Learn to Transfer Statistical Models from-and-to Populations

Chafik SAMIR

UCA, LIMOS

AI Seminar

April, 2024

Outline

Task-based Learning: Traditional configurations

Some key steps in traditional learning:

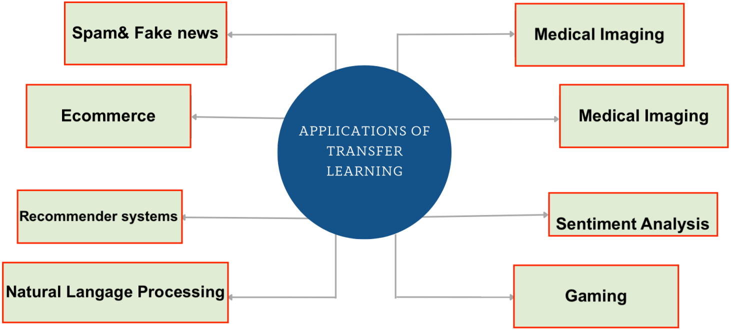
Steps	Task 1	Task 2, ...
1-	Load data D'_1 for T_1	Load data D'_2 for T_2
2-	D_1: Representation	D_2: Representation
3-	Choose and train M_1	Choose and train M_2
4-	μ_1 for optimal $\hat{M}_1$?	μ_2 for optimal $\hat{M}_2$?

Table: We have different configurations (D_i, T_i, M_i, μ_i).

In common classification problems with $T_1 = T_2$:

- D_1 and D_2 belong to the same space: $D_1 = D_2$, $(D_1, D_2 \sim \mathbb{P})$, etc.
- M_1 and M_2 share the same search space M (hyperparameter Θ, loss functions)
- Usually the same evaluation (Precision-Recall) μ

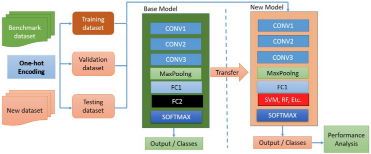
Some Applications of Transfer Learning



TL Example: Object Detection

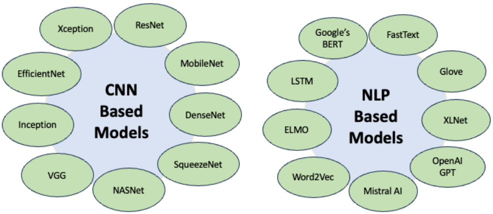
An example of boosting the performance of object detection systems with CNN-based models:

1. Load data and required libraries
2. Select a pre-trained model on large datasets
3. Remove or modify the output layer (or few)
4. Freeze the pre-trained layers (hyperparameter Θ)
5. Fine tuning (start close to the optimum $\hat{\Theta}$?)
6. Evaluate and adjust (available with TensorFlow and PyTorch)



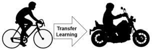
More and more tools

1. It speeds up the learning process
2. It "reduces" the amount of required data (Similarity?)
3. It can provide efficient models as they can be trained "elsewhere" with large datasets
4. Ready to use tools in some applications

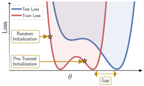


An Overview of reusable knowledge

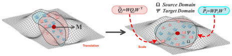
- Before the electric era: Adapt the basic **skill** of balancing



- Build a **prior** to improve the optimization process

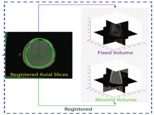


- Transport **data** (domain) or **models** from and to "statistical" populations

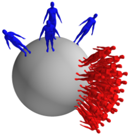


(a) Domain

(b) Distribution



(c) Atlas



(d) Populations

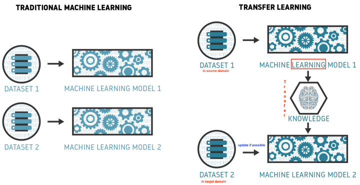
Definitions

A general definition

Given a source $(\mathcal{D}_s, \mathcal{M}_s, \mathcal{T}_s)$, and a target $(\mathcal{D}_t, \mathcal{M}_t, \mathcal{T}_t)$, the transfer aims to improve the learning from target using the learned knowledge (as a prior) from the source.

