A version of the geometry of the multivariate Gaussian model, with applications

Una versione della geometria del modello gaussiano multivariato, con applicazioni

Giovanni Pistone

Abstract We present a version of the classical geometry of the Gaussian multivariate model that has some advantage in the treatment of operations on the tangent bundle on the model. Applications and generalizations are briefly discussed.
Abstract *Presentiamo una versione della classica geometria differenziale del modello gaussiano multivariato che presenta qualche vantaggio quando si considerano operazioni sul fibrato tangente. Discutiamo brevemente applicazioni e generalizzazioni.*

Key words: Information Geometry, Multivariate Gaussian Model, Exponential families

1 Introduction

The geometry of the multivariate Gaussian model $\mathrm{N}(\mu,\Sigma)$ has been studied in detail by Skovgaard in [7], where normal densities are parameterized by the mean parameter μ and the covariance matrix Σ, and the relevant Riemannian geometry is based on an explicit form of the Fisher information. The study of statistical models with tools from differential geometry is frequently called Information Geometry, a name popularized by Amari, see [1]. The main contribution of Amari is the identification of a dually flat connection structure that largely extends the original Riemannian approach. In this paper we present this theory in the framework we reviewed in [5]. It is a non parametric approach where a statistical model is presented in exponential form $\mathscr{M} = \left\{\exp\left(U - K_p(U)\right)\cdot p\right\}$, while the *tangent bundle* is given a concrete form, that is the set of couples (p,u) with $p \in \mathscr{M}$ and u a Fisher score (directional derivative of the log-likelihood) at p. We believe that this approach is conceptually

G. Pistone
Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy e-mail: giovanni.pistone@carloalberto.org

better that the standard abstract parametric presentation because: 1. it allows for non parametric models; 2. presents the tangent bundle as a natural statistical object i.e. the Fisher scores and presents vector fields as statistical pivotal quantities; 3. gives the Amari's dual parallel transports a simple concrete form; 4. it is useful even if the model has actually a finite number of parameters, as it is the case for the multivariate Gaussian model, because connects clearly the geometry on the parameters space with the geometry on the sample space.

In Sec. 2 we discuss the Gaussian model in the form $\exp(U - K_p(U)) \cdot p$, where p is the standard Gaussian density and the second order polynomial U is a linear combination of Hermite polynomials. If $Z \sim \nu_1 = \mathrm{N}(0,1)$ and f, g real smooth functions, then $\mathbb{E}[df(Z)g(Z)] = \mathbb{E}[f(Z)\delta g(Z)]$ where $df(x) = f'(x)$ and $\delta g = xg(x) - g'(x)$ is the Stein operator. It follows that each H_n is a monic polynomial of degree n, $dH_n = nH_{n-1}$, $\mathbb{E}[H_n(Z)H_m(Z)] = 0$ for $n \neq m$, $\mathbb{E}\left[H_n(Z)^2\right] = n!$. In dimension d, for each multi-index α, we define $H_\alpha((x)) = \prod_{i=1}^d H_{\alpha_i}(x_i)$ to get an orthogonal basis of $L^2(\nu_d)$, $\nu_d = \mathrm{N}_d(\mathbf{0}, I_d)$. If we define $d^\alpha = \prod_{i=1}^d d_{x_i}^{\alpha_i}$, $\delta^\alpha = \prod_{i=1}^d \delta_{x_i}^{\alpha_i}$, we have for functions $f, g \colon \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{Z} \sim \nu_d$ that $\mathbb{E}[d^\alpha f(\mathbf{Z})g(\mathbf{Z})] = \mathbb{E}[f(\mathbf{Z})\delta^\alpha g(\mathbf{Z})]$. Sometimes it is convenient to use $\widetilde{H}_\alpha = H_\alpha/\alpha!$.

Second order geometrical object such as Levi-Civita connection and curvature are not discussed in this paper. We restrict , in Sec. 3, to a short presentation of our formalism to model based optimization.

