Submitted to Statistical Science

Bayesian Transfer Learning

Piotr M. Suder, Jason Xu, and David B. Dunson

Abstract. Transfer learning is a burgeoning concept in statistical machine learning that seeks to improve inference and/or predictive accuracy on a domain of interest by leveraging data from related domains. While the term "transfer learning" has garnered much recent interest, its foundational principles have existed for years under various guises. Prior literature reviews in computer science and electrical engineering have sought to bring these ideas into focus, primarily by surveying general methodologies and works from these disciplines. This article highlights Bayesian approaches to transfer learning, which have received relatively limited attention despite their innate compatibility with the notion of drawing upon prior knowledge to guide new learning tasks. Our survey encompasses a wide range of Bayesian transfer learning frameworks applicable to a variety of practical settings. We discuss how these methods address the problem of finding the optimal information to transfer between domains, which is a central question in transfer learning. We illustrate the utility of Bayesian transfer learning methods via a simulation study where we compare performance against frequentist competitors.

Key words and phrases: Bayesian machine learning, domain adaptation, hierarchical model, meta analysis.

1. INTRODUCTION

Transfer learning—applying knowledge gained from training on previous tasks and domains to new tasks—is a burgeoning concept in statistics and machine learning. This natural idea mimics some of the mechanisms of human intelligence where past experience, skills and knowledge are often utilized in learning new topics. It is appealing to apply the same paradigm in developing machine intelligence to extract knowledge from the rapidly growing body of datasets available to scientists which are often related to each other in various ways. If the domains between which the transfer of information occurs are sufficiently related, transfer learning can substantially improve the performance of the target model. This is particularly useful when we have a small target dataset we want to study which does not contain enough datapoints to extract precise inferences or predictions, but have access to a large, related dataset.

Piotr M. Suder: PhD Student, Department of Statistical Science, Duke University (e-mail: piotr.suder@duke.edu). Jason Xu: Associate Professor, Department of Biostatistics, University of California Los Angeles (e-mail: jqxu@g.ucla.edu). David B. Dunson: Arts and Sciences Distinguished Professor, Departments of Statistical Science and Mathematics, Duke University (e-mail: dunson@duke.edu).

For instance, suppose that we want to study brain connectomes of Alzheimer's patients or genomes of people suffering from a rare type of cancer. We may utilize large datasets of brain connectomes or genomes collected from healthy individuals such as the ones provided by the UK Biobank to improve the models fitted to the target data. These related sources may aid in the extraction of, say, a low-dimensional latent representation of the complex data we seek to study, which can be useful toward dimensionality reduction in the target domain.

Although the term *transfer learning* has seen increasing popularity in recent years, some of the ideas undergirding it have been around for much longer, and have appeared under various names. Several recent literature reviews aim to help researchers organize and classify these ideas systematically. To name a few, [80], and more recently [79] and [119], provide general overviews of transfer learning methodology, largely from the computer science and electrical engineering literature. [93] focuses on transfer learning in deep neural networks, an appealing use-case due to the data-hungry nature of deep learning models together with the availability of large datasets for training source models. Areas where deep learning is commonly applied such as computer vision often leverage public datasets such as ImageNet [29] or Open Images V4 [59], with millions of datapoints available for training. Meanwhile, [114] focuses on the phenomenon of negative transfer, where the source domains are too different

from the target domain, so that applying transfer learning *worsens* the performance of the target learner. The existence of the negative transfer phenomenon illustrates the importance of choosing an appropriate amount of information to be transferred (the "strength" of transfer) between domains, which remains one of the key challenges in transfer learning and will be one of the focal topics in this survey.

With the exception of [108], none of these reviews of the literature focuses substantially on Bayesian views. Although the work of Xuan *et al.* [108] explicitly overviews Bayesian transfer learning, its scope is limited to probabilistic graphical models. One can argue that the Bayesian paradigm provides a natural framework for how to incorporate prior information from previous datasets within current inferences, and hence provides a canonical umbrella of approaches for transfer learning. In this paper, we provide an overview of some highlights of the Bayesian transfer learning literature. Our focus is on describing how different classical Bayesian approaches can be either directly applied or easily adapted to transfer learning problems. In doing so, we contribute various ideas toward answering a central question of transfer learning: how do we determine and enforce optimal information transfer between domains utilizing various Bayesian modeling approaches? Our aim is to contribute a broad view of Bayesian transfer learning, while presenting approaches that help surmount the problem of negative transfer.

The rest of the paper is organized as follows. In the following section, we give formal definitions of transfer learning and related areas, and discuss alternative names for related ideas in the literature. In Section 3 we provide an overview of general Bayesian approaches to transfer learning with specific examples and some applications. In Sections 4 and 5 we provide a brief taxonomy of transfer learning and point out several areas where some specific Bayesian approaches introduced here can be particularly useful. Finally, in Section 6 we present a simulation study comparing one of the Bayesian methods introduced here with frequentist transfer learning competitors. We conclude with a discussion in Section 7.

2. DEFINITION AND RELATED AREAS

While approaches for transferring information across statistical tasks have a rich history, use of the "transfer learning" terminology is relatively recent. Perhaps, as a result, there is not yet a standard technical definition of what qualifies as transfer learning. Although some authors adopt a narrow definition of the transfer of parameters between models [47], others welcome broader, more general definitions [80], [119], [93]. In this section, we provide one definition of transfer learning to fix ideas for the rest of the article. We then discuss closely related areas. Here by *domain* we denote the two-element set of the form $\mathcal{D} = \{\mathcal{X}, P\}$, where $\mathcal{X}$ is the *feature space* and P is the marginal probability distribution of the observations $X \in \mathcal{X}$ collected in a dataset associated with $\mathcal{D}$. Given a domain $\mathcal{D}$ and its associated *label space* $\mathcal{Y}$, [80] define a *task* on $\mathcal{D}$ as the set $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where f is a function given by $f = \{(x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$. In this framework, f is the ground truth, the optimal solution to the task that is not observed directly, but whose approximation can be learned from the observed data.

DEFINITION 2.1 (**Transfer Learning**). Consider the *source domains* $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ with respective associated *source tasks* $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$, as well as the *target domain* $\mathcal{D}_0$ with the associated *target task* $\mathcal{T}_0 = \{\mathcal{Y}_0, f_0\}$, where an approximation to f_0 can be learned based on available data (X_0, Y_0) with $X_0 \in \mathcal{X}_0, Y_0 \in \mathcal{Y}_0$. Suppose that $\mathcal{D}_k \neq \mathcal{D}_0$ or $\mathcal{T}_k \neq \mathcal{T}_0$ for any $k = 1, \ldots, K$. **Transfer learning** refers to algorithms that aim to improve the approximation of f_0 by incorporating the knowledge from $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$.

In this setting, by *knowledge* we mean either: (i) the raw data sampled from the source domains, possibly equipped with labels from the source tasks, (ii) learners pre-trained on the data from source domains and tasks, or (iii) oracle models which have complete and true information on the source domains and the source tasks. Among these, cases (i) and (ii) are the most commonly encountered ones in practice.

Our definition follows and generalizes the conventions used by [80]. Note that in the above definition, the source and target domains are not necessarily different, encompassing cases with a common domain but different tasks. Furthermore, when two domains are different, their feature spaces need not differ. In the deep learning literature, the target task is sometimes referred to as the *downstream task* [87].

By allowing the label space to be (a) discrete, (b) one-dimensional, (c) multidimensional, (d) defining only a partition of a dataset associated with $\mathcal{D}$ without giving a specific meaning to how the labels are used for that purpose, this definition comprises, respectively, (a) classification, (b) univariate regression, (c) multivariate regression and dimensionality reduction, and (d) clustering tasks. Finally, allowing the values of f to be probability distributions naturally lends itself to Bayesian posterior learning.

2.1 Related fields

Another closely related problem that follows this paradigm is *multitask learning*. Like transfer learning, its goal is to improve learning in a particular domain based on information from related domains and tasks, but it differs in

the attempt to simultaneously learn each task jointly on all the domains considered. This may improve the performance across tasks by borrowing information between related tasks and domains, in contrast to using a set of tasks and domains only as means to the end of improving performance on a single *target task* [80]. Although some authors regard multitask learning and transfer learning as separate disciplines [80], [109], [98], others consider them as the same field [108], or draw a distinction between the two according to different criteria, as in [40]. Often a multitask learning method can be easily adapted to transfer learning [80], [109]. As we shall see in the following sections, the Bayesian framework elegantly reconciles these notions in many cases.

Continual or *lifelong learning* [52], [115], [58], [34] is a popular concept in machine learning, which combines aspects of transfer and multitask learning. In this setting, an agent faces a sequence of domain-task pairs over time, with the goal being to utilize previously encountered tasks to learn each new task in a more effective way while maintaining the ability to solve the previous tasks [106]. Continual learning attempts to remedy the phenomenon of catastrophic forgetting [75] in transfer learning where the model performs worse on the source tasks after being adjusted to the target task; this commonly occurs in deep learning models [9], [2]. This forgetting can be especially problematic when the number of encountered tasks and model parameters becomes large and it becomes difficult to store the previously encountered datasets and models trained on them.

There is additionally a Bayesian literature on *metalearning*, or *learning to learn* [110], [81], [82], [113]. Vanschoren [97] defines metalearning as methods aiming to improve the "configuration" (e.g. model hyperparameters, network architecture in case of deep learning methods, etc.) of the model for the target task by training on *metadata*. Here metadata refers to information obtained from models trained individually with different configurations; for example, one may vary different aspects of the model and measure its performance via cross validation. While some researchers consider metalearning as distinct from transfer learning [47], following Definition 2.1 we consider it to be a special case of transfer learning. In this case the information from the source domains is utilized by training models with various configurations on these domains and then using the metadata generated from them in improving the model for the target task.

Domain adaptation [104], [116], [8], [37], [102] is another popular term, sometimes used interchangeably with transfer learning as in [56]. However, since knowledge can also be transferred between different tasks in the same domain, we view it as a particular case of transfer learning. Additional terminology for concepts closely related to transfer learning includes cooperative learning [117], [31], knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, incremental, and cumulative learning [80].

3. BAYESIAN APPROACHES TO TRANSFER LEARNING

Two fundamental questions that need to be addressed are: (i) *how* information should be transferred, and (ii) *which* information should be transferred. There are various approaches to answering these questions and they are often related to the models used for solving the source and target tasks. Determining appropriate information transfer between domains is critical, since transferring inappropriate information can result in large bias and suboptimal performance. In extreme cases, one obtains *negative transfer* [114], which corresponds to the case in which transferring information decreases performance.

Some of the existing approaches rely on expert knowledge about the domains considered and their relationships, some introduce statistical measures of similarity between domains, while others rely on more flexible model-based or validation-based approaches to the optimal choice of parameters controlling information transfer. In this section, we discuss different ideas based on Bayesian methodology which can be used to tackle questions (i) and (ii).

3.1 Shared parameters

One of the most prevalent approaches is to use common parameters in the source and target domains. For exposition, throughout this subsection we assume only one source dataset X_S and one target dataset X_T. For convenience of notation, by X_d we denote both the datapoints and their associated labels (when applicable) for domain $d \in \{S, T\}$. We parameterize the likelihood of the data for the source and target domains as $p(X_S \mid \theta_C, \theta_S)$ and $p(X_T \mid \theta_C, \theta_T)$, respectively, where θ_C is the common vector of parameters, shared by the source and target data, while θ_S and θ_T are vectors of parameters unique to the datasets. Let $\pi(\theta_C)$, $\pi(\theta_S)$, $\pi(\theta_T)$ be the prior distributions for, respectively, θ_C, θ_S, θ_T.

