Optimal Bayesian Transfer Learning

Xiaoning Qian,
Alireza Karbalayghareh, and Edward R Dougherty

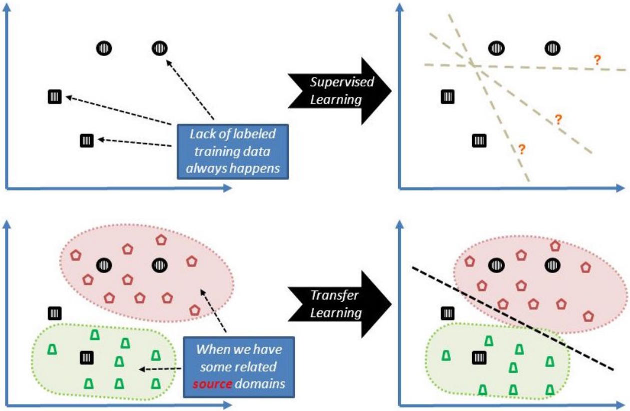
Texas A&M University

August 7, 2018

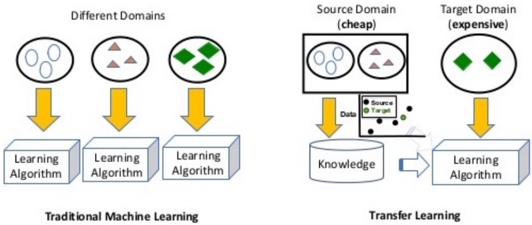
Outline

Transfer Learning Basics

- Traditional machine learning vs. transfer learning



Transfer Learning Basics



- Suppose we want to do a supervised learning but there is lack of labeled data in the domain of interest (target domain).
- Therefore, the classifier cannot be trained well and error rate would be high.
- At the same time, suppose we have plenty of labeled data in a different but relevant domain (source domain).
- The problem of transfer learning is to answer when and how to employ those source data in order to design a more accurate classifier in the target domain.

Domain Adaptation

- Distributions of source and target data are different (not i.i.d. as in traditional machine learning).
- Domain adaptation [1] aims to find a common domain where both source and target data can be transformed to have similar distributions.
- Often, transformation is forced to source and target data but no theoretical guarantee that the prediction performance in the target domain will be enhanced.
- There is no rigorous reasoning for “transferability” and it does not answer if the two domains are actually relevant.
- More critically, is there a way to optimally transfer the relevant knowledge and data from source to target?

Optimal Bayesian Classifier

- Feature-label distribution: $p(\mathbf{x}, l|\theta) = p(\mathbf{x}|l;\theta)p(l|\theta)$
- Prior distribution: $p(\theta)$
- Likelihood: $p(\mathcal{D}|\theta) = \prod_n p(\mathbf{x}^n, l^n|\theta)$
- Posterior: $p(\theta|\mathcal{D}) = \frac{p(\theta)\prod_n p(\mathbf{x}^n, l^n|\theta)}{p(\mathcal{D})}$
- Posterior predictive (effective class-conditional) distribution given a new feature vector $\mathbf{x}^*$: $p(x^*|l;\mathcal{D}) \propto \int d\theta p(x^*|l;\theta)p(\theta|\mathcal{D})$
- Optimal Bayesian classifier:

$$\arg\max_{l\in\{1,\cdots,L\}} p(\theta_l|\mathcal{D})p(x^*|l;\mathcal{D})$$

Bayesian Transfer Learning

- We formulate a Bayesian transfer learning framework to transfer source domain knowledge and data for learning in target domain.
- Our Bayesian framework directly models the feature-label distributions in source and target domains.
- The "transferability" across domains can be characterized by a joint prior distribution on model parameters of feature-label distributions across domains.
- The relevance of source and target problems can be studied through the joint posterior distribution of model parameters.
- Under such a Bayesian framework, we show how to optimally transfer abundant source data to the target domain and define the Optimal Bayesian Transfer Learning (OBTL) classifier.

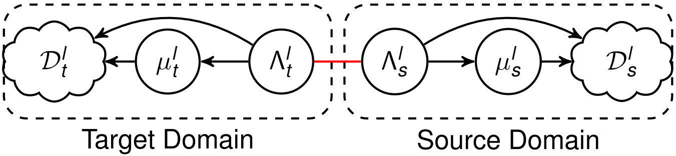
Multivariate Gaussian Data

- Distributions of data in source and target domains:

$$\mathbf{x}_s^l \sim \mathcal{N}\left(\mu_s^l, \left(\Lambda_s^l\right)^{-1}\right), \qquad \mathbf{x}_t^l \sim \mathcal{N}\left(\mu_t^l, \left(\Lambda_t^l\right)^{-1}\right), \quad l \in \{1, \cdots, L\},$$

- Joint prior for the parameters of the two domains:

$$p\left(\mu_s^l, \mu_t^l, \Lambda_s^l, \Lambda_t^l\right) = p\left(\mu_s^l | \Lambda_s^l\right) p\left(\mu_t^l | \Lambda_t^l\right) p\left(\Lambda_s^l, \Lambda_t^l\right), \quad l \in \{1, \cdots, L\},$$
$$\mu_s^l | \Lambda_s^l \sim \mathcal{N}\left(\mathbf{m}_s^l, \left(\kappa_s^l \Lambda_s^l\right)^{-1}\right), \qquad \mu_t^l | \Lambda_t^l \sim \mathcal{N}\left(\mathbf{m}_t^l, \left(\kappa_t^l \Lambda_t^l\right)^{-1}\right),$$



- In the case of one domain, Wishart matrices are used for a conjugate prior for the distribution of precision matrices.
- The main question here is: how to define a joint distribution between two Wishart matrices $p(\Lambda_t^l, \Lambda_s^l)$?

Theorem ([2])

If $\Lambda \sim W_d(\mathbf{M}, \nu)$, and $\mathbf{A}$ is an $r \times d$ matrix of rank r, where $r \leq d$, then $\mathbf{A}\Lambda\mathbf{A}^{'} \sim W_r(\mathbf{A}\mathbf{M}\mathbf{A}^{'}, \nu)$.

Corollary

If $\Lambda \sim W_d(\mathbf{M}, \nu)$ and $\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12}^{'} & \Lambda_{22} \end{pmatrix}$, where Λ_{11} and Λ_{22} are $d_1 \times d_1$ and $d_2 \times d_2$ submatrices, respectively, and if $\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}^{'} & \mathbf{M}_{22} \end{pmatrix}$ is the corresponding partition of $\mathbf{M}$ with $\mathbf{M}_{11}$ and $\mathbf{M}_{22}$ being two $d_1 \times d_1$ and $d_2 \times d_2$ submatrices, respectively, then $\Lambda_{11} \sim W_{d_1}(\mathbf{M}_{11}, \nu)$ and $\Lambda_{22} \sim W_{d_2}(\mathbf{M}_{22}, \nu)$.

