Information Geometric Structure on Positive Definite Matrices and its Applications

Atsumi Ohara　　University of Fukui

2013 March 2-3 at Hokkaido University

ミニワークショップ「統計多様体の幾何学とその周辺（4）」

Outline

1. Introduction
2. Standard information geometry on positive definite matrices
3. Extension via the other potentials (Bregman divergence)
   - Joint work with S. Eguchi (ISM)
4. Conclusions

1. Introduction

$PD(n, \mathbf{R})$: the set of positive definite real symmetric matrices

related to branches in math.

- Riemannian symmetric space
- Symmetric cone (Jordan algebra)
- Symplectic geom. (Siegel-Poincare)
- Information, Hessian geom.
- affine, Kahler, …,C-H, B-T,…?

1. Introduction

$PD(n, \mathbf{R})$: the set of positive definite real symmetric matrices

related to branches in applications

- matrix (in)eq. (Lyapunov,Riccati,…)
- mathematical programming (SDP)
- Statistics, signal processing, time series analysis (Gaussian, Covariance matrix)
- …

Our interests

- stable matrices and IG [O,Amari Kybernetika93]
- standard IG [O,Suda,Amari LAA96]
  - dual conections and Jordan alg. [O,Uohashi Positivity04]
  - means on sym. cones [O IEOT04]
- complexity analysis of IPM
  [O 統計数理98], [Kakihara,O,Tsuchiya JOTA?]
- deformed IG [O,Eguchi ISM_RM05]
- update formula for Q-Newton [Kanamori,O OMS13]

Information geometry on $\mathcal{M}$

Dualistic geometric structure

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z)$$

X, Y and Z : arbitrary vector fields on $\mathcal{M}$

g :Riemannian metric

∇ , ∇^* :a pair of dual affine connections

A simple way to introduce a dualistic structure (1)

- $\mathcal{M}$: open domain in $\mathbf{R}^n$

  φ :strongly convex on $\mathcal{M}$ (i.e., positive definite Hessian mtx.) Cf. Hessian geometry

- Riemannian metric

$$g_{ij} = \frac{\partial^2 \varphi}{\partial x^i \partial x^j}$$

- Dual affine connections

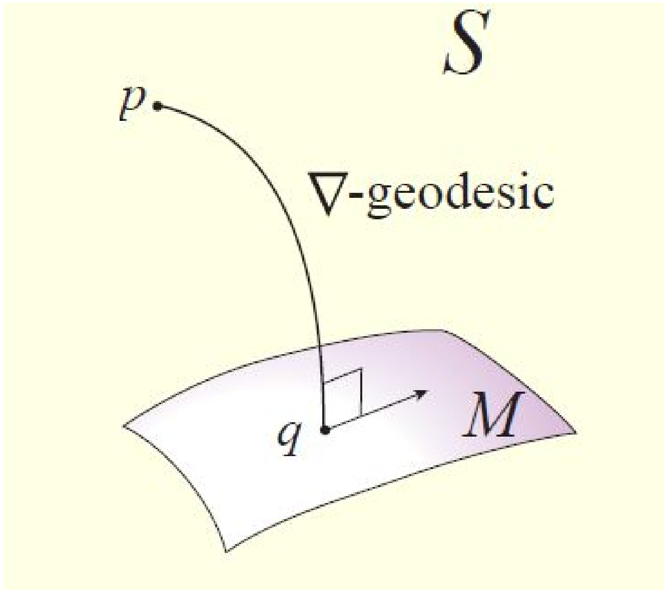
$$\Gamma_{ijk} = 0, \qquad \Gamma^*_{ijk} = \frac{\partial^3 \varphi}{\partial x^i \partial x^j \partial x^k}$$

A simple way to introduce a dualistic structure (2)

- divergence

$$D(p,q) = \varphi(x(p)) - \varphi(x(q)) - \sum_{i=1}^{n} \frac{\partial \varphi}{\partial x^i}(x(q))\{x^i(p) - x^i(q)\}$$

