Improved linear regression prediction by transfer learning

David Obst, Badih Ghattas, Sandra Claudel, Jairo Cugliari, Yannig Goude, Georges Oppenheim

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167947322000792
Manuscript_9b9a055ff09bb9afab2a7dfce7367ade

Improved Linear Regression Prediction by Transfer Learning

David Obst[a,b,*], Badih Ghattas[b], Sandra Claudel[a], Jairo Cugliari[c], Yannig Goude[a], Georges Oppenheim[d]

[a] *EDF R&D, Palaiseau, France*
[b] *Institut de Mathématiques de Marseille, Aix-Marseille Université, Marseille, France*
[c] *Laboratoire ERIC, Université de Lyon 2, Bron, France*
[d] *Laboratoire d'Analyse et de Mathématiques Appliquées Université Paris-Est, Champs-sur-Marne, France*

Abstract

Transfer learning, also referred as knowledge transfer, aims at reusing knowledge from a source dataset to a similar target one. While several studies address the problem of what to transfer, the very important question of *when* to answer remains mostly unanswered, especially from a theoretical point-of-view for regression problems. A new theoretical framework for the problem of parameter transfer for the linear model is proposed. It is shown that the quality of transfer for a new input vector depends on its representation in an eigenbasis involving the parameters of the problem. Furthermore, a statistical test is constructed to predict whether a fine-tuned model has a lower prediction quadratic risk than the base target model for an unobserved sample. Efficiency of the test is illustrated on synthetic data as well as real electricity consumption data.

Keywords: Linear regression, Transfer learning, Statistical test, Fine-tuning, Transfer theory

*Corresponding author
Email address: david.obst@edf.fr (David Obst)

1. Introduction

The traditional statistics and machine learning approach is to learn a model on training data and then perform inference on some new unseen data. This paradigm supposes two main underlying hypotheses, which are not necessarily true in practice. The first one is that enough samples are available to learn a good model which will be used to perform the predictions. However, in many real-life situations, data will be scarce (either because it is difficult or expensive to label or only newly available). Take for instance the problem of predicting orders for a newly contracted customer. Without a sufficiently large order history, a learned model may yield poor forecasts. Another hypothesis is that the data on which inference will be performed stems from the same underlying distribution as the one used for training. In practice, data often evolves with time and space, and thus rigorously speaking it will rarely be true. For example, in natural language processing (NLP), words and their frequency of usage change over time (as well as new words being introduced). Therefore, a state-of-the-art text classification model learned on a corpus a decade ago may not be perfectly suited anymore and have deteriorating results.

Nonetheless, in the previous example the model learned on the 10 year-old corpus will probably still hold some truth and would simply require to be "updated" on a recent corpus. As for the customer order prediction problem, information from long-term customers could be leveraged as they should hold at least some similarities in behavior patterns with the more recent ones. This is the setting of *transfer learning* (TL), which gained a lot of attention in the statistics and machine learning communities in the past decades. Just like humans who generally use their experience and knowledge to adapt to a new task instead of learning from scratch (e.g. learning how to play the guitar after knowing how to play another instrument), the concept of transfer learning is to use a *source* task to improve the results on a *target* task which is of main interest (Weiss, Khoshgoftaar & Wang, 2016). Pan & Yang (2009) and Yang, Zhang, Dai & Pan (2020) are two grounding references in transfer learning where the authors enumerate three key questions. **What kind of knowledge can be transferred:** this aspect mostly focuses on finding the information that common between tasks and what can be brought from the source to the target. **How transfer can be achieved:** it deals with the specific method by which the transfer is performed. The authors cite 4 types of approaches: instance-based algorithms, where the source samples are added to the target dataset with a certain weight (Cai, Gu, Ma & Jin, 2019); feature-based

algorithms, where features are crafted with the help of the source for the target (Yin, Yu, Sohn, Liu & Chandraker, 2019); model or parameter-based transfer, where a source model or are part of it is transferred to the target one; and relation-based transfer, where the associations within the source data are propagated to the target samples (Mihalkova, Huynh & Mooney, 2007). **When to use transfer learning**: if source and target are too dissimilar, the transfer procedure can be detrimental. Thus, this aspect deals with finding the situations when transfer should be performed, but has received significantly less attention than the two first ones. Moreover, the question is crucial from a practical aspect in order to make the decision of using a transfer scheme or not. Of course, the transfer can not be beneficial in every situation (for instance is the source and target data sets are too dissimilar) and thus a tool to help the practitioner to make the decision is useful.

This is why in our work we focus on when to perform parameter transfer between two linear regression tasks. We have at our disposal a target dataset $\mathcal{D}_T = \{X_T, \boldsymbol{Y}_T\}$ with

$$\boldsymbol{Y}_T = X_T \boldsymbol{\beta}_T + \boldsymbol{\varepsilon}_T,$$

where $\boldsymbol{Y}_T = (y_1, y_2, \ldots, y_{N_T})$ is the response vector of size N_T, $X_T \in \mathbb{R}^{N_T \times D}$ is the design matrix with D predictors (including the eventual intercept) and $\boldsymbol{\varepsilon}_T \sim \mathcal{N}(0, \sigma_T^2 I_{N_T})$ is a vector of identically distributed and independent (i.i.d.) Gaussian noise (with I_N denoting the identity matrix of size N). Supposing that X_T has full column rank, the commonly used estimator is the ordinary least squares (OLS) one defined by $\hat{\boldsymbol{\beta}}_T = (X_T^\top X_T)^{-1} X_T^\top \boldsymbol{Y}_T$. Predictions for a new sample $\boldsymbol{x}$ is then achieved by using $\hat{y}_T = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_T$. However, if the target sample size N_T is too small, the quadratic risk of the prediction using $\hat{\boldsymbol{\beta}}_T$ defined by $\mathcal{R}(\hat{y}_T) = \mathbb{E}[(y - \hat{y}_T)^2]$ will be high as it decreases in $\mathcal{O}(1/N_T)$. Let us then suppose we have a *source* dataset $\mathcal{D} = \{X_S, \boldsymbol{Y}_S\}$ with:

$$\boldsymbol{Y}_S = X_S \boldsymbol{\beta}_S + \boldsymbol{\varepsilon}_S,$$

where $\boldsymbol{Y}_S \in \mathbb{R}^{N_S}$, $X_S \in \mathbb{R}^{N_S \times D}$ and $\boldsymbol{\varepsilon}_S \sim \mathcal{N}(0, \sigma_S^2 I_{N_S})$ for which $N_S \gg N_T$. Intuitively, if the source and target tasks are "close" (in a sense to refine), using this source data available will compensate the lack of target data. Hence, our goal is to construct an estimate of the coefficients $\hat{\boldsymbol{\beta}}_{TL}$ such that the prediction error of $\hat{y}_{TL} = x^\top \hat{\boldsymbol{\beta}}_{TL}$ defined by $\mathcal{R}(\hat{y}_{TL}) = \mathbb{E}[(y - \hat{y}_{TL})^2]$ will be lower than the target one $\mathcal{R}(\hat{y}_T)$. The procedure is illustrated in Figure 1.

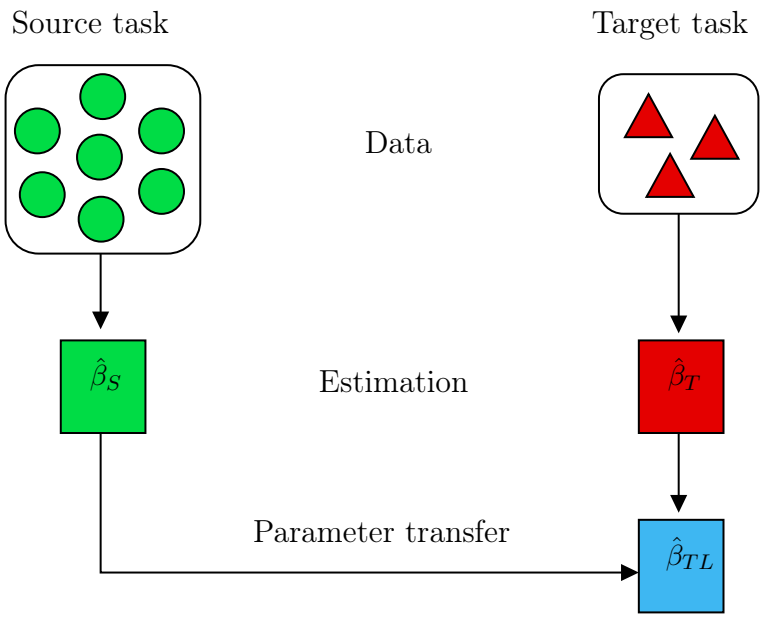


Figure 1: Parameter transfer procedure for the linear model

The framework of parameter transfer for the linear regression has been extensively studied in the literature. On the theoretical level, Maurer (2006) establishes bounds on the average prediction error over m tasks for a linear predictor in a Hilbert space, but do not investigate in which situation learning on multiple sources could be beneficial for a specific target. In Lounici, Pontil, Tsybakov & Van De Geer (2009); Lounici, Pontil, Van De Geer, Tsybakov et al. (2011) the authors consider the setting of sparse multi-task learning in high dimension with a common sparsity pattern within the regression vectors. They obtain oracle inequalities on the prediction error, albeit for the same data on which the parameters were learned. A more practical study is proposed in Bouveyron & Jacques (2010) to transfer the parameters of a linear model. After obtaining an estimate $\hat{\boldsymbol{\beta}}_S$ of the coefficients on the source data, the enriched estimation is obtained by a linear transformation $\hat{\boldsymbol{\beta}}_{TL} = \hat{\Lambda}\hat{\boldsymbol{\beta}}_S$ where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{D-1})$ ($\hat{\lambda}_0$ corresponding to the intercept) is calculated on the target set only. Since learning all the coefficients of $\hat{\Lambda}$ would erase all the benefits of transfer, the authors rather constraint $\hat{\Lambda}$ to have only 1 or 2 coefficients to learn. For example it could be $\hat{\lambda}_1 = \cdots = \hat{\lambda}_{D-1}$. Their results showed significant improvement when the number of target samples is small on two real datasets. Nonetheless, the most successful approach introduced in their paper corresponds to the situation where $\hat{\Lambda}$ has only two free coefficients, which gives very low adaptation freedom. Chen, Owen, Shi et al. (2015) propose another transfer method

(referred to as "enrichment") where the transferred estimator is obtained by a matrix combination of the source and target ones $\hat{\boldsymbol{\beta}}_{TL} = W\hat{\boldsymbol{\beta}}_T + (I_D - W)\hat{\boldsymbol{\beta}}_S$ where W is a matrix that can take different forms. In their paper, the two forms of interests are $W = \omega\, I_D$ (convex combination of source and target estimations) and $W = W_\lambda$ obtained by adding a ridge penalty accounting for the difference between source and target tasks. Their main result consists in proving that under certain hypotheses, a certain choice of ω yields a $\hat{\boldsymbol{\beta}}_{TL}$ that has lower prediction error than $\hat{\boldsymbol{\beta}}_T$, and they propose a plug-in estimator for it. They also propose a plug-in estimator for an optimal W_λ. However, their results remain mainly theoretical. Very recently, new results have been obtained for the problem of transfer for linear regression in Dar & Baraniuk (2020, 2021). Both focus on the problem of leveraging source coefficients for the estimation of the target in the theoretical framework when the features are i.i.d. according to $\mathcal{N}(0, I_D)$. In the first paper they transfer coefficients directly, whereas in the second one they use a "fine-tuning" scheme that consists in adding a Ridge penalty similarly to Chen et al. (2015). They prove that transfer is beneficial in the under or overparametrized cases. Furthermore, they prove that transfer learning can overperform the Bayesian framework even when using the true prior that served to generate the coefficients in their experiments.