A simple Bayesian transfer learning approach would compute the posterior for θ_C based on the prior $\pi(\theta_C)$ and the source data X_S via

$$p(\theta_C \mid X_S) \propto p(X_S \mid \theta_C)\pi(\theta_C) \tag{1}$$
$$= \left(\int p(X_S \mid \theta_C, \theta_S)\pi(\theta_S) d\theta_S \right) \pi(\theta_C),$$

and then use $\pi^*(\theta_C, \theta_T) \propto p(\theta_C \mid X_S)\pi(\theta_T)$ as the prior for (θ_C, θ_T) in the analysis of X_T to obtain the posterior

$$p^*(\theta_C, \theta_T \mid X_T) \propto p(X_T \mid \theta_C, \theta_T)\pi^*(\theta_C, \theta_T).$$

It is straightforward to see that if $X_T \perp\!\!\!\perp X_S \mid (\theta_C, \theta_T)$ and $\theta_S \perp\!\!\!\perp \theta_T$ a priori, then this is equivalent to obtaining

a posterior for (θ_C, θ_T) based on the data (X_T, X_S) with the prior $\pi(\theta_C)\pi(\theta_T)$ on (θ_C, θ_T), i.e.

$$p^*(\theta_C, \theta_T \mid X_T) = p(\theta_C, \theta_T \mid X_T, X_S), \tag{2}$$

where

$$p(\theta_C, \theta_T \mid X_T, X_S) \propto p(X_T, X_S \mid \theta_C, \theta_T)\pi(\theta_C)\pi(\theta_T).$$

Hence, this approach is equivalent to giving equal weights to the source and target data in computing the posterior of the shared parameters. This is an appropriate approach when the model is specified correctly and the true parameters θ_C are indeed exactly the same in the source and target populations.

However, in practice, it is likely that the assumption of exactly equivalent values of θ_C is an oversimplification. As the true values of θ_C vary more widely between the source and target domains, the above approach can have suboptimal performance, particularly when the sample size of the source data is larger than that of the target, which is often the case. A simple and commonly used heuristic solution is to specify the prior for θ_C in the target posterior as a variance-inflated version of the posterior $p(\theta_C \mid X_S)$ from the source data analysis.

Shwartz-Ziv *et al.* [87] apply a related approach to Bayesian deep neural networks (DNNs). First, a Gaussian approximation to the posterior for the DNN fitted to the source data is obtained. The authors assume that the "feature extractor" layers of the DNN are common to the source and target DNN. The variance of the Gaussian approximation to the source posterior for the weights in these layers is scaled up by a constant factor and then used as a prior for the feature extractor component in the target data DNN. The remaining weights characterizing the "head" of the DNN are given an isotropic Gaussian prior. To learn an appropriate amount of information sharing between the source and target domains, the scaling factor is chosen on held-out validation data from the target training dataset.

An alternative approach to controlling the influence of the source data on the target domain posterior distribution is the *power prior* [22], [23], [50], [49], [32]. The power prior for the target parameters is proportional to an initial prior multiplied by the source data likelihood raised to a fractional power. The fractional power serves to diminish the information provided by the likelihood of the source data. In our transfer learning setting, the joint prior for (θ_C, θ_T) in the target model is given by

$$\pi_{a_0}(\theta_C, \theta_T \mid X_S) \propto p(X_S \mid \theta_C)^{a_0}\pi(\theta_C)\pi(\theta_T), \tag{3}$$

where the strength of information transfer ranges between no transfer at $a_0 = 0$ to "full" transfer at $a_0 = 1$. In the latter case, the source data are given equal weight to those in the target domain. This setup generalizes the partial borrowing power prior of [23], where the source model parameters are a subset of those used in the target domain.

Several appealing theoretical properties of the power prior were established in [50] for the case when all the parameters are shared. In that case the posterior for θ reduces to

$$\pi_{a_0}(\theta \mid X_T, X_S) \propto p(X_T \mid \theta)p(X_S \mid \theta)^{a_0}\pi(\theta), \tag{4}$$

where θ determines the distribution of both source and target data. Ibrahim *et al.* [50] show that for a fixed a_0, (4) minimizes the weighted sum of Kullback–Leibler (KL) divergences between the posterior with no information transfer and one with full information transfer, i.e.

$$\pi_{a_0}(\theta \mid X_T, X_S) = \\ = \arg\min_g \{(1-a_0)KL(g \,||\, f_0) + a_0 KL(g \,||\, f_1)\},$$

where f_0 and f_1 are probability densities given by

$$f_0(\theta) = \pi_0(\theta \mid X_S, X_T) \propto p(X_T \mid \theta)\pi(\theta)$$

and

$$f_1(\theta) = \pi_1(\theta \mid X_S, X_T) \propto p(X_T \mid \theta)p(X_S \mid \theta)\pi(\theta).$$

Like in the other transfer learning approaches, choosing the right amount of information to be transferred, in this case governed by the value of a_0—is a key challenge. An approach is to treat a_0 as fixed and perform a sensitivity analysis on a set of values that ideally should include $a_0 = 0$ and $a_0 = 1$, as recommended by [49]. In generalized linear models (GLMs) the choice of a_0 can be better informed with the help of model selection criteria such as those proposed in [50], [49], [48], and [92]. Ibrahim *et al.* [50] propose a penalized likelihood-type criterion (PLC) that chooses $a_0 \in (0, 1]$ to be the minimizer of

$$-2\log \int p(X_T \mid \theta)p(X_S \mid \theta)^{a_0}\pi(\theta)d\theta + \frac{\log(n_S)}{a_0},$$

where n_S is the sample size of the source dataset.

Alternatively, we can treat a_0 as random and, in turn, assign it a prior distribution. We can directly define the joint prior for (θ, a_0) as in [49], i.e.

$$\pi(\theta, a_0 \mid X_S) \propto p(X_S \mid \theta)^{a_0}\pi(\theta)\pi(a_0), \tag{5}$$

or the *normalized power prior* as in [32]

$$\pi(\theta, a_0 \mid X_S) = \pi(\theta \mid X_S, a_0)\pi(a_0) \\ = \frac{p(X_S \mid \theta)^{a_0}\pi(\theta)}{\int p(X_S \mid \theta')^{a_0}\pi(\theta')d\theta'}\pi(a_0). \tag{6}$$

The normalized power prior first specifies a marginal prior for a_0 and then a conditional prior for θ given a_0.

Taking $\pi(a_0)$ to be a beta or Dirichlet distribution depending on the number of source domains is a natural choice, with theoretical support proved in [50] under fixed a_0 that extends to the random a_0 case under (5). Other priors with appropriate support, such as gamma or Gaussian truncated to $[0, 1]$, can also be utilized [22]. However, it is

not clear how the data inform about an appropriate value for a_0, since a_0 is not a traditional parameter. It may be that this approach can be used to represent the prior uncertainty in a_0 but will not adapt to the information in the data to focus on the optimal amount of borrowing from the source data.

The model introduced by Hickey et al. [45] leads to another interesting example of the shared variables approach. Specifically, given source and target datasets $\{(y_{S,i}, \boldsymbol{x}_{S,i})\}_{i=1}^{n_S}$ and $\{(y_{T,j}, \boldsymbol{x}_{T,j})\}_{j=1}^{n_T}$, each consisting of labeled i.i.d. samples, they specify the following model

$$y_{S,i} = h\left[f(\boldsymbol{\theta}_S, \boldsymbol{x}_{S,i}), U_{S,i}\right]$$
$$y_{T,j} = h\left[g(\boldsymbol{\theta}_T, \boldsymbol{x}_{T,j}), U_{T,j}\right],$$

where $U_{S,1}, \ldots U_{S,n_S}$ and $U_{T,1}, \ldots, U_{T,n_T}$ are independent random variables and f, g and h are functions. Under the assumption that $f(\boldsymbol{\theta}_S, \boldsymbol{x}_T) \neq 0$ almost surely, they let

$$\beta_j = g(\boldsymbol{\theta}_T, \boldsymbol{x}_{T,j}) / f(\boldsymbol{\theta}_S, \boldsymbol{x}_{T,j})$$

and then reformulate the target data model as

$$y_{T,j} = h\left[\beta_j f(\boldsymbol{\theta}_S, \boldsymbol{x}_{T,j}), U_{T,j}\right].$$

They assume that $\beta_j \overset{\text{iid}}{\sim} \text{Cauchy}(\delta, \gamma)$ and specify a prior for the parameters δ and γ. Thus, while in the original formulation the true parameters for both domains are not necessarily equal, by explicitly modeling a relevant measure of discrepancy between them represented by β_j's, Hickey et al. [45] are able to easily reformulate the model in a way that allows them to take advantage of the shared parameter approach.

Similar effect is achieved in the work of Li et al. [67], who model the parameter vector that governs the categorical distribution of the target domain data as a mixture of the corresponding parameter vectors from the source domains. In both these works transfer via common parameters is achieved by introducing an additional layer to the model, which brings us to the next approach commonly employed in transfer learning problems, namely that of hierarchical modeling.

3.2 Hierarchical models and random effects

The methods mentioned in the previous section rely on sharing parameters in the likelihood specification for source and target data. Alternatively, we can allow the parameters of the source and target data models to differ, instead imposing the assumption that they come from a jointly specified or identical prior distribution acting as a bridge for information flow between the domains.

As a simple example, consider the Gaussian linear model. Let the datasets $(\boldsymbol{X}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{X}_K, \boldsymbol{y}_K)$ denote the source data and $(\boldsymbol{X}_0, \boldsymbol{y}_0)$ be the target data, where $\boldsymbol{X}_d \in \mathbb{R}^{n_d \times p}$ and $\boldsymbol{y}_d \in \mathbb{R}^{n_d}$ for $d \in \{0, 1, \ldots, K\}$. Under this model we assume that

$$\boldsymbol{y}_d = \boldsymbol{X}_d \boldsymbol{\beta}_d + \boldsymbol{\epsilon}_d, \quad \boldsymbol{\epsilon}_d \sim \mathcal{N}(0, \sigma_d^2 I_{n_d}), \tag{7}$$

with the prior on the coefficients given by $\boldsymbol{\beta}_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $d \in \{0, 1, \ldots, K\}$. Here, domain-specific parameters $\boldsymbol{\beta}_d$ are drawn from a common prior distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is often referred to as a random effects distribution. Model (7) is a common type of hierarchical regression model for data nested within groups (domains in our terminology). Data from all domains are used to inform the mean of random effects $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, causing the borrowing of information.

We can either treat $\sigma_0, \sigma_1, \ldots, \sigma_K$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as fixed, taking a frequentist approach to inference, or specify hyperpriors for them to obtain a Bayesian hierarchical model. In either case, the random effects covariance $\boldsymbol{\Sigma}$ controls how much information transfer there is, analogously to a_0 in the power prior approach. Large covariance implies less shrinkage of the $\boldsymbol{\beta}_d$ values towards the random effects mean $\boldsymbol{\mu}$. In practice, the prior for the random effects mean and covariance will be updated based on information in the data about the variability in the regression coefficients across domains.

For fixed $\sigma_0, \sigma_1, \ldots, \sigma_K$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, [21] showed a direct analytic relationship between $\boldsymbol{\Sigma}$ and the tuning parameter a_0 in the power prior approach, establishing duality between these methods for the Gaussian linear model.

