Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: http://www.tandfonline.com/loi/lqen20

Statistical transfer learning: A review and some extensions to statistical process control

Fugee Tsung, Ke Zhang, Longwei Cheng & Zhenli Song

To cite this article: Fugee Tsung, Ke Zhang, Longwei Cheng & Zhenli Song (2018) Statistical transfer learning: A review and some extensions to statistical process control, Quality Engineering, 30:1, 115-128, DOI: 10.1080/08982112.2017.1373810

To link to this article: https://doi.org/10.1080/08982112.2017.1373810

Accepted author version posted online: 12 Sep 2017.
Published online: 12 Sep 2017.

QUALITY ENGINEERING
2018, VOL. 30, NO. 1, 115–128
https://doi.org/10.1080/08982112.2017.1373810

Taylor & Francis
Taylor & Francis Group

Statistical transfer learning: A review and some extensions to statistical process control

Fugee Tsung, Ke Zhang, Longwei Cheng, and Zhenli Song

Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ABSTRACT
The rapid development of information technology, together with advances in sensory and data acquisition techniques, has led to the increasing necessity of handling datasets from multiple domains. In recent years, transfer learning has emerged as an effective framework for tackling related tasks in target domains by transferring previously-acquired knowledge from source domains. Statistical models and methodologies are widely involved in transfer learning and play a critical role, which, however, has not been emphasized in most surveys of transfer learning. In this article, we conduct a comprehensive literature review on statistical transfer learning, i.e., transfer learning techniques with a focus on statistical models and statistical methodologies, demonstrating how statistics can be used in transfer learning. In addition, we highlight opportunities for the use of statistical transfer learning to improve statistical process control and quality control. Several potential future issues in statistical transfer learning are discussed.

KEYWORDS
3D printing; bayesian modeling; landslides; quality control; regularization; statistical process control; statistical transfer learning; transfer learning; urban rail transit

Introduction

With the remarkable development of information technology in recent years, data mining and machine learning techniques have been widely and successfully applied in various domains and data sources. However, traditional machine learning approaches usually perform well only for single tasks and within the same data distribution (Pan and Yang 2010, Weiss, Khoshgoftaar and Wang 2016). To this end, transfer learning provides an efficient framework for combining multiple sources and allows the transfer of previously acquired knowledge to tackle related tasks in new domains. With the assistance of transfer learning, information transferred from source domains could improve a learner's performance in the target domain (Weiss, Khoshgoftaar and Wang 2016). For instance, learning to play the electronic organ may help facilitate learning the piano (Pan and Yang 2010). Similarly, babies first learn to recognize human faces and then build on this knowledge to recognize other objects (Zhang and Yeung 2014). Transfer learning techniques have been demonstrated to be truly beneficial in many real-world applications such as warranty prediction (Tseng, Hsu, and Lin 2016), surface shape prediction (Shao et al. 2017), WiFi localization (Pan et al. 2008), sentiment classification (Blitzer et al. 2007), and collaborative filter (Pan et al. 2010).

Recent decades have witnessed rapid development in statistical models and methodologies with applications in a variety of fields. Many of these applications increasingly require describing data in different structures via statistical models and methodologies. Many statistical models have been actively studied in a transfer learning framework to integrate multiple data sources and transfer knowledge in specific data types. For example, Jin et al. (2011) investigate a hierarchical Bayesian model to cluster short text messages via transfer learning from auxiliary long text data. Shao et al. (2017) propose a multi-task learning approach for Gaussian processes to predict surface shapes by integrating similar manufacturing processes. In addition to statistical models, statistical methodologies are widely involved and play a critical role in connecting each individual statistical model in transfer learning

CONTACT Fugee Tsung ✉ season@ust.hk Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.

studies. Although transfer learning techniques have been extensively summarized in the field of data mining and machine learning (see Pan and Yang, 2010; Lu et al., 2015; Weiss, Khoshgoftaar and Wang 2016), there currently appears to be a lack of review articles on transfer learning from a statistical perspective.

This study provides a comprehensive review of statistical transfer learning, which refers to the transfer learning literature with a focus on statistical models and methodologies adopted. Apart from reviewing literature, we investigate how statistical transfer learning can be utilized in the field of statistical process control (SPC) and quality control via several real-world applications: landslide monitoring using slope sensor systems, passenger inflow forecasting and monitoring in urban rail transit systems, and shape deformation modeling for 3D printed products.

The rest of the article is organized as follows. In the next section, we provide a brief overview and categorization of transfer learning methods. Then statistical models and methodologies contained in transfer learning papers are discussed and summarized. After that we introduce three applications in SPC and quality control based on statistical transfer learning. The conclusion is presented at the last section.

A brief overview of transfer learning

There has been a great deal of research in recent years on transfer learning techniques and applications and several survey papers of transfer learning have been published on data mining and machine learning. For example, Pan and Yang (2010) introduce a brief history and the categorization of transfer learning techniques and present a comprehensive overview of transfer learning for classification, regression and clustering problems; Taylor and Stone (2009) survey transfer learning for reinforcement learning; Lu et al. (2015) examine transfer learning approaches in the computational intelligence field and cluster them into several categories.

Before we proceed with the detailed review and categorization of transfer learning, it is necessary to clarify the relationship between the closely related concepts of transfer learning, multi-task learning, and self-taught learning. Transfer learning is adopted to extract knowledge from source domains to improve performance in target domains, while in multi-task learning, the roles of the source and target tasks are symmetric: learning the task in each domain can be improved by using shared knowledge gained through other tasks. Multi-task learning approaches are covered in this paper as they are sometimes viewed as a subarea of transfer learning (Xu and Yang 2011) and widely termed as transfer learning techniques (Huang et al. 2012, Zou et al. 2015). Lastly, self-taught learning is transfer learning with an emphasis on utilizing unlabeled data in source domains for predictions in target domains.

Generally speaking, approaches to transfer learning can be divided into three categories based on the form of transferring information from source to target: instance-based, feature-based, and parameter-based transfer learning. A brief review of each transfer learning approach will be given in the following.

Instance-based transfer learning is used when the source and the target instances are generated from two different but closely related distributions so that parts of the source data can be reused in the target task. Dai et al. (2007) present a boosting algorithm, TrAdaBoost, to select the most useful source instances as additional training data for the target task by iteratively reweighting. Their procedure enables the construction of a high-quality model for the target task by integrating only a tiny amount of new data and a large amount of old data. Jiang and Zhai (2007) propose a general instance weighting framework to remove misleading training instances from source data and assign additional weight to instances in target data than those in source data. Liao et al. (2005) adopt an active learning strategy to improve task performance by introducing auxiliary variables for each instance in source data. Wu and Dietterich (2004) implement knowledge transfer by minimizing a weighted sum of two separate loss functions corresponding to source and task data, respectively.

The feature-based transfer approach attempts to learn a common feature structure between source and target data, which can be treated as a bridge for knowledge transfer. Argyriou et al. (2007) propose a regularization-based method to learn a low-dimensional function representation shared by source and target tasks. Lee et al. (2007) adopt a probabilistic approach to learn an informed meta-prior over feature relevance. Their model transfers meta-priors between source and target tasks and can thus deal with cases where tasks have non-overlapping features or the relevance of the features varies between tasks. Blitzer et al. (2006) suggest structural correspondence learning (SCL) to identify correspondences between features from source and target domains by modeling their correlations with pivot features. Although it is shown experimentally that SCL can reduce the difference between domains, selecting the pivot features remains challenging. Pan et al. (2008) utilize maximum mean discrepancy embedding (MMDE) to find a low-dimensional latent feature space in which the distributions of data in different domains are close to each other. Dai et al. (2008) exploit large amounts of auxiliary data to uncover an improved feature representation to enhance the clustering performance of a small amount of target data.

Parameter-based transfer learning assumes that source and target tasks should share parameters or hyper-parameters of prior distributions. Most research has been focused on two aspects: hierarchical Bayesian (HB) frameworks and regularization-based approaches. For the former, the parameters of models for individual tasks are often assumed to be generated from a common prior distribution. Thus, knowledge can be transferred across domains by learning the common information through the abundant auxiliary data from source domains. Gaussian processes are widely used and appropriate for this situation (Lawrence and Platt 2004, Schwaighofer et al. 2005, Bonilla et al. 2007). Researchers have also implemented parameter transfers with regularization-based approaches. Evgeniou and Pontil (2004) separate the parameter in support vector machines (SVMs) for each task into a task-common term and a task-specific term. They present an approach for knowledge transfer based on the minimization of regularization functionals.

Review of statistical transfer learning

As mentioned in the last section, existing surveys of transfer learning have been mainly conducted in the data mining and machine learning fields and focused mostly on methods of transferring information. Unlike existing surveys, transfer learning literature is reviewed from a statistical perspective in this article. Particularly, transfer learning papers in the fields of statistics and industrial engineering will gain additional attention. Recent progress is reviewed and organized from two perspectives: statistical models and statistical methodologies. First, transfer learning approaches for many real-world applications are reviewed based on their underlying statistical models for each single task/domain, including linear models, Gaussian process, network models, and statistical language models. Various statistical techniques are then exploited to transfer information across multiple statistical models and data sources such as boosting-based methods, bagging, Bayesian modeling, and regularization-based approaches are summarized.

Statistical models

A variety of statistical models have been developed in past decades to handle different data types in a diverse range of real-world applications. In an application aimed at transferring knowledge across multiple tasks, the first issue is to select an appropriate statistical model for each single task. As fundamental elements for statistical transfer learning approaches, the underlying statistical models need to be summarized with corresponding applications.

Many transfer learning studies have been conducted based on linear and generalized linear models. For example, to combine Earth System Model outputs for land surface temperature prediction in both South and North America, Gonçalves et al. (2016) adopt a simple linear model for each geographic location and suggested a multi-task learning approach to allow them to share dependencies. The transfer learning approach is more effective than conducting an ordinary least squares regression for each linear model since it can capture dependences across locations. For generalized linear models, Zou et al. (2015) propose a transfer learning method for a logistic regression with an application in degenerate biological systems. Similarly, Zhang et al. (2014) investigate a regularization-based transfer learning approach to capture task relationships for generalized linear models with the incorporation of new tasks. Moreover, Samarov et al. (2015, 2016) consider a linear mixture model to combine multiple outputs, which are applied to handle hyperspectral biomedical images.