Thus most of the aforementioned papers remain theoretical or are too restrictive for real estimation or prediction problems. Moreover, they are lacking one important transfer approach, namely fine-tuning by gradient descent. It consists in reusing a part of the learned parameters on the source (for instance neural network layers or linear model coefficients) and adjusting them on the target with a few gradient iterations (Shin, Roth, Gao, Lu, Xu, Nogues, Yao, Mollura & Summers, 2016). The main question with transfer is hence when to perform it, i.e. when it will be beneficial. Ben-David, Blitzer, Crammer, Kulesza, Pereira & Vaughan (2010) indirectly address the issue of negative transfer for the problem of binary classification. Considering the transfer problem as a special case of a multi-task objective, not only do they obtain an upper bound on the transfer prediction error, but they also prove the existence of phases depending on the number of samples N_S and N_T available for source and target respectively. Fawaz, Forestier, Weber, Idoumghar & Muller (2018) approach the problem empirically: after defining a distance between datasets based on the dynamic time warping (DTW), they show that in general negative transfer will happen when the defined distance between source and target is large.

We propose to address this issue practically albeit with theoretical considerations in the case of linear regression. Transfer is said beneficial for a new sample $(\boldsymbol{x}, y)$ when $\mathbb{E}[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_{TL})^2] < \mathbb{E}[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_T)^2]$. We want a practical decision rule that tells us that fine-tuning will be beneficial for this instance $\boldsymbol{x}$. In our work we derive a new quantity referred as *gain* quantifying the benefits of transfer, without any assumption except the one of the linear model. While the hypothesis of a linear model may seem restrictive, it includes many variants such as generalized linear models (GAM) (Wood, 2017) that make it possible to capture highly nonlinear effects through the use of spline bases. We will also show that it is possible to derive a hypothesis test to *predict in practice* whether the transfer is positive or not. The contributions of the paper are the following:

1. We formalize the problem of negative transfer for the fine-tuning of a linear regression model. Our framework is valid for a broad class of transfer procedures for the linear model found in the literature.
2. We show that the transfer gain for a new feature vector $\boldsymbol{x}$ depends on its representation on an eigenbasis depending on the parameters of the linear model.
3. We suggest a statistical test to choose for a new observation $\boldsymbol{x}$ between the target model or a fine-tuned one.
4. The statistical test has been applied on synthetic data as well as two sets of real electricity consumption data sets, proving the benefits it brings.

The rest of the paper is organized as follows. Section 2 introduces the theoretical framework and methodology leading up to the test of transfer. In Section 3 we illustrate the benefits brought on synthetic data as well as real data, while Section 4 concludes our work and suggests further research possibilities. Finally, in the Appendix proofs of theoretical results are given.

2. Framework & Methodology

We suppose that the matrices X_ν ($\nu \in \{S, T\}$) are full-rank such that $\Sigma_\nu = X_\nu^\top X_\nu$ are both invertible. The transfer methodology on which we focus is fine-tuning. The idea is to start from the source estimator $\hat{\boldsymbol{\beta}}_S = \Sigma_S^{-1} X_S^\top \boldsymbol{Y}_S$ and to perform batch gradient descent (GD) of stepsize α on the least-squares objective $J_T(\boldsymbol{\beta}) = \frac{1}{2} \|Y_T - X_T \boldsymbol{\beta}\|^2$. Intuitively the idea is hence to incorporate

information of the few target samples available with the source estimate as basis. For the linear model the following result can be proven (see Appendix A.1).

Proposition 1. *At iteration $k \in \mathbb{N}$ the fine-tuned estimator of $\boldsymbol{\beta}_T$ is:*

$$\hat{\boldsymbol{\beta}}_k = A^k \hat{\boldsymbol{\beta}}_S + (I - A^k)\hat{\boldsymbol{\beta}}_T, \tag{1}$$

where $A = I_D - \alpha\Sigma_T$ and I_D is the identity matrix of size D.

Therefore the fine-tuned estimator is a matrix combination of source and target estimators. In fact this observation can be taken further in the right vector basis to give more insight on this expression. Since Σ_T is symmetric and real-valued, let P be an orthogonal diagonalization basis matrix such that $\Sigma_T = P\Lambda P^\top$ with $\Lambda = \text{diag}(\lambda_i,\, i = 1, \ldots, D)$ the diagonal matrix of eigenvalues of Σ_T. Let $\tilde{\boldsymbol{\beta}}_\nu$ denote the coordinate of $\hat{\boldsymbol{\beta}}_\nu$ in Σ_T's eigenbasis. Hence $\hat{\boldsymbol{\beta}}_\nu = P\tilde{\boldsymbol{\beta}}_\nu$. As detailed in Appendix A.2, reusing equation (1) yields:

$$\tilde{\boldsymbol{\beta}}_k = (I_D - \alpha\Lambda)^k \tilde{\boldsymbol{\beta}}_S + (I_D - (I_D - \alpha\Lambda)^k)\tilde{\boldsymbol{\beta}}_T, \tag{2}$$

which means that for every coordinate i in this basis we have:

$$\tilde{\beta}_k^{(i)} = (1 - \alpha\lambda_i)^k \tilde{\beta}_S^{(i)} + \big(1 - (1 - \alpha\lambda_i)^k\big)\tilde{\beta}_T^{(i)}. \tag{3}$$

Hence when α is small enough and in the right basis, each coordinate of the fine-tuned coefficient is a convex combination of the source and target coefficients, albeit with different weights depending on the eigenvalues λ_i. For small eigenvalues of Σ_T the fine-tuning procedure will give a larger weight to the source whereas it is the opposite for larger ones.

Note that these expressions relate this transfer strategy to the ones introduced in Chen et al. (2015), where they consider two types of transfer for the linear model. The first one is the pooling of source and target data, leading to estimators of the form $\hat{\boldsymbol{\beta}}_\lambda = W_\lambda \hat{\boldsymbol{\beta}}_S + (I_D - W_\lambda)\hat{\boldsymbol{\beta}}_T$ where W_λ is a matrix depending on the penalty parameter $\lambda > 0$. The second one is a simple convex combination $\hat{\boldsymbol{\beta}}(\omega) = \omega\hat{\boldsymbol{\beta}}_S + (1-\omega)\hat{\boldsymbol{\beta}}_T$ for a constant weight $\omega \in [0, 1]$. Hence transfer by fine-tuning is between those two approaches: in the right basis and for α small enough each coefficient is a convex combination of the source and target ones, albeit with different weights depending on the eigenvalue λ_i which allows for more adaptability than for a constant ω. It is interesting to note that in the end two popular transfer approaches, namely data pooling and fine-tuning yield estimators

of the same class $\hat{\boldsymbol{\beta}}(W) = W\hat{\boldsymbol{\beta}}_S + (I_D - W)\hat{\boldsymbol{\beta}}_T$ with specific forms of $W \in \mathbb{R}^{D\times D}$. In the case of data pooling the expression of W is more complex and it is generally not symmetric (see Chen et al. (2015)). To our knowledge such a strong relationship between the approaches has never been highlighted in literature before.

2.1. Transfer Gain

The quality of a model will be evaluated for a new independent sample $(\boldsymbol{x}, y)$ drawn from the underlying distribution of the target data. We want to know if for this given $\boldsymbol{x}$ the fine-tuned model $\mathcal{M}_k$ relying on the estimator $\hat{\boldsymbol{\beta}}_k$ learned on the source data but fine-tuned on the target one is better than the pure target model $\mathcal{M}_T$ using the basic estimator $\hat{\boldsymbol{\beta}}_T$. Following what was discussed in the introduction we introduce the *algebraic gain* $\Delta\mathcal{R}_k(\boldsymbol{x})$ for sample $(\boldsymbol{x}, y)$ defined by: $\Delta\mathcal{R}_k(\boldsymbol{x}) = \mathbb{E}[(y - \hat{y}_T)^2] - \mathbb{E}[(y - \hat{y}_k)^2]$ where $\hat{y}_T = \boldsymbol{x}^\top\hat{\boldsymbol{\beta}}_T$ and $\hat{y}_k = \boldsymbol{x}^\top\hat{\boldsymbol{\beta}}_k$. We have the following result in the case of fine-tuning.

Proposition 2. *For transfer by fine-tuning as presented by equation (1), at iteration k the gain is:*

$$\Delta\mathcal{R}_k(\boldsymbol{x}) = \boldsymbol{x}^\top H_k \boldsymbol{x} \quad \textit{where} \quad H_k = \sigma_T^2(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k) - \sigma_S^2 A^k\Sigma_S^{-1}A^k - A^k B A^k, \tag{4}$$

with $\Omega_k = \frac{1}{\alpha}\Sigma_T^{-1}(I_D - A^k)$, $B = (\beta_T - \beta_S)(\beta_T - \beta_S)^\top$. When it is positive, the transfer is beneficial for the sample $(\boldsymbol{x}, y)$, and negative otherwise.

Proof. See Appendix A.3. □

Therefore it can be seen that the matrix H_k plays a significant role for the transfer problem. The gain will be positive for vectors in the span of the eigenvectors of H_k associated to positive eigenvalues. The role of the noise in the data as well as the distance between the regression parameters also becomes clear with this formula and seems intuitive. When $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|$ is large, i.e. the means of y_ν's will differ significantly, transfer is likely to be detrimental. When σ_T^2 is large (the target data is noisy), the gain will increase since learning from the target data may be difficult. Note that this expression of the gain does not require any hypothesis on $\boldsymbol{x}$, which is a major difference with previous works. We also see that a uniformly positive transfer may be impossible, and that the benefits of transfer are a local property: therefore for some $\boldsymbol{x}$ it may be beneficial to use a fine-tuned

model, whereas for others not. From (4) bounds on the prediction error can easily be derived:

$$\begin{aligned}\mathbb{E}[(y-\hat{y}_k)^2] &\leq \mathbb{E}[(y-\hat{y}_T)^2] - \lambda_{\min}\left(H_k\right)\|\boldsymbol{x}\|^2, \\ \mathbb{E}[(y-\hat{y}_k)^2] &\geq \mathbb{E}[(y-\hat{y}_T)^2] - \lambda_{\max}\left(H_k\right)\|\boldsymbol{x}\|^2,\end{aligned} \tag{5}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ respectively denote the minimum and maximum eigenvalues of Σ_T. Again those bounds do not require any assumptions and hold for any $\boldsymbol{x} \in \mathbb{R}^D$ and only require H_k to be symmetric, which is the case when performing transfer by fine-tuning. As one can see the transfer is always positive when $\lambda_{\min}(H_k) > 0$. More generally, a similar expression to (4) is possible for any estimator of the form $\hat{\boldsymbol{\beta}}(W) = W\hat{\boldsymbol{\beta}}_S + (I_D - W)\hat{\boldsymbol{\beta}}_T$. However when W is not symmetric, interpretability of transfer in terms of eigenvector direction is lost and the inequalities of (5) cannot be established in the same way. Consequently if H_k was accessible, one would know which model to use exactly for a given $\boldsymbol{x}$. However the issue is that many quantities in the matrix are unknown, namely the true regression parameters β_ν and the true variances of the noise σ_ν^2. A naive approach would consist of considering the "plug-in" estimate $\hat{H}_k$ by replacing the parameters by their estimates, but experiments have shown that this is a rather poor choice in most situations. Another strategy is therefore proposed in the next section. Finally we emphasize again that $\boldsymbol{x}$ is potentially a *novel* observation on which we require no hypothesis. In the aforementioned papers the bounds hold only under specific conditions that did not allow for any $\boldsymbol{x} \in \mathbb{R}^D$, making our result broader.

