Geometric structure of statistical models

Zhengchao Wan

Peking University

June 10, 2016
Overview

• Statistical models
• The Fisher metric
• Chentsov’s theorem
• The $\alpha$-connection
In this lecture, **probability distributions** on a set $\mathcal{X}$ is represented in the following way. If $\mathcal{X}$ is a discrete set, then $p : \mathcal{X} \rightarrow R$ satisfies

$$p(x) \geq 0 (\forall x \in \mathcal{X}) \quad \text{and} \quad \sum_{x \in \mathcal{X}} p(x) = 1 \quad (1)$$

If $\mathcal{X} = R^n$, then $p : \mathcal{X} \rightarrow R$ satisfies

$$p(x) \geq 0 (\forall x \in \mathcal{X}) \quad \text{and} \quad \int p(x)dx = 1. \quad (2)$$

We shall use the integral expression to consider both cases.
Consider a family $S$ of probability distributions on $\mathcal{X}$. Suppose each element of $S$ may be parameterized using $n$ real-valued variables $[\xi^1, \cdots, \xi^n]$ so that

$$S = \{ p_\xi = p(x; \xi) | \xi = [\xi^1, \cdots, \xi^n] \in \Xi \},$$

where $\Xi$ is a subset of $R^n$ and the mapping $\xi \mapsto p_\xi$ is injective. We call such $S$ an $n$-dimensional **statistical model** on $\mathcal{X}$. 
There are several assumptions for the statistical model.

- $\Xi$ is open in $\mathbb{R}^n$.
- $\forall x \in \mathcal{X}$, the function $\xi \mapsto p(x; \xi)(\Xi \to \mathbb{R})$ is $C^\infty$, which allows $\partial_i p(x; \xi)$ to be defined ($\partial_i \triangleq \frac{\partial}{\partial \xi^i}$).
- The order of integration and differentiation may be freely rearranged. For example, we shall often use formulas such as

\[
\int \partial_i p(x; \xi) dx = \partial_i \int p(x; \xi) dx = \partial_i 1 = 0. \tag{4}
\]
Let $\text{supp}(p) \triangleq \{x|p(x) > 0\}$ (the support of $p$), then we only consider the case when $\text{supp}(p_\xi)$ is constant for $\xi$ and redefine $\mathcal{X}$ as $\text{supp}(p)$, i.e., $p(x; \xi) > 0$ holds for all $\xi \in \Xi$ and all $x \in \mathcal{X}$. This means that the model $S$ is a subset of

$$
\mathcal{P}(\mathcal{X}) \triangleq \left\{ p: \mathcal{X} \to R \bigg| p(x) > 0 (\forall x \in \mathcal{X}), \int p(x)dx = 1 \right\}.
$$

(5)
Examples

- Normal Distribution

\[ \mathcal{X} = \mathbb{R}, n = 2, \xi = [\mu, \sigma], \]
\[ \Xi = \{[\mu, \sigma]| -\infty < \mu < \infty, 0 < \sigma < \infty\} \]
\[ p(x; \xi) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \]

- Multivariate Normal Distribution

\[ \mathcal{X} = \mathbb{R}^k, n = k + \frac{k(k + 1)}{2}, \xi = [\mu, \Sigma] \]
\[ \Xi = \{[\mu, \Sigma]| \mu \in \mathbb{R}^k, \Sigma \in \mathbb{R}^{k \times k} : \text{positive definite}\} \]
\[ p(x; \xi) = (2\pi)^{-k/2}(\text{det}\Sigma)^{-1/2}\exp\left\{ -\frac{1}{2}(x - \mu)^t\Sigma^{-1}(x - \mu) \right\} \]
Examples

- **Poisson Distribution**

\[ \mathcal{X} = \{0, 1, 2, \cdots\}, \; n = 1, \; \Xi = \{\xi | \xi > 0\} \]

\[ p(x; \xi) = e^{-\xi} \frac{\xi^x}{x!} \]