Transfer learning has also been employed for Gaussian processes. The key idea of such approaches is to impose a shared prior to connect similar but not identical Gaussian processes. The prediction for each individual Gaussian process benefits from utilizing observations from different but related processes. Many Gaussian process methods based on transfer learning are investigated under various assumptions ranging from block to non-block design (Schwaighofer et al. 2004, Bonilla et al. 2007, Yu et al. 2005) with practical applications in compiler performance predictions and exam score predictions. Furthermore, Shao et al. (2017) integrate cutting force variation modeling with a multi-task learning approach to improve surface prediction accuracy by incorporating engineering insight, in which an iterative multitask Gaussian process learning algorithm is proposed to learn the model parameters.

For network models and graphical models, Huang et al. (2012) propose a transfer learning approach for a Gaussian graphical model. The goal of this method is to learn the brain connectivity network for Alzheimer's disease patients based on functional magnetic resonance image (fMRI) data.

Finally, statistical language models (see Zhai et al. 2008 for more details) can also be extended by transfer learning. Although task-specific statistical language models such as Latent Dirichlet allocation (Blei et al. 2003) for long text or Twitter-LDA (Zhao et al. 2011) for short text have different structures, the words and corresponding language models naturally share linguistic similarities. In this sense, to improve the performance of Twitter clustering, Jin et al. (2011) suggest an extended Twitter clustering scheme by transferring the knowledge learned from long texts to short Twitter texts.

Statistical methodologies

In transfer learning literature, assorted statistical methodologies are recommended to connect and transfer knowledge among multiple statistical models and data sources. Here most of the mainstream statistical methodologies are reviewed to build connections between statistical models in each domain.

Boosting-based weighing schemes are often used to conduct instance-based transfer learning. Based on the well-known adaptive boosting method (Freund and Schapire 1995), Dai et al. (2007) present the boosting algorithm TrAdaBoost, reducing the distribution differences between domains by adjusting the weights of instances. The traditional adaptive boosting is an ensemble method that creates a strong classifier from a number of basic classifiers like decision trees, where the basic classifiers are combined sequentially, by carefully adjusting the weights of training instances in each iteration. To extend the adaptive boosting method to conduct transfer learning on both source domains and target domains, TrAdaBoost adopts a different weighting mechanism which decreases weights of the instances in source domains that are dissimilar to target domains. By doing so, TrAdaBoost consequently allows to enhance the predictive performance in target domains by using the data in source domains. Following this line, boosting-based methods are considered under various scenarios, including multimodal TrAdaBoost (Wei et al. 2016) and multisource TrAdaBoost (Yao and Doretto 2010). Bagging (Breiman 1996) and bootstrap (Efron and Tibshirani 1994) are also extended as TrBagg (Kamishima et al. 2009) and double-bootstrapping source data selection (Lin et al. 2013) to construct an ensemble of learners in the context of instance transfer.

Bayesian modeling is one of the most common statistical techniques for transferring information across different tasks and models. Information on parameters can be easily transferred between models through shared prior distributions and common hyper-parameters. The use of Bayesian priors enables the transfer of information on parameters from one model to another. For instance, to predict products' field return rate during the warranty period, Tseng et al. (2016) propose a hierarchical Bayesian approach to model laboratory and field data collected from multiple products with a similar design. The information sharing between products is efficiently performed using a Dirichlet prior distribution, which leads to improved predictive performance especially for the products with few or even no failures in laboratory. Hierarchical Bayesian based transfer learning methods are widely investigated under different settings and in various applications. For linear models, Zhang et al. (2010) formulate l_q regularization-based multi-task feature selection in a Bayesian framework, which devises expectation-maximization (EM) algorithms to learn model parameters for every task. To link the model coefficients of old and new domains in degenerate biological systems, Zou et al. (2016) adopt a hierarchical structure to characterize the degeneracy and correlation structure of the domains. Moreover, Bouveyron et al. (2014) suggest a Bayesian approach for a mixture of linear regression models in which Dirichlet distribution and inverse-gamma distribution are imposed as priors. For Gaussian processes, the hierarchical Bayesian framework is considered to connect multiple related processes (Yu et al. 2005, Bonilla et al. 2007) and can thus benefit from joint estimation and knowledge transfer. In addition, it is appropriate to adopt Bayesian prior to transfer information between different statistical language models and tasks (Jin et al. 2011), since many language models, such as Latent Dirichlet Allocation (Blei et al. 2003), are themselves probabilistic generative models constructed in a Bayesian manner. Apart from parameter transfer, Bayesian probabilistic models are also exploited for transfer learning in terms of instance weighting in natural language process (NLP) applications (Jiang and Zhai 2007).

Additionally, regularization-based methods provide an effective framework of transfer learning by assuming similar patterns for the model parameters between sources, which are commonly exploited to build connections between linear models through numerous penalties. Specifically, suppose that there are K tasks in total contained in target and source domains. A regularization-based method typically considers the following penalized estimation:

$$\min_{\boldsymbol{B}_k} \sum_k g_k \, (\boldsymbol{X}_k, \, \boldsymbol{Y}_k, \, \boldsymbol{B}_k) \, + \, penalty\,(\boldsymbol{B}) \ ,$$

where g_k is an application-specified loss function with coefficients $\boldsymbol{B}_k$, the k-th row of $\boldsymbol{B}$, and $(\boldsymbol{X}_k, \, \boldsymbol{Y}_k)$ are training data of k-th task. A penalty term is chosen so as to make information on coefficients transferred across all tasks and domains. Many penalties are investigated for different applications under this framework. For instance, Liu, Ji and Ye (2009) adopt L_{21}-norm regularization for linear models to conduct feature selection across multiple domains by encouraging multiple predictors to share similar sparsity patterns, where the penalty is taken as the sum of l_2-norm over each variable, i.e., $penalty(\boldsymbol{B}) \ = \sum_i \sqrt{\sum_k B_{ki}^2}$. Similarly, Liu, Palatucci and Zhang (2009) propose the multi-task Lasso to select significant variables across related linear regression models by replacing l_1-norm regularization with the sum of l_∞ regularization, i.e., $penalty(\boldsymbol{B}) \ = \sum_i \max_k |B_{ki}|$. Following this line, many variants considering other penalties are investigated. For example, an adjusted l_1-norm regularization weighted by spatial information is given by Samarov et al. (2015) and a mixture of l_∞/l_1/ridge-based penalty is discussed in Samarov et al. (2016), where l_1 based penalty is $\sum_i \sum_k |B_{ki}|$ and ridge-based penalty is $\sum_i \sum_k B_{ki}^2$. Gonçalves et al. (2014) design a regularization-based approach induced by a Gaussian prior that characterizes a sparse dependency structure of tasks and extended linear and logistic regressions under this framework and then provide a flexible Gaussian copula model that relaxes the Gaussian marginal assumption (Gonçalves et al., 2016). Kernel extensions are considered in nonlinear models in the context of transfer learning. Zhang et al. (2014) investigate a regularization approach by imposing a matrix-variate Gaussian prior distribution and extend it using kernel methods. Nonlinear multi-task kernels for SVMs have also been studied (Evgeniou and Pontil 2004).

For unsupervised approaches, Song et al. (2015) investigate a PCA-based transfer learning approach and apply it on speech emotion recognition. As a popular PCA-like algorithm in data mining field, sparse coding-based transfer learning has also been extensively studied (Raina et al. 2007, Wei et al. 2016, Maurer et al. 2013). The key idea of sparse coding is to represent data vectors as sparse linear combinations of basic elements to allow homogenous representation structures to be shared between tasks.

For better illustration of the connection among statistical models, methodologies and transfer learning, we've summarized the relationship between transfer learning categories and statistical methodologies in Table 1. In Table 2, we list transfer learning literature that adopts various statistical models and methodologies.

Statistical models and methodologies are both critical in transfer learning and have been widely investigated. However, there remains a limited variety of statistical models extended using transfer learning despite their broad applications. As such, statistical transfer learning extensions for SPC and quality control are demonstrated in the following three sections. Firstly, autoregressive models are extended to a transfer learning version to describe non-contemporaneous relationships between sensors and for the rapid detection of landslides and slope failures (Zhang et al. 2017). Then, the prediction and monitoring of passenger inflows

Table 1. Relationship between transfer learning categories and statistical methodologies

	Boosting	Bagging/ bootstrap	Regularization	Bayesian framework
Instance-transfer	√	√		√
Feature-transfer			√	
Parameter-transfer			√	√

Table 2. Relationship between statistical models and statistical methodologies in transfer learning.

	Linear models and generalized linear models	Gaussian processes	Graphical/ network models	Hierarchical Bayesian models

Boosting, bagging and bootstrap	Dai et al. (2007);
	Wei et al. (2016);
	Yao and Doretto, (2010); …			

Regularization-based methods	Zhang and Tsung, (2017);
	Liu et al. (2009);
	Samarov et al. (2016);
	Gonçalves et al. (2016); …			

Bayesian frameworks	Zhang et al. (2010);
	Zou et al (2016);
	Bouveyron et al. (2014); …
		Shao et al. (2017);
		Yu et al. (2005); Bonilla et al. (2007); …
			Huang et al. (2012)	Song et al. (working paper);
				Jin et al. (2011); …

in a rail transit system will be considered in a statistical transfer learning framework (Song et al. working paper). After that, we introduce a parameter-based transfer learning approach for shape deviation prediction and to control 3D-printed products with distinct shapes based on geometric error decomposition and modeling by incorporating engineering knowledge and experimental design (Cheng et al. 2017).

Statistical transfer learning for landslide monitoring

Landslides are common geographical activities that result in large quantities of rock, earth and debris flowing down hillslopes, leading to thousands of casualties and billions of dollars in infrastructure damage every year around the world (Yang et al. 2010). To detect and predict such abnormal geographical behavior, accelerometer-based sensor systems are widely used in landslide-prone sites. Autocorrelated time series are often used to describe sensor readings over time and autoregressive (AR) models are used for prediction (Pu et al. 2015). SPC procedures for monitoring such autocorrelated processes have also been widely studied (Psarakis and Papaleonida 2007, Castagliola and Tsung 2005).