2.2. Statistical Test for the Positiveness of Transfer

We simplify our problem to knowing in advance whether the transfer will be beneficial or not, i.e. if $\Delta\mathcal{R}_k(\boldsymbol{x}) > 0$. Therefore an alternative is to define the problem as hypothesis testing. Considering that $\mathcal{M}_k$ is likely to be biased, we choose the null hypothesis $H_0 : \{\Delta\mathcal{R}_k(\boldsymbol{x}) \leq 0\}$ against the alternative $H_1 : \{\Delta\mathcal{R}_k(\boldsymbol{x}) > 0\}$. This boils down to choosing between two models, the pure target one and the fine-tuned one for a given target sample. The idea of achieving the best performance in transfer learning by taking advantage of multiple models could be related to Gao, Fan, Jiang & Han (2008) where they weighted classifiers according to the local properties of target observations. The main result of the paper is given in Theorem 1.

Theorem 1. *Let $\boldsymbol{x} \in \mathbb{R}^D$ be **any** observation. Let $\hat{\sigma}_S^2$ and $\hat{\sigma}_T^2$ be the estimations of the noise variances defined by $\hat{\sigma}_\nu^2 = \left\|Y_\nu - X_\nu\hat{\boldsymbol{\beta}}_\nu\right\|^2/(N_\nu - D)$. Let ρ be such that $\rho \geq \|\boldsymbol{\beta}_T - \boldsymbol{\beta}_S\|/\sigma_T$. Then*

the following test is of approximate level a to test H_0 against H_1:

$$\mathbb{1}\Big(\underbrace{\frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2} \frac{\boldsymbol{x}^\top(\Sigma_T^{-1} - \alpha^2 \Omega_k \Sigma_T \Omega_k)x - \rho^2 \left\|A^k \boldsymbol{x}\right\|^2}{\boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}}}_{:=\psi_k(\boldsymbol{x})} > q^{1-a} \Big), \tag{6}$$

where q^{1-a} is the quantile of order $1-a$ of the $\mathcal{F}(N_T - D, N_S - D)$ Fisher-Snedecor distribution of degrees of freedom $N_T - D$ and $N_S - D$[1]*. The p-value for the observed data is:*

$$p_k(\boldsymbol{x}) = \mathbb{P}_{F \sim \mathcal{F}(N_T - D, N_S - D)} \Big(F \geq \psi_k(\boldsymbol{x}) \Big). \tag{7}$$

Proof. See Appendix A.5. □

The parameter ρ can be seen as a prior on the distance between the source and target distributions. Indeed, in the gaussian case one can easily prove that $2D_{KL}\Big(\mathcal{N}(\boldsymbol{x}^\top \beta_S, \sigma_S^2) || \mathcal{N}(\boldsymbol{x}^\top \beta_T, \sigma_T^2)\Big) \leq g\big(\frac{\sigma_S^2}{\sigma_T^2}\big) + \rho^2 \left\|\boldsymbol{x}\right\|^2$ where D_{KL} denotes the Kullback-Leiber (KL) divergence and $g(u) = u - \log(u) - 1$. The larger ρ is, the more significant the difference between source and target distributions is allowed to be and thus the less likely the transfer will be beneficial. When $\rho = 0$ (i.e. $\boldsymbol{\beta}_S = \boldsymbol{\beta}_T$) only the variances differ. Note that the Cauchy-Schwarz approximation lowers the power of the test (see appendix for more details). An issue is that $p_k(\boldsymbol{x}) \longrightarrow 0$ when $k \to +\infty$. Hence when the number of gradient iterations goes to infinity, the test will almost systematically reject the null hypothesis, despite the gain converging to 0. Elements of mathematical proof are given in the Appendix A.6, as well as numerical illustrations of the phenomenon. Therefore a choice of a reasonable k is of crucial importance. Finally the test can only be obtained when using a symmetric weight matrix W, making it not possible to generalize to Chen et al. (2015) for now for instance.

Hence with this test we can define a new model $\mathcal{M}_k^*$ which uses $\mathcal{M}_T$ when the null hypothesis is kept (typically $p_k(\boldsymbol{x}) > 0.05$) and the fine-tuned model $\mathcal{M}_k$ otherwise. This procedure of use of the test is summarized in Figure 2

[1]The Fisher-Snedecor distribution of degrees of freedom d_1 and d_2 is defined by $F = \frac{F1/d_1}{F_2/d_2}$ where the S_i are independent chi-squared distributed of degree of freedom d_i.

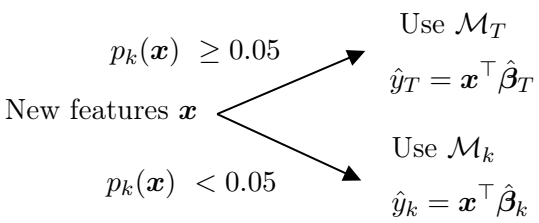


Figure 2: Usage of the test defining $\mathcal{M}_k^*$ in practice.

2.3. Choice of α, k and ρ

Three quantities must be tuned before usage of the test: the gradient step size α, the number of iterations k and the approximation parameter ρ. Equation (3) suggests that $0 < \alpha < 1/\lambda_{\max}(\Sigma_T)$ so that the coordinates remain a convex combination of the source and target ones. Additionally according to Bertsekas (2015), a step size $\alpha^* = 2/\big(\lambda_{\max}(\Sigma_T) + \lambda_{\min}(\Sigma_T)\big)$ allows to converge at maximal speed. However in our case convergence to $\hat{\boldsymbol{\beta}}_T$ is not desirable since it would erase benefits from $\hat{\boldsymbol{\beta}}_S$. Taking $\alpha = \alpha^*/5$ or $\alpha^*/10$ has proven to be a good choice in practice since it ensures the condition $0 < \alpha < 1/\lambda_{\max}(\Sigma_T)$ while remaining close to α^*. Experimentally we observed that a too low value of α could be compensated by a larger k, making the choice of the gradient step size not crucial. More recent results in a theoretical framework similar to Dar & Baraniuk (2020) have even proven that the choice of the gradient step size α has no impact on the maximum value of the gain, thus comforting our experimental observations.

Ideally, one would choose the smallest k such that $\lambda_{\min}(H_k) \geq 0$ (whose existence is not ensured, but would ensure an exclusively positive gain). However it depends on unknown parameters, and again a plug-in estimate yields poor results. We suggest two empirical approaches to determine the number of iterations k, although it remains a work in progress. The first one is to proceed by leave-one-out cross-validation (LOOCV) on the few samples of target data available. We let $X_T^{(-i)}$ and $\boldsymbol{Y}_T^{(-i)}$ denote the data where the sample i has been removed and $\hat{\boldsymbol{\beta}}_T^{(-i)}$ the corresponding estimator. Thus for each sample $i = 1, \ldots, N_T$ we compute the estimator $\hat{\boldsymbol{\beta}}_T^{(-i)}$ and then perform the fine-tuning procedure with gradient descent, obtaining an estimator $\hat{\boldsymbol{\beta}}_k^{(-i)}$. The prediction error for $Y^{(i)}$ can then be calculated, yielding in the end the LOOCV error (8):

$$\frac{1}{N_T} \sum_{i=1}^{N_T} (Y_T^{(i)} - \hat{Y}_T^{(i)})^2 - (Y_T^{(i)} - \hat{Y}_k^{(i)})^2. \tag{8}$$

The number of gradient iterations k is taken to maximize this quantity, with an eventual

compromise to avoid an excessive number of iterations (akin to an elbow rule). The following approach yields satisfactory results as well and completes the LOOCV method. Let us denote by $\mathcal{N}_T = \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}_T; \sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x})$ and $\mathcal{N}_k = \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}_k; \boldsymbol{x}^\top V_k \boldsymbol{x})$ the distributions of the predictions, where $\boldsymbol{\beta}_k = \mathbb{E}[\hat{\boldsymbol{\beta}}_k]$ and $V_k = \sigma_S^2 A^k \Sigma_S^{-1} A^k + \sigma_T^2 \alpha^2 \Omega_k \Sigma_T \Omega_k$. It can be proved that (see Appendix A.7):

$$\Delta\mathcal{R}_k(x) = -2\sigma_T^2\, x^\top \Sigma_T^{-1} x\, D_{KL}(\mathcal{N}_k || \mathcal{N}_T) - \sigma_T^2\, x^\top \Sigma_T^{-1} x \ln\Big(\frac{x^\top V_k x}{\sigma_T^2\, x^\top \Sigma_T^{-1} x}\Big) \tag{9}$$

Let $U_k(\boldsymbol{x})$ denote the second term of the right-hand side. Ideally one would choose k^* maximizing the gain accross all possible $\boldsymbol{x}$ according to their distribution $\mathbb{P}_x$, i.e. $k^* = \text{argmax}_k \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_x}\Big[\Delta\mathcal{R}_k(\boldsymbol{x})\Big]$. It can be shown that there exists a constant $C > 0$ such that:

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_x}\Big[\Delta\mathcal{R}_k(\boldsymbol{x})\Big] \geq -2\,C\,\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_x}\Big[\boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}\Big] + \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_x}\Big[U_k(\boldsymbol{x})\Big].$$

Considering the first term on the right hand side does not depend on k, maximizing $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_x}\Big[U_k(\boldsymbol{x})\Big]$ will maximize the gain as well. Since the true distributions are unknown, as usual we approximate it with an empirical average on both source and target to have more samples at disposal, thus yielding $\overline{U}_k$. Finally since the amount of target data is limited, we cannot afford to perform this procedure on a hold-out set. Therefore k is selected by maximizing $\overline{U}_k := \frac{1}{N_S+N_T}\sum_{i=1}^{N_T+N_S} U_k(\boldsymbol{x}_i)$ where the true variances have been replaced by their empirical counterparts. In case of absence of a local maximum, the elbow rule is applied instead. The two method will be tested, and as the results will show the criteria mostly coincide.

Finally, the choice of ρ is performed by considering a range of possible values (typically between 10^{-5} and 1) and checking the precision and recall of the test when used on the joint training data $\mathcal{D}_S \cup \mathcal{D}_T$. We refer by $\hat{k}$ and $\hat{\rho}$ the choices of k and ρ made with this procedure. The whole transfer procedure has been summarized in Algorithm 1.

Algorithm 1: Full proposed transfer procedure.

Choice of hyperparameters α and k (Section 2.3):

Set α to between $\alpha^*/10$ and $\alpha^*/5$.

Calculate the LOOCV and $\overline{U}_k$ to select $\hat{k}$.

Perform $\hat{k}$ GD iterations (Section 2).