2 Gaussian model in the Hermite basis

Given a vector of means $\mu \in \mathbb{R}^m$ and and a full-rank covariance matrix $\Sigma \in S_m^+$, with $\Sigma = [\sigma_{ij}]$ and $\Sigma^{-1} = [\sigma^{ij}]$, the exponent $-(1/2)(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)$ in the Gaussian density $\mathrm{N}(\mu, \Sigma)$ can be written

$$-\frac{1}{2}\left(\mu^t\Sigma^{-1}\mu + \mathrm{Tr}\left(\Sigma^{-1}\right) + 2\sum_i (\mu^t\Sigma^{-1})_i H_i(\mathbf{x}) + 2\sum_{i<j}\sigma^{ij}H_{ij}(\mathbf{x}) + \sum_i \sigma^{ii}H_{ii}(\mathbf{x})\right),$$

where $H_i(\mathbf{x}) = H_1(x_i) = x_i$ and $H_{ii}(\mathbf{x}) = H_2(x_i) = x_i^2 - 1$ for $i = 1, \ldots, m$, and $H_{ij} = H_1(x_i)H_1(x_j) = x_i x_j$ for $1 \leq i < j \leq m$. The likelihood of $\mathrm{N}(\mu, \Sigma)$ with respect to the standard Gaussian with density $w(\mathbf{x}) = (2\pi)^{-1/2}\exp(-\mathbf{x}^t\mathbf{x}/2)$ has exponent

$$-\frac{1}{2}\mu^t\Sigma^{-1}\mu - \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\right) - \frac{m}{2} + \sum_i (\mu^t\Sigma^{-1})_i H_i(\mathbf{x}) - \sum_{i<j}\sigma^{ij}H_{ij}(\mathbf{x}) - \sum_i (\sigma^{ii}-1)\frac{H_{ii}(\mathbf{x})}{2}$$

Vice-versa, given $I - \Theta \in S_m^+$ and $\theta \in \mathbb{R}^n$, then

$$p(\mathbf{x};\theta_i,\theta_{ij}\colon i\le j)=$$
$$\exp\left(\sum_i \theta_i H_i(\mathbf{x})+\sum_{i<j}\theta_{ij}H_{ij}(\mathbf{x})+\sum_i \theta_{ii}\frac{H_{ii}(\mathbf{x})}{2}-\psi(\theta_i,\theta_{ij}\colon i\le j)\right)w(\mathbf{x}) \quad (1)$$

is the multivariate Gaussian density with $\Sigma^{-1}\mu=\theta=(\theta_i\colon i=1,\dots,n)$, $I-\Sigma^{-1}=\Theta$ with upper entries $(\theta_{ij}\colon i<j)$, and cumulant generating function

$$\psi(\theta_i,\theta_{ij}\colon i\le j)=\frac{1}{2}\theta^t(I-\Theta)^{-1}\theta-\frac{1}{2}\operatorname{Tr}(\Theta)-\frac{1}{2}\log\det(I-\Theta). \quad (2)$$

In Eq. (1) the Gaussian model is presented as an exponential family with natural parameters $(\theta_i\colon i=1,\dots,m;\theta_{ij}\colon 1\le i\le j\le m)$ in the open convex set $\mathbb{R}^n\times(I+S_m^-)$ and w-orthogonal sufficient statistics. From $(\partial/\partial\theta_i)\theta=\mathbf{e}_i$, $(\partial/\partial\theta_{ij})\Theta=E^{ij}$ and Eq. (2) we can compute the first derivatives of the cumulant generating function ψ, that is the expected values of the sufficient statistics,

$$\frac{\partial}{\partial\theta_i}\psi=\theta^t(I-\Theta)^{-1}\mathbf{e}_i=\mu_i, \quad (3)$$
$$\begin{aligned}\frac{\partial}{\partial\theta_{ij}}\psi&=\frac{1}{2}\operatorname{Tr}\left((I-\Theta)^{-1}E^{ij}\right)+\frac{1}{2}\theta^t(I-\Theta)^{-1}E^{ij}(I-\Theta)^{-1}\theta\\&=\sigma_{ij}+\mu_i\mu_j,\quad i<j,\end{aligned} \quad (4)$$
$$\begin{aligned}\frac{\partial}{\partial\theta_{ii}}\psi&=\frac{1}{2}\operatorname{Tr}\left((I-\Theta)^{-1}E^{ii}\right)+\frac{1}{2}\theta^t(I-\Theta)^{-1}E^{ii}(I-\Theta)^{-1}\theta-\frac{1}{2}\\&=\frac{1}{2}\left(\sigma_{ii}+\mu_i^2-1\right).\end{aligned} \quad (5)$$

The second derivatives, that is the covariances of the sufficient statistics, are

