Multi-Task Output Space Regularization

Sergey Feldman SERGEYF@U.WASHINGTON.EDU
Bela A. Frigyik FRIGYIK@U.WASHINGTON.EDU
Maya R. Gupta GUPTA@EE.WASHINGTON.EDU
Dept. of Electrical Engineering
University of Washington
Seattle, WA 98195, USA

Luca Cazzanti LUCA@APL.WASHINGTON.EDU
Applied Physics Lab
University of Washington
Seattle, WA 98195, USA

Peter Sadowski PETERJSADOWSKI@GMAIL.COM
Dept. of Computer Science
University of California, Irvine
Irvine CA 92697, USA

Editor:

Abstract

We investigate multi-task learning from an output space regularization perspective. Most multi-task approaches tie together related tasks by constraining them to share input spaces and function classes. In contrast to this, we propose a multi-task paradigm which we call output space regularization, in which the only constraint is that the *output* spaces of the multiple tasks are related. We focus on a specific instance of output space regularization, multi-task averaging, that is both widely applicable and amenable to analysis. The multi-task averaging estimator improves on the single-task sample average under certain conditions, which we detail. Our analysis shows that for a simple case the optimal similarity depends on the ratio of the task variance to the task differences, but that for more complicated cases the optimal similarity behaves nonlinearly. Further, we show that the estimates produced are a convex combination of the tasks' sample averages. We discuss the Bayesian viewpoint. Three applications of multi-task output space regularization are presented: multi-task kernel density estimation, multi-task-regularized empirical moment constraints in similarity discriminant analysis, and multi-task local linear regression. Experiments on real data sets show statistically significant gains.

Keywords: multi-task learning, similarity-based regularization, output space regularization, manifold regularization, kernel density estimation

1. Introduction

The motivating hypothesis of multi-task learning (MTL) is that combining data from related estimation tasks can yield superior estimates over learning each task separately. A central question of MTL is: how should the tasks be combined or tied together? Consider the following visualization of a T-task MTL estimation problem:

$$\begin{tikzcd}
\mathcal{X}_1 \arrow[r, "f_1(x_{1i};\beta_1)"] & \mathcal{Y}_1 \\
\vdots \quad \vdots & \vdots \\
\mathcal{X}_t \arrow[r, "f_t(x_{ti};\beta_t)"] & \mathcal{Y}_t \\
\vdots \quad \vdots & \vdots \\
\mathcal{X}_T \arrow[r, "f_T(x_{Ti};\beta_T)"] & \mathcal{Y}_T,
\end{tikzcd}$$

where $\mathcal{X}_t$ is the tth tasks' input space, $\mathcal{Y}_t$ is the tth tasks' output space, and $f_t(x_{ti}, \beta_t)$ is the tth tasks' function that maps input $x_{ti} \in \mathcal{X}_t$ to output $y_{ti} \in \mathcal{Y}_t$ and parameterized by set β_t. From the above figure, one observes that the tasks can be tied together through

(a) the input spaces $\{\mathcal{X}_1, \dots, \mathcal{X}_T\}$;

(b) the parameter or function spaces, and/or,

(c) the output spaces $\{\mathcal{Y}_1, \dots, \mathcal{Y}_T\}$.

Prior MTL approaches have focused on (a) and (b), assuming that tasks are related in that they all share an input space (i.e. $\mathcal{X}_t = \mathcal{X}$ for all t) and/or function class (i.e $f_t(x; \beta_t) = f(x; \beta_t)$ for all t). Some examples:

- Joint dimensionality reduction/selection with respect to all tasks at once (input space). E.g. Obozinski et al. (2010).
- Jointly regularizing the function parameters β_t (parameter space). E.g. Argyriou et al. (2007).
- Placing joint priors on the β_t for all t (parameter space). E.g. Bakker and Heskes (2003).
- Defining multi-task kernels (implicit input space). E.g. Micchelli and Pontil (2004).

The possibility of (c) has been largely neglected; however, in this work we argue that it is a natural way to think about the relatedness of tasks. For example, two tasks that we believe should benefit from multi-task learning are "How long do trees live in the vicinity of a nuclear meltdown?" and "How long do humans live in the vicinity of a nuclear meltdown?" These two tasks share an output space (and are likely correlated), but the tasks' domains (trees and humans) may be described by unrelated features. In this paper we investigate cross-task regularization of the task-specific functions' outputs, a paradigm of multi-task learning that we call *output space regularization* (OSR). We show that OSR is both powerful and widely applicable. OSR enables MTL for tasks that do not share an input space or a function class; the tasks need only be related in that they aim to estimate related outputs. The task-specific input spaces and task-specific functions can be completely different.

To elucidate the difference between standard MTL parametric regularization approaches and OSR, consider the following two example MTL objectives. Let (x_{ti}, y_{ti}) be input-output training pairs for the tth task, with N_t pairs per task. Let $A \in \mathbb{R}^{T \times T}$ be the task similarity matrix, where A_{rs} is the similarity between tasks r and s. The following MTL objective explicitly ties the tasks together in the parameter space:

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} (y_{ti} - f(x_{ti}; \beta_t))^2 + \gamma \sum_{r=1}^T \sum_{s=1}^T A_{rs} \|\beta_r - \beta_s\|_2^2. \tag{1}$$

This regularization results in parameter estimates that are more similar. Consider on the other hand the following OSR objective

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} (y_{ti} - f_t(x_{ti}; \beta_t))^2 + \gamma \sum_{r=1}^T \sum_{s=1}^T \sum_{j=1}^{N_r} \sum_{k=1}^{N_s} \mathbb{A}_{rj;sk} (f_r(x_{rj}; \beta_r) - f_s(x_{sk}; \beta_s))^2, \tag{2}$$

where $\mathbb{A}_{rj;sk}$ is the task-and-sample similarity between x_{rj} and x_{sk}. The non-parametric form of (2) regularizes the outputs $\{f(x_{ti}; \beta_t)\}$ of the different tasks, resulting in parameter estimates that correspond to output estimates that are more similar.

In this paper, we will show that OSR is a flexible, widely applicable, and effective method with which to approach multi-task learning problems. In the next section we give the general form of OSR, after which we review related work in multi-task learning, as well as discuss how OSR is related to manifold regularization (Belkin et al., 2006) in Section 3.

Through the lens of OSR, many estimation problems can be cast in a MTL framework, and potentially benefit from additional data sharing. In Section 4, we investigate the simplest application of OSR, which we term*multi-task averaging* (MTA). The tractability of MTA enables an extensive analysis, given in Section 5, which provides some new intuition and insights about OSR in general. MTA is broadly applicable; we illustrate its use for multi-task kernel density estimation (MT-KDE) in Section 6, and for specifying the empirical moment constraints in similarity discriminant analysis in Section 7. Then, in Section 8, we present a second formulation of OSR - *multi-task linear regression.* We interpret local linear regression as a multi-task estimation problem, and show that OSR local linear regression produces a smoother function over the domain of printer color management look-up-tables. We note that, with OSR, any local learning problem can be formulated as a MTL problem. The paper ends with conclusions and open questions.

2. Output Space Regularization

In multi-task learning, we are interested in learning T functions $f_t(x_{ti}; \beta_t)$, where β_t is a set of parameters for the tth task, and x_{ti} is the ith input from task t. Suppose one is given N_t training input-output pairs $\{(x_{ti}, y_{ti})\}_{i=1}^{N_t}$ for the tth task for $t = 1, \ldots, T$. Note that x and y can be any objects, as long as the function outputs are comparable with y. Let L be a loss function, J be a regularization function, and γ be a scalar regularization factor. Additive regularization in multi-task learning is predominantly parametric, that is, it can be formulated as

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti}; \beta_t)) + \gamma J(\{\beta_r\}_{r=1}^T). \tag{3}$$

In this paper we focus on the alternative approach where the regularization is a function of the output values, such that the objective function can be written

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{f_r(x_{rj};\beta_r)\}_{j=1}^{N_r}), \text{ for } r \in \{1,\dots,T\}. \tag{4}$$

OSR is complementary to many parametric multi-task regularization strategies, and can be used in conjunction with them. Also, OSR is easy to generalize to transductive or semi-supervised learning by regularizing function outputs for unlabeled training data and/or test samples. In such cases, let $\{z_{rj}\}_{r=1}^{M_r}$ be a set of samples that includes some combination or subset of the training samples, test samples (transductive), and unlabeled training samples (semi-supervised), where M_r is the total number of samples from the rth tasks used in the regularizer. Then (4) generalizes to

$$\arg\min_{\{\beta_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} L(y_{ti}, f_t(x_{ti};\beta_t)) + \gamma J(\{f_r(z_{rj};\beta_r)\}_{j=1}^{M_r}), \text{ for } r \in \{1,\dots,T\}.$$

All regularization functions J used in this paper require information about the relatedness of multiple tasks which we specify as a similarity matrix between tasks and sometimes samples.

3. Related Work in Multi-Task Learning

A related multi-task approach that focuses on function outputs is the work of Bonilla et al. (2008), where a Gaussian process prior is placed over the latent task functions to constrain their inner products. This method is limited only to Gaussian process regression.

Next, we review other approaches to tying together tasks in various machine learning problems.

One of the dominant approaches is the use of an explicit multi-task parameter regularizer. Evgeniou and Pontil (2004), for instance, penalized the distance of model parameters β_t to $\frac{1}{T}\sum_r \beta_r$, the average of all tasks' model parameters. Argyriou et al. (2007) studied regularizers in which the spectral norm of the matrix β (the tth column is β_t) is penalized. Kato et al. (2008) formulated a multi-task SVM by imposing a constraint on the difference between task parameters.

Also common are Bayesian frameworks in which shared statistical structures of the parameters are learned or imposed. In the work of Bakker and Heskes (2003) some of the model parameters are shared amongst tasks, while others are connected through a joint prior. Xue et al. (2007) drew the β_t from a Dirichlet process prior, and Liu et al. (2009) placed a joint distribution over the tasks' parameters in a semi-supervised framework.

Another approach is the construction of MT kernels. Micchelli and Pontil (2004) explored the linear combination of kernels for multi-task learning, and Evgeniou et al. (2005) showed that the problem of estimating T task functions with certain types of regularization can be cast as a single-task problem using multi-task kernels. Sheldon (2008) built on this work by proposing a graphical multi-task learning framework using graph Laplacians.

Other work focuses on tying tasks together by jointly selecting features or learning subspaces. Argyriou et al. (2008) use a $(2,1)$-norm on the matrix β to encourage a small-number of non-zero rows. Obozinski et al. (2010) build on this work to obtain a computationally efficient joint subspace selection.

Another major question in multi-task learning is how to estimate the similarity (or task relatedness) between tasks and/or samples if it is not provided. The standard approach, taken by many of the above-cited papers, is to estimate the similarity matrix jointly with the task parameters (Zhang and Yeung, 2010; Argyriou et al., 2007; Bonilla et al., 2008; Xue et al., 2007). As a more detailed example, Zhang and Yeung (2010) assumed that there exists a covariance matrix for the task relatedness, and proposed a convex optimization approach to estimate the task covariance matrix and the task parameters in an alternating way.

Our work does not focus on the question of similarity matrix estimation, and, for all of our experiments, the task similarity matrix was either provided by domain experts or specified from domain knowledge. It is straightforward to modify any of the OSR formulations in this paper so that the similarity matrix between the tasks and/or samples is learned in conjunction with the task parameters.