Theorem ([3])

Let $\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12}^{'} & \Lambda_{22} \end{pmatrix}$ *be a* $(d_1 + d_2) \times (d_1 + d_2)$ *partitioned Wishart random matrix, where the diagonal partitions are of sizes* $d_1 \times d_1$ *and* $d_2 \times d_2$*, respectively. The Wishart distribution of* Λ *has* $\nu \geq d_1 + d_2$ *degrees of freedom and positive-definite scale matrix* $\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}^{'} & \mathbf{M}_{22} \end{pmatrix}$ *partitioned in the same way as* Λ*. The joint distribution of the two diagonal partitions* Λ_{11} *and* Λ_{22} *have the density function given by*

$$p(\Lambda_{11}, \Lambda_{22}) = K \operatorname{etr}\left(-\frac{1}{2}\left(\mathbf{M}_{11}^{-1} + \mathbf{F}^{'}\mathbf{C}_2\mathbf{F}\right)\Lambda_{11}\right) \operatorname{etr}\left(-\frac{1}{2}\mathbf{C}_2^{-1}\Lambda_{22}\right) \times |\Lambda_{11}|^{\frac{\nu - d_2 - 1}{2}} \, |\Lambda_{22}|^{\frac{\nu - d_1 - 1}{2}} \; {}_0F_1\left(\frac{\nu}{2}; \frac{1}{4}\mathbf{G}\right), \tag{1}$$

where $\mathbf{C}_2 = \mathbf{M}_{22} - \mathbf{M}_{12}^{'}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$, $\mathbf{F} = \mathbf{C}_2^{-1}\mathbf{M}_{12}^{'}\mathbf{M}_{11}^{-1}$, $\mathbf{G} = \Lambda_{22}^{\frac{1}{2}}\mathbf{F}\Lambda_{11}\mathbf{F}^{'}\Lambda_{22}^{\frac{1}{2}}$, $K^{-1} = 2^{\frac{(d_1+d_2)\nu}{2}}\Gamma_{d_1}\left(\frac{\nu}{2}\right)\Gamma_{d_2}\left(\frac{\nu}{2}\right)|\mathbf{M}|^{\frac{\nu}{2}}$*, and* ${}_0F_1$ *is the generalized matrix-variate hypergeometric function.*

Multivariate gamma function given by $\Gamma_d(\alpha) = \pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\left(\alpha - \frac{i-1}{2}\right)$.

Hypergeometric functions of matrix arguments

Definition ([4])

The generalized hypergeometric function of one matrix argument is defined by

$$ {}_pF_q(a_1, \cdots, a_p; b_1, \cdots, b_q; \mathbf{X}) = \sum_{k=0}^{\infty} \sum_{\kappa \vdash k} \frac{(a_1)_\kappa \cdots (a_p)_\kappa}{(b_1)_\kappa \cdots (b_q)_\kappa} \frac{C_\kappa(\mathbf{X})}{k!}, \tag{2} $$

where a_i, $i = 1, \cdots, p$, and b_j, $j = 1, \cdots, q$, are arbitrary complex (real in our case) numbers, $C_\kappa(\mathbf{X})$ is the zonal polynomial of $d \times d$ symmetric matrix $\mathbf{X}$ corresponding to the ordered partition $\kappa = (k_1, \cdots, k_d)$, $k_1 \geq \cdots \geq k_d \geq 0$, $k_1 + \cdots k_d = k$ and $\sum_{\kappa \vdash k}$ denotes summation over all partitions κ of k. The generalized hypergeometric coefficient $(a)_\kappa$ is defined by

$$ (a)_\kappa = \prod_{i=1}^{d} \left(a - \frac{i-1}{2} \right)_{k_i}, \tag{3} $$

where $(a)_r = a(a+1)\cdots(a+r-1)$, $r = 1, 2, \cdots$, with $(a)_0 = 1$.

Hypergeometric functions of matrix arguments

Most special cases are:

$$
\begin{aligned}
{}_0F_0(\mathbf{X}) &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{C_\kappa(\mathbf{X})}{k!} = \sum_{k=0}^{\infty}\frac{(\mathrm{tr}(\mathbf{X}))^k}{k!} = \mathrm{etr}(\mathbf{X}),\\
{}_1F_0(a;\mathbf{X}) &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{(a)_\kappa C_\kappa(\mathbf{X})}{k!} = |\mathbf{I}_m-\mathbf{X}|^{-a}, \quad ||\mathbf{X}||<1,\\
{}_0F_1(b;\mathbf{X}) &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{C_\kappa(\mathbf{X})}{(b)_\kappa k!}, \qquad \text{(Confluent hypergeometric limit function)}\\
{}_1F_1(a;b;\mathbf{X}) &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{(a)_\kappa}{(b)_\kappa}\frac{C_\kappa(\mathbf{X})}{k!}, \qquad \text{(Confluent hypergeometric function of the first kind)}\\
{}_2F_1(a,b;c;\mathbf{X}) &= \sum_{k=0}^{\infty}\sum_{\kappa\vdash k}\frac{(a)_\kappa(b)_\kappa}{(c)_\kappa}\frac{C_\kappa(\mathbf{X})}{k!}, \quad ||\mathbf{X}||<1, \qquad \text{(Gauss hypergeometric function)}
\end{aligned}
\tag{4}
$$

Hypergeometric functions of matrix arguments

Theorem ([2])

Let $\mathbf{Z}$ *be a complex symmetric matrix whose real part is positive-definite, and let* $\mathbf{X}$ *be an arbitrary complex symmetric matrix. Then*

$$\int_{\mathbf{R}>0} \operatorname{etr}(-\mathbf{Z}\mathbf{R})|\mathbf{R}|^{\alpha-\frac{d+1}{2}} C_\kappa(\mathbf{R}\mathbf{X})d\mathbf{R} = \Gamma_d(\alpha)(\alpha)_\kappa|\mathbf{Z}|^{-\alpha} C_\kappa(\mathbf{X}\mathbf{Z}^{-1}), \tag{5}$$

the integration being over the space of positive-definite $d \times d$ *matrices, and valid for all complex numbers* α *satisfying* $\mathrm{Re}(\alpha) > \frac{d-1}{2}$. $\Gamma_d(\alpha)$ *is the multivariate gamma function defined in (**??**).*

Theorem ([5])

If $\mathbf{Z} > 0$ *and* $\mathrm{Re}(\alpha) > \frac{d-1}{2}$*, and* $\mathbf{X}$ *is a* $d \times d$ *symmetric matrix, we have*

$$\begin{aligned} &\int_{\mathbf{R}>0} \operatorname{etr}(-\mathbf{Z}\mathbf{R})|\mathbf{R}|^{\alpha-\frac{d+1}{2}} {}_pF_q(a_1,\cdots,a_p;b_1,\cdots,b_q;\mathbf{R}\mathbf{X})d\mathbf{R} \\ &= \int_{\mathbf{R}>0} \operatorname{etr}(-\mathbf{Z}\mathbf{R})|\mathbf{R}|^{\alpha-\frac{d+1}{2}} {}_pF_q(a_1,\cdots,a_p;b_1,\cdots,b_q;\mathbf{R}^{1/2}\mathbf{X}\mathbf{R}^{1/2})d\mathbf{R} \\ &= \Gamma_d(\alpha)|\mathbf{Z}|^{-\alpha} {}_{p+1}F_q(a_1,\cdots,a_p,\alpha;b_1,\cdots,b_q;\mathbf{X}\mathbf{Z}^{-1}). \end{aligned}$$

