Tatra Mt. Math. Publ. **17** (1999), 45–53

TATRA MOUNTAINS
Mathematical Publications

A PARAMETRIC MODEL WITH SEPARATE ESTIMATION OF PARAMETERS

Martin Janžura

ABSTRACT. A curved exponential parametric family of probability distributions is introduced. Within this family a block-wise estimation of a large parametric vector is possible.

1. Introduction

A standard procedure of parameter estimation consists in maximizing a function

$$\widehat{F(\theta)} = \int \mathcal{L}^\theta \, \mathrm{d}\widehat{P} \,,$$

where $\widehat{P}$ is the empirical distribution obtained from data, $\mathcal{L}^\theta$ is a suitable "contrast" function, and $\theta \in \Theta \subset \mathbb{R}^K$ is the (vector) parameter. Here $\mathcal{L}^\theta$ is the log-likelihood function in the case of maximum likelihood estimation, or it may be a pseudo-log-likelihood function imitating the properties of log-likelihood. For M-estimates the function can be chosen rather generally. In any case, we have

$$\mathcal{L}^\theta \colon \mathcal{X} \to \mathbb{R}$$

for every $\theta \in \Theta$, where $\mathcal{X}$ is the state space.

When considering the multivariate models, we usually have

$$\mathcal{X} = \bigotimes_{s \in S} X_s \,,$$

where S is a finite set of indices (sites, elements) and for each $s \in S$, X_s is the state space of the sth variable. For simplicity we shall assume that every X_s, $s \in S$, is finite. Within this paper we have in mind the case of a very large set S, i.e.,

$$|S| \gg 1 \,,$$

2000 Mathematics Subject Classification: 62H12.
Keywords: exponential distribution, separate estimation, proper base.

Supported by GA ČR Grant No. 202/0731.

where $|\cdot|$ denotes the cardinality. Consequently, then, in order to deal with non-trivial parametric models we also have a large number of parameters

$$|K| \gg 1\,.$$

Such models frequently occur, for example, with large contingency tables, probabilistic expert systems, or in image processing problems (cf. [1,2]). Then the above mentioned procedure of parameter estimation encounters some serious problems which make it extremely complicated (if not impossible). First, there is a computational problem in expressing the function $\mathcal{L}^\theta$ which usually involves a normalizing constant obtained by summing over the state space $\mathcal{X} = \bigotimes_{s\in S} X_s$. Any numerical maximization procedure may also fail for the multivariate function $\widehat{F(\theta)}$. Secondly, and this can be even more limiting, there is a purely statistical problem, namely any sample size n available in practice (we have $\widehat{P} = \widehat{P}_n$, in fact) will inevitably be too small to compare with the dimensions $|S|$ or $|K|$ respectively. Then the empirical distribution $\widehat{P}_n$ may be by no means a good estimate of the true distribution. Problems concerning overparametrization can also follow.

Therefore, it would be more convenient if it were possible to separate the vector parameter $\theta \in \mathbb{R}^K$ into blocks $\theta_i \in \mathbb{R}^{K_i}$ for $i = 1,\dots,M$, where $\{K_i\}_{i=1,\dots,M}$ is a partition of the set K, and each K_i for $i = 1,\dots,M$ is rather "small", and to estimate the blocks of parameters separately, by maximizing functions

$$\widehat{F_i(\theta_i)} = \int \mathcal{L}_i^{\theta_i}\,\mathrm{d}\widehat{P}_{A_i} \qquad \text{for} \quad \theta_i \in \mathbb{R}^{K_i}\,,$$

where we now have

$$\mathcal{L}_i^{\theta_i} : \mathcal{X}_{A_i} \to \mathbb{R}\,, \qquad \widehat{P}_{A_i} = \widehat{P}/\mathcal{X}_{A_i}$$

and

$$\mathcal{X}_{A_i} = \bigotimes_{s\in A_i} X_s$$

with $A_i \subset S$ again "small" sufficiently for every $i = 1,\dots,M$. Thus, we would obtain a collection of well defined and solvable problems. Let us recall directly that the above scheme works under the additional assumption of the independence between the variables corresponding to the coordinates $s \in S$, i.e., under the product of the parametric distributions

$$P_S^\theta = \bigotimes_{s\in S} P_s^{\theta_s}\,.$$

Here we have $M = |S|$, $A_s = \{s\}$ and, for example,

$$\mathcal{L}_s^{\theta_s} = \log P_s^{\theta_s}$$

for every $s \in S$.