Multiple time series are collected from several landslide-prone sites with multiple sensors assigned. The relationship between such time series can be mainly divided into two categories. The first is contemporaneous relationships, which contain spatially correlated residuals and time-lagged effects. Existing models such as vector autoregressive (VAR) models (Lütkepohl 2005) and spatial-temporal models (Cressie and Wikle 2015) are capable of capturing the contemporaneous relationship between multiple time series. The second is non-contemporaneous relationships, which means the autoregressive structure of sensors in different sites may share similarities. The similarities may result from the same type of sensors adopted and similar geographical activities on site.

In this landslide application, sensor readings are usually recorded during different periods, varying from site to site. Thus existing statistical models such as AR and VAR can only be used separately for sensors in each landslide-prone site and may fail to provide accurate modeling for sensors at a site in the early stages with fewer observations. To improve modeling accuracy, it is helpful to jointly model the time series from different sites by considering the non-contemporaneous relationships, which cannot be conducted using the above existing methods. To this end, a transfer learning approach is proposed to extend autoregressive models for slope failure monitoring, where both contemporaneous and non-contemporaneous relationships of multiple autocorrelated time series are considered. Intuitively, it is expected that information will be transferred from "experienced" sites to early-stage sites by discovering non-contemporaneous dependency structure in AR coefficients. The detailed statistical model is as follows.

Suppose that data are collected from K different landslide-prone sites. For the k-th site, p_k sensors are assigned to collect measurements simultaneously. Sensor readings over time are denoted as time series $\{y_{i,t}^{[k]}\}_{t=1}^{T_k}$ for the i-th sensor. Consider an $AR(L)$ model for i-th sensor at the k-th site:

$$y_{i,t}^{[k]} = \sum_{l} \beta_{i,l}^{[k]} y_{i,t-l}^{[k]} + \varepsilon_{i,t}^{[k]}. \qquad [1]$$

Here, $\beta_{i,l}^{[k]}$ refers to the autoregressive coefficient representing the relationship between the i-th sensor's current measurements and its lag l measurements at site k. We assume that $\boldsymbol{\varepsilon}_t^{[k]} = (\varepsilon_{1,t}^{[k]}, \ldots, \varepsilon_{p_k,t}^{[k]})'$ is a $p_k \times 1$

vector of error terms, following $N\ (0,\ \mathbf{\Sigma}^{[k]})$, where the covariance matrix $\mathbf{\Sigma}^{[k]}$ is adopted to characterize the contemporaneous spatial correlation between sensors within site k. On the other hand, similarly to Gonçalves et al. (2016), a Gaussian prior is imposed over the time-lagged coefficients across the AR(L) models of sensors to transfer information across sites and sensors. Specifically, assume that

$$\boldsymbol{\beta}_l = \left(\beta_{1,l}^{[1]}, \ldots, \beta_{p_1 l}^{[1]}, \ldots, \beta_{1,l}^{[K]}, \ldots \beta_{p_K l}^{[K]}\right) \sim N\left(0,\ \mathbf{\Sigma}_M e^{-\theta l}\right) \qquad [2]$$

independently for each l. Here, Σ_M denotes a hidden dependency structure among the autoregressive models on all sites. Moreover, the decreasing time-lagged effect is accounted for using the exponential function $e^{-\theta l}$, which depends on the lag term l and parameter θ.

The joint likelihood over all sites can be derived to infer the above model parameters from historical data. For site k, let AR coefficients $\boldsymbol{B}^{[k]} = \{\beta_{i,l}^{[k]}$ for $i = 1, \ldots p; l = 1, \ldots L\}$. The log-likelihood conditional on $\boldsymbol{B}^{[k]}$ and $\mathbf{\Sigma}^{[k]}$ is written as $g_k(\mathbf{\Sigma}^{[k]},\ \boldsymbol{B}^{[k]}; \boldsymbol{Y}^{[k]})$. Considering the imposed prior distribution [2] that characterizes the non-contemporaneous dependency structure, in a Bayesian perspective, if we assume non-informative priors for $(\mathbf{\Sigma},\ \mathbf{\Sigma}_M,\ \theta)$, the posterior log-likelihood for the entire transfer learning framework is

$$\sum_k g_k(\mathbf{\Sigma}^{[k]},\ \boldsymbol{B}^{[k]}; \boldsymbol{Y}^{[k]}) + \sum_l \log(\pi(\boldsymbol{\beta}_l | \mathbf{\Sigma}_M, \theta)) + const, \qquad [3]$$

where $\pi(\cdot)$ denotes the prior distribution of $\boldsymbol{\beta}_l$ in Eq. [2]. To get estimation, we can iteratively update the parameters by maximizing the posterior above. Specifically, with $\mathbf{\Sigma}_M$ and θ fixed, $\{\boldsymbol{B}^{[k]}, \mathbf{\Sigma}^{[k]}\}_{k=1}^K$ can be obtained by maximizing [3] through a coordinate ascent algorithm; with $\{\boldsymbol{B}^{[k]}, \mathbf{\Sigma}^{[k]}\}_{k=1}^K$ fixed, we can estimate $\mathbf{\Sigma}_M$ and θ. The latter term in [3] can be viewed as a regularization term that links parameters in all models together.

A Monte Carlo simulation is performed to show the performance of the transfer-learning extended method in Phase I estimation of autoregressive processes. Assume that there are only two sites in total with three sensors at each site. The lengths of observed time series are different: sensors in the first site have only 50 observations while those in the second have 500 observations. Here the first sensor can be regarded as an early-stage site with fewer observations, while the second is an “experienced” site with enough historical data. Let the maximum lag be 4. Suppose that the true coefficients of AR(4) models are generated from a 6 × 6 matrix $\mathbf{\Sigma}_M$, where

$$\mathbf{\Sigma}_M = \begin{bmatrix} 1 & \rho & \ldots & \rho & \rho \\ \rho & 1 & \ldots & \rho & \rho \\ \ldots & \ldots & & \ldots & \ldots \\ \rho & \rho & \ldots & \rho & 1 \end{bmatrix}_{6\times 6}.$$

A larger ρ in $\mathbf{\Sigma}_M$ means the time series of sensors in the simulation tend to evolve in a more similar manner. When $\rho = 1$, the true AR coefficients in all of the sensors are the same. Here, the non-transfer baseline method is taken as separately estimating an AR model within each site. In this simulation, 10,000 replicates are conducted for each ρ ranging from 0–0.99. The transfer learning method and non-transfer method are compared via their averaged root mean square errors (RMSEs) at the first site. The result in Zhang et al. (working paper) show that AR models extended by transfer learning significantly outperform the non-transfer baseline approach and their estimation performance is improved with ρ increasing, which illustrates the necessity and effectiveness of investigating non-contemporaneous relationships and extending the AR model in a statistical transfer learning framework. Further evaluation studies will be conducted to compare predictive performance with more baseline methods considered.

For the monitoring part, statistical transfer learning can provide all-round improvements to extend the existing SPC schemes: more accurate Phase I estimation and more rapid detection in Phase II online monitoring.

For Phase I analysis, the simulation result has shown the effectiveness of estimating in-control AR coefficients for sensors and landslides-prone sites, especially for those with limited historical data. Outlier detection and change-point diagnosis can also be investigated in this transfer learning framework with reasonable assumptions.

Adopting statistical transfer learning in Phase II landslide monitoring is more challenging than for Phase I analysis but worthwhile. An improved understanding of the target site/sensor can be obtained with the help of source sites/sensors, thus leading to rapid detection. For Phase II landslide monitoring, shifts in the autoregressive structure $\boldsymbol{B}^{[k]}$ and the spatial covariance $\mathbf{\Sigma}^{[k]}$ represent abnormal states and need to be monitored. To monitor the autoregressive structure of the k-th site $\boldsymbol{B}^{[k]}$, a generalized likelihood ratio (GLR)-based SPC scheme may be considered where the transferred parameters $\{\boldsymbol{\Sigma}_M,\ \theta\}$ obtained in Phase I can provide a more precise GLR statistic. For the covariance matrix $\boldsymbol{\Sigma}^{[k]}$, a residual-based control chart may be considered in which the improved estimation of $\boldsymbol{B}^{[k]}$ using transfer learning can reduce the noise when constructing statistics.

Statistical transfer learning for monitoring in urban rail transit systems

With the proliferation of smart cities, public transportation services such as urban railway transit (URT) systems are playing an increasingly important role in commuter mobility. For instance, Hong Kong's MTR carries more than five million passengers every day. Aperiodic incidents and events, such as traffic accidents, traffic controls, celebrations, protests and disasters, can lead to abnormal passenger streams on public transportation systems, which can result in serious accidents such as stampedes due to overcrowding in extreme cases. It is important to predict passenger flows and conduct monitoring schemes to prevent accidents due to excessive passenger flow within URT systems. A large body of transportation engineering research has analyzed passenger flows to estimate travel times (Jaiswal et al. 2010), to simulate the distribution and movement of passengers in an area (Setti and Hutchinson 1994), and to predict passenger selection behavior (Ren et al. 2012). However, there has been a dearth of research on the influence of passenger crowding on entire URT systems. It is therefore crucial to develop a statistical methodology to understand, predict, and monitor the number of passengers and the degree of crowding in a URT system in real time.

However, there remain several challenges to be addressed. Most notably, the early warning problem aims to make proactive decisions based on predicted future passenger flows, which are significantly different from conventional statistical process control problems monitoring current processes based on past data. Hence the modeling and predictive performance of the methodology are crucial. Furthermore, the large number of stations in real URT systems particularly requires scalable early warning schemes. In this section, a scalable and predictive-based SPC scheme is sought. To begin with, a single model will be built for each station, allowing for high flexibility and satisfying predictive ability. Moreover, due to the special properties of count data, conventional methods such as functional data analysis cannot be applied to inflow passenger profiles. To tackle this, a hierarchical model is adopted to describe inflow counting data in each station, where $x_k(t)$ represents the number of passengers entering station k at time t. Particularly, $x_k(t)$ is assumed to be a Poisson random variable with intensity parameter $\lambda_k(t)$. After this, the focus is on $\log\lambda_k(t)$, which is inspired by the log-linear model for categorical data (Li et al. 2009). Specifically, the following state space model is proposed for each station $k,\ \ k = 1, 2, \ldots, N$

$$log\lambda_k\,(t) = \alpha_0^{(k)} + \sum_l \alpha_l^{(k)}\, log\lambda_k\,(t-l) + \epsilon_k\,(t)\,, \tag{4}$$

where $\epsilon_k(t)$'s are independent across stations and follow $N(0, \sigma_k^2)$.