Choice of hyperparameter ρ (Section 2.3):

Set ρ by checking the recall and precision of the test on $\mathcal{D}_S \cup \mathcal{D}_T$.

for *every new sample* $\boldsymbol{x}$ **do**

Calculate the p-value $p_{\hat{k}}(\boldsymbol{x})$

if $p_{\hat{k}}(\boldsymbol{x}) < 0.05$ **then**

$\hat{y} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_k$ (reject H_0);

else

$\hat{y} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_T$ (keep H_0);

end

end

3. Numerical Experiments

In this section, the benefits of our framework, with the test in particular, are illustrated on synthetic and real-world datasets. Our goal is twofold: first we want to assess the fine-tuning procedure for the linear model itself, ensuring that it yields a strong prediction that performs better than the one obtained using the pure target model $\mathcal{M}_T$. To make the assessment even more relevant, we will compare our results with two enriched estimators from the literature:

- The model $\mathcal{M}_2$ of Bouveyron & Jacques (2010), which yielded the best results in their study. It obtains the fine-tuned estimator by $\hat{\boldsymbol{\beta}}_{FT} = \hat{\Lambda}\hat{\boldsymbol{\beta}}_S$ where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_1)$ with the two coefficients $\hat{\lambda}_0$ (for the intercept) and $\hat{\lambda}_1$ calculated on the target data.

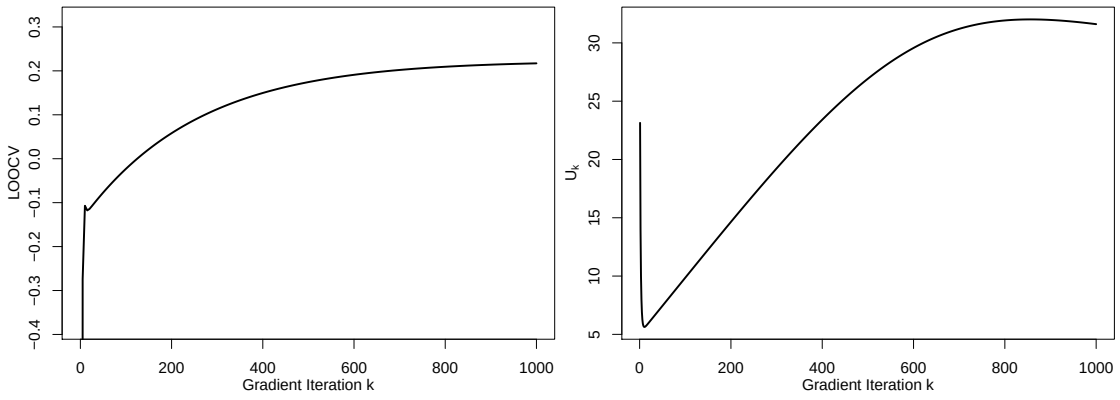
- The model $\mathcal{M}_{\hat{\lambda}}$ based on the estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}} = W_{\hat{\lambda}}\hat{\boldsymbol{\beta}}_T + (I_D - W_{\hat{\lambda}})\hat{\boldsymbol{\beta}}_S$ with $\hat{\lambda}$ as given in Chen et al. (2015).

The second objective is to assess how beneficial the test on the gain introduced in this paper is, i.e. how much improvement is brought with $\mathcal{M}_k^*$ over $\mathcal{M}_k$.

3.1. Synthetic - Polynomial Data

First we consider the problem of the estimation of the coefficients of a target polynomial $P_T(x) = \beta_{T,0} + \beta_{T,1}x + \beta_{T,2}x^2 + \beta_{T,3}x^3$ where $\beta_T = (-1, -1.8, 1.2, 1)^\top$. The advantage of this example lies in how it can be visualized, as one will see afterwards. We have $N_T = 20$ independent target observations $y_{T,i} = P_T(x_{T,i}) + \varepsilon_{T,i}$ with $x_{T,i}$ randomly sampled in $[-3, 1]$ and $\varepsilon_{T,i} \sim \mathcal{N}(0, \sigma_T^2)$. Additionally we have $N_S = 100$ independent source observations $y_{S,i} = P_S(x_{S,i}) + \varepsilon_{S,i}$ with $x_{S,i}$ sampled in $[0, 3]$ and $\varepsilon_{S,i} \sim \mathcal{N}(0, \sigma_S^2)$. The coefficients of P_S are the ones of P_T plus a gaussian noise of mean 0 and standard-deviation 0.3. Finally the noise variances are set to $\sigma_T^2 = \sigma_S^2 = 0.5$. This situation corresponds to mostly disjoint supports of source and target, for which transfer can be beneficial. Considering the locations of the samples for the source and target, the transfer is expected to be beneficial mostly for $x \geq 1$.

The choice for the three hyperparameters follows the strategies proposed in Section 2.3. The step size is set to $\alpha = \alpha^*/5$. For the number of gradient iterations, we aim at maximizing the LOOCV error on the target set and the criterion U_k calculated on $\mathcal{D}_S \cup \mathcal{D}_T$ which have been represented in Figure 3. While both criteria reach a maximum around $k = 1000$ iterations, their growth significantly slows down after 500 iterations. Since $\hat{\boldsymbol{\beta}}_k$ converges towards $\hat{\boldsymbol{\beta}}_T$, taking a too high value of k risks to erase the benefits brought by the transfer procedure. This is why we take $\hat{k} = 500$ here for instance.



(a) LOOCV (b) U_k

Figure 3: Criteria for k ($N_S = 100, N_T = 20$)

Finally the value of ρ is obtained by applying the test on the training data $\mathcal{D}_S \cup \mathcal{D}_T$ and finding

a good compromise between recall and precision on it. It must not be taken too small as else it will systematically reject H_0, and usually we set it such that it corresponds to the "edge of the cliff" where the recall drops and precision soars as represented in Figure. 4.

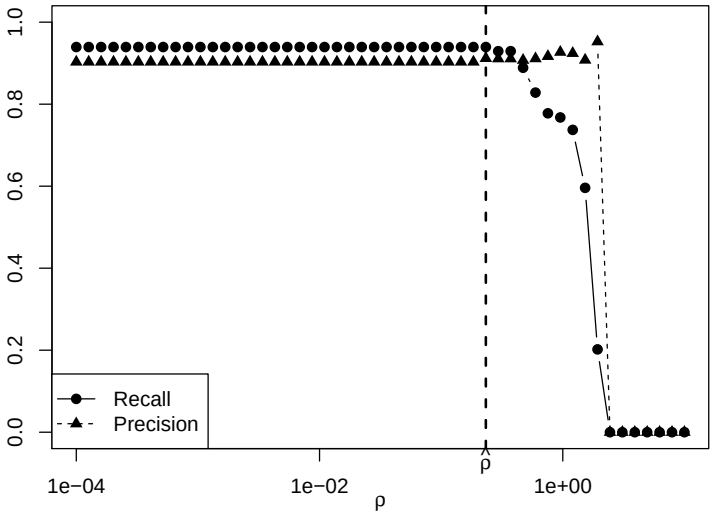


Figure 4: Choice of $\hat{\rho}$.

The true polynomial P_T as well as different estimates have been represented Fig. 5 with the gain defined by (4). As one can see, the pure target estimate is good for $x \leq 0$ typically but very off the true curve for $x > 0$. The estimates obtained by Chen et al.'s approach and ours are very close in this instance and significantly improve upon the pure target estimation. However Bouveyron & Jacques (2010)'s estimate is mostly off as well (except for high values of x) because of the lack of adaptability inherent to the method as discussed in introduction. These observations concur with Figure 6 (a) where the gain is positive for the corresponding x's. One also sees the benefits brought by our tuning method for k over a choice of $k = 50$ for instance, which does not leverage the target set enough. Finally the p-value of our test has been represented in function of x Figure 6 (b). It is high (i.e. the null hypothesis is kept) for $x < 0$, which corresponds to the domain where the target estimate was decent. However on the domain where the target model faired poorly and where $\mathcal{M}_{\hat{k}}$ was close to the true polynomial the p-value is small, which is conformed to intuition.

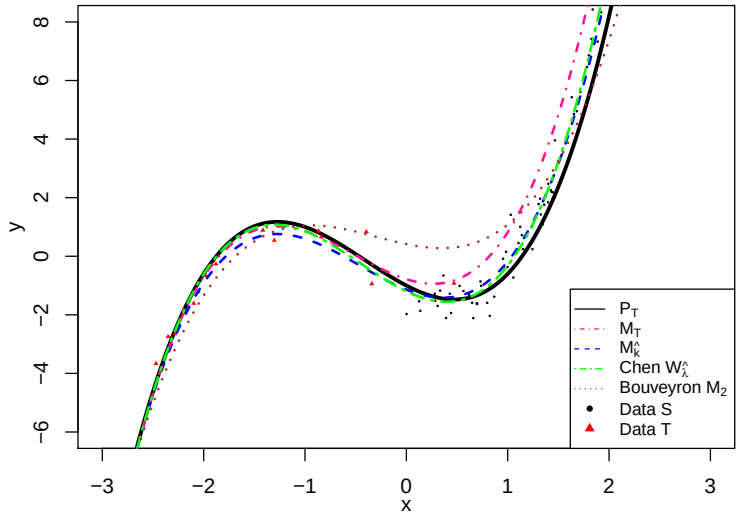
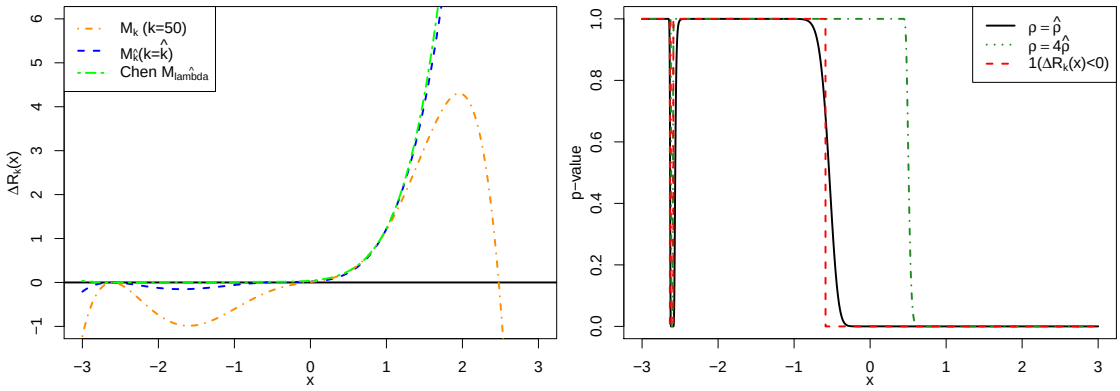


Figure 5: Polynomial P_T and its different estimates.



(a) Theoretical gain for different models.

(b) P-value of the test in function of x ($k = \hat{k}$).

Figure 6: Overlap of P_T and its different estimates.

In order to assess the results, we also compare the coefficient estimation error $\left\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_T\right\|$ as well as the prediction root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2},$$

where T the number of (new) test samples. Thus in total five different predictions errors are given corresponding to the three aforementioned benchmarks ($\mathcal{M}_T$, $\mathcal{M}_2$ from Bouveyron & Jacques (2010) and $\mathcal{M}_{\hat{\lambda}}$ from Chen et al. (2015)) as well as our fine-tuned model $\mathcal{M}_{\hat{k}}$, and the one constructed through the statistical test $\mathcal{M}^*_{\hat{k}}$. In order to evaluate the performance brought by our test, we consider the oracle prediction which knows in advance whether to use $\mathcal{M}_T$ or $\mathcal{M}_{\hat{k}}$ (i.e. when $\mathbb{E}[(y - \hat{y}_k)^2] < \mathbb{E}[(y - \hat{y}_T)^2]$, $\hat{y}_k$ is used for prediction, and $\hat{y}_T$ otherwise). Thus the closer $\mathcal{M}^*_{\hat{k}}$ is to the oracle, the better. The results are given in Table 1. Both the pure target model and the one of Bouveyron & Jacques (2010) yield very poor results as was expected. The best results are obtained here with our fine-tuned model for the prediction, albeit Chen et al.'s estimate is better for estimating the coefficient $\boldsymbol{\beta}_T$ itself. Finally the benefits of our test is illustrated on the final line of the table, improving upon $\mathcal{M}_{\hat{k}}$ by 0.02 points. The oracle is only marginally better, with a prediction RMSE of 0.902.

