Rend. Sem. Mat. Univ. Pol. Torino
Vol. 59, 4 (2001)

C. Macci*

MIXED PARAMETRIZATION FOR CURVED EXPONENTIAL MODELS IN HOMOGENEOUS MARKOV PROCESSES WITH A FINITE STATE SPACE

Abstract. In this paper we deal with homogeneous continuous-time Markov chains (namely homogeneous Markov processes with a finite state space) and we apply the so-called geometrical theory of statistical models to analyze their structure. The results presented here can be considered as a continuous-time version of the results of Rogantin ([10]).

1. Introduction

In this paper we consider the so-called geometrical theory of statistical models and we present a geometrical model of homogeneous continuous-time Markov chains (namely homogeneous Markov processes with a finite state space). The analogous results concerning the discrete-time case are presented by Rogantin ([10]); in particular, for the binary case (namely the case with 2 states only), there is another reference (Rogantin, [9]).

The results presented in this paper and the analogous results concerning the discrete-time case are similar; however there are some differences which will be pointed out in section 6.

In this paper we find a parametrization such that the set of parameters splits in two orthogonal blocks: the first one represents the marginal distributions, the second one represents the intensity matrix which plays the role of the transition matrix in the discrete-time case. We remark that the maximum likelihood estimators of two orthogonal blocks of parameters are asymptotically independent: see e.g. Barndorff-Nielsen and Cox ([4] p. 98) and Murray and Rice ([7] p. 216). We also remark that orthogonal parametrizations are considered in several different situations: for instance see Cox and Reid ([5]).

Section 2 is devoted to recall some preliminaries.

In section 3 we present the log-likelihood (with respect to a suitable law Q) concerning a homogeneous continuous-time Markov chain $(J_t)_{t\in[0,T]}$ and we still have a so-called curved exponential model (see e.g. [4] p. 65).

In section 4 we generalize the mixed parametrization for the curved exponential

*The author thanks G. Pistone and M.P. Rogantin for useful discussions.

model presented in section 3; we recall that the mixed parametrization is a way to obtain an orthogonal parametrization between two subsets of parameters in exponential models: for instance see Amari ([1]) and Barndorff-Nielsen and Cox ([4] p. 62).

In section 5 we consider two submodels: the first one is a particular exponential submodel, the second one concerns the stationary case.

Finally section 6 is devoted to present some concluding remarks.

An infinite-dimensional version of the mixed parametrization is presented in an article of Pistone and Rogantin ([8]) where one can find a wide bibliography concerning the geometrical theory of statistical models; in particular we point out here the reference of Amari ([2]).

2. Preliminaries

In this section we give a short overview of the basic definitions in this paper. For a detailed presentation the reader should consult one of the sources cited above.

Let $\mathcal{X}$ be a measurable space (called sample space) and let ν be a σ-finite measure on $\mathcal{X}$. Then a statistical model is a family of probability densities $(p(x;\theta) : \theta \in \Theta)$ with respect to ν (ν is called dominating measure) where Θ is an open subset of $\mathbb{R}^d$ (for some $d \geq 1$) and $p(x;\theta)$ is sufficiently smooth in θ.

Given a statistical model $(p(x;\theta) : \theta \in \Theta)$, we have a submodel when θ belongs to a suitable subset Θ_0 of Θ.

Now let $T : \mathcal{X} \to \mathbb{R}^d$ be a measurable function and let us denote the usual scalar product in $\mathbb{R}^d$ by $\langle \cdot, \cdot \rangle$. Then $(p(x;\theta) : \theta \in \Theta)$ is an exponential model if the log-likelihood $\log p(x;\theta)$ can be written as

$$\log p(x;\theta) \equiv \langle T(x), \theta \rangle - \Psi(\theta) \tag{1}$$

for all $\theta \in \Theta$, where Ψ is the normalizing factor

$$\Psi(\theta) \equiv \int_{\mathcal{X}} e^{\langle T(x), \theta \rangle} \nu(dx).$$

Similarly $(p(x;\theta) : \theta \in \Theta_0)$ is an exponential submodel if (1) holds for all $\theta \in \Theta_0$.

In view of presenting another concept, let the log-likelihood (1) of an exponential model and an open subset Θ' of $\mathbb{R}^{d'}$ with $d' < d$ be given; then a statistical model $(q(x;u) : u \in \Theta')$ is said to be a curved exponential model if we have

$$\log q(x;u) \equiv \langle T(x), \theta(u) \rangle - \Psi(\theta(u))$$

for all $u \in \Theta'$, where $\theta = \theta(u)$ is satisfies suitable conditions.

Before concluding this section, in view of presenting the topics below, we point out that we use capital letters for the random variables and small letters for the corresponding sample values.