OSR is similar to *manifold regularization* (Belkin et al., 2006) in that the function outputs are tied together, instead of the function parameters. One can view manifold regularization (without the RKHS function regularizer) as a special 1-task case of OSR. For example, Belkin et al.'s Laplacian-regularized least squares objective for semi-supervised regression solves

$$\arg\min_{f\in\mathcal{H}} \quad \textstyle\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda||f||_{\mathcal{H}}^2 + \gamma\sum_{i,j=1}^{N+M} A_{ij}^F(f(x_i) - f(x_j))^2,$$

where f is the regression function to be estimated, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), N is the number of labeled training samples, M is the number of unlabeled training samples, A_{ij}^F is the similarity (or weight in an adjacency graph) between feature samples x_i and x_j, and $||f||_{\mathcal{H}}$ is the norm of the function f in the RKHS. In OSR, as opposed to manifold regularization, we are estimating a separate function for each of the T tasks. Given that OSR allows for T unrelated input spaces and functions, it is an open question whether the mathematics that describe the manifold in Belkin et al.'s work still hold.

A different approach to using manifold regularization concepts for MTL was recently proposed by Agarwal et al. (2010) for parametric multi-task learning. They assume that the task *parameters* lie on a low-dimensional manifold, and alternate estimating the manifold with learning the tasks in an iterative way. Their approach is restricted to a RKHS and assumes a global manifold structure for all tasks, which share a common parametric model.

4. MTA Formulation

In this section, we apply OSR to the problem of estimating multiple means, yielding what we call *multi-task averaging* (MTA). At first glance, MTA may strike one as a trivial formulation of OSR. We have found, however, that it leads to fruitful analysis, and is surprisingly applicable. One useful application of MTA is the novel *multi-task kernel density estimation*, which we propose in Section 6. Also, we found that MTA alleviates the problem of high variance in similarity discriminant analysis, as we show in Section 7.

With squared-loss, MTA has a closed-form solution, and we are able to state some analytic results in the next section that, while important in their own right, also provide intuition for more complicated problems. We give conditions for which the MTA estimates will be better than the task sample averages. We explore the key question, "How should the task similarities be defined?" We show the intuitive result that for the simplest two-task case the optimal similarity between the two tasks depends on the ratio of the task variance to the task mean differences. However, we show that for more complicated cases, the optimal similarity behaves unintuitively. Our main theorem is that the MTA estimates are a *convex* combination of the individual tasks' samples averages. We discuss the Bayesian viewpoint on MTA in Section 5.6. Key notation is given in Table 1.

4.1 MTA Objective

Consider the T-task problem of estimating the means of T random variables. To obtain the MTA objective from OSR, we simply replace f_t with $\hat{y}_t$ in (2) We assume that for the tth random variable one is given N_t samples $\{y_{ti}\}_{i=1}^{N_t}$, where each $y_{ti} \in \mathbb{R}$ is an IID draw from the task-specific random variable Y_t. In addition, we assume the $T \times T$ matrix A describes the relatedness or similarity of any pair of the T tasks, with $A_{tt} = 0$ for all t. The proposed MTA estimates are

$$\{y_t^*\}_{t=1}^T = \operatorname*{arg\,min}_{\{\hat{y}_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} (y_{ti} - \hat{y}_t)^2 + \gamma \sum_{r=1}^T \sum_{s=1}^T A_{rs}(\hat{y}_r - \hat{y}_s)^2. \tag{5}$$

The first term minimizes the empirical loss, and the second term jointly regularizes the estimates. The regularization parameter γ trades-off the empirical risk term versus smoothing the task estimates together. Note that if $\gamma = 0$, (5) produces the T sample averages. The MTA problem can be interpreted as learning a constant function from noisy samples where the "parameter" β and the function output are the same thing. In other words, for the MTA case, output space regularization and function space regularization coincide.

A more general formulation of MTA is

$$\{y_t^*\}_{t=1}^T = \operatorname*{arg\,min}_{\{\hat{y}_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} L(y_{ti}, \hat{y}_t) + \gamma J\left(\{\hat{y}_t\}_{t=1}^T\right),$$

where L is some loss function and J is a regularization function. If L is chosen to be any Bregman loss, then setting $\gamma = 0$ will produce the T sample averages. For the analysis and experiments in this paper, we restrict our attention to the squared error formulation given in (5).

4.2 Closed-form Solution for the Scalar Case

We provide a closed-form solution for (5). Let $\tilde{y} \in \mathbb{R}^T$ and $\tilde{A} \in \mathbb{R}^{T \times T}$ have components

$$\tilde{y}_t = \frac{\sum_{i=1}^{N_t} y_{ti}}{N_t + \gamma \sum_{s \neq t}(A_{ts} + A_{st})} \tag{6}$$

$$\tilde{A}_{tr} = \frac{\gamma(A_{tr} + A_{rt})}{N_t + \gamma \sum_{s \neq t}(A_{ts} + A_{st})}. \tag{7}$$

Table 1: Key Notation

y_{ti}	ith sample from tth task
y_t^*	Multi-task averaging estimate for tth task
$y_t^\dagger$	Multi-task regularized sample average for tth task
$\bar{y}_t$	Sample average for tth task
$\hat{y}_t$	An estimate of the mean of tth task
$\mathbb{Y}^*$	Matrix with columns y_t^*

The closed-form solution given below in (8) will hold for any A such that $I - \tilde{A}$ is invertible. The following lemma shows that, for invertibility of $I - \tilde{A}$, it is sufficient (though not necessary) for A to have non-negative entries.

Lemma: Suppose all the entries of matrix A are non-negative and for all tasks we have at least one sample, that is, $N_t \geq 1$ for all $t \in \{1, \ldots, T\}$. Then matrix $I - \tilde{A}$ is invertible. In addition, for the $T = 2$ case, $I - \tilde{A}$ is invertible for all A, except when $a = \frac{-N_1 N_2}{2(N_1+N_2)\gamma}$.

Proof: See Appendix A.

If $(I - \tilde{A})$ is invertible, then it can be shown that the MTA objective given in (5) is convex and has the closed-form solution

$$y^* = (I - \tilde{A})^{-1}\tilde{y}, \tag{8}$$

where I is the $T \times T$ identity matrix.

Should one find that $I - \tilde{A}$ is not invertible (if e.g. some entries of A are negative), that problem can be solved by a slight perturbation of the matrix A, or by optimizing (5) directly.

4.3 Closed-form Solution for the Vector Case

MTA can also be applied to vectors, with a closed-form solution. For this section only, let $y_{ti} \in \mathbb{R}^d$ and $\hat{y}_t \in \mathbb{R}^d$ be vectors. The MTA objective in (5) generalizes to:

$$\{y_t^*\}_{t=1}^T = \operatorname*{arg\,min}_{\{\hat{y}_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} (y_{ti} - \hat{y}_t)^T (y_{ti} - \hat{y}_t) + \gamma \sum_{r=1}^T \sum_{s=1}^T A_{rs} (\hat{y}_r - \hat{y}_s)^T (\hat{y}_r - \hat{y}_s). \tag{9}$$

Using the fact that $\frac{\partial}{\partial b}(a-b)^T(a-b) = -2(a-b)$, we can obtain a closed-form solution for (9). The definition of $\tilde{y}_t$ is the same as in (6), except with vectors replacing scalars. The definition of $\tilde{A}$ is unchanged from the scalar case. Let $\mathbb{Y}^* \in \mathbb{R}^{d \times T}$ be a matrix with y_t^* as its tth column and let $\tilde{\mathbb{Y}} \in \mathbb{R}^{d \times T}$ be a matrix with $\tilde{y}_t$ as its tth column. Then (9) is solved by

$$\mathbb{Y}^* = \tilde{\mathbb{Y}}(I - \tilde{A})^{-1},$$

which generalizes the scalar solution (8) if column vectors are changed to row vectors and vice-versa.

5. Analysis of MTA

In this section we present some analysis of MTA. Throughout, $\mu \in \mathbb{R}^T$ denotes the $T \times 1$ vector of means for the T tasks, assumed to be finite, and capital Y's will be used to denote random variables. We take as given N_t IID random variables $\{Y_{ti}\}$ for the tth task with $i = 1, \ldots, N_t$. We compare with the vector of sample averages $\bar{Y} \in \mathbb{R}^T$, where

$$\bar{Y}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_{ti}.$$

We do not assume any parametric form for the underlying sample distributions, but note that for many common parametric assumptions (such as Gaussian or Laplacian), the sample average is the maximum likelihood estimate.

We first present in-depth analysis of the $T = 2$ case, starting with a special symmetric case, and then treating the general $T = 2$ case. We give conditions when the MTA estimates are better than the sample averages, provide a formula for the optimal task-similarity matrix A, and show the optimal A sometimes behaves unintuitively. We also prove for $T = 2$ that MTA has a smaller MSE than multi-task regularizing of the sample averages; we hypothesize that this is true in general.

Then we turn to the general case for T tasks. A surprising result is that, under certain conditions, the MTA estimates are convex combinations of the T sample averages. Lastly, we note that MTA is asymptotically unbiased.

5.1 Analysis for Symmetric $T = 2$ Case

We first analyze the simplest case of $T = 2$ symmetric tasks, with N samples per task. Specifically, suppose Y_1 is distributed with finite mean μ_1 and finite variance σ^2, and Y_2 is distributed with finite mean $\mu_2 = \mu_1 + \Delta$ and finite variance σ^2. Let the task-relatedness matrix be $A = [0\ a; a\ 0]$. For a fixed A, the regularization parameter γ in (5) is a useful degree of freedom to control the power of the regularizer. But for this analysis we treat A as a variable (that is, we treat a as a variable), and thus the γ is an unnecessary degree of freedom, so without loss of generality we set $\gamma = 1/2$. For this simple choice of A, the matrix inversion in the closed-form solution (8) can be solved analytically, producing the following properties.

Regularized Estimate: The MTA estimate is a convex combination of the sample averages (for $a \geq 0$):

$$Y_1^* = \left(\frac{N + a}{N + 2a}\right) \bar{Y}_1 + \left(\frac{a}{N + 2a}\right) \bar{Y}_2. \tag{10}$$

Bias: From (10) it follows that the MTA estimate is biased:

$$\mathrm{E}[Y_1^*] - \mu_1 = \left(\frac{a}{N + 2a}\right) \Delta. \tag{11}$$

Variance: From (10) it follows that the MTA estimate has a smaller variance than the sample average (for $a > 0$):

$$\mathrm{Var}[Y_1^*] = \frac{\sigma^2}{N}\left(\frac{N^2 + 2aN + 2a^2}{N^2 + 4aN + 4a^2}\right) < \mathrm{Var}[\bar{Y}_1]. \tag{12}$$

Mean-Squared Error: From (11) and (12), the mean-squared error for estimating only μ_1 (or only μ_2, because the two tasks are completely symmetric for this case) is

$$\mathrm{MSE}[Y_1^*] = \left(\frac{\sigma^2}{N}\right)\frac{N^2 + 2aN + 2a^2}{(N+2a)^2} + \Delta^2\frac{a^2}{(N+2a)^2}. \tag{13}$$

Comparing to $\mathrm{MSE}[\bar{Y}_1] = \sigma^2/N$, one can conclude from (13) that

$$\mathrm{MSE}[Y_1^*] < \mathrm{MSE}[\bar{Y}_1] \text{ if } \Delta^2 < 2\sigma^2\left(\frac{1}{N} + \frac{1}{a}\right) \tag{14}$$

Thus the MTA estimate has lower MSE if the mean-separation Δ is small compared to the variance weighted by a factor dependent on the similarity and the number of samples. Note that as a approaches 0 from above, the RHS of (14) approaches infinity. Thus a small amount of regularization can be helpful even when the mean difference Δ is large. We illustrate this relationship in Fig. 1. Note further that the fact $a > 0$ was not used in derivation of (14).