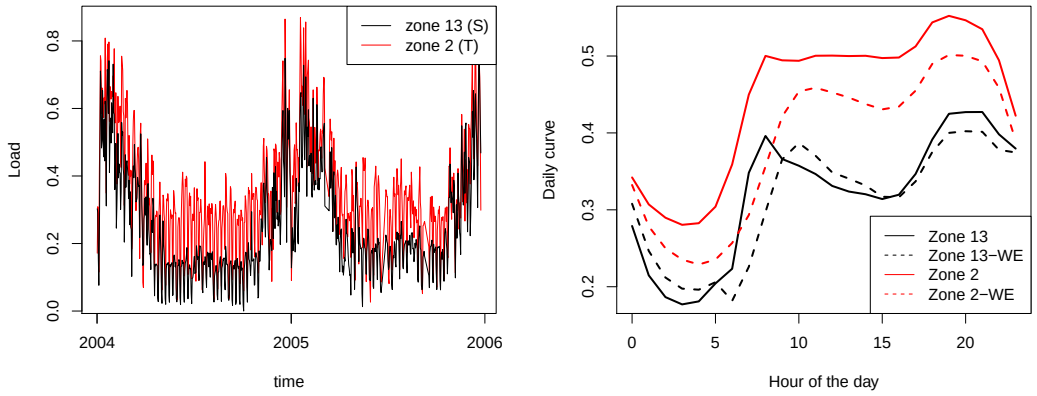
Model	$\left\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_T\right\|$	Prediction RMSE
$\mathcal{M}_T$	0.427	2.679
$\mathcal{M}_2$	0.944	2.023
$\mathcal{M}_{\hat{\lambda}}$	0.181	1.407
$\mathcal{M}_{\hat{k}}$	0.233	0.945
$\mathcal{M}^*_{\hat{k}}$		0.925
Oracle	-	0.902

Table 1: Errors for the different approaches.

3.2. GEFCOM2012 electricity consumption

This dataset was used during the GEFCOM2012 electricity consumption forecasting competition (Hong, Pinson & Fan, 2014). It consists of the electricity consumption of 21 areas (zones) located in the United-States available from the 1$^{\text{st}}$ of January 2004 to the 31$^{\text{st}}$ of December 2007 with a 1 hour temporal resolution (yielding 38,070 samples in total) that we normalized. Input variables include calendar ones such as the day of the week, time of the year, but also the temperature measurements of 10 meteorological stations over the same period. Typically, demand has an annual periodicity with peaks in winter. Mean demand is around 0.35($\pm$0.12) for the target data and 0.25($\pm$0.10) for the source. The nature of our transfer is twofold: across time (period on which a model has been

trained) and space (from one area to another). We will use the measured load at 8a.m. of zone 13 as source and zone 2 as target. To focus on the benefits of our test and to avoid time series stationarity issues, both load time series have been detrended. The overall load as well as the daily profiles have been represented in Figure. 7.



(a) Load over 2004-2005.

(b) Daily Load.

Figure 7: Comparison of the load demand for zones 13 (source) and 2 (target).

In our work we focus on the use of our hypothesis test and not achieving pure predictive performance. Therefore the model we consider for both source and target is very simple:

$$y_{\nu,t} = \beta_{\nu,0} + \beta_{\nu,1}|\sin(\omega t)| + \beta_{\nu,2} WE_t + \sum_{j=3}^{5} \beta_{\nu,j}\theta_t \, \mathbb{1}\big(\theta_t \in I_j\big) + \varepsilon_{\nu,t}, \quad \nu \in \{S,T\}, \tag{10}$$

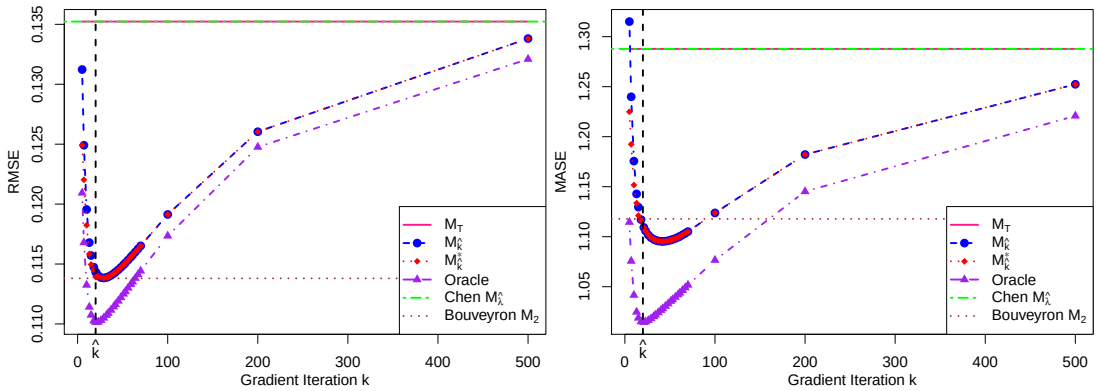
where $y_{\nu,t}$ is the load demand at 8a.m. for day t, $\omega = \frac{2\pi}{365}$. The sine term is used for the annual periodicity, WE_t is a binary variable whose value is on 1 on weekends. θ_t is the temperature and its effect has been cut into three intervals to translate the impact of heating and cooling on the electricity demand (Pierrot & Goude, 2011). Whether it is the source or the target data, the training data will be included within the year 2004, whereas the test data on which performance is finally evaluated will be the whole 2005 year of zone 2. The metrics of evaluation will be again the RMSE introduced earlier, as well as the mean absolute scaled error (MASE) defined by:

$$\text{MASE} = \frac{\frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|}{\frac{1}{N-1}\sum_{i=2}^{N}|y_i - y_{i-1}|},$$

where the denominator is calculated on the train data of size N.

3.2.1. First scenario

In this scenario we suppose that the data for the source $\mathcal{S}$ is available for the whole year 2004. The target training data will only be available from October the 1st to the end of the year. Hence $N_S = 366$ and $N_T = 92$. The RMSE and MASE for different values of k are represented in Figure 8, with a vertical line corresponding to our chosen $\hat{k}$ with the strategy described in Section 2.3. Precise numerical values are given in Table 2. Here the improvement brought by the test is only marginal, for k below a threshold. This is due to the discussed phenomenon at the end of Section 2.2 where when $k \to \infty$ the test tends to systematically reject H_0. Note that the RMSE is close to being minimal for $\hat{k}$. The errors for $\mathcal{M}_2$ and $\mathcal{M}_{\hat{\lambda}}$ have been calculated and represented as dashed horizontal lines. While the former yields results very similar to the fine-tuning approach, the latter in this case yields poor results, as in this situation $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is almost identical to $\hat{\boldsymbol{\beta}}_T$. Most importantly the model $\mathcal{M}^*_{\hat{k}}$ is always as good as $\mathcal{M}_{\hat{k}}$: the test can thus be used safely in practice. The p-value over time on the test set is also represented Fig. 9. One sees that it's almost always close to 0, except locally for cold months. Since $\mathcal{M}_T$ was trained on a similar period the year before, such a behavior is logical.



(a) Test RMSE in function of k.

(b) Test MASE in function of k.

Figure 8: Results on the test data (first scenario).

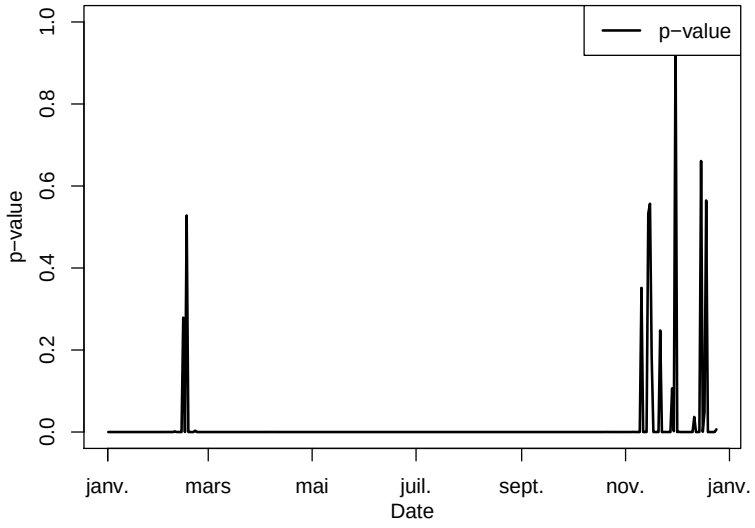


Figure 9: P-value over time for the first scenario ($k = \hat{k}$).

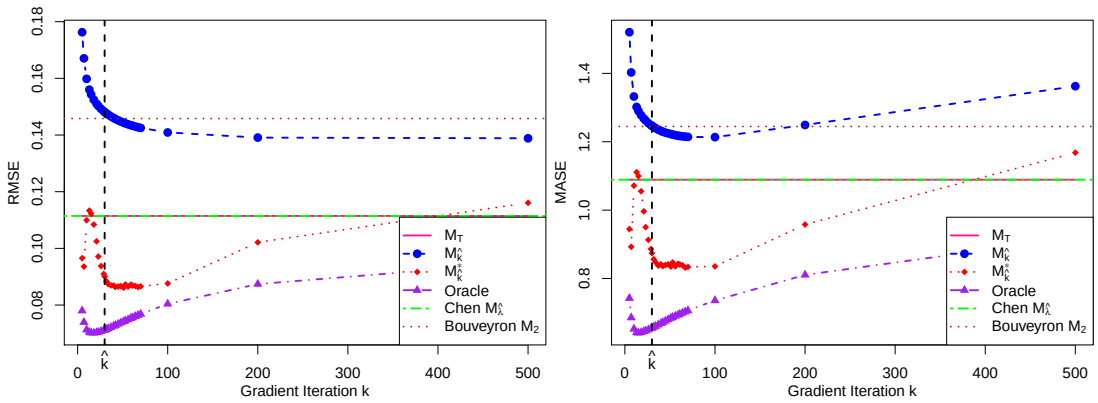
3.2.2. Second scenario

We consider the case where the training data from the source zone is available between April the 1st and September the 30th 2004. The training data from the target one is available between the 1st of September and the 31st of December 2004, and thus $N_S = 182$ and $N_T = 122$. In practice it

could correspond to the case where a customer breaks his contract, and a new one arrives.