- projection
  - MLE, MaxEnt and so on
- Pythagorean relations



2. Standard IG on $PD(n, \mathbf{R})$

- $PD(n, \mathbf{R})$: the set of positive definite real symmetric matrices

- logarithmic characteristic func. on $PD(n, \mathbf{R})$

$$\varphi(P) = -\log \det P, \quad P \in PD(n; \mathrm{R})$$

- The standard case -

-log det *P* appears as

- Semidefinite Programming (SDP)

  self-concordant barrier function
- Multivariate Analysis (Gaussian dist.)

  log-likelihood function

  (structured covariance matrix estimation)
- Symmetric cone: log characteristic function
- Information geometry on $PD(n, \mathbf{R})$

  potential function

Standard dualistic geometric structure on $PD(n, \mathbf{R})$ (1) [O,Suda,Amari LAA96]

- $Sym(n; \mathbf{R})$: the set of *n* by *n* real symmetric matrix
  vec. sp. of dimension $N(= n(n+1)/2)$
- $\{E_i\}_{i=1}^{N}$:arbitrary set of basis matrices
- (primal) affine coordinate system

$$Sym(n; \mathbf{R}) \ni X = \sum_{i=1}^{N} x^i E_i$$

- Identification

$$T_P PD(n) \ni (\partial / \partial x^i)_P \equiv E_i \in Sym(n)$$

Standard dualistic geometric structure on $PD(n, \mathbf{R})$ (2)

$\varphi(P)$ plays a role of potential function

g: Riemannian metric (Fisher for Gaussian)

$$g(X,Y) = \mathrm{tr}(P^{-1}XP^{-1}Y)$$

∇, ∇^* :dual affine connections

$$\left(\nabla_{\partial_i}\partial_j\right)_P \equiv 0, \ \left(\nabla^*_{\partial_i}\partial_j\right)_P \equiv -E_iP^{-1}E_j - E_jP^{-1}E_i$$

Jordan product (mutation)

Properties ➔ symmetric cones

- $GL(n, \mathbf{R})$ -invariant
- $\iota : P \mapsto P^{-1}$: isometric involution
- dual affine coordinate system (Legendre tfm.)

$$P^* := -P^{-1} = \sum_{i=1}^{N} y_i E^i, \ \langle E_i, E^j \rangle = \mathrm{tr}(E_i E^j) = \delta_i^j$$

- divergence

$$D(P, Q) = \mathrm{tr}(PQ^{-1}) - \log\det(PQ^{-1}) - n$$

- self-dual

Invariance of the structure

- Automorphism group, i.e., congruent transformation: $\tau_G P = GPG^T, G \in GL(n, \mathbf{R})$,

  the differential: $(\tau_G)_* X = GXG^T$

Ex) Riemannian metric

$$g_{P'}(X', Y') = g_P(X, Y)$$

$$P' = \tau_G P, X' = \tau_{G*} X \ \ and \ Y' = \tau_{G*} Y$$

Connections represented by Jordan product [Uohashi O Positivity04]

- Recall the dual affine connections:

$$\left(\nabla_{\partial_i}\partial_j\right)_P \equiv 0, \ \left(\nabla^*_{\partial_i}\partial_j\right)_P \equiv -E_iP^{-1}E_j - E_jP^{-1}E_i$$

Hence, $\left(\nabla^*_{\partial_i}\partial_j\right)_I \equiv -E_iE_j - E_jE_i = -2E_i * E_j$

- By the invariance, it follows that

$$\left(\nabla^*_{\partial_i}\partial_j\right)_P \equiv -2(\tau_Q)_*^{-1}\left[\left((\tau_Q)_*E_i\right) * \left((\tau_Q)_*E_j\right)\right]$$

$$Q := P^{-1/2}$$

- Rem. the Levi-Civita is

$$\left(\hat{\nabla}_{\partial_i}\partial_j\right)_P \equiv -\frac{1}{2}(E_iP^{-1}E_j + E_jP^{-1}E_i)$$

Doubly autoparallel submanifold

- <u>Def.</u> Submanifold $\mathcal{L}_{\mathrm{DA}} \subset PD(n;\mathbf{R})$ is doubly autoparallel when it is both ∇ - and ∇^*-autoparallel,

equivalently,

$\mathcal{L}_{\mathrm{DA}} \subset PD(n;\mathbf{R})$ is both linearly and inverse-linearly constrained.