When the above hierarchical model of URT systems is extended in a transfer learning framework, several advantages immediately appear and show promising potential. A URT system typically covers an entire city and connects various areas in a network structure. Hence, the autoregressive structure of Poisson intensities of different stations can be highly related as the stations are spatially close or of the same category, such as downtown, business districts, and residential areas. Consequently, the parameters of log-linear models in Eq. [4] for similar stations are expected to be closely related to each other. Isolating each station is not enough to fully utilize the useful information from other related stations. Instead, it is beneficial to learn multiple tasks (stations) simultaneously under a transfer learning framework.

In more detail, coefficients across different stations are assumed to share a common prior distribution, i.e.,

$$\alpha_l = \left(\alpha_l^{(1)}, \alpha_l^{(2)}, \ldots, \alpha_l^{(N)}\right)^T \sim N\,(0, \Sigma)\,.$$

Here, Σ depicts the inherent relatedness structure among stations. Accordingly, stations can be clustered into different categories by treating Σ as a similarity measure. In this sense, transferring knowledge across stations is expected to improve estimation and thus predictive performance as well as reveal the hidden inherent structure among stations. A similar idea has been raised in the disease mapping problem by borrowing information from neighboring regions (Blangiardo and Cameletti 2015). The major difference in the URT problem is that the knowledge can be shared among stations from similar category of functional zones in addition to stations which are spatially close.

To monitor the passenger inflow in URT, a SPC scheme is required to estimate parameters in Phase I and implement monitoring in Phase II. There are technical challenges in each phase to applying SPC procedures for passenger inflow monitoring applications: during Phase I, a state space model is adopted for each station to describe the latent Poisson intensity where the parameters are estimated together across various stations via a common prior distribution. The estimation of the hierarchical model in transfer learning can improve performance, especially for newly activated stations with less data, where inherent relatedness structure Σ is exploited over the entire URT system. EM algorithms or particle filtering are potential implementations of Phase I parameter inference. During Phase II, a predictive-based monitoring approach is recommended because proactive decisions must be made to avoid accidents. Specifically, for each station, based on a prediction with a given lead time, e.g., 20 min, the personnel in a URT system can decide on a course of action. If the predicted values exceed the prespecified limit, this signifies that upcoming passenger inflow poses a considerable challenge to the operation of the station. Hence, an early warning scheme should be activated, and an alarm may be signaled if necessary. Here, the control limits are determined by each station's specific capacity and infrastructure characteristics in collaboration with Phase I analysis. On the other hand, if the predicted values differ considerably from the later observed actual values, model calibration and online updating are needed. Moreover, to solve the scalability issue, the approach of Mei (2010) can be adopted to combine transfer learning predictions in multiple stations to obtain SPC statistics for detection throughout the URT network, where a local statistic is generated for monitoring based on the predicted passenger flow at each station, and all of these local statistics are then integrated to make a final decision. Further work following this line needs to be conducted.

Statistical transfer learning for 3D printing quality control

3D printing is one of the most promising manufacturing techniques since it enables the direct fabrication of products of complex shapes with few design constraints. However, dimensional inaccuracy remains one of the most concerned quality issues limiting the technology's application. Many shape deviation modeling and compensation methods have been proposed to improve the geometric accuracy of fabricated products, including those devised by Huang et al. (2014, 2015), Luan and Huang (2015), and Wang et al. (2017), among others. However, it has been shown that these methods only perform well for products with specific shapes and usually require re-estimating model parameters for new shapes. There are three major challenges to predicting geometric errors and deriving effective compensation plans for new products before fabrication. First, the geometric error-generating mechanism is very complex and there are multiple error sources, which make it difficult to build an effective model from the first principle. Second, as there are a wide variety of complex shapes, it is only feasible to fabricate limited products for limited shapes due to resource constraints; it is therefore unfeasible to build a single comprehensive model based on data-driven methods that require large amounts of data. Third, it is hard to establish connections between the shape deviation of products fabricated with distinct shapes.

To tackle the above challenges in quality control, Cheng et al. (2017) propose an in-plane shape deviation modeling scheme from a statistical transfer learning perspective. In this scheme, a parameter-based transfer learning approach is adopted based on geometric error decomposition and modeling by incorporating engineering knowledge and experimental design.

Although the error-generating mechanism is complex, the geometric error of a fabricated product can be generally decomposed into two components: (i) a shape-independent error component, which means the model parameters for this component are the same for different shapes, and (ii) a shape-specific error component, corresponding to a specific term for the deviation model of each different shape. The motivation of the above decomposition is based on the observation that the deviations of the same point located on the boundaries of two different in-plane shapes are usually different. One cause of the difference in the fused deposition modeling (FDM) process is that the error induced by depositing material is highly related to the moving path of the extruder. The error at a boundary point is thus expected to have two components: one is generally shared by all shapes and the other is highly related to the shape features. Suppose the input shape of a designed product is ψ_0 and the final shape of the fabricated product is ψ, then

$$\psi = \psi_0 + e_0\left(\psi_0\right) + e_1\left(\psi_0\right) + \varepsilon,$$

where $e_0(\psi_0)$ is the shape-independent deviation, $e_1(\psi_0)$ is the shape-specific deviation, and ε represents the random error.

Since the measured deviation always contains both error components, it is difficult to isolate the shape-independent error from the shape-specific error with simple data-driven methods. To tackle this, Cheng et al. (2017) propose approximating the shape-independent error with the deviation of a point inside a product since the shape-specific error is majorly incurred by shape boundary features. Based on this assumption, an experiment was designed to investigate and model the shape-independent error $e_0(\psi_0)$. First, a circular plate with a radius of 60 mm and a square plate with a side length of 100 mm were fabricated via an FDM 3D printer, as shown in Figure 1. On each plate, 81 marks are designed and fabricated in an 9 × 9 grid pattern. The interval of the grid is 10 mm. Each mark is designed as a circular hole with a radius of 1 mm and its center represents the mark's position. Such a mark is small enough to rarely affect the material shrinkage, and its circular shape can facilitate the measurement process for obtaining its position. Since these marks are inside the product, the measured deviations at these locations are rarely related to the shape features and can hence be used to approximate the shape-independent errors and model $e_0(\psi_0)$ in the Cartesian coordinate system. Suppose the designed marks are fabricated at (x_i, y_i) and their measured locations are denoted as (x_i', y_i'), $i = 1, \ldots M$. The measurement of the shape-independent error at (x_i, y_i) can then be represented as $(e_{0x}(x_i, y_i), e_{0y}(x_i, y_i)) = (x_i' - x_i, y_i' - y_i)$. Based on the significant linear pattern observed in the measurements of shape-independent error, the following linear regression models are applied to model the shape-independent error in the x-direction and y-direction separately:

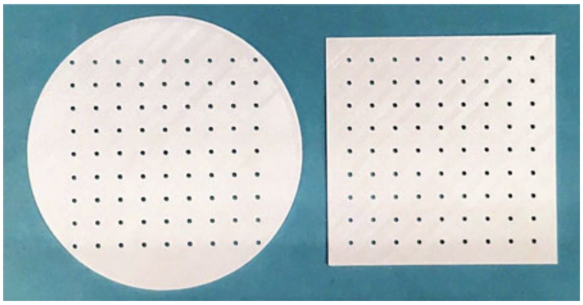

Figure 1. Fabricated circular plate and square plate for modeling shape-independent error. Adapted from Cheng et al. (2017).

$$\begin{cases} e_{0x}\left(x, y\right) = \beta_{1x}x + \beta_{2x}y + e_x \\ e_{0y}\left(x, y\right) = \beta_{1y}x + \beta_{2y}y + e_y \end{cases}.$$

The linear coefficients in the above model can be estimated using the data from our measurements. The result shows that this model can accurately predict the shape-independent error. After this step, the above shape-independent error model can be transferred to predict the shape-independent error component for any new shape. Suppose the input shape is ψ_0. The shape incorporating the predicted shape-independent error can then be represented as

$$\psi' = \left\{\left(x + e_{0x}\left(x, y\right), y + e_{0y}\left(x, y\right)\right) \mid \left(x, y\right) \in \psi_0\right\}.$$

To investigate the shape-specific error, the input shape ψ_0, the shape incorporating the shape-independent error ψ' and the final product shape ψ are represented as $r_0(\theta)$, $r'(\theta)$, and $r(\theta)$ in the polar coordinate system, respectively. $y(\theta) = r(\theta) - r_0(\theta)$ denotes the measured deviation profile; $f_0(\theta) = r'(\theta) - r_0(\theta)$ denotes the predicted shape-independent deviation profile. The shape-specific error is then isolated from the shape-independent error by

$$y\left(\theta\right) - f_0\left(\theta\right) = r\left(\theta\right) - r'\left(\theta\right) = f_1\left(\theta\right) + \varepsilon\left(\theta\right),$$

where $f_1(\theta)$ denotes the shape-specific error and $\varepsilon(\theta)$ is the random error. To demonstrate this, two circular products with radii of 10 mm and 30 mm and two square products with side lengths of 20 mm and 60 mm are fabricated and the corresponding deviation profiles are measured. For each product, the shape-independent deviation profile is predicted and the shape-specific deviation profile is calculated. It is observed that the shape-independent deviation model can capture a major part of the total deviation, and the remaining shape-specific deviation profile has a shape-specific pattern around the 0 line and appears to be rarely affected by the size of a product. This preliminary result shows that the proposed parameter-based transfer learning approach greatly improves the extendibility of the shape deviation model to infer new shapes. Future studies will focus on modeling the relationship between the shape-specific deviation profiles and the shape features, which may further improve the model's predictive performance and thus increase the shape fidelity of fabricated products via deviation compensation.