Results on the test data for the different approaches are given Fig. 10. We see that this time the test significantly improves upon the individual forecasts, lowering the RMSE by up to 0.04 RMSE compared to $\mathcal{M}_{\hat{k}}$ in general. The test efficiently detects the situations of positive and negative transfer, thus taking advantage of each model's specificities. However in this case the value of $\hat{k}$ obtained by the tuning procedure is too low (yielding $\hat{k} = 30$ despite a better value being 40). Nonetheless the efficiency of the test is shown as the RMSE is reduced by about 0.02 compared to $\mathcal{M}_T$ and is very close to the oracle. Chen et al.'s estimator is once more very close to $\hat{\boldsymbol{\beta}}_T$, while this time Bouveyron et al.'s approach yields poor results. Note that before the test all transfer approaches struggled: this is because of the important differences between source and target data. The test efficiently allows to take advantage of the specificities of each model.

The p-value over time on the test set for $k = 40$ is also plotted Figure 11. It is close to 0 on a period similar to the one the source model was trained the year before, and large during the cold months where the model $\mathcal{M}_T$ is expected to be better.



(a) Test RMSE in function of k.

(b) Test MASE in function of k.

Figure 10: Results on the test data (second scenario).

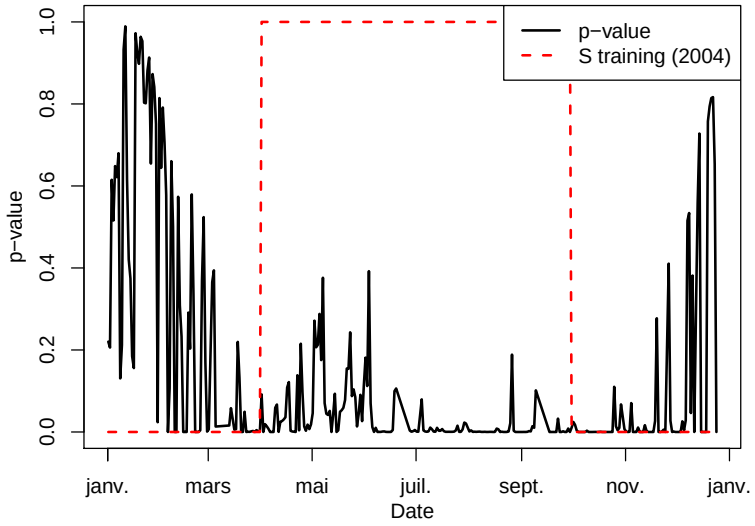


Figure 11: P-value over time for the second scenario ($k = 40$).

Model	RMSE	MASE
$\mathcal{M}_T$	0.135	1.29
$\mathcal{M}_2$	0.114	1.12
$\mathcal{M}_{\hat{\lambda}}$	0.135	1.29
$\mathcal{M}_{\hat{k}}$	0.114	1.11
$\mathcal{M}^*_{\hat{k}}$	0.114	1.10
Oracle	0.110	1.01

Table 2: Errors for the first scenario.

Model	RMSE	MASE
$\mathcal{M}_T$	0.111	1.09
$\mathcal{M}_2$	0.146	1.245
$\mathcal{M}_{\hat{\lambda}}$	0.111	1.089
$\mathcal{M}_{k=40}$	0.146	1.23
$\mathcal{M}^*_{k=40}$	0.087	0.841
Oracle	0.073	0.669

Table 3: Errors for the second scenario.

Finally it is important to assess the sensibility of our approach with respect to the hyperparameters. To achieve that, we represented the difference of RMSE of $\mathcal{M}^*_k$ and $\mathcal{M}_T$ in function of k and ρ in Figure 12. The area surrounding our choice of $\hat{k}$ and $\hat{\rho}$ corresponds to one of significant benefits of the test, even if a couple of gradient iterations more would have been beneficial as highlighted previously. Ergo the results are not excessively sensitive to the values of the hyperparameters. Additionally this plot also shows that the use of the test is almost exclusively beneficial and has very low inherent risk.

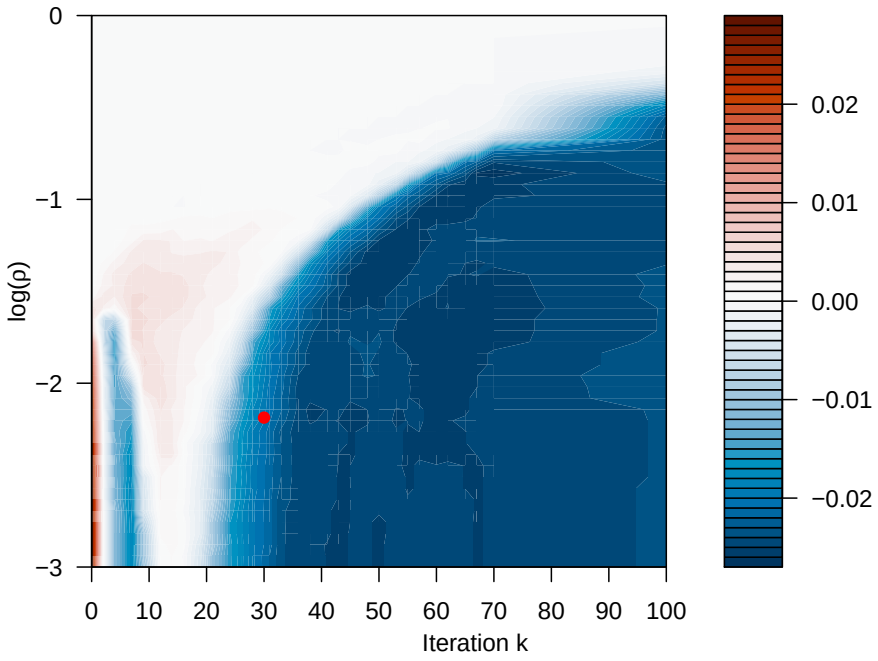
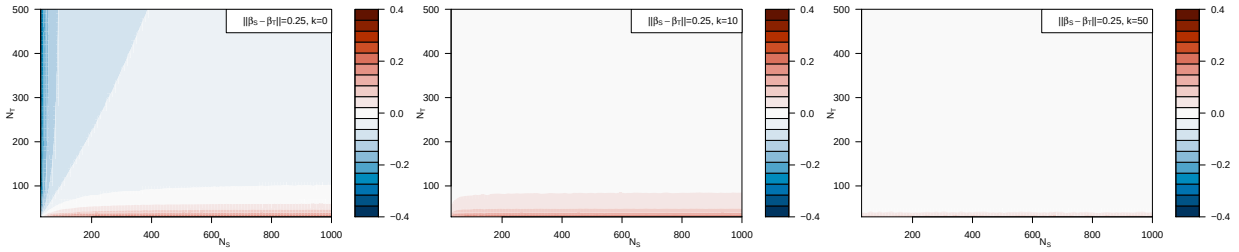


Figure 12: Difference of RMSE between $\mathcal{M}_k^*$ and $\mathcal{M}_T$ in function of k and ρ. Our choice is represented by the red dot. Blue areas correspond to a higher accuracy of the test-based model.

4. Interpretation of the gain with sample sizes

A natural question to ask is how the gain evolves with the sample sizes N_S and N_T. One would for instance expect the gain to increase when the number of source samples is order of magnitudes higher than the target one. In order to analyze these dependencies, we consider the following experimental framework. We suppose that the source and target data are i.i.d. $x_{\nu,i} \sim \mathcal{N}(0, I_D)$ (thus $\Sigma_\nu \sim \mathcal{W}_D(I_D, N_\nu)$ where $\mathcal{W}_D(\Psi, n)$) denotes the Wishart distribution of scale matrix $\Psi \in \mathbb{R}^{D\times D}$ and degrees of freedom n. For (N_S, N_T) in a grid $I_S \times I_T$, we calculate and average the gain $\Delta\mathcal{R}_k(x)$ over $B = 50$ simulations for $x \sim \mathcal{N}(0, I_D)$ as well. Algorithm 2 summarizes the procedure. This experiment is conducted for a dimension size $D = 15$, $k \in \{0, 10, 50\}$, $\alpha = \alpha^*/5$ and $\|\beta_S - \beta_T\| = 0.25$ (both coefficients have been randomly sampled). In order to improve the readability, the gain has been thresholded to the range $[-0.4, 0.4]$. The results are represented in Fig. 13.

Phases are observed depending on the values of N_S and N_T, and follow the intuition. When no fine-tuning is performed (i.e. $k = 0$) the gain will be positive only when the number of target samples



(a) $k = 0$ (b) $k = 10$ (c) $k = 50$

Figure 13: Transfer phases in function of N_S, N_T and k.

N_T is small and the number of source ones N_S is large enough. For N_T above a certain threshold, negative transfer will systematically happen. When k increases, the blue areas corresponding to negative gain fade away thus meaning that at worst the transfer procedure will have a neutral impact, even for large values of N_T. The benefits of transfer through fine-tuning are particularly visible for $k = 10$ with an increase of the size of the positive transfer areas: the fine-tuning procedure allows to take advantage of both source and target samples. However as emphasized before, an excessive number of gradient iterations may erase the benefits of transfer as seen in Figure 13 (c) obtained for $k = 50$ where only for extremely small values of N_T transfer can be beneficial. This is because the fine-tuned estimator $\hat{\boldsymbol{\beta}}_k$ has come too close to the pure target one $\hat{\beta}_T$. Note that these figures remind of Fig. 1 from Ben-David et al. (2010).

Algorithm 2: Gain simulation for varying N_S & N_T

Initialization: $D, \beta_\nu, \sigma_\nu^2, k$. $I_S = \{30, 40, \ldots, 1000\}$ and $I_T = \{30, 40, \ldots, 500\}$.

for (N_S, N_T) *in* $I_S \times I_T$ **do**

  $\overline{\Delta\mathcal{R}_k}(N_S, N_T) \leftarrow 0$;

  for $b = 1, \ldots, B$ **do**

    Generate $X_\nu \sim \mathcal{N}(0, I_D)$. Deduce Σ_ν.;

    Generate $\boldsymbol{x} \sim \mathcal{N}(0, I_D)$;

    Calculate $\overline{\Delta\mathcal{R}_k}(N_S, N_T) \leftarrow \overline{\Delta\mathcal{R}_k}(N_S, N_T) + (1/B)\, \boldsymbol{x}^\top H_k \boldsymbol{x}$;

  end

end

5. Conclusion and future work

In this paper a novel framework for the problem of transfer learning for the linear model is proposed. By defining the gain of transfer by a difference of quadratic prediction errors, we obtain a quantity that measures how beneficial or detrimental transfer by gradient descent is for a new (potentially unobserved) $\boldsymbol{x}$. However the framework of the gain is applicable for any estimator of the form $\hat{\boldsymbol{\beta}}(W) = W\hat{\boldsymbol{\beta}}_S + (I_D - W)\hat{\boldsymbol{\beta}}_T$, which encompasses many found in the literature. Since this gain depends on unknown parameters in practice, we derived a statistical test relying of the Fisher-Snedecor $\mathcal{F}$ distribution to predict negative transfer. The test was applied on synthetic as well as real-world electricity demand data, where it proved its ability to predict negative transfer for new observations. Furthermore our fine-tuning approach proved to be reliable no matter the situation, never being completely off such as the benchmarks from the literature sometimes.
However despite its success, some points remain to investigate. How to choose the right number of gradient iterations k remains problematic, although an empirical approach has been suggested. Furthermore in order to obtain a tractable calculation and satisfying empirical results, we had to rely on an approximation. Another possibility would be to transfer only a subset of parameters. This is often the case for neural networks where only certain layers are transferred Laptev, Yu & Rajagopal (2018), but could be adapted for linear models.