3. Homogeneous and non-stationary case

Let $(J_t)_{t\in[0,T]}$ be a continuous-time Markov chain, namely a homogeneous Markov process with a finite space $E = \{1, \ldots, s\}$, let us denote its initial distribution by $(p_1^{(0)}, \ldots, p_s^{(0)})$ and let us denote its intensity matrix by $G = (\alpha(i, j))_{i,j\in E}$. More precisely we assume that $\alpha(i, j) > 0$ for all $i, j \in E$ with $i \neq j$ and

$$\sum_{j\in E} \alpha(i, j) = 0 \ \ (\forall i \in E);$$

in what follows it is useful to refer to the positive values

$$\alpha(i) = -\alpha(i, i) = \sum_{j\in E,\ j\neq i} \alpha(i, j) \ \ (\forall i \in E). \tag{2}$$

Moreover, for each $t \in [0, T]$, the marginal distribution $(p_1^{(t)}, \ldots, p_s^{(t)})$ of J_t satisfies the obvious condition $\sum_{i\in E} p_i^{(t)} = 1$ and we have

$$(p_1^{(t)}, \ldots, p_s^{(t)}) = (p_1^{(0)}, \ldots, p_s^{(0)})e^{tG} \tag{3}$$

where e^{tG} is the matrix exponential of tG.

The papers of Rogantin ([9] and [10]) concerning the discrete-time case deals with a n-sample; here, in order to have a simpler presentation, we always consider a 1-sample of $(J_t)_{t\in[0,T]}$.
In what follows the ensuing random variables are needed:
for each state $i \in E$, let $N_i^{(0)}$ be the indicator of the event $\{J_0 = i\}$ (namely $N_i^{(0)} = 1_{J_0=i}$);
for each state $i \in E$, let $T_{(i)}$ be the sampling occupation time of $(J_u)_{u\in[0,T]}$ in i;
for $i, j \in E$ with $i \neq j$, let K_{ij} be the sampling number of transitions of $(J_u)_{u\in[0,t]}$ from i to j.
Moreover let K be defined as

$$K = \sum_{i,j\in E,\ i\neq j} K_{ij} \tag{4}$$

and let $(T_h)_{h\geq 0}$ be the epochs of the jumps of $(J_t)_{t\geq 0}$, so that in particular we have

$$0 = T_0 < T_1 < \ldots < T_K \leq T < T_{K+1}.$$

Then we can consider a version of the likelihood with respect to a dominant law Q for $(J_t)_{t\in[0,T]}$ having $q^{(0)} = (q_1^{(0)}, \ldots, q_s^{(0)})$ as the initial distribution and $G_Q = (\beta(i, j))_{i,j\in E}$ as the intensity matrix; in particular we can consider the positive values $(\beta(i))_{i\in E}$ which play the role of the values $(\alpha(i))_{i\in E}$ in (2) for the matrix G and we have

$$\beta(i) = -\beta(i, i) = \sum_{j\in E,\ j\neq i} \beta(i, j) \ \ (\forall i \in E).$$

Thus a version of likelihood is

$$\frac{f(p^{(0)}, G)}{f(q^{(0)}, G_Q)}$$

where

$$f(p^{(0)}, G) = p^{(0)}_{j_0} \underbrace{\prod_{h=1}^{k} \alpha(j_{t_{h-1}}) \exp(-\alpha(j_{t_{h-1}})(t_h - t_{h-1})) \frac{\alpha(j_{t_{h-1}}, j_{t_h})}{\alpha(j_{t_{h-1}})}}_{=1 \text{ if } k=0}$$

$$\cdot \exp(-\alpha(j_{t_k})(T - t_k)) =$$

$$= \prod_{i \in E} (p_i^{(0)})^{n_i^{(0)}} \prod_{i,j \in E,\ j \neq i} \alpha(i,j)^{k_{ij}} \prod_{i \in E} e^{-\alpha(i) t_{(i)}}$$

and, obviously,

$$f(q^{(0)}, G_Q) = q^{(0)}_{j_0} \underbrace{\prod_{h=1}^{k} \beta(j_{t_{h-1}}) \exp(-\beta(j_{t_{h-1}})(t_h - t_{h-1})) \frac{\beta(j_{t_{h-1}}, j_{t_h})}{\beta(j_{t_{h-1}})}}_{=1 \text{ if } k=0}$$

$$\cdot \exp(-\beta(j_{t_k})(T - t_k)) =$$

$$= \prod_{i \in E} (q_i^{(0)})^{n_i^{(0)}} \prod_{i,j \in E,\ j \neq i} \beta(i,j)^{k_{ij}} \prod_{i \in E} e^{-\beta(i) t_{(i)}}.$$

If we consider a choice of the matrix G_Q such that $\beta(i) = 1$ for all $i \in E$ and if we set $p^{(0)} = q^{(0)}$ (namely $p_i^{(0)} = q_i^{(0)}$ for all $i \in E$), we obtain

$$f(p^{(0)}, G_Q) = \prod_{i \in E} (p_i^{(0)})^{n_i^{(0)}} \prod_{i,j \in E,\ j \neq i} \beta(i,j)^{k_{ij}} e^{-T}$$

whence we have

$$\log \frac{f(p^{(0)}, G)}{f(p^{(0)}, G_Q)} = -\sum_{i \in E} t_{(i)} \alpha(i) + \sum_{i,j \in E,\ j \neq i} k_{ij} \log \frac{\alpha(i,j)}{\beta(i,j)} + T =$$

$$= \sum_{i,j \in E,\ j \neq i} k_{ij} \log \frac{\alpha(i,j)/\alpha(i)}{\beta(i,j)/\beta(i)} + \sum_{i,j \in E,\ j \neq i} k_{ij} \log \alpha(i) + \sum_{i \in E} (1 - \alpha(i)) t_{(i)}$$

because $\sum_{i \in E} t_{(i)} = T$. This expression agrees with the expression presented by Dacunha-Castelle and Duflo ([6] p. 286) which concerns a counting point process with marks (see [6] p. 264).