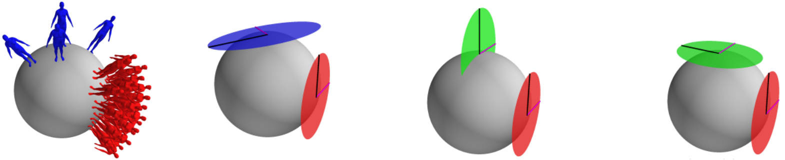
Context for a fixed task (classification, regression)

Given a large population $\mathcal{P}_L$ and a small (labeled or poorly labeled) population $\mathcal{P}_S$, transfer the learned model M_L to be applicable on $\mathcal{P}_S$.



TL on Manifolds: Motivations

- TL can assist us in reusing a well trained model or existing observations to build/improve a new one
- TL was successfully applied for $\mathbb{R}^d$-valued data
- Limitations due to the intrinsic structure from manifold-valued data



(a) 2 Populations (b) Tangent spaces (c) $\mathbb{R}^3$-Translation (d) Parallel transport on $\mathbb{S}^2$

Figure: Illustration of transfer learning on $\mathbb{S}^2$ (Freifeld et al, 2014).

Problem Formulation: TL on Manifolds

To reach such goal, we need some tools:

- Intrinsic distance: Geodesic
- Statistical populations : Mean, variance, covariance, distribution, etc.
- Tangent space at each point
- Parallel transport

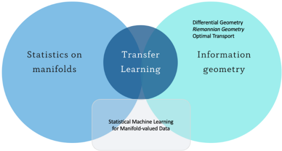


Figure: Generalization of machine learning models for "non-linear" data.

Problem Formulation: TL on Manifolds

To reach such goal, we need some tools:

- Intrinsic distance: Geodesic
- Statistical populations : Mean, variance, covariance, distribution, etc.
- Tangent space at each point
- Parallel transport

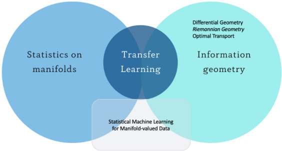


Figure: Generalization of machine learning models for "non-linear" data.

Illustration and applications: Explore the geometry of $\mathcal{P}_+$ and develop a transfer learning algorithm for some statistical models.

The Manifold of Probability Measures

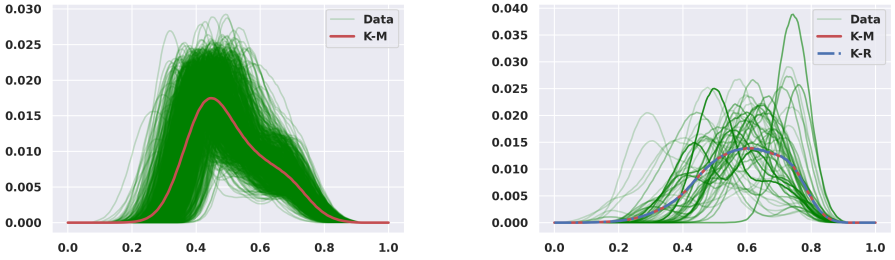


Figure: An illustration with P_L left and P_S right.

Manifold structure

Let $I = \{1, \dots, n, n+1\}, n \in \mathbb{N}$.

- The space of strictly positive probability measures:

$$\mathcal{P}_+(I) = \left\{ \mu = \sum_{i \in I} \mu_i \delta^i \mid \mu_i > 0, \quad \forall i \in I, \text{ and } \sum_{i \in I} \mu_i = 1 \right\}.$$

- Tangent space:

$$T_\mu \mathcal{P}_+(I) = \{\mu\} \times \mathcal{S}_0(I), \text{ where } \mathcal{S}_0(I) \quad = \left\{ \mu = \sum_{i \in I} \mu_i \delta^i \mid \sum_{i \in I} \mu_i = 0 \right\}$$