In the present paper we shall introduce a much more general class of parametric models for which the separate estimation of parameters can be performed. The approach extends the idea of maximum pseudo-likelihood estimation (cf., e.g., [3]).

2. Model

Let us denote by $\mathcal{P}$ the set of all probability distributions on the state space $\mathcal{X}_S = \bigotimes_{s \in S} X_s$, where again S is a finite set of indices, and X_s is a finite state space for every $s \in S$.

For $P, Q \in \mathcal{P}$ we denote by

$$H(P) = \sum_{x_S \in \mathcal{X}_S} \left[-\log P(x_S)\right] P(x_S) \geq 0$$

the entropy, and by

$$I(P|Q) = \sum_{x_S \in \mathcal{X}_S} \left[\log \frac{P(x_S)}{Q(x_S)}\right] P(x_S) \geq 0$$

the I-divergence.

Further, we denote by

$$\mathcal{F}_S = \{f \colon \mathcal{X}_S \to \mathbb{R}\}$$

the set of all real-valued functions on $\mathcal{X}$, and by $\mathcal{F}_V$ for $V \subset S$ we mean the set of functions which actually depend only on the coordinates for V, i.e.,

$$\mathcal{F}_V = \{f = f^* \circ \mathcal{P}r_V \colon \quad f^* \colon \mathcal{X}_V \to \mathbb{R}\},$$

where $\mathcal{P}r_V \colon \mathcal{X} \to \mathcal{X}_V$ is the projection.

Now, suppose a fixed collection

$$\boldsymbol{f} = (f_1, \ldots, f_K)$$

is given with $f_j \in \mathcal{F}_S$ for every $j = 1, \ldots, K$. We shall understand $\boldsymbol{f}$ as a column vector. Then, we may write $\alpha^{\mathrm{T}} \boldsymbol{f}$ for $\sum_{i=1} \alpha_i f_i$.

Let us consider the exponential family

$$\{P_S^\alpha\}_{\alpha \in \mathbb{R}^K} \subset \mathcal{P}$$

of probability distributions, where

$$P_S^\alpha(x_S) = \exp\left\{\alpha^{\mathrm{T}} \boldsymbol{f}(x_S)\right\} / c(\alpha)$$

for every $\alpha \in \mathbb{R}^K$, $x_S \in \mathcal{X}_S$, and

$$c(\alpha) = \sum_{x_S \in \mathcal{X}_S} \exp\left\{\alpha^{\mathrm{T}} \boldsymbol{f}(x_S)\right\}$$

is a normalizing constant.

Further, let us introduce the notation

$$\mathcal{P}(\boldsymbol{f}, 0) = \left\{P \in \mathcal{P}\colon \int f_i \, \mathrm{d}P = 0 \text{ for } i = 1, \ldots, K\right\}.$$

Lemma 2.1. *The function* $\log c(\alpha)$ *attains its minimum at some* $\alpha^0 \in \mathbb{R}^K$ *iff* $P_S^{\alpha^0} \in \mathcal{P}(\boldsymbol{f}, 0)$.

P r o o f. This is obvious by the convex properties of the function $\log c(\alpha)$. □

Proposition 2.2. *The following statements are equivalent*

i) $\log c(\alpha)$ *attains its unique minimum at some* $\alpha_0 \in \mathbb{R}^K$;

ii) $\lim\limits_{\lambda \to \infty} \log c(\lambda\alpha) = +\infty$ *for every* $\alpha \in B_1^K = \left\{\alpha \in \mathbb{R}^K\colon \|\alpha\| = 1\right\}$;

iii) $\left(\alpha^{\mathrm{T}} \boldsymbol{f}\right)_{\max} = \max\limits_{x_S \in \mathcal{X}_S} \alpha^{\mathrm{T}} \boldsymbol{f}(x_S) > 0$ *for every* $\alpha \in B_1^K$.

P r o o f. We may observe

$$\log c(\lambda\alpha) = \lambda\left(\alpha^{\mathrm{T}} \boldsymbol{f}\right)_{\max} + \log \sum_{x_S \in \mathcal{X}_S} \exp\left\{\lambda\left[\alpha^{\mathrm{T}} \boldsymbol{f}(x_S) - \alpha^{\mathrm{T}} \boldsymbol{f}\right]_{\max}\right\},$$

where the latter term belongs to the interval $\left(0, \log|\mathcal{X}_S|\right)$. Therefore iii) $\Rightarrow$ ii), and the existence of the minimum is obvious. The uniqueness follows from the convex property.