- **\( \mathcal{P}(\mathcal{X}) \) for finite \( \mathcal{X} \)**

\[ \mathcal{X} = \{x_0, x_1, \cdots, x_n\}, \]
\[ \Xi = \{[\xi^1, \cdots, \xi^n] | \xi^i > 0 (\forall i), \sum_{i=1}^{n} \xi^i < 1\} \]

\[ p(x_i; \xi) = \begin{cases} \xi^i & (1 \leq i \leq n) \\ 1 - \sum_{i=1}^{n} \xi^i & (i = 0) \end{cases} \]
Given a statistical model $S = \{p_\xi | \xi \in \Xi\}$, the mapping $\varphi : S \to R^n$ defined by $\varphi(p_\xi) = \xi$ allows us to consider $\varphi = [\xi^i]$ as a coordinate system for $S$. Suppose we have a $C^\infty$ diffeomorphism $\psi$ from $\Xi$ to $\psi(\Xi)$, the latter being an open subset of $R^n$. Then if we use $\rho = \psi(\xi)$ as parameters, we obtain $S = \{p_{\psi^{-1}(\rho)} | \rho \in \psi(\Xi)\}$, which expresses the same family of probability distributions as $S = \{p_\xi\}$. We consider $S$ as a $C^\infty$ differential manifold or a statistical manifold.
Fisher information matrix

Let $S = \{p_\xi | \xi \in \Xi \}$ be an $n$-dimensional statistical model. Given a point $\xi$, the Fisher information matrix of $S$ at $\xi$ is the $n \times n$ matrix $G(\xi) = [g_{ij}(\xi)]$, where the $(i,j)^{th}$ element $g_{ij}(\xi)$ is defined by the equation below; in particular, when $n = 1$, we call this the Fisher information.

$$g_{ij}(\xi) \triangleq E_\xi[\partial_i l_\xi \partial_j l_\xi] = \int \partial_i l(x; \xi) \partial_j l(x; \xi) p(x; \xi) dx, \quad (6)$$

where $\partial_i \triangleq \frac{\partial}{\partial \xi_i}$,

$$l_\xi(x) = l(x; \xi) = \log p(x; \xi), \quad (7)$$

and $E_\xi$ denotes the expectation of $p_\xi$, $E_\xi[f] \triangleq \int f(x)p(x; \xi)dx$. 
Fisher information matrix

We assume in Equation (6) that $g_{ij}(\xi)$ is finite for all $\xi$ and $i, j$, and that $g_{ij} : \Xi \to R$ is $C^\infty$.

From Equation (4), we have

$$E_{\xi}[\partial_i l_{\xi}] = 0,$$

(8)

and thus by applying $\partial_j$ to both sides we have

$$g_{ij}(\xi) = -E_{\xi}[\partial_i \partial_j l_{\xi}].$$

(9)

Also we have another representation

$$g_{ij}(\xi) = 4 \int \partial_i \sqrt{p(x; \xi)} \partial_j \sqrt{p(x; \xi)} dx.$$ 

(10)
Fisher information matrix

The matrix \( G(\xi) \) is symmetric (\( g_{ij}(\xi) = g_{ji}(\xi) \)).

\[
c^t G(\xi) c = c^i c^j g_{ij}(\xi) = \int \left\{ \sum_{i=1}^{n} c^i \partial_i l(x; \xi) \right\}^2 p(x; \xi) dx \geq 0,
\]