Joint prior for two precision matrices

- We define the following joint prior for the precision matrices of the source and target domains:

$$p(\Lambda_t^l, \Lambda_s^l) = K^l \mathrm{etr}\left(-\frac{1}{2}\left(\left(\mathbf{M}_t^l\right)^{-1} + \mathbf{F}^{l'}\mathbf{C}^l\mathbf{F}^l\right)\tilde{\Lambda}_t^l\right)$$
$$\mathrm{etr}\left(-\frac{1}{2}\left(\mathbf{C}^l\right)^{-1}\Lambda_s^l\right)\left|\Lambda_t^l\right|^{\frac{\nu^l-d-1}{2}}\left|\Lambda_s^l\right|^{\frac{\nu^l-d-1}{2}}\ {}_0F_1\left(\frac{\nu^l}{2};\frac{1}{4}\mathbf{G}^l\right), \tag{6}$$

where $\mathbf{M}^l = \begin{pmatrix} \mathbf{M}_t^l & \mathbf{M}_{ts}^l \\ \mathbf{M}_{ts}^{l\,'} & \mathbf{M}_s^l \end{pmatrix}$ is an $2d \times 2d$ positive definite scale matrix, and $\nu^l \geq 2d$ is degrees of freedom. $\mathbf{C}^l = \mathbf{M}_s^l - \mathbf{M}_{ts}^{l\,'}\left(\mathbf{M}_t^l\right)^{-1}\mathbf{M}_{ts}^l$, $\mathbf{F}^l = \left(\mathbf{C}^l\right)^{-1}\mathbf{M}_{ts}^{l\,'}\left(\mathbf{M}_t^l\right)^{-1}$, $\mathbf{G}^l = \Lambda_s^{l\,\frac{1}{2}}\mathbf{F}^l\Lambda_t^l\mathbf{F}^{l'}\Lambda_s^{l\,\frac{1}{2}}$, $(K^l)^{-1} = 2^{d\nu^l}\Gamma_d^2\left(\frac{\nu^l}{2}\right)|\mathbf{M}^l|^{\frac{\nu^l}{2}}$.

- The marginal distributions are Wishart for each domain (we are interested in understanding how source data may help better learn the marginal distribution in target domain):

$$\Lambda_z^l \sim W_d(\mathbf{M}_z^l, \nu^l), \quad l \in \{1, \cdots, L\}, \quad z \in \{s, t\}. \tag{7}$$

Posteriors

- Joint likelihood of source and target:

$$\begin{aligned} p(\mathcal{D}_t, \mathcal{D}_s | \mu_t, \mu_s, \Lambda_t, \Lambda_s) &= p(\mathcal{D}_t | \mu_t, \Lambda_t) p(\mathcal{D}_s | \mu_s, \Lambda_s) \\ &= p(\mathcal{D}_t^1, \cdots, \mathcal{D}_t^L | \mu_t^1, \cdots, \mu_t^L, \Lambda_t^1, \cdots, \Lambda_t^L) \\ &\quad \times p(\mathcal{D}_s^1, \cdots, \mathcal{D}_s^L | \mu_s^1, \cdots, \mu_s^L, \Lambda_s^1, \cdots, \Lambda_s^L) \\ &= \prod_{l=1}^{L} p(\mathcal{D}_t^l | \mu_t^l, \Lambda_t^l) \prod_{l=1}^{L} p(\mathcal{D}_s^l | \mu_s^l, \Lambda_s^l). \end{aligned} \tag{8}$$

- Joint posterior of source and target:

$$\begin{aligned} & p(\mu_t, \mu_s, \Lambda_t, \Lambda_s | \mathcal{D}_t, \mathcal{D}_s) \\ & \propto p(\mathcal{D}_t, \mathcal{D}_s | \mu_t, \mu_s, \Lambda_t, \Lambda_s) p(\mu_t, \mu_s, \Lambda_t, \Lambda_s) \\ & \propto \prod_{l=1}^{L} p(\mathcal{D}_t^l | \mu_t^l, \Lambda_t^l) \prod_{l=1}^{L} p(\mathcal{D}_s^l | \mu_s^l, \Lambda_s^l) \prod_{l=1}^{L} p(\mu_t^l, \mu_s^l, \Lambda_t^l, \Lambda_s^l) \\ & \propto \prod_{l=1}^{L} p(\mathcal{D}_t^l | \mu_t^l, \Lambda_t^l) p(\mathcal{D}_s^l | \mu_s^l, \Lambda_s^l) p\left(\mu_s^l | \Lambda_s^l\right) p\left(\mu_t^l | \Lambda_t^l\right) p\left(\Lambda_s^l, \Lambda_t^l\right) \end{aligned} \tag{9}$$

Posteriors of Target Parameters

- Posterior of target given both the source and target data:

$$
\begin{aligned}
p(\mu_t, \Lambda_t | \mathcal{D}_t, \mathcal{D}_s) &= \int_{\mu_s, \Lambda_s} p(\mu_t, \mu_s, \Lambda_t, \Lambda_s | \mathcal{D}_t, \mathcal{D}_s) d\mu_s d\Lambda_s \\
&= \prod_{l=1}^{L} \int_{\mu_s^l, \Lambda_s^l} p(\mu_t^l, \mu_s^l, \Lambda_t^l, \Lambda_s^l | \mathcal{D}_t^l, \mathcal{D}_s^l) d\mu_s^l d\Lambda_s^l \\
&= \prod_{l=1}^{L} p(\mu_t^l, \Lambda_t^l | \mathcal{D}_t^l, \mathcal{D}_s^l),
\end{aligned}
$$

where

$$
\begin{aligned}
& p(\mu_t^l, \Lambda_t^l | \mathcal{D}_t^l, \mathcal{D}_s^l) \\
&= \int_{\mu_s^l, \Lambda_s^l} p(\mu_t^l, \mu_s^l, \Lambda_t^l, \Lambda_s^l | \mathcal{D}_t^l, \mathcal{D}_s^l) d\mu_s^l d\Lambda_s^l \\
&\propto p(\mathcal{D}_t^l | \mu_t^l, \Lambda_t^l) p\left(\mu_t^l | \Lambda_t^l\right) \\
&\times \int_{\mu_s^l, \Lambda_s^l} p(\mathcal{D}_s^l | \mu_s^l, \Lambda_s^l) p\left(\mu_s^l | \Lambda_s^l\right) p\left(\Lambda_s^l, \Lambda_t^l\right) d\mu_s^l d\Lambda_s^l.
\end{aligned} \tag{10}
$$