Optimal Task Relatedness Information: The MSE given in (13) is a convex function of a, thus the first derivative can be used to specify the optimal value for a. This gives in the surprisingly simple result that the $\mathrm{MSE}[Y_1^*]$ is minimized by the task-similarity value

$$a^* = \frac{\sigma^2}{\Delta^2}. \tag{15}$$

This result is key because it specifies that the task-similarity a ideally should measure the variance of the task samples relative to the task mean-difference. In fact, a^* is the inverse of the squared Mahalanobis distance between μ_1 and μ_2. Further, the ideal task-similarity is independent of the number of samples N. Intuitively, in the limit case that the difference in the task means $\Delta \to 0$, the optimal task-relatedness $a^* \to \infty$, and the weights in (10) on $\bar{Y}_1$ and $\bar{Y}_2$ become $1/2$ each. We illustrate the effect of choice of a and the optimal a^* in Fig. 2.

5.2 Analysis for General $T = 2$ Case

Next we generalize the above analysis to the general case for two tasks. Consider the same set-up as Section 5.1, except let there be N_1 samples of Y_1 and N_2 samples of Y_2, and let the corresponding variances be σ_1^2 and σ_2^2. For shorthand, let

$$z = N_1N_2 + aN_1 + aN_2.$$

The MTA estimate is again a convex combination of the sample averages (for $a \geq 0$):

$$Y_1^* = \frac{N_1(N_2 + a)}{z}\bar{Y}_1 + \frac{N_2 a}{z}\bar{Y}_2,$$

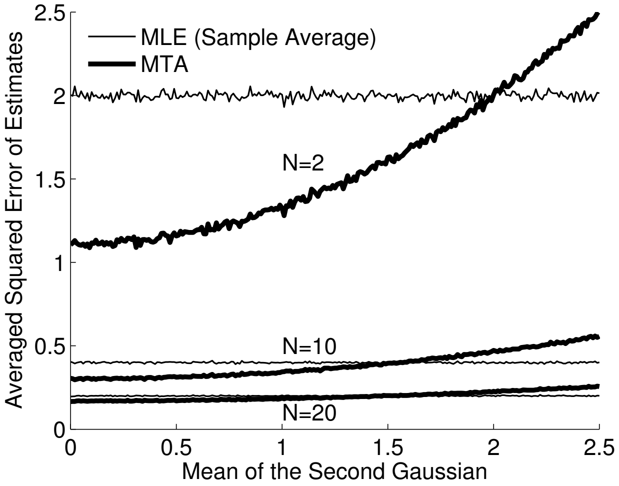


Figure 1: Plot shows squared error summed for two tasks, averaged over 10,000 runs of the simulation. For each task there are N IID samples, for $N = 2, 10, 20$. One task generates samples from a standard Gaussian. The other task generates samples from a Gaussian with $\sigma^2 = 1$ and varying mean value, as marked on the x-axis. The symmetric task-relatedness value was fixed at $a = 1$ (note this is generally not the optimal value, see Fig. 2). One sees that given only $N = 2$ samples from each Gaussian, the MTA estimate is better if the Gaussians are closer than 2 units apart. Given $N = 20$ samples from each Gaussian, the MTA estimate is better if the Gaussians are closer than 1.5 units apart. In the extreme case that the two Gaussians have the same mean ($\mu_1 = \mu_2 = 0$), then with this suboptimal choice of $a = 1$, MTA provides a 45% win for $N = 2$ samples, and a 15% win for $N = 20$ samples.

and the MSE is

$$\mathrm{MSE}[Y_1^*] = \left(\frac{a^2 N_2^2}{z^2}\right)\Delta^2 + \frac{\sigma_1^2 N_1 (N_2 + a)^2 + \sigma_2^2 N_2 a^2}{z^2}. \tag{16}$$

Once again we compare to $\mathrm{MSE}[\bar{Y}_1]$ and get

$$\mathrm{MSE}[Y_1^*] < \mathrm{MSE}[\bar{Y}_1] \text{ if } \Delta^2 < \frac{2\sigma_1^2}{a} + \frac{2\sigma_1^2}{N_2} + \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2}.$$

We illustrate the MSE given in (16) in Fig. 3.

The optimal task-relatedness value a depends on whether one's goal is: *(i)* to minimize the MSE only on the first task regardless of error on the second task; or *(ii)* to minimize the summed MSE over both tasks. For Case *(i)*, analyzing the first derivative of the MSE in (16) shows that the MSE for the first task is minimized by

$$a_1^* = \frac{\sigma_1^2}{\Delta^2 - \frac{\sigma_1^2 + \sigma_2^2}{N_2}}, \tag{17}$$

with swapped indices to compute a_2^*. In this case the MSE is not convex, but an analysis of the second derivative shows that (17) holds in all cases except if the number of samples from the second task $N_2 = (\sigma_1^2 - \sigma_2^2)/\Delta^2$.

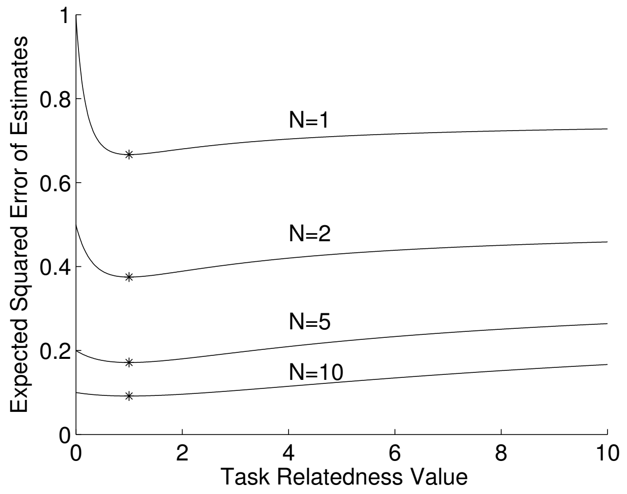


Figure 2: Plot shows the expected squared error summed for two tasks as given in (13), where the task samples were drawn IID from Gaussians $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$. The task-relatedness value a was varied as shown on the x-axis. The minimum expected squared error is marked by a $*$, is independent of N and matches the optimal task-relatedness value given by (15). Given $N = 1$ samples from each task, using the optimal task-relatedness a for this set-up rather than $a = 0$ results in a 1/3 reduction in expected squared error; given $N = 10$ samples from each task, using the optimal a results in a 9% reduction in error.

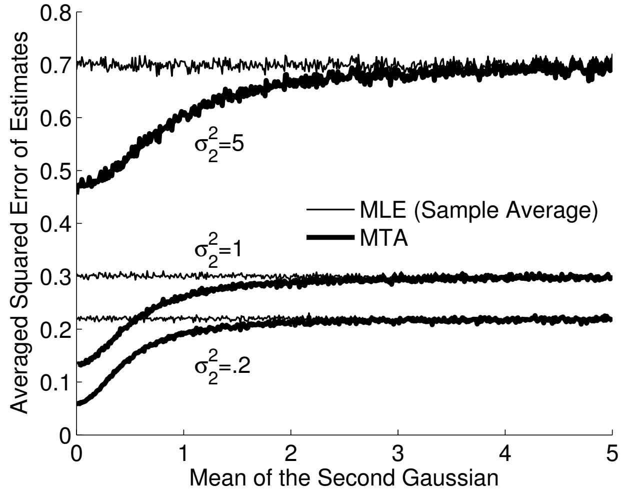


Figure 3: Plot shows the average over 10,000 simulation runs of the sum of the two squared errors for two asymmetric tasks. For all the results shown, $N_1 = 5$ samples are drawn from a standard normal, and $N_2 = 10$ samples are drawn from a second Gaussian whose mean was varied and plotted on the x-axis, and whose variance was either $\sigma_2^2 = .2, 1, 5$, corresponding to the three sets of lines shown. The similarity for each simulation was set to the optimal similarity as given by (18).

For Case *(ii)*, analyzing the first derivative of the sum MSE shows that the sum MSE is minimized by

$$a^* = \frac{N_2^2\sigma_1^2 + N_1^2\sigma_2^2}{\Delta^2(N_1^2 + N_2^2) + (\sigma_1^2 - \sigma_2^2)(N_1 - N_2)}. \tag{18}$$

Analysis of the second derivative shows that this minimizer always hold for the cases of interest (that is, for $N_1, N_2 \geq 1$).

A key consequence of (18) is that the optimal similarity a^* is higher if there are more samples from the lower variance task, that is, if $N_1 > N_2$, then a smaller σ_1^2 leads to a higher a^*. Fig. 3 shows results with the optimal similarity given by (18). In Fig. 3 task two has more samples, and when its variance is $\sigma_2^2 = .2$, the advantage of MTA is proportionally greater than when the task two variance is $\sigma_2^2 = 5$. The dominant effect in such cases is that MTA greatly decreases the error for the higher-variance less-sampled task, sometimes at the cost of a small increase in error of the more-sampled class. Even lower error can be achieved by using the optimal similarity for each task from (17) rather than using one similarity optimized for both tasks. We compare these two strategies in Fig. 4, which shows that a moderate gain is possible if one has the knowledge to optimally set the task-relatedness independently for two tasks.

Surprisingly, while the MTA MSE varies smoothly as a function of any of its parameters, the optimal task-relatedness value a^* can show large jumps and can be negative. To illustrate the unintuitiveness of (18), we plot the optimal similarity given in (18) for two slightly differing parameter sets in Fig. 5.

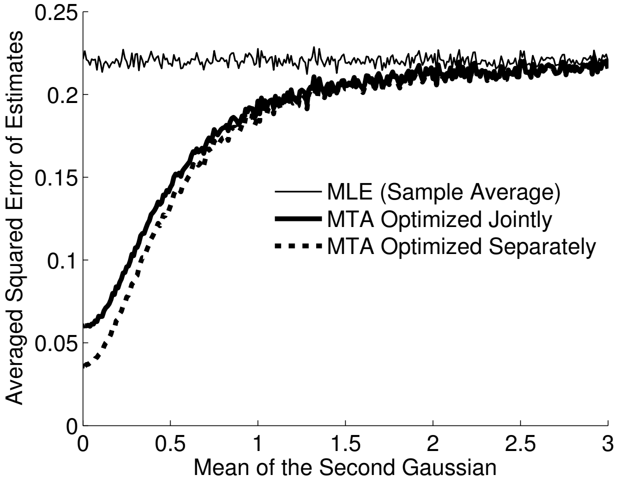


Figure 4: Plot shows the average over 10,000 simulation runs of the sum of the squared errors for two asymmetric tasks where the optimal similarity was set the tasks jointly using (18), or set for each task separately using (17). Other parameters were set to: $N_1 = 5$ samples are drawn from a standard normal, and $N_2 = 10$ samples are drawn from a second Gaussian whose mean was varied and plotted on the x-axis, and whose variance was $\sigma_2^2 = .2$.