(11)

for all \( n \)-dimensional vector \( c \), thus \( G \) is also positive semidefinite. We assume further that \( G \) is positive definite, which is equivalent to stating that \( \{ \partial_i l(x; \xi), \cdots, \partial_n l(x; \xi) \} \) or \( \{ \partial_i p_\xi, \cdots, \partial_n p_\xi \} \) are linearly independent functions on \( \mathcal{X} \).
Define the inner product of the natural basis of the coordinate system \([\xi^i]\) by \(g_{ij} = \langle \partial_i, \partial_j \rangle\).
This uniquely determines a Riemannian metric \(g = \langle , \rangle\). We call this the **Fisher metric** or **information metric**.
It’s easy to see that the Fisher metric is invariant over the choice of coordinate system. Indeed, we may write \(\langle X, Y \rangle_\xi = E_\xi[(X_I)(Y_I)]\) for all tangent vectors \(X, Y \in T_\xi(S)\).
Suppose $F$ is a measurable map from $\mathcal{X}$ to a measure space $\mathcal{Y}$. $\mathcal{Y} = F(\mathcal{X})$. $P_Y = P_X F^{-1}$.

$$(\mathcal{X}, \mathcal{S}, P_X) \xrightarrow{F} (\mathcal{Y}, \mathcal{G}, P_Y)$$

We assume that $P_X$ and $P_Y$ are both dominated by the $L$-measure with density $p(x), q(y)$, for $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean or discreet spaces.
Conditional probability

A conditional probability $P(A|F = y) : \mathcal{S} \times \mathcal{Y} \to R$ of the transform $F$ satisfies:

- $\forall y \in \mathcal{Y}, P(\cdot|F = y)$ is a measure on $\mathcal{X}$;
- $\forall A \in \mathcal{S}, P(A|F = \cdot)$ is a measurable function on $(\mathcal{Y}, \mathcal{G})$;
- It’s the unique function (under the zero-measure meaning) that satisfies:

$$
\int_{F^{-1}B} 1_A dP_X = \int_B P(A|F = y) dP_Y, \quad \forall B \in \mathcal{G} \tag{12}
$$

Particularly, if $F$ is discrete, with image $\{a_1, a_2, \cdots\}$ and distribution $\{p_n = P_X(F = a_n) > 0, n = 1, 2, \cdots\}$, then we have

$$
P(A|F = a_n) = \frac{P_X(A \cap \{F = a_n\})}{P_X(F = a_n)}. \tag{13}
$$
Sufficient statistic

Given the distribution $p(x; \xi)$, this determines the distribution $q(y; \xi)$ on $\mathcal{Y}$. Then rewrite Equation (12) as

$$\int_{A \cap F^{-1}(B)} p(x; \xi) dx = \int_B P(A|y; \xi) q(y; \xi) dy$$  \hspace{1cm} (14)

If the value of $P(A|y; \xi)$ does not depend on $\xi$ for all $A$ and $y$, then we say that $F$ is a **sufficient statistic** for the model $S$. 
In addition, letting

\[ r(x; \xi) = \frac{p(x; \xi)}{q(F(x); \xi)}, \]

\[ p(x|y; \xi) = r(x; \xi)\delta_{F(x)}(y), \quad (15) \]

\[ P(A|y; \xi) = \int_A p(x|y; \xi) dx, \quad A \in \mathcal{S} \]

Then \( P(A|y; \xi) \) satisfies the condition of conditional probability. Thus, that \( F \) is sufficient statistic is equivalent to that \( r(x; \xi) \) does not depend on \( \xi \) for all \( x \).
**Information loss**

**Theorem**

The Fisher information matrix $G_F(\xi) = [g_{ij}^F(\xi)]$ of the induced model $S_F \triangleq \{q(y, \xi)\}$ satisfies $G_F(\xi) \leq G(\xi)$, where $G(\xi) = [g_{ij}(\xi)]$ is the Fisher information matrix of the original model $S$, in the sense that $\Delta G(\xi) \triangleq G(\xi) - G_F(\xi)$ is positive semidefinite. A necessary and sufficient condition for the equality $G_F(\xi) = G(\xi)$ to identically hold is that $F$ is a sufficient statistic for $S$. In general, the information loss caused by summarizing the data $x$ into $y=F(x)$ is given by

$$\Delta g_{ij}(\xi) = E_\xi[\partial_i \log r(X; \xi) \partial_j \log r(X; \xi)]$$

$$= E_\xi[\text{Cov}_\xi[\partial_i l(X; \xi), \partial_j l(X; \xi) | Y]],$$

(16)

where $E_\xi[\text{Cov}_\xi[\cdot, \cdot | Y]] = \int \text{Cov}_\xi[\cdot, \cdot | y]q(y, \xi)dy$, and $\text{Cov}_\xi[\cdot, \cdot | y]$ for a fixed $y$ denotes the covariance with respect to the conditional distribution $p(x | y; \xi)$. 
Markov kernel

