading to inconsistent Bayes factors. Following their suggestions, to overcome this problem, in place of the fractional likelihood we use the *geometric training likelihood*, defined as the geometric mean of $\ell(\theta; \mathbf{x}(l))$ as $\mathbf{x}(l)$ varies over the set of all training samples of size l . The so obtained trained likelihood is denoted by $\ell_{GT}^*(\psi, \psi_0)$. As in the previous case, $\hat{\psi}_0 \equiv \arg \sup_{\psi} \ell_{GT}^*(\psi, \hat{\psi}_0)$. It can be shown that $\hat{\psi}_0 \equiv \arg \sup_{\psi} \frac{1}{L} \sum_{\mathbf{x}(l)} \ell_{\mathbf{x}(l)}^*(\psi, \hat{\psi}_0)$, as $\mathbf{x}(l)$ varies over the set of l -dimensional training samples composed by L elements.

Example 2. *Normal distributions with common standard deviation.* This example is better known as the Neyman-Scott problem. Let (X_j, Y_j) , $j = 1, \dots, n$, denote pairs of random variables with mean μ_j and standard deviation σ , which is the parameter of interest. In this case the symmetric structure of the data is such that the geometric training likelihood coincides with the standard fractional likelihood. Hence in our approach the likelihood for σ^2 is:

$$\ell_{GT}^*(\sigma^2, \hat{\sigma}_0^2) \equiv \ell_F^*(\sigma^2, \hat{\sigma}_0^2) = \sigma^{-(n-l)} \left(\frac{2n-l}{l} \hat{\sigma}_0^2 + \sigma^2 \right)^{-\frac{n}{2}} \exp \left\{ -\frac{2n-l}{4n\sigma^2} \text{RSS} \right\},$$

where RSS is the residual sum of square and $\hat{\sigma}_0^2 = \text{RSS}/n$ is an unbiased estimate of σ^2 . $\square$

Another way to handle the problem of arbitrariness of the training sample $\mathbf{x}(l)$ is to resort to the idea of *imaginary training sample* (Perez and Berger (2002), Iwaki (1997)). It is denoted by $\mathbf{x}^I(l)$ and it is a random variable, independent from the sample elements, that has the same distribution of a generic training sample $\mathbf{x}(l)$. As usual we want to use $\mathbf{x}^I(l)$ to construct the posterior conditional distribution of λ . The resulting $\pi(\lambda|\mathbf{x}^I(l), \psi_0)$ depends on $\mathbf{x}^I(l)$, the unobservable training sample. Following the approach proposed by Iwaki (1997), in order to obtain a posterior distribution $\pi^I(\lambda|\psi_0)$ not dependent on $\mathbf{x}^I(l)$, we can average the posterior conditional distribution using the predictive distribution $f(\mathbf{x}^I(l)|\mathbf{x}, \psi_0)$ of $\mathbf{x}^I(l)$, that is

$$\pi^I(\lambda|\psi_0) = E_{f(\mathbf{x}^I(l)|\mathbf{x}, \psi_0)} \pi(\lambda|\mathbf{x}^I(l), \psi_0).$$

Finally we define the *imaginary trained likelihood* to be:

$$\ell_I^*(\psi, \psi_0) = \int_{\Lambda} f(\mathbf{x}|\psi, \lambda) \pi^I(\lambda|\psi_0) d\lambda, \quad (4)$$

where the estimate of ψ_0 is $\hat{\psi}_0 : \hat{\psi}_0 \equiv \arg \sup_{\psi} \ell_I^*(\psi, \hat{\psi}_0)$.

Example 1. (Continued). The *imaginary trained likelihood* of σ^2 , when $l = 1$, is

$$\ell_I^*(\sigma^2, \hat{\sigma}^2) = \sigma^{-n} \left(\frac{1}{\sigma^2} + \frac{1}{\hat{\sigma}_0^2(2n+1)} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{ns^2}{2\sigma^2} \right\},$$

where $\hat{\sigma}_0^2 = \frac{2n(n+1)}{2n^2-1} s^2$. $\square$

We want to discuss here the choice of l , the dimension of the training sample. This choice is restricted to those values that satisfy $l \geq l_{min}$, where l_{min} is the minimum $l \in \{1, n-1\}$ which let us obtain a proper posterior distribution for the nuisance parameter. An empirical choice, the most common for the alternative Bayes factors, is $l = l_{min}$. Otherwise, the choice can be made on the basis of some optimality criterion. For instance, l can be fixed to the value that maximizes $I^*(\hat{\psi}_0, l)$, the observed Fisher information of the obtained ℓ^* . Of course other possibilities can be considered (e.g. choose l in such a way that ℓ^* has the best frequentist coverage probabilities).

Example 3. Gamma distribution. Consider an i.i.d. sample $(x_1, \dots, x_n)$ from a gamma random variable with index ψ and rate parameter λ . Choosing for λ the reference prior, we obtain:

$$\ell_F^*(\psi, \hat{\psi}_0) = \frac{\Gamma((n-l)\psi + l\hat{\psi}_0)}{\Gamma(\psi)^{n-l}} \left(\frac{(\prod_i x_i)^{1/n}}{\sum_i x_i} \right)^{(n-l)\psi}$$

When $n = 20$, $\psi = 1.5$, $\lambda = 1$, a simulation study shows that the maximum of $I^*(\hat{\psi}_0, l)$ for $l \in \{1, n-1\}$ is attained at $\hat{l} = 6 \neq l_{min} = 1$. We observe that in this case $I^*(\hat{\psi}_0, \hat{l}) = 6.5$ is greater than $I^C(\hat{\psi}_C) = 5.2$, the observed Fisher information of the conditional likelihood $\ell_C(\psi)$, where $\hat{\psi}_C$ is the MLE of ψ obtained using ℓ_C . $\square$

In conclusion, we want to remark that in the context of integration methods, the trained likelihood has the advisable property to be an approximation of a genuine likelihood in spite of the introduction of an improper prior conditional distribution for the nuisance parameters. Compared with the conditioning or marginalizing strategy our method let us recover information on the parameter of interest that otherwise could be lost.

References

- Berger J.O., Liseo B. and Wolpert R.L. (1999) Integrated likelihood methods for eliminating nuisance parameters., *Stat. Science*, 14, 1–28.
- Berger J.O. and Pericchi L.R. (1996) The intrinsic bayes factor for model selection and prediction., *J. Am. Stat. Ass.*, 91, 109–122.
- De Santis F. and Spezzaferri F. (2001) Consistent fractional bayes factor for nested normal linear models., *J. Stat. Plan. Inf.*, 97, 305–321.
- Iwaki K. (1997) Posterior expected marginal likelihood for testing hypotheses., *J. Econ. Asia Univ.*, 21, 105–134.
- O'Hagan A. (1995) Fractional bayes factors for model comparison (with discussion), *J. Royal. Stat. Soc., Ser. B*, 57, 99–138.
- Perez J.M. and Berger J.O. (2002) Expected-posterior prior distributions for model selection., *Biometrika*, 89, 491–511.
- Severini T.A. (2000) *Likelihood methods in statistics*, Oxford University Press.