We have also supposed that the matrices Σ_ν are invertible. However defining the gain without this hypothesis is still possible although its form is slightly more complex, which makes it difficult to adapt the test directly. Finally in this paper we made the hypothesis of linearity, which could seem restrictive. However nonlinearity can be achieved through generalized additive models (GAM) for instance. Since they boil down to a linear model, the formula of the gain is valid for it as well. However as such, the test we introduced cannot be used with GAM yet, and how to extrapolate it is currently under investigation.

Appendix A. Appendix

The appendix presents detailed proofs of the results from Section 2.

Appendix A.1. Proposition 1

Proof. We proceed by mathematical induction.

- For $k = 0$ the property is trivial. $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_S = A^0\hat{\boldsymbol{\beta}}_S + (I_D - A^0)\hat{\boldsymbol{\beta}}_T$.
- Let $k \in \mathbb{N}$ be. We suppose the property true at rank k. We have

$$\hat{\boldsymbol{\beta}}_{k+1} = \hat{\boldsymbol{\beta}}_k - \alpha\nabla J_T(\hat{\boldsymbol{\beta}}_k) = \hat{\boldsymbol{\beta}}_k - \alpha\Sigma_T\hat{\boldsymbol{\beta}}_k + \alpha X_T^\top \boldsymbol{Y}_T.$$

  By definition of $A = I_D - \alpha\Sigma_T$ and because $X_T^\top \boldsymbol{Y}_T = \Sigma_T\hat{\boldsymbol{\beta}}_T$ we obtain:

$$\hat{\boldsymbol{\beta}}_{k+1} = A\hat{\boldsymbol{\beta}}_k + \alpha\Sigma_T\hat{\boldsymbol{\beta}}_T.$$

  Finally by induction hypothesis:

$$\hat{\boldsymbol{\beta}}_{k+1} = A[A^k\hat{\boldsymbol{\beta}}_S + (I_D - A^k)\hat{\boldsymbol{\beta}}_T] + \alpha\Sigma_T\hat{\boldsymbol{\beta}}_T = A^{k+1}\hat{\boldsymbol{\beta}}_S + (I_D - A^{k+1})\hat{\boldsymbol{\beta}}_T,$$

which concludes the induction. □

Appendix A.2. Equations (2) and (3)

Proof. Let P be the orthogonal matrix of eigenvectors of Σ_T be, i.e. such that $\Sigma_T = P\Lambda P^\top$ with $\Lambda = \mathrm{diag}(\lambda_i, i = 1, \ldots, D)$ and $PP^\top = P^\top P = I_D$. Thus $\hat{\boldsymbol{\beta}}_\nu = P\tilde{\boldsymbol{\beta}}_\nu \Leftrightarrow \tilde{\boldsymbol{\beta}}_\nu = P^\top\hat{\boldsymbol{\beta}}_\nu$. One can also write that $A = P(I_D - \alpha\Lambda)P^\top$. Hence reinjecting in (1) gives:

$$\hat{\boldsymbol{\beta}}_k = P(I_D - \alpha\Lambda)^k P^\top\hat{\boldsymbol{\beta}}_S + P(I_D - (I_D - \alpha\Lambda)^k)P^\top\hat{\boldsymbol{\beta}}_T,$$

and applying $P^\top$ on the left of this equation yields:

$$\tilde{\boldsymbol{\beta}}_k = (I_D - \alpha\Lambda)^k\tilde{\boldsymbol{\beta}}_S + (I_D - (I_D - \alpha\Lambda)^k)\tilde{\boldsymbol{\beta}}_T.$$

Finally the matrices involved are diagonal with respective terms $(1 - \alpha\lambda_i)^k$ and $1 - (1 - \alpha\lambda_i)^k$, thus resulting in equation (3). □

Appendix A.3. Proof of Proposition 2

Proof. We remind that $\hat{\boldsymbol{\beta}}_S \sim \mathcal{N}(\boldsymbol{\beta}_S, \sigma_S^2\Sigma_S^{-1})$ and $\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}(\boldsymbol{\beta}_T, \sigma_T^2\Sigma_T^{-1})$. By independence of $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_T$ we thus have:

$$\hat{\boldsymbol{\beta}}_k \sim \mathcal{N}\Big(A^k\boldsymbol{\beta}_S + (I_D - A^k)\boldsymbol{\beta}_T, \sigma_S^2A^k\Sigma_S^{-1}A^k + \sigma_T^2(I_D - A^k)\Sigma_T^{-1}(I_D - A^k)\Big).$$

It is easy to see that $\sigma_T^2(I_D - A^k)\Sigma_T^{-1}(I_D - A^k) = \sigma_T^2\alpha^2\Omega_k\Sigma_T\Omega_k$. We will note $\boldsymbol{\beta}_k = \mathbb{E}[\hat{\boldsymbol{\beta}}_k]$ and $V_k = \mathrm{Var}(\hat{\boldsymbol{\beta}}_k)$. For an independent $y = \boldsymbol{x}^\top\boldsymbol{\beta}_T + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_T^2)$ we obtain that:

$$y - \hat{y}_T = \boldsymbol{x}^\top(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T) + \varepsilon \sim \mathcal{N}\Big(0, \sigma_T^2(1 + \boldsymbol{x}^\top\Sigma_T^{-1}\boldsymbol{x})\Big),$$

$$y - \hat{y}_k = \boldsymbol{x}^\top(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_k) + \varepsilon \sim \mathcal{N}\Big(\boldsymbol{x}^\top(\boldsymbol{\beta}_T - \boldsymbol{\beta}_k), \sigma_T^2 + \boldsymbol{x}^\top V_k\boldsymbol{x}\Big).$$

Thus $\mathcal{R}(\mathcal{M}_T) = \mathbb{E}[(y - \hat{y}_T)^2] = \mathrm{Var}(y - \hat{y}_T) + \mathbb{E}[y - \hat{y}_T]^2 = \sigma_T^2(1 + \boldsymbol{x}^\top\Sigma_T^{-1}\boldsymbol{x})$ and $\mathcal{R}(\mathcal{M}_{T|S}) = \mathbb{E}[(y - \hat{y}_k)^2] = \sigma_T^2 + \boldsymbol{x}^\top V_k\boldsymbol{x} + \Big(\boldsymbol{x}^\top(\boldsymbol{\beta}_T - \boldsymbol{\beta}_k)\Big)^2$. Therefore:

$$\Delta\mathcal{R}_k(\boldsymbol{x}) = \sigma_T^2\boldsymbol{x}^\top\Sigma_T^{-1}\boldsymbol{x} - \boldsymbol{x}^\top V_k\boldsymbol{x} - \boldsymbol{x}^\top B_k\boldsymbol{x}.$$

Finally noticing that $\boldsymbol{\beta}_T - \boldsymbol{\beta}_k = A^k(\boldsymbol{\beta}_T - \boldsymbol{\beta}_S)$, we obtain that $\Big(\boldsymbol{x}^\top(\boldsymbol{\beta}_T - \boldsymbol{\beta}_k)\Big)^2 = \boldsymbol{x}^\top A^k B A^k\boldsymbol{x}$ where $B = (\boldsymbol{\beta}_T - \boldsymbol{\beta}_S)(\boldsymbol{\beta}_T - \boldsymbol{\beta}_S)^\top$ thus yielding the expected result. □

Appendix A.4. Proof of the equations of (5)

Proof. H_k is symmetric. Hence we can introduce $\{\boldsymbol{u}_i\}_{i=1..D}$ an orthonormal basis of eigenvectors of it with $\lambda_i(H_k)$ the associated eigenvalues. Let $\boldsymbol{x} \in \mathbb{R}^D$ be with coordinates x_i in this basis. Thus $\boldsymbol{x}$ can be rewritten $\boldsymbol{x} = \sum_{i=1}^D x_i\boldsymbol{u}_i$. Since $\{\boldsymbol{u}_i\}$ is orthonormal (i.e. $\boldsymbol{u}_i^\top\boldsymbol{u}_j = 1$ if $i = j$ and 0 else) it follows that:

$$\boldsymbol{x}^\top H_k\boldsymbol{x} = \sum_{i,j=1}^D \lambda_i(H_k)\, x_i x_j\, \boldsymbol{u}_i^\top\boldsymbol{u}_j = \sum_{i=1}^D \lambda_i(H_k)\, x_i^2.$$

Since $\lambda_{\min}(H_k) \leq \lambda_i(H_k) \leq \lambda_{\max}(H_k)$ we get that $\lambda_{\min}(H_k)\,\|\boldsymbol{x}\|^2 \leq \boldsymbol{x}^\top H_k\boldsymbol{x} \leq \lambda_{\max}(H_k)\,\|\boldsymbol{x}\|^2$. Finally remembering that $\boldsymbol{x}^\top H_k\boldsymbol{x} = \mathbb{E}[(y - \hat{y}_T)^2] - \mathbb{E}[(y - \hat{y}_k)^2]$ yields (5). □

Appendix A.5. Proof of Theorem 1

Proof. It would be natural to reject H_0 if an estimator $\hat{\delta}(\boldsymbol{x})$ of the gain is above a certain threshold. Hence a natural form of such a decision rule is $\mathbb{1}(\hat{\delta}(\boldsymbol{x}) > K_a)$, where K_a is a constant depending on the desired level a of the test. We consider the estimator of $\Delta\mathcal{R}_k(\boldsymbol{x})$:

$$\hat{\delta}(\boldsymbol{x}) = \hat{\sigma}_T^2\,\boldsymbol{x}^\top\big(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k\big)\boldsymbol{x} - \hat{\sigma}_S^2\boldsymbol{x}^\top A^k\Sigma_S^{-1}A^k\boldsymbol{x} - \boldsymbol{x}^\top A^k B A^k\boldsymbol{x}.$$

While the matrix B is not accessible in practice, we start from this estimator for the sake of the simplicity of the calculations. We will address this issue later. It can be proved (see hereafter) that the type I error, the probability of wrongly rejecting the null hypothesis, is the largest at the boundary $\Delta\mathcal{R}_k(\boldsymbol{x}) = 0$. Thus $\hat{\delta}(\boldsymbol{x}) > K_a$ is equivalent to:

$$\frac{\hat{\sigma}_T^2/\sigma_T^2}{\hat{\sigma}_S^2/\sigma_S^2} + \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{\boldsymbol{x}^\top A^k B A^k \boldsymbol{x}}{\sigma_T^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}} > \frac{K_a + \boldsymbol{x}^\top A^k B A^k \boldsymbol{x} + \hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}}{\hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}}.$$

Since $\frac{\hat{\sigma}_T^2/\sigma_T^2}{\hat{\sigma}_S^2/\sigma_S^2} \sim \mathcal{F}(N_T - D, N_S - D)$, taking $K_a = q^{1-a}\hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x} - \boldsymbol{x}^\top A^k B A^k \boldsymbol{x} - \hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x} + \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2} \boldsymbol{x}^\top A^k B A^k \boldsymbol{x}$ (where q^{1-a} is the quantile of order $1-a$ of the $\mathcal{F}(N_T - D, N_S - D)$ distribution) yields the test of level a:

$$\mathbb{1}\Big(\phi_k(\boldsymbol{x}) := \frac{\hat{\sigma}_T^2 \boldsymbol{x}^\top \big(\Sigma_T^{-1} - \alpha^2 \Omega_k \Sigma_T \Omega_k\big)\boldsymbol{x} - (\hat{\sigma}_T/\sigma_T)^2 \boldsymbol{x}^\top A^k B A^k \boldsymbol{x}}{\hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}} > q^{1-a}\Big).$$