On the other hand, let us assume i), i.e., that $\log c(\alpha) > \log c(\alpha_0)$ for every $\alpha \in \mathbb{R}^K \setminus \{\alpha_0\}$. Then by convexity

$$\log c(\alpha) - \log c(\alpha_0) \geq \delta_\varepsilon \|\alpha - \alpha_0\|$$

for every $\|\alpha - \alpha_0\| > \varepsilon > 0$ with some $\delta_\varepsilon > 0$, and ii) follows.

If $(\alpha_1^{\mathrm{T}} \boldsymbol{f})_{\max} \leq 0$ for some $\alpha_1 \in \mathbb{R}^K$, we obtain $\log c(\lambda\alpha_1) \leq \log|\mathcal{X}_S|$, which is a contradiction, and therefore ii) $\Rightarrow$ iii) holds also. □

Definition 2.3. If some of the above equivalent conditions are satisfied for a collection $\boldsymbol{f} = (f_1, \ldots, f_K)$, it will be called as a "proper base".

Thus, for a proper base $\boldsymbol{f}$ we have a unique minimum α_0 of the function $\log c(\alpha)$ and a uniquely given $P_S^{\alpha_0} \in \mathcal{P}(\boldsymbol{f}, 0)$ (i.e., $P_S^{\alpha} \notin \mathcal{P}(\boldsymbol{f}, 0)$ for $\alpha \neq \alpha_0$).

Lemma 2.4. *If* $P_S^{\alpha_0} \in \mathcal{P}(\boldsymbol{f}, 0)$ *then*

$$H(P_S^{\alpha_0}) > H(P)$$

for every $P \in \mathcal{P}(\boldsymbol{f}, 0) \setminus \{P_S^{\alpha_0}\}$.

P r o o f . For $P_S^{\alpha_0}$, $P \in \mathcal{P}(\boldsymbol{f}, 0)$, we have $H(P_S^{\alpha}) - H(P) = I\big(P_S^{\alpha_0}|P\big) > 0$ whenever $P \neq P_S^{\alpha_0}$. □

Thus, we may either define the model by considering the exponential family $\{P_S^{\alpha}\}_{\alpha\in\mathbb{R}^K}$ and specify α_0 by the condition $P_S^{\alpha} \in \mathcal{P}(\boldsymbol{f}, 0)$ or we may find $P_S^{\alpha_0}$ in $P(\boldsymbol{f}, 0)$ by imposing the additional condition of maximum entropy.

For a collection $\boldsymbol{f} = (f_1, \ldots, f_K)$, let us introduce

$$\|\boldsymbol{f}\| = \left(\sum_{i=1}^{K} \|f_i\|_\infty^2\right)^{1/2}$$

with the supremum norm $\|\cdot\|_\infty$.

Proposition 2.5. *The set of proper bases with respect to the above norm is open.*

P r o o f . Let $\boldsymbol{f}$ be a proper base. The function $\alpha \mapsto (\alpha^{\mathrm{T}}\boldsymbol{f})_{\max}$ is continuous, and therefore $\boldsymbol{f}_{\max}^{\min} = \min_{\alpha\in B^1} (\alpha^{\mathrm{T}}\boldsymbol{f})_{\max} > 0$. Suppose $\|\boldsymbol{f} - \mathbf{g}\| < \varepsilon < \boldsymbol{f}_{\max}^{\min}$ for some collection $\mathbf{g} = (g_1, \ldots, g_K)$. Then $(\alpha^{\mathrm{T}}\mathbf{g})_{\max} \geq (\alpha^{\mathrm{T}}\boldsymbol{f})_{\max} - \|\boldsymbol{f} - \mathbf{g}\| > \boldsymbol{f}_{\max}^{\min} - \varepsilon > 0$ for every $\alpha \in B_1^K$. □

3. Parametric family

Let us consider the family of collections

$$\{\boldsymbol{f}^\theta\}_{\theta\in\mathbb{R}^K},$$

where $\boldsymbol{f}^\theta = (f_1^\theta, \ldots, f_K^\theta)$ for every $\theta \in \mathbb{R}^K$, for which the mapping

$$\theta \mapsto \boldsymbol{f}^\theta$$

is continuous with the above defined norm.