Throughout this paper we consider a different choice of the dominant law Q, namely

$$q_i^{(0)} = \frac{1}{s} \quad (\forall i \in E)$$

and

$$\beta(i, j) = 1 \ \ (\forall i, j \in E \text{ with } i \neq j).$$

Then the positive values $(\beta(i))_{i \in E}$ which play the role of the values $(\alpha(i))_{i \in E}$ in (2) are

$$\beta(i) = -\beta(i, i) = \sum_{j \in E,\ j \neq i} \beta(i, j) = s - 1 \ \ (\forall i \in E).$$

Thus it is easy to check that a version of the log-likelihood is

$$\log \frac{f(p^{(0)}, G)}{f(q^{(0)}, G_Q)} = \log s + \sum_{i \in E} n_i^{(0)} \log p_i^{(0)} - \sum_{i \in E} t_{(i)} \alpha(i) +$$

$$+ \sum_{i,j \in E,\ j \neq i} k_{(ij)} \log \alpha(i, j) + (s-1)T;$$

indeed we have $f(q^{(0)}, G_Q) = \frac{1}{s} \exp(-(s-1)T)$.

By taking into account $\sum_{i \in E} n_i^{(0)} = 1$ and $\sum_{i \in E} t_{(i)} = T$, the latter can be rewritten in a different way; more precisely we choose the elements with index s to play the role of pivot (other choices lead to analogous results) and we have

$$\begin{aligned}
\log \frac{f(p^{(0)}, G)}{f(q^{(0)}, G_Q)} &= \sum_{i=1}^{s-1} n_i^{(0)} \log \frac{p_i^{(0)}}{p_s^{(0)}} + \sum_{i=1}^{s-1} t_{(i)} (\alpha(s) - \alpha(i)) + \\
&+ \sum_{i,j \in E,\ j \neq i} k_{ij} \log \alpha(i, j) + \log s + \log p_s^{(0)} - \alpha(s)T + (s-1)T = \\
&= \sum_{i=1}^{s-1} n_i^{(0)} \log \frac{p_i^{(0)}}{p_s^{(0)}} + \sum_{i=1}^{s-1} t_{(i)} (\alpha(s) - \alpha(i)) + \\
&+ \sum_{i,j \in E,\ j \neq i} k_{ij} \log \alpha(i, j) - [(\alpha(s) - (s-1))T - \log s p_s^{(0)}].
\end{aligned} \tag{5}$$

We remark that we should write down $1 - \sum_{k=1}^{s-1} p_k^{(0)}$ in place of $p_s^{(0)}$.

Now let us consider the following parameters:
θ as $\theta_{ij} = \log \alpha(i, j)$ for $i, j \in E$ with $i \neq j$;
ζ as $p_i^{(0)}$ for $i = 1, \ldots, s-1$.
Then the model (5) can be parametrized with θ and ζ; indeed, by (2), we have

$$\alpha(i) = \sum_{j \in E,\ j \neq i} e^{\theta_{ij}} \ \ (\forall i \in E)$$

which define a full rank transformation (see Appendix). The model (5) is curved because the relations between the parameters (θ, ζ) and the canonical parameters are not linear and the dimension of the sufficient statistics is larger than the dimension of the parameters. Indeed the sufficient statistics is

$$((n_i^{(0)}, t_{(i)})_{i=1,\dots,s-1}, (k_{(ij)})_{i,j\in E,\ i\neq j})$$

so that its dimension is

$$2(s-1)+s(s-1)=s-1+s^2-1,$$

while the dimension of the parameters (θ, ζ) is obviously

$$s(s-1)+s-1=s^2-1.$$

Now let us consider the smallest exponential model which contains the model (5). For this exponential model we refer to the usual notation of the log-likelihood

$$\langle r_1, \theta_1\rangle + \langle r_2, \theta_2\rangle - \psi(\theta_1, \theta_2) \tag{6}$$

where ψ is the normalizing factor; more precisely here we have
$r_1 = (n_i^{(0)}, t_{(i)})_{i=1,\dots,s-1}$ and $r_2 = (k_{ij})_{i,j\in E,\ i\neq j}$ so that the dimensions of θ_1 and θ_2 are $2(s-1)$ and $s(s-1)$ respectively. Moreover, for better explaining the structure of the curved exponential model concerning (5), in (6) we have $\theta_1 = \theta_1(\theta, \zeta)$ and $\theta_2 = \theta_2(\theta, \zeta)$ defined by

$$\begin{cases} \theta_1 = ((\log \frac{p_i^{(0)}}{p_s^{(0)}})_{i=1,\dots,s-1}, (\sum_{j=1}^{s-1} e^{\theta_{sj}} - \sum_{j=1,\ j\neq i}^{s} e^{\theta_{ij}})_{i=1,\dots,s-1}) \\ \theta_2 = (\theta_{ij})_{i,j\in E,\ i\neq j} \end{cases} \tag{7}$$

where, as before, $p_s^{(0)}$ stands for $1-\sum_{k=1}^{s-1} p_k^{(0)}$.