3.3 Shared latent space

Rather than imposing shared parameters on data generating processes for source and target domains, whether explicitly in the likelihoods or at higher levels in a hierarchical model, we can also specify or seek to learn a shared latent space. This approach can be particularly useful in more complex datasets with a large number of dimensions.

3.3.1 Factor analysis In the Bayesian context, many such examples can be found in the factor analysis literature. Under the classical factor model specification outlined in [70] the i-th observation $\boldsymbol{y}_i \in \mathbb{R}^p$ is given by

$$\boldsymbol{y}_i = \boldsymbol{\Lambda} \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q)$ are the vectors of latent factors, $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times q}$ is the factor loading matrix, $\boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Delta})$ are random noise terms with $\boldsymbol{\Delta} = \text{diag}(\delta_1^2, \ldots, \delta_p^2)$, and $\boldsymbol{\eta}_i, \boldsymbol{\epsilon}_j$ are independent for any i, j. It is commonly assumed that $q \ll p$, that is, the high-dimensional data can be explained using a latent structure of much lower-dimensional factors. This model can be equivalently written as a Gaussian distribution with a constrained covariance structure, i.e.

$$\boldsymbol{y}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Delta}.$$

The mean zero assumption on $\boldsymbol{y}_i$ comes from the standard practice of centering the data and does not limit the generality of the model.

In [27] and [28] this setup is generalized to the situation with data coming from multiple domains by letting

$$\boldsymbol{y}_{k,i} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k), \quad \boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^T + \boldsymbol{\Delta}_k,$$

where $\boldsymbol{\Delta}_k = \text{diag}(\delta_{k,1}^2, \ldots, \delta_{k,p}^2)$ is the error variance matrix, $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Phi}_k \in \mathbb{R}^{p\times q_k}$ for domain $k = 1, \ldots, K$. Here, $\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^T$ accounts for the domain-specific dependencies between the datapoints, and $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ is the underlying shared covariance structure which allows for information transfer between domains.

Analogously to the single domain case above, this model has the equivalent representation

$$\boldsymbol{y}_{k,i} = \boldsymbol{\Lambda}\boldsymbol{\eta}_{k,i} + \boldsymbol{\Phi}_k\boldsymbol{\zeta}_{k,i} + \boldsymbol{\epsilon}_{k,i}, \tag{8}$$

with

$$\boldsymbol{\eta}_{k,i} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q), \quad \boldsymbol{\zeta}_{k,i} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{q_k}), \quad \boldsymbol{\epsilon}_{k,i} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Delta}_k),$$

where $\boldsymbol{\eta}_{k,i}$ is a latent factor in the q dimensional shared subspace, $\boldsymbol{\zeta}_{k,i}$ are q_k dimensional domain-specific latent factors and $\boldsymbol{\epsilon}_{k,i}$ are error terms. Thus, $\boldsymbol{\Lambda}$ is the shared factor loading matrix and $\boldsymbol{\Phi}_k$ are $p \times q_k$ domain-specific factor loading matrices. In this model the transfer of knowledge between domains occurs through information borrowing in the estimation of $\boldsymbol{\Lambda}$.

The above model allows for a lot of flexibility between the domains, but suffers from an identifiability issue known as information switching: the data can be fitted equally well with the shared columns in factor loading matrices transferred from $\boldsymbol{\Lambda}$ to $\boldsymbol{\Phi}_k$'s. De Vito *et al.* [27] solve this problem by restricting the augmented matrix $[\boldsymbol{\Lambda}\ \boldsymbol{\Phi}_1 \ldots\ \boldsymbol{\Phi}_K]$ to be lower triangular. One limitation of this approach is that it imposes an ordering on the domains. Often when we have multiple source domains there is no natural ordering between them and hence this approach would not be preferred in such a scenario.

A recent paper proposes a different solution to the problem of information switching by restricting factor loading matrices to linear transforms of the shared factor loading matrix and imposing a shared covariance of error terms between domains [18]. That is, they assume $\boldsymbol{\Phi}_k = \boldsymbol{\Lambda}\boldsymbol{A}_k$, where $\boldsymbol{A}_k \in \mathbb{R}^{q\times q_k}$ and $\boldsymbol{\Delta}_k = \boldsymbol{\Delta} = \text{diag}(\delta_1^2, \ldots, \delta_p^2)$ for every $k = 1, \ldots, K$ in (8). The authors show that under any non-degenerate continuous prior on $\boldsymbol{A}_k$ the information switching does not occur almost surely provided that $0 < \sum_{k=1}^K q_k \leq q$.

This result provides some guidance for choosing the dimensions of shared and domain-specific latent spaces, since these are not known in most practical applications. These dimensions influence the amount of information transferred between domains, as higher values of $q_1, \ldots, q_K$ give more flexibility to domain-specific latent features, thus reducing the influence of shared latent factors. An approach to choosing these dimensions would be to place priors in $q_1, \ldots, q_K$ and q and use the reversible jump algorithms outlined in [41]. However, such algorithms can be computationally prohibitive.

Chandra *et al.* [18] provide an alternative solution by fixing $q, q_1, \ldots, q_K$ at an upper limit, and then utilizing appropriate priors to shrink the excess columns in $\boldsymbol{\Lambda}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_K$. Specifically, they obtain approximate singular values and eigenvectors of the pooled dataset via the augmented implicitly restarted Lanczos bidiagonalization [3]. Then they choose $\hat{q}$ to be the lesser of: the smallest number of factors explaining 95% of the variability in the data; the largest integer smaller than $(2p - \sqrt{8p+1})/2$, in order to ensure identifiability of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ from $\boldsymbol{\Delta}$, following [6]. Then they fix $\hat{q}_k = \hat{q}/K$ for $k = 1, \ldots, K$ to ensure that information switching does not occur. The strength of information transfer between domains is thus directly influenced by the amount of shrinkage induced by the priors on the factor loading matrices. In [18], the fixed priors $\text{vec}(\boldsymbol{\Lambda}) \sim \text{DL}(1/2)$ and $a_{k,i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ are used, where $a_{k,i,j}$ is the (i, j)-th entry of $\boldsymbol{A}_k$ and DL denotes the Dirichlet-Laplace distribution [7]. Thus, a possible extension in this line of work would be to propose some methods for choosing these hyperparameters in an adaptive way depending on how related the domains are.

3.3.2 Mixture models Latent space models are commonly used as a dimensionality reduction tool, including when dealing with non-standard data structures such as networks [46], [10]. As a vignette showing how mixture models can be used in combination with latent space models for flexible transfer learning, we focus on the approach proposed in [33]. They were motivated by data on brain networks for individuals in different groups.

Specifically, given n observed networks each belonging to one of K groups and consisting of V labeled vertices, denote the i-th network together with its group label as $\{y_i, \boldsymbol{\mathcal{L}}(\boldsymbol{A}_i)\}$, where $\boldsymbol{A}_i \in \{0, 1\}^{V\times V}$ is the adjacency matrix and $\boldsymbol{\mathcal{L}}(\boldsymbol{A}_i) \in \{0, 1\}^{V(V-1)/2}$ denotes the lower triangular entries

$$(A_{i[2,1]}, \ldots, A_{i[V,1]}, A_{i[3,2]}, \ldots, A_{i[V,2]}, \ldots, A_{i[V,V-1]})^T$$

of $\boldsymbol{A}_i$. We ignore the main diagonal and the upper triangular part of $\boldsymbol{A}_i$ since the network is an undirected graph and the self-relationships of the nodes are not of interest. In [33] subjects fall into a low- and high-creativity group, so we have $K = 2$ domains. The network representation $\boldsymbol{\mathcal{L}}(\boldsymbol{A})$ conditional on the group membership y is modeled as

$$\mathbb{P}(\boldsymbol{\mathcal{L}}(\boldsymbol{A}_i) = \boldsymbol{a} \mid y = k) = \\ = \sum_{h=1}^{H} \nu_h^{(k)} \prod_{l=1}^{V(V-1)/2} \left(\pi_l^{(h)}\right)^{a_l} \left(1 - \pi_l^{(h)}\right)^{1-a_l}$$

for any $\boldsymbol{a} \in \{0,1\}^{V(V-1)/2}$ with the probability vector

$$\boldsymbol{\pi}^{(h)} = \left(\pi_1^{(h)}, \ldots, \pi_{V(V-1)/2}^{(h)}\right)^T \in (0,1)^{V(V-1)/2}$$

in the h-th mixture component given by

$$\pi_l^{(h)} = \left[1 + \exp\left(-Z_l - D_l^{(h)}\right)\right]^{-1},$$

with

$$\boldsymbol{D}^{(h)} = \boldsymbol{\mathcal{L}}(\boldsymbol{X}^{(h)}\boldsymbol{\Lambda}^{(h)}\boldsymbol{X}^{(h)T}), \quad h = 1, \ldots, H,$$

where $\boldsymbol{X}^{(h)} \in \mathbb{R}^{V \times R}$, $\boldsymbol{\Lambda}^{(h)} = \text{diag}(\lambda_1^{(h)}, \ldots, \lambda_R^{(h)})$ with $\lambda_1^{(h)}, \ldots, \lambda_R^{(h)} \geq 0$, and $\boldsymbol{Z} \in \mathbb{R}^{V(V-1)/2}$.

The model supposes that there are H different types of brain structures. The probability of an edge between the l-th pair of brain regions follows a logistic model having an intercept Z_l characterizing the baseline log odds of a connection and a low-rank deviation that differs according to the individual's brain type. To enable information transfer across the creativity groups (domains), the model assumes the brain structure types do not differ across the groups (referred to as "common atoms" in the mixture modeling literature). However, the proportion of individuals having brain type h, $\nu_h^{(k)}$, does differ across domains $k = 1, \ldots, K$.

Although the goal in [33] was inference on group differences, this model can be used directly to transfer learning from the source domain to a target domain, the source data allowing a more accurate estimation of the shared network types. In addition, the baseline log-odds of an edge between each pair of nodes is also shared across the groups, leading to information sharing about common topological properties of the graphs, including block structures, homophily behaviors, and transitive edge patterns [46]. An important application would be to transferring information from large brain imaging repositories, such as the Human Connectome Project (HCP) and UK Biobank, to small neuroimaging studies in targeted populations.

In [68], shared kernels are used to model complex distributions of multiple variables in different domains. Motivating applications are studies investigating how DNA methylation profiles vary according to the cancer subtype. For samples $i = 1, \ldots, n$, data consist of $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ with x_{ij} denoting the methylation level at site j, for $j = 1, \ldots, p$ with p very large (e.g., $p = 450,000$) and $y_i \in \{1, \ldots, K\}$ denoting the group membership. The density of the data in group k for the jth variable is $f_j^{(k)}(\cdot) = \sum_{h=1}^H \nu_{jh}^{(k)} \mathcal{K}(\cdot; \boldsymbol{\theta}_h)$, with $\mathcal{K}(\boldsymbol{\theta})$ a family of densities parameterized by $\boldsymbol{\theta}$ and $\nu_{jh}^{(k)}$ a probability weight on kernel h specific to site j and group k.

Although the motivation in [68] is testing for differences in methylation between different groups, the proposed approach can be directly applied to transfer learning focused on inferring the marginal densities of very high-dimensional data within a particular domain. Data from all domains are used to infer the shared kernel parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H$ and further information borrowing occurs through a hierarchical model for the weights $\{\nu_{jh}^{(k)}\}$. Even in using shared kernels, this approach allows for highly flexible differences in distribution across groups.