Posteriors of Target Parameters

Lemma

If $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ where $\mathbf{x}_i$ is a $d \times 1$ vector and $\mathbf{x}_i \sim \mathcal{N}(\mu, (\Lambda)^{-1})$, for $i = 1, \cdots, n$, and (μ, Λ) has a Normal-Wishart prior, such that, $\mu|\Lambda \sim \mathcal{N}(\mathbf{m}, (\kappa\Lambda)^{-1})$ and $\Lambda \sim W_d(\mathbf{M}, \nu)$, then the posterior of (μ, Λ) upon observing $\mathcal{D}$ is also a Normal-Wishart distribution:

$$\begin{aligned} \mu|\Lambda, \mathcal{D} &\sim \mathcal{N}(\mathbf{m}_n, (\kappa_n\Lambda)^{-1}), \\ \Lambda|\mathcal{D} &\sim W_d(\mathbf{M}_n, \nu_n), \end{aligned} \tag{11}$$

where

$$\begin{aligned} \kappa_n = \kappa + n, \quad \nu_n = \nu + n, \quad \mathbf{m}_n &= \frac{\kappa\mathbf{m} + n\bar{\mathbf{x}}}{\kappa + n}, \\ \mathbf{M}_n^{-1} = \mathbf{M}^{-1} + \mathbf{S} &+ \frac{\kappa n}{\kappa + n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})', \end{aligned} \tag{12}$$

depending on the sample mean and covariance matrix

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \tag{13}$$

Posteriors of Target Parameters

- Using the previous lemma and theorems, we can find the closed-form posterior distribution of mean and precision matrix of the target domain, which is a function of matrix-variate Confluent hypergeometric function of first kind:

$$
\begin{aligned}
& p(\mu_t^l, \Lambda_t^l | \mathcal{D}_t^l, \mathcal{D}_s^l) = \\
& A^l \left|\Lambda_t^l\right|^{\frac{1}{2}} \exp\left(-\frac{\kappa_{t,n}^l}{2}\left(\mu_t^l - \mathbf{m}_{t,n}^l\right)' \Lambda_t^l \left(\mu_t^l - \mathbf{m}_{t,n}^l\right)\right) \\
& \times \left|\Lambda_t^l\right|^{\frac{\nu^l + n_t^l - d - 1}{2}} \operatorname{etr}\left(-\frac{1}{2}\left(\mathbf{T}_t^l\right)^{-1} \Lambda_t^l\right) \\
& \times {}_1F_1\left(\frac{\nu^l + n_s^l}{2}; \frac{\nu^l}{2}; \frac{1}{2}\mathbf{F}^l \Lambda_t^l \mathbf{F}^{l'} \mathbf{T}_s^l\right),
\end{aligned}
\tag{14}
$$

- We see that as opposed to one-domain posterior which is Normal-Wishart, here the posterior is Normal-Hypergeometric.

where

$$\left(A^l\right)^{-1} = \left(\frac{2\pi}{\kappa^l_{t,n}}\right)^{\frac{d}{2}} 2^{\frac{d\left(\nu^l+n^l_t\right)}{2}} \Gamma_d\left(\frac{\nu^l+n^l_t}{2}\right) \\ \times \left|\mathbf{T}^l_t\right|^{\frac{\nu^l+n^l_t}{2}} {}_2F_1\left(\frac{\nu^l+n^l_s}{2}, \frac{\nu^l+n^l_t}{2}; \frac{\nu^l}{2}; \mathbf{T}^l_s\mathbf{F}^l\mathbf{T}^l_t\mathbf{F}^{l'}\right), \tag{15}$$

and

$$\begin{aligned} \kappa^l_{t,n} &= \kappa^l_t + n^l_t, & \kappa^l_{s,n} &= \kappa^l_s + n^l_s, \\ \mathbf{m}^l_{t,n} &= \frac{\kappa^l_t\mathbf{m}^l_t + n^l_t\bar{\mathbf{x}}^l_t}{\kappa^l_t + n^l_t}, & \mathbf{m}^l_{s,n} &= \frac{\kappa^l_s\mathbf{m}^l_s + n^l_s\bar{\mathbf{x}}^l_s}{\kappa^l_s + n^l_s}, \\ \left(\mathbf{T}^l_t\right)^{-1} &= \left(\mathbf{M}^l_t\right)^{-1} + \mathbf{F}^{l'}\mathbf{C}^l\mathbf{F}^l + \mathbf{S}^l_t + \frac{\kappa^l_t n^l_t}{\kappa^l_t + n^l_t}(\mathbf{m}^l_t - \bar{\mathbf{x}}^l_t)(\mathbf{m}^l_t - \bar{\mathbf{x}}^l_t)', \\ \left(\mathbf{T}^l_s\right)^{-1} &= \left(\mathbf{C}^l\right)^{-1} + \mathbf{S}^l_s + \frac{\kappa^l_s n^l_s}{\kappa^l_s + n^l_s}(\mathbf{m}^l_s - \bar{\mathbf{x}}^l_s)(\mathbf{m}^l_s - \bar{\mathbf{x}}^l_s)', \end{aligned} \tag{16}$$

depending on the corresponding sample mean vectors and sample covariance matrices as follows:

$$\begin{aligned} \bar{\mathbf{x}}^l_t &= \frac{1}{n^l_t}\sum_{i=1}^{n^l_t}\mathbf{x}^l_{t,i}, \quad \bar{\mathbf{x}}^l_s = \frac{1}{n^l_s}\sum_{i=1}^{n^l_s}\mathbf{x}^l_{s,i}, \\ \mathbf{S}^l_t &= \sum_{i=1}^{n^l_t}\left(\mathbf{x}^l_{t,i} - \bar{\mathbf{x}}^l_t\right)\left(\mathbf{x}^l_{t,i} - \bar{\mathbf{x}}^l_t\right)', \\ \mathbf{S}^l_s &= \sum_{i=1}^{n^l_s}\left(\mathbf{x}^l_{s,i} - \bar{\mathbf{x}}^l_s\right)\left(\mathbf{x}^l_{s,i} - \bar{\mathbf{x}}^l_s\right)'. \end{aligned} \tag{17}$$

Effective Class-Conditional Densities

- The effective class-conditional densities (thereafter posterior predictive):

$$p(\mathbf{x}|l) = \int_{\mu_t^l, \Lambda_t^l} p(\mathbf{x}|\mu_t^l, \Lambda_t^l)\pi^\star(\mu_t^l, \Lambda_t^l) d\mu_t^l d\Lambda_t^l \tag{18}$$

$$\begin{aligned} O_{\mathrm{OBTL}}(\mathbf{x}|l) = p(\mathbf{x}|l) = \\ \pi^{-\frac{d}{2}} \left(\frac{\kappa_{t,n}^l}{\kappa_{\mathbf{x}}^l}\right)^{\frac{d}{2}} \Gamma_d\left(\frac{\nu^l + n_t^l + 1}{2}\right)\Gamma_d^{-1}\left(\frac{\nu^l + n_t^l}{2}\right) \\ \times \left|\mathbf{T}_{\mathbf{x}}^l\right|^{\frac{\nu^l + n_t^l + 1}{2}} {}_2F_1\left(\frac{\nu^l + n_s^l}{2}, \frac{\nu^l + n_t^l + 1}{2}; \frac{\nu^l}{2}; \mathbf{T}_s^l \mathbf{F}^l \mathbf{T}_{\mathbf{x}}^l \mathbf{F}^{l\prime}\right) \\ \times \left|\mathbf{T}_t^l\right|^{-\frac{\nu^l + n_t^l}{2}} {}_2F_1^{-1}\left(\frac{\nu^l + n_s^l}{2}, \frac{\nu^l + n_t^l}{2}; \frac{\nu^l}{2}; \mathbf{T}_s^l \mathbf{F}^l \mathbf{T}_t^l \mathbf{F}^{l\prime}\right), \end{aligned} \tag{19}$$

where

$$\begin{aligned} \kappa_{\mathbf{x}}^l &= \kappa_{t,n}^l + 1 = \kappa_t^l + n_t^l + 1, \\ \mathbf{m}_{\mathbf{x}}^l &= \frac{\kappa_{t,n}^l \mathbf{m}_{t,n}^l + \mathbf{x}}{\kappa_{t,n} + 1}, \\ \left(\mathbf{T}_{\mathbf{x}}^l\right)^{-1} &= \left(\mathbf{T}_t^l\right)^{-1} + \frac{\kappa_{t,n}^l}{\kappa_{t,n}^l + 1}\left(\mathbf{m}_{t,n}^l - \mathbf{x}\right)\left(\mathbf{m}_{t,n}^l - \mathbf{x}\right)'. \end{aligned} \tag{20}$$

Optimal Bayesian Transfer Learning (OBTL) Classifier

- Let c_t^l be the prior probability that the target sample $\mathbf{x}$ belongs to the class $l \in \{1, \cdots, L\}$. Since $0 < c_t^l < 1$ and $\sum_{l=1}^{L} c_t^l = 1$, a Dirichlet prior is assumed for the c_t^l:

$$(c_t^1, \cdots, c_t^L) \sim \mathrm{Dir}(L, \xi_t), \tag{21}$$

where $\xi_t = (\xi_t^1, \cdots, \xi_t^L)$ are the concentration parameters, where $\xi_t^l > 0$ for all $l \in \{1, \cdots, L\}$.