In addition to predicting shape deviation and controlling for 3D printed products with distinct shapes, there is also a great need for statistical transfer learning for 3D printing quality control from source machines to new target machines. When a 3D printing machine changes, the shape deviation model must be revised, which indicates the need to regenerate training data. It is thus of great importance to transfer the knowledge acquired from a source machine to a target machine.

Conclusion

The rapid development of information technology, together with advances in sensory and data acquisition techniques, have made it possible to conduct statistical inferences based on multiple domains. To share domain knowledge and combine multiple data sources, transfer learning techniques have been investigated and utilized in many real-world applications. In this article, besides a general review of transfer learning, a summary of transfer learning literature is provided based on statistical models adopted in individual domains and statistical techniques that are exploited to conduct knowledge transfer. Furthermore, transfer learning techniques are applied to SPC and quality control applications with various data types: autocorrelated sensor readings for landslide detection, Poisson counting processes for urban rail transit monitoring, and shape deviation prediction and control for 3D-printed products.

Several research issues in the context of statistical transfer learning remain to be addressed. First, transfer learning techniques have been mainly applied in a limited variety of applications. As stated in the above sections a great number of domain-specific statistical models have the potential to be extended for transfer learning in further applications. For example, as two major dimensions in the information quality framework (Kenett and Shmueli, 2016), integrating data and generalizing findings are commonly required for applications that integrate complex surveys (Kenett, 2016) and statistics data (Dalla and Kenett, 2015). Developing application-specified transfer learning methods should be of great value for practical use. Second, incorporating engineering knowledge into a statistical transfer learning work is also of great interest to the fields of statistics and industrial engineering. For example, in the 3D printing application the shape error decomposition and modeling by incorporating engineering knowledge makes it possible to transfer the model to infer new shapes. A tailor-made statistical model for such applications can pave the way to integrating engineering insight and transfer learning approaches. Third, since the above extensions will inevitably lead to more complex models, efforts at both theoretical analysis and numerical studies are increasingly desired for model inference and parameter estimations when conducting statistical transfer learning approaches. For example, in the landslide monitoring application, the transfer learning model for time-lagged regressions can be inferred through empirical Bayesian and maximum a posteriori (MAP). However, it is also possible to undertake model inference in a full Bayesian manner through Markov chain Monte Carlo methods (Gilks et al. 1995, Rubinstein et al. 2016) or variational inference (Wainwright et al. 2008, Blei et al. 2016). Further statistical analysis is required to make a choice in such transfer learning applications. Specifically, asymptotic properties of the estimators and convergence properties of algorithms are needed to support the choices of statistical methodologies for transfer learning applications. Finally, for SPC extensions, as mentioned earlier, there is substantial room for improvement using statistical transfer learning in both Phase I analysis and Phase II monitoring. In Phase I analysis, it is challenging but worthwhile to conduct statistical transfer learning for in-control parameter estimation, outlier detection and change-point diagnosis, for an improved understanding of in-control situations. During Phase II monitoring, one of the major problems is constructing statistics with the assistance of statistical transfer learning for rapid anomaly detection. Issues such as the online updating of transferred parameters are worth considering.

About the authors

Fugee Tsung is Professor of the Department of Industrial Engineering and Logistics Management (IELM), Director of the Quality and Data Analytics Lab, at the Hong Kong University of Science and Technology (HKUST). He is a Fellow of the Institute of Industrial Engineers (IIE), Fellow of the American Society for Quality (ASQ), Fellow of the American Statistical Association (ASA), Academician of the International Academy for Quality (IAQ) and Fellow of the Hong Kong Institution of Engineers (HKIE). He is Editor-in-Chief of *Journal of Quality Technology* (JQT), Department Editor of the *IIE Transactions*, and Associate Editor of *Technometrics*. He has authored over

100 refereed journal publications, and is the winner of the Best Paper Award for the *IIE Transactions* in 2003 and 2009. He received both his M.Sc. and Ph.D. from the University of Michigan, Ann Arbor and his B.Sc. from National Taiwan University. His research interests include quality engineering and management to manufacturing and service industries, statistical process control and monitoring, industrial statistics, and data analytics.

Ke Zhang is a Ph.D. candidate in Department of Industrial Engineering and Logistics Management at Hong Kong University of Science and Technology. He received a Bachelor's degree in Statistics from University of Science and Technology of China in 2014. His research interests include statistical modeling, process control and data mining.

Longwei Cheng is a Ph.D. candidate in Department of Industrial Engineering and Logistics Management at Hong Kong University of Science and Technology. He received a Bachelor's degree in Automation from University of Science and Technology of China in 2014. His research interests include statistical modeling and quality control.

Zhenli Song is a Ph.D. candidate in Department of Industrial Engineering and Logistics Management at Hong Kong University of Science and Technology. He received a Bachelor's degree in Statistics from University of Science and Technology of China in 2015. His research interests include statistical modeling and data mining.

Acknowledgments

The authors are grateful to professor Steven Rigdon, professor Xiaoming Huo and the two anonymous reviewers for their comments and suggestions that greatly improved our article.

Funding

Professor Tsung's research was supported by the RGC GRF 16203917.

References

Argyriou, A., T. Evgeniou, and M. Pontil. 2007a. Multi-task feature learning. In *Advances in neural information processing systems*, B., Schölkopf, J., Platt, T., Hoffman (Eds.), Vol. 19, pp. 41–48. Cambridge: MIT Press.

Blangiardo, M., and M. Cameletti. 2015. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2016. Variational inference: A review for statisticians. *arXiv preprint arXiv* 1601:00670.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

Blitzer, J., M. Dredze, and F. Pereira. 2007, June. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL* (Vol. 7, pp. 440–447) Stroudsburg: Association for Computational Linguistics.

Blitzer, J., R. McDonald, and F. Pereira. 2006, July. Domain adaptation with structural correspondence learning. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 120–128). Stroudsburg: Association for Computational Linguistics.

Bonilla, E. V., K. M. A. Chai, and C. K. Williams. 2007, December. Multi-task Gaussian process prediction. In *NIPs* (Vol. 20, pp. 153–160). Vancouver: Neural Information Processing Systems.

Bouveyron, C., and J. Jacques. 2014. Adaptive mixtures of regressions: Improving predictive inference when population has changed. *Communications in Statistics-Simulation and Computation* 43 (10):2570–2592.

Breiman, L. 1996. Bagging predictors. *Machine learning* 24 (2):123–140.

Castagliola, P., and F. Tsung. 2005. Autocorrelated SPC for non-normal situations. *Quality and Reliability Engineering International* 21 (2):131–161.

Cressie, N., and C. K. Wikle. 2015. *Statistics for spatio-temporal data*. Hoboken: John Wiley & Sons.

Cheng, L., F. Tsung, and A. Wang. 2017. A transfer learning perspective for modeling shape deviations in additive manufacturing. *IEEE Robotics and Automation Letters* 2 (4):1988–1993.

Dai, W., Q. Yang, G. R. Xue, and Y. Yu. 2008, July. Self-taught clustering. In Proceedings of the 25th International Conference on Machine Learning (pp. 200–207). New York: ACM.

Dai, W., Q. Yang, G. R. Xue, and Y. Yu. 2007, June. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning (pp. 193–200). New York: ACM.

Dalla Valle, L., and R. S. Kenett. 2015. Official Statistics Data Integration for Enhanced Information Quality, *Quality and Reliability Engineering International* Vol. 31, No. 7:pp. 1281–1300.

Efron, B., and R. J. Tibshirani. 1994. *An introduction to the bootstrap*. CRC Press.

Evgeniou, A., and M. Pontil. 2007. Multi-task feature learning. *Advances in Neural Information Processing Systems* 19:41.

Evgeniou, T., and M. Pontil. 2004, August. Regularized multi–task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 109–117). New York: ACM.

Freund, Y., and R. E. Schapire. 1995, March. A desicion-theoretic generalization of on-line learning and an application to boosting. In European conference on computational learning theory (pp. 23–37). Heidelberg: Springer Berlin Heidelberg.

Gilks, W. R., S. Richardson, and D. Spiegelhalter. Eds.).1995. *Markov chain Monte Carlo in practice*. Boca Raton: CRC Press.

Gonçalves, A. R., P. Das, S. Chatterjee, V. Sivakumar, F. J. Von Zuben, and A. Banerjee. 2014, November. Multi-task sparse structure learning. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 451–460). New York: ACM.

Gonçalves, A. R., F. J. Von Zuben, and A. Banerjee. 2016. Multitask sparse structure learning with Gaussian copula models. *Journal of Machine Learning Research* 17 (33):1–30.

Huang, S., J. Li, K. Chen, T. Wu, J. Ye, X. Wu, and L. Yao. 2012. A transfer learning approach for network modeling. *IIE Transactions* 44 (11):915–931.

Huang, Q., H. Nouri, K. Xu, Y. Chen, S. Sosina, and T. Dasgupta. 2014. Statistical predictive modeling and compensation of geometric deviations of three-dimensional printed products. *Journal of Manufacturing Science and Engineering* 136 (6):061008.

Huang, Q., J. Zhang, A. Sabbaghi, and T. Dasgupta. 2015. Optimal offline compensation of shape shrinkage for three-dimensional printing processes. *IIE Transactions* 47 (5):431–441.

Jaiswal, S., J. Bunker, and L. Ferreira. 2010. Influence of platform walking on BRT station bus dwell time estimation: Australian analysis. *Journal of Transportation Engineering*, 136 (12), 1173–1179.

Jiang, J., and C. Zhai. 2007, June. Instance weighting for domain adaptation in NLP. In ACL (Vol. 7, pp. 264–271). Stroudsburg: Association for Computational Linguistics.

Jin, O., N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. 2011, October. Transferring topical knowledge from auxiliary long texts for short text clustering. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 775–784). New York: ACM.

Kamishima, T., M. Hamasaki, and S. Akaho. 2009, December. TrBagg: A simple transfer learning method and its application to personalization in collaborative tagging. In Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on (pp. 219–228). Piscataway: IEEE.