5.3 Estimating the Optimal Task Similarity

Examining the formulas for the optimal task relatedness value a^* in Section 5, one observes that the similarity matrix A should reflect the relatedness of the underlying tasks' *statistics* and not necessarily their *semantic relatedness*. This observation predates MTL by a few decades; Efron and Morris (1977) noted that shrinking baseball statistics towards imported

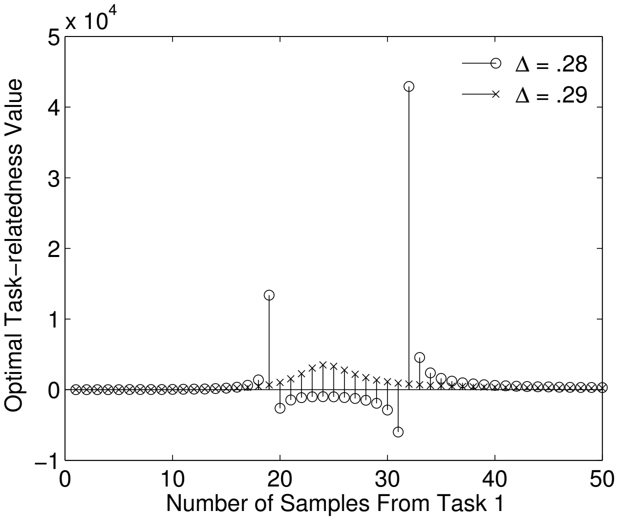


Figure 5: Plot shows the optimal similarity for two tasks as given by (18) as a function of the number of samples N_1 drawn from a standard normal. The other parameters were fixed at $N_2 = 10$, $\sigma_2^2 = 5$, and two choices of mean separation are shown, $\Delta = .28$ and $.29$.

car prevalence data can decrease estimation error, though few would call these two tasks semantically similar.

In practice, however, we expect that the given task-similarity matrix A encodes domain knowledge reflecting semantic similarity and is not (necessarily) statistically optimal. We illustrate in our experiments that such task relatedness information is often available, and advantageous to incorporate into the estimation. If no task relatedness information is provided, it may be necessary to use the above formulas to estimate the optimal task similarity matrix. If, on the other hand, an expert-generated A is available, it may be of further benefit to combine it with an estimate of the theoretically optimal A. For example, based on (15), we hypothesize that it may be effective to estimate σ as the average standard deviation of each task's samples, and estimate Δ as the difference of the sample means. This data-dependent approach is analogous to empirical Bayesian methods in which prior parameters are estimated from data (Casella, 1985).

5.4 MTA Estimates Are Convex Combinations of Sample Averages

From (8) and (6) it follows that the MTA estimates are *linear* combinations of the task-specific sample averages $\{\bar{y}_t\}$. However, as the special case of $T = 2$ above suggests, we can make an even stronger statement (with the addition of some constraints): the MTA estimates are *convex* combinations of the T sample averages. An important consequence of this theorem is that the MTA estimates preserve convexity constraints; that is, if the samples from all T tasks are from a convex set C, then the theorem guarantees that the MTA estimates will also be from the convex set C.

Theorem: Suppose all the entries in matrix A are non-negative. Then if $N_t \geq 1$ for all t, the MTA estimates are convex combinations of the sample averages.

Proof: The proof is given in Appendix B.

5.5 Asymptotic Analysis

The (q, r) entry of matrix $\tilde{A}$ given by (7) goes to 0 as N_t goes to infinity. This means that the whole matrix $\tilde{A}$ converges to 0 in norm and since matrix inversion is a continuous operation, $(I - \tilde{A})^{-1} \to I$ in norm. Because $\tilde{Y}_t - \bar{Y}_t \to 0$ as $N_t \to \infty$, by the law of large numbers we can conclude that Y^* asymptotically approaches the true mean μ.

5.6 Bayesian Interpretation of MTA

In this section, we consider MTA in a Bayesian framework, and discuss related work in James-Stein and empirical Bayesian estimation.

The MTA objective given in (5) is a structural risk minimization. Many structural risk minimization problems can be interpreted as maximum a posteriori (MAP) estimates. For example, consider the ridge regularized analogue to (5):

$$\begin{aligned} \arg\min_{\{\hat{y}_t\}_{t=1}^T} \sum_{t=1}^T \sum_{i=1}^{N_t} (y_{ti} - \hat{y}_t)^2 + \gamma \sum_{r=1}^T \hat{y}_r^2 &= \arg\max_{\{\hat{y}_t\}_{t=1}^T} p(\{y_{ti}\}|\hat{y}) p(\hat{y}) \\ &= \arg\max_{\{\hat{y}_t\}_{t=1}^T} p(\hat{y}|\{y_{ti}\}), \end{aligned} \tag{19}$$

where the likelihood $p(\{y_i\}|\hat{y})$ is the normal distribution $\mathcal{N}(\hat{y}, I)$ and the prior $p(\hat{y})$ is the normal distribution $\mathcal{N}(0, I/\gamma)$.

The Bayesian interpretation of (5) has the same form as (19), where again the likelihood $p(\{y_i\}|\hat{y})$ is the normal distribution $\mathcal{N}(\hat{y}, I)$, but $p(\hat{y})$ is now the improper prior

$$\begin{aligned} p(\hat{y}) &\propto e^{-\gamma \sum_{r=1}^T \sum_{s=1}^T A_{rs}(\hat{y}_r - \hat{y}_s)^2} \\ &= \prod_{r=1}^T \prod_{s=1}^T e^{-\gamma A_{rs}(\hat{y}_r - \hat{y}_s)^2}. \end{aligned} \tag{20}$$

The improper prior (20) is a product of Gaussian priors on the differences. Note that in general this does not imply that the differences are independent. Consider for example the case of $T = 3$ tasks. Then, assuming the differences are independent, (20) becomes

$$\hat{Y}_1 - \hat{Y}_2 \sim \mathcal{N}(0, 1/2\gamma A_{12}) \tag{21}$$

$$\hat{Y}_2 - \hat{Y}_3 \sim \mathcal{N}(0, 1/2\gamma A_{23}) \tag{22}$$

$$\hat{Y}_1 - \hat{Y}_3 \sim \mathcal{N}(0, 1/2\gamma A_{13}), \tag{23}$$

and note

$$\hat{Y}_1 - \hat{Y}_3 = (\hat{Y}_1 - \hat{Y}_2) + (\hat{Y}_2 - \hat{Y}_3). \tag{24}$$

Here, since the differences are independent, the variances on the right-hand side of (24) must add, and (21)-(23) would imply that

$$\frac{1}{A_{13}} = \frac{1}{A_{12}} + \frac{1}{A_{23}} \quad \text{and} \quad \frac{1}{A_{12}} = \frac{1}{A_{13}} + \frac{1}{A_{23}} \quad \text{and} \quad \frac{1}{A_{23}} = \frac{1}{A_{12}} + \frac{1}{A_{13}}. \tag{25}$$

The three constraints in (25) cannot be satisfied by any positive A, and for any $T > 2$ the analogous set of $T(T-1)/2$ constraints will also be impossible to satisfy.

The proposed MTA estimator in (5) differs from the related work in standard Bayesian and empirical Bayesian multi-task estimators. For example, a standard Bayesian multi-task estimator would tie the tasks together through a prior on the unknown means to be estimated, e.g.,

$$Y_{ti} \sim \mathcal{N}(\hat{Y}_t, \sigma^2) \tag{26}$$

$$\hat{Y}_t \sim \mathcal{N}(z, \tau^2), \tag{27}$$

where the prior parameters z and τ and the sample deviation σ^2 would be assumed known. This Gaussian model results in the estimate

$$\hat{y}_t = \frac{\sigma^2}{\sigma^2 + \tau_2} z + \frac{\tau^2}{\sigma^2 + \tau_2} \hat{\mu}_t,$$

such that the tying between the tasks all occurs through the (assumed) known prior parameters.

Empirical Bayesian estimators use standard Bayesian models as in (26) and (27), but differ in that all the necessary parameters are estimated from the data rather than assumed (Casella, 1985). Analogously, we could estimate the matrix A from data; this is discussed further in Section 5.3.

The James-Stein estimator (Casella, 1985) makes the assumption (26) with a known σ^2, then estimates

$$\hat{y} = \left(1 - \frac{(T-2)\sigma^2}{\|\hat{\mu}\|^2}\right) \hat{\mu}.$$

Thus, the James-Stein estimate is significantly different from the proposed estimator, and as with many empirical Bayesian multi-task estimates, does not jointly estimate the tasks for small $T = 2$.

In summary, the proposed MTA estimator differs from the related work in Bayesian and Stein estimation in what information is assumed given (or to be estimated), and in the assumed statistical model.

5.7 Comparison of MTA and Multi-Task Regularized Sample Averages

Consider a variant of MTA in which one first obtains the sample mean $\bar{y}_t$ for each task, and then applies (5) as follows:

$$\{y_t^\dagger\}_{t=1}^T = \arg\min_{\{\hat{y}_t\}_{t=1}^T} \sum_{t=1}^T (\bar{y}_t - \hat{y}_t)^2 + \gamma \sum_{r=1}^T \sum_{s=1}^T A_{rs} (\hat{y}_r - \hat{y}_s)^2. \tag{28}$$

We refer to the resulting estimates $\{y_t^\dagger\}$ as *multi-task regularized sample averages*. This variant is related to domain adaptation methods and transfer learning (Daumé III and Marcu, 2006), which compute some estimates for some tasks, and then regularize estimates for new tasks to the previous tasks' estimates. Using similar analysis as in Section (5.1), it is straightforward to show that

$$\mathrm{MSE}[Y^\dagger] = \frac{\sigma^2(1+2a+2a^2)}{N(1+2a)^2} + \Delta^2\frac{a^2}{(1+2a)^2}. \tag{29}$$

To gain some insight into when MTA performs better than multi-task regularized sample averages, we examine the difference between (29) and (13). After some algebraic manipulations, this difference can be written as:

$$\mathrm{MSE}[Y^\dagger] - \mathrm{MSE}[Y^*] = \frac{a(N-1)\left(a\Delta^2 N(4a+N+1) - 2\sigma^2(aN+a+N)\right)}{N(1+2a)^2(N+2a)^2}. \tag{30}$$

When the difference in (30) is positive, MTA performs better than the multi-task regularized sample averages. This occurs when

$$\frac{\Delta^2}{2\sigma^2} > \frac{1+\frac{1}{a}+\frac{1}{N}}{N+4a+\frac{1}{N}},$$

which is true for large values of N and/or a, and when the Mahalanobis distance between the two tasks is large.

6. Multi-Task Kernel Density Estimation

In this section we present *multi-task kernel density estimation* (MT-KDE), as an application of OSR.