- The posterior of c_t^l's is also another Dirichlet distribution:

$$\begin{aligned} \pi^\star = (c_t^1, \cdots, c_t^L | \mathbf{n}) &\sim \mathrm{Dir}(L, \xi_t + \mathbf{n}) \\ &= \mathrm{Dir}(L, \xi_t^1 + n_t^1, \cdots, \xi_t^L + n_t^L), \end{aligned} \tag{22}$$

with the posterior mean of c_t^l as

$$\mathrm{E}_{\pi^\star}(c_t^l) = \frac{\xi_t^l + n_t^l}{N_t + \xi_t^0}, \tag{23}$$

where $N_t = \sum_{l=1}^{L} n_t^l$ and $\xi_t^0 = \sum_{l=1}^{L} \xi_t^l$.

- The optimal Bayesian transfer learning (OBTL) classifier for any new unlabeled sample $\mathbf{x}$ in the target domain is defined as:

OBTL

$$\Psi_{\mathrm{OBTL}}(\mathbf{x}) = \arg \max_{l \in \{1, \cdots, L\}} \mathrm{E}_{\pi^\star}(c_t^l) O_{\mathrm{OBTL}}(\mathbf{x}|l). \tag{24}$$

OBC in Target Domain

- The effective class-conditional densities $p(\mathbf{x}|l) = O_{\mathrm{OBC}}(\mathbf{x}|l)$ for OBC are derived as:

$$O_{\mathrm{OBC}}(\mathbf{x}|l) = \pi^{-\frac{d}{2}} \left(\frac{\kappa^l_{t,n}}{\kappa^l_{t,n}+1}\right)^{\frac{d}{2}} \Gamma_d\left(\frac{\nu^l + n^l_t + 1}{2}\right) \Gamma_d^{-1}\left(\frac{\nu^l + n^l_t}{2}\right) \left|\mathbf{M}^l_{\mathbf{x}}\right|^{\frac{\nu^l+n^l_t+1}{2}} \left|\mathbf{M}^l_{t,n}\right|^{-\frac{\nu^l+n^l_t}{2}}, \tag{25}$$

where

$$\left(\mathbf{M}^l_{\mathbf{x}}\right)^{-1} = \left(\mathbf{M}^l_{t,n}\right)^{-1} + \frac{\kappa^l_{t,n}}{\kappa^l_{t,n}+1}(\mathbf{m}^l_{t,n} - \mathbf{x})(\mathbf{m}^l_{t,n} - \mathbf{x})', \tag{26}$$

$$\begin{aligned} &\kappa^l_{t,n} = \kappa^l_t + n^l_t, \quad \nu^l_{t,n} = \nu^l + n^l_t, \quad \mathbf{m}^l_{t,n} = \frac{\kappa^l_t \mathbf{m}^l_t + n^l_t \bar{\mathbf{x}}^l_t}{\kappa^l_t + n^l_t}, \\ &\left(\mathbf{M}^l_{t,n}\right)^{-1} = \left(\mathbf{M}^l_t\right)^{-1} + \mathbf{S}^l_t + \frac{\kappa^l_t n^l_t}{\kappa^l_t + n^l_t}(\mathbf{m}^l_t - \bar{\mathbf{x}}^l_t)(\mathbf{m}^l_t - \bar{\mathbf{x}}^l_t)', \end{aligned} \tag{27}$$

with the corresponding sample mean and covariance: $\bar{\mathbf{x}}^l_t = \frac{1}{n^l_t}\sum_{i=1}^{n^l_t} \mathbf{x}^l_{t,i}$, $\quad \mathbf{S}^l_t = \sum_{i=1}^{n^l_t}\left(\mathbf{x}^l_{t,i} - \bar{\mathbf{x}}^l_t\right)\left(\mathbf{x}^l_{t,i} - \bar{\mathbf{x}}^l_t\right)'$.

- The OBC is defined as: $\Psi_{\mathrm{OBC}}(\mathbf{x}) = \mathrm{argmax}_{l\in\{1,\cdots,L\}}\ \mathrm{E}_{\pi^\star}(c^l_t) O_{\mathrm{OBC}}(\mathbf{x}|l)$.

Theorem

If $\mathbf{M}^l_{ts} = \mathbf{0}$ *for all* $l \in \{1,\cdots,L\}$*, then*

$$\Psi_{\mathrm{OBTL}}(\mathbf{x}) = \Psi_{\mathrm{OBC}}(\mathbf{x}), \tag{28}$$

meaning that if there is no interaction between the source and target domains in all the classes a priori, then the OBTL classifier turns to the OBC classifier in the target domain.

Laplace Approximation of Gauss Hypergeometric

- The Gauss hypergeomeric function has the following integral representation:

$$
\begin{aligned}
{}_2F_1(a, b; c; \mathbf{X}) &= B_d^{-1}(a, c-a) \\
&\times \int_{0_d < \mathbf{Y} < \mathbf{I}_d} |\mathbf{Y}|^{a-\frac{d+1}{2}} |\mathbf{I}_d - \mathbf{Y}|^{c-a-\frac{d+1}{2}} |\mathbf{I}_d - \mathbf{X}\mathbf{Y}|^{-b} d\mathbf{Y},
\end{aligned}
\tag{29}
$$

which is valid under the following conditions: $\mathbf{X} \in \mathbf{C}^{d \times d}$ is symmetric and satisfies $\mathrm{Re}(\mathbf{X}) < \mathbf{I}_d$, $\mathrm{Re}(a) > \frac{d-1}{2}$, and $\mathrm{Re}(c-a) > \frac{d-1}{2}$. $B_d(\alpha, \beta)$ is the multivariate beta function

$$
B_d(\alpha, \beta) = \frac{\Gamma_d(\alpha)\Gamma_d(\beta)}{\Gamma_d(\alpha+\beta)}. \tag{30}
$$

- The Laplace approximation is one common solution to approximate the integral

$$
I = \int_{y \in D} h(y) \exp(-\lambda g(y)) dy, \tag{31}
$$

where $D \subseteq \mathbf{R}^d$ is an open set and λ is a real parameter. If $g(\lambda)$ has a unique minimum over D at point $\hat{y} \in D$, then the Laplace approximation to I is given by

$$
\tilde{I} = (2\pi)^{\frac{d}{2}} \lambda^{-\frac{d}{2}} |g^{''}(\hat{y})|^{-\frac{1}{2}} h(\hat{y}) \exp(-\lambda g(\hat{y})), \tag{32}
$$

where $g^{''}(y) = \frac{\partial^2 g(y)}{\partial y \partial y^T}$ is the Hessian of $g(y)$.