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\psi=\mathbf{e}_j^t(I-\Theta)^{-1}\mathbf{e}_i,\quad \frac{\partial^2}{\partial\theta_i\partial\theta_{jh}}\psi=\theta^t(I-\Theta)^{-1}E^{jk}(I-\Theta)^{-1}\mathbf{e}_i,$$

and

$$\begin{aligned}\frac{\partial^2}{\partial\theta_{ij}\partial\theta_{hk}}\psi=\frac{1}{2}\operatorname{Tr}&\left((I-\Theta)^{-1}E^{hk}(I-\Theta)^{-1}E^{ij}\right)\\&+\frac{1}{2}\theta^t(I-\Theta)^{-1}E^{hk}(I-\Theta)^{-1}E^{ij}(I-\Theta)^{-1}\theta\\&+\frac{1}{2}\theta^t(I-\Theta)^{-1}E^{ij}(I-\Theta)^{-1}E^{hk}(I-\Theta)^{-1}\theta.\end{aligned}$$

This formulæ are to be compared with the expression of the Riemannian metric in [7]. Yo Sheena [6] has a different parameterization in which the Fisher matrix is diagonal.

We have used up to now standard change-of-parameter computations. We turn now to exploit specific properties of the Hermite system. Let us write $U(\mathbf{x};\theta,\Theta)=$

$\sum_i \theta_i H_i(\mathbf{x}) + \sum_{i<j} \theta_{ij} H_{ij}(\mathbf{x}) + \sum_i \theta_{ii} \frac{H_{ii}(\mathbf{x})}{2}$. The vector space generated by the sufficient statistics Span $\left(U_{\theta,\Theta} :\ \theta, \Theta\right)$ is the space of polynomials up to degree 2 in the variables $X_1, \ldots, X_n$ that are centered with respect to w. In the geometrical picture, it is the tangent space at w of the Gaussian model, while the tangent space at $p_{\theta,\Theta}$ is generated by the Fisher's scores, i.e. the partial derivatives of the log-density, see the discussion in [5]. We have

$$\begin{aligned}
&\partial U(\mathbf{x};\theta,\Theta)/\partial x_i = \\
&\frac{\partial}{\partial x_i}\left(\theta_i H_1(x_i) + H_1(x_i)\sum_{j<i}\theta_{ji}H_1(x_j) + \frac{1}{2}\theta_{ii}H_2(x_i) + H_1(x_i)\sum_{i<j}\theta_{ij}H_1(x_j)\right) \\
&\qquad = \theta_i + \sum_{j<i}\theta_{ji}H_1(x_j) + \theta_{ii}H_1(x_i) + \sum_{i<j}\theta_{ij}H_1(x_j)
\end{aligned}$$

and $\partial^2 U(\mathbf{x};\theta,\Theta)/\partial x_i \partial x_j = \theta_{ij}$. In matrix form, the basic relation between parameters of the Gaussian model and Hermite polynomials is

$$\nabla_{\mathbf{x}} U(\mathbf{x};\theta,\Theta) = \theta + \Theta\mathbf{x}, \quad \mathrm{Hess}_{\mathbf{x}} U(\mathbf{x};\theta,\Theta) = \Theta. \tag{6}$$

Let us write the expectation parameters as $\eta_i = \mathbb{E}_{\theta,\Theta}[H_i]$, $\eta_{ij} = \mathbb{E}_{\theta,\Theta}[H_{ij}]$, $i<j$, $\eta_{ii} = \mathbb{E}_{\theta,\Theta}[H_{ij}]/2$, and $\mathbb{E}_{\mathbf{0},I} = \mathbb{E}$. We can compute the η's as moments, instead of derivatives of the cumulant generating function. From $H_i = \delta_i 1$,

$$\begin{aligned}
\eta_i = \mathbb{E}\left[H_i \mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] &= \mathbb{E}\left[\partial_i \mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] \\
&= \mathbb{E}\left[(\theta_i + \sum_j \theta_{ij}H_j)\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] = \theta_i + \sum_j \theta_{ij}\eta_j,
\end{aligned}$$

or $\eta = \theta + \Theta\eta$, $\eta = (I-\Theta)^{-1}\theta$, cf. Eq. (3). For η_{ij} we need

$$\begin{aligned}
\partial_i\partial_j \mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)} &= \left(\theta_{ij} + (\theta_i + \sum_h \theta_{ih}H_h)(\theta_j + \sum_k \theta_{jk}H_k)\right)\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)} \\
&= \left(\theta_{ij} + \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})H_h + \sum_{h,k}\theta_{ik}\theta_{jh}H_hH_k\right)\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}.
\end{aligned}$$