6.1 MT-KDE Objective

Recall that for standard kernel density estimation (KDE) (Silverman, 1986), a set of random samples $x_i \in \Omega, i \in \{1,\ldots,N\}$ are assumed to be drawn IID from an unknown distribution p_X over some set Ω, and the problem is to estimate the density over Ω. Given a kernel function $K : \Omega \times \Omega \to \mathbb{R}$, the standard KDE estimate for any $z \in \Omega$ is

$$\hat{p}(z) = \frac{\bar{y}(z)}{\int_{z'\in\Omega}\bar{y}(z')}, \text{ where}$$
$$\bar{y}(z) = \frac{1}{N}\sum_{i=1}^{N} K(x_i, z). \tag{31}$$

Note that $\bar{y}(z)$ in (31) is the sample average of N kernel functions. When multiple density estimates are desired over the same Ω, we can replace the kernel function sample average in (31) with the MTA estimate, which we term MT-KDE. Suppose one is given T sets of samples $\{x_{ti} \in \Omega\}_{i=1}^{N_t}$ for $t \in \{1,\ldots,T\}$ drawn IID from each task, and T kernel functions $\{K_t(x_{ti}, x_{tj})\}$, and a $T \times T$ task similarity matrix A. Then form the MTA estimates using (5):

$$\{y_t^*(z)\} = \underset{\{\hat{y}_t(z)\}_{t=1}^T}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{N_t} (K_t(x_{ti}, z) - \hat{y}_t(z))^2 + \gamma \sum_{r=1}^{T} \sum_{s=1}^{T} A_{rs} \left(\hat{y}_r(z) - \hat{y}_s(z)\right)^2 .$$

The normalized MT-KDE density estimate for the tth task is then $p_t^*(z) = y_t^*(z) / \int_\Omega y_t^*(z')$. Note that often the normalizer is not needed in practice. If needed, it may be non-trivial to compute if Ω is not a finite discrete set.

6.2 Density Estimation for Terrorism Risk Assessment

We compare KDE and MT-KDE on a problem of estimating the probability of terrorist events in Jerusalem using the Naval Research Laboratory's Adversarial Modeling and Exploitation Database (NRL AMX-DB). The NRL AMX-DB combined multiple open primary sources[1] to create a rich representation of the geospatial features of urban Jerusalem and the surrounding region, and accurately geocoded locations of terrorist attacks. Density estimation models are used to analyze the behavior of such violent agents, and to allocate security and medical resources. In related work, Brown et al. (2004) also used a Gaussian kernel density estimate to assess risk from past terrorism events.

We estimate a risk density for 40,000 geographical locations (samples) in a 20km × 20km area of interest in Jerusalem. Each geographical location is represented by a $D = 76$-dimensional feature vector. Each of the 76 features is the distance in kilometers to the nearest instance of some geographic location of interest, such as the nearest market or bus stop. Locations of past events are known for three different density estimation problems: 17 suicide bombings, 11 non-suicide bombings (e.g. car bombs), and 34 shootings. All the events are attributed to one of seven terrorist groups. For each of the three problems (suicide bombing, non-suicide bombings, and shootings), the density estimates for these seven groups are expected to be related, and we treat them as $T = 7$ tasks.

We compared KDE and MT-KDE for this problem as follows. The set Ω consists of the 40,000 grid points over which the density is estimated. The kernel function K was taken to be a Gaussian kernel on the two 76-dimensional feature vectors corresponding to any two grid points. Each Gaussian kernel was constrained to have diagonal covariance, where the variances were chosen automatically using the *normal reference density* technique (Venables and Ripley, 2002). For the task-similarity matrix A needed for MTA, we asked terrorism expert Mohammed M. Hafez of the Naval Postgraduate School to assess the similarity between the seven groups during the Second Intifada (the time period of the dataset); these similarities are shown in Table 2. The similarity for the "Unknown" group in each

[1] Primary sources included the NRL Israel Suicide Terrorism Database (ISD) cross referenced with numerous open sources (including the Israel Ministry of Foreign Affairs, BBC, CPOST, Daily Telegraph, Associated Press, Ha'aretz Daily, Jerusalem Post, Israel National News), as well as the University of New Haven (UNH) Institute for the Study of Violent Groups (ISVG), the University of Maryland (UMD) Global Terrorism Database (GTD), and the National Counter Terrorism Center (NCTC) Worldwide Incident Tracking System (WITS). The terrorist event data used took place in Jerusalem during the Second Intifada, between October 2000 and July 2003. Only events that are believed to have occurred at their planned location were included, that is, bombs that were detonated early or on accident were not included.

Table 2: Hafez's Similarity Matrix A

	AAMB	Hamas	PIJ	PFLP	Fatah	Force17	Unknown
AAMB	0	.2	.2	.6	.8	.8	.6
Hamas	.2	0	.8	.2	.2	.2	.4
PIJ	.2	.8	0	.2	.2	.2	.4
PFLP	.6	.2	.2	0	.6	.6	.5
Fatah	.8	.2	.2	.6	0	1	.6
Force17	.8	.2	.2	.6	1	0	.6
Unknown	.6	.4	.4	.5	.6	.6	0

column/row was set to be the average similarity of the known groups in that column/row. We were interested in investigating the usefulness of the provided expert similarity values, and thus, noting that the regularization parameter γ acts a as multiplier on the similarities, we set $\gamma = 1$.

After computing the KDE and MT-KDE density estimates using all but one of the training examples $\{x_{ti}\}$ for each task, we sort the resulting 40,000 estimated probabilities for each of the seven tasks, and extract the rank of the left-out known event. The median of the left-out rank is reported in Table 3. Ideally, the rank of the left-out known event would be very low, indicating that the location of the left-out event is at high-risk. The results show that the median ranks for MT-KDE are much lower than those for KDE for all three problems.

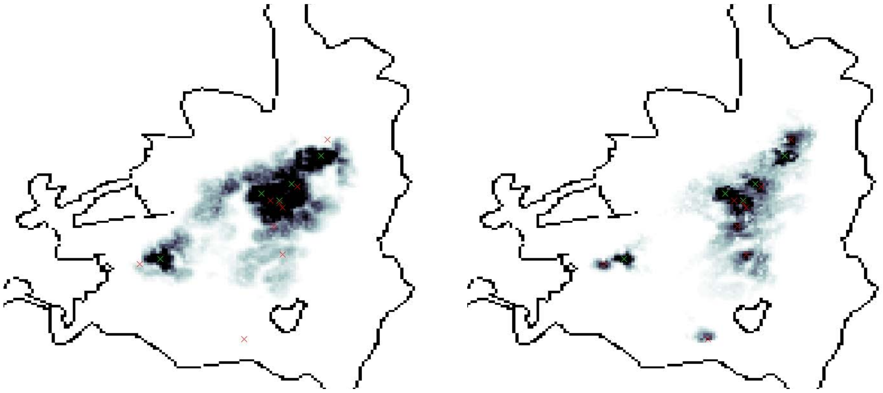
Fig. 6 shows a spatial map of the density estimates for KDE and MT-KDE for the AAMB group, where for this figure all the past events were used as training. The green x's mark AAMB events. Two differences are notable. First, the MT-KDE predicts higher risk of AAMB events where the red x's are. Second, MT-KDE predicts less risk for many locations, analysis of the risk estimates for the non-AAMB groups suggests that the lowered risk corresponds to locations estimated to be low-risk for the other groups.

Table 3: Median Rank of Left-Out Event Out of 40,000

Method	Suicide Bombings	Bombings	Shootings
KDE	308	6,512	11,160
MT-KDE	88	900	4,752

6.3 MT-KDE For Density Estimation Over Different Domains

We generalize MT-KDE to the problem of producing T density estimates where the tth density estimate is over a task-specific domain Ω_t, with U_t domain elements that we index by z_{tu} for $u = 1, \ldots, U_t$. Over the T tasks there are a total of $U = \sum_t U_t$ domain elements. As the T domains share a feature space, we use the task-and-domain-element $U \times U$ similarity matrix that we implicitly index as $\mathbb{A}_{rv;sw}$ to specify the similarity between the vth domain



KDE for AAMB

MT-KDE for AAMB

Figure 6: Spatial maps of the risk density estimated for suicide bombing events by AAMB, trained on all the past events. The green x's denote the six AAMB suicide bombings, the red x's denote the eleven suicide bombings by the other groups (specifically, by Hamas and PIJ).

element of the rth task and the wth domain element of the sth task. Then we estimate

$$\begin{aligned} \operatorname*{arg\,min}_{\{\{\hat{y}_t(z_{tu})\}_{u=1}^{U_t}\}_{t=1}^{T}} & \sum_{t=1}^{T} \sum_{u=1}^{U_t} \sum_{i=1}^{N_t} \left(K_t(x_{ti}, z_{tu}) - \hat{y}_t(z_{tu})\right)^2 \\ & + \gamma \sum_{r=1}^{T} \sum_{v=1}^{U_r} \sum_{s=1}^{T} \sum_{w=1}^{U_s} \mathbb{A}_{rv;sw} \left(\hat{y}_r(z_{rv}) - \hat{y}_s(z_{sw})\right)^2 . \end{aligned} \quad (32)$$

The resulting set of estimates $\{\{y_t^*(z_{tu})\}_{u=1}^{U_t}\}_{t=1}^{T}$ are then normalized within each task to form the T density estimates. That is, for the tth task, the MT-KDE density estimate is,

$$\hat{p}_t(z_{tu}) = \frac{y_t^*(z_{tu})}{\sum_{v=1}^{U_t} y_t^*(z_{tv})}.$$

6.4 Density Estimation for Mass Spectrometry Charge Classification

We apply MT-KDE for density estimation over different domains to the problem of guessing the charge for mass spectrometry samples. The dataset consists of $T = 25$ runs of the LTQ-FT Ultra Hybrid Mass Spectrometer using data-dependent acquisition, and we treat each run as a task. Each of the runs is the result of a single protein sample of one of six species: C. elegans, S. cerevisiae, mouse, Leishmania, bovine, or human. Each sample is a tandem mass spectrum, from which was extracted a set of $D = 18$ features that were shown to be useful for a related mass spectrometry filtering problem (Feldman, 2009). Each spectrum is the result of extensive bombardment and fragmentation of a peptide ion. There are 24752 spectra total, and within each task there are spectra of peptides of various charges from +1 to +5, which we consider classes. To compare MT-KDE with KDE, we split the samples randomly in half into a training set (samples x_{ti} in (32)) and a test set (samples z_{tu} in

(32)). To perform randomized 10-fold cross-validation, the training set was further split 75/25 into training and validation sets. For each of the five charge-classes, we use the MT-KDE objective to simultaneously estimate the probability density for all the test samples across the $T = 25$ tasks. Given the five estimated class-conditional densities for MT-KDE or KDE, each test sample is classified as the maximum likelihood charge-class.

Further experimental details are as follows. Each task's data is separately standard normalized to have zero mean and a standard deviation of one. The similarity function in this case is both task and sample dependent and can be factorized as $\mathbb{A}_{rv;sw} = A_{rs}A^F_{rv;sw}$, where $A_{rs} = 1$ if the species of run in task r and s are the same, and 0 otherwise, and $A^F_{rv;sw}$ measures the similarity between the feature vectors z_{rv} and z_{sw} with a radial basis function with fixed bandwidth $\sigma^2 = 10$. This σ value was chosen such that the mean similarity of the resulting training similarity matrix is approximately 0.5, so that the resulting similarities are well scaled and not skewed too far towards zero or one. We consider the following range of γ: $\{0,\ 0.001,\ 0.01,\ 0.1,\ 1,\ 10\}$ for the purposes of cross-validation; $\gamma = 1$ was chosen as the optimal parameter. Note that $\gamma = 0$ is the standard KDE case.

The resulting test classification errors are 41.1% for KDE and 33.3% for MT-KDE. The difference in these errors is statistically significant according to two one-sided Wilcoxon non-parametric tests; the tests were performed on the resulting zero-one vector of misclassifications (zero when the maximum likelihood class is correct, and one otherwise).