Thus, if we denote the manifold corresponding to (6) by $\mathcal{M}$, the model (5) corresponds to a submanifold $\mathcal{S}_{omo}$ embedded in $\mathcal{M}$. Moreover, as far as the dimensions are concerned, we have

$$\dim \mathcal{M} = 2(s-1)+s(s-1)=s-1+s^2-1 \tag{8}$$

and

$$\dim \mathcal{S}_{omo} = s(s-1)+s-1=s^2-1; \tag{9}$$

we remark that, as for the discrete-time case, the difference between $\dim \mathcal{M}$ and $\dim \mathcal{S}_{omo}$ is equal to $s-1$.
The first $2(s-1)$ elements of $\nabla\psi(\theta_1, \theta_2)$ will be denoted by $(\nabla\psi(\theta_1, \theta_2))_1$ and they correspond to the parameters which depend on the marginal distributions. Then

$$\mathcal{M} : \begin{cases} \eta_1 = (\nabla\psi(\theta_1, \theta_2))_1 \\ \theta_2 = \theta \end{cases}$$

represents a mixed parametrization for the exponential model (6).
We remark that the parametrization of the marginal distributions in (6) emphasizes the initial distribution and the integral of the marginal distributions on $[0, T]$; indeed, for all $i \in E$, we have

$$\mathbb{E}_{p^{(0)},G}[N_i^{(0)}] = p_i^{(0)} \tag{10}$$

and

$$\mathbb{E}_{p^{(0)},G}[T_{(i)}] = \int_0^T p_i^{(t)} dt. \tag{11}$$

We also remark that

$$\frac{d}{dt}(p_1^{(t)}, \dots, p_s^{(t)}) = (p_1^{(0)}, \dots, p_s^{(0)})e^{tG}G = (p_1^{(t)}, \dots, p_s^{(t)})G$$

by (3); thus, by taking into account (11), we obtain

$$\begin{aligned} \mathbb{E}_{p^{(0)},G}[(T_{(1)}, \dots, T_{(n)})]G &= \int_0^T \frac{d}{dt}(p_1^{(t)}, \dots, p_s^{(t)})dt \\ &= (p_1^{(T)}, \dots, p_s^{(T)}) - (p_1^{(0)}, \dots, p_s^{(0)}). \end{aligned}$$

As pointed out in the papers of Rogantin ([9] and [10]), the parametrization of marginal distributions in the smallest exponential model which contains the model of a homogeneous discrete-time Markov chain $(J_t)_{t=0,1,\dots,T}$ emphasizes the following quantities: the initial distribution $(p_i^{(0)})_{i\in E}$, the final distribution $(p_i^{(T)})_{i\in E}$ and the sum of the intermediate marginal distributions $(\sum_{t=1}^{T-1} p_i^{(t)})_{i\in E}$.

Thus (10) and (11) lead us to similar conclusions for the continuous-time case; indeed here we have the integral which plays the role of the sum and the main difference is that the final distribution $(p_i^{(T)})_{i\in E}$ is not emphasized. This means that, with respect to this parametrization, the final state j_T can be neglected; this can be motivated by noting that j_T is determined by the initial state j_0 and the transitions numbers $(k_{ij})_{i,j\in E,\ i\neq j}$ and this leads us to think that it is possible to consider a different parametrization with respect to which the final distribution $(p_i^{(T)})_{i\in E}$ is emphasized. For better explaining how we can determine j_T by knowing j_0 and $(k_{ij})_{i,j\in E,\ i\neq j}$, for each state $i \in E$ let A_i and B_i be the random variables

$$A_i = \sum_{j\in E,\ j\neq i} K_{ji} \text{ and } B_i = \sum_{j\in E,\ j\neq i} K_{ij};$$

then we have two different situations: if $j_T = j_0$ we have

$$a_i - b_i = 0 \text{ for all } i \in E;$$

if $j_T \neq j_0$ we have

$$a_i - b_i = \begin{cases} 0 & \text{if } i \neq j_0 \text{ and } i \neq j_T \\ +1 & \text{if } i = j_T \\ -1 & \text{if } i = j_0 \end{cases}.$$

Finally we point out another difference: in the continuous-time case the total number of transitions of $(J_t)_{t\in[0,T]}$ is a random variable (namely K in (4)), while in the discrete-time case the total number of transitions of $(J_t)_{t=0,1,\dots,T}$ is not random because it is equal to T. Some further differences between discrete-time case and continuous-time case are presented below (section 6).