- The calibrated Laplace approximation of Gauss hypergeometic functions of matrix argument:

$$
\begin{aligned}
{}_2\hat{F}_1(a,b;c;\mathbf{X}) = \frac{{}_2\tilde{F}_1(a,b;c;\mathbf{X})}{{}_2\tilde{F}_1(a,b;c;\mathbf{0})} &= c^{cd-\frac{d(d+1)}{4}} R_{2,1}^{-\frac{1}{2}} \\
&\times \prod_{i=1}^{d} \left\{ \left(\frac{\hat{y}_i}{a}\right)^{a} \left(\frac{1-\hat{y}_i}{c-a}\right)^{c-a} (1-x_i\hat{y}_i)^{-b} \right\},
\end{aligned}
\tag{33}
$$

where

$$
R_{2,1} = \prod_{i=1}^{d}\prod_{j=i}^{d} \left\{ \frac{\hat{y}_i\hat{y}_j}{a} + \frac{(1-\hat{y}_i)(1-\hat{y}_j)}{c-a} - \frac{b x_i x_j \hat{y}_i \hat{y}_j (1-\hat{y}_i)(1-\hat{y}_j)}{(1-x_i\hat{y}_i)(1-x_j\hat{y}_j)a(c-a)} \right\},
\tag{34}
$$

where $\hat{y}_i$ is defined as

$$
\hat{y}_i = \frac{2a}{\sqrt{\tau^2 - 4ax_i(c-b)} - \tau},
\tag{35}
$$

where $\tau = x_i(b-a) - c$ and $\mathbf{X} = \mathrm{diag}\{x_1, \cdots, x_d\}$.

Experiment results: synthetic data

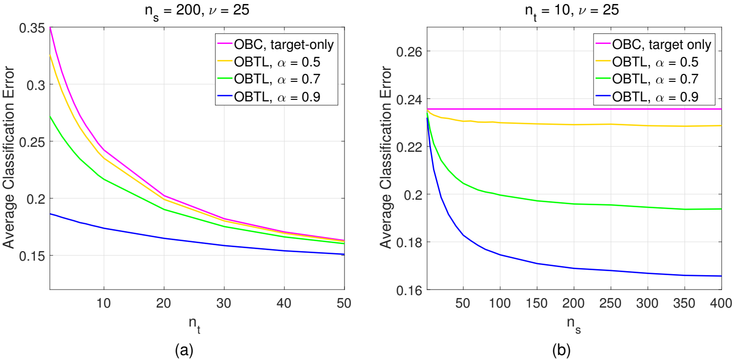


Figure: (a) Average classification error versus the number of target training data per class, n_t. The dimension is $d = 10$, number of source training data per class is $n_s = 200$, and there are $L = 2$ classes in each domain, (b) Average classification error versus the number of source training data per class, n_s. The dimension is $d = 10$, number of target training data per class is $n_t = 10$, and there are $L = 2$ classes in each domain.

Experiment results: synthetic data

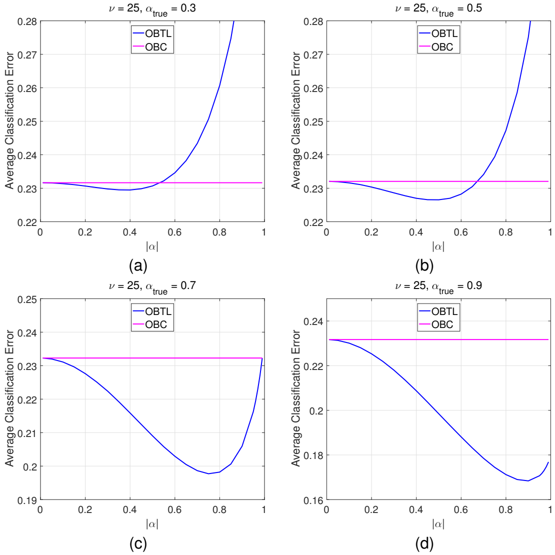
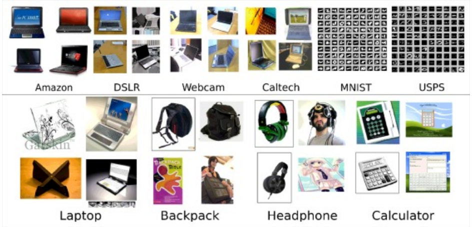


Figure: Average classification error vs $|\alpha|$

Experiment results: benchmark image datasets

- Office and Caltech dataset
- Labeled images from four domains: Amazon website, DSLR camera, Webcam, and Caltech dataset
- Labels are office stuff like laptop, backpack, calculator, ...



Experiment results: benchmark image datasets

Table: Accuracy for different source and target domains in *Office+Clatech256* dataset. Domain names are denoted as a: *amazon*, w: *webcam*, d: *dslr*, c: *Caltech256*. Red shows the best accuracy and blue shows the second best accuracy in each column. The results of the first six methods has been adopted from [1]. Similar to [1], we also used the simulation setup of [6] for the OBTL's results.

	a $\to$ w	a $\to$ d	a $\to$ c	w $\to$ a	w $\to$ d	w $\to$ c	d $\to$ a	d $\to$ w	d $\to$ c	c $\to$ a	c $\to$ w	c $\to$ d	Mean
1-NN-t	34.5	33.6	19.7	29.5	35.9	18.9	27.1	33.4	18.6	29.2	33.5	34.1	29.0
SVM-t	63.7	57.2	32.2	46.0	56.5	29.7	45.3	62.1	32.0	45.1	60.2	56.3	48.9
HFA [7]	57.4	55.1	31.0	**56.5**	56.5	29.0	42.9	60.5	30.9	43.8	58.1	55.6	48.1
MMDT [6]	64.6	56. 7	36.4	47.7	67.0	32.2	46.9	74.1	34.1	49.4	63.8	56.5	52.5
CDLS [8]	**68.7**	**60.4**	35.3	51.8	60.7	33.5	50.7	68.5	34.9	50.9	**66.3**	**59.8**	53.5
ILS (1-NN) [1]	59.7	49.8	**43.6**	54.3	**70.8**	**38.6**	**55.0**	**80.1**	**41.0**	**55.1**	62.9	56.2	**55.6**
OBTL	**72.1**	**60.5**	**42.4**	**54.7**	**76.5**	**37.7**	**53.9**	**84.8**	**40.2**	**54.8**	**70.6**	**61.2**	**59.1**