7. Multi-Task Regularized Similarity Discriminant Analysis

Similarity discriminant analysis (SDA) is a generative classifier that models the class-conditional probability distribution of the similarities between samples using discrete exponentials (Cazzanti et al., 2008). Like quadratic discriminant analysis, SDA may produce biased class models, and it has been shown that applying SDA locally (called *local SDA*) to a set of k neighboring (most similar) training samples for each test sample generally decreases model bias and improves performance (Cazzanti and Gupta, 2007). However, estimating the required discrete exponential parameters from only k neighbors can have problematically high variance, and in some cases the maximum likelihood parameter estimate is infinite. Here we show that multi-task averaging can be profitably applied to this parameter estimation task.

7.1 MT-SDA Formulation

In similarity-based classification, the training data is a $N \times N$ matrix S of pairwise similarities between N training samples, and their corresponding class labels $y_i \in \{1, 2, \ldots, G\}$ for $i = 1, \ldots, N$. The similarities are assumed to have a discrete (perhaps after quantization) domain Ω. The test data is a $N \times 1$ vector s of similarities between a test sample and the N training samples. SDA produces a probability that the test sample belongs to each of the G classes. In our experiments we use local SDA, but for notational simplicity we restrict our explanation to the SDA; the only difference is in practice the SDA model is trained anew for the k nearest-neighbors of each test sample.

The pairwise SDA model (Sadowski et al., 2010) requires estimating G^2 discrete exponential distributions of the probability of seeing a certain similarity value between two samples if one sample is from class g and the other sample is from class h. That is, the gth-hth distribution has the form:

$$P(s_{gh}) = \gamma_{gh} e^{\lambda_{gh} s_{gh}}, \tag{33}$$

where s_{gh} is a similarity between a sample from class g and a sample from class h.

We focus on the problem of estimating the parameter λ_{gh} in (33); given λ_{gh} the normalizer γ_{gh} will be implied. The maximum likelihood estimate $\hat{\lambda}_{gh}$ satisfies the empirical mean constraint (Sadowski et al., 2010)

$$v_{gh} = \sum_{s_{gh} \in \Omega} s_{gh} \gamma_{gh} e^{\lambda_{gh} s_{gh}}, \tag{34}$$

where $v_{gh} = \frac{\sum_{a:y_a=g} \sum_{b:y_b=h} S(a,b)}{k_g k_h}$ is the empirical mean with k_g, k_h are the number of training samples in class g and class h, respectively. To find the maximum likelihood estimate $\hat{\lambda}_{gh}$, one can compute the sample average on the left-hand side of (34), and then numerically solve for the best $\hat{\lambda}_{gh}, \hat{\gamma}_{gh}$ pair. A particular problem arises when $S(a,b)$ is at an extreme value of the similarity domain Ω for all a, b. For example, if the similarities can only be 0 or 1, and all training sample pairs from class g and class h happen to have zero similarity, then the $\hat{\lambda}_{gh}$ will be ∞.

We apply MTA to this problem and compute multi-task regularized sample averages to replace the left-hand side of (34), and then numerically solve (34) to produce the *multi-task SDA* (MT-SDA) estimates of the discrete exponential parameters. Here, the multiple tasks are the G^2 class-pairings.

In this application, the multi-task regularization operates across pairs of classes: the average similarity of samples from class g to samples of class h is regularized toward the average similarity of the samples from class l to class m. In our experiments, we define the task-similarity matrix A as a Gaussian kernel on the pairwise class sample averages, as follows. Let v_{gh} denote the gth-hth sample average given by the left-hand side of (34). The task similarity between the gth-hth task and the l-mth task is taken to be:

$$A_{g-h;l-m} = e^{-(v_{gh} - v_{lm})^2/\sigma}.$$

7.2 MT-SDA Experiments

We compare the classification performance of the local pairwise SDA classifier to the above MT-SDA variant. Other similarity-based classifiers have been proposed Chen et al. (2009), but we restrict our comparison to local SDA to focus on the value of adding the multi-task regularization.

We report classification results for six different benchmark similarity datasets (Chen et al., 2009). Table 4 shows the mean test error rates computed from 20 random train/test splits of the data, with 20% of the data held out for testing and 80% used for training. For each split, the training set was further split into 10 disjoint cross-validation partitions to select the multi-task parameter γ and the Gaussian kernel parameter σ among the possible values $\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$, and the neighborhood size k among the possible values $\{2, 4, 8, 16, 32, 64, \max(128, N)\}$). Across five datasets multi-task local SDA outperforms the standard local SDA and one dataset is a tie. The improved performance is statistically significant according to a Wilcoxon sign rank test ($p = 0.05$).

Table 4: Percent test error averaged over 20 random test/train splits for the benchmark similarity datasets.

	Amazon	Sonar	Patrol	Protein	Voting	FaceRec
	2 classes	2 classes	8 classes	4 classes	2 classes	139 classes
Local SDA	11.32	15.25	11.56	10.00	6.15	4.23
Multi-task Local SDA	8.95	14.50	11.56	9.77	5.52	3.44

8. Multi-Task Local Linear Regression

In this section, we discuss how OSR can be used for multi-task regularization of some parametric models. We focus on the local linear regression model, where $f(x, \beta_t) = x^T \beta_t + \beta_{t0}$ for some local neighborhood of $x \in \mathbb{R}^d$, with parameters $\beta_t \in \mathbb{R}^d$ and $\beta_{t0} \in \mathbb{R}$. For local linear regression, we combine the OSR regularizer with a standard Tikhonov regularization of each $\beta_t \in \mathbb{R}^d$ to find a matrix of regression coefficients β with tth column β_t that solve

$$\begin{aligned}\arg\min_{\beta} & \sum_{t=1}^{T} \sum_{i=1}^{N_t} (y_{ti} - \beta_t^T x_{ti} - \beta_{t0})^2 \\ & + \frac{\gamma}{2} \sum_{r,s=1}^{T} A_{rs} (\beta_r^T z_r + \beta_{r0} - \beta_s^T z_s - \beta_{s0})^2 + \lambda \sum_{t=1}^{T} (\beta_t - \beta_G)^T (\beta_t - \beta_G), \qquad (35)\end{aligned}$$

where z_r is the rth task or test sample, $\beta_G \in \mathbb{R}^d$ is a fixed vector of linear regression weights that the local regression coefficients are regularized towards, and γ and λ are regularization parameters. We set $\beta_{t0} = \mathbf{mean}(\{y_t\})$, which is the optimal offset for Tikhonov regularization. For our experiments, we set $\beta_G \in \mathbb{R}^d$ to be the global least-squares linear regression coefficients found by fitting all the training data with one hyperplane (rather than locally fitting some of the training data). Tikhonov regularization often increases accuracy, but it does not explicitly enforce smoothness of the outputs across the local neighborhoods. The OSR regularization may add further error reduction, and should create a smoother function over the domain because it explicitly forces similar (e.g. "nearby") test samples to have close output values.

The OSR-LLR estimate defined by (35) has a closed-form, given in Appendix C.

8.1 OSR-LLR Color Management Experiment

We demonstrate the OSR-LLR on the problem of color management of a printer, which consists of estimating what RGB color should be input to a printer to best reproduce a desired CIELab color (Bala, 2003). Because such estimations must be fast, color management is internationally standardized by the International Color Consortium to use a look-up table (LUT) or lattice of points in CIELab space that can be efficiently interpolated for any particular test CIELab color. The color management estimation problem is to best estimate the RGB values corresponding to the CIELab LUT. Local Tikhonov regression, as defined by the first and third terms of (35), has been shown to be a state-of-the-art method for this estimation problem (Hrustemovic and Gupta, 2008), where the β_G are the regression coefficients for the least-squares global hyperplane fit. For those experiments and these, "local" is defined by the enclosing k-NN (Gupta et al., 2008) such that X_t is the matrix of the N_t enclosing k-NN training points for test point z_t.

In this work, we treat the estimation of the RGB value for each gridpoint in the lattice as a task, and simultaneously regress many of them at a time. For problems such as color management, minimizing the MT trace-norm, for example, is not expected to be helpful since the problem is already low-dimensional and the colorspace dimensions are all meaningful. Most importantly however, focusing on tying together the model parameters does not explicitly enforce smoothness between the lattice points, which the "manifold regularization" term in OSR-LLR does explicitly. For this application, smooth rendering of color gradients in the reproduced images is an important dual goal alongside estimation accuracy (Morovič et al., 2008).

In our experiments, the training samples were (CIElab,RGB) pairs measured from the 918-sample Gretag Macbeth TC9.18 RGB target for a laser printer (Samsung CLP300) and an inkjet printer (HP D7260). In each case, the LUT to be estimated is a regular $17 \times 17 \times 17$ lattice in CIElab space. For each of the 17^3 lattice points, the problem is to estimate an RGB triplet of outputs. As standard, we treat the R, G, and B problems independently, and we use a standard ICC profile architecture of 1D LUTs for linearization before the estimated 3D LUT (Bala, 2003).

For OSR-LLR, we regularized all the in-gamut lattice points towards each other, where the in-gamut points were estimated to be all lattice points within three ΔE_{76} (that is, Euclidean distance in CIELab space) of the convex hull of the training samples. For the Samsung and HP printers there were $T = 611$ and $T = 756$ estimated-in-gamut lattice points. For lattice points that fell outside the estimated gamut, we set $\gamma = 0$ in (35) which is equivalent to only using Tikhonov regression for those points. For the 3D LUTs, to ensure that distant samples do not overly influence one another, we used a simple adjacency test such that $A_{rs} = 1$ whenever two lattice points are directly next to each other (each lattice point has 6 such neighbors), and 0 otherwise, resulting in a very sparse task-similarity matrix A. The spacing of the 1D LUTs is much finer, so we use the radial basis function kernel $A_{rs} = e^{\frac{||z_r - z_s||_2^2}{2}}$.

For Tikhonov regression, we cross-validated the regularization parameter λ from a range of $\{0, 1, 5, 10, 20\}$ and chose it by minimizing the difference between the training points' RGB values and their estimated values for each choice of λ. For the Samsung and HP printers, $\lambda = 10$ and $\lambda = 5$ were chosen for Tikhonov for each printer. For OSR, we used the λ chosen for Tikhonov to ensure a direct comparison. To choose γ, we ran the estimated LUTs on a benchmark test image, and chose the largest γ from the set $\{0, 0.1, 1, 10, 100\}$ that preserved the black point (too large a γ will turn black to gray because the output colors are smoothed towards each other). For the Samsung and HP printers, these values were $\gamma = 10$ and $\gamma = 1$, respectively.

For all estimation methods, each dimension of X_t was normalized to have mean 0 and standard deviation 1, and the associated z_t was normalized accordingly. For OSR, the tth column block of B_G was normalized by the standard deviation of X_t.

Because perceptions of smoothness are strongly dictated by uneven jumps in luminance (Wyszecki and Stiles, 2000), the smoothness is measured by the square root of the average

ℓ_2 norm of the estimated LUT's RGB differences along the L^* dimension for adjacent nodes, compared to the smoothness baseline of even RGB differences (Morovič et al., 2008). These roughness results are reported in Table 5. The results show that the OSR+Tikhonov LUTs provided greater than just-noticeable decreased roughness (increased smoothness) as expected, and the statistically indistinguishable or better accuracy than only using Tikhonov regularization. Specifically, OSR is statistically significantly more accurate than Tikhonov for the HP printer according to a Wilcoxon one-sided significance test (with a p-value of .05), and there is no statistically significant difference between OSR and Tikhonov for the Samsung printer.