4. Generalization of mixed parametrization for curved exponential models

First of all let us consider a notation used in what follows: the identity matrix of dimension m ($m \geq 1$) will be denoted by I_m.
Then let us start with a parametric representation of $\mathcal{S}_{omo}$ in $\mathcal{M}$: for the function F defined by (7) we have

$$\mathcal{S}_{omo} \; : \; \begin{cases} \theta_1 = F(\zeta,\theta) \\ \theta_2 = \theta \end{cases} . \tag{12}$$

Then let us consider $\hat{P} = (\hat{\zeta},\hat{\theta})$ in $\mathcal{S}_{omo}$ and let $\gamma_{\hat{P}}(\theta)$ be the coordinate surface through $\hat{P}$ lying on $\mathcal{S}_{omo}$; a parametric representation is

$$\gamma_{\hat{P}}(\theta) \; : \; \begin{cases} \theta_1 = F(\hat{\zeta},\theta) \\ \theta_2 = \theta \end{cases}$$

Now let $F_2(\hat{\zeta},\hat{\theta})$ be the matrix with the derivatives of $F(\zeta,\theta)$ with respect to the parameters θ and evaluated in $\hat{P}$; we remark that the matrix $F_2(\hat{\zeta},\hat{\theta})$ has $2(s-1)$ rows and $s(s-1)$ columns. Then the columns of the matrix $\gamma'_{\hat{P}}(\hat{\theta}) = \binom{F_2(\hat{\zeta},\hat{\theta})}{I_{s(s-1)}}$ generate the tangent space to $\gamma_{\hat{P}}(\theta)$ at the point $\hat{P}$; obviously the matrix $\gamma'_{\hat{P}}(\theta)$ has $2(s-1)+s(s-1)$ rows and $s(s-1)$ columns.
It is possible to write down an exponential model which corresponds to $\mathcal{M}$ such that in $\hat{P}$ has the latter $s(s-1)$ coordinates proportional to $\gamma'_{\hat{P}}(\hat{\theta})$ and the parameters - denoted by (τ_1,τ_2) - are linear combinations of (θ_1,θ_2). More precisely we have

$$\langle r_1,\tau_1\rangle + \langle (r_1,r_2)\gamma'_{\hat{P}}(\hat{\theta}),\tau_2\rangle - \tilde{\psi}(\tau_1,\tau_2)$$

which can be rewritten as follows

$$\langle r_1,\tau_1 + F_2(\hat{\zeta},\hat{\theta})\tau_2\rangle + \langle r_2,\tau_2\rangle - \tilde{\psi}(\tau_1,\tau_2)$$

so that, by taking into account (6), we can consider the function h defined as follows

$$(\tau_1,\tau_2) \mapsto (\theta_1,\theta_2) = h(\tau_1,\tau_2) = \begin{cases} \theta_1 = \tau_1 + F_2(\hat{\zeta},\hat{\theta})\tau_2 \\ \theta_2 = \tau_2 \end{cases} ;$$

then in particular we have

$$\psi(h(\tau_1,\tau_2)) \equiv \tilde{\psi}(\tau_1,\tau_2)$$

and

$$\nabla_\theta \psi(h(\tau_1, \tau_2)) J_h(\tau_1, \tau_2) \equiv \nabla_\tau \tilde{\psi}(\tau_1, \tau_2)$$

where

$$J_h(\tau_1, \tau_2) = \begin{pmatrix} I_{2(s-1)} & F_2(\hat{\zeta}, \hat{\theta}) \\ 0 & I_{s(s-1)} \end{pmatrix}$$

is the Jacobian matrix of h in which we exhibit the blocks of dimension $2(s-1)$ and $s(s-1)$. Thus, as far as the first $2(s-1)$ components are concerned, the gradients coincide:

$$(\nabla_\theta \psi(h(\tau_1, \tau_2)))_1 \equiv (\nabla_\tau \tilde{\psi}(\tau_1, \tau_2))_1. \tag{13}$$

In conclusion, if we consider the mixed parametrization on $\mathcal{S}_{omo}$, the parameters η_1 which depend on the marginal distributions are orthogonal to the parameters θ in $\hat{P}$.

Now let $\mathcal{H}$ be the coordinate surface corresponding to η_1 in $\hat{P}$; we remark that $\mathcal{H}$ corresponds to an exponential submodel in $\mathcal{M}$ so that it is not curved. We can say that $\mathcal{H}$ is orthogonal to the tangent space to $\gamma_{\hat{P}}(\theta)$ at the point $\hat{P}$.
Moreover let us consider the surface $\mathcal{C}_{omo} = \mathcal{H} \cap \mathcal{S}_{omo}$. We can say that $\dim \mathcal{C}_{omo} = s-1$. Indeed, for each $\hat{P} \in \mathcal{S}_{omo}$, the tangent space to $\mathcal{H}$ (which is $\mathcal{H}$ itself) at the point $\hat{P}$ is orthogonal to the tangent space to $\gamma_{\hat{P}}(\theta)$ at the point $\hat{P}$ and the sum between the dimensions of these two tangent spaces is equal to $\dim \mathcal{M}$; then, by (9) and (8),

$$\dim \mathcal{C}_{omo} = \dim \mathcal{H} + \dim \mathcal{S}_{omo} - \dim \mathcal{M} = \tag{14}$$

$$= 2(s-1) + (s^2-1) - (s-1+s^2-1) = s-1$$

follows from Grassmann formula.
This fact does not depend on the selected point $\hat{P}$. Indeed, for any other point $\tilde{P}$ in $\mathcal{C}_{omo}$, let $\gamma_{\tilde{P}}(\theta)$ be the coordinate surface through $\tilde{P}$ and the orthogonal space to the tangent space $\gamma'_{\tilde{P}}(\tilde{\theta})$ is still $\mathcal{H}$ by (13).
In conclusion the surface $\mathcal{C}_{omo}$ can be represented with $s-1$ parameters function of η_1.