Given two measure spaces \((X, \mathcal{I}, \mu)\) and \((Y, \mathcal{G}, \nu)\), suppose \((X, \mathcal{I}, \mu)\) has a probability measure \(P\) dominated by \(\mu\) with a density \(p\). Then a measurable function \(\kappa : X \times Y \rightarrow R\) (written as \(\kappa(y|x)\)) with respect to the product \(\sigma\)-algebra \(\mathcal{I} \otimes \mathcal{G}\) is the **Markov kernel** if it satisfies:

- \(\kappa \geq 0\)

- Let \(k(x, B) = \int_B \kappa(y|x)\nu(dy), \forall x \in X, B \in \mathcal{G}\). Then \(k(\cdot, B)\) is \(\mathcal{I}\)-measurable and \(k(x, \cdot)\) is a probability measure on \((Y, \mathcal{G})\).
Markov kernel

Now define another probability measure induced from $P$ on $(\mathcal{Y}, \mathcal{G})$:

$$Q(B) = \int k(x, B) dP = \int k(x, B)p(x)\mu(dx),$$  \hspace{1cm} (17)

which has a density

$$q(y) \triangleq \frac{dQ}{d\nu} = \int k(y|x)p(x)\mu(dx).$$  \hspace{1cm} (18)
Monotonicity

Given the Markov kernel $\kappa$, then we have

$$G_{\kappa}(\xi) \leq G(\xi), \quad (19)$$

where $G_{\kappa}(\xi)$ is the Fisher information matrix of the induced model: $q(y; \xi) = \int \kappa(y|x)p(x; \xi)dx$.

The previous case for a deterministic mapping $F$ corresponds to $\kappa(y|x) = \delta_{F(x)}y$. 
Chain rule

The equation following from the above relation

$$G(\xi) = G_\kappa(\xi) + \Delta G(\xi)$$  \hspace{1cm} (20)

is called the **chain rule**.

As a special case of the chain rule, the **additivity**

$$G_{12}(\xi) = G_1(\xi) + G_2(\xi)$$  \hspace{1cm} (21)

holds for a product model: $p_{12}(x_1, x_2; \xi) = p_1(x_1; \xi)p_2(x_2; \xi)$. 
Convexity

Given two models \( \{ p_1(x; \xi) \} \) and \( \{ p_2(x; \xi) \} \) having common sample space \( \mathcal{X} \) and parameter space \( \Xi \), we have the following convexity:

\[
G_\lambda(\xi) \leq \lambda G_1(\xi) + (1 - \lambda) G_2(\xi), \quad 0 \leq \forall \lambda \leq 1,
\]

(22)

where \( G_1(\xi) \), \( G_2(\xi) \) and \( G_\lambda(\xi) \) are the Fisher information matrices of \( \{ p_1(x; \xi) \} \) and \( \{ p_2(x; \xi) \} \) and \( \{ \lambda p_1(x; \xi) + (1 - \lambda) p_2(x; \xi) \} \), respectively.
Let $S = \{p_\xi\}$ be an $n$-dimensional model, and consider the function $\Gamma_{ij,k}^{(\alpha)}$:

$$
\Gamma_{ij,k}^{(\alpha)} \triangleq E_\xi[(\partial_i \partial_j l_\xi + \frac{1-\alpha}{2} \partial_i l_\xi \partial_j l_\xi)(\partial_k l_\xi)],
$$