- Fisher-Rao metric:

$$\mathfrak{g}_\mu(X, Y) = \sum_{i \in I} \frac{X_i Y_i}{\mu_i}, \forall X = \sum_{i \in I} X_i \delta^i, \quad X = \sum_{i \in I} Y_i \delta^i \in T_\mu \mathcal{P}_+(I).$$

Riemannian calculus on $\mathcal{P}_+$

- The Fisher Rao distance d^{FR}: Given $\mu, \nu \in \mathcal{P}_+(I)$, we have

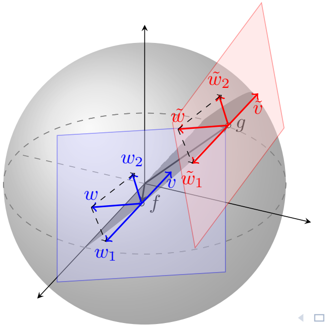
$$d^{FR}(\mu, \nu) = 2 \arccos \left(\sum_{i \in I} \sqrt{\mu_i \nu_i} \right).$$

Riemannian calculus on $\mathcal{P}_+$

- The Fisher Rao distance d^{FR}: Given $\mu, \nu \in \mathcal{P}_+(I)$, we have

$$d^{FR}(\mu,\nu) = 2 \arccos \left(\sum_{i \in I} \sqrt{\mu_i \nu_i} \right).$$

$\Rightarrow$ Isometry: By the map $\Phi(\mu) = 2\sum_{i\in I}\sqrt{\mu_i}e_i$, $\mathcal{P}_+(I)$ is isometric to the sphere $\mathbb{S}^+_{(0,2)}(I) = \{ f \in \mathbb{R}^{n+1} \mid f^i > 0, \forall i \in I \text{ and } \sum_{i\in I}(f^i)^2 = 4 \}$



Riemannian calculus on $\mathcal{P}_+$

- Geodesic path: Starting at μ with direction X.

$$\alpha_i(t) = \left(\cos\frac{t}{2} + \frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\sin\frac{t}{2}\right)^2 \mu_i, \quad \alpha(t) = \sum_{i\in I}\alpha_i(t)\delta^i$$

- Log map: From $\mathcal{P}_+$ to tangent space

$$\log_\mu(\nu) = \frac{l}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\delta^i.$$

- Exponential map: From tangent space to $\mathcal{P}_+$

$$\exp_\mu(X) = \sum_{i\in I}\left(\cos\frac{\|X\|_\mu}{2} + \frac{X_i}{\mu_i\|X\|_\mu}\sin\frac{\|X\|_\mu}{2}\right)^2\mu_i\delta^i, \quad \forall(\mu, X)\in\varepsilon,$$

- Levi-Civita parallel transport:

$$\Gamma_{\mu\rightarrowtail\nu}(X) = \sum_{i\in I}\sqrt{\nu_i}\left(-C\sqrt{\mu_i}\left(2\sin\frac{l}{2} - 2\frac{\tau_i}{\mu_i}\cos\frac{l}{2}\right)\right.$$

Mean and Variance

- The intrinsic mean on $\mathcal{P}_+$ Using Riemannian geodesic distance, the Riemannian mean of a set of probability measures $\{\mu_i\}_{i=1}^N$ on $\mathcal{P}_+(I)$ is by the minimizer of the Fréchet variance:

$$\mu^* = \text{argmin}_\mu \sum_{i=1}^{N} d^{FR}(\mu, \mu_i)^2 \tag{1}$$

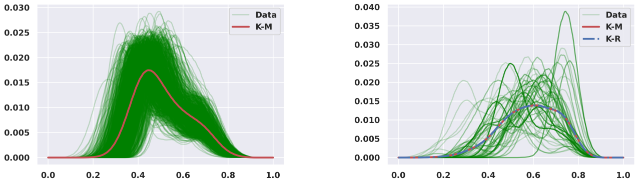


Figure: An illustration with P_L left, P_S right, and their corresponding means in red.