Kenett, R. S. 2016. On generating high InfoQ with Bayesian networks, *Quality Technology and Quantitative Management* 13 (3).

Kenett, R. S., and G.. Shmueli. 2016. *Information quality: the potential of data and analytics to generate knowledge*, Hoboken: John Wiley and Sons.

Lawrence, N. D., and J. C. Platt. 2004, July. Learning to learn with the informative vector machine. In Proceedings of the Twenty-First International Conference on Machine Learning (p. 65). New York: ACM.

Lee, S. I., V. Chatalbashev, D. Vickrey, and D. Koller. 2007, June. Learning a meta-level prior for feature relevance from multiple related tasks. In Proceedings of the 24th International Conference on Machine Learning (pp. 489–496). New York: ACM.

Li, J., C. Zou, and F. Tsung. 2009, August. Monitoring multivariate binomial data via log-linear models. Proceedings of the 1 st INFORMS International Conference on Service Science. Catonsville: INFORMS.

Liao, X., Y. Xue, and L. Carin. 2005, August. Logistic regression with an auxiliary data source. In Proceedings of the 22 nd International Conference on Machine Learning (pp. 505–512). New York: ACM.

Lin, D., X. An, and J. Zhang. 2013. Double-bootstrapping source data selection for instance-based transfer learning. *Pattern Recognition Letters* 34 (11):1279–1285.

Liu, J., S. Ji, and J. Ye. 2009, June. Multi-task feature learning via efficient l 2, 1-norm minimization. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (pp. 339–348). Corvallis: AUAI Press.

Liu, H., M. Palatucci, and J. Zhang. 2009, June. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 649–656). New York: ACM.

Lu, J., V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. 2015. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems* 80:14–23.

Luan, H., and Q. Huang. 2015, August. Predictive modeling of in-plane geometric deviation for 3D printed freeform products. In Automation Science and Engineering (CASE), 2015 IEEE International Conference on (pp. 912–917). Piscataway: IEEE.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maurer, A., M. Pontil, and B. Romera-Paredes. 2013, January. Sparse coding for multitask and transfer learning. In *ICML* (2) (pp. 343–351). Princeton: The International Machine Learning Society.

Mei, Y. 2010. Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* 97 (2):419–433.

Pan, S. J., and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (10):1345–1359.

Pan, S. J., V. W. Zheng, Q. Yang, and D. H. Hu. 2008, July. Transfer learning for wifi-based indoor localization. In Association for the advancement of artificial intelligence (AAAI) workshop (p. 6). Palo Alto: The Association for the Advancement of Artificial Intelligence.

Pan, S. J., J. T. Kwok, and Q. Yang. 2008, July. Transfer Learning via Dimensionality Reduction. In *AAAI* (Vol. 8, pp. 677–682). Palo Alto: The Association for the Advancement of Artificial Intelligence.

Pan, W., E. W. Xiang, N. N. Liu, and Q. Yang. 2010, July. Transfer Learning in Collaborative Filtering for Sparsity Reduction. In *AAAI* (Vol. 10, pp. 230–235). Palo Alto: The Association for the Advancement of Artificial Intelligence.

Pu, F., J. Ma, D. Zeng, X. Xu, and N. Chen. 2015. Early warning of abrupt displacement change at the Yemaomian landslide of the Three Gorge Region, China. *Natural Hazards Review* 16 (4):04015004.

Psarakis, S., and G. E. A. Papaleonida. 2007. SPC procedures for monitoring autocorrelated processes. *Quality Technology and Quantitative Management* 4 (4):501–540.

Raina, R., A. Battle, H. Lee, B. Packer, and A. Y. Ng. 2007, June. Self-taught learning: transfer learning from unlabeled data. In Proceedings of the 24th International Conference on Machine Learning (pp. 759–766). New York: ACM.

Ren, H., J. Long, Z. Gao, and P. Orenstein. 2012. Passenger assignment model based on common route in congested transit networks. *Journal of Transportation Engineering* 138 (12):1484–1494.

Rubinstein, R. Y., and D. P. Kroese. 2016. *Simulation and the Monte Carlo method*. Hoboken: John Wiley and Sons.

Samarov, D. V., D. Allen, J. Hwang, Y. Lee, and M. Litorja. 2016. A Coordinate Descent based approach to solving the Sparse Group Elastic Net. *Technometrics* (just-accepted).

Samarov, D. V., J. Hwang, and M. Litorja. 2015. The spatial lasso with applications to unmixing hyperspectral biomedical images. *Technometrics* 57 (4):503–513.

Schwaighofer, A., V. Tresp, and K. Yu. 2004. Learning Gaussian process kernels via hierarchical Bayes. In *Advances in Neural Information Processing Systems* (pp. 1209–1216). Vancouver: Neural Information Processing Systems.

Setti, J. R., and B. G. Hutchinson. 1994. Passenger-terminal simulation model. *Journal of Transportation Engineering* 120 (4):517–535.

Shao, C., J. Ren, H. Wang, J. J. Jin, and S. J. Hu. 2017. Improving machined surface shape prediction by integrating multitask learning with cutting force variation modeling. *Journal of Manufacturing Science and Engineering* 139 (1): 011014.

Song, P., W. Zheng, J. Liu, J. Li, and X. Zhang. 2015, November. A novel speech emotion recognition method via transfer PCA and sparse coding. In *Chinese Conference on Biometric Recognition* (pp. 393–400). Berlin: Springer International Publishing.

Song, Z., K. Zhang, and F. Tsung. 2017. A Multi-task learning approach for improved forecast of passenger inflow rates into a URT System. *Working paper*.

Taylor, M. E., and P. Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10 (Jul):1633–1685.

Tseng, S. T., N. J. Hsu, and Y. C. Lin. 2016. Joint modeling of laboratory and field data with application to warranty prediction for highly reliable products. *IIE Transactions* 48 (8):710–719.

Wainwright, M. J., and M. I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1 (1–2):1–305.

Wang, A., S. Song, Q. Huang, and F. Tsung. 2017. In-plane shape-deviation modeling and compensation for fused deposition modeling processes. *IEEE Transactions on Automation Science and Engineering* 14 (2): 968–976.

Wei, Y., Y. Zheng, and Q. Yang. 2016, August. Transfer knowledge between cities. In Proceedings of the 22 nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1905–1914). New York: ACM.

Weiss, K., T. M. Khoshgoftaar, and D. Wang. 2016. Transfer Learning Techniques. In *Big Data Technologies and Applications* (pp. 53–99). Springer International Publishing.

Wu, P., and T. G. Dietterich. 2004, July. Improving SVM accuracy by training on auxiliary data sources. In Proceedings of the Twenty-first International Conference on Machine Learning (p. 110). New York: ACM.

Xu, Q., and Q. Yang. 2011. A survey of transfer and multitask learning in bioinformatics. *Journal of Computing Science and Engineering* 5 (3):257–268.

Yang, X., and L. Chen. 2010. Using multi-temporal remote sensor imagery to detect earthquake-triggered landslides. *International Journal of Applied Earth Observation and Geoinformation* 12 (6):487–495.

Yao, Y., and G. Doretto. 2010, June. Boosting for transfer learning with multiple sources. In Computer vision and pattern recognition (CVPR), 2010 IEEE conference on (pp. 1855–1862). Piscataway: IEEE.

Yu, K., V. Tresp, and A. Schwaighofer. 2005, August. Learning Gaussian processes from multiple tasks. In Proceedings of the 22 nd international conference on Machine learning (pp. 1012–1019). New York: ACM.

Zhai, C. 2008. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies* 1 (1):1–141.

Zhang, K., and F. Tsung. 2017. A Multi-task learning framework integrating non-contemporaneous autoregressive models. *Working paper*.

Zhang, Y., and D. Y. Yeung. 2014. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8 (3):12.

Zhang, Y., D. Y. Yeung, and Q. Xu. 2010. Probabilistic multi-task feature selection. In Advances in neural information processing systems (pp. 2559–2567). Vancouver: Neural Information Processing Systems.

Zhao, W. X., J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li. 2011, April. Comparing twitter and traditional media using topic models. In European Conference on Information Retrieval (pp. 338–349). Heidelberg: Springer Berlin Heidelberg.

Zou, N., Y. Zhu, J. Zhu, M. Baydogan, W. Wang, and J. Li. 2015. A Transfer Learning Approach for Predictive Modeling of Degenerate Biological Systems. *Technometrics* 57 (3):362–373.

QUALITY ENGINEERING
2018, VOL. 30, NO. 1, 129–132
https://doi.org/10.1080/08982112.2018.1388053

Taylor & Francis
Taylor & Francis Group

Discussion on "Statistical transfer learning: A review and some extensions to statistical process control"

Xuemin Zi[a] and Changliang Zou[b]

[a]School of Science, Tianjin University of Technology and Education, China; [b]Institute of Statistic and LPMC, Nankai University, China

We would like to congratulate the authors on an interesting and important review of industrial statistics applications of transfer learning techniques. This is a timely discussion because there is an urgent need for dealing with datasets from multiple domains in various industries (not only in manufacturing, but also in service) and handling related tasks in target domains by transferring previously acquired knowledge from source domains. In this discussion, we will focus on some recent developments in the field of statistics that may essentially be categorized as transfer learning and thus similar methodologies can be applied in related applications.