Linearly constrained ➔ ∇ -autoparallel
Inverse-linearly ➔ ∇^* -autoparallel

Both Linearly and Inverse-linearly Constrained matrices $\mathcal{L}_{\mathsf{DA}}$ in $PD(n)$

Given $E_0, \cdots E_m, F^0, \cdots, F^m \in Sym(n)$,

$\{E_i\}_{i=1}^m, \{F^i\}_{i=1}^m$: linearly independent

$$P \in \mathcal{L}_{\mathsf{DA}} \Leftrightarrow \begin{cases} P = E_0 + \sum_{i=1}^m x^i E_i \geq O, \ \exists x \in \mathbf{R}^m \\ P^{-1} = F^0 + \sum_{i=1}^m y_i F^i \geq O, \ \exists y \in \mathbf{R}^m \end{cases}$$

Set $\mathcal{V} = \mathrm{span}\{E_i\}_{i=1}^{m}$.

conditions for Doubly Autoparallelism

Let $\mathcal{L}$ be linearly constrained in $PD(n)$.

The followings are equivalent:

i) $\mathcal{L}$ is ∇^*-autoparallel (hence, **D.A.**),

ii) ∇^*-imbedding curvature H^* vanishes on $\mathcal{L}$

iii) $E_i P^{-1} E_j + E_j P^{-1} E_i \in \mathcal{V}, \quad \forall i, j, \forall P \in \mathcal{L}$

ii) and iii) are difficult to check for all $P \in \mathcal{L}$

Doubly autoparallelism (special case)

- Jordan product for $Sym(n)$

$$X * Y = (XY + YX)/2$$

Cf. Malley 94

Let both E_0 and I are in $\mathcal{V} = \text{span}\{E_i\}_{i=1}^m$.

The followings are equivalent:

i) $\mathcal{L}$ is D. A.

ii) $\mathcal{V}$ is Jordan subalgebra of $Sym(n)$

$$E_i * E_j \in \mathcal{V}, \quad \forall i, j \quad \textbf{(easy to check)}$$

Rem. $\mathcal{L} = PD(n) \cap \mathcal{V}$ is a subcone in $PD(n)$

Doubly autoparallelism - Examples – (1)

1) Doubly symmetric matrices:

symmetric w.r.t. both main and anti-main diagonal entries

2) Matrices with the prescribed eigenvectors

— Ex. circulant matrices etc.

These examples are Jordan subalgebras.

Doubly autoparallelism - Examples - (2)

4) Let $\mathcal{JS}$ be any Jordan subalgebra in $Sym(n)$

$$\mathcal{A}_2 := \{A - BXB^T | X \in \mathcal{JS},\ \det A \neq 0,\ B^T A^{-1} B \in \mathcal{JS}\},$$

Then $\mathcal{L}_2 := \mathcal{A}_2 \cap PD(n)$ is doubly autoparallel.

$\mathcal{A}_1$ and $\mathcal{A}_2$ are generally affine subspace,

hence, not Jordan subalgebras

Applications of DA

- Nearness, matrix approximation,
  - $GL(n)$-invariance, convex optimization
- Semidefinite Programming
  - If a feasible region is DA, an explicit formula for the optimal solution exists.
- Maximum likelihood estimation of structured covariance matrix
  - GGM, Factor analysis, signal processing (AR model)

MLE of str. cov. matrix (1)

- n samples of random variable z

$$z_i \sim N(0, P), \quad P \in \mathcal{S} \subset PD(n)$$

$\mathcal{S}$: **linearly constrained in many cases ($\mathcal{S} = \mathcal{L}$),**

$\longrightarrow$ **signal processing, factor analysis etc.**

- **main term of logarithmic likelihood function**

$$h(P) = -\log\det P - \mathrm{tr}(P^{-1}S), \quad S = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^T.$$

ML estimation of $P \Leftrightarrow \max h(P)$, **s.t.** $P \in \mathcal{L}$

$\Leftrightarrow \min D(S, P)$, s.t. $P \in \mathcal{L}$

MLE of str. cov. matrix (2)

$$h(P) \to \max \ \text{s.t.} \ P \in \mathcal{L}$$

$$\Updownarrow$$

$$\tilde{h}(Q) = -\log\det Q + \operatorname{tr}(QS) \to \min, \ \text{s.t.} \ Q^{-1} = P \in \mathcal{L}$$