OBTL for Count Data

- We use the Negative Binomial model for the feature-label distribution in each domain:
$$\mathbf{x}^l_{z,i,j} \sim \mathrm{NB}(\mu^l_{z,i}, r^l_{z,i}), \tag{36}$$
with the probability mass function (PMF)

$$P(\mathbf{x}^l_{z,i,j} = k | \mu^l_{z,i}, r^l_{z,i}) = \frac{\Gamma(k + r^l_{z,i})}{\Gamma(r^l_{z,i})\Gamma(k+1)} \left(\frac{\mu^l_{z,i}}{\mu^l_{z,i} + r^l_{z,i}} \right)^k \left(\frac{r^l_{z,i}}{\mu^l_{z,i} + r^l_{z,i}} \right)^{r^l_{z,i}}, \tag{37}$$

where $z \in \{s, t\}$ denotes the source, s, or target, t, domains; $\mu^l_{z,i}$ and $r^l_{z,i}$ are respectively the mean and shape of the gene i in domain z and class l. The shape parameter is the inverse of the dispersion parameter in Negative Binomial model, which controls the amount of variance. The mean and variance of $\mathbf{x}^l_{z,i,j}$ are

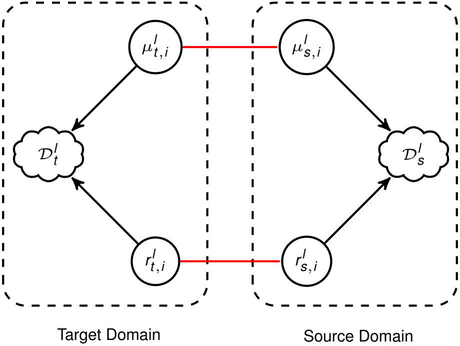
$$\begin{aligned} \mathrm{E}(\mathbf{x}^l_{z,i,j}) &= \mu^l_{z,i}, \\ \mathrm{Var}(\mathbf{x}^l_{z,i,j}) &= \mu^l_{z,i} + \frac{\left(\mu^l_{z,i}\right)^2}{r^l_{z,i}}. \end{aligned} \tag{38}$$

Priors and Posteriors

- Let $\mu = \left\{\mu_{\{s,t\},\{1:d\}}^{\{1:L\}}\right\}$ and $r = \left\{r_{\{s,t\},\{1:d\}}^{\{1:L\}}\right\}$ denote respectively all the mean and shape parameters of the d genes in L classes and two domains s and t. The prior is factorized as

$$p(\mu, r) = \prod_{l=1}^{L} \prod_{i=1}^{d} p\left(\mu_{s,i}^{l}, \mu_{t,i}^{l}\right) p\left(r_{s,i}^{l}, r_{t,i}^{l}\right). \tag{39}$$

- No closed-form posteriors in this model.
- Hamiltonian Monte Carlo (HMC) method is used for posterior sampling, which outperforms other MCMC methods in that it eliminates all the tuning steps.



Joint Prior

Lemma

If $\Lambda \sim W_2(\mathbf{M}, \nu)$, $\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix}$, *and* $\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix}$, *then* $\lambda_{ii} \sim m_{ii}\chi^2_\nu$ *for* $i = 1, 2$, *where* χ^2_ν *denotes the Chi-squared distribution with* ν *degrees of freedom. As a result, the their mean and variance are* $\mathrm{E}(\lambda_{ii}) = \nu m_{ii}$ *and* $\mathrm{Var}(\lambda_{ii}) = 2\nu m_{ii}^2$ *for* $i = 1, 2$. *The covariance and correlation between* λ_{11} *and* λ_{22} *are respectively*

$$\mathrm{Cov}(\lambda_{11}, \lambda_{22}) = 2\nu m_{12}^2, \quad \rho_\lambda = \frac{m_{12}^2}{m_{11} m_{22}}. \tag{40}$$

Theorem ([3])

Let $\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix}$ *be a* 2×2 *Wishart random matrix with* $\nu \geq 2$ *degrees of freedom and positive-definite scale matrix* $\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix}$. *The joint distribution of the two diagonal entries* λ_{11} *and* λ_{22} *have the density function given by*

$$\begin{aligned} p(\lambda_{11}, \lambda_{22}) = {} & K \exp\left(-\frac{1}{2}\left(m_{11}^{-1} + c_2 f^2\right)\lambda_{11}\right) \exp\left(-\frac{1}{2} c_2^{-1}\lambda_{22}\right) \\ & \times (\lambda_{11})^{\frac{\nu}{2}-1} (\lambda_{22})^{\frac{\nu}{2}-1} \, {}_0F_1\left(\frac{\nu}{2}; \frac{1}{4} g\right), \end{aligned} \tag{41}$$

where $c_2 = m_{22} - m_{12}^2 m_{11}^{-1}$, $f = c_2^{-1} m_{12} m_{11}^{-1}$, $g = f^2 \lambda_{11}\lambda_{22}$, $K^{-1} = 2^\nu \Gamma^2\left(\frac{\nu}{2}\right) |\mathbf{M}|^{\frac{\nu}{2}}$, *and* ${}_0F_1$ *is the generalized hypergeometric function.*

Joint Prior

Here ${}_0F_1(b;x) = \sum_{k=0}^{\infty} \frac{x^k}{(b)_k k!}$ is called confluent hypergeometric limit function, which is closely related to the Bessel functions:

$$J_\alpha(x) = \frac{(\frac{x}{2})^\alpha}{\Gamma(\alpha+1)}\, {}_0F_1\left(\alpha+1; -\frac{1}{4}x^2\right). \tag{42}$$

Now, we can define the joint priors of both mean and shape parameters in terms of correlations between two domains:

$$\begin{aligned} p(\mu^l_{s,i}, \mu^l_{t,i}) &= K^l_{\mu,i} \exp\left(-\frac{\mu^l_{s,i}}{2m^l_{s,i}(1-\rho^l_{\mu,i})}\right) \exp\left(-\frac{\mu^l_{t,i}}{2m^l_{t,i}(1-\rho^l_{\mu,i})}\right) \left(\mu^l_{s,i}\right)^{\frac{\nu_\mu}{2}-1} \left(\mu^l_{t,i}\right)^{\frac{\nu_\mu}{2}-1} \\ &\times\ {}_0F_1\left(\frac{\nu_\mu}{2}; \frac{\rho^l_{\mu,i}}{4m^l_{s,i}m^l_{t,i}\left(1-\rho^l_{\mu,i}\right)^2}\mu^l_{s,i}\mu^l_{t,i}\right), \end{aligned} \tag{43}$$

$$\begin{aligned} p(r^l_{s,i}, r^l_{t,i}) &= K^l_{r,i} \exp\left(-\frac{r^l_{s,i}}{2s^l_{s,i}(1-\rho^l_{r,i})}\right) \exp\left(-\frac{r^l_{t,i}}{2s^l_{t,i}(1-\rho^l_{r,i})}\right) \left(r^l_{s,i}\right)^{\frac{\nu_r}{2}-1} \left(r^l_{t,i}\right)^{\frac{\nu_r}{2}-1} \\ &\times\ {}_0F_1\left(\frac{\nu_r}{2}; \frac{\rho^l_{r,i}}{4s^l_{s,i}s^l_{t,i}\left(1-\rho^l_{r,i}\right)^2} r^l_{s,i}r^l_{t,i}\right), \end{aligned} \tag{44}$$

Effective Class-Conditional Densities

- Effective class-conditional density for any new test data in target domain is defined as:

$$p(\mathbf{x}|l) = \int_{\mu_t^l, r_t^l} p(\mathbf{x}|\mu_t^l, r_t^l)\pi^\star(\mu_t^l, r_t^l)d\mu_t^l dr_t^l \tag{45}$$

  for $l \in \{1, \cdots, L\}$, where $\pi^\star(\mu_t^l, r_t^l) = p(\mu_t^l, r_t^l|\mathcal{D}_t^l, \mathcal{D}_s^l)$ is the posterior of (μ_t^l, r_t^l) upon observation of $\mathcal{D}_t^l$ and $\mathcal{D}_s^l$.
- There is no closed form solution for the effective densities. Posterior samples from HMC sampling are used to approximate these effective densities. Suppose we have N posterior samples from all of d genes in L classes. Then the approximation is given by:

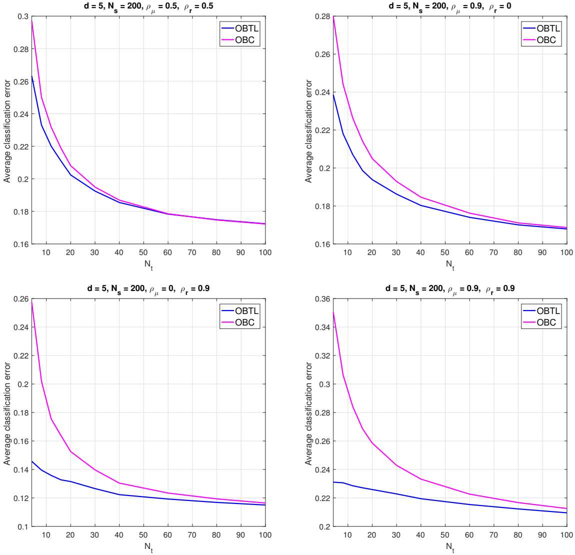
$$p(\mathbf{x}|l) = \frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{d} p(\mathbf{x}_i|\bar{\mu}_{t,i,j}^l, \bar{r}_{t,i,j}^l) \tag{46}$$

  where $\bar{\mu}_{t,i,j}^l$ and $\bar{r}_{t,i,j}^l$ are the $j-th$ posterior sample of gene i in class l of target domain for the mean and shape parameters, respectively.
- The OBTL is given by:

OBTL

$$\Psi_{\mathrm{OBTL}}(\mathbf{x}) = \arg\max_{l\in\{1,\cdots,L\}} \mathrm{E}_{\pi^\star}(c_t^l)p(\mathbf{x}|l). \tag{47}$$

Experiment results: synthetic data



Experiment results: RNA-seq data

- We classify two kinds of lung cancer: LUAD and LUSC
- Data are extracted from The Cancer Genome Atlas (TCGA)
- Two RNA-seq measurements: RNA-seq and RNA-seq-v2. These have different distributions for each genes, so assume two domains:
- Target domain: RNA-seq. LUAD: 125 tumor samples. LUSC: 223 tumor samples
- Source domain: RNA-seq-v2. LUAD: 515 tumor samples. LUSC: 501 tumor samples
- Experiment setup: we randomly generate 50 splits of training (from source and target) and test (only from target) data. We assume $n_s^l = 100$ and $n_t^l = 5$ and number of test data per target class is 100 in each split.
- The average classification error is given for different values of correlations of mean and shape parameters between source and target domains.
- The average error of the OBC is also given for the sake of comparisons.
- Two different sets of features of size $d = 10$ are picked.

Experiment results: RNA-seq data

Case 1:
OBC error = 0.1453
OBTL error:

	$\rho_r = 0.5$	$\rho_r = 0.7$	$\rho_r = 0.9$	$\rho_r = 0.99$
$\rho_\mu = 0.5$	0.1187	0.1184	0.1153	0.1136
$\rho_\mu = 0.7$	0.1193	0.1175	0.1149	0.1139
$\rho_\mu = 0.9$	0.1162	0.1141	0.1130	0.1122
$\rho_\mu = 0.99$	0.1167	0.1127	0.1107	0.1111

Case 2:
OBC error = 0.1936
OBTL error:

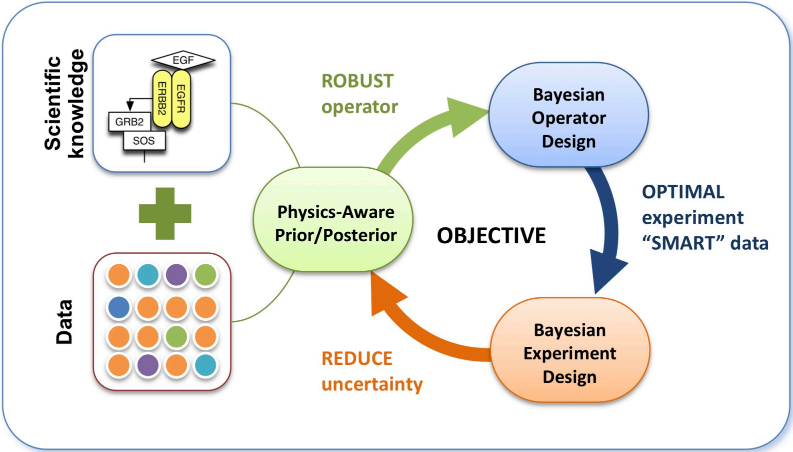
	$\rho_r = 0.5$	$\rho_r = 0.7$	$\rho_r = 0.9$	$\rho_r = 0.99$
$\rho_\mu = 0.5$	0.1654	0.1640	0.1588	0.1543
$\rho_\mu = 0.7$	0.1678	0.1628	0.1571	0.1540
$\rho_\mu = 0.9$	0.1646	0.1619	0.1569	0.1531
$\rho_\mu = 0.99$	0.1631	0.1607	0.1561	0.1513

Conclusions [9]

- We formulate a Bayesian transfer learning framework to transfer source domain knowledge and data for learning in target domain.
- Our Bayesian framework directly models the feature-label distributions in source and target domains.
- The “transferability” across domains can be characterized by a joint prior distribution on model parameters of feature-label distributions across domains.
- We derive the Optimal Bayesian Transfer Learning (OBTL) classifier for both continuous and count data with efficient computational solutions.

Future Research

- Such a Bayesian transfer learning framework enables the closed-loop learning to design experiments for “smart” data and scientific knowledge acquisition.



References

S. Herath, M. Harandi, and F. Porikli, "Learning an invariant Hilbert space for domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3956–3965.

R. J. Muirhead, *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.

K. Halvorsen, V. Ayala, and E. Fierro, "On the marginal distribution of the diagonal blocks in a blocked Wishart random matrix," *International Journal of Analysis*, vol. 2016, pp. 1–5, 2016.

D. K. Nagar and J. C. Mosquera-Benıtez, "Properties of matrix variate hypergeometric function distribution," *Applied Mathematical Sciences*, vol. 11, no. 14, pp. 677–692, 2017.

A. K. Gupta, D. K. Nagar, and L. E. Sánchez, "Properties of matrix variate confluent hypergeometric function distribution," *Journal of Probability and Statistics*, vol. 2016, 2016.

J. Hoffman, E. Rodner, T. Darrell, J. Donahue, and K. Saenko, "Efficient learning of domain-invariant image representations," in *nternational Conference on Learning Representations (ICLR)*, 2013.

L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," *ICML*, 2012.

Y.-H. Hubert Tsai, Y.-R. Yeh, and Y.-C. Frank Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5081–5090.

A. Karbalayghareh, X. Qian, and E. R. Dougherty, "Optimal Bayesian transfer learning," *IEEE Transactions on Signal Processing*, vol. 66, pp. 3724–3739, 2018.