Transfer of Learned Models

Formulation

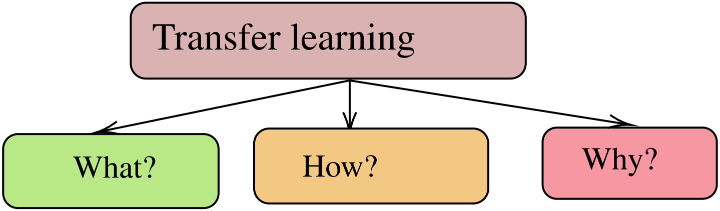
Population $P_1 = P_L$

- $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$
- $\dim P_1 = N_1$.
- $\mu_1^* = \arg\min_\mu \sum_{i=1}^{N_1} d^{FR}(\mu, \mu_i)^2$.
- $a_i = \log_{\mu_1^*}(\mu_i) \in T_{\mu_1^*}\mathcal{P}_+(I)$.
- Statistical model S_1 on $T_{\mu_1^*}\mathcal{P}_+(I)$.

Population $P_2 = P_S$

- $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$
- $\dim P_2 = N_2$, $N_2 \ll N_1$.
- $\mu_2^* = \arg\min_\mu \sum_{i=1}^{N_2} d^{FR}(\mu, \mu_i)^2$.
- $b_i = \log_{\mu_2^*}(\mu_i) \in T_{\mu_2^*}\mathcal{P}_+(I)$.
- Statistical model S_2 on $T_{\mu_2^*}\mathcal{P}_+(I)$

Knowledge as a Model



Linear regression

Logistic regression
Covariance &
PCA

Parallel transport

Isometry

Inner product

To apply the model
on the target population

Step-by-Step TL

1. Project the set of probability measure P_{N_1} to the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, lift the set of probability measure P_{N_2} to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$.

Step-by-Step TL

1. Project the set of probability measure P_{N_1} to the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, lift the set of probability measure P_{N_2} to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$.
2. Learn a statistical model S_1 on $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, learn a statistical model S_2 on $T_{\mu_2^*}\mathcal{P}_+(I)$.

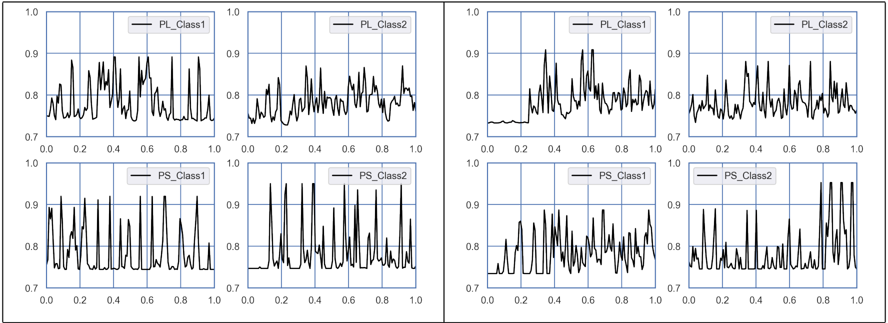
Step-by-Step TL

1. Project the set of probability measure P_{N_1} to the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, lift the set of probability measure P_{N_2} to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$.
2. Learn a statistical model S_1 on $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, learn a statistical model S_2 on $T_{\mu_2^*}\mathcal{P}_+(I)$.
3. Parallel transport S_1 to $T_{\mu_2^*}\mathcal{P}_+(I)$ along the geodesic curve α by computing $S_T = \Gamma_{\mu_1^* \mapsto \mu_2^*}(S_1)$.

Step-by-Step TL

1. Project the set of probability measure P_{N_1} to the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, lift the set of probability measure P_{N_2} to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$.
2. Learn a statistical model S_1 on $T_{\mu_1^*}\mathcal{P}_+(I)$. Similarly, learn a statistical model S_2 on $T_{\mu_2^*}\mathcal{P}_+(I)$.
3. Parallel transport S_1 to $T_{\mu_2^*}\mathcal{P}_+(I)$ along the geodesic curve α by computing $S_T = \Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(S_1)$.
4. Compute the fused model on $T_{\mu_2^*}\mathcal{P}_+(I)$ using shrinkage estimation: $S_\lambda = \lambda S_T + (1-\lambda)S_2,\ 0 \leq \lambda \leq 1$.