- If $\mathcal{L}$ is also inverse-linearly constrained, i.e., $\mathcal{L}$ is DA, then MLE is a convex optimization problem with a solution formula:

$$P = E_0 + \sum_{i=1}^{m} x^i E_i,$$

$$x = A^{-1}b, \quad a_j^i = \operatorname{tr}(E_j F^i), b^i = \operatorname{tr}(E_0 - S)F^i$$

MLE of str. cov. matrix (3)

Furthermore,

- Imbedding method with the EM algorithm

[Rubin & Szatrovski 82], [Malley 94]

$p \times p$ Toeplitz mtxs. $\rightarrow q \times q$ circulant mtxs. $\exists q > p$

Ex. $p = 3,\ q = 4$

$$T = \begin{pmatrix} y_0 & y_1 & y_2 \\ y_1 & y_0 & y_1 \\ y_2 & y_1 & y_0 \end{pmatrix}, \quad C = \left(\begin{array}{ccc|c} y_0 & y_1 & y_2 & y_1 \\ y_1 & y_0 & y_1 & y_2 \\ y_2 & y_1 & y_0 & y_1 \\ \hline y_1 & y_2 & y_1 & y_0 \end{array}\right).$$

MLE of str. cov. matrix (4)

T: covariance of imcomplete data

C: covariance of complete data

— S: sample covariance for T (not Toeplitz)

— $\hat{C}$: estimate for C (circulant)

— $\tilde{S}$: expected value of C (not circulant)

Initialize $\hat{C}$.

E-step: Compute $\tilde{S}$ from S and $\hat{C}$

M-step: Compute new $\hat{C}$ from $\tilde{S}$

MLE of str. cov. matrix (5)

- E-step: Explicit formula for simple imbedding (e.g., upper-left corner etc)

- M-step: reduces to solving a linear equation if the structure of C is DA.

3. Extension via the other potentials (Bregman divergence)

[O,Eguchi ISM_RM05]

- The other convex potentials

V-potential functions

$$\varphi^{(V)}(P) = V(\det P)$$

- Study their different and common geometric natures
- Application to multivariate statistics?

Contents

- V-potential function
- Dualistic geometry on $PD(n, \mathbf{R})$
- Foliated Structure
- Decomposition of divergence
- Application to statistics

  geometry of a family of multivariate
  elliptic distributions

Def. V-potential function

$$\varphi^{(V)}(P) = V(\det P), \qquad V(s): \mathbf{R}_+ \to \mathbf{R}$$

-The standard case:

$$V(s) = -\log s \Rightarrow \varphi(P) = -\log\det P$$

Characteristic function on $PD(n, \mathbf{R})$

(strongly convex)

Def.

$$\nu_i(s) = \frac{d\nu_{i-1}(s)}{ds}s, \quad i = 1, 2, \cdots, \quad \text{where } \nu_0(s) = V(s)$$

Rem. The standard case:

$$\nu_1(s) = -1, \nu_k(s) = 0, \quad k \geq 2$$

Prop. (Strong convexity condition)

The Hessian matrix of the V-potential is positive definite on $PD(n, \mathbf{R})$ if and only if

For $\forall s > 0$,

$$\text{i)}\nu_1(s) < 0, \quad \text{ii)}\beta^{(V)}(s) < \frac{1}{n}, \text{ where } \beta^{(V)}(s) = \frac{\nu_2(s)}{\nu_1(s)}$$

Prop.

When two conditions in Prop.1 hold, Riemannian metric derived from the V-potential is

$$g_P^{(V)}(X,Y)$$
$$= -\nu_1(\det P)\,\mathrm{tr}(P^{-1}XP^{-1}Y) + \nu_2(\det P)\,\mathrm{tr}(P^{-1}X)\,\mathrm{tr}(P^{-1}Y)$$

Here,

X,Y : vector field ~ symmetric matrix-valued func.