(23)

where $\alpha$ is some arbitrary real number. Hence we have an affine connection $\nabla^{(\alpha)}$ on $S$ defined by

$$
\langle \nabla_{\partial_i} \partial_j, \partial_k \rangle = \Gamma_{ij,k}^{(\alpha)}
$$

(24)

where $g = \langle , \rangle$ is the Fisher metric. We call this the $\alpha$-connection.
Properties of the $\alpha$-connection

- The $\alpha$-connection is symmetric.
- The relation between the $\alpha$-connection and the $\beta$-connection

\[
\Gamma_{ij,k}^{(\beta)} = \Gamma_{ij,k}^{(\alpha)} + \frac{\alpha - \beta}{2} T_{ijk},
\]

(25)

where $(T_{ijk})_\xi \triangleq E_\xi[\partial_i l_\xi \partial_j l_\xi \partial_k l_\xi]$.

- \[
\nabla^{(\alpha)} = (1 - \alpha)\nabla^{(0)} + \alpha \nabla^{(1)} \\
= \frac{1 + \alpha}{2} \nabla^{(1)} + \frac{1 - \alpha}{2} \nabla^{(-1)}.
\]

(26)
Properties of the $\alpha$-connection

- For a submanifold $M$ of $S$, the $\alpha$-connection on $M$ is the projection with respect to $g$ of the $\alpha$-connection on $S$.
- By taking the partial derivative of $g_{ij}$, we obtain

$$\partial_k g_{ij} = \Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i},$$

which leads to:

**Theorem**

*The 0-connection is the Riemannian connection of the Fisher metric.*
Exponential family

Definition (Exponential family)

If an n-dimensional model \( S = \{ p_\theta | \theta \in \Theta \} \) can be expressed in terms of functions \( \{ C, F_1, \cdots, F_n \} \) on \( \mathcal{X} \) and a function \( \psi \) on \( \Theta \) as

\[
p(x; \theta) = \exp \left[ C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) \right], \quad (28)
\]

then we say that \( S \) is an exponential family, and that the \( [\theta^i] \) are its canonical parameters.

From the normalization condition \( \int p(x; \theta) dx = 1 \) we obtain

\[
\psi(\theta) = \log \int \exp \left[ C(x) + \sum_{i=1}^{n} \theta^i F_i(x) \right] dx. \quad (29)
\]
Exponential family

The parametrization $\theta \mapsto p_\theta$ is one-to-one iff the $n + 1$ functions $\{F_1, \cdots, F_n, 1\}$ are linearly independent, where 1 denotes the constant function which identically takes the value 1. From now on, we always assume the linear independence for the exponential families.
Examples

- **Normal Distribution**

  \[ C(x) = 0, \quad F_1(x) = x, \quad F_2(x) = x^2, \quad \theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2} \]

  \[ \psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma}) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2} \log(-\frac{\pi}{\theta^2}) \]

- **Poisson Distribution**

  \[ C(x) = -\log x!, \quad F(x) = x, \quad \theta = \log \xi \]

  \[ \psi(\theta) = \xi = \exp \theta \]
Examples

- $\mathcal{P}(\mathcal{X})$ for finite $\mathcal{X}$

  $$C(x) = 0, \quad F_i(x) = \begin{cases} 1 & (x = x_i) \\ 0 & (x \neq x_i) \end{cases}$$

  $$\theta^i = \log \frac{p(x_i)}{p(x_0)} = \log \frac{\xi^i}{1 - \sum_{j=1}^{n} \xi^j}, \quad (i = 1, \cdots, n)$$

  $$\psi(\theta) = -\log p(\theta) = -\log (1 - \sum_{i=1}^{n} \xi^i) = \log (1 + \sum_{i=1}^{n} \exp \theta^i)$$
Exponential connection

Letting $\partial_i = \frac{\partial}{\partial \theta^i}$, we may obtain

$$\partial_i l(x; \theta) = F_i(x) - \partial_i \psi(\theta) \quad (30)$$

$$\partial_i \partial_j l(x; \theta) = -\partial_i \partial_j \psi(\theta). \quad (31)$$