Comparing Two Populations of Manifold-valued Data



Example 4: Mean per class	Example 10: Mean per class

Table: Test statistics on different datasets

	1	2	3	4	5	6	7	8	9	10
d	1.5	**1.3**	1.9	**2.2**	1.8	**2.2**	1.9	1.5	2.0	**2.6**
σ_L	**2.8**	**2.8**	2.5	2.4	2.6	2.4	2.7	2.5	**2.3**	2.4
σ_S	3.2	**3.4**	3.2	3.2	2.8	**2.6**	3.2	3.1	2.9	2.8
$p\%$	2	3	**0**	14	1	**0**	9	**0**	**0**	**0**

Linear Regression on $T_{\mu_1^*}\mathcal{P}_+(I)$

- **Inner product on $T_{\mu_1^*}\mathcal{P}_+(I)$:**
  $\mathfrak{g}_{\mu_1^*} : T_{\mu_1^*}\mathcal{P}_+(I) \times T_{\mu_1^*}\mathcal{P}_+(I) \to \mathbb{R}, (v, w) \to \mathfrak{g}_{\mu_1^*}(v, w) = v^T G_{\mu_1^*} w.$
- **Data:** $\mathcal{D} = \{(a_i, t_i), i = 1, ..., N_1\}$, $a_i = \log_{\mu_1^*}(\mu_i)$, $t_i \in \mathbb{R}$.
- **Model:** $y_i : T_{\mu_1^*}\mathcal{P}_+(I) \to \mathbb{R}$,

$$y_i = a_i^T \beta + \beta_0 = \mathfrak{g}_{\mu_1^*}(a_i, G_{\mu_1^*}^{-1}\beta) + \beta_0,$$

  where $\beta_0 \in \mathbb{R}$, $\beta \in T_{\mu_1^*}\mathcal{P}_+(I)$.
- **Least-squares estimation** of β_0 and β:

$$(\widehat{\beta}_0, \widehat{\beta}) = \underset{\beta \in T_{\mu_1^*}\mathcal{P}_+(I), \beta_0 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{N_1} l_i(a_i^T \beta + \beta_0).$$

  where $l_i : \mathbb{R} \to \mathbb{R}_+$, $l_i(y_i) = (y_i - t_i)^2 = (a_i^T \beta + \beta_0 - t_i)^2$.

Algorithm: Transfer of the linear regression model

- **Input:** $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$, $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$ with $N_2 \ll N_1$.
  1. **Compute** μ_1^* and μ_2^* from P_{N_1} and P_{N_2}.
  2. **Project** P_{N_1} on $T_{\mu_1^*}\mathcal{P}_+(I)$ and $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$ on $T_{\mu_2^*}\mathcal{P}_+(I)$.
  3. **Find** $(\widehat{\beta}_0, \widehat{\beta})$ the least squares estimates parameters of the linear regression model on $T_{\mu_1^*}\mathcal{P}_+(I)$.
  4. **Find** (η, η_0) the least squares estimates parameters of the linear regression model on $T_{\mu_2^*}\mathcal{P}_+(I)$.
  5. **Apply** $\Gamma_{\mu_1^* \to \mu_2^*}$ to parallel transport tangent vectors a_i and $G_{\mu_1^*}^{-1}\beta$ to $T_{\mu_2^*}\mathcal{P}_+(I)$.
  6. **Compute** $\widehat{\delta} = G_{\mu_2^*}\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(G_{\mu_1^*}^{-1}\widehat{\beta})$. $(\widehat{\delta}, \beta_0)$ is the solution of the linear regression model $\tilde{y}_i$ on $T_{\mu_2^*}\mathcal{P}_+(I)$.

$$\widehat{\delta} = \underset{\delta \in T_{\mu_2^*}\mathcal{P}_+(I)}{\arg\min} \sum_{i=1}^{N_2} l_i\left(\left(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)\right)^T \delta + \beta_0\right)$$

- **Ouput:** The fused solution $\eta_\lambda = \lambda\widehat{\delta} + (1-\lambda)\widehat{\eta}$, $0 \leq \lambda \leq 1$.