Further, for some $\theta^0 \in \mathbb{R}^K$ let the collection $\boldsymbol{f}^{\theta^0}$ be a proper base. Then by Proposition 2.5, there exists an open set of proper bases $\{\boldsymbol{f}^\theta\}_{\theta\in\Theta}$, and by Lemma 2.1 and Proposition 2.2 there is a family of uniquely defined probability distributions

$$\{P_S^{\alpha(\theta)}(\cdot|\theta)\}_{\theta\in\Theta},$$

where

$$P_S^{\alpha(\theta)}(x_S|\theta) = \exp\left\{\alpha(\theta)^{\mathrm{T}}\, \boldsymbol{f}^\theta(x_S)\right\}/c\big(\alpha(\theta);\theta\big)\,,$$
$$c(\alpha(\theta);\theta) = \sum_{x_S\in\mathcal{X}_S} \exp\left\{\alpha(\theta)^{\mathrm{T}}\, \boldsymbol{f}^\theta(x_S)\right\}$$

is the normalizing constant, and $\alpha(\theta)$ is implicitly defined by the condition $P_S^{\alpha(\theta)}(\cdot|\theta) \in \mathcal{P}(\boldsymbol{f}^\theta, 0)$.

Proposition 3.1. *The mapping*

$$\alpha\colon \Theta \to \mathbb{R}^K,$$

defined by $P_S^{\alpha(\theta)}(\cdot|\theta) \in \mathcal{P}(\boldsymbol{f}^\theta, 0)$, *is continuous.*

Proof. We have

$$\left|\log c(\alpha;\theta^0) - \log c(\alpha;\theta^1)\right| \le \|\alpha\|\cdot\left\|f^{\theta^0} - f^{\theta^1}\right\|$$

and

$$\log c(\alpha;\theta^0) - \log c\big(\alpha(\theta^0);\theta^0\big) \ge \left\|\alpha - \alpha(\theta^0)\right\|\,\delta_\varepsilon > 0$$

for $\left\|\alpha - \alpha(\theta^0)\right\| > \varepsilon > 0$. Therefore

$$\log c\big(\alpha(\theta^0);\theta^1\big) \le \log c\big(\alpha(\theta^0);\theta^0\big) + \left\|\alpha(\theta^0)\right\|\cdot\left\|f^{\theta^0} - f^{\theta^1}\right\|$$

and

$$\log c(\alpha;\theta^1) \ge \log c\big(\alpha(\theta^0);\theta^0\big) + \left\|\alpha - \alpha(\theta^0)\right\|\cdot\delta_\varepsilon - \|\alpha\|\cdot\left\|f^{\theta^1} - f^{\theta^0}\right\|$$

for $\left\|\alpha - \alpha(\theta^0)\right\| > \varepsilon$. If

$$\left\|f^{\theta^0} - f^{\theta^1}\right\| < \frac{\varepsilon\cdot\delta_\varepsilon}{2\|\alpha(\theta^0)\| + \varepsilon}$$

we obtain finally $\|\alpha(\theta^1) - \alpha(\theta^0)\| < \varepsilon$. □

Let us notice that the above defined distributions are "curved exponential" since the exponent $\alpha(\theta)^{\mathrm{T}}\,\boldsymbol{f}^\theta$ depends nonlinearly on the parameter.

4. Estimation

For the sample size $n \in \mathbb{N}$ let us introduce the empirical distribution $\widehat{P}_S^n$ on $\mathcal{X}_S$ in the standard way. Then, we shall define the estimate $\widehat{\theta}^n$ as a solution of the system of equations

$$\int \boldsymbol{f}^\theta\,\mathrm{d}\widehat{P}^n = 0\,. \tag{S}$$

For the sake of simplicity, we shall make an additional technical assumption, namely

$$\mathrm{Det}\left[\int D^\theta \,\mathrm{d}P_S\right] \neq 0 \qquad \text{for every } \theta \in \Theta \text{ and every positive } P_S \text{ on } \mathcal{X}_S\,, \tag{A}$$

where

$$D^\theta_{ij} = \frac{\partial}{\partial \theta_i} \boldsymbol{f}^\theta_j \qquad \text{for} \quad i,\, j = 1, \ldots, K\,.$$

Theorem 4.1. *Under assumption* (A), *the solution* $\widehat{\theta}^n$ *of the system* (S) *exists with probability tending to 1 as* $n \to \infty$, *and* $\widehat{\theta}^n$ *is a consistent estimate of the true parameter* $\theta^0 \in \Theta$.