Table 5: Mean roughness and ΔE_{94} error for HP D7260 and Samsung CLP300 printers.

	HP D7260		Samsung CLP300	
	Roughness	Mean Error	Roughness	Mean Error
Tikhonov Regularization	18	2.12	18.5	4.02
OSR + Tikhonov Regularization	16.2	2.09	16.3	4.05

9. Conclusions and Open Questions

The presented OSR formulation is powerful and can be applied to a broad class of estimation problems. We illustrated these attributes with three applications of OSR. First we presented a multi-task kernel density estimation. Then, the proposed multi-task similarity discriminant analysis showed the use of OSR for a complicated parameter estimation. Lastly, we showed how parametric models like local linear regression can benefit from the smoothing action of OSR.

Because it is the outputs that are regularized together rather than the parameters, OSR can be used with *different models* for each of the different tasks. For example, consider the two-task problem of detecting anomalies from an audio stream and a synchronized video stream. The input feature spaces for the two tasks are different. It would be nonsensical to impose any constraint tying the task-specific function parameters together numerically or insist on a shared low-dimensional feature space. Thus, this two-task problem is difficult to even formulate in the MTL paradigm that dominates today's literature, but naturally fits the OSR paradigm. Despite the input spaces not being identical, one can nevertheless specify the task-sample similarity matrix $\mathbb{A}$ by preferentially tying together cross-task samples that are closer in time.

A key question that we have only partially answered is, "How should one specify a task-sample similarity in general? Our $T = 2$ MTA analysis showed that the task similarity matrix should capture the relative in-task variance to between-task difference. Our experiments successfully relied on side information and kernel computations to quantify task-and-sample relatedness. The anomaly detection example above uses the known variable of time as side information for sample relatedness. Another approach is to let the task-sample similarity matrix be an additional variable to optimize over. A third approach is to estimate optimal task-sample similarities. In the context of MTA, we touched on the question of how to estimate the optimal task similarity matrix from data, analogous to an empirical Bayesian method, but much theoretical and experimental research remains in that direction.

We found that a negative two-task similarity may be optimal in terms of minimizing mean-squared error. The interpretation of a negative similarity remains an open question. We did not explore the potential of a negative similarity in any of our applications. Further, some of the other theoretical results do not hold for negative similarity, such as the convexity result.

We focused on squared error for the loss and regularization functions, because of its usefulness and tractability. Alternative loss functions remain to be explored. Multi-task medians, for instance, resulting from an L_1 loss may be very useful in practice. Multi-task learning of vectors and matrices may also benefit from the proposed regularization. We established some theoretical foundations for MTA, perhaps the simplest OSR formulation. Our main theoretical result is that the proposed MTA estimates are convex combinations of sample averages. Our proof directly analyzes the Laplace expansion of the matrix inverse. An open question is whether there is a simpler proof of this result.

Acknowledgments

We thank Amol Kapila for insightful discussions on OSR from a Bayesian perspective. We thank Carol Chang, Brian Sandberg, and Ruth Willis of the Naval Research Lab for the terrorist event dataset and helpful discussions. We thank Mohammed M. Hafez of the Naval Postgraduate School for the matrix of similarities of the terrorist groups. We thank Michael J. MacCoss and Barbara Frewen at the University of Washington for the tandem mass spectrometry dataset. This work was funded by a United States PECASE Award and by the United States Office of Naval Research.

Appendix A: Proof of Lemma

Let $B = I - \tilde{A}$. The (t, s)th entry of B is

$$B_{ts} = \begin{cases} 1 & \text{if } t = s \\ -\frac{\gamma(A_{ts}+A_{st})}{N_t+\pi_t} & \text{if } t \neq s, \end{cases}$$
$$\text{where } \pi_t = \gamma \sum_{s \neq t} (A_{ts} + A_{st}).$$

Next, we use the Gershgorin Circle Theorem (Horn and Johnson, 1990). Let the sum of the off-diagonal entries for the tth row be

$$R_t = \sum_{s \neq t} |B_{ts}| = \frac{\pi_t}{N_t + \pi_t}. \tag{36}$$

The Gershgorin disk $D(B_{tt}, R_t)$ is the closed disk in $\mathbb{C}$ with center B_{tt} and radius R_t. From (36), $R_t < 1$ if all the entries of A are non-negative. Then since $B_{tt} = 1$ for all t, every Gershgorin disk is contained within the positive half-plane of $\mathbb{C}$. Thus by the Gershgorin Circle Theorem, the real part of every eigenvalue of matrix B is positive, and therefore the determinant of B is positive. This completes the proof.

For the $T = 2$ case, we have that $\tilde{A}_{12} = \tilde{A}_{21} = \frac{2a\gamma}{N+2a\gamma}$. The matrix $I - \tilde{A}$ is not invertible if and only if its determinant is zero:

$$\begin{aligned}
\mathbf{Det}(I - \tilde{A}) = 1 - \left(\frac{-2a\gamma}{N_1 + 2a\gamma}\right)\left(\frac{-2a\gamma}{N_2 + 2a\gamma}\right) &= 0 \\
N_1N_2 + 2(N_1 + N_2)a\gamma + 4a^2\gamma^2 &= 4a^2\gamma^2 \\
a &= \frac{-N_1N_2}{2(N_1 + N_2)\gamma},
\end{aligned}$$

as was to be shown.

Appendix B: Proof of Theorem

By the theorem's assumptions, the entries of A are non-negative, and thus by the lemma in Section 4.2, $(I - \tilde{A})^{-1}$ exists. Recall that a matrix inverse can be written as the product of the determinant and the transpose of the matrix of the cofactors. Because there is no set ordering of the tasks, we can simultaneously swap the rth and sth row and column of $I - \tilde{A}$; this means that, without loss of generality, it is enough to prove the theorem for the first task estimate:

$$((I - \tilde{A})^{-1}\tilde{y})_1 = \frac{1}{\det(I - \tilde{A})} \sum_{r=1}^{T} C_{r1} \frac{N_r}{N_r + \pi_r} \bar{y}_r, \tag{37}$$

where C_{rs} is the (r, s)th cofactor, and by the theorem's assumptions $N_r \geq 1$ for any $r = 1, \ldots, T$. To show that (37) is a convex combination of the T sample averages, we must show that the coefficients on the $\bar{y}_r$ are non-negative and sum to 1.

That is, we must show that

$$\frac{1}{\det(I - \tilde{A})} \sum_{r=1}^{T} C_{r1} \frac{N_r}{N_r + \pi_r} = 1,$$

or equivalently, that

$$\sum_{r=1}^{T} C_{r1} \frac{N_r}{N_r + \pi_r} - \det(I - \tilde{A}) = 0. \tag{38}$$

For ease of notation, let $S_{st} = \gamma(A_{st} + A_{ts})$ and $S_{tt} = 0$. Since

$$\det(I - \tilde{A}) = C_{11} - \sum_{r=2}^{T} \frac{S_{1r}}{N_r + \pi_r} C_{r1},$$

the condition (38) can be re-stated

$$\frac{\pi_1 C_{11}}{N_1 + \pi_1} - \sum_{r=2}^{T} \frac{(S_{1r} + N_r)C_{r1}}{N_r + \pi_r} = 0. \tag{39}$$

Factor out $1/(N_1 + \pi_1)$ from (39) to produce the equivalent condition

$$\pi_1 C_{11} - \sum_{r=2}^{T} \frac{(S_{1r} + N_r)(N_1 + \pi_1) C_{r1}}{N_r + \pi_r} = 0. \tag{40}$$

As a first step to showing (40), let's look at C_{11}. It is the determinant of the $(T-1) \times (T-1)$ minor

$$M_{11} = \begin{bmatrix} 1 & -\frac{S_{23}}{N_2+\pi_2} & -\frac{S_{24}}{N_2+\pi_2} & \cdots & -\frac{S_{2T}}{N_2+\pi_2} \\ -\frac{S_{23}}{N_3+\pi_3} & 1 & -\frac{S_{34}}{N_3+\pi_3} & \cdots & -\frac{S_{3T}}{N_3+\pi_3} \\ & & & \ddots & \\ -\frac{S_{2T}}{N_T+\pi_T} & -\frac{S_{3T}}{N_T+\pi_T} & \cdots & -\frac{S_{(T-1)T}}{N_T+\pi_T} & 1 \end{bmatrix}$$

and therefore it is

$$C_{11} = \prod_{s=2}^{T} \frac{1}{N_s + \pi_s} \begin{vmatrix} N_2 + \pi_2 & -S_{23} & -S_{24} & \cdots & -S_{2T} \\ -S_{23} & N_3 + \pi_3 & -S_{34} & \cdots & -S_{3T} \\ & & & \ddots & \\ -S_{2T} & -S_{3T} & \cdots & -S_{(T-1)T} & N_T + \pi_T \end{vmatrix}. \tag{41}$$

We will call the matrix whose determinant featured in Eq. (41) by D and the corresponding determinant by C'_{11}. That is,

$$D = \begin{bmatrix} N_2 + \pi_2 & -S_{23} & -S_{24} & \cdots & -S_{2T} \\ -S_{23} & N_3 + \pi_3 & -S_{34} & \cdots & -S_{3T} \\ & & & \ddots & \\ -S_{2T} & -S_{3T} & \cdots & -S_{(T-1)T} & N_T + \pi_T \end{bmatrix}$$

and $C'_{11} = \det(D)$.

Next let's look at the cofactors C_{r1}. They are the product of $(-1)^{r+1}$ and the determinant of the $(T-1) \times (T-1)$ minor M_{r1} which we get by removing the rth row and the 1st column of matrix $I - \tilde{A}$. If we look at the determinant of M_{r1} we see that we can factor out $\frac{1}{N_k + \pi_k}$ from the rows $k < r$ and $\frac{1}{N_{k+1} + \pi_{k+1}}$ from the rows $k \geq r$. As a result we get that

$$\begin{aligned} &\frac{N_1 + \pi_1}{N_r + \pi_r} C_{r1} \\ &= \prod_{s=2}^{T} \frac{1}{N_s + \pi_s} \begin{vmatrix} N_2 + \pi_2 & -S_{23} & -S_{24} & \cdots & \cdots & -S_{2T} \\ & & & \vdots & & \\ S_{12} & S_{13} & \cdots & \cdots & \cdots & S_{1T} \\ & & & \vdots & & \\ -S_{2(T-1)} & \cdots & \cdots & -S_{(T-2)(T-1)} & N_{T-1} + \pi_{T-1} & -S_{T(T-1)} \\ -S_{2T} & \cdots & \cdots & -S_{(T-2)T} & -S_{(T-1)T} & N_T + \pi_T \end{vmatrix}, \end{aligned} \tag{42}$$

where the row $[S_{12}\ S_{13}\ \cdots\ S_{1T}]$ is in the $(r-1)$th row of the determinant. We will call the determinant in Eq. (42) C'_{r1}. Notice, that we can construct C'_{r1} from C'_{11} by replacing the

$(r-1)$th row of C'_{11} with the vector $[S_{12}\ S_{13}\ \cdots\ S_{1T}]$. Using the expressions for C'_{11} and C'_{r1}, we can derive a simpler constraint that is equivalent to Eq. (40):

$$\pi_1 C'_{11} - \sum_{r=2}^{T} (S_{1r} + N_r) C'_{r1} = 0,$$

or equivalently

$$\sum_{s=2}^{T} S_{1s} C'_{11} - \sum_{r=2}^{T} (S_{1r} + N_r) C'_{r1} = 0. \tag{43}$$

For each s value we will express the matrix that produces the determinant C'_{11} slightly differently. For any s let C'_{11} be written as the determinant of a matrix where we replace the $(s-1)$th row by the sum of all the rows (which does not change the determinant), that is, let C'_{11} be written

$$\begin{vmatrix} N_2+\pi_2 & -S_{23} & -S_{24} & \cdots & -S_{2T} \\ & & \vdots & & \\ N_2+S_{12} & N_3+S_{13} & N_3+S_{13} & \cdots & N_T+S_{1T} \\ & & \vdots & & \\ -S_{2T} & -S_{3T} & \cdots & -S_{(T-1)T} & N_2+\pi_2 \end{vmatrix}.$$

As the last step we show below that for each r and each s in $\{2,\ldots,T\}$ the $(s-1)$th term of $(S_{1r}+N_r)C'_{r1}$ cancels the $(r-1)$th term of $S_{1s}C'_{11}$. This gives a one-to-one correspondence between the terms of the sums in Eq. (43) and as a result the difference in Eq. (43) is 0.