Logistic Regression on $T_{\mu_1^*}\mathcal{P}_+(I)$

- **Inner product on $T_{\mu_1^*}\mathcal{P}_+(I)$:**
  $\mathfrak{g}_{\mu_1^*} : T_{\mu_1^*}\mathcal{P}_+(I) \times T_{\mu_1^*}\mathcal{P}_+(I) \to \mathbb{R}; (v, w) \to \mathfrak{g}_{\mu_1^*}(v, w) = v^T G_{\mu_1^*} w$.
- **Data:** $\mathcal{D} = \{(a_i, t_i)\}_{i=1}^{N_1}$, $a_i = \log_{\mu_1^*}(\mu_i)$ and $t_i \in \{0, 1\}$.
- **Model:** The probability of t_i being in class 1, $P(t_i = 1|a_i)$ is

$$p(a_i) = \frac{1}{1 + e^{-\left(a_i^T \omega + \omega_0\right)}} = \frac{1}{1 + e^{-\left(g_{\mu_1^*}(a_i, G_{\mu_1^*}^{-1}\omega) + \omega_0\right)}}$$

- **Maximum Likelihood Estimation (MLE):** Let $\hat{\theta} = (\widehat{\omega}_0, \widehat{\omega})$ be the maximum likelihood estimators of $\theta = (\omega_0, \omega)$.

Covariance Matrices

- Let $A = [a_1, ..., a_{N_1}] \in T_{\mu_1^*}\mathcal{P}_+(I)$, with $a_i = \log_{\mu_1^*}(\mu_i)$ and let $B = [b_1, ..., b_{N_2}] \in T_{\mu_2^*}\mathcal{P}_+(I)$, with $b_i = \log_{\mu_2^*}(\mu_i)$.
- **The covariance matrix estimator** is defined as

$$C_{N_1} = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} \log_{\mu_1^*}(\mu_i) \log_{\mu_1^*}(\mu_i)^T = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} a_i a_i^T = \frac{1}{N_1 - 1} AA^T$$

and

$$C_{N_2} = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} \log_{\mu_2^*}(\mu_i) \log_{\mu_2^*}(\mu_i)^T = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} b_i b_i^T = \frac{1}{N_2 - 1} BB^T$$

Covariance Matrices

- Let $A = [a_1, \ldots, a_{N_1}] \in T_{\mu_1^*}\mathcal{P}_+(I)$, with $a_i = \log_{\mu_1^*}(\mu_i)$ and let $B = [b_1, \ldots, b_{N_2}] \in T_{\mu_2^*}\mathcal{P}_+(I)$, with $b_i = \log_{\mu_2^*}(\mu_i)$.
- **The covariance matrix estimator** is defined as

$$C_{N_1} = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} \log_{\mu_1^*}(\mu_i)\log_{\mu_1^*}(\mu_i)^T = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} a_i a_i^T = \frac{1}{N_1 - 1}AA^T$$

and

$$C_{N_2} = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} \log_{\mu_2^*}(\mu_i)\log_{\mu_2^*}(\mu_i)^T = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} b_i b_i^T = \frac{1}{N_2 - 1}BB^T$$

➢ C_{N_2} may be a poor estimate of the true covariance matrix of P_{N_2}.