However B and σ_T are unknown in practice, we will thus have to rely on a lower bound of $\phi_k(\boldsymbol{x})$ for the test. By hypothesis, we have $\|\boldsymbol{\beta}_T - \boldsymbol{\beta}_S\|/\sigma_T \leq \rho$. Since B is symmetric $\boldsymbol{x}^\top A^k B A^k \boldsymbol{x} \leq \lambda_{\max}(B) \left\|A^k \boldsymbol{x}\right\|^2$. Moreover B is a rank 1 matrix and thus its sole nonzero eigenvalue is $\lambda_{\max}(B) = \|\boldsymbol{\beta}_T - \boldsymbol{\beta}_S\|^2$. The aforementioned hypothesis leads to $\frac{1}{\sigma_T^2}\boldsymbol{x}^\top A^k B A^k \boldsymbol{x} \leq \rho^2 \left\|A^k \boldsymbol{x}\right\|^2$. Therefore we have the following lower bound $\psi_k(\boldsymbol{x})$ of $\phi_k(\boldsymbol{x})$ that can be used in practice:

$$\psi_k(\boldsymbol{x}) = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{\boldsymbol{x}^\top(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k)\boldsymbol{x} - \rho^2\left\|A^k\boldsymbol{x}\right\|^2}{\boldsymbol{x}^\top A^k \Sigma_S A^k \boldsymbol{x}}.$$

What remains to prove is that the type I error is maximum at the frontier, i.e. where $\Delta\mathcal{R}_k(\boldsymbol{x}) = 0$. If $\Delta\mathcal{R}_k(\boldsymbol{x}) \leq 0$ then:

$$\phi_k(\boldsymbol{x}) \leq \frac{\hat{\sigma}_T^2 \frac{\sigma_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x} + \boldsymbol{x}^\top A^k B A^k \boldsymbol{x}}{\sigma_T^2} - (\hat{\sigma}_T^2/\sigma_T^2)\boldsymbol{x}^\top A^k B A^k \boldsymbol{x}}{\hat{\sigma}_S^2 \boldsymbol{x}^\top A^k \Sigma_S^{-1} A^k \boldsymbol{x}},$$

with an equality on the frontier. Finally the r.h.s. can be simplified in $F = \frac{\hat{\sigma}_T^2/\sigma_T^2}{\hat{\sigma}_S^2/\sigma_S}$. Thus finally:

$$\mathbb{P}_{\Delta\mathcal{R}_k(\boldsymbol{x})\leq 0}(\phi_k(\boldsymbol{x}) \geq q^{1-a}) \leq \mathbb{P}_{F\sim\mathcal{F}(N_T-D,N_S-D)}\Big(F \geq q^{1-a}\Big) = \mathbb{P}_{\Delta\mathcal{R}_k(\boldsymbol{x})=0}(\phi_k(\boldsymbol{x}) \geq q^{1-a}\Big) = a,$$

which proves that the type I error is maximum at $\Delta\mathcal{R}_k(\boldsymbol{x}) = 0$ and that the level of the test is a.

Thus the p-value of the test relying on $\phi_k(\boldsymbol{x})$ can thus be upper bounded by $\mathbb{P}_{F\sim\mathcal{F}(N_T-D,N_S-D)}\big(F \geq \psi_k(\boldsymbol{x})\big)$, proving all the results of the theorem. □

Appendix A.6. Proof of $p_k(x) \to 0$ when $k \to \infty$

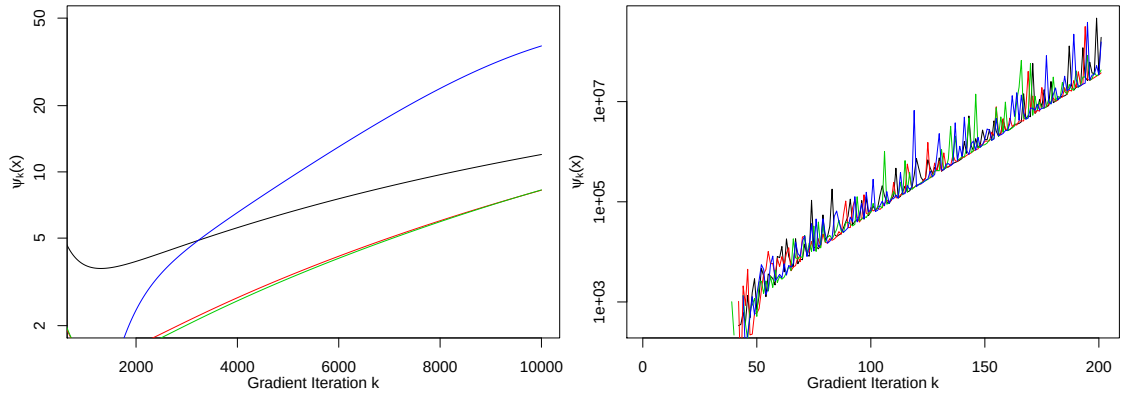
When $k \to \infty$, the gain $\Delta\mathcal{R}_k(\boldsymbol{x})$ converges towards 0 (i.e. transfer becomes neutral, as $\hat{\boldsymbol{\beta}}_k \to \hat{\boldsymbol{\beta}}_T$). Hence when the number of gradient iterations k becomes large, it would be logical for the test to keep the null hypothesis that transfer is not beneficial. However in fact $\psi_k(\boldsymbol{x}) \to +\infty$, and since the test has the form $\mathbb{1}(\psi_k(\boldsymbol{x}) > q_{1-a})$ (Eq. (6)), the null hypothesis H_0 (transfer is negative) will systematically be rejected, no matter $\boldsymbol{x}$. Elements of proof for the limit of $\psi_k(\boldsymbol{x})$ are given hereafter, as well as numerical illustrations in Figure A.14:

$$\begin{aligned}\psi_k(\boldsymbol{x}) &= \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{\boldsymbol{x}^\top(\Sigma_T^{-1}-\alpha^2\Omega_k\Sigma_T\Omega_k)\boldsymbol{x}-\rho^2\left\|A^k\boldsymbol{x}\right\|^2}{\boldsymbol{x}^\top A^k\Sigma_S^{-1}A^k\boldsymbol{x}} \\ &= \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{\boldsymbol{x}^\top(2A^k\Sigma_T^{-1}-A^{2k}\Sigma_T^{-1})\boldsymbol{x}-\rho^2\left\|A^k\boldsymbol{x}\right\|^2}{\boldsymbol{x}^\top A^k\Sigma_S^{-1}A^k\boldsymbol{x}}.\end{aligned}$$

Using asymptotic notations, it can be noted that $\boldsymbol{x}^\top A^{2k}\Sigma_T^{-1}\boldsymbol{x} = o\Big(\boldsymbol{x}^\top A^k\Sigma_T^{-1}\boldsymbol{x}\Big)$ as well as $\left\|A^k\boldsymbol{x}\right\|^2 = o\Big(\boldsymbol{x}^\top A^k\Sigma_T^{-1}\boldsymbol{x}\Big)$. Hence:

$$\psi_k(\boldsymbol{x}) \underset{k\to\infty}{\sim} \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2}\frac{2\boldsymbol{x}^\top A^k\Sigma_T^{-1}\boldsymbol{x}}{\boldsymbol{x}^\top A^k\Sigma_S^{-1}A^k\boldsymbol{x}}.$$

Now it is visible that the denominator converges towards 0 with higher speed, and thus $\psi_k(\boldsymbol{x}) \to +\infty$. We illustrated those results in two numerical settings: the first is the one presented in Section 3.1 with the polynomial dataset and yields Fig. A.14a. In the second one we consider a simple linear regression problem where $\boldsymbol{\beta}_T$ is randomly sampled on the uniform sphere of dimension D. The source coefficients $\boldsymbol{\beta}_S$ are obtained by $\boldsymbol{\beta}_S = \boldsymbol{\beta}_T + r\boldsymbol{d}$ where r corresponds to the desired $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|$ and $\boldsymbol{d}$ is a direction sampled on the uniform sphere again. The design matrices X_S and X_T are obtained by i.i.d. sampling the rows respectively from $\mathcal{N}(0,1)$ and $\mathcal{U}[-1,1]$. Figure A.14b corresponds to this situation. In both instances the growth of the test statistic $\psi_k(\boldsymbol{x})$ towards infinity is clear.



(a) Polynomial setting.

(b) Linear setting.

Figure A.14: Behavior of the test statistic $\psi_k(\boldsymbol{x})$ when k goes to infinity.

Appendix A.7. Proof of equation (9)

Proof. The KL divergence between two univariate gaussians directly yields:

$$\begin{aligned} 2D_{KL}(\mathcal{N}_k||\mathcal{N}_T) &= \frac{\boldsymbol{x}^\top V_k \boldsymbol{x}}{\sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}} + \frac{\left(\boldsymbol{x}^\top \left(\boldsymbol{\beta}_T - \boldsymbol{\beta}_k\right)\right)^2 - \sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}}{\sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}} - \ln\left(\frac{\boldsymbol{x}^\top V_k \boldsymbol{x}}{\sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}}\right) \\ &= \frac{-\Delta\mathcal{R}_k(\boldsymbol{x})}{\sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}} - \ln\left(\frac{\boldsymbol{x}^\top V_k \boldsymbol{x}}{\sigma_T^2 \boldsymbol{x}^\top \Sigma_T^{-1} \boldsymbol{x}}\right) \quad \text{hence the result.} \end{aligned}$$

□

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, *79*, 151–175.

Bertsekas, D. (2015). *Convex optimization algorithms*. Athena Scientific.

Bouveyron, C., & Jacques, J. (2010). Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters*, *31*, 2237–2247.

Cai, L., Gu, J., Ma, J., & Jin, Z. (2019). Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees. *Energies*, *12*, 159.

Chen, A., Owen, A. B., Shi, M. et al. (2015). Data enriched linear regression. *Electronic journal of statistics*, *9*, 1078–1112.

Dar, Y., & Baraniuk, R. G. (2020). Double double descent: On generalization errors in transfer learning between linear regression tasks. *arXiv preprint arXiv:2006.07002*, .

Dar, Y., & Baraniuk, R. G. (2021). Transfer learning can outperform the true prior in double descent regularization. *arXiv preprint arXiv:2103.05621*, .

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2018). Transfer learning for time series classification. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 1367–1376). IEEE.

Gao, J., Fan, W., Jiang, J., & Han, J. (2008). Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 283–291).

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012.

Laptev, N., Yu, J., & Rajagopal, R. (2018). Reconstruction and regression loss for time-series transfer learning. In *Proc. SIGKDD MiLeTS*.

Lounici, K., Pontil, M., Tsybakov, A. B., & Van De Geer, S. (2009). Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, .

Lounici, K., Pontil, M., Van De Geer, S., Tsybakov, A. B. et al. (2011). Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, *39*, 2164–2204.

Maurer, A. (2006). Bounds for linear multi-task learning. *Journal of Machine Learning Research*, *7*, 117–139.

Mihalkova, L., Huynh, T., & Mooney, R. J. (2007). Mapping and revising markov logic networks for transfer learning. In *Aaai* (pp. 608–614). volume 7.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*, 1345–1359.

Pierrot, A., & Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, *2011*.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, *35*, 1285–1298.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 9.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). *Transfer Learning*. Cambridge University Press. doi:10.1017/9781139061773.

Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2019). Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5704–5713).