5. Two submodels in $\mathcal{S}_{omo}$

The first submodel concerns the case in which all the values $\alpha(i, j)$ ($i, j \in E$ with $i \neq j$) are equal to a fixed positive value; in such a case we obtain an exponential submodel. The second case concerns the stationary case.

5.1. An exponential submodel in $\mathcal{S}_{omo}$

Let us consider the case in which all the values $\alpha(i, j)$ ($i, j \in E$ with $i \neq j$) are equal to a positive value α; in such a case (2) gives

$$\alpha(i) = (s-1)\alpha \quad (\forall i \in E).$$

Then, by referring to the random variable K defined in (4), the log-likelihood in (5) becomes

$$\sum_{i=1}^{s-1} n_i^{(0)} \log \frac{p_i^{(0)}}{p_s^{(0)}} + k \log \alpha - [(s-1)(\alpha-1)T - \log s p_s^{(0)}];$$

we still remark that we should write down $1 - \sum_{k=1}^{s-1} p_k^{(0)}$ in place of $p_s^{(0)}$.
Thus we have an exponential submodel of $\mathcal{S}_{omo}$ (so that it corresponds to a not curved surface) and its dimension is s; more precisely

$$\langle v_1, \theta_1 \rangle + \langle v_2, \theta_2 \rangle - \psi^*(\theta_1, \theta_2)$$

where:
ψ^* is the normalizing factor;
v_1 and θ_1 have dimension $s-1$, $v_1 = (n_i^{(0)})_{i=1,\dots,s-1}$ and

$$\theta_1 = (\theta_1^{(i)})_{i=1,\dots,s-1} = (\log \frac{p_i^{(0)}}{p_s^{(0)}})_{i=1,\dots,s-1}; \tag{15}$$

v_2 and θ_2 have dimension 1, $v_2 = k$ and $\theta_2 = \log \alpha$.
Then, as far as the expression of $\psi^*(\theta_1, \theta_2)$ is concerned, we have

$$\psi^*(\theta_1, \theta_2) = (s-1)(e^{\theta_2} - 1)T - \log[s(1 + \sum_{i=1}^{s-1} e^{\theta_1^{(i)}})^{-1}]; \tag{16}$$

we remark that, in order to obtain (16), $\theta_2 = \log \alpha$ and (15) give

$$\alpha = e^{\theta_2} \tag{17}$$

and

$$\sum_{i=1}^{s-1} e^{\theta_1^{(i)}} = \frac{\sum_{i=1}^{s-1} p_i^{(0)}}{p_s^{(0)}} = \frac{1 - p_s^{(0)}}{p_s^{(0)}} = \frac{1}{p_s^{(0)}} - 1,$$

whence we obtain

$$p_s^{(0)} = (1 + \sum_{i=1}^{s-1} e^{\theta_1^{(i)}})^{-1}; \tag{18}$$

thus (16) follows by replacing (17) and (18) in $(s-1)(\alpha-1)T - \log s p_s^{(0)}$.
Finally let us consider the gradient of $\psi^*(\theta_1, \theta_2)$ in (16); we obtain

$$\partial_{\theta_1^{(i)}} \psi^*(\theta_1, \theta_2) = \frac{e^{\theta_1^{(i)}}}{1 + \sum_{j=1}^{s-1} e^{\theta_1^{(j)}}} \quad (\forall i \in \{1, \dots, s-1\})$$

and

$$\partial_{\theta_2} \psi^*(\theta_1, \theta_2) = (s-1)e^{\theta_2} T,$$

so that we have

$$\partial_{\theta_1^{(i)}} \psi^* = \frac{\frac{p_i^{(0)}}{p_s^{(0)}}}{1 + \sum_{j=1}^{s-1} \frac{p_j^{(0)}}{p_s^{(0)}}} = p_i^{(0)} \quad (\forall i \in \{1, \dots, s-1\})$$

and

$$\partial_{\theta_2} \psi^* = (s-1)\alpha T.$$

In conclusion the parameters $(p_i^{(0)})_{i=1,\dots,s-1}$ concerning the initial distribution are orthogonal to $(s-1)\alpha T$.

5.2. Stationary case

In such a case the initial distribution and the other marginal distributions coincide; these distributions can be expressed in terms of the entries of the matrix G introduced at the beginning of section 3; thus we have $s(s-1)$ parameters. In such a case the log-likelihood in (5) can be rewritten in an analogous way, with $(p_i^{(0)})_{i=1,\dots,s-1}$ replaced by suitable functions of the entries of the intensity matrix G.
This model is represented by a submanifold $\mathcal{S}_{sta}$ embedded in $\mathcal{M}$ and let us consider $\mathcal{C}_{sta} = \mathcal{H} \cap \mathcal{S}_{sta}$; then, since we have $\dim \mathcal{S}_{sta} = s(s-1)$, we can still employ Grassmann formula as in section 4 and we obtain