Rem. The standard case:

$$g_P^{(V)}(X,Y) = \mathrm{tr}(P^{-1}XP^{-1}Y)$$

Prop. (affine connections)

Let ∇ be the canonical flat connection on $PD(n,\mathbf{R})$. Then the V-potential defines the following dual connection ${}^*\nabla^{(V)}$ with respect to $g^{(V)}$:

$$\left({}^*\nabla^{(V)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)_P = -E_i P^{-1} E_j - E_j P^{-1} E_i - \Phi(E_i, E_j, P) - \Phi^{\perp}(E_i, E_j, P),$$

$$\Phi(X,Y,P) = \frac{\nu_2(s)\,\mathrm{tr}(P^{-1}X)}{\nu_1(s)} Y + \frac{\nu_2(s)\,\mathrm{tr}(P^{-1}Y)}{\nu_1(s)} X,$$

$$\Phi^{\perp}(X,Y,P)$$

$$= \frac{(\nu_3(s)\nu_1(s) - 2\nu_2^2(s))\,\mathrm{tr}(P^{-1}X)\,\mathrm{tr}(P^{-1}Y) + \nu_2(s)\nu_1(s)\mathrm{tr}(P^{-1}XP^{-1}Y)}{\nu_1(s)(\nu_1(s) - n\nu_2(s))} P$$

Rem. the standard case:

$$\left({}^{*}\nabla^{(V)}_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \right)_P = -E_i P^{-1} E_j - E_j P^{-1} E_i$$

"mutation" of the Jordan product of E_i and E_j

divergence function

Divergence function derived from $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$

$$D^{(V)}(P,Q) = \varphi^{(V)}(P) + \varphi^{(V)*}(Q^*) - \langle Q^*, P\rangle$$
$$= V(\det P) - V(\det Q) + \langle Q^*, Q - P\rangle.$$

$$P^* = \mathrm{grad}\varphi^{(V)}(P) = \nu_1(\det P)P^{-1}$$

- a variant of relative entropy,
- Pythagorean type decomposition

Prop.

The largest group that preserves the dualistic structure $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ invariant is

$$\tau_G \quad \text{with} \quad G \in SL(n, \mathbf{R})$$

except in the standard case.

Rem. the standard case: τ_G with $G \in GL(n, \mathbf{R})$

Rem. The power potential of the form:

$$V(s) = (1 - s^{\beta})/\beta$$

has a special property.

Special properties for the power potentials

- Orthogonality is $GL(n)$-invariant.
- The dual affine connections derived from the power potentials are $GL(n)$-invariant.

Hence,

- Both ∇ - and ${}^{*}\nabla^{(V)}$ -projection are $GL(n)$ -invariant.

Foliated Structures

The following foliated structure features the dualistic geometry $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ derived by the V-potential.

$$PD(n, \mathbf{R}) = \bigcup_{s>0} \mathcal{L}_s, \quad \mathcal{L}_s = \{P | P > 0, \det P = s\}.$$

$$PD(n, \mathbf{R}) = \bigcup_{P \in \mathcal{L}_s} \mathcal{R}_P. \ \mathcal{R}_P = \{Q | Q = \lambda P, 0 < \lambda \in \mathbf{R}\}$$

Prop.
Each leaf $\mathcal{L}_s$ and $\mathcal{R}_P$ are orthogonal each other with respect to $g^{(V)}$.