There is a rich literature on alternative Bayesian mixture models for borrowing information across grouped data while also allowing distinct characteristics of each group. Suppose we let $\boldsymbol{x}_i$ denote feature data for subject i with $y_i \in \{1, \ldots, K\}$ denoting the subject's group membership. Then, a common approach is to incorporate subject-specific parameters $\boldsymbol{\theta}_i$ within the likelihood function for $\boldsymbol{x}_i$ and then let $\boldsymbol{\theta}_i \sim P_{y_i}$, with the collection of group-specific random effects distributions $(P_1, \ldots, P_K) \sim \Pi$ given an appropriate prior. Popular choices of Π include the hierarchical Dirichlet process (HDP) [95] and the nested Dirichlet process (NDP) [83], both of which fall into the broad class of hierarchical processes [12]. These approaches characterize each P_k as almost surely discrete while incorporating statistical dependence between P_k and P_l for all $k \neq l$, leading to dependence in clustering.

Alternatively, analogously to the multi-group factor models of [27], [28], and [18], Müller *et al.* [78] modeled the group-specific random effects distributions P_k as a mixture of a common cross-group distribution P_0 and group-specific distributions Q_k, with a hyperprior chosen for the mixture weight on P_0 to allow data adaptivity. This approach and the above approaches assume *a priori* exchangeability between groups. In the future, it will be interesting to adapt these approaches and develop appropriate extensions explicitly targeting the transfer learning case in which one domain is the particular focus. A relevant recent advance is the graphical Dirichlet process of Chakrabarti *et al.* [16], which incorporates a known directed acyclic graph (DAG) characterizing the dependence structure across groups.

3.4 Network transfer

As an alternative to viewing each source domain as providing exchangeable information about the target domain *a priori*, there is often expert knowledge about directed relationships between the different domains. Incorporating a network of relationships among domains in transfer learning is termed *network transfer*, as opposed to *direct transfer*.

In Bayesian network meta-analysis [71] the goal is often to compare the efficacy of a pair of treatments based on multiple studies, some of which may involve arms with other treatments. Let W, X, Y, Z be four available treatments among which we want to compare the efficacy of X and Y. Suppose that we have the dataset $\mathcal{D}_{XY}$ formed

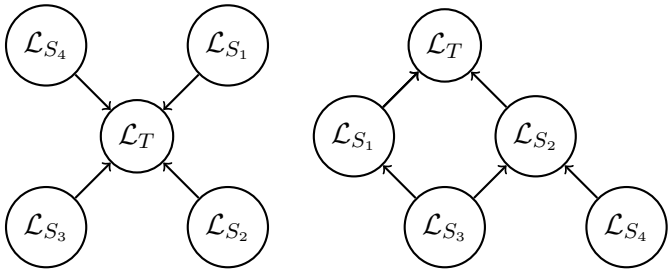

FIG 1. *Direct transfer (left) and network transfer learning (right). In direct transfer all the source learners are used directly in supporting the training of the target learner, whereas in network transfer we can have a more complex structure with some of the source learners supporting other source learners rather than the target learner directly.*

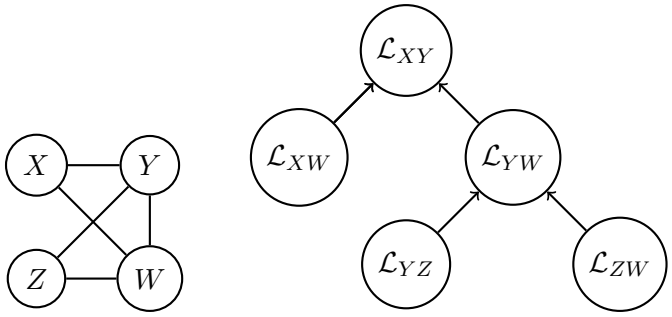

FIG 2. *Evidence network for the treatment comparison (left) and the network transfer of information between the associated learners in the meta analysis (right).*

based on studies comparing X and Y and that we also have access to datasets $\mathcal{D}_{XW}$, $\mathcal{D}_{YW}$, $\mathcal{D}_{YZ}$, and $\mathcal{D}_{WZ}$ comparing, respectively X to W, Y to W, Y to Z, and W to Z. We may have several different trials for certain of these comparisons. The knowledge extracted from the trials comparing other treatments can be used to indirectly improve the analysis of the X vs Y trials. Figure 2 shows a graph representing the observed comparisons between treatments, sometimes referred to as the evidence network [72], and the associated network transfer graph.

The general framework for Bayesian network meta-analysis is outlined in [72]. Denote the mean difference observed in the efficacy of treatments k and l in study i by $\delta_{i,k,l}$ and the baseline difference in efficacy between treatments k and l by $d_{k,l}$. We refer to $d_{k,l}$ as effect parameters. In [72] they are divided into basic parameters $\boldsymbol{d}_b$ and functional parameters $\boldsymbol{d}_f$. Any set of effect parameters can be treated as basic parameters if the edges associated with them create a spanning tree of the evidence network. The functional parameters are the remaining effect parameters.

Network meta-analysis assumes that functional parameters can be represented as linear functions of basic parameters, that is, $\boldsymbol{d}_f = \boldsymbol{F}\boldsymbol{d}_b$ for some matrix $\boldsymbol{F}$. This assumption is referred to as *evidence consistency*. Usually, these relations take the form $d_{j,k} = d_{j,l} - d_{k,l}$ for any treatments j, k, l. In our example, we can choose $d_{X,W}, d_{Y,W}, d_{Z,W}$ as the basic parameters and then relate the functional parameters to them through $d_{X,Y} = d_{X,W} - d_{Y,W}$ and $d_{Y,Z} = d_{Y,W} - d_{Z,W}$. Leveraging this assumption is analogous to utilizing the shared parameter strategy outlined in Section 3.1. We can use these identities to increase the precision of the estimation of $d_{Y,W}$, which, in turn, together with the estimates of $d_{X,W}$, can increase the precision of the estimation of $d_{X,Y}$. This is represented by the network transfer in Figure 2.

The linear relationship $\boldsymbol{d}_f = \boldsymbol{F}\boldsymbol{d}_b$ can be used to model the vector of observed differences in treatments $\boldsymbol{\delta}$ conditionally on $\boldsymbol{d}_b$ and the covariance of $\boldsymbol{\delta}_b$, denoted by $\mathrm{Cov}(\boldsymbol{\delta}_b) = \boldsymbol{V}_b$. Using the Gaussian distribution $\boldsymbol{\delta} \sim \mathcal{N}\left(\left(\boldsymbol{d}_b^T, \boldsymbol{d}_b^T\boldsymbol{F}^T\right)^T, \boldsymbol{V}\right)$ is standard [71], [30], [64], where

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_b & \boldsymbol{V}_b\boldsymbol{F}^T \\ \boldsymbol{F}\boldsymbol{V}_b^T & \boldsymbol{F}\boldsymbol{V}_b\boldsymbol{F}^T \end{pmatrix}.$$

This prior can be incorporated within a hierarchical model for the individual observations in each study. Further borrowing of information can be facilitated by placing a common random-effects distribution on the basic treatment effect parameters as in [64].

Additional flexibility in information transfer can come from allowing for violations in evidence consistency. Lu *et al.* [72] provide such a framework through $\boldsymbol{d}_f = \boldsymbol{F}\boldsymbol{d}_b + \boldsymbol{w}$, where $\boldsymbol{w}$ represents inconsistencies between studies. In our example

$$d_{X,Y} = d_{X,W} - d_{Y,W} + w_{X,Y,W}$$

and

$$d_{Y,Z} = d_{Y,W} - d_{Z,W} + w_{Y,Z,W}.$$

The inferred size of $\boldsymbol{w}$ directly measures how related the domains are and determines how much information transfer should occur between them. There can be various sources of inconsistencies between the pairwise comparisons. They can stem from limitations in the design of individual studies and from changes in the baseline efficacy of treatments over time, for example, due to increasing antibiotic resistance. This problem has recently been addressed in [64] where the basic parameters are assumed to vary over time according to a Gaussian process. Thus, information transfer between domains is corrected for the times at which the associated datasets were collected.

Often, the appropriate transfer network joining the domains is not known and needs to be inferred. One can take a brute-force approach to select the best transfer network under some quality measure by checking every possible graph. However, this approach quickly becomes intractable as the number of domains grows with millions of possible transfer networks on just eight domains. Zhou *et al.* [117] provide a greedy algorithm

which starts from the target learner and at each step includes a source learner yielding the highest conditional marginal likelihood for the target task. Specifically, suppose that we have learners $\mathcal{L}_1, \ldots, \mathcal{L}_K$ operating on datasets $\boldsymbol{D}^{(1)} = (\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)}), \ldots, \boldsymbol{D}^{(K)} = (\boldsymbol{y}^{(K)}, \boldsymbol{X}^{(K)})$, respectively, where $\mathcal{L}_1$ is the target learner. Let $G = (V, E)$ be the (connected) network transfer graph with $V = \{1, \ldots, K\}$. Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ be the parameters in $\mathcal{L}_1, \ldots, \mathcal{L}_K$, where for every $(i, j) \in E$ there exist subvectors $\boldsymbol{\theta}_{i,\mathcal{C}_i}$ and $\boldsymbol{\theta}_{j,\mathcal{C}_j}$ of, respectively, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ which are restricted to be equal (shared parameter approach). Then at each step, given the chosen set of learners $Q \subset V$, which is known as the *linkage set*, let $N_G(Q)$ be the set of neighbors of Q, consisting of all learners adjacent to at least one learner in Q. The new learner j^* to be added to Q is then selected via

$$j^* = \arg\max_{j \in N_G(Q)} p(\cup_{k \in Q} \boldsymbol{y}^{(k)} \mid \cup_{k \in Q} \boldsymbol{X}^{(k)}, \boldsymbol{D}^{(j)}). \tag{9}$$

The algorithm ends once adding a new learner no longer increases the conditional likelihood in (9). The complexity of this algorithm is $O(K^2)$ under the assumption that the conditional likelihood in (9) can be obtained in constant time. This can be further reduced to $O(K \log K)$ if the likelihood computation is parallelized between the learners in Q. Zhou *et al.* [117] provide theoretical guarantees for the recovery of the optimal transfer subnetwork of G.

Having explored a variety of Bayesian approaches to transferring information between domains in a flexible manner, we now discuss which among these are applicable to particular types of transfer learning problem depending on (i) feature spaces of source and target domains; (ii) the availability of labels and samples in both source and target datasets. We note that there is an immense Bayesian literature that provides relevant models for transfer learning that we did not mention above. Instead, we have chosen to highlight some approaches that we find particularly interesting and useful.

4. HOMOGENEOUS VS HETEROGENEOUS TRANSFER

Following criterion (i), transfer learning problems can be dichotomized on the basis of whether or not the observations in the source and target domains live in the same feature spaces. In the literature, the former case is typically referred to as *homogeneous transfer* [103], [53] and the latter as *heterogeneous transfer* [118], [26]. This classification often determines which of the approaches presented in Section 3 is appropriate or even feasible to use.

4.1 Homogeneous transfer

Homogeneous transfer occurs when the source and target data and labels have the same meaning in the different domains but may follow different distributions. For example, the same variables are collected for each of the study subjects, but subjects in different groups may have considerably different attributes; hence, the distribution of the variables being collected may vary across groups. In such cases, it typically makes sense to define the same form of likelihood for the data in each domain, though there may be systematic differences in the parameters. All of the methods described in Section 3 can be applied directly to problems with identical feature spaces across domains.