$$\dim \mathcal{C}_{sta} = \dim \mathcal{H} + \dim \mathcal{S}_{sta} - \dim \mathcal{M} =$$

$$= 2(s-1) + s(s-1) - (s-1+s^2-1) = 0.$$

In next section 6 $\dim \mathcal{C}_{sta} = 0$ will be explained in a more direct way.
Finally we remark that the choice of initial distribution in order to have the stationary case gives $s-1$ further conditions; this explains in a direct way the following:

$$\dim \mathcal{C}_{omo} - \dim \mathcal{C}_{sta} = s - 1. \tag{19}$$

6. Concluding remarks

6.1. Comparison with the discrete-time case

Let us start by recalling some results derived in the previous sections:

$$\dim \mathcal{C}_{sta} = 0; \quad \dim \mathcal{C}_{omo} = s-1; \quad \dim \mathcal{H} = 2(s-1)$$

where $\dim \mathcal{H}$ coincides with the number of components of η_1.
The analogous results concerning the discrete-time case are (see Rogantin, [10]):

$$\dim \mathcal{C}_{sta} = s-1; \quad \dim \mathcal{C}_{omo} = 2(s-1); \quad \dim \mathcal{H} = 3(s-1).$$

Thus (19) still holds in the discrete-time case and it seems to be natural; indeed, in order to have the stationary case, the initial distribution can be expressed in terms of suitable functions of the entries of the transition matrix and we still have $s-1$ further conditions.

In both the cases (continuous-time and discrete-time) we have (8) and (9). Moreover we have

$$\dim \mathcal{S}_{omo} = s - 1 + s(s-1)$$

(see (9)) which seems to be natural; indeed we have $s-1$ conditions for the initial distribution and $s(s-1)$ further conditions which concern the intensity matrix G in the continuous-time case and the transition matrix in the discrete-time case.

As far as $\dim \mathcal{H}$ is concerned, (14) holds in both the cases and the difference between the continuous-time case and the discrete-time case concerns the different dimensions of the sufficient statistics. More precisely, by referring to the log-likelihood in (6), in the continuous-time case we have

$$r_1 = (n_i^{(0)}, t_{(i)})_{i=1,\dots,s-1}$$

while, in the discrete-time case, its analogous is

$$r_1 = (n_i^{(0)}, \sum_{t=1}^{T-1} n_i^{(t)}, n_i^{(T)})_{i=1,\dots,s-1}$$

where, for each state $i \in E$ and for each time $t \in \{0, 1, \dots, T\}$, the random variable $N_i^{(t)}$ is the indicator of the event $\{J_t = i\}$ (namely $N_i^{(t)} = 1_{J_t = i}$).

Finally by (10) and (11) we have

$$\eta_1 = (\eta_1^{(1)}, \eta_1^{(2)}) = ((p_i^{(0)})_{i=1,\dots,s-1}, (\int_0^T p_i^{(t)} dt)_{i=1,\dots,s-1});$$

thus, by (12) with F defined by (7), the surface $\mathcal{C}_{omo}$ can be represented with $s-1$ parameters functions of $\eta_1^{(1)}$ only (namely the initial distribution only). This fact also explains in a more direct way that the surface $\mathcal{C}_{sta}$ concerning the stationary case is reduced to a single point.

6.2. A possible future work

An idea for a possible future work concerns Markov additive processes (see e.g. [3] pp. 39–47), namely bivariate Markov processes (J_t, S_t) where (J_t) is a Markov process with state space E and the increments of the $\mathbb{R}^d$-valued process (S_t) satisfy a suitable condition. In particular one could refer to the case in which E is finite because, in such a case, the structure of Markov additive process is completely understood with some differences between discrete-time case and continuous-time case.

Thus we could expect to find out some differences in terms of geometrical theory of statistical models. These differences could have some connection with the differences presented in this paper between discrete-time case and continuous-time case for (J_t).

Appendix

Let us consider the model (5) and the parameters introduced in section 3:
θ as $\theta_{ij} = \log\alpha(i, j)$ for $i, j \in E$ with $i \neq j$;
ζ as $p_i^{(0)}$ for $i = 1, \dots, s-1$.

We concentrate our attention on the function defined as follows: the image of a point $((p_i^{(0)})_{i=1,\dots,s-1}, \theta)$ in the domain is

$$((\log \frac{p_i^{(0)}}{1 - \sum_{k=1}^{s-1} p_k^{(0)}})_{i=1,\dots,s-1}, (\sum_{j=1}^{s-1} e^{\theta_{sj}} - \sum_{j=1,\ j\neq i}^{s} e^{\theta_{ij}})_{i=1,\dots,s-1}, \theta).$$