Regularization-based model estimation across multiple classes

As authors surveyed in the third section indicate, an effective and popular framework of transfer learning is to apply regularization-based methods, by assuming models from different sources share certain characteristics. Some commonly used penalties were summarized. Recently, the combination of fused lasso (Tibshirani et al. 2005) and standard lasso penalties is shown to be quite efficient for model estimation across multiple classes (Danaher, Wang, and Witten 2014). For example, consider the problem of estimating multiple related Gaussian graphical models from a high-dimensional dataset with observations belonging to distinct classes. Graphical models are especially of interest in the analysis of modern social network data and can provide a useful tool for visualizing the relationships between individuals and for generating social hypotheses. The standard formulation for estimating a Gaussian graphical model assumes that each observation is drawn from the same distribution (Friedman, Hastie, and Tibshirani 2007). However, in many datasets the observations may correspond to several distinct classes (sources or domains), so the assumption that all observations are drawn from the same distribution is inappropriate. Estimating separate graphical models for the distinct classes does not exploit the similarity between the true graphical models. In addition, estimating a single graphical model with the pooled samples ignores the fact that we do not expect the true graphical models to be identical, and that the differences between the graphical models may be of interest. Guo et al. (2011) proposed to take a penalized log-likelihood approach and the penalty term is essentially in the form of $P(\mathbf{B}) = \lambda \sum_i (\sum_k B_{ki}^2)^{1/2}$, where we slightly abuse notation here by using the same symbols as the authors' in the third section, but this should not cause any confusion. Clearly, this is basically an idea of statistical transfer learning and the penalty function is similar to the penalties mentioned by the authors.

Danaher et al. (2014) pointed that the approach of using $P(\mathbf{B})$ has some disadvantages. One is that it uses just one tuning parameter and cannot control separately the sparsity level and the extent of network similarity. The other is that in cases where we expect edge values as well as the network structure to be similar between classes, the proposal of Guo et al. (2011) may not be well suited because it encourages shared patterns of sparsity but does not encourage similarity in the signs and values of the non-zero edges. Danaher et al. (2014) employ generalized fused lasso

$$\tilde{P}(\mathbf{B}) = \lambda_1 \sum_i \sum_k |B_{ki}| + \lambda_2 \sum_i \sum_{k<k'} |B_{ki} - B_{k'i}|,$$

CONTACT Changliang Zou nk.chlzou@gmail.com Institute of Statistic and LPMC, Nankai University, Nankai Qu 300071, China.

where λ_1 and λ_2 are non-negative tuning parameters. Like the lasso, the use of $\tilde{P}(\cdot)$ would result in sparse estimates $\mathbf{B}_1, \ldots, \mathbf{B}_k$ when the tuning parameter λ_1 is large. In addition, many elements of $\mathbf{B}_1, \ldots, \mathbf{B}_k$ will be identical across classes when the tuning parameter λ_2 is large (Tibshirani et al. 2005). Thus, $\tilde{P}(\mathbf{B})$ borrows information aggressively across classes, encouraging not only a similar network structure but also similar edge values.

Similar treatments can also be applicable when the underlying assumption that graphical models are not changing over time is violated. There is growing evidence that network structures are often non-stationary. As a result there is a clear need to quantify dynamic changes in network structure over time. Specially, there is a need to estimate a network at each observation in order to accurately quantify temporal diversity. Clearly, estimating time-varying networks or graphical models could be again viewed as a statistical transfer learning problem because we need to borrow information of data structure across time. To date, the most commonly used approach to achieve this goal involves the use of sliding windows or kernel-based methods (Esposito et al. 2003). However, as pointed by some authors, such as Monti et al. (2014), while sliding windows are a valuable tool for investigating high-level dynamics of networks there are two main issues associated with its use. First, the choice of the window length can be a difficult parameter to tune. Second, the use of sliding windows needs to be accompanied by an additional mechanism to determine if variations in edge structure are significant. In light of this, Monti et al. (2014) used a variant of $\tilde{P}(\mathbf{B})$

$$\check{P}(\mathbf{B}) = \lambda_1 \sum_i \sum_k |B_{ki}| + \lambda_2 \sum_i \sum_{k \geq 2} |B_{ki} - B_{k-1,i}|,$$

which encourages the estimation procedure to produce estimates with the two properties; sparsity and temporal homogeneity. With the help of $\check{P}(\mathbf{B})$, we are able to obtain individual estimates of graphical models at each time point as opposed to a model for the entire time series, allowing one to fully characterize the dynamic evolution of networks over time.

Inference with transfer-learning techniques

Aforementioned studies focus mainly on model estimation by connecting and transferring knowledge among multiple statistical models and data sources. We would like to briefly discuss some potential use of transfer learning in statistical hypothesis testing. Here, we take large-scale simultaneous hypothesis testing problems as example, in which thousands or even tens of thousands of cases are considered together. This problem has become a familiar feature in scientific fields such as biology, medicine, genetics, neuroscience, economics, and finance and has been well studied, and many solutions have been proposed (Storey and Tibshirani 2003). The most commonly used approaches start with a list of N p-values p_i, one for each hypothesis $\mathbb{H}_i$, and reject all hypotheses with a p-value below a (possibly random, i.e., data-dependent) threshold $q*$. The goal is to control a measure of Type I error at level α. Traditionally, this measure has been the Family-Wise Error Rate (FWER), but for many applications this is too stringent, and over the last 20 years the False Discovery Rate (FDR) has become a popular choice, as it is more permissive and adaptive (Benjamini and Hochberg 1995).

In many real-world applications, beyond p-values, side information, represented by covariates $\mathbf{X}_i$, is often available for each hypothesis. This side-information may be related to the different power of the tests, or to different prior probabilities of the null hypothesis being true. For example, previous studies may suggest that some null hypotheses are more or less likely to be false; similarly, in spatially structured problems, non-null hypotheses are more likely to be clustered than true null hypotheses. It is thus anticipated that exploiting structural prior information will improve the performance of conventional multiple testing procedures. Such covariates are often apparent to domain scientists. While side information is likely to be irrelevant in the context of single hypothesis testing, without taking into account this information in large-scale simultaneous hypothesis testing problems would tend to lack the detection ability of statistical significance. The covariate-adjusted or side-information-exploited FDR estimation and control is thus essentially a statistical transfer learning problem. It has recently become an active research topic and several attempts have been made in the literature to incorporate side information. For instance, methods that up-weight or down-weight hypotheses appeared in Genovese, Roeder, and Wasserman (2006) and Hu, Zhao, and Zhou (2010), among many others, say using

$$Q_i = p_i/\omega_i, \quad i = 1, \ldots, N,$$

instead of p_i's themselves with the popular BH procedure (Benjamini and Hochberg 1995), where ω_i's are some pre-chosen weights (related to $\mathbf{X}_i$) satisfying $\sum \omega_i = 1$. If the weights are chosen a priori, that is, without looking at the p-values, then the weighted BH procedure also controls the FDR. In particular, most of the literature to date only considers the case of deterministic weights. These results are very valuable, since weighted multiple testing procedures have been shown to be robust to weight misspecification (Genovese, Roeder, and Wasserman 2006): choosing good weights can lead to huge increases in power, yet bad weights will only slightly decrease power compared to the unweighted procedure. In contrast, some works allow the weights to depend also on the p-values in a data-driven way, while still controlling the FDR (Scott et al. 2013). Another different approach, based on a two-stage approach mainly arising from the microarray literature, extracts the prior information to remove a subset of features which seems to generate uninformative signals in the filtering stage, followed by applying some multiple testing procedure to the remaining features which have passed the filter in the selection stage; see, for example, Bourgon, Gentleman, and Huber (2010), Sarkar, Chen, and Guo (2013), and the references therein. The success of these solutions is mainly due to the assistance of transfer learning; information transferred from source domains could improve the performance of statistical inference in the target domain. In the literature of statistical science, recent interests concentrate upon optimal selection of tuning parameters, such like the filtering thresholds in the two-stage procedure, for example, see Du and Zhang (2014).

In summary, given the fact that advances in data acquisition techniques have led to the increasing necessity of handling datasets from multiple domains/classes, statistical modeling and inference that are able to make efficient use of previously acquired knowledge from source domains or across classes, have become critical in a variety of industrial applications. The review of statistical models and methodologies in this direction is very timely, and more research efforts are needed for establishing certain statistical properties in a systematic framework to ensure the right use of transferred knowledge.

About the authors

Xuemin Zi is Professor of the School of Science at Tianjin University of Technology and Education. Her research interests include statistical process control and design of experiments.

Changliang Zou is Professor of the Institute of Statistics at the Nankai University. He has authored over 90 refereed journal publication in Statistics and Industrial Engineering, such as *Journal of the American Statistical Association, Annals of Statistics, Biometrika, Technometrics, Journal of Quality Technology, Naval Reserch Logisitics*, and *IIE Transactions*. He currently serves as an associate editor of Technometrics and Statistica Sinica, and is on the editorial board of the Journal of Quality Technology.

Acknowledgments

Finally, we would like to thank the authors, and the editors, for the opportunity to comment.

Funding

We would like express our appreciation of the support of NNSF of China, Grant 11771332.

References

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* 57:289–300.

Bourgon, R., R. Gentleman, and W. Huber. 2010. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences* 107:9546–51.

Danaher, P., P. Wang, and D. M. Witten. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B* 76:373–97.

Du, L., and C. M. Zhang. 2014. Single-index modulated multiple testing. *Annals of Statistics* 42:1262–1311.

Esposito, F., E. Seifritz, E. Formisano, R. Morrone, T. Scarabino, G. Tedeschi, and F. Di Salle. 2003. Real-time independent component analysis of fMRI time-series. *Neuroimage* 20:2209–24.

Friedman, J., T. Hastie, and R. Tibshirani. 2007. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–41.

Genovese, C. R., K. Roeder, and L. Wasserman. 2006. False discovery control with p-value weighting. *Biometrika* 93:509–24.

Guo, J., E. Levina, G. Michailidis, and J. Zhu. 2011. Joint estimation of multiple graphical models. *Biometrika* 98:1–15.

Hu, J. X., H. Zhao, and H. Zhou. 2010. False discovery rate control with groups. *Journal of the American Statistical Association* 105:1215–27.

Monti, R. P., P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, and G. Montana. 2014. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage* 103:427–43.

Sarkar, S. K., J. Chen, and W. Guo, 2013. Multiple testing in a two-stage adaptive design with combination tests controlling FDR. *Journal of the American Statistical Association* 108:1385–1401.

Scott, J. G., R. C. Kelly, M. A. Smith, P. Zhou, and P. E. Kass. 2015. False discovery-rate regression: An application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* 110:459–71.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100:9440–5.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B* 67: 91–108.