Examples of homogeneous transfer include clinical studies with patients divided according to their health status or subtype of disease. Researchers are often interested in improving the accuracy of inference and predictions for a particular group of patients by utilizing information collected from the other groups or from healthy individuals. This can be especially useful when dealing with rare diseases, where it is often difficult to collect measurements on a large sample of affected patients. In [8] and [43], the authors use RNA sequencing datasets for different types of lung, kidney, head, and neck cancer as source domains to improve the precision of subtype identification for particular types of lung cancer. Bayesian models that borrow strength between classes and types of cancer have been applied in other contexts, including survival analysis [74], [85], and protein network inference [94], [4].

4.2 Heterogeneous transfer

This case is more challenging as the source and target datasets do not consist of measurements on exactly the same variables for different study subjects. In order for transfer learning to apply, there has to be something in common across the domains. A typical setting is when the domains have overlapping but not completely identical sets of variables. For example, there may be a common focus across the domains in studying the impact of key predictors of interest on a response, measured under different covariates.

For regression or classification models, the coefficients for the key predictors are not directly comparable across models adjusting for different covariates. Hence, shared parameter models are not appropriate. However, it may make sense to assume that the domain-specific coefficients for the key predictors are drawn from a common random-effects distribution, thereby enabling borrowing of information. Shared latent space models are even more natural in this case. By jointly modeling the response, key predictors, and covariates as conditionally independent given latent factors, we induce a parsimonious *latent factor regression/classification* model [14], [36]. Multistudy variants of such models can seamlessly handle cases in which the covariates differ across domains. The multigroup variants of Bayesian mixture models discussed in Section 3 similarly apply for the joint distributions of key

predictors, covariates, and response(s) to induce flexible transfer learning.

Bayesian methods enjoy a key advantage in this context in their ability to rely on joint latent feature models to transfer information across domains with partially overlapping variables. The majority of non-Bayesian approaches such as deep transfer learning ([111], [93], [86], [96], [73], [69]) rely on both domains having observations in the same feature space for pre-training and fine-tuning the learners. We detail some examples of Bayesian transfer with overlapping variables in the following.

4.2.1 Multi-study latent factor regression Recall the multi-study latent factor model in equation (8). Previously, we considered the observed data on subject i in study (domain) k, $\boldsymbol{y}_{k,i}$, to be p dimensional, with p fixed between subjects and domains. To generalize this, we instead consider the p-dimensional data $\boldsymbol{y}_{k,i}$ to be the *complete data* for the subject i in the study k that could potentially have been measured. Then, we define $\boldsymbol{m}_{k,i} = (m_{k,i,j}, j = 1, \ldots, p)^T$ as the *missingness pattern* for the subject (k, i), with $m_{k,i,j} = 1$ if the jth variable is not observed for that subject and $m_{k,i,j} = 0$ otherwise. A variable is missing for a subject if the study they participated in does not collect that variable, or if the study planned to collect that variable, but it was not available.

Let $\boldsymbol{y}_{k,i}^{(obs)} = \{y_{k,i,j}, j : m_{k,i,j} = 0, j = 1, \ldots, p\}$ denote the $p_{k,i} = \sum_{j=1}^{p}(1 - m_{k,i,j})$ dimensional observed data vector for subject (k, i). The Gaussian multi-study latent factor model characterizes the complete data vector as $\boldsymbol{y}_{k,i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k)$. This in turn induces a $p_{k,i}$-dimensional multivariate Gaussian distribution for the observed data vector $\boldsymbol{y}_{k,i}^{(obs)}$ having covariance corresponding to the appropriate submatrix of $\boldsymbol{\Sigma}_k$. In fitting Bayesian multi-study factor models, it is not necessary to impute the missing data. Instead, one can simply take into account the differing observed data contributions for each subject in implementing a Gibbs sampler or alternative Markov chain Monte Carlo (MCMC) algorithm for posterior sampling.

This approach can be used for transfer learning about the covariance structure in multivariate data specific to the target domain. Alternatively, when the focus is on regression, one can concatenate outcomes, predictors of interest, and covariates together in the $\boldsymbol{y}_{k,i}$ data vector for subject (k, i). A Gaussian linear regression model for the outcome given the predictors of interest and covariates can then be obtained directly from the covariance $\boldsymbol{\Sigma}_k$ using standard multivariate Gaussian theory. This type of approach is straightforward to extend to mixed categorical and continuous data by following the popular approach of linking categorical variables to underlying Gaussian variables.

4.2.2 Nonlinear and nonparametric extensions A limitation of the above multi-study factor analysis model is the assumption of multivariate Gaussianity. It is therefore useful to consider extensions that incorporate shared and study-specific latent factors while relaxing these restrictive distribution assumptions, which also imply linear relationships among the variables.

In the single modality case, there is a rich literature on nonlinear factor models. For example, we could let

$$\boldsymbol{y}_i = f(\boldsymbol{\eta}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p), \tag{10}$$

where $\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q)$ are the vectors of latent factors and $f(\cdot)$ is an unknown and potentially non-linear function. Gaussian process latent variable models (GP-LVMs) place a GP prior on the function f mapping from the latent to ambient space [63], [35], [89], [107]. Alternatively, the popular class of variational autoencoders (VAEs) characterize f using deep neural networks and takes a variational approach to inference [77], [51], [15].

While these highly flexible nonlinear latent variable models have exhibited appealing practical performance as black-box models for generating new data that resemble the training data, they are prone to a number of vexing issues in reproducing statistical inferences. One major challenge is the curse of dimensionality resulting from the fact that the function f is an unknown mapping from q to p dimensional space; the space of such functions is immense, necessitating an enormous amount of training data for adequate performance. Furthermore, these models are not identifiable without substantial additional constraints. Another common problem is termed posterior collapse [24], [99], [100], in which there is a lack of learning about the latent variables based on the data. Although there have been some attempts at addressing these problems, there remains a lack of a practically useful methodology for performing reproducible dimensionality reduction.

The above challenges are exacerbated in considering extensions to the multi-study (transfer learning) case. Hence, we recommend starting with more parsimonious nonlinear latent factor models in future work developing such extensions. One promising point of departure is the recently proposed NIFTY framework of Xu *et al.* [107], which lets

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$
$$\eta_{ih} = g_h(u_{ik_h}), \quad h = 1, \ldots, q,$$

where $\boldsymbol{\Lambda}$ is a factor loading matrix, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, and $u_{ik} \overset{iid}{\sim} U(0, 1)$ for $k = 1, \ldots, K \leq q$. Each latent factor η_{ih} is a transformation of a latent u_{ik_h} through an unknown non-decreasing function g_h. The subscript k_h allows the same latent uniforms to be used for multiple factors, inducing dependence.

This model induces a flexible multivariate density for $\boldsymbol{y}_i$, while massively reducing the dimensionality relative to the model (10). In their paper, they provided theory on identifiability, leveraging on pre-training with state-of-the-art nonlinear dimensionality reduction algorithms. They also showed excellent performance for a wide variety of complex examples. They were even able to train a realistic generative model for bird songs based on few training examples; audio recordings of bird songs provide an example of massive dimensional data with low intrinsic dimension. NIFTY can exploit the complex low-dimensional structure in the data for highly efficient performance.

In conducting inference for latent variable models, the NIFTY authors noticed a common problem of *distributional shift*. In particular, many of the current models assume that the latent variables are i.i.d. $\mathcal{N}(0,1)$ or $U(0,1)$. Inferences on the parameters, such as the induced covariance in the Gaussian linear factor model case, critically depend on this assumption holding not just *a priori* but also *a posteriori*. Xu *et al.* [107] propose a general approach to solving the distributional shift of latent variables by forcing the posterior distribution of latent variables to be very close to i.i.d. $U(0,1)$.

4.2.3 Mixture models An alternative direction towards building more flexible models for transfer learning, including in the partially overlapping variables case, is to rely on mixture models, building on the developments in Section 3.3. Such models also have the advantage of clustering subjects within the different domains. In the partially overlapping variable transfer learning case, it is appealing to define a joint model, as motivated above. However, Chandra *et al.* [17] recently noted a pitfall of mixture models in high-dimensional cases in which the posterior tends to concentrate on trivial clusterings of the observations that place all subjects into one cluster or in singleton clusters.

As a solution in the single domain case, they proposed a latent mixture model formulation that lets

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \\ \boldsymbol{\eta}_i &\sim \sum_{h=1}^{H} \nu_h \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Delta}_h), \end{aligned} \tag{11}$$

so that a mixture of Gaussians model is used for the latent variables in a linear factor model. They prove that this model solves the above-mentioned pitfall in Bayesian clustering in high dimensions. The trick is to model the variation across clusters in a lower-dimensional latent space to address the curse of dimensionality.

With this single domain specification as the starting point, there are multiple promising directions forward in terms of extensions to the multiple domain transfer learning case. One possibility is to define a multi-study factor model as in Chandra *et al.* [18] but instead of assuming Gaussian shared and study-specific latent factors, use Gaussian mixture models to induce a flexible distribution on the latent factors while also producing separate clusters of subjects in each domain with respect to the shared and study-specific components. An alternative is to rely on the model in equation (11) but with domain-specific distributions for the latent factors defined as

$$f_k(\boldsymbol{\eta}_i) = \int \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\theta}_i) dP_k(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Delta}_i\}$ are the Gaussian parameters and P_k is a mixing distribution on these parameters that is specific to domain k.

In the special case in which $P_k = P = \sum_{h=1}^{H} \nu_h \delta_{\boldsymbol{\theta}_h}$ with $\delta_{\boldsymbol{\theta}}$ a degenerate distribution concentrated at $\boldsymbol{\theta}$ we obtain the original model in (11). However, by using the different priors $(P_1, \ldots, P_K) \sim \Pi$ considered in Section 3.3 we can allow differences across the domains while borrowing information; further borrowing is achieved through the implicit assumption of a latent space that is shared across domains - this is induced through the use of a common factor loading matrix $\boldsymbol{\Lambda}$.

4.2.4 Multiresolution transfer learning Closely related to the overlapping variables case is the setting in which data are collected for each domain on related random functions or stochastic processes. For example, let $f_k : \mathcal{T} \to \mathbb{R}$ denote a latent smooth continuously differentiable function for domain k, and suppose that we have

$$y_{k,i} = f_k(t_{k,i}) + \epsilon_{k,i}, \quad \epsilon_{k,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{12}$$

with $\boldsymbol{y}_k = \{y_{k,i}, i = 1, \ldots, n_k\}$ the observed data and $\boldsymbol{t}_k = \{t_{k,i}, i = 1, \ldots, n_k\}$ the observation locations for domain k. Often, observation locations do not line up across domains and certain domains may have lower resolution data than others, with the *resolution* referring to the density of the observation points $\boldsymbol{t}_k$ over the support $\mathcal{T}$.

Model (12) represents a type of *functional data analysis* (FDA). In many FDA settings, we observe noisy realizations of a subject-specific function at a finite set of points, but here we are considering the case in which we have one function for each domain and are particularly interested in inference on the function for the target domain. There are many applications in which this type of problem arises. We may have domain-specific regression functions f_k and want to borrow information across domains in a nonparametric regression context without assuming any common parameters. Alternatively, $\boldsymbol{y}_k$ may correspond to a domain k-specific time series and we want to borrow information across related time series to enhance prediction for a target series.

A natural Bayesian approach to inference under model (12) is to consider a functional data extension of the hierarchical and random effects modeling approaches highlighted in Section 3. For example, one could use a hierarchical GP that lets $f_k \sim \text{GP}(f_0, c)$ with f_0 in turn given a GP prior [5], [57]. For articles on using GPs in transfer learning settings closely related to that of (12), refer to [101], [112], [44]. These approaches can automatically accommodate the case in which the observations are denser in some domains than in others.