Thus the jacobian matrix is

$$J = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \\ 0 & I_{s(s-1)} \end{pmatrix}$$

in which we exhibit three blocks:

$$B_1 = (\frac{\partial}{\partial p_j^{(0)}} \log \frac{p_i^{(0)}}{1 - \sum_{k=1}^{s-1} p_k^{(0)}})_{i,j\in 1,\dots,s-1};$$

$$B_2 = (\frac{\partial}{\partial\theta_{ij}} (\sum_{j=1}^{s-1} e^{\theta_{sj}} - \sum_{j=1,\ j\neq k}^{s} e^{\theta_{kj}}))_{k=1,\dots,s-1,\ i,j\in E\ i\neq j};$$

$$I_{s(s-1)}.$$

Then the transformation has full rank if we find an invertible minor in J of order $s - 1 + s(s-1)$. To this aim we remark that

$$\det \begin{pmatrix} B_1 & 0 \\ 0 & I_{s(s-1)} \end{pmatrix} = \det B_1;$$

then we need to prove that $\det B_1 \neq 0$. This follows from the next

PROPOSITION 1. *We have*

$$\det B_1 = \frac{1}{p_1^{(0)}} \cdots \frac{1}{p_{s-1}^{(0)}} \frac{1}{1 - \sum_{k=1}^{s-1} p_k^{(0)}}.$$

Proof. In this proof it will be useful to write down $p_s^{(0)}$ in place of $1 - \sum_{k=1}^{s-1} p_k^{(0)}$. Let us start an explicit expression for the entries of B_1; for $i, j = 1, \dots, s-1$ we have

$$\frac{\partial}{\partial p_j^{(0)}} \log \frac{p_i^{(0)}}{1 - \sum_{k=1}^{s-1} p_k^{(0)}} = \begin{cases} \frac{1}{p_i^{(0)}} + \frac{1}{1-\sum_{k=1}^{s-1} p_k^{(0)}} = \frac{1}{p_i^{(0)}} + \frac{1}{p_s^{(0)}} & (i = j) \\ \frac{1}{1-\sum_{k=1}^{s-1} p_k^{(0)}} = \frac{1}{p_s^{(0)}} & (i \neq j) \end{cases}.$$

Then, in order to write down the columns of B_1, let us consider the following notation: let be C the column having all the entries equal to $\frac{1}{p_s^{(0)}}$ and, for all $k = 1, \dots, s-1$, let C_k be the column having $\frac{1}{p_k^{(0)}}$ in the k-th place and all the other entries equal to zero. Then we have

$$B_1 = (C + C_1, \dots, C + C_{s-1});$$

namely, for all $k = 1, \dots, s-1$, the k-th column is $C + C_k$. It is known that the determinant of a matrix is multilinear with respect to the columns of matrix and, when at least two columns of a matrix coincide, its determinant is equal to zero; thus we obtain

$$\det B_1 = \det(C_1, \dots, C_{s-1}) + \sum_{k=1}^{s-1} \det(C_1, \dots, C_{k-1}, C, C_{k+1}, \dots, C_{s-1}) =$$

$$= \frac{1}{p_1^{(0)}} \cdots \frac{1}{p_{s-1}^{(0)}} + \sum_{k=1}^{s-1} \frac{1}{p_1^{(0)}} \cdots \frac{1}{p_{k-1}^{(0)}} \frac{1}{p_s^{(0)}} \frac{1}{p_{k+1}^{(0)}} \cdots \frac{1}{p_{s-1}^{(0)}} =$$

$$= \frac{p_s^{(0)} + \sum_{k=1}^{s-1} p_k^{(0)}}{p_1^{(0)} \cdots p_s^{(0)}} = \frac{1}{p_1^{(0)} \cdots p_s^{(0)}}.$$

□

References

[1] AMARI S.I., *Differential geometry of curved exponential families-curvatures and information loss*, Ann. Stat. **10** (1982), 357–385.

[2] AMARI S.I., *Differential-geometrical methods in statistics*, Lecture Notes in Statistics **28**, Springer-Verlag, Berlin 1985.

[3] ASMUSSEN S., *Ruin Probabilities*, World Scientific, London 2000.

[4] BARNDORFF-NIELSEN O.E. AND COX D.R., *Inference and asymptotics*, Monographs in Statistics and Applied Probability **52**, Chapman & Hall, London 1994.

[5] COX D.R. AND REID N., *Parameter orthogonality and approximate conditional inference (with a discussion)*, J. Roy. Stat. Soc., Ser. B **49** (1987), 1–39.

[6] DACUNHA-CASTELLE D. AND DUFLO M., *Probability and Statistics* vol. II, Springer-Verlag, New York 1986.

[7] MURRAY M.K. AND RICE J.W., *Differential geometry in statistics*, Monographs in Statistics and Applied Probability **48**, Chapman & Hall, London 1993.

[8] PISTONE G. AND ROGANTIN M.P., *The exponential statistical manifold: mean parameters, orthogonality and space transformations*, Bernoulli **5** (1999), 721–760.

[9] ROGANTIN M.P., *Geometrical modelling of Markovian dependence*, Metron **54** (1996), 45–65.

[10] ROGANTIN M.P., *Parametrizzazione mista per modelli esponenziali curvi nel caso di processi Markoviani discreti*, Atti della XXXVIII Riunione Scientifica della Società Italiana di Statistica (Rimini 9-13 Aprile 1996), 235–242.

AMS Subject Classification: 62M99, 60J27.

Claudio MACCI
Dipartimento di Matematica
Università di Torino
Via Carlo Alberto 10
10123 Torino, ITALIA
e-mail: `claudio.macci@unito.it`

Lavoro pervenuto in redazione il 18.10.2000 e, in forma definitiva, il 20.06.2001.