QUALITY ENGINEERING
2018, VOL. 30, NO. 1, 133–134
https://doi.org/10.1080/08982112.2018.1386956

Taylor & Francis
Taylor & Francis Group

Discussion of "Statistical transfer learning: A review and some extensions to statistical process control"

Panagiotis Tsiamyrtzis

Department of Statistics, Athens University of Economics and Business, Athens, Greece

I would like to start by congratulating the authors for the nice presentation on the very interesting topic of Statistical Transfer Learning (STL) and especially for bringing to the fore its relation to the area of Statistical Process Control (SPC). In what follows some further aspects of statistical transfer learning will be presented while others will be discussed further, aiming to provide a more spherical point of view of this area.

Aspects of statistical transfer learning

Transfer learning attempts to carry over information among tasks and improve the learning procedure. This sharing of information across tasks (depending on the problem under study) can be done either in parallel or sequentially (see Figure 1).

> **Parallel STL**: When various tasks need to be learned simultaneously. These tasks might be different spatial/temporal cases of the same task or tasks from different domains. So, knowledge transfer is multi-directional, like the landslide monitoring example.
>
> **Sequential STL**: When we have one or more tasks that we have learned and we are interested in transferring this available knowledge in the learning of a new task. So, we have unidirectional transfer, like the 3D printing quality control example.

The former (also known as multi-task learning in machine learning) has the advantage that we have a bigger set to work with but the learning is done simultaneously. On the other hand, the latter allows knowledge attained from past (source) data to be carried over to new (target) data analysis.

Negative transfer learning

In learning a new (target) task one can try to either do it:

(i) from scratch, that is, use just the available data and ignore any relative knowledge;
(ii) use transfer learning to carry over information from source task(s) and attempt to have improved performance in learning the target task.

Is transfer learning always preferable? The answer is no. The scenario where transfer learning can potentially make things worst is known as negative transfer learning (Torrey and Shavlik 2009), where attempting to transfer knowledge from source to target task not only does not improve performance but it may actually decrease it. To avoid negative transfer the user needs to be careful, that the tasks are similar enough and that the transfer method is well leveraged. This is quite demanding for an autonomous system and methods that will be able to protect against negative transfer (allowing only "safe" transfer) will most likely reduce the benefit of transfer learning, compared to a method that does "aggressive" transfer learning, which will have excellent performance in similar tasks but will allow negative transfer in dissimilar scenarios.

Statistical methods in transfer learning

Transfer learning is not really a new concept in the area of statistics. The Bayesian approach, for example, can be seen as a transfer learning mechanism. The idea of a prior distribution along with the hierarchical modeling naturally fits this purpose. As a representative example

CONTACT Panagiotis Tsiamyrtzis ✉ pt@aueb.gr Department of Statistics, Athens University of Economics and Business, 76 Patission Street, Athens 10434, Greece.

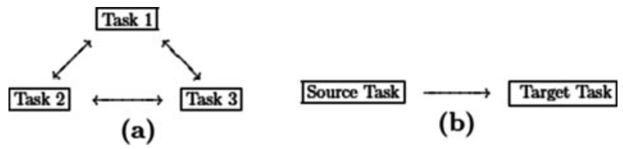

Figure 1. Parallel (a) versus sequential (b) statistical transfer learning.

one can refer to the power priors (Ibrahim and Chen 2000) which play the role of (sequential) statistical transfer learning mechanism: if D_0 are the source task data we form the prior:

$$\pi(\theta|D_0, \alpha_0) \propto [L(\theta|D_0)]_0^{\alpha} \pi_0(\theta)$$

and then upon observing the target data D_1 we obtain the posterior:

$$\pi(\theta|D_0, D_{1,}\alpha_0) \propto [L(\theta|D_1)][L(\theta|D_0)]_0^{\alpha} \pi_0(\theta)$$

where the value of $\alpha_0 \in [0, 1]$ will determine the effect (reflecting the similarity between the source and the target task) of the source data (D_0) in determining the posterior distribution of the parameter θ, once we use the target data (D_1).

Can the Bayesian approach allow negative transfer? If the source and target tasks have been generated from very different values of parameters, then the posterior in the target task can be negatively affected from the prior (that was set from the source task), mainly when we have low volumes of data, as with big data the effect of the prior diminishes. In any case, prior sensitivity analysis can be helpful to examine whether the prior used, affects (negatively) the posterior or not.

Transfer learning and SPC

The article demonstrates ways where SPC methods can provide tools in statistical transfer learning. An interesting question though comes if we inverse the above and ask whether SPC methods can benefit from the use of statistical transfer learning. For example, in frequentist-based control charting (Shewhart charts, CUSUM, EWMA, etc.) a standard practice is to employ a Phase I/II split, where in Phase I we perform learning (calibration) while in Phase II we perform testing. So, learning stops at the end of Phase I. Transfer learning philosophy would suggest to carry over learning in Phase II, incorporating the information from new data as they become available. Such a proposal is feasible via a Bayesian SPC scheme (see, e.g., Tsiamyrtzis and Hawkins 2005, 2010) which can be set as a sequentially updated mechanism allowing the parameter transfer learning, as data become available progressively, providing solutions even when we have small amounts of data and braking free form the usual Phase I/II constraint.

About the author

Panagiotis Tsiamyrtzis received the BSc degree in mathematics from the Aristotle University of thessaloniki, Greece. He obtained the MSc and PhD degrees in statistics from the school of statistics at University of Minnesota, USA, working with Douglas M. Hawkins in the field of SPC. In 2004 he joined the department of statistics at Athens University of Economics and Business, where he is currently Associate Professor. He is also a Research Associate Professor at the Department of Computer Science, University of Houston, TX, USA. His research interests include Bayesian statistical process control and statistical aspects of Computational Physiology, where he published more than 50 papers.

References

Ibrahim, J. G., and M. H. Chen. 2000. Power prior distributions for regression models. *Statistical Science* 15 (1): 46–60.

Torrey, L., and J. Shavlik. 2009. Transfer learning. In *Handbook of research on machine learning applications*, eds. E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, 242–64. IGI Global. http://pages.cs.wisc.edu/~shavlik/abstracts/torrey.handbook09.abstract.html

Tsiamyrtzis, P., and D. M. Hawkins. 2005. A Bayesian scheme to detect changes in the mean of a short-run process. *Technometrics* 47 (4):446–56. doi:10.1198/004017005000000346.

Tsiamyrtzis, P., and D. M. Hawkins. 2010. Bayesian startup phase mean monitoring of an autocorrelated process that is subject to random sized jumps. *Technometrics* 52 (4):438–52. doi:10.1198/TECH.2010.08053.

QUALITY ENGINEERING
2018, VOL. 30, NO. 1, 135–136
https://doi.org/10.1080/08982112.2018.1390957

Rejoinder

Fugee Tsung, Ke Zhang, Longwei Cheng, and Zhenli Song

Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

We would like to thank the discussants for their insightful comments on a number of important areas in Statistical Transfer Learning (STL). Here we would like to take this opportunity to summarize some of their key ideas for further discussion and analysis.

Initially, Tsiamyrtzis pointed out that the STL techniques can be categorized as parallel STL and sequential STL by the way they share information across tasks. We agree with the summary that the parallel STL is also known as multi-task learning and normally presents as multi-directional while in sequential STL the learning methods have specific unidirectional transfer. We have also discussed the difference of multi-task learning and typical transfer learning in our original article.

Both discussion articles mention the statistical methods adopted in transfer learning. For Bayesian-based methods in STL, Tsiamyrtzis indicates that the idea of a prior distribution along with the hierarchical Bayesian modeling naturally fit the purpose of transfer learning to share information among parameters. We present the same idea in our article and we expect an integration of Bayesian methods and STL applications for SPC. For regularization-based methods, we acknowledge Zi and Zou's (2018) introduction on some recent developments of regularization-based methods in high-dimensional statistics that may be further adopted in STL applications, like fused Lasso (Tibshirani et al. 2005) for network structured data.

Tsiamyrtzis raises an important issue of transfer learning: negative transfer (Torrey and Shavlik 2009). In fact, inappropriate transfer across tasks may have negative effects on the learning performance, which have been widely discussed in machine learning studies (Pan and Yang 2010). The negative effects normally originate the unconformity between the dataset and the assumed model/pattern. For example, as Tsiamyrtzis specifies, in Bayesian-based transfer learning, the transfer is allowed by a prior distribution that assumes similarity between source and target domains. If the source and target tasks have been generated from very different parameters, the posterior in the target task may be negatively affected by the prior. Similarly, in regularization-based transfer learning, if the tasks do not present a common sparsity pattern as assumed, the transfer learning performance might be compromised or even worsen.

Discussants also mention the further applications of STL in statistical inference and SPC. Zi and Zou (2018) suggest the potential use of STL in hypothesis testing. We agree with their comments and are glad to see the potential improvement that STL could make for multiple testing problems. On the other hand, Tsiamyrtzis notes that SPC methods may benefit from the use of STL by incorporating information in Phase I and Phase II, like in a Bayesian manner. We appreciate this important suggestion on future research. This idea seems in a similar sense to online learning (Hoffman, Bach, and Blei 2010).

In summary, STL provides an efficient framework to extend and improve SPC methods by combining previously acquired knowledge across classes/sources, which meets the increasing necessity of fusing different datasets in modern industrial and service applications. We thank the discussants for their valuable comments, and hope the discussion article will identify some directions that warrant future research for SPC and quality improvement.

CONTACT Fugee Tsung ✉ season@ust.hk Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

References

Hoffman, M., F. R. Bach, and D. M. Blei. 2010. Online learning for latent Dirichlet allocation. In *Advances in neural information processing systems*, 856–64. Vancouver: Neural Information Processing Systems.

Pan, S. J., and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (10):1345–59. doi:10.1109/TKDE.2009.191.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* 67:91–C108. doi:10.1111/j.1467-9868.2005.00490.x.

Torrey, L., and J. Shavlik. 2009. Transfer learning. In *Handbook of research on machine learning applications*, eds. E. Soria, J. Martin, R. Magdalena, M. Martinez, and A. Serrano, 242–64. IGI Global.

Zi, X., and C. Zou. 2018. Discussion on "Statistical transfer learning: A review and some extensions to statistical process control." *Quality Engineering* 30 (1): 129–132.