Wilson *et al.* [101] introduce GP regression networks (GPRNs) which use latent GPs to transfer information between different continuous-time processes. Specifically, given K time series $y_1, \ldots, y_K$, [101] let

$$y_k(t) = \sum_{q=1}^{Q} W_{k,q}(t)\,[f_q(t) + \eta_q] + \epsilon_k,$$

where $f_q \sim \mathcal{GP}(0, \boldsymbol{K}_q^f)$ are latent basis GPs evaluated at the observation times, $W_{k,q} \sim \mathcal{GP}(0, \boldsymbol{K}_w)$ are domain-specific weights of the latent GPs, while $\eta_q = \eta_q(t)$ and $\epsilon_k = \epsilon_k(t)$ are respectively $\mathcal{N}(0, \sigma_f^2)$ and $\mathcal{N}(0, \sigma_y^2)$ white noise processes. The matrix of weights $\boldsymbol{W}(t)$ determines the strength and structure of the information transfer between domains, analogously to the learning network introduced in [117]. However, unlike most of the work in Section 3.4, GPRN allows the strength of the transfer within the network of learners to vary over time.

Although [101] simplify inference by assuming identical measurement times across domains, [44] extend the approach to allow different measurement times and increase flexibility by using deep GPs. By using a shared latent space approach with GPs serving as the basis for the latent space, these models are able to achieve transfer both across resolutions and different classes of observations, corresponding to different air pollutants in the application presented in [44].

5. AVAILABILITY OF LABELED DATA

As we have seen, the Bayesian paradigm provides a fertile ground for developing a rich variety of techniques relevant to transfer learning in both supervised [1], [87], [22], [49], [53], [54], [55], [74] and unsupervised settings [27], [28], [18], [91], [90]. When the focus is on prediction, there are often challenges presented by the limited availability of labeled data. The term *semi-supervised* learning refers to the case in which labels are only available for a subset of the samples. The joint modeling approaches for heterogeneous transfer learning described in the previous section can trivially handle semi-supervised settings, as missing labels are just one type of missing data.

Particularly challenging are cases of *one-shot* and *few-shot* learning, which refers to having only a single or a few labeled samples in the target domain, respectively. For articles proposing Bayesian approaches to handle semi-supervised learning and these cases of a tiny number of labeled target data, refer to [110], [81], [84], [60], [61], [62]. In such cases, performance is critically dependent on borrowing of information from labeled data in the source domains. Common examples include classification based on images or audio [60], [61], [62]. For example, we may have many labeled examples of different individuals handwriting but only a tiny number for the individual of interest.

In the transfer learning literature, *inductive transfer learning* refers to the case where target domain labels are available, while *transductive transfer learning* has labels available only in the source domains [80]. Of course, if there are certain systematic differences between the source and target domains, accurate transductive transfer may be impossible [25]. However, there is a rich PAC-Bayesian literature on this topic ([37], [38], [39], [88]) which specifies conditions that allow for successful training of the target learner in the absence of target labels in classification settings. They provide theoretical upper bounds on the expected error on the target domains of a Gibbs classifier depending on various measures of divergence between the distributions of predictors and labels of both domains as well as the properties of the set of voter algorithms from which the Gibbs classifier is constructed.

6. SIMULATION ILLUSTRATION

In order to illustrate Bayesian transfer learning in practice, we run a simulation experiment focused on the problem of transfer learning targeting the covariance and/or precision matrix of a high-dimensional multivariate Gaussian distribution. In particular, there are data collected on the same set of variables for subjects in different groups, and we would like to allow the covariance/precision to vary across groups while borrowing information. This is a natural setting for both Bayesian multi-study factor analysis models and frequentist competitors focused on transfer learning in precision matrix estimation.

Specifically, we compare Bayesian Subspace Factor Analysis (SUFA) [18] presented in Section 3.3 with the frequentist Trans-CLIME method proposed by Li *et al.* [66]. Trans-CLIME is a transfer learning extension of constrained L_1 minimization for inverse matrix estimation (CLIME) [11]. We also include the frequentist estimator proposed by Guo *et al.* [42], which Li *et al.* [66] refer to as multitask graphical lasso (MT-Glasso). While [18] focused on inferring covariance, SUFA can be used just as easily to infer any functional of the covariance, such as precision.

We focus our comparisons on precision estimation, as this was the emphasis in [66] and [42]. Here, we generate synthetic p-dimensional data in the S (source) and T

(target) domains, with sample sizes n_S and n_T respectively. We consider four experiments where we alternatively vary the sample sizes and dimensionality under two different data generating models. For both data generating mechanisms, we first fix $n_S = 140$, $n_T = 70$ and examine performance as dimension varies $p \in \{40, 60, \ldots, 280\}$. Then we fix $p = 100$, and vary $n_S \in \{40, 60, \ldots, 280\}$, while keeping $n_T = n_S/2$ for each case. In order to obtain the average performance of each method we generate 100 replicated datasets with fixed true precision matrices for both domains in each setting across all the replicates. We then report the average Frobenius and L_1 norm of the error in the estimated target precision matrix.

In the first setting, the data in the source and target domains come from factor models which share some latent factors, but for which SUFA is misspecified. These results are displayed in Figure 3. Next, we consider the data generation mechanism used in the paper introducing Trans-CLIME [66], where source and target data are generated from Gaussian distributions with precision matrices $\mathbf{\Omega}_S, \mathbf{\Omega}_T$, respectively. Here $\mathbf{\Omega}_T$ is a sparse banded matrix with

$$(\mathbf{\Omega}_T)_{i,j} = 2 \cdot 0.6^{|i-j|} \cdot \mathbb{1}(|i-j| \leq 7)$$

and

$$\mathbf{\Omega}_S = (\boldsymbol{I}_p + \mathbf{\Delta})^{-1}\mathbf{\Omega}_T;$$

each entry of $\mathbf{\Delta}$ is set to zero with probability 0.9 and it is sampled from $U(-10/p, 10/p)$ with probability 0.1. Following [66], if the resulting $\mathbf{\Omega}_S$ matrix is not positive semidefinite, it is symmetrized and projected onto the positive semidefinite cone. We present the results in Figure 4.

Performance tends to improve with growing dimension for SUFA and in most cases for Trans-CLIME and MT-Glasso as well. This somewhat counterintuitive phenomenon, known as the blessing of dimensionality [65], is commonly encountered in the literature on covariance and precision estimation [76], [105], [13], [20]. SUFA clearly outperformed its frequentist competitors in all settings for data sampled from the factor model. There is significant instability in the solutions provided by Trans-CLIME and MT-Glasso. In particular, under the data generation mechanism from [66], SUFA remains competitive at least in certain settings and provides stable solutions across the board. It is important to note that SUFA was created for estimating covariance, rather than precision. In addition, the L_1 error norm could be a more favorable metric for Trans-CLIME and MT-Glasso, as they are built on minimization of L_1.

Hence, we find that this particular Bayesian approach to transfer learning based on a shared latent space model can outperform its frequentist counterparts even on tasks it was not built for (estimating the precision instead of the covariance), and unlike its counterparts, it does not become highly unstable on data generating mechanisms it was not explicitly designed for. Indeed, as we have partially illustrated, since we have a posterior for the covariance that has support on the space of positive semidefinite covariances, we can give Bayes estimates and posterior credible intervals providing uncertainty quantification for any functional of the covariance of interest. Hence, from a single Bayesian analysis, we can obtain multiple results of interest that are all internally coherent.

7. DISCUSSION

Transfer learning is a timely problem given the abundance of data sets from related domains. In many applications, there are simply not enough data from the domain of interest to support reliable inference and accurate predictions as we seek to fit increasingly complex models. Hence, it becomes critical to cleverly borrow information from available "source" datasets.

Choosing the appropriate strength and structure of information transfer between domains remains one of the key challenges. The Bayesian paradigm offers a wide variety of approaches to transfer learning, including shared parameters, hierarchical and random-effects models, shared latent space, and network transfer methods. There is a rich literature developing and applying these approaches in transfer learning settings, although most often "transfer learning" is not mentioned in the associated papers.

This article has focused on providing a flavor for some of the interesting directions that are possible in terms of Bayesian transfer learning, but has not attempted a comprehensive overview of the massive relevant literature. Most of the transfer learning literature has focused on the simplest homogeneous transfer case, where data from different domains consist of the same variables measured in different subjects. Bayesian ideas applied to transfer learning can be particularly useful in the more challenging settings presented in Section 4, where these existing methods largely do not apply.

We have purposely focused much of our attention on shared latent space-type models for Bayesian transfer learning, ranging from multi-study factor analysis to multi-group Bayesian nonparametric models. We focused on these areas because the associated models are not as well known to the broad community but are very practically useful, including in challenging high-dimensional and complex structured data cases. In addition, there have been interesting recent developments that we have highlighted, while sketching out some promising directions for ongoing research. This includes extending Bayesian continuous latent factor modeling approaches to transfer learning settings. Our view is that more careful statistical models will tend to dominate over-parametrized black boxes, such as VAEs, in many settings.

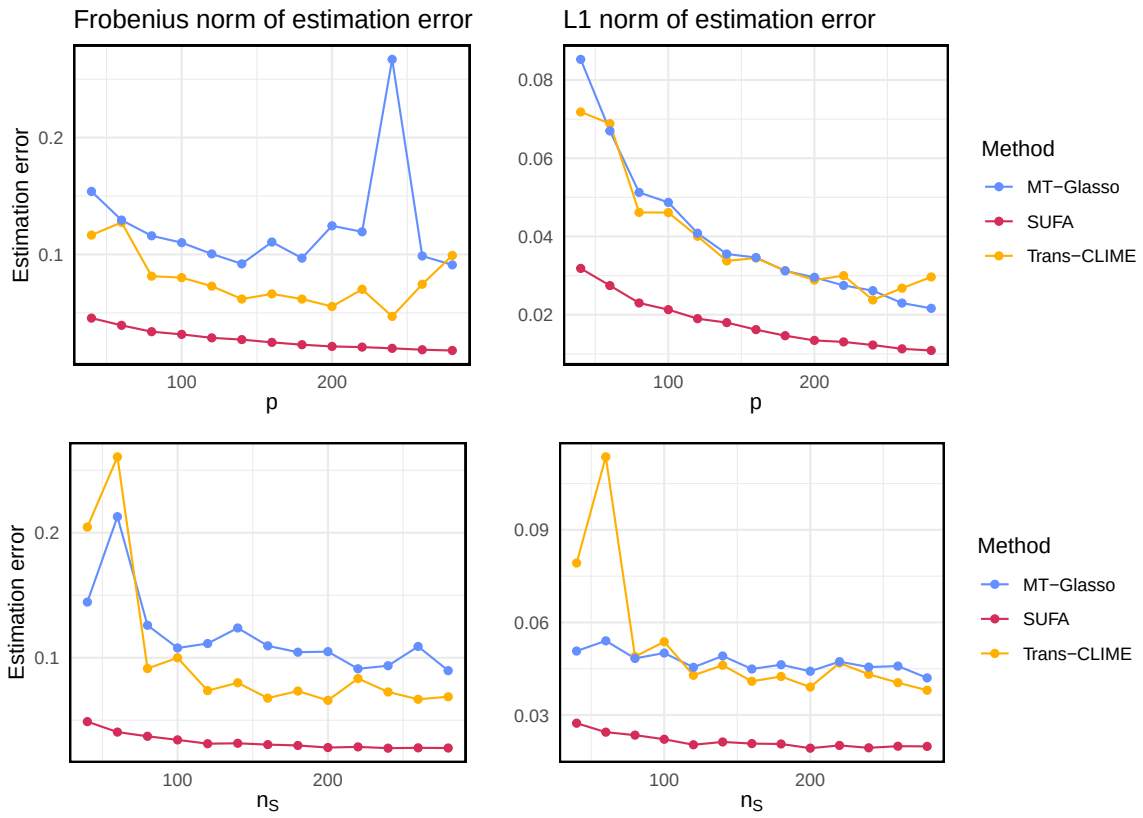

FIG 3. *Frobenius and L_1 norm errors of target precision matrix estimates for SUFA, Trans-CLIME and MT-Glasso over varying dimension p, and sample sizes of source and target domains. Samples for both domains come from factor models which share some of the latent factors.*

An additional interesting area for future research is Bayesian transfer learning involving deep neural networks. While in recent years there have been papers taking some early steps in this area [1], [19], [87], there is plenty of potential for further impactful developments in this field, especially given the importance of transfer learning to deep neural networks training due to their data-hungry nature.

8. ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 856506), United States National Institutes of Health (R01ES035625), Office of Naval Research (N00014-21-1-2510), National Institute of Environmental Health Sciences (R01ES027498), and National Science Foundation (DMS-2230074). The contributions of Suder were supported in part by the Duke Summer Research Fellowship.

REFERENCES

[1] ABBA, M. A., WILLIAMS, J. P. and REICH, B. J. (2023). A penalized complexity prior for deep Bayesian transfer learning with application to materials informatics. *The Annals of Applied Statistics* **17** 3241–3256.

[2] AVRAHAMI, O., LISCHINSKI, D. and FRIED, O. (2022). GAN cocktail: mixing GANs without dataset access. In *European Conference on Computer Vision* 205–221.

[3] BAGLAMA, J. and REICHEL, L. (2005). Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing* **27** 19–42.

[4] BALADANDAYUTHAPANI, V., TALLURI, R., JI, Y., COOMBES, K. R., LU, Y., HENNESSY, B. T., DAVIES, M. A. and MALLICK, B. K. (2014). Bayesian sparse graphical models for classification with application to protein expression data. *The Annals of Applied Statistics* **8** 1443.

[5] BEHSETA, S., KASS, R. E. and WALLSTROM, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika* **92** 419–434.

[6] BEKKER, P. A. and TEN BERGE, J. M. (1997). Generic global indentification in factor analysis. *Linear Algebra and its Applications* **264** 255–263.

[7] BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.

[8] BOLUKI, S., QIAN, X. and DOUGHERTY, E. R. (2021). Optimal Bayesian supervised domain adaptation for RNA sequencing data. *Bioinformatics* **37** 3212–3219.

[9] BOSCHINI, M., BONICELLI, L., PORRELLO, A., BELLITTO, G., PENNISI, M., PALAZZO, S., SPAMPINATO, C. and CALDERARA, S. (2022). Transfer without forgetting. In *European Conference on Computer Vision* 692–709.

[10] BU, F., KAGAAYI, J., GRABOWSKI, M. K., RATMANN, O. and XU, J. (2024). Inferring HIV transmission patterns from

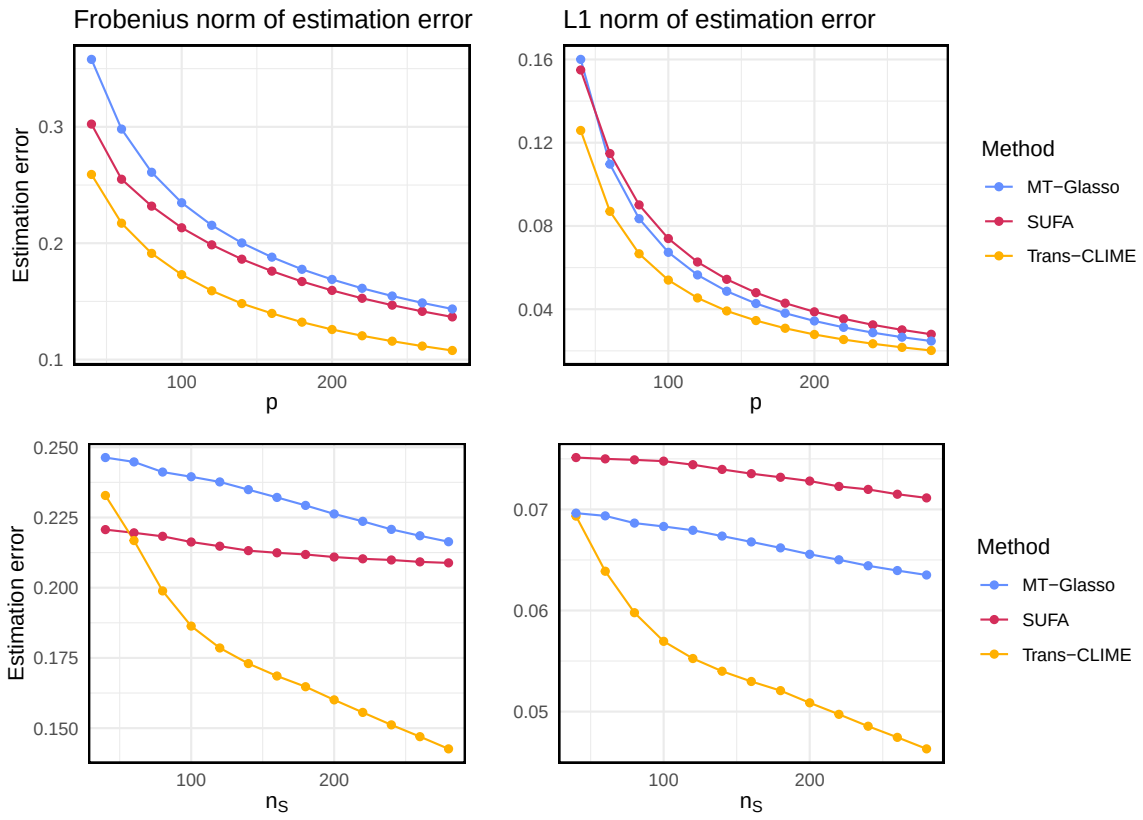

FIG 4. *Frobenius and L_1 norm errors of target precision matrix estimation for SUFA, Trans-CLIME and MT-Glasso over varying dimension p and sample sizes of source and target domains. Samples for both domains come from the data generating mechanism described in the paper introducing Trans-CLIME [66].*

viral deep-sequence data via latent typed point processes. *Biometrics* **80**.

[11] CAI, T., LIU, W. and LUO, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

[12] CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics* **47** 67 – 92.

[13] CAO, Y., LIN, W. and LI, H. (2019). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association* **114** 759–772.

[14] CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103** 1438–1456.

[15] CEMGIL, T., GHAISAS, S., DVIJOTHAM, K., GOWAL, S. and KOHLI, P. (2020). The autoencoding variational autoencoder. In *Advances in Neural Information Processing Systems* **33**.

[16] CHAKRABARTI, A., NI, Y., MORRIS, E. R. A., SALINAS, M. L., CHAPKIN, R. S. and MALLICK, B. K. (2024). Graphical dirichlet process for clustering non-exchangeable grouped data. *Journal of Machine Learning Research* **25** 1–56.

[17] CHANDRA, N. K., CANALE, A. and DUNSON, D. B. (2023). Escaping the curse of dimensionality in Bayesian model-based clustering. *Journal of Machine Learning Research* **24** 1–42.

[18] CHANDRA, N. K., DUNSON, D. B. and XU, J. (2024). Inferring covariance structure from multiple data sources via subspace factor analysis. *Journal of the American Statistical Association* 1–15.

[19] CHANDRA, R. and KAPOOR, A. (2020). Bayesian neural multi-source transfer learning. *Neurocomputing* **378** 54–64.

[20] CHATTOPADHYAY, S., ZHANG, A. R. and DUNSON, D. B. (2024). Blessing of dimension in Bayesian inference on covariance matrices. *arXiv preprint arXiv 2404.03805*.

[21] CHEN, M.-H. and IBRAHIM, J. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis* **1** 551-574.

[22] CHEN, M.-H. and IBRAHIM, J. G. (2000). Power prior distributions for regression models. *Statistical Science* **15** 46 – 60.

[23] CHEN, M.-H., IBRAHIM, J. G., LAM, P., YU, A. and ZHANG, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics* **67** 1163–1170.

[24] DAI, B., WANG, Z. and WIPF, D. (2020). The usual suspects? Reassessing blame for VAE posterior collapse. In *Proceedings of the 37th International Conference on Machine Learning* **119** 2313–2322.

[25] DAVID, S. B., LU, T., LUU, T. and PÁL, D. (2010). Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* **9** 129–136.

[26] DAY, O. and KHOSHGOFTAAR, T. M. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data* **4** 1–42.

[27] DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337-346.

[28] DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *The Annals of Applied Statistics* **15** 1723–1741.

[29] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[30] DIAS, S., WELTON, N. J., SUTTON, A. J. and ADES, A. E. (2014). *NICE DSU technical support document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials*. National Institute for Health and Care Excellence (NICE).

[31] DING, D. Y., LI, S., NARASIMHAN, B. and TIBSHIRANI, R. (2022). Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences* **119**.

[32] DUAN, Y., YE, K. and SMITH, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics: The Official Journal of the International Environmetrics Society* **17** 95–106.

[33] DURANTE, D. and DUNSON, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis* **13** 29-58.

[34] EBRAHIMI, S., ELHOSEINY, M., DARRELL, T. and ROHRBACH, M. (2020). Uncertainty-guided continual learning with Bayesian neural networks. *International Conference on Learning Representations*.

[35] ELEFTHERIADIS, S., RUDOVIC, O. and PANTIC, M. (2014). Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing* **24** 189–204.

[36] FERRARI, F. and DUNSON, D. B. (2021). Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association* **116** 1521–1532.

[37] GERMAIN, P., HABRARD, A., LAVIOLETTE, F. and MORVANT, E. (2013). A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning* **28** 738–746.

[38] GERMAIN, P., HABRARD, A., LAVIOLETTE, F. and MORVANT, E. (2016). A new PAC-Bayesian perspective on domain adaptation. In *Proceedings of The 33rd International Conference on Machine Learning* **48** 859–868.

[39] GERMAIN, P., HABRARD, A., LAVIOLETTE, F. and MORVANT, E. (2020). PAC-Bayes and domain adaptation. *Neurocomputing* **379** 379–397.

[40] GÖNEN, M. and MARGOLIN, A. Kernelized Bayesian transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* **28** 1831-1839.

[41] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

[42] GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15.

[43] HAJIRAMEZANALI, E., ZAMANI DADANEH, S., KARBALAYGHAREH, A., ZHOU, M. and QIAN, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. In *Advances in Neural Information Processing Systems* **31**.

[44] HAMELIJNCK, O., DAMOULAS, T., WANG, K. and GIROLAMI, M. (2019). Multi-resolution multi-task Gaussian processes. In *Advances in Neural Information Processing Systems* **32**.

[45] HICKEY, J., WILLIAMS, J. P. and HECTOR, E. C. (2024). Transfer learning with uncertainty quantification: Random effect calibration of source to target (RECaST). *Journal of Machine Learning Research* **25** 1–40.

[46] HOFF, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* **20**.

[47] HOSPEDALES, T., ANTONIOU, A., MICAELLI, P. and STORKEY, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44** 5149–5169.