Hence we have $\Gamma_{ij,k}^{(1)} = -\partial_i \partial_j \psi(\theta) E_\theta[\partial_k l_\theta]$, which is 0 from Equation (8). Therefore, $[\theta^i]$ is a 1-affine coordinate system, and $S$ is 1-flat. We call $\nabla^{(1)}$ the exponential connection, or the e-connection, and shall write $\nabla^{(1)} = \nabla^{(e)}$. 
Definition (Mixture family)

If an $n$-dimensional model $S = \{p_\theta | \theta \in \Theta \}$ can be expressed in terms of functions $\{C, F_1, \cdots, F_n\}$ on $\mathcal{X}$ as

$$p(x; \theta) = C(x) + \sum_{i=1}^{n} \theta^i F_i(x), \quad (32)$$

then we say that $S$ is an mixture family, and that the $[\theta^i]$ are its mixture parameters.
Examples

Given $n + 1$ distributions $\{p_0, p_1, \cdots, p_n\}$, we have a mixture family:

$$p(x; \theta) = \sum_{i=1}^{n} \theta^i p_i(x) + (1 - \sum_{i=1}^{n} \theta^i)p_0(x)$$

$$= p_0(x) + \sum_{i=1}^{n} \theta^i \{p_i(x) - p_0(x)\}$$

(33)

where $[\theta^i]$ are subject to $\theta^i > 0$ and $\sum_i \theta^i < 1$. $\mathcal{P}(\mathcal{X})$ itself is a mixture family when $\mathcal{X}$ is finite, for $\mathcal{P}(\{x_0, \cdots, x_n\})$ may be expressed in the form above by letting $p_i(x_j) = \delta_{ij}$. 
Mixture connection

For a mixture family, we have

\[ \partial_i l(x; \theta) = \frac{F_i(x)}{p(x; \theta)} \quad \text{and} \quad \partial_i \partial_j l(x; \theta) = -\frac{F_i(x)F_j(x)}{p(x; \theta)^2}, \quad (34) \]

from which \( \partial_i \partial_j l + \partial_i l \partial_j l = 0 \), and hence \( \Gamma^{(-1)}_{ij,k} = 0 \). Therefore \( [\theta^i] \) is a \((-1)\)-affine coordinate system, and \( S \) is \((-1)\)-flat. We call \( \nabla^{(-1)} \) the mixture connection or the m-connection, and we write \( \nabla^{(-1)} = \nabla^{(m)} \).
Theorem

Let $S$ be an exponential family (a mixture family, respectively) and $M$ be a submanifold of $S$. Then $M$ is an exponential family (a mixture family) iff $M$ is $e$-autoparallel ($m$-autoparallel) in $S$. 
Proof.

We only prove that if $S$ and $M$ are e-families then $M$ is e-autoparallel in $S$. Let $S = \{p(x; \theta)\}$, $M = \{q(x; u)\}$ be given by

$$p(x; \theta) = \exp \left[ C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) \right],$$

$$q(x; u) = p(x; \theta(u)) = \exp \left[ D(x) + \sum_{a=1}^{m} u^a G_a(x) - \varphi(u) \right].$$

Then we have

$$G_a(x) - \partial_a \varphi(u) = \partial_a \log q(x; u)$$

$$= (\partial_a \theta^i) u \partial_i \log p(x; \theta(u))$$

$$= (\partial_a \theta^i) u \{F_i(x) - \partial_i \psi(\theta(u))\},$$
Proof.
and hence 
\[(\partial_a \theta^i)_u F_i(x) + \lambda_a(u) = G_a(x),\]
where $\lambda_a(u)$ is constant of $x$. Since $G_a(x)$ does not depend on $u$ and since $\{F_1, \cdots, F_n, 1\}$ are assumed to be linearly independent, we see that $(\partial_a \theta^i)_u$ is constant of $u$. This, combined Theorem in Chapter 1, implies that $M$ is e-autoparallel in $S$. \qed
Infinite-dimensional manifold