From $H_{ij} = \delta^i\delta_j 1$, $\mathbb{E}_{\theta,\Theta}[H_hH_k] = \eta_{hk}$ if $h \neq k$, $\mathbb{E}_{\theta,\Theta}\left[H_h^2\right] = 2\eta_{hh} + 1$, we obtain

$$\eta_{ij} = \theta_{ij} + \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})\eta_h + \sum_{h\neq k}\theta_{ik}\theta_{jh}\eta_{hk} + \sum_h \theta_{ih}\theta_{jh}(2\eta_{hh}+1),$$

to be compared with Eqs. (4) and (5).

3 Optimization

Let $f\colon \mathbb{R}^m \to R$ be a continuous bounded function, with maximum at a point $\mathbf{m} \in \mathbb{R}^m$. We define the *relaxed function* $F(\theta,\Theta) = \mathbb{E}_{\theta,\Theta}[f] = \mathbb{E}\left[f\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right]$. Then $F(\theta,\Theta) \leq f(\mathbf{m})$ and for each sequence (θ_n,Θ_n), $n = 1,2,\ldots$, such that $\lim_{n\to\infty}(I-\Theta_n)^{-1} = \lim_{n\to\infty}\Sigma_n = 0$ and $\lim_{n\to\infty}(I-\Theta_n)^{-1}\theta_n = \lim_{n\to\infty}\mu_n = \mathbf{m}$, we have $\lim_{n\to\infty}F(\theta_n,\Theta_n) = f(\mathbf{m})$. This remark has been used in Optimization when the function f is a *black box* that is when no analytic expression is known, but the function can be computed at each point $\mathbf{x}$, see for example [2]. In fact, the gradient of the relaxed function has components

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}F &= \mathrm{Cov}_{\theta,\Theta}(f,H_i),\\
\frac{\partial}{\partial\theta_{ij}}F &= \mathrm{Cov}_{\theta,\Theta}(f,H_{ij}), \quad i<j,\\
\frac{\partial}{\partial\theta_{ii}}F &= \frac{1}{2}\mathrm{Cov}_{\theta,\Theta}(f,H_{ii}),
\end{aligned}$$

so that the direction of steepest ascent at (θ,Θ) can be learned from a sample of $\mathrm{e}^{U_{\theta,\Theta}-\phi(\theta,\Theta)}\nu$ for example from sample covariances. This method does not require any smoothness in the original function and it is expected to have a better robustness vs local maxima than the ordinary gradient search because mean values of the function f are used. A reduction of dimensionality is obtained by considering submodels, for example Θ diagonal.

We note that the gradient of the relaxed function is related with the f, ∇f, Hess f as follows. We have $\mathrm{Cov}_{\theta,\Theta}(f,H_i) = \mathbb{E}_{\theta,\Theta}[fH_i] - \mathbb{E}_{\theta,\Theta}[f]\,\mathbb{E}_{\theta,\Theta}[H_i]$ and

$$\begin{aligned}
&\mathbb{E}_{\theta,\Theta}[fH_i] = \mathbb{E}\left[H_i f\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] = \mathbb{E}\left[\partial_i\left(f\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right)\right]\\
&= \mathbb{E}\left[(\partial_i f + f\partial_i U)\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] = \mathbb{E}_{\theta,\Theta}[\partial_i f] + \theta_i\mathbb{E}_{\theta,\Theta}[f] + \sum_j \theta_{ij}\mathbb{E}_{\theta,\Theta}[fH_j].
\end{aligned}$$

If $\mathbf{H}_1$ is the vector with components $H_1,\ldots,H_m$,

$$\mathbb{E}_{\theta,\Theta}[f\mathbf{H}_1] = \mathbb{E}_{\theta,\Theta}\left[(I-\Theta)^{-1}\nabla f\right] + \mathbb{E}_{\theta,\Theta}[f]\,(I-\Theta)^{-1}\theta,$$

so that $\nabla_\theta F = \mathbb{E}_{\theta,\Theta}\left[(I-\Theta)^{-1}\nabla f\right] = \mathbb{E}_{\mu,\Sigma}[\Sigma\nabla f]$.