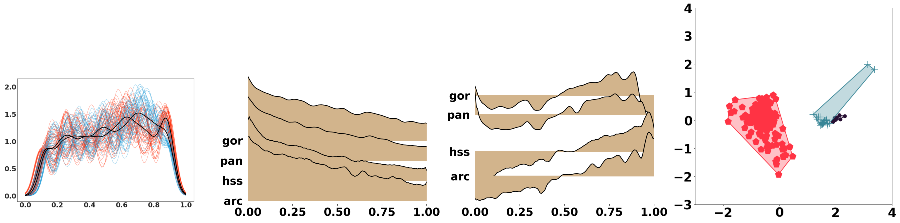
Transfer of covariance matrices and PCs

- **Input:** $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$, $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$ with $N_2 \ll N_1$.
  1. **Compute** μ_1^* and μ_2^* from P_{N_1} and P_{N_2}.
  2. **Compute** A and B by projecting P_{N_1} on $T_{\mu_1^*}\mathcal{P}_+(I)$ and P_{N_2} on $T_{\mu_2^*}\mathcal{P}_+(I)$.
  3. **Compute** covariance matrices C_{N_1} and C_{N_2}.
  4. **Compute** the SVD of A: $A = VDU^T$.
  5. **Compute** the eigen-value decomposition C_{N_1}: VD^2V^T.
  6. **Parallel transport** C_{N_1} to $T_{\mu_2^*}\mathcal{P}_+(I)$:
     $\tilde{A} = \Gamma_{\mu_1^* \to \mu_2^*}(\{a_i\}_{i=1}^{N_1}) \in T_{\mu_2^*}\mathcal{P}_+(I)$,
     $\tilde{C}_{N_1} = \dfrac{1}{N_1 - 1}\tilde{A}\tilde{A}^T = \dfrac{1}{N_1 - 1}\tilde{V}D^2\tilde{V}^T$.
  7. Fix $k_1 \in \mathbb{N}, k_1 < n$. **Compute** k_1-dimensional PC as k_1 eigen-vectors of V and their corresponding k_1 eigen-values D.
  8. Compute k_1-dimensional PC of $\tilde{A}$ as k_1 eigen-vectors of $\tilde{V}$.
- **Output** $C_\lambda = \Pi^r(\lambda \tilde{C}_{N_1} + (1 - \lambda)C_{N_2})$, V, $\tilde{V}$, D.

Transfer of TPCA

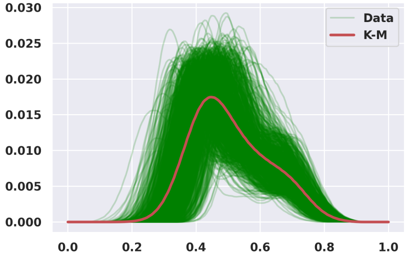
Also called geodesic PCA, for dimensionality reduction and visualization:

1. $\overline{V} = [\overline{v}_1, ..., \overline{v}_n] \in \mathcal{R}^{n \times n}$ is the orthogonal matrix which the eigenvectors of BB^T.
2. $\{\tilde{v}_i\}_{i=1}^{k_1}$ and $\{D_{i,i}/\sqrt{N_1 - 1}\}_{i=1}^{k_1}$ as a PCA model on $T_{\mu_2^*}\mathcal{P}_+(I)$.
3. Fusion: A Gram-Schmidt Orthonormalisation of $\{\tilde{v}_i, \overline{v}_j\}$ and their corresponding eigen-values.



Experiments: Transfer of TPCA

Results on functional data: Populations P_L and P_S with $N_1 = 998$ and $N_2 = 100$ observations, respectively. Each sample belongs to $\mathcal{P}_+(I)$, with $|I| = 100$.



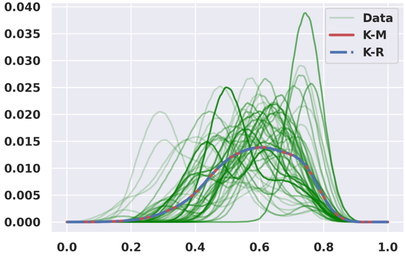
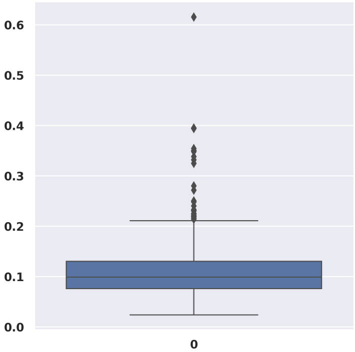


Figure: Some observations and their corresponding Karcher mean. for : Population 1(Left) and Population 2, (Right). In both cases, K-M denotes the Karcher Mean from original and K-R the Karcher Mean from reconstructions with 2 tangent TPCs.