[48] IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* **57** 383–388.

[49] IBRAHIM, J. G., CHEN, M.-H., GWON, Y. and CHEN, F. (2015). The power prior: theory and applications. *Statistics in Medicine* **34** 3724–3749.

[50] IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* **98** 204–213.

[51] ILSE, M., TOMCZAK, J. M., LOUIZOS, C. and WELLING, M. (2020). DIVA: domain invariant variational autoencoders. In *Proceedings of the 3rd Conference on Medical Imaging with Deep Learning* **121** 322–348.

[52] KAPOOR, S., KARALETSOS, T. and BUI, T. D. (2021). Variational auto-regressive Gaussian processes for continual learning. In *Proceedings of the 38th International Conference on Machine Learning* **139** 5290–5300.

[53] KARBALAYGHAREH, A., QIAN, X. and DOUGHERTY, E. R. (2018). Optimal Bayesian transfer learning. *IEEE Transactions on Signal Processing* **66** 3724–3739.

[54] KARBALAYGHAREH, A., QIAN, X. and DOUGHERTY, E. R. (2018). Optimal Bayesian transfer regression. *IEEE Signal Processing Letters* **25** 1655–1659.

[55] KARBALAYGHAREH, A., QIAN, X. and DOUGHERTY, E. R. (2021). Optimal Bayesian transfer learning for count data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **18** 644–655.

[56] KOUW, W. M. and LOOG, M. (2019). An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv 1812.11806*.

[57] KOWAL, D. R., MATTESON, D. S. and RUPPERT, D. (2019). Functional autoregression for sparsely sampled data. *Journal of Business & Economic Statistics* **37** 97–109.

[58] KUMAR, A., CHATTERJEE, S. and RAI, P. (2021). Bayesian structural adaptation for continual learning. In *Proceedings of the 38th International Conference on Machine Learning* **139** 5850–5860.

[59] KUZNETSOVA, A., ROM, H., ALLDRIN, N., UIJLINGS, J., KRASIN, I., PONT-TUSET, J., KAMALI, S., POPOV, S., MALLOCI, M., KOLESNIKOV, A., DUERIG, T. and FERRARI, V. (2020). The open images dataset V4. *International Journal of Computer Vision* **128** 1956–1981.

[60] LAKE, B. M., SALAKHUTDINOV, R., GROSS, J. and TENENBAUM, J. B. (2011). One shot learning of simple visual concepts. *Cognitive Science* **33**.

[61] LAKE, B. M., SALAKHUTDINOV, R. and TENENBAUM, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science* **350** 1332–1338.

[62] LAKE, B. M., SALAKHUTDINOV, R. R. and TENENBAUM, J. (2013). One-shot learning by inverting a compositional causal process. In *Advances in Neural Information Processing Systems* **26**.

[63] LAWRENCE, N. D. and MOORE, A. J. (2007). Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th International Conference on Machine Learning* 481–488.

[64] LEBLANC, P. M. and BANKS, D. (2023). Time-varying Bayesian network meta-analysis. *arXiv preprint arXiv 2211.08312*.

[65] LI, Q., CHENG, G., FAN, J. and WANG, Y. (2018). Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association* **113** 380–389.

[66] LI, S., CAI, T. T. and LI, H. (2023). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association* **118** 2171–2183.

[67] LI, Z. R., WU, Z., CHEN, I. and CLARK, S. J. (2024). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *The Annals of Applied Statistics* **18** 1137 – 1159.

[68] LOCK, E. F. and DUNSON, D. B. (2015). Shared kernel Bayesian screening. *Biometrika* **102** 829–842.

[69] LONG, M., CAO, Y., WANG, J. and JORDAN, M. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning* **37** 97–105.

[70] LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14** 41–67.

[71] LU, G. and ADES, A. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23** 3105–3124.

[72] LU, G. and ADES, A. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101** 447–459.

[73] LUO, Z., ZOU, Y., HOFFMAN, J. and FEI-FEI, L. F. (2017). Label efficient learning of transferable representations across domains and tasks. In *Advances in Neural Information Processing Systems* **31**.

[74] MAITY, A. K., BHATTACHARYA, A., MALLICK, B. K. and BALADANDAYUTHAPANI, V. (2020). Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data. *Biometrics* **76** 316–325.

[75] MCCLOSKEY, M. and COHEN, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation* **24** 109–165.

[76] MOLSTAD, A. J., EKVALL, K. O. and SUDER, P. M. (2024). Direct covariance matrix estimation with compositional data. *Electronic Journal of Statistics* **18** 1702 – 1748.

[77] MORAN, G. E., SRIDHAR, D., WANG, Y. and BLEI, D. (2022). Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*.

[78] MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66** 735–749.

[79] NIU, S., LIU, Y., WANG, J. and SONG, H. (2020). A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence* **1** 151–166.

[80] PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22** 1345–1359.

[81] PATACCHIOLA, M., TURNER, J., CROWLEY, E. J., O' BOYLE, M. and STORKEY, A. J. (2020). Bayesian meta-learning for the few-shot setting via deep kernels. In *Advances in Neural Information Processing Systems* **34**.

[82] RAVI, S. and BEATSON, A. (2019). Amortized Bayesian meta-learning. *International Conference on Learning Representations*.

[83] RODRIGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103** 1131–1154.

[84] SALAKHUTDINOV, R., TENENBAUM, J. and TORRALBA, A. (2012). One-shot learning with a hierarchical nonparametric Bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* **27** 195–206.

[85] SAMORODNITSKY, S., HOADLEY, K. A. and LOCK, E. F. (2020). A pan-cancer and polygenic Bayesian hierarchical model for the effect of somatic mutations on survival. *Cancer Informatics* **19**.

[86] SHIN, H.-C., ROTH, H. R., GAO, M., LU, L., XU, Z., NOGUES, I., YAO, J., MOLLURA, D. and SUMMERS, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging* **35** 1285–1298.

[87] SHWARTZ-ZIV, R., GOLDBLUM, M., SOURI, H., KAPOOR, S., ZHU, C., LECUN, Y. and WILSON, A. G. (2022). Pre-train your loss: Easy Bayesian transfer learning with informative priors. In *Advances in Neural Information Processing Systems* **36**.

[88] SICILIA, A., ATWELL, K., ALIKHANI, M. and HWANG, S. J. (2022). PAC-Bayesian domain adaptation bounds for multiclass learners. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence* **180** 1824–1834.

[89] SONG, G., WANG, S., HUANG, Q. and TIAN, Q. (2019). Harmonized multimodal learning with Gaussian process latent variable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** 858–872.

[90] SOTIROPOULOS, S. N., HERNÁNDEZ-FERNÁNDEZ, M., VU, A. T., ANDERSSON, J. L., MOELLER, S., YACOUB, E., LENGLET, C., UGURBIL, K., BEHRENS, T. E. and JBABDI, S. (2016). Fusion in diffusion MRI for improved fibre orientation estimation: An application to the 3T and 7T data of the Human Connectome Project. *Neuroimage* **134** 396–409.

[91] SOTIROPOULOS, S. N., JBABDI, S., ANDERSSON, J. L., WOOLRICH, M. W., UGURBIL, K. and BEHRENS, T. E. (2013). RubiX: Combining spatial resolutions for Bayesian inference of crossing fibers in diffusion MRI. *IEEE Transactions on Medical Imaging* **32** 969–982.

[92] SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64** 583–639.

[93] TAN, C., SUN, F., KONG, T., ZHANG, W., YANG, C. and LIU, C. (2018). A survey on deep transfer learning. In *International Conference on Artificial Neural Networks* 270–279.

[94] TAN, L. S. L., JASRA, A., DE IORIO, M. and EBBELS, T. M. D. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics* **11** 2222–2251.

[95] TEH, Y.-W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581.

[96] TZENG, E., HOFFMAN, J., DARRELL, T. and SAENKO, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision* 4068–4076.

[97] VANSCHOREN, J. (2019). Meta-learning. *Automated Machine Learning: Methods, Systems, Challenges* 35–61.

[98] WANG, B. and PINEAU, J. (2015). Online boosting algorithms for anytime transfer and multitask learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* **29**.

[99] WANG, Y., BLEI, D. and CUNNINGHAM, J. P. (2021). Posterior collapse and latent variable non-identifiability. In *Advances in Neural Information Processing Systems* **35**.

[100] WANG, Z. and ZIYIN, L. (2022). Posterior collapse of a linear latent variable model. In *Advances in Neural Information Processing Systems* **36**.

[101] WILSON, A. G., KNOWLES, D. A. and GHAHRAMANI, Z. (2012). Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning* 1139–1146.

[102] WOOD, F. and TEH, Y. W. (2009). A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* **5** 607–614.

[103] WU, Q., WU, H., ZHOU, X., TAN, M., XU, Y., YAN, Y. and HAO, T. (2017). Online transfer learning with multiple homogeneous or heterogeneous sources. *IEEE Transactions on Knowledge and Data Engineering* **29** 1494–1507.

[104] WU, Z., LI, Z. R., CHEN, I. and LI, M. (2024). Tree-informed Bayesian multi-source domain adaptation: Cross-population probabilistic cause-of-death assignment using verbal autopsy. *Biostatistics* **25** 1233–1253.

[105] XU, J. and LANGE, K. (2022). A proximal distance algorithm for likelihood-based sparse covariance estimation. *Biometrika* **109** 1047–1066.

[106] XU, J. and ZHU, Z. (2018). Reinforced continual learning. In *Advances in Neural Information Processing Systems* **32**.

[107] XU, M., HERRING, A. H. and DUNSON, D. B. (2023). Identifiable and interpretable nonparametric factor analysis. *arXiv preprint arXiv 2311.08254*.

[108] XUAN, J., LU, J. and ZHANG, G. (2021). Bayesian transfer learning: An overview of probabilistic graphical models for transfer learning. *arXiv preprint arXiv 2109.13233*.

[109] YANG, Q., ZHANG, Y., DAI, W. and PAN, S. J. (2020). *Transfer learning*. Cambridge University Press.

[110] YOON, J., KIM, T., DIA, O., KIM, S., BENGIO, Y. and AHN, S. (2018). Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems* **32**.

[111] YOSINSKI, J., CLUNE, J., BENGIO, Y. and LIPSON, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* **28**.

[112] YOUSEFI, F., SMITH, M. T. and ÁLVAREZ, M. (2019). Multitask learning for aggregated data using Gaussian processes. In *Advances in Neural Information Processing Systems* **33**.

[113] ZHANG, Q., FANG, J., MENG, Z., LIANG, S. and YILMAZ, E. (2021). Variational continual Bayesian meta-learning. In *Advances in Neural Information Processing Systems* **35**.

[114] ZHANG, W., DENG, L., ZHANG, L. and WU, D. (2022). A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica* **10** 305–329.

[115] ZHAO, T., WANG, Z., MASOOMI, A. and DY, J. (2022). Deep Bayesian unsupervised lifelong learning. *Neural Networks* **149** 95–106.

[116] ZHOU, A. and LEVINE, S. (2021). Bayesian adaptation for covariate shift. In *Advances in Neural Information Processing Systems* **35**.

[117] ZHOU, J., DING, J., TAN, K. M. and TAROKH, V. (2021). Model linkage selection for cooperative learning. *Journal of Machine Learning Research* **22** 1–44.

[118] ZHOU, J., PAN, S., TSANG, I. and YAN, Y. (2014). Hybrid heterogeneous transfer learning through deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* **28**.

[119] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE* **109** 43–76.