In a similar way, $\mathrm{Cov}_{\theta,\Theta}(f,H_{ij}) = \mathbb{E}_{\theta,\Theta}[fH_{ij}] - \mathbb{E}_{\theta,\Theta}[f]\,\mathbb{E}_{\theta,\Theta}[H_{ij}]$ and

$$\begin{aligned}
\mathbb{E}_{\theta,\Theta}[fH_{ij}] &= \mathbb{E}\left[H_{ij}f\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right] = \mathbb{E}\left[\partial_i\partial_j\left(f\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right)\right]\\
&= \mathbb{E}\left[\partial_i\left[(\partial_j f + f\partial_j U_{\theta,\Theta})\mathrm{e}^{U_{\theta,\Theta}-\psi(\theta,\Theta)}\right]\right]\\
&= \mathbb{E}_{\theta,\Theta}\left[\partial_i\partial_j f + \partial_i f\partial_j U_{\theta,\Theta} + \partial_j f\partial_i U_{\theta,\Theta} + f\partial_i\partial_j U_{\theta,\Theta} + \partial_i U_{\theta,\Theta}\partial_j U_{\theta,\Theta}\right].
\end{aligned}$$

Now we can substitute in the equation above $\partial_i U_{\theta,\Theta} = \theta_i + \sum_h \theta_{ih} H_h$, $\partial_j U_{\theta,\Theta} = \theta_j + \sum_h \theta_{jh} H_h$, $\partial_i \partial_j U_{\theta,\Theta} = \theta_{ij}$ and

$$\begin{aligned}\partial_i U_{\theta,\Theta} \partial_j U_{\theta,\Theta} &= (\theta_i + \sum_h \theta_i h H_h)(\theta_j + \sum_k \theta_{ik} H_k) \\ &= \theta_i \theta_j + \sum_h (\theta_i \theta_{jh} + \theta_j \theta_{ih}) H_h + \sum_{h,k} \theta_{ih} \theta_{jk} H_h H_k \\ &= \theta_i \theta_j + \sum_h (\theta_i \theta_{jh} + \theta_j \theta_{ih}) H_h + 2 \sum_{h<k} \theta_{ih} \theta_{jk} H_{hk} + \sum_h \theta_{ih} \theta_{jh} (H_{hh} + 1),\end{aligned}$$

to obtain the required relation. We leave the rest of the computation to the reader.

4 Conclusions

We have presented the Gaussian model of Eq. (1) in a way that connects the parameter space with the sample space, see Eq. (6). This formalism has a number of advantages: *1*. The geometry of the vector bundles of the model is connected with statistical objects e.g., estimating functions. *2*. The use of Hermite polynomials as sufficient statistics allows to use properties of this class of orthogonal polynomials. *3*. The splitting of the log density into simple effects and interactions is reduced to computations on multivariate Hermite polynomials. *4*. The form of the exponential family in Eq. (1) suggests a possible generalization to a higher order expansion e.g., up to order 4, that is generalized Gaussian distributions.

Acknowledgements We thank the Referee for his valuable comments. The Author is supported by the de Castro Statistics Initiative, Collegio Carlo Alberto, and is a member of GNAMPA–INdAM. This paper is part of an ongoing collaboration with L. Malagò, Università di Milano e.g., [3, 4].

References

1. Amari, S., Nagaoka, H.: Methods of information geometry. American Mathematical Society, Providence, RI (2000).
2. Arnold, L., Auger, A., Hansen, N., Ollivier, Y.: Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles (2011v1; 2013v2). ArXiv:1106.3708
3. Malagò, L.: On the geometry of optimization based on the exponential family relaxation. Ph.D. thesis, Politecnico di Milano (2012)
4. Malagò, L., Matteucci, M., Pistone, G.: Stochastic natural gradient descent by estimation of empirical covariances. In: IEEE Congress on Evolutionary Computation, pp. 949–956. (2011).
5. Pistone, G.: Nonparametric information geometry. In: F. Nielsen, F. Barbaresco (eds.) Geometric Science of Information, 5–36. Springer-Verlag, Berlin Heidelberg (2013).
6. Sheena, Y.: Inference on the eigenvalues of the covariance matrix of a multivariate normal distribution–geometrical view– (2012). ArXiv1211.5733
7. Skovgaard, L.T.: A Riemannian geometry of the multivariate normal model. Scand. J. Statist. **11**(4), 211–223 (1984)