Experiment: Results with transferred TPCA



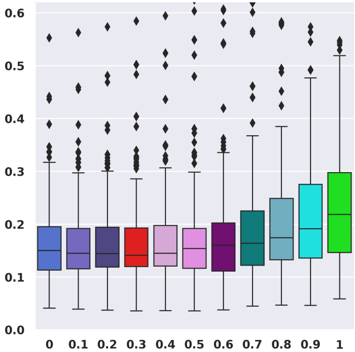


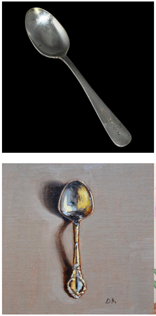
Figure: The reconstruction error (geodesic distance) on P_L (left) and on TP_S (right) for $\lambda \in \{0, 0.1, 0.2, \dots, 1\}$.

Step-by-Step of Model Transfer

- Learn from P_L only:
  1. Compute μ^* and ν^*
  2. Project the elements of P_L to $T_{\mu^*}\mathcal{P}_+(I)$.
  3. Project the elements P_S to $T_{\nu^*}\mathcal{P}_+(I)$.
  4. Learn a statistical model S_1 on $T_{\mu^*}\mathcal{P}_+(I)$.
  5. Parallel transport S_1 to $T_{\nu^*}\mathcal{P}_+(I)$ along the geodesic curve α by computing $S_T = \Gamma_{\mu^* \rightarrowtail \nu^*}(S_1)$.
- If P_S is informative, we can update the statistical model:
  1. Learn a statistical model S_2 on $T_{\nu^*}\mathcal{P}_+(I)$.
  2. Compute the fused model on $T_{\nu^*}\mathcal{P}_+(I)$
  3. Fusion.
     A simple example for shrinkage estimation if valid (Models' space):
     $S_\lambda = \lambda S_T + (1-\lambda)S_2,\ 0 \leq \lambda \leq 1$.

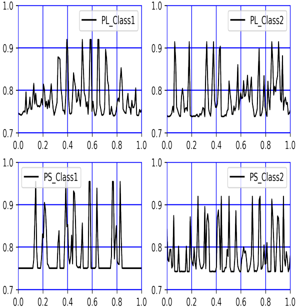
Example: Results with Logistic Regression

- Histograms (SIFT: scale invariant feature transform) from real and painting images[1]
- $|P_L| = 1000$ and several $|P_S| = 86$ for test (total 260 and split 0.33) with 100 samplings.

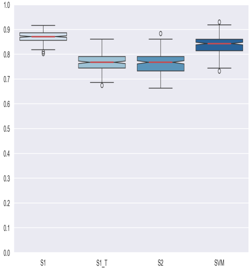


(a) Real

(b) Painting



(c) Means



(d) Boxplots

[1] https://www.hemanthdv.org/officeHomeDataset.html

Conclusion

1. Many successful solutions exist for vector spaces.
2. A new framework for some manifolds.
3. Model Transfer (MT) as Transfer Learning (TL) for probability measures.
4. This framework can be adapted and extended using the analytic expression of the parallel transport (better, or approximations).
5. The proposed methods enjoy several important benefits:
    - The solution is designed for the space of probability measures $\mathcal{P}_+$.
    - The analytic expressions $\mathcal{P}_+$ are easy to implement and escapes the computational requirement of Schild's Ladder approximation.
    - Can be applied for discrete PDFs, prior & posterior distributions (open problem).

Questions?

Joint works with Anis FRADI and Tien Tam TRAN

Thank you for your attention !!

Some references

- O. Freifeld, S. Hauberg, M. J. Black. Model Transport: Towards Scalable Transfer Learning on Manifolds, IEEE CVPR 2014.
- N.Jean, S. M. Xie, and S. Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. NIPS, 2018.
- D.Hafner, D.Tran, T. Lillicrap, A.Irpan, and J. Davidson. Noise contrastive priors for functional uncertainty. In Uncertainty in Artificial Intelligence, PMLR, 2020.
- A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. arXiv:2002.08791, 2020.
- R. Shwartz-Ziv, M. Goldblum, H. Souri, S. Kapoor, C. Zhu, Y. LeCun, A. G. Wilson. Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors, arXiv:2205.10279, 2022.