If $\mathcal{X}$ is finite, we know $\mathcal{P}(\mathcal{X})$ is both $e$- and $m$-family. For the continuous case, given two distributions $p_1(x), p_2(x)$, the $e$-family connecting them is written as

$$p_{exp}(x, t) = \exp\{(1 - t)\log p_1(x) + t\log p_2(x) - \psi(t)\}, \quad (35)$$

while the $m$-family connecting them as

$$p_{mix}(x, t) = (1 - t)p_1(x) + tp_2(x). \quad (36)$$

Then, we can regard this infinite-dimensional $\mathcal{P}(\mathcal{X})$ as an $e$- and a $m$-family.
Example of $\alpha$-flat model

Given a smooth probability density function $q$ on $\mathbb{R}$, let $q^{(k)}$ be the $k$th i.i.d. extension; i.e., for $y = (y_1, \ldots, y_k)^t$, $q^{(k)}(y) = q(y_1) \cdots q(y_k)$. For a regular matrix $A \in \mathbb{R}^{k \times k}$ and a vector $\mu = (\mu_1, \cdots, \mu_k)^t \in \mathbb{R}^k$, define the probability density function $p_{A,\mu}$ on $\mathbb{R}^k$ by

$$p(x; A, \mu) = q^{(k)}(A^{-1}(x - \mu))/|\text{det}A|,$$

which gives the distribution for $AY + \mu$ when $Y$ is supposed to distribute according to $q^{(k)}(y)$. For instance, $q \sim \mathcal{N}(0, 1)$, we obtain

$$p(x; A, \mu) = (2\pi)^{-k/2}(\text{det}\Sigma)^{-1/2}\exp\{-\frac{1}{2}(x - \mu)^t\Sigma^{-1}(x - \mu)\},$$

where $\Sigma \triangleq AA^t$. 
Example of $\alpha$-flat model

Fix $q$ and $A$ arbitrarily and consider the model $S \triangleq \{p_{A,\mu} | \mu \in R^k\}$. This model is not in general an e-family nor a m-family, but is always $\alpha$-flat for all $\alpha$ and is a Euclidean space for the Fisher metric. This can be explained by the fact that $S$ is essentially the direct product, including its $\alpha$-connections and Fisher metric, of $k$ copies of the 1-dimensional statistical model $\{q(y - \nu | \nu \in R)\}$ on which affine aonnections are always flat.
Invariance of Fisher metric

Let \( S = \{p(x; \xi)\} \) be a model on \( X \) and \( F : X \to Y \) be some mapping, which induces a model \( S_F = \{q(y; \xi)\} \) on \( Y \). If \( F \) is a sufficient statistic for \( S \), then \( \partial_i \log p(x; \xi) = \partial_i \log q(F(x); \xi) \), and hence \( g_{ij}, \Gamma^{(\alpha)}_{ij,k} \) are the same on both \( S \) and \( S_F \). We refer to this as the invariance of Fisher metric and the \( \alpha \)-connection with respect to \( F \).
Consider a manifold with finite points. Let $X_n \triangleq \{0, 1, \cdots, n\}$ and $P_n \triangleq P(X_n)$. Suppose that we have a sequence $\{(g_n, \nabla_n)\}_{n=1}^{\infty}$ on $P_n$ for each $n$.

**Theorem (Chentsov)**

Assume that $\{(g_n, \nabla_n)\}_{n=1}^{\infty}$ is invariant with respect to sufficient statistics; i.e., for all $n \geq m$, $S \subset P_n$, and $F : X_n \rightarrow X_m$ s.t. $F$ is a sufficient statistic for $S$, the induced metrics and connections on $S$ and $S_F$ are assumed to be invariant. Then there exist a positive real number $c$ and a real number $\alpha$ s.t., for all $n$, $g_n$ coincides with the Fisher metric on $P_n$ scaled by a factor of $c$, and $\nabla_n$ coincides with the $\alpha$-connection on $P_n$. 