To show that, let $r, s \in \{2,\ldots,T\}$ and consider the Laplace expansion of C'_{r1} along the $(r-1)$th row. The $s-1$th term in that expansion is $(-1)^{r+s-2}(S_{1r}+N_r)S_{1s}\det\left(M'_{(r-1)(s-1)}\right)$ where $M'_{(r-1)(s-1)}$ is the minor of D we get by erasing the $(r-1)$th row and the $(s-1)$th column of D. Consider also the Laplace expansion of $S_{1s}C'_{11}$ along the $(s-1)$th row and consider the $(r-1)$th term of this expansion: $(-1)^{r+s-2}S_{1s}(N_r+S_{1r})\det\left(M'_{(s-1)(r-1)}\right)$. Since D is symmetric $M'_{(s-1)(r-1)} = \left(M'_{(r-1)(s-1)}\right)^T$ and since the determinant of a matrix and the determinant of its transpose are the same we have that

$$(-1)^{r+s-2}(S_{1r}+N_r)S_{1s}\det\left(M'_{(r-1)(s-1)}\right) = (-1)^{r+s-2}S_{1s}(N_r+S_{1r})\det\left(M'_{(s-1)(r-1)}\right).$$

At this point, we have shown that the coefficients involved in Eq. (37) sum to 1. Until now, we have only needed the fact that $S_{st} = S_{ts}$ and that $I - \tilde{A}$ is invertible. The theorem's assumption that A has non-negative entries is sufficient but not necessary for the invertibility of $I - \tilde{A}$.

In order to complete the proof, we must also show that the coefficients in (37) are non-negative, and for that we explicitly need the theorem's assumption that the entries of A are non-negative.

From the proof of the lemma, we already know that the determinant in (37) is positive, and by the theorem's assumptions $N_r/(N_r+\pi_r) > 0$. Thus, we only need to show that all the C_{r1} are non-negative for $r \in \{2, \dots, T\}$. If we look at the Eq. (42) we see that it is enough to show that C'_{r1} is non-negative in order to show that C_{r1} is non-negative. To show that C'_{r1} is non-negative we rewrite it as follows. Replace the $(r-1)$th column with the sum of all the columns:

$$\begin{vmatrix} N_2+\pi_2 & -S_{23} & \cdots & N_2+S_{12} & \cdots & -S_{2T} \\ & & \vdots & & & \\ S_{12} & S_{13} & \cdots & \pi_1 & \cdots & S_{1T} \\ & & \vdots & & & \\ -S_{2(T-1)} & S_{2(T-1)} & \cdots & N_{T-1}+S_{1(T-1)} & \cdots & -S_{T(T-1)} \\ -S_{2T} & S_{2T} & \cdots & N_T+S_{1T} & \cdots & N_T+\pi_T \end{vmatrix}. \tag{44}$$

This replacement does not change the determinant. Let's call the matrix featured in Eq. (44) by F. If we apply the Gershgorin Circle Theorem to matrix F, and again denote the sum of the absolute values of the off-diagonal elements in row t by R_t, we see that $F_{tt} - R_t = S_{tr} \geq 0$ for all $t \neq r-1$ and $F_{(r-1)(r-1)} - R_{r-1} = S_{1r} \geq 0$. Thus the distance from the center of each of the Gershgorin discs to the origin is at least as large as the radius, and thus by the Gershgorin Circle Theorem, the real part of each eigenvalue of F must be non-negative, which implies that the determinant of F (which is C'_{r1}) is non-negative.

Appendix C: OSR-LLR Solution

We give a closed-form solution to (35). Additional notation is needed. We define the following:

$X_t \in \mathbb{R}^{D \times N_t}$ the matrix with columns x_{ti}, for $i = \{1, \dots, N_t\}$

$y_t \in \mathbb{R}^{1 \times N_t}$ the row vector with entries y_{ti}

$X \in \mathbb{R}^{DT \times N}$ a block diagonal matrix with blocks X_t, for $t = \{1, \dots, T\}$, with $N = \sum_{t=1}^{T} N_t$

$Z \in \mathbb{R}^{DT \times T}$ a block diagonal matrix with single column blocks z_t

$Y \in \mathbb{R}^{T \times N}$ a block diagonal matrix with single row blocks y_t, for $t = \{1, \dots, T\}$

$B \in \mathbb{R}^{DT \times T}$ a block diagonal matrix with single column blocks β_t

$L = A^d - A$ the graph Laplacian matrix (Chung, 2004) of A, where A^d is a diagonal matrix with entries $A^d_{rr} = \sum_{t=1}^{T} A_{rt}$

$B_0 \in \mathbb{R}^{T \times N}$ a block diagonal matrix with single row blocks containing β_{t0} repeated N_t times

$\beta_0 \in \mathbb{R}^{T \times T}$ a diagonal matrix, with (t,t)th entry β_{t0}

$B_G \in \mathbb{R}^{dT \times T}$ a block diagonal matrix with single column blocks β_G

$\mathbf{1}$ a column vector of ones.

The OSR-LLR objective given by (35) can be re-written in this matrix notation:

$$\arg\min_B \quad \mathbf{1}^T(Y - (B^T X + B_0))(Y - (B^T X + B_0))^T \mathbf{1} + \gamma \mathbf{1}^T (B^T Z + \beta_0) L (B^T Z + \beta_0)^T \mathbf{1} \\ + \lambda \mathbf{1}^T (B - B_G)^T (B - B_G) \mathbf{1}.$$

The OSR-LLR problem is convex, and thus solved by taking the partial derivative of the objective with respect to B and setting it equal to zero. To do so, we use the following two identities (Petersen and Pedersen, 2008):

$$\frac{\partial a^T B^T D B a}{\partial B} = (D^T + D)Baa^T \text{ and } \frac{\partial a^T B^T b}{\partial B} = ba^T.$$

Thus the minimizer matrix B^* must solve:

$$\begin{aligned} 0 &= -2XY^T\mathbf{1}\mathbf{1}^T + 2XX^TB^*\mathbf{1}\mathbf{1}^T + 2XB_0^T\mathbf{1}\mathbf{1}^T + 2\gamma ZLZ^TB^*\mathbf{1}\mathbf{1}^T + 2\gamma ZL\beta_0^T\mathbf{1}\mathbf{1}^T \\ &+ 2\lambda B^*\mathbf{1}\mathbf{1}^T - 2\lambda B_G\mathbf{1}\mathbf{1}^T. \end{aligned}$$

Because of the special structure of the terms, the above can be simplified without destroying information by right-multiplying both sides of the equation by $\frac{1}{T}\mathbf{1}$, getting rid of the rightmost $\mathbf{1}^T$. Note that the remaining $B^*\mathbf{1}$ is actually a $dT \times 1$ column vector of all the β_t, stacked. Let $\hat{B}^* = B^*\mathbf{1}$, then,

$$\hat{B}^* = (XX^T + \gamma ZLZ^T + \lambda I)^{-1}(XY^T - XB_0^T - \gamma ZL\beta_0^T + \lambda B_G)\mathbf{1}, \tag{45}$$

where the inverse exists as long as $\lambda > 0$ because λI is positive definite, while $XX^T + \gamma ZLZ^T$ is PSD for symmetric A.

References

A. Agarwal, H. Daumé III, and S. Gerber. Learning multiple tasks using manifold regularization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54. 2010.

A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *Journal Machine Learning Research*, 4:83–99, December 2003.

R. Bala. *Device Characterization, Ch. 5 of Digital Color Handbook.* CRC Press, 2003.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal Machine Learning Research*, 7:2399–2434, 2006.

E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2008.

D. Brown, J. Dalton, and H. Hoyle. Spatial forecast methods for terrorist events in urban environments. *Lecture Notes in Computer Science*, 3073:426–435, 2004.

G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, pages 83–87, 1985.

L. Cazzanti and M. R. Gupta. Local similarity discriminant analysis. In *Proc. Intl. Conf. Machine Learning*, 2007.

L. Cazzanti, M. R. Gupta, and A. J. Koppal. Generative models for similarity-based classification. *Pattern Recognition*, 41(7):2289–2297, July 2008.

Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal Machine Learning Research*, 10:747–776, March 2009.

F. R. K. Chung. *Spectral Graph Theory.* 2004.

H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *Journal Artificial Intelligence Research*, 26:101–126, 2006.

B. Efron and C. N. Morris. Stein's paradox in statistics. *Scientific American*, 236(5): 119–127, 1977.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD '04*, pages 109–117, 2004.

T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal Machine Learning Research*, (6), April 2005.

S. Feldman. Filtering tandem mass spectra for quality. *University of Washington Master's Thesis*, 2009.

M. R. Gupta, E. K. Garcia, and E. M. Chin. Adaptive local linear regression with application to printer color management. *IEEE Trans. Image Processing*, 17(6):936–945, 2008.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990. Corrected reprint of the 1985 original.

N. Hrustemovic and M. R. Gupta. Multiresolutional regularization of local linear regression over adaptive neighborhoods for color management. *Proc. IEEE Intl. Conf. Image Processing*, 2008.

T. Kato, H. Kashima, M. Sugiyama, and K. Asai. Multi-task learning via conic programming. In *Advances in Neural Information Processing Systems (NIPS)*, pages 737–744. 2008.

Q. Liu, X. Liao, H. Li, J. R. Stack, and L. Carin. Semisupervised multitask learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, (6), June 2009.

C. A. Micchelli and M. Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

J. Morovič, A. Albarran, J. Arnabat, Y. Richard, and M. Maria. Accuracy-preserving smoothing of color transformation LUTs. *Proc. Color Imaging Conf.*, 2008.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 231–252, 2010.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. Version 20081110 (online).

P. Sadowski, L. Cazzanti, and M. R. Gupta. Bayesian and pairwise local similarity discriminant analysis. In *Proc. IEEE Conf. Cognitive Information Processing*, 2010.

D. Sheldon. Graphical multi-task learning, 2008. Advances in Neural Information Processing Systems (NIPS) Workshops.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.

W. Venables and B. Ripley. *Modern Applied Statistics.* Springer Verlag, New York, 2002.

G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae.* John Wiley & Sons, second edition, 2000.

Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal Machine Learning Research*, 8:35–63, 2007.

Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships. In *Proc. of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.