P r o o f . The statement follows directly from the well-known implicit function theorem, which guarantees the existence of the continuous mapping

$$P \mapsto \theta\,,$$

given implicitly by the system (S), in a neighbourhood of $P^{\alpha(\theta^0)}(\cdot\,|\theta^0)$. □

5. Separate estimation

As a simple example of separate estimation, we consider a particular form of the collection

$$\boldsymbol{f}^\theta = \left(f^\theta_1, \ldots, f^\theta_K\right)$$

namely we assume

$$f^\theta_i = f^{\theta_\ell}_i \in \mathcal{F}_{A_\ell}$$

for every $i \in K_\ell$, where $A_\ell \subset S$ is "small", $\{K_\ell\}_{\ell=1,\ldots,M}$ is a partition of K with each $K_\ell \subset K$ again "sufficiently small", and where each

$$f^{\theta_\ell}_i : \mathcal{X}_{A_\ell} \to \mathbb{R} \qquad \text{for} \quad i \in K_\ell$$

actually depends only on the vector parameter $\theta_\ell \in \mathbb{R}^{K_\ell}$.

We may, for example, consider a class of conditional distributions

$$\left\{Q^{\theta_\ell}_{V_\ell\,|\,W_\ell}\right\}_{\ell=1,\ldots,M}$$

where $\theta_\ell \in \mathbb{R}^{K_\ell}$, $V_\ell \subset S$, $W_\ell \subset S \setminus V_\ell$ for every $\ell = 1, \ldots, M$. Let us emphasize that these conditional distributions may be inconsistent, i.e., they may not be obtained by any global joint distribution. Further, we assume the likelihood functions

$$\mathcal{L}^{\theta_\ell}_\ell = \log Q^{\theta_\ell}_{V_\ell\,|\,W_\ell}$$

to be twice differentiable with respect to the parameters. We put

$$f_i^{\theta_\ell} = \frac{\partial}{\partial \theta_i} \mathcal{L}_\ell^{\theta_\ell} \qquad \text{for} \quad i \in K_\ell$$

and

$$D_{ij}^{\theta_\ell} = \frac{\partial^2}{\partial \theta_i \, \partial \theta_j} \mathcal{L}_\ell^{\theta_\ell} \qquad \text{for} \quad i, j \in K_\ell \,, \quad \ell = 1, \dots, M \,.$$

Then we obtain the above scheme with $A_\ell = V_\ell \cup W_\ell$ for $\ell = 1, \dots, M$.

Remark 5.1. For $P_S^{\theta^0} = P_S^{\alpha(\theta^0)}(\cdot \,|\, \theta^0)$ the following property we may observe, namely

$$\theta_\ell^0 = \arg \min_{\theta_\ell \in \mathbb{R}^{K_\ell}} \int I\Big(P_{V_\ell|W_\ell}^{\theta^0} \Big| Q_{V_\ell|W_\ell}^{\theta_\ell}\Big) \, \mathrm{d}P_{W_\ell}^{\theta^0}$$

simultaneously for every $\ell = 1, \dots, M$. Thus, we conclude that the conditional distributions $P_{V_\ell|W_\ell}^{\theta^0}$ of the parametric model $P_S^{\theta^0}$ is, in the above sense, as close as possible (but not equal in general) to the given conditional distributions $Q_{V_\ell|W_\ell}^{\theta_\ell^0}$, $\ell = 1, \dots, M$.

Now the estimation procedure consists in solving the sequence of systems

$$0 = \int f_i^{\theta_\ell} \, \mathrm{d}\widehat{P}_{A_\ell}^n \qquad \text{for} \quad i \in K_\ell \tag{S$_\ell$}$$

separately for $\ell = 1, \dots, M$.

Moreover, let us assume

$$-\int D^{\theta_\ell} \, \mathrm{d}P_{A_\ell} > 0$$

for every positive probability measure P_{A_ℓ} and $\ell = 1, \dots, M$, which, equivalently, means that

$$\theta \mapsto -\sum_{\ell=1}^{M} \int \mathcal{L}_\ell^{\theta_\ell} \, \mathrm{d}P_{A_\ell}$$

is a strongly convex function. Then the crucial conditions of Sections 2 and 4 are satisfied.

REFERENCES

[1] BISHOP, Y. M. M.—FLENBERG, S. E. F.—HOLLAND, P. W.: *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass, 1975.

[2] HÁJEK, P.—HAVRÁNEK, T.—JIROUŠEK, R.: *Uncertain Information Processing in Expert Systems*, CRC Press, London, 1992.

[3] JANŽURA, M.: *Maximum pseudo–likelihood estimators*, Tatra Mt. Math. Publ. **7** (1996), 195–201.

Received April 27, 1998

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
CZ–182 08 Prague
CZECH REPUBLIC

E-mail: janzura@utia.cas.cz