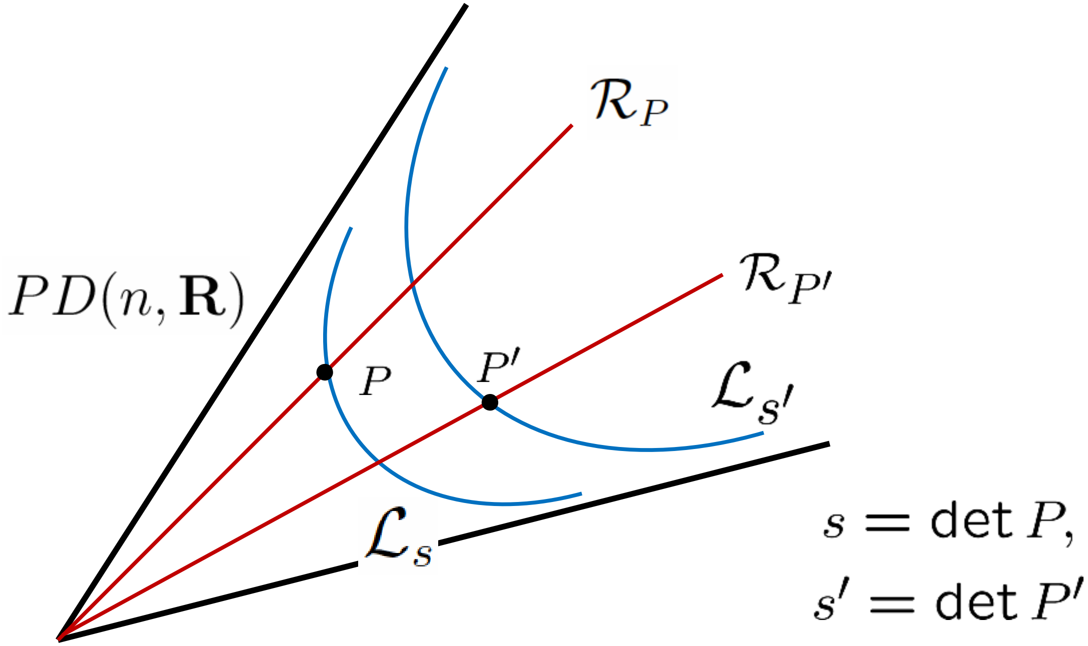
Prop.
Every $\mathcal{R}_P$ is simultaneously a ∇ - and $^*\nabla^{(V)}$-geodesic for an arbitrary V-potential.

Prop.

a

Each leaf $\mathcal{L}_s$ is a homogeneous space with the constant negative curvature $k_s = 1/(\nu_1(s)n)$.

$$R(X,Y)Z = k\{g(Y,Z)X - g(X,Z)Y\}.$$



Application to multivariate statistics

- Non Gaussian distribution
  (generalized exponential family)
  - Robust statistics
    - beta-divergence,
    - Machine learning, and so on
  - Nonextensive statistical physics
    - Power distribution,
    - generalized (Tsallis) entropy, and so on

Application to multivariate statistics

- Geometry of U-model

Def.

Given a convex function U and set $u=U'$, U-model is a family of elliptic (probability) distributions specified by P:

$$\mathcal{M}_U = \left\{ f(x,P) = u\left(-\frac{1}{2}x^T P x - c_U(\det P) \right) : P \in PD(n, \mathbf{R}) \right\}$$

$c_U(\det P)$:normalizing const.

Rem. When U=exp, the U-model is the family of Gaussian distributions.

U-divergence:

Natural closeness measure on the U-model

$$D_U(f,g) = \int \{U(\xi(g(x))) - U(\xi(f(x))) - f(x)[\xi(g(x)) - \xi(f(x))]\}\, dx.$$

where ξ is the inverse function of u.

Rem. When U=exp, the U-divergence is the Kullback-Leibler divergence (relative entropy).

Prop.

Geometry of the U-model equipped with the U-divergence coincides with $(g^{(V)}, \nabla, {}^*\nabla^{(V)})$ derived from the following V-potential function:

$$V(s) = \varphi_U(s) := s^{-\frac{1}{2}} \int U\left(-\frac{1}{2}x^T x - c_U(s)\right) dx + c_U(s), \quad s > 0.$$

Conclusions

Sec. 2

- DA submanifold: needs a tractable characterization or the classification

Sec. 3

- Derived dualistic geometry is invariant under the $SL(n,\mathbf{R})$ -group actions
- Each leaf is a homogeneous manifold with a negative constant curvature
- Decomposition of the divergence function (skipped)
- Relation with the U-model with the U-divergence

Main References

A. Ohara, N. Suda and S. Amari, Dualistic Differential Geometry of Positive Definite Matrices and Its Applications to Related Problems, *Linear Algebra and its Applications*, Vol.**247**, 31-53 (1996).

A. Ohara, Information geometric analysis of semidefinite programming problems, *Proceedings of the institute of statistical mathematics* (統計数理), Vol.**46**, No.2, 317-334 (1998) in Japanese.

A. Ohara and S. Eguchi, Geometry on positive definite matrices and V-potential function, Research Memorandum No. 950, The Institute of Statistical Mathematics, Tokyo, July (2005).